VARIAN ASSOCIATES
611 Hansen Way, Palo Alto, California

Engineering Report
No. 151-18

Copy No. 6

MINE DETECTION STUDY PROGRAM
FINAL REPORT

(15 January 1954 - 31 January 1957)

Prepared for: Engineer Research and
Development Laboratories
United States Army
Fort Belvoir, Virginia
Attention: Mr. Wesley F. Pape
Contracting Officer

On: Contract No. DA-44-009 eng-2004
Project No. 8-07-04-009

By: E. T. Jaynes
H. A. Osborn

Approved: Robert L. Jepsen
Robert L. Jepsen
Director of Research

LIST OF ILLUSTRATIONS

Figure No.		Page No.
2-1	Family of Curves $F(k) = 1 - \|k\|^n$, $\|k\| > 2$, With Corresponding Improvement Factors I(n)	12
2-2	Family of Curves $F(k) = \|1-k\|^n$, $\|k\| < 1$, With Corresponding Improvement Factors I(n)	13
3-1	Conditional and Average Losses as Functions of the Detection Threshold V_o. The $L(S_1)$ Curve Is Symmetric About the Point $\{S_g^{1/2} L_r\}$.	31
3-2	Probability of Detection, Neyman-Pearson Case	33
6-1	Antenna Pattern Through 8-inch Diameter Aperture in Ground Plane	63
6-2	Antenna Pattern From 24-inch Diameter Parabolic Reflector	64

I. INTRODUCTION

The purpose of this contract was to investigate theoretically a number of aspects and possibilities of non-metallic land mine detection. The work was divided into three main tasks:

1. Extensions of the statistical theory of anomaly suppression as started in Fort Belvoir Contract No. DA-44-009 eng-795 at Stanford University (this task has been completed and a final report has been submitted).

2. Investigation of possible microwave mine detectors (this task has been completed and a final report has been submitted).

3. Consideration of the problem of mine detection from the standpoint of information theory and data processing (the present report is the final one on task 3).

In 1951, when work on the mine detection problem started at Stanford University, it was generally believed that mine detection was basically an electromagnetic problem. The work accordingly was directed first toward calculation and measurement of field strengths and amplitudes of target signals. These problems are so difficult that anyone starting on them is protected for a long time from realizing the following fact: Even if we had at hand complete quantitative answers to all electromagnetic questions, this in itself would not tell us how to design a good mine detector. Furthermore, since even the simplest configurations that could be imagined for antennas still lead to hopeless boundary-value problems, one is led to a somewhat different conception of the place of electromagnetic theory in the mine detector program. By means of electromagnetic theory one can learn:

1. The physical mechanism by which a detector finds mines.
2. The significant things that should be measured (<u>not</u> calculated!) when testing a system.

Beyond these factors (which may be regarded as answered for most systems presently contemplated by the δY_{mn} theorem), electromagnetic theory can help us in a qualitative sense, as a guide to intuition. The potentially most useful application of electromagnetic theory, although it has not been productive in the past, is in connection with the synthesis problem, the formulation of which is one of the main objects of this report.

Consider now the two very important questions, approximately inverse to each other:

1. What is the proper way of interpreting experimental data so as to reduce it to some kind of statement about expected behavior in the field?

2. In the electromagnetic problem of antenna design, what is the desired characteristic toward which we should work?

These questions have nothing to do with electromagnetic theory, but until they are answered, no amount of electromagnetic theory can raise the level of development work above that of random trial-and-error. The reason for this is that the mine detector situation is at best marginal, so that the statistical aspect assumes an importance greater than in any other detection problem. For example, a study of statistical factors can lead to appreciable improvements in the reliability of radar; nevertheless, radar was practical and useful even before its statistical aspect was clearly recognized. In non-metallic land mine detection, the development of any practical system at all will require full exploitation of statistical knowledge. There is at present no guarantee that even this will actually result in a practical mine detector.

Suppose now that, with the benefit of all the hindsight accumulated thus far, we go back to the beginning and reconsider the whole mine detector problem. The first thing we notice is that in order to detect an object, the detector must be able to place an appreciable amount of energy in the region occupied by the object. Energy placed in other regions, while not necessarily harmful to detection, cannot contribute anything to the desired signal. Therefore, we build a system in such a way that as much energy as possible is placed in the ground at the expected depth of a mine and try it out. We now discover that although we get satisfactory signals from mines, we obtain equally large ones from random inhomogeneities in the soil, which prevent us from recognizing the mine signals. It is at this point that the problem becomes basically one of statistics rather than electromagnetic theory. Until enough has been learned about the statistics of the problem to give at least provisional answers to the two questions above, we cannot even state what the electromagnetic problem is, and there is no sense of direction to guide further development work. In what follows, we give such provisional answers, and state a definite electromagnetic synthesis problem whose solution would contribute much to the mine-detection art. No claim is made about this criterion being the ultimate one, since much remains to be done in the statistical part of the problem; however, it appears unlikely that it can be appreciably changed in the future unless experiments should demonstrate that soil statistics are far from gaussian.

II. SUMMARY OF PREVIOUS THEORY OF ANOMALY SUPPRESSION

A. Formal Relations

The following is a condensed summary of the most important equations arising in the theory of anomaly suppression as developed under Contract No. DA-44-009 eng-795 at Stanford, and somewhat extended in task 1 of this contract. Although the purpose of task 3 is to supplant the earlier theory by a more fundamental one, the original theory retains an important place. In many cases the new theory reduces, as far as final results and design criteria are concerned, to the previous relations with a new interpretation. Thus most of the work done previously remains applicable.

We operate with several functions:

$f(x)$ = antenna function = $(E_1 \cdot E_2)$.

$g(x)$ = ground function = $\epsilon_{ground} - \langle \epsilon_{ground} \rangle$

$h(x)$ = mine function = $\begin{cases} \epsilon_{mine} - \langle \epsilon_{ground} \rangle & \text{inside a mine} \\ 0 & \text{elsewhere} \end{cases}$

$\gamma(x,x')$ = ground covariance function = $\langle g(x)\, g^*(x') \rangle$.

In terms of these functions, the ratio of $|\text{peak signal}|^2$ to mean square anomaly signal is

$$\eta = \frac{\left| \int h(x)\, f(x)\, dx \right|^2}{\iint f(x)\, \gamma(x,x')\, f^*(x')\, dx\, dx'} \tag{2-1}$$

and the antenna design problem is that of choosing $f(x)$ so as to make this as large as possible. Writing

$$p(x) = \int \gamma(x,x')\, f^*(x')\, dx'$$

we have, from the Schwarz inequality:

$$\left| \int h(x)\ f(x)\ dx \right|^2 = \left| \int h(x) \sqrt{\frac{f(x)}{p(x)}} \cdot \sqrt{f(x)\ p(x)}\ dx \right|^2$$

$$\leq \int |h(x)|^2 \left| \frac{f(x)}{p(x)} \right| dx \cdot \int f(x)\ p(x)\ dx$$

or

$$\eta \leq \int |h(x)|^2 \left| \frac{f(x)}{p(x)} \right| dx$$

with equality if and only if $h(x) = K\ p(x)$, where K is some constant. Thus the maximum possible value of η is attained if $f(x)$ satisfies the integral equation

$$h^*(x) = \int f(x')\ \gamma(x',x)\ dx' \tag{2-2}$$

the solution of which was called the "unconditional optimum" antenna function in the previous work.

These relations may also be stated in terms of the fourier transforms of the original functions. Let

$$\left.\begin{aligned} h(x) &= \int H(k)\ e^{ikx} dk \\ f(x) &= \frac{1}{2\pi} \int F(k)\ e^{-ikx} dk \\ \gamma(x,x') &= \int |G(k)|^2\ e^{ik(x-x')} dk \end{aligned}\right\} \tag{2-3}$$

where $|G(k)|^2$ is the spectral density of $g(x)$. Then

$$\eta = \frac{\left|\int H(k)\ F(k)\ dk\right|^2}{\int |F(k)|^2 |G(k)|^2\ dk} \tag{2-4}$$

and η is maximized by the choice

$$F(k) = \frac{H(k)^*}{|G(k)|^2} \tag{2-5}$$

whereupon η becomes

$$\eta_{max} = \int \frac{|H(k)|^2}{|G(k)|^2} \, dk \; . \tag{2-6}$$

B. Practical Constraints and the Improvement Factor

In practice the integral (2-6) will diverge, indicating an infinite signal-to-anomaly ratio. This arises physically because the signal ~~mine function~~ is assumed to have sharp edges, so that $|H(k)|^2$ goes down only like $(1/k)^2$ while $|G(k)|^2$ may decrease this rapidly, or even more rapidly, at large k. The resulting f(x), however, then has singularities (such as the derivative of a δ-function) and this condition is not physically realizable. In practice we cannot produce "wiggles" in f(x) of arbitrarily short wavelength; F(k) must be zero outside a certain range ($-K < k < K$) of wave-numbers for any physically realizable antenna. At frequencies such that the mine is largely in the antenna radiation field, rather than the induction field, the maximum possible wave-number K is determined by the frequency:

$$K = \frac{2\omega \sqrt{\epsilon_g}}{c} = \frac{4\pi}{\lambda_g} \tag{2-7}$$

the factor 2 arising because, in $f(x) = (E_1 \cdot E_2)$, the phase factor $\exp\left[i \frac{\omega}{c} x\right]$ appears in both E_1 and E_2. Since λ_g is the wavelength in the ground, we have roughly

$$K \simeq \frac{4\pi \gamma \sqrt{5}}{30} \simeq \gamma \, cm^{-1} \tag{2-8}$$

where γ is the frequency in kmc.

Because of these practical considerations, the limits of integration in (2-4) and (2-6) must be taken as ($\pm$ K) or the appropriate

generalization to 2 and 3 dimensions, and (2-5) must be understood as holding only within this range. We can now define a figure of merit $[\eta/\eta_{max}]$ which tells how far a given antenna function F(k) is from the best that could be attained within the same range of wave-numbers. We prefer to consider its reciprocal, the <u>improvement factor</u>

$$I\left[F(k)\right] = \frac{\eta_{max}}{\eta} \tag{2-9}$$

which shows how much the signal-to-anomaly ratio could be improved if one could obtain the correct design (2-5). For general mine and ground functions this is rather complicated, but we wish to point out a practical case where it simplifies greatly. Suppose we wish to find very small mines in soil of short correlation distance. Then within the restricted range of wave-numbers ($-K < k < K$) accessible to our antenna, $|H(k)|$ and $|G(k)|^2$ are practically constant. In this case the improvement factor reduces to

$$I\left[F(k)\right] = \frac{\int_R dk \cdot \int_R |F(k)|^2\, dk}{\left|\int_R F(k)\, dk\right|^2} = \frac{\overline{|F(k)|^2}}{\overline{|F(k)|}^2} \tag{2-10}$$

where the bar denotes an average over the region R of accessible wave-numbers. The factor $I[F(k)]$ now depends only on the shape of the function F(k). If F(k) is constant within R, then $I[F(k)] = 1$ and we are already at the optimum condition. Any variations in F(k), whether of magnitude or phase, within R, will increase $I[F(k)]$. As a numerical example, consider a one-dimensional model with $F(k) = 1 - \left|\frac{k}{K}\right|^n, |k| < K$. Working out the integrals, we find

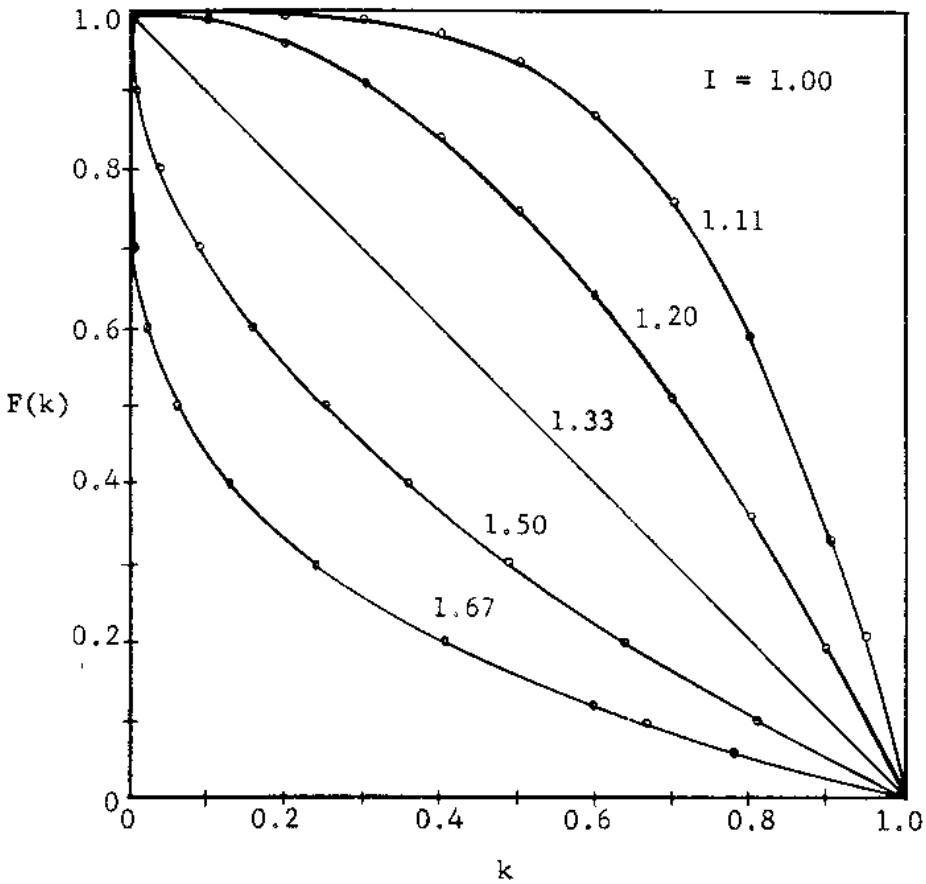
$$I\left[F(k)\right] = I(n) = \frac{2n+2}{2n+1} \tag{2-11}$$

This case is tabulated and sketched in Figure 2-1. Note that, although variations in F(k) are penalized, the improvement factor actually varies surprisingly little for different shapes.

Somewhat greater variations are found for the function $F(k) = \left| 1 - \frac{k}{K} \right|^n$, $|k| < K$. Here one finds

$$I(n) = \frac{(n+1)^2}{2n+1} \tag{2-12}$$

The curves for this case are given in Figure 2-2.

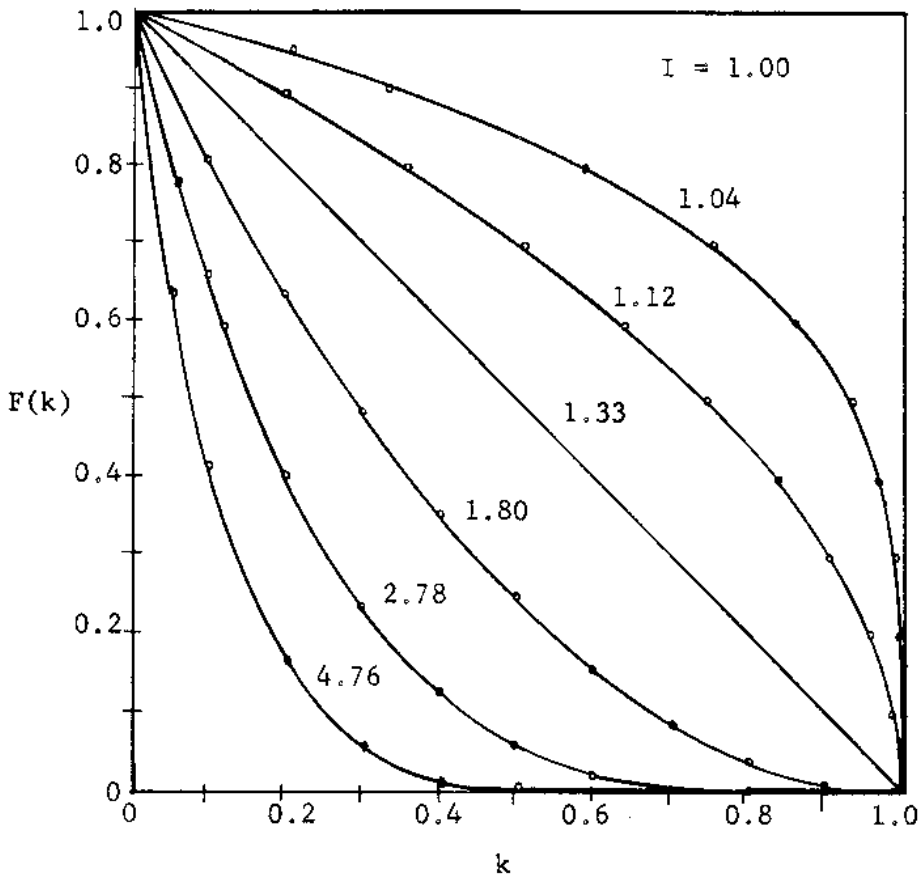


$$F(k) = 1 - |k|^n$$

n	I(n)
∞	1.00
4	1.11
2	1.20
1	1.33
0.5	1.50
0.25	1.67

FIGURE 2-1

FAMILY OF CURVES $F(k) = 1 - |k|^n$, $|k| > 2$, WITH CORRESPONDING IMPROVEMENT FACTORS I(n)



$$F(k) = \left|1 - k\right|^n$$

n	I(n)
0	1.00
0.25	1.04
0.50	1.12
1.00	1.33
2.0	1.80
4.0	2.78
8.0	4.76

FIGURE 2-2

FAMILY OF CURVES $F(k) = \left|1 - k\right|^n, \left|k\right| < 1$, WITH CORRESPONDING IMPROVEMENT FACTORS I(n)

If F(k) oscillates, very large values of $I[F(k)]$ may result, since oscillations reduce $\bar{F}$ much more than they do $\overline{F^2}$. For example, if

$$F(k) = e^{i\frac{\pi}{2}k^2a^2}, \quad |k| < K \tag{2-13}$$

we find I involves the Fresnel integrals (real and imaginary parts of the Cornu spiral), and for Ka > 1, a fair approximation is

$$I \simeq 2K^2a^2 \tag{2-14}$$

As another example, take

$$F(k) = \sin\left|\frac{(2n+1)\ \pi k}{K}\right|, \quad |k| < K \tag{2-15}$$

Now we find

$$I(n) = \frac{\pi^2}{8}\ (2n+1)^2 \tag{2-16}$$

The last two examples are important in showing that failure to control the phase of F(k), even though its amplitude is quite uniform, can reduce reliability very drastically. In fact, from all the above numerical examples we may draw a qualitative conclusion that in the antenna design problem as considered in the previous theory, control of phase is more important than control of amplitude. This conclusion is not altered by consideration of the more general case where H(k), $|G(k)|^2$ vary appreciably within the range of wave-numbers "seen" by the antenna, although the optimum phase of F(k) is not necessarily a constant as in the above examples.

The importance of these considerations for the present study lies in the following result, obtained in Section VI below:

<u>Let $|F(k)| \neq 0$ almost everywhere in the range $(-K < k < K)$. Then a data-processing computer can be built which will restore the full</u>

improvement factor I $[F(k)]$ and give the same signal-to-anomaly ratio as if the antenna function were the unconditional optimum (2-5) within this range. No linear operation on the signal can do better than this.

The implications of this result for the antenna design problem are considered in Section VI.

III. SUMMARY OF EXISTING THEORY OF SIGNAL DETECTION AND EXTRACTION

The following is a summary of some elementary aspects of the existing statistical theory of reception. It is essentially a distillation of the literature, leaning especially heavily on the recent summary of Van Meter and Middleton,[1] and a paper by R. C. Davis.[2] It contains no new results, although there are some simplifications in arguments and notation, as well as differences in interpretation. We are here concerned with exposition of concepts and principles, in a way that arrives as quickly as possible at the important results of the theory. In order to do this, mathematical rigor is dispensed with entirely; for that one must go back to the original papers.

A. Definitions

Notation: Let A,B,C,D,.... stand for various propositions, such as "a mine is present", "the observed signal is S," etc. Then

$(A|B)$ = Conditional probability of A, given B.

$(AB|CD)$ = Joint conditional probability of A and B, given C and D . . . . etc.

For our purposes, everything follows from the single fundamental rule of calculation,

$$(AB|C) = (A|BC)\ (B|C) = (B|AC)\ \ (A|C) \qquad (3\text{-}1)$$

If the propositions B, C are not mutually contradictory, this may be rearranged to give the rule of learning by experience, called Bayes' Theorem:

- - - - - - -

1 Van Meter, D. and Middleton, D., "Modern Statistical Approaches to Reception in Communication Theory", Trans. I.R.E., Professional Group on Information Theory (PGIT-4, Sept. 1954)

2 Davis, R.C., J. A. P. 25, 76 (1954)

$$(A|BC) = (A|B)\,\frac{(C|AB)}{(C|B)} = (A|C)\,\frac{(B|AC)}{(B|C)} \tag{3-2}$$

of which the last expression follows from the symmetry of $(A|BC)$ in B and C. Equation (3-2) shows how the probability of an event changes as a result of acquiring new information. It forms the cornerstone of our theory, when used to tell us the probability of the presence of a mine on the basis of the observed signal.

Summing (3-1) over B, we obtain the chain rule

$$(A|C) = \sum_B (A|BC)\,(B|C)\ . \tag{3-3}$$

Now let

X = prior knowledge, of any kind

S = signal

N = noise (anomaly signal)

V = V(S,N) = observed voltage

D = decision about the nature of the signal.

We adopt the view that there is no such thing as an "absolute" or "correct" probability; all probabilities are conditional on X at least, and X may vary with different observers or different situations. This is simply a recognition of the fact that all probabilities are, in the last analysis, expressions of human ignorance.

The purpose of probability theory is to aid us in forming plausible conclusions in situations where we do not have enough information to arrive at certain conclusions; as Laplace put it, probability theory is "common-sense reduced to calculation." It is important to avoid any impression that X is some kind of universally valid proposition about nature; it is simply whatever initial information we have at our disposal for attacking the problem. If X happens to be irrelevant to estimation of some quantity Y, then this attitude is unnecessary but harmless. Alternatively, we can equally well regard X as

a set of hypotheses whose consequences we wish to investigate, so that all equations may be read, "If X were true, then ----". It makes no difference in the formal theory. We often suppress X for brevity, with the understanding that even when it does not appear explicitly it is still "built into" all bracket symbols.

Any probabilities conditional on X alone are called <u>a priori</u> probabilities. Thus, we have

$(S|X)$ = a-priori probability of the particular signal S

$(N|X) = W(N)$ = a-priori probability of the particular sample of noise N.

In a linear system, $V = S+N$, and

$$(V|SX) = W(V-S) \tag{3-4}$$

The reader may be disturbed by the absence of density functions, dS's, dN's, etc., which might be expected in case of continuously variable S,N. Note, however, that our equations are homogeneous in these quantities, so they cancel out anyway. By $\sum_A$ we mean ordinary summation if A is discrete, integration with appropriate density functions if A is continuous.

A <u>decision rule</u> $(D|V)$ represents the process of drawing inferences about the signal from the observed voltage V. If it is always made in a definite way, then $(D|V)$ has only the values 0, 1 for any D and V; however, we may also have a randomized decision rule according to which $(D|V)$ is a true probability distribution. The essence of any decision rule which can be built into automatic equipment is that the decision must be made on the basis of V alone; V is, by definition, the quantity which contains all the information actually used (in addition to the ever-present X) in arriving at the decision. Thus for all Y, we have

$$(D|V) = (D|VY) \tag{3-5}$$

An equivalent statement is that D depends on any quantity Y only through the intermediate influence of Y on V:

$$(D|Y) = \sum_V (D|V)\ (V|Y) \tag{3-6}$$

B. Sufficiency and Information

Equation (3-5) has interesting consequences; suppose we wish to judge the correctness of the proposition Y on the basis of knowledge of V and D. From (3-1)

$$(Y|VD) = (D|VY)\ (Y|V)$$

and, using (3-5), this reduces to

$$(Y|VD) = (Y|V). \tag{3-7}$$

Thus, if V is known, knowledge of D is redundant and cannot help us in estimating any other quantity. The reverse is not true, however; we could equally well use (3-1) in another way:

$$(Y|VD)\ (V|D) = (Y|D)\ (V|YD)$$

Combining this with (3-7), there results the

Theorem - Let D be a possible decision, given V. Then

$$(V|D) \neq 0, \text{ and}$$

$$(Y|V) = (Y|D) \text{ if and only if } (V|D) = (V|YD). \tag{3-8}$$

In words: Knowledge of D is as good as knowledge of V for estimating Y if and only if Y is irrelevant for the estimation of V given D. Stated differently, in the "environment" produced by knowledge of D, the ~~quantities~~ propositions Y and V appear to be independent random quantities, that is,

$$(YV|D) = (Y|D)\ (V|D). \tag{3-9}$$

In this case, D is said to be a sufficient statistic for estimating Y. Evidently a decision rule which makes D a sufficient statistic for estimating the signal S is in some sense superior to one without this property. However, such a rule does not necessarily exist; equation (3-9) is a very restrictive condition, since it must be satisfied for all values of Y, V, and all D for which $(D|V) \neq 0$.

The concept of sufficiency is closely related to that of information. The definition of sufficiency could equally well be stated as: D is a sufficient statistic for estimating Y if it contains all the information about Y which V contains. Since D is determined from V, if it is not a sufficient statistic it necessarily contains less information about Y than does V. In this statement, the term "information" was used in a loose, intuitive sense; does it remain valid if we adopt Shannon's measure of information? The entropy, or degree of uncertainty, represented by a probability distribution (P_i) is, according to Shannon,[3] the expectation value of the "surprisal" $\log (1/P_i)$. Thus, the entropy of Y with a specific value of D given is

$$H_D (Y) = - \sum_Y (Y|D) \log (Y|D) \tag{3-10}$$

and its average over all values of D is

$$\bar{H}_D(Y) = \sum_D (D|X) H_D (Y) . \tag{3-11}$$

If

$$\bar{H}_C(Y) < H_D(Y) \tag{3-12}$$

we say that C contains, on the average, more information about Y than does D. Note, however, that it may be otherwise for specific values of C and D.

- - - - - - -

3 These notions are discussed further in Section IV below, where a derivation of Shannon's Theorem is given.

Acquisition of a new information can never increase $\overline{H}$; let D,V,Y be, for the moment, any three quantities and form the expression

$$\overline{H}_V(Y) - \overline{H}_{DV}(Y) = \sum_{DVY} (DV|X)\,(Y|DV)\log(Y|DV)$$

$$- \sum_{VY} (V|X)\,(Y|V)\log(Y|V)$$

$$= \sum_{DVY} (DV|X)\,(Y|DV)\log\left[\frac{(Y|DV)}{(Y|V)}\right]$$

$$\geq \sum_{DVY} (DV|X)\left[(Y|DV) - (Y|V)\right] = 0 \qquad (3\text{-}13)$$

Here we have used the fact that $\log X \geq \left(1 - \frac{1}{X}\right)$, with equality if and only if $X = 1$. Thus, $\overline{H}_{DV}(Y) \leq \overline{H}_V(Y)$, with equality if and only if equation (3-7) holds for all D,V,Y for which $(DV|X) \neq 0$. Since (3-13) holds regardless of the meaning of D and V, we can equally well conclude that for all D,V,Y,

$$\overline{H}_D(Y) \geq \overline{H}_{DV}(Y) \leq \overline{H}_V(Y)$$

Now letting D,V,Y resume their original meanings, we have in consequence of (3-7) $H_V = H_{DV}$, so that

$$\overline{H}_V(Y) \leq \overline{H}_D(Y) \qquad (3\text{-}14)$$

with equality if and only if equation (3-9) holds. Thus, if by "information" we mean minus the average entropy of Y over the a-priori distribution of D, zero information loss in going from V to D is equivalent to sufficiency of D for estimating Y. Note that inequalities of the form (3-13) hold only for the averages $\overline{H}$, not for the H. Acquisition of a specific piece of information (that an event previously considered improbable had in fact occurred) may in some cases increase the entropy of Y. However, we expect this to happen only rarely, and on the average the entropy can only be lowered by additional information. These considerations show that the

term "information" is not altogether a happy choice for entropy expressions; in spite of the entropy increase, the situation just described could hardly be called one of less information, but rather one of less certainty.

C. Loss Functions and Criteria of Optimum Performance

In order to say that one decision rule is better than another, we need some specific criterion of what we want our detection system to accomplish. The criterion will vary with the application, and obviously no single decision rule can be best for all purposes. A very general type of criterion is obtained by assigning a loss function $L(D,S)$ which represents our judgment of how serious it is to make decision D when signal S is in fact present. In case there are only two possible signals, $S_o = 0$ (i.e., no signal), and $S_1 \neq 0$, and consequently two decisions, D_o, D_1, there are two types of error: the false alarm $A = \left\{ D_1, S_o \right\}$ and the false rest $R = \left\{ D_o, S_1 \right\}$. In some applications one type of error might be much more serious than another: Suppose that a false rest is considered ten times as serious as is a false alarm, while a correct decision of either type represents no "loss"; we could then take $L(D_o, S_o) = 0$, $L(D_o, S_1) = 10$, $L(D_1, S_o) = 1$, $L(D_1, S_1) = 0$.

Whenever the possible signals and the possible decisions form discrete sets, the loss function becomes a loss matrix: In the above example,

$$L_{ij} = \begin{pmatrix} 0 & 10 \\ 1 & 0 \end{pmatrix}$$

The loss matrix plays approximately the same role in detection theory as does the payoff matrix in game theory.[4] A player in a game may adopt

- - - - - - -

4 Blackwell, D., and Girschick, M.A., Theory of Games and Statistical Decisions, Wiley and Sons, Inc., New York (1954)

that strategy which maximizes his expected gain; correspondigly, we may adopt that decision rule which minimizes the expected loss.

Instead of arbitrarily assigning a certain loss $L(D,S)$ to each possible type of error, we may consider information loss by the assignment $L(D,S) = -\log (S|D)$. A decision rule which minimizes information loss is one which makes the decision D in some sense as close as possible to a sufficient statistic for estimation of the signal. The information loss criterion is difficult to apply, because the function $L(D,S)$ depends on the decision rule.

The conditional loss $L(S)$ is the average loss incurred when the specific signal S is present:

$$L(S) = \sum_{D} L(D,S)\ (D|S) \tag{3-15}$$

which may in turn be expressed in terms of the decision rule and the properties of the noise by using (3-6). The average loss is the expected value of this over all possible signals:

$$\langle L \rangle = \sum_{S} L(S)\ (S|X) \quad . \tag{3-16}$$

Two different criteria of optimum performance now suggest themselves:

1. The Minimax Criterion

For a given decision rule $(D|V)$, consider the conditional loss $L(S)$ for all possible signals, and let $\left[L(S)\right]_{max}$ be the maximum value attained by $L(S)$. We seek that decision rule which makes $\left[L(S)\right]_{max}$ as small as possible. This criterion concentrates attention on the worst possible case regardless of the probability of occurrence of that case, and is thus in a sense the most conservative one. If the worst possible case is extremely unlikely to arise, one would probably call it too conservative. It has, however, the practical advantage that it does

not involve the a-priori probabilities of the signals $(S|X)$, and so can be applied when no information is available on which to base such a probability assignment.

2. The Bayes Criterion

We seek the decision rule for which the average loss $\langle L \rangle$ is minimized. In order to apply this, a distribution $(S|X)$ must be available. In Section IV below we describe methods for obtaining such a-priori distributions.

* * * * *

Other criteria have been proposed. Siegert's "Ideal Observer" minimizes the total probability of error regardless of type, but this is evidently a special case of the Bayes criterion in which each type of error is assigned the same loss. In case there are only two possible signals, the Neyman-Pearson criterion may be applied. Here one fixes the probability of one type of error at some small value ϵ, and then minimizes the probability of the other type of error subject to this constraint. This also is a special case of the Bayes criterion, where we find, not an absolute minimum, but a constrained minimum of $\langle L \rangle$. As shown below, the minimax criterion may be regarded as a supplemented version of the Bayes, in which we choose the worst possible $(S|X)$ after having found the Bayes solution for given $(S|X)$.

Substituting in succession equations (3-15), (3-6), and (3-3) into (3-16), we obtain for the average loss

$$\langle L \rangle = \sum_{DV} \left[\sum_{S} L(D,S) \; (VS|X) \right] (D|V) \qquad (3\text{-}17)$$

If $L(D,S)$ is a definite function independent of $(D|V)$, there is no function $(D|V)$ for which this expression is stationary in the sense of calculus of variations. We then minimize $\langle L \rangle$ merely by choosing for each possible V that decision $D_1(V)$ for which

$$K(D_1,V) = \sum_S L(D_1, S)\ (VS|X) \tag{3-18}$$

is a minimum; that is, we adopt the decision rule

$$(D|V) = \delta(D-D_1) \quad . \tag{3-19}$$

In general, there will be only one such D_1, and the best decision rule is non-random. However, in case of "degeneracy", $K(D_1,V) = K(D_2,V)$, any randomized rule of the form

$$(D|V) = a\ \delta(D-D_1) + b\ \delta(D-D_2)\ ,\ a+b = 1 \tag{3-20}$$

is just as good. This degeneracy occurs at "threshold" values of V where we change from one decision to another.

D. A Discrete Example

Consider the case already mentioned where there are two possibilities, $S_o = 0$, $S_1 \neq 0$, and a loss matrix

$$L_{ij} = \begin{pmatrix} L_{00} & L_{01} \\ L_{10} & L_{11} \end{pmatrix} = \begin{pmatrix} 0 & L_a \\ L_r & 0 \end{pmatrix}$$

where L_a and L_r are the losses incurred by a false alarm and false rest, respectively. Then

$$\begin{aligned} K(D_0,V) &= L_a(VS_1|X) \\ K(D_1,V) &= L_r(VS_0|X) \end{aligned} \tag{3-21}$$

and the decision rule that minimizes $\langle L \rangle$ is

$$\left\{\begin{array}{l} \text{choose } D_1 \text{ if } \quad \dfrac{(VS_1|X)}{(VS_0|X)} > \dfrac{L_a}{L_r} \\[2ex] \text{choose } D_0 \text{ if } \quad \dfrac{(VS_1|X)}{(VS_0|X)} > \dfrac{L_a}{L_r} \\[2ex] \text{choose either at random in case of equality} \end{array}\right\} \tag{3-22}$$

If the a-priori probabilities of signal and no signal are

start here

$$(S_1|X) = p, \quad (S_0|X) = q = 1-p \tag{3-23}$$

respectively, we have $(VS_1|X) = (V|S_1)\ (S_1|X) = p(V|S_1)$, etc., and the decision rule becomes

$$\text{choose } D_1 \text{ if } \quad \frac{(V|S_1)}{(V|S_0)} > \frac{qL_a}{pL_r} \quad \text{, etc.} \tag{3-24}$$

The left-hand side of (3-24) is called a likelihood ratio. It depends only on the statistical properties of the noise, and is the quantity which should be computed by the optimum receiver according to the Bayes criterion. The same quantity is the essential one regardless of the assumed loss function and regardless of the probability of occurrence of the signal; these affect only the threshold of detection. Furthermore, if the receiver merely computes this likelihood ratio and delivers it at the output without making any decision, it provides us with all the information we need to make optimum decisions in the Bayes sense. Note particularly the generality of this result, which is one of the most important ones for our applications; no assumptions are needed as to the type of signal, linearity of the system, or statistical properties of the noise.

We now work out, for purposes of illustration, the decision rules and their degree of reliability, for several of the above criteria. To make the problem as simple as possible, imagine a linear system in which the voltage is observed at a single instant, and we are to decide

whether a signal, which can have only amplitude S_1, is present in noise, which is gaussian with mean square value $\langle N^2 \rangle$:

$$W(N) = \frac{1}{\sqrt{2\pi \langle N^2 \rangle}} \exp\left[- \frac{N^2}{2 \langle N^2 \rangle} \right] \qquad (7\text{-}25)$$

The likelihood ratio in (7-24) then becomes

$$\frac{(V|S_1)}{(V|S_0)} = \frac{W(V-S_1)}{W(V)} = \exp\left[\frac{2VS_1 - S_1^2}{2 \langle N^2 \rangle} \right] \qquad (7\text{-}26)$$

and since this is a monotonic function of V, the decision rule can be written as

$$\text{choose} \begin{Bmatrix} D_1 \\ D_0 \end{Bmatrix} \text{when } V \begin{Bmatrix} > \\ < \end{Bmatrix} V_b \qquad (7\text{-}27)$$

with

$$\frac{V_b}{\sqrt{2 \langle N^2 \rangle}} = \frac{1}{2s} \left[\log\left(\frac{qL_a}{pL_r} \right) + s^2 \right] = v_b \qquad (7\text{-}28)$$

in which

$$s = \frac{S_1}{\sqrt{2 \langle N^2 \rangle}} \quad \text{is the signal-to-noise ratio,* and}$$

$$v = \frac{V}{\sqrt{2 \langle N^2 \rangle}} \quad \text{is the normalized voltage.}$$

Now we find for the probability of a false rest:

$$(R|X) = (D_0 S_1 | X) = p \sum_V (D_0|V)\,(V|S_1) = p \int_{-\infty}^{V_b} dV\, W(V-S_1)$$
$$= \frac{1}{2}\, P \left[1 + \operatorname{erf}(v_b - s) \right], \qquad (7\text{-}29)$$

* The factor $\frac{1}{\sqrt{2}}$ in this definition is perhaps unusual, but it makes the following equations especially simple.

and for a false alarm,

$$(A|X) = (D_1 S_0|X) = q \sum_V (D_1|V)\ (V|S_0) = q \int_{V_b}^{\infty} dV\ W(V) \qquad (7\text{-}30)$$

$$= \frac{1}{2}\, q \left[1 - \operatorname{erf} v_b \right] .$$

Here erf(X) is the error function

$$\operatorname{erf}(X) \equiv \frac{2}{\sqrt{\pi}} \int_0^X e^{-y^2} dy \qquad (7\text{-}31)$$

tabulated in Pierce.** For $X > 2$, a good approximation is

$$1 - \operatorname{erf}(X) \simeq \frac{e^{-X^2}}{X\sqrt{\pi}} \qquad (7\text{-}32)$$

As a numerical example, if $L_r = 10\ L_a$, $q = 10p$, these expressions reduce to

$$(A|X) = 10(R|X) = \frac{5}{11} \left[1 - \operatorname{erf}\left(\frac{1}{2}\, s\right) \right] \qquad (7\text{-}33)$$

The probability of a false alarm is less than 3×10^{-3}, and of a false rest less than 3×10^{-4}, for $s > 4$.

Let us see what the minimax criterion would give in this problem. The conditional losses are

$$L(S_0) = L_a \sum_V (D_1|V)\ (V|S_0) = L_a \int (D_1|V)\ W(V)\ dV$$

$$L(S_1) = L_r \sum_V (D_0|V)\ (V|S_1) = L_r \int (D_0|V)\ W(V - S_1)\ dV \qquad (7\text{-}34)$$

Writing $f(V) \equiv (D_1|V) = 1 - (D_0|V)$, the only restriction on $f(V)$ is $0 \leq f(V) \leq 1$. Since L_a, L_r, and $W(V)$ are all positive, a change

** Pierce, B.O., A Short Table of Integrals, Ginn and Co. (1929)

$\delta f(V)$ in the neighborhood of any given point V will always increase one of the quantities (7-34) and decrease the other. Thus when the maximum L(S) has been made as small as possible, we will certainly have $L(S_0) = L(S_1)$, and the problem is thus to minimize $L(S_0)$ subject to this constraint. Suppose that for some particular $(S|X)$ the Bayes solution happened to give $L(S_0) = L(S_1)$. Then this particular solution must be identical with the minimax solution, for with the above constraint, $\langle L \rangle = [L(S)]_{max}$, and if the Bayes solution minimizes $\langle L \rangle$ with respect to all admissible variations $\delta f(V)$ in the decision rule, it <u>a fortiori</u> minimizes it with respect to the smaller class of variations which keep $L(S_0) = L(S_1)$. Therefore our optimum decision rule will have the same form as before: There is some threshold V_m such that

$$f(V) = \begin{cases} 0, & V < V_m \\ 1, & V > V_m \end{cases} \qquad (7\text{-}36)$$

Any change in V_m from the value which makes $L(S_0) = L(S_1)$ necessarily increases one or the other of these quantities. The equation determining V_m is therefore

$$L_a \int_{V_m}^{\infty} W(V)\, dV = L_r \int_{-\infty}^{V_m} W(V-S_1)\, dV$$

or, in terms of normalized quantities,

$$L_a \left[1 - \text{erf } v_m\right] = L_r \left[1 + \text{erf } (v_m - s)\right] \qquad (7\text{-}37)$$

Note that (7-30), (7-31) give the conditional probabilities of false rest and false alarm for any decision rule of the type (7-36), regardless of whether the threshold was determined from (7-28) or not; for the arbitrary threshold V_0

$$\begin{aligned} (R|S_1) &= (V < V_0|S_1) = \tfrac{1}{2}\left[1 + \text{erf } (v_0 - s)\right] \\ (A|S_0) &= (V > V_0|S_1) = \tfrac{1}{2}\left[1 - \text{erf } v_0\right] \end{aligned} \qquad (7\text{-}38)$$

From (7-28) we see that there is always a particular ratio (p/q) which makes the Bayes threshold V_b equal to the minimax threshold V_m. For values of (p/q) other than this worst value, the Bayes criterion gives a lower average loss than does the minimax, although one of the conditional losses $L(S_0)$, $L(S_1)$ will be greater than the minimax value.

These relations and several previous remarks are illustrated geometrically in Figure 7-1, in which we plot the conditional losses $L(S_0)$, $L(S_1)$ and the average loss $\langle L \rangle$ as functions of the threshold V_0, for the case $L_a = \frac{3}{2} L_r$, $p = q = \frac{1}{2}$. The minimax threshold is at the common crossing-point of these curves, while the Bayes threshold occurs at the lowest point of the $\langle L \rangle$ curve. One sees how the Bayes threshold moves as the ratio (p/q) is varied, and in particular that the value of (p/q) which makes $V_b = V_m$ also leads to the maximum value of the $\langle L \rangle_{min}$ obtained by the Bayes criterion. Thus we could also define a "maximin" criterion; first find the Bayes decision rule which gives minimum $\langle L \rangle$ for a given $(S|X)$, then vary the a-priori probabilities $(S|X)$ until the maximum value of $\langle L \rangle_{min}$ is attained. This is the worst possible (in the Bayes sense) a-priori probability, and the design thus obtained is identical with the one resulting from the minimax criterion.

The Neyman-Pearson criterion is easily discussed in this example: Suppose the conditional probability of a false alarm $(D_1|S_0)$ is held fixed at some small value ϵ, and we wish to minimize the conditional probability $(D_0|S_1)$ of a false rest, subject to this constraint. Now the Bayes criterion minimizes the average loss

$$\langle L \rangle = pL_r\,(D_0|S_1) + qL_a(D_1|S_0)$$

with respect to any admissible variation $\delta(D|V)$ in the decision rule. In particular, therefore, it minimizes it with respect to the smaller class of variations which hold $(D_1|S_0)$ constant at the value finally obtained. Thus it minimizes $(D_0|S_1)$ with respect to these variations

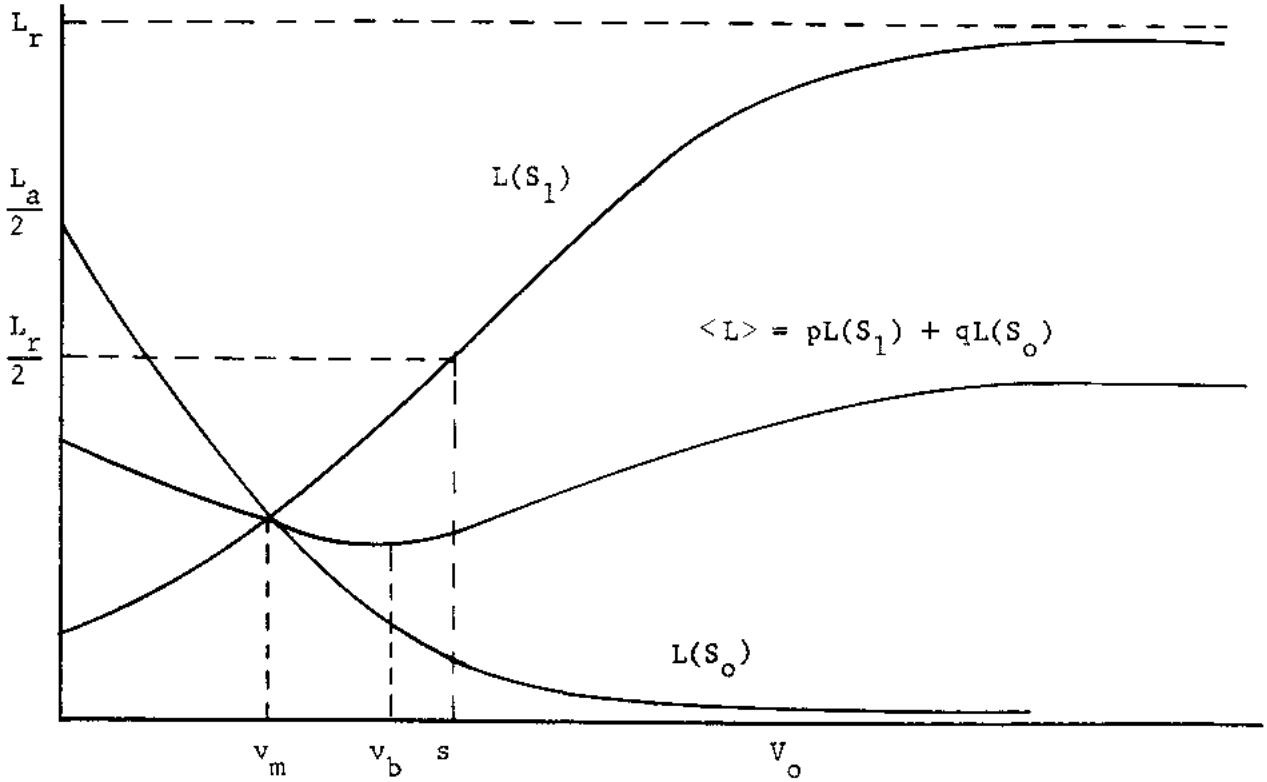


FIGURE 3-1

CONDITIONAL AND AVERAGE LOSSES AS FUNCTIONS OF THE DETECTION THRESHOLD V_O. THE $L(S_1)$ CURVE IS SYMMETRIC ABOUT THE POINT $\left\{ S_g \ 1/2 \ L_r \right\}$

and solves the Neyman-Pearson problem; we need only choose the particular value of the ratio (qL_a/pL_r) which results in the assumed value of ϵ according to equations (7-28), (7-30), and our previously found solution in the desired one.

We find for the Neyman-Pearson threshold, from (7-38)

$$v_{np} = \mathrm{erf}^{-1}\,(1-2\,\epsilon) \qquad (7\text{-}39)$$

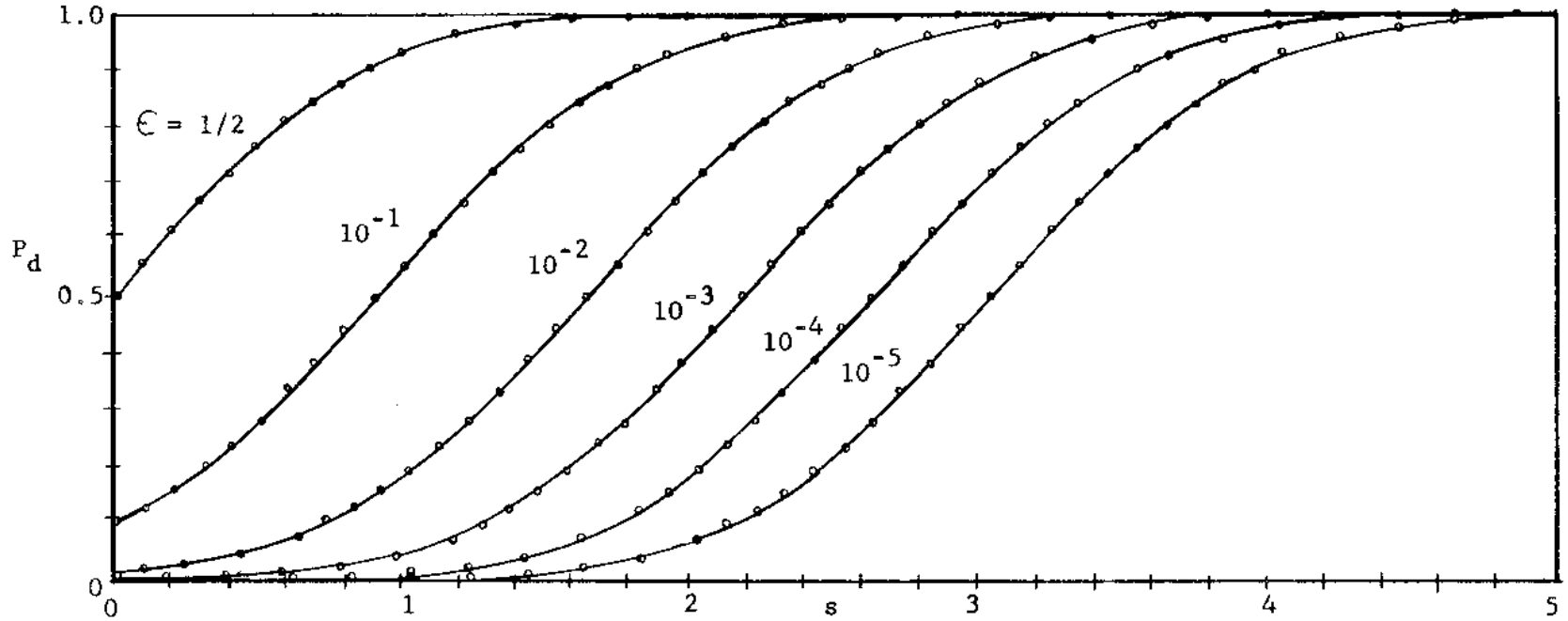
and the conditional probability of detection is

$$\begin{aligned}(D_1|S_1) &= 1 - (D_0|S_1) \\ &= \frac{1}{2}\left\{1 + \mathrm{erf}\left[s - \mathrm{erf}^{-1}(1 - 2\epsilon)\right]\right\}\end{aligned} \qquad (7\text{-}40)$$

This is plotted in Figure 7-2 as a function of s, for several values of ϵ. From it we see that if $\epsilon = 10^{-3}$, a detection probability of 99 per cent or better is attained for $s > 4$.

E. Gaussian Noise

It is important to note that these numerical examples depend critically on our assumption of gaussian noise. If the noise if not gaussian, the actual situation may be either more or less favorable than indicated by the above relations. It is well known that in one sense gaussian noise is the worst possible kind; because of its maximum entropy properties, gaussian noise can obscure a weak signal more completely than can any other noise of the same average power. On the other hand, gaussian noise is a very favorable kind from which to extract a fairly strong signal, because the probability that the noise will exceed a few times the RMS value $\langle N^2 \rangle^{1/2}$ becomes vanishingly small. Consequently, the probability of making an incorrect decision on the presence or absence of a signal goes to zero very rapidly as the signal strength is increased.



$$P_d = 1/2 \left\{ 1 + \mathrm{erf}\left[s - \mathrm{erf}^{-1}(1 - 2\epsilon) \right] \right\}$$

ϵ = probability of false alarm

FIGURE 3-2

PROBABILITY OF DETECTION, NEYMAN-PEARSON CASE

The high reliability of operation found above for $s > 4$ would not be found for noise possessing a probability distribution with wider "tails".

The type of noise distribution to be expected in any particular case depends, of course, on the physical mechanism which gives rise to the noise. When the noise is the resultant of a large number of small, independent effects, the central limit theorem of probability theory tells us that it will be gaussian regardless of the nature of the individual sources. ~~In the application to mine detectors, the anomaly signal is a resultant of several small signals due to individual regions of the soil with different dielectric constants. It is not known whether these regions are sufficiently small and numerous so that gaussian conditions are approached, but this appears highly doubtful, especially in the case of focused-beam systems with very small illuminated area.~~ However, it may be that the variations in individual sources ~~regions~~ are themselves gaussian, in which case no appeal to the central limit theorem is needed. Whether or not this is the case, in the absence of any information to the contrary, it is best to assume a gaussian noise distribution, ~~of anomaly signals, for reasons to be presented in Section IV below.~~

for reasons already given in Chapter 6.

VARIAN ASSOCIATES
611 Hansen Way, Palo Alto California

6. STATISTICAL INFERENCE BASED ON INFORMATION THEORY

A. Philosophical Digression

A large portion of statistical inference, and in particular the ~~The problem of signal detection, viewed statistically, is one~~ problem of detection of signals in noise, is one of reasoning from observed effects to probable causes. Historically, this type of problem is one of the oldest in probability theory, and for nearly two centuries it was considered solved by the application of Bayes' Theorem [equation (4-27)], also called the principle of inverse probability. Here one has a certain a-priori probability $(A|X)$ representing our judgment as to the truth of proposition A, and this judgment is altered by acquisition of new information B, according to the equation

$$(A|BX) = (A|X) \frac{(B|AX)}{(B|X)} \quad . \qquad (6\text{-}1)$$

In recent years, the situation has become obscured by rise of the positivist philosophy, according to which any probability is considered meaningless unless it can be measured numerically by observation of frequency ratios in a random experiment. The probability of an event is thus considered to be an objective property of that event, which exists independently of human knowledge.* Concurrently with this, the notion of a-priori probabilities has become a bête-noire to be avoided at all costs, except in the case that the information X consists of the result of a random experiment in which $(A|X)$ was measured. In the latter case, equation (6-1) is tolerated, the additional information B being interpreted as selecting out of the original population of events a

- - - - - - -

certain sub-class, in which the frequency of event A may be different from its value in the population as a whole. Evidently this is not the case in the problem of signal detection, and so many modern statisticians would not consider equation (6-1) as applicable.

To most people the natural way of approaching the detection problem would be to ask for a receiver which computes the probability that a signal is present, and either delivers this probability at the output, or else makes a decision that the signal is present (that is, sounds an alarm) whenever the probability exceeds a certain preassigned level. To such a program our hypothetical modern statistician would say, "This does not make sense; it is meaningless to speak of the probability that a signal is present because the signal is not a random variable; either it is there or it is not." He would then attack the problem in a quite different (and more complicated) way, for example, by introduction of confidence intervals. However, in so doing he would deprive himself of any way of incorporating prior information into the problem, and we have already seen how greatly the prior information can affect the interpretation of data.

In retaining equation (6-1) as the fundamental relation of detection theory we are rejecting the modern positivist views as to the nature of probability, and returning to those of Laplace. In our theory, the probability of an event is a formal means of expressing our expectation that the event will or did occur, and is not necessarily derived from any random experiment. In this subjective interpretation we recognize that probability theory and statistical inference are merely tools to aid us in forming plausible conclusions, on the basis of whatever information is available. Thus we regard <u>a-priori</u> probabilities, however arrived at, as perfectly respectable. The theory must, of course, break down when the available information is not sufficient to operate with, but that is as it should be.

In applying our theory, one of the first problems that confronts us is that of translating the initial information, which may be of very diverse nature, into an assignment of <u>a-priori</u> probabilities. This is the problem of "inventing thermometers" discussed in Chapter 4. As already noted, it is

not an easy problem, and from a psychological point of view it is probably safe to conclude that the difficulty of finding any unique assignments in cases where very little information is available is the real reason why the principle of inverse probability has fallen into disrepute in recent years. The theory has suffered from lack of any constructive principle which would give us a reason for preferring one probability assignment to another in cases where both agree equally well with the available information. We wish to show that, to a large extent, this principle has been supplied by the development of information theory. The application of the information principle is particularly simple and elegant when the initial information consists of average values of certain quantities, and we now turn to the problem of inference in this special case.

B. The Information Principle

The quantity x is capable of taking on the discrete values x_i, $(i = 1,2,\cdots,n)$. We are not given the corresponding probabilities p_i; all we know is the expectation value of the function f(x):

$$\langle f(x) \rangle = \sum_i p_i \; f(x_i) \qquad (6\text{-}2)$$

On the basis of this information, what can be said about the expected value of the function g(x)? At first glance, the problem seems insoluble because the given information is insufficient to determine the probabilities p_i. Equation (6-2) and the normalization condition

$$\sum_i p_i = 1 \qquad (6\text{-}3)$$

give only two relations; it would require specification of (n-2) more conditions before the quantities necessary for calculation of $\langle g(x) \rangle$ could be found.

In assigning the probabilities p_i we wish to make full use of whatever information is given, but in order to avoid any bias in our judgments we must be careful not to assume anything that is not given. The great advance provided by information theory lies in the discovery that there is a unique, unambiguous criterion for the amount of information contained in a discrete probability distribution, which agrees with our intuitive notions that a sharply peaked distribution represents more information than a broad one, and satisfies all other conditions which make it reasonable. It turns out to be somewhat more natural mathematically to consider uncertainty, or lack of information. At the end of this section we sketch Shannon's proof that the quantity which increases with increasing uncertainty and is additive for independent sources of uncertainty is

$$H(p_1 \ldots p_n) = -K \sum_i p_i \log p_i \qquad (6\text{-}4)$$

where K is a positive constant. Since this is just the expression found for entropy in statistical mechanics, it will be called the entropy of the probability distribution p_i; henceforth we shall consider the terms "entropy" and "uncertainty" as synonymous.

It is now evident how to solve our problem; in making estimates we must use that probability distribution which has maximum entropy subject to the given information. This is the only unbiased assignment we can make;

to use any other would amount to arbitrary assumption of information which by hypothesis we do not have. From this point on, the problem is elementary; we have to maximize (6-4) subject to the constraints of equation (6-2) and (6-3). The solution is, in fact, to be found in any textbook on statistical mechanics; by the method of Lagrangian multipliers we have

$$\delta \sum_i \left[p_i \log p_i + \lambda p_i + \mu p_i \; f(x_i) \right] = 0$$

from which

$$\log p_i + \lambda + \mu f(x_i) = 0$$

or,

$$p_i = e^{-\lambda - \mu f(x_i)} \tag{6-5}$$

The constants λ, μ are determined by substituting into (6-2) and (6-3). The results may be written as

$$\begin{aligned} \langle f(x) \rangle &= -\frac{\partial}{\partial \mu} \log Z(\mu) \\ 1 &= Z(\mu)\, e^{-\lambda} \end{aligned} \tag{6-6}$$

where

$$Z(\mu) = \sum_i e^{-\mu f(x_i)} \tag{6-7}$$

will be called the partition function.

This may be generalized to any number of functions f(x): Given the averages

$$\langle f_r(x) \rangle = \sum_i p_i f_r(x_i) \quad , \quad r = (1, 2 \ldots, m) \tag{6-8}$$

form the partition function

$$Z(\lambda_1, \ldots, \lambda_m) = \sum_i \exp - \left[\lambda_1 f_1(x_i) + \ldots + \lambda_m f_m(x_i) \right] \tag{6-9}$$

Then the maximum-entropy probability distribution is given by

$$p_i = \exp - \left[\lambda_o + \lambda_1 \; f_1(x_i) + \ldots + \lambda_m \; f_m(x_i) \right] \tag{6-10}$$

in which the constants are determined from

$$\left\langle f_r(x) \right\rangle = - \frac{\partial}{\partial \lambda_r} \log Z \tag{6-11}$$

$$\lambda_o = \log Z \tag{6-12}$$

The entropy of the distribution (6-10) then reduces to

$$S_{max} = \lambda_o + \lambda_1 \quad \langle f_1(x) \rangle \quad + \ldots + \lambda_m \quad \langle f_m(x) \rangle \tag{6-13}$$

where the constant K in (6-4) has been set equal to unity. The variance of the distribution of $f_r(x)$ is found to be

$$\Delta^2 f_r = \langle f_r{}^2 \rangle - \langle f_r \rangle^2 = \frac{\partial^2}{\partial \lambda_r{}^2} (\log Z) \; . \tag{6-14}$$

More generally, all central moments of order 2 or higher are obtained in this way; for example,

$$\left\langle \left(f_k - \langle f_k \rangle \right)^m \left(f_r - \langle f_r \rangle \right)^n \right\rangle = \frac{\partial^m}{\partial \lambda_k{}^m} \; \frac{\partial^n}{\partial \lambda_r{}^n} (\log Z) \tag{6-15}$$

In addition to its dependence on x, the function f_k may contain other parameters α_i, and it is easily verified that the maximum-entropy estimates of the derivatives are given by

$$\left\langle \frac{\partial f_k}{\partial \alpha_i} \right\rangle = \frac{1}{\lambda_k} \; \frac{\partial}{\partial \alpha_i} \; (\log Z) \quad , \tag{6-16}$$

etc.

The principle of maximum entropy may be regarded as an extension of the principle of insufficient reason (to which it reduces in case no

information is given except enumeration of the possibilities x_i), with the following essential difference. The maximum-entropy distribution may be asserted for the positive reason that it is uniquely determined as the one which is maximally non-committal with regard to missing information, instead of the essentially negative one that there was no reason to think otherwise. Thus the concept of information supplies the missing criterion of choice which Laplace needed to remove the apparent arbitrariness of the principle of insufficient reason, and in addition it shows precisely how this principle is to be modified in case there are reasons for "thinking otherwise."

Note A

It will be apparent from the above equations that the theory of maximum-entropy inference is identical in mathematical form with the rules of calculation provided by statistical mechanics. Specifically, let the energy levels of a system be

$$E_i \ (\alpha_1, \alpha_2, \ldots)$$

where the external parameters α_i may include the volume, strain tensor, applied electric or magnetic fields, number of molecules of each type, etc. Then if we know only the average energy $\langle E \rangle$, the maximum-entropy probabilities of the levels E_i are given by a special case of (6-10), which we recognize as the Boltzmann distribution, and such quantities as pressure, chemical potentials, etc., are given by special cases of (6-16). Thus, we can regard statistical mechanics, not as a physical theory, but as an example of statistical inference based on the information principle.

C. Shannon's Theorem

Evidently the crucial point of this theory of inference is the demonstration that the expression (6-4) actually represents a reasonable measure of information. We give here a condensed version of Shannon's proof of this fundamental result.

Note A

The maximum-entropy distribution has the important property that no possibility is ignored; it assigns positive weight to every situation that is not absolutely excluded by the given information. In its practical effect, this property is equivalent to an "ergodic theorem."

The variable x can assume the discrete values $(x_1, \ldots x_n)$. Our partial understanding of the processes which determine the value of x can be represented by assigning corresponding probabilities $(p_1, \ldots, p_n)$. We ask, with Shannon, whether it is possible to find any quantity $H(p_1 \ldots p_n)$ which measures in a unique way the amount of information (or, what amounts to the same thing, the amount of uncertainty) in this probability distribution. It might at first seem very difficult to specify conditions for such a measure which would ensure both uniqueness and consistency, to say nothing of usefulness. Accordingly, it is a remarkable fact that the most elementary conditions of consistency, amounting really to only one composition law, already determine the function $H(p_1 \ldots p_n)$ to within a constant factor. The three conditions are:

1. H is a continuous function of the p_i.

2. If all values of p_i are equal, the quantity $A(n) = H(1/n, \ldots 1/n)$ is a monotonic, increasing function of n.

3. The composition law applies. Instead of giving the probabilities of the events $(x_1 \ldots x_n)$ directly, we might group the first k of them together as a single event, and give its probability $w_1 = (p_1 + \ldots + p_k)$. Then the next m possibilities are assigned the total probability $w_2 = (p_{k+1} + \ldots + p_{k+m})$, etc. When this much has been specified, the amount of uncertainty as to the composite events is $H(w_1 \ldots w_r)$. Then we give the conditional probabilities $(p_1/w_1, \ldots, p_k/w_1)$ of the ultimate events $(x_1 \ldots x_k)$ knowing that the first composite event had occurred, the conditional probabilities for the second composite event, and so on. We arrive ultimately at the same state of knowledge as if the $(p_1 \ldots p_n)$ had been given directly. Therefore, if our information measure is to be consistent, we must obtain the same ultimate uncertainty no matter how the choices were broken down in this way. Thus, we must have

$$H(p_1 \ldots p_n) = H(w_1 \ldots w_r) + w_1 H(p_1/w_1, \ldots, p_k/w_1) + w_2 H(p_{k+1}/w_2, \ldots, p_{k+m}/w_2) + \ldots \qquad (6\text{-}17)$$

The weighting factor w_1 appears in the second term because the additional uncertainty $H(p_1/w_1, \ldots, p_k/w_1)$ is encountered only with probability w_1. For example, $H(1/2, 1/3, 1/6) = H(1/2, 1/2) + \frac{1}{2} H(2/3, 1/3)$.

From condition (1), it is sufficient to determine H for all rational values

$$p_i = \frac{n_i}{\sum n_i}$$

with n_i integers. But then condition (3) implies that H is determined already from the symmetrical quantities $A(n)$. For we can regard a choice of one of the alternatives $(x_1 \ldots x_n)$ as a first step in the choice of one of

$$\sum_{i=1}^{n} n_i$$

equally likely alternatives, the second step of which is also a choice between n_i equally likely alternatives. As an example, with $n = 3$, we might choose $(n_1, n_2, n_3) = (3, 4, 2)$. For this case the composition law becomes

$$H\left(\frac{3}{9}, \frac{4}{9}, \frac{2}{9}\right) + \frac{3}{9} A(3) + \frac{4}{9} A(4) + \frac{2}{9} A(2) = A(9)$$

In general, it could be written

$$H(p_1 \ldots p_n) + \sum_i p_i A(n_i) = A(\sum_i n_i). \qquad (6\text{-}18)$$

Particularly, we could choose all values of n_i equal to m, whereupon (6-18) reduces to

$$A(m) + A(n) = A(mn) \qquad (6\text{-}19)$$

Evidently this is solved by setting

$$A(n) = K \log n \qquad (6\text{-}20)$$

(Shannon has shown that (6-20) is in fact the only solution of (6-19).)

where, by condition (2), $K > 0$. Substituting (6-20) into (6-18), we have the desired result,

$$H(p_1 \ldots p_n) = K \log \left(\sum n_i\right) - K \sum p_i \log n_i$$
$$= - K \sum p_i \log p_i \, . \qquad (6\text{-}21)$$

In quantum statistical mechanics, this same expression is obtained for entropy by certain combinatorial arguments associated with an ensemble of N systems, the essential mathematical fact being the identity

$$\lim_{N \to \infty} \frac{1}{N} \log \left[\frac{N!}{(Np_1)! \, (Np_2)! \, \ldots \, (Np_n)!} \right] = - \sum p_i \log p_i \, . \qquad (6\text{-}22)$$

With Shannon's theorem, however, the expression (6-21) not only acquires a new significance, but it can "stand on its own feet" as the appropriate information measure without recourse to any ensemble or limiting process.

8 A GENERALIZED DETECTION PROBLEM

In this section we consider the problem of signal detection in a general way which does not refer to any specific physical problem. The results may then be interpreted in various ways so that they represent the theory of optimum filters, optimum receivers, optimum antennas, or optimum data-processing computers. We will use the term "filter" in a generic sense, to include all these possibilities.

$\mathcal{G}$ small G

g = small G

There is a certain region R in which a function $\mathcal{G}(x)$ is observed. It may be one-dimensional (as in an ordinary "time-domain" filter), or multi-dimensional (as in an antenna which scans a two-dimensional region). The observed function $\mathcal{G}(x)$ will consist of "noise" $g(x)$, with or without a "signal" $h(x)$. Thus

$$\mathcal{G}(x) = \begin{cases} g(x) + h(x), & \text{if a signal is present} \\ g(x), & \text{if no signal is present.} \end{cases} \qquad (8\text{-}1)$$

Certain statistical properties of the noise $g(x)$ are known, and the signal $h(x)$ contains certain parameters. We are to find the best way of deciding whether or not the signal is present, and the best estimates of its parameters if it <u>is</u> present.

It is customary to distinguish between an autocorrelation function, where one averages over all translations, and a covariance function, in which the average is taken instead over some ensemble of different functions at a given point. If there actually exists any definite ensemble, these averages may well be different. The existence of an ensemble implies that one has some way of enumerating the different, mutually exclusive possibilities in a way independent of the study of a single function $g_a(x)$, and some criterion by which a weighting is assigned to each possibility. Now in ~~any~~ most applications of statistics to a physical "random" function, there is no such ensemble; we have one and only one physical situation and the consideration of other situations than the one in fact existing could not possibly have any bearing on the problem.

If in spite of this we continue to use the notion of an ensemble, we have to recognize the following: The only source of information which is available for setting up an ensemble is the single function $g(x)$. By the probability p_k of a particular function $g_k(x)$ defined in a region R, we could mean only the average frequency, over some much larger region, with which a configuration locally, like g_k, appears in the single function $g(x)$. The use of an ensemble is a great convenience mathematically, since the average of some functional $F\left[g(x)\right]$ can then be written in a way that appears to avoid difficult mathematical questions in the limiting process

small subsets

$$\langle F \rangle = \lim_{X\to\infty} \frac{1}{2X} \int_{-X}^{X} F\left[g(x)\right] dx \qquad (8\text{-}3)$$

of the direct calculation. However, the important point we wish to make here is that the difficulty is in no sense avoided by using the ensemble; it is merely pushed back into the problem of determining the p_k and $g_k(x)$ in the ensemble. These probabilities must now, of course, be recognized as entirely subjective quantities expressing the fact that in any given run of data we do not know in advance which particular local sample of the function $g(x)$ will be present.

It is apparent from the foregoing that in our problem the distinction between an autocorrelation function and a covariance function is usually entirely artificial; they are necessarily the same function. We will use an ensemble-type notation because it is easy to handle and makes the equations look simpler. We must frankly recognize that nothing is really gained thereby; if there were any real difficulties about the existence of limits of the type (8-3), those difficulties would still be with us but merely hidden from view by a trick of notation. In practice no such difficulties can exist, because we have only a finite sample of the function $g(x)$ to operate with; thus (8-3) does not correctly describe the process actually used to find averages. The quantity X does not tend

to infinity, but represents the size of some finite interval over which we have measured g(x). This is, furthermore, never as large as the entire domain of existence of g(x), otherwise we would have no need of statistical methods at all. If g(x) were measured over its entire domain of existence, it would make no sense to call it a random function. The practical use of a statistical approach is to enable us to infer certain things about g(x) in regions where it has not been measured, on the basis of measured values of g(x) in some smaller region. Thus, the correct interpretation of (8-3) is not the mathematical passage to a limit; the notation is merely a crude way of indicating that we do not want to commit ourselves to any particular value of X, but that, other things being equal, we should prefer to use the largest possible value of X in our calculations. It would be a good idea to invent some new notation specifically for this purpose.

The quantities $g_{\blacksquare}(x)$, $h_{\blacksquare}(x)$ are in general complex-valued. ~~since we can measure both the amplitude and phase of the voltage at the antenna terminals~~. This reminds us of a problem not considered here, but worth studying: Suppose we measure only the amplitude, only the phase, or only the real component, etc., of the function g(x). ~~antenna voltage~~. Which of these contains the most information about the signal, and how much reliability of detection is lost through failure to measure the entire complex function?

The statistical properties of $g_{\alpha}(x)$ would be rather awkward to state in terms of the probabilities of certain particular functions $g_{a}(x)$, or the probabilities of various values at various points. Since it is a quadratically integrable ~~continuous~~ function, however, (from conservation of energy), its statistical properties can be described in terms of a denumerable set of ~~number~~ random variables as follows: Let $\phi_n(x)$ be a complete set of orthonormal functions in the region R, and expand

$$g_{\alpha}(x) = \sum_n g_n \, \phi_n(x) \quad . \tag{8-4}$$

The coefficients g_n are now the random variables. We can choose the ϕ_n so that the g_n ~~values~~ are uncorrelated:

$$\langle g_n \, g_m^* \rangle = \frac{\delta_{nm}}{\lambda_n} \tag{8-5}$$

where the brackets $\langle \; \rangle$ stand for averages over the "ensemble". The covariance function then becomes

$$\gamma_{\alpha}(x,x') \equiv \langle g(x) \, g^*(x') \rangle = \sum_{mn} \langle g_n g_m^* \rangle \, \phi_n(x) \, \phi_m^*(x')$$

$$= \sum_n \frac{\phi_n(x) \, \phi_n^*(x')}{\lambda_n} \tag{8-6}$$

which we recognize as the bilinear form of a Green's function, so that ϕ_n and λ_n are the eigenfunctions and eigenvalues of the integral equation

$$\phi(x) = \lambda \int_R \gamma_{\alpha}(x,x') \, \phi(x') \, dx' \tag{8-7}$$

This equation leaves an arbitrary constant (that is, independent of x) phase factor $e^{i\theta_n}$ in each ϕ_n. Since the g_n are complex we write

$$g_n = a_n + ib_n \tag{8-8}$$

and choose θ_n so that a_n and b_n are uncorrelated:

$$\langle a_n b_n \rangle = 0 \qquad (8\text{-}9)$$

This amounts to a rotation of coordinates so the concentration ellipse of g_n is aligned with the real and imaginary axes. Now define

$$\left.\begin{aligned} \langle a_n^2 \rangle &= A_n^2 \quad , \quad \langle b_n^2 \rangle = B_n^2 \\ u_n &= \frac{a_n}{A_n} \quad , \quad v_n = \frac{b_n}{B_n} \end{aligned}\right\} \qquad (8\text{-}10)$$

Then u_n, v_n are real, uncorrelated random variables with variance unity. If the ~~ground statistics are~~ noise is gaussian, they are therefore independent random variables with probability distributions

$$\begin{aligned} p(u_n)\, du_n &= \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2} u_n^2}\, du_n \\ p(v_n)\, dv_n &= \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2} v_n^2}\, dv_n \end{aligned} \qquad (8\text{-}11)$$

According to the principles of ~~Section IV, at the present state of knowledge we~~ Chapter 6, we usually have no rational basis for using any distribution other than gaussian. If experiments should demonstrate that ~~actual soil~~ the actual noise ~~statistics are~~ far from gaussian, then new possibilities would be opened up in which different types of filters than those found here would give greater reliability of detection than we will obtain. The new filters would, however, be non-linear and undoubtedly more difficult to realize.

We write the ~~ground~~ noise function as

$$g_n(x) = \sum_n (A_n u_n + iB_n v_n)\, \phi_n(x) \qquad (8\text{-}12)$$

with probability distribution $P\left[g(x)\right]$, which is a symbolic shorthand expression for (8-11). Also there may or may not be a signal centered at position x_o:

$$h(x) = M\, s(x-x_o) \qquad (8\text{-}13)$$

where M is a real amplitude so defined that

$$\int_R |s(x)|^2\, dx = 1 \qquad (8\text{-}14)$$

We expand s(x) also in the orthogonal functions $\phi_n(x)$:

$$s(x-x_o) = \sum_n \left[A_n S_n'(x_o) + iB_n S_n''(x_o)\right] \phi_n(x) \qquad (8\text{-}15)$$

We wish to find the procedure ~~data-processing computer~~ which enables us to make the best possible inferences about the value of x_o. ~~location of mines~~. More specifically, given the function ~~resultant voltage~~ q(x) ~~at the antenna terminals~~ over a certain region R, we wish to test the hypothesis that a signal ~~mine~~ with parameters M, x_o is in R against the hypothesis that no signal ~~mine~~ is in R. The "observed voltage" is also expanded in the functions $\phi_n(x)$:

$$q(x) = \sum_n \left[A_n q_n' + iB_n q_n''\right] \phi_n(x) \qquad (8\text{-}16)$$

The theory of Section 7 ~~III~~ makes it clear that regardless of the particular criterion of optimum performance we adopt, we are going to find that the fundamental quantity needed is the likelihood ratio

$$L(M,x_o) = \frac{P\left[q - h(M,x_o)\right]}{P\left[q\right]} \qquad (8\text{-}17)$$

$$= \frac{\left[\begin{array}{l}\text{Probability that } q_A(x) \text{ would be observed if a signal of amplitude } M \\ \quad \text{were present at position } x_o\end{array}\right]}{\left[\text{Probability that } q_A(x) \text{ would be observed if no signal were present}\right]}$$

From (8-11), we obtain

$$\log L(M,x_o) = \frac{1}{2}\sum_n \left\{ q_n'^2 + q_n''^2 - \left[q_n' - M\, S_n'(x_o)\right]^2 - \left[q_n'' - M\, S_n''(x_o)\right]^2 \right\}$$

$$= M\,\alpha(x_o) - \frac{M^2}{2}\,\beta(x_o) \qquad (8\text{-}18)$$

where

$$\alpha(x_o) \equiv \sum_n \left[q_b'\, S_n'(x_o) + q_n''\, S_n''(x_o)\right] \qquad (8\text{-}19)$$

$$\beta(x_o) \equiv \sum_n \left\{ \left[S_n'(x_o)\right]^2 + \left[S_n''(x_o)\right]^2 \right\} \qquad (8\text{-}20)$$

As a function of M, $L(M,x_o)$ reaches a maximum given by

$$\log L_{max}(x_o) = \frac{\alpha^2}{2\beta} \qquad (8\text{-}21)$$

when

$$\alpha - M\beta = 0$$

We define the total likelihood that a signal is at x_o, regardless of M, by

$$L(x_o) = \int L(M,x_o)\, dM = \frac{2\pi}{\beta(x_o)}\, L_{max}(x_o)\ . \qquad (8\text{-}22)$$

Equations (8-19) and (8-20) are cast into a more unified form if we define a function

$$\Phi(x_o,x) = \sum_n \left[\frac{S_n'(x_o)}{A_n} + i\,\frac{S_n''(x_o)}{B_n}\right] \phi_n(x) \qquad (8\text{-}23)$$

which might be called the "reconstituted signal". It is a filtered version of $s(x)$, in which parts of $s(x)$ that are likely to be highly contaminated with noise are suppressed, while parts that are unlikely to be so affected are amplified. In terms of $\Phi(x_o,x)$ we find

$$\alpha(x_o) = \mathrm{Re} \int \Phi^*(x_o,x)\ q(x)\ dx \qquad (8\text{-}24)$$

$$\beta(x_o) = \mathrm{Re} \int \Phi^*(x_o,x)\ s(x-x_o)\ dx \qquad (8\text{-}25)$$

For given noise ~~ground~~ properties and shape of signal, $\Phi(x_o x)$ and $\beta(x_o)$ are determined once and for all.

The function $\beta(x_o)$ provides a measure of the mean-square noise level, weighted according to how much a particular noise pattern resembles the signal $s(x-x_o)$. If $s(x)$ very much resembles $\emptyset_n(x)$, then noise which also resembles $\emptyset_n(x)$ is very serious, but noise resembling some other $\emptyset_m(x)$ does not hurt so much because it can be filtered out without losing much of the signal.

The above relations were developed in the most general way of the "ensemble" language. They are greatly simplified if now we make use of the fact that $\gamma(x,x')$ is ~~nearly~~ usually an autocorrelation function; therefore by definition it can depend only on $(x-x')$. This determines the nature of the $\emptyset_n(x)$; for (8-6) then implies that

$$\frac{d}{dx''}\left[\emptyset_n(x+x'')\ \emptyset_n^*(x'+x'')\right] = 0$$

and this can be true only if the $\emptyset_n$ are plane waves:

$$\left.\begin{aligned} \emptyset_n &= (\text{const.})\ e^{i\,k_n x} && \text{1 dimension} \\ \emptyset_n &= (\text{const.})\ e^{i\,\vec{k}_n\cdot\vec{x}} && \text{2 dimensions} \end{aligned}\right\} \qquad (8\text{-}26)$$

Now the probability distribution of $g(x)$ must be the same at all positions; that is, $g(x)$ is a "stationary" random function [in view of the above discussion of ensembles, this is necessary by definition; otherwise we would have no basis for setting up any probability distribution at all for $g(x)$]. But a translation $\phi_n(x) \longrightarrow \phi_n(x+x'')$ merely shifts the phase of ϕ_n, therefore of g_n. Thus the random variable $(g_n e^{i\theta_n})$, $\theta_n \equiv (\vec{k}_n \cdot \vec{x}'')$, must have exactly the same probability distribution as does g_n, and so we must have

$$A_n = B_n = \frac{1}{\sqrt{2\lambda_n}} \quad . \tag{8-27}$$

The reconstituted signal is now

$$\Phi(x_o,x) = \Phi(x_o-x) = \sum_n \frac{S_n}{A_n} \phi_n(x-x_o) \tag{8-28}$$

where

$$S_n \equiv S_n'(0) + iS_n''(0) \tag{8-29}$$

so that $\beta(x_o)$ no longer depends on x_o but is the appropriate mean-square noise level.

The most important fact for intuitive understanding of this theory,

is the observation that in consequence of the simplification (8-28), the "reconstituted signal" (which is the weighting function for the linear filter whose output is $\alpha(x_o)$) itself satisfies an integral equation:

$$2S^*(x) = \int_R \Phi^*(x') \, \gamma(x',x) \, dx' \tag{8-30}$$

This is readily verified by substituting (8-6) and (8-28) into (8-30). Define a function

$$f(x) \equiv \frac{M}{2} \Phi^*(x)$$

then we have

TO HERE

$$h_a^*(x) = \int f_a(x')\ \gamma_a(x', x)\ dx' \qquad (8\text{-}31)$$

which is identical with (2-2) except for the subscript "a". It seems astonishing at first that the problem of the computer which maximizes the probability of detection and the problem of the antenna response function which maximizes signal-to-anomaly ratio should turn out to have identical solutions (note that they are not the same in the more general treatment where $A_n \neq B_n$), but consider the following two statements:

1. The computed likelihood of the presence of a mine has a sharp peak at x_o, much higher than in surrounding regions.

2. The signal due to a mine has a sharp peak at x_o that rises far above the background noise.

On second thought, one sees that these statements are not really very different, and it is quite plausible that engineering designs of receivers, based only on the attempt to increase the signal-to-noise ratio, actually come fairly close to the best design that sophisticated statistical analysis can give. Once again, however, it must be emphasized that the simplicity of the final results of the statistical theory depends entirely on the assumption of gaussian statistics; for non-gaussian statistics much more complicated, non-linear filters would be found.

Once the basic identity of these problems is recognized, we can adopt a global view and consider the antenna design problem and the data-processing problem simultaneously; this is done in the next section.

VI. A GLOBAL VIEW OF THE PROBLEM

In the work on Fort Belvoir Contract No. DA-44-009 eng-795 at Stanford University, a solution was found to the following problem:

Problem A - For a given mine and given statistical properties of the soil, find the antenna function which maximizes the signal-to-anomaly ratio as seen at the antenna terminals.

In task 3 of the present contract, we have solved a different problem:

Problem B - For a given mine, given soil statistics, and given antenna function, find the best way of processing the received data so as to achieve the maximum reliability of detection.

The surprising fact emerged that, in the case of gaussian soil statistics, the solutions to these problems were mathematically identical. The meaning of this is seen most clearly in terms of the fourier transforms of the various functions. For convenience, we list here the needed functions and their fourier transforms:

	function	fourier transform
Mine function	$h(x)$	$H(k)$
Ground autocorrelation function	$\gamma(x)$	$\lvert G(k)\rvert^2$
Antenna function	$f(x)$	$F(k)$
Mine function as seen at antenna terminals	$h_a(x)$	$H_a(k) = H(k)\ F(k)$
Anomaly autocorrelation function as seen at antenna terminals	$\gamma_a(x)$	$\lvert G_a(k)\rvert^2 = \lvert G(k)\rvert^2 \lvert F(k)\rvert^2$
Weight function of filter in computer	$f_a(x) = 1/2\ M\ \phi^*(x)$	$F_a(k)$

(6-1)

Now, just as the solution of the integral equation (2-2) for the optimum antenna function assumes the form (2-5),

$$F(k)_{opt} = \frac{H^*(k)}{|G(k)|^2} \tag{6-2}$$

the solution of the integral equation (5-3) for the optimum weight function of the data-processing computer is

$$F_a(k) = \frac{H^*_a(k)}{|G_a(k)|^2} = \frac{H^*(k)\ F^*(k)}{|G(k)|^2\ |F(k)|^2} \tag{6-3}$$

But this reduces to

$$F_a(k)\ F(k) = \frac{H^*(k)}{|G(k)|^2} \quad . \tag{6-4}$$

Thus if F(k) were actually equal to the "unconditional optimum" (6-2), the solution of the computer problem would reduce to $F_a(k) = 1$, or $f_a(x) = \delta(x)$; that is, no computer is needed. If the antenna function is not the unconditional optimum, then (6-4) shows that the computer acts as a "corrective filter" to restore at the computer output exactly the same signal that would have appeared at the antenna terminals, had the antenna been the unconditional optimum. Viewed in this way, the computer problem assumes the more familiar form of construction of a filter with prescribed amplitude and phase characteristics.

In the usual theory of filters in the time domain one cannot prescribe the amplitude and phase responses independently; they are connected by certain physical realizability conditions (the Bode relations) which ensure that no signal can appear at the output before it arrives at the input. The output at time t must depend only on the input at times prior to t. In the present case, however, there is no such restriction on physical realizability since the independent variable is position rather than time. Both the antenna and the computer can "see" forward as well as backward. This greatly simplifies the theory, for it means that, whatever finite amplitude and phase may be specified for $F_a(k)$ by (6-4), it is always possible in principle to construct a

computer which will accomplish this. However, in regions where $F(k) = 0$, the filter receives no input to work with, and the value of $F_a(k)$ is irrelevant, so long as its magnitude is not great enough to introduce noise (true thermal or shot noise, as distinguished from anomalies) in the computer output.

More generally, the practical ~~filter~~ design criterion for a computer is: In all regions of k-space where the available signal $F(k)\,H(k)$ is above the thermal noise level, $F_a(k)$ should be chosen to satisfy (6-4). In other regions $F_a(k)$ is irrelevant to detection and may as well be taken as zero. Since we have seen in Section II above (equation 2-7) that $F(k)$ will be different from zero in only a finite range of k-values, this will be true also of the optimum $F_a(k)$. Thus the optimum computer response will never involve singular functions such as derivatives of δ-functions, but will always be well-behaved.

We are now in a position to state the final conclusions of this study of data-processing computers. They are essentially all contained in (6-4), which is the fundamental result of this work.

a. If one decides that a computer is not going to be used, then the antenna design criterion remains the same as that found in the Stanford reports; the antenna response function that maximizes signal-to-anomaly ratio at its terminals also maximizes the reliability of detection. Of several antennas, the best one is the one for which the expression (2-4) is a maximum. Here the integration is taken over the region R of k-space for which $H(k)\,F(k)$ is measurably large. Thus two things are desirable: the region R should be as large as possible, and within this region the antenna function should come as close as possible to satisfying (6-2).

b. If one decides that a computer is going to be used, the antenna design criterion is considerably changed. The phase of F(k) now has no importance, since the computer can always straighten it out again. For the same reason, variations in the magnitude of F(k) are unimportant as long as this magnitude is not too small. The only important thing is the size of the region R of k-space within which F(k) H(k) is sufficiently large to override thermal noise. Of several antennas, the one for which $|F(k)|^2$ is large over the greatest range of wave-numbers will provide the computer with the greatest amount of information, and lead to the best detection.

c. For any antenna which delivers a signal F(k) H(k) above thermal noise in a certain region R of k-space, it is possible to build a computer which corrects phase and amplitude, and resynthesizes the same signal and the same signal-to-anomaly ratio as would be produced by the unconditional optimum antenna operating in the same region R. Thus the full improvement factor of Section II can be restored by a computer.

d. In the case of gaussian soil statistics, it is impossible to build a computer which does better than the computer of c.

e. If typical soil correlation functions were known, it would be possible, merely from laboratory measurements of the antenna function F(k), to predict in advance both the signal-to-anomaly ratio obtainable with the "bare" antenna, and also the amount of improvement which is possible with a computer.

f. An antenna for which F(k) maintains essentially constant phase over the region R can never be helped very much by a computer. This is demonstrated by the numerical values in Figures 2-1 and 2-2.

VII. AN EXPERIMENTAL EXAMPLE

In Figures 6-1 and 6-2 we give some fourier transforms of antenna functions obtained by Mr. R. E. Gang on a microwave antenna (3.4 kmc) developed at Varian Associates. In these measurements the antenna patterns were obtained by moving the antenna past a small probe antenna, which fed a square-law crystal detector, the output of which was recorded. This did not give the true antenna function $f(x) = (E_1 \cdot E_2)$, but, if one considers using a single-dish system for which $E_1 = E_2$, only the magnitude $|f(x)|$ was obtained. Nevertheless, the results are of interest because they represent the only available measurements of this sort, and because other existing mine detectors also give only magnitudes; thus if one were to consider adding a computer to existing mine detectors, this example is a realistic one. The fourier transform of $|f(x)|$ was obtained numerically, with an estimated accuracy of about 5 per cent of the maximum value.

From these measurements several conclusions can be obtained. First, consider the maximum wave-number K to be expected in F(k) according to equation (2-7). It has the value

$$K = \frac{2\omega}{c} = \frac{4\pi \times 3.4 \times 10^9}{3 \times 10^{10}} = 1.42 \text{ cm}^{-1} \qquad (7\text{-}1)$$

The maximum value found in the measurements was $K = 0.7 \text{ cm}^{-1}$ for the plain antenna, and $K = 0.4 \text{ cm}^{-1}$ for the antenna when equipped with a metallic sheet in the ground plane, in which an 8-inch diameter aperture was cut. Thus, the range of wave-numbers actually attained is about one-half of the maximum theoretically possible, or in two-dimensional k-space, the region R has only about one-quarter of the area possible. Thus for detection of very small mines, a considerably higher efficiency antenna is theoretically possible at this frequency.

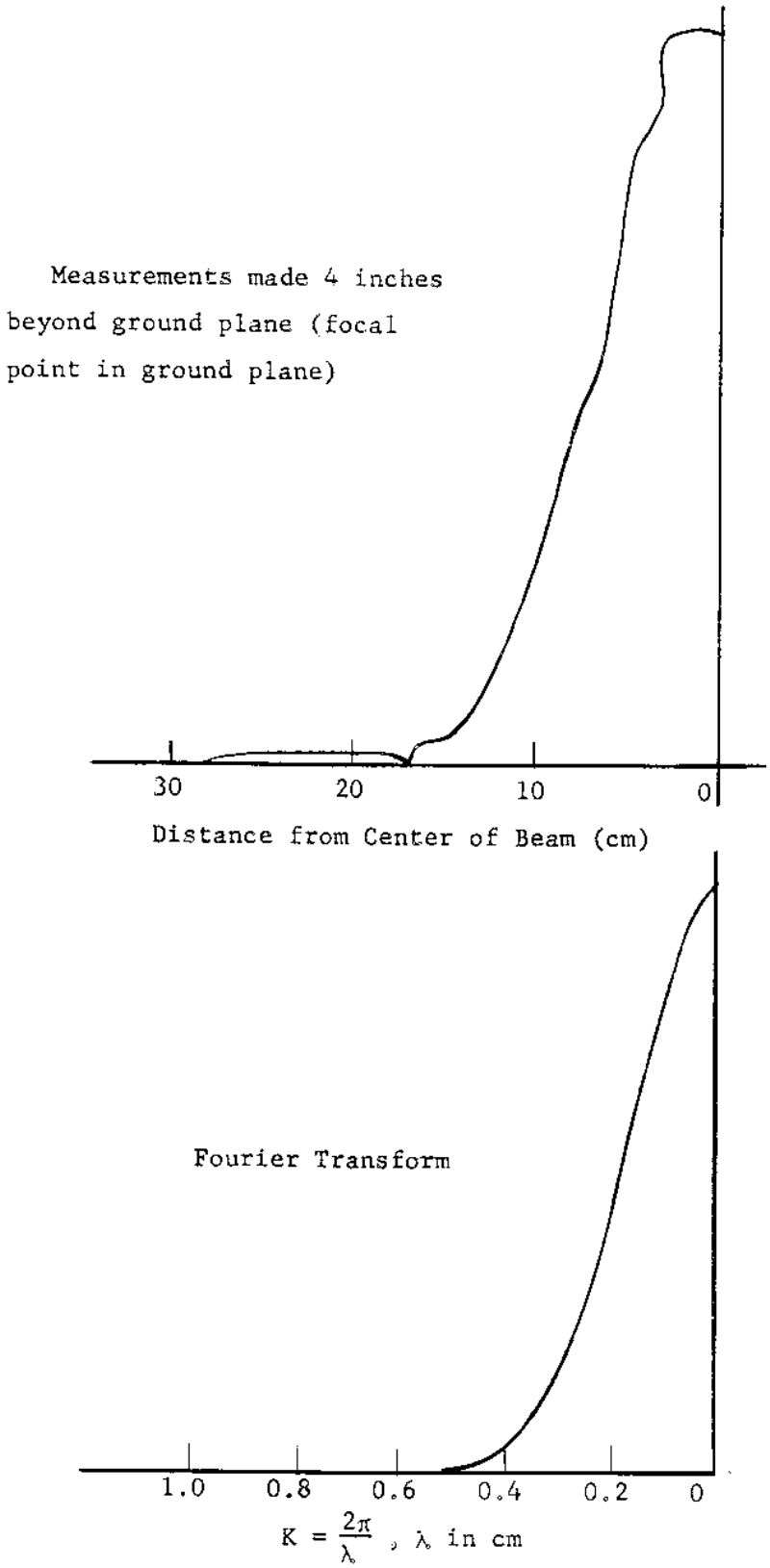


FIGURE 6-1
ANTENNA PATTERN THROUGH 8-INCH
DIAMETER APERTURE IN GROUND PLANE

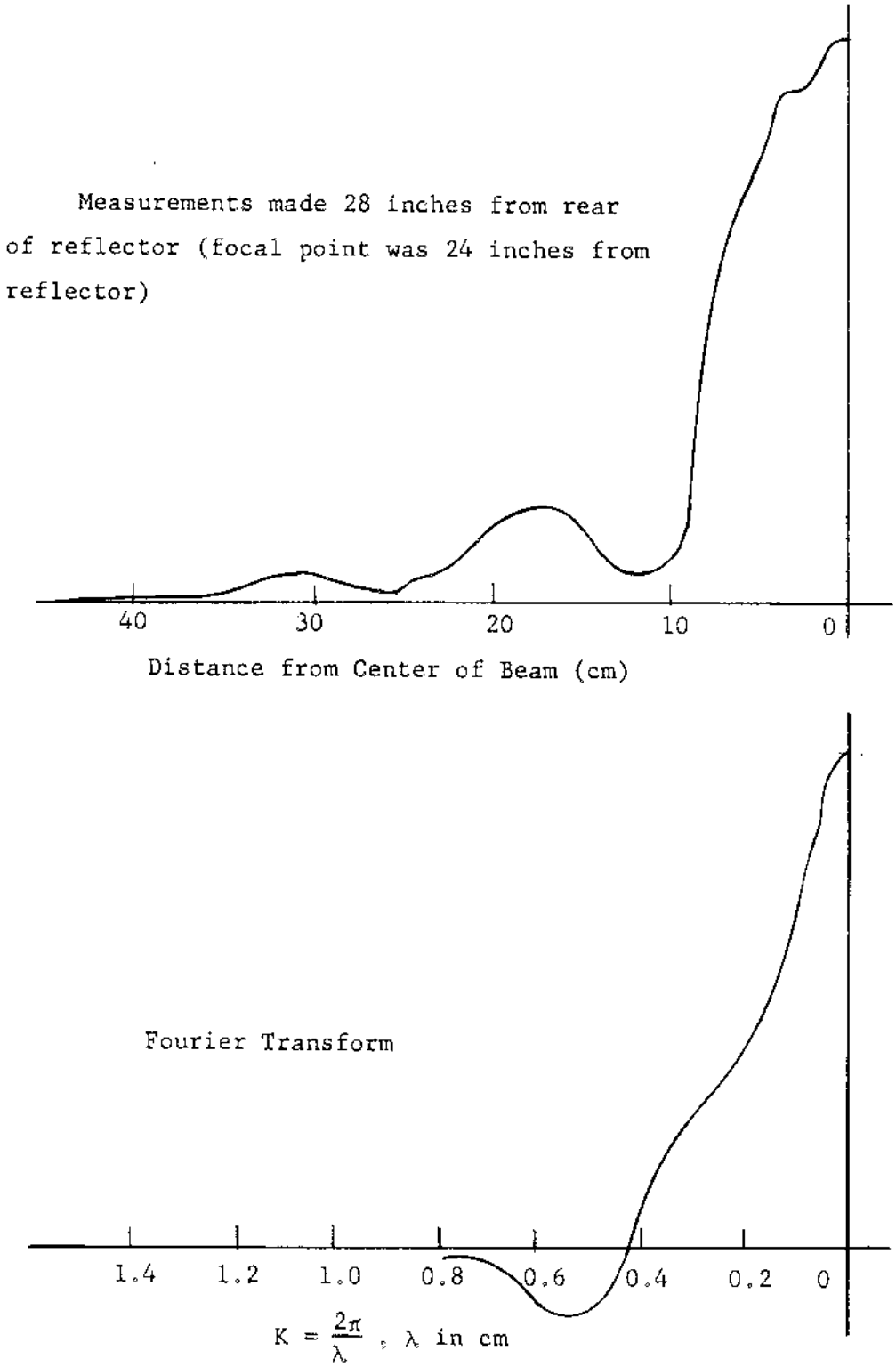


FIGURE 6-2

ANTENNA PATTERN FROM 24-INCH DIAMETER PARABOLIC REFLECTOR

It is seen that the effect of the ground plane with aperture was to cut off the "side lobes" of the antenna pattern. As long as one looks only at the pattern f(x) and thinks in terms of a small illuminated area, the aperture sheet appears to be very beneficial. However, when one looks at the fourier transforms F(k), it is apparent that cutting off the side-lobes reduces the range of available wave-numbers by almost 50 per cent; therefore if one were to use a computer with this antenna, it would be better to eliminate the ground-plane sheet. The side-lobes, although weak and spread out, contain "high-frequency" components that deliver useful information to the computer. This gives a good example of how radically the antenna design requirements are changed by the possibility of using a computer.

Continuing with interpretation of these measurements, we note that if one is forced to use a ground plane with the antenna, a computer cannot give much improvement; comparing with Figures 2-1 and 2-2, we note that the improvement factor for the antenna with aperture plane is only about 1.3. Therefore, the most that a computer could do is to increase the signal-to-anomaly ratio by about 1 db. What this means in reliability of detection depends, of course, on what the signal-to-anomaly ratio was before adding the computer. However, it could not amount to much. On the other hand, a computer used on the antenna without aperture plane would give a very substantial improvement in performance.

VIII. CONCLUSIONS AND RECOMMENDATIONS

With the completion of this work, the theory of mine detection appears to have reached a "plateau" from which little further progress can be made until certain experimental results are obtained. We now have a theory which relates the properties of mine, soil, antenna, and computer, and shows how all these factors affect the reliability of detection. Nevertheless, because of lack of experimental data it is impossible at the present time to answer two crucial questions:

a. How close are present mine detectors to the best that could ever be made?

b. Is the expected improvement from a computer sufficiently great to justify a computer development program?

For further progress in the mine detection field it is imperative that we obtain experimental values of the soil function $|G(k)|^2$ and the antenna function F(k) for various soils and various antennas now in use. In this connection we may note that the answer to question (a) above involves very difficult measurements for which new techniques will have to be developed. On the other hand, the answer to question (b) is relatively easy to obtain from apparatus and methods already in use, for the theory of the optimum computer can be developed entirely in terms of the quantities $H_a(k)$, $|G_a(k)|^2$ measurable directly at the antenna terminals, as in Section V above, no reference to the ultimate quantities H(k), F(k), $|G(k)|^2$ being necessary. The quantity $h_a(x)$ is just the "ideal marble-block mine signal", while $\gamma_a(x)$ is obtainable from statistical analysis of responses over unmined soil.

There is one direction, however, in which further theoretical work involving only electromagnetic theory, instead of statistics, could help to advance the mine-detection art even without new experiments. This concerns the new antenna design requirements found for an antenna which

is to be followed by a computer. The problem might be stated in the form of two more questions:

c. What are the ultimate restrictions imposed by electromagnetic theory on the range R of wave-numbers over which $|F(k)|^2$ can be large, for an antenna operating at a single frequency?

d. How does one design an antenna so as to achieve the maximum size of the region R?

From the reasoning of Section II above, it may be conjectured that the answer to question (c) is that the maximum R is a circle of radius $K = 4\pi/\lambda$. Question (d) remains entirely unexplored.