FC

U.S. AIR FORCE
Project RAND
RESEARCH MEMORANDUM

This is a working paper. It may be expanded, modified, or withdrawn at any time. The views, conclusions, and recommendations expressed herein do not necessarily reflect the official views or policies of the U. S. Air Force.

The RAND Corporation
SANTA MONICA • CALIFORNIA

U. S. AIR FORCE
PROJECT RAND

RESEARCH MEMORANDUM

THE THEORY OF INFORMATION

Edgar Reich

RM-454

ASTIA Document Number AD 116555

20 September 1950

Assigned to ______________

This is a working paper. It may be expanded, modified, or withdrawn at any time. The views, conclusions, and recommendations expressed herein do not necessarily reflect the official views or policies of the United States Air Force.

The RAND Corporation
1700 MAIN ST. • SANTA MONICA • CALIFORNIA

The Theory of Information

Table of Contents

The Theory of Information

Table of Contents (Con't)

THE THEORY OF INFORMATION

I. Introductory Remarks

I. Scope and Continuity of this Report

This report discusses the modern theory of 2-point unidirectional communication that is associated with the names of Shannon and Wiener in the light of Shannon's Theory of information. While being for the most part an outline of Shannon's classical paper (22), the report also sketches some applications and presents a discussion on the question of uniqueness of formulation of the theory of information.

In an attempt not to obscure the underlying train of thought, some of the mathematical proofs are heuristic in nature. The theory's present state makes this ine.itable anyway.

The block diagram below summarizes the continuity of the paper

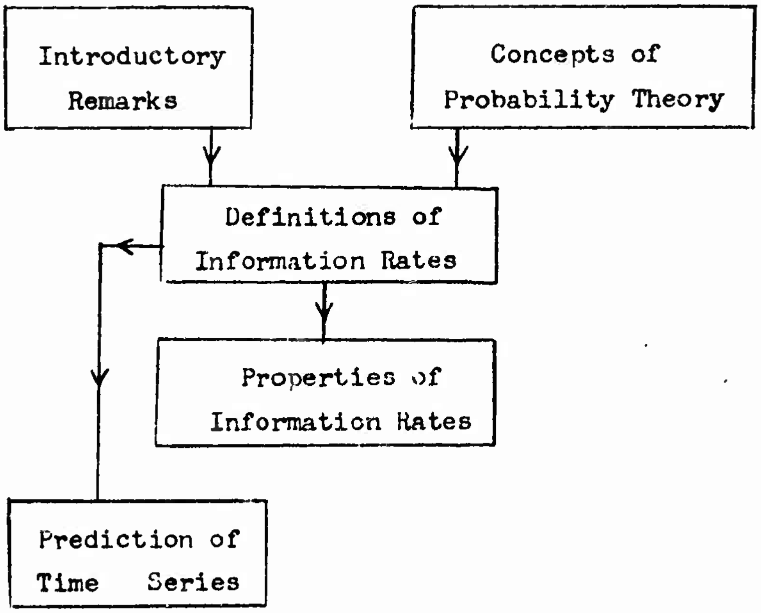
Introductory Remarks
Concepts of Probability Theory
Definitions of Information Rates
Properties of Information Rates
Prediction of Time Series

Fig. 1 Continuity Diagram

2. General Remarks on the Theory of Information

During the last decade or so it has been realized that communication in the presence of noise is a problem susceptible to treatment by the methods of probability theory. In such treatments we have all been accustomed to the frequent use of such scalars as the second moments of distributions, etc. Shannon has shown the great usefulness of defining another scalar, called the information rate, and has built up a theory of communication in which information rate plays a fundamental part. The crux of the theory is that information rate is a scalar capable of characterizing a source in such a manner as to specify the speed at which source messages are to be transmitted in order that they may be received without error in spite of the presence of a given intervening noise. (Actually it is usually possible only to transmit with an arbitrarily small, but non-zero, probability of error, but this is a fine point of the type that will henceforth be overlooked.)

In Shannon's paper information rate is introduced by first defining a quantity called information which is shown to warrant that name because it satisfies many of the intuitive requirements for such a quantity. For the sake of variety, a different approach is used in this report. We postulate certain requirements for a scalar which is to be called information rate, and show, by assuming certain restrictions, that the postulates imply a unique formulation. This "uniqueness theorem" gives some insight into the fact that information rate seems to be of such fundamental importance not only for the problem of two-point communication, but for broader fields as well [1].

[1] Cf. references 4, 7, 19

As regards applicability of the theory to design of specific communication links as well as appraisal of existing links, such attempts usually turn out to be discouragingly difficult. In this respect information theory can be compared to electromagnetic theory where the analytic work involved in solving specific problems is often forbidding. In information theory the fundamental "undefined" variables are "emitted symbol" and "received symbol". The existence of noise in the transmitting channel is taken care of in the theory by not requiring that the received symbols be the same as the symbols emitted by the source, but only that there be a statistical dependence between the two. The fundamental problem is to "code" the emitted symbols in such a way as to best combat the noise. In order to take into account the fact that the recipient may not be interested in all the detail of the emitted symbols the concept of "fidelity" is introduced. It is evident that a vast number of problems arising in technology can be described in terms of information theory by posing the problem of how to best modify (i.e. "code" or "modulate") the output of some source (i.e. "emitted symbol") so as to best suit the destined recipient, but where there is a chance that the article transmitted will be distorted along the way. The following may briefly be cited as examples.

(a) A source emits real numbers between 0 and 1 at the rate of ten per second, the distribution of the number being known. The recipient is interested in knowing the output correct to three decimal places. The transmitting facilities are capable of transmitting only 0's and 1's at the rate of 25 per second, and are disturbed by noise in such a way that if a certain symbol (i.e. a 0 or 1) is transmitted there is a probability of

1/4 that the wrong symbol is received. The question arises: Is it possible to satisfy the recipient's desire of 3-place accuracy, and if so, how should be source output be coded? Note that although it will be necessary to represent the real number in terms of a sequence of 0's and 1's, this does not necessarily mean that the representation should be of the binary type (i.e. base 2 representation). The fact that the noise corrupts 0's and 1's indiscriminately and independently would make it likely that a binary representation, where some digits carry more weight than others, is not as good as a more hybrid type of representation.

(b) Speech is to be transmitted over a channel having a bandwidth of 10 cps. The transmitter is capable of delivering an (average) power of 1 kw. The channel (including input circuit of receiver) is permeated by white noise of 2 watts intensity. What type of modulation system should be used if the only criterion of fidelity is that the transmitted speech is received in intelligible form?

(c) A photo-electric device equipped with telescope is to be capable of indicating on a 3-position dial whether a cloud at which the telescope is pointed is predominantly of the cirrus, stratus, or cumulus type. How should such a device be made? To fit this situation into the mathematical model of information theory ~~it~~ is necessary to make the following interpretations:

sky → source

device → coder

space between dial and eye of observer (possibly also nervous system of observer, etc.) → channel fact that channel can transmit only the words "cirrus", "stratus", and "cumulus", and that these words are transmitted without error when so indicated on the dial ⟶ channel "noise" characteristic

To give an indication of the potentialities of information theory we will now outline what information theory "tells us to do" in each of the three above cases. The appendix will show more specifically how the statements below follow from the theory developed in the body of the report.

(a) Information theory gives a mathematical scheme for obtaining the optimum representation of the real numbers in the system using only 0's and 1's. This scheme requires minimizing functions of several variables, solving equations, etc., and could be achieved by a great deal of horse work. The resulting optimum system will require an "infinite" delay at the transmitter, and thus would have to involve a storage tube or equivalent device. It is likely that if a common-sense coding scheme were used instead, the resulting system, although not strictly optimum, would have a much greater chance of being practically physically realizable.

(b) Information theory tells us to build a detector capable of recognizing speech sounds, and a coder to code the detector output into samples of white noise. Information theory does not give any technically valuable hints as how to build the speech sound detector.

(c) Information theory tells us to build the indicating device but gives no worthwhile indication of how to go about it.

From the above we see that information theory is in most cases unfortunately merely a device for rephrasing already well-realized technical difficulties into more generalized form. The main selling point of information theory is that in reducing difficulties to a more generalized form it may of conceptual help in their solution.

II Concepts of Probability Theory

1. Summary

Probability distributions necessary for a statistical description of 2-point communication in the presence of noise are defined.

2. Statistical Description of Unidirectional 2-Point Communication

We are concerned with the description of a link made up of a source, producing symbols, x, which are corrupted by noise into received symbols, y.

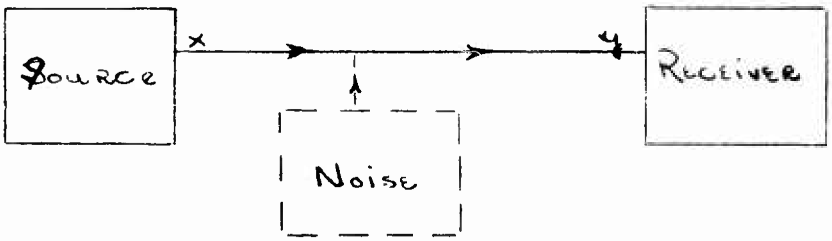

Fig. 2 Fundamental Communication Link

It is convenient to define "symbol" in terms of the actual output of the source in such a way that successive such symbols are independent and are affected independently by noise. For instance, if the source is one that produces letters of written English in the presence of a noise that affects successive letters independently, a symbol should be defined as a group of ten or more consecutive letters, because successive such groups are practically independent in written English[2].

[2] See reference 22. To be on the safe side it might be necessary to use considerably longer groups to eliminate context.

With the foregoing in mind, a message can be defined as a sequence of independent symbols. Messages can be described in terms of the probabilities of their symbols, the probability of a certain symbol being the fraction of time it occurs in a long message. The usefulness of the probabilistic approach is that for many statistical sources occurring in nature the probabilities associated with long messages from a given source are the same for all long messages from that source.

In the presence of noise the emission of a certain symbol, x, by the source may result in the situation that the corresponding received symbol, y, is not the same of x. Physically, the noise occurs in the transmission link, or "channel". The channel is described statistically by associating a family of transition probabilities with the noise. We define

$q_x(y)dy$ = probability that the received symbol will be in the region (y,y+dy) of the symbol space <u>if</u> the emitted symbol is x.

Let us also define

$p(x)dx$ = probability that emitted symbol is in (x,x+dx).

These two distributions determine a joint probability:

$p(x,y)dxdy = p(x)dxq_x(y)dy$, the probability that a symbol in the range (x,x+dx) will be emitted <u>and</u> (as a result of this) a symbol in the range (y,y+dy) received.

Focusing our attention of the received symbols without reference to their prime cause, we see a statistical situation described by

$q(y)dy$ = probability that a symbol in the range (y,y+dy) is received.

It is also possible to define the inverse transfer probability

$p_y(x)dx$ = probability that the emitted symbol was in (x,x+dx) <u>if</u>

the received symbol was y.

All the distributions can be expressed in terms of $p(x,y)$:

$$\left\{\begin{aligned} p(x) &= \int p(x,y)dy \\ q(y) &= \int p(x,y)dx \\ q_x(y) &= p(x,y)/p(x) \\ p_y(x) &= p(x,y)/q(y) \quad . \end{aligned}\right.$$

It should be noted that before the receipt of a symbol the recipient's knowledge of what *will be* emitted is characterized by the distribution $p(x)$, while after the receipt of say the symbol $y = m$ the relevant distribution as to what *was* sent is $p_m(x)$. It is therefore natural to think of $p(x)$ as the *a-priori* distribution, and $p_m(x)$ as the distribution of what was emitted *a-posteriori* to receipt of m. To emphasize this we will often write $p(x) = p_o(x)$.

For cases in which the variables assume only a discrete set of values the distributions can be obtained by use of the Dirac delta function. We have

$$p(x,y) = P(i,j)\delta(x-i)\delta(y-j)$$

$$p(x) = P(i)\delta(x-i)$$

$$q(y) = Q(j)\delta(y-j)$$

$$q_x(y) = \frac{p(x,y)}{p(x)} = \frac{P(i,j)}{P(i)}\,\delta(y-j) = Q_i(j)\delta(y-j)$$

$$p_y(x) = P_j(i)\delta(x-i)$$

where $P(i,j)$, $P(i)$, $Q(j)$, $Q_i(j)$, $P_j(i)$ are the analogous probabilities for the discrete symbols.

III Definitions of Information Rates

1. Summary

A definition of information-receipt rate is evolved from fundamental postulates. Derived definitions are then formulated for the concepts of information rate of a source and information-transmitting rate of a channel.

2. Information-Receipt Rate

Let I(m) denote the information obtained as the result of receiving the particular message $y = m$. The following postulate suggests itself:

Postulate I: I(m) is a scalar which depends on the a-priori and a-posteriori distributions of what the source emitted:

$$I(m) = \Phi\left[p_o(x),\ p_m(x)\right]$$

with the property that

$$\Phi\left[p_o(x),\ p_o(x)\right] = 0 .$$

The second part of the postulate implies that no information was gained if the a-posteriori probability as to what was transmitted is the same as the a-priori one as to what will be transmitted.

As the entire process under consideration is a statistical one it is to be expected that statistical functions of I will play a more important part than I itself. We define the "information-receipt" rate R as the average amount of information received per symbol, i.e. as the expected value of I:

$$R = E\left[I(m)\right] = \int q(m)I(m)dm.$$

R should be invariant under any transformation that merely amounts to a one-to-one relabeling of the message symbols without changing the fundamental physical process; otherwise the information obtainable from a message could be changed by restating the message in a logically equivalent way. For instance, suppose the received message is read from a meter calibrated according to y^3 instead of y. If the distributions p_o and p_m are recalculated on the basis of x^3 instead of x the resulting value of R should be the same. Now a relabeling of the variable $x \rightarrow f(x)$ transforms a distribution p(x) into the distribution $p(x)/f'(x)$ where $x = g(z)$ is the function inverse to $z = f(x)$. Therefore we have the following postulate:

Postulate II: The transformation

$p(x) \rightarrow p(g(x))/f'(g(x))$ where g is the inverse of f, and p generically represents all the probability distributions entering into the definition of R, leaves R invariant.

Actually we will not consider the problem of finding the most general functional Φ that satisfies the postulates, because this problem is too difficult, and has not yet been solved to the author's knowledge.

Assumption I: I is of the form

$$I = \int F(p_o(x), x)dx - \int F(p_m(x), x)dx$$

where $F = F(u,v)$ is some real function of two real variables.

We can think of $\int F(p_i(x), x)dx$ as the "uncertainty" associated with the distribution $p_i(x)$. Then the restricted class of definitions of I determined by assumption I is one in which the received information is taken as the difference between an a-priori uncertainty and an a-posteriori uncertainty. Note that the assumed fashion by which the distribution function determines the associated uncertainty is a common one for assigning scalars to distribution functions; for instance, the k'th moment of a distribution p(x) can be written in the form $\int F(p(x), x)dx$ if we take $F(u,v) = uv^k$.

It will be assumed that F has continuous partial derivatives through the second order, and if fact, we will from now on assume all sorts of "good behavior", including interchangeability of order of integration, etc. With these limitations in mind the following uniqueness theorem will be derived:

Theorem I: If the definition of information is restricted as in assumption I, then in order to satisfy postulates I and II it is necessary and sufficient that

$$R = \text{constant} \cdot \int\int p(x,y) \log \frac{p(x,y)}{p_o(x)q(y)}\, dxdy \qquad (3)$$

Proof:

Since Assumption I automatically implies that postulate I is satisfied, and therefore it is only necessary to subject R to the conditions of postulate II. We have

$$(1) \qquad R = \int F(p_o, x)dx - \int\int q(m)F(p_m, x)dxdm.$$

The invariance condition implies that

[3] This is the formula for information-receipt rate proposed by Shannon on the basis of considerations other than those employed here. For definitions of symbols see page 8

$$\int F\left[\frac{p_o(g(x))}{f'(g(x))}, x\right]dx - \int\int \frac{q(g(m))}{f'(g(m))} \; F\left[\frac{p_m(g(x))}{f'(g(x))}, x\right] dxdm =$$

(2) $$= \int F\left[\frac{p_o(x)}{f'(x)}, f(x)\right] f'(x)dx - \int\int q(m)F\left[\frac{p_m(x)}{f'(x)} , f(x)\right] f'(x)dxdm$$

is independent of the choice of f.

(3) Let $F(u,v) = uG(u,v)$. Then (2) becomes

(4) $\int p_oG(p_o/f',f)dx - \int\int qp_mG(p_m/f',f)dxdm.$

(5) Subject f to the variation $\Delta f(x) = \epsilon w(x)$. Since (4) is independent of f the corresponding variation of (4) must vanish:

(6.1) $$-\int p_o^2G_u(p_o/f',f)w'/f'^2dx - \int\int qp_m^2G_u(p_m/f',f)w'/f'^2dxdm +$$

(6.2) $$+\int p_oG_v(p_o/f',f)wdx + \int\int qp_mG_v(p_m/f',f)wdxdm = 0$$

Since w' can be very large compared to w lines (6.1) and (6.2) must vanish separately. The vanishing of (6.1) implies in turn that

(7) $p_o^2G_u(p_o/f',f) + \int qp_m^2G_u(p_m/f',f)dm = 0$

(8) Setting $G_u(u,v) = r(u,v)/u^2$ we obtain

(9) $r(p_o/f'f) + \int qr(p_m/f',f)dm = 0.$

Note that a variation $\Delta p(x,y) = \epsilon h(x)k(y)$, where $\int hdx = \int kdx = 0$, is an admissible variation of p(x,y) providing h and k are appropriately bounded. Such a variation produces the following variations in the associated distributions:

(10) $$\begin{cases} \Delta p_o(x) = 0 \\ \Delta q(m) = 0 \\ \Delta p_m(x) = \epsilon h(x)k(m)/q(m). \end{cases}$$

Subjecting the respective quantities of (9) to the variations prescribed by (10) yields

(11) $\int r_u(p_m/f',f)k(m)dm = 0.$

Due to the arbitrariness of k(m) (11) implies that

(12) $r_u(p_m/f',f)$ = function independent of m.

Subjecting (12) to another variation of the type (10) we obtain

(12) + $\triangle$(12) = independent of m; therefore

(13) $f' \triangle (12) = 6\, r_{uu}(p_m/f',f)h(x)k(m)/q(m)$ = indep. of m.

From the arbitrariness of k(m) it clearly follows that

(14) $r_{uu}(u,v) = 0$

Combining (14) and (8) gives

(15) $G(u,v) = a(v)\ln u+b(v)/u+c(v)$

for some functions a(v), b(v), c(v).

Returning now to (6.2) we see that it implies

(16) $p_o G_v(p_o/f',f) + \int qp_m G_v(p_m/f',f)dm = 0.$

(17) Let $G_v(u,v) = s(u,v)/u$. Then (16) becomes

(18) $s(p_o/f',f) + \int qs(p_m/f'f)dm = 0.$ As (18) is of exactly the same form as (9) it similarly implies that

(19) $s_{uu}(u,v) = 0.$

Combining (19) and (17) yields

(20) $G_v(u,v) = A(v) + B(v)/u.$

But (15) implies

(21) $G_v(u,v) = a'(v)\ln u+b'(v)/u+c'(v).$

Combining (20) and (21):

(22) $a'(v) = 0.$

Thus

(23) $G(u,v) = \text{const} \cdot \ln u+b(v)/u+c(v)$ and

(11) $\int r_u(p_m/f',f)k(m)dm = 0.$

Due to the arbitrariness of k(m) (11) implies that

(12) $r_u(p_m/f',f)$ = function independent of m.

Subjecting (12) to another variation of the type (10) we obtain

$(12) + \triangle(12)$ = independent of m; therefore

(13) $f'\triangle(12) = \epsilon\, r_{uu}(p_m/f',f)h(x)k(m)/q(m)$ = indep. of m.

From the arbitrariness of k(m) it clearly follows that

(14) $r_{uu}(u,v) = 0$

Combining (14) and (8) gives

(15) $G(u,v) = a(v)\ln u+b(v)/u+c(v)$

for some functions a(v), b(v), c(v).

Returning now to (6.2) we see that it implies

(16) $p_oG_v(p_o/f',f) + \int qp_mG_v(p_m/f',f)dm = 0.$

(17) Let $G_v(u,v) = s(u,v)/u$. Then (16) becomes

(18) $s(p_o/f',f) + \int qs(p_m/f'f)dm = 0.$ As (18) is of exactly the same form as (9) it similarly implies that

(19) $s_{uu}(u,v) = 0.$

Combining (19) and (17) yields

(20) $G_v(u,v) = A(v) + B(v)/u.$

But (15) implies

(21) $G_v(u,v) = a'(v)\ln u+b'(v)/u+c'(v).$

Combining (20) and (21):

(22) $a'(v) = 0.$

Thus

(23) $G(u,v) = \text{const} \cdot \ln u+b(v)/u+c(v)$ and

(24) $F(u,v) = \text{const} \cdot u\ln u + uc(v) + b(v).$

Substituting (24) in (1):

(25) $$R/\text{const} = \int p_o \ln p_o dx - \iint q p_m \ln p_m dx dm + \int (p_o - \int q p_m dm) c(x) dx.$$

But $\int q(m) p_m(x) dm = \int q(m) p(x,m)/q(m) dm = p_o(x).$

Therefore the last integral of (25) vanishes, making it

(26) $$R/\text{const} = \iint p(x,y) \ln p_o dx dy - \iint p(x,m) \ln \frac{p(x,m)}{q(m)} dx dm =$$

$$= - \iint p(x,y) \ln \frac{p(x,y)}{p_o(x) q(y)} dx dy,$$

as was to be proved.

Arbitrarily setting the const of (26) equal to -1, we obtain four equivalent representations of R:[4]

$$R = \iint p(x,y) \ln \frac{p(x,y)}{p(x)q(y)} dx dy =$$

$$= -\int p(x) \ln p(x) dx - \int q(y) \ln q(y) dy = \iint p(x,y) \ln p(x,y) dx dy =$$

$$= -\int p(x) \ln p(x) dx + \iint p(x,y) \ln p_y(x) dx dy =$$

$$= -\int q(y) \ln q(y) dy + \iint p(x,y) \ln q_x(y) dx dy.$$ [5]

In the discrete case the formulas reduce to

$$R = \sum_{i,j} P(i,j) \ln \frac{P(i,j)}{P(i)Q(j)} =$$

$$= -\sum_i P(i) \ln P(i) - \sum_j Q(j) \ln Q(j) + \sum_{i,j} P(i,j) \ln P(i,j) =$$

[4] Sometimes it is convenient to use the base 2 for the logarithm; in that case const = $-1/\ln 2$.

[5] More compact representations of these relations will be given in section IV.

$$= -\sum_{i} P(i)\ln P(i)+\sum_{i,j} P(i,j)\ln P_{j}(i) =$$

$$= -\sum_{j} Q(j)\ln Q(j)+\sum_{i,j} P(i,j)\ln Q_{i}(j).$$

3. Information Rate of a Source

The information rate of a source is defined in terms of the per-symbol rate at which information produced by the source is capable of being received.

Consider the expression for information receipt rate

$$R = -\int p(x)\ln p(x)dx + \int\int p(x,y)\ln p_{y}(x)dxdy$$

derived in the last paragraph. On first thought one would be inclined to define the information rate of the source as the value of R that would be obtained if symbols emitted by the source (described by the distribution p(x)) were received in the absence of noise, that is with $p_y(x) = \delta(x-y)$. In general, however, the right side of the above expression for R will become infinite for this type of transmission.(6) In order to make it unnecessary to set $p_y(x) = \delta(x-y)$ information rate of a source will be defined <u>relative to a fidelity criterion</u>.

Let $\rho(x,y)$ be a continuous function of x and y whose value is a measure of the punishment meted out if the symbol y is received as a result of the source emitting the symbol x. (Presumably $\rho(x,x) = 0$; that is, there is no punishment if the emitted symbol is also the received symbol.)

(6) The expression for R will become infinite if $p_y(x) = \delta(x-y)$, providing the source is not completely discrete.

The average amount of punishment per transmitted symbol is

$$v = \int\int p(x,y)\, \rho\, (x,y)dxdy.$$

Let us call v the quality of the system [7]. The information rate of the source with given p(x) relative to the fidelity criterion $v = \int\int p \rho\, dxdy$ is defined as the minimum information-receipt rate necessary to preserve the quality v. The minimum is taken over all possible noise conditions:

$$R_{\text{source}} = \min_{q_x(y)} \int\int p(x,y) \ln \frac{p(x,y)}{p(x)q(y)} dxdy \text{ with}$$

$$\int\int p(x,y)\, \rho\, (x,y)dxdy = v = \text{const.}$$

For discrete transmission systems the rate of the source is

$$R_{\text{source}} = \min_{Q_i(j)} \sum_{i,j} P(i,j) \ln \frac{P(i,j)}{P(i)Q(j)} \quad \text{with}$$

$$\sum_{i,j} P(i,j)\, \rho\, (x_i, y_j) = v = \text{const.}$$

In this case it is possible to obtain the rate of the source in an absolute sense by requiring perfect fidelity; i.e., by requiring $P(i,j) = P(i)\delta_{ij}$. This means $Q_i(j) = \delta_{ij}$, and $Q(j) = P(j)$. Therefore

$$R_{\text{source absolute}} = - \sum_{i} P(i) \ln P(i)$$

In order to clarify the remarks of page (16) we can think of the case where the source symbols have a continuous distribution as a limiting case of the discrete situation, with the help of the substitution $P(i) = p(x_i)\Delta x$. The

[7] "Infidelity" would be a better word.

fact that $R_{s.\ ab.}$ becomes infinite as $\overline{\Delta x} \to 0$ indicates that from the absolute point of view (i.e. without reference to a fidelity criterion) continuous sources have an infinite information rate per emitted symbol.

A formal, although not very useful, expression for the rate can be obtained by carrying out the minimization procedure indicated in the definition. We will carry it out for the discrete case, and then state the analogous results for the more general case.

It is desired to minimize

$$(1) \qquad - \sum_j Q(j) \ln Q(j) + \sum_{i,j} P(i,j) \ln Q_i(j) =$$

$$= D = - \sum_{i,j} P(i) Q_i(j) \ln \sum_m P(m) Q_m(j) + \sum_{i,j} P(i) Q_i(j) \ln Q_i(j)$$

for given P(i)'s over all $Q_i(j)$, subject to

$$\begin{cases} (2.1) \quad E_o = \sum_{i,j} P(i) Q_i(j) \rho(i,j) = v = \text{const} \\ (2.2) \quad E_i = \sum_j Q_i(j) = 1 \quad (i = 1,2,\dots) \end{cases}$$

According to the method of Lagrangian multipliers, the minimum of D will be obtained when

$$(3) \qquad \frac{\partial D}{\delta Q_k(1)} + \sum_i \lambda_i \frac{\partial E_i}{\delta Q_k(1)} = 0 \qquad (k,1 = 1,2,\dots)$$

where the λ_i are adjusted to satisfy (2). From (1):

$$(4) \qquad \frac{\partial D}{\delta Q_k(1)} = P(k) \log \left[\frac{Q_k(1)}{\sum_m P(m) Q_m(1)} \right] = P(k) \log(P_1(k)/P(k))$$

From (2)

$$(5) \qquad \sum_i \lambda_i \frac{\partial E_i}{\delta Q_k(1)} = \lambda_o P(k) \rho(k,1) + \lambda_k$$

Putting (4), (5) into (3):

(6) $P_1(k) = P(k) \exp\left[-\lambda_o \rho(k,1) - \lambda_k/P(k)\right] =$

$= A(k) \exp(-\lambda_o \rho(k,1)).$

where the A(k)'s are determined as functions of λ_o by

(7) $\sum_k A(k)\exp(-\lambda_o \rho(k,1)) = 1 \qquad (1 = 1,2,\ldots)$

under the restriction that $A(k) = 0$ if $P(k) = 0$. λ_o is adjusted to satisfy (2.1).

Note that (7) determines A(k) as the solution of a non-homogeneous system of linear algebraic equations. Unfortunately, D cannot be evaluated directly from a knowledge of P(k), and $P_1(k)$. It is first necessary to evaluate some one of the quantities, P(i,j), $Q_1(j)$, or Q(j), and this requires the solution of a system of linear algebraic equations. This is the reason why the expression (6) has only limited practical value for evaluating specific information rates of sources.

In the special case where $P(k) \neq 0$ for any $k = 0, \pm 1, \pm 2, \ldots \to \pm\infty$, and $\rho(i,j) = \tau(i-j)$ the solution of (7) is

(8) A(k) = indep of k $= \alpha(\lambda_o)$, making the a-posteriori probability that i was transmitted an exponentially decaying function of the error metric $\tau(i-j)$:

(9) $P_j(i) = \alpha(\lambda_o)\exp(-\lambda_o \tau(i-j)).$

Solutions for the continuous case are obtained by replacing the probabilities in the above formulas by the corresponding distributions, and the sigma signs by integrals. This transforms the linear algebraic equations into integral equations.

4. Channel Capacity

In the theory of information the ability of a channel to transmit information produced by a source to the receptor is described by a quantity known as channel capacity. The concept of the channel is needed to take into account the fact that the symbols emitted by the source are not necessarily the symbols arriving at the receiver. Loosely speaking, therefore, the channel is that part of a 2-point one-way communication system where the noise occurs.

Since the physical nature of transmission links is often of such a nature as to limit the number of symbols per second that can be transmitted through it, channel capacity will be defined on a per-unit time, instead of a per-unit symbol basis. Let M be the number of symbols per second, and let $q_x(y)$ be the transition probability distribution describing the noise; then the channel will be operating at its "capacity" C when the source is properly "matched" to the channel:

$$C = \max_{p(x)} M \iint p(x,y) \ln \frac{p(x,y)}{p(x)q(y)} dxdy.$$

The right side of the above equation will be maximized for some distribution p(x). The channel will be able to transmit the maximum amount of information per second if it is fed by a source governed by the distribution p(x). This concept is valuable because it is always possible to code the output of a source to give the encoded symbols an arbitrary given distribution.[8] It should be noted that under certain conditions it may be desirable to maximize the channel over only a restricted class of permissible p(x)'s [9]. In that case the channel capacity is relative to the permissible set of input symbols.

[8] Details will be given in a later section.

[9] For instance we may permit only p(x)'s with a given second moment (a power limitation)

5. Example: Capacity of a Band-limited Channel with White Noise

The restriction of band limitation of say, from 0 to W cycles per second, means that the spectra of both the function emitted by the source and the noise are limited to the interval (0,W). Such functions can be written in the form

$$(1)\qquad f(t) = \sum_{k=-\infty}^{\infty} f(k/2W)\,\Phi(t-k/2W)$$

where $\Phi(t) = \dfrac{\sin 2\pi Wt}{2\pi Wt}$. [10]

Since $\int_{-\infty}^{\infty} \Phi(t-m/2W)\Phi(t-n/2W)dt = \delta_{mn}/2W$ for integral m and n

$$(2)\qquad \text{power of } f(t) = \lim_{T\to\infty} (1/2T)\int_{-T}^{T} f^2(t)dt =$$

$$= (1/2W) \lim_{T\to\infty} (1/2T) \sum_{k=-2WT}^{2WT} f^2(k/2W) =$$

$$= \lim_{n\to\infty} (1/2n) \sum_{k=-n}^{n} f^2(k/2W) = \overline{f^2(k/2W)}.$$

From (1) we see that f(t) can be thought of as produced by a source that emits a pulse shape Φ with amplitude $x_k = f(k/2W)$ at instants of time 1/2W seconds apart. If the x_k are picked from a distribution p(x) then (2) indicates that the power of f(t) will be the second moment of p:

$$(3)\qquad \text{power of } f(t) = \int_{-\infty}^{\infty} x^2 p(x)dx$$

A representation of band-limited white noise of power N can be obtained by means of the concept that it results when a large number of correspondingly

[10] This formula can be obtained by expanding the spectrum of f(t) in a Fourier series, and then using the Fourier integral representation for f(t).

band-limited functions are added at random. Let $g(t)$ represent the noise, and $f_i(t)$ typify the functions that add to produce the noise. Evidently $g(k/2W) = \sum_i f_i(k/2W)$. By the central limit theorem $x_k = g(k/2W)$, will have a Gaussian distribution, which by (3) must have a second moment equal to N:

(4) $r(x) = (1/\sqrt{2\pi N})\exp(-(x^2/2N))$=distribution of g. x_k's corresponding to two different values of k are independent.

In the preceding paragraphs we have spoken of $f(k/2W)$ as the coefficient of the elementary pulse shapes that make up the signal. It is apparent that the pulse shapes themselves merely act as carriers. A model for the entire process is obtained if we consider the source to emit a sequence of real numbers picked from a distribution, say $p(x)$, with second moment S (the power of the source). These real numbers are the "symbols" produced by the source, the symbol-producing rate being

(5) $M = 2W$ symbols per second.

The effect of the noise is to add a second sequence term by term to the source sequence, with the terms of the second sequence picked at random from the distribution (4).

Due to the additive nature of the noise

(6) $p(x,y) = p(x)r(y-x)$. Therefore

(7) $$R = -\int q(y)\ln q(y)dy + \iint p(x,y)\ln q_x(y)dxdy =$$

$$= -\int q(y)\ln q(y)dy + \int r(z)\ln r(z)dz.$$

By (4)

(8) $$\int r(z)\ln r(z)dz = (-\tfrac{1}{2})\ln(2\pi eN)$$

The problem now is to maximize $-\int q(y)\ln q(y)dy$ over all $p(x)$. Since the total power at the receiver is S+N the second moments of $p(x)$ and $q(y)$ must be fixed at S and S+N respectively. If the maximization of (7) were to be carried out over all possible q(y) instead of p(x) (as it actually must be) we could use the easily proved theorem that

(9) $\max\limits_{q(y)} - \int q(y) \ln q(y) dy$ with $y^2 q(y) dy$ = fixed is obtained when q(y) is Gaussian; i.e., $\max\limits_{q(y)} - \int q(y) \ln q(y) dy = (\tfrac{1}{2}) \ln 2\pi e(S+N)$.

It is, however, certainly true in view of the preceding that

(10) $\left[\max\limits_{p(x)} - \int q(y) \ln q(y) dy \text{ with } x^2 p(x) dx = S \right] \leq$ value obtained when q(y) were Gaussian = $(\tfrac{1}{2}) \ln 2\pi e(S+N)$.

Now from the equation

(11) $q(y) = \int p(x,y) dx = \int p(x) r(y-x) dx$

and the fact that r(z) is Gaussian it happens to follow fortuitously that it is possible to make q(y) Gaussian by taking p(x) Gaussian:

(12) $p(x) = (1/\sqrt{2\pi S}) \exp(-(x^2/2S)$.

Therefore the inequality of (10) becomes an equality, and we have, combining (5), (7), (8), (10)

(13) $C = W \log(S+N/N)$

as the capacity of the model channel. But the model channel was obtained from the real channel by a relabeling process, namely by relabeling sequences of pulses as sequences of real numbers. Since (7) was derived under the postulate that it is invariant under relabeling [11] (13) is also the capacity of the real channel. According to (12) the channel

[11] In the derivation of R it was actually only postulated that invariance held if real numbers were relabeled as other real numbers, and only one-dimensional distributions were considered. If the distributions had been taken multi-dimensional the above statement would have followed rigorously.

is maximized when the source emits white noise.

IV. Properties of Information Rates.

1. Summary

The various information rates are expressed in terms of the entropy and conditional-entropy functions which are defined and studied. It is shown that the number of highly probable long sequences of symbols emitted by a source is closely related to the information rate of the source.

In the last paragraph the fundamental theorem for 2-point, 1-way communication is derived. This states that with a proper en- and decoding equipment the output of a source can always be transmitted in the presence of noise, without error, at a rate determined by the channel capacity and the information rate of the source.

2. Entropy Functions

Let the entropy G of a distribution function f(x) be defined as

(1) $G = -\int f(x)\log f(x)dx.$

Therefore the entropy of the source is

(2) $G(S) = -\int p(x)\log p(x)dx$, where S stands for Source.

and the entropy of the received symbols is

(3) $G(T) = -\int q(x)\log q(x)dx$, where T stands for Receiver.

We also define the mixed or relative entropies

(4) $G_T(S) = -\int\int p(x,y)\log p_y(x)dxdy$ and

(5) $G_S(T) = -\int\int p(x,y)q_x(y)dxdy.$

(4) is spoken of as the "entropy of S knowing T", and (5) the "entropy of T knowing S". By thinking of the pair (x,y) as one symbol, we can extend (1) to cover the concept of joint entropy:

(6) $G(T,S) = -\iint p(x,y)\log p(x,y)dxdy.$

It can easily be shown that

(7) $G(T,S) = G(T) + G_T(S) = G(S) + G_S(T) = G(S,T).$

It also follows from (4) and (5) that if x and y are independent then

(8) $G_S(T) = G(T)$ and $G_T(S) = G(S).$

Thus if T and S are independent

(9) $G(T) = G(T) + G(S).$

For the discrete case it is desirable to introduce analogous quantities:

(10) $H(S) = -\sum_i P_i \log P_i$

(11) $H(T) = -\sum_j Q_j \log Q_j$

(12) $H_S(T) = -\sum_{i,j} P(i,j)\log Q_i(j)$

(13) $H_T(S) = -\sum_{i,j} P(i,j)\log P_j(i)$

(14) $H(T,S) = H(S,T) = H_T(S) + H(T) = H(S) + H_S(T) =$

$$= -\sum_{i,j} P(i,j)\log P(i,j).$$

It is possible to express information rate in terms of the quantities defined above. The expression is in the continuous case

(15) $R = G(S) - G_T(S) = G(T) - G_S(T).$

According to III, 5, (7) when the noise symbols are "additive" and independent of the source symbols (15) becomes

(16) $R = G(T) - G(N)$

where $G(N) = -\int r(x)\log r(x)dx$ is the entropy of the noise.

For the discrete case (15) degenerates into

(17) $R = H(S)-H_T(S) = H(T)-H_S(T).$ [12]

The reader may have noticed that G(S) is actually the uncertainty function U(p) arrived at in III, 2, (24). (III,2,(25) shows that b(v), and c(v) appearing in III,2,(24) are irrelevant.). In other words (1) is a measure of the uncertainty associated with the distribution f(x) [13]. More generally, for instance, $G_T(S)$ is the uncertainty of the symbol at S, knowing the symbol at T. With this interpretation we can easily "derive" relation (16). One need merely note that $G_S(T)$, the uncertainty of what was received, knowing what was emitted, is, in the case of independent ad-

[12] This is true even though none of the G's individually degenerate into the corresponding H's. A G can be thought of as differing from the corresponding H by an infinite additive constant, these constants cancelling out when the difference of two G's or H's is taken.

[13] In the discrete case $H = -\sum_I f_i \log f_i$ is a measure of the uncertainty associated with the probabilities $(f_1,f_2,\ldots,f_n)$ in quite an absolute sense. It can be shown that H will be a maximum when all the f's are equal, and it is obvious that H is zero if and only if one of the f's is unity and all others vanish.

ditive noise, the uncertainty of received signal plus noise with the emitted signal known, this being simply the uncertainty of the noise, G(N). Substituting $G_S(T) = G(N)$ into (15) yields (16).

3. Laws of Long Sequences

This paragraph lists some properties of long sequences of output symbols from a discrete source, transmitted over a noisy channel.

Law I:

Every emitted sequence of length L>>l symbols has w.h.p.[14] exp $(H_S(T)L)$ received sequences of length L as possible consequences.

Proof:

If the sequence $(x_1, x_2, \ldots x_L)$ is emitted it will w.h.p. contain the symbol x_i P(i)L times (i = 1,2,...,n where n is the number of possible symbols). The emitted message can therefore be considered to consist of n (possibly interlaced) blocks of P(i)L symbols each. Each such i'th block will produce a block of P(i)L received symbols, containing the j'th symbol $Q_i(j)P(i)L = P(i,j)L$ times. The probability of a particular block of received symbols is therefore w.h.p.

$$\prod_{j=1}^{n} \left[Q_i(j)\right]^{P(i,j)L}$$

[14] The phrases, "with high probability" (w.h.p.), and "with probability zero" (w.p.z.) are to be interpreted as meaning that the probabilities referred to approach 1 and 0 respectively as L→∞. Sometimes when elements of a set V are w.h.p. also in the set W, we will say "All elements of V are in W".

The probability of the entire received sequence is therefore

$$\prod_{i,j=1}^{n} \left[Q_i(j)\right]^{P(i,j)L} = \exp(-H_S(T)L).$$

The desired result now follows because each of the h.p. received messages are equally likely.

Corollary I: (The dual of Law I) [15]

Corollary II:

The number of h.p. emitted sequences of length L is exp(H(S)L).

Corollary III: (The dual of Corollary II)

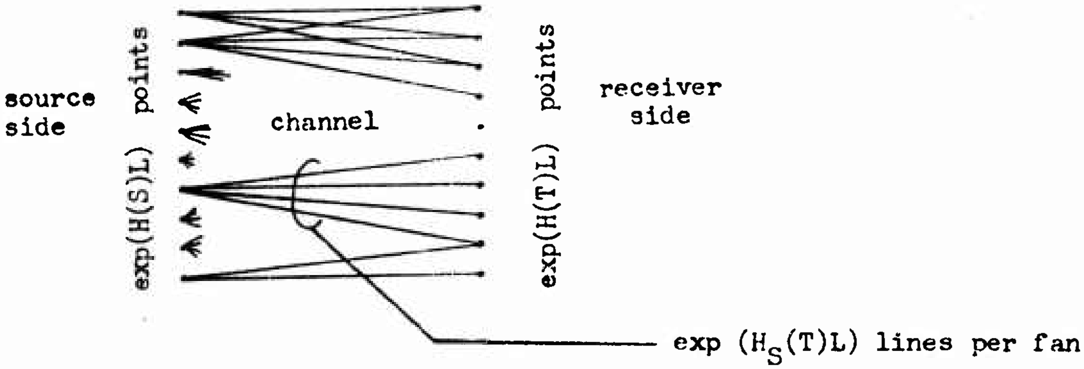

Fig. 3 Transmission of Sequences of L>>1 Symbols Over Noisy Channel

Figure 3 illustrates the situation occurring when long sequences are transmitted over a noisy channel. Received and transmitted sequences are represented as points on the right and left respectively. Each fan shows

[15] The dual is obtained by interchanging the words "source" and "receiver", the symbols i and j, P(i) and Q(j), and $Q_i(j)$ and $P_j(i)$.

how many received sequences a given emitted sequence can result in. Since the fans, in general, overlap the receiver cannot know exactly what was transmitted. However, if only a few of the possible points on the left were actually used to represent messages it is conceivable that the resulting fans might not overlap. A necessary condition for this to occur is certainly that no more than

$$\exp(H(T)L)/\exp(H_S(T)L) = \exp\left[(H(T)-H_S(T))L\right]$$

of the points on the left are used.

<u>Law II</u>:

If less than $\exp\left[(H(T)-H_S(T)-\delta)L\right]$ $(\delta > 0)$ points are selected at random from the source side of the channel the resulting fans will overlap w.p.z. [16]

Proof:

Suppose $\exp\left[(H(T)-H_S(T)-\delta)L\right]$ points are selected at random from the left, making the probability that a particular point is a selected point

$$\exp\left[(H(T)-H_S(T)-\delta)L\right]/\exp(H(S)L) = \exp(-H_T(S)L-\delta L).$$

No two fans emanating from selected points will overlap if any given point on the right cannot be "caused" by more than one selected point. Each point on the right can a-priori (i.e. if no selection of points on the left were used) be caused by $\exp(H_T(S)L)$ left points. The probability P that at least two of these points are selected points is less than 1-A, where A is the probability that none of the $\exp(H_T(S)L)$ points is a

[16] This is of course a much weaker theorem than one giving specific instructions as how to pick the points on the left to get the minimum possible overlap. Stronger theorems have been obtained for specific channels. See Refs. 2, 11, 20.

selected point.

$$A = \left[1-\exp(-H_T(S)L-\delta L)\right]^{\exp H_T(S)L} \longrightarrow 1 \quad \text{as} \quad L \longrightarrow \infty .$$

Therefore $P \rightarrow 0$ as $L \longrightarrow \infty$; q.e.d.

Corollary IV: (The dual of Law II)

Corollary V:

If $\exp\left[(H(T)-H_S(T)+\delta)L\right]$ points ($\delta > 0$) are selected at random from the receiver side of a channel the fans emanating from them[17] cover w.h.p. all the $\exp(H(S)L)$ points at the source side.

Proof:

By Corollary IV if $\exp\left[(H(T)-H_S(T)-\delta)L\right]$ were selected at the right there would be no overlapping of fans so that $\exp\left[(H(S)-\delta)L\right]$ of the $\exp(H(S)L)$ points on the left are covered. The desired result follows easily.

4. Fundamental Theorem for Transmission over Noisy Channel

Theorem II:

It is possible to match a source producing R units of information per symbol (relative to a fidelity criterion) to a channel of capacity C by means of coders in such a way that if less than C/R symbols per unit time are transmitted the transmission quality will satisfy the fidelity criterion.

Proof:

The proof will consist in describing various coders and decoders

[17] These fans originate at the right and spread out toward the left. They are the duals of the ones shown in Fig. 3, and indicate the number of emitted sequences that could have caused the received sequence from which they emanate.

by means of which it is possible to attain the objective announced. The pertinent block diagram is shown in Figure 4.

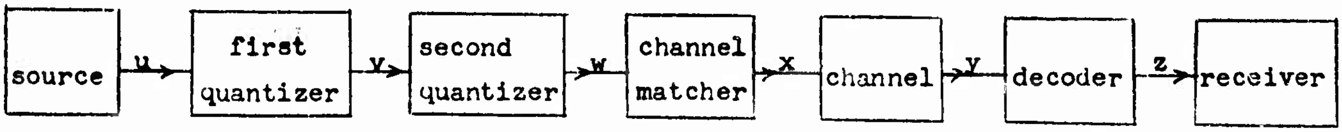

Fig. 4 Block Diagram for Transmission System With Coding Equipment

(a) The first quantizer.

The FQ (first quantizer) is not needed if the source is discrete. If the source is not discrete the FQ is used (purely for the sake of mathematical convenience) to quantize it into very fine but discrete levels. It is intuitively obvious that very fine quantizing has no appreciable effect on the rate of the source. Thus the information rates at u and v are the same.

(b) The second quantizer

The SQ (second quantizer) is not needed if the fidelity criterion requires perfect transmission. If, on the other hand, it is not dictated that the symbols at v be transmitted with perfect fidelity (i.e. if the rate R at v is not the absolute rate[18] at v) the SQ quantizes the symbols at v in such a way that the quantized symbols put out at w have an absolute rate R. (Therefore from w onward there must be no more distortion in the transmission system.)

Fundamentally, the SQ operates by first ascertaining which of an equivalent number of classes a given sequence v belongs to, and then transmitting a code number for that class; for instance, the code number might simply be the "central" sequence of the particular class. Specifically, these classes and their code symbols can be determined with

[18] Cf. III, 3 as reference for this section

the help of Corollary V as follows: We consider v the "emitted" symbols and w the "received" symbols. With this notation the rate of information per symbol at v

$$R = \min_{Q_i(j)} \sum_{i,j} P(i,j) \ln \frac{P(i,j)}{P(i)Q(j)}$$

with $\sum_{i,j} P(i,j)\rho(i,j) = \text{const.}$

Suppose P'(i,j) is the P(i,j) for which the minimum in the above definition of R is obtained. Select, according to the method of Cor. 5, exp(RL+δL) points on the "receiver" side of the transmission system obtained with P(i,j) = P'(i,j). The SQ is to be constructed so that it will use a particular selected point as the code for the class of points caught in the fan emanating from the selected point. The SQ obtained by this construction satisfies the fidelity criterion, and has the property that, looking into its output terminal w, we see a source of absolute rate R units of information per symbol.

(c) The channel matcher [19]

The CM (channel matcher) is, as its name indicates, a device for encoding the symbols arriving at w into symbols that are best able to combat the noise present in the channel. Since it must be possible to recover the symbols w with perfect accuracy at the receiver, the CM must be a one-to-one coder; that is, it must be reversible.

For purposes of discussing the CM consider x to be the "emitted" symbols and y the "received" symbols. Assume that the symbols x are

[19] Cf. III, 4 as reference for this section

produced according to the distribution $P''(i)$ and transmitted at the rate M' symbols per second, where $P''(i)$ and M' maximize the channel, i.e. assume that

$$C = \max_{P(i)} M \sum_{i,j} P(i,j) \ln \frac{P(i,j)}{P(i)Q(j)} \text{ , where } M \text{ is the number of}$$

symbols per second, is obtained for $P(i) = P''(i)$ and $M = M'$. If $H(S'') = -\sum_i P''(i) \log P''(i)$, then, according to Cor. II, if the channel is operated with $P(i) = P''(i)$ there will be $\exp(H(S'')M'T)$ possible h.p. long sequences of length T seconds at point x. According to Law II if less than $\exp(CT)$ of these sequences are used as messages the "receiver" at y will be able to ascertain exactly which message was sent. The problem for the CM is therefore only to code the symbols arriving at w into the $\exp(CT)$ symbols that are available for transmission without error. Since $\exp(RL)$ symbols of length L arrive at w such coding will obviously be possible if and only if $RL < CT$, i.e. if and only if no. of symbols per sec. produced by source $= L/T < C/R$,

where R is the rate of the source, and C the capacity of the channel.

(d) The decoder

The decoder performs the operation inverse to the CM, so that we end up with the same symbols at z that originated at w.

V. Prediction of Time Series.

I. Summary

This section outlines the philosophy behind the prediction problem for time series chosen from an ensemble of time series for which a certain set of multi-dimensional set of probability functions exists, and is a-priori known.

2. Multi-dimensional Probability Distributions

(1) Let $z_1, z_2, \ldots, z_1, \ldots$ be a typical time series of an ensemble of time series.

(2) Let $V_k(y_1, y_2, \ldots, y_k)dy_1 dy_2 \ldots dy_k$ $(k = 1,2,\ldots)$ be the probability that if a block of k consecutive z's, beginning with z_{j+1}, is selected at random from (1) the z's will lie in the region

$$y_i \leq z_{i+j} \leq y_i + dy_i \quad (i = 1,2,\ldots,k),$$

relation (2) being postulated to hold independent of j, and independent of which particular time series is chosen from the ensemble.

(3) Let $W_k(y_1, y_2, \ldots, y_k; y_{k+1})dy_{k+1}$ $(k \geq 1)$

be the probability that z_{j+k+1} will lie in the region

$$y_{k+1} \leq z_{j+k+1} \leq y_{k+1} + dy_{k+1} \quad \underline{\text{if}} \quad z_{i+j} = y_i \quad (i = 1,2,\ldots,k).$$

If we arbitrarily set

(4) $W_o(y) = V_1(y)$ it follows that the V and W functions are related through

(5) $V_k(y_1, y_2, \ldots, y_k) = V_{k-1}(y_1, y_2, \ldots, y_{k-1}) W_{k-1}(y_1, y_2, \ldots, y_{k-1}; y_k)$

if $k \geq 2$.

To obtain a complete statistical description of the stochastic process in question all the W_k's (or what is easier experimentally, all the V_k's) must be found. In most practical cases there will be no "influence" extending further than, say j signals. This simply means that

$$(6) \qquad W_k(y_1,y_2,\ldots,y_k;y_{k+1}) = F(y_{k-j+1},y_{k=j+2},\ldots,y_k,y_{k+1})$$

for k larger than some sufficiently large j.

3. Predictability

Loosely speaking, the more redundant a time series is, i.e. the less uncertainty there is about the next signal, knowing a certain number of previous signals, the more easily predictable will the time series be. Some of the terms used in the preceding sentence can be defined exactly.

(a) k-derived uncertainty = H_k

Analogously to III, 3 let $\rho(x,z)$ measure the punishment meted out if a signal x is predicted to be the symbol z, and let v measure the amount by which two signals must differ in order to become practically distinguishable.

Define R_k to be the rate of a mathematically artificial source that produces symbols x independently according to the distribution $p(x) = W_k(\vec{y};x)$ relative to the criterion

$$\iint p(x,z)\rho(x,z)dxdy = v; \text{ i.e.}$$

$$(8) \qquad R_k(\vec{y}) = \min_{q_x(z)} \iint p(x,z)\log\frac{p(x,z)}{p(x)q(z)}\, dxdz$$

with

$$\int\int p(x,z)\,\rho(x,z)dxdy = v,$$

where $p(x,z) = p(x)q_x(z)$, $q(z) = \int p(x,z)dx$.

Let H_k be the average of R_k over the possible $y_1, y_2, \ldots, y_k$:

(9) $$H_k = \int R_k(\vec{y})V_k(\vec{y})d\vec{y} \quad \text{(a k-fold integral)}$$

H_k is evidently the average amount of information needed to specify a signal if the previous k signals are known. Thus it is a measure of the uncertainty with which we know what a signal will be if we know the previous k signals.

(b) redundancy

(10) $$\text{Let } H_\infty = \lim_{k \to \infty} H_k$$

The redundancy of the time series can be defined as

(11) $$\mu = 1 - H_\infty / H_o$$

If successive symbols are independent we will have

(12) $W_k(\vec{y};x) = W_o(x)$ (all k), and therefore

(13) $H_k = H_o$, so that $\mu = 0$.

If the next signal is, on the other hand, completely determined once a sufficient number of preceding signals are known $\mu = 1$.

It should not be forgotten that, in general, the redundancy is relative to the punishment function $\rho(x,y)$ and the distinguishability criterion v.

4. The Mechanism of Prediction

(a) choice of the punishment function $\rho(x,y)$

In order to design a predictor, it is, in principle, necessary to first specify the function of two variables $\rho(x,y)$ that measures the punishment meted out if the next signal is predicted to be "y" but actually turns out to be "x". Although the choice of $\rho(x,y)$ will be dictated by the application of the predictor, its selection is ultimately a psychological problem.

The predictor is designed so as to minimize the expected value of $\rho(x,y)$.[20]

A common choice for $\rho(x,y)$ is

$$(13) \qquad \rho(x,y) = f(x-y),$$

in which case the punishment depends only on the error. For instance the (for reasons of analytic simplicity) popular least-squares criterion

$$(14) \qquad \rho(x,y) = f(x-y) = (x-y)^2$$

is of this type.

(b) unrestricted versus restricted prediction

The most general form of predictor is a computer which, on the basis of all information at hand, predicts a signal so as to minimize the expected value of the punishment. According to the

[20] This might be called a "rational" prediction criterion. It is conceivable (in fact the motivation for gambling) to have the punishment function dependent not merely on x and y, but also on the probability that x will occur. Maximizing the expected value of ρ in such a case would amount to an "irrational" criterion. With irrational criteria it may be desirable for the predictor to play a mixed strategy against the time series, i.e. to "toss a coin". With rational criteria it is pointless to play a mixed strategy.

aforelying formulation of the prediction problem the computer can do this if it remembers all previous signals, computes the a-priori distribution of the signal to be predicted according to the W_k functions, and then minimizes the expected value of ϱ . A process such as this can be called "unrestricted" prediction.

On the other hand, consider the case where, for practical reasons, it is necessary to place theoretically artificial restrictions on the storage mechanism and permissible operations assigned to the computer. When this situation arises we speak of "restricted" prediction. An example is the case of so-called linear prediction where the computer is permitted to evaluate only linear combinations (with permanently fixed coefficients) of amplitudes of past signals. Although a time series of redundancy $\mu = 1$ is perfectly predictable in the unrestricted sense it may not be so in the restricted sense.

The more restricted a predictor is the larger the error of prediction will be. On the other hand the predictor may be applicable to a larger ensemble of time series if it is more restricted. Thus restriction of predictors has among other things the effect of trading error for versatility.

5. Examples

(a) sine wave samples

Consider a source producing signals z_i at discrete time instants (1 = 1,2,...) according to the recursion formula

(15) $f(z_i) = \sin i \qquad (i = 1,2,\ldots)$

It can be shown that the points i mod (2π) cover the interval

$(0,2\pi)$ in an everywhere dense fashion and in such a way that the probability that i mod(2π) is between two real numbers exists and is flat over $(0,2\pi)$. Therefore the distribution $W_o(z)$ for $f(z_i)$ is the same as that obtained for sin t if t is picked at random from a distribution flat over $(0,2\pi)$. This latter is [21]

(16) $$W_o(z) = \begin{cases} 1\Big/\left(\pi\sqrt{1-z^2}\right) & \text{if} \quad |z| < 1 \\ 0 & \text{if} \quad |z| \geq 1 \end{cases}$$

(17) Let a_k = sink, b_k = cosk.

Then if a given signal has the amplitude z_j it is equally likely that z_{j+1} be

(18) $$b_1 z_j + a_1 \sqrt{1-z_j^2} \quad \text{or} \quad b_1 z_j - a_1 \sqrt{1-z_j^2} \, .$$ Thus

(19) $$W_1(y_1;y_2) = (1/2)\delta\left[y_2-(b_1y_1+a_1\sqrt{1-y_1^2}\right] + (1/2)\delta\left[y_2-(b_1y_1-a_1\sqrt{1-y_1^2})\right] .$$

If two or more consecutive samples are known all future samples can be predicted perfectly because $f(z_i)$ satisfies a difference equation of the second order. The distributions are

(20) $$W_k(y_1,y_2,\ldots,y_k;y_{k+1}) = \delta\left[y_{k+1}-(a_k y_2/a_1 - a_{k-1}y_1/a_1)\right] .$$

(b) redundancy of English

According to an estimate given by Shannon [22] the redundancy μ of written English relative to a criterion requiring perfect distinguishability of different letters is $\mu = 0.5$. This figure probably neglects long-term context.

[21] Cf., for example, ref. 21

[22] Ref. 22

(c) Wiener predictor

The Wiener predictor is a restricted predictor of the linear type with a least-squares error criterion. To design such a predictor it turns out to be unnecessary to know all the W_k functions. It is sufficient to have the autocorrelation function of the time series:

$$(21) \qquad \varphi_{11}(k) = \lim_{N\to\infty} 1/(2N+1) \sum_{i=-N}^{N} z_i z_{i+k}$$

which is expressible in terms of the W's.

(d) restricted prediction of digital expansions of irrational numbers

As an example of the fundamental difference between restricted and unrestricted prediction consider the problem of predicting the (k+1)st digit in the decimal expansion of an irrational number, say π[23], knowing the first k digits.

Since π is defined by a recursion formula it is obviously possible to predict the next digit exactly, providing there are no restrictions on the computations permitted. It is merely necessary to use one of the standard series expansions. On the other hand, it would be a miraculous mathematical coincidence if, say, the Wiener predictor were able to yield future digits unerringly.

[23] This is connected with the problem of the bandwidth required to transmit π over a noisy channel. We assume, for the sake of this discussion, that the W_k functions actually exist for π. There is some empirical evidence to support such a conjecture.

VI APPENDIX

In I,2 the limitations of information theory were illustrated by three examples [24]. It will now be shown how the statements made there follow more specifically from the theory presented in the body of the report.

(a) The problem is to construct the FQ, SQ, and CM of Figure 4. Assume that the output has, say a flat distribution over (0,1), i.e.

$$(1) \qquad p(u) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Imagine the FQ to convert u to a finely quantized form, say

$$(2) \qquad P(v_i) = 10^{-10} i \qquad (i = 0,1,2,\ldots,10^{-10})$$

Evaluate

$$\begin{cases} R = \min\limits_{Q_{v_i}(w_j)} \sum\limits_{i,j=0}^{10^{-10}} P(v_i,w_j) \log \dfrac{P(v_i,w_j)}{P(v_i)Q(w_j)} \\ \text{with} \quad Q_{v_i}(w_j) = 0 \quad \text{if} \quad |v_i - w_j| > 5 \times 10^{-4} \\ \text{and } P(v_i) \text{ as defined by (2).} \end{cases}$$

Let the minimum be achieved say for $Q_{v_i}(w_j) = Q'_{v_i}(w_j)$.

In order to build (on paper) a proper SQ consider a system with input statistic $P(v_i)$, and noise conditions described by $Q'_{v_i}(w_j)$. The SQ should be designed to operate on long messages, say 100 seconds (= 1000 symbols) long. If we arbitrarily pick $10^3 R$ of the possible high-

[24] Cf. I,2 for this section

probability received messages of the system whose transfer statistic is $Q'_{v_i}(w_j)$ and construct the fans from each of these selected received messages to the corresponding high-probability emitted messages, then, according to Corollary V, most of the emitted messages will be covered by fans. Let a fan be called by the received message from which it originates. The SQ is then to be constructed in such a way as to code an emitted message into the name of one of the fans that covers that emitted message. Evidently the SQ will involve storage facilities as well as reading and comparison circuits.

In order to build the CM it is necessary to find the channel capacity.

(4)
$$\begin{cases} C = 25 \max\limits_{P(w_i)} \sum\limits_{i,j=1}^{2} P(x_i,y_j)\log\dfrac{P(x_i,y_j)}{P(x_i)Q(y_j)} \\ \text{with } Q_{x_i}(y_j) = 3/4\ \delta_{x_i y_j} \\ \text{where say } x_1 \text{ and } y_1 \text{ represent the binary digit 0, and } x_2 \end{cases}$$

and y_2 represent the digit 1.

Let the maximum be achieved for say $P(x_i) = P''(x_i)$. If $1000R < 100C$, i.e. if $R < C/10$, it is possible to code the sequences at w into sequences at x in such a way that, according to Theorem II there will be no error in transmission. The transmitted messages must have a statistic $P''(x_i)$ and the required CM will again involve storage, reading, and comparison circuits.

(b) From the fact that the transmission is band-limited and subject to an average power limitation it follows that the speech should be coded into white noise. T.king 100 words per minute as a reasonable rate of speaking, the information rate of speech comes out to be about 10 units/second relative to a fidelity criterion that requires only intelligibility [25].

The combined FQ, SQ, and CM necessary would be a device that stores long speech-sound groups, say sentences, and looks up the appropriate white noise representation in a code book. Building such a coder is a purely technical problem outside the scope of information theory.

If the speech code is to be transmitted without error over a 10 cps. band then, according to III,5,(13) the received signal-to-noise ratio, S/N must be at least as great as the root of

(5) $10 = 10\log(1+S/N)$ or

(6) Required $S/N \geq 2$.

(c) This problem can be formulated mathematically but the formulation is actually quite useless. If we assume the device to take photographs of the sky, and if only a finite number of photographs are possible (e.g. if different photographs differ only in that different squares of a rectangular grid are filled in) there will be only a finite number of "source symbols", and it is only necessary to build an appropriate SQ. Let the possible photographs be enumerated by $i = 1,2,\ldots,n$, and the possible cloud types by $j = 1,2,3$. In order to build the SQ it is first necessary to calculate the information rate of the source subject to the fidelity criterion $\sum_{i,j} \rho(i,j)P(i,j) = \text{const.}$ Thus it is necessary to have an a-priori

[25] A result of experiments carried out to determine the redundancy of written English. See ref. 22.

set of probabilities as to the types of photographs expected, and it is also necessary to know $\rho(i,j)$ in terms of i and j. The latter requirement simply means that it is necessary to know a decision method for determining whether an arbitrary fixed value of i corresponds to a cloud of the cirrus, stratus, or cumulus types before it is possible to go ahead with the calculations necessary to obtain the SQ. However, finding such a decision method is the entire essence of the posed problem.

VII BIBLIOGRAPHY

1. Books

1. G.H. Hardy, J.E. Littlewood, G. Ploya, "Inequalities", Cambridge University Press, London, 1934.

2. J.L. Lawson, G.E. Uhlenbeck, "Threshold Signals", Vol. 24, Radiation Laboratcry Series, McGraw-Hill, New York, 1950.

3. C.E. Shannon, W. Weaver, "The Mathematical Theory of Communication", University of Illinois Press, 1949.

4. N. Wiener, "Cybernetics", John Wiley, New York, 1948; "Extrapolation, Interpolation, and Smoothing of Stationary Time Series", John Wiley, New York, 1949.

2. Periodical Literature and Miscellaneous Reports

5. P. Aigrain, "Theory of Communication", Ann. Telecommun., 4, (Dec. 1949), 406-411.

6. A.S. Besicovitch, "On the Sum of Digits of Real Numbers Represented in the Dyadic System", Math. Annalen, 110, (1934), 321-330.

7. L. Brillouin, "Life, Thermodynamics, and Cybernetics", Am. Scientist, 37, 4, (Oct. 1949), 554-568.

8. R. Cohen, "Analytical and Practical Aspects of Wiener's Theory of Prediction", Tech. Rep. 69, M.I.T. Res. Lab. of Electronics, (2 June 1948).

9. H.G. Eggleston, "The Fractional Dimension of a Set Defined by Decimal Properties", 20, (1949), 31-36.

10. R.M. Fano, "The Transmission of Information", Tech. Report 65, Res. Lab. Electronics, M.I.T., 17 March 1949.

11. D. Gabor, "Theory of Communication", J.I.E.E., Part III, 93, 26, (Nov. 1946), 429-457; "New Possibilities in Speech Transmission", 94, 32 (Nov. 1947), 369-390.

12. M.J.E. Golay, "Notes on Digital Coding", Proc. I.R.E., 37, 6, (June 1949), 657.

13. R.W. Hamming, "Error Detecting and Error Correcting Codes", BSTJ, 29, 2, (April 1950), 147-160.

14. Y.W. Lee, C.A. Stutt, "Statistical Prediction of Noise", Proc. National Electronics Conf., Chicago, (1949), 342-365.

15. Y.W. Lee, T.P. Cheatham, J.B. Wiesner, "Application of Correlation Functions in the Detection of Small Signals in Noise", Tech. Report 141, Res. Lab. Electronics, M.I.T., Oct. 1949.

16. N. Levinson, "The Wiener RMS Error Criterion in Filter Design and Prediction", J. Math. Phys., 25, 4, (Jan. 1947), 261-278; "A Heuristic Exposition of Wiener's Mathematical Theory of Prediction and Filtering", J. Math. Physics, 26, 2, (July 1947), 110-119.

17. D.M. MacKay, "Quantal Aspects of Scientific Information", Phil. Mag., ser. 7, 41, 314, (March 1950), 289-311.

18. N.C. Metropolis, G. Reitwiesner, J. von Neumann, "Statistical Treatment of Values of First 2000 Decimal Digits of e and π Calculated on the Eniac", MTAC, 4, 30, (April 1950), 109-

19. R.C. Raymond, "Communication, Entropy, and Life", Am. Scientist, 38, 2, (April 1950), 273-278.

20. S.O. Rice, "Communication in the Presence of Noise", BSTJ, 29, 1, (Jan. 1950), 60-93.

21. "Mathematical Analysis of Random Noise", BSTJ, 23, (1944), 282; 25, (1945), 46.

22. C.E. Shannon, "A Mathematical Theory of Communication", BSTJ, 27, 3 (July 1948), 379-423, and 27, 4, (Oct. 1948), 623-656.

23. C.E. Shannon "Communication in the Presence of Noise", Proc. I.R.E., 37, 1, (Jan. 1949), 10-21.

24. C.E. Shannon, "Communication Theory of Secrecy Systems", BSTJ, 28, 4, (Oct. 1949), 656-715.

25. H. Weyl, "Uber die Gleichverteilung von Zahlen modulo Eins", Mathematische Annalen, 77, (1916), 313-352.

26. H.O.A. Wold, "On Prediction in Stationary Time Series", Ann. Math. Stat., 19, 4, (Dec. 1948), 558-567.

27. L.A. Zadeh, "An Extension of Wiener's Theory of Prediction", J. Appl. Phys., 21, 7, (July 1950), 645-655.

3. Government Reports

28. M.J.E. Golay, "Notes on the Relations between Bandwidth, Available Energy and Reception of Information in Communication Systems", Ft. Monmouth, N.J., (20 Aug. 1948), Tech. Memorandum M-1108.

29. A.E. Laemmel, "General Theory of Communication, Report R-208-49, PIB-152, Contract W28-099-ac-481, Microwave Research Institute, Polytechnic Institute of Brooklyn, (13 July 1949).