#1

# A UNIFIED THEORY OF INFORMATION

KERNS H. POWERS

2 week LOAN COPY only

TECHNICAL REPORT 311

FEBRUARY 1, 1956

RESEARCH LABORATORY OF ELECTRONICS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS

The Research Laboratory of Electronics is an interdepartmental laboratory of the Department of Electrical Engineering and the Department of Physics.

The research reported in this document was made possible in part by support extended the Massachusetts Institute of Technology, Research Laboratory of Electronics, jointly by the Army (Signal Corps), the Navy (Office of Naval Research), and the Air Force (Office of Scientific Research, Air Research and Development Command), under Signal Corps Contract DA36-039 sc-42607, Project 102B; Department of the Army Project 3-99-10-022.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

RESEARCH LABORATORY OF ELECTRONICS

Technical Report 311 February 1, 1956

# A UNIFIED THEORY OF INFORMATION

Kerns H. Powers

This report is indentical with a thesis submitted to the Department of Electrical Engineering, M.I.T., 1956, in partial fulfillment of the requirements for the degree of Doctor of Science.

## Abstract

The probabilistic theory of information is extended to processes involving the most general probability distributions. A change of probability measure on an abstract space serves as the appropriate mathematical model for the fundamental information process. A unified definition for the amount of concomitant information, which takes the form of a functional of the a priori and a posteriori measures, is introduced. This definition is sufficiently general to be applied to a theory that includes both the discrete and continuous theories as special cases.

The definition is applied in a study of the information associated with the realizations of a stochastic process. For the evaluation of mutual information rates between stationarily correlated multivariate gaussian time series, the techniques of linear prediction are employed. A brief investigation is made of the problems of communication in the presence of noise and through linear networks.

## TABLE OF CONTENTS

Introduction v

I. Probability Measure 1

1. The probability measure space 1

2. Cartesian products of probability spaces 6

Product measure on the Cartesian product of independent spaces 8

Cartesian products of nonindependent spaces 11

3. Some important theorems 16

4. Infinite-dimensional product spaces 21

II. The Stochastic Process 22

5. Ergodic theory 22

Strictly-stationary discrete process 24

Ergodic processes 25

6. The autocorrelation coefficients of a discrete process 28

7. Continuous parameter processes 32

III. A Unified Definition of Information 35

8. History of the problem 36

Contributions of Wiener 36

Contributions of Shannon 39

Contributions of Woodward 41

9. The information process 42

10. The communication problem 49

11. Symmetry relations 51

12. Additivity of information 54

13. Communication in the presence of additive gaussian noise 62

IV. Prediction Theory 64

14. Spectral theory of a discrete stochastic process 64

Spectrum factorization 68

Crosscorrelation 70

15. Simple prediction 71

Pure prediction 71

The general prediction problem 73

The Wold-Kolmogorov decomposition 77

Uniqueness 81

16. Multiple prediction 82

V. Information Rates in Stochastic Processes 87

17. Mutual rate of a pair of gaussian sequences 87

18. Extension to the continuous case 95

19. Information rates in discrete networks 97

20. The rate of information about future values 100

Bibliography 104

# INTRODUCTION

Our purpose is to provide a unified mathematical theory for the treatment of the statistical processes by which information is conveyed in communication systems. As Wiener[19] pointed out, the problems of the communications engineer are closely related to the problems that have occupied the statistician for many years. In many communication systems, the messages that are of interest are analogous to the random functions of time that are commonly known in statistics as time series. These time series have two distinguishing features: (1) they may be defined either for discrete instants of time or on a time continuum; and (2) they may either take on a discrete set of values or be distributed over an amplitude continuum.

Perhaps, then, we should say that four basic types of time series are encountered in communication theory, although mixtures of these types are sometimes found. A "unified" theory of information should be sufficiently general to include the four basic types and their mixtures as special cases of the most general information process. It is precisely this generalization with which this report is concerned.

In order to handle the most general distribution in amplitude, it has been necessary to use rather advanced mathematical concepts. Consequently, the reader will find that this report is primarily a mathematical study. An early attempt was made to treat the theory in the language of the engineer, but it was found that much of its value as a unified theory was lost. Hence, it was decided that the full mathematical flavor would be retained — supplemented by a number of examples and physical interpretations that would make the results of more immediate use to the communications engineer.

A large part of this work is expository; and much work is included which is not original with the author. Included, for example, are discussions of the probability measure space, stochastic processes, ergodic theory, and the spectral theory of the discrete stochastic process. The author's approach is sufficiently different to make these theories more readily applicable to the communication problem. Our treatment of the theories is designed not to repeat those of the literature but rather to supplement them.

For example, the concept of the probability measure space is well known, but it is felt that the concept of the independence of spaces, introduced herein, represents a divergence from customary treatments. This concept was suggested by one of the thought-provoking problems in Halmos' Measure Theory (Problem 3, section 36).

Similarly, in the section on ergodic theory, a slightly different but equivalent definition of an ergodic process better illustrates the relation between our physically intuitive notions concerning ergodicity and the purely mathematical notion concerning the metric-transitivity of set transformations in a measure space. Although there is a very extensive bibliography on ergodic theory, the published works are, for the most part, purely mathematical in nature, with little or no reference to applications in the communication problem. The ergodic theorem is stated here in its mathematical form, but we try to point out more clearly its application to the interchange of statistical averages and time averages in a certain class of time series.

# I. PROBABILITY MEASURE

In order to give a unified treatment of information theory that includes the continuous and discrete theories as special cases, the problem becomes less formidable and the solution more general if an appeal is made to the measure theoretic concepts of probability theory. The foundation for such a theory has been presented by Kolmogorov,[1] who was one of the first to give to probability theory the firm mathematical foundation on which it now stands. The methods, notations, and terminology of this chapter follow rather closely those given by Halmos,[2] although a thorough comprehension of the material in a forthcoming report by Wernikoff[3] constitutes a sufficient background for the reading of this portion of the report. Wernikoff, too, has employed the notation of Halmos. Although Wernikoff devoted a very restricted portion of his report to probability measure itself, his treatment of measure theory in general includes the probability measure space as the special case of a space whose total measure is one.

We shall make free interchange of the terms "measure" and "probability," and any statement with regard to measure will imply probability measure unless stated otherwise. The occasional use of measure in the sense of Lebesgue will be qualified by employing the term Lebesgue Measure.

## 1. The Probability Measure Space

Following Kolmogorov, we assume axiomatically the existence of the following entities:

I. An abstract space X of elements x.

II. A $\sigma$-algebra[*] $\mathcal{S}$ of subsets E of X. If the space X is the space of real values, $\mathcal{S}$ is assumed to include the intervals.

III. For every set $E \epsilon \mathcal{S}$, a real valued, nonnegative, countably additive set function $\mu(E)$ such that $\mu(0) = 0$, and $\mu(X) = 1$.

[*] A $\sigma$-algebra $\mathcal{S}$ is a nonempty class of sets with the following properties:
(a) if $E \epsilon \mathcal{S}$ and $F \epsilon \mathcal{S}$, then $(E - F) \epsilon \mathcal{S}$,

(b) if $E_i \epsilon \mathcal{S}$ $(i = 1,2,\ldots)$, then $\left( \bigcup_{i=1}^{\infty} E_i \right) \epsilon \mathcal{S}$,

(c) if $E \epsilon \mathcal{S}$, then $(X - E) \epsilon \mathcal{S}$.

It is clear that a $\sigma$-algebra is simply a $\sigma$-ring which includes, in addition, the entire space X.

The elements x are called elementary events or *contingencies*; sets of these contingencies are called *events*. The set function $\mu(E)$ is the probability of the event E.

Assumptions I, II, and III above are essentially equivalent to the six axioms given by Kolmogorov, although they are slightly more restrictive in that he assumed only an algebra of sets rather than a $\sigma$-algebra. However, these assumptions do indeed satisfy all six axioms and the restriction is for all practical applications a minor one. Most authors of modern probability theory base their work on a set of postulates equivalent to I, II, and III.

These three assumptions define what in measure theory is called a *measure space* or more specifically, since $\mu(X) = 1$, a *probability measure space*. It is customary to denote a measure space by the triplet $(X, \mathcal{S}, \mu)$ which implies the existence of a space X and a $\sigma$-algebra (or $\sigma$-ring) $\mathcal{S}$ of subsets of X on which is defined the measure $\mu$. The sets of $\mathcal{S}$ are called the *measurable* sets; and, by definition, a set is measurable if and only if it is an element of the $\sigma$-algebra on which the measure is defined.

In order to see how the concept of a measure space includes our intuitive notions concerning probability, let us examine the relationships of set theory from a purely probabilistic point of view. We note that if the set A is regarded as the occurrence of an event A, the complementary set X - A is its nonoccurrence. The set $A \cup B$ represents the occurrence of either the event A or B or both, while the set $A \cap B$ is the occurrence of both A and B. The difference A - B represents the occurrence of A but not B. The set inclusion $A \subset B$ is interpreted to mean that the occurrence of A implies that of B. The empty set 0 is the impossible event; the space X is the certain event. It should be noted, however, that there will exist, in general, sets of probability zero which are not empty as well as sets of probability one which do not consist of the entire space. Thus probability zero and probability one do not imply impossibility and certainty, respectively, although the converse is indeed true.

Let us now consider two disjoint sets A and B. Since $A \cap B = 0$, it is an impossible event that both A and B occur simultaneously. In probability language, the sets A and B are said to be incompatible or *mutually exclusive*. From the countably additive property of the measure $\mu$, it follows that $\mu$ is also finitely additive and, since $A \cap B = 0$,

$$\mu(A \cup B) = \mu(A) + \mu(B). \tag{1.1}$$

This is the well-known axiom that the probability of occurrence of either of two mutually exclusive events is the sum of the probabilities of the occurrence of each of the events. The countably additive property of $\mu$ is used to deduce the result that if $\{E_i\}$ $(i = 1,2,\ldots)$ is a sequence of mutually exclusive events, then

$$\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i). \tag{1.2}$$

One might conceivably ask a question of the form: What is the probability of occurrence of the event E when it is known already that the event F has occurred? Clearly, if $E \cap F = 0$, this probability must be zero. On the other hand, if $E \supset F$, the occurrence of F implies that of E, and the probability we have asked for must be one. This question introduces the concept of the conditional probability $\mu_F(E)$ of the event E relative to the event F. Our intuitive notions suggest the definition

$$\mu_F(E) = \frac{\mu(E \cap F)}{\mu(F)}, \tag{1.3}$$

provided the set F is not null. Certainly, should F be of measure zero, our original question has doubtful meaning. It may be seen that if $E \cap F = 0$, then $\mu_F(E) = 0$; and if $E \supset F$, then $E \cap F = F$ and $\mu_F(E) = 1$.

Interchanging the roles of E and F in (1.3) above, we obtain

$$\mu_E(F) = \frac{\mu(F \cap E)}{\mu(E)} = \frac{\mu(F)\mu_F(E)}{\mu(E)}, \tag{1.4}$$

which contains the essence of Bayes' Theorem. The concept of conditional probability may be regarded to be more general than that of absolute probability; in fact, by setting F = X in (1.3), we have

$$\mu_X(E) = \mu(E). \tag{1.5}$$

Hence the probability of the event E is its conditional probability relative to the space.

A case may well occur in which the conditional probability $\mu_F(E)$ is independent of the set F. Then we have

$$\mu_F(E) = \mu(E). \tag{1.6}$$

However, from (1.4), it is seen that E and F must be mutually independent in that we also have

$$\mu_E(F) = \mu(F). \tag{1.7}$$

The usual condition for independence can be obtained from (1.3), and we can state that a necessary and sufficient condition for a pair of sets E,F to be statistically independent is that

$$\mu(E \wedge F) = \mu(E)\mu(F). \tag{1.8}$$

Repeated application of the countably additive property of $\mu$ may be employed to extend these results to countable intersections. We say that the members $E_i$ $(i = 1,2,\ldots)$ of a sequence of events are statistically independent if and only if

$$\mu\Big(\bigcap_{k=1}^{n} E_{i_k}\Big) = \prod_{k=1}^{n} \mu(E_{i_k}) \qquad (n = 2,3,\ldots) \tag{1.9}$$

for every finite subset of the sequence.

It should be remarked that pairwise independence of the members of a (finite or countable) sequence of sets is not a sufficient condition for independence of the sequence except, of course, when the sequence consists of only two sets. However, if a sequence is independent, all subsequences are, by definition, also independent.

Let us now consider a random variable f which takes on, at random, real values x in $(-\infty,\infty)$. The space X is the infinite real line, and our $\sigma$-algebra $\mathcal{S}$ is assumed to include the intervals. The event E is interpreted to mean that the value x of f falls in the set E. If the set E is measurable $(E \in \mathcal{S})$, then the measure $\mu(E)$ is simply the probability that $f \in E$. If we denote the semiclosed interval $(-\infty, x]$ by $I_x$, then

$$\mu(I_x) = \text{probability that } f \leqslant x \tag{1.10}$$

defines a nondecreasing point function, continuous on the right, and bounded by zero and one. This function is the well-known <u>distribution function</u> of the random variable f. In order to emphasize the fact that $\mu(I_x)$ is a point function, we shall usually write the distribution function as $\mu(x)$. This should cause no confusion as long as we adhere to the convention of denoting arguments that are points by lower-case letters and those that are sets of points by capital letters.

It is clear that the function $\mu(x)$ generates the measure of intervals in that we may write

$$\begin{aligned}
\mu\{x\colon a < x \leqslant b\} &= \mu(b) - \mu(a)\\
\mu\{x\colon a < x < b\} &= \mu(b-) - \mu(a)\\
\mu\{x\colon a \leqslant x \leqslant b\} &= \mu(b) - \mu(a-)\\
\mu\{x\colon a \leqslant x < b\} &= \mu(b-) - \mu(a-).
\end{aligned} \tag{1.11}$$

In fact we can obtain from $\mu(x)$ the measure of any set consisting of a finite or countable union of intervals or even countable intersections and differences. In other words, the function $\mu(x)$ generates the measure

$\mu$ on the class of Borel sets of the real line.

If the function $\mu(x)$ is the integral of its derivative, then $\mu'(x)$ is called the probability density distribution of the random variable f.

These considerations apply equally well to an n-dimensional random variable whose range X of values is the n-dimensional Euclidean space. The distribution function in such a case is, of course, an n-dimensional point function $\mu(x_1,x_2,\ldots,x_n)$ nondecreasing in each variable, while the density function, if it exists, is given by

$$\frac{\partial^n \mu(x_1,x_2,\ldots,x_n)}{\partial x_1 \, \partial x_2 \ldots \partial x_n}.$$

Let us now consider functions of the random variable f. For example, we might wish to investigate the statistical behavior of the square of the random variable or its logarithm, its absolute magnitude, and so on. We should require, naturally, that a function F(f) of a random variable f be a random variable itself; that is, if f has a distribution $\mu(x)$, there should exist a distribution $\nu(\xi)$ for the values $\xi$ of F(f). Such a distribution is defined in terms of the measure $\mu$ by

$$\nu(\xi) = \mu\left\{x\colon F(x) \leqslant \xi\right\} \quad . \tag{1.12}$$

It is seen that $\nu(\xi)$ will exist for all $\xi$ if and only if F is a measurable function of f. The expectation or mean value of a measurable function F(f) is defined to be its integral with respect to the measure $\mu$

$$\overline{F(f)} = \int_X F \, d\mu = \int_{-\infty}^{\infty} F(x) \, d\mu(x) = \int_{-\infty}^{\infty} \xi \, d\nu(\xi). \tag{1.13}$$

The mean of f itself is given by

$$\bar{f} = \int_X x \, d\mu, \tag{1.14}$$

its mean square by

$$\overline{f^2} = \int_X x^2 d\mu, \tag{1.15}$$

and its variance by

$$\sigma_f^2 = \overline{(f - \bar{f})^2} = \int_X (x - \bar{f})^2 d\mu = \int_X x^2 d\mu - 2\bar{f} \int_X x \, d\mu + \bar{f}^2 = \overline{f^2} - \bar{f}^2. \tag{1.16}$$

A pair of functions F(f) and G(f) are said to be statistically independent if

$$\mu(\{x: F(x)\epsilon M_1\} \cap \{x: G(x)\epsilon M_2\})$$
$$= \mu(\{x: F(x)\epsilon M_1\})\mu(\{x: G(x)\epsilon M_2\}) \qquad (1.17)$$

for every pair of Borel sets $M_1$ and $M_2$. If F and G are integrable $[\mu]$, a necessary and sufficient condition that they be independent is that

$$\int_X FG\, d\mu = \int_X F\, d\mu \int_X G\, d\mu. \qquad (1.18)$$

The proof of this last result can be found in Halmos.[2]

## 2. Cartesian Products of Probability Spaces

In the preceding section we were concerned with the probability theory of a single random variable f taking values x in a probability measure space $(X,\mathcal{S},\mu)$. In this section, we extend that study to a pair of random variables f and g taking values on measurable spaces $(X,\mathcal{S})$ and $(Y,\mathcal{T})$, respectively. The events of interest will be sets of ordered pairs of values (x,y) ($x\epsilon X$, $y\epsilon Y$) where (x,y) represents possible values of the joint random variable (f,g). The variables f and g may represent the outcomes of two successive performances of a single experiment (such as successive rolls of a die) in which case the range of values of f and g are identical. On the other hand, we may consider the pair (f,g) to represent the outcome of a single experiment in which the range of values of f and g need not be the same. For example, if we draw a single card from a standard deck of 52 cards the variable (f,g) may represent the rank and suit of the card drawn. The range of f (the space X) contains thirteen contingencies; that of g (the space Y), contains only four.

Let the random variables f and g be defined on measure spaces $(X,\mathcal{S},\mu)$ and $(Y,\mathcal{T},\nu)$. For almost every value g = y, let there exist a conditional measure $\mu_y$ on $\mathcal{S}$; and for almost every value f = x, a conditional measure $\nu_x$ on $\mathcal{T}$. The random variable f is said to be independent of the random variable g if the measure $\mu_y$ is independent of y; that is, if $\mu_y = \mu$ for almost every y. It will be shown later that independence of f on g implies that of g on f. An example of independence is given by the representation of the pair (f,g) by rank and suit of a playing card drawn from a complete deck. Specification of the suit of the card drawn does not affect the probability distribution of the rank nor does specification of the rank change the distribution of the suit. However, let us remove one known card from the deck and draw from the remainder. In this case, the variables f and g are no longer independent. Certainly, the specification of the suit of the card drawn in this case changes the weight of probabilities over the rank.

We shall have a great deal to do with measure spaces $(X,\mathcal{S},\mu_y,\mu)$ on which more than one measure is defined. A concept that will be very important to our study is that of absolute continuity. Since the set functions with which we deal will usually be measures (hence nonnegative), it is of value to state the definition of absolute continuity as it applies specifically to measures.

DEFINITION 2.1. Given a measurable space and a pair of measures $\mu, \nu$ defined on that space, the measure $\nu$ is said to be <u>absolutely continuous</u> with respect to the measure $\mu$ (in symbols, $\nu \rightsquigarrow \mu$) if for every $\epsilon > 0$ there exists a $\delta > 0$ such that whenever $\mu(E) < \delta$ , $\nu(E) < \epsilon$.

In simpler terms, we may say that $\nu \rightsquigarrow \mu$ if and only if $\mu(E) = 0$ implies $\nu(E) = 0$. It should be noted that the symbol $(\rightsquigarrow)$ is not, in general, symmetric. When we have both $\nu \rightsquigarrow \mu$ and $\mu \rightsquigarrow \nu$ , we write $\nu \leftrightsquigarrow \mu$. Absolute continuity is, however, both reflexive ( $\nu \rightsquigarrow \nu$ ) and transitive ( $\nu \rightsquigarrow \mu \rightsquigarrow \lambda$ implies $\nu \rightsquigarrow \lambda$). When we use the term "absolutely continuous" to describe a point function, we imply (unless stated otherwise) that the set function which it generates is absolutely continuous with respect to Lebesgue measure.

We shall have need also for the theorem of Radon-Nikodym, which has some important additional conclusions when applied specifically to probability measures. It is of value to state here a restricted form of that theorem. The proof of a more general form of the theorem can be found in Halmos[2].

THEOREM 2.1. (Radon-Nikodym). Given a measurable space $(X,\mathcal{S})$ and a pair of probability measures $\nu$ and $\mu$ defined on $\mathcal{S}$ with $\nu \rightsquigarrow \mu$, there exists a nonnegative, finite-valued function $\vartheta$, integrable with respect to $\mu$ on X so that for every measurable set E

$$\nu(E) = \int_E \vartheta \, d\mu. \qquad (2.1)$$

The function $\vartheta$ which is defined uniquely except on a set of $\mu$-measure zero, is called the Radon-Nikodym derivative and is frequently written

$d\nu/d\mu$. The nonnegativeness of $\vartheta$ follows from the fact that $\nu$ is a measure; its integrability, from the fact that $\nu$ is a probability measure with $\nu(E) \leqslant 1$.

We turn our attention now to the pair of measurable spaces $(X,\mathcal{S})$ and $(Y,\mathcal{T})$. Let $E\in\mathcal{S}$ be a subset of X; and $F\in\mathcal{T}$, a subset of Y. The set of all ordered pairs $(x,y)$ with $x\in E$ and $y\in F$ is called a rectangle and is denoted by $E \times F$.

$$E \times F = \{(x,y)\colon x\in E,\ y\in F\}. \tag{2.2}$$

A typical rectangle is shown in Fig. 1. Note that we do not assume the sets E and F to be intervals but rather measurable sets in general. Every such rectangle is a subset of the rectangle $X \times Y$, which is given the special name of the Cartesian product space. Even though $\mathcal{S}$ and $\mathcal{T}$ are $\sigma$-algebras of subsets of X and Y, respectively, the class of all rectangles $E \times F$ with $E \in \mathcal{S}$ and $F \in \mathcal{T}$ does not form a ring. Although the intersection of a pair of rectangles is always a rectangle*, neither the union nor the difference need be. However, let us consider the class $\mathcal{R}$ of sets that are finite unions of disjoint rectangles. That is, if $A \in \mathcal{R}$, then

$$A = \bigcup_{i=1}^{n} E_i \times F_i \qquad (E_i \times F_i)\cap(E_k \times F_k) = 0,\ k \neq i \tag{2.3}$$

where $E_i\in\mathcal{S}$ and $F_i\in\mathcal{T}$. It can be shown without difficulty (see Halmos[2], p. 39) that the class $\mathcal{R}$ is closed under unions and differences and hence forms a ring. Since the space $X \times Y$ is itself a rectangle belonging to $\mathcal{R}$, it follows that $\mathcal{R}$ is an algebra. Now if we consider the extended class of sets whose members are all those subsets of $X \times Y$ that can be constructed by a countable set of operations of unions, differences, and intersections applied to the class of rectangles, this extended class forms a $\sigma$-algebra. We denote by $\mathcal{S}\times\mathcal{T}$ the $\sigma$-algebra of subsets of $X \times Y$ generated in this manner by the class of rectangles. Clearly $(X \times Y, \mathcal{S} \times \mathcal{T})$ is a measurable space.

Product measure on the Cartesian product of independent spaces. We consider a pair of measure spaces $(X,\mathcal{S},\mu)$ and $(Y,\mathcal{T},\nu)$ and the measurable space $(X \times Y, \mathcal{S}\times\mathcal{T})$ formed by their Cartesian product. We say the spaces X and Y are independent if and only if the conditional measures

*We have, in fact, the identity

$$(E_1 \times F_1)\cap(E_2 \times F_2) = (E_1\cap E_2) \times (F_1\cap F_2).$$

$\mu_y$ and $\nu_x$ are independent of y and x, respectively.

Let us define the product measure on the Cartesian product of a pair of independent spaces in terms of the measures on the component spaces. A good example of such a measure is Lebesgue measure on the plane; that is, the measure of the "area" of sets in the Euclidean two-space. This example is clearly one of independent spaces, since the Lebesgue measure or "length" of sets on the X-axis is certainly independent of the values y on the Y-axis.

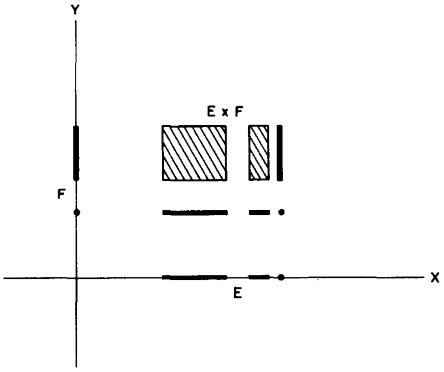

Fig. 1. The rectangle E × F.

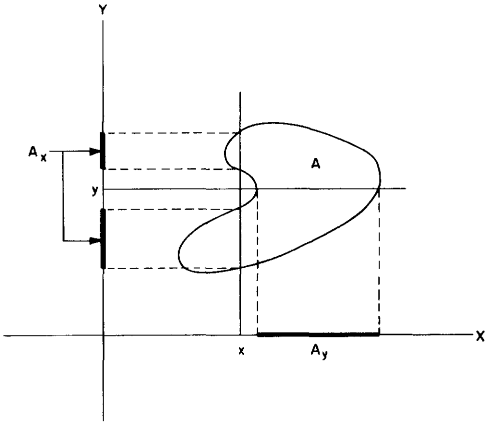

Fig. 2. The sections of a set A.

Let us consider a subset A of X × Y, which we assume, of course, to be an element of the $\sigma$-algebra $\mathscr{S} \times \mathscr{T}$ . In other words, we assume A to be a measurable set. As is illustrated in Fig. 2, the sections of the set A are defined as follows: For any fixed x, let the set

$$A_x = \{y: (x,y) \epsilon A\} \tag{2.4}$$

be called the x-section of A, while for a given y,

$$A_y = \{x: (x,y) \epsilon A\} \tag{2.5}$$

is its y-section. Notice that $A_x \subset Y$ and $A_y \subset X$. Let $\nu(A_x)$ be the measure of the x-section, and consider the set function $\rho(A)$ on $\mathscr{S} \times \mathscr{T}$ defined by

$$\rho(A) = \int_X \nu(A_x)\, d\mu. \tag{2.6}$$

It is a rather simple matter to show that $\rho(A)$ is a nonnegative, countably-additive set function such that $\rho(0) = 0$ and $\rho(X \times Y) = \mu(X)\,\nu(Y)$. In other words, $\rho$ is a measure on the product $\sigma$-algebra $\mathscr{S} \times \mathscr{T}$ . If A is the rectangle E × F, it follows that

$$A_x = \begin{cases} F & x \epsilon E \\ 0 & x \not\epsilon E, \end{cases} \tag{2.7}$$

hence for every measurable rectangle E × F,

$$\rho(E \times F) = \int_E \nu(F)\, d\mu = \nu(F)\mu(E). \tag{2.8}$$

From the definition of $\mathcal{S} \times \mathcal{T}$, every measurable set A of the product space X × Y can be covered by the union of a countable sequence of disjoint rectangles of finite $\rho$-measure. (See Halmos[2], Chaps. II and III.) In other words, we may write

$$A \subset \bigcup_{i=1}^{\infty} (E_i \times F_i) \qquad (E_j \times F_j) \cap (E_k \times F_k) = 0, \qquad k \neq j. \tag{2.9}$$

By the extension theorems, the $\rho$-measure of A is defined to be the greatest lower bound of the measures of all possible coverings of A.

$$\begin{aligned} \rho(A) &= \inf \rho\left[\bigcup_{i=1}^{\infty} (E_i \times F_i)\right] \\ &= \inf \sum_{i=1}^{\infty} \rho(E_i \times F_i) \qquad (2.10) \\ &= \inf \sum_{i=1}^{\infty} \mu(E_i)\, \nu(F_i). \end{aligned}$$

Here we have used the countable additivity of $\rho$ and the fact that the $(E_i \times F_i)$ are disjoint.

Now let $\mu(A_y)$ be the measure of the y-section and consider the set function $\rho'(A)$ on $\mathcal{S} \times \mathcal{T}$ defined by

$$\rho'(A) = \int_Y \mu(A_y)\, d\nu\,. \tag{2.11}$$

Following an identical argument to that used above, we find that for every $A \epsilon \mathcal{S} \times \mathcal{T}$,

$$\rho'(A) = \inf \sum_{i=1}^{\infty} \mu(E_i)\, \nu(F_i). \tag{2.12}$$

Thus $\rho'(A) = \rho(A)$ for every measurable A. We define the product measure $\mu \times \nu$ by the relations

$$(\mu \times \nu)(A) = \rho(A) = \int_X \nu(A_x)\, d\mu = \int_Y \mu(A_y)\, d\nu . \qquad (2.13)$$

For every measurable rectangle $E \times F$,

$$(\mu \times \nu)(E \times F) = \mu(E)\, \nu(F). \qquad (2.14)$$

It might be noted that if $\mu$ and $\nu$ are both Lebesgue measure on X and Y, the relation $\rho(A) = \rho'(A)$ reduces to the trivial conclusion that "area equals area". However, since these results apply to more general measures, the equivalence of the two definitions is not trivial. We note also that if $\nu$ and $\mu$ are probability measures (not necessarily the same) we have

$$\rho(X \times Y) = \mu(X)\, \nu(Y) = 1 \qquad (2.15)$$

and the product measure is also a probability measure.

Cartesian products of nonindependent spaces. In the preceding subsection, we saw that a product measure can be defined on the Cartesian product of a pair of spaces which are assumed to be independent. The product measure of any measurable set was defined in terms of the measures on the component spaces. The resulting measure had the property that the product measure of a rectangle is simply the product of the component measures of its sides. In this subsection, we shall extend those results to the Cartesian product of nonindependent spaces. In this case, there will exist in general, in addition to the measures $\mu$ and $\nu$, conditional measures $\mu_y$ and $\nu_x$ on the component spaces. It is clear, however, that the conditional measures may not be defined independently of one another but that there must exist some sort of Bayes relation between them and the absolute measures $\mu$ and $\nu$. It is well known that if a probability density exists on the product space, we can obtain all absolute (or marginal) densities as well as conditional densities on the component spaces by simple operations on the joint density.

From these considerations, then, we shall work the problem in the reverse order from that of the previous subsection. We shall assume the existence of a measure $\lambda$ on the measure space $(X \times Y, \mathcal{S} \times \mathcal{T}, \lambda)$ which need not have the product property posessed by $\rho$ with regard to rectangles. We shall then show how the component measures may be obtained in terms of the general measure $\lambda$. Restriction will be made in this development to probability measures.

For every measurable subset E of X, we call the set

$$S_E = \{(x,y): x \in E\} \tag{2.16}$$

the strip over E; for every $F \in \mathcal{T}$, we call

$$S_F = \{(x,y): y \in F\} \tag{2.17}$$

the strip over F (see Fig. 3). Clearly both $S_E$ and $S_F$ are measurable subsets of X × Y and, in fact, are rectangles. We can write the strips as

$$\begin{aligned} S_E &= E \times Y \\ S_F &= X \times F. \end{aligned} \tag{2.18}$$

It can be seen from Fig. 3 that $S_E \cap S_F = E \times F$.

Now for all sets of $\mathcal{S} \times \mathcal{T}$ let a measure $\lambda$ be so defined that $\lambda(X \times Y) = 1$. We define the absolute measures $\mu$ and $\nu$ on the component spaces $(X, \mathcal{S})$ and $(Y, \mathcal{T})$ by

$$\begin{aligned} \mu(E) &= \lambda(S_E) \\ \nu(F) &= \lambda(S_F) \end{aligned} \tag{2.19}$$

for every $E \in \mathcal{S}$ and $F \in \mathcal{T}$, respectively. Since the strips $S_X$ and $S_Y$ are both simply X × Y, it follows that $\mu(X) = \nu(Y) = 1$.

Let $I_y = (-\infty, y]$ be a semiclosed interval on the space Y. For any fixed y, define the nonnegative set function $\varphi_y(E)$ by

$$\varphi_y(E) = \lambda(E \times I_y) \tag{2.20}$$

for every measurable set $E \subset X$. If E is a null set, that is, if $\mu(E) = 0$, it follows that $\lambda(S_E) = 0$. Further, since $(E \times I_y) \subset S_E$, it also follows that $\lambda(E \times I_y)$, hence $\varphi_y(E)$, equals zero for all y. Thus $\varphi_y \sim \mu$, and by the Radon-Nikodym theorem there exists uniquely $[\mu]$ a nonnegative,

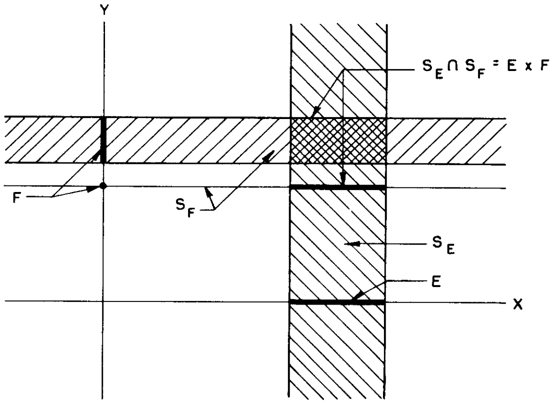

Fig. 3. The strips over E and F.

finite-valued function $\vartheta(x,y)$ such that

$$\varphi_y(E) = \int_E \vartheta(x,y)\ d\mu(x) \tag{2.21}$$

for every measurable set E. We note that since $\varphi_y(E) = \lambda(E \times I_y)$ is a nondecreasing function of y for any fixed E, $\vartheta(x,y)$ must also be nondecreasing with y for almost all x. Also, since

$$\varphi_\infty(E) = \lambda(E \times Y) = \lambda(S_E) = \mu(E),$$

we have

$$\mu(E) = \int_E \vartheta(x,\infty)\ d\mu(x) \tag{2.22}$$

for every E. Hence $\vartheta(x,\infty) = 1\ [\mu]$. A similar argument shows that $\vartheta(x, -\infty) = 0\ [\mu]$; hence, for almost every x, $\vartheta(x,y)$ is a distribution function on the space Y.

We shall call $\vartheta(x,y)$ the conditional distribution function $\nu_x(y)$ which generates, for almost every x, a conditional measure $\nu_x(F)$ of every Borel subset F of Y. By setting $E = I_x = (-\infty, x]$, (2.21) becomes

$$\varphi_y(I_x) = \lambda(x,y) = \int_{-\infty}^{x} \nu_t(y)\ d\mu(t), \tag{2.23}$$

and we can write

$$\nu_x(y) = \frac{\partial\lambda(x,y)}{\partial\mu(x)} \qquad [\mu]. \tag{2.24}$$

The partial derivative is, of course, to be interpreted in the Radon-Nikodym sense. We use the partial derivative notation simply to emphasize the fact that y plays the rôle of a parameter in the integrand (2.21) which the derivative represents.

For every Borel set F, we can write (2.21) as

$$\lambda(E \times F) = \int_E \nu_x(F)\ d\mu. \tag{2.25}$$

If we set E = X, then $\lambda(X \times F) = \lambda(S_F) = \nu(F)$, and the absolute measure $\nu$ can be expressed in terms of the conditional measure $\nu_x$ as

$$\nu(F) = \int_X \nu_x(F)\ d\mu. \tag{2.26}$$

If we consider the semiclosed interval $I_x = (-\infty, x]$ and the nonnegative set function

$$\psi_x(F) = \lambda(I_x \times F) \tag{2.27}$$

defined on $\mathfrak{T}$, the identical reasoning leads to the definition of the conditional measure $\mu_y$

$$\mu_y(x) = \frac{\partial \psi_x}{\partial \nu} = \frac{\partial \lambda(x,y)}{\partial \nu(y)} \quad [\nu] , \tag{2.28}$$

where the Radon-Nikodym derivative is taken for a fixed x. We then have for every Borel set $E \epsilon \mathcal{S}$ and every measurable set $F \epsilon \mathfrak{T}$,

$$\lambda(E \times F) = \int_F \mu_y(E) \, d\nu \tag{2.29}$$

and

$$\mu(E) = \int_X \mu_y(E) \, d\nu . \tag{2.30}$$

If the sets E and F are both Borel sets, that is, if E × F is a Borel rectangle, then (2.25) and (2.29) give equivalent definitions for the measure of such a rectangle in terms of the absolute and conditional measures on the component spaces. We must enlarge these expressions in order to obtain a similar relation for arbitrary Borel sets in the product space.

Let A be an arbitrary Borel set in X × Y. It follows that the sections $A_y$ and $A_x$ will be Borel sets in X and Y, respectively. Consider the expression

$$\lambda'(A) = \int_X \nu_x(A_x) \, d\mu . \tag{2.31}$$

If A is a Borel rectangle E × F, then

$$A_x = \begin{cases} F & x \epsilon E \\ 0 & x \not\epsilon E, \end{cases} \tag{2.32}$$

and in this case, we have, from (2.25),

$$\lambda'(E \times F) = \lambda(E \times F). \tag{2.33}$$

Now let us consider the measure

$$\lambda''(A) = \int_Y \mu_y(A_y) \, d\nu . \tag{2.34}$$

As before, if $A = E \times F$,

$$A_y = \begin{cases} E & y \epsilon F \\ 0 & y \not\epsilon F \end{cases} \tag{2.35}$$

and, from (2.29),

$$\lambda''(E \times F) = \lambda(E \times F). \tag{2.36}$$

With the observation that every Borel set A may be covered by a countable union of disjoint Borel rectangles of finite measure, it follows that we can write

$$\lambda(A) = \int_X \nu_x(A_x)\, d\mu = \int_Y \mu_y(A_y)\, d\nu \tag{2.37}$$

for the measure of any arbitrary Borel set in X × Y.

We show now how the results of this section are related to the customary treatments of probability theory when the probability density functions exist. If the measure $\lambda$ is absolutely continuous with respect to the Lebesgue product measure, the probability density function

$$\frac{\partial^2 \lambda(x,y)}{\partial x \partial y} \tag{2.38}$$

exists. The distribution functions

$$\mu(x) = \lambda(I_x \times Y) = \lambda(x,\infty) \tag{2.39}$$

$$\nu(y) = \lambda(X \times I_y) = \lambda(\infty,y) \tag{2.40}$$

are also absolutely continuous with respect to Lebesgue measure; hence $\mu'(x)$ and $\nu'(y)$ exist and represent the absolute probability densities on the component spaces. In the absolutely continuous case, the Radon-Nikodym derivatives in the definitions of the conditional measures become derivatives in the ordinary sense, and we have

$$\nu_x(y) = \frac{\partial \lambda(x,y)}{\partial \mu(x)} = \frac{1}{\mu'(x)} \frac{\partial \lambda(x,y)}{\partial x} \tag{2.41}$$

$$\mu_y(x) = \frac{\partial \lambda(x,y)}{\partial \nu(y)} = \frac{1}{\nu'(y)} \frac{\partial \lambda(x,y)}{\partial y} . \tag{2.42}$$

Taking the partial derivative of both sides of these two expressions with respect to y and x, respectively, we obtain

$$\frac{\partial \nu_x(y)}{\partial y} = \frac{1}{\mu'(x)} \; \frac{\partial^2 \lambda(x,y)}{\partial x \partial y} \qquad (2.43)$$

$$\frac{\partial \mu_y(x)}{\partial x} = \frac{1}{\nu'(y)} \; \frac{\partial^2 \lambda(x,y)}{\partial x \partial y} \; . \qquad (2.44)$$

If we let $p(x,y) = \partial^2\lambda(x,y)/\partial x \partial y$ denote the joint probability density distribution of the values (x,y) of the pair of random variables f and g, then $p(x) = \mu'(x)$ is the probability density of x and $p(y) = \nu'(y)$ is that of y. Similarly, $p_x(y) = \partial \nu_x(y)/\partial y$ corresponds to the conditional density of x for a given value y, and $p_y(x) = \partial \mu_y(x)/\partial x$ represents the conditional density of y for a given x. If we make these substitutions into (2.43) and (2.44), the familiar relations follow:

$$p(x,y) = p_x(y) \; p(x) \qquad (2.45)$$

$$p(x,y) = p_y(x) \; p(y). \qquad (2.46)$$

## 3. Some Important Theorems

In this section we shall make more precise the concept of independence and shall treat certain theorems concerning product spaces which will be useful to our study of information theory. Some of these theorems are so well known in integration theory that we shall simply state them without proof.

THEOREM 3.1. Consider a measure space $(X \times Y, \mathcal{S} \times \mathfrak{T}, \lambda)$ with component measures defined as in the preceding section. Let $E \times F$ be a Borel rectangle in $\mathcal{S} \times \mathfrak{T}$. Then the following expressions are equivalent in the sense that each implies the other two:

(a) $\mu_y(E) = \mu(E) \qquad [\nu]$

(b) $\nu_x(F) = \nu(F) \qquad [\mu]$

(c) $\lambda(E \times F) = \mu(E)\,\nu(F)$.

PROOF. Assuming $\mu_y(E) = \mu(E)$ $[\nu]$, we can write

$$\lambda(E \times F) = \int_F \mu_y(E)\, d\nu = \mu(E)\,\nu(F).$$

Thus (a) implies (c). However, from (2.26),

$$\int_E \nu_x(F)\, d\mu = \lambda(E \times F) = \mu(E)\, \nu(F) = \int_E \nu(F)\, d\mu, \tag{3.1}$$

and since this must hold for every Borel set E, we have

$$\nu_x(F) = \nu(F) \qquad [\mu]. \tag{3.2}$$

Thus (c) implies (b); and (a) implies (b). The proof that (b) implies (c) which in turn implies (a), follows from a similar argument.

DEFINITION 3.1. The component spaces X and Y of the product space X × Y are said to be independent if, for every Borel rectangle E × F, the relations (a),(b), and (c) hold (or, of course, if any one of them holds).

COROLLARY. The component spaces X and Y are independent if and only if the distribution functions satisfy

$$\lambda(x,y) = \mu(x)\, \nu(y). \tag{3.3}$$

PROOF. Apply expression (c) to the rectangle $I_x \times I_y$.

DEFINITION 3.2. The random variables f and g are said to be independent random variables if and only if the component spaces on which they are defined are independent.

It followš from the definition given in section 2 and from Theorem 3.1 that if f is independent of g, then g is also independent of f; and we say simply that f and g are independent.

We turn now to an important theorem which relates the integral of a function defined on a product space to the iterated integrals of the function over the component spaces. This theorem is the well-known Fubini theorem. We shall simply state it here, referring the reader to Halmos[2] or any other standard treatise on integration for proof.

THEOREM 3.2 (Fubini). Let $(X, \mathcal{S}, \mu)$ and $(Y, \mathcal{T}, \nu)$ be independent measure spaces, and let $\vartheta(x,y)$ be a function integrable on the measurable rectangle $(E \times Y)$. Then

$$\int_{E \times F} \vartheta\, d(\mu \times \nu) = \int_E \int_F \vartheta\, d\nu\, d\mu = \int_F \int_E \vartheta\, d\mu\, d\nu. \tag{3.4}$$

Before treating a similar theorem for nonindependent spaces, we shall have need for the following lemmas that will also prove useful in our study of information theory.

LEMMA 3.3. Let $(X \times Y, \mathcal{S} \times \mathcal{T}, \lambda)$ be a Cartesian product space with components $(X, \mathcal{S}, \mu, \mu_y)$ and $(Y, \mathcal{T}, \nu, \nu_x)$. Let the product measure $\rho = \mu \times \nu$ be so defined on $\mathcal{S} \times \mathcal{T}$ that X and Y are independent spaces. If either of the relations

$$\text{(a)} \quad \lambda \rightsquigarrow \rho$$

$$\text{(b)} \quad \mu_y \rightsquigarrow \mu \qquad [\nu]$$

$$\text{(c)} \quad \nu_x \rightsquigarrow \nu \qquad [\mu]$$

holds, then the other two also hold, and

$$\text{(d)} \quad \frac{\partial \mu_y}{\partial \mu} = \frac{\partial \nu_x}{\partial \nu} = \frac{d\lambda}{d\rho} \qquad [\lambda].$$

PROOF. We shall show first that (b) implies (a), which in turn implies (c). An identical argument shows that (c) implies (a), which implies (b). Assuming that $\mu_y \rightsquigarrow \mu \ [\nu]$, then there exists uniquely $[\nu]$ by the Radon-Nikodym theorem a nonnegative, finite-valued function $\vartheta(x,y)$, integrable $[\mu]$ on X such that, for every Borel set E,

$$\mu_y(E) = \int_E \vartheta(x,y)\, d\mu \qquad [\nu]. \tag{3.5}$$

From the expression for $\lambda$ given in (2.38),

$$\lambda(A) = \int_Y \mu_y(A_y)\, d\nu = \int_Y \int_{A_y} \vartheta(x,y)\, d\mu\, d\nu, \tag{3.6}$$

where we have used (3.5). Letting $\chi_{A_y}(x)$ be the characteristic function[*] of the set $A_y$, we have

[*] Let E be a subset of the space X. The characteristic function $\chi_E(x)$ is defined by

$$\chi_E(x) = \begin{cases} 1 & x \epsilon E \\ 0 & x \epsilon (X - E). \end{cases}$$

$$\lambda(A) = \int_Y \int_X \chi_{A_y}(x)\, \vartheta(x,y)\ d\mu\ d\nu$$

$$= \int_{X\times Y} \chi_{A_y}(x)\ \vartheta(x,y)\ d\rho \tag{3.7}$$

by Fubini's theorem. From the definition of the sections $A_x$ and $A_y$, we note that $x\epsilon A_y$ implies and is implied by $(x,y)\epsilon A$. Similarly, $y\epsilon A_x$ is equivalent to $(x,y)\epsilon A$. Thus, for all $(x,y)$,

$$\chi_{A_y}(x) = \chi_A(x,y) = \chi_{A_x}(y). \tag{3.8}$$

It follows that

$$\lambda(A) = \int_A \vartheta(x,y)\ d\rho \tag{3.9}$$

for every Borel set $A \subset (X \times Y)$, and conclusion (a) $\lambda \rightsquigarrow \rho$ is valid.

Let $A = E \times F$ be a Borel rectangle.

$$\lambda(E \times F) = \int_{E\times F} \vartheta(x,y)\ d\rho = \int_E \int_F \vartheta(x,y)\ d\nu\ d\mu. \tag{3.10}$$

But from Equation (2.26), we have

$$\lambda(E \times F) = \int_E \nu_x(F)\ d\mu.$$

Since these expressions must hold for all Borel sets E, it follows that

$$\nu_x(F) = \int_F \vartheta(x,y)\ d\nu \qquad [\mu]. \tag{3.11}$$

Therefore conclusion (c) is valid. From (3.9), $\vartheta(x,y)$ is unique $[\rho]$, but since $\lambda \rightsquigarrow \rho$, it must also be unique $[\lambda]$. From a comparison of (3.5), (3.9), and (3.11) conclusion (d) follows at once.

The proof of the following lemma, which follows from the Radon-Nikodym theorem, may be found in Halmos.[2]

LEMMA 3.4. Let $\sigma$ and $\mu$ be finite measures on a measure space with $\sigma \rightsquigarrow \mu$. If $\vartheta$ is a finite-valued function, integrable $[\sigma]$ on a set E, then

$$\int_E \vartheta \, d\sigma = \int_E \vartheta \frac{d\sigma}{d\mu} \, d\mu. \tag{3.12}$$

Note that if $\sigma$ is a conditional measure, let us say $\mu_y$, and $\mu_y \rightsquigarrow \mu$ $[\nu]$, the conclusion of this lemma may be written

$$\int_E \vartheta(x) \, d\mu_y = \int_E \vartheta(x) \frac{\partial \mu_y}{\partial \mu} \, d\mu \qquad [\nu]. \tag{3.13}$$

We repeat again that the partial derivative notation is employed simply to emphasize the fact that y is only a parameter in the integrand which $\partial \mu_y / \partial \mu$ represents.

We now have the tools to prove a theorem similar to that of Fubini but which applies in addition to the Cartesian product of nonindependent spaces. In the case of independence, this theorem reduces to that of Fubini.

THEOREM 3.5. Let $(X \times Y, \mathscr{S} \times \mathscr{T}, \lambda)$ be a product space with component measures as defined in the preceding section. Further, if $\rho = \mu \times \nu$, let $\lambda \rightsquigarrow \rho$. If $h(x,y)$ is a function integrable $[\lambda]$ on the Borel rectangle $(E \times F) \epsilon \mathscr{S} \times \mathscr{T}$, then

$$\int_{E\times F} h \, d\lambda = \int_E \int_F h \, d\nu_x \, d\mu = \int_F \int_E h \, d\mu_y \, d\nu . \tag{3.14}$$

PROOF. From the absolute continuity condition on the product measures, $\partial \nu_x / \partial \nu$ and $\partial \mu_y / \partial \mu$ exist by Lemma 3.3 and we may write

$$\begin{aligned} \int_E \int_F h \, d\nu_x \, d\mu &= \int_E \int_F h \frac{\partial \nu_x}{\partial \nu} \, d\nu \, d\mu \\ &= \int_E \int_F h \frac{d\lambda}{d\rho} \, d\nu \, d\mu \\ &= \int_{E\times F} h \frac{d\lambda}{d\rho} \, d\rho \\ &= \int_{E\times F} h \, d\lambda , \end{aligned} \tag{3.15}$$

where we have made use of Lemmas 3.3 and 3.4 as well as Fubini's theorem. From a similar argument, it follows that

$$\int_F \int_E h \, d\mu_y \, d\nu = \int_{E\times F} h \, d\lambda. \tag{3.16}$$

If the component spaces are independent, we have $\nu_x = \nu \ [\mu]$, $\mu_y = \mu \ [\nu]$, and $\lambda = \rho$. From the reflexive property of absolute continuity, Theorem 3.5 applies. Its conclusion is, in this case, identical with that of Fubini's theorem.

## 4. Infinite-Dimensional Product Spaces

Before terminating our discussion of product spaces, let us remark that the definitions and theorems given here may be extended by iteration to an n-dimensional product space. It should be noted that the product space $(X \times Y, \mathcal{S} \times \mathcal{T}, \lambda)$ is a measure space in much the same sense as is the component space $(X, \mathcal{S}, \mu)$. There is no inherent property of a measure space which limits its dimension. It follows that if $(Z, \mathcal{U}, \sigma)$ is a measure space, there exists by the methods of this chapter another measure space of triplets $(x,y,z)$ consisting of the Cartesian product of the spaces $X \times Y$ and $Z$. Also, for probability measure spaces with measure one, any n-dimensional product of such spaces will be another probability space of measure one. This process may be continued indefinitely to an infinite dimensional space whose total measure is still unity. If we let $X_i (i = 1,2,3\ldots)$ be a sequence of measure spaces, the space formed by

$$X_1 \times X_2 \times X_3 \times \cdots = \mathop{\times}_{i=1}^{\infty} X_i \tag{4.1}$$

consists of the space of all infinite sequences of random variables $(x_1, x_2, x_3, \ldots)$.

Since the details and modifications necessary for this extension are treated admirable by Halmos,[2] we shall not be concerned further with this development. Let us remark, however, that the results given in this chapter for a product space $X \times Y$ are valid even though either or both $X$ and $Y$ are themselves infinite-dimensional probability measure spaces.

# II. THE STOCHASTIC PROCESS

Let $\{f(t)\}$ be a family or ensemble of random functions of an argument $t \epsilon A$, where A is a subset of the real line. That is, for every value of the parameter t, the function f(t) is a random variable taking values from a probability measure space. Consider an arbitrary finite subset $(t_1, t_2, \ldots, t_n)$ of values of the set A and the corresponding values of the n-tuple $[f(t_1), f(t_2), \ldots, f(t_n)]$ taken on by the members of the ensemble. Let $z = (z_1, z_2, \ldots, z_n)$ be the value of a particular n-tuple and Z the n-dimensional space of all possible values over the ensemble. If there is defined a probability measure $\lambda$ on a $\sigma$-algebra of subsets of Z, the family $\{f(t)\}$ is called a stochastic process (see Khintchine[4]). A specific member f(t) is called a realization of the process.

If the set A of parameters t is a continuum or the entire real line, the ensemble $\{f(t)\}$ is called a continuous-parameter process. If t represents real time, the realizations f(t) may be regarded as random time functions. When, on the other hand, the set A consists of the positive and negative integers only, the ensemble is called a discrete, or integral-parameter process. Accordingly, the realizations of a discrete process are sequences of (not necessarily independent) random variables, and we denote by $f = \{f_i\}$ $(i = \ldots, -1, 0, 1, \ldots)$ a specific realization of such a process. In this paper we shall be concerned primarily with discrete processes, although we shall obtain certain specific results in the information theory of a continuous-parameter process.

## 5. Ergodic Theory

The ergodic theory had its origins in the classical studies of statistical mechanics, wherein it was desirable to relate time averages associated with a particular system to statistical averages associated with a universe of realizations of the system. We shall not be concerned with the mechanical systems themselves but rather with the mathematical model of such systems, namely, the stochastic process. For our purposes, then, the ergodic theory is concerned with the relationship between the average (over the parameter) of some function of a particular realization and its probability average over all possible realizations of the process.

The ergodic theorems themselves are primarily theorems concerning point and set transformations on a measure space, hence are a part of the Lebesgue theory. We shall be concerned only with that theorem of Birkhoff-Hopf-Khintchine[5,6,7], known as the Individual Ergodic Theorem, that deals in particular with one-to-one measure-preserving transformations on a space of finite measure. We shall state this theorem in its measure-theoretic form and shall show how it applies to the relationship between averages of a stochastic process.

Let $(X, \mathscr{B}, \mu)$ be a probability measure space and T a one-to-one point transformation of the space X onto itself. That is, if $x \epsilon X$, then $Tx \epsilon X$ and $T^{-1} x \epsilon X$. Since T is one-to-one, $T(T^{-1}x) = x$. Iterations of the transformation are represented by integral powers of T:

$$T(Tx) = T^2 x. \tag{5.1}$$

It follows that

$$T^k(T^j x) = T^{k+j} x = T^j(T^k x). \tag{5.2}$$

Let E be a measurable subset of X. Then the set

$$TE = \left\{ x: \ T^{-1} x \epsilon E \right\} \tag{5.3}$$

is well defined. We assume here that T is a measurable transformation; that is, if $E \epsilon \mathscr{B}$, then $T^k E \epsilon \mathscr{B}$ for every integer k.

DEFINITION 5.1. A transformation T is called measure-preserving if it is measurable and if $\mu(TE) = \mu(E)$ for every $E \epsilon \mathscr{B}$.

It follows from the measurability of T that, if T is measure-preserving, then

$$\mu(T^k E) = \mu(E) \tag{5.4}$$

for every integer k.

DEFINITION 5.2. A set M is said to be invariant under the transformation T if

$$\mu(M \cup TM) = \mu(M \cap TM). \tag{5.5}$$

More simply, we may state that a set M is invariant if it differs from its image set by a set of measure zero; that is, if both sets (M - TM) and (TM - M) are of measure zero. If, in particular, TM = M, then M is invariant under T.

DEFINITION 5.3. A transformation T is said to be metrically-transitive if it is measure-preserving and if, in addition, it leaves invariant no set of measure other than zero or one.

In simpler terms, if T is metrically-transitive and there exists a set M such that (5.5) holds, then the measure of M must be either zero or one. It is clear that a transformation is metrically-transitive only with respect to some defined measure.

The ergodic theorem of interest to us can be stated as follows:

THEOREM 5.1. (Individual Ergodic Theorem). Let T be a one-to-one transformation of the measure space $(X, \mathcal{S}, \mu)$ onto itself, and $\varphi(x)$ an integrable function defined on X.

(A) If T is measure-preserving, then the function

$$\hat{\varphi}(x) = \lim_{N \to \infty} \frac{1}{N+1} \sum_{k=0}^{N} \varphi(T^k x) \tag{5.6}$$

exists almost everywhere $[\mu]$ and

$$\int_X \hat{\varphi}(x) \, d\mu = \int_X \varphi(x) \, d\mu. \tag{5.7}$$

(B) Furthermore, if T is metrically-transitive, then $\hat{\varphi}(x)$ is constant $[\mu]$ and

$$\lim_{N \to \infty} \frac{1}{N+1} \sum_{k=0}^{N} \varphi(T^k x) = \int_X \varphi(x) \, d\mu \qquad [\mu]. \tag{5.8}$$

For a proof of the theorem as stated here, the reader is referred to Wiener[6] or Riesz[7].

Strictly-stationary discrete process. Let us consider a discrete stochastic process whose realizations are random sequences $\{f_i\}$ $(i = \ldots, -1, 0, 1, \ldots)$. Let $f = (f_{i_1}, f_{i_2}, \ldots, f_{i_n})$ be an arbitrary finite set of elements of a particular realization of the process taking values $z = (z_1, z_2, \ldots, z_n)$ on an n-dimensional probability measure space $(Z, \mathcal{U}, \lambda)$. For any integer k, the translated n-tuple $(f_{i_1+k}, f_{i_2+k}, \ldots, f_{i_n+k})$ assumes values on a space $(Z, \mathcal{U}, \lambda')$. The process is said to be strictly-stationary, or stationary in the strict sense, if for every $E \epsilon \mathcal{U}$, $\lambda'(E) = \lambda(E)$ independently of the index k. In other words, the sequence $\{f_i\}$ is a realization of a strictly-stationary discrete process if, for every integer k, the distribution functions associated with $\{f_i\}$ are identical with those corresponding distributions associated with the sequence $\{f_{i+k}\}$.

We now define a transformation T which transforms each element of $\{f_i\}$ into the next succeeding element. That is

$$T f_k = f_{k+1}. \tag{5.9}$$

This transformation is clearly one-to-one, and successive iterations of the transformation form a simple translation of the sequence

$$T^m \{f_i\} = \{f_{i+m}\}. \tag{5.10}$$

We can thus choose a particular element, say $f_o$, from the sequence $\{f_i\}$ and represent the stochastic process by an ensemble of sequences of the form $\{T^i f_o\}$ $(i = \ldots,-1,0,1,\ldots)$.

If $z = (z_1,z_2,\ldots,z_n)$ represents the value of the n-tuple $f = (f_{i_1},f_{i_2},\ldots,f_{i_n})$, then $T^k z$ is the value of the n-tuple $(f_{i_1+k},f_{i_2+k},\ldots,f_{i_n+k})$. Let E be a subset of the space Z of values z and

$$\lambda(E) = \text{Probability that } (f_{i_1},f_{i_2},\ldots,f_{i_n}) \in E. \tag{5.11}$$

Then

$$\begin{aligned}\lambda(T^k E) &= \text{Probability that } (f_{i_1},f_{i_2},\ldots,f_{i_n}) \in T^k E \\ &= \text{Probability that } (f_{i_1-k},f_{i_2-k},\ldots,f_{i_n-k}) \in E.\end{aligned} \tag{5.12}$$

If the process is strictly-stationary, it follows that these probabilities are equal, therefore

$$\lambda(T^k E) = \lambda(E) \tag{5.13}$$

for every integer k. That is, the transformation T which performs a translation of a strictly-stationary discrete process is a probability measure-preserving transformation.

Let $\varphi(f)$ be a measurable function of the n-tuple f. By part (A) of the ergodic theorem, the parametric average of $\varphi$ for a particular realization exists for almost every value z of the n-tuple f. This is given by

$$\hat{\varphi}(z) = \lim_{N\to\infty} \frac{1}{N+1} \sum_{k=0}^{N} \varphi(T^k z) \qquad [\lambda]. \tag{5.14}$$

<u>Ergodic</u> <u>processes</u>. A case of particular interest is the process for which $\hat{\varphi}(z)$ has the same value for almost every realization of the process.

This is the so-called ergodic case.

DEFINITION 5.4. A discrete stochastic process is called ergodic if it is strictly-stationary and if every measurable function $\varphi(f)$ of an arbitrary n-tuple of the process assumes the same parametric average for almost all realizations of the process.

We have shown that a translation of a strictly-stationary process is a measure-preserving transformation. We shall show later that, in the ergodic case, this transformation is metrically-transitive.

We consider a discrete ergodic process whose realizations are sequences $\{f_i\}$ of random variables. Let the elements $f_o$ of the realizations take values x on the one-dimensional space $(X,\mathcal{S},\mu)$. If $F(x)$ is any integrable function of x, then the sequence $\{F(f_i)\} = \{F(T^i f_o)\}$ is a sequence of random variables itself. Since an ergodic process is also strictly-stationary, the transformation T is measure-preserving. From part (A) of the ergodic theorem follows the existence of the parametric average

$$\hat{F}(x) = \lim_{N\to\infty} \frac{1}{N+1} \sum_{k=0}^{N} F(T^k x) \qquad [\mu] \qquad (5.15)$$

for almost every value x of the element $f_o$. The statistical average of F over the ensemble for any particular index k is

$$\overline{F} = \int_X F(x)\, d\mu, \qquad (5.16)$$

since, from the stationarity condition, the element $f_k$ is defined also on $(X,\mathcal{S},\mu)$.

In order for the parametric average (5.15) to have any meaning as a true average, it should be required that its value remain unchanged by starting the summation at some element other than $f_o$. In other words, we require that $\hat{F}(T^n x) = \hat{F}(x)$ for every integer n. To see that stationarity guarantees this, we consider the function

$$\begin{aligned}\hat{F}(Tx) &= \lim_{N\to\infty} \frac{1}{N+1} \sum_{k=0}^{N} F(T^{k+1}x) \\ &= \lim_{M\to\infty} \left\{\frac{M+1}{M}\,\frac{1}{M+1} \sum_{m=0}^{M} F(T^m x) - \frac{1}{M} F(x)\right\} \qquad [\mu],\end{aligned} \qquad (5.17)$$

wherein we have set m = k + 1 and M = N + 1.

The integrability of F demands that it be finite-valued almost everywhere, hence the second term vanishes $[\mu]$ in the limit. The limit of the first term is simply $\hat{F}(x)$ thus, by iteration,

$$\hat{F}(T^n x) = \hat{F}(x) \qquad [\mu] \tag{5.18}$$

for every n. In the stationary case, then, the parametric average $\hat{F}(x)$ of a particular realization has the same value for almost all values x of the elements $f_k$ of that realization. In the ergodic case, moreover, its value remains the same for almost all realizations. Therefore, in the ergodic case,

$$\hat{F}(x) = c \qquad [\mu], \tag{5.19}$$

where c is a constant independent of x. Applying (5.7) we have

$$c = \int_X F(x)\, d\mu = \overline{F} \tag{5.20}$$

and, equating (5.15) and (5.16), we have

$$\lim_{N \to \infty} \frac{1}{N+1} \sum_{k=0}^{N} F(T^k x) = \int_X F(x)\, d\mu \qquad [\mu] \tag{5.21}$$

which is the assertion of part (B) of the ergodic theorem.

Now we can show that the translation transformation T of an ergodic process is metrically-transitive. To do this, we choose a particular function of x, namely, the characteristic function $\chi_E(x)$ of some measureable set E. That is, we consider the random sequence $\{\chi_E(f_i)\}$ whose elements are 1 when $f_i \epsilon E$, and zero otherwise. This function is clearly $\mu$-integrable and its integral over the space X is simply the measure $\mu(E)$ of the set E. From (5.21) above,

$$\mu(E) = \lim_{N \to \infty} \frac{1}{N+1} \sum_{k=0}^{N} \chi_E(T^k x) \qquad [\mu]. \tag{5.22}$$

Remembering that x is the value taken on by a particular element, say $f_o$, we see that the expression

$$\frac{1}{N+1} \sum_{k=0}^{N} \chi_E(T^k x)$$

represents simply the proportionate number of times the value of the elements $\{f_i\}$ falls in the set E in a succession of length (N + 1) following the element $f_o$. It is quite reasonable that the limit of this expression should represent, in the ergodic case, the probability that any element fall in the set E.

In order to show the metric-transitivity of the transformation T, we postulate the existence of a set M invariant under T. For almost every $x \epsilon M$, $x \epsilon TM$ also. We may state invariance by writing for all k,

$$\chi_M(x) = \chi_{T^k M}(x) \qquad [\mu], \tag{5.23}$$

since the set of points x of X for which the equality fails to hold must be of measure zero. From (5.22) it follows that

$$\mu(M) = \lim_{N \to \infty} \frac{1}{N+1} \sum_{k=0}^{N} \chi_M(T^k x) \qquad [\mu]. \tag{5.24}$$

However, for almost every x, $T^k x \epsilon M$ implies $x \epsilon T^{-k}M$. Consequently,

$$\chi_M(T^k x) = \chi_{T^{-k}M}(x) = \chi_M(x) \qquad [\mu] \tag{5.25}$$

is independent of the index k. Equation (5.24) becomes

$$\mu(M) = \chi_M(x) \qquad [\mu], \tag{5.26}$$

and the measure of the invariant set M is either zero or one, depending on whether or not the initial element $f_o$ lies in M. Thus, in the ergodic case, the translation transformation is metrically-transitive.

## 6. The Autocorrelation Coefficients of a Discrete Process

In this section, we show that the autocorrelation coefficients associated with a discrete ergodic process can be expressed in two equivalent forms. These forms are (a) in terms of a statistical average over the ensemble and (b) in terms of a parametric average in a particular realization.

We consider a discrete stochastic process whose realizations are random sequences $\{f_i\}$. Let the element $f_k$ take values x from a measure space $(X, \mathcal{S}, \mu)$ and $f_j$ take values y from the space $(Y, \mathcal{T}, \nu)$. Since, in general, the elements of the sequence $\{f_i\}$ will not be statistically independent, the pair $(f_k, f_j)$ will take values on the product space $(X \times Y, \mathcal{S} \times \mathcal{T}, \lambda)$ with component spaces X and Y. The measures $\mu$ and $\nu$ are then the absolute measures on the component spaces. The autocorrelation coefficient $R_{kj}$ is defined to be the statistical average of the product $f_k f_j$. Thus

$$R_{kj} = \overline{f_k f_j} = \int_{X\times Y} xy \, d\lambda, \tag{6.1}$$

when this integral exists. It follows from Fubini's theorem that

$$R_{kk} = \overline{f_k f_k} = \int_{X\times Y} x^2 \, d\lambda = \int_X x^2 d\mu = \overline{f_k^2} \,. \tag{6.2}$$

If the elements $f_k$ and $f_j$ are statistically independent, then $\lambda = \mu \times \nu$, and from Fubini's theorem,

$$R_{kj} = \int_{X\times Y} xy \; d(\mu \times \nu) = \int_X \int_Y xy \; d\mu \; d\nu$$

$$= \int_X x \; d\mu \int_Y y \; d\nu = \overline{f_k} \; \overline{f_j}. \tag{6.3}$$

Furthermore, if the process is strictly-stationary, the distributions associated with the pair $(f_k, f_j)$ are identical to those associated with the translate $(f_{k+m}, f_{j+m})$, for all integers m. Thus

$$R_{kj} = \overline{f_k f_j} = \overline{f_{k+m} f_{j+m}} = \overline{f_m f_{m+j-k}} \quad , \tag{6.4}$$

wherein we have translated again by -k. From the strict stationarity, the latter average is independent of m, hence $R_{kj}$ is a function of the difference (j - k) only. If we denote by n the difference (j - k), the autocorrelation coefficient of a stationary process becomes

$$R_n = \overline{f_m f_{m+n}} = \overline{f_{m-n} f_m} = R_{-n} \quad , \tag{6.5}$$

wherein we have translated again by -n. Accordingly, the autocorrelation coefficient of a stationary discrete process is an even function of the index n. Clearly, $R_o$ is the mean-square $\overline{f_k^2}$ which is, of course, independent of k.

From the ergodic theorem, every realization of a stationary process posesses a parametric average

$$\lim_{N \to \infty} \frac{1}{N+1} \sum_{k=0}^{N} f_k f_{k+n}$$

whose value is independent of the value of the initial element $f_o$. In the ergodic case (that is, if the process is also ergodic), this average is the same for almost all realizations and in fact is equal to $R_n$. Thus we have

$$R_n = \lim_{N \to \infty} \frac{1}{N+1} \sum_{k=0}^{N} f_k f_{k+n} = \int_{X \times Y} xy \, d\lambda \tag{6.6}$$

where X × Y is the space of values of the pair $(f_k, f_{k+n})$ for any k.

This expression relates two equivalent definitions for the autocorrelation coefficients of an ergodic process. We can obtain these coefficients either by an average over the ensemble of the product of any particular pair $(f_k, f_{k+n})$, or by an average of that product over the parameter k in a particular realization of the process. From a practical point of view, this is a very important relation. While the ensemble is a sort of fictional entity, the realization, or at least a finite part of the realization, represents a finite sequence of numbers which may have been obtained by experiment. The parametric average represents a value which we might well obtain from a finite number of measurements in the laboratory on a particular random sequence.

From the communication point of view, a particular realization of the process may represent a message or, perhaps, some received signal conveying information about that message. In such a situation, we shall not, in general, have a knowledge of all the elements of a realization but rather a knowledge of only those elements in a finite past history of the sequence. From these considerations, it is of interest to show that such a knowledge of a particular realization is sufficient for obtaining any necessary statistical characteristics concerning the future of the sequence.

Let us consider again a particular realization $\left\{ T^i f_o \right\}$ of an ergodic ensemble. Let $f = \left( T^{i_1} f_o, T^{i_2} f_o, \ldots, T^{i_n} f_o \right)$ represent an arbitrary n-tuple taking values z from the n-dimensional measure space $(Z, \mathcal{U}, \lambda)$. Let $\varphi(z)$ be an integrable $[\lambda]$ function of the values z of the n-tuple f. As was shown in section 5, the translation transformation $T^k$, when applied to an ergodic process, is metrically-transitive as well as measure-preserving for every integer k. It follows, in particular, that $T^{-1}$ has these properties. Applying the ergodic theorem to the function $\varphi(z)$ and to the inverse transformation $T^{-1}$, we obtain

$$\lim_{N \to \infty} \frac{1}{N+1} \sum_{k=0}^{N} \varphi(T^{-k} z) = \int_Z \varphi(z) \, d\lambda. \tag{6.7}$$

If $E \epsilon \mathcal{U}$ is a subset of Z and $\chi_E(z)$ is its characteristic function, the expression above becomes, with $\varphi(z) = \chi_E(z)$,

$$\lambda(E) = \lim_{N \to \infty} \frac{1}{N+1} \sum_{k=0}^{N} \chi_E(T^{-k}z). \tag{6.8}$$

We note that if z represents the present value of the n-tuple f in a particular realization, then $T^{-k}z$ for $k > 0$ represents all past values. Thus we can always obtain all joint distribution functions connected with the realizations of an ergodic process by making a sufficiently large number of measurements on the past history of any one of its realizations.

If we set $f = (f_k, f_{k-n})$, it follows from the even property of the autocorrelation coefficients that $R_n$ is given by the average of the product xy of values (x,y) taken on by the pair $(f_k, f_{k-n})$. Applying Eq. (6.7), we obtain

$$R_n = \lim_{N \to \infty} \frac{1}{N+1} \sum_{k=0}^{N} f_{-k} f_{-k-n}$$

$$= \lim_{N \to \infty} \frac{1}{N+1} \sum_{k=0}^{-N} f_k f_{k-n}. \tag{6.9}$$

Since $f_o$ represents the present value of the sequence $\{ f_i \}$, it is seen that the autocorrelation coefficients also can be obtained from an observation of past values only of the elements of the sequence.

If we remember that in the strictly-stationary case the parametric average is invariant under translations, i.e., $\hat{\varphi}(T^k z) = \hat{\varphi}(z)$, it follows that

$$R_n = \lim_{N \to \infty} \frac{1}{N+1} \sum_{k=0}^{-N} f_{k-m} f_{k-m-n} \tag{6.10}$$

is independent of m.

Now let us suppose that we have at our disposal a very large number N of ordered random variables. That is, we know precisely the values of the finite sequence $(f_{-N}, f_{-N+1}, \ldots, f_{-1}, f_o)$. If the two averages

$$\hat{f} = \frac{1}{N - m + 1} \sum_{k=0}^{-N+m} f_{k-m} \tag{6.11}$$

and

$$\hat{R}_n = \frac{1}{N - m - n + 1} \sum_{k=0}^{-N+m+n} f_{k-m} f_{k-m-n} \tag{6.12}$$

are independent of the index m for small (compared to N) values of m and n, the sequence will be said to be <u>stationary</u> (in the wide sense) <u>in the past</u>. It is clear that the class $\mathfrak{F}$ of all infinite sequences $\{f_i\}$ $(i = \ldots,-1,0,1,\ldots)$ with average $\bar{f}$ and correlation coefficients $R_n$ equal to $\hat{f}$ and $\hat{R}_n$, respectively, contains a subclass of such sequences having the finite sequence above as the values of its past N elements. The class $\mathfrak{F}$ of sequences, which is said to be ergodic in the wide-sense, represents the ensemble from which our finite sequence was selected. The statistics associated with the unknown future of the sequence whose finite past we know are precisely those statistics associated with the future of the realizations of the class $\mathfrak{F}$ .

## 7. Continuous Parameter Processes

In this section, we shall review the results of the preceding section as they apply to the continuous-parameter process. The details of the development will be omitted, since the arguments are, with minor modifications, similar to those encountered in the study of discrete processes.

The ergodic theorem concerns a measure space $(X,\mathcal{S},\mu)$ and an Abelian group of continuous-parameter one-to-one transformations $T^{\gamma}$ of the space X onto itself. In this case, the parameter $\gamma$ is any real number in $(-\infty,\infty)$. The transformations have the property that

$$T^{\tau}(T^{\gamma}x) = T^{\tau+\gamma}x \tag{7.1}$$

for all $\gamma$ and $\tau$ . The ergodic theorem in the continuous-parameter case is slightly weaker than that in the discrete case in that hypotheses concerning measurability, measure-preservation, and metric-transitivity of a transformation $T^{\gamma}$ are made for <u>almost all</u> real values of the parameter $\gamma$ , that is, with the exception of a set of values of Lebesgue measure zero.

The ergodic theorem for continuous-parameter transformations may be stated as follows:

THEOREM 7.1 (Individual Ergodic Theorem). Let $T^{\gamma}$ be an element of an Abelian group of one-to-one transformations of the measure space $(X,\mathcal{S},\mu)$ onto itself, and $\varphi(x)$ a function, integrable on X, so defined that $\varphi(T^{\gamma}x)$ is measurable in both x and $\gamma$ .

(A) If $T^{\gamma}$ is measure-preserving for almost all $\gamma$, then the function

$$\hat{\varphi}(x) = \lim_{A \to \infty} \frac{1}{A} \int_0^A \varphi(T^{\gamma} x) \, d\gamma \tag{7.2}$$

exists almost everywhere $[\mu]$, and

$$\int_X \hat{\varphi}(x) \, d\mu = \int_X \varphi(x) \, d\mu. \tag{7.3}$$

(B) Furthermore, if $T^{\gamma}$ is metrically-transitive for almost all $\gamma$, then $\hat{\varphi}(x)$ is constant $[\mu]$, and

$$\lim_{A \to \infty} \frac{1}{A} \int_0^A \varphi(T^{\gamma} x) \, d\gamma = \int_X \varphi(x) \, d\mu \qquad [\mu]. \tag{7.4}$$

Let the random time function f(t) be a realization of a continuous-parameter stochastic process so that any arbitrary n-tuple $f = [f(t_1), f(t_2), \ldots, f(t_n)]$ assumes values $z = (z_1, z_2, \ldots, z_n)$ on the product space $(Z, \mathcal{U}, \lambda)$. For any real number $\tau$, the translation $[f(t_1 + \tau), f(t_2 + \tau), \ldots, f(t_n + \tau)]$ assumes values on $(Z, \mathcal{U}, \lambda')$. In a manner analogous to that for the discrete case, the process $\{f(t)\}$ is said to be <u>strictly-stationary</u> if, for every $E \in \mathcal{U}$, $\lambda'(E) = \lambda(E)$, independently of the value of $\tau$. In other words, f(t) is a realization of a strictly-stationary process if every distribution function associated with f(t) is identical with the corresponding distribution function associated with $f(t + \tau)$.

It follows as before that, if we define the time translation transformation $T^{\tau}$ by

$$T^{\tau} f(t) = f(t + \tau), \tag{7.5}$$

such a transformation preserves measure in a strictly-stationary process. From part (A) of the ergodic theorem, the time average $\hat{\varphi}$ of any integrable function of f exists for almost all realizations of the process and is given by

$$\hat{\varphi}(z) = \lim_{A \to \infty} \frac{1}{A} \int_0^A \varphi(T^{\tau} z) \, d\tau \qquad [\lambda], \tag{7.6}$$

where z is the value of the n-tuple f. Although this average will be independent of time for a particular realization, its value may well be different for different realizations, hence may depend on the particular value z of f.

In the ergodic case, however, the time average is the same for almost all realizations, therefore is a constant for almost all z. Thus if the process $\{ f(t) \}$ is ergodic,

$$\hat{\varphi} = \lim_{A\to\infty} \frac{1}{A} \int_0^A \varphi(T^{\tau} z)\, d\tau = \int_Z \varphi(z)\, d\lambda \qquad [\lambda]. \qquad (7.7)$$

As in the discrete case, we can obtain all probability distribution functions (as well as autocorrelation functions) from an observation of the past history only of a particular realization of an ergodic continuous-parameter process. For example, if we let $t_o$ denote the present time, the measure $\lambda(E)$ of any measurable set in the product space $(X \times Y, \mathscr{S} \times \mathscr{T}, \lambda)$ of values $(x,y)$ of the pair $[f(t_o), f(t_o - \tau)]$ is obtained by the application of (7.7) to the characteristic function $\chi_E(z)$ of the set E and the use of the metrically-transitive transformation $T^{-\gamma}$. Thus

$$\lambda(E) = \lim_{A\to\infty} \frac{1}{A} \int_0^A \chi_E \left[T^{-\gamma}(x,y)\right] d\gamma, \qquad (7.8)$$

where

$$\chi_E\left[T^{-\gamma}(x,y)\right] = \begin{cases} 1 & \text{if } \left[f(t_o - \gamma), f(t_o - \tau - \gamma)\right] \epsilon E \\ 0 & \text{otherwise.} \end{cases} \qquad (7.9)$$

Similarly, the autocorrelation function becomes

$$R(\tau) = \int_{X\times Y} xy\, d\lambda$$

$$= \lim_{A\to\infty} \frac{1}{A} \int_{-A+t_o}^{t_o} f(t)\, f(t - \tau)\, dt, \qquad (7.10)$$

which, in the stationary case, is independent of $t_o$.

# III. A UNIFIED DEFINITION OF INFORMATION

The primary objective of a communication system is to obtain from a source a certain amount of information and to convey it to a receiver at the other end of the system. In order for the information conveyed to have any value, it must in some manner add to the knowledge of the receiver. Should the receiver receive a statement which it already knew to be true, certainly no information will have been conveyed. In order for any statement to convey information, there must exist, a priori in the receiver, an uncertainty concerning the subject of the statement. The more uncertain the receiver is about the content of the statement, the larger is the amount of information received.

It was Wiener, perhaps, who first recognized that communication is in reality a statistical problem. His pioneering work,[8] and that of Shannon,[9] first introduced statistical methods into the communication problem. They extended the earlier work of Nyquist and Hartley concerning information measures to include the probability concepts that are necessary in handling more general classes of communication problems. From this study, they evolved a statistical definition for the measure of the information concept, placing information theory on a firmer mathematical foundation.

A very general communication system is one in which ideas in some form or other are conveyed from a source to a receiver in order to add to the knowledge at the receiver. Such a general system does not lend itself directly to treatment by mathematical methods because of the rather obscure concepts of ideas and knowledge. One way to make the transition to a more amenable system is to assume that the messages or ideas are coded in such a way that they represent either a discrete sequence of values of a random variable or a continuously varying random time function. Teletype, telephone, television and many other forms of communication systems are examples of this coding. If we assume the existence of one-to-one transducers which transform these sequences and functions back to their original forms, practically no loss of generality results from the transition.

In the design of a communication system, one does not optimize the system in order to maximize the information conveyed about a particular message but rather designs the system to handle a specific class of messages. For example, a telephone system must be capable of handling messages conveyed by a large variety of voices varying in their individual characteristic. Thus a telephone system is optimized to convey the maximum information _on the average_ about any message selected from an ensemble of possible messages. In coding the particular messages in the form of random sequences or random time functions, the appropriate mathematical model for the study of information is the stochastic process. The a priori knowledge of the receiver at any instant may be expressed in terms of a set of probability distribution functions associated with the ensemble of possible messages. On the reception of a certain amount of additional information in a time interval, the a posteriori knowledge is characterized by a new set of these distributions. It is clear that the amount of information conveyed in the interval should be expressed in terms of the a priori and a posteriori distribution functions. We see, then, that the fundamental process by which information is conveyed has as its mathematical model a change of a set of distribution functions.

## 8. History of the Problem

In order to develop a general theory of the information associated with a stochastic process, we shall start with rather simple problems and gradually extend the results in a more general direction. At the same time, we shall review the major fundamental contributions of the early work of Wiener and Shannon and point out their connections with the development given here.

Contributions of Wiener. We consider a random variable f which takes on, at random, real values x from a probability measure space $(X,\mathcal{S},\mu)$. The measure $\mu(E)$ of any measurable set E is then the probability that the value of f falls in the set E.

$$\mu(E) = \text{probability that } f\epsilon E. \qquad (8.1)$$

Wiener[8] has considered the following problem: If we know a priori that the value of the variable f lies in some set A and we are told, in addition, that $f\epsilon B$, how much information have we obtained? Clearly, our a posteriori knowledge is that f belongs to both A and B, hence lies in their intersection $A\cap B$. As introduced by Wiener, a reasonable measure of the information received is

$$I_1 = \log \frac{\mu(A)}{\mu(A\cap B)} \quad , \qquad (8.2)$$

where the base of the logarithm determines the units of I. The logarithmic measure is chosen in order to make the information from independent sources additive. For example, if we receive further information to the effect that $f\epsilon C$, that is, an additional amount given by

$$I_2 = \log \frac{\mu(A \cap B)}{\mu(A \cap B \cap C)} , \tag{8.3}$$

the total information given us is

$$I = I_1 + I_2 = \log \frac{\mu(A)}{\mu\left[A \cap (B \cap C)\right]} , \tag{8.4}$$

which may be seen to represent the amount of information conveyed by the equivalent statement $f \in B \cap C$.

We note that if $B \supset A$, the amount of information $I_1$ is zero. On the other hand, if $A \cap B$ is a set of small measure relative to A, the amount of information obtained becomes quite large. If $A \cap B$ is a nonempty set of zero measure, we see that the information obtained is infinite. This is the value that our intuition would prefer. On the other hand, suppose that the set $A \cap B$ is empty. Our expression gives an infinite value to such information although it is doubtful in this case whether any information at all has been obtained. Our a priori and a posteriori knowledge become contradictory. If A and B are disjoint, the value of f cannot possibly lie in both A and B; thus, either our a priori knowledge is false or the information given us is false. Without further information, we do not know which to discard.

We shall interpret (8.2) to have meaning only when $A \cap B$ is nonempty and shall take the point of view that our a priori knowledge is unquestionably correct. This assumption allows us to conclude that $f \in A \cap B$ when we have received only the information that $f \in B$. With this interpretation, it follows that since $A \supset A \cap B$, then $\mu(A) \geqslant \mu(A \cap B)$, and the information obtained is always nonnegative.

The definition of information given by (8.2) treats only the special case in which the information received serves to reduce the range of values of f to some subset of its a priori range but does not provide a measure for the more general information processes. Wiener has considered also the following, more general problem:

If the random variable f is known a priori to be distributed according to the probability distribution function $\rho(x)$ and information is received which permits the formulation of a new a posteriori distribution $\nu(x)$, how much information does this change of distribution represent? Here, the information has not merely reduced the range of values of f to a subset, but has changed the defining measure on the space. Later we shall see that this is not a different situation from the first problem considered but simply represents a more general process of which the first is a special case.

Using a suggestion of von Neumann, Wiener proposed, as a solution to this general problem, that the information received should measure the change of our uncertainty about the random variable in question. He then introduced an expression of the form

$$\int_{-\infty}^{\infty} \mu'(x) \log \mu'(x)\, dx$$

to represent the uncertainty, or entropy, associated with the distribution density function $\mu'(x)$ by which the random variable is distributed. For the information resulting from a change of distribution of the variable, he proposed that we take the difference of the entropies of the a priori and a posteriori distributions. Such a definition looks promising at first glance and, in fact, guarantees that the information given by each of a sequence of distributions adds to give the total information. However, there are several shortcomings in the implications of this definition.

In the first place, the definition for entropy in terms of density functions is capable of treating only the special case in which the probability distribution functions are absolutely continuous. Furthermore, it forces sequences of distributions to provide additive information with no regard for the question of independence of the sources. It provides information which may be either positive or negative, giving no answer to the question of whether the a priori or a posteriori knowledge is more correct. Finally, it does not necessarily give a higher information value to less likely events.

This last statement becomes clear when we consider the information processes of Fig. 4. Suppose that $\rho(x)$ is an a priori distribution function, and let $\nu_1(x)$ and $\nu_2(x)$ represent two possible a posteriori distributions, differing only by their mean values. Clearly, $\nu_2(x)$ is less likely to follow $\rho(x)$ than is $\nu_1(x)$, because its most probable value lies in a neighborhood of low a priori probability density. It follows that $\nu_2(x)$ should provide more information than $\nu_1(x)$. However, it is easily seen that the definition of entropy gives a value invariant under translation of the distribution, with the result that the entropies of both a posteriori distributions are equal. According to the difference-of-entropy

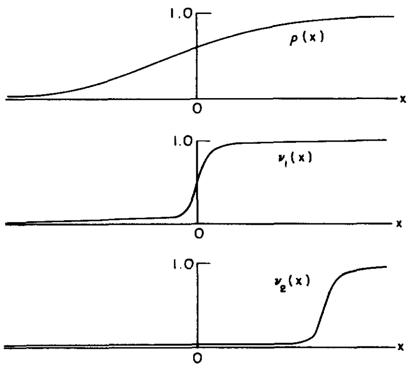

Fig. 4

definition, they both provide equal values of information even though one is much less likely to occur than the other.

Contributions of Shannon. In his classical work[9], Shannon did not use the definition of information proposed by Wiener, nor did he provide any definition at all for information itself. In spite of this omission, Shannon's paper represents one of the most significant single contributions to the modern theory of communication. He made use of the entropy concept as a natural measure of information and derived expressions in terms of entropies for the rate of transmission of information and the information capacity of both noiseless and noisy channels. Shannon treated in detail, however, the communication problem in only the two cases represented by pure step and absolutely continuous distribution functions. He proposed an attack for the general mixed case based on his expression for information rate but presented few details of such an extension.

Shannon defined the entropy associated with a finite set of probabilities $\{p_i\}$, with

$$\sum_i p_i = 1, \tag{8.5}$$

by the expression

$$H = -\sum_i p_i \log p_i, \tag{8.6}$$

which may be considered to represent the average amount of information required to single out any one element $x_k$, with probability $p_k$, from the entire set of all possible x. It thus represents the average rate of information (per symbol) generated by a source that produces a sequence $\{x_i\}$ of independent random variables according to a discrete probability distribution. He extended this definition to include nonindependent sequences by representing more general sources as Markov processes and evolved a definition for the average rate of information generated by any source of ergodic character.

A simple form of the communication problem considered by Shannon may be stated as follows: Let X represent a finite set of n symbols $x_i$ with each of which is associated a probability p(i). Let an ergodic source generate an infinite sequence of symbols selected independently at random from the set X, and let this source feed a channel whose output consists of an independent sequence of elements $y_j$ from an m-element set Y. Let a set of nm joint probabilities p(i,j) be defined on the product space X × Y. For every element $x_i$, there exists a probability $p_i(j)$ for the occurrence of the symbol $y_j$. Clearly, the set of probabilities $p_i(j)$ characterize the properties of the channel, and

$$p(i,j) = p_i(j)\ p(i) \tag{8.7}$$

is determined by the characteristics of both the channel and source. From (8.6) the source entropy is given by

$$\begin{aligned} H(x) &= -\sum_i p(i) \log p(i) \\ &= -\sum_{i,j} p(i,j) \log \sum_j p(i,j). \end{aligned} \tag{8.8}$$

The entropy of the received symbols $y_j$ is given by

$$\begin{aligned} H(y) &= -\sum_j p(j) \log p(j) \\ &= -\sum_{i,j} p(i,j) \log \sum_i p(i,j). \end{aligned} \tag{8.9}$$

Shannon defined a <u>conditional</u> <u>entropy</u> $H_x(y)$ as the average of the entropy of y for each value of x, weighted according to the probability associated with that particular x.

$$H_x(y) = -\sum_{i,j} p(i,j) \log p_i(j). \tag{8.10}$$

He then defined the rate (per symbol) at which information is conveyed through the channel to be

$$\begin{aligned} R &= H(y) - H_x(y) \\ &= \sum_{i,j} p(i,j) \log \frac{p(i,j)}{p(i)p(j)}\ . \end{aligned} \tag{8.11}$$

It is clear that the quantity p(i,j)/p(i)p(j) is some sort of measure of the dependence between the input and output of the channel; hence the average value of its logarithm is a reasonable measure of the average information conveyed per symbol. The channel capacity is then defined as the maximum value of R for all possible source distributions.

Shannon treated also the case of an infinite alphabet X consisting of the entire real line $(-\infty, \infty)$ whose elements x are distributed by an absolutely continuous distribution function $\mu(x)$. By analogy with his discrete theory, he defined the entropy of an absolutely continuous distribution function by the expression

$$H = -\int_{-\infty}^{\infty} \mu'(x) \log \mu'(x)\, dx, \tag{8.12}$$

which is the negative of the quantity used by Wiener. For the continuous analogy of the discrete communication problem treated above, Shannon considered a pair of such alphabets X and Y on whose Cartesian product X × Y the joint distribution density p(x,y) is defined. If we denote by p(x) the density $\mu'(x)$ and by p(y) the corresponding density on y, the average rate (per symbol) conveyed through a continuous channel from an infinite alphabet source generating independently a sequence of such symbols becomes

$$R = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}\, dxdy. \tag{8.13}$$

Contributions of Woodward. A slightly different interpretation of Shannon's theory is represented in the work of Woodward, who developed the entire theory from a pair of additivity axioms. Woodward introduced the interpretation of the bi-variate random variable

$$I_{xy} = \log \frac{p(x,y)}{p(x)p(y)} \tag{8.14}$$

as a measure of the mutual information between the random variables x and y. He then studied the properties of various averages of $I_{xy}$ over the various distribution functions involved. If x and y are considered to be the transmitted message and received signal, respectively, the average information about x provided by a particular y received is, from the receiver's point of view,

$$I_y = \int_{-\infty}^{\infty} p_y(x) \log \frac{p(x,y)}{p(x)p(y)}\, dx \tag{8.15}$$

if p(x,y) is a probability density, and

$$I_y = \sum_x p_y(x) \log \frac{p(x,y)}{p(x)p(y)} \qquad (8.16)$$

if p(x,y) is purely discrete.

It is clear that the average of $I_{xy}$ over all possible (x,y) is simply the rate of information given by Shannon.

## 9. The Information Process

Although the entropy concept is a very useful one and, as a measure of uncertainty, has a rather illuminating physical interpretation, it is a concept that does not lend itself to an abstract generalization. In fact, the definitions given by Shannon for the entropy of discrete and absolutely continuous distributions represent entirely different entities that have similar but by no means identical properties. For example, if we consider a sequence of step distribution functions that converges uniformly to some absolutely continuous distribution $\mu(x)$, it will follow that the entropy of the members of the sequence becomes unbounded in the limit, even though the integral (8.12) is finite. This will be true, in fact, if $\mu(x)$ has a positive derivative over any interval of continuity.

Similarly, the definition given in (8.12) has no meaning if the distribution is not absolutely continuous, and it cannot be applied to a general monotonic distribution. If our information process is one that involves a distribution function containing a set of discontinuities with total variation less than one, there exists no definition for the uncertainty associated with such a distribution. Any attempt to treat such a process from the entropy point of view would necessarily present a formidable problem.

In order to circumvent these difficulties, we shall simply introduce another "reasonable" definition for the information resulting from a change of distribution. Our purpose in introducing a new definition is to divorce the theory of information from a dependence on the entropy concept. In order to justify this definition, we shall merely show that it has the properties demanded by our intuition and that its application to specific processes gives results in complete agreement with those obtained by Wiener with the difference-of-entropy approach. The unified theory given here will be shown, in fact, to be a true generalization of the theory of Shannon and Woodward, and will include both their discrete and continuous theories as special cases. Although the definition of information given here will be essentially new in form, it will be shown to be in complete harmony with an already well-established theory of communication. In the communication problem, it will become an abstract generalization of the expression of Woodward for the information evaluated from the viewpoint of the receiver.

In the case of pure step distributions, the entropy will exist and will have the same physical interpretation as that given by Shannon, although its expression in terms of probabilities will be a derived one rather than a defining one.

We consider again the problem of the random variable f distributed a priori by $\rho(x)$ and a posteriori by $\nu(x)$. Let us suppose that the true value of f is y. A "reasonable" measure for the information associated with the fact that f = y is

$$I(y) = \lim_{\epsilon \to 0} \log \frac{\nu(y + \epsilon) - \nu(y - \epsilon)}{\rho(y + \epsilon) - \rho(y - \epsilon)} \quad . \tag{9.1}$$

in order to get the total information about f, we average I(y) over all possible y with respect to the a posteriori distribution; that is, with respect to our best knowledge concerning the value of f.

$$I = \int_{-\infty}^{\infty} \lim_{\epsilon \to 0} \log \frac{\nu(y + \epsilon) - \nu(y - \epsilon)}{\rho(y + \epsilon) - \rho(y - \epsilon)} \, d\nu(y) . \tag{9.2}$$

It is apparent that under certain conditions the integrand in (9.2) will fail to exist. First, consider the case in which $\rho(x)$ contains discontinuities not in common with those of $\nu(x)$. The integrand becomes infinitely negative at such points. However, since $\rho(x)$ is monotonic and bounded, the set of all its discontinuities must be countable — hence also the subset of those not in common with discontinuities of $\nu$. Clearly, this subset is of $\nu$-measure zero and the value of the integral is unaffected by the divergence of the integrand.

Next, consider the case in which $\nu(x)$ has discontinuities not in common with those of $\rho(x)$. The set of these discontinuities is of positive $\nu$-measure and, since the integrand becomes positively infinite on this set, the integral diverges. Thus a necessary condition for the finiteness of I is that any point of discontinuity of $\nu(x)$ be a point of discontinuity of $\rho(x)$ also.

These considerations suggest that the integral in (9.2) can be finite only if the measure $\nu$ is absolutely continuous with respect to the measure $\rho$, in which case there exists by the Radon-Nikodym theorem a derivative $d\nu/d\rho$. The form of the integral in (9.2) can be made more compact by interpreting

$$\lim_{\epsilon \to 0} \frac{\nu(y + \epsilon) - \nu(y - \epsilon)}{\rho(y + \epsilon) - \rho(y - \epsilon)}$$

as a derivative in the Radon-Nikodym sense, and our definition takes the more general form

$$I = \int_X \log \frac{d\nu}{d\rho} \, d\nu \, . \tag{9.3}$$

Let us now show that this definition is a valid generalization of (8.2). Hence it reduces to the fundamental definition given by Wiener when applied to an information process in which the range of a random variable is reduced to a subset of its a priori range. We consider again a random variable f taking values on a measure space $(X, \mathcal{S}, \mu)$. If $\mu(x)$ is absolutely continuous, the fact that $f \epsilon A$ may be expressed by an a priori density

$$\rho'(x) = \begin{cases} \mu'(x) \;/\mu(A) & x \epsilon A \\ 0 & \text{otherwise} \end{cases} \tag{9.4}$$

Or, more generally, we can relax the absolute continuity condition and write

$$\frac{d\rho}{d\mu} = \begin{cases} 1/\mu(A) & x \epsilon A \\ 0 & \text{otherwise.} \end{cases} \tag{9.5}$$

Similarly, the additional knowledge $f \epsilon A \cap B$ may be formulated in terms of the a posteriori measure $\nu$:

$$\frac{d\nu}{d\mu} = \begin{cases} 1/\mu(A \cap B) & x \epsilon A \cap B \\ 0 & \text{otherwise.} \end{cases} \tag{9.6}$$

Since we may write

$$\frac{d\rho}{d\mu} = \chi_A(x)/\mu(A), \tag{9.7}$$

and

$$\frac{d\nu}{d\mu} = \chi_{A \cap B}(x)/\mu(A \cap B), \tag{9.8}$$

it follows that

$$\rho(E) = \frac{1}{\mu(A)} \int_E \chi_A(x) \, d\mu = \frac{\mu(E \cap A)}{\mu(A)} \, , \tag{9.9}$$

and

$$\nu(E) = \frac{1}{\mu(A \cap B)} \int_E \chi_{A \cap B}(x)\ d\mu = \frac{\mu(A \cap B \cap E)}{\mu(A \cap B)} \tag{9.10}$$

for every measurable set E. Hence $\mu(E) = 0$ implies both $\rho(E) = 0$ and $\nu(E) = 0$ (provided, of course, that $\mu(A \cap B) \neq 0$). Thus $\rho \sim \mu$ and $\nu \sim \mu$. Also, if $\rho(E) = 0$, then (with $\mu(A) \neq 0$) $\mu(E \cap A) = 0$. But $A \cap B \cap E$ is a subset of $E \cap A$ thus $\mu(A \cap B \cap E)$, hence $\nu(E)$, is zero also. We have then $\nu \sim \rho \sim \mu$, and the information for this case becomes

$$I = \int_X \log \frac{d\nu}{d\rho}\, d\nu = \int_X \frac{d\nu}{d\mu} \left[ \log \frac{d\nu}{d\mu} - \log \frac{d\rho}{d\mu} \right] d\mu$$

$$= \int_{A \cap B} \frac{1}{\mu(A \cap B)} \left[ \log \frac{1}{\mu(A \cap B)} - \log \frac{1}{\mu(A)} \right] d\mu$$

$$= \log \frac{\mu(A)}{\mu(A \cap B)}, \tag{9.11}$$

in agreement with Equation (8.2). Thus (9.3) is a valid generalization of the first definition of Wiener.

Before showing that our information is nonnegative, we shall have need for the following logarithmic inequality:

LEMMA 9.1. Given a measure space $(X, \mathcal{S}, \mu)$ and a nonnegative function f, defined and integrable on a set E, it will follow that

$$\int_E f \log f\ d\mu \geqslant \int_E (f - 1)\ d\mu.$$

PROOF. Consider the following decomposition of the set E: Let $A = \{x: f(x) \geqslant 1\}$.

$$\int_E f \log f\ d\mu = \int_A |f \log f|\ d\mu - \int_{E-A} |f \log f|\ d\mu$$

$$\geqslant \int_A |f - 1|\ d\mu - \int_{E-A} |f - 1|\ d\mu \tag{9.12}$$

$$= \int_A (f - 1)\ d\mu + \int_{E-A} (f - 1)\ d\mu = \int_E (f - 1)\ d\mu.$$

THEOREM 9.2. The information resulting from a change of defining measure on a probability space is nonnegative.

PROOF. It suffices to consider only the case in which $\nu \sim \rho$.

$$I = \int_X \log \frac{d\nu}{d\rho}\, d\nu = \int_X \frac{d\nu}{d\rho} \log \frac{d\nu}{d\rho}\, d\rho$$

$$\geqslant \int_X \left(\frac{d\nu}{d\rho} - 1\right) d\rho = \nu(X) - \rho(X) = 0.$$

Here we have made use of the fact that $\rho$ and $\nu$ are probability measures; hence from the Radon-Nikodym theorem that $d\nu/d\rho$ is nonnegative and integrable. We used also Lemmas 3.7 and 9.1.

We shall employ the term <u>information process</u> to represent that physical process whose mathematical model may be regarded as a change of defining measure on an abstract probability space. In other words, the mathematical model of an information process is a probability measure space $(X, \mathcal{S}, \rho, \nu)$, where $\rho$ and $\nu$ are the a priori and a posteriori measures associated with the process. We shall frequently speak of "the information process $(X, \mathcal{S}, \rho, \nu)$" wherein we imply the existence of an actual physical process involving a change of probability distribution.

DEFINITION 9.1. The <u>information</u> resulting from an information process $(X, \mathcal{S}, \rho, \nu)$ is defined by

$$I = \int_X \log \frac{d\nu}{d\rho}\, d\nu , \qquad (9.13)$$

if $\nu \sim \rho$, and $+\infty$ otherwise.

This definition is sufficiently general to cover a very large number of special cases. In the first place, the formulation in terms of an abstract space does not in any way limit the dimensionality of the process. In an n-dimensional process, the space X simply becomes an n-dimensional product space. Furthermore, since the definition is independent of any fixed coordinate system, the value of the information is invariant under transformation of coordinates. It applies equally well to information processes in which the distribution functions are either absolutely continuous, purely discrete step functions, or even functions with discontinuities whose total variation is less than one.

It is perhaps of value to show by means of certain examples just what part is played by the Radon-Nikodym derivative in the evaluation of the information resulting from specific processes.

EXAMPLE 9.1. Let $\rho(x)$ and $\nu(x)$ represent a priori and a posteriori distribution functions which are absolutely continuous with respect to Lebesgue measure. Under these conditions, the densities $\rho'(x)$ and $\nu'(x)$ exist. Assuming, in addition, that $\nu \rightsquigarrow \rho$, the Radon-Nikodym derivative becomes a derivative in the ordinary sense, and is simply the ratio of the a posteriori and a priori probability densities. The information is then given by the improper Riemann integral

$$I = \int_{-\infty}^{\infty} \nu'(x) \log \frac{\nu'(x)}{\rho'(x)} dx. \tag{9.14}$$

EXAMPLE 9.2. Let $\rho(x)$ be a monotonic step-function with a countable number of discontinuities of magnitude $p_k$ on a set S of elements $x_k$. Let $\nu(x)$ be a similar function except that its discontinuities have magnitude $q_k$ and occur on a subset M of S. Since every discontinuity of $\nu(x)$ occurs in common with one of $\rho(x)$, and since every subset of X disjoint with S is of both $\rho$- and $\nu$- measure zero, it follows that $\nu \rightsquigarrow \rho$. The Radon-Nikodym derivative is a function whose value is $q_k/p_k$ at every $x_k \epsilon M$ and zero at every $x_k \epsilon (S - M)$. This function is left undefined at all other points of X which form, of course, a set of $\rho$-measure zero. The Stieltjes integral (9.13) becomes in this case the simple summation

$$I = \sum_{x_k \epsilon M} q_k \log \frac{q_k}{p_k} \tag{9.15}$$

over all the points $x_k$ of M.

EXAMPLE 9.3. For this example, we treat a mixed process involving the distributions $\rho(x)$ and $\nu(x)$ of Fig. 5. For simplicity, we have chosen a small number of discontinuities. It is seen that every discontinuity of $\nu(x)$ occurs in common with one of $\rho(x)$. Also, the interval $(x_3 < x \leqslant x_4)$, which is of $\rho$-measure zero, may be seen to be of $\nu$-measure zero also. We assume that, in the vicinity to the left of $x_3$, the derivative $\rho'(x_3 - \epsilon) > 0$ for all $\epsilon > 0$ but that

$$\lim_{\epsilon \to 0} \rho'(x - \epsilon) = 0.$$

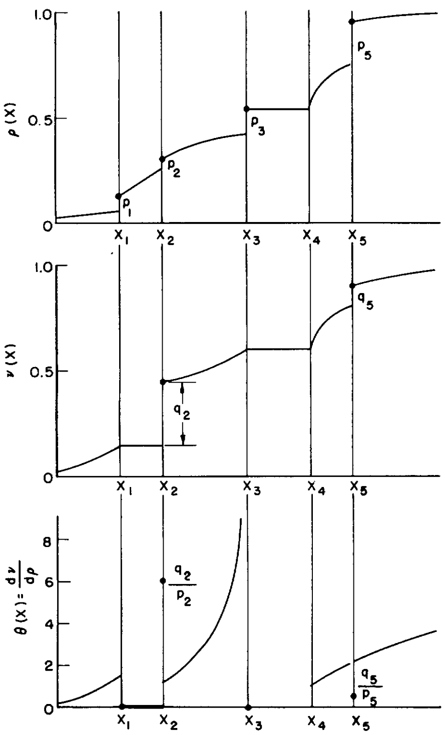

Fig. 5. A mixed information process.

Under these assumptions, it is easy to verify that $\nu \rightsquigarrow \rho$. Thus the Radon-Nikodym derivative $d\nu/d\rho$ is defined $[\rho]$. This function, which we call $\vartheta(x)$, is plotted in Fig. 5. We note first of all that $\vartheta(x)$ must be well defined at all points of discontinuity of $\rho(x)$ and must have the value $p_k/q_k$ at such points $x_k$. These values are indicated by the heavy dots in the figure. With the recognition that any monotonic function can be expressed as the sum of a continuous function and a step-function, $\vartheta(x)$ is given at all other points by the ratio $\nu_c'(x)/\rho_c'(x)$, where the subscript c denotes the continuous part. In this manner, we define $\vartheta(x)$ at all points with the possible exception of a set of $\rho$-measure zero. From Fig. 5, we see that $\vartheta(x)$ is left undefined in the interval $(x_3 < x \leqslant x_4)$, which is indeed of zero $\rho$-measure. However, from the absolute continuity condition, any set of $\rho$-measure zero must be also of $\nu$-measure zero — hence the integral in (9.13) with respect to $\nu$ is unaffected by any values we might assign to $\vartheta(x)$ in the undefined interval. After determining the Radon-Nikodym derivative, it is an easy matter to verify that the Stieltjes integral

$$\nu(x) = \int_{-\infty}^{x} \vartheta(\xi)\, d\rho(\xi) \qquad (9.16)$$

is valid.

For the mixed case, the information resulting from the process of this example is given by the Stieltjes integral

$$I = \int_{-\infty}^{\infty} \log \vartheta(x)\, d\nu(x)$$

$$= q_2 \log \frac{q_2}{p_2} + q_5 \log \frac{q_5}{p_5} + \int_{-\infty}^{\infty} \nu_c'(x) \log \frac{\nu_c'(x)}{\rho_c'(x)}\, dx. \qquad (9.17)$$

Thus the information received in a mixed process is simply the sum of a discrete and a continuous part.

This example emphasizes the fact that the assertion of the Radon-Nikodym theorem that the function $\vartheta(x)$ be finite-valued does not imply its boundedness. In Fig. 5, it can be seen that $\vartheta(x)$ does become unbounded in the neighborhood to the left of the point $x_3$. However, $\vartheta(x)$ has the value zero at the point $x_3$ and is indeed finite-valued everywhere.

We might also note from this example that, although the condition $\nu \sim \rho$ is a necessary one for the finiteness of the information I, it is not a sufficient one. Since $\vartheta(x)$ is unbounded, the convergence of the integral in (9.17) depends upon the behavior of log $\vartheta(x)$ in the neighborhood to the left of $x_3$. It is an easy matter to draw examples for which divergence results even though $\nu$ is assumed absolutely continuous with respect to $\rho$.

## 10. The Communication Problem

Let us consider again the simple communication problem of section 8 in the light of the unified definition and show how the results agree with those obtained by Shannon in both the discrete and absolutely continuous cases. We consider a source which generates a sequence $\{f_i\}$ of independent random variables of values x selected from the measure space $(X, \mathcal{S}, \mu)$. These values are transformed by the channel into an output sequence $\{g_i\}$ of independent elements taking values y on the space $(Y, \mathfrak{T}, \nu)$. The channel is defined in terms of the conditional measure $\nu_x$ on the Y space. We then consider the product space $(X \times Y, \mathcal{S} \times \mathfrak{T}, \lambda)$ of pairs (x,y) representing the possible values of the pair $(f_k, g_k)$. Here $\lambda$ is the measure defined, for an arbitrary Borel set $A \in \mathcal{S} \times \mathfrak{T}$, by

$$\lambda(A) = \int_X \nu_x(A_x) \, d\mu, \tag{10.1}$$

where $A_x$ is the x-section of A as defined in section 2. By the methods used in that section, the measure $\nu$ may be found from the relation

$$\nu(F) = \int_X \nu_x(F) \, d\mu. \tag{10.2}$$

Having determined the distribution functions $\lambda(x,y)$ and $\nu(y)$, we obtain the conditional distribution $\mu_y(x)$ from

$$\mu_y(x) = \frac{\partial \lambda(x,y)}{\partial \nu(y)} \quad [\nu]. \tag{10.3}$$

Accordingly, from a knowledge of $\mu$ and $\nu_x$, which are defined by the source and channel, we can determine all product and component measures on the product space X × Y. It is clear that, for almost every value y of the channel output, there is defined an information process $(X, \mathscr{S}, \mu, \mu_y)$. The information resulting from such a process is (for $\mu_y \sim \mu\ [\nu]$)

$$I(y) = \int_X \log \frac{\partial \mu_y}{\partial \mu}\, d\mu_y \qquad [\nu]. \tag{10.4}$$

I(y) represents the amount of information about the value x from the source provided by a particular value y in the channel output. The average information (per element) provided by the sequence g about the source sequence f is the average of I(y) over all possible y. Thus the average rate of information becomes

$$R(f;g) = \int_Y \int_X \log \frac{\partial \mu_y}{\partial \mu}\, d\mu_y\, d\nu . \tag{10.5}$$

However, from Lemma 3.3 and Theorem 3.5, this can be expressed in terms of the measure $\lambda$ and the product measure $\rho$:

$$R(f;g) = \int_{X\times Y} \log \frac{d\lambda}{d\rho}\, d\lambda. \tag{10.6}$$

If we apply this result to the special case wherein $\mu(x)$ and $\nu_x(y)$ are absolutely continuous with respect to Lebesgue measure on X and Y, respectively, it will follow that both $\lambda(x,y)$ and $\rho(x,y)$ are absolutely continuous with respect to the Lebesgue product measure on X × Y. The Radon-Nikodym derivative becomes the ratio of the probability densities,

$$\begin{aligned} \frac{d\lambda}{d\rho} &= \frac{\partial^2 \lambda(x,y)}{\partial x\, \partial y} \Bigg/ \frac{\partial^2 \rho(x,y)}{\partial x\, \partial y} \\ &= \frac{p(x,y)}{p(x)p(y)} \quad , \end{aligned} \tag{10.7}$$

where $p(x) = \mu'(x)$ and $p(y) = \nu'(y)$. In the absolutely continuous case, the rate becomes

$$R(f;g) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}\, dx\, dy, \tag{10.8}$$

in complete agreement with the result of Shannon for such a case.

In the discrete case, the functions $\mu(x)$ and $\nu_x(y)$ are step-functions with discontinuities $p(i)$ and $p_i(j)$ on a countable set of elements $(x_i,y_j)$. The joint distributions $\lambda(x,y)$ and $\rho(x,y)$ are step-functions in both variables with discontinuities of magnitudes $p(i,j)$ and $p(i)p(j)$, respectively, on the same set of values $(x_i,y_j)$. The Radon-Nikodym derivative $d\lambda/d\rho$ becomes the ratio

$$\frac{p(i,j)}{p(i)p(j)}$$

at the points $(x_i,y_j)$ and is left undefined elsewhere. "Elsewhere" is, of course, a set of $\rho$-measure zero. For the discrete case, the Lebesgue-Stieltjes integral of (10.6) becomes the summation

$$R(f;g) = \sum_{i,j} p(i,j) \log \frac{p(i,j)}{p(i)p(j)} \qquad (10.9)$$

over the countable set of points $(x_i, y_j)$. This is, of course, Shannon's definition of the rate for discrete distributions.

It is easily seen that the integrand

$$\log \frac{d\lambda}{d\rho}$$

of (10.6), which is a function of the values $(x,y)$ of the joint random variable $(f,g)$, is the abstract generalization of Woodward's mutual information $I_{xy}$, and extends his results to the more general information processes.

## 11. Symmetry Relations

We consider a discrete stochastic process of the type studied by Khintchine[4]; that is, a family or ensemble of sequences $\xi = \{\xi_i\}$ $(i = \ldots,-1,0,1,\ldots)$ such that for any finite subset $x = (x_1,x_2,\ldots,x_n)$ of elements of $\xi$ there is defined a probability measure $\mu$ on an $n$-dimensional product space $X$. Let $x = (x_1,x_2,\ldots,x_n)$ and $y = (y_1,y_2,\ldots,y_m)$ be two such subsequences, disjoint from each other, taking values on the $n$- and $m$- dimensional spaces $(X,\mathcal{S},\mu)$ and $(Y,\mathfrak{T},\nu)$, respectively. Let the space $(X \times Y, \mathcal{S} \times \mathfrak{T}, \lambda)$ represent the $(m + n)$-dimensional space of all pairs $(x,y)$. In addition to the absolute measures $\mu$ and $\nu$, there will exist, for almost every $y$ and $x$, conditional measures $\mu_y$ and $\nu_x$ on the component $\sigma$-algebras $\mathcal{S}$ and $\mathfrak{T}$. It will be of interest to formulate the information about the sequence $x$ that is provided on the average by the specification of the value of the sequence y. From the discussion in section 10 it is clear that, for almost every y, the measure space $(X, \mathscr{B}, \mu, \mu_y)$ forms an information process resulting in an amount of information given by

$$I(y) = \int_X \log \frac{\partial \mu_y}{\partial \mu} \, d\mu_y \qquad [\nu] \tag{11.1}$$

provided, of course, that $\mu_y \sim \mu \; [\nu]$. That is, any particular value of y (with the exception of a set of values of $\nu$-measure zero) provides an information of amount I(y) about the value of the sequence x. In order to obtain the amount of the information about x given on the average by the sequence y, we take the mean value of I(y) over all possible values of y. Consequently,

$$\overline{I(x;y)} = \int_Y \int_X \log \frac{\partial \mu_y}{\partial \mu} \, d\mu_y \, d\nu \tag{11.2}$$

is the average information about x provided by y.

LEMMA 11.1. If x and y are a pair of disjoint subsequences of a particular realization of a discrete stochastic process, and $\overline{I(x;y)}$ is the average information about x provided by y, then

$$\overline{I(x;y)} = \overline{I(y;x)} \tag{11.3}$$

PROOF. From Lemma 3.3 and Theorem 3.5,

$$\begin{aligned} \overline{I(x;y)} &= \int_Y \int_X \log \frac{\partial \mu_y}{\partial \mu} \, d\mu_y \, d\nu \\ &= \int_{X \times Y} \log \frac{\partial \nu_x}{\partial \nu} \, d\lambda \\ &= \int_X \int_Y \log \frac{\partial \nu_x}{\partial \nu} \, d\nu_x \, d\mu \\ &= \overline{I(y;x)}. \end{aligned} \tag{11.4}$$

Let $z = (z_1, z_2, \ldots, z_\ell)$ be another subsequence of $\xi$, disjoint with both x and y and taking values on an $\ell$-dimensional space $(Z, \mathcal{U}, \sigma)$.

We wish to formulate the expression for the average information about x provided by y *when the value of z is known a priori*. Clearly, for almost every z, there will exist a conditional measure $\lambda_z$ on the measurable space $(X \times Y, \mathcal{S} \times \mathcal{T})$. For almost every (y,z) and (x,z), there will exist conditional measures $\mu_{yz}$ and $\nu_{xz}$ on the component spaces X and Y. The information about x provided by a particular value y, given a priori a fixed value z, becomes

$$\int_X \log \frac{\partial \mu_{yz}}{\partial \mu_z} \, d\mu_{yz}.$$

The average of this expression over all y, but for a fixed z, becomes

$$\int_Y \int_X \log \frac{\partial \mu_{yz}}{\partial \mu_z} \, d\mu_{yz} \, d\nu_z.$$

Taking the additional average over the Z-space, we obtain

$$\overline{I(x;y|z)} = \int_Z \int_Y \int_X \log \frac{\partial \mu_{yz}}{\partial \mu_z} \, d\mu_{yz} \, d\nu_z \, d\sigma, \qquad (11.5)$$

which represents the information about x provided on the average by y when the value of z is known.

LEMMA 11.2. If x,y, and z are disjoint subsequences of *a particular* realization of a discrete stochastic process, and I(x;y|z) is the average information about x provided by y when z is known, then

$$\overline{I(x;y|z)} = \overline{I(y;x|z)}.$$

PROOF. From the application of Lemma 3.3 to the space $(X \times Y, \mathcal{S} \times \mathcal{T}, \lambda_z)$, which exists for almost every z, it follows that for almost every (x,y,z)

$$\frac{\partial \mu_{zy}}{\partial \mu_z} = \frac{\partial \nu_{zx}}{\partial \nu_z}. \qquad (11.6)$$

Also, from Theorem 3.5, we have for almost every z

$$\int_Y \int_X \log \frac{\partial \mu_{zy}}{\partial \mu_z} \, d\mu_{zy} \, d\nu_z = \int_X \int_Y \log \frac{\partial \mu_{zy}}{\partial \mu_z} \, d\nu_{zy} \, d\mu_z. \qquad (11.7)$$

Thus

$$\int_Z \int_Y \int_X \log \frac{\partial \mu_{zy}}{\partial \mu_z} \, d\mu_{zy} \, d\nu_z \, d\sigma = \int_Z \int_X \int_Y \log \frac{\partial \nu_{zx}}{\partial \nu_z} \, d\nu_{zx} \, d\mu_z \, d\sigma , \tag{11.8}$$

which is equivalent to the assertion of the lemma.

Lemma 11.2 is merely a restatement of Lemma 11.1 with the inclusion of an a priori condition. It is not surprising that the symmetry relation expressed in Lemma 11.1 is independent of the a priori knowledge, hence that an auxilliary condition does not destroy this symmetry.

It should be noted that in the proofs of Lemmas 11.1 and 11.2, no explicit use was made of the fact that the spaces X,Y, and Z are finite-dimensional. In fact, from the extension of the product space theory to a space of infinite dimensions, it follows that both of these lemmas are valid, even though any of the three sequences x,y, and z are infinite subsequences of $\xi$. The requirement that they be disjoint, however, is necessary for preserving the meaning of an ordered triplet (x,y,z).

## 12. Additivity of Information

One of the very important properties of the logarithmic measure of information is that under suitable conditions the joint information provided about a pair of independent events is the sum of the informations given separately about the individual events. For example, it was shown by Wiener and Shannon that the entropy associated with the joint random variable with values (x,y) is, in the case of independence, the sum of the entropies associated with each variable. It is of interest to determine the conditions under which we can make similar statements concerning the information defined in section 9.

Let us consider the pair (f,g) of random variables defined independently on the spaces $(X, \mathcal{S}, \mu_1)$ and $(Y, \mathfrak{T}, \nu_1)$. The product measure $\rho_1 = \mu_1 \times \nu_1$ is defined on the Cartesian product X × Y of the <u>independent</u> spaces X and Y. Let us suppose that a certain amount of information is received which allows the formulation of a new a posteriori measure $\rho_2 = \mu_2 \times \nu_2$, which we assume to retain posession of the product property with regard to rectangles. That is, we assume that the information received has not destroyed the independence of the pair (f,g). The information processes $(X \times Y, \mathcal{S} \times \mathfrak{T}, \rho_1, \rho_2)$, $(X, \mathcal{S}, \mu_1, \mu_2)$, and $(Y, \mathfrak{T}, \nu_1, \nu_2)$ are associated with the pair (f,g), the random variable f, and the random variable g, respectively. We shall now prove the following theorem:

THEOREM 12.1. Let $(X \times Y, \mathcal{S} \times \mathcal{T}, \rho_1, \rho_2)$ be an information process associated with the pair (f,g) of random variables which are both a priori and a posteriori independent. If $\overline{I(f,g)}$ is the information resulting from the above process, with $\overline{I(f)}$ and $\overline{I(g)}$ that resulting from the processes $(X, \mathcal{S}, \mu_1, \mu_2)$ and $(Y, \mathcal{T}, \nu_1, \nu_2)$, respectively, where $\rho_1 = \mu_1 \times \nu_1$ and $\rho_2 = \mu_2 \times \nu_2$, it follows that

$$\overline{I(f,g)} = \overline{I(f)} + \overline{I(g)}. \tag{12.1}$$

PROOF. Let $E \times F$ be an arbitrary rectangle in $X \times Y$. Then

$$\begin{aligned} \rho_2(E \times F) &= \int_E \nu_2(F)\, d\mu_2 = \int_E \nu_2(F) \frac{d\mu_2}{d\mu_1}\, d\mu_1 \\ &= \int_E \int_F \frac{d\nu_2}{d\nu_1} \frac{d\mu_2}{d\mu_1}\, d\nu_1\, d\mu_1 \\ &= \int_{E\times F} \frac{d\nu_2}{d\nu_1} \frac{d\mu_2}{d\mu_1}\, d\rho_1 \end{aligned} \tag{12.2}$$

must hold for all $E \times F$. Since any measurable set A in $X \times Y$ can be covered by a countable union of disjoint rectangles, it follows that

$$\rho_2(A) = \int_A \frac{d\nu_2}{d\nu_1} \frac{d\mu_2}{d\mu_1}\, d\rho_1 \tag{12.3}$$

for every $A \in \mathcal{S} \times \mathcal{T}$. Thus

$$\frac{d\rho_2}{d\rho_1} = \frac{d\nu_2}{d\nu_1} \frac{d\mu_2}{d\mu_1} \qquad [\rho_1, \rho_2]. \tag{12.4}$$

The information becomes

$$\begin{aligned} \overline{I(f,g)} &= \int_{X \times Y} \log \frac{d\rho_2}{d\rho_1}\, d\rho_2 = \int_X \int_Y \log \frac{d\nu_2}{d\nu_1} \frac{d\mu_2}{d\mu_1}\, d\nu_2\, d\mu_2 \\ &= \int_X \log \frac{d\nu_2}{d\nu_1}\, d\nu_2 + \int_Y \log \frac{d\mu_2}{d\mu_1}\, d\mu_2 \\ &= \overline{I(f)} + \overline{I(g)}, \end{aligned} \tag{12.5}$$

which was to be proven.

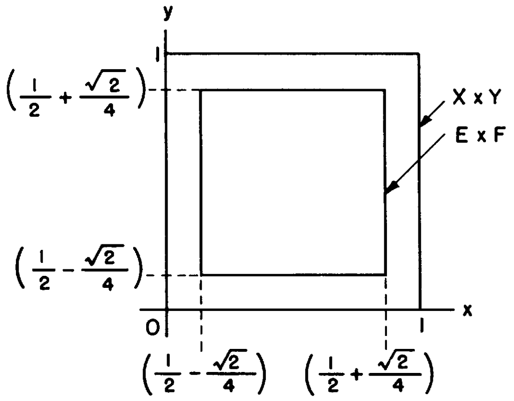
y
1
$\left(\frac{1}{2}+\frac{\sqrt{2}}{4}\right)$
$\left(\frac{1}{2}-\frac{\sqrt{2}}{4}\right)$
X x Y
E x F
x
0
1
$\left(\frac{1}{2}-\frac{\sqrt{2}}{4}\right)$
$\left(\frac{1}{2}+\frac{\sqrt{2}}{4}\right)$

Fig. 6

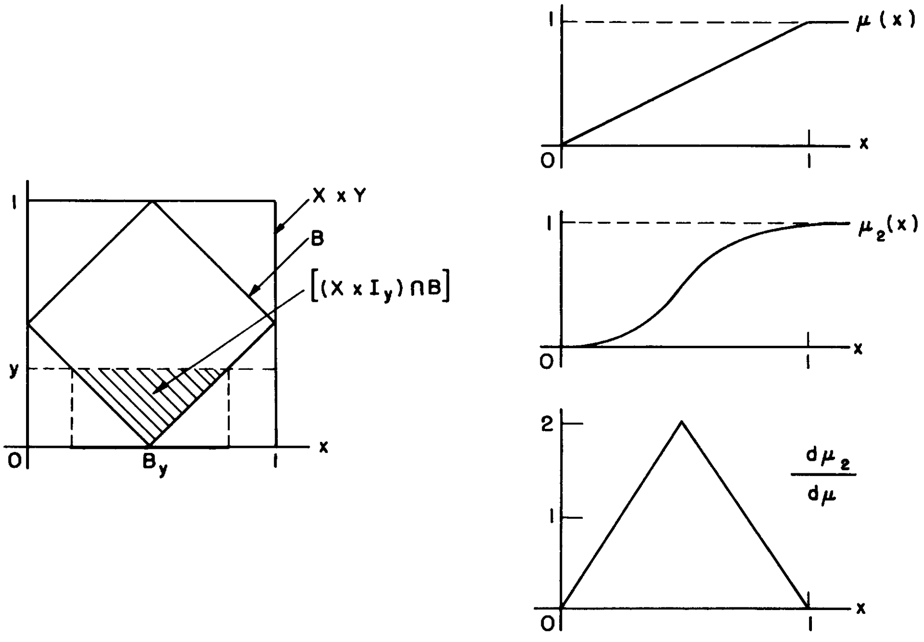
1
X x Y
B
$[(X \times I_y) \cap B]$
y
0
$B_y$
1
x
1
$\mu(x)$
0
1
x
1
$\mu_2(x)$
0
1
x
2
1
$\frac{d\mu_2}{d\mu}$
0
1
x

Fig. 7

In order to show the necessity for the assumption that f and g are both a priori and a posteriori independent, it is perhaps best to illustrate by means of the following examples:

EXAMPLE 12.1. Let f and g be independent random variables distributed uniformly in the interval (0,1). The spaces $(X, \mathcal{S}, \mu)$ and $(Y, \mathfrak{T}, \nu)$ are then the unit interval and the measures $\mu$ and $\nu$ are both Lebesgue measure. The Cartesian product space $(X \times Y, \mathcal{S} \times \mathfrak{T}, \rho)$ is then the unit square with $\rho = \mu \times \nu$ Lebesgue measure on the plane.

With the foregoing as a priori knowledge, let us assume that we are told, in addition, that the value (x,y) of the pair (f,g) lies inside the set $E \times F$, which is shown in Fig. 6. From (9.11), the information provided by the pair (f,g) is given by

$$\overline{I(f,g)} = \log \frac{\rho(X \times Y)}{\rho(E \times F)} = - \log \mu(E) \ \nu(F)$$

$$= - \log \frac{\sqrt{2}}{2} \cdot \frac{\sqrt{2}}{2} = 1 \text{ bit.} \tag{12.6}$$

In order to evaluate the information obtained about f and g, we note that the a posteriori measure $\lambda$, which is defined on $\mathcal{S} \times \mathfrak{T}$, is given by

$$\lambda(A) = \frac{\rho\left[A \cap (E \times F)\right]}{\rho(E \times F)} \tag{12.7}$$

for every measurable set A. If we let $\mu_2$ and $\nu_2$ be the absolute measures on the components X and Y of the product space $(X \times Y, \mathcal{S} \times \mathfrak{T}, \lambda)$, it follows that for every $E_1 \epsilon \mathcal{S}$

$$\mu_2(E_1) = \lambda(E_1 \times Y) = \frac{\rho\left[(E_1 \times Y) \cap (E \times F)\right]}{\rho(E \times F)}$$

$$= \frac{\rho\left[(E_1 \cap E) \times (F \cap Y)\right]}{\rho(E \times F)}$$

$$= \frac{\mu(E_1 \cap E) \ \nu(F \cap Y)}{\mu(E) \ \nu(F)} = \frac{\mu(E_1 \cap E)}{\mu(E)} = \mu_E(E_1), \tag{12.8}$$

which is the conditional measure of $E_1$ relative to E. Similarly for every $F_1 \epsilon \mathfrak{T}$,

$$\nu_2(F_1) = \nu_F(F_1). \tag{12.9}$$

Since

$$\lambda(E_1 \times F_1) = \frac{\rho\left[(E_1 \times F_1) \cap (E \times F)\right]}{\rho(E \times F)}$$

$$\lambda(E_1 \times F_1) = \frac{\rho\left[(E_1 \cap E) \times (F_1 \cap F)\right]}{\rho(E \times F)}$$

$$= \frac{\mu(E_1 \cap E) \quad \nu(F_1 \cap F)}{\mu(E) \quad \nu(F)}$$

$$= \mu_E(E_1) \quad \nu_F(F_1), \tag{12.10}$$

it is clear that $\lambda$ is a product measure of independent spaces. Therefore f and g are a posteriori, as well as a priori, independent.

We might note at this point that for every $y \epsilon F$, the a posteriori conditional distribution $\mu_{2y}(x)$ exists and is uniform on E but generates zero probability on the set X - E. This, of course, is identical with the absolute measure $\mu_2 = \mu_E$. For $y \not\epsilon F$, no conditional measure is defined on X, but the set Y - F is of zero $\nu_2$-measure. We have, then, $\mu_{2y} = \mu_2 \quad [\nu_2]$.

The information provided about f is

$$\overline{I(f)} = \int_X \log \frac{d\mu_E}{d\mu} \; d\mu_E. \tag{12.11}$$

We note that, since for every $E_1 \epsilon \mathscr{S}$,

$$\int_{E_1} \frac{d\mu_E}{d\mu} \, d\mu = \mu_E(E_1) = \frac{\mu(E_1 \cap E)}{\mu(E)} = \int_{E_1 \cap E} \frac{1}{\mu(E)} \quad d\mu$$

$$= \int_{E_1} \frac{\chi_E(x)}{\mu(E)} \, d\mu, \tag{12.12}$$

we have

$$\frac{d\mu_E}{d\mu} = \begin{cases} 1/\mu(E) & x \epsilon E \\ 0 & x \not\epsilon E. \end{cases} \qquad [\mu]. \tag{12.13}$$

Thus

$$\overline{I(f)} = \int_X \frac{d\mu_E}{d\mu} \log \frac{d\mu_E}{d\mu} \; d\mu$$

$$= \int_E \frac{1}{\mu(E)} \log \frac{1}{\mu(E)} \; d\mu$$

$$\overline{I(f)} = -\log \mu(E) = -\log \frac{\sqrt{2}}{2} = \frac{1}{2} \text{ bit.} \qquad (12.14)$$

From a similar treatment,

$$\overline{I(g)} = -\log \nu(F) = \frac{1}{2} \text{ bit,} \qquad (12.15)$$

and in this case

$$\overline{I(f,g)} = \overline{I(f)} + \overline{I(g)}. \qquad (12.16)$$

EXAMPLE 12.2. Let us assume the same a priori knowledge as in Example 12.1, and suppose that we receive information which tells us $(x,y) \in B$, where B is the set shown in Fig. 7. It is seen that B is simply the set $E \times F$ of the previous example rotated by 45 degrees. In this case, the a posteriori measure $\lambda(A)$ of sets $A \in \mathscr{S} \times \mathfrak{T}$, is given by

$$\lambda(A) = \frac{\rho(A \cap B)}{\rho(B)}. \qquad (12.17)$$

The a posteriori absolute distribution functions on the component spaces are given by

$$\mu_2(x) = \lambda(I_x \times Y) = \frac{\rho\left[(I_x \times Y) \cap B\right]}{\rho(B)} \qquad (12.18)$$

$$\nu_2(y) = \lambda(X \times I_y) = \frac{\rho\left[(X \times I_y) \cap B\right]}{\rho(B)} \quad . \qquad (12.19)$$

In this example, however, for every value y, there exists a conditional measure $\mu_{2y}$ on the space X which is not equal to $\mu_2$. For every value y, the conditional distribution for the value x is uniform on the section $B_y$ and generates zero probability on $(X - B_y)$. Thus the information received has effectively destroyed the statistical independence between f and g. It is found from (12.18) above and from the fact that the a priori distribution is given by $\mu(x) = x$ on (0,1) that the Radon-Nikodym derivative $d\mu_2/d\mu$ is the triangular-shaped function of Fig. 7. By the symmetry of the problem in x and y, it follows that

$$\overline{I(g)} = \overline{I(f)} = \int_X \frac{d\mu_2}{d\mu} \log \frac{d\mu_2}{d\mu} \, d\mu$$

$$= \int_0^{\frac{1}{2}} 4x \log 4x \, dx + \int_{\frac{1}{2}}^1 (-4x + 4) \log (-4x + 4) \, dx$$

$$\overline{I(g)} = \overline{I(f)} = \log 2 - \frac{1}{2}\log e$$

$$= 1 - \log_2\sqrt{e} \quad \text{bits.} \tag{12.20}$$

However, the information about the pair (f,g) is given by

$$\overline{I(f,g)} = \log \frac{\rho(X \times Y)}{\rho(B)} = 1 \text{ bit,} \tag{12.21}$$

which is the same as that of Example 12.1. Thus in this case,

$$\overline{I(f,g)} > \overline{I(f)} + \overline{I(g)}, \tag{12.22}$$

even though f and g were a priori independent.

It is desirable to develop a slightly more general additivity relation which requires no condition concerning independence. Such a relation does exist and we shall state it as a theorem concerning the information associated with a stochastic process.

Let $\xi = \{\xi_i\}$ $(i = \ldots,-1,0,1,\ldots)$ be a particular realization of a stochastic process and let w, x, y, and z represent disjoint subsequences of $\xi$ taking values on the spaces $(W,\mathcal{V},\omega)$, $(X,\mathcal{S},\mu)$, $(Y,\mathcal{T},\nu)$, and $(Z,\mathcal{U},\sigma)$, respectively. The information about the pair (x,y) provided on the average by the sequence z, when the sequence w is known a priori, is given by

$$\overline{I(x,y;z|w)} = \int_W \int_Z \int_{X\times Y} \log \frac{\partial\lambda_{wz}}{\partial\lambda_w}\, d\lambda_{wz}\, d\sigma_w\, d\omega, \tag{12.23}$$

where $\lambda_{wz}$ and $\lambda_w$ are conditional measures on the space $(X \times Y, \mathcal{S} \times \mathcal{T}, \lambda)$. We can write (12.23) as

$$\overline{I(x,y;z|w)} = \int_W \int_Z \int_X \int_Y \log \frac{\partial\lambda_{wz}}{\partial\lambda_w}\, d\nu_{wzx}\, d\mu_{wz}\, d\sigma_w\, d\omega. \tag{12.24}$$

In order to decompose the integrand, we make use of the following lemma:

LEMMA 12.2. Let $(X \times Y \times Z, \mathcal{S} \times \mathcal{T} \times \mathcal{U})$ be the Cartesian product of the measure spaces $(X \times Y, \mathcal{S} \times \mathcal{T}, \lambda)$ and $(Z, \mathcal{U}, \sigma)$. Let $\lambda_z$ represent the conditional measure on $X \times Y$. Then

$$\frac{\partial\lambda_z}{\partial\lambda} = \frac{\partial\nu_{zx}}{\partial\nu_x}\,\frac{\partial\mu_z}{\partial\mu} = \frac{\partial\mu_{zy}}{\partial\mu_y}\,\frac{\partial\nu_z}{\partial\nu} \quad [\lambda], \tag{12.25}$$

where $\nu_{zx}$, $\nu_x$ and $\nu$ are conditional and absolute measures on Y, while $\mu_{zy}$, $\mu_y$ and $\mu$ are the corresponding measures on X.

PROOF. Let E × F be an arbitrary rectangle in X × Y. Then

$$\int_{E\times F} \frac{\partial\lambda_z}{\partial\lambda}\, d\lambda = \lambda_z(E \times F) = \int_E \int_F d\,\nu_{zx}\, d\mu_z$$

$$= \int_E \int_F \frac{\partial\nu_{zx}}{\partial\nu_x}\, \frac{\partial\mu_z}{\partial\mu}\, d\,\nu_x\, d\mu$$

$$= \int_{E\times F} \frac{\partial\nu_{zx}}{\partial\nu_x}\, \frac{\partial\mu_z}{\partial\mu}\, d\lambda. \qquad (12.26)$$

Since this expression must hold for all rectangles E × F, and since every measurable set in X × Y can be covered by a countable union of disjoint rectangles, the integrands must be equal $[\lambda]$. A similar treatment proves the second assertion.

We may thus write

$$\overline{I(x,y;z|w)} = \int_W \int_Z \int_X \int_Y \log \frac{\partial\nu_{wzx}}{\partial\nu_{wx}}\, \frac{\partial\mu_{wz}}{\partial\mu_w}\, d\,\nu_{wzx}\, d\mu_{wz}\, d\,\sigma_w\, d\omega$$

$$= \int_W \int_Z \int_X \int_Y \log \frac{\partial\nu_{wzx}}{\partial\nu_{wx}}\, d\,\nu_{wzx}\, d\mu_{wz}\, d\,\sigma_w\, d\omega$$

$$+ \int_W \int_Z \int_Y \log \frac{\partial\mu_{wz}}{\partial\mu_w}\, d\mu_{wz}\, d\,\sigma_w\, d\omega$$

$$= \overline{I(x;z|w)} + \overline{I(y;z|w,x)}. \qquad (12.27)$$

The following theorem has been established:

THEOREM 12.3. Let w, x, y, and z be disjoint subsequences of a particular realization $\xi = \{\xi_i\}$ of a discrete stochastic process. Let $\overline{I(x,y;z|w)}$ be the information about the pair (x,y) provided on the average by z when w is known a priori. Then

$$\overline{I(x,y;z|w)} = \overline{I(x;z|w)} + \overline{I(y;z|w,x)}. \qquad (12.28)$$

This theorem allows a decomposition of the information provided about a pair (x,y) of random variables with no restriction concerning their independence. If these variables are both a priori and a posteriori independent, the second term will become

$$\overline{I(y;z|w,z)} = \overline{I(y;z|w)} \tag{12.29}$$

and we shall have obtained a complete decomposition of the joint information into individual informations.

From the symmetry property of the average information, we can write (12.28) as

$$\overline{I(z;x,y|w)} = \overline{I(z;x|w)} + \overline{I(z;y|w,x)}.$$

Hence, the information from a pair (x,y) of sources may be decomposed similarly into the sum of the individual informations from each source.

## 13. Communication in the Presence of Additive Gaussian Noise

We include in this section, for the sake of completeness, a result of Shannon concerning a channel in which a gaussian noise signal is added to the message. We assume the message and noise to be statistically independent. Let $x = (x_1,x_2,\ldots,x_n)$ be the value of a particular subsequence of the message; and $n = (n_1,n_2,\ldots,n_n)$, the value of a corresponding noise subsequence. It is clear that the channel output y will have the value $(x_1 + n_1, x_2 + n_2, \ldots, x_n + n_n)$. Let the values x be governed by the probability density distribution p(x), while the noise has gaussian density q(n). Since the message and noise are statistically independent, the conditional distribution density of the values y for a given x becomes simply

$$p_x(y) = q(y - x). \tag{13.1}$$

The average rate (per symbol) at which the output y gives information about the input x becomes

$$\begin{aligned}\overline{I(x;y)} &= \lim_{n\to\infty} \frac{1}{n} \int_Y \int_X p(x,y) \log \frac{p_x(y)}{p(y)}\, dx\, dy \\ &= \lim_{n\to\infty} \frac{1}{n} \int_Y \int_X p(x)\, q(y - x) \log \frac{q(y - x)}{p(y)}\, dx\, dy \\ &= \lim_{n\to\infty} \frac{1}{n} \left[\int_Z q(z) \log q(z)\, dz - \int_Y p(y) \log p(y)\, dy\right].\end{aligned} \tag{13.2}$$

The first term is simply the negative of what Shannon calls the entropy of the noise; the second term is the entropy of the channel output. It is of interest to determine the capacity of such a channel; that is, to maximize $\overline{I(x;y)}$ with respect to the message distribution p(x). It is clear that, since the entropy of the noise is independent of p(x), the problem is simply one of maximizing

$$- \int_Y p(y) \log p(y)\, dy$$

for any fixed n, where

$$p(y) = \int_X p(x)\, q(y - x)\, dx. \tag{13.3}$$

However, Shannon showed that the entropy of any n-dimensional absolutely continuous distribution is a maximum when that distribution is gaussian. Thus, since q(y - x) is gaussian, it follows that if p(y) is gaussian then p(x) must also be gaussian. We have then

THEOREM 13.1. The information conveyed by a message in the presence of an independent additive gaussian noise has its maximum value when the message itself is gaussian.

Since gaussianly distributed noises occur quite commonly in communication systems, the study of gaussian stochastic processes is of profound interest in the theory of information. We shall have more to say about these processes in section 17, where we shall evaluate the average rate at which any gaussian sequence conveys information about another similar sequence correlated in some manner with it. This problem, of course, includes as a special case the problem of communication in the presence of an independent, additive, gaussian noise.

# IV. PREDICTION THEORY

## 14. Spectral Theory of a Discrete Stochastic Process

Let us consider a discrete stochastic process and let $f = \{f_i\}$ be a realization of that process. Let the element $f_k$ of $f$ take on real values $x$ from a probability measure space $(X, \mathscr{S}, \mu)$. The mean or expected value of the element $f_k$ is then given by

$$\overline{f_k} = \int_X x \, d\mu. \tag{14.1}$$

Let $f_{k+m}$ be another element of $f$ which assumes real values $y$ on the space $(Y, \mathfrak{T}, \nu)$. We then consider the product space $(X \times Y, \mathscr{S} \times \mathfrak{T}, \lambda)$ of all values $(x,y)$ taken on by the ordered pair $(f_k, f_{k+m})$. Thus $\lambda(x,y)$ represents the joint probability distribution function for the pair $(f_k, f_{k+m})$. The mean value of the product $f_k f_{k+m}$ is

$$\overline{f_k f_{k+m}} = \int_{X \times Y} xy \, d\lambda(x,y). \tag{14.2}$$

If the means $\overline{f_k}$ and $\overline{f_k f_{k+m}}$ are independent of the index k, the process is said to be <u>stationary</u> in the <u>wide</u> <u>sense</u> of Khintchine.

In the case of stationarity, the mean

$$R_m = \overline{f_k f_{k+m}} \tag{14.3}$$

is called the <u>autocorrelation</u> <u>coefficient</u> of the process. In the remainder of this section, we shall be concerned with stationary sequences only.

It has been shown, originally by Wiener[11] and later by Wold[13], that in the stationary case, there exists a bounded, nondecreasing, spectral function $W(\vartheta)$ defined on $(-\pi, \pi)$ with $W(-\pi) = 0$ such that

$$R_m = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-im\vartheta} \, dW(\vartheta). \tag{14.4}$$

This is the discrete analog of the Wiener-Khintchine theorem[4,12] for continuous parameter processes. By the Lebesgue decomposition theorem, the function $W(\vartheta)$ may be expressed as the sum

$$W(\vartheta) = W_1(\vartheta) + W_2(\vartheta) \tag{14.5}$$

of two nondecreasing functions, the first of which is absolutely continuous, while the second has an almost everywhere vanishing derivative. This decomposition is unique if we set $W_1(-\pi) = 0$. Wold has shown that this spectral decomposition is accompanied by a corresponding decomposition of the sequence $\{f_i\}$ by which each element $f_k$ is expressed as the sum

$$f_k = f_k^{(1)} + f_k^{(2)}. \tag{14.6}$$

The sequences $\{f_i^{(1)}\}$ and $\{f_i^{(2)}\}$ have, as spectra, the functions $W_1(\vartheta)$ and $W_2(\vartheta)$, respectively. Wold has shown that sequences posessing the latter type of spectra are purely deterministic; that is, the entire future of the sequence is completely specified by its values in the past. In other words, if we know the past history of a sequence of this type, we may predict its future perfectly (in the sense of zero mean-square error) by an operation on that past history. It is clear that if the spectrum is a pure step-function, the sequence is almost periodic, and such a sequence is certainly predictable. Furthermore, the so-called singular* spectra are of the class represented by $W_2(\vartheta)$, hence correspond to purely deterministic sequences.

Kolmogorov[14] made an extensive study of the class of stationary sequences with absolutely continuous spectra and proved that a necessary and sufficient condition for such a sequence to be nondeterministic is that the integral

$$\int_{-\pi}^{\pi} \left|\log W_1'(\vartheta)\right| \, d\vartheta$$

be finite. He has thus shown that if the above integral diverges even a sequence with an absolutely continuous spectrum is deterministic. Sequences with absolutely continuous spectra for which this integral is finite are termed regular by Kolmogorov, and it is only these sequences which are useful as information carriers. For the remainder of this thesis, we shall be concerned primarily with regular sequences.

We may express the autocorrelation coefficient of a regular sequence as

$$R_m = \frac{1}{2\pi} \int_{-\pi}^{\pi} W'(\vartheta) e^{-im\vartheta} \, d\vartheta, \tag{14.7}$$

hence, the nonnegative spectral density $W'(\vartheta)$ is equivalent to the

*A singular function is a continuous function with an almost everywhere vanishing derivative having, in addition, the property that it has positive variation on a set of Lebesgue measure zero. The so-called Cantor function is an example of a singular function. (See Munroe[26], p. 196.)

Fourier series expansion of the autocorrelation coefficient:

$$W'(\vartheta) \sim \sum_{m=-\infty}^{\infty} R_m e^{im\vartheta}. \tag{14.8}$$

We note that in general we may write

$$\overline{f_k f_{k+m}} = \overline{f_{(k+m)} f_{(k+m)-m}}. \tag{14.9}$$

If the process is stationary, the value of this mean is independent of k, hence of k + m. It follows that

$$R_m = R_{-m}, \tag{14.10}$$

and

$$W'(\vartheta) \sim R_o + 2 \sum_{m=1}^{\infty} R_m \cos m\vartheta \tag{14.11}$$

is an even function of $\vartheta$. That is, $W'(\vartheta) = W'(-\vartheta)$.

The Fourier series on the right-hand side of (14.8) may be regarded as the boundary values on the unit circle of a real function $\Lambda(z)$ of the complex variable $z = re^{i\vartheta}$. It follows that

$$\Lambda(e^{i\vartheta}) \sim W'(\vartheta) \tag{14.12}$$

and, since $W'(\vartheta)$ is even, we have

$$\Lambda(z) = \Lambda(\tfrac{1}{z}). \tag{14.13}$$

The coefficients $R_m$ may be obtained from $\Lambda(z)$ by the contour integral equivalent of (14.7):

$$R_m = \frac{1}{2\pi i} \oint \Lambda(z) \frac{dz}{z^{m+1}} , \tag{14.14}$$

where the integral is performed on the unit circle. From the Parseval relation for Fourier series, it follows that

$$\sum_{m=-\infty}^{\infty} R_m^2 = \frac{1}{2\pi i} \oint \Lambda^2(z) \frac{dz}{z}. \tag{14.15}$$

Note that we might write $\Lambda(z)$ as a Laurent series expansion of the coefficients $R_m$

$$\Lambda(z) = \sum_{m=-\infty}^{\infty} R_m z^m, \tag{14.16}$$

but the only assumption we need make concerning convergence is that this series converge in mean-square on the unit circle. Accordingly, $\Lambda(e^{i\vartheta})$ is well defined on the unit circle; we define $\Lambda(z)$ elsewhere in the plane to be simply the function $\Lambda(e^{i\vartheta})$ with $e^{i\vartheta}$ replaced by z. $\Lambda(z)$ is thus the analytic continuation into the plane of the function $\Lambda(e^{i\vartheta})$ on the unit circle. It will follow that

$$\Lambda(e^{i\vartheta}) = \underset{r\to 1}{\text{l.i.m.}}\ \Lambda(re^{i\vartheta}). \tag{14.17}$$

EXAMPLE 14.1. As an example of this procedure, let us consider a sequence whose autocorrelation coefficients are given by $R_m = a^{-|m|}$ with $a > 1$ and real. From (14.16),

$$\Lambda(z) = 1 + \sum_{m=1}^{\infty} a^{-m} z^{m} + \sum_{m=1}^{\infty} a^{-m} z^{-m}. \tag{14.18}$$

The first series converges for $|z| < a$, while the second converges for $|z| > 1/a$. Since $a > 1$, the expansion converges absolutely in the annular ring $1/a < |z| < a$, which includes the unit circle. Summing these series, we obtain

$$\Lambda(z) = 1 + \frac{\frac{z}{a}}{1 - \frac{z}{a}} + \frac{\frac{1}{az}}{1 - \frac{1}{az}}$$

$$= \frac{(\frac{1}{a} - a)\ z}{(z - a)(z - \frac{1}{a})}, \tag{14.19}$$

with the result that $\Lambda(z)$ is defined throughout the extended plane with the exception of its poles at $z = a$ and $z = 1/a$.

From (14.13) and the reflection principle it follows that the poles and zeros of $\Lambda(z)$ have mirror symmetry about the unit circle. That is, if a zero or pole occurs a $z = \gamma$, there will exist also a zero or pole, respectively, at $z = 1/\gamma$ for all complex $\gamma$.

EXAMPLE 14.2. As a second example, let us consider the set of autocorrelation coefficients given by

$$R_m = \begin{cases} 1 & m = 0 \\ \frac{1}{2|m|} & m \neq 0. \end{cases} \tag{14.20}$$

The spectral function for the stationary sequence having these coefficients becomes

$$\Lambda(z) = 1 + \frac{1}{2}\sum_{m=1}^{\infty} \frac{z^m}{m} + \frac{1}{2}\sum_{m=1}^{\infty} \frac{z^{-m}}{m} . \tag{14.21}$$

We note that the first series diverges, in particular, for $z = 1$; hence also for $|z| > 1$. Similarly, the second series diverges for $|z| < 1$. Thus the Laurent series expansion of the correlation coefficients fails to converge absolutely in any region. However, on the unit circle we can write

$$\Lambda(e^{i\vartheta}) = 1 + \sum_{m=1}^{\infty} \frac{1}{m} \cos m\vartheta, \tag{14.22}$$

where the Fourier series converges in mean-square. Summing this series, we have

$$\Lambda(e^{i\vartheta}) = 1 - \frac{1}{2} \log (4 \sin^2 \frac{\vartheta}{2}). \tag{14.23}$$

If $e^{i\vartheta}$ is replaced by z, then

$$\Lambda(z) = 1 - \frac{1}{2} \log (1 - z)(1 - \frac{1}{z}) \tag{14.24}$$

defines the spectral function at all points in the plane where the right-hand side exists.

Spectrum factorization. If $\Lambda(z)$ is the spectrum of a regular sequence, we can employ a theorem of Szegö[15] to factor $\Lambda(z)$ as follows:

$$\Lambda(z) = \lambda(z)\ \lambda(\frac{1}{z}), \tag{14.25}$$

where the function $\lambda(z)$ is analytic and nonvanishing inside $|z| < 1$. Since the theorem given by Szegö is somewhat more general than we shall have need for, we shall express his results in a more restricted form. The following theorem is a consequence of his theorem; we refer the reader to Szegö's paper[15] for its proof.

THEOREM 14.1 (Szegö). Let the real nonnegative function $F(\vartheta)$ be even and integrable on $(-\pi,\pi)$. Then a necessary and sufficient condition for the existence of a function $\lambda(z)$, analytic in $|z| < 1$, with

$$F(\vartheta) \sim \left|\lambda(e^{i\vartheta})\right|^2, \tag{14.26}$$

is that $\int_{-\pi}^{\pi} |\log F(\vartheta)|\ d\vartheta$ be finite.

The expansion

$$\lambda(e^{i\vartheta}) \sim \sum_{k=0}^{\infty} \alpha_k e^{ik\vartheta} \tag{14.27}$$

converges in mean-square, and

$$\sum_{k=0}^{\infty} |\alpha_k|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(\vartheta)\, d\vartheta. \tag{14.28}$$

The coefficients $\alpha_k$ may be so chosen that $\lambda(z)$ is nonvanishing in $|z| < 1$.

It is clear that, for a regular sequence, the spectral density $W'(\vartheta)$ may be identified with the function $F(\vartheta)$, and Szegö's theorem applies. We shall show how the function $\lambda(z)$ may be obtained, in general, from the spectral function $\Lambda(z)$. Clearly, if $\Lambda(z)$ is a rational function in z as in Example 14.1, we can factor $\Lambda(z)$ by inspection simply by associating with $\lambda(z)$ the poles and zeros of $\Lambda(z)$ which lie outside the unit circle. In that example, we have

$$\lambda(z) = \sqrt{a^2 - 1}\, \frac{1}{z - a}. \tag{14.29}$$

A simple computation verifies that $\Lambda(z) = \lambda(z)\lambda(1/z)$ agrees with (14.19).

More generally, when $\Lambda(z)$ is not rational we have, since

$$\int_{-\pi}^{\pi} |\log \Lambda(e^{i\vartheta})|\, d\vartheta < \infty, \tag{14.30}$$

that $\log \Lambda(e^{i\vartheta})$ may be expanded in a Fourier series in $(-\pi,\pi)$:

$$\log \Lambda(e^{i\vartheta}) \sim \sum_{k=-\infty}^{\infty} c_k e^{ik\vartheta}, \tag{14.31}$$

from which

$$c_k = c_{-k} = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ik\vartheta} \log \Lambda(e^{i\vartheta})\, d\vartheta$$

$$= \frac{1}{2\pi i} \oint \log \Lambda(z) \frac{dz}{z^{k+1}}\, . \tag{14.32}$$

Let it be noted that since

$$|c_k| \leqslant \frac{1}{2\pi} \int_{-\pi}^{\pi} |\log \Lambda(e^{i\vartheta})|\, d\vartheta,$$

the coefficients $c_k$ are bounded.

We now form the function

$$\lambda(z) = e^{\left[\frac{1}{2}(c_o + 2\sum_{k=1}^{\infty} c_k\, z^k)\right]}$$

$$= \sum_{k=0}^{\infty} \alpha_k\, z^k. \tag{14.33}$$

It follows directly that

$$\lambda(z)\lambda(\tfrac{1}{z}) = e^{\left[c_o + \sum_{k=1}^{\infty} c_k(z^k + z^{-k})\right]} = e^{\log \Lambda(z)} = \Lambda(z). \tag{14.34}$$

Furthermore, from Parseval's relation, we have

$$\sum_{k=0}^{\infty} |\alpha_k|^2 = \frac{1}{2\pi i}\oint \lambda(z)\lambda(\tfrac{1}{z})\, \frac{dz}{z}$$

$$= \frac{1}{2\pi i}\oint \Lambda(z)\, \frac{dz}{z} < \infty. \tag{14.35}$$

Since the expansion (14.33) for $\lambda(z)$ contains no negative powers of z, and since the sum of the squares of its coefficients converges, $\lambda(z)$ is analytic in $|z| < 1$. Similarly, the expansion

$$g(z) = 2 \log \lambda(z) = c_o + 2\sum_{k=1}^{\infty} c_k\, z^k \tag{14.36}$$

contains no negative powers of z and, because the $c_k$ are bounded, it does not diverge inside the unit circle. Therefore,

$$\lambda(z) = \exp \tfrac{1}{2}\, g(z) \tag{14.37}$$

cannot vanish inside the unit circle. It follows from (14.32) and (14.33) that

$$\lambda(0) = e^{\frac{1}{2} c_o} = \exp\left\{\frac{1}{4\pi i}\oint \log \Lambda(z)\, \frac{dz}{z}\right\}. \tag{14.38}$$

<u>Crosscorrelation</u>. Let $f = \{f_i\}$ and $g = \{g_i\}$ be a pair of real stationary sequences posessing autocorrelation coefficients $R_m^{(ff)}$ and $R_m^{(gg)}$. The sequences f and g are said to be <u>stationarily correlated</u>

if the coefficient

$$R_m^{(fg)} = \overline{f_k\, g_{k+m}} = R_{-m}^{(gf)} \tag{14.39}$$

does not depend on k. Cramér[16] has shown that there exists a function $W_{fg}(\vartheta)$ of bounded variation in $(-\pi,\pi)$, with $W_{fg}(-\pi) = 0$, such that

$$R_m^{(fg)} = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-im\vartheta}\, dW_{fg}(\vartheta). \tag{14.40}$$

If the function $W_{fg}(\vartheta)$ is absolutely continuous, there exists a function $\Lambda_{fg}(z)$ which on the unit circle is equivalent to $W'_{fg}(\vartheta)$. We then have

$$R_m^{(fg)} = \frac{1}{2\pi i} \oint \Lambda_{fg}(z)\, \frac{dz}{z^{m+1}} \tag{14.41}$$

and

$$\Lambda_{fg}(z) = \sum_{m=-\infty}^{\infty} R_m^{(fg)}\, z^m = \Lambda_{gf}(\tfrac{1}{z}), \tag{14.42}$$

where the Laurent series is assumed to converge only in the mean-square sense on the unit circle. The Parseval relation for this case becomes

$$\sum_{m=-\infty}^{\infty} \left[R_m^{(fg)}\right]^2 = \frac{1}{2\pi i} \oint \Lambda_{fg}(z) \Lambda_{fg}(\tfrac{1}{z})\, \frac{dz}{z}. \tag{14.43}$$

## 15. Simple Prediction

Let us review the pure prediction problem of a discrete sequence which was treated first by Wold[13] and later, in more detail, by Kolmogorov[18] and Wiener[19].

Pure prediction. We consider a random sequence $\{f_i\}$, which is one realization of a stochastic process. Let the element $f_k$ represent the present value of f and the elements $f_{k+p}$ represent future values when $p > 0$ and past values when $p < 0$.

The prediction problem may be formulated as follows: Let us assume that we know precisely all values of f in the past and present; that is, we have at our disposal the subsequence $(\ldots, f_{k-2}, f_{k-1}, f_k)$. On the basis of this knowledge, we want to find the expected value of the element $f_{k+p}$ which lies p-units ahead in the sequence. It is clear that there will exist a conditional distribution function, let us say $\mu_p(x)$, for the value of $f_{k+p}$, given the past history of f. In order to obtain this distribution function explicitly, it would be necessary to begin with the joint distribution function for the subsequence $(\ldots f_{k-2}, f_{k-1}, f_k, f_{k+p})$ defined on an infinite-dimensional product space and obtain from this the conditional measure on the component space on which the element $f_{k+p}$ is defined, conditioned by the values $(\ldots f_{k-2}, f_{k-1}, f_k)$ on the remaining component spaces. Having once determined the conditional distribution we could find the expected value. It is evident that such an approach represents a rather formidable problem. It is possible, however, to get an approximation for the expected value without actually obtaining the distribution function.

Under the assumption that the distribution $\mu_p(x)$ exists, let us see what value $\alpha$ we should predict for $f_{k+p}$ in order to minimize the mean-square error resulting from our prediction. Letting the true value of $f_{k+p}$ be x, we wish to minimize the function

$$\overline{\varepsilon^2} = \int_X (x - \alpha)^2 \, d\mu_p(x) \tag{15.1}$$

with respect to $\alpha$. Expanding the expression for $\overline{\varepsilon^2}$ and setting its derivative with respect to $\alpha$ equal to zero, we obtain

$$\frac{d}{d\alpha} \overline{\varepsilon^2} = - 2 \int_X x \, d\mu_p + 2\alpha = 0 \tag{15.2}$$

or

$$\alpha = \int_X x \, d\mu_p . \tag{15.3}$$

It is a simple matter to verify that this value of $\alpha$ results in a true minimum of $\overline{\varepsilon^2}$.

Since $\int x \, d\mu_p$ represents the expected value of $f_{k+p}$, the expected value is that prediction which gives the smallest mean-square error of all possible predictions. Substituting the expected value for $\alpha$ in (15.1), we find that the minimum value for the mean-square error is the variance of the conditional distribution.

The preceding discussion indicates that we may obtain the expected value of $f_{k+p}$ by performing that operation on the past of f which minimizes the mean-square error between the true value of $f_{k+p}$ and the result of the operation. In other words, mean-square prediction is equivalent to prediction of the expected value. If we want something other than the expected value, however, mean-square prediction should not be used.

For example, if the conditional distribution density has appreciable skew and yet is unimodal, we might choose to predict the mode or most-probable value which may differ appreciably from the mean. In such a case, some prediction other than mean-square is called for. However, if the density is unimodal and symmetric about the mode, then the mean and mode coincide, and mean-square prediction also gives the most-probable value. To the author's knowledge, little work has been done in other than mean-square prediction.

Even if we restrict ourselves to the prediction of the expected value, it is not quite clear just what sort of operation we need make on the past in order to minimize the mean-square error of prediction. The work of Wold, Kolmogorov, and Wiener represents an approximation to the prediction problem in all but the gaussian case. They treated only linear prediction; that is, they obtained an approximation for the expected value of $f_{k+p}$ by a linear operation on the past of f. It has been pointed out by Wiener[19] and proved by Singleton[27] that, when the sequence f is multivariate gaussian, the optimum operator to be applied to the past of f for obtaining the expected value of the future element $f_{k+p}$ is a linear operator. Also in the gaussian case, the mean and mode for the conditional distribution coincide; hence the only reasonable prediction is mean-square linear prediction. Consequently, mean-square linear prediction is optimal in the gaussian case.

The linear mean-square prediction problem, in which the future of a sequence is predicted by a linear operation on its past, has been treated in detail by Kolmogorov[18], Wiener[19], and Doob[20]. We shall treat in some detail a slightly more general problem which includes pure prediction as a special case. Also subsumed under this discussion is the problem of filtering in the presence of noise. We shall show that the results obtained apply to the pure prediction problem and agree with those given by Kolmogorov.

The general prediction problem. Let $\{f_i\}$ and $\{g_i\}$ be a pair of real regular sequences stationarily correlated in the wide sense of Khintchine. That is, the correlation functions, $R_m^{(ff)}, R_m^{(gg)}, R_m^{(fg)}$, and their associated spectra $\Lambda_{ff}(z), \Lambda_{gg}(z)$, and $\Lambda_{fg}(z)$, exist. We formulate the prediction problem as follows: Let us assume that the values of the past and present of the sequence f are known precisely. On the basis of this knowledge we would like to obtain the best mean-square approximation of the element $g_{k+p}$ by a linear operation on the past and present of f. That is, we want to find that set of coefficients $\{a_i\}$ $(i = 0,1,2,\ldots)$ which minimizes the mean-square error

$$\overline{\varepsilon^2} = \overline{\left[g_{k+p} - \sum_{i=0}^{\infty} a_i\ f_{k-i}\right]^2}. \tag{15.4}$$

Setting $\partial\overline{\varepsilon^2}/\partial a_m = 0$, we find that the optimum a's are those that satisfy the set of equations

$$\sum_{i=0}^{\infty} a_i\ R_{m-i}^{(ff)} = R_{m+p}^{(fg)} \qquad m \geqslant 0. \tag{15.5}$$

The minimum mean-square error for prediction of g p-units ahead becomes

$$\sigma_p^2 = R_o^{(gg)} - \sum_{i=0}^{\infty} a_i\ R_{i+p}^{(fg)} \tag{15.6}$$

with the a's satisfying (15.5). In order to obtain the coefficients, we write (15.5) in spectral form:

$$\frac{1}{2\pi i}\oint A(z)\ \Lambda_{ff}(z)\ \frac{dz}{z^{m+1}} = \frac{1}{2\pi i}\oint \Lambda_{fg}(z)\ \frac{dz}{z^{m+p+1}} \qquad m \geqslant 0, \tag{15.7}$$

where

$$A(z) = \sum_{i=0}^{\infty} a_i\ z^i. \tag{15.8}$$

Using methods analogous to the solution of the Wiener-Hopf integral equation, we find the solution of (15.7) to be

$$\begin{aligned} A(z) &= \frac{1}{\lambda_{ff}(z)} \sum_{k=0}^{\infty} z^k \left[\frac{1}{2\pi i}\oint \frac{\Lambda_{fg}(u)}{\lambda_{ff}(\frac{1}{u})}\ \frac{du}{u^{k+p+1}}\right] \\ &= \frac{1}{z^p\lambda_{ff}(z)} \sum_{k=p}^{\infty} z^k \left[\frac{1}{2\pi i}\oint \frac{\Lambda_{fg}(u)}{\lambda_{ff}(\frac{1}{u})}\ \frac{du}{u^{k+1}}\right], \end{aligned} \tag{15.9}$$

where $\lambda_{ff}(z)$ is that factor of $\Lambda_{ff}(z)$ analytic and nonvanishing inside the unit circle. In order to show that A(z) in (15.9) satisfies (15.7), we write

$$\begin{aligned} &\frac{1}{2\pi i}\oint A(z)\ \Lambda_{ff}(z)\ \frac{dz}{z^{m+1}} \\ &= \frac{1}{2\pi i}\oint \lambda_{ff}(\tfrac{1}{z}) \sum_{k=p}^{\infty} z^k \left[\frac{1}{2\pi i}\oint \frac{\Lambda_{fg}(u)}{\lambda_{ff}(\frac{1}{u})}\ \frac{du}{u^{k+1}}\right] \frac{dz}{z^{p+m+1}}. \end{aligned} \tag{15.10}$$

If we replace the sum $\sum_{k=p}^{\infty}$ by the difference of sums $\sum_{k=-\infty}^{\infty} - \sum_{k=-\infty}^{p-1}$, the right-hand side of (15.10) becomes

$$\frac{1}{2\pi i} \oint \Lambda_{fg}(z) \frac{dz}{z^{m+p+1}}$$

$$- \sum_{k=-\infty}^{p-1} \left[\frac{1}{2\pi i} \oint \lambda_{ff}(\tfrac{1}{z}) \frac{dz}{z^{m+p-k+1}}\right]\left[\frac{1}{2\pi i} \oint \frac{\Lambda_{fg}(u)}{\lambda_{ff}(\frac{1}{u})} \frac{du}{u^{k+1}}\right].$$

The first bracketed factor becomes, on setting $z = 1/\zeta$,

$$\frac{1}{2\pi i} \oint \lambda_{ff}(\zeta)\, \zeta^{m+p-k-1}\, d\zeta ,$$

but since $\lambda_{ff}(z)$ is analytic in $|z| < 1$, this integral vanishes when the exponent satisfies $m + p - k - 1 \geqslant 0$. Since the summation of k is over $(-\infty, p - 1)$, we always have $k \leqslant p - 1$; hence the integral vanishes for all $m \geqslant 0$. Clearly, (15.7) is satisfied.

Expressing (15.6) in spectral form and using the expression for A(z) given in (15.9) we obtain, for the minimum value of the mean-square error,

$$\sigma_p^2 = R_o^{(gg)} - \frac{1}{2\pi i} \oint A(\tfrac{1}{z}) \Lambda_{fg}(z) \frac{dz}{z^{p+1}}$$

$$= R_o^{(gg)} - \frac{1}{2\pi i} \oint \frac{\Lambda_{fg}(z)}{\lambda_{ff}(\frac{1}{z})} \sum_{k=p}^{\infty} z^{-k} \frac{1}{2\pi i} \oint \frac{\Lambda_{fg}(u)}{\lambda_{ff}(\frac{1}{u})} \frac{du}{u^{k+1}} \frac{dz}{z}$$

$$= R_o^{(gg)} - \sum_{k=p}^{\infty} \left[\frac{1}{2\pi i} \oint \frac{\Lambda_{fg}(z)}{\lambda_{ff}(\frac{1}{z})} \frac{dz}{z^{k+1}}\right]^2 . \qquad (15.11)$$

In order to interpret this result for the pure prediction problem, we simply set g = f; hence $\Lambda_{gg}(z) = \Lambda_{ff}(z)$, and $\Lambda_{fg}(z) = \Lambda_{ff}(z)$. The operator A(z) for pure prediction may be expressed in the following form:

$$A(z) = \frac{1}{z^p \lambda_{ff}(z)} \sum_{k=p}^{\infty} z^k \left[\frac{1}{2\pi i} \oint \lambda_{ff}(u) \frac{du}{u^{k+1}}\right]$$

$$= \frac{1}{z^p} \left\{1 - \frac{1}{\lambda_{ff}(z)} \sum_{k=0}^{p-1} z^k \left[\frac{1}{2\pi i} \oint \lambda_{ff}(u) \frac{du}{u^{k+1}}\right]\right\}$$

$$A(z) = \frac{1}{z^p} \left[ 1 - \frac{1}{\lambda_{ff}(z)} \frac{1}{2\pi i} \oint \lambda_{ff}(u) \frac{1 - (\frac{z}{u})^p}{1 - \frac{z}{u}} \frac{du}{u} \right]$$

$$= \frac{1}{\lambda_{ff}(z)} \frac{1}{2\pi i} \oint \lambda_{ff}(u) \frac{du}{u^p(u - z)} . \qquad (15.12)$$

The minimum mean-square error for prediction of $f_{k+p}$ by a linear operation on the past and present of f becomes

$$\sigma_p^2 = R_o^{(ff)} - \sum_{k=p}^{\infty} \left[ \frac{1}{2\pi i} \oint \lambda_{ff}(z) \frac{dz}{z^{k+1}} \right]^2 . \qquad (15.13)$$

Since $\lambda_{ff}(z)$ is analytic within the unit circle, the integral in the bracket vanishes for $k < 0$. Hence, for $p \leqslant 0$,

$$\sigma_p^2 = R_o^{(ff)} - \sum_{k=0}^{\infty} \left[ \frac{1}{2\pi i} \oint \lambda_{ff}(z) \frac{dz}{z^{k+1}} \right]^2 . \qquad (15.14)$$

But by the Parseval relation, the right-hand side becomes zero and, as we would naturally expect, the past and present of f can be predicted perfectly. For $p > 0$, we replace in (15.13) the sum $\sum\limits_{k=p}^{\infty}$ by the difference of sums $\left[ \sum\limits_{k=0}^{\infty} - \sum\limits_{k=0}^{p-1} \right]$, and the minimum mean-square error for prediction of $f_{k+p}$ which lies p-units ahead is

$$\sigma_p^2 = \sum_{k=0}^{p-1} \left[ \frac{1}{2\pi i} \oint \lambda_{ff}(z) \frac{dz}{z^{k+1}} \right]^2 , \qquad (15.15)$$

a result obtained by Kolmogorov[17]. It follows, in particular, that

$$\sigma_1^2 = \lambda_{ff}^2(0) = \exp \left[ \frac{1}{2\pi i} \oint \log \Lambda_{ff}(z) \frac{dz}{z} \right] . \qquad (15.16)$$

Since $\Lambda_{ff}(z) \geqslant 0$ on the unit circle, it follows from the inequality between arithmetic and geometric means* that

*Let f be a nonnegative function, integrable $[\mu]$ on a set E of finite measure. Then

$$\exp \left\{ \frac{1}{\mu(E)} \int_E \log f \, d\mu \right\} \leqslant \frac{1}{\mu(E)} \int_E f \, d\mu ,$$

where the equality holds if and only if f is constant $[\mu]$. For proof, see Hardy, Littlewood, Polya[22].

$$\sigma_1^2 = \exp\left[\frac{1}{2\pi i}\oint \log \Lambda_{ff}(z)\,\frac{dz}{z}\right] \leqslant \frac{1}{2\pi i}\oint \Lambda_{ff}(z)\,\frac{dz}{z} = R_0^{(ff)}, \tag{15.17}$$

where the equality holds when and only when $\Lambda_{ff}(z) = R_0^{(ff)}$ almost everywhere on $|z| = 1$. Thus, if the sequence f has finite mean-square $(R_0^{(ff)} < \infty)$, the logarithmic integral can diverge only toward $(-\infty)$ if it is to diverge at all. This divergence does occur if $\Lambda_{ff}(e^{i\vartheta})$ vanishes on a set of values of $\vartheta$ of positive measure.

It is seen that if the logarithmic integral should diverge, $\sigma_1^2 = 0$; and by iteration we can predict any future element $f_{k+p}$ perfectly from a knowledge of the past history of f. Accordingly, a necessary condition for a sequence to be nondeterministic is that the logarithmic integral be finite, hence that the sequence be regular.

We can conclude, then, that if a sequence posesses a spectrum that vanishes on any set in $(-\pi,\pi)$ of positive measure, it is deterministic. We note from (15.15) that

$$\sigma_p^2 = \sum_{k=0}^{p-1}\left[\frac{1}{2\pi i}\oint \lambda(z)\,\frac{dz}{z^{k+1}}\right]^2 \leqslant \sum_{k=0}^{\infty}\left[\frac{1}{2\pi i}\oint \lambda(z)\,\frac{dz}{z^{k+1}}\right]^2$$

$$= R_0^{(ff)}. \tag{15.18}$$

The mean-square error for any p is, as we would expect, always bounded by the mean-square of the sequence.

If $\Lambda_{ff}(z) = R_0^{(ff)}$ almost everywhere on $|z| = 1$, then

$$R_m^{(ff)} = \frac{1}{2\pi i}\oint R_0^{(ff)}\,\frac{dz}{z} = R_0^{(ff)}\;\;\delta_m^0, \tag{15.19}$$

where $\delta_m^0$ is the Kroneker delta. We then have $\lambda(z) = \sqrt{R_0^{(ff)}}$, hence

$$\sigma_p^2 = \sum_{k=0}^{p-1} R_0^{(ff)}\;\;\delta_k^0 = R_0^{(ff)} \qquad p > 0, \tag{15.20}$$

and no prediction is possible for $p > 0$.

A sequence whose correlation coefficient is given by (15.19) is called _purely random_ or an _orthogonal_ sequence.

_The Wold-Kolmogorov decomposition._ If we re-examine the pure prediction problem with $p = 1$, the error term becomes

$$f_{k+1} - \sum_{i=0}^{\infty} a_i \, f_{k-i} , \tag{15.21}$$

with the a's obtained from

$$a_k = \frac{1}{2\pi i} \oint A(z) \, \frac{dz}{z^{k+1}} \ . \tag{15.22}$$

For this case, however, (15.9) becomes

$$\begin{aligned} A(z) &= \frac{1}{z \, \lambda_{ff}(z)} \sum_{k=1}^{\infty} z^k \left[ \frac{1}{2\pi i} \oint \lambda_{ff}(u) \, \frac{du}{u^{k+1}} \right] \\ &= \frac{1}{z} \left[ 1 - \frac{\lambda_{ff}(0)}{\lambda_{ff}(z)} \right] , \end{aligned} \tag{15.23}$$

where we have replaced $\sum\limits_{k=1}^{\infty}$ by $\left[ \sum\limits_{k=0}^{\infty} - \sum\limits_{k=0}^{0} \right]$ and used Parseval's theorem. Let us re-write (15.21), replacing k by k - 1, and consider the error

$$\varepsilon_k = f_k - \sum_{i=0}^{\infty} a_i \, f_{k-1-i} \tag{15.24}$$

which forms the k-th element of a sequence $\{\varepsilon_i\}$. Forming the cross-correlation of $\varepsilon$ with f, we have

$$\begin{aligned} R_m^{(\varepsilon f)} &= \overline{\varepsilon_k \, f_{k+m}} = R_m^{(ff)} - \sum_{k=0}^{\infty} a_k \, R_{m+1+k} \\ &= \frac{1}{2\pi i} \oint \Lambda_{ff}(z) \left[ 1 - \frac{1}{z} A(\tfrac{1}{z}) \right] \frac{dz}{z^{m+1}} . \end{aligned} \tag{15.25}$$

But from (15.23),

$$\frac{1}{z} A(\tfrac{1}{z}) = 1 - \frac{\lambda_{ff}(0)}{\lambda_{ff}(\frac{1}{z})} ; \tag{15.26}$$

and

$$R_m^{(\varepsilon f)} = \frac{\lambda_{ff}(0)}{2\pi i} \oint \lambda_{ff}(z) \, \frac{dz}{z^{m+1}} , \tag{15.27}$$

which, from the analyticity of $\lambda_{ff}(z)$, must vanish for m $<$ 0. We say that $\mathcal{E}$ is orthogonal to the past of f in the sense that $\overline{\mathcal{E}_k f_{k+m}} = 0$ when $m < 0$. However from (15.24), it is seen that the past of $\mathcal{E}$ is a linear combination of the past of f; hence $\mathcal{E}$ must be orthogonal to its own past. We thus obtain

$$\overline{\mathcal{E}_k \; \mathcal{E}_{k+m}} = \lambda^2_{ff}(0) \quad \delta^o_m \; , \tag{15.28}$$

and the sequence $\{\mathcal{E}_i\}$ is purely random. The fact that $\mathcal{E}$ is also orthogonal to its own future follows, of course, from the even property of the autocorrelation coefficients.

If we write (15.24) as

$$f_k = \sum_{i=0}^{\infty} a_i f_{k-1-i} + \mathcal{E}_k \; , \tag{15.29}$$

the sequence f has been decomposed into two parts, the first of which is a linear combination of its own past, while the second is orthogonal to that past. This expression is known as the Wold-Kolmogorov decomposition of a sequence $\{f_i\}$. The element $\mathcal{E}_k$ represents the "innovation" or new information carried by the element $f_k$. In other words, $\{\mathcal{E}_i\}$ represents that part of $\{f_i\}$ which cannot be predicted from a knowledge of the past elements. Clearly, if the sequence $\{\mathcal{E}_i\}$ has zero variance, the sequence $\{f_i\}$ is deterministic.

We might also write (15.24) in the form

$$\mathcal{E}_k = \sum_{i=0}^{\infty} \beta_i \; f_{k-i} \tag{15.30}$$

where

$$\beta_o = 1$$

$$\beta_k = - a_{k-1} \qquad k > 0. \tag{15.31}$$

Thus for $k > 0$,

$$\beta_k = \frac{-1}{2\pi i} \oint A(z) \frac{dz}{z^k} = \frac{-1}{2\pi i} \oint z \; A(z) \; \frac{dz}{z^{k+1}}$$

$$= \frac{1}{2\pi i} \oint \left[ \frac{\lambda_{ff}(0)}{\lambda_{ff}(z)} - 1 \right] \frac{dz}{z^{k+1}} \; . \tag{15.32}$$

We note that for $k \neq 0$,

$$\frac{1}{2\pi i} \oint 1 \; \frac{dz}{z^{k+1}} = 0,$$

and for k = 0,

$$\frac{1}{2\pi i}\oint \frac{\lambda_{ff}(0)}{\lambda_{ff}(z)}\,\frac{dz}{z} = 1.$$

Therefore, we can write for all k,

$$\beta_k = \frac{1}{2\pi i}\oint \frac{\lambda_{ff}(0)}{\lambda_{ff}(z)}\,\frac{dz}{z^{k+1}} \;. \tag{15.33}$$

Since $\lambda_{ff}(z)$ is analytic and nonvanishing in $|z| < 1$, the integral on the right vanishes for $k < 0$.

We might expect from the orthogonality property of the sequence $\{\varepsilon_i\}$ that the element $f_k$ can be expressed as a linear combination of the values of the sequence $\varepsilon$. That is,

$$f_k \sim \sum_{i=-\infty}^{\infty} \alpha_i\ \varepsilon_{k-i}, \tag{15.34}$$

in which mean-square convergence of the series is required. That such a representation is possible, can be seen from the following considerations:

If we take the crosscorrelation of f and $\varepsilon$, (15.34) yields

$$\overline{f_k\ \varepsilon_{k-m}} = \sum_{i=-\infty}^{\infty} \alpha_i\ \overline{\varepsilon_{k-i}\,\varepsilon_{k-m}}$$

$$= \lambda_{ff}^2(0)\ \alpha_m \;, \tag{15.35}$$

from which we obtain the coefficient

$$\alpha_m = \frac{R_m^{(\varepsilon f)}}{\lambda_{ff}^2(0)} = \frac{1}{2\pi i}\oint \frac{\lambda_{ff}(z)}{\lambda_{ff}(0)}\,\frac{dz}{z^{m+1}} \;, \tag{15.36}$$

where we have used (15.27). From the analyticity of $\lambda_{ff}(z)$ in the unit circle, the integral on the right vanishes for $m < 0$, and (15.34) becomes

$$f_k \sim \sum_{i=0}^{\infty} \alpha_i\ \varepsilon_{k-i} \;. \tag{15.37}$$

Let us consider the partial sum of n terms of the series and the following limit:

$$\lim_{n\to\infty} \overline{\left[f_k - \sum_{i=0}^{n} \alpha_i \,\varepsilon_{k-i}\right]^2}$$

$$= \lim_{n\to\infty}\left[R_o^{(ff)} - 2\sum_{i=0}^{n}\alpha_i\, R_i^{(\varepsilon f)} + \lambda_{ff}^2(0)\sum_{i=0}^{n}\alpha_i^2\right]$$

$$= R_o^{(ff)} - \sum_{i=0}^{\infty}\frac{\left[R_i^{(\varepsilon f)}\right]^2}{\lambda_{ff}^2(0)}. \tag{15.38}$$

From Parseval's relation and (15.27),

$$\sum_{i=0}^{\infty}\left[R_i^{(\varepsilon f)}\right]^2 = \frac{1}{2\pi i}\oint \lambda_{ff}^2(0)\,\lambda_{ff}(z)\,\lambda_{ff}(\tfrac{1}{z})\,\frac{dz}{z}$$

$$= \lambda_{ff}^2(0)\,\frac{1}{2\pi i}\oint \Lambda_{ff}(z)\,\frac{dz}{z}$$

$$= \lambda_{ff}^2(0)\,R_o^{(ff)}. \tag{15.39}$$

The limit of Equation (15.38) is zero; hence the series in (15.37) converges in mean-square. We see then that a regular sequence $\{f_i\}$ may always be expressed in terms of the past history of its innovation. This representation is referred to in statistics as a one-sided moving average of a purely random sequence.

The minimum mean-square error for prediction of $f_{k+p}$ which was given in (15.15) can be expressed in terms of the coefficients $\alpha_i$ as

$$\sigma_p^2 = \sum_{i=0}^{p-1}\lambda_{ff}^2(0)\,\alpha_i^2$$

$$= \lambda_{ff}^2(0)\left[1 + \alpha_1^2 + \dots + \alpha_{p-1}^2\right]. \tag{15.40}$$

In this form it is easily seen that if $\lambda_{ff}^2(0) = 0$, that is, if

$$\frac{1}{2\pi i}\oint \log \Lambda_{ff}(z)\,\frac{dz}{z} = -\infty, \tag{15.41}$$

then the sequence f may be predicted perfectly for any finite p.

Uniqueness. To summarize the results of the Wold-Kolmogorov decomposition: any regular sequence $\{f_i\}$ is related to its innovation $\{\varepsilon_i\}$

by the pair of relations (15.30) and (15.37). If we employ the transformation U which transforms each element of a sequence into the preceding element, these relations can be expressed in the operational form,

$$\mathcal{E}_k = B(U) f_k \tag{15.42}$$

$$f_k \sim A(U)\ \mathcal{E}_k = B^{-1}(U)\ \mathcal{E}_k , \tag{15.43}$$

where B(U) and A(U) are power series expansions in positive powers of U of the coefficients $\beta_i$ and $\alpha_i$. These operations are represented in the block diagram form of Fig. 8. In order that only positive powers of U occur (that is, in order for the operations to apply to past values only), the function B(z) must be analytic within the unit circle. Also, since $A(z) = B^{-1}(z)$ must satisfy the same analyticity condition, it follows that B(z) must, moreover, be nonvanishing in $|z| < 1$.

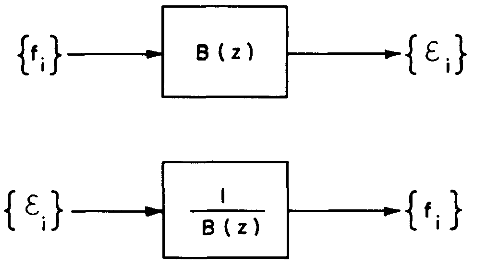

Fig. 8. The orthogonalization scheme.

The spectral conditions,

$$B(z)\ B(\tfrac{1}{z})\ F_{ff}(z) = F_{\mathcal{E}\mathcal{E}}(z) = \text{constant}, \tag{15.44}$$

along with the boundary condition B(0) = 1 $[B(0) = \beta_o$ is the coefficient of the element $f_k$ in (15.30)] guarantees uniquely the following relations:

$$B(z) = \frac{\lambda_{ff}(0)}{\lambda_{ff}(z)} \tag{15.45}$$

$$\overline{\mathcal{E}^2} = \lambda_{ff}^2(0). \tag{15.46}$$

The Wold-Kolmogorov decomposition of a single random sequence is, hence, unique.

## 16. Multiple Prediction

Rather than treat in detail the multiple prediction problem, we shall simply discuss the Wold-Kolmogorov decomposition in its connection with multidimensional processes. Although there is only one such decomposition of a single-dimensional process, we shall show that this property of uniqueness is not shared with the multidimensional process. It is for this reason that the multiple prediction problem has not yet been completely solved in closed form. We include this section partly for the sake of completeness, but primarily to point out the need for additional study along these lines.

Kolmogorov's work was first extended to multidimensional processes by Zasuhin[21], who demonstrated the nonuniqueness of the Wold-Kolmogorov decomposition. More recent studies by Whittle[27] have formalized the multiple prediction problem but only Wiener has provided a usable solution. Employing the Gram-Schmidt orthogonalization procedure, Wiener and Rankin[23] obtained a scalar series solution which, although not in closed form, lends itself to machine computation. In certain specific problems, Wiener's method of undetermined coefficients[19] can yield a satisfactory solution. However, the important general closed-form solution is, to the knowledge of this author, yet to be found.

Following Zasuhin, we consider an n-dimensional vector process $\{F_i\} = \{f_{1i}, f_{2i}, \ldots, f_{ni}\}$ with correlation coefficients

$$R_m^{(ij)} = \overline{f_{ik}\ f_{j,k+m}} \tag{16.1}$$

and spectral functions

$$\Lambda_{ij}(z) = \sum_{k=-\infty}^{\infty} R_k^{(ij)}\ z^k. \tag{16.2}$$

These spectral functions form the elements of an n x n Hermitian matrix $\mathcal{A}(z)$, whose determinant

$$\Delta(z) = \left|\mathcal{A}(z)\right| \tag{16.3}$$

is real, nonnegative, and even on the unit circle. Zasuhin has shown that a necessary and sufficient condition that the vector process be regular and of rank n is that

$$\int_{-\pi}^{\pi} \left|\log\ \Delta(e^{i\vartheta})\right|\ d\vartheta \tag{16.4}$$

be finite.

Under the assumption that this latter condition is fulfilled, the Wold-Kolmogorov decomposition consists in finding the set of coefficients $\{\eta_{ijk}\}$ and $\{\gamma_{ijk}\}$ for which

$$f_{im} \sim \sum_{j=1}^{n} \sum_{k=0}^{\infty} \eta_{ijk} \ \varepsilon_{j,m-k} \qquad (16.5)$$

$$\varepsilon_{im} = \sum_{j=1}^{n} \sum_{k=0}^{\infty} \gamma_{ijk} \ f_{j,m-k}, \qquad (16.6)$$

subject to the constraints

$$\overline{\varepsilon_{ik} \ \varepsilon_{jm}} = \delta_{km} \ \delta_{ij} \ \sigma_j^2 \qquad (16.7)$$

$$\gamma_{ijo} = \eta_{ijo} = \delta_{ij}. \qquad (16.8)$$

These latter coefficients correspond to $\beta_o$ and $\alpha_o$ in the single-dimensional process. The variance $\sigma_j^2$ is the mean-square error for prediction of the sequence $\{f_{ji}\}$ one step ahead. Using the translation transformation U, (16.5) and (16.6) can be expressed in vector form:

$$F_m \sim \mathcal{H}(U) \ E_m \qquad (16.9)$$

$$E_m = \mathcal{G}(U) \ F_m = \mathcal{H}^{-1}(U) \ F_m, \qquad (16.10)$$

where $F_m$ and $E_m$ are n-dimensional vectors representing the m-th elements of the process, and $\mathcal{H}(U)$ is a matrix of operators with elements

$$H_{ij}(z) = \sum_{k=0}^{\infty} \eta_{ijk} \ z^k. \qquad (16.11)$$

In order for $\mathcal{H}(U)$ to contain only positive powers of U and at the same time have an inverse with these properties, its determinant $|\mathcal{H}(z)|$ must be analytic and nonvanishing inside the unit circle. The constraint on the coefficient $\eta_{iko}$ imposes the additional condition $|\mathcal{H}(0)| = 1$.

A straightforward calculation from (16.1) and the orthogonality constraint yields

$$\Lambda_{ij}(z) = \sum_{r=1}^{n} H_{jr}(z) \ H_{ir}(\tfrac{1}{z}) \ \sigma_r^2 \qquad (16.12)$$

where

$$\sigma_r^2 = \overline{\varepsilon_{rk}^2}. \qquad (16.13)$$

In matrix form, (16.12) becomes

$$\mathcal{A}(z) = \mathcal{H}(\tfrac{1}{z}) \cdot \mathcal{S} \cdot \tilde{\mathcal{H}}(z) \qquad (16.14)$$

where $\tilde{\mathscr{A}}$ denotes the transpose of $\mathscr{A}$, and the matrix $\mathscr{D}$ is a diagonal matrix of elements $\sigma_r^2$. Setting

$$\left|\mathscr{D}\right| = \prod_{r=1}^{n} \sigma_r^2 = \sigma^2, \tag{16.15}$$

we obtain the determinantal equation

$$\left|\mathscr{A}(z)\right| \left|\mathscr{A}\left(\frac{1}{z}\right)\right| \sigma^2 = \Delta(z). \tag{16.16}$$

By the method of Szegö, we let $\Delta(z) = \delta(z)\ \delta(1/z)$, with $\delta(z)$ analytic and nonvanishing in the unit circle. From the analyticity requirements of $|\mathscr{A}(z)|$, we obtain

$$\left|\mathscr{A}(z)\right| = \frac{\delta(z)}{\sigma}. \tag{16.17}$$

But since $|\mathscr{A}(0)| = 1$,

$$\left|\mathscr{A}(z)\right| = \frac{\delta(z)}{\delta(0)} \tag{16.18}$$

and

$$\sigma^2 = \delta^2(0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log\ \Delta(e^{i\vartheta})\ d\vartheta. \tag{16.19}$$

The determinant $\sigma^2$ is often called the prediction variance or intrinsic variance of the process.

Using the inverse $\mathscr{G}(z)$ of the matrix $\mathscr{A}(z)$, we can write (16.14) as

$$\mathscr{G}\left(\frac{1}{z}\right) \cdot \mathscr{M}(z) \cdot \tilde{\mathscr{G}}(z) = \mathscr{D}, \tag{16.20}$$

which illustrates another interpretation of the prediction problem. The Wold-Kolmogorov decomposition consists of determining that matrix $\mathscr{G}(z)$ by which the spectral matrix is pre- and post-multiplied in accordance with (16.20) in order to reduce it to a diagonal matrix with constants along the diagonal. The components of the prediction variance are then those diagonal elements. Although a first impulse is to make $\mathscr{G}(z)$ the matrix of the eigenvectors of the spectral matrix, this procedure gives a $\mathscr{G}$-matrix that does not satisfy the analyticity requirements; furthermore, the eigenvalues are functions of z.

Before the preceding considerations convey to the reader the impression that the decomposition is a straightforward procedure, let it be remarked again that the solution is hampered by the nonuniqueness mentioned earlier. In this regard, let us suppose that in a given case the optimum operators have been found, and that their operations applied to the set $\{f_{ri}\}$ produce an innovation sequence $\{\mathcal{E}_{ki}\}$. Let $a_{ij}$ be a set of $n^2$ numbers satisfying

$$\sum_{j=1}^{n} a_{ij}\, a_{kj}\, \sigma_j^2 = \delta_{ik}\, \sigma_k^2. \tag{16.21}$$

Consider the sequence $\{\zeta_{ik}\}$ formed by the operation

$$\zeta_{ik} = \sum_{m=1}^{n} a_{im}\, \mathcal{E}_{mk}. \tag{16.22}$$

This operation is a unitary transformation of the <u>present</u> <u>value</u> <u>only</u> of the innovation vector. Consider the correlation function:

$$\begin{aligned}\overline{\zeta_{ik}\;\zeta_{jm}} &= \sum_{p=1}^{n} a_{ip} \sum_{q=1}^{n} a_{jq}\, \overline{\mathcal{E}_{pk}\;\mathcal{E}_{qm}} \\ &= \sum_{p=1}^{n} a_{ip}\, a_{jp}\, \overline{\mathcal{E}_{pk}\;\mathcal{E}_{pm}} \\ &= \delta_{km} \sum_{p=1}^{n} a_{ip}\, a_{jp}\, \sigma_p^2 \\ &= \delta_{km}\, \delta_{ij}\, \sigma_j^2. \end{aligned} \tag{16.23}$$

On comparing (16.23) with (16.7), we see that the sequence $\{\zeta_{ri}\}$ is a perfectly good innovation sequence having identical statistical characteristics with $\{\mathcal{E}_{ri}\}$. If the matrix $\mathcal{Y}(z)$ satisfies (16.20) so also does the matrix $\mathcal{A} \cdot \mathcal{Y}(z)$, where $\mathcal{A}$ is a unitary matrix of elements $a_{ij}$. Thus, for the multidimensional processes, the Wold-Kolmogorov decomposition is unique only up to a unitary transformation. It is for this reason that the solution of the multiple prediction problem is difficult. In order to find the optimum prediction operator, we need to impose an additional constraint to guarantee uniqueness. And this constraint must be consistent with the constraints already imposed.

## 17. Mutual Rate of a Pair of Gaussian Sequences

Let $f = \{f_i\}$ and $g = \{g_i\}$ be a pair of stationarily correlated sequences which we assume to be governed by a set of multivariate gaussian distribution functions. It is of interest to determine the rate (per element) at which the sequence f conveys information about the sequence g. That is, given the past history of f up to the element $f_{k-1}$, how much additional information about the entire sequence g is provided on the average by the next element $f_k$? Denoting by p the subsequence $(\ldots,f_{k-2},f_{k-1})$ of f, we wish to formulate

$$\overline{I(g;f_k|p)}, \tag{17.1}$$

that is, the average information about g provided by the element $f_k$ when the past of that element is known. It is not quite clear at this stage whether or not the information (17.1) really represents the average rate per element. By the average rate we mean a number R so defined that, from a sufficiently large number N of successive elements of f, we obtain an amount of information which is of the order of NR. To see that (17.1) does satisfy this intuitional requirement for the average rate, let us consider the pair of elements $(f_k,f_{k+1})$. From Theorem 12.3, the average information about g provided by the pair $(f_k,f_{k+1})$ when the past p is known becomes

$$\overline{I(g;f_k,f_{k+1}|p)} = \overline{I(g;f_k|p)} + \overline{I(g;f_{k+1}|p,f_k)}. \tag{17.2}$$

However the pair $(p,f_k)$ in the last term is the subsequence $(\ldots,f_{k-2},f_{k-1},f_k)$, which is, simply, the past of the element $f_{k+1}$. Since the process is assumed stationary, the information (17.1) is independent of the index k, hence the two terms in the right-hand side of (17.2) are equal. By iteration, it follows that

$$\overline{I(g;f_k,f_{k+1},\ldots,f_{k+N-1}|p)} = N\ \overline{I(g;f_k|p)} \tag{17.3}$$

and $\overline{I(g;f_k|p)}$, which is independent of k, does indeed represent the average rate R per element provided by f about the sequence g. We call

$$R(g;f) = \overline{I(g;f_k|f_{k-1},f_{k-2},\ldots)} \tag{17.4}$$

the average rate at which f conveys information about g.

In order to evaluate the information in (17.1), we need to determine a priori and a posteriori distributions for the sequence g. The a priori distribution is that distribution of g conditioned by the subsequence p; the a posteriori distribution is that of g conditioned by the pair $(p,f_k)$. Since the sequence g is composed of an infinite set of elements, both of these distributions are infinite-dimensional. When we take into consideration the correlation of the elements of g, it becomes apparent that the evaluation of these distributions represents a rather formidable problem. Here we can make use of Lemma 11.2 to great advantage. According to that lemma, we can write

$$\overline{I(g;f_k|p)} = \overline{I(f_k;g|p)}. \tag{17.5}$$

The right-hand side may be interpreted as the information about the element $f_k$ provided on the average by the specification of the value of every element of g in the past, present, and future. In order to evaluate this information, we need determine a priori and a posteriori distributions of the <u>single</u> <u>element</u> $f_k$ alone. These distributions are clearly one-dimensional. The effect of Lemma 11.2 and Theorem 12.1 is to reduce an infinite-dimensional problem to one of a single dimension.

Some question may conceivably arise concerning a physical justification for the prediction of a known element $f_k$ from the past, present, and future of a completely unknown sequence $\{g_i\}$. Let us remember that here we are using the mathematical artifice of solving a simple hypothetical problem whose solution is identical with that of a more difficult physical one. The only justification necessary for such a procedure is the fact that if the solution to the physical problem exists, then by Lemma 11.2 the solution to the hypothetical one exists also, and these solutions are identical.

As the sequences f and g have been assumed to be multivariate gaussian, the a priori and a posteriori distributions for the element $f_k$ will be simple gaussian distributions. Thus, in order to completely specify these distributions, we need determine only their means and variances. It is now quite clear that the evaluation of the average rate is a problem of prediction. In the gaussian case, in fact, it becomes one of linear prediction.

In order to find the a priori distribution function for $f_k$, we need find the optimum linear operation to be applied to the past of f in order to obtain the best mean-square approximation of the element $f_k$. The result of that operation becomes the mean of the distribution; the minimum value of the mean-square error becomes its variance. From the results given in section 15 for pure prediction one-step ahead, the a priori distribution of a particular element $f_k$ has a mean

$$\alpha_k = \sum_{i=0}^{\infty} a_i f_{k-1-i} \tag{17.6}$$

with

$$a_m = \frac{1}{2\pi i} \oint \frac{1}{z} \left[ 1 - \frac{\lambda_{ff}(0)}{\lambda_{ff}(z)} \right] \frac{dz}{z^{m+1}} , \tag{17.7}$$

and variance

$$\sigma_1^2 = \lambda_{ff}^2(0). \tag{17.8}$$

Thus the a priori probability density for a particular $f_k$ is

$$\rho'(x) = \frac{1}{\sqrt{2\pi\, \sigma_1^2}} \exp \left\{ - \frac{(x - \alpha_k)^2}{2\, \sigma_1^2} \right\}. \tag{17.9}$$

Similarly, to evaluate the a posteriori distribution for that same element $f_k$, we need determine the sets of coefficients $\{b_i\}$ $(i = 0,1,\ldots)$ and $\{c_i\}$ $(i = \ldots,-1,0,1,\ldots)$ for which

$$\overline{\mathcal{E}^2} = \overline{\left[ f_k - \sum_{i=0}^{\infty} b_i\, f_{k-1-i} - \sum_{i=-\infty}^{\infty} c_i g_{k-i} \right]^2} \tag{17.10}$$

is a minimum. Note that the index of the coefficient $c_i$ runs over all of the positive and negative integers. Hence, although the function

$$B(z) = \sum_{i=0}^{\infty} b_i\, z^i \tag{17.11}$$

will be analytic within the unit circle, the function

$$C(z) = \sum_{i=-\infty}^{\infty} c_i\, z^i \tag{17.12}$$

will not be, in general.

Setting

$$\frac{\partial \overline{\mathcal{E}^2}}{\partial b_m} = \frac{\partial \overline{\mathcal{E}^2}}{\partial c_m} = 0$$

in (17.10), we obtain the pair of equations

$$\sum_{i=0}^{\infty} b_i \, R_{m-i}^{(ff)} + \sum_{i=-\infty}^{\infty} c_i \, R_{m+1-i}^{(fg)} = R_{m+1}^{(ff)} \qquad m \geqslant 0 \qquad (17.13)$$

$$\sum_{i=0}^{\infty} b_i \, R_{m-1-i}^{(gf)} + \sum_{i=-\infty}^{\infty} c_i \, R_{m-i}^{(gg)} = R_{m}^{(gf)}, \qquad (17.14)$$

where the second equation must hold for all m. The minimum value of the mean-square error becomes

$$\sigma_2^2 = \min \; \overline{\mathcal{E}^2} = R_0^{(ff)} - \sum_{i=0}^{\infty} b_i \, R_{i+1}^{(ff)} - \sum_{i=-\infty}^{\infty} c_i \, R_i^{(gf)}. \qquad (17.15)$$

We can express (17.14) in spectral form

$$\frac{1}{2\pi i} \oint \left[ z \, B(z) \Lambda_{gf}(z) + C(z) \; \Lambda_{gg}(z) \right] \frac{dz}{z^{m+1}}$$

$$= \frac{1}{2\pi i} \oint \Lambda_{gf}(z) \, \frac{dz}{z^{m+1}} \, , \qquad (17.16)$$

but since this expression must be valid for all integers m, we can equate the integrands. Solving for C(z), we have

$$C(z) = \frac{\Lambda_{gf}(z) - z \, B(z) \; \Lambda_{gf}(z)}{\Lambda_{gg}(z)}. \qquad (17.17)$$

Equation (17.13), expressed in spectral form, becomes

$$\frac{1}{2\pi i} \oint \left[ B(z) \; \Lambda_{ff}(z) + \frac{1}{z} \, C(z) \; \Lambda_{fg}(z) \right] \frac{dz}{z^{m+1}}$$

$$= \frac{1}{2\pi i} \oint \Lambda_{ff}(z) \, \frac{dz}{z^{m+2}} \qquad m \geqslant 0. \qquad (17.18)$$

Substituting (17.17) in (17.18) to eliminate C(z), we have

$$\frac{1}{2\pi i} \oint B(z) \left[ \Lambda_{ff}(z) - \frac{\Lambda_{fg}(z) \; \Lambda_{gf}(z)}{\Lambda_{gg}(z)} \right] \frac{dz}{z^{m+1}}$$

$$= \frac{1}{2\pi i} \oint \left[ \Lambda_{ff}(z) - \frac{\Lambda_{fg}(z) \; \Lambda_{gf}(z)}{\Lambda_{gg}(z)} \right] \frac{dz}{z^{m+2}} \qquad m \geqslant 0. \; (17.19)$$

It follows that the function

$$\Lambda(z) = \Lambda_{ff}(z) - \frac{\Lambda_{fg}(z)\,\Lambda_{gf}(z)}{\Lambda_{gg}(z)} \qquad (17.20)$$

is even and nonnegative on the unit circle, hence is a spectral function. Comparing

$$\frac{1}{2\pi i}\oint B(z)\,\Lambda(z)\,\frac{dz}{z^{m+1}} = \frac{1}{2\pi i}\oint \Lambda(z)\,\frac{dz}{z^{m+2}} \qquad m \geqslant 0 \qquad (17.21)$$

with (15.7) of section 15, we see that the operator B(z) is the pure prediction operator for prediction one-step ahead of a sequence having the spectrum $\Lambda(z)$. From (15.23), it follows that the solution of (17.21) is

$$B(z) = \frac{1}{z}\left[1 - \frac{\lambda(0)}{\lambda(z)}\right], \qquad (17.22)$$

where $\lambda(z)$ is that factor of $\Lambda(z)$ analytic and nonvanishing inside the unit circle. The minimum value of the mean-square error (17.15) is then

$$\sigma_2^2 = \lambda^2(0) = \exp \frac{1}{2\pi i}\oint \log \Lambda(z)\,\frac{dz}{z}. \qquad (17.23)$$

The a posteriori distribution density for the particular element $f_k$, given its past and the entire sequence g, becomes

$$\nu'(x) = \frac{1}{\sqrt{2\pi\,\sigma_2^2}} \exp\left\{\frac{(x - \beta_k)^2}{2\,\sigma_2^2}\right\} \qquad (17.24)$$

with

$$\beta_k = \sum_{i=0}^{\infty} b_i\, f_{k-1-i} + \sum_{i=-\infty}^{\infty} c_i\, g_{k-i}. \qquad (17.25)$$

The coefficients $b_m$ and $c_m$ are given by

$$b_m = \frac{1}{2\pi i}\oint \frac{1}{z}\left[1 - \frac{\lambda(0)}{\lambda(z)}\right]\frac{dz}{z^{m+1}} \qquad (17.26)$$

and

$$c_m = \frac{1}{2\pi i}\oint \frac{\Lambda_{gf}(z)}{\Lambda_{gg}(z)}\,\frac{\lambda(0)}{\lambda(z)}\,\frac{dz}{z^{m+1}} \qquad (17.27)$$

The information about the sequence g provided by a particular element

$f_k$ becomes, from (9.14),

$$I_k = \int_{-\infty}^{\infty} \nu'(x) \log \frac{\nu'(x)}{\rho'(x)} \, dx$$

$$= \frac{1}{2} \log \frac{\sigma_1^2}{\sigma_2^2} - \frac{\sigma_1^2 - \sigma_2^2}{2\sigma_1^2} + \frac{(\alpha_k - \beta_k)^2}{2\sigma_1^2} \qquad (17.28)$$

when natural units of information are employed.

To obtain the average information given by all the elements of f, we take the average of $I_k$ over all possible k:

$$R(g;f) = \overline{I_k} = \frac{1}{2} \log \frac{\sigma_1^2}{\sigma_2^2} - \frac{\sigma_1^2 - \sigma_2^2}{2\sigma_1^2} + \frac{\overline{(\alpha_k - \beta_k)^2}}{2\sigma_1^2} \; . \qquad (17.29)$$

In order to evaluate the mean-square of the quantity $(\alpha_k - \beta_k)$, we note that

$$\overline{(\alpha_k - \beta_k)^2} = \overline{\alpha_k^2} + \overline{\beta_k^2} - 2\,\overline{\alpha_k \beta_k}. \qquad (17.30)$$

From (17.6),

$$\overline{\alpha_k^2} = \sum_{i=0}^{\infty} a_i \sum_{m=0}^{\infty} a_m R_{i-m}^{(ff)}$$

$$= \sum_{i=0}^{\infty} a_i R_{i+1}^{(ff)}$$

$$= R_o^{(ff)} - \sigma_1^2, \qquad (17.31)$$

where we have used (15.5) and (15.6) of section 15 in the special case of pure prediction with p = 1.
Similarly, from (17.25),

$$\overline{\beta_k^2} = \sum_{i=0}^{\infty} b_i \sum_{m=0}^{\infty} b_m R_{m-i}^{(ff)} + \sum_{i=-\infty}^{\infty} c_i \sum_{m=-\infty}^{\infty} c_m R_{m-i}^{(gg)}$$

$$+ 2 \sum_{i=0}^{\infty} b_i \sum_{m=-\infty}^{\infty} c_m R_{i+1-m}^{(fg)} = \sum_{i=0}^{\infty} b_i R_{i+1}^{(ff)} + \sum_{i=-\infty}^{\infty} c_i R_i^{(gf)}$$

$$= R_o^{(ff)} - \sigma_2^2 , \qquad (17.32)$$

where we have used (17.13), (17.14), and (17.15). The cross-moment $\overline{\alpha_k\beta_k}$ becomes

$$\overline{\alpha_k\beta_k} = \sum_{i=0}^{\infty} a_i \sum_{m=0}^{\infty} b_m \, R^{(ff)}_{m-i} + \sum_{i=0}^{\infty} a_i \sum_{m=-\infty}^{\infty} c_m \, R^{(fg)}_{i+1-m}$$

$$= \sum_{i=0}^{\infty} a_i \, R^{(ff)}_{i+1}$$

$$= R^{(ff)}_{0} - \sigma_1^2. \tag{17.33}$$

Combining (17.31), (17.32) and (17.33), with (17.30), we have

$$\overline{(\alpha_k - \beta_k)^2} = \sigma_1^2 - \sigma_2^2, \tag{17.34}$$

and the average rate becomes*

$$R(g;f) = \frac{1}{2} \log \frac{\sigma_1^2}{\sigma_2^2}$$

$$= \frac{1}{2} \log \frac{\lambda^2_{ff}(0)}{\lambda^2(0)}$$

$$= \frac{1}{4\pi i} \oint \log \frac{\Lambda_{ff}(z)}{\Lambda(z)} \frac{dz}{z}$$

$$= \frac{1}{4\pi i} \oint \log \frac{\Lambda_{ff}(z)\,\Lambda_{gg}(z)}{\Lambda_{ff}(z)\,\Lambda_{gg}(z) - \Lambda_{fg}(z)\,\Lambda_{gf}(z)} \frac{dz}{z} \tag{17.35}$$

From the symmetry of this expression with respect to f and g, it is seen that the sequence g provides information about f at the same rate as that provided by f about g.

It is of interest to note that the information rate is invariant under linear operations on the past of the sequences involved. Let a sequence $\{h_i\}$ be derived from the sequence $\{f_i\}$ in the following manner: If $\{\eta_i\}$ is a set of coefficients for which

*It has been pointed out to the author by Dr. R. A. Silverman that Equation (17.35), as well as its extension to the continuous case (section 18), has been given in a recent Russian publication by M.S. Pinsker[28].

$$\sum_{m=0}^{\infty} \eta_m f_{k-m} \tag{17.36}$$

converges in mean-square, let that sum be equivalent to the element $h_k$. Thus the sequence h is the result of an operation on the past of the sequence f. Setting

$$H(z) = \sum_{m=0}^{\infty} \eta_m z^m, \tag{17.37}$$

we obtain

$$\Lambda_{hh}(z) = H(z)\, H(\tfrac{1}{z})\, \Lambda_{ff}(z), \tag{17.38}$$

$$\Lambda_{hg}(z) = H(z)\, \Lambda_{fg}(z). \tag{17.39}$$

The average rate at which the sequence h provides information about g is then

$$\begin{aligned} R(g;h) &= \frac{1}{4\pi i} \oint \log \frac{\Lambda_{hh}(z)\, \Lambda_{gg}(z)}{\Lambda_{hh}(z)\, \Lambda_{gg}(z) - \Lambda_{hg}(z)\, \Lambda_{gh}(z)} \frac{dz}{z} \\ &= \frac{1}{4\pi i} \oint \log \frac{H(\frac{1}{z})\, H(z)\, \Lambda_{ff}(z)\, \Lambda_{gg}(z)}{H(z)\, H(\frac{1}{z}) \left[\Lambda_{ff}(z)\, \Lambda_{gg}(z) - \Lambda_{fg}(z)\, \Lambda_{gf}(z)\right]} \frac{dz}{z} \\ &= R(g;f). \end{aligned} \tag{17.40}$$

Thus the linear operation on the past of f has left the rate unchanged.

The expression (17.35) for the mutual rate of an arbitrary pair of stationary gaussian sequences is sufficiently general to handle a large variety of problems. For example, we can examine the problem treated by Shannon[25] concerning the capacity of a channel in which a gaussian noise is added to a band-limited gaussian message.

Let the message be essentially limited to the low-frequency band (0,W) and let $\{g_i\} = \{m_i\}$ represent the values of the message at sample points 1/2W seconds apart. Similarly, let $\{n_i\}$ represent the corresponding noise samples. If $\{f_i\}$ is the sequence of elements

$$f_k = m_k + n_k \;, \tag{17.41}$$

where message and noise are assumed uncorrelated, it follows that

$$\Lambda_{gg}(z) = \Lambda_{mm}(z)$$

$$\Lambda_{ff}(z) = \Lambda_{mm}(z) + \Lambda_{nn}(z)$$
$$\Lambda_{fg}(z) = \Lambda_{mm}(z) \tag{17.42}$$

and

$$R(m;f) = \frac{1}{4\pi i} \oint \log \left[ 1 + \frac{\Lambda_{mm}(z)}{\Lambda_{nn}(z)} \right] \frac{dz}{z}, \tag{17.43}$$

which, in essence, is the result obtained by Shannon. In section 18, where we treat functions of a continuous time, we shall obtain an expression which resembles his more closely in notation.

To obtain the capacity of the channel subject to a message power constraint, we simply maximize (17.43) with respect to $\Lambda_{mm}(e^{i\vartheta})$, subject to the constraint

$$\frac{1}{2\pi i} \oint \Lambda_{mm}(z) \frac{dz}{z} = \sigma_m^2. \tag{17.44}$$

Such a maximization leads to the result that the sum $\Lambda_{mm}(z) + \Lambda_{nn}(z)$ should be constant on the unit circle except when the total message power is too low to permit such a choice while maintaining $\Lambda_{mm}(e^{i\vartheta})$ non-negative. In such a case, the sum should be made constant whenever $\Lambda_{nn}(e^{i\vartheta}) < \sigma_m^2 + \sigma_n^2$, and $\Lambda_{mm}(e^{i\vartheta})$ should be set near zero otherwise.

## 18. Extension to the Continuous Case

The results of the previous section can be effectively applied to the study of random time functions which are realizations of a continuous-parameter stochastic process. In order to make the transition from the discrete to the continuous-parameter case, we utilize a technique of sampling in the time domain. That is, the random functions, which are defined continuously in time, are supposed to be sampled at equal intervals T, providing a discrete sequence of random variables. We can then employ our previous results to determine the mutual rate of information between such sequences. We obtain the time rate for the random functions by allowing the sampling interval T to approach zero.

Let us consider a pair of multivariate gaussian time functions f(t) and g(t) which are assumed to be stationarily correlated in the wide-sense of Khintchine. Thus the correlation functions

$$\varphi_{ff}(\tau) = \overline{f(t)\ f(t + \tau)}$$
$$\varphi_{gg}(\tau) = \overline{g(t)\ g(t + \tau)}$$

$$\varphi_{fg}(\tau) = \overline{f(t)\ g(t + \tau)} \tag{18.1}$$

exist and are independent of t. The spectra are the Fourier transforms of the correlation functions:

$$\Phi_{fg}(\omega) = \int_{-\infty}^{\infty} \varphi_{fg}(\tau) e^{-i\omega\tau} d\tau, \tag{18.2}$$

with a similar definition for $\Phi_{ff}(\omega)$ and $\Phi_{gg}(\omega)$. If we sample both f(t) and g(t) at equal time intervals T, the sequences $\{f_i\}$ and $\{g_i\}$ with

$$\begin{aligned} f_k &= f(kT) \\ g_k &= g(kT) \end{aligned} \tag{18.3}$$

are well defined and have correlation coefficients given by

$$\begin{aligned} R_m^{(ff)} &= \overline{f_k\ f_{k+m}} = \overline{f(kT)\ f(kT + mT)} \\ &= \varphi_{ff}(mT) \end{aligned} \tag{18.4}$$

$$R_m^{(gg)} = \varphi_{gg}(mT) \tag{18.5}$$

$$R_m^{(fg)} = \varphi_{fg}(mT). \tag{18.6}$$

The spectra of the sequences becomes

$$\Lambda_{fg}(e^{i\vartheta}) = \sum_{m=-\infty}^{\infty} \varphi_{fg}(mT) e^{im\vartheta}, \tag{18.7}$$

with similar expressions for $\Lambda_{ff}(e^{i\vartheta})$ and $\Lambda_{gg}(e^{i\vartheta})$.

Letting $\vartheta = \omega T$, we note that

$$T\Lambda_{fg}(e^{i\omega T}) = \sum_{m=-\infty}^{\infty} \varphi_{fg}(mT) e^{im\omega T}\ T. \tag{18.8}$$

Passing to the limit as T → 0, the right-hand side becomes, formally, an integral:

$$\begin{aligned} \lim_{T \to 0} T\Lambda_{fg}(e^{i\omega T}) &= \int_{-\infty}^{\infty} \varphi_{fg}(\tau) e^{i\omega\tau} d\tau \\ &= \Phi_{fg}^{*}(\omega). \end{aligned} \tag{18.9}$$

The average rate (per element) at which the sequence $\{f_i\}$ conveys information about the sequence $\{g_i\}$ becomes

$$R_T(g;f) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \log \frac{\Lambda_{ff}(e^{i\vartheta})\ \Lambda_{gg}(e^{i\vartheta})}{\Lambda_{ff}(e^{i\vartheta})\ \Lambda_{gg}(e^{i\vartheta}) - \Lambda_{fg}(e^{i\vartheta})\ \Lambda_{gf}(e^{i\vartheta})}\, d\vartheta$$

$$= \frac{T}{4\pi}\int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} \log \frac{\Lambda_{ff}(e^{i\omega T})\ \Lambda_{gg}(e^{i\omega T})}{\Lambda_{ff}(e^{i\omega T})\ \Lambda_{gg}(e^{i\omega T}) - \left|\Lambda_{fg}(e^{i\omega T})\right|^2}\, d\omega. \tag{18.10}$$

The time rate (per second) at which the time function f(t) conveys information about the time function g(t) is then

$$R(g;f) = \lim_{T\to 0} \frac{1}{T} R_T(g;f)$$

$$= \frac{1}{4\pi}\int_{-\infty}^{\infty} \log \frac{\Phi_{ff}(\omega)\ \ \Phi_{gg}(\omega)}{\Phi_{ff}(\omega)\ \ \Phi_{gg}(\omega) - \left|\Phi_{fg}(\omega)\right|^2}\, d\omega. \tag{18.11}$$

If we treat the special case in which g(t) is a message m(t), and f(t) is the sum of that message and a noise n(t), uncorrelated with m, the rate at which f gives information about m becomes

$$R(m;f) = \frac{1}{4\pi}\int_{-\infty}^{\infty} \log \left[1 + \frac{\Phi_{mm}(\omega)}{\Phi_{nn}(\omega)}\right] d\omega$$

$$= \int_{0}^{\infty} \log \left[1 + \frac{\Phi_{mm}(2\pi f)}{\Phi_{nn}(2\pi f)}\right] df, \tag{18.12}$$

which is seen to agree with the results obtained by Wiener[8] and Shannon[25]. Correlation between the message and noise offers no additional difficulty, since more generally we have

$$R(m;f) = \frac{1}{4\pi}\int_{-\infty}^{\infty} \log \frac{\Phi_{mm}(\omega)\ \left[\Phi_{mm}(\omega) + \Phi_{nn}(\omega) + 2\ \mathrm{Re}\ \Phi_{mn}(\omega)\right]}{\Phi_{mm}(\omega)\ \ \Phi_{nn}(\omega) - \left|\Phi_{mn}(\omega)\right|^2}\, d\omega, \tag{18.13}$$

where $\Phi_{mn}(\omega)$ is the cross-spectrum of the message and noise.

## 19. Information Rates in Discrete Networks

In section 17, we obtained an expression for the average rate per element at which a discrete gaussian sequence $\{f_i\}$ provides information about the past, present, and future of another gaussian sequence $\{g_i\}$. A good example for the application of this expression is provided by the sampled-data filter. Let the sequence $\{g_i\}$ represent the input to a linear sampled-data network whose output is the sequence $\{f_i\}$. In effect, the filter performs a linear operation on the past history of its input in such a way that any element $f_k$ of the output is expressed by

$$f_k = \sum_{m=0}^{\infty} \gamma_m \, g_{k-m}, \tag{19.1}$$

where the coefficients $\{\gamma_i\}$ define the filter characteristic. It is well known that if $\{g_i\}$ is multivariate gaussian in its distribution, so also will be the sequence $\{f_i\}$, and the results of section 17 apply.

Letting

$$G(z) = \sum_{m=0}^{\infty} \gamma_m \, z^m, \tag{19.2}$$

we note that

$$\begin{aligned} \overline{f_k \, f_{k+n}} &= \sum_{m=0}^{\infty} \gamma_m \sum_{i=0}^{\infty} \gamma_i \, \overline{g_{k-m} \, g_{k+n-i}} \\ &= \sum_{m=0}^{\infty} \gamma_m \sum_{i=0}^{\infty} \gamma_i \, R^{(gg)}_{m+n-i} \\ &= \sum_{m=0}^{\infty} \gamma_m \sum_{i=0}^{\infty} \gamma_i \left[ \frac{1}{2\pi i} \oint \Lambda_{gg}(z) \, \frac{dz}{z^{m+n-i+1}} \right] \\ &= \frac{1}{2\pi i} \oint G(z) \, G(\tfrac{1}{z}) \, \Lambda_{gg}(z) \, \frac{dz}{z^{n+1}} \, , \end{aligned} \tag{19.3}$$

from which we conclude

$$\Lambda_{ff}(z) = G(z) \, G(\tfrac{1}{z}) \, \Lambda_{gg}(z). \tag{19.4}$$

A similar treatment yields

$$\Lambda_{fg}(z) = G(\tfrac{1}{z}) \, \Lambda_{gg}(z). \tag{19.5}$$

It is easily seen that

$$\Lambda_{ff}(z) \, \Lambda_{gg}(z) = \Lambda_{fg}(z) \, \Lambda_{gf}(z), \tag{19.6}$$

with the result that R(g;f) becomes infinite. We thus conclude that the output of a linear sampled-data filter provides information about the past, present, and future of the input at an infinite rate.

This result becomes clear when we note from Lemma 11.2 that this rate is identical to the rate of information about $f_k$ provided by the past, present, and future of the input sequence $\{g_i\}$. Since the past, present, and future of g (or in fact, the past and present only of g) is sufficient to completely specify the present value of the output, we must conclude that the output of any linear discrete network provides information about the past, present, and any portion of the future of the input at an infinite rate.

On the other hand, we might conceivably want to know the rate at which the output of such a network provides information about the past history only of the input. Since that past history may not completely specify the present value of the output, such a rate may well be finite. This problem is one of multiple prediction. That is, the a posteriori variance is the minimum mean-square error for prediction of the element $f_k$ from a linear operation on the pasts of the sequences $\{f_i\}$ and $\{g_i\}$. However, from (19.6), the determinant of the spectral matrix of the vector process vanishes; hence the process is of rank one. This should be expected from the fact that a knowledge of the past history of the input allows the complete specification of the past history of the output. Thus we need determine only that mean-square error resulting from an optimum prediction of $f_k$ from a linear operation on the past history alone of the input. The a priori variance is simply

$$\sigma_1^2 = \lambda_{ff}^2(0), \tag{19.7}$$

and the a posteriori variance becomes

$$\sigma_2^2 = \inf \overline{\left[ f_k - \sum_{i=0}^{\infty} \alpha_i \, g_{k-1-i} \right]^2}. \tag{19.8}$$

Using the methods of section 15 for the solution of (15.4), we find

$$\begin{aligned} \sigma_2^2 &= R_o^{(ff)} - \sum_{m=1}^{\infty} \left[ \frac{1}{2\pi i} \oint \frac{\Lambda_{gf}(z)}{\lambda_{gg}(\frac{1}{z})} \frac{dz}{z^{m+1}} \right]^2 \\ &= R_o^{(ff)} - \sum_{m=1}^{\infty} \left[ \frac{1}{2\pi i} \oint G(z) \, \lambda_{gg}(z) \frac{dz}{z^{m+1}} \right]^2 \\ &= G^2(0) \, \lambda_{gg}^2(0), \end{aligned} \tag{19.9}$$

where G(z) is the filter characteristic which is, of course, analytic in the unit circle.

If we form the function $G(z)\ G(\frac{1}{z})$, such a function has the properties of a spectrum, hence is nonnegative and even on the unit circle. By the methods of Szegö, we can then express

$$G(z)\ G(\tfrac{1}{z}) = \Gamma(z)\ \Gamma(\tfrac{1}{z}), \tag{19.10}$$

with $\Gamma(z)$ analytic and nonvanishing within the unit circle. Let it be noted that $G(z) = \Gamma(z)$ if and only if G(z) does not vanish in $|z| < 1$. The class of discrete filters whose transfer functions have this latter property corresponds to the class of "minimum-phase" networks in the continuous network theory.

From (19.4),

$$\lambda_{ff}(z) = \Gamma(z)\ \lambda_{gg}(z); \tag{19.11}$$

and (19.7) becomes

$$\sigma_1^2 = \Gamma^2(0)\ \lambda_{gg}^2(0). \tag{19.12}$$

Thus, from (19.9) and (19.12), the rate of information provided by the output about the past only of the input is

$$R = \frac{1}{2}\log\frac{\sigma_1^2}{\sigma_2^2} = \frac{1}{2}\log\frac{\Gamma^2(0)}{G^2(0)}$$

$$= \frac{1}{4\pi i}\oint \log\frac{G(z)\ G(\frac{1}{z})}{G^2(0)}\ \frac{dz}{z}\ . \tag{19.13}$$

If G(z) is nonvanishing within the unit circle (hence has a physically realizable inverse), the present value of the output provides no information about the past history of the input. Certainly, in such a case, we can perform the operation $G^{-1}(z)$ on the past history of the output and recover the past history of the input exactly and with no delay. However, when the inverse is not realizable (except, of course, with delay) we need the information provided by the present and, possibly, future values of the output in order to reconstruct the past of the input. Thus (19.13) is the sort of expression our intuition would expect.

## 20. The Rate of Information About Future Values

An interesting problem for the application of the foregoing ideas is the following: Given the past history of a gaussian sequence $\{f_i\}$ up to the element $f_{k-1}$, how much information about the future of the sequence is given on the average by the next element $f_k$? Clearly, if the sequence is purely random, the element $f_k$ is statistically independent of the future, hence can be expected to provide no information about it. On the other hand, if the sequence f is deterministic, its past history completely specifies its future, and again the information must be zero. It is quite reasonable, then, to question whether or not there exists a class of sequences which provide a maximum amount of information about the future. Unfortunately, we shall see that no such class does exist. In fact we shall provide an example of a sequence for which this information is infinite.

This problem is one involving both prediction and interpolation. Once again, we employ Lemma 11.2 and determine the average information about the element $f_k$ provided by the future when the past is known. As before, the a priori variance is given by the minimum mean-square error for prediction of f one-step ahead:

$$\sigma_1^2 = \lambda_{ff}^2(0). \tag{20.1}$$

The a posteriori variance is the minimum mean-square error for interpolating the element $f_k$ by a linear operation on the elements

$$(\ldots, f_{k-2}, f_{k-1}, f_{k+1}, f_{k+2}, \ldots).$$

This latter problem has been treated in detail by Kolmogorov,[14,17,18] and we shall simply make direct use of his results. He showed that the minimum value of that mean-square error is given by

$$\sigma_2^2 = \frac{\pi}{\displaystyle\int_0^{\pi} \frac{d\vartheta}{\Lambda_{ff}(e^{i\vartheta})}}. \tag{20.2}$$

The average information provided by $f_k$ about the future of the sequence f when its past is known is thus

$$I = \frac{1}{2} \log \frac{\sigma_1^2}{\sigma_2^2}$$

$$= \frac{1}{2} \log \frac{\lambda_{ff}^2(0)}{\pi} \int_0^{\pi} \frac{d\vartheta}{\Lambda_{ff}(e^{i\vartheta})}$$

$$I = \frac{1}{2\pi}\int_0^\pi \log \Lambda_{ff}(e^{i\vartheta})\, d\vartheta + \frac{1}{2}\log \frac{1}{\pi}\int_0^\pi \frac{d\vartheta}{\Lambda_{ff}(e^{i\vartheta})}. \qquad (20.3)$$

If the sequence $\{f_i\}$ is purely random, that is, if $\Lambda_{ff}(e^{i\vartheta})$ is constant in $(0,\pi)$, this expression becomes zero. It is quite natural to question whether or not there exists a spectral function that maximizes the information in (20.3). However, let us consider a sequence whose spectrum is given by

$$\Lambda_{ff}(e^{i\vartheta}) = \vartheta \qquad (20.4)$$

in $(0,\pi)$. The first integral in (20.3) is finite for this case, whereas the second integral diverges. Such a sequence thus provides information about its future at an infinite rate.

It is interesting to note the existence of a class of sequences which can be extrapolated only with error, but which can be interpolated perfectly (in the sense of zero mean-square error).

~ ~ ~

The extension of prediction theory techniques to the evaluation of information rates in many processes follows immediately the solution of the more general prediction problem. For example, to obtain the rate at which one gaussian time series gives information about the past history of a similar series correlated with it is to solve the problem of multiple linear prediction. If this solution can be achieved with uniqueness, our study of information in a linear network extends directly to the evaluation of the rate at which the output of the network provides information about the past history of the input, up to any fixed time in the past.

Similarly, to obtain the mutual rate between nongaussian time series is to solve the problem of multiple nonlinear prediction. It is also true here that the mean value of the conditional distribution for a random variable is given by the optimum <u>mean</u>-<u>square</u> prediction from an operation on the condition. If the series are nongaussian, that operation is, of course, nonlinear. When the distribution for the error of prediction can be found, the conditional distribution for the random variable is simply that error distribution translated by the value of the optimum prediction.

The solutions to the problems of sections 19 and 20 are not included for the sake of their value as concrete results, but rather to illustrate the relation between information theory and the theory of prediction.

# ACKNOWLEDGMENT

The writer is deeply indebted to Professor Y. W. Lee for his constant guidance and encouragement through the course of this research, and to the members of his group for the many informative discussions on the problems of communication theory. He is also indebted to Professor Peter Elias; and to Dr. S. D. Pezaris, who aided the author in rectifying a fundamental error in an early approach to the study.

This acknowledgment would be incomplete without a mention of Professor Norbert Wiener, whose stimulating lectures are reflected in much of the present work.

# Bibliography

1. A. Kolmogorov, Foundations of Probability (Ergeb. Math. u. Grenzgeb., 2, No. 3, Berlin, 1933)(Chelsea, New York, 1950).

2. P. R. Halmos, Measure Theory (D. Van Nostrand Company, Inc., New York, 1950).

3. R. E. Wernikoff, Outline of Lebesgue theory: A heuristic introduction, Technical Report 310, Research Laboratory of Electronics, M.I.T., to be published.

4. A. Khintchine, Korrelationstheorie der stationären stochastischen Prozesse, Math. Annalen, 109, 604-615 (1934).

5. G. D. Birkhoff, Proof of the ergodic theorem, Proc. Nat. Acad. of Sci., U.S.A., 17, 656-660 (1932).

6. N. Wiener, The ergodic theorem, Duke Math. Jour., 15, 1, 1-18 (1939).

7. F. Riesz, Sur la théorie ergodique, Comm. Math. Helv., 17, 221-239 (1945).

8. N. Wiener, Cybernetics (John Wiley and Sons, Inc., New York, 1948).

9. C. E. Shannon, A Mathematical theory of communication, Bell System Tech. J., 27, 3, 379-423 (July 1948); 27, 4, 623-656 (October 1948).

10. P. M. Woodward and I. L. Davies, Information theory and inverse probability in telecommunication, Proc. I.E.E., 99, Part III, 37-44 (1952).

11. N. Wiener, The spectrum of an array, J. Math. Phys., 6, 145-157 (1926).

12. N. Wiener, Generalized harmonic analysis, Acta Math., 55, 117-258 (1930).

13. H. A. Wold, A Study in the Analysis of Stationary Time Series (Dissertation, Uppsala, 1938), 2nd ed. (Almqvist and Wiksell, Stockholm, 1954).

14. A. Kolmogorov, Stationary sequences in Hilbert space, Bull. State Univ., Moscow, Ser. Math., 2, 6, (1941).

15. G. Szegö, Über die Randwerte einer analytischen Funktion, Math. Annalen, 84, 232-244 (1921).

16. H. Cramér, On the theory of stationary random processes, Annals of Math., 41, 215-230 (1940).

17. A. Kolmogorov, Sur l'interpolation et extrapolation des suites stationaries, Compt. rend. Acad. des Sci., Paris, 208, 5, 2043-2045 (June 1939).

18. A. Kolmogorov, Interpolation und Extrapolation von stationären zufälligen Folgen, Bull. Acad. Sci., U.S.S.R., Ser. math., 5, 3-14 (1941).

19. N. Wiener, Extrapolation, Interpolation, and Smoothing of Stationary Time Series, NDRC Report, 1942 (John Wiley and Sons, Inc., New York, 1949).

20. J. L. Doob, Stochastic Processes (John Wiley and Sons, Inc., 1953).

21. V. Zasuhin, On the theory of multidimensional stationary random processes, Compt. rend. Acad. Sci., U.S.S.R., 33, 435-437 (1941).

22. P. Whittle, The analysis of multiple stationary time series, J. Roy. Stat. Soc., Ser. B, 15, 125-139 (1953).

23. N. Wiener and B. Rankin, Multiple prediction, Annals of Math. Statistics, to be published.

24. G. H. Hardy, J. E. Littlewood, and G. Polya, Inequalities (Cambridge Univ. Press, Cambridge, 1934).

25. C. E. Shannon, Communication in the presence of noise, Proc. I.R.E., 37, 1, 10-21 (January 1949).

26. M. E. Munroe, Measure and Integration Theory, (Addison-Wesley Publishing Co., Inc., Cambridge, Massachusetts, 1953).

27. H. E. Singleton, Theory of nonlinear transducers, Technical Report 160, Research Laboratory of Electronics, M.I.T., August 12, 1950.

28. M. S. Pinsker, Quantity of information in a gaussian stationary process contained in a second process stationarily connected with it, Compt. rend. Acad. Sci., U.S.S.R., 99, 213-216 (1954).