XIII. STATISTICAL COMMUNICATION THEORY

Prof. Y. W. Lee
Prof. A. G. Bose
Prof. A. H. Nuttall
D. A. Chesler
W. M. Cowan, Jr.
D. A. George
A. D. Hause
I. M. Jacobs
K. L. Jordan, Jr.
J. W. Pan
M. Schetzen
D. W. Tufts
C. E. Wernlein, Jr. (Absent)
G. D. Zames

RESEARCH OBJECTIVES

This group is interested in a variety of problems in statistical communication theory. Current research is primarily concerned with: continuous nonlinear systems that are representable by a series of generalized convolution integrals, the polynomial representation for nonlinear systems with well-defined inputs, the characterization of random processes by series expansions, properties of correlation functions, the connectivity and capacity of random graphs, and the effect of uncertainties on radar detection.

1. The study of continuous nonlinear systems is directed toward the representation and analysis of nonlinear systems that are composed of linear subsystems and continuous nonlinear no-memory subsystems. One objective is the development of an optimum nonlinear system.

2. The polynomial representation for nonlinear systems is being developed for linear memory and nonlinear no-memory devices alternating in cascade, particularly multistage amplifiers and the saturating servo. An important part of this work is the search for inequalities which bound the behavior of complex nonlinear systems by that of simpler ones.

3. In the study of the characterization of random processes by series expansions it has been found that a certain set of orthonormal functions minimizes the truncation error. This fact is of importance in practical application of these expansions. Further investigation will be made into their application to filtering and coding.

4. Some inequalities for correlation functions are being studied. The properties of autocorrelation functions are being investigated from the standpoint that they belong to the class of functions whose Fourier transforms are positive.

5. It is felt that random graphs provide abstract models of communication nets in which the links between stations are not perfect and have a nonzero probability of failure. Present research is being directed toward determining link redundancy and desirable link configurations that can provide a desired degree of reliable communication between any two stations or between all stations in the net. The available mathematical tools are limited, and effort is also being directed toward obtaining general techniques. We appreciate the assistance of Professor E. Arthurs in this study.

6. Our ability to detect and resolve radar targets may depend strongly on our full knowledge of the noise encountered and the radar circuit. Incomplete knowledge of radar noise may be the result of measurement difficulties or of nonstationariness. Incomplete knowledge of the radar circuit may be attributable to manufacturing tolerances of the components or to the effect of temperature on the components. A study of the effects of the uncertainties in radar problems is being carried out. This work has been suggested to a member of this group by Professor W. M. Siebert.

Y. W. Lee

A. NOISE-LEVEL ESTIMATION

A model for radar noise y(t) is shown in Fig. XIII-1. In this model a stationary-Gaussian-noise generator has an output x(t) which is passed through an amplifier with gain $(N_o)^{1/2}$ to produce the radar noise y(t). Our problem is that of estimating N_o from a sample of the noise y(t) which is of finite duration.

Two cases are considered. In the first, it is assumed that the power density

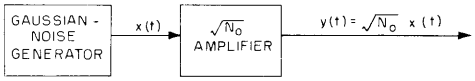

Fig. XIII-1. Radar-noise model.

spectrum of x(t) is known exactly. For this case it will be shown that the value of N_o can be determined exactly from a finite sample of y(t). In the second case it is assumed that there is some uncertainty in our knowledge of the spectrum of x(t). For this case a measure of the obtainable accuracy of the estimate of N_o will be given.

This work was suggested by Professor W. M. Siebert.

1. Spectrum Known Exactly

It is assumed that the power density spectrum of x(t) in Fig. XIII-1 is known exactly. A sample of noise y(t) for $0 \leq t \leq T_o$ is given. It will be shown that it is possible to estimate N_o exactly from this sample.

Consider the ergodic ensemble $\{x(t)\}$ each of whose members has the same known power density spectrum. Let the ensemble average of the product of x(t) at $t = t_1$ and x(t) at $t = t_2$ be given by

$$E\{x(t = t_1)\, x(t = t_2)\} = R_1(t_1, t_2) \tag{1}$$

Let the autocorrelation function corresponding to the known spectrum of x(t) be given by $R_o(\tau)$, where $\tau = t_2 - t_1$. The ergodic condition implies that

$$R_1(t_1, t_2) = R_o(t_2 - t_1) = R_o(\tau) \tag{2}$$

Grenander (1) has shown that the expression

$$\lim_{M \to \infty} \sum_{n=1}^{M} \frac{1}{\lambda_n} \left[\int_0^{T_o} y(t)\, \phi_n(t)\, dt \right]^2 \tag{3a}$$

converges in probability to N_o. Now $\{\phi_n(t)\}$ and $\{\lambda_n\}$ are defined by the integral equation

$$\int_0^{T_o} R_1(t_1, t_2)\, \phi_n(t_1)\, dt_1 = \lambda_n \phi_n(t_2) \qquad 0 \leq t_2 \leq T_o \tag{3b}$$

The normalized members of the set $\{\phi_n(t)\}$ are the characteristic functions of $R_1(t_1, t_2)$, and the $\{\lambda_n\}$ are its characteristic values.

After we have solved Eq. 3b for the sets $\{\phi_n(t)\}$ and $\{\lambda_n\}$, we can use them, together with the given noise sample y(t), for computing the summation of expression 3a. The summation is our estimate of N_o, and this estimate converges in probability to N_o.

2. Spectrum Not Known Exactly

The results of the previous section are unrealistic because the assumption of exact knowledge of the spectrum of x(t) would not be valid in practice. In the following discussion some uncertainty in the knowledge of the spectrum of x(t) is assumed. A method is then derived for estimating N_o from a sample of y(t) of finite duration. The variance of this estimate is then given.

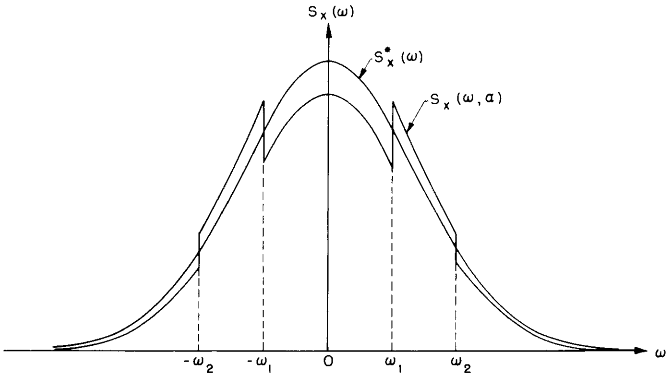

Fig. XIII-2. Power density spectrum of x(t).

Instead of assuming exact knowledge of the spectrum of x(t), we now assume that the spectrum of x(t) is one of an ensemble of spectra $\{S_x(\omega, a)\}$ with the average value $S_x^*(\omega)$. Figure XIII-2 shows $S_x^*(\omega)$ and a typical member of the ensemble $\{S_x(\omega, a)\}$. The members of the ensemble are generated by the following process

$$S_x(\omega, a) = S_x^*(\omega)\ [1 + a_i] \qquad \begin{aligned} \omega_{i-1} < |\omega| < \omega_i \\ i = 1, 2, 3, \ldots \end{aligned} \tag{4}$$

where $\{a_i\}$ are random variables with the properties

$$\left.\begin{aligned} & a_i > -1 \\ & E[a_i] = 0 \\ & E[a_i a_j] = \begin{cases} \sigma_i^2 & i = j \\ 0 & i \neq j \end{cases} \end{aligned}\right\} \tag{5}$$

In order to estimate N_o, we are given: $S_x^*(\omega)$, the average spectrum of x(t); $\{\sigma_i^2\}$ and

$\{\omega_i\}$, the sets that describe the uncertainty in the spectrum of x(t); and a sample of the received noise y(t), $0 \leq t \leq T_o$, where T_o is the duration of the sample. Then we make an estimate of N_o, based on these quantities. We shall want our estimate to be unbiased and to have minimum variance. These conditions will now be stated more precisely. Let Z be the estimate of N_o. Z will be a function of the received sample of noise, y(t). By definition,

$$E_{\alpha, x}\{Z\} = \iint p(\alpha, x)\ Z\ d\alpha\ dx \tag{6}$$

"Unbiased condition" means that

$$E_{\alpha, x}\{Z\} = N_o \qquad \text{for all } N_o \tag{7}$$

"Minimum variance condition" means that, subject to the unbiased constraint, the estimate must satisfy

$$E_{\alpha, x}\{(Z-N_o)\} = \text{minimum} \tag{8}$$

In general, the estimation process will be some nonlinear operation on the received sample of noise, y(t), with $0 \leq t \leq T_o$. We consider only those nonlinear processes for which we can use the Wiener expansion (3); that is

$$Z = k_o + \int_0^{T_o} k_1(\tau_1)\ y(\tau_1)\ d\tau_1 + \int_0^{T_o}\int_0^{T_o} k_2(\tau_1, \tau_2)\ y(\tau_1)\ y(\tau_2)\ d\tau_1\ d\tau_2$$
$$+ \int_0^{T_o}\int_0^{T_o}\int_0^{T_o} k_3(\tau_1, \tau_2, \tau_3)\ y(\tau_1)\ y(\tau_2)\ y(\tau_3)\ d\tau_1\ d\tau_2\ d\tau_3 + \dots \tag{9}$$

in which the k_n are to be determined.

When we write expression 9 in terms of x(t) with

$$y(t) = (N_o)^{1/2} \cdot x(t) \tag{10}$$

and take the expected value of Z, we obtain

$$\begin{aligned} E_{x, \alpha}\{Z\} = k_o &+ (N_o)^{1/2} \int_0^{T_o} k_1(\tau_1)\ E_{x, \alpha}\{x(\tau_1)\}\ d\tau_1 \\ &+ N_o \int_0^{T_o}\int_0^{T_o} k_2(\tau_1, \tau_2)\ E_{x, \alpha}\{x(\tau_1)\ x(\tau_2)\}\ d\tau_1\ d\tau_2 \\ &+ N_o^{3/2} \int_0^{T_o}\int_0^{T_o}\int_0^{T_o} k_3(\tau_1, \tau_2, \tau_3)\ E_{x, \alpha}\{x(\tau_1)\ x(\tau_2)\ x(\tau_3)\}\ d\tau_1\ d\tau_2\ d\tau_3 \end{aligned} \tag{11}$$

Only the quadratic term (K_2) is linear in N_o. The unbiased estimator constraint implies that the expected value of Z will be linear in N_o. Therefore only the quadratic term in the Wiener expansion can be used. The estimation problem becomes that of finding a quadratic kernel $K_2(T_1, T_2)$ that satisfies Eqs. 7 and 8.

The quadratic kernel $K_2(T_1, T_2)$ is assumed to be symmetric in T_1 and T_2. The kernel can be expanded in the following form:

$$\begin{aligned} K_2(\tau_1, \tau_2) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} a_{mn} \cos \frac{m 2\pi \tau_1}{T_o} \cos \frac{n 2\pi \tau_2}{T_o} \\ + \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} b_{mn} \sin \frac{m 2\pi \tau_1}{T_o} \sin \frac{n 2\pi \tau_2}{T_o} \\ + \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{1}{2} c_{mn} \left(\cos \frac{m 2\pi \tau_1}{T_o} \sin \frac{n 2\pi \tau_2}{T_o} + \sin \frac{n 2\pi \tau_1}{T_o} \cos \frac{m 2\pi \tau_2}{T_o} \right) \end{aligned} \tag{12}$$

Equations 7 and 8 are to be satisfied by a proper choice of the sets $\{a_{mn}\}$, $\{b_{mn}\}$, and $\{c_{mn}\}$.

Let us define

$$\begin{aligned} A_m = \int_0^{T_o} \cos \frac{m 2\pi t}{T_o} y(t) \, dt \\ B_m = \int_0^{T_o} \sin \frac{m 2\pi t}{T_o} y(t) \, dt \end{aligned} \tag{13}$$

Then $\{A_m\}$ and $\{B_m\}$ are the Fourier coefficients of the sample of y(t). Combining Eqs. 9, 12, and 13, we obtain

$$\begin{aligned} Z = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} a_{mn} A_m A_n \\ + \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} b_{mn} B_m B_n \\ + \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} c_{mn} A_m B_n \end{aligned} \tag{14}$$

By definition,

$$E_x\{A_m\} = \int p(x|\alpha)\, A_m\, dx \tag{15}$$

It is assumed that the sample duration T_o is large enough so that for all α the Fourier coefficients can be considered to be uncorrelated. That is,

$$\begin{aligned} &E_x\{A_m A_n\} = 0 && m \neq n \\ &E_x\{B_m B_n\} = 0 && m \neq n \\ &E_x\{A_m B_n\} = 0 && \text{for all m and n} \end{aligned} \tag{16}$$

The determination of the sets $\{a_{mn}\}$, $\{b_{mn}\}$, and $\{c_{mn}\}$ that satisfy Eqs. 7 and 8 is a long but straightforward algebraic process, which will not be presented here. Two of the results will be given without proof:

$$\begin{aligned} \text{(i)}\quad &a_{mn} = 0 && m \neq n \\ &b_{mn} = 0 && m \neq n \\ &c_{mn} = 0 && \text{all m and n} \end{aligned} \tag{17}$$

This result depends on the assumptions given in Eqs. 16. Equation 14 then simplifies to

$$Z = \sum_{m=0}^{\infty} a_{mm} A_m^2 + \sum_{m=1}^{\infty} b_{mm} B_m^2 \tag{18}$$

(ii) The variance of the estimation procedure is

$$E_{\alpha,x}\{(Z - N_o)^2\} = \frac{N_o}{T_o}\left[\sum_{i=1} \frac{\omega_i - \omega_{i-1}}{2\pi(1+\sigma_i^2) + \sigma_i^2 T_o(\omega_i - \omega_{i-1})}\right]^{-1} \tag{19}$$

The expression

$$\sum_{i=1} \frac{\omega_i - \omega_{i-1}}{2\pi(1 + \sigma_i^2) + \sigma_i^2 T_o(\omega_i - \omega_{i-1})}$$

can be defined as an effective bandwidth. The motivation for such a definition is found in the following example.

Consider the case in which

$$\begin{aligned} \sigma_1^2 &= 0 \\ \sigma_i &= \infty \qquad\qquad i = 2, 3, 4, \ldots \end{aligned} \tag{20}$$

The variance of the estimate of N_o as given by Eq. 19 then becomes

$$E_{a,x}\{(Z-N_o)^2\} = \frac{N_o}{T_o}\frac{1}{\frac{\omega_1}{2\pi}}$$

The quantity $\omega_1/2\pi$ is the useful bandwidth in cycles per second. The variance of the estimate of N_o is seen to be inversely proportional to the time effective-bandwidth product.

D. A. Chesler

References

1. U. Grenander, Stochastic processes and statistical inference, Ark. Mat., Bd. 1, No. 17, p. 221 (1950).
2. D. Slepian, Some comments on the detection of Gaussian signals in Gaussian noise Trans. IRE, PGIT, vol. IT-4, no. 3, pp. 65-68 (June 1958).
3. N. Wiener, Nonlinear Problems in Random Theory (The Technology Press, Cambridge, Mass., and John Wiley and Sons, Inc., New York, 1958).

B. CONNECTIVITY IN RANDOM GRAPHS – TWO LIMIT THEOREMS

The two theorems presented in this report are the first results in a study of connectivity in random graphs. This initial research has been devoted to limit theorems for two reasons: to provide insight into the link-redundancy that is necessary for obtaining a desired path or tree probability, and to allow the use of convenient but weak bounds on the probabilities. These bounds sharpen only in the limit. Exact probabilities of paths and trees are difficult to calculate and can not usually be reduced to workable form.

Before stating the theorems we shall define the problem. A graph is a collection of labeled nodes and associated links. In this report, the links are directed. The following restriction is assumed at all times: At most, one link may exist between two nodes in a given direction. A graph is made into a random graph by examining each link in the graph independently, destroying a link with a probability q, and retaining it with a probability, $p = 1 - q$ $(0 < p < q)$. In connection with the random graph, two probabilities are of present interest. One is the probability of path-connectivity; that is, the probability that if any ordered pair of nodes is chosen, a directed path exists from one to the other. The second is the probability of tree-connectivity; that is, the probability that all pairs of nodes are connected in both directions.

We introduce limit statements as follows. A sequence $\{n_i\}$ of positive integers is chosen in which, as i gets larger, n_i increases without bound. For each member of the sequence, a number ℓ_i is assigned, and a scheme is given for constructing a graph with n_i nodes and ℓ_i links. The graph is then randomized, and path- and tree-probabilities are calculated. Thus a sequence of path-probabilities and tree-probabilities is obtained. The question arises, Is there some minimal manner in which the number of links, ℓ_i, can be assigned so that either or both of the probability sequences tend to one in the limit? This question is answered in Theorems I and II.

Theorem I. For a graph with directed links, it is necessary and sufficient for the path-probability to tend to one that the average number of links per node, $\bar{\ell}_i = \ell_i/n_i$, tend to infinity, independent of the way in which n_i tends to infinity.

Theorem II. For a graph with directed links, it is necessary and sufficient that the average number of links per node, $\bar{\ell}_i$ grow faster than log n_i. In particular, it is necessary that $\bar{\ell}_i$ grow as $\log_{1/q}(n_i) + A_i$, with A_i tending to infinity as slowly as desired. It is sufficient that $\bar{\ell}_i$ grow as $\log n + (1+\epsilon) \log \log n_i$, where ϵ is any number greater than zero, and the logarithms are all to one base $(1 - p^2)^{-1}$.

First, consider the proofs of the necessity statements. For the path probability to tend to one, the number of links in every cut-set must tend to infinity; if not, the probability that the graph is separated at the cut-set remains nonzero. In particular, the number of links emanating from each node must tend to infinity, since the nodes involved in the path are not prechosen. This completes the necessity proof of Theorem I.

The necessity proof for Theorem II is a bit more difficult. Since all nodes must be connected for a tree, the probability of a tree, P_T, is less than the probability that every node have at least one outgoing link intact. Let ℓ_j be the number of links emanating from node j, before randomization, at some stage in the sequence, i. (The subscript i will be suppressed hereafter.) Then,

$$P_T \leq \left(1 - q^{\ell_1}\right)\left(1 - q^{\ell_2}\right) \dots \left(1 - q^{\ell_n}\right) \tag{1}$$

where q is the probability that a link fail. The right-hand side of Eq. 1 can be bounded above by use of the following lemma.

Lemma I. If $\ell = \ell_1 + \ell_2 + \dots + \ell_n$ is held fixed, the right-hand side of Eq. 1 is maximized by setting

$$\ell_1 = \ell_2 = \dots = \ell_n = \bar{\ell} \tag{2}$$

Stated another way, the probability of tree-connectivity, P_T, for any graph of n nodes and ℓ links is bounded above by Eq. 1 when $\ell_j = \ell/n$ for all j.
Therefore

$$P_T \leq (1 - q^{\bar{\ell}})^n \tag{3}$$

The proof of Lemma I follows from the observation that the second derivative of $(1 - q^x)$ with respect to x is negative and nonzero for all finite positive x.

An additional bound is now needed for the right-hand side of Eq. 3.

Lemma II. $a \geq 0$ implies $1 - a \leq e^{-a}$.

Lemma II is obvious for $a \geq 1$ and can be proved for $a < 1$ by drawing a graph or by taking logarithms of both sides and expanding the logarithm on the left-hand side about $a = 0$.

Thus, with Lemma II, Eq. 3 can be rewritten:

$$P_T \leq \left(1 - q^{\bar{\ell}}\right)^n \leq \exp\left(-nq^{\bar{\ell}}\right) \tag{4}$$

If P_T is to tend to one, then the exponent on the right-hand side of Eq. 4 must tend to zero. If we take the logarithm of the exponent, the condition becomes

$$\log n - \bar{\ell} \log \frac{1}{q} \rightarrow -\infty \tag{5}$$

which is the desired result. Equation 5 states that the average number of links per node must grow faster than log n.

To prove the sufficiency statements of Theorems I and II, it is only necessary to supply a structure with the proper behavior. One structure that can be used is the hierarchical structure of order j, H_j. H_j is constructed as follows.

First, a complete graph of N_1 nodes is constructed. A complete graph is one in which every node is connected to every other node with a directed link, and therefore requires a total of $N_1(N_1 - 1)$ links. A favored node is chosen from the N_1 nodes, and the complete graph is viewed as a supernode, with external connections to the favored node. Now, N_2 of these supernodes are connected in a complete graph, and this complete graph is treated as a (super) supernode. Again, a node is chosen for external connections (from one of the N_2 originally chosen nodes), and N_3 of these supernodes of level 2 are connected into a complete graph. This process is continued until N_j supernodes of level j - 1 are connected into a complete graph. The final structure is H_j. The choice of the number of nodes to be used at each level is dictated by the following lemma.

Lemma III. For an n-node hierarchical graph, and for $N_1 \gg 1$, the required number of links is minimized by setting

$$N_{i+1} = \frac{1}{2} N_i^2 \qquad i = 1, 2, \ldots, j-1 \tag{6}$$

The proof of this lemma is involved, algebraically, and is not presented here.

If Lemma III is obeyed, the average number of links per node in the entire graph is related to the number of links in the first-level structures, N_1, by

$$\bar{\ell}(j) = N_1 + \frac{N_2}{N_1} + \frac{N_3}{N_1N_2} + \dots + \frac{N_j}{N_1 \dots N_{j-1}} \tag{7a}$$

$$\bar{\ell}(j) = \left[2 - \left(\frac{1}{2}\right)^{j-1}\right] N_1 \tag{7b}$$

Furthermore, the total number of nodes in H_j, $n(j)$, is related to the number of nodes in the first-order structure, N_1, by

$$n(j) = N_j N_{j-1} \dots N_1 \tag{8a}$$

$$n(j) = \left(\frac{1}{2}\right)^{2^j-j-1} N_1^{2^j-1} \tag{8b}$$

Both Eqs. 7a and 8a are obtained by a brief consideration of the structure of H_j, and Eqs. 7b and 8b result from a substitution of Eq. 6 in Eqs. 7a and 8a, respectively. Finally, Eqs. 7b and 8b can be combined to obtain

$$\bar{\ell}(j) = 4(2^j - 1)\left(\frac{1}{2}\right)^{\frac{j2^j}{2^j-1}} n(j)^{\frac{1}{2^j-1}} \tag{9}$$

Equation 9 is important in the proof of sufficiency in Theorems I and II. It states that in the j^{th}-order hierarchy, the average number of links per node grows as a fractional power of the number of nodes. It can be shown that the probability of a tree involving paths between nodes of length equal to 2j, at most, tends to one in H_j as $n(j)$ tends to infinity. Thus, in the hierarchical structure, bounding the maximum length of paths requires that the average number of links per node grow as a fractional power of the number of nodes.

However, in the hypotheses of Theorems I and II, no bound is put on maximum path length. Thus, as the number of nodes is increased, the order of the hierarchy can be increased also. It is necessary to determine how fast the order can be increased and still provide that the probability of path-connectivity or of tree-connectivity tend to one. This is done implicitly in the following proof of sufficiency for Theorem I.

Let $P(n)$ be the minimum probability of a path taken over all pairs of nodes in the n-node graph, H_j.

Lemma IV. The probability of a path between two nodes, P_2, in a complete graph is greater than the probability of a path of length 2 in the complete graph. But there are n - 1 possible paths of length 2, all of which are statistically independent. Thus,

$$P_2 \geqslant 1 - (1 - p^2)^n \tag{10}$$

In a j^{th}-order hierarchy, the paths of lowest probability occur for nodes that are so distant that they can only be joined by climbing up from the first level to the second level, and up to the j^{th} level, and then back down to the first level again. At each change of level, a path is required through a complete graph. Thus, the minimum probability of a path over all pairs of nodes, P(n), is, by Eq. 10,

$$P(n) \geqslant \left[1-(1-p^2)^{N_1}\right]\left[1-(1-p^2)^{N_2}\right]\dots\left[1-(1-p^2)^{N_{j-1}}\right]\dots\left[1-(1-p^2)^{N_1}\right] \tag{11a}$$

$$P(n) \geqslant \left\{\left[1-(1-p^2)^{N_1}\right]\left[1-(1-p^2)^{N_2}\right]\dots\left[1-(1-p^2)^{N_j}\right]\right\}^2 \tag{11b}$$

Multiplying out the factors in Eq. 11b, and dropping all but the first two terms in the resulting sum, we obtain

$$[P(n)]^{1/2} \geqslant 1 - \sum_{k=1}^{j} (1-p^2)^{N_k} \tag{12a}$$

since the part discarded is positive. Equation 12a can be rewritten with the aid of Eq. 6 as

$$[P(n)]^{1/2} \geqslant 1 - \sum_{k=1}^{j} (1-p^2)^{\left(\frac{1}{2}\right)^{2^k-k-1} N_1^{2^k-1}} \tag{12b}$$

It can be easily shown that the right-hand side of Eq. 12b can be bounded by

$$[P(n)]^{1/2} \geqslant 1 - \sum_{k=1}^{j} (1-p^2)^{kN_1} \tag{13a}$$

$$[P(n)]^{1/2} \geqslant 1 - \frac{(1-p^2)^{N_1}}{1-(1-p^2)^{N_1}} \tag{13b}$$

However, the right-hand side of Eq. 13b tends to one if N_1 tends to infinity. Since, by Eq. 7b,

$$\bar{\ell}(j) = \left[2-\left(\frac{1}{2}\right)^{j-1}\right] N_1 < 2N_1 \tag{14}$$

it is sufficient that $\bar{\ell}$ tend to infinity, however slowly. This is the desired result.

It is possible in the same manner to prove the sufficiency statement of Theorem II, by using the hierarchical structure. However, another structure yields the result more easily, and also provides variety.

Consider a graph with its nodes arranged in a rectangular array, with s columns and k rows. Each row is made into a complete subgraph by use of $s(s-1)$ links. Successive rows are connected together by having a single directed link from a node in one row to the node directly beneath it, and from nodes in the bottom row to the respective nodes in the top row. Note that $\bar{\ell} = (s-1) + 1 = s$. ($\bar{\ell}$ will be used in place of s in the following discussion.)

One way of obtaining a tree in this structure is to have a tree existing in every row, and have at least one link existing between successive rows. This is not the only way of obtaining a tree; if a row does not have a tree, the different pieces in the row can still be connected by paths extending through the remainder of the graph. However, it is convenient to consider those cases in which there is a tree in every row, and thereby obtain a useful lower bound to the probability of an unrestricted tree in the entire graph. Call the probability of an unrestricted tree P_T. A lower bound to the probability of a tree in a row is obtained from Eq. 10 as follows. Equation 10 states that the probability of a path in an n-node complete graph, P_2, is bounded by $P_2 \geq 1 - (1 - p^2)^n$.

Lemma V. The probability of a tree in an n-node complete graph, P_3, is bounded by

$$P_3 \geq P_2^{n^2} \qquad (15)$$

The proof of Lemma V follows from noting that there are n^2 pairs of nodes in the complete graph, each of which must have a connecting path. However, the paths for all of these pairs are not independent and, to obtain the correct probability of a tree, conditional probabilities must be used. However, the conditional probability of a path connecting a pair of nodes, given that other paths are in existence, is certainly greater than the probability without such information. (The knowledge effectively increases the reliability of all links.) Hence, ignoring this knowledge results in the lower bound specified in the lemma.

Now, if $\bar{\ell}$ is the number of columns and k is the number of rows, k can be replaced by $n/\bar{\ell}$, where n is the total number of nodes in the graph.

If we apply Eqs. 10 and 15 for each of the $n/\bar{\ell}$ rows, and require the existence of at least one of the $\bar{\ell}$ links between rows, the following bound on the probability of a tree, P_T, in the entire graph is obtained. Thus

$$P_T \geq \left[1 - q^{\bar{\ell}}\right]^{\left(\frac{n}{\bar{\ell}}\right)} \left[1 - (1 - p^2)^{\bar{\ell}}\right]^{\bar{\ell}^2 \left(\frac{n}{\bar{\ell}}\right)} \qquad (16)$$

where the first factor takes into account connections between rows, and the second factor takes into account trees in each row.

The following inequality is now useful. For $0 \leq x \leq 1$ and $n > 1$

$$(1-x)^n \geq 1 - nx \qquad (17)$$

and Eq. 16 becomes

$$P_T \geq \left[1 - \frac{n}{\bar{\ell}} q^{\bar{\ell}}\right]\left[1 - n\bar{\ell}(1 - p^2)^{\bar{\ell}}\right] \tag{18a}$$

$$P_T = 1 - \frac{n}{\bar{\ell}} q^{\bar{\ell}} - n\bar{\ell}(1 - p^2)^{\bar{\ell}} + n^2 q^{2\bar{\ell}}(1 + p)^{\bar{\ell}} \tag{18b}$$

It is only necessary to find conditions on $\bar{\ell}$ to insure that $n/\bar{\ell}\ q^{\bar{\ell}}$ and $n\bar{\ell}(1 - p^2)^{\bar{\ell}}$ go to zero as n increases. Taking the logarithm of $n/\bar{\ell}\ q^{\bar{\ell}}$, we obtain the first condition.

$$\log n - \log \bar{\ell} - \bar{\ell} \log \frac{1}{q} \to -\infty \tag{19}$$

Likewise, taking the logarithm of $n\bar{\ell}(1 - p^2)^{\bar{\ell}}$, we obtain the second condition, which is stronger than the first.

$$\log n + \log \bar{\ell} - \bar{\ell} \log \frac{1}{1 - p^2} \to -\infty \tag{20}$$

It is therefore sufficient the $\bar{\ell}$ grow faster than log n in order that the probability of a tree tend to one. It can be shown that a sharper simple bound than Eq. 20 cannot be obtained from the hierarchical structure, even by varying the number of supernodes at various levels.

It has been shown that if a structure is intelligently chosen, the sufficiency requirements of Theorems I and II can be met. It is not known, at the present time, whether an unintelligent choice – for example, a random choice – will suffice. This is being investigated.

I want to acknowledge the advice and encouragement of Prof. E. Arthurs, who anticipated the existence of Theorem I and II.

I. M. Jacobs

C. MINIMIZATION OF TRUNCATION ERROR IN SERIES EXPANSIONS OF RANDOM PROCESSES

In certain problems in the theory of random signals and noise it is convenient to be able to represent a random process in a particular interval of time by a countable set of random variables or coordinates (1). One possible set is the set of Fourier coefficients obtained by expanding the process in a series with respect to some complete orthonormal set of functions $\{\psi_n(t)\}$. If a finite series is used to approximate the process, a mean-square error that is a function of the set $\{\psi_n(t)\}$ is introduced. In this report we endeavor to find that set $\{\phi_n(t)\}$ which reduces this truncation error to a minimum.

For the expansion

$$x(t) = \sum_{n=1}^{\infty} a_n \psi_n(t)$$

the coordinates are given by

$$a_n = \int_a^b x(t) \psi_n(t) dt$$

where [a, b] is the interval of interest. The mean and covariances are given by

$$E[a_n] = \int_a^b E[x(t)] \psi_n(t) dt$$

and

$$E[a_n a_m] = E\left[\int_a^b x(t) \psi_n(t) dt \int_a^b x(s) \psi_m(s) ds\right]$$

$$= \int_a^b \int_a^b R(t, s) \psi_n(t) \psi_m(s) dt ds$$

where E denotes expectation, and R(t, s) is the correlation function. It is convenient to think of such a representation as a point in "signal space" with the a_n's as coordinates on an infinite set of orthogonal axes. Distance between points can be defined in a way that is analogous to physical space; that is

$$d^2 = \sum_{n=1}^{\infty} (a_n - b_n)^2$$

where d is distance, and $(a_1, a_2, \ldots)$ and $(b_1, b_2, \ldots)$ are the coordinates of the two points. Because of Parseval's theorem this is equivalent to the rms difference between functions.

There are two variants of the problem in which we are interested: (a) Given an integer N, what set of orthonormal functions minimizes the truncation error of using only N coordinates? (b) Given a truncation error ξ_o, what set of functions minimizes the number of coordinates, N, that is required to attain this error? We shall prove that the answer to the first question is the set of orthonormal eigenfunctions, $\{\phi_n(t)\}$, of the integral equation

$$\int_a^b R(t, s) \phi_n(t) dt = \beta_n \phi_n(s) \qquad a \leqslant s \leqslant b \qquad (1)$$

with the eigenvalues so arranged that

$$\beta_1 \geqslant \beta_2 \geqslant \beta_3 \geqslant \dots \tag{2}$$

Then we shall show that the second question is equivalent to the first. This set of functions is also the solution to the problem of finding that set of functions for which the coordinates are independent (2) so that for

$$x(t) = \sum_{n=1}^{\infty} a_n \phi_n(t)$$

where

$$a_n = \int_a^b x(t)\, \phi_n(t)\, dt$$

we have

$$E[a_n a_m] = \begin{cases} \beta_n & n = m \\ 0 & n \neq m \end{cases} \tag{3}$$

and

$$\sum_{n=1}^{\infty} \beta_n = E\left[\int_a^b x^2(t)\, dt\right] = \int_a^b R(t,t)\, dt \tag{4}$$

We shall assume that R(t, s) is positive definite, so that the solutions to Eq. 1 form a complete orthonormal set (3). It will be safe to assume that most physical processes satisfy this restriction; perfectly bandlimited processes are ruled out, however.

Denoting the mean-square truncation error by

$$\xi_N\left[\{\psi_n(t)\}\right] = E\left[\int_a^b \left[x(t) - x_{N,\psi}(t)\right]^2 dt\right]$$

where

$$x_{N,\psi}(t) = \sum_{n=1}^{N} a_n \psi_n(t)$$

we shall show that

$$\xi_N\left[\{\psi_n(t)\}\right] \geqslant \xi_N\left[\{\phi_n(t)\}\right] \tag{5}$$

for any orthonormal set $\{\psi_n(t)\}$ and where $\{\phi_n(t)\}$ are the solutions to Eq. 1.

Suppose we expand the process in terms of $\{\phi_n(t)\}$ so that

$$x_{N,\phi}(t) = \sum_{n=1}^{N} a_n\phi_n(t) \tag{6}$$

where

$$a_n = \int_a^b x(t)\ \phi_n(t)\ dt \tag{7}$$

The truncation error is then given by

$$\xi_N\left[\{\phi_n(t)\}\right] = E\left[\int_a^b \left[x(t) - x_{N,\phi}(t)\right]^2 dt\right] \tag{8}$$

The integral in brackets is the random variable

$$\int_a^b \left[x(t) - x_{N,\phi}(t)\right]^2 dt = \int_a^b x^2(t)\ dt - 2\int_a^b x(t)\ x_{N,\phi}(t)\ dt$$

$$+ \int_a^b x_{N,\phi}^2(t)\ dt$$

$$= \int_a^b x^2(t)\ dt - 2\int_a^b x(t) \sum_{n=1}^{N} a_n\phi_n(t)\ dt + \int_a^b \sum_{n=1}^{N} a_n\phi_n(t) \sum_{m=1}^{N} a_m\phi_m(t)\ dt$$

If we interchange the order of summation and integration, we have

$$= \int_a^b x^2(t)\ dt - 2\sum_{n=1}^{N} a_n \int_a^b x(t)\ \phi_n(t)\ dt + \sum_{n=1}^{N}\sum_{m=1}^{N} a_na_m \int_a^b \phi_n(t)\ \phi_m(t)\ dt$$

and from Eq. 7 and orthogonality

$$= \int_a^b x^2(t)\ dt - 2\sum_{n=1}^{N} a_n^2 + \sum_{n=1}^{N} a_n^2 = \int_a^b x^2(t)\ dt - \sum_{n=1}^{N} a_n^2$$

From Eq. 8

$$\xi_N\left[\{\phi_n(t)\}\right] = \int_a^b E\left[x^2(t)\right] dt - \sum_{n=1}^{N} E\left[a_n^2\right]$$

and from Eqs. 3 and 4, we have

$$= \int_a^b R(t,t)\,dt - \sum_{n=1}^{N} \beta_n$$

$$= \sum_{n=1}^{\infty} \beta_n - \sum_{n=1}^{N} \beta_n = \sum_{n=N+1}^{\infty} \beta_n \tag{9}$$

Suppose, now, that we create any other complete set of orthonormal functions $\{\psi_n(t)\}$ from the original set by means of an orthogonal transformation η

$$\eta\phi_n = \psi_n \qquad n = 1, 2, \ldots$$

that is

$$\psi_i(t) = \sum_{j=1}^{\infty} \eta_{ij}\phi_j(t); \qquad i = 1, 2, \ldots$$

with

$$\left.\begin{aligned} \sum_{j=1}^{\infty} \eta_{ij}\eta_{kj} &= \delta_{ik} \\ \sum_{j=1}^{\infty} \eta_{ji}\eta_{jk} &= \delta_{ik} \end{aligned}\right\} \qquad i, k = 1, 2, \ldots, M \tag{10}$$

where δ_{ik} is the Kronecker delta. The transformation can be considered as a rotation of axes in signal space. It can be represented by the square array

$$\begin{bmatrix} \eta_{11} & \eta_{12} & \cdots \\ \eta_{21} & \eta_{22} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}$$

For simplicity, we can consider an M-dimensional signal space in which M is as large as we wish. The η_{ij} are not independent quantities; there are the $[(M+1)\ M]/2$ (Eqs. 10) that they must satisfy. Therefore there are

$$M^2 - \frac{(M+1)\ M}{2} = \frac{M(M-1)}{2} = \binom{M}{2}$$

independent quantities that characterize η. Such a set of independent quantities are the rotations, θ_i, in each of the mutually perpendicular planes in M-dimensional space that correspond to the Eulerian angles of three dimensions (4). Therefore

$$\eta = \eta\left(\theta_1, \theta_2, \ldots, \theta_{\binom{M}{2}}\right)$$

where the θ_i are independent variables. For the new set of orthonormal functions, we have

$$x_{n,\psi}(t) = \sum_{n=1}^{N} b_n \psi_n(t)$$

where

$$b_n = \int_a^b x(t)\,\psi_n(t)\,dt = \int_a^b x(t) \sum_{m=1}^{M} \eta_{nm}\phi_m(t)\,dt$$

$$= \sum_{m=1}^{M} \eta_{nm} \int_a^b x(t)\,\phi_m(t)\,dt = \sum_{m=1}^{M} \eta_{nm} a_m$$

or, in matrix notation, b = Ha, where

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_M \end{bmatrix} \qquad a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_M \end{bmatrix} \qquad H = \begin{bmatrix} \eta_{11} & \eta_{12} & \cdots & \eta_{1M} \\ \eta_{21} & \eta_{22} & \cdots & \eta_{2M} \\ \cdots & \cdots & \cdots & \cdots \\ \eta_{M1} & \eta_{M2} & \cdots & \eta_{MM} \end{bmatrix}$$

The covariances of the b's are

$$\lambda_{nm} = E[b_n b_m] = E\left[\sum_{i=1}^{M} \eta_{ni} a_i \sum_{j=1}^{M} \eta_{mj} a_j\right]$$

$$= \sum_{i=1}^{M} \sum_{j=1}^{M} \eta_{ni}\eta_{mj}\, E[a_i a_j]$$

and from Eq. 3 we have

$$\lambda_{nm} = \sum_{i=1}^{M} \eta_{ni}\eta_{mi}\beta_i$$

or

$$\Lambda = H^T\, B\, H \tag{11}$$

for which

$$\Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1M} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2M} \\ \dots & \dots & \dots & \dots \\ \lambda_{M1} & \lambda_{M2} & \dots & \lambda_{MM} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} \beta_1 & 0 & \dots & 0 \\ 0 & \beta_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \beta_M \end{bmatrix}$$

are the covariance matrices of the b's and a's, respectively.

Relation 9 also holds for the new expansion

$$\xi_N\left[\{\psi_n(t)\}\right] = \int_a^b R(t,t)\,dt - \sum_{n=1}^{N} \lambda_{nn}$$

To prove Eq. 5 it will, therefore, be sufficient to show that

$$\sum_{n=1}^{N} \lambda_{nn} \leqslant \sum_{n=1}^{N} \beta_n \tag{12}$$

where λ_{nn} and β_n are the diagonal terms of Λ and B in relation 11, and H is any orthogonal matrix. As we have mentioned, the transformation η can be considered as a series of rotations in each of the $\binom{M}{2}$ mutually perpendicular planes. We have, then,

$$\sum_{n=1}^{N} \lambda_{nn} = F\left(\theta_1, \theta_2, \dots, \theta_{\binom{M}{2}}\right)$$

where

$$F(0,0,\dots,0) = \sum_{n=1}^{N} \beta_n$$

The matrix H will be of the form

$H = [\theta_1][\theta_2]\dots[\theta_i]\dots\left[\theta_{\binom{M}{2}}\right]$ where $[\theta_i]$ is the elementary rotation matrix

$$[\theta_i] = \begin{array}{c} \\ \\ \\ \\ k \\ \\ \ell \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccc} & & & k & & \ell & \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & & \dots & & \dots & 0 \\ 0 & 1 & 0 & & \dots & & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ . & . & . & \cos\theta_i & \dots & \sin\theta_i & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ . & . & . & -\sin\theta_i & \dots & \cos\theta_i & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & & \dots & & \dots & 1 \end{bmatrix} \end{array}$$

corresponding to a rotation in the plane formed by the k^{th} and ℓ^{th} axes. The inequality (Eq. 12) can easily be proved for a single rotation, θ_i. Using only the pertinent 2 × 2 submatrices we have for Eq. 11

$$\Lambda = \begin{bmatrix} \lambda_{kk} & \lambda_{k\ell} \\ \lambda_{\ell k} & \lambda_{\ell\ell} \end{bmatrix}$$

$$= \begin{bmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{bmatrix} \begin{bmatrix} \beta_k & 0 \\ 0 & \beta_\ell \end{bmatrix} \begin{bmatrix} \cos\theta_i & \sin\theta_i \\ -\sin\theta_i & \cos\theta_i \end{bmatrix}$$

$$= \begin{bmatrix} \beta_k \cos^2\theta_i + \beta_\ell \sin^2\theta_i & (\beta_\ell - \beta_k)\cos\theta_i \sin\theta_i \\ (\beta_\ell - \beta_k)\cos\theta_i \sin\theta_i & \beta_k \sin^2\theta_i + \beta_\ell \cos^2\theta_i \end{bmatrix}$$

There are three cases to consider here: $k > N$; $k < N$, $\ell \leqslant N$; and $k \leqslant N$, $\ell > N$. For $k > N$, it is seen that the situation is unchanged and $F(0, \ldots, \theta_i, \ldots, 0) = F(0, \ldots, 0, \ldots, 0)$.

For $k < N$, $\ell \leqslant N$

$$\lambda_{ii} = \beta_i \qquad \text{all} \qquad i \neq k, \ell \qquad\qquad i = 1, 2, \ldots, N$$

$$\lambda_{kk} = \beta_k \cos^2\theta_i + \beta_\ell \sin^2\theta_i$$

$$\lambda_{\ell\ell} = \beta_k \sin^2\theta_i + \beta_\ell \cos^2\theta_i$$

and it is seen that

$$F(0, \ldots, \theta_i, \ldots, 0) = \sum_{n=1}^{N} \lambda_{nn} = \sum_{n=1}^{N} \beta_n = F(0, \ldots, 0, \ldots, 0)$$

For $k \leqslant N$, $\ell > N$, we have

$$F(0, \ldots, \theta_i, \ldots, 0) = \sum_{n=1}^{N} \lambda_{nn} = \sum_{\substack{n=1 \\ n\neq k}}^{N} \beta_n + \beta_k \cos^2\theta_i + \beta_\ell \sin^2\theta_i$$

$$= \sum_{\substack{n=1 \\ n\neq k}}^{N} \beta_n + \beta_k - (\beta_k - \beta_\ell)\sin^2\theta_i = F(0, \ldots, 0, \ldots, 0) - (\beta_k - \beta_\ell)\sin^2\theta_i \tag{13}$$

From the first two cases, we see that $F\left(\theta_1, \ldots, \theta_{\binom{M}{2}}\right)$ is invariant under a rotation either completely outside or completely inside the N-dimensional subspace formed by the first N coordinates. The first result is obvious, and the second means that any orthonormal set of N functions formed by linear combinations of the first N solutions of Eq. 1 is equally as good as the first N solutions themselves. From Eq. 13 we also see that any rotation in a plane formed by two axes corresponding to equal eigenvalues leaves $F\left(\theta_1, \ldots, \theta_{\binom{M}{2}}\right)$ unchanged, but this rotation also gives an eigenfunction of Eq. 1, since any linear combination of eigenfunctions corresponding to equal eigenvalues is an eigenfunction.

The proof of inequality 12 can be extended to the general case of $\binom{M}{2}$ rotations but the proof is lengthy and will not be presented here.

The second question at the beginning of this report can be shown to be equivalent to the first, as follows. By the previous proof we have maximized

$$F(\eta, N) = \sum_{n=1}^{N} \lambda_{nn} \tag{14}$$

for all $N \geqslant 1$; that is,

$$F(\eta_o, N) > F(\eta_1, N) \tag{15}$$

for any $\eta_1 \neq \eta_o$ and for any N, with $\eta_o = \eta(0, 0, \ldots, 0)$. Given a particular $A = \int_a^b R(t, t)\, dt - \xi_o$, where ξ_o is the truncation error, we want to show that for $F(\eta, N) = A$, if $F(\eta_o, N_o) = F(\eta_1, N_1)$ for an $\eta_1 \neq \eta_o$, then $N_o < N_1$. Suppose that $N_1 < N_o$, then, since $F(\eta_o, N_1) < F(\eta_o, N_o)$, as seen from Eq. 14, it is true that $F(\eta_1, N_1) > F(\eta_o, N_1)$. This is a contradiction, however, as seen from Eq. 15.

K. L. Jordan, Jr.

References

1. U. Grenander, Stochastic processes and statistical inference, Ark. Mat. 1, 207-209 (1950).
2. W. B. Davenport and W. L. Root, An Introduction to the Theory of Random Signals and Noise (McGraw-Hill Publishing Company, New York, 1958), pp. 96-101.
3. Ibid., p. 374, Theorem 8.
4. J. C. Slater and N. H. Frank, Mechanics (McGraw-Hill Book Company, Inc., New York, 1947), p. 107.

D. FOURIER TRANSFORMS OF POSITIVE FUNCTIONS

1. Types of Positive Functions

Functions that are the Fourier transforms of positive functions are important in engineering work. The characteristic function, which is the Fourier transform of a probability density distribution, and the autocorrelation function, which is the Fourier transform of a power-density distribution, are examples. By noting that, for a passive circuit, the real part of the impedance or admittance must be positive, we may conclude that the even part of the impulse response of a two-terminal passive network is also a function of this class. In fact, it can be rigorously shown that these three functions are, mathematically speaking, really all the same function. Thus, given a characteristic function, we can always find a random process that has this function for its autocorrelation, and we can also find a passive network that has this same function for the even part of its impulse response.

Let us consider a probability density distribution of the form

$$P(x) = \frac{a}{\pi} \frac{1}{a^2 + x^2} \tag{1}$$

The characteristic function is then

$$\begin{aligned} f(t) &= \int_{-\infty}^{\infty} P(x) \exp(jxt)\, dx \\ &= \frac{2a}{\pi} \int_0^{\infty} \frac{\cos xt}{a^2 + x^2}\, dx \\ &= \exp(-a|t|) \end{aligned} \tag{2}$$

We can find an ensemble of random waves which has the function f(t) for its ensemble autocorrelation by assuming the random waves to be sinusoidal with statistically independent random phase and frequency. For our example, we choose each member, X(t), of the ensemble as

$$X(t) = \sqrt{2} \cos[\phi + \omega t] \tag{3}$$

where the probability density distribution of ϕ over the ensemble is

$$P(\phi) = \begin{cases} \frac{1}{2\pi} & 0 \leqslant \phi < 2\pi \\ 0 & \text{elsewhere} \end{cases} \tag{4}$$

and the probability density distribution of ω over the ensemble is

$$P(\omega) = \frac{a}{\pi} \frac{1}{a^2 + \omega^2} \tag{5}$$

The ensemble autocorrelation of X(t) is then

$$\begin{aligned} R(\tau) &= \overline{X(t)\,X(t+\tau)}^{x} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 2\cos[\phi + \omega t] \cos\,[\phi + \omega(t+\tau)]\, P(\phi)\, P(\omega)\, d\phi\, d\omega \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cos(\omega\tau)\, P(\phi)\, P(\omega)\, d\phi\, d\omega \\ &= \frac{a}{\pi} \int_{-\infty}^{\infty} \frac{\cos\omega\tau}{a^2 + \omega^2}\, d\omega \\ &= \exp(-a|\tau|) \end{aligned} \tag{6}$$

which is identical with the characteristic function, Eq. 2.

Since the ensemble (Eq. 3) is not ergodic, the time autocorrelation of any one member is not equal to the ensemble autocorrelation. The time autocorrelation of any one member is, in fact, $2\cos(\omega\tau)$. However, we can always find a Gaussian random process for which the time autocorrelation is the desired f(t). Let us consider the random process, X(t), for which the probability distribution of the M random variables, $X(t_k)$, with $k = 1, 2, 3, \ldots, M$, is a normal distribution. Since a normal distribution in any number of variables is completely defined by its first- and second-order moments, we shall choose them as

$$\left.\begin{aligned} &E[X(t_k)] = 0 \\ &E[X(t_k)\,X(t_j)] = R(t_j - t_k) \end{aligned}\right\} \tag{7}$$

These moments then define a probability distribution with the characteristic function $\exp[-1/2\, Q(a)]$ with

$$Q(a) = E\left[\sum_{k=1}^{M} a_k\, X(t_k)\right]^2 = \sum_{k=1}^{M} \sum_{j=1}^{M} a_i a_j\, R(t_j - t_k) \tag{8}$$

We shall show in section 3 that Q(a) is necessarily non-negative and thus the moments (Eqs. 7) do indeed define a normal distribution. Since the time instants, t_k, are arbitrary, X(t) is an ergodic Gaussian random process (1) whose autocorrelation is

$$E[X(t_k)\,X(t_k+\tau)] = R(\tau) \tag{9}$$

We should now be able to find a network for which the even part of the impulse response is $\exp(-\alpha|t|)$. Thus, the desired impulse response is

$$h(t) = \left\{\begin{matrix} 0 & \text{for} & t < 0 \\ 2e^{-\alpha t} & \text{for} & t \geqslant 0 \end{matrix}\right\} \tag{10}$$

The total impedance can now be calculated.

$$\begin{aligned} Z(\omega) &= \int_0^\infty h(t)\,\exp(-j\omega t)\,dt \\ &= \int_0^\infty 2\,\exp[-(\alpha + j\omega)\,t]\,dt \\ &= \frac{2}{\alpha + j\omega} \end{aligned} \tag{11}$$

A network that has this function for its impedance is shown in Fig. XIII-3. The dual of this network (see Fig. XIII-4) will have the function of Eqs. 11 for its admittance.

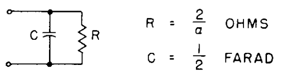

Fig. XIII-3.

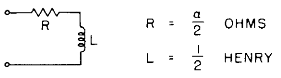

Fig. XIII-4.

In general, given the Fourier transform, f(t), of any even postive function, we can follow the procedure that has been indicated for finding a random process whose autocorrelation function is f(t), and for realizing a two-terminal passive network for which f(t) is the even part of the impulse response. Of course, the required network may not always be composed entirely of lumped parameters, as in our example, but may contain distributed parameters.

The correspondence between the autocorrelation function and the impulse response implies a more general correspondence. It can be shown that it implies a one-to-one correspondence between the M^2 impulse responses (2) of an M-port passive network and the M^2 correlation functions of M stationarily correlated random processes. Thus, if we are given an M-port network with its M^2 impulse responses, $h_{mn}(t)$, we can always find M stationarily correlated random processes whose correlation functions are

$$R_{mn}(t) = \begin{cases} h_{mn}(t) & \text{for} \quad t \geq 0 \\ h_{nm}(t) & \text{for} \quad t \leq 0 \end{cases} \tag{12}$$

and vice versa. Thus there exists a complete correspondence between network functions and correlation functions. This correspondence is illustrated with the simple

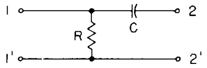

Fig. XIII-5.

example of a two-port network, as shown in Fig. XIII-5. The impedance impulse responses (3) of this network are

$$\left.\begin{aligned} h_{11}(t) &= R u_o(t) \\ h_{12}(t) &= h_{21}(t) = R u_o(t) \\ h_{22}(t) &= R u_o(t) + \frac{1}{C} u_1(t) \end{aligned}\right\} \tag{13}$$

Thus, we should be able to find two random processes whose correlation functions are

$$\left.\begin{aligned} R_{11}(\tau) &= R u_o(\tau) \\ R_{12}(\tau) &= R_{21}(\tau) = R u_o(\tau) \\ R_{22}(\tau) &= R u_o(\tau) + \frac{1}{C} \end{aligned}\right\} \tag{14}$$

Let us choose one random process, $X_1(t)$, as white noise with a power spectral density function, $S(f) = R$ watts/cycle. Then

$$R_{11}(\tau) = R u_o(\tau) \tag{15}$$

We choose our second random process, $X_2(t)$, as

$$X_2(t) = X_1(t) + \frac{1}{C^{1/2}} \tag{16}$$

That is, $X_2(t)$ is $X_1(t)$ with a dc value of $1/C^{1/2}$ added. Then

$$\left.\begin{aligned} R_{22}(\tau) &= R u_o(\tau) + \frac{1}{C} \\ R_{21}(\tau) &= R_{12}(\tau) = R u_o(\tau) \end{aligned}\right\} \tag{17}$$

2. Properties of Positive Functions

Since all three functions are mathematically identical, a property of any one of them is also a property of the other two. In discussing some of these properties we shall refer to the class of functions, f(t), that are the Fourier transforms of the positive functions, P(ω), as functions of the class P. We shall normalize P(ω) so that

$$f(o) = \int_{-\infty}^{\infty} P(\omega)\, d\omega = 1 \tag{18}$$

We shall also restrict ourselves to the case of real f(t); then, P(ω) and f(t) are even functions (4). It is seen that if $f_1(t)$ and $f_2(t)$ are functions of the class P, then so are

a) $g_1(t) = f_1(t) + f_2(t)$

b) $g_2(t) = f_1(t)\, f_2(t)$

c) $g_3(t) = f_1(t) * f_2(t) = \int_{-\infty}^{\infty} f_1(x)\, f_2(t-x)\, dx$

A number of bounds on f(t) can also be derived. Two of the bounds (5) which we have found most useful are

d) $1 \geqslant f(t) \geqslant 2f(\infty) - 1$

e) $f(2t) \geqslant 2f^2(t) - 1$

That $|f(t)| \leqslant 1$ is known. The Kolmogorov inequality, $1 - |f(2t)|^2 \leqslant 4[1 - |f(t)|^2]$, is implied by inequality e), but not conversely.

From these results a number of interesting properties of f(t) can be derived. For example, by direct use of inequality e), it is easily seen that the functions $\exp(-a\,|t|^n)$ and $[1 + a\,|t|^n]^{-1}$ are of class P only for $n \leqslant 2$. By use of b), we also find that for $n \leqslant 2$, the function $[1 + a\,|t|^n]^{-k}$ is of class P for k = 1, 2, ...

As another illustration, since, from inequality d), $f(t) \leqslant 1$, let us write f(t) in the form

$$f(t) = 1 - \epsilon(t), \qquad \epsilon(t) \geqslant 0 \tag{19}$$

Inequality e) can then be written as

$$4\epsilon(t) > \epsilon(2t), \qquad \epsilon(t) \neq 0 \tag{20}$$

To illustrate an application of this result, let us write, for small t,

$$f(t) = 1 - k|t|^n \tag{21}$$

Then

$$\epsilon(t) = k|t|^n \tag{22}$$

From Eq. 20, we require that

$$4k|t|^n > k2^n|t|^n \qquad t \neq 0$$

or

$$4 > 2^n$$

and then

$$n < 2 \tag{23}$$

Thus, near the origin, f(t) may only assume a shape that is of lower order than a parabola, as shown in Fig. XIII-6. The shaded region is disallowed for any f(t) of the form $[1 - k|t|^n]$. If we did assume a parabolic shape for f(t) near the origin, P(ω) would be negative for some ω. Wernikoff (6) has considered bounds similar to this.

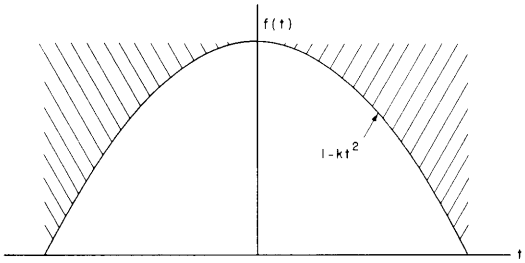

Fig. XIII-6.

However, f(t) may equal one for t ≠ 0, but it can be shown that

f) If f(t) is equal to one over any interval, then it must be one everywhere and P(ω) is an impulse at the origin. Also, the second derivative of f(t) at a point where f(t) = 1 may not be zero unless f(t) = 1 everywhere.

g) If f(t) is equal to one at some point other than the origin, then it must be periodic with a period equal to the time that it takes for it to first reach one after t = 0.

This result is useful in computing autocorrelation functions. For, if the computation yields a value of the autocorrelation equal to the initial value, R(0), we can stop computing, for the rest of the curve is determined because it must be periodic.

What can be said about f(t) if it ever reaches its maximum negative value, minus one? It can be shown that

h) f(t) may never equal minus one over any interval.

i) If f(t) is ever equal to minus one at a point, then it must be periodic. The period for this case, however, is twice the time that it takes for f(t) to first reach minus one.

3. Proof of Inequality e)

Let us now derive inequality e) for other useful inequalities can be obtained from this derivation. Since

$$f(t) = \int_{-\infty}^{\infty} P(\omega) \exp(j\,\omega t)\, d\omega \tag{24}$$

it is seen that

$$\begin{aligned}\sum_{n=1}^{K} \sum_{m=1}^{K} A_n A_m^* f(t_n - t_m) &= \int_{-\infty}^{\infty} P(\omega) \sum_{n=1}^{K} \sum_{m=1}^{K} A_n A_m^* \exp[j\omega(t_n - t_m)]\, d\omega \\ &= \int_{-\infty}^{\infty} P(\omega) \left| \sum_{n=1}^{K} A_n \exp(j\,\omega t_n) \right|^2 d\omega\end{aligned} \tag{25}$$

Since the integrand is non-negative, we find that for any set of points, t_n, and every number, A_n, $(n = 1, 2, \dots K)$ $(K = 1, 2, \dots)$,

$$\sum_{n=1}^{K} \sum_{m=1}^{K} A_n A_m^* f(t_n - t_m) \geq 0 \tag{26}$$

If we now define the matrix $[f_{nm}]$ in which $f_{nm} = f(t_n - t_m)$, then a necessary and sufficient condition that f(t) satisfy Eq. 26 is that the matrix $[f_{nm}]$ be non-negative definite. This result can be used to obtain a number of general bounds on f(t). For example, by considering the third discriminant of the matrix $[f_{nm}]$, letting $\tau_1 = t_1 - t_2$ and $\tau_2 = t_2 - t_3$, and noting that $f(0) = 1$, we have

$$\begin{vmatrix} 1 & f(\tau_1) & f(\tau_1 + \tau_2) \\ f(\tau_1) & 1 & f(\tau_2) \\ f(\tau_1 + \tau_2) & f(\tau_2) & 1 \end{vmatrix} \geq 0 \tag{27}$$

By expanding this determinant, we obtain the inequality

$$1 + 2f(\tau_1)\, f(\tau_2)\, f(\tau_1 + \tau_2) - f^2(\tau_1 + \tau_2) - f^2(\tau_1) - f^2(\tau_2) \geq 0 \tag{28}$$

which can be written in the form

$$[1 - f(\tau_1 + \tau_2)]\,[1 + f(\tau_1 + \tau_2) - 2f(\tau_1)\, f(\tau_2)] \geq [f(\tau_1) - f(\tau_2)]^2 \tag{29}$$

Since $1 - f(\tau_1 + \tau_2) \geq 0$ and $[f(\tau_1) - f(\tau_2)]^2 \geq 0$, we have

$$1 + f(\tau_1 + \tau_2) - 2f(\tau_1)\, f(\tau_2) \geq 0 \tag{30}$$

or

$$f(\tau_1 + \tau_2) \geq 2f(\tau_1)\, f(\tau_2) - 1 \tag{31}$$

By considering the special case for which $\tau_1 = \tau_2$, we obtain $f(2t) \geq 2f^2(t) - 1$, which is inequality e).

Inequality 31 is true for τ_2 negative. Thus, by making use of the eveness of f(t), we obtain the inequality

$$f(|\tau_1| - |\tau_2|) \geq 2f(\tau_1)\, f(\tau_2) - 1 \tag{32}$$

Consideration of other discriminants of the matrix $[f_{nm}]$ yields other inequalities which f(t) must satisfy.

M. Schetzen

References

1. Ergodicity is insured, since $\int_{-\infty}^{\infty} |R(\tau)|\, d\tau < \infty$. See W. B. Davenport and W. L. Root, An Introduction to the Theory of Random Signals and Noise (McGraw-Hill Publishing Company, Inc., New York, 1958), p. 67.
2. We refer here only to the impedance (or admittance) impulse response; that is, the impulse response whose Fourier transform is a transfer impedance (or admittance).
3. We define $u_0(t)$ as the unit impulse function, and $u_1(t)$ as the unit step function.
4. The results that will be discussed have been generalized to complex f(t). For the moment, however, complex f(t) is of little practical significance.
5. We, of course, assume here that $\lim_{t \to \infty} f(t) = f(\infty)$ exists. It will not exist in those cases for which f(t) contains a periodic component. For such cases, we subtract the periodic component and apply the bound given by d) to the aperiodic component that is left. This can be done because the aperiodic part of f(t) can be shown to be also of class P.
6. R. E. Wernikoff, Quarterly Progress Report, Research Laboratory of Electronics, M.I.T., April 15, 1955, pp. 38-41.

E. RESOLUTION OF A RANDOM PROCESS AND PROPERTIES OF THE COMPONENTS

In his paper "Methods of Solving Noise Problems," W. R. Bennett (1) discusses the consequences of resolving a random process $[y(t)]$ into parts $[\alpha(t)]$, $[\beta(t)]$ and $[\lambda(t)]$, whose cross spectra with a given process $[x(t)]$ are, respectively, a pure real function of frequency, a pure imaginary function of frequency, and zero. This note is intended as a heuristic justification of these consequences.

Let [x(t)] and [y(t)] be stationary processes with power-density spectra $W_x(\omega)$ and $W_y(\omega)$, respectively. The cross-power density spectrum between [x(t)] and [y(t)] is

$$W_{xy}(\omega) = U_{xy}(\omega) + i\, V_{xy}(\omega) \qquad (i = \sqrt{-1})$$

where U_{xy} and V_{xy} are real functions of ω. (Functional dependence on ω will be assumed unless otherwise stated.)

We wish to prove that, if there exists a process [z(t)] with the property that $W_{xz} = 0$, then there exists at least one set of auxiliary processes [α(t)], [β(t)], and [λ(t)], with the properties

$$W_y = W_\alpha + W_\beta + W_\lambda$$

$$W_{x\alpha} = U_{xy}$$

$$W_{x\beta} = i\, V_{xy}$$

$$W_{x\lambda} = 0$$

If, in addition, W_α and W_β are as small as possible for all ω (subject to the constraints listed above), the set is unique in the sense that W_α, W_β, and W_λ are unique. The following spectrum relations, that are due to Bennett, are then valid:

$$\text{(a)}\quad W_\alpha = \frac{U_{xy}^2}{W_x}$$

$$\text{(b)}\quad W_\beta = \frac{V_{xy}^2}{W_x}$$

$$\text{(c)}\quad W_\lambda = W_y - \frac{|W_{xy}|^2}{W_x}$$

$$\text{(d)}\quad W_{\alpha\beta} = \frac{iU_{xy}V_{xy}}{W_x}$$

$$\text{(e)}\quad W_{\alpha\lambda} = W_{\beta\lambda} = 0$$

Lemma. For any two processes [x(t)] and [y(t)], as described above, $W_x W_y \geqslant |W_{xy}|^2$.

Proof. Two well-known theorems concerning linear operations on stationary processes will be used.

Theorem I. If the random input to a linear, time-invariant system with system function Y_1 has power-density spectrum W_i, then the output will have the power-density spectrum

$$W_o = |Y_1|^2 \cdot W_i$$

(By system functions we mean voltage or current transfer ratios, which are expressed in terms of the (complex) Fourier transform of the impulse response of the system, and $\bar{Z}$ indicates the complex conjugate of Z.)

Theorem II. If the cross-power density between two random processes, [a(t)] and [b(t)], is W_{ab}, and if [a(t)] is the input to linear, time-invariant system 1 with system function Z_1 and [b(t)] is the input to linear, time-invariant system 2 with system function Z_2, then the cross-power density spectrum between the output of system 1 and the output of system 2 is

$$\bar{Z}_1 \cdot Z_2 \cdot W_{ab}$$

Proof: Consider the model of Fig. XIII-7.

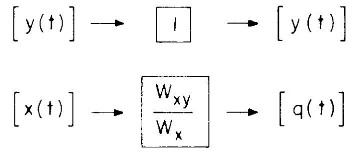

Fig. XIII-7.

This figure represents linear operations on the processes [x(t)] and [y(t)]. For example, on the second line of Fig. XIII-7 [x(t)] is the random input to a linear, time-invariant system with system function W_{xy}/W_x, whose output is [q(t)]. By Theorem I we have

$$W_q = \frac{|W_{xy}|^2}{W_x^2} \cdot W_x = \frac{|W_{xy}|^2}{W_x} \tag{1}$$

Rewriting Eq. 1, we have

$$W_x W_q = |W_{xy}|^2 \tag{2}$$

To prove our lemma it suffices to prove that $W_y(\omega) \geqslant W_q(\omega)$ for all values of ω.

Assume that there exists a radian frequency ω_1, with the property that $W_q(\omega_1) > W_y(\omega_1)$. Using Fig. XIII-7 and Theorem II, we have

$$W_{qy} = \frac{W_{yx}}{W_x} \cdot 1 \cdot W_{xy} = \frac{|W_{xy}|^2}{W_x} = W_{yq} = W_q \qquad (3)$$

since $W_{yx} = \overline{W_{xy}}$, and $W_{qy} = \overline{W_{yq}}$

Let us now consider

$$W_{y-q} = W_y - W_{qy} - W_{yq} + W_q$$

From Eqs. 3,

$$W_q - W_{yq} = 0 \qquad \text{for all } \omega$$

$$W_y - W_{qy} = W_y - W_q \qquad \text{for all } \omega$$

But by assumption

$$W_y(\omega_1) - W_q(\omega_1) < 0$$

Therefore, $W_{y-q}(\omega_1)$ is negative. This is impossible because a self-power density spectrum is always positive. Our assumption is then false, and $W_y \geqslant W_q$ for all ω. This proves our lemma.

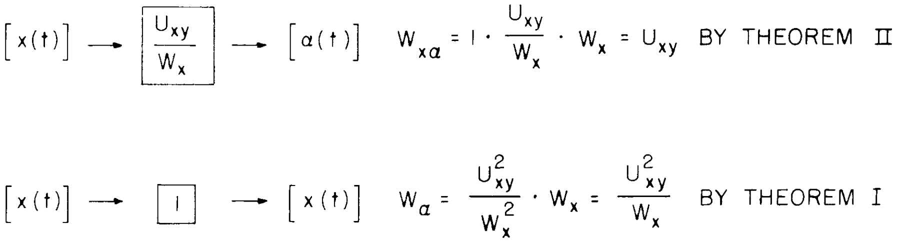
[x(t)]
$\frac{U_{xy}}{W_x}$
[α(t)]
$W_{x\alpha} = 1 \cdot \frac{U_{xy}}{W_x} \cdot W_x = U_{xy}$ BY THEOREM II
[x(t)]
1
[x(t)]
$W_\alpha = \frac{U_{xy}^2}{W_x^2} \cdot W_x = \frac{U_{xy}^2}{W_x}$ BY THEOREM I

Fig. XIII-8.

Given the ensembles [x(t)] and [y(t)], let us now consider choosing the ensemble [α(t)] so that $W_{x\alpha} = U_{xy}$. Such an [α(t)] may be constructed from [x(t)] by passing [x(t)] through the linear, time-invariant filter of Fig. XIII-8. Our lemma tells us that, in general,

$$W_\alpha \geqslant \frac{|W_{\alpha x}|^2}{W_x}$$

For our choice of [α(t)],

$$W_\alpha = \frac{U_{xy}^2}{W_x} = \frac{W_{x\alpha}^2}{W_x} = \frac{|W_{x\alpha}|^2}{W_x}$$

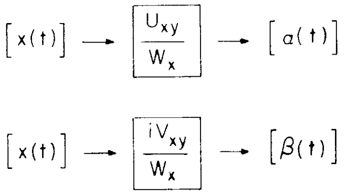

Fig. XIII-9.

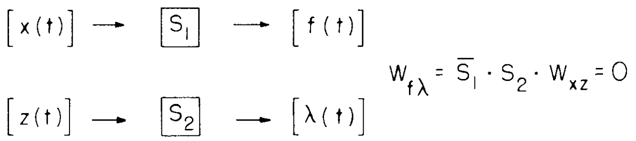

Fig. XIII-10.

Thus W_α is as small as possible for all ω. Formula (a) then follows from the fact that $[\alpha(t)]$ is selected so that W_α is a minimum for all values of ω under the condition that $W_{x\alpha} = U_{xy}$. Note that W_α is unique, because U_{xy} and W_x are unique. In exactly the same manner formula (b) follows from the like constraints on our choice of $[\beta(t)]$, and w_β is also unique.

Next consider the model of Fig. XIII-9. Theorem II tells us that

$$W_{\alpha\beta} = \frac{U_{xy}}{W_x} \cdot \frac{iV_{xy}}{W_x} \cdot W_x = \frac{iU_{xy}V_{xy}}{W_x}$$

This gives us formula (d). Adding formulas (a) and (b), we obtain

$$W_\alpha + W_\beta = \frac{U_{xy}^2 + V_{xy}^2}{W_x} = \frac{|W_{xy}|^2}{W_x} \tag{4}$$

It follows from Eq. 4 and from our lemma that

$$W_y \geqslant W_\alpha + W_\beta \tag{5}$$

Thus, in general, we must add some process $[\lambda(t)]$ to $[\alpha(t)] + [\beta(t)]$ so that $W_{\alpha+\beta+\lambda} = W_y$. The process $[\lambda(t)]$ cannot be correlated with $[x(t)]$, for, if it were, $W_{x(\alpha+\beta+\lambda)}$ would not be W_{xy}, which is required.

We originally postulated that there exists a process $[z(t)]$ with the property that $W_{xz} = 0$. Theorem II and Fig. XIII-10 (with $S_1 = 1$, $[f(t)] = [x(t)]$, and S_2 an arbitrary linear system function) show that any linear operation on $[z(t)]$ gives us a possible choice for $[\lambda(t)]$, since $W_{x\lambda} = 0$.

Now let $S_1 = U_{xy}/W_x$ and $[f(t)] = [\alpha(t)]$. Theorem II tells us that

$$W_{\alpha\lambda} = \frac{U_{xy}}{W_x} \cdot S_2 \cdot W_{xz} = \frac{U_{xy}}{W_x} \cdot S_2 \cdot 0 = 0.$$

If we again change Fig. XIII-10 so that $S_1 = iV_{xy}/W_x$ and $[f(t)] = [\beta(t)]$, it follows that $W_{\beta\lambda} = 0$. This gives us formula (e). Let us now consider

$$W_{\alpha+\beta+\lambda} = W_\alpha + W_\beta + W_\lambda + W_{\alpha\beta} + W_{\beta\alpha} + W_{\alpha\lambda} + W_{\lambda\alpha} + W_{\beta\lambda} + W_{\lambda\beta}$$

Using formulas (d) and (e), we have

$$W_{\alpha\lambda} = W_{\beta\lambda} = W_{\lambda\alpha} = W_{\lambda\beta} = 0$$

$$W_{\alpha\beta} = -W_{\beta\alpha} = \overline{W_{\beta\alpha}}$$

Therefore,

$$W_{\alpha+\beta+\lambda} = W_{\alpha} + W_{\beta} + W_{\lambda} \tag{6}$$

Equations 5 and 6 tell us that we can satisfy the condition that $W_y = W_\alpha + W_\beta + W_\lambda$ if we pick W_λ so that

$$W_\lambda = W_y - (W_\alpha + W_\beta) \tag{7}$$

Substitution of Eq. 4 in Eq. 7 gives us formula (c)

$$W_\lambda = W_y - \frac{|W_{xy}|^2}{W_x}$$

We cannot, of course, conclude that

$$[y(t)] = [\alpha(t)] + [\beta(t)] + [\lambda(t)] \tag{8}$$

However, our usual calculations involving stationary processes only utilize second-order statistics (i.e., lowest-order correlation functions and their Fourier transforms, the power-density spectra). For such purposes, we may consider Eq. 8 to be valid.

Thus, when we are interested in second-order statistics only, we can "decompose" a stationary random process $[y(t)]$ with respect to another process $[x(t)]$. (This is analogous to picking one axis of a Cartesian three-dimensional coordinate system along a vector x, and then decomposing a vector y into its projections on each axis.) The relations (a) through (e) can then be used to simplify subsequent manipulations.

D. W. Tufts

References

1. W. R. Bennett, Methods of solving noise problems, Proc. IRE 44, 609-638 (1956).

F. CANONICAL FORMS FOR NONLINEAR STATISTICAL ESTIMATORS

This report deals with a method for developing nonlinear statistical estimators in canonical forms. The optimum predictor, filter, coder, and decision operator are considered for processes that are random and stationary. Discrete processes are studied first, and then Wiener's representation (1) is used for continuous processes in conjunction with a definition of "probability density." In a sense, the method is an extension of that used by Bose (2).

N. Wiener has introduced a complete set of functionals that are orthogonal with respect to a Brownian motion. No such functionals are known for arbitrary random processes, so that it is difficult to design optimum nonlinear predictors (or other operators) for them.

However, each stationary random process has a characteristic operator that is easily computed — the operator which when it is operating on any "input" belonging to the process yields the probability density of this "input" as the "output." Such an operator is useful because an optimum predictor — in fact almost any statistical estimator — is derivable from it.

1. Discrete Processes: Synthesis of the Probability-Density Operator

Consider a time sequence: $x_o, x_1, \dots x_n$, as shown in Fig. XIII-11. We express the probability density $p(x_o, x_1, \dots x_n)$ as a Fourier sum of orthogonal functions each of which is multiplied by a suitable coefficient. The coefficients are simply the time averages of the corresponding functions.

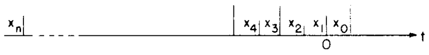

Fig. XIII-11. A time sequence.

Suppose that $p(x_o, x_1, \dots x_n)$ is well enough behaved; for example,

(a) $0 \leqslant x_i \leqslant 1$ $\qquad i = 1, 2, \dots n.$

(b) $p(x_o, x_1, \dots x_n)$ is continuous in x_i $\qquad i = 1, 2, \dots n.$

(c) $p(x_o, x_1, \dots x_n) = 0$ when any $x_i = 0$ or 1

Let $\phi_1(x), \phi_2(x) \dots$ be any complete set of functions of x that are orthonormal over $0 \leqslant x \leqslant 1$. That is,

$$\int_0^1 \phi_i(x)\, \phi_j(x)\, dx = \begin{cases} 0 & \text{if} \quad i \neq j \\ 1 & \text{if} \quad i = j \end{cases}$$

Then the set of products corresponding to all possible permutations of functions and variables forms a complete orthonormal set over the n-space $[0, 1]^n$:

$$\Phi_i(x_o, \dots x_n) = \phi_{i_o}(x_o)\, \phi_{i_1}(x_1) \dots \phi_{i_n}(x_n), \qquad i = 1, 2, \dots$$

Hence

$$p(x_o, x_1, \ldots x_n) = \sum_{i=1}^{\infty} a_i \Phi_i(x_o, \ldots x_n)$$

and

$$a_i = \int_{x_o} \cdots \int_{x_n} \Phi_i(x_o, \ldots x_n)\, p(x_o, x_1, \ldots x_n)\, dx_o \ldots dx_n$$

but for an ergodic process this expression for a_i is almost always the time average of $\Phi_i(x_o, \ldots x_n)$.

Thus, to synthesize any such operator we build any set of orthonormal (in the ordinary sense) function generators with adjustable multiplying constants a_i. See Fig. XIII-12.

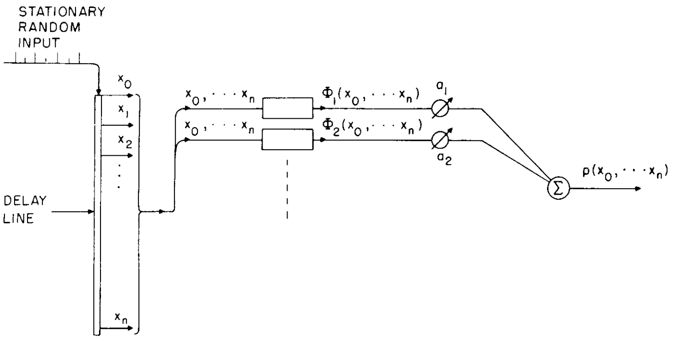

Fig. XIII-12. Canonical form for $p(x_o, x_1, \ldots x_n)$.

To determine a_i we average over time the output of Φ_i and set a_i equal to this average. (In a quasi-stationary case such an operator could adjust itself automatically to changing statistics.)

a. Effect of Truncation

Because the base functions used here are not orthogonal in the statistical sense, a realization of the density operator in terms of a finite number, say m, base functions does not yield an optimum choice of coefficients.

However, bounds may be calculated for the truncation error. If, for example, the probability distribution has a derivative that is less than M, then the density is bounded:

$$p(x_o, \ldots x_n) < M$$

whereupon

$$\overline{(\text{Error})^2} = \overline{\left[p(x_o, \ldots x_n) - \sum_{i=1}^{m} a_i \Phi_i(x_o, \ldots x_n)\right]^2}$$

$$= \int_0^1 \ldots \int_0^1 p(x_o, \ldots x_n) \left[p(x_o, \ldots x_n) - \sum_{i=1}^{m} a_i \Phi_i(x_o, \ldots x_n)\right]^2 dx_o \ldots dx_n$$

$$< M \int_0^1 \ldots \int_0^1 \left[p(x_o, \ldots x_n) - \sum_{i=1}^{m} a_i \Phi_i(x_o, \ldots x_n)\right]^2 dx_o \ldots dx_n$$

= M × [mean error squared in the approximation of the volume of $p(x_o, \ldots x_n)$ by m Fourier terms]

Our choice of coefficients is therefore optimum for approximating the given upper bound.

Alternatively, if

$$\int_0^1 \ldots \int_0^1 p(x_o, \ldots x_n)^2 \, dx_o \ldots dx_n = N$$

then

$$\overline{(\text{Error})^2} = \int_0^1 \ldots \int_0^1 p(x_o, \ldots x_n) \left[p(x_o, \ldots x_n) - \sum_{i=1}^{m} a_i \Phi_i(x_o, \ldots x_n)\right]^2 dx_o \ldots dx_n$$

$$\leq N^{1/2} \left\{ \int_0^1 \ldots \int_0^1 \left[p(x_o, \ldots x_n) - \sum_{i=1}^{m} a_i \Phi_i(x_o, \ldots x_n)\right]^4 dx_o \ldots dx_n \right\}^{1/2}$$

If $p(x_o, \ldots x_n)$ is smooth enough, it can be represented well by using a small number of terms. It may be useful here to consider "bandlimited" probability densities that can be represented without error by a finite number of terms.

2. Least-Mean-Square Prediction

We predict the future, x_o, on the basis of n samples in the past, $x_1, \ldots x_n$. For least-mean-square prediction the expected squared error is minimized. Hence

$$\overline{(\text{Error})^2} = \overline{(x_o - \hat{x}_o)^2}$$

$$= \int_0^1 (x_o - \hat{x}_o)^2 \, p(x_o, x_1, \ldots x_n) \, dx_o$$

where $\hat{x}_o$ is the predicted value of x_o based on $x_1, \ldots x_n$. Differentiating this expression with respect to $\hat{x}_o$, and setting the derivative equal to zero, we obtain

$$-2 \int_0^1 (x_o - \hat{x}_o) \, p(x_o, x_1, \ldots x_n) \, dx_o = 0$$

whence

$$\hat{x}_o = \frac{\int_0^1 x_o \, p(x_o, x_1, \ldots x_n) \, dx_o}{p(x_1, \ldots x_n)}$$

The denominator is a probability density and is realized, as we have shown, if a sample record of the process is available. The numerator is constructed by first synthesizing $p(x_o, x_1, \ldots x_n)$, as shown above, then replacing the $\Phi_i(x_o, \ldots x_n)$ terms by $\int_0^1 x_o \, \Phi_i(x_o, \ldots x_n) \, dx_o$, and leaving the coefficients unchanged. The adjustment makes use of the "future" which is available on the sample record. For adjustment, the entire sequence is shifted back in time so that all terms are realizable. In operation, the predictor does not depend on the future, x_o, since all terms containing it have been replaced by constants. See Figs. XIII-13 and XIII-14.

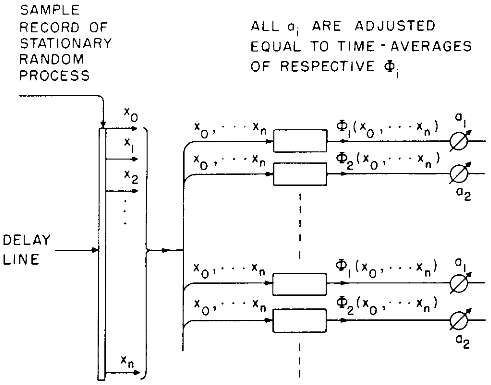
SAMPLE RECORD OF STATIONARY RANDOM PROCESS
ALL a_i ARE ADJUSTED EQUAL TO TIME-AVERAGES OF RESPECTIVE Φ_i
x_0
x_1
x_2
x_n
DELAY LINE
$x_0, \cdots x_n$
$\Phi_1(x_0, \cdots x_n)$
$\Phi_2(x_0, \cdots x_n)$
a_1
a_2

Fig. XIII-13. Canonical form for predictor: adjustment.

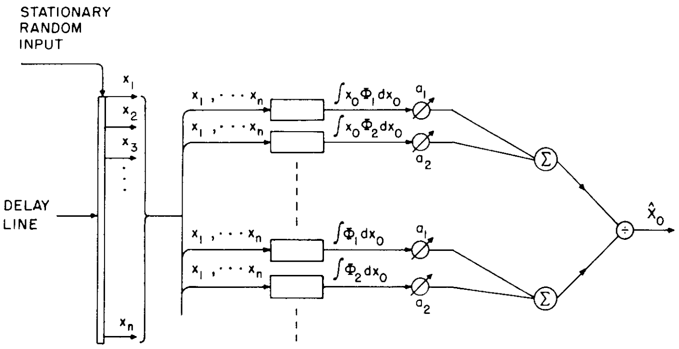

Fig. XIII-14. Canonical form for predictor: operation.

3. Separation of Signals from Noise

Consider a signal sequence, $x_n \dots x_1$, which has been contaminated by noise (such as additive or multiplicative) into a received sequence $y_n \dots y_1$. The least-mean-square estimate, $\hat{x}_1$, of the present signal, x_1, based on the last n received samples, $y_n \dots y_1$, is, similarly to the prediction case

$$\hat{x}_1 = \frac{\int_0^1 x_1\, p(x_1, y_1, \dots y_n)\, dx_1}{p(y_1, \dots y_n)}$$

This operator is realized in much the same way as the predictor. A sample record of the pasts of the received signal and the corresponding presents of the transmitted signal is fed through a bank of $\Phi_i(x_1, y_1 \dots y_n)$ operators, and coefficients are adjusted to be equal to time averages, in order to form the joint density $p(x_1, y_1 \dots y_n)$. The factors containing x_1 are then replaced by their moments, which yields the numerator. The denominator is a density operator.

4. Coding

We have a sequence of symbols

$$--- x_3, x_2, x_1$$

The symbol x_1 is occurring now, and we wish to code it into y_1 in such a way that y_1 is statistically independent of the past up to y_n. We accomplish this by making $p(y_1 | y_2, \dots y_n)$ constant, that is,

$$\begin{aligned} p(y_1 | y_2, \dots y_n) &= 1 \qquad && 0 \leqslant y_1 \leqslant 1 \\ &= 0\ . && \text{elsewhere} \end{aligned}$$

Suppose we derive y_1 from x_1 by a nonlinear no-memory operation $F_{x_2, \dots x_n}$ whose form depends on the past, $x_2, \dots x_n$. Suppose that $F_{x_2, \dots x_n}$ has an inverse and is differentiable. Now

$$y_1 = F_{x_2 \dots x_n}(x_1)$$

Therefore $p(y_1 | y_2, \dots y_n)$ is derived from $p(x_1 | x_2, \dots x_n)$ by the Jacobian relation:

$$p(y_1 | y_2 \dots y_n) = \frac{p(x_1 | x_2, \dots x_n)}{\frac{d}{dx_1} F_{x_2, \dots x_n}(x_1)} \begin{cases} = 1 & 0 \leqslant y_1 \leqslant 1 \\ = 0 & \text{elsewhere} \end{cases}$$

whence

$$\frac{dF_{x_2, \dots x_n}}{dx_1} = p(x_1 | x_2, \dots x_n) \qquad 0 \leqslant y_1 \leqslant 1$$

Since $x_1 = 0$ when $y_1 = 0$, upon integrating,

$$F_{x_2, \dots x_n}(x_1) = \int_0^{x_1} p(x_1' | x_2, \dots x_n)\, dx_1'$$

and, since this equals y_1,

$$y_1 = \frac{\int_0^{x_1} p(x_1', x_2, \dots x_n)\, dx_1'}{\int_0^{1} p(x_1', x_2, \dots x_n)\, dx_1'}$$

We realize this in the usual manner, as shown in Fig. XIII-15.

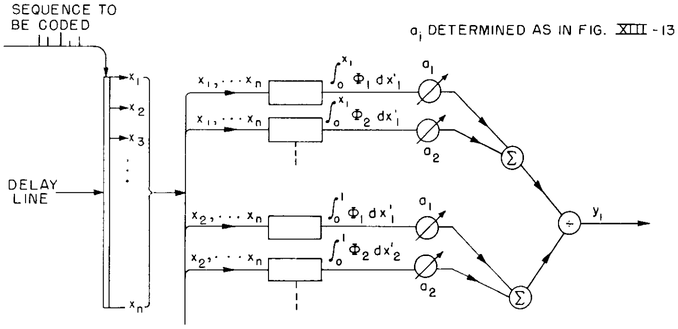

Fig. XIII-15. Canonical form for coding.

5. Computation of Averages

In order to calculate the ensemble average of the output of an operator, $F(x_1, \dots x_n)$, when the input is a stationary random process, we express both the operator and the density in orthonormal expansion:

$$p(x_1, \dots x_n) = \sum_i a_i \, \Phi_i(x_1, \dots x_n)$$

$$F(x_1, \dots x_n) = \sum_i b_i \, \Phi_i(x_1, \dots x_n)$$

Then

$$\overline{F(x_1, \dots x_n)} = \int_0^1 \dots \int_0^1 F(x_1, \dots x_n)\, p(x_1, \dots x_n)\, dx_1 \dots dx_n$$

$$= \sum_i a_i b_i$$

that is, the scalar product of the two (since the Φ_i are orthonormal).

6. Application to Decision Theory

It is required that we decide on the basis of a received signal which of two sources is transmitting. Moreover, the decision must be made as rapidly as is consistent with a probability less than ϵ of making a wrong decision.

Assume that the sources generate stationary random discrete processes, as follows:

Source	A Priori probability	Characteristic probability density for a sequence of n samples
s_1	k	$p_1(x_1, \dots x_n)$
s_2	1 - k	$p_2(x_1, \dots x_n)$

From Bayes' Rule

$$P(s_1|x_1, \dots x_n) = \frac{kp_1(x_1, \dots x_n)}{kp_1(x_1, \dots x_n) + (1-k)\, p_2(x_1, \dots x_n)}$$

and

$$P(s_2|x_1, \dots x_n) = 1 - P(s_1|x_1, \dots x_n)$$

where $P(s_1|x_1, \dots x_n)$, $P(s_2|x_1, \dots x_n)$ are the a posteriori probabilities of s_1 and s_2. If we decide in favor of the larger probability, then the smaller is the probability of error, P(error).

Suppose that the characteristic source densities are different enough and n is large enough so that for a decision based on n samples, almost always,

$$P(\text{error}) < \epsilon$$

Very often, however, an adequate decision can be made after receiving only m samples with $m \ll n$. We should like, therefore, to keep track of, say, $P(s_1|x_1, \dots x_m)$ as m increases from 1 to n, and decide only when this quantity comes within ϵ of either 0 or 1.

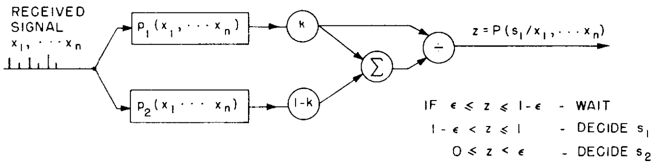
RECEIVED
SIGNAL
$x_1, \cdots x_n$
$p_1(x_1, \cdots x_n)$
$p_2(x_1 \cdots x_n)$
k
$1-k$
Σ
$\div$
$z = P(s_1/x_1, \cdots x_n)$
IF $\epsilon \leq z \leq 1-\epsilon$ - WAIT
$1-\epsilon < z \leq 1$ - DECIDE s_1
$0 \leq z < \epsilon$ - DECIDE s_2

Fig. XIII-16. Decision operator.

To build such a decision operator, we synthesize a density operator for each source experimentally, as described above, and use them as shown in Fig. XIII-16.

7. An Alternative Method for Estimators

In the method that has been outlined for putting estimators into canonical form the same set of coefficients is used in every case, while the canonical form is varied. The coefficients characterize the process.

Alternatively, the same canonical form might be used for every process, and the coefficients varied. (This is the more usual approach.)

Thus for the optimum predictor,

$$\hat{x}_o = \frac{\int_0^1 x_o\, p(x_o, \ldots x_n)\, dx_o}{p(x_1, \ldots x_n)}$$

Now if $\hat{x}_o$ is well enough behaved, then we can expand it as follows:

$$\hat{x}_o = \sum_{i=1}^{\infty} b_i\, \Phi_i(x_1, \ldots x_n)$$

with

$$b_i = \int_0^1 \cdots \int_0^1 \hat{x}_o\, \Phi_i(x_1, \ldots x_n)\, dx_1 \ldots dx_n$$

$$= \int_0^1 \cdots \int_0^1 x_o \frac{\Phi_i(x_1, \ldots x_n)}{p(x_1, \ldots x_n)}\, p(x_o, x_1 \ldots x_n)\, dx_o\, dx_1 \ldots dx_n$$

$$= \overline{\frac{x_o\, \Phi_i(x_1, \ldots x_n)}{p(x_1, \ldots x_n)}}$$

For an ergodic process, this is a weighted time average of Φ_i that can be obtained experimentally if a $p(x_1, \ldots x)$ operator is available. The method is outlined

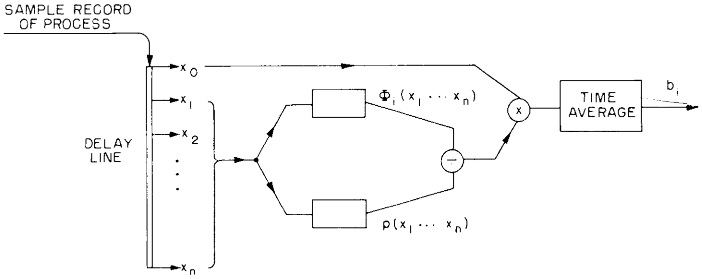

Fig. XIII-17. Determination of coefficient b_i for predictor.

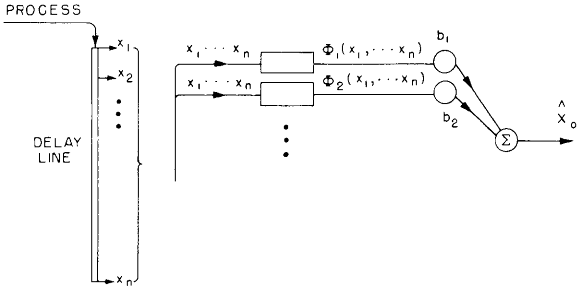

Fig. XIII-18. Canonical form for an optimum predictor.

in Figs. XIII-17 and XIII-18.

8. Estimators for Continuous Signals (3)

In the discrete case the optimum predictor is a function of the past which, for any past, gives the moment of the conditional probability of the future. It is tempting to look for a similar "functional" in the continuous case.

Our method for putting the discrete predictor into canonical form depends essentially on having

(a) A complete set of orthonormal functions in a coordinate system, $(x_o, x_1, \ldots x_n)$, and

(b) A joint probability density $p(x_o, x_1, \ldots x_n)$ of the past, $(x_1, \ldots x_n)$, and future, x_o, in this coordinate system.

In the continuous case we shall use the coordinate system for classifying pasts that N. Wiener (1) has developed. We shall define a probability density of pasts with respect to this coordinate system, and put statistical estimators – in particular, the optimum predictor – into a canonical form by using a complete set of orthonormal functionals (Wiener's "Hermite-Laguerre functionals") in this system. Although functionals are used in place of functions, and the coordinate system is different from that used in the discrete case, the procedure is entirely parallel to that in the discrete case, and the same diagrams are applicable.

9. Probability Density for a Continuous Process

The definition, which will be given in this section, of the probability density of a function of time which belongs to a continuous process X amounts, roughly, to this: Given any such function x(t), we enclose it in a sequence of decreasing "strips" each of which is delineated by a series of marker points. See Fig. XIII-19.

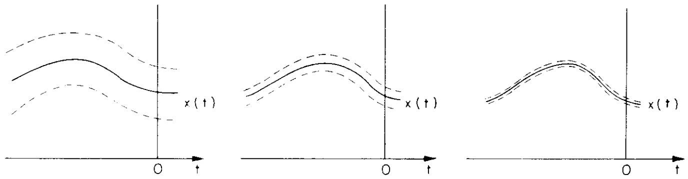

Fig. XIII-19. A function of time with decreasing "strips" which surround it.

The "weight" of any "strip" is the probability that a Brownian motion will lie inside its marker points. The "probability density" of x(t) is the ratio of the probability that a member of X will lie inside the markers of a "strip" to the "weight" of that "strip" in the limit as the "strip" width approaches zero.

Using the notation of Wiener (1), we map each past into a point α on the unit interval $[0, 1]$. Aside from excluded sets of pasts and points in α (all of measure zero), this mapping is 1-1 and onto. The length of any subset of points α is the probability with which the corresponding past occurs in Brownian motion.

A similar mapping, β, is made, by using the probability of occurrence in process X. Then, if α is an excluded point, define the probability density of α as $p(\alpha) = 0$. If α is not such a point define it as

$$p(\alpha) = \lim_{n \to \infty} \frac{\Delta\beta_n}{\Delta\alpha_n}$$

in which $\Delta\alpha_n$ is an open interval in the α line, formed in the n^{th} stage of the subdivision of pasts (as described by Wiener) to which α belongs, and $\Delta\beta_n$ is the corresponding interval in β.

Only those processes are considered for which $p(\alpha)$ exists for all α.

10. Averages of Functionals

If $F(\alpha)$ is a functional of pasts, α, belonging to a Brownian motion, then its ensemble average is

$$\overline{F(\alpha)} = \int_0^1 F(\alpha)\, d\alpha$$

Similarly, for process X,

$$\overline{F(\beta)} = \int_0^1 F(\beta)\, d\beta$$

Formally, therefore,

$$\overline{F(\beta)} = \int_0^1 F(\alpha)\, \frac{d\beta}{d\alpha}\, d\alpha$$

Unfortunately $d\beta/d\alpha$ does not, in general, exist. However, if $p(\alpha)$ exists, then

$$\overline{F(\beta)} = \int_0^1 F(\alpha)\, p(\alpha)\, d\alpha$$

11. The Probability-Density Operator in Canonical Form

Wiener's Hermite-Laguerre functionals, denoted by $\Phi_i(\alpha)$, are a complete set that is orthonormal over α. Hence, provided that $p(\alpha)$ is L^2, we have

$$p(\alpha) \sim \sum_{i=1}^{\infty} a_i\, \Phi_i(\alpha)$$

where ~ denotes that the representation is valid in the "limit-in-the-mean" sense. The Fourier coefficient a_i is given by

$$a_i = \int_0^1 p(\alpha)\, \Phi_i(\alpha)\, d\alpha$$

This is an ensemble average, and, for an ergodic process, it equals the time average of $\Phi_i(\alpha)$.

We can therefore construct $p(\alpha)$ exactly as $p(x_1, \dots x_n)$ was constructed in the discrete case, substituting $\Phi_i(\alpha)$ for $\Phi_i(x_1, \dots x_n)$.

12. The Joint Probability Density of Past and Future

For prediction, a joint probability aensity of past and future is required. This is defined (see Fig. XIII-20) as

$$p(\alpha, z) = 0 \qquad \text{if } \alpha \text{ is an excluded point}$$

$$= \lim_{\substack{n \to \infty \\ \Delta z \to 0}} \frac{P(\Delta\beta_n, \Delta z)}{\Delta\alpha_n, \Delta z} \qquad \text{elsewhere}$$

where $\Delta\alpha_n$, $\Delta\beta_n$ are intervals of order n [as in the definition for $p(\alpha)$] for Brownian motion and X, respectively; Δz is an interval to which z belongs; and $P(\Delta\beta_n, \Delta z)$ is the probability that process X has a member whose past, β, is in $\Delta\beta_n$, and whose future, z, is in Δz.

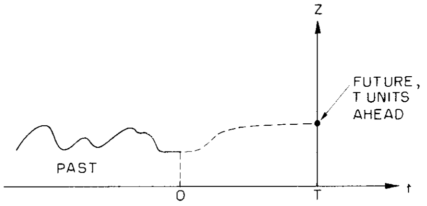

Fig. XIII-20. Past and future of a member of X.

Only those processes are considered for which $p(\alpha, z)$ exists and is unique for all α.

13. Mixed Functionals

The function $p(\alpha, z)$ depends on a function (the past) and on a single coordinate (the future). Such a function is referred to as a "mixed functional."

14. Canonical Forms for Mixed Functionals

A "mixed functional" $F(\alpha, z)$ can be expressed in a polynomial-integral form as

$$F(\alpha, z) \sim \sum_{m, n=1}^{\infty} x^m \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} K_{mn}(\tau_1, \dots \tau_n)\, dx(\alpha, \tau_1) \dots dx(\alpha, \tau_n)$$

where $x(\alpha, t)$ is the function of time, t, which corresponds to α. An orthonormal expansion is

$$F(\alpha, z) \sim \sum_{i, j=1}^{\infty} a_{ij}\, \Phi_i(\alpha)\, \Psi_j(z)$$

where $\Phi_i(\alpha)$ are the Hermite-Laguerre functionals; $\Psi_j(z)$ is any complete set of orthonormal functions of z; and

$$a_{ij} = \int_{-\infty}^{\infty} \int_{0}^{1} F(\alpha, z)\, \Phi_i(\alpha)\, \Psi_j(z)\, d\alpha\, dz$$

In particular, when $F(\alpha, z) = p(\alpha, z)$ in these expressions, a canonical form is obtained for the joint probability density of past and future; a_{ij} is the time average of $\Phi_i(\alpha)\, \Psi_j(z)$; and the realization of $p(\alpha, z)$ is entirely analogous to that for $p(x_o, \dots x_n)$ in the discrete case.

15. Canonical Form for the Predictor

The expected, squared, prediction error is

$$\overline{(\text{error})^2} = \int_z [z - \hat{z}(\alpha)]^2\, p(\alpha, z)\, dz$$

in which $\hat{z}(\alpha)$ is the predicted value of the future, z, based on the past, α.

This error is minimum, just as in the discrete case, when

$$\hat{z}(\alpha) = \frac{\int_z zp(\alpha, z)\, dz}{p(\alpha)}$$

The canonical form for $\hat{z}(\alpha)$ is derived from the forms for $p(\alpha, z)$ and $p(\alpha)$ in a manner that, again, parallels the derivation of $\hat{x}_o$ from $p(x_o, x_1, \dots x_n)$ and $p(x_1, \dots x_n)$ in the discrete case.

16. Conclusion

The method of synthesizing statistical estimators which is presented here is useful in that it allows any set of orthonormal functions (in the discrete case, or functionals in the continuous case) to be used as the basis for expansion. The same set of coefficients which characterizes the process appears in the canonical forms.

The definition of probability density for an arbitrary process emphasizes the importance of Wiener's representation as being not merely a method of analyzing nonlinear systems subjected to Gaussian inputs, but a "coordinate system" for nonlinear problems.

G. D. Zames

References

1. N. Wiener, Nonlinear Problems in Random Theory (The Technology Press, Cambridge, Mass. and John Wiley and Sons, Inc., New York, 1958).
2. A. G. Bose, A theory of nonlinear systems, Technical Report 309, Research Laboratory of Electronics, M.I.T., May 15, 1956.
3. The statements of this section should not be considered final.