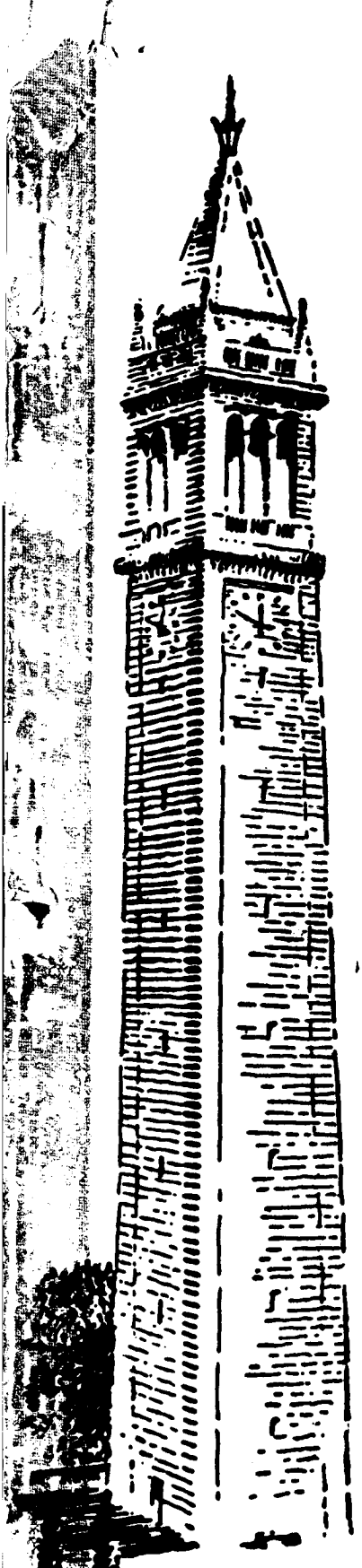


NOTES ON SYSTEM THEORY

VOLUME VII
May 1965

Report No. 65-14
AF-AFOSR-139-64
AF-AFOSR-292-64
AF-AFOSR-639-64
NSF GP-2684
NASA NsG-354 (S-1) / (S-2)

ELECTRONICS RESEARCH LABORATORY

UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA

Electronics Research Laboratory
University of California
Berkeley, California
Report No. 65-14

NOTES ON SYSTEM THEORY

VOLUME VII

The research herein reported is made possible through support received from the Air Force Office of Scientific Research under Grants AF-AFOSR-292-64, AF-AFOSR-639-64; by the Joint Services Electronics Programs (U.S. Army, U.S. Navy, and U.S. Air Force) under Grant No. AF-AFOSR-139-64; by National Science Foundation under Grants NSF GP-2684; and by National Aeronautics and Space Administration under Grant NsG-354(S-1)(S-2).

May 1965

TABLE OF CONTENTS

TABLE OF CONTENTS(Cont'd) Page

THE DISTRIBUTION OF MATRICES RESULTING FROM NEWTON'S IDENTITIES IN A FIELD OF CHARACTERISTIC TWO*

E. R. Berlekamp

Introduction: In order to decode a binary group code, one considers the set of all error patterns which have the same syndrome (pattern of parity check failures) as the received syndrome. From that set one chooses an error pattern which has the fewest errors. A solution to this problem is a set of error-location numbers $\{\beta_i\}$. Bose-Chaudhuri (1961a, b) have suggested that, if the block length, N, is one less than a power of 2, then the positions of the code (and hence the error-location numbers) may be taken as the nonzero elements of a finite field, $GF(N + 1) = GF(2^k)$. Bose-Chaudhuri have demonstrated methods of choosing a parity check matrix for the code so that the syndrome of the received sequence may be interpreted as the first e odd power-sum symmetric functions, $S_j = \sum_i \beta_i^j$; $j = 1, 3, \ldots, 2e - 1$. The decoding problem for Bose-Chaudhuri codes consists of finding the $\{\beta_i\}$, given the S_j, $j = 1, 3, \ldots, 2e - 1$. One form of a solution is the polynomial $\sigma(x) = \prod_i (x - \beta_i) = \sum_{i=0}^{m} \sigma_{m-i} x^i$. Chien (1964) has recently demonstrated an elegant, easily instrumented method for finding all the roots in $GF(2^k)$ of any polynomial $\sigma(x)$, but this procedure still requires one to evaluate several determinants of matrices relating the σ's to the S's. Massey (1965) has suggested a method to reduce this work to the evaluation of a single determinant of the type we consider here.

The relations between the σ's and the S's are given by Newton's Identities. (Peterson, p. 176):

* The research herein was supported by the Air Force Office of Scientific Research under Grant AF-AFOSR-639-64.

$$S_1 - \sigma_1 = 0 \qquad (1)$$

$$S_2 - S_1\sigma_1 + 2\sigma_2 = 0 \qquad (2)$$

$$S_3 - S_2\sigma_1 + S_1\sigma_2 - 3\sigma_3 = 0 \qquad (3)$$

$$S_4 - S_3\sigma_1 + S_2\sigma_2 - S_1\sigma_3 + 4\sigma_4 = 0 \qquad (4)$$

$$\vdots \qquad \vdots$$

In a field of characteristic two, we have additional equations, $S_{2i} = S_i^2$ for all i. Every even numbered Newton's identity is equivalent to one of these equations. When the dependent equations are omitted and the remaining equations are converted into matrix form, we have

$$\begin{bmatrix} S_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ S_3 & S_2 & S_1 & 1 & 0 & 0 & 0 & 0 & 0 \\ S_5 & S_4 & S_3 & S_2 & S_1 & 1 & 0 & 0 & 0 \\ S_7 & S_6 & S_5 & S_4 & S_3 & S_2 & S_1 & 1 & 0 \\ & & & & \vdots & & & & \\ S_{2e-1} & S_{2e-2} & \cdots & & & & & & \end{bmatrix} \begin{bmatrix} 1 \\ \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \vdots \\ \sigma_{2e-1} \end{bmatrix} = 0 \qquad (7)$$

It is convenient to consider $S_0 = \sigma_0 = 1$; $S_{-i} = \sigma_{-i} = 0$ for all $i > 0$. Then the above equations become

$$\sum_j S_{2i-j-1}\sigma_j = 0 \qquad \text{for } i = 1, 2, \ldots, e. \tag{8}$$

The decoding procedure of Peterson (1962) first finds a polynomial $\sigma(x)$ of lowest degree which satisfies equations (8), if there is any such polynomial of degree m not more than e. If this polynomial has m distinct roots in $GF(2^k)$, then these roots are the error positions. Notice that if this polynomial has m roots in $GF(2^k)$, then they are distinct. For if $\sigma(x) = (x - \beta_0)^2\sigma_0(x)$, then $\sigma_0(x)$ is a polynomial of degree (m - 2), and the coefficients of $\sigma_0(x)$ also satisfy (8). In general, however, the solution of (8) with the least degree may have nonlinear factors which are irreducible over $GF(2^k)$.

Bose-Chaudhuri and Peterson have shown that Peterson's decoding procedure will correct any error pattern containing no more than e errors, and no others. There are $\sum_{i=0}^{e}\binom{N}{i}$ such correctable patterns. But the total number of correctable error patterns is equal to the total number of syndromes. Except for the trivial (Hamming) case when $e = 1$, the number of correctable error patterns is much larger than $\sum_{i=0}^{e}\binom{N}{i}$, typically $(N + 1)^e$. The additional correctable error patterns have more than e errors, and each of them is less likely than any of the $\sum_{i=0}^{e}\binom{N}{i}$ patterns which Peterson's algorithm can correct. But there are so many correctable error patterns with more than e errors that the overall error probability of the code would be substantially reduced by any improvements which enabled the decoder to tackle these cases.

It is for this reason that we begin here an investigation of the properties of equations (8) over the entire ensemble of possible matrices corresponding to the $(N + 1)^e$ possible values of $S_1, S_3, \ldots, S_{2e-1}$. We introduce an invertible transformation of equations (8) which reduces the

enumeration problem to one which has previously been solved by Daykin (1960) and Berlekamp (1963) independently. Since this transformation reduces the size of the matrix to be inverted by a factor of two, it might well prove useful in the construction of decoding circuits for this problem. We do not examine this possibility here.

Triangular transformations: We now introduce an arbitrary triangular transformation of the equations (8). Let

$$T_k = \sum_j A_j S_{k-2j}, \quad \text{where } A_0 = 1;\ A_{-i} = 0 \quad \text{for } i > 0. \tag{9}$$

Then if we add to each equation (8) A_1 times the previous equations, A_2 times the equation before that, ..., we transform (7) into

$$\begin{bmatrix} T_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ T_3 & T_2 & T_1 & 1 & 0 & 0 & 0 & 0 & \\ T_5 & T_4 & T_3 & T_2 & T_1 & 1 & 0 & 0 & \\ T_7 & T_6 & T_5 & T_4 & T_3 & T_2 & T_1 & 1 & \\ \vdots & & & & & & & & \\ T_{2e-1} & T_{2e-2} & \cdots & & & & & & \end{bmatrix} \begin{bmatrix} 1 \\ \sigma_1 \\ \sigma_2 \\ \vdots \\ \\ \\ \\ \\ \sigma_{2e-1} \end{bmatrix} = 0 \tag{10}$$

or equivalently

$$\sum_j T_{2i-j-1}\sigma_j = 0 \quad \text{for } i = 1, 2, \ldots, e. \tag{11}$$

Notice that this transformation is readily invertible:

$$S_k = T_k + \sum_{j=1}^{k} A_j S_{k-2j}. \tag{12}$$

We now choose the A's in such a way that $T_{2i} = 0$, for all i. This is readily accomplished by setting

$$0 = \sum_{j=0}^{i} A_j S_{2(i-j)} \quad \text{or} \quad A_i = \sum_{j=0}^{i-1} A_i S_{2(i-j)}. \tag{13}$$

Given the initial condition that $A_0 = 1$, equation (13) defines the rest of the A_i consecutively.

Notice that $\{T_k\}$, like $\{S_k\}$ and $\{\sigma_k\}$, are homogeneous symmetric functions.

Having chosen $T_{2i} = 0$, equations (10) can now be solved:

$$\sigma_{2i-1} = \sum_{k} T_{2i-2k-1} \sigma_{2k}$$

or

$$\begin{bmatrix} T_1 & 0 & 0 & 0 & 0 \dots \\ T_3 & T_1 & 0 & 0 & 0 \\ T_5 & T_3 & T_1 & 0 & 0 \\ T_7 & T_5 & T_3 & T_1 & 0 \\ T_9 & T_7 & T_5 & T_3 & T_1 \\ \vdots & \vdots & & & \ddots \\ T_{2e-1} & & & & \end{bmatrix} \begin{bmatrix} 1 \\ \sigma_2 \\ \sigma_4 \\ \sigma_6 \\ \sigma_8 \\ \vdots \end{bmatrix} = \begin{bmatrix} \sigma_1 \\ \sigma_3 \\ \sigma_5 \\ \sigma_7 \\ \sigma_9 \\ \vdots \end{bmatrix}. \tag{14}$$

If, following Peterson, we now wish to find a solution of (14) with degree at most e, we set $\sigma_{e+1} = \sigma_{e+2} = \dots = 0$. Equations (14) can then be partitioned into

$$\begin{bmatrix} & \cdots & T_5 & T_3 & T_1 \\ & & & T_5 & T_3 \\ & & & & T_5 \\ \vdots & & & & \vdots \\ T_{2e-1} & \cdots & & & \end{bmatrix} \begin{bmatrix} 1 \\ \sigma_2 \\ \vdots \\ \\ \sigma_{2\left[\frac{e}{2}\right]} \end{bmatrix} = \underline{0}; \tag{15}$$

$$\begin{bmatrix} T_1 & 0 & 0 & 0 \cdots \\ T_3 & T_1 & 0 & 0 \\ T_5 & T_3 & T_1 & 0 \\ \vdots & & & \ddots \end{bmatrix} \begin{bmatrix} \sigma_1 \\ \sigma_3 \\ \sigma_5 \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 \\ \sigma_2 \\ \sigma_4 \\ \vdots \end{bmatrix} . \tag{16}$$

The problem is thereby reduced to the solution of equation (15). This is equivalent to Peterson's equations in theorem 9.4, page 176. But our matrix of equation (15) is of dimension only $(2[e/2]) \times (2[e/2]+1)$, and, furthermore, it is parasymmetric, having a constant value along any minor diagonal. This enables us to make certain statements about the existence and number of solutions to (15), because the enumeration of persymmetric matrices by rank has been accomplished by Daykin (1960) and Berlekamp (1963). The major result is that the number of $m \times n$ persymmetric matrices of rank less than or equal to r over GF(q) is q^r, for any $r < m \leq n$. This enables us to conclude, for example, that equations (15) will have a unique solution for $(q-1)/q = N/(N+1)$ of the possible cases. This happens only when the square submatrix formed by omitting the first column of equations (15) is nonsingular.

In fields with characteristic $p > 2$, it is very difficult to get similar results, for several reasons. One may introduce triangular transformations which reduce the dimensions of the matrices resulting from Newton's identities by a factor of p, but the resulting smaller

matrices are not parasymmetric and the problem of enumerating them has not been solved. The relevance of these cases to coding is also less direct, because one now needs both error-location numbers and an error number for each location to specify an error pattern.

References

E. R. Berlekamp (1963), "The enumeration of matrices by rank," unpublished BTL memo.

R. C. Bose and D. K. Ray-Chaudhuri (1960a), "On a class of error-correcting binary group codes," Inf.and Control 3, pp. 68-79.

R. C. Bose and D. K. Ray-Chaudhuri (1960b), "Further results on error-correcting binary group codes," Inf. and Control 3, pp. 279-290.

R. T. Chien (1964), "Cyclic decoding procedures for Bose-Chaudhuri-Hocquenghem codes," PGIT-10, pp. 357-362.

G. E. Daykin (1960), "Distribution of bordered OK_{ERB} symmetric matrices in a finite field," Journal für die reine und angewandte Mathematik 203, 47-54.

J. L. Massey (1965), unpublished correspondence.

W. W. Peterson (1961), Error-Correcting Codes, MIT Press and John Wiley and Sons.

MULTIPARAMETER SENSITIVITY IN LINEAR NETWORKS

R. N. Biswas and E. S. Kuh

Introduction

In the analysis and synthesis of linear networks, a knowledge of the sensitivity of the network function under consideration with respect to the physical parameters is often extremely important. The first definition of sensitivity was given by Bode[1]; Mason[2] defined sensitivity by the inverse of Bode's definition, and that has been the accepted definition of sensitivity since then. The sensitivity of a scalar transfer function $w(s)$ with respect to a parameter x is defined to be

$$S_x^w = \frac{\partial \ell n\, w}{\partial \ell n\, x} = \frac{\partial w}{\partial x}\frac{x}{w}.$$

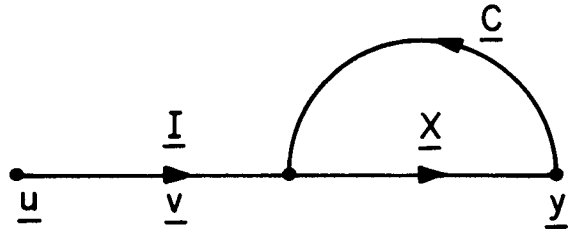
However, this measures only the first order approximation of the change in $w(s)$ due to variation in x, as is readily seen from the Taylor series expansion of $w(s)$ around the nominal value of x.

The purpose of this note is to give a compact formulation of the first order sensitivity functions, and use this formulation to obtain expressions for the "higher order sensitivity functions," which may be used to estimate the deviation of a transfer function from the nominal value with greater accuracy.

Flow Graph Representation

A convenient technique for investigating the sensitivities of a network function with respect to n varying parameters $x_1, x_2, \ldots, x_n$ is to represent the network by the following vector flow graph.

The research herein was supported by the National Science Foundation under Grant GP-2684.



$\underline{I}$ = Unit matrix of order n

$\underline{X}$ = Diag. $(x_1, x_2, \ldots, x_n)$

$\underline{C}$ = Transfer matrix from the outputs of $x_1, x_2, \ldots, x_n$ to their inputs

(c_{ij} = transfer function from the output of the j-th varying element to the input of the i-th varying element.)

The output vector $\underline{y} = \underline{X}\,\underline{v}$, where $\underline{v} = \underline{u} + \underline{C}\,\underline{y}$, $\underline{u}$ being the unit vector. The overall transfer matrix $\underline{W}$, defined by the relation $\underline{y} = \underline{W}\,\underline{u}$, is then given by

$$\underline{W} = (\underline{I} - \underline{X}\,\underline{C})^{-1}\underline{X}.$$

If u_j and y_i are respectively the scalar input and the scalar output of interest, the scalar transfer function is obviously given by w_{ij} (i-th row-, j-th column-element of the matrix $\underline{W}$).

The $\underline{C}$ matrix contains the fixed parameters of the network; if sensitivities with respect to all the parameters are sought, then $\underline{C}$ depends upon the network topology only.

Let us briefly illustrate how a network can be put in the desired flow graph form.

Example 1: Multiloop Feedback Amplifier

Suppose we are interested in finding the variation in the overall transfer function due to changes in the individual stage gains. The stage gains are therefore labelled as $x_1, x_2, \ldots, x_n$. c_{ij} is then the net feedback factor from the output of the j-th stage to the input of the i-th stage. The form of the $\underline{C}$ matrix depends upon the configuration of the amplifier.

Example 2: Passive RLC Circuit

In this case all the parameters are considered to be subject to variations. If we pick a tree such that the scalar input is either a current source in parallel with a tree-branch or a voltage source in series with a link, and the scalar output is either a tree-branch voltage or a link current, and let

$\underline{v}$ = Col. (tree-branch currents, link voltages)

$\underline{y}$ = Col. (tree-branch voltages, link currents)

so that

$\underline{X}$ = Diag. (tree-branch impedances, link admittances),

then it follows from Bashkow's results[3] that

$$\underline{C} = \begin{pmatrix} 0 & -D \\ D^t & 0 \end{pmatrix}$$

where the fundamental cut-set matrix = $(\underline{I} \vdots \underline{D})$. $\underline{C}$ is thus a description of the network topology only.

The Sensitivity Functions

The multiparameter Taylor series expansion of the transfer function w_{ij} around the nominal values of $x_1, x_2, \ldots, x_n$ gives

$$\delta w_{ij} = \sum_{k=1}^{n} \frac{\partial w_{ij}}{\partial x_k} \delta x_k + \frac{1}{2!} \sum_{k=1}^{n} \sum_{\ell=1}^{n} \frac{\partial^2 w_{ij}}{\partial x_k \partial x_\ell} \delta x_k \delta x_\ell + \ldots$$

or

$$\frac{\delta w_{ij}}{w_{ij}} = \sum_{k=1}^{n} \frac{\partial w_{ij}}{\partial x_k} \frac{x_k}{w_{ij}} \frac{\delta x_k}{x_k} + \frac{1}{2!} \sum_{k=1}^{n} \sum_{\ell=1}^{n} \frac{\partial^2 w_{ij}}{\partial x_k \partial x_\ell} \frac{x_k x_\ell}{w_{ij}} \frac{\delta x_k}{x_k} \frac{\delta x_\ell}{x_\ell} + \ldots$$

or

$$\frac{\partial w_{ij}}{\partial x_k} = \frac{w_{ik} w_{kj}}{x_k^2} \quad \Rightarrow \quad \boxed{S_k^{ij} = \frac{w_{ik} w_{kj}}{x_k w_{ij}}} .$$

The first order sensitivity function is thus inversely proportional to the transfer function w_{ij}, and directly proportional to w_{ik} and w_{kj}, which clearly have the following significances:

w_{ik} = extraneous signal gain from the input of the k-th varying element to the scalar output

w_{kj} = signal gain from the scalar input to the output of the k-th varying element.

Now, since $\underline{W} = (\underline{I} - \underline{X}\underline{C})^{-1}\underline{X} = (\underline{X}^{-1} - \underline{C})^{-1}$, it follows that

$$w_{ij} = \Delta_{ji} / \Delta ,$$

where $\Delta = \det(\underline{X}^{-1} - \underline{C})$ and Δ_{ji} = the cofactor of Δ corresponding to the j-th row and the i-th column. Then

$$S_k^{ij} = \frac{\Delta_{ki} \Delta_{jk}}{x_k \Delta \Delta_{ji}} .$$

This expression, which can be used for direct computation, clearly shows that both the poles and the zeros of the scalar transfer function are poles of the sensitivity functions. Thus in a synthesis problem, once the transfer function has been specified, the poles of the (first order) sensitivity functions are fixed; one has some control, however, over the zeros, and this fact may be utilized to minimize sensitivity while synthesizing a specified transfer function, by proper choice of the network configuration.

The higher order sensitivity functions are obtained directly from definition, by repeated use of the relation

$$\frac{\partial w_{ij}}{\partial x_k} = \frac{w_{ik} w_{kj}}{w_{ij}} \qquad \text{for all } i, j, \text{ and } k.$$

The results are the following.

$$S_{k\ell}^{ij} = \hat{S}_{k\ell}^{ij} + \hat{S}_{\ell k}^{ij}, \quad \text{where } \hat{S}_{k\ell}^{ij} = \frac{w_{ik} w_{k\ell} w_{\ell j}}{x_k x_\ell w_{ij}} \quad (k \neq \ell),$$

$$S_{kk}^{ij} = \frac{w_{ik} w_{kj}}{x_k w_{ij}} \left(\frac{w_{kk}}{x_k} - 1 \right), \text{ etc.}$$

The coefficient of the first order terms are the usual (first order) sensitivity functions as defined earlier. We shall use the notation:

$$S_k^{ij} \triangleq \frac{\partial w_{ij}}{\partial x_k} \frac{x_k}{w_{ij}} .$$

The logical choices for the higher order sensitivity functions would then be the following.

2nd order: $S_{k\ell}^{ij} \triangleq \frac{\partial^2 w_{ij}}{\partial x_k \partial x_\ell} \frac{x_k x_\ell}{w_{ij}} : k \neq \ell,$

$$S_{kk}^{ij} \triangleq \frac{1}{2!} \frac{\partial^2 w_{ij}}{\partial x_k^2} \frac{x_k^2}{w_{ij}}$$

3rd order: $S_{k\ell m}^{ij} \triangleq \frac{\partial^3 w_{ij}}{\partial x_k \partial x_\ell \partial x_m} \frac{x_k x_\ell x_m}{w_{ij}} : k \neq \ell \neq m \neq k$ and so on.

The factors $\frac{1}{2!}$, $\frac{1}{3!}$ etc. are left out when the subscripts are different because the same combinations of different subscripts appear 2! times in the 2nd order terms, 3! times in the 3rd order terms and so on; physically, all permutations of the subscripts for the same combination have to be associated with one and the same sensitivity function.

In terms of the sensitivity functions thus defined, then, the fractional change in the transfer function may be expressed as follows.

$$\frac{\delta w_{ij}}{w_{ij}} = \sum_k S_k^{ij} \frac{\delta x_k}{x_k} + \sum_{k,\ell}' S_{k\ell}^{ij} \frac{\delta x_k \delta x_\ell}{x_k x_\ell} + \sum_{k,\ell,m}' S_{k\ell m}^{ij} \frac{\delta x_k \delta x_\ell \delta x_m}{x_k x_\ell x_m} + \ldots$$

The primed summation signs indicate that the summations are to be extended over all combinations of the subscripts, no distinction being made between terms involving the same subscripts in different orders.

Expressions for the sensitivity functions are obtained through the use of the well-known matrix sensitivity relation:

$$\underline{\delta W} = \underline{S}\, \underline{\delta X}\, \underline{X}^{-1}\, \underline{W},$$

relating the change $\underline{\delta W}$ in the overall transfer matrix to the change $\underline{\delta X}$ in the forward transfer matrix. For small $\underline{\delta X}$, the (first order) sensitivity matrix is given by

$$\underline{S} = (\underline{I} - \underline{X}\,\underline{C})^{-1} = \underline{W}\,\underline{X}^{-1}$$

$$\therefore\ \underline{\delta W} = \underline{W}\,\underline{X}^{-1}\,\underline{\delta X}\,\underline{X}^{-1}\,\underline{W}$$

$$\Rightarrow \delta w_{ij} = \sum_k w_{ik} \frac{1}{x_k}\,\delta x_k \frac{1}{x_k}\, w_{kj}\,.$$

The expressions, though not very simple, clearly show a definite pattern; the higher order sensitivity functions are multilinear combinations of the first order ones, as shown in the following.

$$S_{k\ell}^{\ ij} = S_k^{\ ij} S_\ell^{\ kj} + S_\ell^{\ ij} S_k^{\ \ell j} \quad (k \neq \ell),$$

$$S_{kk}^{\ ij} = S_k^{\ ij} \left(S_k^{\ kk} - 1 \right),$$

$$S_{k\ell m}^{\ ij} = S_k^{\ ij} S_{\ell m}^{\ kj} + S_\ell^{\ ij} S_{km}^{\ \ell j} + S_m^{\ ij} S_{k\ell}^{\ mj} \quad (k \neq \ell \neq m \neq k)$$

and so on.

These expressions clearly show that low values of $S_k^{\ ij}$ for all k do not necessarily guarantee low $S_{k\ell}^{\ ij}$, $S_{k\ell m}^{\ ij}$ etc. for all k, ℓ, ..., except in the ideal case where $S_k^{\ ij} = 0$ for all k, in which case all the sensitivity functions are identically zero, so that the transfer function w_{ij} is completely independent of all the network parameters. In general, therefore, unless the fractional changes $\delta x_k / x_k$ are all small enough to justify a first order approximation, one has to take the higher order sensitivities into consideration. The advantage of the foregoing formulation is that this can be done without much complication, since the sensitivity functions are all expressed in terms of scalar transfer functions between different points in the network, which can be found directly from a given network topology and the nominal values of the parameters.

References

1. H. W. Bode, Network Analysis and Feedback Amplifier Design, D. Van Nostrand Co. Inc.

2. S. J. Mason, "Feedback theory - some properties of signal flow graphs," Proc. I.R.E., Vol. 41; Sept. 1953.

3. T. R. Bashkow, "The A-matrix, a new network description," I.R.E. Transactions on Circuit Theory; Sept. 1957.

ON THE DEFINITION AND ANALYSIS OF PULSE-FREQUENCY MODULATED SYSTEMS*

Jean Gabriel Blanchard

I. DEFINITION OF A GENERAL PULSE-FREQUENCY MODULATOR (GPFM)

1. Introduction

A GPFM is defined as a system operating on piecewise continuous inputs and converting them into sequences of asynchronous pulses of identical shape but different signs. The output of the modulator is completely characterized by the sequences $\{t_p\}$, $\{\epsilon_p\}$, and the function $P_0(t)$ (Fig. 1). The input-output relationship of the GPFM, $m(t) = M[x(t)]$, where M is a nonlinear operator, is in fact a relation between $x(t)$ and the pair $(\{t_p\}, \{\epsilon_p\})$.

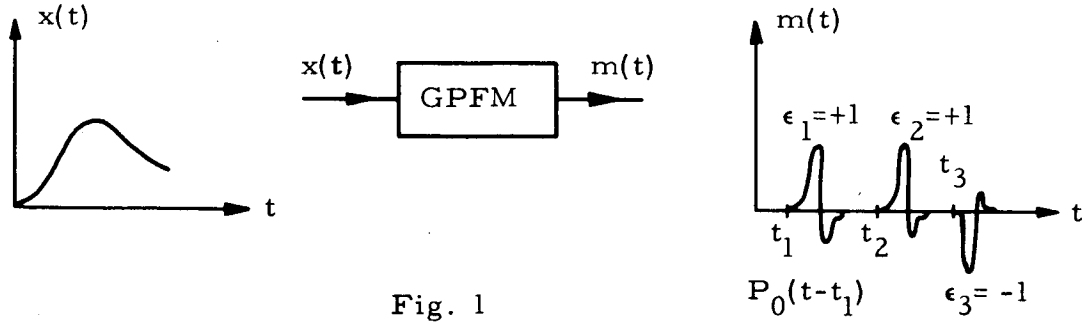


Fig. 1

* The research herein was supported by the University of California Fellowship Program.

2. Definitions

A. The input $x(t)$ is assumed to be piecewise continuous. We define $x(t, \theta'/\theta)$ as follows (Fig. 2):

$$x(t, \theta'/\theta) = \begin{cases} 0 & \text{for } t < \theta \\ x(t) & \text{for } \theta \leq t \leq \theta' \\ 0 & \text{for } t > \theta' \end{cases} \qquad \theta' > \theta$$

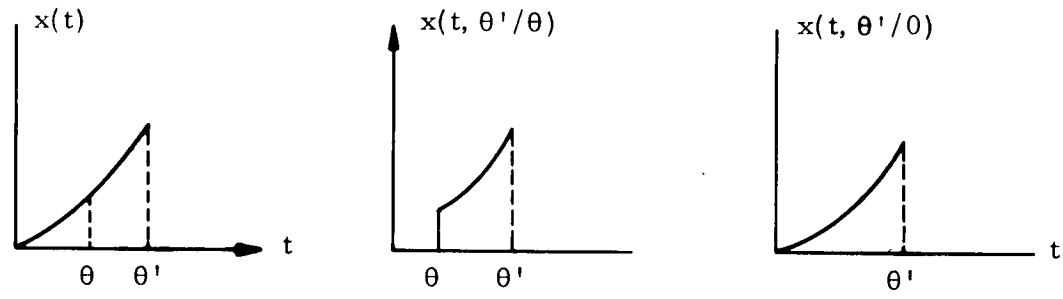


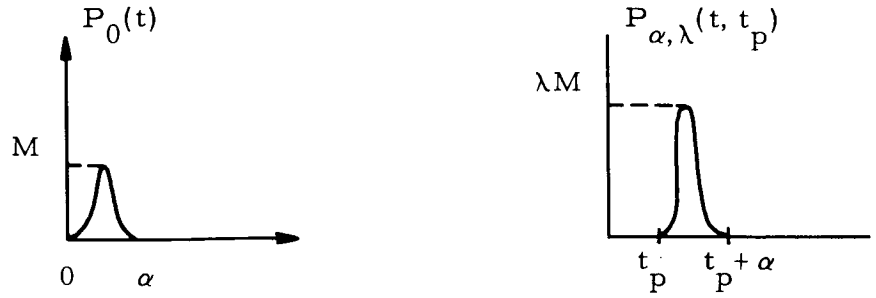
Fig. 2

B. We shall denote by $P_{\alpha, \lambda}(t, t_p)$ a function (or generalized function) where t_p is a variable and (α, λ) are parameters. Assuming t_p to be known, we have:

$$P_{\alpha, \lambda}(t, t_p) = \lambda P_{\alpha, 1}(t, t_p)$$

$$P_{\alpha, 1}(t, t_p) = \begin{cases} 0 & \text{for } t < t_p \\ P_0(t - t_p) & \text{for } t_p \leq t \leq t_p + \alpha \\ 0 & \text{for } t > \alpha + t_p \end{cases}$$

a. $P_0(t)$ may be a given function of t equal to zero for $t < 0$ and $t > \alpha$. It is continuous in the interval $(0, \alpha)$ (Fig. 3).



$$P_0(t) = e^{-\frac{1}{t^2}} \cdot e^{-\frac{1}{(t-\alpha)^2}} \qquad \text{for } 0 \le t \le \alpha$$

Fig. 3

b. $P_0(t)$ may also be a generalized function: $P_0(t) = \delta(t)$. This case can be considered as a limiting case of a. (see above). Take $P_0(t) = \lambda[u(t) - u(t - \alpha)]$ where α is small and λ large. Assume $\varphi(t)$ to belong to space D. We have then:

$$\int_{-\infty}^{+\infty} \varphi(\tau) P_0(\tau) d\tau = \lambda \int_0^{\alpha} \varphi(\tau) d\tau = \lambda\alpha\varphi(\xi) \qquad 0 \le \xi \le \alpha$$

and $P_0(t)$ is equivalent to $H\delta(t)$ where $\lambda\alpha = H$.

C. Consider now the class of all functions of the form $x(t, \theta'/\theta)$ where θ, θ' are arbitrary but finite and satisfy the following inequality: $0 < \theta < \theta'$. We define then a real functional F whose domain is the previous class and which takes values on the real line. We denote the value of this functional by $I_{x(t)}(\theta, \theta')$. The following assumptions are made on $I_{x(t)}(\theta, \theta')$:

$I_{x(t)}(\theta, \theta') = 0$ for $x(t) \equiv 0$ in (θ, θ')

$I_{x(t)}(\theta_0, \theta')$, for fixed θ_0, is a continuous function of θ'

$I_{x(t)}(\theta, \theta) = 0$ for all $x(t)$.

Example: $I_{x(t)}(\theta, \theta') = \int_{-\infty}^{+\infty} K(\tau)x(\tau, \theta', \theta)d\tau = \int_{\theta}^{\theta'} K(\tau)x(\tau)d\tau$.

D. $\left\{T_k\right\}_{K=0, 1, \ldots}$ is a given sequence of positive numbers.

3. Definition of M

Assume x(t) to be given. Consider an arbitrary t_0. We define the output $m_{t_0}(t)$ of the GPFM, corresponding to the input $x(t, t_0/0)$, as follows:

$$m_{t_0}(t) = \sum_{t_n} \epsilon(t_n)P_{\alpha, \lambda}(t, t_n) \qquad (1)$$

where $P_0(t)$, α, λ are given and the sequences $\{t_n\}$ and $\epsilon\{t_n\}$ are determined by the following relations:

A. $0 < t_1 < t_2 < \ldots < t_n < \ldots \leq t_0$.

B. Assuming t_{n-1} to be known, t_n is defined as the first value of t for which $|I_{x(t)}(t_{n-1}, t)| = T_{n-1}$. Furthermore, $\epsilon(t_p)$ is taken equal to +1 if $I_{x(t)}\{t_{n-1}, t_n\} > 0$ or -1 if $I_{x(t)}\{t_{n-1}, t_n\} < 0$.

C. Conditions A and B determine uniquely $m_{t_0}(t)$ for a given x(t).

D. The output m(t) corresponding to the total input x(t) is defined as limit of $m_{t_0}(t)$ when $t_0 \rightarrow \infty$.

4. Example

Consider the system represented in Fig. 4. Assume the GPFM to be defined as follows:

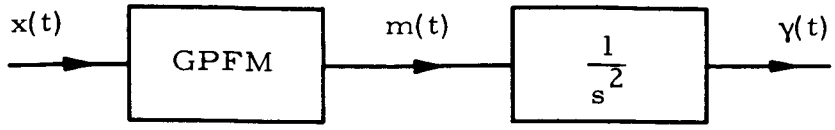


Fig. 4

$$I(\theta, \theta') = \int_{\theta}^{\theta'} x(\tau)d\tau$$

$$T_k = \frac{T}{(k+1)^2}$$

$$x(t) = \begin{cases} 0 & \text{for } t < 0 \\ e^{-\beta t} & \text{for } t \geq 0 \end{cases}$$

$$P_0(t) = \delta(t)$$

Determination of the sequences $\{\epsilon_p\}$ and $\{t_p\}$.

$$I(\theta', \theta) = \int_{\theta}^{\theta'} e^{-\beta\tau}d\tau = \frac{1}{\beta}\left\{e^{-\beta\theta} - e^{-\beta\theta'}\right\}$$

$\theta' > \theta \rightarrow I(\theta', \theta) > 0$ so $\epsilon(t_k) = \epsilon_k = +1$ for all k.

The t_k are determined recursively by the relation: $e^{-\beta t_{k-1}} - e^{-\beta t_k} = \frac{T\beta}{k^2}$.

We have therefore: $1 - e^{-\beta t_p} = \sum_{k=1}^{k=p} \frac{T\beta}{k^2}$ or $t_p = \frac{1}{\beta}\log_n\left\{1 - \beta\sum_{k=1}^{k=p} \frac{T}{k^2}\right\}$.

Two cases may be distinguished:

A. $T \leq \frac{1}{\beta}\frac{6}{\pi^2}$ or $1 - \sum_{k=1}^{k=\infty} \frac{\beta T}{k^2} \geq 0$: there is an infinite number of emitted pulses.

B. $T > \frac{6}{\beta\pi^2}$ or $1 - \sum_{k=1}^{k=+\infty} \frac{\beta T}{k^2} < 0$: there is a finite number of emitted pulses at the instants t_k, k = 1, N, where N is the greatest integer such that $\sum_{k=1}^{k=N} \frac{1}{k^2} \leq \frac{1}{T\beta}$.

II. GENERAL PULSE FREQUENCY MODULATED SYSTEMS (GPFMS)

A GPFMS is a system having one or more GPFM as components. It contains linear and eventually nonlinear plants. Feedback loops may also exist (Fig. 5).

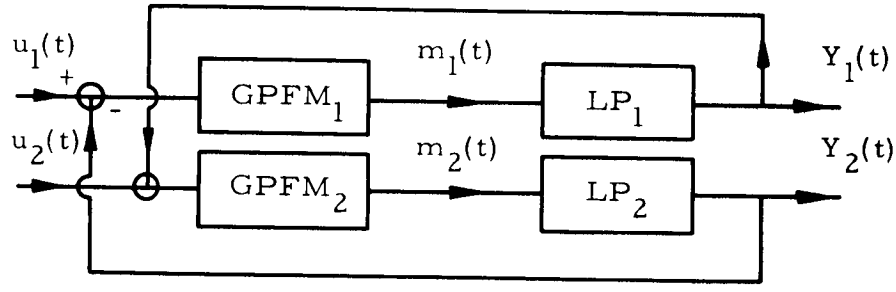


Fig. 5

Large systems as defined previously are difficult to analyze. Only simulations on digital computers or approximation methods seem to be of interest in their study. Some simpler cases (Fig. 6) will be studied in details.

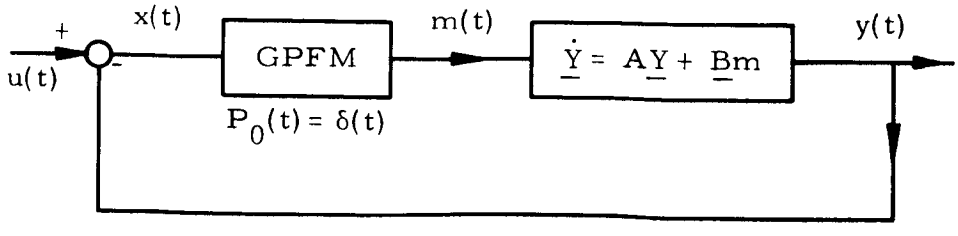


Fig. 6

III. CASE OF A GPFM FOLLOWED BY A LINEAR PLANT DESCRIBED BY A VECTOR DIFFERENTIAL EQUATION OF THE FORM $\underline{Y} = A\underline{Y} + \underline{B}m(t)$

1. Determination of the linear plant output $[P_0(t) = \delta(t)]$

Assume the linear plant to admit an input-output relationship of the form $\sum_{k=0}^{k=N} a_k y^{(k)} = m(t)$ $\qquad a_N = 1$

Take $A = \underbrace{\begin{bmatrix} 0 & 1 & 0 & \cdots & \cdots \\ 0 & 0 & 1 & \cdots & \cdots \\ & & & & \\ 0 & 0 & \cdots & \cdots & 1 \\ -a_0 & -a_1 & & & -a_{N-1} \end{bmatrix}}_{N}$ $\Big\}$ N-1, $\underline{B} = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix}$

$$\underline{Y}(t) = e^{A(t - t_0)}\,\underline{Y}(t_0) + e^{At}\int_{t_0}^{t} e^{-A\tau}\underline{B}\,m(\tau)d\tau$$

but

$$e^{At} = \sum_{k=0}^{k=m} \alpha_k(t)A^k \qquad 0 \le m \le N - 1$$

take

$$t_0 = t_p^- \quad \text{and} \quad t_p^+ \le t \le t_{p+1}^-$$

$$m(t) = \epsilon(t_p)\delta(t - t_p)$$

$$\underline{Y}(t) = e^{A(t - t_p)}\underline{Y}(t_p^-) + e^{At}\sum_{k=0}^{k=m} A^k\underline{B}\underbrace{\int_{t_p^-}^{t} \epsilon(t_p)\alpha_k(\tau)\delta(\tau - t_p)d\tau}_{\epsilon(t_p)\alpha_k(t_p)}.$$

So

$$\underline{Y}(t) = e^{A(t - t_p)}[\underline{Y}(t_p^-) + \epsilon(t_p)\underline{B}] \qquad (2)$$

Now set $t = t_{p+1}^-$, $\underline{Y}_p = \underline{Y}(t_p^-)$. We obtain from (2):

$$\underline{Y}_{p+1} = e^{A(t_{p+1} - t_p)}(\underline{Y}_p + \epsilon_p \underline{B}). \tag{3}$$

This discrete system is completely defined if we know the sequences $\{t_p\}$ and $\{\epsilon_p\}$. Conditions 3.A. enable us (at least in theory) to determine t_{p+1} and ϵ_p as a function of $x(t)$. Furthermore, Eq. (3) suffices to characterize the output of the linear plant (Eq. 2).

2. Determination of $\{t_p\}$ and $\{\epsilon_p\}$

The main difficulty lies in the solution of the equation $I(t_n, t) = \epsilon_{n+1} Tn$.

A. Example:

Take $x(t) = a + bt + ct^2$. Assume

$$I(t_n, t) = \int_{t_n}^{t} k(\tau)x(\tau)d\tau .$$

From 3.A. we deduce:

$$\int_{t_n}^{t} k(\tau)x(\tau)d\tau = \epsilon_{n+1} T_n$$

or if we set

$$\int_{t_n}^{t} k(\tau)d\tau = A(t, t_n), \quad \int_{t_n}^{t} tk(\tau)d\tau = B(t, t_n), \quad \int_{t_n}^{t} t^2 k(\tau)d\tau = C(t, t_n)$$

$$aA(t, t_n) + bB(t, t_n) + cC(t, t_n) = \epsilon_{n+1} T.$$

Finally we obtain a relation of the form $f(t, t_n) = \epsilon_{n+1} T_n$ where t_{n+1} is the smallest positive root of the previous equality. In general, it is impossible to find a closed form expression for t_{n+1} as a function of t_n (similarily for the computation of ϵ_n).

B. If there is a feedback loop, the problem of finding the t_n and ϵ_n becomes also unsolvable in general if we do not use approximation methods, graphical solutions, analog or digital simulations. For instance,

$$I(t_n, t) = \int_{t_n}^{t} f[e(\tau)]d\tau .$$

We then have:

$$\underline{Y}_t = \sum_{k=0}^{k=m} \alpha_k(t - t_n)A^k[\underline{Y}_n + \epsilon_n B] \qquad \text{for } t_n + \leq t \leq t_{n+1}^{-}.$$

Take

$$\begin{cases} \underline{Y}_n + \epsilon_n B = \underline{z}_n \\ \underline{a}_k = \text{first row of } A^k \end{cases}$$

$$y_t = \sum_{k=0}^{k=m} \alpha_k(t - t_n) < \underline{a}_k, \underline{z}_n >$$

owing to the form of A, $\underline{a}_k$ reduces to

$$\begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 1 \\ \cdot \\ \cdot \\ 0 \end{bmatrix} \longrightarrow (k+1)\text{-th row}$$

and $y_t = \sum_{k=0}^{k=m} \alpha_k(t - t_n)z_{k,n}$. The t_n and ϵ_n are obtained from the relation

$$\int_{t_n}^{t_{n+1}} f[u(\tau) - \sum_{k=1}^{k=m} \alpha_k(\tau - t_n)z_{k,n}]d\tau = \epsilon_{n+1} T_n. \qquad (4)$$

This equation in t_{n+1} is not in general solvable:

$$\text{in theory} \begin{cases} t_{n+1} = f(\underline{z}_n,\ t_n) \\ \epsilon_{n+1} = g(\underline{z}_n,\ t_n). \end{cases}$$

An interesting case arises when $I(t_n, t)$ is taken equal to $\int_{t_n}^{t} x(\tau)d\tau$ and the sequence T_p reduces to a constant threshold T. This particular case has been previously studied under the name of IPFM (see references). Few general results appear to have been found. From Eq. (4), it may be seen that for zero input $[u(t) = 0]$, the relation determining the t_p and ϵ_p is of the form $-\sum_{k=1}^{k=m} \int_{t_n}^{t_{n+1}} \alpha_k(\tau - t_n) z_{k,n} d\tau = \epsilon_{n+1} T$ or if we set $-\int_{t_n}^{t_{n+1}} \alpha_k(\tau - t_n) = A_k(t_{n+1} - t_n) - A_k(0)$ and $t_{n+1} - t_n = \Delta t_n$

$$\sum_{k=1}^{k=m} \{A_k(\Delta t_n) - A_k(0)\} z_{k,\,n} = \epsilon_{n+1} T.$$

Therefore, we may write:

$$\Delta t_n = f(\underline{z}_n) \qquad \epsilon_{n+1} = g(\underline{z}_n). \tag{5}$$

The discrete system considered in part III.1. becomes then an autonomous homogeneous system:

$$\underline{Y}_{p+1} = e^{Af(\underline{z}_p)}(\underline{Y}_p + \epsilon_p \underline{B})$$

or

$$\underline{z}_{p+1} = e^{Af(\underline{z}_p)}(\underline{z}_p) + g(\underline{z}_p)\underline{B}.$$

The stability of this system may then be studied using an extension of the Liapunov theorems if a closed form expression or a valid approximation may be found for Eq. (5).

IV. ZERO INPUT ANALYSIS AND STABILITY IN THE CASE OF A FIRST ORDER LINEAR PLANT WHEN THE GPFM REDUCES TO AN IPFM

Consider the system represented in Fig. 7.

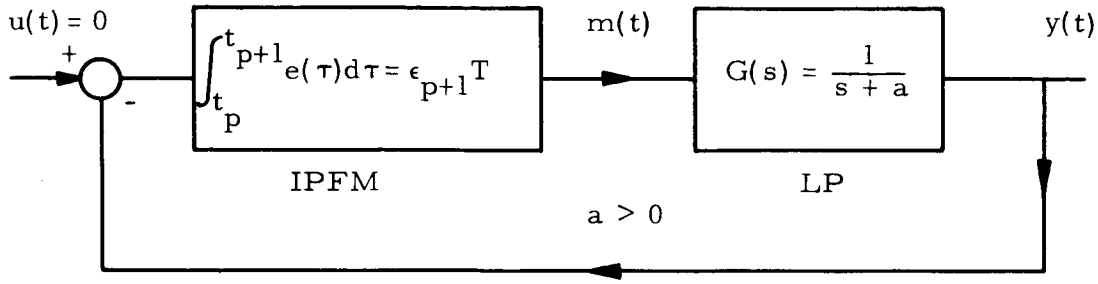


Fig. 7

For $t_p^+ \leq t \leq t_{p+1}^-$

$$y(t) = e^{-a(t - t_p)}(y_p + \epsilon_p)$$

$$y_{p+1} = e^{-a(t_{p+1} - t_p)}(y_p + \epsilon_p) \qquad (6)$$

1. Determination of the $\{t_p\}$ and $\{\epsilon_p\}$

$$-\int_{t_p}^{t_{p+1}} e^{-a(\tau - t_p)}(y_p + \epsilon_p)d\tau = \epsilon_{p+1} T \rightarrow \frac{1}{a}\left\{e^{-a(\tau - t_p)}\right\}_{t_p}^{t_{p+1}} = \frac{\epsilon_{p+1} T}{y_p + \epsilon_p}$$

or

$$e^{-a(t_{p+1} - t_p)} = 1 + \frac{aT}{y_p + \epsilon_p}\epsilon_{p+1} .$$

Since e^{-at} (for $t \geq 0$) is ≤ 1 we have

$$\epsilon_{p+1} = - \text{ sign } (y_p + \epsilon_p) \tag{7}$$

and if $|y_p + \epsilon_p| > aT$

$$e^{-a(t_{p+1} - t_p)} = 1 - \text{sign } (y_p + \epsilon_p)\frac{aT}{y_p + \epsilon_p} .$$

Finally:

$$t_{p+1} - t_p = - \frac{1}{a} \log_N \left(1 - \text{sign } (y_p + \epsilon_p)\frac{aT}{y_p + \epsilon_p} \right). \tag{8}$$

2. Determination of the "equivalent" discrete system

From Eqs. (6), (7), and (8), we deduce

$$y_{p+1} = y_p + \epsilon_p - \Big(\text{sign } (y_p + \epsilon_p)\Big) aT.$$

Now set $y_p + \epsilon_p = z_p$ ($z_p = y(t_p^+)$). We obtain

$$z_{p+1} = z_p - (aT + 1) \text{ sign } z_p. \tag{9}$$

$$\text{with } \begin{cases} \epsilon_{p+1} = -\text{sign } z_p \\ t_{p+1} - t_p = -\frac{1}{a} \log \left(1 - \frac{aT}{|z_p|} \right) \end{cases} \qquad \begin{cases} \epsilon_0 = 0 \\ t_0 = 0 \\ z_0 = y_0 \end{cases}$$

3. Stability

Two cases may be distinguished.

A. $z_0 > aT + 1$ or $< - (aT + 1)$. For instance take $z_0 > 0$. Owing to Eq. (9), sign $(z_p) = + 1$ ($\epsilon_p = -1$) during p steps where p is the largest integer such that

$$\frac{z_0}{aT + 1} - 1 \leq t \leq \frac{z_0}{aT + 1}$$

at the p-th step, $z_p = E_1$ is in the interval $[0, aT + 1]$

a) $E_1 < aT$.

No other pulses will be emitted. The system stops at the p-th instant.

b) $aT < E_1 < aT + 1$ and $aT > 1$.

z_{p+1} is < 0 and is in the interval $[-aT, 0]$ let $E_2 = z_{p+1}$. No other pulses are emitted; the system stops at the (p+1)-th instant.

c) $aT < E_1 < aT + 1$ and $aT < 1$. We see that

$z_{p+2} = z_{p+1} + aT + 1 = z_p = E_1$. z_p will take therefore the values E_1 and E_2 alternatively and the total output will oscillate between E_1 and E_2 $\left(y_t = e^{-a(t-t_p)} z_p \quad \text{for } t_p + \leq t \leq t_{p+1}^- \right)$.

The period of the oscillations is

$$p = -\frac{1}{a} \log_N \left[\left(1 - \frac{aT}{|E_1|} \right) \left(1 - \frac{aT}{|E_2|} \right) \right] \quad \text{(Fig. 8).}$$

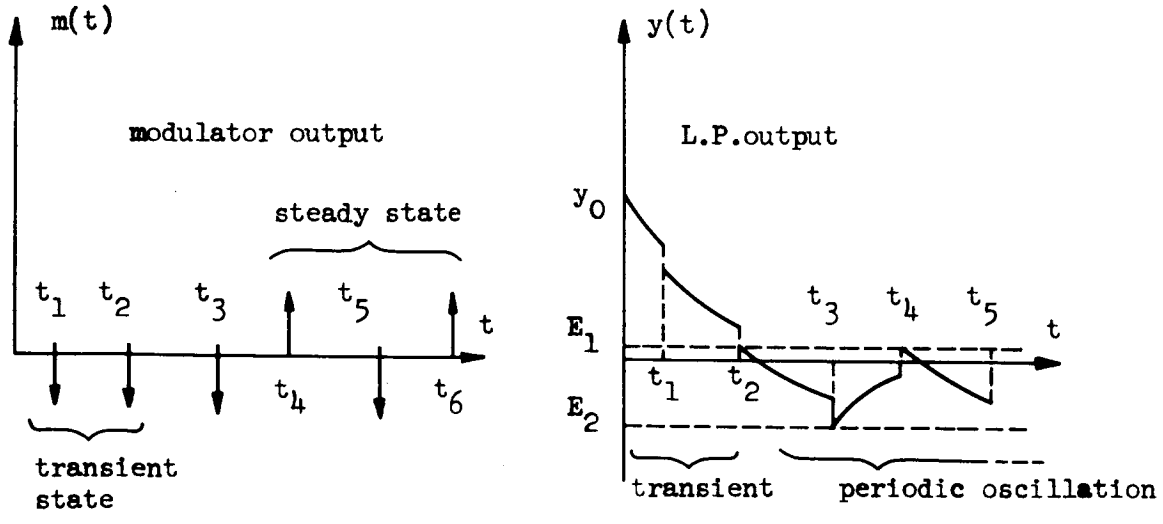


Fig. 8

B. $|z_0| < aT + 1$

Similar conclusion without transient part.

Reference

R. L. Farrenkopf, A. E. Sabroff, and P. C. Wheeler, "An integral pulse frequency on-off attitude control system," Technical Report, Space Technology Laboratories.

THE PROBLEM OF NEURON MODELING

J. G. Blanchard and E. I. Jury

I. THE PROBLEM OF SIMULATION OF NEURONAL BEHAVIOR

Several attempts have been made in the past to simulate by electrical networks the behavior of neural elements. Physiologists[1] have tried to match their experimental results with partly empirical formulas describing the variations of the main chemical or electrical factors taken into consideration. Engineers in the field of bioengineering have directed their efforts toward developing models reproducing qualitatively the main electrical properties of the neuron.[2, 3]

A brief survey of the literature shows that connections between the work of those physiologists and bioengineers are not always as strong as it would be desirable. For some reasons due mainly to the complexity of the phenomena to be studied, the different models presented by engineers appear to lack a certain unity which would have been preserved if more attention could have been devoted to the assumptions, approximations, and limitations involved. Furthermore, as the models become more and more sophisticated, the risk of making speculative assumptions increases unless the physiological aspects of the problem are constantly kept in mind. It appears finally that the models in question do not display enough degrees of freedom to take into account most of the parameters influencing the neuronal behavior.

It is our opinion that the problem of neuron modeling ought to be rethought if the work of bioengineers is to take a larger place in this particular domain of neurophysiology. Further research could be oriented in two different directions:

The research herein was supported by the Air Force Office of Scientific Research under Grant AF-AFOSR-292-64.

A. Development of accurate models based on quantitative and qualitative experimental data about the chemical and electrical behavior of the precise neural element to be studied. Owing to the nonlinear and time-varying character of the phenomena involved, it is likely that any model of this sort will necessarily be complex and defy any deep analysis with mathematical tools now available. However, a computer study is always possible. Any such attempt should require the combined efforts of physiologists and engineers. It would be particularily helpful if further investigation were made of the neuronal properties as a function of different parameters, such as temperature or external chemical concentration.

B. Development of simplified models, in which more initiative could be left to the bioengineer, in order to investigate very complex problems such as those arising in the behavior of groups of neurons or synapses. Any approach to those problems should require definitions and methods not always available to the physiologist. Optimal models could then be defined with respect to certain performance criteria corresponding to the particular behavioral function (or functions) to be studied.

II. A MODEL BASED ON THE HODGKIN AND HUXLEY EQUATIONS

In 1952, A. L. Hodgkin and A. F. Huxley proposed an electrical circuit (Fig. 1) to represent the membrane of an axon.[1] It is our opinion that this model is as important today.

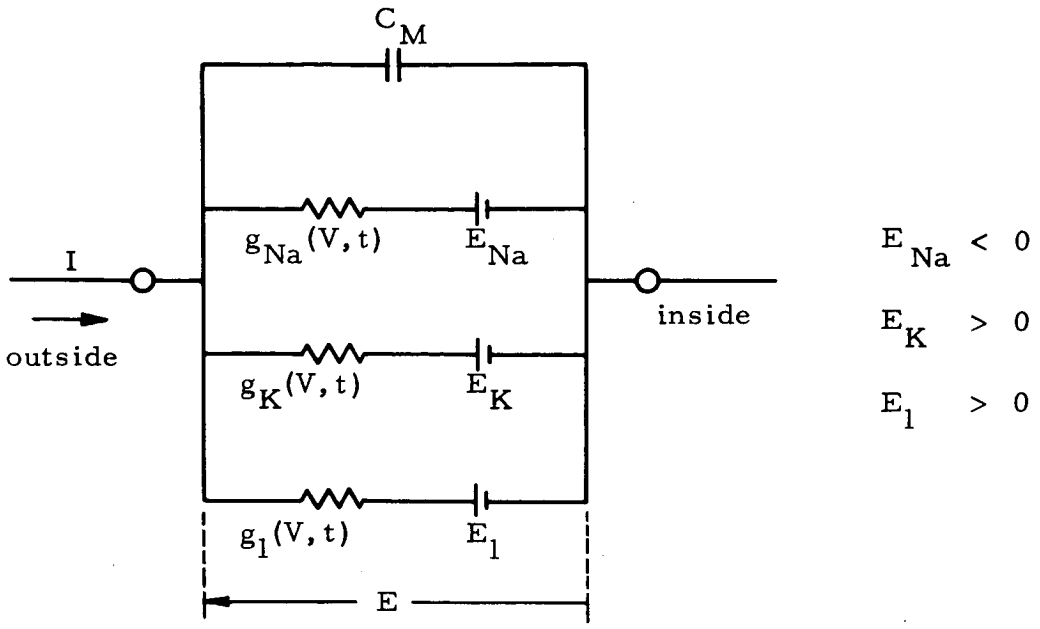


Fig. 1

A. Definitions and Notations

I and E are, respectively, the transmembrane current and potential. E_{Na} and E_K denote the sodium and potassium equilibrium potential as calculated from the Nernst equation. E_l is known as the leakage potential and takes into account other ions such as clorine ions. E_r represents the resting transmembrane potential. g_{Na}, g_K, and g_l are the sodium, potassium, and leakage conductances. C_M is the transmembrane capacitance. We now set:

$$V = E - E_r$$
$$V_{Na} = E_{Na} - E_r$$
$$V_K = E_K - E_r \qquad V_{Na} < 0,\ V_K > 0,\ V_l < 0$$
$$V_l = E_l - E_r$$

B. Analysis of the Circuit

I and V are variables. g_{Na} and g_K are function of V and t. The other quantities are parameters.

Basic equation:

$$I = C_M\frac{dV}{dt} + g_{Na}(V, t)(V - V_{Na}) + g_K(V, t)(V - V_K) + g_l(V - V_l). \qquad (1)$$

The quantities g_{Na} and g_K are determined as follows:

$$g_{Na} = g_l m^3 h$$

$$g_K = g_2 n^4$$

where g_l and g_2 are constants and m, h, n, are variables determined by a set of time-varying first-order differential equations of the form:

$$\frac{dx}{dt} = a_x(V(t))\cdot(I - x) - b_x(V(t))\cdot x \qquad \text{for } x = m,\ n,\ h, \qquad (2)$$

where a_x and b_x are known functions of V(t) determined empirically (see appendix). The constants x(0) and (x = m, n, h) are also determined empirically in order to fit experimental results.

The main difficulty arising in the analysis of these equations is due to the variation of g_{Na} and g_K. Theoretically, it is possible to solve equations(2) and express their solution as a function of V(t), t and the initial values x(0). If we now insert these solutions into equation (1), we obtain the following results:

i) $\dot{V}(t) = H(V(t), t) + kI(t)$.

This form is of interest if I(t) is considered as being the input and V(t) is the output of the system.

ii) $I = G(V(t), t) + C_M \dot{V}$.

This relation would give the output I(t) as a function of the input V(t). For example, this equation would enable us to determine the transmembrane current when V(t) is constant (Simulation of Voltage Clamp experiment).

C. Model Properties

It has been shown by Huxley and Hodgkin that equations (1) and (2) describe satisfactorily the membrane behavior when a stimulus (current stimulus or initial value $V_0 \neq 0$ imposed to the transmembrane potential) is applied. The phenomena of potential threshold, current stimulus threshold, refractoriness (absolute and relative), accommodation anode break excitation, and so on, have been correctly reproduced by these equations.

D. The Problem of Local Excitation by Current Stimuli and Its Analog Computer Simulation

This problem is of particular interest in the study of the receptor (or generator) potential and repetitive firing relationship. It is generally assumed that the current induced by the generator potential plays the role of stimulus in generating the repetitive discharge.

An analog computer model based on equations (1) and (2) enables us to simulate the I(t), V(t) relationship (see below). The interest of such a simulation is obvious. The main chemical and electrical factors, such as V_K or g_l or even some empirical relations such as $a_m(V)$, can be easily modified and adjusted. This type of model could be considered as belonging to the class A models described in Part I of this paper.

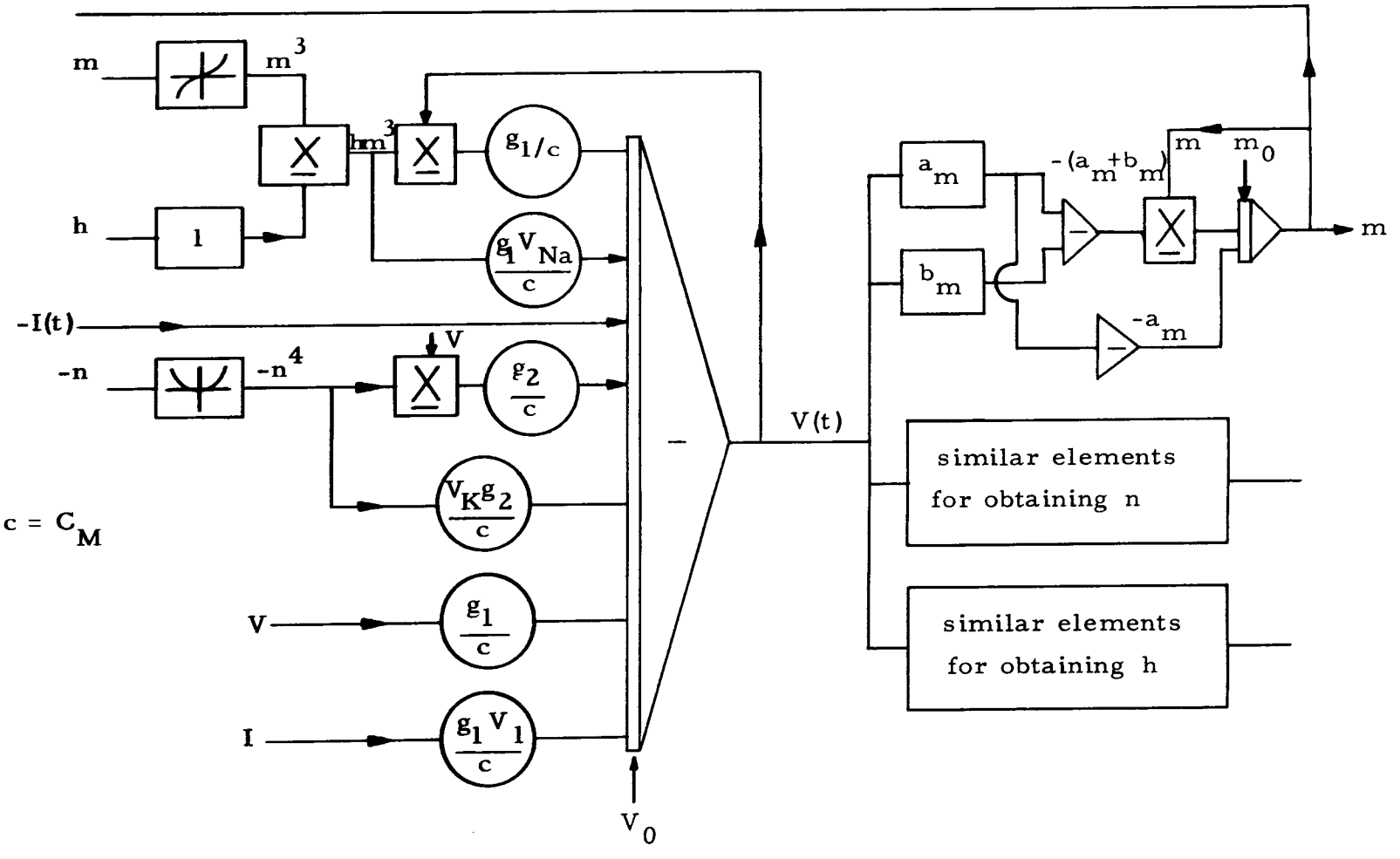

m
m^3
h
1
hm^3
$g_{1/c}$
$\frac{g_1 V_{Na}}{c}$
-I(t)
-n
$-n^4$
V
$\frac{g_2}{c}$
$\frac{V_K g_2}{c}$
V
$\frac{g_1}{c}$
I
$\frac{g_1 V_1}{c}$
V_0
V(t)
$c = C_M$
a_m
b_m
$-(a_m+b_m)$
m
m_0
$-a_m$
m
similar elements for obtaining n
similar elements for obtaining h

Fig. 2

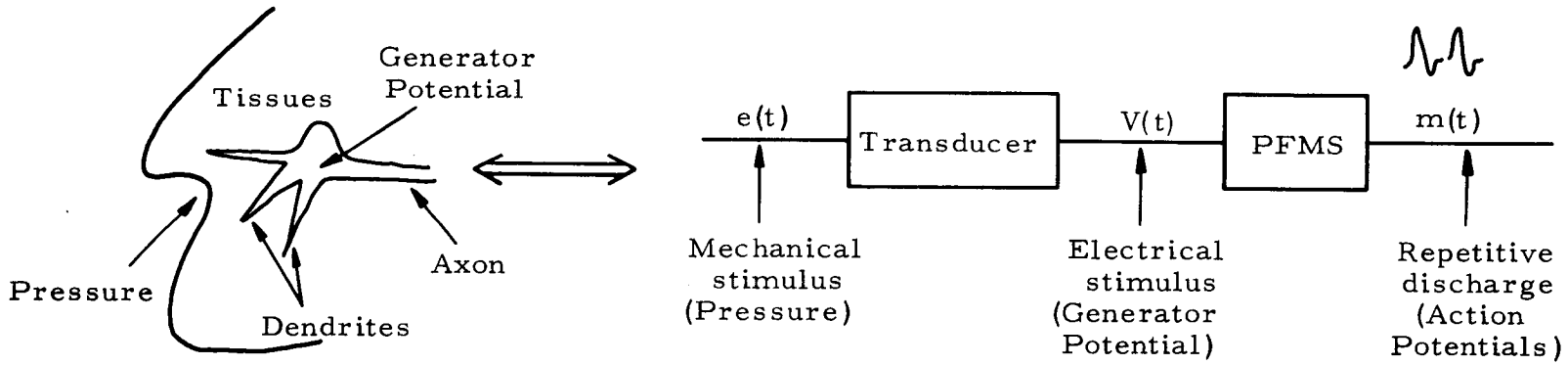

Tissues
Generator Potential
Pressure
Dendrites
Axon
e(t)
Transducer
V(t)
PFMS
m(t)
Mechanical stimulus (Pressure)
Electrical stimulus (Generator Potential)
Repetitive discharge (Action Potentials)

Fig. 3

III. A SIMPLIFIED MODEL USING THE PROPERTIES OF PULSE FREQUENCY MODULATED SYSTEMS[4, 5, 6]

We present now a simplified model which could be of interest in the study of sensory receptors and the transmission of information to axons. It is generally accepted that the original stimulus causes a deformation of the dendritic branches of the sensory receptor. A long-lasting depolarization of the resting potential of the receptor known as generator potential follows. The generator potential acts as a persistent cathodal stimulus generating currents which cause the initial segment of the axon to respond repetitively to the depolarization.[7] (See below)

A. Analysis of the Model

i) Definition of the transducer: The transducer is assumed to be a linear system converting the original stimulus (e.g., mechanical) into an electrical stimulus (generator potential). For large e(t), there is a saturation.

$$V(t) = \int_0^t h(t - u) f(e(u)) du,$$

where f(e(u)) is represented in Fig. 4.[8] In a first approximation, h(t) can be taken as e^{-at} with $a > 0$.

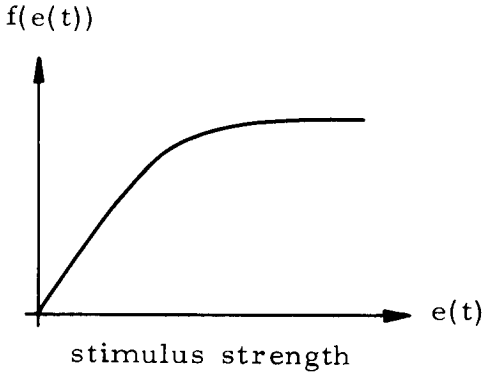


Fig. 4

ii) Definition of the PFMS (Pulse Frequency Modulated System).

This element converts the information provided by the generator potential into a sequence of pulses (action potentials) called the repetitive discharge.

a) Notations:

$$I(t, t_k) = \int_{t_k}^{t} e^{-c(u - t_k)} \cdot V(u)du$$

$$T_k(t) = T_0 \cdot e^{bk} / \left(1 - e^{-q(t - t_k - t_r)} \cdot f(t_k)\right)$$

$$P_{l,\,h}(t - t_k) = h \cdot P(t - t_k)$$

where c, b, q, h, t_r and T_0 are positive constants,

t_k is the time at which the k-th pulse (action potential) is emitted,

V(t) is the generator potential,

P(t) is a continuous function of time t, equal to zero outside a certain interval (0, d) where d is positive (Fig. 5), and

$f(t_k) = 0$ if $k = 0$,

$= 1$ if $k \geq 1$.

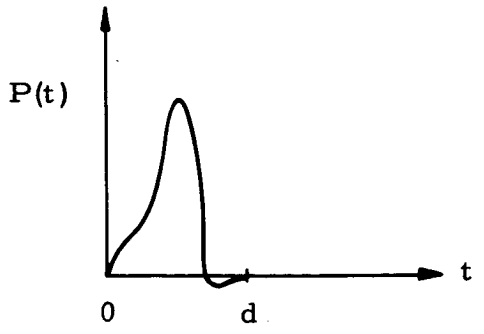


Fig. 5

b) Let us clarify now the meaning of those parameters from the point of view of physiology:

c refers to the delay separating the beginning of a stimulus above threshold and the time of emission of an action potential. Furthermore, it will determine the value of the rheobase if the strength-duration curve is studied.

T_0 can be described as being the value of the membrane threshold. when no recent stimulus has been applied.

b represents a quantity which can be taken as the rate of adaptation of the receptor (b is in general very small).

q and the curve $1/(1 - e^{-q(t-t_k-t_r)})$ represent the variation of excitability (inverse of the threshold) after the period of absolute refractoriness following the emission of an action potential. They refer therefore to the relative refractoriness. The coefficient $f(t_k) = 0$ or 1 enables us to study separately the case where no recent stimulus has been applied (the membrane potential is equal to its resting value). The quantity $t-t_k-t_r$ can be considered as being infinite and the threshold $T_0(t)$ reduces to T_0.

t_r represents the absolute refractory period.

P(t) describes the form of the action potential and the coefficient h enables to adjust its magnitude. The origin of the potential axis is taken at $-E_r$ where E_r is the resting value of the membrane potential.

V(t) is always assumed to be positive. We only consider the case of depolarizing inputs.

c) Analysis of the input-output relationship of the PFMS

Assume an action potential to have been emitted at time t_k. The value of the time t_{k+1} at which the next action potential is to be emitted is the smallest positive root of the equation:

$$I(t, t_{k+1}) = T_k(t).$$

In general, we can write the output as: $m(t) = \sum_{t_k} P_{1,h}(t-t_k)$ where

m(t) represents the repetitive discharge of action potentials.

Example: Set $V(u) = V$ = constant. Consider $k = 0$ (first action potential). We then have $V(1 - e^{-t}) = c \cdot T_0$.

For the following action potentials, we have:

$$V(1 - e^{-(t-t_k)}) = c \cdot T_0/(1 - e^{-q(t-t_k-t_r)}) \ (*), \qquad t_k \leq t \leq t_{k+1}.$$

d) We shall now prove that this model agrees with the main properties of the neuron.

Strength duration curve

For $V < c \cdot T_0$, there is no excitation. The rheobase is therefore equal to $c \cdot T_0$. The equation of the strength duration curve is

$V = c \cdot T_0/(I - e^{-t})$ (see reference 9).

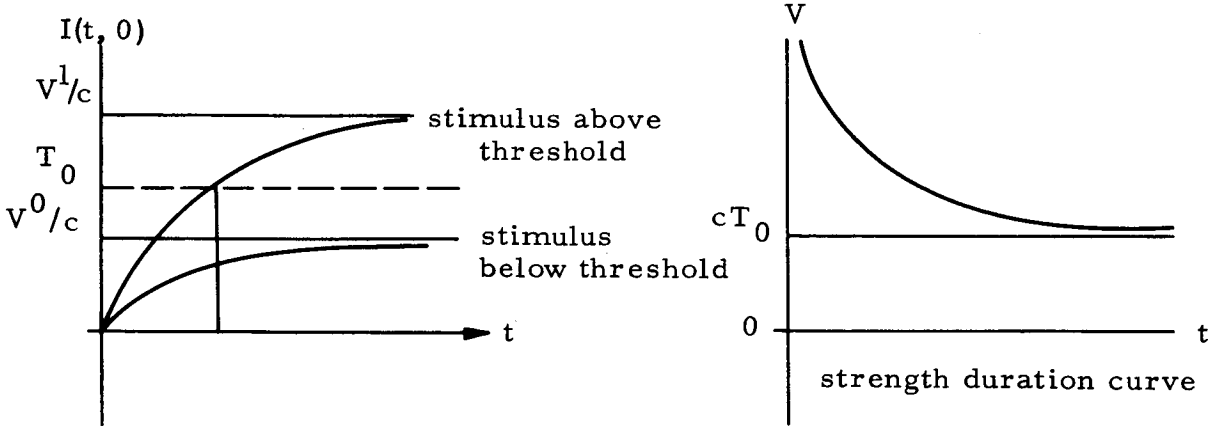


Fig. 6

Gradient-threshold curve

Assume $V(t) = st$ where s is a constant. We have

$$I(t, 0) = (-s/c)te^{-ct} + (s/c^2)(1 - e^{-ct}).$$

$I(t, 0)$ is monotone increasing and its maximum is s/c^2. We see therefore that for $s < c^2T_0$ there cannot be excitation. This value c^2T_0 corresponds to the gradient threshold. See Fig. 7 below.

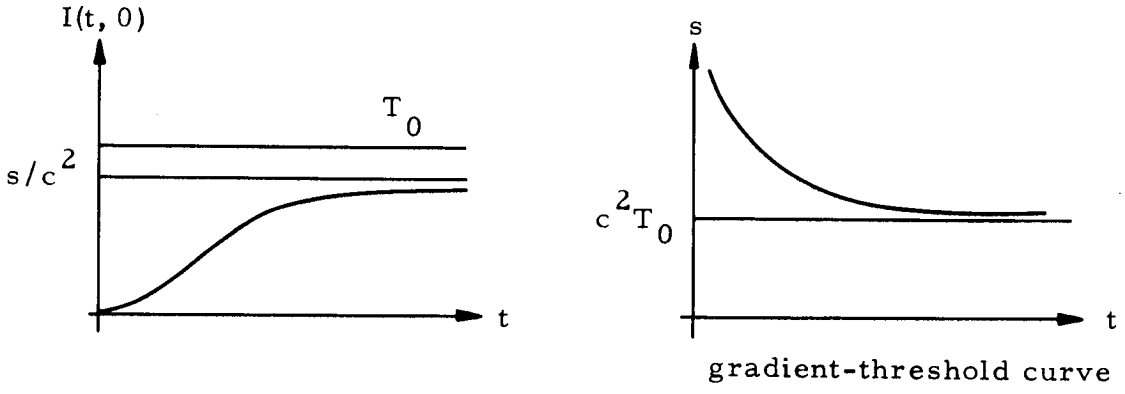


Fig. 7

Absolute refractoriness

Assume that an action potential has been emitted at time t_k and that $V(t)$ is a constant equal to V.

As we have seen, the following equation enables us to compute the time t_{k+1} of emission of the next action potential:

$$V(1 - e^{-(t-t_k)}) = c \cdot T_0/(1 - e^{-q(t-t_k-t_r)}).$$

We see that, whereas the left inside of this equation is always positive for $t > t_k$, the right inside will only be positive if $t > t_k + t_r$. In other words, there cannot be any root of the equation $I(t, t_k) = T_k(t)$ lying in the interval $(t_k, t_k + t_r)$. t_r is the absolute refractory period.

Relative refractoriness

Figure 8 shows the variations of $T_k(t)$ for $t > t_k + t_r$ and the variation of $I(t, t_k)$ for $t > t_k$. The variation of the threshold is in good agreement with experimental data (see reference 9).

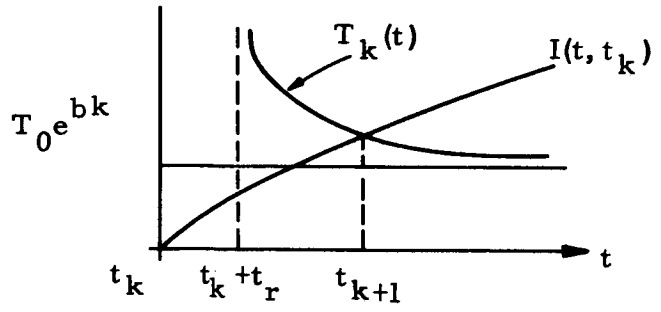


Fig. 8

Adaptation

We see from (*) that when k increases (for V constant), e^{bk} increases and $t_{k+1} - t_k$ increases. This property corresponds to the well-known property of adaptation (Fig. 9).

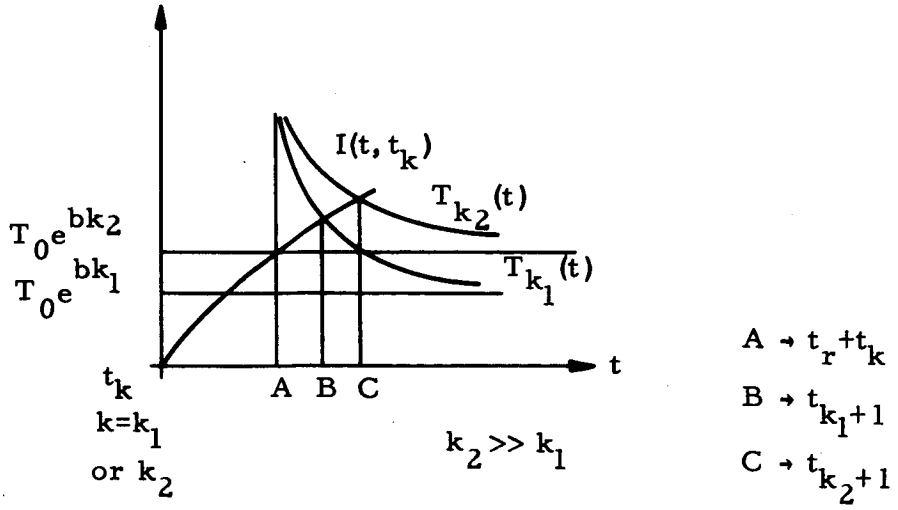


Fig. 9

B. Interest and Limitations of the Model

The digital simulation of the model is extremely easy. Furthermore, the parameters of this model can be adjusted to agree with experimental data (for example, t_r = 2 ms, d = 6 ms, h = 95 mv). Finally this model reproduces the main properties of neuron excitation. It could be noted, however, that this model does not intend to provide the values of a subthreshold potential. However, $I(t, t_k)$ is a good approximation to this potential if accommodation for subthreshold inputs is neglected.

APPENDIX:

$$a_n = 0.01\,(v + 10)/(e^{\frac{v+10}{10}} - 1)$$

$$b_n = 0.125\, e^{v/80}$$

$$a_m = (0.1)\cdot(v + 25)/(e^{\frac{v+25}{10}} - 1)$$

$$b_m = 4\, e^{v/18}$$

$$a_h = 0.07\, e^{v/20}$$

$$b_h = 1/(e^{\frac{v+30}{10}} + 1)$$

v in m volts.

REFERENCES

1. A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane currents and its application to conduction and excitation in nerve." J. Physiol., vol. 117, pp. 500-544; March, 1952.
2. L. D. Harmon, "Artificial neuron," Science, vol. 129; April, 1959.
3. F. F. Hiltz, "Analog computer simulation of a neural element, " IRE Trans. on Bio-medical Engineering; January, 1962.
4. Jones, et. al., "Pulse modulation in physiological systems, phenomenological aspects," IRE Trans. on Bio-medical Engineering; January, 1961.
5. Farrenkopf, et. al., "An integral pulse frequency on-off attitude control system," Technical Report; March, 1963. Space Technology Laboratories.
6. J. G. Blanchard, "On the definition and analysis of pulse frequency modulated systems," Notes on System Theory-vol. VII; to be published ERL, University of California.
7. E. E. Selkurt, Physiology, Little, Brown and Company, Boston; 1962.
8. J. Gray and M. Sato, J. of Physiology, vol. 122, p. 610; 1953.
9. I. Tasaki, "Conduction of nerve impulse," Handbook of Physiology - Section I - Neurophysiology-American Physiological Society, 1959.

ON THE STABILITY OF FEEDBACK CONTROL SYSTEMS WITH PERTURBATION GAIN*

C. T. Chen

Abstract

This paper considers a special class of single-loop feedback systems (such as shown in Fig. 1) both in the continuous and in the sampled-data cases. It is shown that if the gain deviation from the linearity of the memoryless nonlinear element, $\tilde{\varphi}(\sigma, t)$, is smaller than $\lambda(t)\sigma$, then $\lambda(\cdot) \in L^1(0, \infty)$ implies that the system output is bounded for any bounded input. As $t \to \infty$, the system output tends to that of the linearized ($\tilde{\varphi}(\sigma, t) \equiv 0$) system as $t \to \infty$. Similar results hold for the sampled-data case.

I. Continuous Case

Given the system

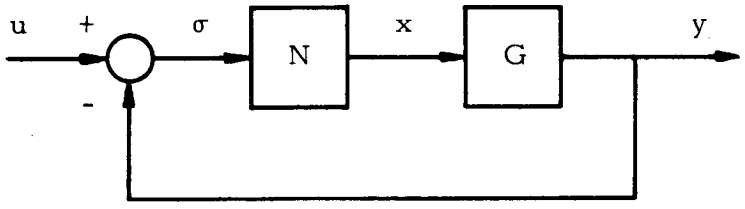


Fig. 1

where N is a memoryless, time-varying, nonlinear element, G is a linear, time-invariant subsystem.

Assumption

(A.1) The input-output relation of G is given by

*This work was supported wholly by the Joint Services Electronics Program (U.S. Army, U.S. Navy and U.S. Air Force) under Grant No. AF-AFOSR-139-64.

$$y(t) = z(t) + \int_0^t g(t - \tau)x(\tau)d\tau \qquad (1)$$

where $y(t)$ is the output, $x(t)$ the input, $z(t)$ the zero-input response, and $g(t)$ the unit impulse response of G.

(A.2) For all initial states, $z(t)$ is bounded on $[0, \infty)$

(A.3) $g(t) = 1(t)[r + g_1(t)]$ where $1(t)$ is the unit step function; $r \geq 0$; $g_1(t)$ is an element of $L^1(0, \infty)$, bounded and $\to 0$ as $t \to \infty$.

(A.4) The nonlinear, time-varying element N is characterized by $x = \varphi(\sigma, t)$, and $\varphi(\sigma, t) = k\sigma + \tilde{\varphi}(\sigma, t)$, $k > 0$ where $\tilde{\varphi}(\sigma, t)$ is the nonlinear effect due to gain deviation.

Theorem 1. Suppose the feedback control system shown in Fig. 1 satisfies (A.1) ~ (A.4), and if

(a) $1 + kG(s) \neq 0$ for all Re $s \geq 0$

(b) there exists a continuous, nonnegative, real valued function $\lambda(t)$ such that

$$|\varphi(\sigma, t)| \leq \lambda(t)|\sigma|, \text{ and } \int_0^\infty \lambda(t)dt \leq N < \infty$$

then

(i) for all bounded inputs u, the output y is bounded

(ii) for any output $y(t)$, there exists a response $\hat{y}(t)$ of a linearized system [by putting $\tilde{\varphi}(\sigma, t) \equiv 0$] which has the same input and initial state, such that

$$\lim_{t \to \infty} |y(t) - \hat{y}(t)| = 0.$$

Proof: If $\tilde{\varphi}(\sigma, t) \equiv 0$, and (A.1) ~ (A.3) and (a) are satisfied, the closed loop unit impulse response, $h(t)$, of the linearized system will

be an element of $L^1(0, \infty)$, bounded, $\to 0$ as $t \to \infty$.[1] It follows, that for any bounded input, the output $\hat{y}(t)$ will be bounded.

Introducing $h(t)$ and $\hat{y}(t)$, Eq. (1) can be written as

$$y(t) = \hat{y}(t) + \frac{1}{k}\int_0^t h(t - \tau)\tilde{\varphi}[u(\tau) - y(\tau), \tau]d\tau \qquad (2)$$

where $\int_0^\infty h(t)e^{-st}dt = \frac{kG(s)}{1 + kG(s)}$.

Let $\hat{y}_M \triangleq \sup_{0\le t\le\infty} |\hat{y}(t)|$, $h_M \triangleq \sup_t |h(t)|$, $u_M \triangleq \sup_t |u(t)|$, then

$$|y(t)| \le \hat{y}_M + \frac{1}{k}\int_0^t |h(t - \tau)|\lambda(\tau)|u(\tau) - y(\tau)|d\tau$$

$$\le \hat{y}_M + \frac{1}{k} h_M u_M N + \frac{h_M}{k}\int_0^t \lambda(\tau)|y(\tau)|d\tau.$$

Apply the Gronwall-Bellman inequality; we get $|y(t)| \le (\hat{y}_M + \frac{1}{k}h_M u_M N)e^{\frac{h_M N}{k}}$ i.e., $y(t)$ is bounded on $[0, \infty)$. Define $y_M \triangleq \sup_t |y(t)|$. From Eq. (2)

$$|y(t) - \hat{y}(t)| \le \frac{1}{k}\int_0^t |h(t - \tau)||\tilde{\varphi}[u(\tau) - y(\tau), \tau]|d\tau$$

$$\le \frac{u_M + y_M}{k}\int_0^t |h(t - \tau)|\lambda(\tau)d\tau.$$

Because $h(t)$ is bounded on $[0, \infty)$, tends to zero as $t \to \infty$, and $\lambda(\cdot) \in L^1_{(0, \infty)}$ from Lemma 1, it can be concluded that

$$|y(t) - \hat{y}(t)| \to 0 \quad \text{as} \quad t \to \infty. \qquad \text{Q.E.D.}$$

Lemma 1. Let f_1 be bounded on $(0, \infty)$, $\to 0$ as $t \to \infty$, and $f_2 \in L^1_{(0, \infty)}$; then $f_1 * f_2 \to 0$ as $t \to \infty$.

Proof: For any $\epsilon > 0$, $\exists T$ and T' such that $|f_1(t)| < \epsilon$ for all $t > T$ and $\int_0^\tau |f_2(t - \tau)|d\tau < \epsilon$, for $t > T' > T$. Let $|f_1(t)| \leq f_{1M}$, then

$$|f_1 * f_2| \leq \int_0^T |f_2(t - \tau)|\,|f_1(\tau)|d\tau + \int_T^t |f_2(t - \tau)|\,|f_1(\tau)|d\tau$$

$$\leq f_{1M}\epsilon + \epsilon\int_T^t |f_2(t - \tau)|d\tau \leq \epsilon(f_{1M} + \int_0^\infty |f_2(t)|dt)$$

for $t > T'$. Q.E.D.

Corollary 1.1. Let (A.1) ~ (A.4), (a) and (b) hold. If, in addition $r > 0$, $z(t) \to z_\infty$ as $t \to \infty$, and if $u(t)$ is bounded on $[0, \infty)$ and tends to u_∞ as $t \to \infty$, then the output $y(t)$ will tend to u_∞ as $t \to \infty$.

Proof: From Ref. (1) $\hat{y}(t) \to u_\infty$ as $t \to \infty$. Apply Theorem 1. $y(t) \to u_\infty$ as $t \to \infty$. Q.E.D.

Corollary 1.2. Let (A.1) ~ (A.4), (a) and (b) hold. If, in addition $r > 0$, $z(t) \to z_\infty$ as $t \to \infty$, and if $u(\cdot) \in L^1_{(0, \infty)}$, then $y(t) \to 0$ as $t \to \infty$.

Proof: This is a direct application of Theorem 1, Lemma 1, and the fact $z(t) \to z_\infty \Rightarrow \int_0^t h(t - \tau)z(\tau)d\tau \to z_\infty$ as $t \to \infty$.[1] Note: $u(\cdot) \in L'_{(0, \infty)}$ $\not\Rightarrow u(t) \to 0$, as $t \to \infty$.

II. Sampled-data System

Given the system

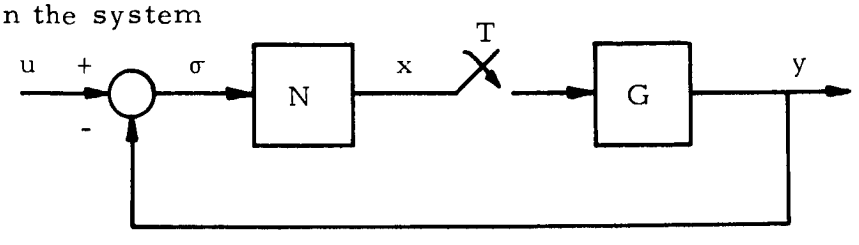


Fig. 2

where N is a memoryless, time-varying nonlinear element, G is a linear, time-invariant subsystem. The sampling period, T, is assumed to be constant.

Assumption

(B.1) The input-output relation of G is characterized by

$$y(n) = z(n) + \sum_{m=0}^{n} g(n - m)\, x(m)$$

where the sequence $\{y(n)\}$ is the sampled output, $\{x(n)\}$ the sampled input, $\{z(n)\}$ the sampled zero-input response, and $\{g(n)\}$ the sampled unit impulse response.

(B.2) For all initial states, $\{z(n)\}$ is a bounded sequence.

(B.3) $$g^*(t) \triangleq \sum_{n=0}^{\infty} g(n)\delta(t - nT) = \sum_{n=0}^{\infty} [r + g_1(n)]\delta(t - nT)$$

where r is a nonnegative real-valued constant; $\sum_{n=0}^{\infty} |g_1(n)| < \infty$,

a condition which may be written $\{g_1(n)\} \in \ell^1$.

(B.4) The nonlinear element is characterized by $x = \varphi(\sigma, n)$ and $\varphi(\sigma, n) = k\sigma + \tilde{\varphi}(\sigma, n), \quad k > 0$

where $\tilde{\varphi}(\sigma, n)$ is the nonlinear effect due to gain deviation.

Definition:

$$G^*(s) \triangleq \mathcal{L}[g^*(t)] = \sum_{n=0}^{\infty} g(n)e^{-nTs}$$

$$H^*(s) \triangleq \frac{kG^*(s)}{1 + kG^*(s)}$$

$$h^*(t) = \mathcal{L}^{-1}[H^*(s)] = \sum_{n=0}^{\infty} h(n)\delta(t - nT)$$

Theorem 2. Suppose the sampled-data system shown in Fig. 2 satisfies (B.1) ~ (B.4), and if

(a) $1 + kG^*(s) \neq 0$ for all Re $s \geq 0$

(b) there exists a nonnegative real-valued sequence $\{\lambda(n)\}$

such that $|\tilde{\varphi}(\sigma, n)| \leq \lambda(n) |\sigma|$ and $\sum_{n=0}^{\infty} \lambda(n) \leq M < \infty$; furthermore,

$0 \leq \frac{h_M}{k}\lambda(n) < 1$ for $n = 0, 1, 2, \ldots$, where

$h_M \triangleq \max_n |h(n)|$

then

(i) for any bounded input sequence $\{u(n)\}$, the output sequence $\{y(n)\}$ is bounded.

(ii) for any output $\{y(n)\}$, there exists a response $\hat{y}(n)$ of a linearized system [by putting $\tilde{\varphi}(\tau, n) \equiv 0$] which has the same input $\{u(n)\}$ and initial state, such that

$\lim_{n\to\infty} |y(n) - \hat{y}(n)| = 0.$

In proving this theorem, we need the following lemma.

Lemma 2. Let $\{g(n)\}$, $\{\lambda(n)\}$ be real-valued sequences, c, d be real constants. If $\{\lambda(n)\} \in \ell^1$ and $0 \leq d\lambda(n) < 1$ for $n = 0, 1, 2, \ldots$, and if

$$g(n) \leq c + d \sum_{m=0}^{n} \lambda(m)g(m)$$

then there exists a positive constant $\beta > 0$, such that

$$g(n) \leq \frac{c}{\prod_{m=0}^{n} [1 - d\lambda(m)]} \leq \frac{c}{\beta} < \infty \qquad \text{for } n = 0, 1, 2, \ldots$$

Proof: This lemma will be proved in two steps.

(i) $g(n) \le c + d \sum_{m=0}^{n} \lambda(m)g(m) \le \dfrac{c}{\prod_{m=0}^{n} [1 - d\lambda(m)]}$ for $n = 0, 1, 2, \ldots$

If $n = 0$, it is obvious. Suppose when $n = k$, it is true, then when

$n = k + 1$, $[1 - d\lambda(k + 1)]g(k + 1) \le c + d \sum_{m=0}^{k} \lambda(m)g(m)$

$\le \dfrac{c}{\prod_{m=0}^{k} [1 - d\lambda(m)]}$; (i) is proved by induction.

(ii) $0 < \beta < \prod_{m=0}^{\infty} [1 - d\lambda(m)]$.

Because $\{\lambda(n)\} \in l^1$, there exists an N, such that $d \sum_{m=N}^{\infty} \lambda(m) < \frac{1}{2}$.

It implies $\prod_{m=N}^{\infty} [1 - d\lambda(m)] \ge 1 - d \sum_{m=N}^{\infty} \lambda(m) > \frac{1}{2}$.

Let $\prod_{m=0}^{N-1} [1 - d\lambda(m)] = 2\beta > 0$

then

$$\prod_{m=0}^{\infty} [1 - d\lambda(m)] > \beta > 0.$$

Because $\prod_{m=0}^{n} [1 - d\lambda(m)]$ is monotonically nonincreasing and it is bounded from below, the limit exists. By (i), (ii) the desired result follows. Q.E.D.

Proof of Theorem 2. Before we start to prove the theorem, some well-known facts will be recalled. Consider the linearized system shown in Fig. 3.

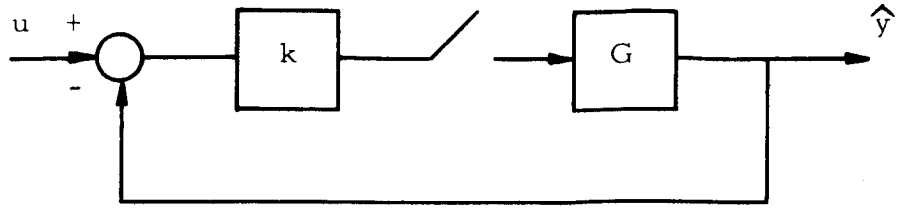


Fig. 3

If (B.1) ~ (B.3) and (a) are satisfied, $H^*(s) = \frac{kG^*(s)}{1 + kG^*(s)}$ is analytic in the closed right half plane, it follows $\{h(n)\} \in \ell^1$. Let $\hat{y}(n)$ be its output, then $\hat{y}(n) = z(n) + \sum_{m=0}^{n} h(n - m)[u(m) - z(m)]$. Since $\{h(n)\} \in \ell^1$, for any bounded input sequence and bounded sampled zero-input response of G, the output $\{\hat{y}(n)\}$ is bounded.

Introducing h(n) and $\hat{y}(n)$, Eq. (3) can be written as

$$y(n) = \hat{y}(n) + \frac{1}{k} \sum_{m=0}^{n} h(n - m)\tilde{\varphi}[u(m) - y(m), m]. \tag{4}$$

Let $\hat{y}_M \triangleq \max_n |y(n)|$, $h_M \triangleq \max_n |h(n)|$, then

$$|y(n)| \leq \hat{y}_M + \frac{1}{k} h_M u_M M + \frac{h_M}{k} \sum_{m=0}^{n} \lambda(m)|y(m)|.$$

Direct application of Lemma 2: $\{y(n)\}$ is bounded. Let $y_M \triangleq \max_n |y(n)|$.

To prove (ii), we need the following well-known fact: if $\{\lambda(n)\} \in \ell^1$, $\{h(n)\} \in \ell^1$, then $\{\lambda(n) * h(n)\} \in \ell^1$ and $\lambda(n) * h(n) \to 0$ as $n \to \infty$. Since

$$|y(n) - \hat{y}(n)| \leq \frac{u_M + y_M}{k} \sum_{m=0}^{n} |h(n - m)|\lambda(m).$$

It follows $|y(n) - \hat{y}(n)| \to 0$ as $n \to \infty$. Q.E.D.

Corollary 2.1. Let (B.1) ~ (B.4), (a) and (b) hold. If, in addition $r > 0$, $z(n) \to z_\infty$ as $n \to \infty$, and if $\{u(n)\}$ is bounded and $u(n) \to u_\infty$ as $n \to \infty$, then $y(n)$ tends to u_∞ as $n \to \infty$.

Proof: For any $\epsilon > 0$, there exists N, such that $|z(n) - z_\infty| < \epsilon$, $|u(n) - u_\infty| < \epsilon$ for $n \geq N$. Since $\{h(n)\} \in \ell^1$, for the same ϵ, there exists $N' > N$, such that $\sum_{m=0}^{N} |h(n-m)| < \epsilon$, for $n > N' > N$. If $r > 0$, $\sum_{n=0}^{\infty} h(n) = G^*(0) = 1$, therefore, for $\epsilon > 0$, there exists N'', such that $1 - \epsilon \leq \sum_{m=N}^{n} h(n-m) \leq 1$ for $n > N''$. Let $P \triangleq \sum_{n=0}^{\infty} |h(n)|$, then

$$\hat{y}(n) = z(n) + \sum_{m=0}^{N} h(n-m)u(m) + \sum_{m=N}^{n} h(n-m)u(m) - \sum_{m=0}^{N} h(n-m)z(m)$$

$$- \sum_{m=N}^{n} h(n-m)z(m)$$

$$\leq z_\infty + \epsilon + u_M \sum_{m=0}^{N} |h(n-m)| + u_\infty \sum_{m=N}^{n} h(n-m) + \epsilon \sum_{m=N}^{n} |h(n-m)|$$

$$+ z_M \sum_{m=0}^{N} |h(n-m)| - z_\infty \sum_{m=N}^{n} h(n-m)$$

$$+ \epsilon \sum_{m=N}^{n} |h(n-m)| \qquad \text{for } n > \max(N', N'')$$

$$\leq u_\infty + \epsilon(1 + u_M + 2P + z_M + z_\infty) \qquad \text{for } n > \max(N', N''). \quad (5)$$

Similarly

$$\hat{y}(n) \geq z(n) - u_M \sum_{m=0}^{N} |h(n-m)| + u_\infty \sum_{m=N}^{n} h(n-m) - \epsilon \sum_{m=N}^{n} |h(n-m)|$$

$$- z_M \sum_{m=0}^{N} |h(n-m)| - z_\infty \sum_{m=N}^{n} h(n-m)$$

$$- \epsilon \sum_{m=N}^{n} |h(n-m)| \qquad \text{for } n > \max(N', N'')$$

$$\geq u_\infty - \epsilon (1 + u_M + 2P + z_M + u_\infty) \qquad \text{for } n > \max(N', N''). \quad (6)$$

Eqs. (5) and (6) imply $\hat{y}(n) \to u_\infty$ as $n \to \infty$. By Theorem 2, $y(n) \to u_\infty$ as $n \to \infty$ Q. E. D.

Acknowledgment

This work was carried out under the supervision of Professor C. A. Desoer; the author wishes to express his gratitude for his guidance and encouragement.

References

1. C. A. Desoer, "A general formulation of the Nyquist criterion," Internal Technical Memorandum M-104, Electronics Research Lab., University of California, Berkeley; Oct., 1964. To appear in Trans. IEEE, CT; June, 1965.

2. W. Kaplan, "Operational methods for linear systems," Addison, Wesley Pub. Co., Reading, Mass., 1962.

SIMPLE CODING FOR K-ARY UNILATERAL CHANNELS

D. Daetz

The problem under consideration is related to one first treated by Shannon.[1] We wish to find the maximum rate at which information can be transmitted over a given channel with zero probability of error ($P_e \equiv$ probability of error). An algorithm has been obtained for generating a code which achieves the maximum rate of information transfer subject to the constraints $P_e \equiv 0$ and $N = 2$ (N is the number of letters in each code word) for a particular class of channels, viz., the K-ary unilateral channel shown in Fig. 1.

Since the rate $R \equiv \frac{\ell n M(N)}{N}$, where $M(N)$ is the number of code words of length N in the code, when $N = 2$ we have $R = \frac{1}{2} \ln M(2)$. To maximize R one must maximize $M(2)$. Let $M^*(N)$ be the maximum value of $M(N)$ consistent with $P_e = 0$ that can be obtained with optimal coding, i.e.,

$$M^*(N) \equiv \max M(N)$$

all codes of length $N \ni P_e = 0$.

This work was supported by the National Science Foundation Cooperative Fellowship Program.

The claim is that for the K-ary unilateral channel

$$M^*(2) = \frac{K^2}{4} \qquad K \text{ even integer}$$

$$M^*(2) = \left[\frac{K(K-1)}{4}\right] \qquad K \text{ odd integer}^{\dagger}$$

The method for finding an optimal or suitable code (i.e., one with $N = 2$, $P_e = 0$, $M(2) = M^*(2)$) is based on the equivalence of the problem of finding a suitable code to the problem of fitting the maximum number of blocks of four squares (hereafter "blocks" will be equivalent to "blocks of four squares") without overlap onto the surface of a torus which is partitioned into K^2 squares (K is the number of inputs and outputs of the channel). First we will demonstrate that the claimed equivalence is valid and that $M^*(2)$ is in fact what we stated it to be. Then the algorithm will be stated and examples of its use displayed.

Consider the K-ary unilateral channel of Fig. 1. Suppose the code word $a_i a_j$ has been chosen for transmission. The possible outputs of the channel when this word is sent are

$$b_i b_j, \quad b_i b_{j+1}, \quad b_{i+1} b_j, \quad b_{i+1} b_{j+1}$$

---

† $[x] \equiv$ largest integer contained in X. Hereafter, brackets will be used only in this connection.

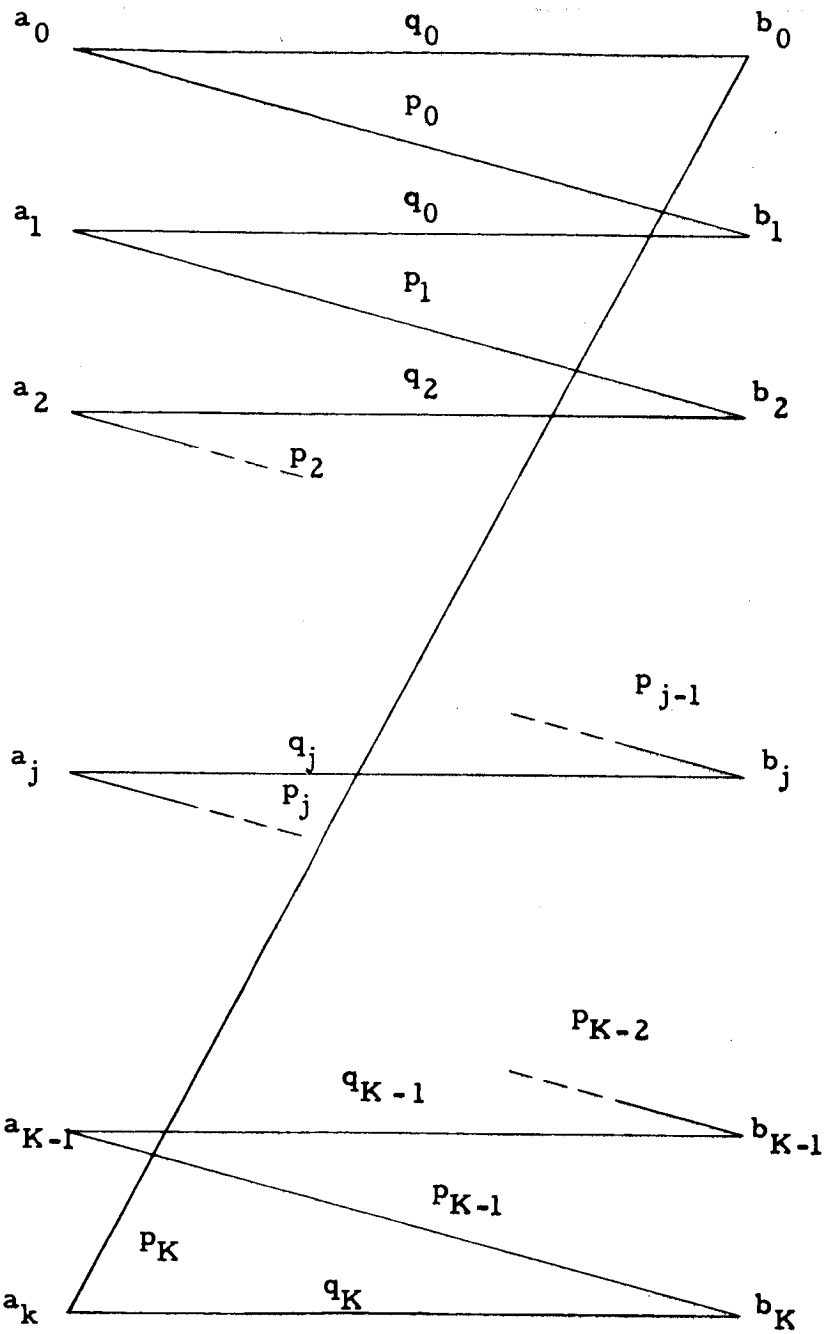


Fig. 1.

$q_j,\ p_j > 0,\ q_j + p_j = 1,\ j = 0, 1, \ldots, K$

where all indices are taken mod K. Let us construct an output table where the rows correspond to the outputs occurring in response to the first letter of the code word and the columns correspond to the outputs occurring in response to the second letter of the code word. Note that the output table is equivalent to squares on the surface of a torus because all indices are taken mod K. Fig. 2 shows an output table. We put x's in the squares corresponding to possible outputs when $a_i a_j$ is sent. This has been done in Fig. 2, and it is seen that a 2x2 block of four squares is formed. The indices of the square in the upper left-hand corner of the block are the same as the indices of the letters in the code word sent. Hence, if we specify the row and column of the square in the upper left-hand corner of the block we specify the code word sent: $i, j \longrightarrow a_i a_j$. If we fit blocks onto the torus without overlap then we can always say unequivocally what word was sent given any received sequence $b_i b_j$. One can be convinced of the need to have nonoverlapping blocks for $P_e = 0$ by considering what happens if we allow overlap. Suppose that both code words $a_i a_j$ and $a_{i+j} a_{j+1}$ are in the code. The outputs corresponding to $a_i a_j$ are $b_i b_j$, $b_i b_{j+1}$, $b_{i+1} b_j$, $b_{i+1} b_{j+1}$ and those corresponding to $a_{i+1} a_{j+1}$ are $b_{i+1} b_{j+1}$, $b_{i+1} b_{j+2}$, $b_{i+2} b_{j+1}$, b_{i+2}, b_{j+2}. Suppose the word $b_{i+1} b_{j+1}$, which is common to the output table blocks of $a_i a_j$ and $a_{i+1} a_{j+1}$, is received. How shall one determine what word was sent? At best one can only make a guess based on the probability that $a_i a_j \longrightarrow b_{i+1} b_{j+1}$ and $a_{i+1} a_{j+1} \longrightarrow b_{i+1} b_{j+1}$. It is clear that in such a situation $P_e > 0$. Therefore, we conclude that $P_e = 0$ requires the blocks to be fit

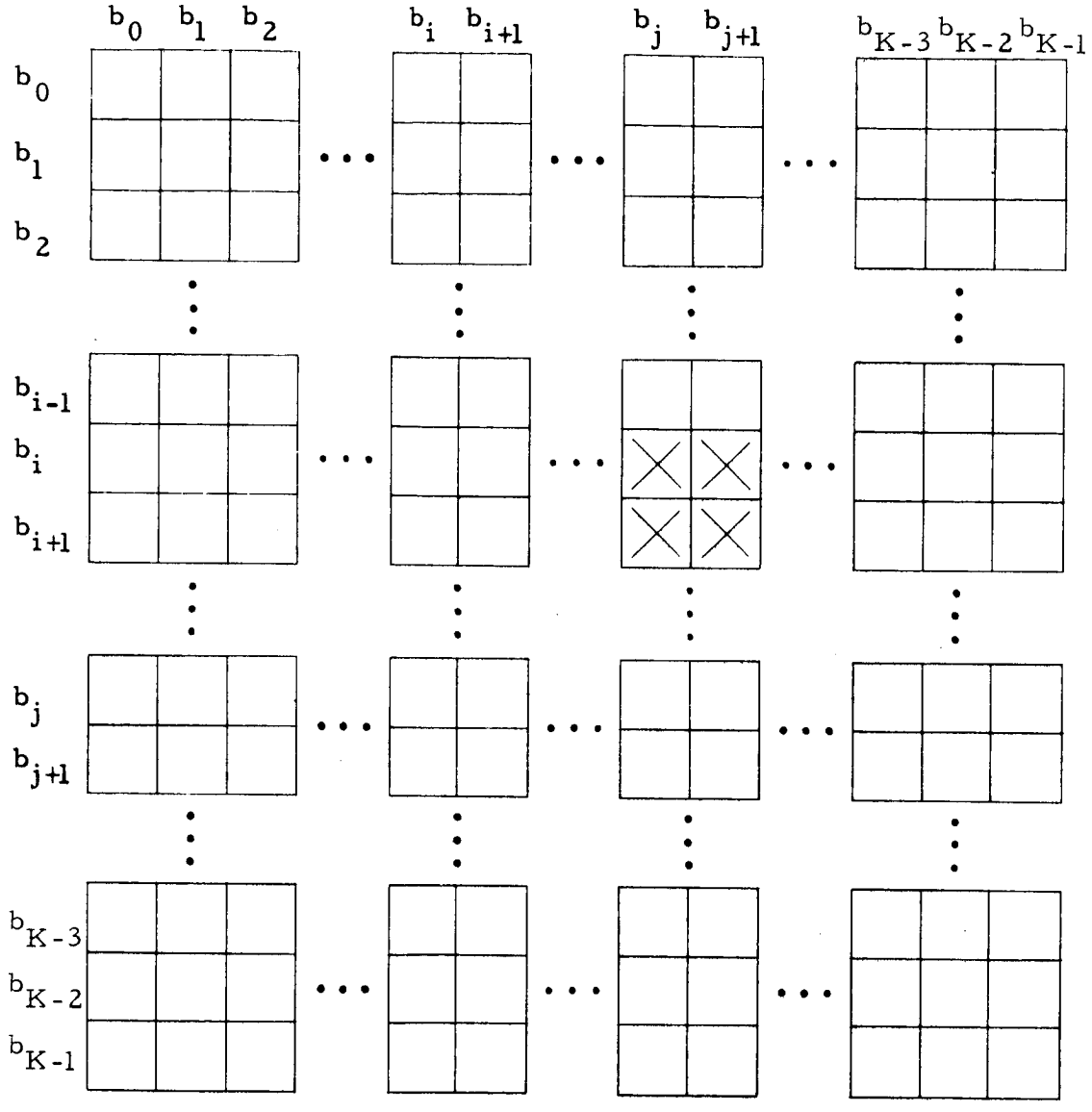

b_0 b_1 b_2 b_i b_{i+1} b_j b_{j+1} b_{K-3} b_{K-2} b_{K-1}
b_0
b_1
b_2
b_{i-1}
b_i
b_{i+1}
b_j
b_{j+1}
b_{K-3}
b_{K-2}
b_{K-1}

Fig. 2.

onto the torus without overlap.

We claimed at the beginning that

$$M^*(2) = \frac{K^2}{4} \qquad K \text{ even integer}$$

$$M^*(2) = \left\lceil \frac{K(K-1)}{4} \right\rceil \qquad K \text{ odd integer}$$

The validity of these claims can be shown as follows:

a. K even.

When K is even we can write $K = 2m$, m integer. Then $K^2 = (2m)^2 = 4m^2$ is the number of squares on the surface of the torus (K rows and K columns in the output table). If we could use every square on the torus when fitting the blocks without overlap then we could place $\frac{K^2}{4}$ blocks on the torus since each block covers four squares. But there is a 1:1 correspondence between code words and blocks. Hence, the maximum number of code words we could hope to find is $\frac{K^2}{4}$. Therefore, $M(2) \leq \frac{K^2}{4}$. Since the algorithm can arrange $\frac{K^2}{4}$ blocks without overlap on the torus, we get in fact $M^*(2) = \frac{K^2}{4}$.

b. K odd.

When K is odd we note that we must always have at least one square in each row and in each column which cannot be used since each block is two squares on a side and two does not divide an odd number evenly. Therefore, at best, we will still be unable to use K of the squares. Since there are K^2 squares on the surface of the torus we are left with only $K^2 - K = K(K-1)$ squares in which to fit the blocks. Now each block uses up four squares so we see that $M(2) \leq \left\lceil \frac{K(K-1)}{4} \right\rceil$. Since the algorithm gives a suitable code with $M(2) = \left\lceil \frac{K(K-1)}{4} \right\rceil$ when K is odd, we have $M^*(2) = \left\lceil \frac{K(K-1)}{4} \right\rceil$.

The algorithm for finding a code with M*(2) code words such that the probability of error is zero

Case I: The number of inputs K is even. $M^*(2) = \frac{K^2}{4}$. One such code satisfying $M(2) = \frac{K^2}{4}$ is formed by using only even inputs (counting zero as even) and letting the code words be the $(K/2)^2$ different two-letter sequences of the letters corresponding to the inputs chosen.

Example: K = 4. See Fig. A. Use inputs "0" and "2". A suitable code has code words: a_0a_0, a_0a_2, a_2a_0, a_2a_2.

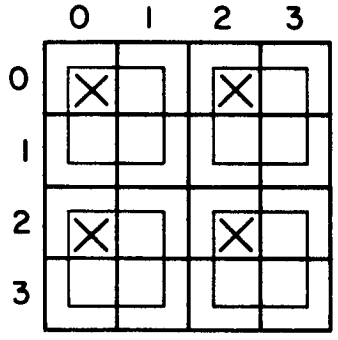


Fig. A: K = 4

Case II: The number of inputs K is of the form K = 4i + 1, i positive integer. $M^*(2) = \frac{K(K-1)}{4}$. To achieve a code with M*(2) code words the only unused squares can be chosen as successive "knight's moves" (see Fig. B), i.e., if $A_{i,j}$ denotes the jth square in the ith row, then the unused squares are given by

$$A_{i,j} \qquad j = 0, 1, 2, \ldots, K-1$$

$$i = 2j \bmod K$$

The square which is the upper left-hand corner of each block of four squares

(the squares containing "x" in Fig. B) is given by

$$A_{i,j} \qquad i = 0, 1, 2, \ldots, K-1$$

$$j = \left(\frac{i}{2} + 2\ell - 1 + \frac{K}{2}(i \bmod 2)\right) \bmod K, \quad \ell = 1, 2, \ldots, \frac{K-1}{4}$$

A set of suitable code words is then

$$a_i a_j \qquad i = 0, 1, 2, \ldots, K-1$$

$$j = \left(\frac{i}{2} + 2\ell - 1 + \frac{K}{2}(i \bmod 2)\right) \bmod K, \quad \ell = 1, 2, \ldots, \frac{K-1}{4}$$

For example, for K = 9, a set of 18 code words formed by this rule is shown in Fig. B.

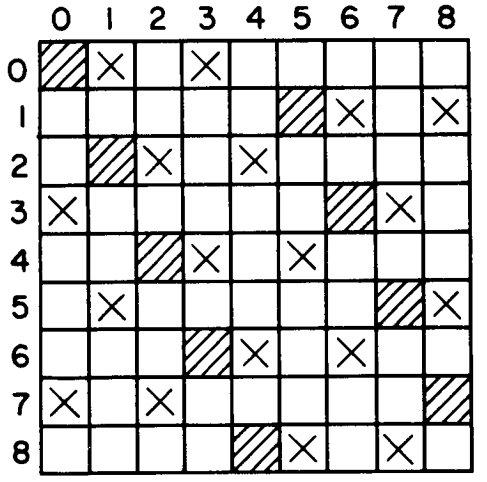


Fig. B: K = 9

Case III: The number of inputs K is of the form K = 4i-1, i positive integer, $M^*(2) = \left[\frac{K(K-1)}{4}\right]$. Most of the unused squares can be selected as successive knight's moves. More precisely, if $A_{i,j}$ denotes the jth square in the ith row, then the unused squares may be chosen by

$$A_{0,j} \qquad j = \left(\frac{K-1}{2}\right)\ell \qquad \ell = 0, 1, 2.$$

$$A_{i,j} \qquad i = 1, 2, \ldots, K-1$$

$$j = \frac{i}{2} + \frac{K-2}{2}(i \bmod 2)$$

The square which is the upper left-hand corner of each block of four squares is given by

$$A_{i,j} \qquad i = 0, 1, 2, \ldots, K-1$$

$$j = \left(\frac{i}{2} + 2\ell - 1 + \frac{K-2}{2}(i \bmod 2)\right) \bmod K,$$

$$\ell = 1,\ 2,\ \ldots, \left(\frac{K-3}{4} + i \bmod 2\right)$$

A set of code words which achieves $M^*(2)$ with $P_e = 0$ is

$$a_i a_j \qquad i = 0, 1, 2, \ldots, K-1$$

$$j = \left(\frac{i}{2} + 2\ell - 1 + \frac{K-2}{2}(i \bmod 2)\right) \bmod K,$$

$$\ell = 1,\ 2,\ \ldots, \left(\frac{K-3}{4} + i \bmod 2\right)$$

An example of this case (consider K=11) is shown in Fig. C.

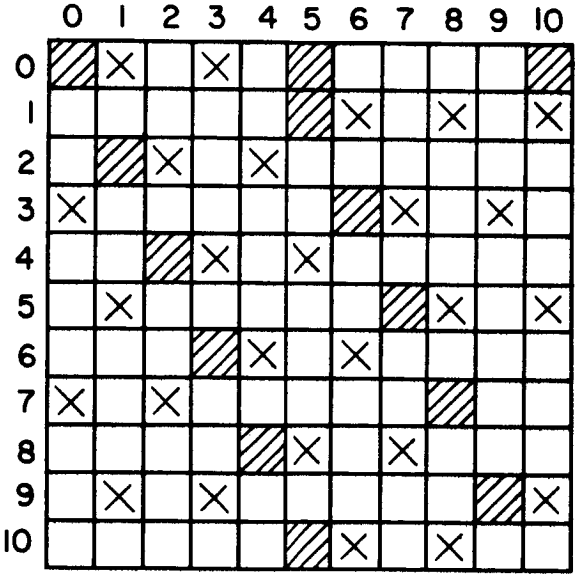


Fig. C: K = 11

Acknowledgment:

I wish to express my gratitude to Prof. Elwyn Berlekamp, without whose guidance and encouragement this paper would not have been written.

Reference

1. C. E. Shannon, "Zero-Error Capacity of Noisy Channels," IRE Trans IT-2, p. 8, (1956).

ON THE NONPARAMETRIC ESTIMATION OF SHIFT IN THE TWO-SAMPLE PROBLEM*

Terrence Fine

I. INTRODUCTION

The problem discussed in this note arises from the situation in which we are presented with two sets of samples, one from each of two independent populations of independent elements. We assume that one population has been shifted by an amount Δ with respect to the other and wish to estimate Δ without assuming more than that the population distribution possesses a density. This problem might arise, for example, in designing a highly adaptive radar system; observations made without a transmitted pulse would establish the environmental background and could be compared with observations made following a transmitted pulse.

More precisely, we are given a sample $X_1, \ldots, X_n$ of observations that are independent and identically distributed (i.i.d.) as $F(x)$ and a second sample of observations $Y_1, \ldots, Y_m$ that are i.i.d. as $G(x)$ and independent of the first sample. For some, initially unknown, shift Δ, $F(x + \Delta) \equiv G(x) \in \mathcal{G}$, where $\mathcal{G}$ is the class of all absolutely continuous distributions and G is otherwise unspecified.

We shall present some results concerning a class of estimators $\overline{\Delta}_\psi$ for Δ and discuss three members of this class having, in some measure, the property of robustness with respect to $\mathcal{G}$. The basis of the criterion we employ to determine robustness is the asymptotic relative efficiency of an estimator $\overline{\Delta}_\psi$ with respect to another estimator $\overline{\Delta}_\phi$, for a given distribution $G \in \mathcal{G}$, defined by

$$e_{\overline{\Delta}_\psi, \overline{\Delta}_\phi}(G) \equiv \lim_{N \to \infty} \frac{\operatorname{var} \overline{\Delta}_\phi}{\operatorname{var} \overline{\Delta}_\psi}, \tag{1}$$

* The research herein was supported by the Adolph C. and Mary Sprague Miller Institute.

where $N = m + n$ and $n/N \to \lambda$ $(0 < \lambda < 1)$. It is also possible to analogously define the asymptotic absolute efficiency $e_{\bar{\Delta}_\psi}(G)$ by comparison of $\bar{\Delta}_\psi$ with the best translation invariant estimator based upon knowledge of G. In our usage an estimator $\bar{\Delta}_\psi$ is said to be robust relative to Δ_ϕ if $\inf_G e_{\bar{\Delta}_\psi, \bar{\Delta}_\phi}(G)$ is not much less than unity, and it is absolutely robust if $\inf_G e_{\bar{\Delta}_\psi}(G)$ is not much less than unity.

It is known that for any finite $N (=m + n)$ there does not exist a uniformly minimum variance (UMV) estimator of Δ. If $N \to \infty$ then a UMV estimator of Δ exists based upon the estimation of the underlying distribution G and is discussed by Hajek. However, Hajek's most robust estimator is robust only for very large samples and is known to be sometimes inferior to the normal scores or Hodges-Lehmann estimators for moderately large N (Huber). These results provide some indication of the nature of the problem and the results to be expected.

II. CHOICE OF A CLASS OF ESTIMATORS

The consideration of possible estimators is arbitrarily restricted to functionals of the individual sample and combined samples empirical distributions. If we define the unit step function U(x) by

$$U(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0, \end{cases} \tag{2}$$

with the property that

$$\int_{-\varepsilon}^{\varepsilon} U(x)dU(x) \equiv 1, \tag{3}$$

then the empirical distributions are defined by

$$F_n(x) \equiv \frac{1}{n} \sum_{i=1}^{n} U(x - X_i),$$

$$G_m(x) \equiv \frac{1}{m} \sum_{i=1}^{m} U(x - Y_i),$$

and

$$H_N(x) \equiv \lambda_N F_n(x + \bar{\Delta}_\psi) + (1 - \lambda_N) G_m(x), \tag{4}$$

where

$$\lambda_N \equiv \frac{n}{N} \quad \text{and} \quad N = m + n.$$

The class of estimators we have decided upon is given by

$$V_N \equiv \int_{-\infty}^{\infty} [F_n(x + \bar{\Delta}_\psi) - G_m(x)] \psi[H_N(x)] dF_n(x + \bar{\Delta}_\psi) \approx 0, \tag{5}$$

where ψ is some nonnegative function of H_N. The choice of V_N is motivated below.

Our initial approach was to adopt a metric in the space of empirical distribution functions and use that shift $\bar{\Delta}_\psi$ which brought the shifted empirical distribution $F_n(x + \bar{\Delta}_\psi)$ as close as possible, in the sense of the metric, to the unshifted empirical distribution $G_m(x)$. The metrics that were considered were those of

Kolmogorov

$$\sup_x \left| F_n(x + \bar{\Delta}) - G_m(x) \right|,$$

Renyi

$$\sup_{0 < a < H_N < 1-a} \frac{\left| F_n(x + \bar{\Delta}) - G_m(x) \right|}{H_N(x)},$$

and Cramer and von Mises

$$\int_{-\infty}^{\infty} [F_n(x + \bar{\Delta}) - G_m(x)]^2 \psi(H_N) dH_N.$$

The metrics of Kolmogorov and Renyi were discarded after a brief and unencouraging examination. An informal analysis of the Cramer and von Mises metric indicated that it did not lead to performance superior to that of the metric W_N^2 given by

$$W_N^2 \equiv \int_{-\infty}^{\infty} [F_n(x + \bar{\Delta}) - G_m(x)]^2 \psi(H_N) dx,$$

and thus it too was discarded in favor of W_N^2.

If we adopt W_N^2, then we are to select $\bar{\Delta}$ so as to minimize W_N^2. Inspection of W_N^2 indicates that W_N^2 is a piecewise linear function of $\bar{\Delta}$ with at most nm vertices corresponding to shift values of $-Y_j + X_i$. Therefore, W_N^2 attains a minimum at a vertex where its derivative with respect to $\bar{\Delta}$ may not be defined. However, if we were to formally differentiate W_N^2 with respect to $\bar{\Delta}$ and attempt to set the result equal to zero we would obtain something like

$$2 \int_{-\infty}^{\infty} [F_n(x + \bar{\Delta}) - G_m(x)] \psi(H_N) dF_n(x + \bar{\Delta})$$

$$+ \lambda_N \int_{-\infty}^{\infty} (F_n - G_m)^2 \psi'(H_N) dF_n \approx 0.$$

Heuristically, we might expect that for large N the second term is negligible in comparison with the first, and this leads us to consider the first term, V_N, alone. We note, though, that V_N is a piecewise constant function of $\bar{\Delta}$, or as we now write $\bar{\Delta}_\psi$, and we may not be able to find a $\bar{\Delta}_\psi$ such that $V_N = 0$. Thus we adopt that $\bar{\Delta}_\psi$ (or one of those $\bar{\Delta}_\psi$, if it is not unique) that brings V_N closest to zero. This choice of V_N can now be discussed without further reference to its origin.

The functional V_N can also be expressed in terms of the ranks S_i of the ordered X_i ($X_1 \le \ldots \le X_n$) in the total ordered sample of $X_1, \ldots, X_n$ and $Y_1 + \overline{\Delta}_\psi, \ldots, Y_m + \overline{\Delta}_\psi$. We employ the observations that

$$F_n(X_i) = \frac{i}{n}, \qquad G_m(X_i - \overline{\Delta}_\psi) = \frac{S_i - i}{m},$$

and

$$H_N(X_i) = i\left(\frac{\lambda_N}{n} - \frac{1 - \lambda_N}{m}\right) + \frac{1 - \lambda_N}{m} S_i, \tag{6}$$

to conclude that

$$V_N = \frac{1}{n} \sum_{i=1}^{n} \left[i\left(\frac{1}{n} + \frac{1}{m}\right) - \frac{S_i}{m} \right] \psi\left(\frac{S_i}{N}\right). \tag{7}$$

From (7) it's easy to see that V_N is, as claimed, piecewise continuous in $\overline{\Delta}_\psi$; the ranks S_i, which are implicit functions of $\overline{\Delta}_\psi$, remain constant for variations of $\overline{\Delta}_\psi$ that don't include the vertices $-Y_j + X_i$.

III. ASYMPTOTIC ANALYSIS OF V_N

We state only the results concerning the distribution of $\overline{\Delta}_\psi$ (proofs may be obtained from the author).

Lemma 1. A sufficient condition for $V_N(\overline{\Delta}_\psi)$ to be a monotonically nondecreasing, piecewise constant function of $\overline{\Delta}_\psi$ is that for $0 \le t \le 1$

$$\psi(t) \ge \begin{cases} -t\psi'(t) & \text{for } \psi' < 0 \\ (1-t)\psi'(t) & \text{for } \psi' > 0. \end{cases} \tag{8}$$

Lemma 2. If $\psi(t)$ satisfies (8), then the distribution of $\overline{\Delta}_\psi$ defined by

$$\bar{\Delta}_\psi = \sup \{\bar{\Delta} : V_N(\bar{\Delta}) < 0\} \tag{9}$$

is given by

$$P(\bar{\Delta}_\psi < \bar{\Delta}) = P(V_N(\bar{\Delta}) \geq 0). \tag{10}$$

Lemma 3. If $F_n(x)$ is the empirical distribution for $X_1, \ldots, X_n$ i.i.d. as $G(x - \Delta)$, $G_m(x)$ is the empirical distribution for $Y_1, \ldots, Y_m$ i.i.d. as $G(x)$, $G \in \mathcal{G}$, the two samples are independent, $\lim_{N\to\infty} \frac{n}{N} = \lambda (0 < \lambda < 1)$,

$$(\exists K)\forall(0 \leq x,\ y \leq 1)\ni |\psi(x) - \psi(y)| \leq \frac{K(x - y)}{xy(1-x)(1-y)}, \tag{11}$$

$\exists(x_0 \neq 0,\ 1)\ni |\psi(x_0)| < \infty$, V_N is defined by (5), and

$$T_N \equiv \int_{-\infty}^{\infty} [F(x + \bar{\Delta}_\psi) - G(x)]\psi[\lambda F(x + \bar{\Delta}_\psi) + (1 - \lambda)G(x)]dF(x + \bar{\Delta}_\psi)$$

$$+ \int_{-\infty}^{\infty} [F_n(x + \Delta) - G_m(x)]\psi(G)dG, \tag{12}$$

then $\bar{\Delta}_\psi \longrightarrow \Delta$ as $N \to \infty \Rightarrow |V_N - T_N| = o_p(N^{-1/2})$.

Lemma 4. If ψ satisfies (8) and (11) then

$$P(\sqrt{N}(\bar{\Delta}_\psi - \Delta) < \bar{\Delta}) \sim P(T_N(\Delta + \frac{\bar{\Delta}}{\sqrt{N}}) \geq 0). \tag{13}$$

Lemma 5. If ψ satisfies (11) then $\int_{-\infty}^{\infty} [F_n(x + \Delta) - G_m(x)]\psi(G)dG$ is asymptotically normally distributed with zero mean and variance σ_ψ^2,

given by

$$\sigma_\psi^2 = [N\lambda_N(1 - \lambda_N)]^{-1}\int_0^1\int [\min(t, S) - tS]\psi(t)\psi(S)dtdS. \tag{14}$$

Theorem 1. If ψ satisfies (8) and (11) then

$$P(\sqrt{N}(\bar{\Delta}_\psi - \Delta) < \bar{\Delta}) \sim \operatorname{erfc}\left[\frac{-m(\Delta + \frac{\bar{\Delta}}{\sqrt{N}})}{\sigma_\psi}\right], \tag{15}$$

where $\operatorname{erfc} x = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-y^2/2}dy$, σ_ψ^2 is given by (14) and

$$m\ (\bar{\Delta}) = \int_{-\infty}^{\infty} [F(x) - G(x - \bar{\Delta})]\psi[\lambda F(x) + (1 - \lambda)G(x - \bar{\Delta})]dF. \tag{16}$$

Corollary 1. If the conditions of Theorem 1 are satisfied and, in addition, for small $\bar{\Delta} - \Delta$

$$m\ (\bar{\Delta}) = (\bar{\Delta} - \Delta)\int_{-\infty}^{\infty} F'\psi(F)dF + o(\bar{\Delta} - \Delta), \tag{17}$$

then $\bar{\Delta}_\psi$ is asymptotically normal with mean Δ and variance $\sigma_{\bar{\Delta}_\psi}^2$ given by

$$\sigma_{\bar{\Delta}_\psi}^2 = [N\lambda_N(1 - \lambda_N)]^{-1}\left[\int_{-\infty}^{\infty} F'\psi(F)dF\right]^{-2}$$

$$\times\int_0^1\int [\min(t, S) - tS]\ \psi(t)\psi(S)dtdS. \tag{18}$$

Theorem 1 and Corollary 1 provide the desired asymptotic analysis of $\bar{\Delta}_\psi$ for a wide variety of ψ.

IV. COMPARISON BETWEEN $\overline{\Delta}_\psi$ AND Δ^* OF HODGES-LEHMANN AND Δ_N OF NORMAL SCORES

The estimator Δ^* defined by

$$\Delta^* = \text{med}\,(X_j - Y_i) \qquad (j = 1, \ldots, n;\ i = 1, \ldots, m) \tag{19}$$

has been proposed and examined by Hodges and Lehmann. The asymptotic variance of Δ^* is known to be given by

$$\text{var}\,\Delta^* = [12\lambda(1-\lambda)N]^{-1}\left[\int_{-\infty}^{\infty} F'\,dF\right]^{-2}. \tag{20}$$

Lemma 6. If ψ is identically constant then it satisfies (8) and (11) and if $Z_K \equiv X_i - Y_j$ $(K = 1, \ldots, nm)$, $Z_K \le Z_{K+1}$ then

$$\overline{\Delta}_\psi = \begin{cases} Z_{\frac{nm}{2}(1-\frac{1}{n})} & \text{if } \frac{nm}{2}(1-\frac{1}{n}) \text{ is an integer} \\ Z_{\frac{nm}{2}(1-\frac{1}{n})+\frac{1}{2}} & \text{if } \frac{nm}{2}(1-\frac{1}{n}) \text{ is not an integer.} \end{cases} \tag{21}$$

Lemma 6 establishes that $\overline{\Delta}_\psi$ and Δ^* are almost identical estimators if ψ is constant.

Theorem 2. For ψ satisfying (8) and (11), and $F \in \mathcal{G}$, $\sup_\psi \inf_F e_{\overline{\Delta}_\psi, \Delta^*}(F) = 1$, and the supremum is attained for ψ identically constant. Theorem 2 establishes that in our class of estimators there is no estimator uniformly better than Δ^*. Nonetheless, there exist estimators occasionally slightly poorer than Δ^* but often much better, and we consider two such in what follows.

The normal scores estimators Δ_N can be implicitly defined in terms of the statistic $h(S)$ defined by

$$h(S) = E\left[w^{(S)}\right], \tag{22}$$

where $w^{(S)}$ is the S-th largest of N i.i.d. as N(0, 1) random variables. If $S_1, \ldots, S_n$ are the ranks of $X_1, \ldots, X_n$ in the combined sample of $X_1, \ldots, X_n$ and $Y_1 + \Delta_N, \ldots, Y_m + \Delta_N$, then the estimation procedure is to choose Δ_N so that $\sum_{i=1}^{n} h(S_i)$ is as close to zero as possible. The asymptotic variance of Δ_N is known to be given by (Hodges and Lehmann (1960))

$$\text{var } \Delta_N = [\lambda(1 - \lambda)N]^{-1}\left[\int_{-\infty}^{\infty} \frac{F'dF}{\phi[\Phi^{-1}(F)]}\right]^{-2}, \tag{23}$$

where ϕ is the N(0, 1) density and Φ^{-1} is the inverse of the N(0, 1) distribution.

Theorem 3. For $F \in \mathcal{G}$, ψ satisfying (8) and (11) and (18) valid,

$$\sup_{\psi} \inf_{F} e_{\Delta_\psi, \Delta_N}(F) = 1, \tag{24}$$

and the supremum is attained for $\psi(t) = 1/(\phi[\Phi^{-1}(t)])$. Theorem 3 establishes that there is no estimator in our class uniformly better than Δ_N, but there is one just as good.

V. A PARTICULAR CHOICE OF $\bar{\Delta}_\psi$

We single out for consideration the particular $\bar{\Delta}_\psi$ corresponding to

$$\psi(t) \equiv \begin{cases} \frac{1}{t} & \text{for } 0 \le t \le \epsilon \\ \frac{1}{\epsilon} & \text{for } \epsilon \le t \le 1 - \epsilon \\ \frac{1}{1-t} & \text{for } 1 - \epsilon \le t \le 1. \end{cases} \tag{25}$$

This choice of ψ satisfies (8) and (11).

Lemma 7. If the expansion (18) is valid, $F \in \mathcal{G}$ and ψ is given by (25) then

$$\text{var } \bar{\Delta}_\psi = [12N\lambda(1-\lambda)]^{-1}[\epsilon^{-2} + 16\epsilon + 12\ell n(1-\epsilon)] \left[\int_{-\infty}^{\infty} F'\psi(F)dF\right]^{-2}. \tag{26}$$

Theorem 4. Under the above conditions,

$$\begin{aligned} &\inf_F e_{\bar{\Delta}_\psi, \Delta^*}(F) = 1 - o(\epsilon), \quad \sup_F e_{\bar{\Delta}_\psi, \Delta^*}(F) = \infty, \\ &\inf_F e_{\bar{\Delta}_\psi, \Delta_N}(F) = o(\epsilon), \text{ and } \sup_F e_{\bar{\Delta}_\psi \Delta_N}(F) = \infty. \end{aligned} \tag{27}$$

Theorem 4 indicates that for small ϵ $\bar{\Delta}_\psi$ is never much worse than Δ^* and sometimes infinitely better. However, we also see that $\bar{\Delta}_\psi$ is occasionally very much worse than Δ_N.

We conclude that within the class of estimators we have developed there exist members performing as well as the Hodges-Lehmann or normal scores estimators but none that are uniformly better.

ACKNOWLEDGMENT

The author is indebted to Professor E. Lehmann of the Statistics Department, University of California, Berkeley, for introducing him to this area of research.

REFERENCES

J. Hajek, "Asymptotically most powerful rank order tests," Ann. Math. Statistics, Vol. 33, pp. 1124-1147; 1962.

J. L. Hodges, Jr. and E. L. Lehmann, "Comparison of the normal scores and Wilcoxon tests," Proc. Fourth Berkeley Symposium, Vol. 1, pp. 307-317; 1960.

J. L. Hodges, Jr. and E. L. Lehmann, "Estimates of location based on rank tests," Ann. Math. Statistics, Vol. 34, pp. 598-611; 1963.

P. J. Huber, "Robust estimation of a location parameter," Ann. Math. Statistics, Vol. 35, pp. 73-101; 1964.

ON THE EQUIVALENCE OF FINITE-STATE SEQUENTIAL MACHINE MODELS

Oscar H. Ibarra

Introduction

The equivalence of sequential machine models proposed by Moore[1] and Mealy[2] referred to here as Moore machine and Mealy machine has been shown by Cadden[3], Gill[4], and Hartmanis[5]. The purpose of this communication is to give a new proof of the equivalence and to present effective procedures for transforming one model to the other which preserves machine minimality. A new upper bound on the number of states of a Moore machine equivalent to a Mealy machine is given and is shown to be the best possible bound.

Notations, Terminology and Definitions

Let $\Sigma(\Delta)$ be a finite non-empty set of input (output) symbols. $\Sigma^*(\Delta^*)$ denotes the free monoid generated by $\Sigma(\Delta)$ with identity Λ (called the null sequence) under the operation of concatenation. $\Sigma^+(\Delta^+)$ is $\Sigma^* - \{\Lambda\}$ $(\Delta^* - \{\Lambda\})$. If $x = \sigma_0\sigma_1 \ldots \sigma_{m-1}$ (σ_i in Σ) and $0 \leq i \leq j \leq m$, then ${}_ix_j = \sigma_i\sigma_{i+1}\ldots\sigma_{j-1}$ (${}_ix_i = \Lambda$). We call x a tape and ${}_ix_j$ a subtape of x. If x is a tape in Σ^*, $\overbrace{xx\ldots x}^{k}$ is written as x^k ($\Lambda^k = \Lambda$).

Definition 1. A Mealy machine is a system $\mathcal{M} = \langle \Sigma, S, \Delta, f, g \rangle$, where

$\Sigma = \{\sigma_1, \ldots, \sigma_p\}$ is a finite non-empty set of input symbols called the input alphabet,

$\Delta = \{\delta_1, \ldots, \delta_q\}$ is a finite non-empty set of output symbols called the output alphabet,

$S = \{s_1, \ldots, s_n\}$ is a finite non-empty set of internal states called the state set,

f is a function (called the next state or transition function) which maps $S \times \Sigma$ into S; and

g is a function (called the output function) which maps $S \times \Sigma$ into Δ.

This work was supported wholly by the Joint Services Electronics Program (U.S. Army, U.S. Navy and U.S. Air Force) under Grant No. AF-AFOSR-139-64.

A Mealy machine with n states, p input symbols, and q output symbols is referred to as an (n, p, q) Mealy machine. Such a machine can be described by a transition diagram with n vertices, drawn as small circles, and oriented branches, drawn as lines between pairs of vertices with arrow signs pointing from one vertex to the other. Each vertex represents a state in S and is labeled the state it represents. A branch leaving state s_i and terminating in state s_k is labeled $(\sigma_{i1}; \delta_{i1}) \vee (\sigma_{i2}; \delta_{i2}) \vee \ldots \vee (\sigma_{im}; \delta_{im})$ if $f(s_i, \sigma_{ij}) = s_k$ and $g(s_i, \sigma_{ij}) = \delta_{ij}$ (for $j = 1, 2, \ldots, m$). $(\sigma_{ij}; \delta_{ij})$ $(j = 1, 2, \ldots m)$ is called an input-output pair. Clearly, the branches originating from any state are labeled with the total number of p input-output pairs. Figure 1 is an example of a (4, 2, 2) Mealy machine.

The functions f and g can be extended to mappings, $f: S \times \Sigma^* \rightarrow S$ and $g: S \times \Sigma^* \rightarrow \Delta^*$ by the following inductive definitions:

$$f(s, \Lambda) = s, \quad f(s, x\sigma) = f(f(s, x), \sigma) \quad (s \text{ in } S, x \text{ in } \Sigma^*, \sigma \text{ in } \Sigma)$$

$$g(s, \Lambda) = \Lambda, \quad g(s, x\sigma) = g(s, x)g(f(s, x), \sigma)$$

The extended functions satisfy the identities:

$$f(s, xy) = f(f(s, x), y)$$

$$g(s, xy) = g(s, x)g(f(s, x), y) \qquad (s \text{ in } S, \ x, y \text{ in } \Sigma^*).$$

For s in S, $x = {}_0x_m$ in Σ^*, we may write:

$$g(s, x) = g(s, {}_0x_m) = g(s, {}_0x_1)g(f(s, {}_0x_1), {}_1x_2) \ldots$$

$$g(f(s, {}_0x_{m-1}), {}_{m-1}x_m).$$

Definition 2. A Moore machine is a system $\mathscr{L} = \langle \Sigma, T, \Delta, M, N \rangle$, where Σ is the input alphabet, Δ is the output

alphabet, T is the state set, M is the transition function and N is the output function which maps T onto B.

As in the Mealy model, a Moore machine with n states, p inputs, and q outputs, briefly an (n, p, q) Moore machine can be represented by a transition diagram. This time, each vertex is labeled t; δ, where t is in T and δ = N(t). A branch labeled $\sigma_1 \vee \sigma_2 \vee \ldots \vee \sigma_m$ leaves vertex t;δ and terminates in vertex t';δ' if $M(t, \sigma_j) = t'$ (for j = 1, 2, . . . , m) and N(t') = δ'. A (4, 2, 2) Moore machine is shown in Fig. 2.

The function M can be extended to a mapping from T x Σ^* into T in the same way as was done for f.

Definition 3. Let $\mathcal{L} = < \Sigma, T, \Delta, M, N >$ be a Moore machine. For t in T, for any x in Σ^*, let $h_t(x) = N(M(t, x))$. Note that for x, y in Σ^*, $h_t(xy) = N(M(t, xy)) = N(M(M(t, x), y)) = h_{M(t, x)}(y)$. Define a function $\hat{h}_t$ which will map Σ^* into Δ^* by $\hat{h}_t(x) = \hat{h}_t({}_0x_m) = h_t({}_0x_1)\, h_t({}_0x_2) \ldots h_t({}_0x_m)$. We have for any σ in Σ and any y in Σ^+ $\hat{h}_t(\sigma y) = h_t(\sigma)\, \hat{h}_{M(t, \sigma)}(y)$.

Definition 4. Let R be an equivalence relation on the set of states S of any sequential machine. R is a right congruence relation if and only if (s, s') in R implies that for all σ in Σ, (f(s, σ), f(s', σ) is in R.

Definition 5. In a Mealy machine, two states s, s' in S are equivalent (denoted by $s \equiv s'$) if and only if for all x in Σ^+, g(s, x) = g(s', x). (Note that by definition 1, $g(s, \Lambda) = \Lambda$ for all s in S). If there exists an x in Σ^+ such that $g(s, x) \neq g(s', x)$, we say that s and s' are distinguishable (denoted by $s \not\equiv s'$) and

x is called a distinguishing sequence for s and s'. In a Moore machine, two states t, t' in T are equivalent ($t \equiv t'$) if and only if for all x in Σ^*, $\hat{h}_t(x) = \hat{h}_{t'}(x)$. If $\hat{h}_t(x) \neq \hat{h}_t(x)$ for some x in Σ^* (possibly Λ), t and t' are distinguishable ($t \not\equiv t'$) and x is a distinguishing sequence for t and t'. Equivalence of states is a right congruence relation.

Definition 6. A machine is minimal if no two states are equivalent. Machine minimization techniques can be found in Moore[1] and Gill[6].

Definition 7. Let $\mathcal{M} = <\Sigma, S, \Delta, f, g>$ be a Mealy machine. Write $g_s(x) = g(s,x)$ for any x in Σ^*. Let $F(\mathcal{M}) = \{ g_s(\ ) \mid s \text{ in } S \}$. $F(\mathcal{M})$ is the set of all functions $\mathcal{M}$ can compute starting from any state s in S and is called the behavior of $\mathcal{M}$. Two Mealy machines $\mathcal{M}_1$ and $\mathcal{M}_2$ are equivalent if and only if $F(\mathcal{M}_1) = F(\mathcal{M}_2)$.

Let $\mathcal{L} = <\Sigma, T, \Delta, M, N>$ be a Moore machine. We know that $\hat{h}_t$ maps Σ^* into Δ^*. Let $F(\mathcal{L}) = \{ \hat{h}_t(\ ) \mid t \text{ in } T \}$. $F(\mathcal{L})$ is the behavior of $\mathcal{L}$. Two Moore machines $\mathcal{L}_1$ and $\mathcal{L}_2$ are equivalent if and only if $F(\mathcal{L}_1) = F(\mathcal{L}_2)$.

In a Moore machine, $\hat{h}_t(\Lambda) = h_t(\Lambda) = N(t)$, i.e., the null sequence produces an output, whereas in a Mealy machine, $g_s(\Lambda) = g(s,\Lambda) = \Lambda$. This raises the question as to how the equivalence of the two models should be defined. The next definition gives the answer.

Definition 8. Let $\mathcal{M}$ be a Mealy machine and $\mathcal{L}$ be a Moore machine. $\mathcal{M}$ and $\mathcal{L}$ are equivalent (denoted by $\mathcal{M} \equiv \mathcal{L}$) if and only

if $F(\mathcal{M}) = F(\mathcal{L})$, where the equality is taken to mean the restriction of the functions $g_s(\ )$ in $F(\mathcal{M})$ and $\hat{h}_t(\ )$ in $F(\mathcal{L})$ to mappings from Σ^+ into Δ^+. Also, input tapes of the form $x \wedge^k y$ (x, y in Σ^+, k an integer) are not considered in comparing the two models.

EQUIVALENCE THEOREMS

Theorem 1. For every Mealy machine $\mathcal{M}$, there exists a Moore machine $\mathcal{L}$ such that $\mathcal{M} \equiv \mathcal{L}$.

Proof.

Let $\mathcal{M} = < \Sigma, S, \Delta, f, g >$ be a Mealy machine.

For every s in S, form Q_s, where

$$Q_s = \{ (s,\delta) \mid \exists\ s' \text{ in } S,\ \exists\ \sigma \text{ in } \Sigma \text{ such that } f(s', \sigma) = s \text{ and } g(s', \sigma) = \delta \}$$

Let $\mathcal{L} = < \Sigma, T, \Delta, M, N >$, where

$$T = \bigcup_{s \in S} Q_s$$

For all (s,δ) in Q_s, for any σ in Σ,

$$M((s,\delta), \sigma) = (f(s,\sigma), g(s,\sigma)) \text{ and } N((s,\delta)) = \delta.$$

M is extended to a function from $T \times \Sigma^*$ into T as follows: For (s,δ) in T, $x = {}_0x_m$ in Σ^+,

$$M((s,\delta), \wedge) = (s,\delta)$$

$$M((s,\delta), {}_0x_m) = (f(s, {}_0x_m),\ g(f(s, {}_0x_{m-1}), {}_{m-1}x_m)).$$

It follows immediately that for (s,δ) in T, $x = {}_0x_m$ in Σ^+,

$$h_{(s,\delta)}({}_0x_m) = N(M((s,\delta), {}_0x_m)) = g(f(s, {}_0x_{m-1}), {}_{m-1}x_m).$$

Now we show that $F(\mathcal{M}) = F(\mathcal{L})$. Let $g_s(\ )$ be in $F(\mathcal{M})$ and take

$\hat{h}_{(s,\ \delta)}(\ \ )$ in F ($\mathcal{L}$). Let $x = {}_0x_m$ in Σ^+.

$$g_s(x) = g_s({}_0x_m) = g(s, {}_0x_1)\, g(f(s, {}_0x_1), {}_1x_2) \ldots g(f(s, {}_0x_{m-1}), {}_{m-1}x_m)$$

$$= h_{(s,\ \delta)}({}_0x_1) h_{(s,\ \delta)}({}_0x_2) \ldots h_{(s,\ \delta)}({}_0x_m)$$

$$= \hat{h}_{(s,\ \delta)}({}_0x_m)$$

$$= \hat{h}_{(s,\ \delta)}(x) \therefore F(\mathcal{K}) \leq F(\mathcal{L}).$$

Let $\hat{h}_{(s,\ \delta)}(\ \ )$ be in F ($\mathcal{L}$) and take $g_s(\ \ )$ in F ($\mathcal{K}$). Let $x = {}_0x_m$ in Σ^+.

$$\hat{h}_{(s,\ \delta)}(x) = \hat{h}_{(s,\ \delta)}({}_0x_m) = h_{(s,\ \delta)}({}_0x_1)\, h_{(s,\ \delta)}({}_0x_2) \ldots h_{(s,\ \delta)}({}_0x_m)$$

$$= g(s, {}_0x_1)\, g(f(s, {}_0x_1), {}_1x_2) \ldots g(f(s, {}_0x_{m-1}), {}_{m-1}x_m)$$

$$= g(s, {}_0x_m)$$

$$= g(s, x) \therefore F(\mathcal{L}) \leq F(\mathcal{K}).$$

Q. E. D.

Corollary 1. If $\mathcal{K}$ has n states and q output symbols, $\mathcal{L}$ has $n' \leq nq$ states.

Remark. Theorem 1 provides an effective algorithm for transforming a Mealy machine to an equivalent Moore machine.

Theorem 2. The Moore machine $\mathcal{L}$ constructed in the proof of Theorem 1 is minimal if $\mathcal{K}$ is minimal.

Proof.

Let $(s, \delta) \equiv (s', \delta')$ in T, $(s, \delta) \neq (s', \delta')$. By definition 5, $\hat{h}_{(s,\ \delta)}(x) = \hat{h}_{(s',\ \delta')}(x)$ for all x in Σ^*. Taking $x = \Lambda$, $\hat{h}_{(s,\ \delta)}(\Lambda) = \hat{h}_{(s',\ \delta')}(\Lambda) \Longleftrightarrow N((s, \delta)) = N((s', \delta')) \Longrightarrow \delta = \delta'$. Hence, it is sufficient to show that $(s, \delta) \equiv (s', \delta)$, $s \neq s'$ leads to a contradiction.

$$\hat{h}_{(s,\ \delta)}(x) = \hat{h}_{(s',\ \delta)}(x) \quad \text{for all } x \text{ in } \Sigma^*$$

$$\Longrightarrow \quad \hat{h}_{(s,\ \delta)}(x) = \hat{h}_{(s',\ \delta)}(x) \quad \text{for all } x \text{ in } \Sigma^+$$

By Theorem 1,

$$\hat{h}_{(s,\ \delta)}(x) = g_s(x) \quad \text{for all } x \text{ in } \Sigma^+$$

and

$$\hat{h}_{(s',\ \delta)}(x) = g_{s'}(x) \quad \text{for all } x \text{ in } \Sigma^+$$

$$\Longrightarrow \quad g_s(x) = g_{s'}(x) \quad \text{for all } x \text{ in } \Sigma^+$$

$$\Longrightarrow \quad s \equiv s' \ (s \neq s')$$

$$\Longrightarrow \quad \text{contradiction since } \mathcal{M} \text{ is minimal.}$$

Q. E. D.

Example. Figure 3(a) is a minimal (3, 2, 2) Mealy machine. Carrying out the construction given in Theorem 1, the equivalent minimal Moore machine is obtained and is shown in Figure 3(b). Figure 3(c) is the same machine with states relabeled.

Theorem 3. For a minimal (n, p, q) Mealy machine $\mathcal{M}$, the bound $n' \leq nq$ (n' = the number of states of the minimal Moore machine $\mathcal{L} \equiv \mathcal{M}$) can be achieved with equality for any n if and only if $p \geq q$.

Proof.

Assume $p < q$. We will show that $n' < nq$ (in fact, $n' \leq np$). The branches leaving each state in the transition diagram of $\mathcal{M}$ are labeled with the total number of p input-output pairs. Since there are n states, there are exactly np input-output pairs. Let q_i be the number of input-output pairs with distinct output symbols terminating in state s_i, (note that q_i is the cardinality of Q_{s_i} in Theorem 1). We assert that $n' = q_1 + q_2 + \ldots + q_n \leq np$. For suppose $n' > np$. Then, since each input symbol is associated with exactly 1 output symbol, there must exist $r \geq np + 1$ input-output pairs. This is a contradiction. Hence, $n' \leq np$ and $n' < nq$.

For $p \geq q$, let $\Sigma = \{\sigma_1, \ldots, \sigma_p\}$, $\Delta = \{\delta_1, \ldots, \delta_q\}$ and let

$$E = (\sigma_3; \delta_3) \vee (\sigma_4; \delta_4) \vee \ldots \vee (\sigma_q; \delta_q) \vee (\sigma_{q+1}; \delta_q) \vee \ldots \vee (\sigma_p; \delta_q).$$

Consider the (n, p, q) Mealy machine shown in Fig. 4. It is easy to check that this machine is minimal. Since the branches terminating in each state are labeled with at least q input-output pairs with distinct output symbols. $n' = nq$.

Q. E. D.

Definition 9. Let $\mathscr{L} = < \Sigma, T, \Delta, M, N >$ be a Moore machine. Define a relation R on the state set T by:

For t, t' in T, (t, t') in R if and only if for any σ in Σ, $M(t, \sigma) = M(t', \sigma)$.

We call t and t' 1-equivalent states. R is a right congruence relation.

Proposition 1. Let $\mathscr{L} = < \Sigma, T, \Delta, M, N >$ be a Moore machine. If for all x in Σ^+, $\hat{h}_t(x) = \hat{h}_{t'}(x)$, then t and t' are 1-equivalent.

Proof.

Suppose there exists a σ in Σ such that $M(t, \sigma) \neq M(t', \sigma)$. Since $\mathscr{L}$ is minimal, there exists a distinguishing sequence y in Σ^+ for $M(t, \sigma)$ and $M(t', \sigma)$. Then $\hat{h}_{M(t, \sigma)}(y) \neq \hat{h}_{M(t', \sigma)}(y)$ and since $h_t(\sigma) = h_{t'}(\sigma)$, we have $h_t(\sigma)\, \hat{h}_{M(t, \sigma)}(y) \neq h_{t'}(\sigma)\, \hat{h}_{M(t', \sigma)}(y) \Longrightarrow$ $\hat{h}_t(\sigma y) \neq \hat{h}_{t'}(\sigma y) \Longrightarrow$ contradiction.

Theorem 4. Let $\mathscr{L} = < \Sigma, T, \Delta, M, N >$ be a Moore machine. There exists a Mealy machine $\mathscr{M}$ such that $\mathscr{M} \equiv \mathscr{L}$. Furthermore, $\mathscr{M}$ is minimal if $\mathscr{L}$ is minimal.

Proof

Let R be the relation of 1-equivalence on the state set T. Define a Mealy machine $\mathscr{M} = <\Sigma, S, \Delta, f, g, >$, where

$S = \{ [t] \mid t \text{ in } T \}$, $[t]$ denotes the R-equivalence class containing t.

$f([t], \sigma) = [t']$ and $g([t], \sigma) = h_t(\sigma)$ if $M(t, \sigma) = t'$ and $h_t(\sigma) = N(M(t, \sigma))$ (t, t' in T, σ in Σ).

Since R is a right congruence relation, the machine $\mathcal{M}$ is well defined. $F(\mathcal{L}) = \{\hat{h}_t(\ ) \mid t \text{ in } T\}$ and $F(\mathcal{M}) = \{g_{[t]} \mid [t] \text{ in } S\}$. To show that $F(\mathcal{L}) \subseteq F(\mathcal{M})$, let $\hat{h}_t(\ )$ be in $F(\mathcal{L})$ and take $g_{[t]}(\ )$ in $F(\mathcal{M})$. Let $x = {}_0x_m$ in Σ^+.

$$\hat{h}_t(x) = \hat{h}_t({}_0x_m) = h_t({}_0x_1)h_t({}_0x_2)\cdots h_t({}_0x_m)$$
$$= h_t({}_0x_1)\, h_{M(t,\, {}_0x_1)}({}_1x_2)\cdots h_{M(t,\, {}_0x_{m-1})}({}_{m-1}x_m)$$
$$= g([t],\, {}_0x_1)\, g([M(t,\, {}_0x_1)],\, {}_1x_2)\cdots g([M(t,\, {}_0x_{m-1})],\, {}_{m-1}x_m)$$
$$= g([t],\, {}_0x_1)\, g(f([t],\, {}_0x_1),\, {}_1x_2)\cdots g(f([t],\, {}_0x_{m-1}),\, {}_{m-1}x_m)$$
$$= g_{[t]}({}_0x_m)$$
$$= g_{[t]}(x) \quad \therefore \quad F(\mathcal{L}) \subseteq F(\mathcal{M}).$$

Now let $g_{[t]}(\ )$ be in $F(\mathcal{M})$ and take $\hat{h}_t(\ )$ in $F(\mathcal{L})$. Let $x = {}_0x_m$ in Σ^+.

$$g_{[t]}(x) = g_{[t]}({}_0x_m) = g([t],\, {}_0x_1)\, g(f([t],\, {}_0x_1),\, {}_1x_2)\cdots$$
$$\cdots g(f([t],\, {}_0x_{m-1}),\, {}_{m-1}x_m)$$
$$= g([t],\, {}_0x_1)\, g([M(t,\, {}_0x_1)],\, {}_1x_2)\cdots$$
$$\cdots g([M(t,\, {}_0x_{m-1})],\, {}_{m-1}x_m)$$
$$= h_t({}_0x_1)h_{M(t,\, {}_0x_1)}({}_1x_2)\cdots$$
$$\cdots h_{M(t,\, {}_0x_{m-1})}({}_{m-1}x_m)$$
$$= h_t({}_0x_1)\, h_t({}_0x_2)\cdots h_t({}_0x_m)$$
$$= \hat{h}_t({}_0x_m)$$
$$= \hat{h}_t(x) \qquad \therefore \quad F(\mathcal{M}) \subseteq F(\mathcal{L}).$$

If $\mathcal{L}$ is minimal, we assert that $\mathcal{M}$ is minimal. For suppose $[t] \equiv [t']$ in S, $[t] \neq [t']$. Then $g([t], x) = g([t'], x)$ for all x in $\Sigma^+ \Rightarrow \hat{h}_t(x) = \hat{h}_{t'}(x)$ for all x in $\Sigma^+ \Rightarrow (t, t')$ in R (by Proposition 1) $\Rightarrow$ $[t] = [t'] \Longrightarrow$ contradiction.

Q. E. D.

Remark. Theorem 4 provides an effective procedure for transforming a Moore machine to an equivalent Mealy machine.

Example. Figures 5(a) - (c) illustrate the procedure for transforming a Moore machine to an equivalent Mealy machine.

Acknowledgment

The author wishes to thank Professors Michael A. Harrison and Arthur Gill for their comments and suggestions.

References

1. E. F. Moore, "Gedanken-experiments on sequential machines," Automata Studies, edited by C. E. Shannon and J. McCarthy, Princeton University Press; 1956, pp. 129-153.
2. G. H. Mealy, "A method for synthesizing sequential circuits," Bell Systems Technical Journal, vol. 34, pp. 1045-1079; September, 1956.
3. W. J. Cadden, "Equivalent sequential circuits," IRE Transactions on Circuit Theory, vol. CT-6, pp. 30-34; March, 1959.
4. A. Gill, "Comparison of finite-state models," IRE Transactions on Circuit Theory, vol. CT-7, pp. 178-179; June, 1960.
5. J. Hartmanis, "The equivalence of sequential machine models," IEEE Transactions on Electronics Computers, vol. EC-12, pp. 18-19; February, 1963.
6. A. Gill, Introduction to the Theory of Finite-State Machines, McGraw Hill, 1962, Chapter 3.

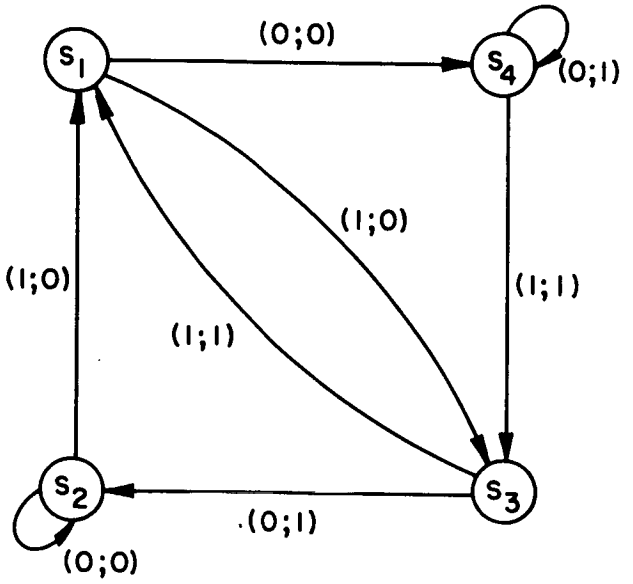


Fig. 1. A (4, 2, 2) Mealy machine.

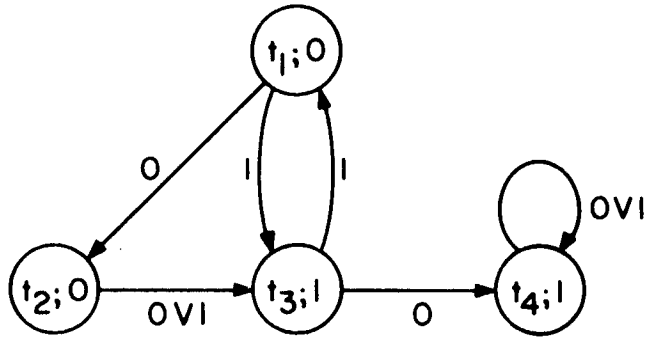


Fig. 2. A (4, 2, 2) Moore machine.

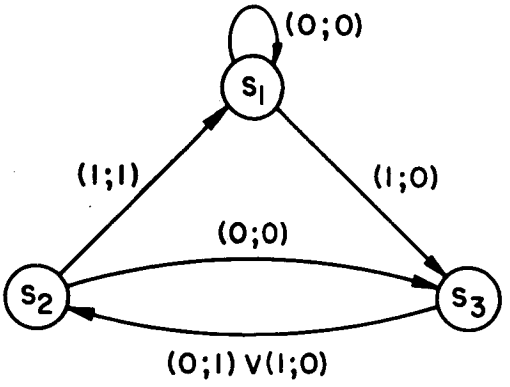


Fig. 3(a). A minimal (3, 2, 2) Mealy machine $\mathcal{A}_1$.

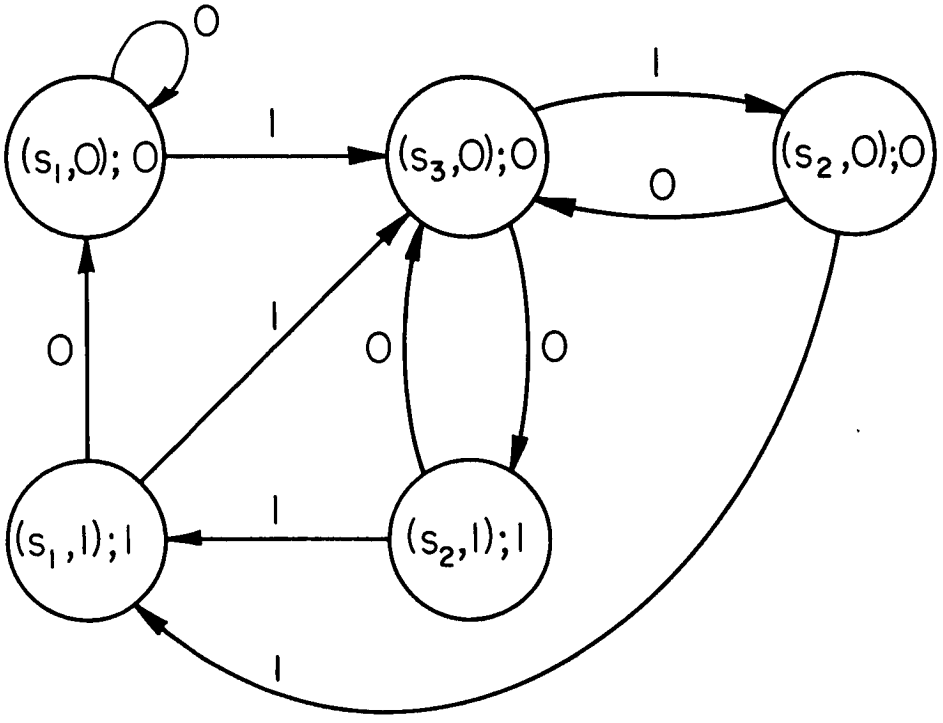


Fig. 3(b). Minimal (5, 2, 2) Moore machine $\mathscr{L}_1 \equiv \mathscr{M}_1$.

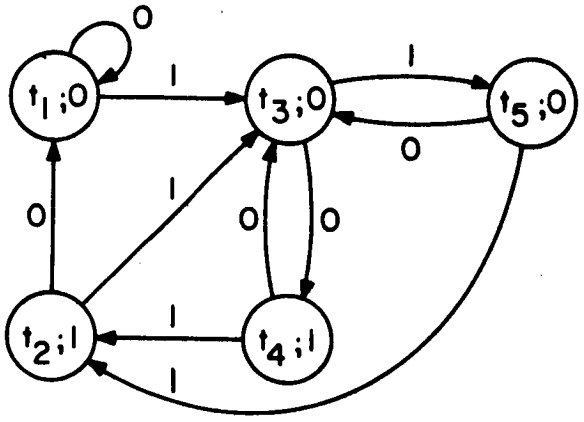


Fig. 3(c). Moore machine of Fig. 3(b) with states relabeled.

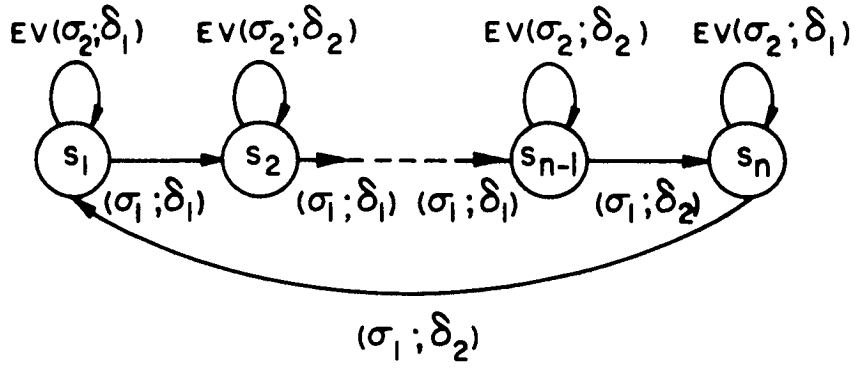


Fig. 4. An (n, p, q) Mealy machine.

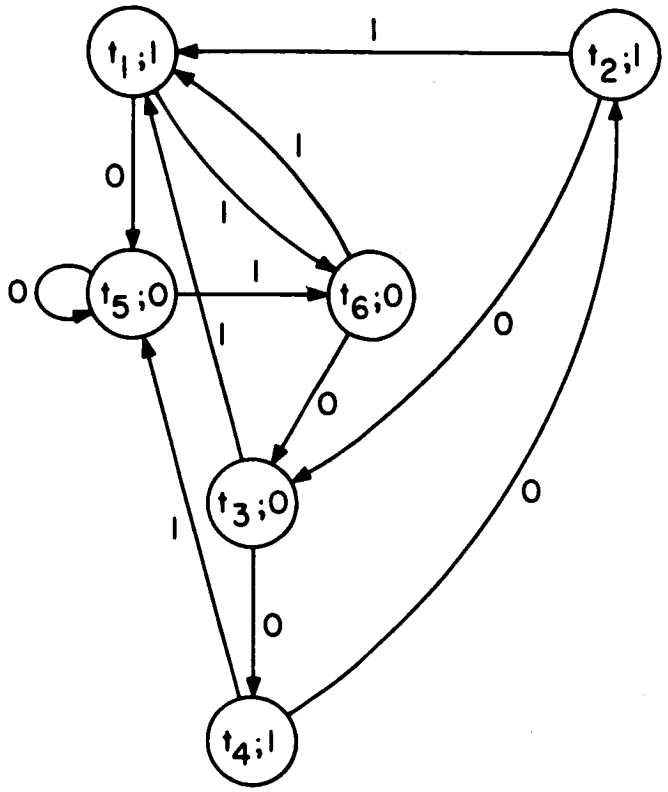


Fig. 5(a). A minimal (6, 2, 2) Moore machine $\mathscr{L}_2$.

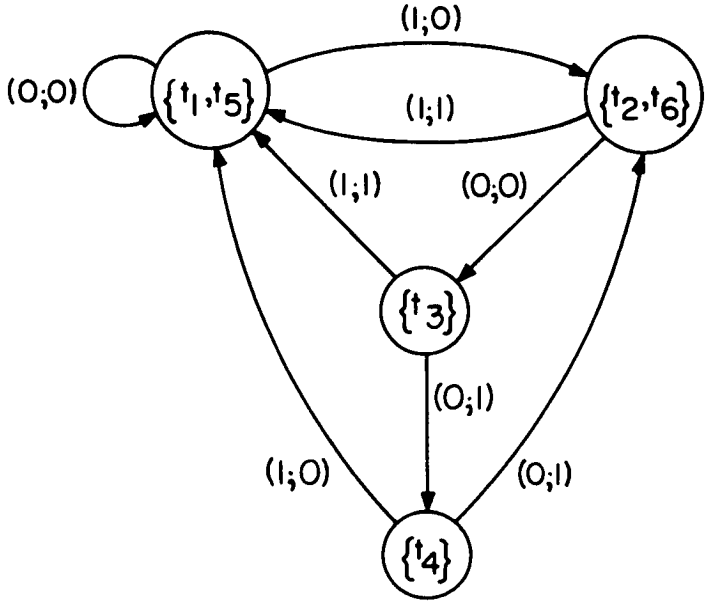


Fig. 5(b). Minimal (4, 2, 2) Mealy machine $\mathcal{A}_2 \equiv \mathcal{L}_2$ ($\{t_i, t_j\}$ is an R-equivalence class).

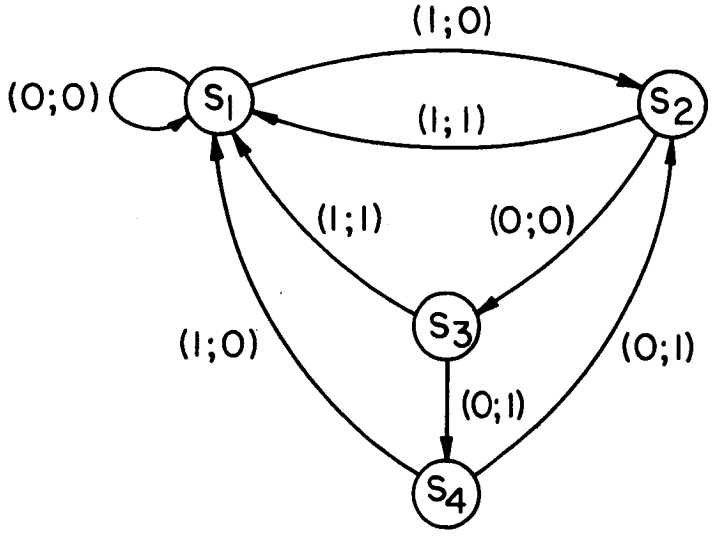


Fig. 5(c). Moore machine of Fig. 5(b) with states relabeled.

A NOTE ON THE PERMANENT OF A MATRIX

Jean-Paul Jacob

The permanent of a matrix, redefined below, is not a new concept.[1] Surprisingly, however, it seems that one of its most trivial applications, namely, the counting of non-zero additive terms in the expansion of the determinant of a matrix, has not yet been used. Our main result using this fact is presented in reference 2.

Definition: Given an $n \times n$ matrix $\underline{A} \triangleq [a_{i,j}]$, the permanent of $\underline{A}$, denoted $\left|\underline{A}\right|^{+}$ is by definition

$$\left|\underline{A}\right|^{+} \triangleq \sum_{P} a_{1,P(1)} a_{2,P(2)} \cdots a_{n,P(n)}$$

where the sum is extended over all possible permutations P of the integers $1, 2, \ldots, n$.

Hence, the permanent of a matrix $\underline{A}$ is the same sum as the determinant of $\underline{A}$ with all terms prefixed by the positive sign.

Lemma 1: Given a square matrix $\underline{A}$ of order n, the number of non-zero summands in the expansion of its determinant is equal to the permanent of $\underline{\tilde{A}}$, where $\underline{\tilde{A}}$ is obtained by substituting 1 for all non-zero entries of $\underline{A}$.

The term "summand" and the proof of lemma 1 can be found in reference 2.

---

*This work was supported wholly by the Joint Services Electronics Programs (U.S. Army, U.S. Navy and U.S. Air Force) under Grant No. AF-AFOSR-139-64.

Example 1: Given $\underline{A}$ below, how many summands are there in the expression of $|A|$?

$$\underline{A} = \begin{bmatrix} a_{11} & 0 & a_{13} \\ 0 & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

The answer is:

$$|\underline{\widehat{A}}| = \begin{vmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{vmatrix} = 1 + 1 + 1 = 3$$

One can trivially verify that, indeed,

$$|A| = a_{11}a_{22}a_{33} - a_{13}a_{22}a_{31} - a_{11}a_{23}a_{32}$$

The counting character of the permanent leads to a natural application, i.e., solving problems in combinatorial analysis. Examples 2 and 3 illustrate this

Example 2: Five gentlemen wearing hats go to a banquet. When handling their hats at the checking counter, the receptionist forgets to give them tickets. At the exit of the banquet the receptionist gives randomly one hat to each of the five gentlemen. Which is the probability p that no gentleman will get his own hat back?

Solution: Consider the 5th order matrix $\underline{A}_5$ below.

$$\underline{A}_5 = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{15} \\ a_{21} & a_{22} & \cdots & a_{25} \\ \vdots & & & \\ a_{51} & a_{52} & \cdots & a_{55} \end{bmatrix}$$

Each additive term in the expansion of $|A_5|$ is of the form:

$$a_{1,P(1)}\, a_{2,P(2)} \cdots a_{5,P(5)} \tag{1}$$

If we associate with the first subscript of $a_{i,P(i)}$ the i^{th} gentleman and with the second subscript the gentleman who gets the i^{th} gentleman's hat, there is a 1 - 1 correspondence between each term in the expansion of $|\underline{A}_5|$ and each possible way in which the receptionist can distribute the 5 hats.

On the other hand, the cases in which at least one gentleman receives his own hat correspond to terms like (1) in which there is at least one index i, $i \leq 5$, such that

$$i = P(i),$$

which implies $a_{i,P(i)}$ to be a diagonal entry. In the numerator of p (the number of "favorable cases") we do not want to count such cases.

Therefore:

$$p = \frac{\begin{vmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{vmatrix}}{\begin{vmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{vmatrix}}$$

In Theorem 1 below we will show how permanents like those above can easily be computed. Theorem 2 gives an elegant solution for the numerator of p.1

Example 3^3: In how many ways N can 8 rooks be placed on a chessboard so that none can take another and that none stands on the white diagonal?

Answer: $N = |\tilde{\underline{A}}_8 - \underline{I}_8|$

Where $\tilde{\underline{A}}_8$ is the 8×8 square matrix with all its entries equal to 1 and I_8 is the 8×8 identity matrix. N is the permanent of the 8×8 matrix in which all diagonal entries are zero and all other are 1 (see conjecture 1)

Again, the answer was obtained by observing the 1 - 1 correspondence between the additive terms in the expansion of the determinant

$$\begin{vmatrix} 0 & a_{12} & \cdots & a_{18} \\ a_{21} & 0 & & a_{28} \\ \vdots & & & \\ a_{81} & a_{82} & \cdots & 0 \end{vmatrix}$$

and the cases described in our problem.

Lemma 2: The permanent of a matrix $\underline{A}$ remains unchanged if one interchanges any two rows (any two columns) of $\underline{A}$.

The proof of lemma 2 is trivial from the definition of permanent.

Theorem 1: Let $\underline{A}$ be an $n \times n$ matrix whose entries are either "1" or "0". If there is at most 1 zero in each row and in each column of $\underline{A}$, then

$$|\underline{A}| = \sum_{k=0}^{Z} (-1)^k \binom{Z}{k} (n-k)!$$

where Z is the total number of zero entries in $\underline{A}$. (note that, from the assumptions, $Z \leq n$)

The proof of Theorem 1 goes by induction on Z. Also, from lemma 2 and the assumptions of Theorem 1, we lose no generality in assuming that all zero entries of $\underline{A}$ are in its main diagonal, namely, $a_{1,1}, a_{2,2}, \cdots, a_{Z,Z}$.

a) $Z = 0$

Then the determinant of $\underline{A}$ has as many terms as the possible permutations of n objects, i.e., n!

$$n! = \sum_{k=0}^{0} (-1)^k \binom{0}{k} (n-k)!$$

b) Assume that Theorem 1 is true for all $0 \leq Z \leq p-1$. To prove that it is true for p, notice that to obtain

$$\begin{vmatrix} 0 & a_{12} & \cdots & a_{1p} & \cdots & a_{1n} \\ a_{21} & 0 & \cdots & a_{2p} & \cdots & a_{2n} \\ \vdots & & & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & 0 & \cdots & a_{pn} \\ \vdots & & & & & \\ a_{n1} & a_{n2} & \cdots & a_{np} & \cdots & a_{nn} \end{vmatrix} \qquad (2)$$

from

$$\begin{vmatrix} 0 & a_{12} & \cdots & a_{1p} & \cdots & a_{1n} \\ a_{21} & 0 & \cdots & a_{2p} & \cdots & a_{2n} \\ \vdots & & & & & \\ a_{p1} & a_{p2} & \cdots & a_{pp} & \cdots & a_{pn} \\ \vdots & & & & & \\ a_{n1} & a_{n2} & \cdots & a_{np} & \cdots & a_{nn} \end{vmatrix} \qquad (3)$$

we have to substract from (3) the number of non-zero summands which contain a_{pp}. A little reasoning shows that this number is

$$(n-1)! - \binom{p-1}{1}(n-2)! + \binom{p-2}{2}(n-3)! + \ldots \pm (n-p)!$$

which can also be written as

$$\sum_{k=1}^{p} (-1)^{k+1} \binom{p-1}{k-1} (n-k)! \tag{4}$$

Using the induction hypothesis to compute (3) and subtracting (4) from it, our answer becomes

$$\sum_{k=0}^{p-1} (-1)^k \binom{p-1}{k} (n-k)! - \sum_{k=1}^{p} (-1)^{k+1} \binom{p-1}{k-1} (n-k)! =$$

$$= n! + \sum_{k=1}^{p-1} (-1)^k \binom{p-1}{k} (n-k)! + \sum_{k=1}^{p} (-1)^k \binom{p-1}{k-1} (n-k)! =$$

$$= n! + \sum_{k=1}^{p-1} (-1)^k (n-k)! \left[\binom{p-1}{k} + \binom{p-1}{k-1} \right] + (-1)^p (n-p)! =$$

$$= n! + \sum_{k=1}^{p-1} (-1)^k (n-k)! \binom{p}{k} + (-1)^p (n-p)! =$$

$$= \sum_{k=0}^{p} (-1)^k \binom{p}{k} (n-k)! \qquad \text{Q. E. D.}$$

A particular case of Theorem 1 is given for $Z = n$. This is the case of Example 3, for instance, in which

$$N = \frac{8!}{2!} - \frac{8!}{3!} + \frac{8!}{4!} - \frac{8!}{5!} + \frac{8!}{6!} - \frac{8!}{7!} + \frac{8!}{8!} \tag{5}$$

Considering that $8! = 40320$ and $4! = 24$, the exact value of N above cannot be computed with a slide rule. Instead, we may try to use ingenuity. Rewrite (5) as

$$N = 8! \left(\frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} + \ldots + \frac{1}{8!} \right) \tag{6}$$

Observe now that between the brackets in (6) are the 7 first non-zero terms in the expansion of e^{-1}. Since our answer N has to be a positive integer, one may suspect that the 7 terms above approximate e^{-1} well enough for N to be the closest integer to $(8!)e^{-1}$. A search of this type originated our

Conjecture 1: Let $\underline{A}$ be an $n \times n$ matrix whose entries are "1" or "0" being zero iff it is a diagonal entry.

Then $\left| \underline{A}^{+} \right|^{+} = \left[(n!)e^{-1} \right]$

where the square brackets in the right-hand side mean "closest integer."

Notice that conjecture 1 is, in fact, a theorem, but its proof would involve steps of direct verification and convergence properties of the sequence

$$\sum_{k=2}^{n} \frac{1}{k!} .$$

The answer to example 3 is then

$$N = \left[40\ 320\ e^{-1} \right] = 14\ 853$$

Final Remark: During the course of our divagations on the utilities of the concept of the permanent we found many other applications. Their importance was, however, minute or nil because easier alternative methods were known. The most sophisticated of such applications is in graph theory, where one is interested in counting the total number of trees in a graph. In the case of the graph of a network, the sum of all tree admittance products is the determinant of the nodal admittance matrix.[4] There is a 1 - 1 correspondence between each tree of the graph and each non-zero subdeterminant (of order k) of a reduced incidence matrix A_r (of order kxe) of the graph. If we compute A_r adding (e - k) rows of ones to it and compute - in a special way - the permanent of this new matrix, we will obtain the total number of trees in the graph multiplied by (e - k)! . By "special computation" we mean: compute the permanent of the complete matrix x by taking a Laplace expansion[2] along the (e - k) rows of ones and computing modulo 2 the subdeterminants of the original k rows.

ACKNOWLEDGMENT

The author acknowledges the work of his colleague, O. Ibarra, who arrived at Theorem 1 independently.

REFERENCES

1. T. Muir, A Treatise on the Theory of Determinants, (Longmans, Green & Co., N. Y., 1933).
2. J-P Jacob, "The number of terms in the general gain formulas for Coates and Mason s signal-flow graphs," Notes on System Theory, Vol. VII, page 109.
3. W. Feller, An Introduction to Probability Theory and Its Applications,

(John Wiley & Sons, N. Y., 1960), page 101, problem 11.

4. S. Seshu & M. B. Reed, Linear Graphs and Electrical Networks, (Addison - Wesley Inc., 1961), page 156.

THE NUMBER OF TERMS IN THE GENERAL GAIN FORMULAS FOR COATES AND MASON SIGNAL-FLOW-GRAPHS

Jean Paul Jacob

This communication is intended to answer questions of the sort: given a Coates flow-graph,[1, 2] or a Mason signal-flow-graph,[3, 4, 5] how many "connection paths," or "loop products," will one have to find in order to compute numerator and denominator of the general gain formula?

We initially recall the definition of the determinant of an $n \times n$ matrix $\underline{A} = [a_{i,j}]$ as being

$$|\underline{A}| = \sum_{P} (\text{sign } P)\, a_{1,P(1)}\, a_{2,P(2)} \cdots a_{n,P(n)}$$

$$= \sum_{P} (\text{sign } P) \prod_{i=1}^{n} a_{i,P(i)}$$

where the summation extends over all possible permutations $\underline{P}$ of the integers 1, 2, ..., n.

Easier than the determinant is the concept of the permanent of a matrix $\underline{A} = [a_{i,j}]$ which is, by definition,[6] the same sum as the determinant of $\underline{A}$ with all terms prefixed by the positive sign, i.e.,

$$\text{permanent of } \underline{A} \triangleq |\underline{A}|^{+} \triangleq \sum_{P}^{+} a_{1,P(1)}\, a_{2,P(2)} \cdots a_{n,P(n)}$$

$$= \sum_{P}^{+} \prod_{i=1}^{n} a_{i,P(i)}.$$

This work was supported wholly by the Joint Services Electronics Program (U.S. Army, U.S. Navy and U.S. Air Force) under Grant No. AF-AFOSR-139-64.

Example 1: Find the permanent of $\underline{A}$ below

$$\begin{vmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{vmatrix}_{+} = \begin{vmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{vmatrix}_{+} + \begin{vmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{vmatrix}_{+}$$

$$= (1 + 1) + (1 + 1 + 1) = 5.$$

Notice that we obtained the permanent above by using Laplace's expansion along row 2, with all signs taken positive. It is not difficult to verify that the Laplace expansion technique is valid for permanents.

Definition 1: The simplified connection matrix $\underline{C}$ of a Coates flow-graph G_0 [2] (graph G with source node 0 deleted) is a square matrix in which the (i, j) entry is 1 or 0, being 1 iff there is, in G_0, an edge from node j to node i: it is understood that the entry (i, i) is 1 iff there is a self-loop at node i. All nodes of G_0 are represented in $\underline{C}$.

Example 2: Find the simplified connection matrix $\underline{C}_1$ of the Coates flow-graph of Fig. 1.

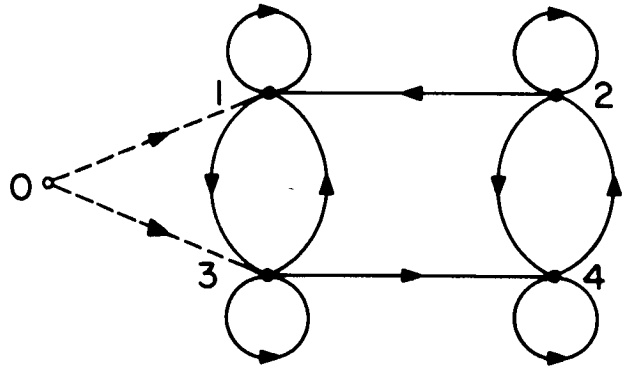


Fig. 1

$$\underline{C}_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \end{array} .$$

Note that, from the definition above, the simplified connection matrix can also be trivially derived from the conventional connection matrix.[7] Change all nonzero entries of the latter for 1.

Below we refer to "summand" as an additive term of a sum containing no parentheses. For instance, in

$$\begin{vmatrix} a+b & c \\ d & e \end{vmatrix} = (a+b)e-cd = ae+be-cd$$

there are 3 summands.

Lemma 1: Given a square matrix $\underline{A}$, the number of nonzero summands in the expansion of its determinant is equal to the permanent of $\tilde{\underline{A}}$, where $\tilde{\underline{A}}$ is obtained by substituting 1 for all nonzero entries of $\underline{A}$.

Proof: $$|\underline{A}| = \sum_{P} (\text{sign } P) \prod_{i=1}^{n} a_{i,\,P(i)} \qquad (1)$$

$$\overset{+}{|}\tilde{\underline{A}}\overset{+}{|} = \sum_{P} \prod_{i=1}^{n} \tilde{a}_{i,\,P(i)} . \qquad (2)$$

By comparing (1) and (2), we see that there is a one-to-one correspondence between the nonzero summands of $|\underline{A}|$ and the nonzero summands of $\overset{+}{|}\tilde{\underline{A}}\overset{+}{|}$. But each of the nonzero summands of $\overset{+}{|}\tilde{\underline{A}}\overset{+}{|}$ is 1,

and we are adding as many 1's as the nonzero summands of $|\underline{A}|$. Hence the lemma.

Lemma 2: The number of summands in the denominator Δ of the general gain formula for a Coates graph G is equal to the permanent of the simplified connection matrix $\underline{C}$ of G_0.

Sketch of Proof: Each additive term of Δ is called a connection gain $C(G_0)$. In a very similar way to the proof of Lemma 1 of [2], one can prove that there is a 1 - 1 correspondence between the connection gains and the nonzero terms of $|\underline{C}|$. By Lemma 1, the number of nonzero terms in the expansion of $|\underline{C}|$ is exactly $\overset{+}{|}\underline{C}\overset{+}{|}$.

Example 3: For the graph of Fig. 1, there are $\overset{+}{|}\underline{C}_1\overset{+}{|}$ connections, and from Example 1, $\overset{+}{|}\underline{C}_1\overset{+}{|} = 5$. The 5 connections of this graph are presented in Fig. 4. (page 117).

Example 4: How many connections are there in the graph of Fig. 2?

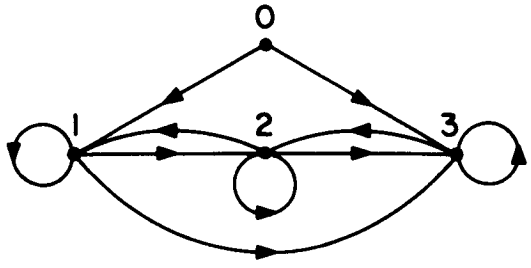


Fig. 2

$$\underline{C}_2 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \end{array} ; \quad \overset{+}{|}\underline{C}_2\overset{+}{|} = 1 + 1 + 1 + 1 = 4.$$

Figure 5 exhibits these 4 connections (page 118).

Lemma 3: The number of one-connections from source node 0 to node j in a graph G is given by $\left|\underline{C}^{j}\right|$ where $\underline{C}^{j}$ is obtained from the simplified connection matrix of G_0 when column j is substituted by a new column 0; an entry (i, 0) in column 0 is 1 iff there is an edge from node 0 to node i.

Illustration of Lemma 3: In Fig. 1 the number of one-connections from node 0 to node 3 is given by

$$\left|\underline{C}^{3}\right| = \begin{array}{c|cccc} & 1 & 2 & 0 & 4 \\ 1 & 1 & 1 & 1 & 0 \\ 2 & 0 & 1 & 0 & 1 \\ 3 & 1 & 0 & 1 & 0 \\ 4 & 0 & 1 & 0 & 1 \end{array} = \begin{vmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} + \begin{vmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{vmatrix} = 2 + 2 = 4.$$

Sketch of Proof: Once more [2] must be cited: Lemma 3 follows from Lemma 2 if one observes the 1 - 1 correspondence between the one-connections of Fig. 6(a)[2] and the connections of Fig. 6(b) of [2].

Example 5: Consider again the example of Fig. 2. How many one-connections are there from node 0 to node 3?

$$\left|\underline{C}^{3}\right| = \begin{vmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{vmatrix} = 1 + 1 + 1 + 1 = 4.$$

Figure 6 shows these 4 connections (page 119).

Definition 2: The simplified connection matrix C of a Mason signal flow-graph G_0 is a square matrix, the (i, j)-th entry of which, $i \neq j$, is 1 or 0, being 1 iff there is, in G_0, a connection

from node j to node i. The (i, i)-th entry is either 1 or 2, being 1 if node i has no self-loop. Once more the index i (or j) runs through all nodes representing variables.

Note that, from the definition above, the simplified connection matrix can be trivially derived from the conventional connection matrix.[7]

For a same graph, the off-diagonal entries of $\underline{C}$ are unique, independently of whether the graph is a Coates or a Mason flow-graph. The diagonal entries of $\underline{C}$ differ by 1, being higher for the Mason flow-graph.

Example 6: Suppose Fig. 1 represents a Mason flow-graph. Then

$$\underline{C}_3 = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 0 & 2 & 0 & 1 \\ 1 & 0 & 2 & 0 \\ 0 & 1 & 1 & 2 \end{bmatrix}.$$

Lemma 4: The total number of terms in the denominator Δ of the general gain formula for a Mason signal flow graph is equal to the permanent of its simplified connection matrix.

Illustration of Lemma 4: For the Mason graph illustrated in Fig. 1, there are

$$\left|\underline{C}_3\right|^{+}_{+} = \begin{vmatrix} 2 & 1 & 1 & 0 \\ 0 & 2 & 0 & 1 \\ 1 & 0 & 2 & 0 \\ 0 & 1 & 1 & 2 \end{vmatrix}^{+}_{+} = 2 \begin{vmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 1 & 2 \end{vmatrix}^{+}_{+} + \begin{vmatrix} 2 & 1 & 1 \\ 1 & 0 & 2 \\ 0 & 1 & 1 \end{vmatrix}^{+}_{+}$$

$$= 2(8 + 2) + (1 + 4 + 1) = 26$$

terms in the denominator of the general gain formula.

Sketch of Proof: If a system of n equations in n unknowns $x_1, x_2, \ldots, x_n$ is given in matrix form $\underline{Cx} = \underline{c}_0$ then a Mason signal flow-graph is a graphical representation of this system when rewritten as

$$\underline{Cx} = (\underline{I} - \underline{C}')\underline{x} = \underline{c}_0$$

with certain rules used to represent the equations above.

I is the n x n identity matrix. (I - C') differs from C only in that it has one more summand in all diagonal entries.

There is a 1 - 1 correspondence[5] between the summands in the expansion $|I - C'|$ and the summands of Δ. Apply Lemma 1 to the matrix (I - C') and notice that $(I - \tilde{\tilde{C}}')$ is exactly the simplified connection matrix of the Mason signal flow-graph.

Lemma 5: The number of terms contained in the expansion of Δ_k^{0j}, where Δ_k^{0j} is the Δ associated with the k-th forward path from node 0 to j, is equal to the permanent of $\underline{C}^k$, where $\underline{C}^k$ is the matrix obtained from the simplified connection matrix by deleting all rows and columns corresponding to nodes in the k-th path.

Illustration of Lemma 5: Consider the Mason signal flow-graph illustrated in Fig. 3. Its simplified connection matrix is:

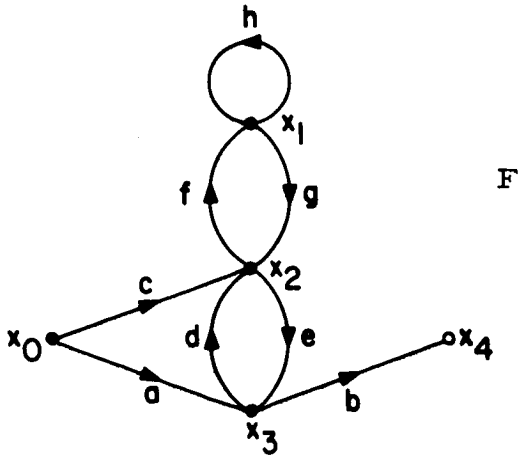


Fig. 3

$$\underline{C} = \left[\begin{array}{cccc|c} 2 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 2 \\ 0 & 1 & 1 & 0 & 3 \\ 0 & 0 & 1 & 1 & 4 \end{array}\right] .$$

We are interested in x_4/x_0. How many terms has the numerator of x_4/x_0? With path ab (which touches nodes 3 and 4) we have $\begin{vmatrix} 2 & 1 \\ 1 & 1 \end{vmatrix}^{+} = 3$ terms. With path ceb (which touches nodes 2, 3 and 4) we have $|2|^{+}_{+} = 2$ terms. The denominator (Lemma 1) has

$$\begin{vmatrix} 2 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{vmatrix}^{+} = \begin{vmatrix} 2 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{vmatrix}^{+}_{+} = 2 + 2 + 1 = 5 \text{ terms.}$$

Indeed, as one can verify, $\dfrac{x_4}{x_0} = \dfrac{ab(1 - h - fg) + ceb(1 - h)}{1 - (h + fg + dc) + hdc}$.

Proof of Lemma 5: Lemma 5 is a trivial consequence of Lemma 4 because Δ_k^{0j} is just the Δ for a graph obtained from G by deleting all nodes (and incident connections) which belong to path k.

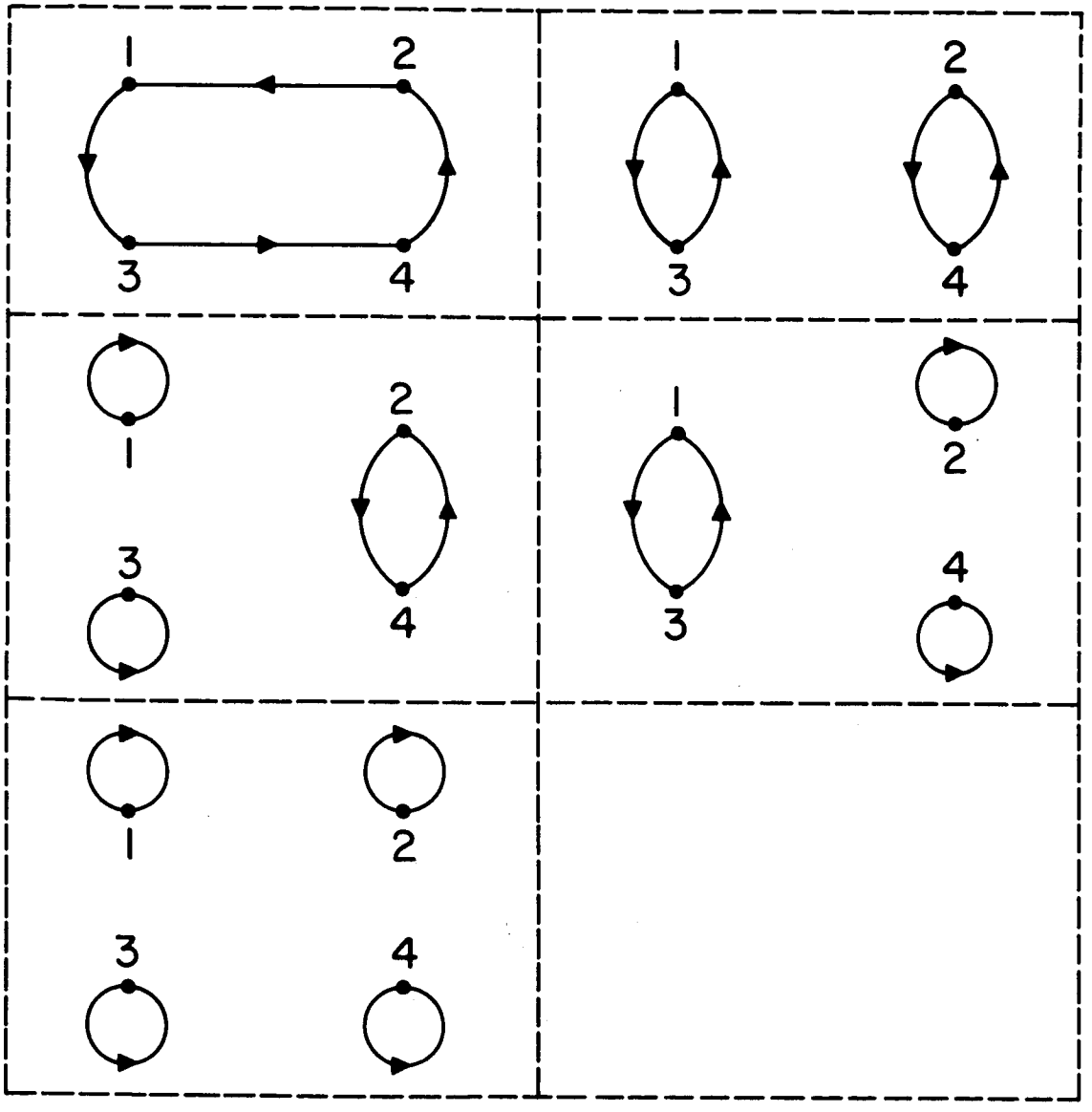


Fig. 4. The 5 connections of the flow-graph shown in Figure 1.

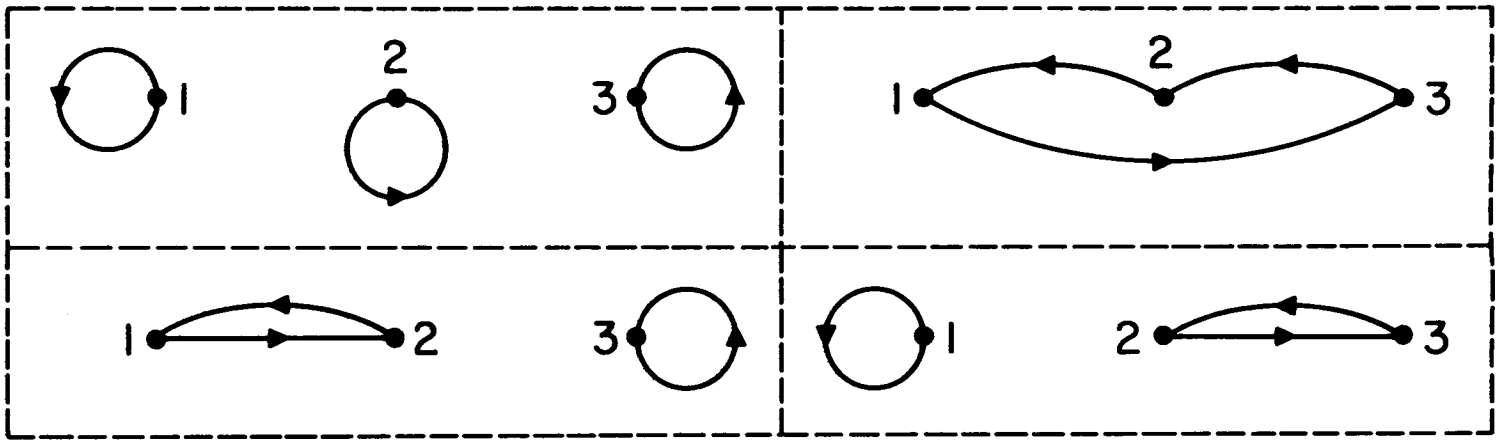


Fig. 5. The 4 connections of the flow-graph shown in Figure 2.

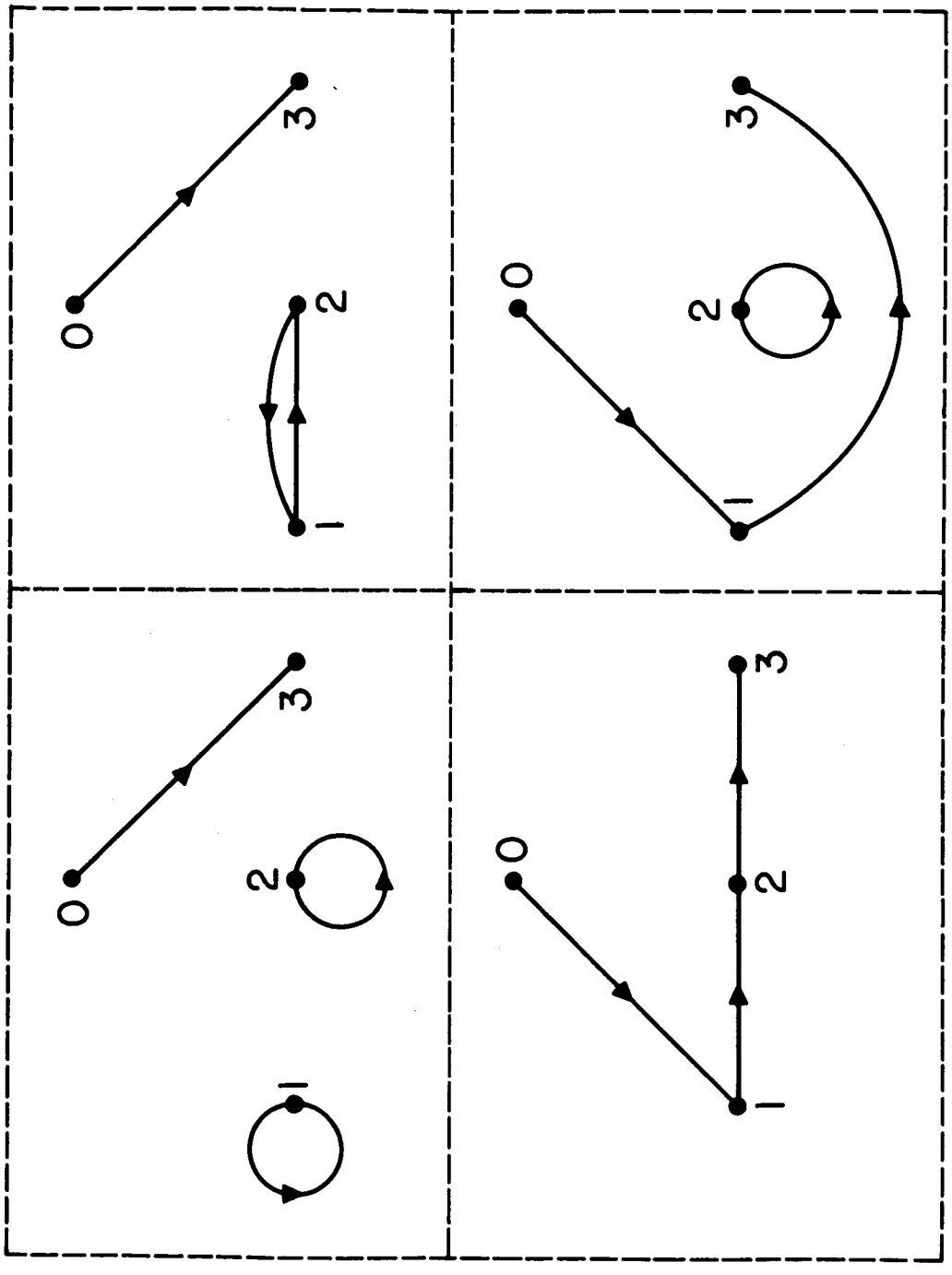


Fig. 6. The 4 one-connections of the flow-graph shown in Figure 2.

References

1. C. L. Coates, "Flow graph solutions of linear algebraic equations," IRE Trans. on Circuit Theory, Vol. CT-6, pp. 170-187; June 1959.

2. C. A. Desoer, "The optimum formula for the gain of a flow graph or a simple derivation of Coates' formula," Proc. IRE, Vol. 48 (5), pp. 883-889; May 1960.

3. S. J. Mason, "Feedback theory - some properties of signal flow graphs," Proc. IRE, Vol. 41, pp. 1144-1156; Sept. 1953.

4. S. J. Mason, "Feedback theory - further properties of signal flow graphs," Proc. IRE, Vol. 44, pp. 920-926; July 1956.

5. E. Carrignol and Y. Chow, Linear Signal Flow Graphs and Applications, John Wiley and Sons, Inc., N. Y., 1962.

6. A. C. Aitken, Determinants and Matrices, 5th ed., Interscience Publishers, N. Y., p. 30; 1948.

7. S. Seshu and N. Balabanian, Linear Network Analysis, John Wiley and Sons, Inc., N. Y., 1959.

ON THE NUMBER OF ROOTS OF A REAL POLYNOMIAL INSIDE (OR OUTSIDE) THE UNIT CIRCLE USING THE DETERMINANT METHOD

E. I. Jury

In this note, a simple procedure is presented which gives the number of roots outside and inside the unit circle of a real polynomial $F(z)$. It requires the evaluation of either odd or even determinants depending on "n" (the degree of the polynomial), i.e., whether n is even or odd. The form of these determinants is presented elsewhere[1] and we present only the results:

n - even

The number of roots outside the unit circle (if none of the appropriate $A'_k \pm B'_k$ vanishes) is given by the number of sign changes of the following determinants:

$$\left.\begin{array}{l} A'_1 + B'_1, \qquad\qquad A'_3 + B'_3, \ \ldots, \ A'_{n-1} + B'_{n-1} \\ A'_1 + B'_1, \ A'_1 - B'_1, \ A'_3 - B'_3, \ \ldots, \ A'_{n-1} - B'_{n-1} \end{array}\right\} \quad \text{or,}$$

$$A'_{n-1} - B'_{n-1}, \ \ldots, \ A'_1 - B'_1,$$

$$A'_1 + B'_1, \ \ldots, \ A'_{n-1} + B'_{n-1}$$

plus one or zero depending on the sign of $F(1)\ F(-1)$.

The research herein was supported by the Air Force Office of Scientific Research under Grant AF-AFOSR-292-64.

n - odd

$$1,\ A'_2 + B'_2,\ A'_4 + B'_4,\ \ldots,\ A'_{n-1} + B'_{n-1}$$
$$1,\ A'_2 - B'_2,\ A'_4 - B'_4,\ \ldots,\ A'_{n-1} - B'_{n-1}$$

or,

$$A'_{n-1} - B'_{n-1},\ \ldots,\ A'_2 - B'_2,\ 1,$$
$$A'_2 + B'_2,\ \ldots,\ A'_{n-1} - B'_{n-1}$$

plus one or zero depending on the sign of $F(1)\ F(-1)$.

The number of roots of $F(z) = 0$ inside the unit circle is n minus the number of the outside roots, provided $F(1)\ F(-1) \neq 0$.

Example

Let $n = 10$, and using the following value

$$A'_1 + B'_1 < 0,\ A'_3 + B'_3 < 0,\ A'_5 + B'_5 < 0,\ A'_7 + B'_7 > 0,\ A'_9 + B'_9 > 0$$

$$A'_1 + B'_1 < 0,\ A'_1 - B'_1 < 0,\ A'_3 - B'_3 > 0,\ A'_5 - B'_5 > 0,$$
$$A'_7 - B'_7 < 0,\ A'_9 - B'_9 < 0$$

and $F(1)\ F(-1) < 0$, we obtain the number of changes of sign of the determinants as 3. Since $n = 10$, and $F(1)\ F(-1) < 0$, the number of roots outside the unit circle is odd, i.e., three. The number inside the unit circle is $10 - 3 = 7$.

Example

To find the root distribution of the following polynomial

$$F(z) = 3 - 2z - \frac{3}{2} z^2 + z^3,\ n = 3$$

In this case:

$$A'_2 + B'_2 = -\frac{11}{2} < 0, \quad A'_2 - B'_2 = -\frac{21}{2} < 0$$

Hence,

$$1 > 0, \quad A'_2 + B'_2 < 0$$

$$1 > 0, \quad A'_2 - B'_2 < 0$$

There are at least two roots outside the unit circle. Since $F(1)\ F(-1) > 0$ (no single root can exist inside the unit circle), hence the third root is outside the unit circle.

The singular cases are discussed in detail in references (1) and (2). The proof of the above simplified counting is based on certain manipulations[2] of the following formula[1]:

$$2(A'^2_k - B'^2_k) = (A'_{k-1} - B'_{k-1})(A'_{k+1} + B'_{k+1})$$

$$+ (A'_{k-1} + B'_{k-1})(A'_{k+1} - B'_{k+1})$$

References

1. E. I. Jury, Theory of Application of the z-Transform Method, John Wiley and Sons, Inc., Ch. 3.
2. B. M. Brown, "On the distribution of the zeros of a polynomial," to be published.
3. E. I. Jury, "A modified stability table for linear discrete systems," to be published in Proc. IEEE; 1965.

STABILITY OF SINGLE-LOOP FEEDBACK SYSTEMS

C. T. Lee and C. A. Desoer

We consider the system S shown in Fig. 1: it is a single-input single-output single-loop feedback system; N is a memoryless time-invariant nonlinear element and G is a nonanticipative time-varying linear system. The purpose of this paper is to extend the sufficient conditions for stability given by Popov[1] and Desoer[2, 3].

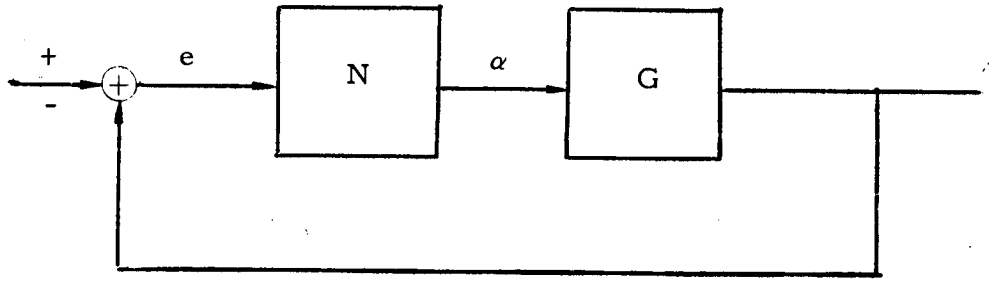


Fig. 1. The System S

Of course, since G is time-varying there is no frequency domain interpretation. However it is of interest to note that the linear system must satisfy some very simple assumptions.

N. 1. The nonlinear element is characterized by

$$\alpha(t) = \varphi[e(t)] \tag{1}$$

where the nonlinear characteristic φ is a piecewise continuous function defined on $(-\infty, \infty)$, i.e., φ may only have a finite number of discontinuities in any finite interval, and at any such discontinuity, e.g., e_i, both $\varphi(e_i-)$ and $\varphi(e_i+)$ are defined and finite. Furthermore,

N. 2. There are two positive constants ε and k such that

$$0 < \varepsilon \leqslant \frac{\varphi(e)}{e} \leqslant k - \varepsilon \quad \text{for all } e \neq 0 \tag{2}$$

The research herein was supported by National Aeronautics and Space Administration under Grant NsG-354 (S-2).

and

$$\varphi(0) = 0.$$

G. 1. The linear subsystem G is characterized by its input-output equation

$$y(t) = z(t) + \int_0^t g(t, t_2)\alpha(t_2)dt_2 \qquad t \geq 0 \tag{3}$$

where $z(\cdot)$ is the zero-input response which depends on the state of G at time 0.

In other words, G is a nonanticipative linear time-varying system represented by a superposition integral and N is a time-invariant memoryless nonlinearity whose characteristic lies in the sector by the line $\alpha = \epsilon e$ and $\alpha = (k-\epsilon)e$. The equation for the feedback system may be written in the operator form as follows:

$$y = z + G\alpha, \tag{4}$$

where α is given in terms of y by (1). To state the main result we must list more assumptions:

G. 2. $g(t_1, t_2) = 0$ whenever $t_1 < t_2$

$$|g(t_1, t_2)| \leq g_M < \infty \qquad \text{for all } t_1 \text{ and } t_2 \geq 0$$

$$\int_0^\infty |g(t_1, t_2)|dt_2 \leq M^2 \qquad \text{for all } t_1 \geq 0$$

$g(t, t_2) \to 0$ as $t \to \infty$ uniformly for t_2 in any finite interval $(0, T)$.

G. 3. For all initial states, $z(0)$ is finite, $z(\cdot)$ and $\dot{z}(\cdot)$ are in $L^2(0, \infty)$,

$$z(t) \to 0 \quad \text{as} \quad t \to \infty.$$

G. 4. The solution of (4) is assumed to satisfy the following requirement: for any finite t

$$y(\cdot) \in L^2(0, t) \text{ and } \dot{y}(\cdot) \in L^2(0, t).$$

In other words, G. 2 requires that the impulse response be bounded for all times, that the response to an impulse applied at any time t_2 goes to zero, that the zero state response of G to any bounded input be bounded. G. 3 requires that the response due to any initial state be square integrable as well as its derivative, and that it goes to zero as $t \to \infty$. G. 4 may be thought of as a condition eliminating certain kinds of finite escape time response. In fact, if $z(\cdot)$ is bounded it is easy to show that y is bounded by an exponential; see the proof of Assertion I in reference 3. The main result is stated in the form of a theorem.

Theorem

Consider the system S shown in Fig. 1. Let $u \equiv 0$. Suppose that the assumptions N. 1-2, G. 1-4 are satisfied. Under these assumptions, if there are two positive numbers q and δ such that

$$\langle x, (G + qDG + \frac{1}{k})x \rangle \geq \delta \langle x, x \rangle \quad \text{for all } x \in L^2(0, \infty), \qquad (5)$$

then the output of y of S is bounded on $[0, \infty)$, belongs to $L^2(0, \infty)$ and goes to zero as $t \to \infty$, whatever might be the state of G at time 0. Here the operator D denotes the time derivative operator, i. e.,

$$Dx = \frac{d}{dt}(x(t)).$$

Before proving this theorem, we shall prove two auxiliary lemmas.

Lemma 1

Let A be an operator (not necessarily linear) mapping from Hilbert space H into itself and let x_i, $i = 1, 2, 3$ be points in H such that

$$x_1 = x_2 + A(x_3). \tag{6}$$

Under the above assumptions, if there exists $\delta > 0$ such that for all $x \in H$

$$\langle A(x), x\rangle \geq \delta \langle x, x\rangle, \tag{7}^{*}$$

then

$$-\langle x_1, x_3\rangle \leq \frac{1}{4\delta} \langle x_2, x_2\rangle.$$

Proof

From (6),

$$\langle x_1, x_3\rangle = \langle x_2 + A(x_3), x_3\rangle = \langle x_2, x_3\rangle + \langle A(x_3), x_3\rangle.$$

From (7),

$$\langle x_1, x_3\rangle \geq \langle x_2, x_3\rangle + \delta\langle x_3, x_3\rangle.$$

Now,

$$\langle x_2, x_3\rangle + \delta \langle x_3, x_3\rangle = \langle \sqrt{\delta}\, x_3 + \frac{1}{2\sqrt{\delta}} x_2, \sqrt{\delta}\, x_3 + \frac{1}{2\sqrt{\delta}} x_2 \rangle - \frac{1}{4\delta} \langle x_2, x_2\rangle.$$

Since the first term of the right hand side of the equality is nonnegative,

$$\langle x_1, x_3\rangle = \langle x_2, x_3\rangle + \delta\langle x_3, x_3\rangle \geq - \frac{1}{4\delta}\langle x_2, x_2\rangle$$

$$\therefore \quad -\langle x_1, x_3\rangle \leq \frac{1}{4\delta}\langle x_2, x_2\rangle.$$

This is the desired inequality. Q. E. D.

Lemma 2

Let $x \in L^2(0, \infty) \cap L^\infty(0, \infty)$

$$|g(t_1, t_2)| \leq g_M < \infty \quad \text{for all } t_1 \text{ and } t_2 \geq 0$$

$$\int_0^\infty |g(t_1, t_2)|\, dt_2 \leq M^2 < \infty \quad \text{for all } t_1 \geq 0$$

* Zames [4] defines the operators satisfying (7) "δ-passive operators"

$g(t, t_2) \to 0$ as $t \to \infty$ uniformly in any finite interval $[0, T]$.

Then $w(t) \triangleq \int_0^t g(t, t_2) x(t_2) dt_2$ goes to zero as $t \to \infty$.

Proof

$$|w(t)| \leq \int_0^t |g(t, t_2)|\,|x(t_2)|\,dt_2 \leq \left(\int_0^t |g(t, t_2)|\,dt_2\right)^{1/2}$$

$$\cdot \left(\int_0^t |g(t, t_2)|\,|x(t_2)|^2 dt_2\right)^{1/2} \leq M \left(\int_0^t |g(t, t_2)|\,|x(t_2)|^2 dt_2\right)^{1/2}$$

Because $x \in L^2(0, \infty)$, for arbitrary $\epsilon_1 > 0$ we can choose T sufficiently large so that

$$\int_T^\infty |x(t_2)|^2 dt_2 < \epsilon_1.$$

Then, since $g(t, t_2) \to 0$ as $t \to \infty$ uniformly, for arbitrary $\epsilon_2 > 0$ there is T' so that

$$\int_0^T |g(t, t_2)|\,dt_2 < \epsilon_2 \qquad \text{for } t \geq T'.$$

Hence,

$$\int_0^t |g(t, t_2)|\,|x(t_2)|^2 dt_2 = \int_0^T |g(t, t_2)|\,|x(t_2)|^2 dt_2 + \int_T^t |g(t, t_2)|\,|x(t_2)|^2 dt_2$$

Since $x \in L^\infty$, there exists a finite number x_M such that $|x(t)| \leq x_M$, $t \geq 0$

$$\therefore \qquad \int_0^t |g(t, t_2)|\,|x(t_2)|^2 dt_2 \leq x_M^2 \int_0^T |g(t, t_2)|\,dt_2 + g_M \int_T^t |x(t_2)|^2 dt_2$$

$$\leq x_M^2 \epsilon_2 + g_M \epsilon_1 \qquad \text{for } t \geq T'$$

$$\therefore \qquad |w(t)| \leq M(x_M^2 \epsilon_2 + g_M \epsilon_1)^{1/2} \qquad \text{for } t \geq T'$$

$\therefore$ $\quad w(t) \to 0$ as $t \to \infty$. O.E.D.

Proof of Theorem

Let a subscript T placed on any function, such as y, denote the function defined by

$$y_T(t) = \begin{cases} y(t) & 0 \leq t < T \\ 0 & \text{elsewhere} \end{cases}$$

From (4)

$$y_T = z_T + G\alpha_T \qquad \text{for } 0 \leq t \leq T.$$

Let us differentiate the above equation and add. We get

$$y_T + q\dot{y}_T + \frac{1}{k}\alpha_T = z_T + q\dot{z}_T + (G + qDG + \frac{1}{k})\alpha_T.$$

From G. 3, G. 4, N. 2 and the assumption of the theorem, it is implied that the assumptions of Lemma 1 hold; hence, for any $T > 0$

$$-\int_0^T \left\{y(t) + q\dot{y}(t) + \frac{1}{k}\alpha(t)\right\}\alpha(t)dt \leq \frac{1}{4\delta}\int_0^T |z(t) + q\dot{z}(t)|^2 dt.$$

Hence, since $u \equiv 0$, $e(t) = -y(t)$, and $\alpha(t) = \varphi[e(t)]$

$$\int_0^T \left\{e(t) - \frac{1}{k}\varphi[e(t)]\right\}\varphi[e(t)]\,dt + q\int_0^T \varphi[e(t)]\dot{e}(t)dt \leq \frac{1}{4\delta}\int_0^\infty |z(t) + q\dot{z}(t)|^2 dt \triangleq C. \quad (8)$$

The constant C is finite and is independent of T. Call J_1 and J_2 the two terms of the left hand side. By (2), $J_1 \geq 0$, hence, $J_2 \leq C$

$$J_2 = q\int_0^T \varphi[e(t)]\,\dot{e}(t)dt = q\int_{e(0)}^{e(T)} \varphi(\xi)d\xi$$

$$\therefore \quad q\int_0^{e(T)} \varphi(\xi)d\xi \leq C + \int_0^{e(0)} \varphi(\xi)d\xi \triangleq C'. \quad (9)$$

Since $z(0)$ is finite by G. 3 and $e(0) = -y(0) = -z(0)$, the constant C' equal to the right-hand side of (9) is finite. Now by (2) we get

$$|e(T)|^2 \leq \frac{2C'}{\epsilon q} \quad \text{for all } T > 0.$$

Therefore $e(\cdot)$, J_2, and $y(\cdot)$ are bounded on $[0, \infty)$. Using (9) in (8) and N. 2 we get

$$J_1 \leq C' \tag{10}$$

Now by N. 2 again

$$J_1 \triangleq \int_0^T \left\{e(t) - \frac{1}{k}\ \varphi[e(t)]\right\} \varphi[e(t)]\, dt \geq \frac{\epsilon^2}{k} \int_0^T |e(t)|^2 dt \quad \text{for all } T > 0.$$

It follows that $e(\cdot) \in L^2(0, \infty)$, and the same is true for $y(\cdot)$. Thus, $y(\cdot) \in L^\infty \cap L^2$ and $\alpha(\cdot) \in L^\infty \cap L^2$. By Lemma 2, $(G\alpha)(t) \to 0$ as $t \to \infty$. Finally, with $z \to 0$ by (G. 3), (4) implies that $y(t) \to 0$ as $t \to \infty$. Thus we have shown that $y \in L^2 \cap L^\infty$ and $\to 0$ as $t \to \infty$. Q. E. D.

It might seem that the conclusions are restricted to statements concerning the zero-input response of the closed loop system, but this is not the case. Suppose that $u \neq 0$ and that it satisfies the condition G. 3, then (4) may be rewritten as

$$-e = -u + z + G\alpha \quad \text{with } e = u - y.$$

Since $u - z$ in the present case satisfies the same conditions as z in the statement of the theorem, we therefore have the corollary:

Corollary

Let the input u satisfy the conditions G. 3, and let all the other assumptions of the theorem hold. Then the output y of S is bounded, on $[0, \infty)$, belongs to $L^2(0, \infty)$ and goes to zero as $t \to \infty$, whatever might

be the initial state of G at time 0.

It is obvious that the results can be extended when input-output relations are represented in a vector form.

Example

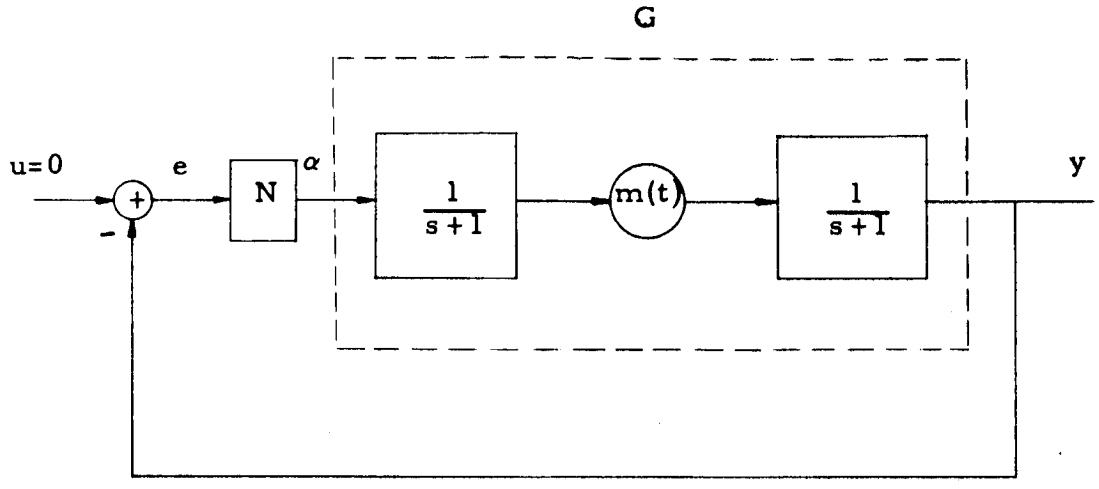


Fig. 2.

N satisfies N.1 and N.2. Assume $|m(t)| \leq \mu < \infty$ for $t \geq 0$. It can easily be shown that

$$g(t_1, t_2) = \left\{ e^{-(t_1 - t_2)} \int_{t_2}^{t_1} m(\tau) d\tau \right\} 1(t_1 - t_2)$$

and $g(t_1, t_2)$ satisfies G.1, G.2. By direct calculation, G.3 is easily shown to hold; G.4 follows by the Gronwall-Bellman lemma (5) applied to (3) where (2) is taken into account. In the following we shall find a stability region for k.

$$(G\alpha)(t) = \int_0^t e^{-(t-\tau)} m(\tau) \int_0^\tau e^{-(\tau-\xi)} \alpha(\xi) d\xi d\tau$$

$$= e^{-t} \int_0^t e^{\tau} m(\tau) \int_0^\tau e^{-(\tau-\xi)} \alpha(\xi) d\xi d\tau$$

$$(DG\alpha)(t) = -(G\alpha)(t) + m(t) \int_0^t e^{-(t-\xi)} \alpha(\xi) d\xi$$

$$((G + DG)\alpha)(t) = m(t) \int_0^t e^{-(t-\xi)} \alpha(\xi) d\xi$$

$$|<\alpha, (G + DG)\alpha>| \leq \mu \int_0^\infty |\alpha(t)| \int_0^t e^{-(t-\xi)} |\alpha(\xi)| d\xi dt$$

$$\leq \mu < \alpha, \alpha >$$

$$\therefore < \alpha, \left(G + DG + \frac{1}{k} \right) \alpha > \geq \left(\frac{1}{k} - \mu \right) < \alpha, \alpha > .$$

Therefore, if $\frac{1}{k} - \mu > 0$, then the conclusions of the theorem hold for the system under consideration.

REFERENCES

1. M. A. Aizerman and F. R. Gantmacher, Absolute Stability of Regulator Systems, Chapter III, San Francisco, Holden Day; 1964.

2. C. A. Desoer, "A generalization of the Popov criterion," Trans. IEEE, AC-10, April 1965.

3. C. A. Desoer, A General Formulation of the Nyquist Criterion, Internal Technical Memorandum M-104, October 26, 1964, Electronics Research Laboratory, University of California, Berkeley. To be published in Trans. IEEE, CT-12; June 1965.

4. G. Zames, "On the stability of nonlinear, time-varying feedback systems," Proceedings of the N.E.C., vol. 20, pp. 725-730; October 1964.

5. E. A. Coddington and N. Levinson, Theory of Ordinary Differential Equations, New York, McGraw-Hill Book Co., 1955, p. 37, Problem 1.

A CONSTRUCTIVE DERIVATION OF THE CAPACITY OF A BANDLIMITED CHANNEL

D. J. Sakrison and L. P. Seidman

I. Introduction

The expression

$$C = W \log_2(1 + \frac{P}{2N_o W})$$

for the capacity of a channel limited to bandwidth 2W in the presence of additive Gaussian white noise of two sided spectral density N_o is well known ([1], Chapt. 5). However, the proof usually given of the positive side of Shannon's coding theorem for this channel is not constructive [2]. That is, it does not demonstrate for a rate less than C the existence of a code with large but finite delay that can achieve some desired small error probability. The major defect of such proofs in this regard is their use of the sin x/x functions for signalling. To make the proof constructive, one needs to replace these functions with suitable functions whose duration is equal to the block length of the code (one half the delay). The suitable functions for use in this regard turn out to be the Fourier transforms of the prolate spheroidal functions discussed by Slepian, Landau and Pollack [3, 4, 5, 6].

Consider a channel over which we are allowed to send signals of time duration T; each signal to have energy less than PT and energy

---

This work was supported wholly by the Joint Services Electronics Program (U. S. Army, U. S. Navy and U. S. Air Force) under Grant No. AF-AFOSR-139-64.

outside a band $\pm$ W c.p.s. less than ϵPT, $0 < \epsilon < 1$. This channel has additive white Gaussian noise of two sided spectral density N_o. We shall show in the next section that the capacity of such a channel is

$$C = W \log_2\left(1 + \frac{(1-\epsilon)P}{2N_oW}\right) + \frac{\epsilon P}{2N_o} \log_2 e$$

by showing that for large WT this channel can be reduced to a collection of amplitude continuous, time discrete channels whose joint capacity is easily computed. Since there exist constructive proofs of the coding theorem for such a collection of time discrete channels, our argument will be constructive.

II. Solution

Taking the Fourier transform of the functions discussed in [3], we have a set of time limited functions $\varphi_i(t)$ $i = 0, 1, \ldots,$ which are complete in $L^2[0, T]$ and orthonormal ([6], p. 70). Further, there exist constants λ_i such that $\lambda_i > 0$, λ_i strictly decreasing with i, and λ_i is the energy of φ_i in the frequency band $|f| \leq W$.

Slepian has shown ([5], p. 11) that if b is fixed and

$$i = 2WT + b\ell n \frac{\pi}{2} 2WT$$

then

$$\lim_{WT \to \infty} \lambda_i = \frac{1}{1 + e^{\pi^2 b}}$$

Now consider a channel with white additive Gaussian noise of spectral density N_o. We are to send a signal S(t), such that (1), (2), and (3) are satisfied:

$$S(t) = 0 \text{ for } t \text{ not in } [0, T] \qquad (1)$$

$$\text{energy } S(t) \leq P \tag{2}$$

$$\{\text{energy of } S(t) \text{ out of band } |f| < W\} \leq \epsilon P \tag{3}$$

Then representing the signal in the basis of prolate spheroidal wave functions we have,

$$S(t) = \sum_{k=1}^{\infty} S_k \varphi_k(t)$$

and the constraints are

$$\sum_{k=1}^{\infty} |S_k|^2 \leq P \tag{4}$$

$$\sum_{k=1}^{\infty} |S_k|^2 (1-\lambda_k) \leq \epsilon P \tag{5}$$

The problem is to choose the S_k, satisfying (4) and (5) to obtain channel capacity.

Arbitrarily choosing d_o, let

$$N_1 = 2WT - d_o \ell n(\tfrac{\pi}{2} 2WT)$$

$$N_2 = 2WT + d_o \ell n(\tfrac{\pi}{2} 2WT)$$

Our channel is then represented by three distinct and statistically independent channels:

$$\text{chan 1} = \{S_1, S_2, \ldots, S_{N_1}\}$$

$$\text{chan 2} = \{S_{N_{1+1}}, \ldots, S_{N_2}\}$$

$$\text{chan } 3 = \{S_{N_{2+1}}, S_{N_{2+2}}, \ldots,\}$$

and we must maximize the channel capacity

$$C = C_1(P_1) + C_2(P_2) + C_3(P_3)$$

where $C_i(P_i)$ is the capacity of the i^{th} channel with energy P_i used on it, subject to the constraints;

$$P_1 + P_2 + P_3 \leq P$$

$$\sum |S_k|^2 (1-\lambda_k) < \epsilon P$$

Suppose $P_o = P_1 + P_2 \leq P$, and $P_1 = \alpha P_o$. Then, as is easily shown, the optimum value of α is

$$\alpha = \frac{2W_1}{2W_1 + 2W_2} = \frac{N_1}{N_1 + (N_2 - N_1)} = \frac{N_1}{N_2} = \frac{2WT - d_o \ell n(\frac{\pi}{2} 2WT)}{2WT + d_o \ell n(\frac{\pi}{2} 2WT)}$$

so $\alpha \to 1$ as $2WT \to \infty$. Thus no power is used in C_2. Now for $k > N_1$,

$$1 - \lambda_k \leq \epsilon_1 \leq 1 - \frac{1}{1 + e^{-d_o \pi^2}}$$

for $k > N_2$, $\quad 1 - \lambda_k \leq 1 - \epsilon_2 \leq 1 - \dfrac{1}{1 - e^{+d_o \pi^2}}$

where ϵ_1 and ϵ_2 may be made arbitrarily small by choosing d_o large. Thus, we must have for large $2WT$

$$\sum_{k=0}^{N_1} |S_k|^2 = P_1 = (1-\epsilon)P$$

$$\sum_{k=N_1+1}^{N_2} |S_k|^2 = P_2 = 0$$

$$\sum_{k=N_2+1}^{\infty} |S_k|^2 = P_3 = \epsilon P$$

and then the capacity of the channel is (from the usual formulas)

$$C = C[(1-\epsilon)P] + C_3[\epsilon P]$$

$$= W \log_2 \left(1 + \frac{(1-\epsilon)P}{2WN_o}\right) + \frac{\epsilon P}{2N_o} \log_2 e.$$

If $\epsilon \to 0$, we have

$$C = W \log_2 \left(1 + \frac{P}{2WN_o}\right)$$

which is the formula obtained by considering the channel as one with $2WT$ coefficients. It must be observed, that while the above conclusions are strictly true only as $WT \to \infty$, agreement is quite good for modest WT, ([1], p. 4).

References

1. R. M. Fano, Transmission of Information, MIT Press and John Wiley and Sons, Inc., New York; 1961.
2. C. E. Shannon, "The mathematical theory of communication," B.S.T.J.; July 1948 and October 1948 (reprinted by The University of Illinois Press, Urbana.)
3. D. Slepian, and H. O. Pollak, "Prolate spheroidal wave functions, Fourier analysis and uncertainty - I," B.S.T.J., Vol. 40, pp. 43-46, Jan. 1961.
4. H. J. Landau, and H.O. Pollak, "Prolate spheroidal wave functions, Fourier Analysis and uncertainty - II," B.S.T.J., Vol. 40, pp. 65-84; Jan. 1961.
5. D. Slepian, "Some Asymptotic Expansions for Prolate Spheroidal Wave Functions," Unpublished Bell Telephone Laboratories, Inc., memorandum.
6. A. Papoulis, The Fourier Integral and its Applications, McGraw-Hill Book Company, Inc., New York; 1962.

STABILITY ANALYSIS OF MONOTONE FEEDBACK SHIFT REGISTERS*

C. J. Tan and A. Gill

In a recent paper Massey and Liu[1] introduced a special class of feedback shift registers, or FSR's, called monotone. The stability of this class of FSR's was examined in that paper. However, Massey and Liu were unable to formulate a general stability test for monotone FSR's. In this note, a general test (which is also the necessary and sufficient condition) for the stability of monotone FSR's is derived, by the use of Boolean matrices.[2]

Description of autonomous FSR's by Boolean matrices.

A general autonomous FSR consists of delayers and logical elements arranged in the form as shown in Fig. 1.

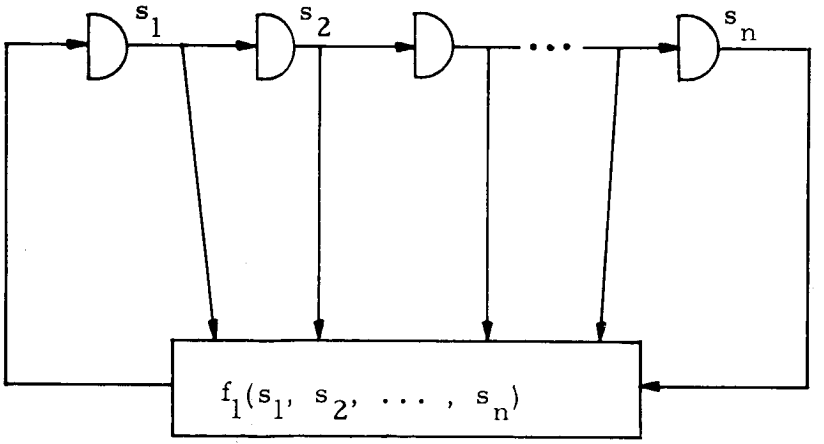


Fig. 1

* The research herein was supported by the Air Force Office of Scientific Research under Grant AF-AFOSR-639-64.

s_i is the binary output symbol of the i-th delayer, and $f_i(s_1, s_2, \ldots, s_n)$ is an arbitrary Boolean input function for the i-th delayer. The binary n-tuple $\underline{s} = (s_1, s_2, \ldots, s_n)$ is the state of the FSR.

A Boolean matrix is, in effect, a Karnaugh map of Boolean functions in its row space. For example, consider the FSR shown in Fig. 2.

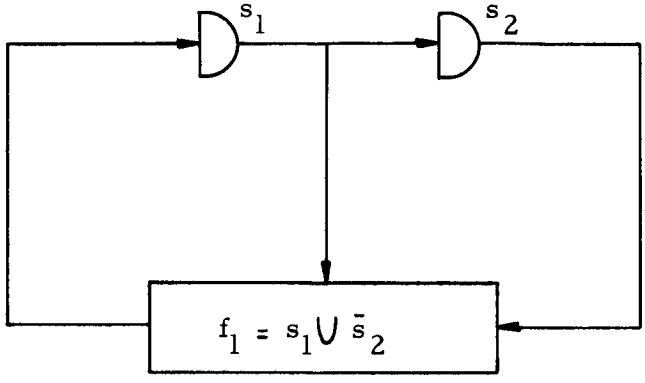


Fig. 2

The feedback function f_1 and f_2 can be expressed as the sum of all minimal polynomials, i.e.,

$$f_1 = s_1 \cup \bar{s}_2 = (1 \cap \bar{s}_1 \cap \bar{s}_2) \cup (1 \cap s_1 \cap \bar{s}_2) \cup (0 \cap \bar{s}_1 \cap s_2) \cup (1 \cap s_1 \cap s_2)$$

$$f_2 = s_1 = (0 \cap \bar{s}_1 \cap \bar{s}_2) \cup (1 \cap s_1 \cap \bar{s}_2) \cup (0 \cap \bar{s}_1 \cap s_2) \cup (1 \cap s_1 \cap s_2)$$

The matrix equation characterizing this FSR is $\underline{B}_2 \, \underline{s} = \underline{f}$, i.e.,

$$\begin{bmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \qquad (1)$$

If each state $\underline{s} = (s_1, s_2, \ldots, s_n)$ of an FSR with n delayers is regarded as a binary number with the least significant digit at the left, or equivalently at the top row in the column matrix $\underline{s}$, then each column in $\underline{B}_n$ can be represented by its decimal equivalent state vector, e.g., for above example,

$$\underline{B}_2 = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \underline{1} & \underline{3} & \underline{0} & \underline{3} \end{bmatrix} \tag{2}$$

Furthermore, in a Boolean matrix $\underline{B}_n$, representing an FSR with n delayers, the state vector at the i-th column, starting from the left, is the successor state of state $\underline{i-1}$. Hence, for the above example, state (0 0) is followed by state (1 0), state (1 0) is followed by state (1 1), etc.

The identity Boolean matrix $\underline{A}$ is one for which $\underline{f} = \underline{A}\,\underline{s} = \underline{s}$. It is the matrix which expresses the set of equations $f_i = s_i$, i = 1, ... , n;

$$\underline{A}_1 = \begin{bmatrix} 0 & 1 \end{bmatrix}, \quad \underline{A}_2 = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \quad \underline{A}_3 = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}, \text{ etc.}$$

In general, expressed in decimal equivalent,

$$\underline{A}_n = \begin{bmatrix} \underline{0} & \underline{1} & \underline{2} \cdots \underline{2^{n-1}-1} & \underline{2^{n-1}} & \underline{2^{n-1}+1} & \cdots & \underline{2^n-1} \end{bmatrix} \tag{3}$$

Monotone FSR's.

An FSR is called monotone[1] if and only if $W(\underline{B}\,\underline{s}) \leq W(\underline{s})$, for all $\underline{s}$, where $W(\underline{s})$ denotes the Hamming weight of the state $\underline{s}$. So, an FSR is monotome if the Hamming weights of its states in every state sequence form a monotonically non-increasing sequence.

For any FSR of the type shown in Fig. 1, the state function $\underline{f} = (f_1, f_2, \ldots, f_n)$ is equal to $(f_1, s_1, s_2, \ldots, s_{n-1})$. Hence, all rows, except the first row, of the corresponding Boolean matrix $\underline{B}$ are fixed; and row i, where $2 \leq i \leq n$, is the same as the (i - 1) st row of the identity matrix $\underline{A}_n$. That is:

$$\underline{B}_n = \begin{bmatrix} X_0 & X_1 & X_2 & X_3 & \cdots & X_{2^{n-1}-1} & X_{2^{n-1}} & X_{2^{n-1}+1} & \cdots & X_{2^n-1} \\ 0 & 1 & 0 & 1 & \cdots & 1 & 0 & 1 & \cdots & 1 \\ 0 & 0 & 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 1 \\ \vdots & & & & & & \vdots & & & \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

$$\underline{A}_n = \begin{bmatrix} 0 & 1 & 0 & 1 & \cdots & 1 & 0 & 1 & \cdots & 1 \\ 0 & 0 & 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 1 \\ \vdots & & & & & & \vdots & & & \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 1 & 1\cdots & 1 \end{bmatrix} \qquad (4)$$

Each one of the 2^n positions on the top row can be filled with either 0 or 1. Each different combination represents one of the 2^{2^n} possible distinct feedback functions f_1. For a monotone FSR, $W(\underline{B}\,\underline{s}) \leq W(\underline{s})$. Hence $X_0 = X_1 = \cdots = X_{2^n-1} = 0$.

If it is further required that $X_{2^{n-1}} = \ldots = X_{2^n-1} = 1$, then the row space of $\underline{B}_n$ is the cyclic shift of the row space of the corresponding identity matrix $\underline{A}_n$, as can be seen from equation (4). The feedback function of this particular Boolean matrix is $f_1 = s_n$. Such an FSR is called a pure cycling register.[3] Each successor state is a cyclic shift of the previous state. Hence all states fall in a cycle which consists of all states $\underline{i}$ such that i is a binary shift modulo 2^n-1 of a binary number j such that $\underline{j}$ is not contained in some other cycle. For example, for n = 3

$$\underline{B}_3 = \begin{array}{c} \begin{matrix} X_0 & & & & & & & X_7 \\ \downarrow & & & & X_4 \downarrow & & & \downarrow \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

and every state is in one of the 4 cycles shown in Fig. 3.

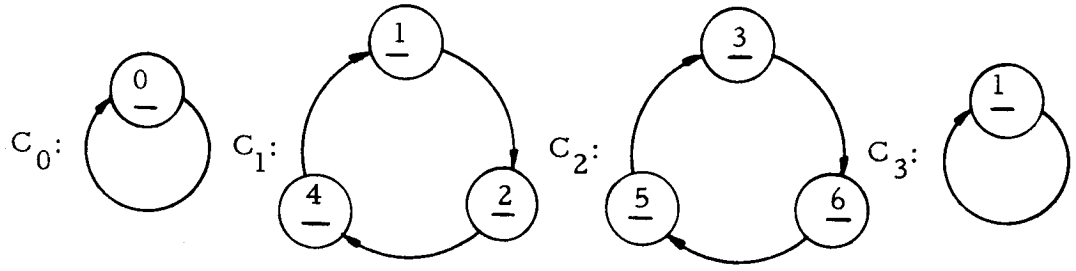


Fig. 3

If X_6, say, is filled with 0 instead of 1, then the successor state of (0 1 1) will not be a cyclic shift of $\underline{6}$ but rather a state belonging to another cycle with weight one less than the original cycle; in this case $\underline{B} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$.

Stability.

A monotone FSR is called stable if and only if $\underline{B}^N \underline{s} = \underline{0}$ for every state of the FSR, and some finite positive integer N. In other words, there cannot be any cycles except the trivial one in the state diagram.

Let the set of states $S = \{\underline{0}, \underline{1}, \underline{2}, \ldots, \underline{2^n-1}\}$ be the index set

S can be decomposed into disjoint subsets, each containing all the states in C_i, where C_i is one of the cycles arising from the pure cycling register. Hence,

$$
\begin{aligned}
H_0 &= \{ \underline{0} \} \\
H_1 &= \{ \underline{1}, \underline{2}, \underline{4}, \cdots, \underline{2^{n-1}} \} \\
H_3 &= \{ \underline{3}, \underline{6}, \underline{12}, \cdots, \underline{3 \times 2^{n-1}} \} \\
&\cdot \\
&\cdot \\
&\cdot \\
H_{2^n-1} &= \{ \underline{2^n-1} \}
\end{aligned}
\tag{5}
$$

Let $X = \{ X_{\underline{0}}, X_{\underline{1}}, \cdots, X_{\underline{2^n-1}} \}$ be the symbols at the respective columns at the first row of $\underline{B}_n$. They fall into subsets according to their respective indices; that is,

$$
\begin{aligned}
G_0 &= \{ X_{\underline{0}} \} \\
G_1 &= \{ X_{\underline{1}}, X_{\underline{2}}, \cdots, X_{\underline{2^{n-1}}} \} \\
&\cdot \\
&\cdot \\
&\cdot \\
G_{2^n-1} &= \{ X_{\underline{2^n-1}} \}
\end{aligned}
\tag{6}
$$

Theorem: A monotone FSR is stable if and only if at least one of the $X_{\underline{i}}$'s, for $\underline{i} \geq \underline{2^{n-1}}$, in each G_s in (6), is equal to zero.

Proof: Let h_s be a subset of H_s, for every H_s in (5), such that $h_s = \{ \underline{i} : X_{\underline{i}} = 0, \underline{i} \geq \underline{2^{n-1}} \}$. Hence for any $\underline{k} \in h_s$, $\underline{B}\,\underline{k} = \underline{\ell}$, where $\underline{\ell} \in H_t$ and $W(H_t) = W(H_s) - 1$. Given any $\underline{i} \in H_s$, $\underline{B}^L \underline{i} = \underline{k}$, for some $\underline{k} \in h_s$, and some positive integer L. Therefore $\underline{B}(\underline{B}^L \underline{i}) = \underline{B}\,\underline{k} = \underline{\ell} \in H_t$. Hence, given any initial state except zero, its weight will

decrease after some finite number of transitions until it reaches state $\underline{0}$. Since $\underline{B}\,\underline{0} = \underline{0}$, this monotone FSR is stable. On the other hand, if the monotone FSR is stable, then there cannot be any cycles in the state diagram, and every state has to reach the zero state. Therefore, at least one of the $X_{\underline{i}}$, for $\underline{i} \geq \underline{2}^{n-1}$, in each G_s has to be zero. Q.E.D.

From (6) it is obvious that $X_{\underline{2}^{n-1}}$ and $X_{\underline{2^n-1}}$ are the only elements with indices larger or equal to $\underline{2}^{n-1}$ in their respective subsets G_1 and G_{2^n-1}. Therefore it follows from the above theorem that $X_{\underline{2}^{n-1}} = X_{\underline{2^n-1}} = 0$.

Golomb and Welch[4] have shown that the number of cycles obtained from the pure cycling register is

$$Z(n) = \frac{1}{n} \sum_{d|n} \phi(d)\, 2^{n/d} \tag{7}$$

where $\phi(d)$ is Euler's ϕ-function, and summation is over all divisors d of n. Since in a Boolean matrix of a monotone FSR we have $X_{\underline{0}} = X_{\underline{1}} = \cdots = X_{\underline{2^n-1}} = 0$, there are exactly 2^{n-1} digits which are not fixed. However, at least Z(n) - 1 digits, one in each G_i, except G_0, in (6), have to be zero. Therefore, the following corollary is immediate:

Corollary: There are at least $2^{2^{n-1}-Z(n)+1}$ distinct stable n-stage monotone FSR's.

A table of Z(n), for n up to 10, is given in page 3, Ref. 3.

Example: Consider the FSR of Fig. 4.

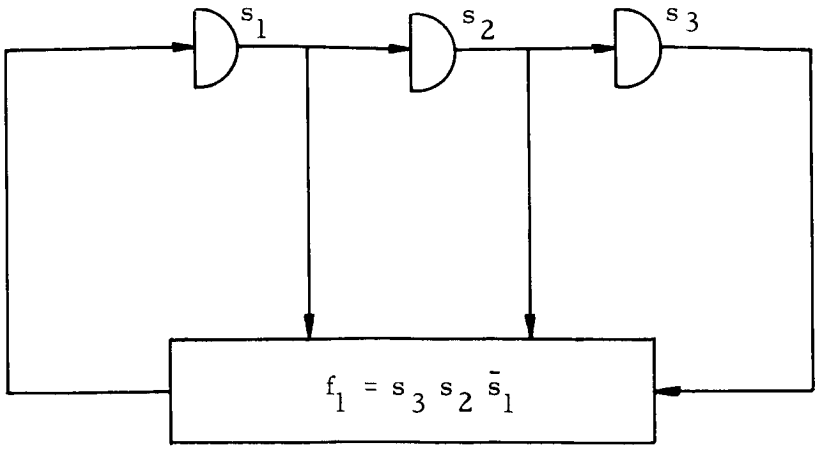


Fig. 4

The Boolean matrix for this FSR is

$$\underline{B}_3 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$G_0 = \{ X_{\underline{0}} \}$

$G_1 = \{ X_{\underline{1}}, X_{\underline{2}}, X_{\underline{4}} \}$

$G_3 = \{ X_{\underline{3}}, X_{\underline{6}}, X_{\underline{5}} \}$

$G_7 = \{ X_{\underline{7}} \}$

$X_{\underline{4}} = X_{\underline{5}} = X_{\underline{7}} = 0$

$\Longrightarrow$ This monotone FSR is stable.

The state diagram of the FSR is shown in Fig. 5.

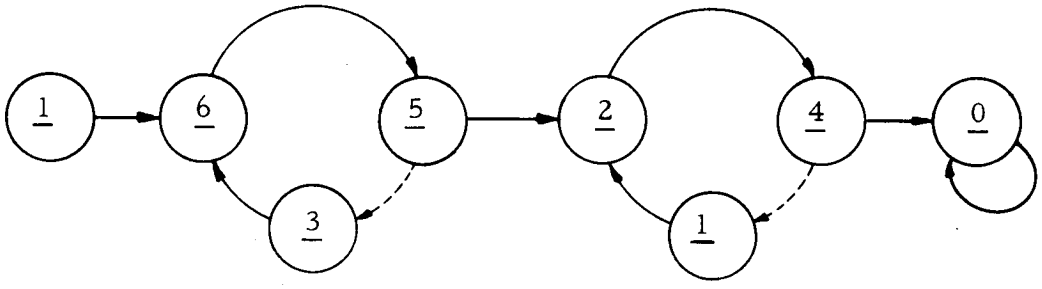


Fig. 5

ACKNOWLEDGMENT

The authors are indebted to Professor E. R. Berlekamp for many useful and stimulating discussions.

REFERENCES

1. J. L. Massey and R. Liu, "Monotone Feedback Shift-Registers," Proceedings of the Second Annual Allerton Conference on Circuit and System Theory, University of Illinois, Sept. 30, 1964.

2. J. O. Campeau, "The Synthesis and Analysis of Digital Systems by Boolean Matrices," IRE Trans., Vol. EC-G(4), Dec. 1957, pp. 231-41.

3. S. W. Golomb, L. R. Welch, and R. M. Goldstein, "Cycles from Nonlinear Shift Registers," JPL Prog. Rpt. No. 20-389, Calif. Inst. of Tech., Aug. 31, 1959.

4. S. W. Golomb and L. R. Welch, "Non-linear Shift Register Sequences," JPL Memo. 20-149, Calif. Inst. of Tech., Oct. 25, 1957.

CODING GRAPHS AND INFORMATION LOSSLESS AUTOMATA†

P. P. Varaiya

Let G be a finite, directed graph with vertices $v \in V$ and labeled arcs satisfying the following conditions: each arc has a label $\sigma \in \Sigma$ attached to it and each label may be used any number of times. There are no restrictions on the number of arcs leaving a vertex, except that there is at least one arc having each vertex (i.e., there are no terminating vertices). A graph satisfying these conditions is called a coding graph.[1]

Let (v, v') be an ordered pair of vertices (not necessarily distinct). We will define three subsets of Σ^+ as follows: *

$x = (\sigma_1, \ldots, \sigma_n) \in F_0(v, v')$ if there is at least one path from v to v' with $(\sigma_1, \ldots, \sigma_n)$ as the labels on each of these paths, in the order in which they take place on this path.

$x = (\sigma_1, \ldots, \sigma_n) \in F_1(v, v')$ if there is exactly one path from v to v' with $(\sigma_1, \sigma_2, \ldots, \sigma_n)$ as the labels on that path

$x = (\sigma_1, \ldots, \sigma_n) \in F_2(v, v')$ if there are at least two paths from v to v' with $(\sigma_1, \ldots, \sigma_n)$ as the labels on each of these paths.

---

*Σ^+ is the free semi-group, under concatenation, of all non-empty finite sequences of elements from Σ.

† The research herein was supported by the National Aeronautics and Space Administration under Grant NsG-354 (S-1).

Clearly $F_0(v, v') = F_1(v, v') \cup F_2(v, v')$ and $F_1(v, v') F_2(v, v') = \phi$: We shall say that $F_1(v, v')$ is the set of tapes of Σ^+ which is information lossless with respect to (v, v'). The reason for this term is that given (v, v') and $(\sigma_1, \ldots, \sigma_n) \in F_1(v, v')$, we can determine the path from v to v' with which that sequence is associated, i.e., $(\sigma_1, \ldots, \sigma_n)$ preserves the information of the path from which it is derived. However, if $(\sigma_1, \ldots, \sigma_n) \in F_2(v, v')$, this information is lost.

We will say that G is information-lossless with respect to (v, v') if $F_2(v, v') = \phi$. We will prove the following theorem.

Theorem: The sets $F_0(v, v')$, $F_1(v, v')$ and $F_2(v, v')$ are regular subsets of Σ^+.

The term "regular subset"[2] is defined as follows:

Definition: A nondeterministic finite automaton (n d f a) is a system $\mathcal{A} = <\Sigma, S, M, A, F>$ where Σ, S are finite nonempty sets, A and F are subsets of S, and M is a map from $\Sigma \times S$ into the power set (set of all subsets) of S.

We will say that a tape $(\sigma_1, \ldots, \sigma_n) \in \Sigma^+$ is accepted by $\mathcal{A}$ if there is a sequence $(s_0, s_1, \ldots, s_n)$ of elements from S such that

a) $s_0 \in A$, $s_n \in F$

b) $s_{i+1} \in M(\sigma_{i+1}, s_i)$ $0 \leq i \leq n-1$.

Let $T(\mathcal{A}) = \{(\sigma_1, \ldots, \sigma_n) \in \Sigma^+ \mid (\sigma_1, \ldots, \sigma_n)$ is accepted by $\mathcal{A}\}$

Definition: A subset $T \subseteq \Sigma^+$ is regular iff $T = T(\mathcal{A})$ for some n d f a.

The theorem will be proved by a sequence of a simple results.

Fact 1. $F_0(v, v')$ is regular.

Proof: Consider the n d f a $\mathcal{A} = <\Sigma, V, M, \{v\}, \{v'\}>$ where $M(\sigma, v_i) = \{v_j \in v \mid \exists$ an arc in G from v_i to v_j with label $\sigma\}$.

Then $(\sigma_1, \ldots, \sigma_n) \in T(\mathcal{U})$ iff $\exists$ a sequence $(v_0, \ldots, v_n)$ with $v_0 = v$, $v_n = v'$, and $\exists$ an arc from v_i to v_{i+1} with label σ_{i+1} for $0 \leq i \leq n-1$; iff $(\sigma_1, \ldots, \sigma_n) \in F_0(v, v')$. Q.E.D.

Now for $n \geq 0$, consider the relations $P_n \subseteq V \times V$ defined below:

$P_0 = \{(v_i, v_i) \mid v_i \in V\}$

$P_1 = \{(v_j, v_\ell) \mid \exists (v_i, v_i) \in P_0; \exists \sigma \in \Sigma;$ $\exists$ two <u>distinct</u> arcs from v_i to v_j and from v_i to v_ℓ with the <u>same</u> label σ

$P_{n+1} = \{(v_j, v_\ell) \mid \exists (v_i, v_k) \in P_n; \exists \sigma \in \Sigma;$ $\exists$ arcs (not necessarily distinct) from v_i to v_j and from v_k to v_ℓ with the <u>same</u> label σ

<u>Fact 2:</u> For $n \geq 1$.

$P_n = \{(v_j, v_\ell) \mid \exists v_i \in V; \exists (\sigma_1, \ldots, \sigma_n) \in \Sigma^+;$ $\exists$ two paths from v_i to v_j and from v_i to v_ℓ such that they have the same labels $(\sigma_1, \ldots, \sigma_n)$ and such that the first arcs on these paths are distinct.

<u>Proof:</u> Follows from the definitions.

<u>Fact 3:</u> a) $P_n = \phi \Rightarrow P_{n+1} = \phi$

b) $P_n = P_{n+k}$ for some $k > 0 \Rightarrow P_{n+j+rk} = P_{n+j}$ $\forall r \geq 0$. $\forall\, 0 \leq j < k$

c) $\exists N \leq 2^{|V|^2}$ ∋ $\forall k > N$, $\exists n \leq N$ such that $P_k = P_n$.

<u>Proof:</u> Follows from the definition of P_n and the observation that

$P_n \subseteq V \times V$ $\forall n$. Q.E.D.

Let $P = \bigcup_{n=1}^{\infty} P_n = \sum_{n=1}^{\infty} P_n$. Let (v_i, v_j) be a fixed pair in V V. Construct an n. d. f. a. (v_i, v_j) as follows:

$(v_i, v_j) = < \Sigma, P_0 \quad P, \{(v_i, v_i)\}, \{(v_j, v_j)\} >$. The function M is defined below:

a) If $(v_1, v_2) = (v_i, v_i)$, then for $\sigma \in \Sigma$,

$M(\sigma, (v_1, v_2)) = \{(v_1', v_2') \mid$ two distinct arcs from v_1 to v_1' and from v_2 to v_2' with the same label $\sigma \}$.

b) If $(v_1, v_2) \neq (v_i, v_i)$, and $(v_1, v_2) \in P$, then for $\sigma \in \Sigma$,

$M(\sigma, (v_1, v_2)) = \{(v_1', v_2') \mid$ arcs (not necessarily distinct) from v_1 to v_1' and from v_2 to v_2' with the same $\sigma \}$.

c) Otherwise, let $M(\sigma, (v_1, v_2) = \phi$. Let

$$U(v_i, v_j) \stackrel{\text{def}}{=\!=} T(\quad (v_i, v_j))$$

Lemma 1: $U(v_i, v_j)$ is the set of all tapes $(\sigma_1, \ldots, \sigma_n) \in \Sigma^+$ such that there exist two paths from v_i to v_j with the same labels $(\sigma_1, \ldots, \sigma_n)$ and such that the first arcs in these paths are distinct.

Proof: Suppose $(\sigma_1, \ldots, \sigma_n) \in U(v_i, v_j)$. Then a sequence

$((\underline{v}_o, \underline{v}'_o), (v_1, v'_1), \ldots, (v_n, v'_n)$ such that

$(\underline{v}_0, \underline{v}'_0) = (v_i, v_i)$, $(\underline{v}_n, \underline{v}_n') = (v_j, v_j)$ and

$(\underline{v}_{i+1}, \underline{v}'_{i+1}) \in M(\sigma_{i+1}, (\underline{v}_i, \underline{v}'_i)) \quad 0 \leq i \leq n-1$.

In particular $(\underline{v}_1, \underline{v}'_1) \in M(\sigma_i, (v_i, v_i))$ so that two distinct arcs with the same label σ_1 from v_i to $\underline{v}_1$ and from v_i to $\underline{v}_1'$. $\therefore$ $(\underline{v}_1, \underline{v}'_1) \in P_1 \quad P$. By induction we can show that $(\underline{v}_i, \underline{v}_i') \in P_i \quad P \quad i$ and that arcs from $\underline{v}_i$ to $\underline{v}_{i+1}$ and from $\underline{v}'_i$ to $\underline{v}'_{i+1}$ with the same label σ_{i+1} $\therefore$ $(\sigma_1, \ldots, \sigma_n)$ satisfies the required condition.

Conversely suppose that $(\sigma_1, \ldots, \sigma_n)$ satisfies the given condition. Then $((\underline{v}_0, \underline{v}'_0), (\underline{v}_1, \underline{v}'_1), \ldots, (\underline{v}_n, \underline{v}'_n))$ such that

$(\underline{v}_o, \underline{v}'_o) = (v_i, v'_i)$; $(\underline{v}_n, \underline{v}'_n = (v_j, v_j$.

$\exists$ two distinct arcs from v_i to $\underline{v}_1$ and from v_i to $\underline{v}'_1$ with the same label σ_1.

$\exists$ arcs from $\underline{v}_i$ to $\underline{v}_{i+1}$ and from $\underline{v}'_i$ to $\underline{v}'_{i+1}$ with the same label σ_{i+1}.

It is clear then that $(\underline{v}_i, \underline{v}'_i) \in P_i \subseteq P$, $i > 0$ and that

$(\underline{v}_{i+1}, \underline{v}'_{i+1}) \in M(\sigma_{i+1}, (\underline{v}_i, \underline{v}'_i))$ $i = 0, \ldots, n-1$

$\therefore (\sigma_1, \ldots, \sigma_n) \in U(v_i, v_j)$ Q.E.D.

We now return to the proof of the theorem.

Suppose $(\sigma_1, \ldots, \sigma_n) \in F_2(v, v')$. Then $\exists$ two <u>distinct</u> paths from v to v' with the same labels $(\sigma_1, \ldots, \sigma_n)$. Since the paths are distinct, there must be an integer

i, $0 \leq i \leq n-1$, such that the paths differ for the first time in their $(i+1)$st arc. It is clear then that

if $i > 1$, $(\sigma_1, \ldots, \sigma_i) \in F_o(v, v_j)$ for some $v_j \in v$

and $(\sigma_{i+1}, \ldots, \sigma_n) \in U(v_j, v')$.

or if

$i = 0$, $(\sigma_1, \ldots, \sigma_n) \in U(v, v')$.

In other words,

$$F_2(v, v') \subseteq U(v, v') \cup \bigcup_{v_j \in v} F_0(v, v_j)\ U(v_j, v')^*$$

Conversely, if $(\sigma_1, \ldots, \sigma_n) \in U(v, v')$, certainly $(\sigma_1, \ldots, \sigma_n) \in F_2(v, v')$, and if $(\sigma_1, \ldots, \sigma_j) \in F_0(v, v_j)$ and $(\sigma_{j+1}, \ldots, \sigma_n) \in U(v_j, v')$, then $(\sigma_1, \ldots, \sigma_n) \in F_2(v, v')$.

<u>Lemma 2</u>: For (v, v') fixed, $F_2(v, v')$ is regular.

Proof: $F_2(v, v') = u(v, v') \cup \bigcup_{v_j \in v} F_o(v, v_j)\ U(v_j, v')$

---

*If A, B are subsets of Σ^+ $AB = \{ xy \mid x \in A, y \in B \}$, AB is called the product of A and B.

Now the sets $U(v, v')$, $F_o(v, v_j)$, and $U(v_j, v')$ are regular. Since the class of regular sets is closed under finite unions and finite products (Ref 2), $F_2(v, v')$ is regular.

Proof of the Theorem: We have shown that $F_o(v, v')$ and $F_2(v, v')$ are regular. Since $F_1(v, v') = F_o(v, v') - F_2(v, v')$ and since regular sets are closed under set differences (Ref. 2), it follows that $F_1(v, v')$ is regular

Q.E.D.

References

1. On Information Lossless Automata, Shimon Even, Sperry Rand Research Report SRRC-RR-63-1.
2. "Finite automata and their decision problems," M. O. Rabin and D. Scott, IBM Journal of Research and Development; April, 1959.

THE SARDINAS AND PATTERSON TEST*

P. P. Varaiya

Let Σ and Δ be two nonempty finite sets. Let Σ^+ and Δ^+ denote the free semigroups generated by Σ and Δ, respectively, i.e., $\Sigma^+(\Delta^+)$ is the free semigroup of all nonempty finite sequences from $\Sigma(\Delta)$ under concatenation. Let $\phi: \Sigma \to \Delta^+$ be an arbitrary one-one map from Σ into Δ^+. Let $\hat{\phi}: \Sigma^+ \to \Delta^+$ be the natural homomorphic extension of ϕ to Σ^+ defined inductively by:

$$\hat{\phi}(\sigma) = \phi(\sigma) \quad \text{for all } \sigma \in \Sigma$$

and

$$\hat{\phi}(x\sigma) = \hat{\phi}(x)\phi(\sigma) \quad \text{for all } x \in \Sigma^+ \ \sigma \in \Sigma.$$

Definition. We say that ϕ is a code iff $\hat{\phi}$ is one-one, i.e., iff $\hat{\phi}$ is an isomorphism of Σ^+ into Δ^+.

Sardinas and Patterson[1] were the first to formulate a test to determine whether or not a given ϕ was a code. Their proof, however, is not completely satisfactory and leaves some questions unanswered. Our proof of their result will answer all these questions.

Our derivation is partly based on an incorrect proof by Schützenberger.[2] With the help of a convenient notation due to Schutzenberger,[2] we formulate the S. and P. test.

Definition. Let $(S, \cdot)$ be an arbitrary semigroup and A, B be (possibly empty) subsets of S. Then

$$A^{-1}B \stackrel{\text{def}}{=\!=} \{s \in S \mid \exists\, a \in A \text{ such that } a \cdot s \in B\}$$

$$=\!= \{s \in S \mid A \cdot s \cap B \neq \phi\}$$

and

$$BA^{-1} \stackrel{\text{def}}{=\!=} \{s \in S \mid \exists\, a \in A \text{ such that } s \cdot a \in B\}$$

$$=\!= \{s \in S \mid s \cdot A \cap B \neq \phi$$

* The research herein was supported by the National Aeronautics and Space Administration under Grant NsG-354 (S-1).

Theorem (Sardinas and Patterson): Let Σ, Δ be nonempty finite sets and let Σ^+, Δ^+ be the free semigroups generated by Σ and Δ, respectively. Let $\phi: \Sigma \to \Delta^+$ be a one-one map and let $\hat{\phi}$ be the natural homomorphic extension of ϕ to Σ^+

Let $P_o = \phi(\Sigma) \subseteq \Delta^+$ and define,

$$P_1 = P_o^{-1}P_o$$

$$P_{n+1} = P_o^{-1}P_n \cup P_n^{-1}P_o \quad \text{for } n \geq 1.$$

Then ϕ is a code $\Leftrightarrow$ $P_n \cap P_o = \phi \; \forall n \geq 1$.

Proof:

I. Notation: Let $x, y \in \Delta^+$. We say that

$$x < y \overset{\text{def}}{\Leftrightarrow} \exists z \in \Delta^+ \ni zx = y$$

We note that $x<y$, $y<w \Rightarrow x < w$.

II. Next. for $n \geq 1$, we define sets $Q_n \subseteq \Delta^+$ as follows:

$q(n) \in Q_n$ iff $q(n)$ satisfies conditions (i) - (vi) below:

(i) $\exists\, p_i \in P_o$ for $1 \leq i \leq n+1$.

(ii) $\exists\, n'$ with $1 \leq n' \leq n$ such that

(iii) $p_1 p_2 \cdots p_{n'} q(n) = p_{n'+1} \cdots p_{n+1}$

(iv) $p_1 \neq p_{n'+1}$

(v) $q(n) < p_{n+1}$ and

(vi) $\forall i > 0, \forall j > 0 \quad p_{1+i} \cdots p_{n'} q(n) \neq p_{n'+j} \cdots p_{n+1}$

III. We shall prove by induction that $P_n = Q_n \; \forall n \geq 1$.

a) $P_1 = Q_1$: By definition of P_1, $q(1) \in P_1$ iff $\exists p_1 \in P_o$, $\exists\, p_2 \in P_o$

such that $p_1 q(1) = p_2$ iff $q(1) \in Q_1$.

b) Assume that $P_n = Q_n$. We shall show that $P_{n+1} = Q_{n+1}$.

Let $q(n+1) \in P_{n+1} = P_o^{-1} P_n \cup P_n^{-1} P_o$. Then $\exists\, p_o \in P_o$, $\exists\, q(n) \in P_n = Q_n$ such that either $p_o q(n+1) = q(n)$ or $q(n)q(n+1) = p_o$. In the first case, by condition (iii) we have,

$$p_1 \cdots p_{n'} p_o q(n+1) = p_1 \cdots p_{n'} q(n) = p_{n'+1} \cdots p_{n+1}.$$

Clearly, $q(n+1)$ satisfies conditions (i) - (iv) and (vi) since $q(n)$ satisfies those conditions. Condition (v) is satisfied because $q(n+1) < q(n)$ and $q(n) < p_{n+1}$, so that $q(n+1) \in Q_{n+1}$.
In the second case, using (iii) again,

$$p_1 \cdots p_{n'} q(n) q(n+1) = p_{n'+1} \cdots p_{n+1} q(n+1) = p_1 \cdots p_{n'} p_o.$$

It is easy to verify that conditions (i) - (vi) are satisfied so that $q(n+1) \in Q_{n+1}$.

Conversely, let $q(n+1) \in Q_{n+1}$. Then,

$p_1 \cdots p_{n'} q(n+1) = p_{n'+1} \cdots p_{n+2}$ where (i) - (vi) are satisfied.

By (v) and (vi), we must have
either $q(n+1) < p_{n+2} < p_{n'} q(n+1)$
or

$$q(n+1) < p_{n'} q(n+1) < p_{n+2}.$$

In the first case let $q(n)q(n+2) = p_{n+2}$. Then,

$$p_{n'+1} \cdots p_{n+1} q(n) = p_1 \cdots p_{n'}.$$

It is clear that conditions (i) - (iv) and (vi) are satisfied by $q(n)$ since $q(n+1)$ satisfies them. Since $p_{n+2} = q(n)q(n+1) < p_{n'} q(n+1)$, we have $q(n) < p_{n'}$ so that (v) holds and hence $q(n) \in Q_n = P_n$. Since $q(n)q(n+1) = p_{n+2} \in P_o$, we have $q(n+1) \in P^{-1} P_o \subsetneq P_{n+1}$.

In the second case, let $q(n) = p_{n'} q(n+1) < p_{n+2}$. Then,

$$p_1 \cdots p_{n'-1} q(n) = p_{n'+1} \cdots p_{n+2}.$$

Surely, $q(n)$ satisfies (i) - (vi) so that $q(n) \in Q_n = P_n$. Since $q(n) = p_{n'} q(n+1)$ we have $q(n+1) \in P_o^{-1} P_n \subseteq P_{n+1}$.

IV. Proof of the theorem:

A. To prove that the condition is necessary, let $q(n) \in P_n \cap P_o = Q_n \cap Q$. Then $\exists\, p_i \in P_o$, $1 \leq i \leq n+1$ such that

$$p_1 \cdots p_{n'} q(n) = p_{n'+1} \cdots p_{n+1} \text{ and } p_1 \neq p_{n'+1}.$$

Let $p_i = \phi(\sigma_i)$ and $q(n) = \phi(\sigma_q)$ with $\sigma_i, \sigma_q \in \Sigma$. Then, $\sigma_1 \neq \sigma_{n'+1}$ and

$$\hat{\phi}(\sigma_1, \ldots, \sigma_{n'}, \sigma_q) = \hat{\phi}(\sigma_{n'+1}, \ldots, \sigma_{n+1}) \text{ so that } \hat{\phi} \text{ is not 1-1.}$$

B. To prove that the condition is sufficient, suppose $\hat{\phi}$ is not one-one. Then $\exists (\sigma_1, \ldots, \sigma_n) \neq (\sigma_1', \ldots, \sigma_n')$ with

$$\hat{\phi}(\sigma_1, \ldots, \sigma_n) = p_1 \cdots p_n = \hat{\phi}(\sigma_1', \ldots, \sigma_n') = p_1' \cdots p_m'.$$

Let i be the integer such that $p_j = p_j'$ $j \leq i$ and $p_{i+1} \neq p_{i+1}'$, then $p_{i+1} \cdots p_n = p_{i+1}' \cdots p_m'$ and $p_{i+1} \neq p'_{i+1}$.

Now let j and k be the smallest positive integers such that $p_{i+1} \cdots p_{i+j} = p_{i+1}' \cdots p_{i+j}'$. It is easy to see that either p_{i+j} or p_{i+k}' belongs to $Q_{j+k-1} \cap P_o$. Q. E. D.

Remarks:

1. The test is valid where Σ and Δ are countably infinite sets.
2. If we define, in a dual manner,

$$\overline{P}_1 = P_o P_o^{-1} \text{ and } \overline{P}_{n+1} = P_o \overline{P}_n^{-1} \cup \overline{P}_n P_o^{-1}$$

we obtain, by using a dual proof that

ϕ is a code $\Leftrightarrow P_o \cap P_n = \phi \quad \forall n \geq 1$.

3. Suppose that Σ is finite and let $P_o = \phi(\Sigma)$.

Let $P = \{x \in \Delta^+ \mid \exists p_o \in P_o$ such that $x \prec p_o\}$. Then P is finite and that $P_n = Q_n \subsetneq P$. From this observation we obtain the following corollary.

Corollary 1:

a) $P_n = \phi \Rightarrow P_{n+1} = \phi$

b) If $P_n = P_{n+k}$ for some n and $k > o$,

then $P_{n+jk+r} = P_{n+r} \quad \forall j \geq o \qquad \forall o \leq r < k$.

c) $\exists N < \infty$ such that $\forall n \geq N \; \exists \; m < N$ such that $P_n = P_m$.

Proof: a) and b) follow from the definition of the P_n's and do not require finiteness of Σ.

c) Let $|P|$ = cardinality of P. Then since $|P| < \infty$ the set R of all subwords of words in P is finite. Let $N = 2|R| + 1$. Then, since each $P_n \subseteq R$ there exist integers ℓ, $k < N$ such that $P_\ell = P_k$. c) then follows from b).

Corollary 2: Let Σ be finite and let ϕ: $\Sigma \to \Delta^+$ be an arbitrary map. Let $\hat{\phi}$ be the homomorphic extension of ϕ to Σ^+. For $n > o$ let $\Sigma_n = \bigcup_{i \leq n} \Sigma^i$. Then $\exists N < \infty$ such that

ϕ is a code $\iff \hat{\phi}$: $\Sigma_N \to \Delta^+$ is one-one.

Proof: "$\Longrightarrow$" is obvious.

"$\Longleftarrow$": Suppose $\hat{\phi}$ is not a code. Then $\exists n < \infty$ such that $q(n) \in P_n \cap P_o$. By Corollary 1 c) we can suppose that $n < N$. It is clear from our proof that $\exists (p_1, \ldots, p_{n+1}) \in P_o$ such that

$p_1 \cdots p_{n'} q(n) = p_{n'+1} \cdots p_{n+1}$ with $p_1 \neq p_{n'+1}$.

Let $\phi(\sigma_i) = p_i$ and $\phi(\sigma_q) = q(n)$. Then $\sigma_1 \neq \sigma_{n'+1}$ and

$\hat{\phi}(\sigma_1, \ldots, \sigma_{n'}, \sigma_q) = \hat{\phi}(\sigma_{n'+1}, \ldots, \sigma_{n+1})$

Since $n < N$, the assertion is proved. Q. E. D.

Definition: Let $x \in \Sigma^+$ and let $x = (\sigma_1, \ldots, \sigma_n)$. We will say that x has length n and denote it by $|x|$.

Definition: Let x and y belong to Σ^+ (Δ^+). We say that

$$x \propto y \iff \exists\, z \in \Sigma^+(\Delta^+) \text{ such that } xz = y$$

and $$x \overset{\propto}{=} y \iff x \propto y \text{ or } x = y.$$

Definition: Let ϕ be a code. Let $\hat{\phi}$ be the extension of ϕ. By the delay of ϕ we mean the smallest integer $\ell \geq o$ such that

$$\forall x \in \Sigma^+, \forall y \in \Sigma^+, \forall z \in \Sigma^+ \text{ with } |z| \geq \ell^*$$

$$\phi(xz) \overset{\propto}{=} \phi(y) \Rightarrow x \overset{\propto}{=} y .$$

If there is no such finite integer, we say that ϕ has infinite delay and write $\ell = \infty$.

Corollary: Let ϕ be a code, and m be the smallest integer such that P_m is empty. Then $m \leq \ell+1$. Also, $\ell = \infty$ iff P_m is nonempty $\forall m$.

Proof: A: "$\ell+1 \leq m$."

If $\ell = o$ the assertion is true because $m \geq 1$.

Suppose $\ell > o$ and suppose that $\ell+1 > m$. Then $\exists x = (\sigma_1, \ldots, \sigma_r)$ $z = (\sigma_{r+1}, \ldots, \sigma_{r+s})$ with $s \geq \ell$, $\exists y = (\sigma'_1, \ldots, \sigma'_n)$ such that

$$\hat{\phi}(xz) \overset{\propto}{=} \hat{\phi}(y) \text{ and}$$

$$x \not\equiv y \text{ i.e., } (\sigma_1, \ldots, \sigma_m) \neq (\sigma'_1, \ldots, \sigma'_j) \quad j.$$

Since $\hat{\phi}$ is 1-1 (2) implies that

$$\hat{\phi}(xz) \propto \hat{\phi}(y).$$

Without loss of generality, we may assume that

$$\sigma_1 \neq \sigma'_1 \text{ and}$$

$$\hat{\phi}(\sigma'_1, \ldots, \sigma'_{n-1}) \propto \hat{\phi}(\sigma_1, \ldots, \sigma_r, \sigma_{r+1}, \ldots, \sigma_{r+s}) \propto \phi(\sigma'_1, \ldots, \sigma'_n). \tag{3}$$

Now let $\phi(\sigma_i) = p_i$ and $\phi(\sigma'_i) = p'_i$. Then (3) can be written as

$$p'_1 \ldots p'_{n-1} \propto p_1 \ldots p_{r+s} \propto p'_1 \ldots p'_n \text{ with } p_1 \neq p'_1.$$

Now let $q \in \Delta^+$ be such that

$$p_1 \ldots p_{r+s} q = p'_1 \ldots p'_n .$$

---

*If $|z| = o$, then $xz = zx = x \; \forall\, x \in \Sigma^+$.

We claim that $q \in Q_{r+s+n-1} = P_{r+s+n-1}$.

Clearly q satisfies conditions (i) - (v) given in the proof of the theorem. Suppose $\exists\, i > o$, $\exists\, j > o$ such that

$$p_{1+i} \cdots p_{r+s} q = p'_{1+j} \cdots p'_n .$$

But then $p_1 \cdots p_i = p'_1 \cdots p'_j$. Since ϕ is a code, this means that $(\sigma_1, \ldots, \sigma_i) = (\sigma'_1, \ldots, \sigma'_j)$ so that $\sigma_1 = \sigma'_1$, which implies $p_1 = p'_1$, which is a contradiction

$$\therefore \quad q \in P_{r+s+n-1} .$$

$$\therefore \quad m \geq r+s+n-1 \geq 1+\ell+1-1 = \ell+1$$

B. "$m = \infty \implies \ell = \infty$"

Suppose P_m is nonempty for all m. Let n be any fixed integer. We will show that $\ell \geq n$.

Let $q(m) \in Q_m = P_m$. Then, $\exists\, p_i \in P_o$ $\quad 1 \leq i \leq m+s$ such that

$$p_1 \cdots p_{m'} q(m) = p_{m'+1} \cdots p_{m+1} \quad \text{and} \quad p_1 \neq p_{m'+1}$$

It is clear that by taking m sufficiently large we can make $m' > n+1$. Let $\phi(\sigma_i) = p_i \;\forall\, i$, and choose $x = \sigma_1$, $z = \sigma_2 \cdots \sigma_{m'}$ and $y = \sigma_{m'+1} \cdots \sigma_{m+1}$. Then $|z| = m'-2 > n$ and

$\phi(xz) \wedge \phi(y)$ but $x \not\wedge y$ since $\sigma_1 \neq \sigma_{m'+1}$. Q.E.D.

References

1. Sardinas and Patterson, "A Necessary and Sufficient Condition for Unique Decomposition of Coded Messages," IRE Convention Record, Part 8, 1953.

2. P. Dubreil et C. Pisot, Algebre et Théorie des Nombres, Seminaire; 1955/56.

SHADOWS OF FUZZY SETS*

L. A. Zadeh

The concept of a fuzzy set and some of its implications were discussed in a recent report.[1] In this note, we shall focus our attention on the properties of shadows of fuzzy sets.

In order to make our discussion self-contained, we shall begin by recapitulating several concepts relating to fuzzy sets which will be needed in the sequel.

1. Roughly speaking, a fuzzy set is a "class" of objects in which there may be grades of membership intermediate between full membership and non-membership. Thus, a fuzzy set A in a space $X = \{x\}$ is characterized by a membership function μ_A which associates with each point x in X a real number $\mu_A(x)$ in the interval $[0,1]$, with $\mu_A(x)$ representing the grade of membership of x in A.

2. A fuzzy set A is said to be contained in a fuzzy set B, written as $A \subset B$, if and only if $\mu_A(x) \leq \mu_B(x)$ for all x in X.

3. The union of two fuzzy sets A and B is denoted by $A \cup B$. If $C = A \cup B$, then by definition $\mu_C(x) = \text{Max}[\mu_A(x), \mu_B(x)]$, $x \in X$, meaning that the grade of membership of x in $A \cup B$ is the larger of the grades of membership of x in A and B.

4. Similarly, the intersection of two fuzzy sets A and B is denoted by $A \cap B$. If $C = A \cap B$, then by definition $\mu_C(x) = \text{Min}[\mu_A(x), \mu_B(x)]$, $x \in X$.

---

*The research herein was supported by National Aeronautics and Space Administration under Grant NsG-354 (S-2).

In what follows, it will be assumed throughout that $X = E^n$ = Euclidean n-space. In such a space, a fuzzy set A is convex if and only if the sets $\Gamma_\alpha = \{x | \mu_A(x) \geq \alpha\}$ are convex for all $\alpha > 0$. Equivalently, A is convex if and only if the inequality

$$\mu_A(\lambda x_1 + (1-\lambda)x_2) \geq \text{Min}(\mu_A(x_1), \mu_A(x_2)) \tag{1}$$

holds for all x_1, x_2 in E^n and all λ in the interval $[0,1]$.

We are now ready to define what is meant by a shadow of a fuzzy set. Thus, let p_0 and H be, respectively, a point and a hyperplane in E^n. Then, a point-shadow of A on H is a fuzzy set S(A) in H whose membership function $\mu_{S(A)}(x)$ is defined as follows: Let L be a line passing through p_0, with L intersecting H at a point h. Then,

$$\begin{aligned} \mu_{S(A)}(h) &= \underset{x \in L}{\text{Sup}}\, \mu_A(x) \\ \mu_{S(A)}(x) &= 0 \quad , \quad x \notin H \end{aligned} \tag{2}$$

Note that we use the suggestive term "point-shadow" to describe this fuzzy set because it bears resemblance to the shadow thrown by a cloud A on a plane H, with p_0 acting as a point source of light.

The transformation S which takes A into S(A) will be referred to as point-projection of A on H with respect to p_0. In the special case where p_0 is a point at infinity and the lines L are orthogonal to H, we shall refer to S(A) and S as orthogonal shadow and orthogonal projection, respectively. For example, if H is the coordinate plane $H = \{x | x_1 = 0\}$, $x = (x_1, \ldots, x_n)$ then the orthogonal shadow of A on H is characterized by the membership function

$$\mu_{S(A)}(x_2, \ldots, x_n) = \operatorname{Sup}_{x_1} \mu_A(x_1, \ldots, x_n) \quad , \quad x \in H$$
$$= 0 \quad , \quad x \notin H \tag{3}$$

In the sequel, we shall frequently use the terms shadow and projection without the adjectives "point" or "orthogonal," relying on the context to indicate the specific meaning in which these terms should be understood.

We proceed to establish several basic properties of shadows of fuzzy sets. Most of these properties are immediate consequences of the defining relation (2).

Homogeneity. Let kA denote a fuzzy set whose membership function is given by

$$\mu_{kA}(x) = k\mu_A(x) \tag{4}$$

where k is a constant, $0 \leq k \leq 1$. Then clearly

$$S(kA) = kS(A) \tag{5}$$

Monotonicity. This property is expressed by the relation

$$A \subset B \Longrightarrow S(A) \subset S(B) \tag{6}$$

and is an immediate consequence of

$$\forall x\,[\mu_A(x) \leq \mu_B(x)] \Longrightarrow \operatorname{Sup}_L \mu_A(x) \leq \operatorname{Sup}_L \mu_B(x)$$

Distributivity. For any fuzzy sets A and B, we have

$$S(A \cup B) = S(A) \cup S(B) \tag{7}$$

which implies that S is distributive with respect to U. This follows at once from the identity

$$\underset{L}{\text{Sup}}\ \text{Max}(\mu_A(x), \mu_B(x)) = \text{Max}\left(\underset{L}{\text{Sup}}\ \mu_A(x), \underset{L}{\text{Sup}}\ \mu_B(x)\right) . \tag{8}$$

In connection with (8), it is natural to raise the question: Is S distributive with respect to $\cap$, that is, is it true that

$$S(A \cap B) = S(A) \cap S(B) . \tag{9}$$

In this case, the corresponding relation in terms of membership functions reads

$$\underset{L}{\text{Sup}}\ \text{Min}\ (\mu_A(x), \mu_B(x)) = \text{Min}\ (\underset{L}{\text{Sup}}\ \mu_A(x), \underset{L}{\text{Sup}}\ \mu_B(A)) \tag{10}$$

This relation is not valid for arbitrary $\mu_A(x)$ and $\mu_B(x)$. However, it can be made valid by suitably restricting $\mu_A(x)$ and $\mu_B(x)$, as is done in the case of the minimax theorem.[2]

Note that by combining (5) and (1), we have for any constants k_1 and k_2 in $[0,1]$,

$$S(k_1 A \cup k_2 B) = k_1 S(A) \cup k_2 S(B) \tag{11}$$

This identity indicates that S is a linear transformation, with the restriction that $k_1, k_2 \in [0,1]$. Note also that S is idempotent, i.e., $S^2(A) = S(S(A)) = S(A)$.

Invariance of convexity. Let A be a convex fuzzy set in E^n. Then S(A) is a convex fuzzy set in H.

Proof. We shall prove this assertion under the simplifying assumption that, for all L, $\sup_L \mu_A(x)$ is attained for some x, so that Sup can be replaced by Max.

Let L_1 and L_2 be two lines passing through p_0, and let h_1 and h_2 be the points at which they intersect H. Let $\mu_{S(A)}(h_1) = M_1$, $\mu_{S(A)}(h_2) = M_2$, and let x_1 and x_2 be points on L_1 and L_2, respectively, at which $\mu_A(x)$ attains the values M_1 and M_2. Then, by the convexity of A, for any point x on the line segment $[x_1, x_2]$, $x = \lambda x_1 + (1 - \lambda)x_2$, $0 \leq \lambda \leq 1$,

$$\mu_A(x) \geq \text{Min}(M_1, M_2)$$

Now, consider a line L passing through p_0 and x. This line will intersect H at the point $h = \lambda h_1 + (1 - \lambda)h_2$, and by the definition of S(A) we can write

$$\mu_{S(A)}(h) \geq \mu_A(x) \geq \text{Min}(M_1, M_2), \quad h \in [h_1, h_2] \tag{12}$$

which shows that S(A) is convex.

Equality of convex fuzzy sets. A useful property of shadows of convex sets is expressed by: If A and B are convex sets and S(A) = S(B) for all p_0 (and a fixed H), then A = B.

Proof. It will be sufficient to show that if $A \neq B$, then there exists a p_0 such that $S(A) \neq S(B)$.

Assuming that $A \neq B$, let x_0 be a point at which $\mu_A(x_0) \neq \mu_B(x_0)$, e.g., for concreteness, $\mu_A(x_0) = \alpha > \mu_B(x_0) = \beta$. Since B is a convex set, the set $\Gamma_\beta = \{x \mid \mu_B(x) > \beta\}$ is a convex set and hence there exists

a hyperplane F supporting Γ_β and passing through x_1. In relation to F, we have $\mu_B(x) \leq \beta$ for all x on F and on the side of F not containing Γ_β.

Now let p_0 be an arbitrarily chosen point on F, and let L be a line passing through p_0 and x_0. At the intersection, h, of this line with H (which may be at infinity), we have

$$\mu_B(h) \leq \beta$$

but on the other hand $\mu_A(h) \geq \alpha$ since $\mu_A(x_0) = \alpha$. Consequently, $\mu_A(h) \neq \mu_B(h)$. Q. E. D.

In the case of orthogonal shadows, the statement of the property in question becomes: If A and B are convex sets and S(A) = S(B) for all H, then A = B. More generally, if A and B are not necessarily convex, then the conclusion A = B would be replaced by the weaker equality conv A = conv B, when conv A denotes the convex hull of A, that is, the smallest convex fuzzy set containing A.

References

1. L. A. Zadeh, "Fuzzy Sets," ERL Report No. 64-44, November 16, 1964. (To be published in Information and Control.)
2. S. Karlin, Mathematical Methods and Theory in Games, Programming and Economics, (Addison-Wesley Publishing Co., Inc., Reading, Mass.), p. 28 et seq.

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Electronics Research Laboratory University of California, Berkeley	UNCLASSIFIED
	2b. GROUP

3. REPORT TITLE

NOTES ON SYSTEM THEORY, VOLUME VII

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Research Report

5. AUTHOR(S) *(Last name, first name, initial)*

Various

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
May 1965	173	67

8a. CONTRACT OR GRANT NO. Various	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO.	65-14
c.	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d.	

10. AVAILABILITY/LIMITATION NOTICES

Qualified requesters may obtain copies of this report from DDC

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Various

13. ABSTRACT

This is the seventh issue of Notes on System Theory. The purpose of these notes is twofold: first, to provide an auxiliary publication medium for short contributions by students and faculty engaged in research in systems and related areas; second, to contribute to the development of system theory as a basic scientific discipline.

DD FORM 1 JAN 64 1473

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
coding						
sensitivity networks						
bio-electronics						
circuits						
stability theory						
matrixes						
detection						
sequential machines						
automata						
signal flow graphs						
polynomials						
channel capacity						
fuzzy sets						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C), or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.