Thesis Entitled

THE OPTIMAL DIGITAL CONTROL AND DESIGN OF DYNAMIC SYSTEMS

Submitted to the University of Glasgow

by

Robert Ronald Duncan, B.Sc.

In Fulfilment of the Requirements for the Degree of

Doctor of Philosophy.

November 1966

Acknowledgements

The author would like to express his appreciation for the guidance of Professor G.D.S. MacLellan and Dr. H.A. Barker, under whom the research reported in this thesis was conducted.

During the three year period of research, the author was financially supported by a Fellowship awarded by the Ford Trust, to the executives of which the author is deeply indebted.

CONTENTS

NOTATION

The author has defined the notation in the thesis at the time of use. Those important variables used universally throughout the thesis are detailed below.

A	Continuous state transition matrix.
$B(N-j)$	Optimal feedback coefficient matrix.
$C(n-1)$	Condition matrix for controllability.
$\bar{C}(n-1)$	Condition matrix for observability.
D	Continuous control transition matrix.
E	Statistical expectation.
$f_{N-j}[x(j)]$	Performance index at time $(N-j)T$.
G	Discrete control transition matrix.
H	Control cost matrix.
H^*	Effective intersample control cost matrix.
I	Unit matrix.
J	Error covariance matrix.
$m(jT)$	Control input at time jT.
N	Optimisation interval.
$P(N-j)$	Equivalent discrete riccati matrix.
Q	State cost matrix.
Q^*	Effective intersample state cost matrix.
Q^{**}	Effective intersample - progressive stability state cost matrix.
$R(j)$	Discrete disturbance covariance matrix at time jT.

$R(N-j)$	Partitioned part of E.D.R. matrix.
$r(jT)$	Discrete disturbance input vector at time jT.
T	Sampling period.
t	Time.
U	Matrix used in intersample theory.
U^*	Matrix used in intersample - progressive stability theory.
$V(N-j)$	Partitioned part of E.D.R. matrix.
$V(j)$	Matrix used in estimation theory.
$W(j)$	Discrete measurement covariance matrix at time jT.
$w(jT)$	Discrete measurement input vector at time jT.
$x(jT)$	State vector or system variable vector at time jT.
$y(jT)$	State vector or system variable vector at time jT.
$Z(N-j)$	Partitioned part of E.D.R. matrix.
$z(jT)$	State vector or system variable vector at time jT.
ϕ	Uncontrolled system's discrete state transition matrix.
Φ	Controlled system's discrete state transition matrix.

CHAPTER 1 - INTRODUCTION

1.1 PRELIMINARY REMARKS

The necessity of improving the efficiency of industrial processes and the ever increasing demand for high performance systems have resulted in a reawakening of research activity in the field of optimal control theory. The classical interrelated problems of Lagrange, Bolza and Mayer, first analysed in the nineteenth century with the aid of the specifically devised techniques of the Calculus of Variations (1), could be considered as the first mathematically orientated study of the general control optimisation problem. The recent resurgence of activity received much of its initial momentum from the publication of the work of two eminent applied mathematicians, namely L. S. Pontryagin and R. E. Bellman. The Maximum Principle of Pontryagin (2) and the Principle of Optimality of Bellman (3), along with the classical theories of the calculus of variations, now forms a rigorous mathematical foundation upon which to base the solution of complex optimisation problems.

The optimisation problem is essentially the determination of the open or closed loop control inputs, which, when applied to the dynamic system, results in a response which minimises some functional which, it has already been decided, characterises the

performance of the system. Certain constraints may also be imposed concerning the admissability of the control inputs or the resulting response of the controlled system. The control inputs so determined are called the optimal control inputs, or the optimal control policy of the system.

In the case of non-linear dynamic systems, the optimisation problem eventually evolves into the determination of the solution of a two point boundary value problem, which is notoriously difficult in all but a few cases. In recent years many papers have been published which, when dealing with the complex optimisation problem, have as their net result the reduction of the problem to a correspondingly complex two point boundary value problem. On the other hand, relatively few papers have dealt with the fundamental problem of immense practical importance, namely, the solution of the typical two point boundary value problems found in optimal control theory. The author has found the most notable papers in this vein have been those of Bryson and Denham (4) (5), Westcott, Florentin and Pearson (6), Noton (7) and Levine (8). The methods presented therein necessitate, in general, the use of large high-speed digital computers. In the solution of the two point boundary value problem, the time taken to converge to an optimal solution, if convergence actually occurs, is not known at the outset of the calculation. This lack of information is a major drawback in the utilisation of

such an optimisation procedure in an on-line capacity. The logical development is the approximation of the non-linear problem by a corresponding linear problem for which a quasi-analytic solution can be derived and hence its on-line use is feasible. Satisfactory results have been reported by Pearson (9) and Davis (10) using such a procedure.

In order to obtain a quasi-analytic solution, not only must one have a linear system, but one is also restricted, at least at present, to a quadratic performance criterion. The continuous optimisation problem then reduces to the determination of the solution of the well known matrix ricatti equation, which is relatively easily accomplished by the adjoint variable technique (9), (11), (12). The optimal control input in this case turns out to be a linear function of, in general, all the state variables of the system. The resulting system therefore has a closed loop feedback structure, as opposed to the open loop structure which results from the implementation of the optimal input of a non-linear system. Classical control engineering analysis (13) has shown that a closed loop system is a desirable structure, a conclusion which has been confirmed as a by-product of calculating the optimal control input of a linear system. Although the optimal non-linear system design results in an open loop structure in practice, it seems reasonable to assume that it is, in fact, a non-

linear feedback system, but the methods of analysis are such that the problem cannot be solved in such a way as to permit feedback implementation. This thesis is concerned with the digital optimal control problem; therefore it is not intended to discuss in depth the optimal control techniques applicable to continuous systems.

A digital control system is defined as a dynamic system in which the control input is applied to the system intermittently. One of the reasons why such a system arrangement is of practical significance is that if any form of digital device is required in the system to calculate the control inputs etc., the output of that device will be intermittent. Assuming the continuation of present trends, it is likely that digital computers will be incorporated in complex control systems, performing such functions as dynamic optimal control generation and system identification. As yet, the majority of computers which have been included in control systems have been used to tackle the static optimisation problem. Nevertheless, recent results would suggest that a further improvement in system efficiency could be made by dynamic optimal control (14). The process control industries will undoubtedly be one of the major users of computer control, and therefore it is desirable that any digital optimal control technique devised should be applicable to systems with typical process dynamics. Although, as in the continuous case, it is necessary to restrict the system's dynamics to be linear and the

criterion to be quadratic, to obtain a quasi-analytic solution, it will be shown in the following chapters that it is possible to include systems with multiple state, control and measurement delays and with additive correlated noise disturbance and measurement inputs.

The approach in this thesis has been to assume the presence of a computer in the system, and hence derive the optimal digital control input for a linear dynamic system. The dynamics are assumed independent of time, although in fact they may be slowly varying with the corresponding intermittent updating of the optimal control calculations. Wherever possible, particular attention has been given to designing good sub-optimal digital control systems where the reduction in system complexity compensates for the loss of optimality.

1.2 - THE OPTIMAL DIGITAL CONTROL PROBLEM

The Maximum Principle was originally devised for continuous systems, but has recently been extended to deal with discrete systems with limited success (15) (16). On the other hand, the Principle of Optimality and the associated Dynamic Programming theory are essentially discrete in concept and therefore easily applied to the solution of the optimal digital control problem. For this reason, the dynamic programming approach has been used in this thesis.

In an optimal digital control system, the choice of control inputs is made at each sampling instant in such a way that the performance criterion governing the response of the system over some specified number of sampling intervals, N, is minimised. The optimal digital control

problem is therefore clearly seen to be an N stage control input decision process, which is, in general, difficult to solve. In analogy with continuous systems, the set of inputs which result from making the N stage decisions is called the optimal control input policy of the system.

The intuitive concept of the Principle of Optimality can be stated as - An optimal control input policy has the property that whatever the initial state and initial control decisions are, the remaining control decisions must constitute an optimal control policy with regard to the state resulting from the preceding control input decisions. Thus by making the control input decisions in reverse order to which they will actually be used, reduces the N stage decision process to N single stage decision processes, which are readily solvable. Since the decision making has to be performed in reverse time, one requires to have a knowledge of the state of the system at each sampling instant before the input decisions at all the preceding sampling instants have been made. The lack of this information about the state of the system is the stumbling block which prevents the direct application of the dynamic programming techniques to non-linear dynamic systems. Although theoretical methods can easily be developed (3), their implementation, for all but a few simple problems, would require a digital computer with storage capacity and speed several magnitudes greater than those even anticipated today.

In mathematical terms, the Principle of Optimality can be stated as

$$f(\overline{N-j}\ T) = g(t) + f(\overline{N-j+1}\ T) \qquad (1.7.1$$

with $f(\overline{N+1}\ T) = 0$

where T is the sampling period,

N is the total number of sampling intervals over which the system has to be optimally controlled,

$f(\overline{N-j}\ T)$ and $f(\overline{N-j+1}\ T)$ are the optimal values of the performance index calculated from the initial sampling instant to j th and j+1 th sampling instant respectively,

and $g(t)$ is the value of optimal performance index during the time interval $jT \leq t \leq \overline{j+1}\ T$.

It was pointed out by Kalman and Koepcke (17) that if one restricted the system to have linear deterministic dynamics and the performance criterion to be quadratic in the state and control variables of the system, thefunctional $f(\overline{N-j}\ T)$ could be expressed as a function of the state $x(jT)$, namely

$$f(\overline{N-j}T) = x'(jT)\ P\ (\overline{N-j}\ T)\ x(jT) \qquad (1.7.$$

where $P(N-j)$ is a square, symmetric, non-negative definite matrix of dimensions equal to the order of the system.

This equality is analogous to a similar expression, which is obtained for continuous systems under the same restrictions. In that case the matrix $P(\overline{N-j}\ T)$ is the ricatti matrix and therefore, for notational convenience, the author has called the matrix $P(N-j)$

the equivalent discrete ricatti matrix or the E.D.R. matrix.

The use of the relationships of equations (1.2.1.) and (1.2.2.), hence-forth expressed in a more precise notation as,

(a) $f_{N-j}\,[x(jT)] = g[x(jT),\ m(jT)] + f_{N-\overline{j+1}}\,[x(\overline{j+1}\ T)]$ (1.2.3.

with $f_{N+1}\,[x(-T)] = 0$

and $x(jT)$ and $m(jT)$ the state and control input of the system at sampling instant jT.

(b) $f_{N-j}[x(j\dot{T})] = x'(jT)\ P(\overline{N-j}T)\ x(jT)$ (1.2.4

enable a quasi-analytic method of determining the optimal control policy for the digitally controlled system as will be shown in Chapter 2. As in the pure dynamic programming approach, the control inputs are again calculated in the reverse order to that in which they would be used.

The original optimisation technique (17) was developed correctly by Gunckel and Franklin (18) and wrongly by Tou (19) and Joseph (20) to include linear systems with additive noise disturbances with zero mean. A further development which allowed the amplitude of the control variable inputs to be constrained within prescribed inputs was indicated by Deley and Franklin (21), but the practicability of the proposed technique is questionable, from a computational point of view, for systems of even modest complexity.

The research reported in this thesis was motivated by

several factors. Firstly, the author was interested in the possibility of extending the basic optimisation technique to a larger class of systems and in particular to carry out the optimisation with respect to a more design orientated performance criterion than those used by the authors of the above mentioned papers (17), (18), (19), (20). Secondly, to the author's knowledge, there was no recorded discussion of the computational aspects of even the simple optimal digital control technique, a fact most certainly due to the very few numerical calculations which have been attempted and reported. Thirdly, there was the problem of implementing the optimal control possibly in a sub-optimal manner if such a procedure was judged advantageous. Finally, the author was interested in exactly how such factors as the systems eigenvalues, the sampling period and the compatibility of the performance criterion with the system's dynamics would affect the generation of the optimal control policy and the resulting performance index of the optimal system. The results of the author's investigations are reported in the remainder of this thesis.

CHAPTER 2 - OPTIMAL DIGITAL CONTROL TECHNIQUES

2.1. - STATE VARIABLE DESCRIPTION OF DYNAMIC SYSTEMS

2.1.1. - CONCEPT OF THE STATE OF A SYSTEM

In the study of the optimisation problem in control theory, it has been found that it is more convenient mathematically to describe the dynamics of a system in the time domain rather than in the frequency domain, the latter being mainly used in more classical control theory. In order to develop a generalised form of the description of the system's dynamics in the time domain, the concept of a system's state and its state vector has been introduced (22), (23). A complete understanding of the meaning of a system's state is essential to the comprehension of the methods of solution to be used throughout this thesis. Therefore it was thought worthwhile to include a concise note on this topic.

The author has found it convenient to define the state of a dynamic system to be the minimum amount of information about the system at time t_o which, along with a knowledge of the dynamic characteristics of the system and the inputs to the system during the time interval $t-t_o$ allows the response of the system at time t, $t \geqslant t_o$, to be calculated. The state vector has then been defined as the column vector of minimum dimensions, a knowledge of the past and present history of whose elements is equivalent to a knowledge of the state of the system.

Consider a system whose dynamics are described by a p th order differential equation. It can be shown that this single equation can be represented by a non-unique set of p 1st order differential equations, linear systems retaining their linear properties (24). Let the p time dependent variables used in these p equations be defined as the system variables. It should be noted that these 1st order equations will also contain any input variables appearing in the p th order equation. In order to calculate the response of the system under discussion at some future time, one would only require a knowledge of the system variables at the present time t_o along with a knowledge of the dynamic characteristics of the system and the inputs to the system in time interval $t-t_o$. The vector of the system variables can therefore be seen to be equivalent to the state vector previously defined. The system variables are thus defined to be the state variable, i.e. variable which describe the system's state. To simplify mathematical manipulation, the p 1st order differential equations are henceforth represented in matrix form,

$$\dot{x}(t) = f(x(t), m(t), t)$$

where $x(t)$ is the state vector,

$m(t)$ is the control input vector,

t is the independent time variable,

$f(x(t), m(t) t)$ is a vector of functions of its arguments.

The above argument has considered the special case of a system described by differential equations, but similar remarks also hold for

systems described by difference equations (25), (26). In the case of systems described by partial differential equations one could also theoretically define a state vector, but the latter would be infinitely large since the elements of the state vector would have to represent continuous distributions rather than a finite number of variables. Approximations to the infinite state vector can be made, as will be shown when the multiple state and control delay system, a subset of the general partial differential class of systems, is solved in subsection (2.1.4.)

2.1.2. - THE AUGMENTED STATE VECTOR

From the definition of the state vector, it can be deduced that if two systems with state variable equations

$$\dot{x}(t) = f\big(x(t),\ m(t),\ t\big) \tag{2.1.}$$

and $$\dot{y}(t) = g\big(y(t),\ u(t),\ t\big) \tag{2.2.}$$

are interconnected through a system with no dynamics, then one of the state vectors of the composite system would be the partitioned vector $z(t)$,

$$z(t) = \big[x(t) \,\vdots\, y(t)\big]'$$

Consider a control system described by equation (2.1.) which is subject to an additive disturbance input described by equation (2.2.) The above statement indicates that the response of the composite

disturbed system can be studied by considering the single augmented state system

$$\dot{z}(t) = h_1\left(z(t),\ m(t),\ u(t),\ t\right)$$

It can also be deduced that if no direct interconnection existed between the two systems, then the two separate systems could be examined together by considering the single augmented state system

$$\dot{z}(t) = h_2\left(z(t),\ m(t),\ u(t),\ t\right)$$

where $h_2\left(z(t),\ m(t),\ u(t),\ t\right) = \begin{bmatrix} f(x(t),\ m(t),\ t) \\ g(y(t),\ u(t),\ t) \end{bmatrix}$

The augmented state variable approach to systems analysis indicated above is used extensively in deriving the general solution of the digital optimisation problem presented later in this chapter.

2.1.3. - SOLUTION OF STATE EQUATION $\dot{x}(t) = f\left(x(t),\ m(t),\ t\right)$

2.1.3. - SOLUTION OF STATE EQUATION $\dot{x}(t) = f\left(x(t),\ m(t),\ t\right)$

The discrete dynamic programming optimisation problem requires the state at one sampling instant to be a known function of the state at the preceding sampling instants. The state equations governing the system must therefore be represented in the form of a matrix difference equation. Such a representation of the general matrix state vector differential equation will now be calculated and, by an extension of the solution technique adopted, an approximate solution of the general state vector difference - differential equation will be obtained.

Consider the matrix differential equation

$$\dot{x}(t) = f\big(x(t),\ m(t),\ t\big) \tag{2.3.}$$

Equation (2.3.) is satisfied by the integral equation

$$x(t) = x(t_o) + \int_{t_o}^{t} f\big(x(r),\ m(r),\ r\big)\, dr \tag{2.4.}$$

If one regards $x(t)$ appearing in equation (2.4.) to be unknown, then equation (2.4) can be solved by the method of successive approximations (24) and thence it can be shown that the function $x(t)$ defined in equation (2.4) is the general solution of equation (2.3). In particular, the linear time varying coefficient differential equation corresponding to equation (2.3) i.e.

$$\dot{x}(t) = A(t)\, x(t) + D(t)\, m(t) \tag{2.5.}$$

has solution

$$x(t) = x(t_o) + \int_{t_o}^{t} \big[A(q)x(q) + D(q)m(q)\big]\, dq \tag{2.6.}$$

Considering t to be a dummy variable, one can obtain expressions

$$x(q) = x(t_o) + \int_{t_o}^{q} \big[A(r)x(r) + D(r)m(r)\big]\, dr$$

$$x(r) = x(t_o) + \int_{t_o}^{r} \big[A(s)x(s) + D(s)m(s)\big]\, ds$$

etc.

The successive substitution of these expressions into equation (2.6) would eventually yield the infinite series solution

$$x(t) = \sum_{i=0}^{\infty} \alpha_i(t, t_o)\, x(t_o) + \sum_{i=0}^{\infty} \beta_i(t, t_o) \tag{2.7.}$$

where $\alpha_o\ (t, t_o) = I$

$$\alpha_i\ (t, t_o) = \int_{t_o}^{t} A(r)\, \alpha_{i-1}\ (r, t_o)\, dr \quad i = 1, 2 \ldots$$

$$\text{and } \beta_o(t,t_o) = \int_{t_o}^{t} D(r)\, m(r)\, dr$$

$$\beta_i(t,t_o) = \int_{t_o}^{t} A(r)\beta_{i-1}(r,t_o)\, dr \qquad i=1,\ 2\ \ldots$$

If the equation (2.5) has time invariant coefficients, then equation (2.7) can be reduced to a simpler form, i.e.

$$x(t) = x(t_o) + AT\,x(t_o) + A^2T^2/2!\ \ x(t_o) + \ldots.$$
$$+ D\int_{t_o}^{t} m(q)dq + AD\int_{t_o}^{t}\int_{t_o}^{q} m(r)\, dr\, dq + \ldots. \qquad (2.8.)$$

where $T = t-t_o$.

If it is further assumed that $t-t_o$ is the sampling period of the digital control system, then, by definition of the latter, the input variable $m(r)$ has a constant value during this period equal to $m(t_o)$. The equation then becomes

$$x(t) = \sum_{i=0}^{\infty}(AT)^i/i!\ x(t_o) + \sum_{i=1}^{\infty} A^{i-1}T^i/i!\ Du(t_o) \qquad (2.9.)$$

Replacing the infinite series by the variables used in later sections of this thesis, the difference equation solution of the differential equation (2.5.) is

$$x(t) = \phi(t-t_o)\,x(t_o) + G(t-t_o)\,m(t_o) \qquad (2.10.)$$

where $\phi(t-t_o) = \sum_{i=0}^{\infty}(AT)^i/i!$ (commonly denoted by $\exp(AT)$)

and $G(t-t_o) = \sum_{i=1}^{\infty} A^{i-1}T^i/i!\ D = \int_{t_o}^{t}\phi(t-r)\, D\, dr$

Having obtained the matrix infinite series $\phi(t-t_o)$, its convergence properties must be established before equation (2.10) can be shown to be practically useful. The scalar form of the

series is well known to be convergent for all finite T, but the matrix form requires proof (22).

Let S_r denote the finite series $\sum_{k=0}^{r} (TA)^k/k!$ where A is a p × p matrix.
Let the (i,j)th element of S_r be denoted by $\sum_{k=0}^{r} a(k)_{ij}\, T^k/k!$
Let $\alpha = \max_i \left(\sum_{j=1}^{p} |a(k)_{ij}| \right)$ where maximisation is with respect to the p rows of matrix S_r.

$$\text{Then } \max_{ij} |a(k+1)_{ij}| = \max_{ij} \left| \sum_{l=1}^{p} a_{il} a(k)_{lj} \right|$$

$$\leq \max_{ij} \left(\sum_{l=1}^{p} |a_{il}|\, |a(k)_{lj}| \right)$$

$$\leq \left(\max_i \sum_{l=1}^{p} |a_{il}|\right)\left(\max_{lj} |a(k)_{lj}|\right)$$

$$\text{Hence } \max_{ij} |a(k+1)_{ij}| \leq \alpha \max_{lj} |a(k)_{lj}|.$$

$$\text{For } n = 0, \quad \max_{ij} |a(o)_{ij}| \leq \alpha$$

Therefore by induction

$$\max_{ij} |a(k+1)_{ij}| \leq \alpha^{k}$$

Thus the infinite series defining the elements of $\phi(t-t_o)$ is dominated by series $\sum_{i=0}^{\infty} \frac{(\alpha T)^i}{i!}$ which is a scalar exponential series and is uniformly convergent. Having therefore proved the convergence of $\phi(t-t_o)$ the convergence of $G(t-t_o)$ automatically follows due to simple relationship between these matrices (cf equation (2.10).

2.1.4. - SOLUTION OF THE GENERAL MATRIX DIFFERENCE-DIFFERENTIAL EQUATION

In the introduction it was pointed out that the type of systems to which optimal digital control were likely to be applied

where those whose dynamics often contain pure time delays, i.e. systems whose dynamics are described by difference-differential equations. Using an extension of the method used in the solution of equation (2.5), the solution of the linear time invariant matrix difference-differential equation

$$\dot{x}(t) = \sum_{i=0}^{\tau_1} A_i\, x(t-i\delta) + \sum_{i=0}^{\tau_2} D_i\, m(t-i\delta) \qquad (2.11.)$$

where τ_1 and τ_2 are specified positive integer quantities and $x(t)$ and $m(t)$ are the system variables and control variable vectors respectively,

will now be derived.

The general solution of equation (2.11) was derived by firstly examining simple cases of the equation and thence deducing the general solution structure.

Consider the unforced system described by the class 1 difference-differential equation

$$\dot{x}(t) = A\, x(t) + Bx(t-\delta) \qquad (2.12.)$$

where δ is a pure time delay.

If one considers $x(t-\delta)$ to be a time varying input, then the above equation has a solution similar in form to equation (2.10), namely,

$$x(t) = \phi\,(t-t_o)\, x(t_o) + \int_{t_o}^{t} \phi\,(t-q)\, B\, x(q-\delta)\, dq \qquad (2.13.)$$

The variable $x(t)$ is a function of $x(t_o)$ and $x(r)$ where $t_o - \delta \leq r \leq t-\delta$. In order to expand $x(t)$ as a function of $x(t_o)$, $x(t_o-\delta)$ etc. it is necessary to find an expression relating $x(q-\delta)$ to $x(t_o-\delta)$ and

$x(s)$ where $t_o-2\delta \leq s \leq t-2\delta$. This can be done by considering t and t_o in equation (2.13) as dummy variables and thence by slight manipulation of equation (2.13) one obtains the expression

$$x(q-\delta) = \phi(q-t_o)\, x(t_o-\delta) + \int_{t_o-\delta}^{q-\delta} (q-\delta-r)B\, x(r-\delta)dr \qquad (2.14.)$$

By changing variable in last term, equation (2.14) becomes

$$x(q-\delta) = \phi(q-t_o)\, x(t_o-\delta) + \int_{t_o}^{q} \phi(q-s)B\, x(s-2\delta)ds \qquad (2.15.)$$

Inspection of equations (2.13) and (2.15) reveals that a general form of equation for $x(q-i\delta)$ has been derived. The successive substitution of these expressions into equation (2.13) in a similar fashion to the method used for solution of equation (2.5), eventually yields an infinite series expression for $x(t)$.

$$x(t) = \sum_{i=0}^{\infty} \phi_i(t-t_o)\, x(t-i\delta) \qquad (2.16.)$$

where $\phi_o(t-t_o)$ is equal to $\phi(t-t_o)$ in equation (2.9)

and $\phi_i(t-t_o) = \int_{t_o}^{t} \phi(t-q)\, B\, \phi_{i-1}(q-t_o)\, dq$

The system variable vector $x(t)$ has therefore been expressed as a function of the delayed system variables. The remaining problem however, is to derive a simple method of evaluating $\phi_i(t-t_o)$ for $i>0$. A particular case of equation (2.12) is that in which $\delta = 0$, i.e.

$$\dot{x}(t) = (A + B)\, x(t)$$

which has a known solution (cf equations (2.5.) and (2.10))

$$x(t) = \left[I + (A+B)T + (A+B)^2T^2/2! + (A+B)^3T^3/3! + \ldots\right]x(t_o) \qquad (2.17.)$$

where $T = t-t_o$.

From equation (2.16) it is apparent that the term $\phi_i(t-t_o)$ contains

matrix B to the cumulative i'th power only. Since equation (2.17) holds for all finite matrices B, the infinite series in equation (2.17) can be arranged as a sum of groups of terms with a common cumulative power of B, the i'th group of terms being equal to the matrix $\phi_i(t-t_o)$. Therefore

$$\phi_o(t-t_o) = I + AT + A^2T^2/2! + A^3T^3/3! + \ldots \qquad (2.18.)$$

$$\phi_1(t-t_o) = 0 + BT + (AB + BA)T^2/2! + (A^2B+ABA+BA^2)T^3/3! + \ldots$$

$$\phi_2(T-t_o) = 0 + 0 + B^2T^2/2! + (AB^2+BAB+B^2A)T^3/3! + \ldots$$

etc.

The j th group, where j denotes power of T, in the expression for $\phi_i(t-t_o)$ will be denoted as $F_{ij}(T)$. From equations (2.17) it can be seen by inspection that $F_{ij}(T)$ consists of a summation of terms of general form

(a) AT/j x (term with common factor $T^{j-1}/(j-1)!$ and B to the cumulative i th power)

or (b) BT/j x (term with common factor $T^{j-1}/(j-1)!$ and B to the cumulative (i-1)th power)

In other words F_{ij} can be formed from a summation of

(a) AT/j x $F_{ij-1}(T)$

and (b) BT/j x $F_{i-1\ j-1}(T)$

Expressed in mathematical terms, the iteration procedure for F_{ij} becomes

$$F_{ij}(T) = \left[AT\ F_{ij-1}(T) + BT\ F_{i-1j-1}(T)\right]/j \qquad (2.19)$$

where $F_{oo}(T) = I$

and $F_{ko}(T) = F_{-1,e}(T) = 0$ for $k > 0$ and all l.

The validity of the above iterative equation was proved by straightforward expansion and then comparison with the groups appearing in equations (2.18). The matrices $\phi_i(t-t_o)$ can now be easily calculated from expression

$$\phi_i(t-t_o) = \sum_{j=o}^{\infty} F_{ij}(T).$$

A simple iterative procedure has thus been developed to calculate the matrices appearing in equation (2.16).

The infinite series solution of equation (2.12) expressed in equation (2.16) is the exact solution. Before a truncated form of the solution can be used as an approximation to the true solution, the infinite series must be proved convergent. Before considering the convergence of $\phi_i(t-t_o)$ to a null matrix, one must first consider the convergence of $F_{ij}(T)$ with increasing j. Assume that $F_{i-1,j-1}(T)$ converges with increasing j, so that for a sufficiently large j, $j \geq n$, $F_{i-1,\ j}(T)$ is approximately a null matrix. The iteration for $F_{ij}(T)$ for $j > n$ can then be simplified to

$$F_{ij}(T) = AT\ F_{i\ j-1}(T)/j \qquad \text{where } j > n$$

$$\text{Hence } (AT)^n \sum_{j=n}^{\infty} F_{ij} = \left[\sum_{j=n}^{\infty} (AT)^j/j!\right] F_{i\ j-1}(T) \times (n-1)!$$

$$< \left[\sum_{j=o}^{\infty} (AT)^j/j!\right] F_{i\ j-1}(T) \times (n-1)! \qquad (2.21)$$

where $<$ means less than all the terms of

The infinite series in square brackets has already been proved to be convergent, thus left hand side of equation (2.2.1) converges to a limit subject to $F_{i-1,j-1}(T)$ converging. But iteration for $F_{oj}(T)$ converges since $\phi_o(t-t_o) = \sum_{j=o}^{\infty} F_{oj}(T)$ which has already been proved

convergent. By induction it can therefore be deduced that $F_{ij}(T)$ is convergent for all finite matrices A and B and for all finite T. The equations (2.18) show that when $j \leq i$ and $i > 0$, $F_{ij}(T) = 0$; thus convergence of $F_{ij}(T)$ as j increases ensures the convergence of $\phi_i(t-t_o)$ as i increases. Truncation of exact solution expressed in equation (2.16) to obtain an approximate solution is therefore valid.

The class 2 difference-differential equation

$$\dot{x}(t) = A\,x(t) + B\,x(t-\delta) + C\,x(t-\varepsilon) \tag{2.22}$$

where δ and ε are pure time delays,

will now be examined in the light of the results already obtained and thence a solution for the general multiple state delay equation will be deduced.

Equation (2.22) has solution (cf equation (2.13)).

$$x(t) = \phi(t-t_o)\,x(t_o) + \int_{t_o}^{t} \phi(t-q)\,B\,x(q-\delta)\,dq + \int_{t_o}^{t} \phi(t-q)\,C\,x(q-\varepsilon)\,dq \tag{2.23}$$

From equation (2.23) an expression for $x(q-\gamma)$, where γ is some arbitrary delay, can be derived (cf equation 2.14)

$$x(q-\gamma) = (q-t_o)\, x(t_o-\gamma) + \int_{t_o}^{t} \phi(q-r)\, B\, x(r-\delta-\gamma)\, dr + \int_{t_o}^{t} \phi(q-r)\, C\, x(r-\varepsilon-\gamma)\, dr \qquad (2.$$

Since both $x(r-\delta-\gamma)$ and $x(r-\varepsilon-\gamma)$ have a similar form to $x(q-\gamma)$, equation (2.24) represents a general expression which could be successively substituted into equation (2.23). The resulting infinite series can be represented as

$$x(t) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \phi_{ij}(t-t_o)\, x(t-i\delta-j\varepsilon) \qquad (2.25)$$

In the general case of multiple delays, the equation corresponding to equation (2.25) would contain r_1 infinite summations, where r_1 is the number of state delays present. To eliminate the duplex infinite summations, the assumption that all delays are an integer multiple of a basic delay is made. In the particular case under consideration, one assumes $\varepsilon = 2\delta$. This type of assumption is not as restrictive as it might at first appear, since in the multiple delay case not all the delays need be present; thus the assumption is equivalent to requiring that the physical delays are integer multiples of a basic delay which itself may be entirely fictitious.

With the assumption that $\varepsilon = 2\delta$, an expression for $x(q-\delta)$ can be found which is a function of $x(t_o-\delta)$, $x(r-\delta)$ and $x(r-2\delta)$, i.e. equation (2.2.4) with $\varepsilon = 2\delta$. The successive substitution of this expression

into equation (2.23) yields the expression for x(t).

$$x(t) = \sum_{i=0}^{\infty} \phi_i (t-t_o) \; \alpha (t-i\delta) \tag{2.26}$$

where $\phi_o(t-t_o)$ is as before

$\phi_i(t-t_o)$ has form $\int_{t_o}^{t} \phi(t-q)\beta_q \int_{t_o}^{r} \phi(q-r)\beta_r \int_{t_o}^{s} \ldots ds\,dq$ (2.27)

with each β_j either matrix B or C.

As in the class 1 system, the particular case of $\delta = \varepsilon = 0$ was considered since exact solution was known to be

$$x(t) = \left[I + (A+B+C)T + (A+B+C)^2T^2/2! + \ldots \right] x(t_o) \tag{2.28}$$

From detailed expansion of equation (2.26) it was noted that the expression for $\phi_i(t-t_o)$ contained matrix B to the cumulative k th power and matrix C to the cumulative l th power, for all non negative integer values of k and l where the relationship

$$i = k + 2l \tag{2.29}$$

was satisfied. Since matrices B and C can have any value, it is possible to equate $\phi_i(t-t_o)$ to the group of terms of equation (2.28) where the relationship of equation (2.29) holds. By a similar argument to that used to derive the iteration procedure for calculating $\phi_i(t-t_o)$ for the class 1 equation, the iteration for class 2 systems can

be proved to be

$$\phi_i\,(t-t_o) = \sum_{j=1}^{\infty} F_{ij}(T) \qquad (2.30)$$

where $F_{ij} = \left[AT\, F_{i\ j-1}(T) + BT\, F_{i-1,j-1}(T) + CT\, F_{i-2,j-1}(T)\right]/j$

and $F_{oo}(T) = I$, $F_{ko}(T) = F_{-1\ell}(T) = F_{-2\ell}(T) = 0$ for $k > 0$ and all ℓ.

The validity of this iteration was proved by straightforward expansion. The series $\phi_i(t-t_o)$ can also be shown to be convergent by the methods already illustrated.

By a close examination of the results expressed in equations (2.16), (2.19), (2.20), (2.26) and (2.30) one can deduce that the solution of the general homogeneous difference-differential equation

$$\dot{x}(t) = \sum_{i=o}^{\tau_1} A_i\, x(t_o - i\delta) \qquad (2.31)$$

can be expressed as

$$x(t) = \sum_{j=o}^{\infty} \phi_i(t-t_o)\, x(t_o - i\delta) \qquad (2.32)$$

where $\phi_i(t-t_o) = \sum_{j=o}^{\infty} F_{ij}(T)$

and $F_{ij}(T) = \sum_{k=o}^{\tau_1} \left[A_k T\, F_{i-k,\ j-1}(T)\right]/j$

with $F_{oo}(T) = I$ and $F_{ko}(T) = F_{-m,\ 1}(T) = 0$ for $m > 0$, all 1, and all k except $k = 0$.

Having considered systems described by homogeneous

difference-differential equations, the theory will now be developed to include systems with multiple delay inputs. Consider the system

$$\dot{x}(t) = A\,x(t) + B\,x(t-\delta) + D_o m(t) \tag{2.33}$$

The solution of this equation is (cf equation (2.13))

$$x(t) = \phi(t-t_o)\,x(t_o) + \int_{t_o}^{t} \phi(t-q)\,B\,x(q-\delta)\,dq + \int_{t_o}^{t} \phi(t-q)\,D_o m(q)\,dq \tag{2.34}$$

Using method of successive substitution as previously, equation (2.34) can be rewritten as the infinite series solution

$$x(t) = \sum_{i=o}^{\infty} \phi_i(t-t_o)\,x(t_o-i\delta) + \sum_{i=o}^{\infty} \gamma_{ii}(t-t_o) \tag{2.35}$$

where $\phi_i(t-t_o)$ is defined as before (cf equation 2.20)

$$\gamma_{io}(t-t_o) = \int_{t_o}^{t} \phi(t-r)\,D_o m(r-i\delta)\,dr,$$

and $$\gamma_{ij}(t-t_o) = \int_{t_o}^{t} \phi(t-r)\,D_o\,\gamma_{i,j-1}(r-t_o)\,dr$$

Since the expressions for $\phi_i(t-t_o)$ is identical to those already obtained, the only problem is the evaluation of $\gamma_{ii}(t-t_o)$. From equation (2.35) it can be seen that $\gamma_{ii}(t-t_o)$ is related to $\phi_i(t-t_o)$ by equation

$$\gamma_{ii}(t-t_o) = \int_{t_o}^{t} \phi_i(t-r)\,D_o m(r-i\delta)\,dr \tag{2.36}$$

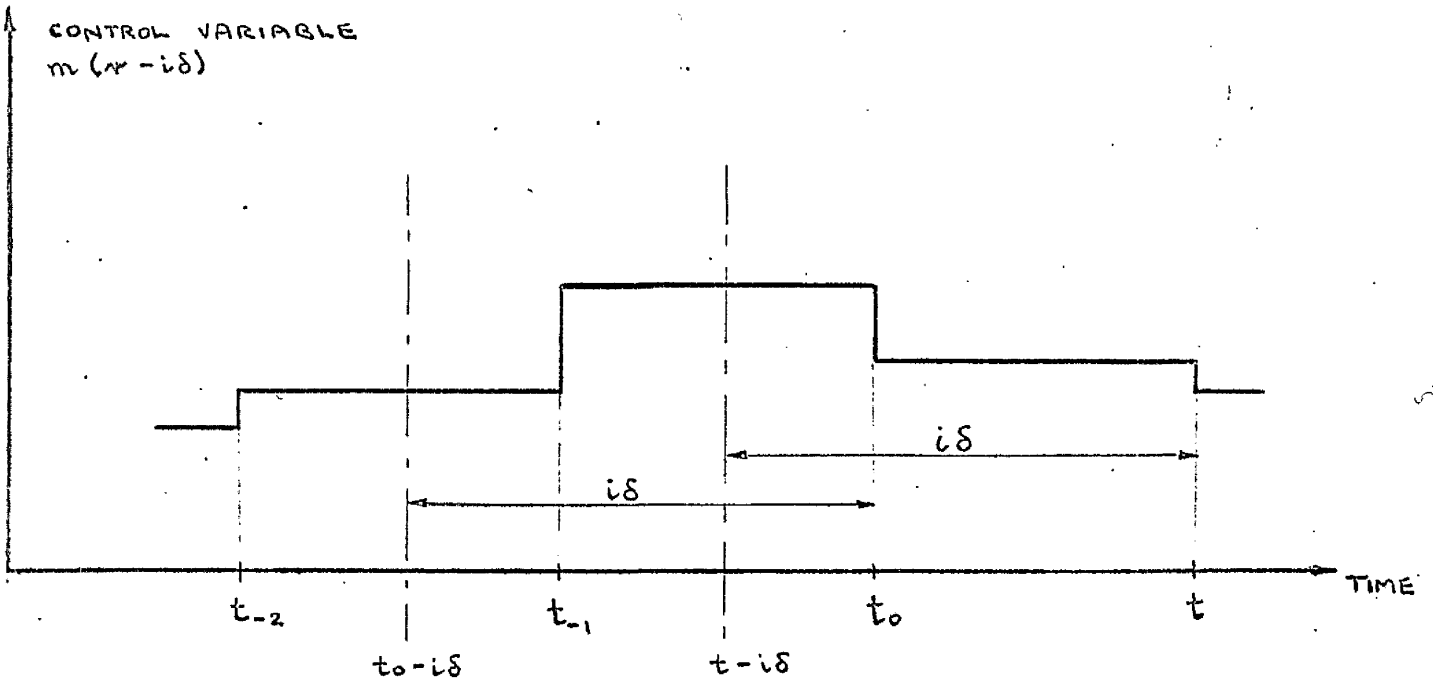


FIGURE 1 DELAYED CONTROL VARIABLE.

By definition of a digital control system $m(r) = m(t_o)$ for $t_o \leq r < t$. On the other hand, variable $m(r-i\delta)$ will adopt at least two values and therefore cannot be directly removed from the integrand in equation (2.36). Figure (1) illustrates the variation of $m(r-i\delta)$. The integral expression can nevertheless be split into two parts, giving

$$\gamma_{ii}(t-t_o) = \int_{t_o}^{t_o+\sigma_i} \phi_i(t-r)\, D_o\, dr\, m(t_o - \overline{\tau_i+1}) + \int_{t_o+\sigma_i}^{t} \phi_i(t-r)\, D_o\, dr\, m(t_o - \tau_i) \qquad (2.37)$$

where $\tau_i = (t-t_o) \times$ integer part of $(i\delta/(t-t_o))$

and $\sigma_i = i\delta - \tau_i$

Using equation (2.20), equation (2.37) can be rewritten as

$$\gamma_{ii}(t-t_o) = \left[\sum_{j=o}^{\infty} \left\{F_{ij}(T) \times T - F_{ij}(T-\sigma_i) \times (T-\sigma_i)\right\} D_o/j+1\right] m(t_o - \overline{\tau_i+1}) + \left[\sum_{j=o}^{\infty} \left\{F_{ij}(T-\sigma_i) \times (T-\sigma_i)\right\} D_o/j+1\right] m(t_o - \tau_i)$$

$$\triangleq S_{2i}(D_o)\, m(t_o - \overline{\tau_i+1}) + S_{1i}(D_o)\, m(t_o - \tau_i) \qquad (2.38)$$

By substituting expression for $\gamma_{ii}(t-t_o)$ from equation (2.38) into equation (2.35) and collecting terms with common factor $m(t_o-iT)$ one obtains the expression

$$x(t) = \sum_{i=o}^{\infty} \phi_i(t-t_o)\, x(t_o-i\delta) + \sum_{i=o}^{\infty} \bar{\epsilon}_{io}(t-t_o)\, m(t_o-iT) \qquad (2.39)$$

where ϕ_i is defined as before (cf equation (2.20))

and $\bar{G}_{io}(t-t_o) = \sum_j S_{1j}(D_o) + S_{2,j-1}(D_o)$ with summation taken over all j where $\tau_j(\delta) = iT$ starting with $j = 0$.

In the case where δ is greater than T, the occasion will arise when $\tau_j = iT$ and $\tau_{j+1} = (i+2)T$. In such cases $\bar{G}_{io}(t-t_o)$ is defined as a null matrix.

Note that since $F_{ij}(r)$ and $\phi_i(r)$ have already been proved convergent for any finite time interval r, the matrix $\bar{G}_{io}(t-t_o)$ will also be convergent.

By examination of the above results, it can be deduced that the solution of equation

$$\dot{x}(t) = A\,x(t) + B\,x(t-\delta) + D_o m(t) + D_1 m(t-\delta) \tag{2.40}$$

is

$$x(t) = \sum_{i=o}^{\infty} \phi_i(t-t_o)\,x(t_o-i\delta) + \sum_{i=o}^{\infty} \bar{G}_{io}(t-t_o)m(t_o-iT) + \sum_{i=o}^{\infty} \bar{G}_{il}(t-t_o)m(t_o-iT) \tag{2.41}$$

where $\phi_i(t-t_o)$ are defined as before (cf equation (2.20)

$\bar{G}_{io}(t-t_o)$ are defined in equation (2.39)

$\bar{G}_{il}(t-t_o) = \sum_j S_{1j}(D_1) - S_{2,j-1}(D_1)$ with summation taken over all j where $\tau_j(\delta) = i$ starting with $j = 1$.

and $S_{1j}(D_1)$ and $S_{2j}(D_1)$ are defined in equation (2.38) with D_1 replacing D_o.

Equation (2.41) can be simplified by collecting terms with common factor $u(t_o - iT)$ resulting in expression

$$x(t) = \sum_{i=0}^{\infty} \phi_i(t-t_o)\, x(t_o - i\delta) + \sum_{i=0}^{\infty} G_i(t-t_o)\, m(t_o - iT) \qquad (2.42)$$

where $G_i(t-t_o) = \overline{G}_{io}(t-t_o) + \overline{G}_{i1}(t-t_o)$

Finally, it is now possible to deduce that the difference equation solution of general difference-differential equation described by equation (2.11) is

$$x(t) = \sum_{i=0}^{\infty} \phi_i(t-t_o)\, x(t_o - i\delta) + \sum_{i=0}^{\infty} G_i(t-t_o)\, m(t_o - iT) \qquad (2.43)$$

where (a) $\phi_i(t-t_o) = \sum_{j=0}^{\infty} F_{ij}(T)$

$$F_{ij}(T) = \sum_{\rho_k=0}^{r_1} A_k T\, F_{i-r,j-1}(T)/j$$

with $F_{oo}(T) = I$, $F_{ko}(T) = F_{-ml}(T) = 0$ for $m > 0$, all 1, and all k except $k = 0$.

and (b) $G_i(t-t_o) = \sum_{\rho_k=0}^{r_2} \overline{G}_{ik}(t-t_o)$

$$\overline{G}_{ik} = \sum_{j} S_{1j}(D_k) + S_{2,j-1}(D_k)$$

with summation taken over all j where

$\tau_j(\delta) = iT$ starting with $j = k$

$$S_{2j}(D_k) = \sum_{j=0}^{\infty} \left[F_{ij}(T) \times T - F_{ij}(T-\sigma_j) \times (T-\sigma_j) \right] D_k/j+1$$

(a) For State at Sampling Period

$$x(\overline{j+1}\ T) = \sum_{i=0}^{n_1} \phi_i(T)\, x(jT - i\ ) + \sum_{i=0}^{n_2} G_i(T)\, m(\overline{j-i}\ T).$$

(b) For State within Sampling period

$$x(\overline{j+1}\ T - i\delta) = \sum_{i=0}^{n_1} \phi_i(T - i\delta)\, x(jT - i\delta) + \sum_{i=0}^{n_2} G_i(T - \sigma)\, m(\overline{j - i - k}\ T)$$

where k = integer part of $i\delta/T$

and $\sigma = i\delta - kT$

FIGURE 2 - APPROXIMATE STATE DIFFERENCE EQUATIONS FOR A DIFFERENCE-DIFFERENTIAL EQUATION.

$$S_{1j}(D_k) = \sum_{j=o}^{\infty} \left[F_{ij}(T) \times (T-\sigma_i) \right] D_k/j+1$$

$$\tau_j(\delta) = (t-t_o) \times \text{integer part of } j\delta/(t-t_o)$$

$$\sigma_i = j\delta - \tau_j(\delta) \text{ and } T = t-t_o$$

It can be inferred from the convergence analysis previously carried out that $\phi_i(t-t_o)$ and $G_i(t-t_o)$ are convergent to null matrices with increasing i. It is thus possible to obtain an approximate truncated solution of equation (2.43), namely,

$$x(t) = \sum_{i=o}^{n_1} \phi_i(t-t_o)\, x(t_o-i\delta) + \sum_{i=o}^{n_2} G_i(t-t_o)\, m(t_o-iT)$$

According to the author's definition of state given in sub-section (2.1.1.), the state of the system described by equation (2.11) with solution described by equation (2.43) is a knowledge of vectors $x(t-i\delta)$ and $u(t-iT)$ for $i \geq 0$. The state vector of general difference differential equation is thus an infinite vector for exact solution, but a finite vector for a truncated approximate solution. For the latter case, the resulting matrix state vector difference equation corresponding to equation (2.11) can then be written as is indicated in figure (2).

After the author had completed this part of his research, a paper by Koepcke (27) was published, in which a similar problem to that solved in this sub-section was tackled. A combined s-plane

z-plane method was used and a solution determined for the extremely restricted case of the delays being equal to multiples of the sampling interval. With this assumption, it was found that both methods yielded identical numerical results.

In conclusion, this sub-section contains the derivation of the general difference-differential equation in the form of a difference equation, which is required by the discrete dynamic programming optimisation technique. The only additional restriction imposed on the linear, time invariant coefficient difference-differential equation was that the delays appearing in that equation where multiples of some delay which may itself be fictitious. The matrix iterative equations developed in this section are ideal for computer handling. An algol program used to derive the various matrices found in the solution of equation (2.11) is detailed in appendix (A)

2.1.4. - ADDITIVE NOISE DISTURBANCES

All physical systems operate in a noisy environment. Nevertheless, for the most part, classical system design is carried out under the assumption that negligible random disturbances are present. In this thesis, the particular case of additive random noise introduced to an otherwise deterministic system will be considered. The introduction of stochastic processes raises the problem of whether the state of the system, previously defined for a deterministic system, is applicable in the stochastic situation, since

the future response of a stochastic system is generally dependent on the manner in which it reached its present condition. If the statistical characteristics of the random disturbances are known and are independent of the previous response of the system, the estimated response of the system at time t could be calculated from a knowledge of the present condition of the system at time t_o and the deterministic inputs to the system in interval t_o to t_1. Therefore the previous definition of state still holds for this subset of all stochastic processes. Such processes are commonly called Markov processes and these will be the only type of stochastic process investigated in this thesis.

2.2. - THE QUADRATIC SUMMATION CRITERION OPTIMAL DIGITAL CONTROL POLICY.

In this section the optimal digital control policy of the digitally controlled Markovian system described by the state vector equations

$$\dot{x} = A\, x(t) + D\, m(t) + u(t) \tag{2.44}$$

where $u(t)$ is an additive independent random vector, will be determined such that when it is applied to the system, it minimises the expected value of the performance criterion defined by equation

$$J_N\left[x(0)\right] = \sum_{i=0}^{N} \left[x'(iT)Q\, x(iT) + m'(iT)\, H\, m(iT)\right] T \tag{2.45}$$

where T is constant sampling period

Q is state cost matrix

and H is control cost matrix.

In order that the performance criterion satisfies the convexity requirement for a unique minimum, the matrices Q and H are restricted to being positive definite or semi-definite.

In Chapter 1 it was pointed out that several papers had been published on the solution of the problem posed above, the most notable being those of Tou (19), Gunckel and Franklin (18) and Joseph (20). Nevertheless, the optimal control law of the above system will be determined below, using the methods proposed in these papers because

(a) the performance criterion differs from those proposed in the beforementioned papers,

(b) the results differ from those obtained in references (19) and (20), even allowing for the different criterion used,

(c) the results will be shown to be a particular case of the results obtained with a more practically orientated performance criterion than that expressed in equation (2.45).

Before attempting to determine the optimal digital control law, some justification is put forward for the choice of the performance criterion

structure. The criterion commonly used by several authors is

$$J_{1N}\left[x(0)\right] = \sum_{i=1}^{N}\left[x'(iT)\ Q\ x(iT) + m'(\overline{i-1}T)\ H\ m(\overline{i-1}\ T)\right]T \qquad (2.46)$$

which does not include any term which is a function of the initial state $x(0)$. The authors, in using such a criterion, argue that since there is no control over the initial state, there is no need to include it in the performance criterion, which is a reasonable argument. At the same time, it is often suggested that equation (2.46) is the discrete analogue of the continuous criterion

$$J_{2N}\left[x(0)\right] = \int_{0}^{NT}\left[x'(t)\ Q\ x(t) + m'(t)\ H\ m(t)\right]dt \qquad (2.47)$$

Considering the limit of the continuous integral to be an infinite summation, it is obvious that equation (2.45) and not equation (2.46) is the discrete analogue of (2.47). It was for this reason that the performance criterion of equation (2.45) was initially adopted as the performance criterion.

It should be noted that the multiplicative sampling period term appearing in equation (2.45) is of no real significance, since minimising the funtional $J_N\left[x(0)\right]$ also minimises $1/T$ times that functional. Although equation (2.45) is the discrete analogue of equation (2.47), the two values of the performance index need bear little relationship to one another when they are calculated for a specific optimal digital control, except in the case as $T \to 0$ when the two indeces will tend to the same value. (28).

Finally, it was stated that the expected value of the performance index will be minimised. It has become accepted practice to use the expected value approach when dealing with stochastic systems (29). In the particular problem to be solved, the expectation is made with respect to the independent random noise input u(t).

2.2.1. - DERIVATION OF OPTIMAL CONTROL POLICY

In Chapter 1 the fundamental concept of dynamic programming, namely the Principle of Optimality, was introduced. The latter will now be used in the solution of the optimisation problem. No additional restrictions will be made concerning the independent noise disturbance's characteristics until it is found necessary to do so.

From the results of subsection (2.1.3.), the equation (2.44) has a difference equation solution

$$x(\overline{j+1}\ T) = \phi(T)\, x(jT) + G(T)\, m(jT) + r(jT) \qquad (2.48)$$

where $\phi(T)$ and $G(T)$ are defined in equation (2.10)

$$r(jT) = \int_{t_j}^{t_{j+1}} (t_{j+1}-q)\, u(q)\, dq \quad \text{with } t_{j+1}-t_j = T.$$

In the following analysis, it is assumed that the sampling period is constant and thus, for simplicity, the argument T will not be included where it is unnecessary to an understanding of the mathematics.

The optimal performance index is defined as

$$f_N[x(0)] = \underset{m(i)}{\text{Min}} \; E\left[\sum_{i=0}^{N} x'(i)\, Q\, x(i) + m'(i)\, H\, m(i)\right] \qquad (2.49)$$

where the N stage minimisation is made with respect to the choice of $m(i)$ for all $0 \leq i \leq N$.

The principle of optimality states, (cf equation (1.1)),

$$f_{N-j}[x(j)] = \underset{m(j)}{\text{Min}} \; E\left[x'(j)\, Q\, x(j) + m'(j)\, H\, m(j) + f_{N-\overline{j+1}}[x(j+1)]\right] \qquad (2.50)$$

where the one stage minimisation is made with respect to $m(j)$ and $f_{-1}[x(N+1)] = 0$.

Let it now be assumed that $f_{N-j}[x(j)]$ can be expressed as

$$f_{N-j}[x(j)] = x'(j)\, P(N-j)\, x(j) + Z(N-j) \qquad (2.51)$$

where $P(N-j)$ is a symmetric, as yet, arbitrary matrix, and $Z(N-j)$ is an unknown scalar quantity.

Substituting the corresponding form of the above expression into equation (2.50) yields

$$f_{N-j}[x(j)] = \underset{m(j)}{\text{Min}} \left\{ x'(j)\, Q\, x(j) + m'(j)\, H\, m(j) + E\left[x'(j+1)\, P(N-\overline{j+1})\, x(j+1) + Z(N-\overline{j+1})\right]\right\} \qquad (2.52)$$

remembering that the expectation is taken with respect to the independent

noise input u(t).

Substituting the expression for x(j+1) from equation (2.48) into equation (2.52) and assuming that the independent noise disturbance u(t) has zero mean, one obtains expression

$$f_{N-j}\left[x(j)\right] = \underset{m(j)}{\text{Min}} \Big\{ x'(j)\left[Q + \phi' P(N-\overline{j+1})\phi\right] x(j) + 2x'(j)\left[\phi' P(N-\overline{j+1})\, G\right] m(j) + m'(j)\left[H + G' P(N-\overline{j+1})\, G\right] m(j) + E\left[r'(j)\, P(N-\overline{j+1}) r(j)\right] + Z(N-j) \Big\} \qquad (2.53)$$

Since the minimisation of equation (2.53) is only over one stage, equating the first variation of $f_{N-j}[x(j)]$ with respect of m(j) to zero will yield the optimal choice of m(j) i.e.

$$2\, x'(j)\, \phi' P(N-\overline{j+1}) G + 2m'(j)\left[H+G' P(N-\overline{j+1})\, G\right] = 0.$$

Hence

$$m(j) = -\left[H+G' P(N-\overline{j+1})G\right]^{-1} G' P(N-\overline{j+1})\phi\; x(j) \qquad (2.54a)$$

$$\triangleq B(N-\overline{j+1})\; x(j) \qquad (2.54b)$$

assuming that the inverse exists.

The control vector m(j) has been found to be a function of the arbitrary matrix $P(N-\overline{j+1})$ under the assumption expressed in equation (2.51). To prove that the structure of the latter is valid, and, as a result, obtain an iterative or difference equation in P(N-j) and Z(N-j), the expression for m(j) will be back-substituted into equation

(2.53), giving

$$f_{N-j}[x(j)] = x'(j)\left[Q + \phi' P(N-\overline{j+1})\phi + \phi' P(N-\overline{j+1})G[H+G'P(N-\overline{j+1})G]^{-1} G'P(N-\overline{j+1})\phi\right] x(j) + E\left[r'(j)\, P(N-\overline{j+1})\, r(j)\right] + Z(N-\overline{j+1}) \quad (2.55)$$

On the other hand, by assumption,

$$f_{N-j}[x(j)] = x'(j)\, P(N-j)\, x(j) + Z(N-j) \quad (2.56)$$

The equation (2.55) and (2.56) are of identical form, thus structure of the assumed expression for $f_{N-j}[x(j)]$ is valid. This is not to say that the assumed structure is the only possible valid one. As it so happens, the control input expressed in equation (2.54) is in fact the optimal input, but the mathematical rigour of the digital optimisation technique concerning this point is not as complete as it might be. It is not intended to analyse this problem, but the appreciation that such a problem exists assists in the understanding of developments of the basic optimisation technique in Chapter 5.

Returning to the main theme of this section, since equations (2.55) and (2.56) hold for all values of x(j), the equivalent parts of these equations can be equated, giving

$$P(N-j) = Q + \phi' P(N-\overline{j+1})\phi + \phi' P(N-\overline{j+1})G\left[H+G'P(N-\overline{j+1})G\right]^{-1} G'P(N-\overline{j+1})\phi \quad (2.57)$$

and $$Z(N-j) = E\left[r'(j)\, P(N-\overline{j+1})\, r(j)\right] + Z(N-\overline{j+1}) \quad (2.58)$$

VARIABLES CALCULATED IN FORWARD TIME. $\longrightarrow$

$x(0)$	$x(1)\ldots$	$x(j-1)$	$x(j)$	$x(j+1)\ldots$	$x(N-1)$	$x(N)$	$x(N+1)$

$\longleftarrow$ OPTIMISATION INTERVAL $\longrightarrow$

VARIABLES CALCULATED IN REVERSE TIME. $\longleftarrow$

$f_N\ x(0)$	$f_{N-1}\ x(1)$	$f_{N-j-1}\ x(j-1)$	$f_{N-j}\ x(j)$	$f_{N-j+1}\ x(j+1)$	$f_1\ x(N-1)$	$f_0\ x(N)$	0
$P(N)$	$P(N-1)$	$P(N-j-1)$	$P(N-j)$	$P(N-j+1)$	$P(1)$	$P(0)$	
$Z(N)$	$Z(N-1)$	$Z(N-j-1)$	$Z(N-j)$	$Z(N-j+1)$	$Z(1)$	$Z(0)$	
$m(0)$	$m(1)$	$m(j-1)$	$m(j)$	$m(j+1)$	$m(N-1)$	$m(N)$	

FIGURE 3 - RELATIONSHIP BETWEEN FORWARD TIME AND VARIABLES USED IN OPTIMISATION PROCEDURE.

The determination of the initial conditions and the relationship between the various variables used in the above analysis and the forward time axis are clarified by an examination of Figure (3). The optimisation interval is $0 \leq t \leq NT$, so that the last control vector to be applied to the system would be m(N). Consider equation (2.50) with j = N, i.e.

$$f_o[x(N)] = \underset{m(N)}{\text{Min}} \; E\left[x'(N)\, Q\, x(N) + m'(N)\, H\, m(N)\right]$$

since $f_{-1}[x(N+1)] = 0$ by definition.

Since both matrices Q and H are positive definite or semi-definite minimisation will be achieved by letting m(N) be a null vector. Therefore

$$f_o[x(N)] = x'(N)\, Q(N)\, x(N)$$
$$= x'(N)\, P(0)\, x(N) \qquad \text{(by assumption)}$$

The initial conditions for the difference equations (2.57) and (2.58) are therefore P(0) = Q and Z(0) = 0.

Figure (3) shows that P(N-j) is calculated in reverse time, thus only when P(N) has been calculated from equation (2.57) with initial conditions P(0) = Q can the first control variable m(0) be calculated.

2.2.2. - CHARACTERISTICS OF OPTIMAL SYSTEM

Firstly, it is noted that the optimal control policy is

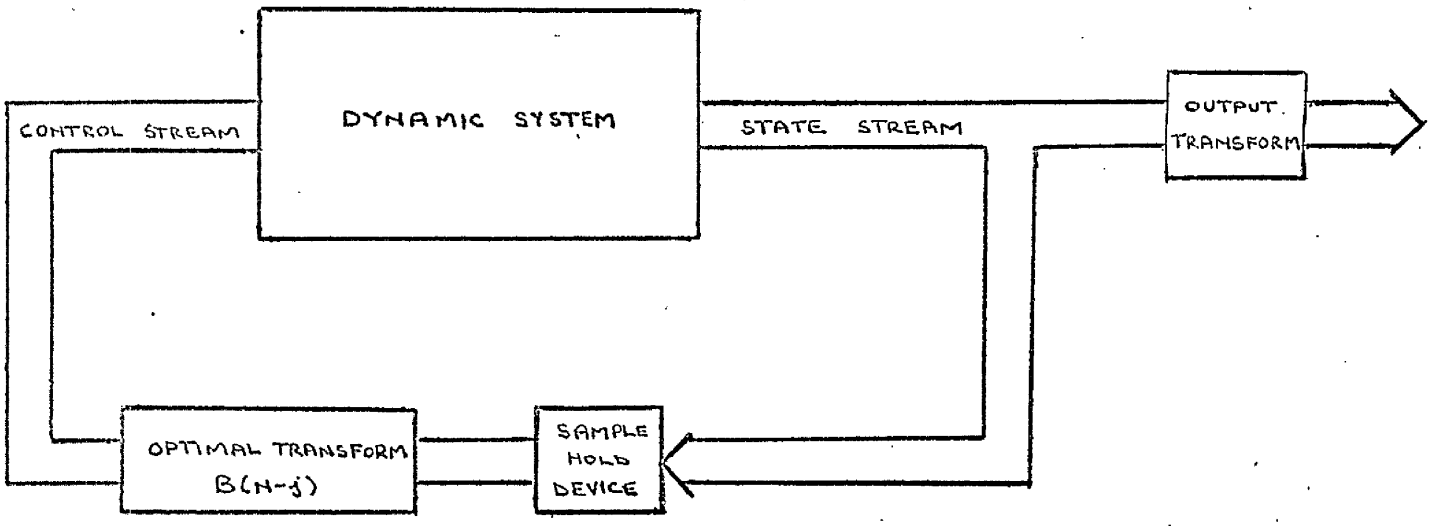


FIGURE 4 OPTIMAL FEEDBACK SYSTEM.

independent of the statistics of the random noise disturbance when the latter is restricted to having a zero mean.

It is seen from equation (2.54) that the control input at time $t = jT$ is a linear time varying function of the state variables of the system at the same time. The resultant system, as indicated in Figure (4) is in the form of a feedback system with, in general, a non-dynamic loop emanating from each state variable and terminating at each control input point. The useful practical characteristics of well designed feedback systems are well known. It would therefore be hoped that the optimally controlled system, which would result from the implementation of the optimal control policy, would have these desirable properties. Such aspects of optimal digital design are discussed in detail in Chapters 4 and 5. For obvious reasons the matrix $B(N-\overline{j+1})$ appearing in equation (2.54b) is called the optimal feedback coefficient matrix. Since the latter is time varying, all the values of $P(N-j)$ for $i \leq j < N$ would have to be stored to allow the calculation of $m(i)$ for $0 \leq i \leq N$. This storage requirement, an inherent property of the dynamic programming approach, is a severe limitation to the usefulness of the optimal digital control policy calculated above. If it is assumed that the difference equation in $P(N-j)$ i.e. equation (2.57), is stable, then, as $N-j$ becomes large, $P(N-j)$ would tend to a constant matrix. For practical purposes let the difference between successive values of $P(N-j)$ be less than some

specified tolerance for $j > n_1$. If matrix $P(N-n_1)$ was then used to calculate $m(j)$ for $0 \leq j \leq n_2$ and the resulting controlled system was stable, with $x(i)$ approximating to a null vector for $i \geq n_2$, then the control input for $m(j)$ for $j \geq n_2$ would be zero. The optimal feedback coefficient matrix is therefore time invariant if the optimisation interval NT, in the definition of the performance criterion, is greater than $(n_1 + n_2)T$. If the upper limit on the optimisation interval was infinite, then the feedback matrix would always be time invariant.

2.3. - GENERALISATION OF OPTIMISATION PROCEDURE.

In this section two particular problems are solved and thence a generalised optimisation procedure is developed, which is applicable to a large class of linear systems. The details of the calculations are given in appendix (B).

2.3.1. - STOCHASTIC SYSTEM WITH NON-ZERO MEAN NOISE DISTURBANCE

The system under consideration is described by equation

$$\dot{x}(t) = A\,x(t) + D\,m(t) + u(t) \qquad (2.59)$$

where $u(t)$ is independent noise disturbance input with non-zero mean.

By splitting $u(t)$ into a zero mean random noise part and a mean part,

it is possible to rewrite equation (2.59) in form

$$\dot{x}(t) = A\,x(t) + D\,m(t) + K\,u_1(t) + u_2(t)$$

where $u_2(t)$ is zero mean disturbance matrix

$u_1(t)$ is vector of mean values of random disturbances,

and K is a suitable matrix.

The above equation also represents a system with zero mean disturbance inputs and a deterministic time varying disturbance input vector $u_1(t)$. All the remarks of this section are thus applicable to such a system.

Assuming that the performance criterion is as defined in equation (2.45) the results of this optimisation problem were found to be (cf appendix B).

Control Policy:

$$m(j) = -\left[H(j) + G'P(N-\overline{j+1})G\right]^{-1} \times \left\{G'P(N-\overline{j+1})\phi\, x(j) + \left[G'P(N-\overline{j+1}) + G'V(N-\overline{j+1})\right] r_1(j)\right\} \qquad (2.60)$$

i.e. feedback from both state variables x(j) and discrete variables $r_1(j)$ corresponding to $u_1(j)$ (cf appendix B).

Performance Index:

$$f_{N-j}\left[x(j),\, r_1(j)\right] = x'(j)\,P(N-j)x(j) + 2x'(j)\,V(N-j)\,r_1(j) + r_1'(j)\,R(N-j)\,r(j) + Z(N-j)$$

where the matrices P(N-j), V(N-j) and R(N-j) and scalar Z(N-j) are determined from difference equations

$$P(N-j) = Q + \phi' P(N-\overline{j+1})\,\phi - \phi' P(N-\overline{j+1})\,G\left[H+G'P(N-\overline{j+1})G\right]^{-1} G'P(N-\overline{j+1})\,\phi \quad (2.61)$$

$$V(N-j) = -\phi' P(N-\overline{j+1})G\left[H+G'P(N-\overline{j+1})G\right]^{-1}\left[G'P(N-\overline{j+1})+G'V(N-\overline{j+1})\right] \phi' V(N-\overline{j+1}) + \phi' P(N-\overline{j+1}) \quad (2.62)$$

$$R(N-j) = 2V(N-\overline{j+1}) + R(N-\overline{j+1}) + P(N-\overline{j+1}) - \left[G'P(N-\overline{j+1})+G'V(N-\overline{j+1})\right]'\left[H+G'P(N-\overline{j+1})G\right]^{-1}\left[G'P(N-\overline{j+1})+G'V(N-\overline{j+1})\right] \quad (2.63)$$

$$Z(N-j) = E\left[r_2'(j)\,P(N-\overline{j+1})\,r_2(j)\right] + Z(N-j+1) \quad (2.64)$$

with initial conditions P(0) = Q and V(0) = R(0) = null matrix and Z(0) = 0.

The assumption necessarily made to obtain a solution was that the vector $u_1(j)$ was constant over interval $0 \leq t \leq NT$, which is the case if the statistics are stationary.

In sub-section (2.1.2.) the augmented state vector was introduced. Using the augmented state formulation of the above problem it will be shown that the two methods give identical results. The same terminology is used as in the above calculation and $u_1(t)$ is assumed constant over the full optimisation interval.

Consider the augmented state vector $\left[x(t) \,;\, u_1(t)\right]'$ with the corresponding matrix dynamic equation

$$\begin{bmatrix} \dot{x}(t) \\ \dot{u}(t) \end{bmatrix} = \left[\begin{array}{c|c} A & K \\ \hline 0 & 0 \end{array}\right]\begin{bmatrix} x(t) \\ u_1(t) \end{bmatrix} + \begin{bmatrix} D \\ 0 \end{bmatrix} u(t) + \begin{bmatrix} u_2(t) \\ 0 \end{bmatrix} \quad (2.65)$$

which has solution

$$\begin{bmatrix} x(j+1) \\ \hline u_1(j+1) \end{bmatrix} = \begin{bmatrix} \phi & \alpha \\ \hline 0 & I \end{bmatrix} \begin{bmatrix} x(j) \\ \hline u_1(j) \end{bmatrix} + \begin{bmatrix} G \\ \hline 0 \end{bmatrix} m(t) + \begin{bmatrix} r_2(t) \\ \hline 0 \end{bmatrix}$$

where α is matrix $\int_{jT}^{\overline{j+1}T} \phi\ (\overline{j+1}T - q)K\ dq$

But $r_1(j) = \int_{jT}^{j+1T} \phi\ (\overline{j+1}T - q)K\ n_1(q)\ dq$

$= \alpha\ n_1(jT)$ under assumption of constant $n_1(t)$.

Thus to give uniform results, let augmented state vector $[x(j) \mid r_1(j)]'$ be considered. The difference equation description of system in terms of this state vector is

$$\begin{bmatrix} x(j+1) \\ \hline r_1(j+1) \end{bmatrix} = \begin{bmatrix} \phi & I \\ \hline 0 & I \end{bmatrix} \begin{bmatrix} x(j) \\ \hline r_1(j) \end{bmatrix} + \begin{bmatrix} G \\ \hline 0 \end{bmatrix} m(t) + \begin{bmatrix} r_2(t) \\ \hline 0 \end{bmatrix} \quad (2.66)$$

The corresponding performance criterion is

$$J_N[x(0)] = E \sum_{i=0}^{\infty} \left\{ \begin{bmatrix} x(i) \\ \hline r_1(i) \end{bmatrix}' \begin{bmatrix} Q & 0 \\ \hline 0 & 0 \end{bmatrix} \begin{bmatrix} x(i) \\ \hline r_1(i) \end{bmatrix} + m'(i)\ H\ m(i) \right\} \quad (2.67)$$

A system with same form of state difference equations and performance criterion has already been solved in sub-section (2.2.1.) with results given in equations (2.51), (2.54a), (2.57) and (2.58). By direct substitution of corresponding matrices, it is found that

Control Policy:

$$m(j) = \left[H(j) + G' P(N-\overline{j+1})G\right]^{-1} \times$$
$$\left[G' P(N-\overline{j+1})\phi \;\vdots\; G' P(N-\overline{j+1}) + G' V(N-\overline{j+1})\right]\begin{bmatrix} x(j) \\ r_1(j) \end{bmatrix} \qquad (2.68)$$

Performance Index:

$$f_{N-j}\begin{bmatrix} x(j) \\ \hline r_1(j) \end{bmatrix} = \begin{bmatrix} x(j) \\ \hline r_1(j) \end{bmatrix}' \left[\begin{array}{c|c} P(N-j) & V(N-j) \\ \hline V'(N-j) & R(N-j) \end{array}\right]\begin{bmatrix} x(j) \\ \hline r_1(j) \end{bmatrix} + Z(N-j) \qquad (2.69)$$

where the matrices P(N-j), V(N-j) and R(N-j) and scalar Z(N-j) are defined in equations (2.61) - (2.64).

Due to the exact equivalence of the above relationships with those obtained by direct method of calculation, the augmented state formulation of the above problem has thus been shown to result in a solution which is identical to that obtained by the more direct method.

2.3.2. - THE DIFFERENCE CRITERION

In many circumstances one is interested in forcing one or more of the state variables of a system to follow a prescribed trajectory. Such systems are the more general case of the regulator problem studied in section (2.2).

Consider the system described by equation

$$\dot{x}(t) = A\,x(t) + D\,m(t) \tag{2.70}$$

The deterministic system has been chosen for no other reason than to simplify the argument, but similar results also hold for stochastic systems.

Let a vector $y(t)$, with same dimensions as $x(t)$ be the vector of the desired value of vector $x(t)$. In the case calculated below it is assumed that the desired value of $x(t)$ is constant during interval $0 \leq t \leq NT$. By a simple argument, but with a considerable increase in numerical complexity, the results of this section can be shown to hold when $y(t)$ is the output vector of a dynamic system with state vector $[y(t) \,|\, w(t)]'$ and dynamic equation of form

$$[\dot{y}(t) \,|\, \dot{w}(t)]' = \bar{A}[y(t) \,|\, w(t)]' + h(t)$$

where $\bar{A}$ is matrix of desired output system,

$h(t)$ is a deterministic constant vector or an independent random noise input.

The quadratic summation performance index which defines the cost of deviation from the desired trajectory while at the same time putting a cost on the control is

$$J_N[x(0)] = \sum_{i=0}^{N} [x(i) - y(i)]' Q [x(i) - y(i)] + m'(i) H m(i) \tag{2.71}$$

The detailed solution of this problem is given in appendix (B), the

important results are,

Control Policy:

$$m(j) = -\left[H + G'P(N-\overline{j+1})G\right]^{-1}\left[G'P(N-\overline{j+1})\phi\, x(j) + G'V(N-\overline{j+1})y(j)\right] \quad (2.72)$$

i.e. feedback from the state variables and the desired trajectory variables.

Performance Index:

$$f_{N-j}[x(j)] = x'(j)\,P(N-j)x(j) + 2x'(j)\,V(N-j)y(j) + y'(j)R(N-j)y(j) \quad (2.73)$$

where matrices P(N-j), V(N-j) and R(N-j) are determined from difference equations,

$$P(N-j) = Q(j) + \phi' P(N-\overline{j+1})\,\phi - \phi' P(N-\overline{j+1})G\left[H+G'P(N-\overline{j+1})G\right]^{-1} G'P(N-\overline{j+1})\,\phi \quad (2.74)$$

$$V(N-j) = -Q + \phi' P(N-\overline{j+1})G\left[H + G'P(N-\overline{j+1})G\right]^{-1} G'V(N-\overline{j+1}) + \phi' V(N-\overline{j+1}) \quad (2.75)$$

$$R(N-j) = Q + R(N-\overline{j+1}) - V'(N-\overline{j+1})\,G\left[H + G'P(N-\overline{j+1})G\right]^{-1} G'V(N-\overline{j+1}) \quad (2.76)$$

with initial conditions P(0) = Q and V(0) = R(0) are null matrices.

As in sub-section (2.3.1), the problem will now be formulated in augmented state vector terminology. Consider the augmented state vector $\left[x(t)\,;\,y(t)\right]'$. The equivalent dynamic

equations and performance criterion are

$$\begin{bmatrix} \dot{x}(t) \\ \hline \dot{y}(t) \end{bmatrix} = \begin{bmatrix} A & 0 \\ \hline 0 & 0 \end{bmatrix} \begin{bmatrix} x(t) \\ \hline y(t) \end{bmatrix} + \begin{bmatrix} D \\ \hline 0 \end{bmatrix} m(t)$$

$$\text{and} \quad J_N[x(0)] = \sum_{i=0}^{\infty} \begin{bmatrix} x(t) \\ \hline y(t) \end{bmatrix}' \begin{bmatrix} Q & -Q \\ \hline -Q & Q \end{bmatrix} \begin{bmatrix} x(t) \\ \hline y(t) \end{bmatrix}$$

These augmented state vector equations are a particular case of the system discussed in sub-section (2.2.1) this the solution can be found by substitution and some simple manipulation to be

Control Policy:

$$m(j) = -\left[H + G'P(N-\overline{j+1})G\right]^{-1}\left[G'P(N-\overline{j+1})\phi\, x(j) + G'V(N-\overline{j+1})y(j)\right] \quad (2.77)$$

Performance Index:

$$f_{N-j}\begin{bmatrix} x(j) \\ \hline y(j) \end{bmatrix} = \begin{bmatrix} x(j) \\ \hline y(j) \end{bmatrix}' \begin{bmatrix} P(N-j) & V(N-j) \\ \hline V(N-j) & R(N-j) \end{bmatrix} \begin{bmatrix} x(j) \\ \hline y(j) \end{bmatrix} \quad (2.78)$$

where matrices P(N-j), V(N-j) and R(N-j) are defined in equations (2.74) - (2.76).

As in section (2.3.1) it has been shown that the augmented state vector formulation of the optimisation problem results in an identical solution to that obtained by a more direct method.

2.3.3. - CONCLUSIONS

In conclusion, the importance of the analysis discussed in section (2.3) is two fold. Firstly, by considering the results of section (2.2) with those of this section, it can be deduced that the direct approach optimisation techniques can be generalised by considering the completely augmented state variable formulation of the problem and the basic technique of section (2.2). This result is of major importance in the study of the optimisation problem, as it makes it necessary to write only one computer optimisation program rather than a multitude of programs, which would be required by the direct unaugmented method. Secondly, and more importantly, the augmented E.D.R. matrix has been shown to be a matrix compounded of three different types of matrices, which have differing properties, resulting from the difference equations which define them. The blind use of the augmented state variable approach, without a detailed knowledge of the direct approach, would lead to a confused understanding of the stability properties of the augmented E.D.R. matrix difference equation, which will be discussed in detail in Chapter 3.

2.4 - SHORTCOMINGS OF THE QUADRATIC SUMMATION CRITERION

One of the objectives of the research reported in this thesis was to develop a practically useful optimal design

procedure. In this respect the quadratic summation criterion

$$J_N[x(0)] = \sum_{i=0}^{\infty} x'(i)\ Q\ x(i) + m'(i)\ H\ m(i)$$

suffers from two severe limitations, which will be discussed below.

In order that the optimally controlled system be acceptable, it must be stable when the optimal control policy is implemented. As a result of this necessary property, the state vector in the regulator problem will tend to a null vector with increasing time. If the sampling period is chosen large enough, the state vector will approximately be a null vector before the second sampling instant. Only the first control input will be non-zero and the total performance index will be made up of terms due to the initial state and/or the initial control input, assuming that matrices Q and/or H are non-null. It would thus appear that increasing the sampling period to a large value would result in a possible decrease in the performance index, which is indicative of an improvement in the system's response. From all the usual practical design considerations, this conclusion of improved performance is fallacious. If the criterion were redefined so as to include the sampling period as a multiplication term i.e.

$$J_N[x(0)] = \sum_{i=0}^{\infty} [x'(i)\ Q\ x(i) + m'(i)\ H\ m(i)]\ T$$

then at least increasing the sampling period to large values would not

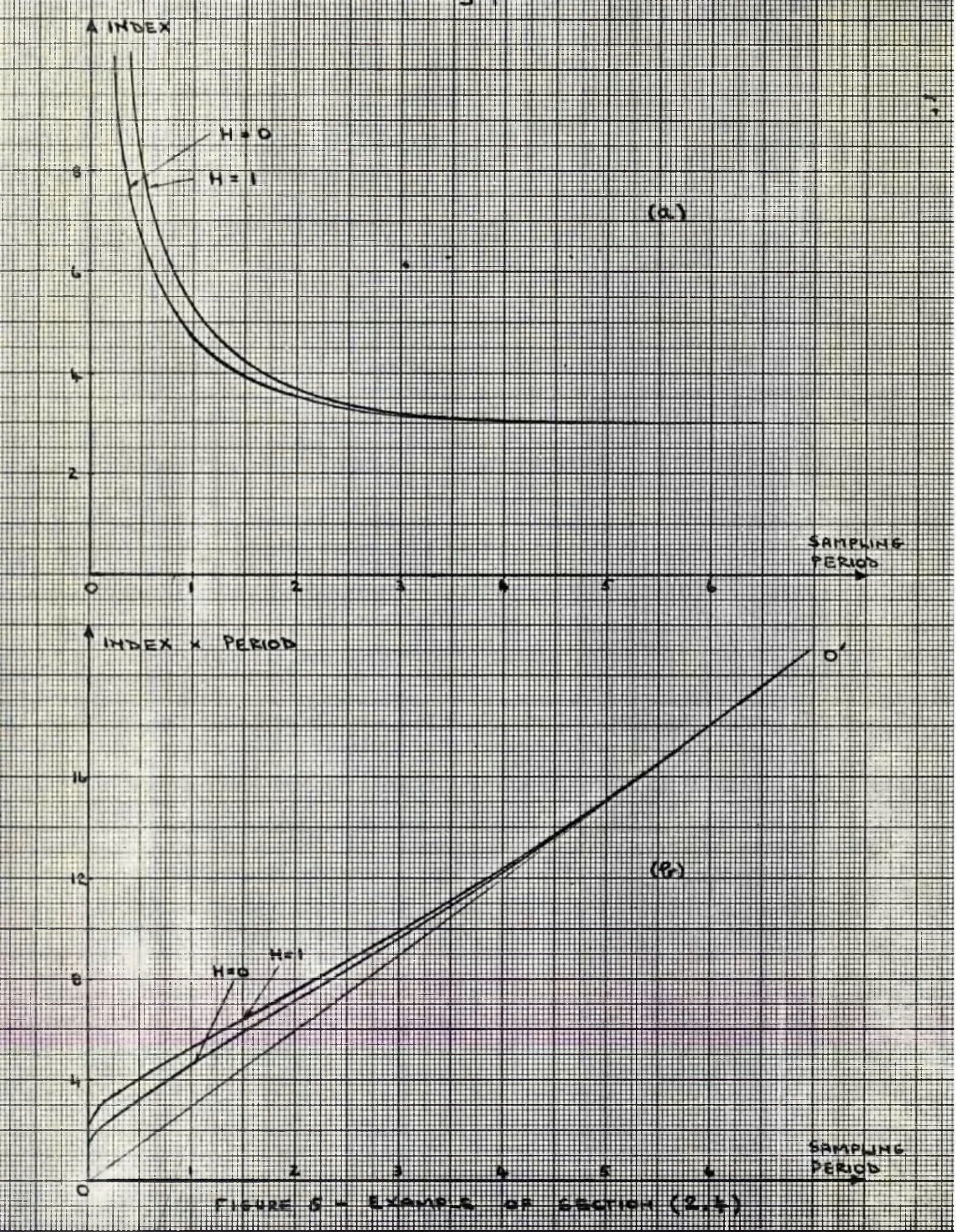


FIGURE 6 – EXAMPLES OF SECTION (2.4)

reduce the value of the index, but the possibility of no change in index still exists if

$$x'(0)\ Q\ x(0) + m'(0)\ H\ m(0) = 0$$

There follows an illustrative example of a 3rd order deterministic system described by

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \end{bmatrix} = \begin{bmatrix} -0.5 & 1 & 0 \\ 0 & -2 & 1 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} m_1(t).$$

The initial conditions are taken as $x_1(0) = x_2(0) = x_3(0) = 1$. For a performance criterion of $\sum_{i=0}^{\infty} [x'(i)\ I\ x(i) + m'(i)\ H\ m(i)]$, with H = 0 and 1, the variation of performance index with sampling period is shown in Figure (5a). As the sampling period tends to zero, the value of the performance index in both cases tends to an infinite value. These results indicate a supposedly better or equally good system response for increasing values of sampling period over the entire range of sampling period.

Figure (5b) shows the variation of performance index multiplied by the sampling period against the sampling period. The line 00′ to which the plotted functions become asymptotic with increasing period is $g(T) = [x'(0)\ I\ x(0)]T$. This alternative measure of the system's performance at least results in an indicated worsening of the system's response with increasing period.

The designer of a digital control system has at some time to decide upon a value of sampling period, basing his choice upon the results of the analysis of the performance system with respect to changes in the latter. The results of the above example do not encourage the belief that the results obtained from analysing the optimal system, even with the criterion including the multiplicative sampling period, would be consistent with classical design procedures. The net result of above discussion is that the quadratic summation criterion is not a practically suitable measure of the system's performance.

The second related objection to the quadratic summation criterion concerns one of the inherent disadvantages of digital control systems, namely, the frequent occurrence of ripple in the state variables during the sampling period. In the author's experience, a particular type of intersample ripple has been found to occur in many cases where the optimal control policy has been calculated by the quadratic summation criterion. The state variables which appear in the expanded performance criterion are forced by the control input to achieve zero value at the sampling instants, but during the sampling interval large deviations occur. The particular case where these deviations have a frequency of half the sampling frequency is the common underdamped oscillatory response. Such responses with large ripple content are not acceptable from the

more classical design standards; thus it must be concluded once again that the quadratic summation criterion is not an acceptable measure of system performance. After the introduction of an improved performance criterion, a system, in which excessive intersample ripple occurs, will be discussed in detail.

2.5 - THE INTERSAMPLE QUADRATIC SUMMATION CRITERION

The shortcomings of the quadratic summation criterion are due to the fundamental limitation of that criterion, which is that it is dependent on the response of the system at sampling instants only and not throughout the sampling period. An alternative criterion, which overcomes this practical objection, has been developed and still allows the solution of the optimal digital control problem. This criterion has been designated the intersample quadratic summation criterion and is defined as

$$J_N[x(0)] = \sum_{j=0}^{N} \left\{ \sum_{i=0}^{\tilde{n}} [x'(k)\ Q\ x(k)] + m'(jT)\ H\ m(jT) \right\}. \qquad (2.79)$$

where $k = jT + is$

and n some prescribed integer quantity such that $ns \leq T$ with

s a time interval.

In the particular case when $n = 0$ the above criterion is in the form of a quadratic summation criterion.

2.5.1. - DERIVATION OF OPTIMAL CONTROL POLICY

Consider the system treated in section (2.2),

$$\dot{x}(t) = A\,x(t) + D\,m(t) + n(t) \tag{2.80}$$

The optimal control policy will now be evaluated, which minimises the expectation of the criterion defined in equation (2.79) using a similar method to that of section (2.2).

The general difference equation solution of equation (2.80) is

$$x(jT + is) = \phi(is)\,x(jT) + G(is)\,m(jT) + r(j,is) \tag{2.81}$$

where $0 \leq is \leq T$

$\phi(is)$ and $G(is)$ are defined in equation (2.10)

and $r(j,is) = \int_{jT}^{jT+is} \phi\,(jT + is-q)\,n(q)\,dq.$

Using equation (2.81) and assuming that the independent noise input has zero mean it can be shown that

$$J_N[x(0)] = x'(jT)\,Q^*\,x(jT) + 2\,x'(jT)\,U\,m(jT) + m'(jT)\,H^*\,m(jT) + E\Big[\sum_{i=1}^{\tilde{n}} r'(j,is)\,Q\,r(j,is)\Big] \tag{2.82}$$

where $Q^* = \sum_{i=0}^{\tilde{n}} \phi'(is)\,Q\,\phi(is)$

$$H^* = H + \sum_{i=1}^{\tilde{n}} G'(is)\,Q\,G(is)$$

$$U = \sum_{i=1}^{\tilde{n}} \phi'(is)\,Q\,G(is).$$

By Principle of Optimality, the multistage decision process can be written as

$$f_{N-j}\left[x(jT)\right] = \underset{m(j)}{\text{Min}}\ E\left\{\sum_{i=0}^{\infty} x'(k)\ Q\ x(k) + m'(iT)\ H\ m(iT) + f_{N-\overline{j+1}}\left[x(\overline{j+1}T)\right]\right\} \tag{2.83}$$

where $f_{-1}\left[x(\overline{N+1}T)\right] = 0$

and $k = jT + is$ (cf equation 2.79).

Let it be assumed that

$$f_{N-j}\left[x(jT)\right] = x'(jT)\ P(N-j)\ x(jT) + Z(N-j) \tag{2.84}$$

Substituting expression for $f_{N-\overline{j+1}}\left[x(\overline{j+1}T)\right]$ from equation (2.84) and thence expression for $x(\overline{j+1}T)$ from equation (2.81), one obtains an expression for $f_{N-j}\left[x(jT)\right]$ in terms of $x(jT)$ and $m(jT)$. The first variation of $f_{N-j}\left[x(jT)\right]$ can then be equated to zero, giving an optimal control policy of

$$m(jT) = -\left[H^{*} + G'(T)\ P(N-\overline{j+1})\ G(T)\right]^{-1} \times \left[G'(T)\ P(N-\overline{j+1})\ \phi(T) + U'\right] x(j) \tag{2.85}$$

Back-substituting this expression for $m(jT)$ into the expanded version of equation (2.83) yields the difference equations governing the E.D.R. matrix $P(N-j)$ and scalar $Z(N-j)$ as

$$P(N-j) = Q^{*} + \phi'(T)\, P(N-j+1)\, \phi(T) - [\phi'(T)\, P(N-j+1)\, G(T) + U]\,[H^{*} + G'(T) P(N-j+1)\, G(T)]^{-1} \times [G'(T)\, P(N-j+1)\, \phi(T) + U'] \qquad (2.86)$$

$$\text{and} \quad Z(N-j) = E\left[r'(j)\, P(N-\overline{j+1})\, r(j) + \sum_{i=1}^{\tilde{\infty}} r'(j,is)\, Q\, r(j,is)\right] + Z(N-j+1) \qquad (2.87)$$

The initial conditions of the above difference equations can be obtained from equation (2.83) since

$$f_{o}[x(N)] = \underset{m(N)}{\text{Min}}\left[\sum_{i=o}^{\tilde{\infty}} x'(NT + is)\, Q\, x(NT + is) + m'(NT)\, H\, m(NT)\right]$$

Hence

$$P(0) = Q^{*} - U\,[H^{*}]^{-1}\, U'$$

and

$$Z(0) = E \sum_{i=1}^{\tilde{\infty}} r'(j,is)\, Q\, r(j,is)$$

The structural properties of the equations (2.85) - (2.87) bear a strong relationship to those derived in solution with a quadratic summation criterion. The system designed by using the intersample criterion will therefore also have the properties noted in sub-section (2.2.2), in particular the state feedback characteristics. A comparable generalised optimisation technique similar to that described in section (2.3) can also be shown to exist.

To illustrate the reduction in intersample ripple, which

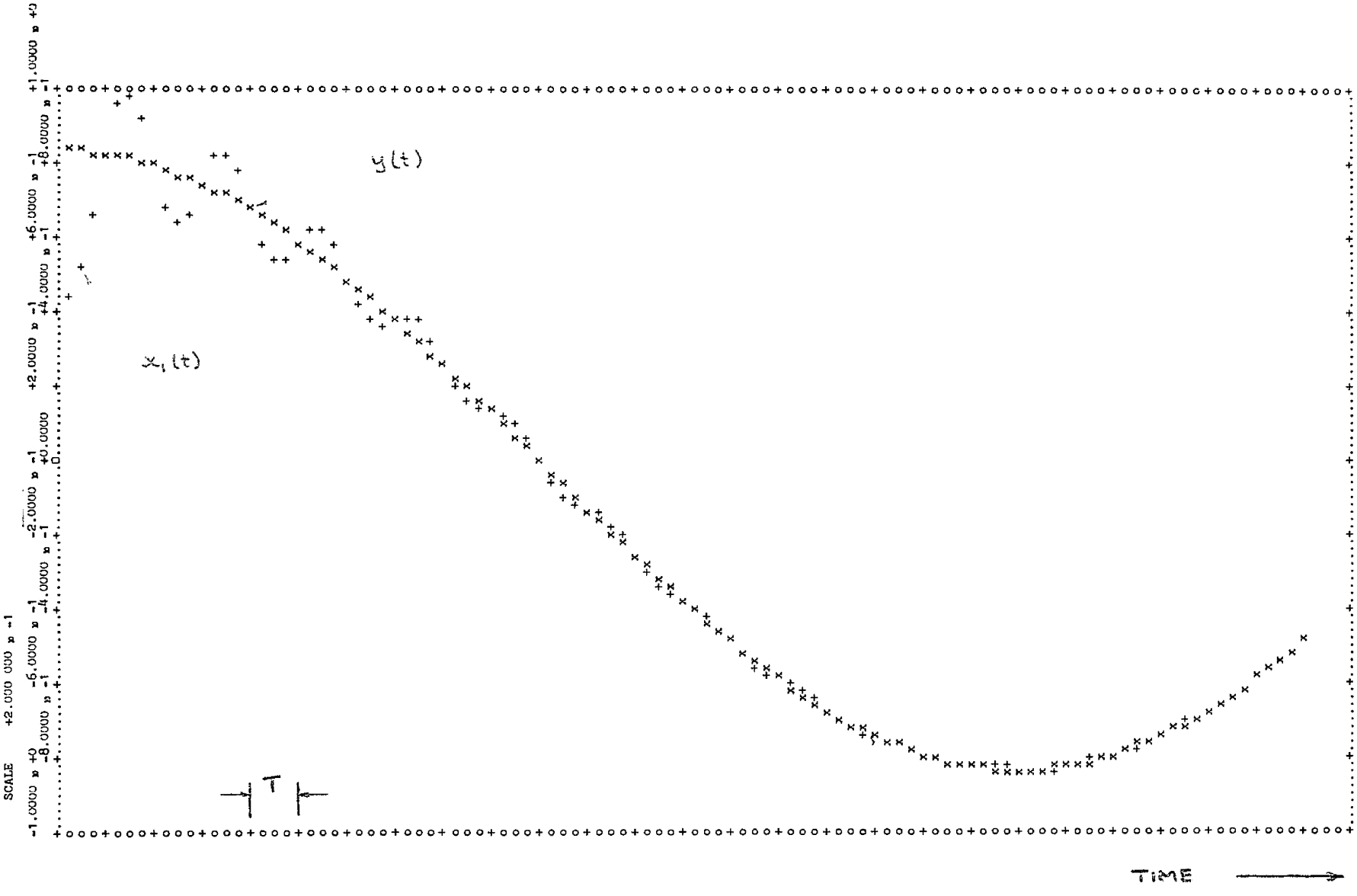


FIGURE 6 - TRAJECTORIES OF $x_1(t)$ AND $y(t)$ FOR QUADRATIC SUMMATION CRITERION.

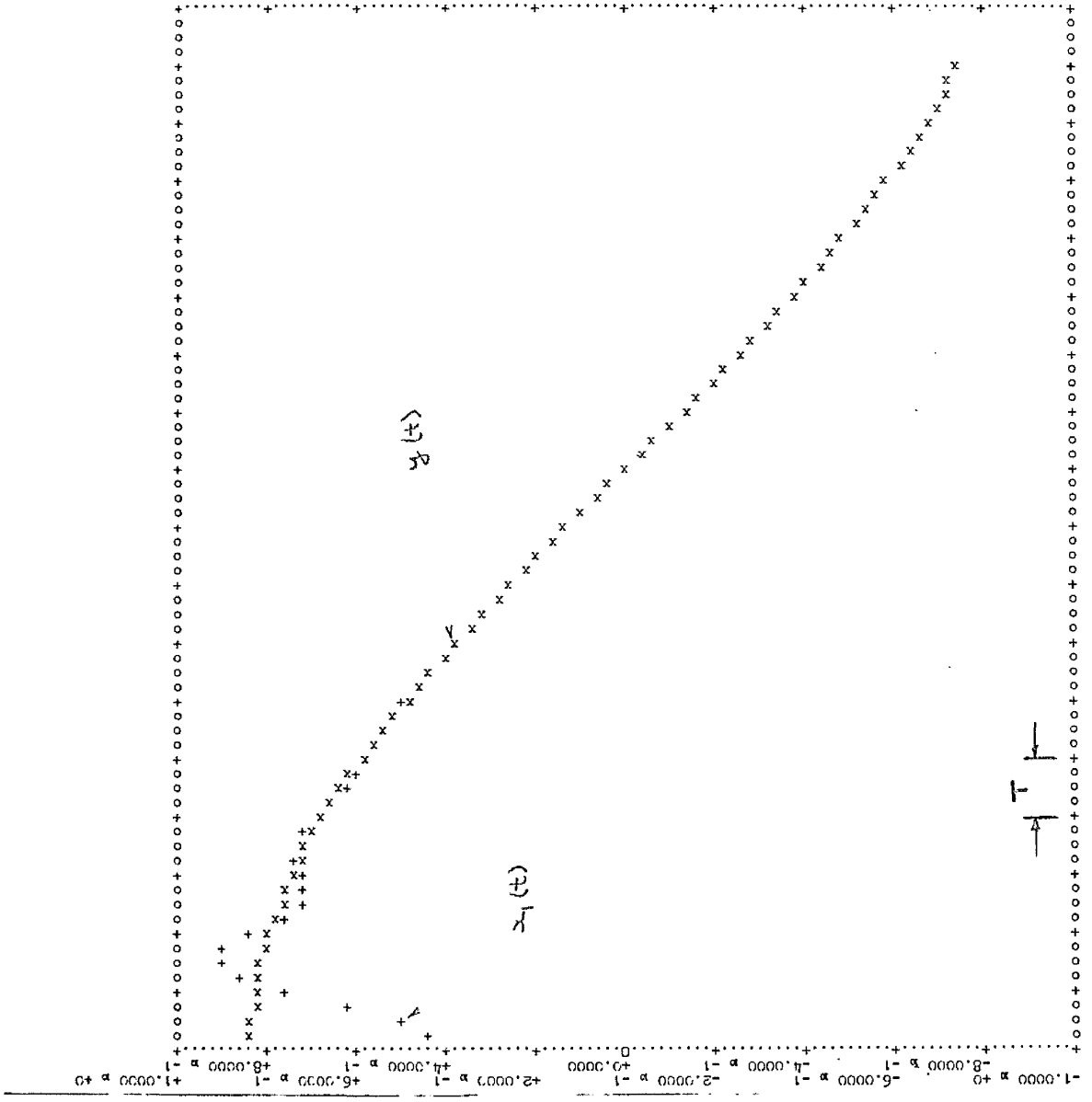


FIGURE 7 - TRAJECTORIES OF $x_1(t)$ AND $y(t)$ FOR INTERSAMPLE CRITERION.

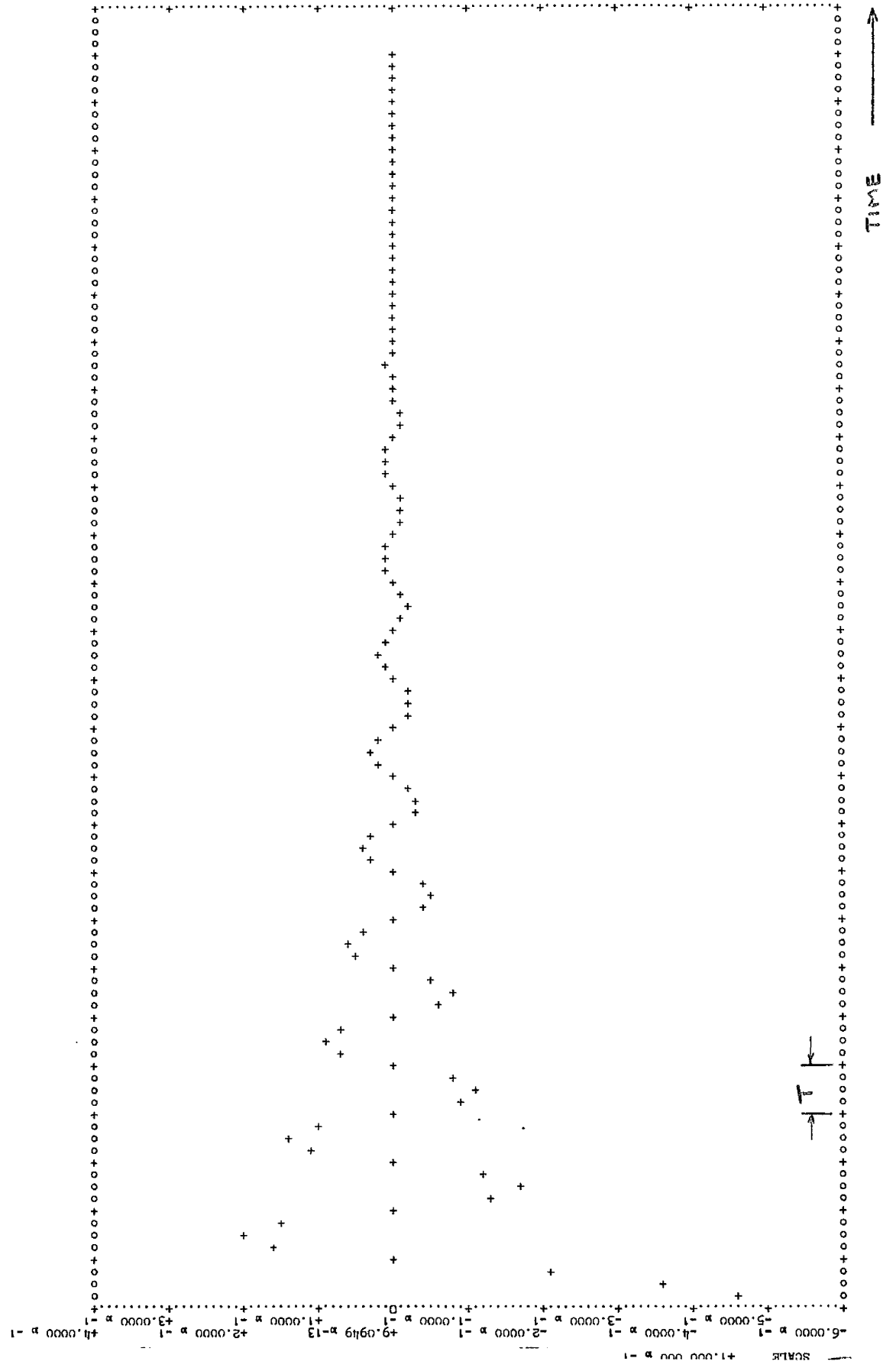


FIGURE 8 - DEVIATION BETWEEN $x_1(t)$ AND $y(t)$ FOR QUADRATIC SUMMATION CRITERION.

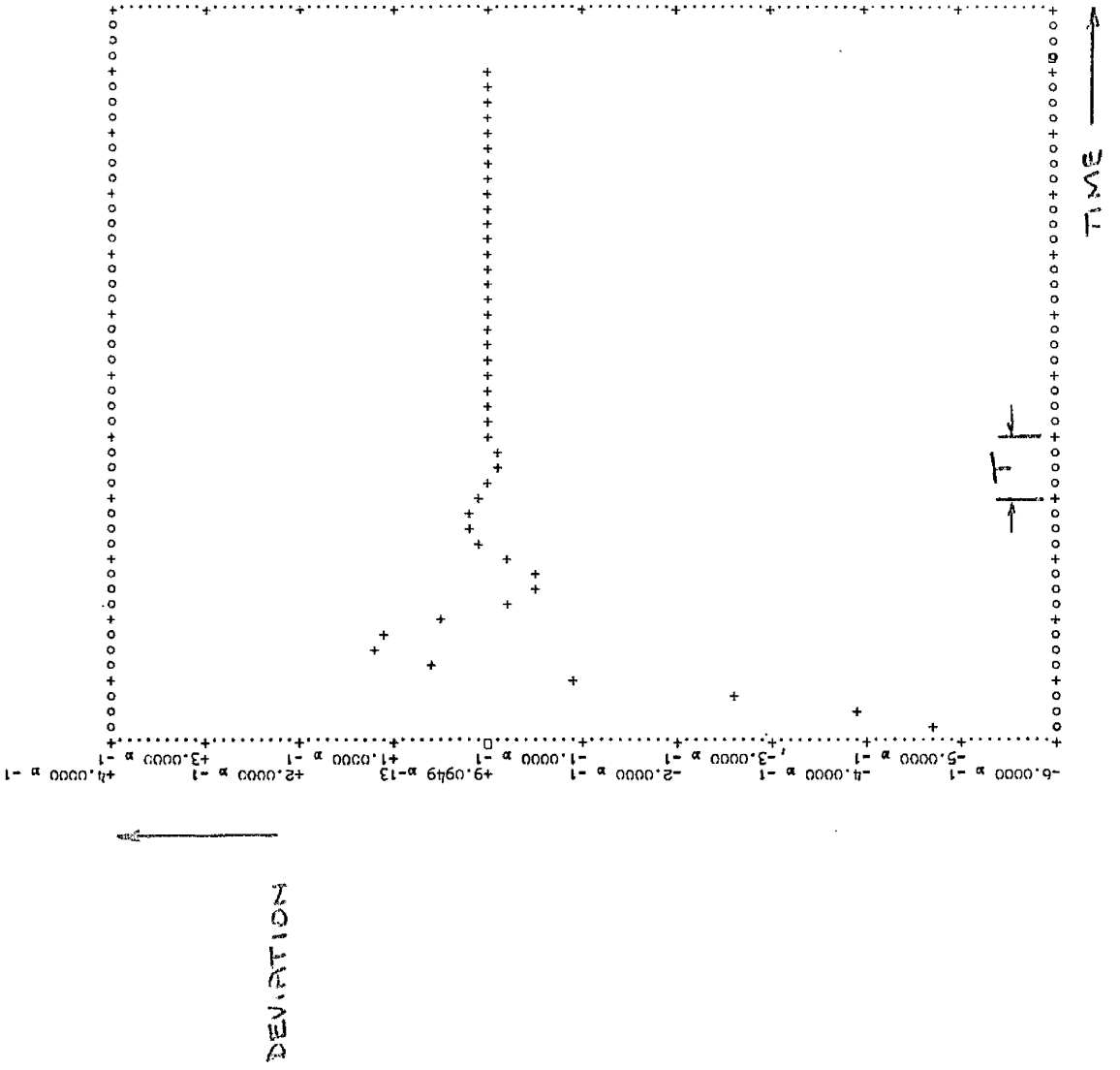


FIGURE 9 - DEVIATION BETWEEN $x_1(t)$ AND $y(t)$ FOR INTERSAMPLE CRITERION.

results from using the intersample criterion, a system will be discussed, which is particularly susceptible to intersample ripple when designed using the quadratic summation criterion. Consider the deterministic 2nd order system described by state equation

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 0 & -2 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} m_1(t)$$

The system is to be forced in such a way that $x_1(t)$ follows a sinusoidal oscillation $y(t)$ with period 8 secs. and amplitude unity. The initial conditions are $x_1(0) = x_2(0) = -0.5$ units and $y = 1$ unit. The two performance criterion chosen were

(a) quadratic summation: $J_{N1}\ x(0) = \sum_{j=0}^{\infty} \left[x_1(jT) - y(jT)\right]^2$

and (b) intersample: $J_{N2}\ x(0) = \sum_{j=0}^{\infty} \sum_{i=0}^{1} \left[x_1(jT + iT/2) - y(jT + iT/2)\right]^2$

i.e. deviation is costed at sampling instant and halfway through period.

Figures (6) and (7) show the optimal trajectories of $x_1(t)$ and $y(t)$ and Figures (8) and (9) show the deviation of the actual from the desired trajectory for both criteria. The figures are reduced copies of the output of a computer program which was written to produce the results of the optimisation program in graphical form on a flexowriter or line printer.

The deviation in the system with criterion J_{N1} was found to be zero at the second and all subsequent sampling instants, i.e. the N stage decision process has degenerated into a single stage decision process. In conventional sampled-data theory, this is called a dead-beat response. The excessive intersample ripple, common to such systems, has a period twice that of the sampling period. The reduction in intersample ripple when the intersample criterion was used is clearly seen from a comparison of Figures (8) and (9). It can therefore be concluded that even one of the simplest types of intersample criterion has resulted in a system's performance, which is more acceptable from the point of view of conventional design criteria than that which resulted from using the quadratic summation criterion.

2.5.2. - THE QUADRATIC INTEGRAL CRITERION.

Well established methods have been devised for the numerical determination of the definite integral

$$\int_b^c g(a)\, da$$

Due to the fact that integration is essentially a smoothing process relatively simple methods give high accuracy results. A series of formulae which only require a knowledge of the value of g(a) at equidistant points over interval b to c, can be determined from the Lagrange interpolation formula (30). The mechanics of the process consist of approximating the function g(a) by a polynomial, which

passes through the known values of g(a). Weddle's rule defined below is such a formula

$$\int_b^c g(a)\,da = 3s/10\ \left(w_0 + 5w_1 + w_2 + 6w_3 + w_4 + 5w_5 + w_6 \right) + \text{error term due to polynomial approximation.}$$

where c-b is divided into 6 equal sub-intervals of length s so that $w_i = g(b + is)$.

Assuming that the sub-interval is small enough, the integration rule without the error term gives accurate results. Where necessary, the interval c-b could be divided into groups of 6 equidistant ordinates and the complete integral calculated by summing the separate calculated values for each group.

Returning to the intersample criterion, it can be rearranged from form of equation (2.79) into

$$J_N\, x(0) = \sum_{j=0}^{N} \left[\sum_{i=1}^{n_1} \sum_{\ell=0}^{6} x'(k)\, Q\, x(k) + m'(jT)\, H\, m(jT) \right] \qquad (2.88)$$

where $n_1 = n/6$

and $k = jT + (1 + i-1)s$

If matrices Q and H are assumed to vary during period according to

$$Q = \bar{Q} \times 3s/10 \times q_r \quad \text{and} \quad H = \bar{H} \times T \qquad (2.89)$$

where Q and H are constant matrices

and $q_r = 1$ for $r = 0, 2, 4$

5 for $r = 1, 5$

6 for $r = 3,$

then, from Weddle's rule, equation (2.88) is an approximation to the continuous integral

$$J_N \; x(0) = \int_0^{NT} \left[x'(t)\, \bar{Q}\, x(t) + m\,(t)\, \bar{H}\, m(t)\right] dt$$

time dependent

The use of/cost matrices does not overcomplicate the digital optimisation procedure, since one need only perform the corresponding substitution of value of Q and H in the evaluation of Q^*, U and H^*. The intersample criterion has therefore been shown to be equivalent to the integral criterion for all values of period under the conditions of equation (2.89), and assuming that the sub-interval s is small enough. The integral criterion is a type of criterion which gives logical results when the performance index is analysed for varying sampling period, since it is a continuous measure of performance equally valid for all sampling periods, whereas the quadratic and intersample criterion are not.

Since performance criterion defined in equation (2.88) results in a reduction of intersample ripple and also gives logical sampling frequency analysis results, it is considered to be a more practical measure of the system's performance and, as such, will be used in the examples detailed in Chapter 6.

CHAPTER 3 - COMPUTATIONAL CONSIDERATIONS

3.1. - MOTIVATION

Although a great deal of effort has gone into developing control calculation techniques for the solution of specific optimal control problems, concentrating mainly on systems of a continuous nature, little work has been published on the examination of the resulting optimal solutions. If the optimal control policies are to be implemented, it is vital that such an analysis should be carried out. The analytic and numerical properties of the iterative optimal procedures devised in the preceeding chapter will be discussed in this chapter.

Kalman in his paper (31) detailed a proof of the existence of the optimal solution for a linear deterministic state unaugmented system, and hence the convergence of the corresponding E.D.R. matrix iterative equation. The proof relied on the complete rather than the partial controllability property of the unaugmented system. Certain aspects of the controllability property of a system will be discussed in section (3.2). The theoretical asymptotic properties of some extremely restricted scalar cases of the iterative matrix equations similar to those defined in section (2.2) were analysed by Adorno in paper (32). It is intended to provide a general proof of convergence for both state-unaugmented and augmented systems

later in this chapter.

The author was unable to find any papers which discussed the numerical or computational aspects of calculating the optimal digital control policy from iterative matrix equations of any type of matrices similar to those detailed in Chapter 2. The reason for this is undoubtedly due to the fact that very few numerical examples have actually been calculated, even for the quadratic summation criterion. Those systems that have been examined have been of an extremely simple and well behaved type, i.e. asymptotically stable uncontrolled systems. The use of the quadratic intersample criterion was found in many cases to increase the difficulties of obtaining an accurate numerical solution; thus a thorough analysis of the computational aspects of the method of generating the optimal control policy was carried out, in an attempt to devise a more computationally accurate method.

3.2. - THE EXISTENCE OF AN OPTIMAL CONTROL POLICY

In this section the author has attempted to stipulate the conditions which, when obeyed, ensure the existence of a control policy. To do so, the system property of controllability is used. The concept of controllability was first introduced by Kalman(31) and has since been considerably elaborated upon in several papers. Those of particular note are reference (33) for continuous systems

and reference (34) for sampled data systems. The definitions of controllability vary widely from paper to paper, so the author has defined a complete and a partial controllability property, which have particular reference to the optimisation problem under discussion in this thesis.

3.2.1. - THE CONTROLLABILITY PROPERTY

The most general and at the same time practical definition of a completely controllable system is a system in which one, some or all of the input variables stimulate all the modes of the system. In an undisturbed optimal control problem, one is interested in whether or not it is possible to force all the state variables and control variables, which appear in the criterion, to zero. The requirements for an optimal system are thus less restrictive than those required for a completely controllable system. If a system is such that the control inputs can be chosen to force only some specified variables to zero, then the system is defined to be partially controllable. Since some of the modes of a system may be unaffected by the control inputs and yet, due to being inherently stable, the associated state variables tend to zero values, the partial controllability property will be seen to be of direct application in the analysis of the optimisation problem.

Consider the system described by the state difference equation

$$x(\overline{j+1}\ T) = \phi\, x(jT) + G\, m(jT) \tag{3.1}$$

where ϕ is of dimensions $p \times p$

and G is of dimensions $p \times c$

By successive substitution, assuming initial condition x(0) and dropping argument T for simplicity,

$$x(n) = \phi^n\, x(0) + C(n-1)\rho\,(n-1) \tag{3.2}$$

where C(n-1) is augmented matrix $\left[\phi^{n-1}G \,\vdots\, \phi^{n-2}G \,\vdots\, \ldots \,\vdots\, G\right]$

and $\rho\,(n-1)$ is augmented vector $\left[m(0) \,\vdots\, m(1) \,\vdots\, \ldots \,\vdots\, m(n-1)\right]$

Letting $y(n) = x(n) - \phi^n\, x(0)$, equation (3.2) becomes

$$C(n-1)\rho\,(n-1) = y(n) \tag{3.3}$$

For equation (3.3) to have any solution for $\rho\,(n-1)$, not necessarily unique, the scalar equations must be consistent, i.e. rank $\left[C(n-1)\right]$ must equal rank $\left[C(n-1) \,\vdots\, y(n)\right]$. If rank $\left[C(n-1)\right]$ equals p, which is the largest value possible for a matrix with dimensions $p \times (pc)$, then rank $\left[C(n-1) \,\vdots\, y(n)\right]$ will also have rank p. If, on the other hand, rank $\left[C(n-1)\right]$ is less than p, then it is possible that rank $\left[C(n-1) \,\vdots\, y(n)\right]$ will differ from rank $\left[C(n-1)\right]$. A sufficient but not necessary condition for the equations to have any solution is therefore

$$\text{Rank}\left[C(n-1)\right] = p \tag{3.4}$$

As a result, a sufficient condition for the system to be completely controllable is expressed in equation (3.4), where n is as yet unspecified.

Since matrix C(n-1) has rank p, the augmented part $\phi^n G$ cannot reduce or increase the rank; thus assuming equation (3.4) is true

$$\text{Rank } [C(n+i)] = \text{Rank } [C(n)] = p \qquad \text{for } i \geq 0$$

Due to Caley-Hamilton theorem $\phi^r G$ can be expressed as a linear combination of $\phi^{r-1}G$, $\phi^{r-2}G$, ... and G where $r \geq p$, the dimension of ϕ. It is therefore only necessary to consider n = p, since $n > p$ does not contribute any further information. The complete controllability condition, which is a sufficient but not necessary condition, then becomes

$$\text{Rank } [C(p-1)] = p$$

Returning to the digital optimal control problem, let it be assumed that only some of the state variables appear in the performance criterion, either directly or indirectly. If the control cost matrix is non null, all the state variables will appear indirectly in criterion in general, since the optimal control inputs are linear functions of all the state variables, except in extremely rare cases. Nevertheless, consider the case when the cost matrix is null and let the vector $\bar{x}(j)$ of dimension $s_1 \times 1$, defined by

$$\bar{x}(j) = M_1\, x(j)$$

where M_1 is a suitable one to one transformation matrix,

be the vector of the variables in the criterion. From equation (3.1)

$$M_1\, x(n) = M_1 \phi^n\, x(0) + M_1\, C(n-1)\, \rho\,(n-1)$$

Therefore $$M_1 C(n-1)\, \rho\,(n-1) = M_1 y(n) \tag{3.6}$$

Let it now be assumed that the uncontrolled system equation has some stable modes, hence

$$\lim_{r \to \infty} M_1 \phi^n \longrightarrow \text{a matrix with } s_1 - s_2 \text{ rows of zero elements}$$

where $n = r \times p$

The objective of the optimal control is to force the state variables $\bar{x}(n)$ to zero and therefore the matrix $M_1 y(n)$ will contain $s_1 - s_2$ zero rows as $r \to \infty$. The corresponding consistency condition for equation (3.6) to have any solution is

$$\text{Rank}\left[M_2 M_1 C(n-1)\right] = s_2 \tag{3.7}$$

where matrix M_2 is a suitable one to one transformation such that $M_2 M_1 y(n)$ is the non null vector of dimensions $s_2 \times 1$ corresponding to the non null rows of $M_1 \phi^n$ as $r \to \infty$.

Since $\text{rank}\left[M_2 M_1 C(n-1)\right] \leq \text{Min}\left\{ \text{rank}\left[M_2\right], \text{rank}\left[M_1\right], \text{rank}\left[C(n-1)\right] \right\}$ and $\text{rank}\left[M_2\right] = s_2$ and $\text{rank}\left[M_1\right] = s_1$, a necessary condition for

equation (3.7) to be true is that rank $[C(n-1)] \geq s_2$. This is not a sufficient condition and therefore one must be satisfied with condition expressed in equation (3.7).

Once again, using the Caley-Hamilton theorem, it can be deduced that one need only consider $n = p$ due to linear dependence for $n > p$. One can therefore conclude that a system has the property of partial controllability if the system obeys the sufficient condition

$$\text{Rank } [M_2 M_1 \; C(p-1)] = s_2. \tag{3.8}$$

Although the conditions for complete and partial controllability are both only sufficient, it is evident that the condition for partial controllability is less restrictive than that for complete controllability. From a practical viewpoint, one would like to be able to stimulate all the modes of the system by the chosen control inputs, since such a system is then easily controlled. In general, the majority of practical systems are completely controllable but, if an augmented state description of the system is considered due to uncontrollable disturbance or reference inputs, the augmented system will not be completely controllable.

3.2.2. - EXISTENCE AND CONVERGENCE THEOREM FOR UNAUGMENTED STATE SYSTEMS.

Having digressed on the subject of partial controllability, this property will now be used to prove the existence of an optimal

control policy and hence to prove the theoretical convergence of the iterative equations used in the calculation of the control policy. For the moment, consider an unaugmented system described by the difference state equation

$$x(j+1) = \phi\, x(j) + G\, m(j).$$

If this system obeys the partial controllability condition, then the performance index, $f_{N-j}[x(j)]$, will have an upper bound, which will be denoted by $\mathcal{F}_{N-j}[x(j), \bar{m}(j)]$, where $\bar{m}(j)$ is some possible non-optimal control input vector. Since the state and control cost matrices have been defined to be at least positive semi-definite, $f_{N-j}[x(j)]$ will also have a lower limit of zero. Hence one has the inequality

$$0 \leq f_{N-j}[x(j)] \leq \mathcal{F}_{N-j}[x(j), \bar{m}(j)].$$

The existence of an optimal solution for a system which is partially controllable and has a performance criterion with positive definite or positive semi-definite cost matrices, has therefore been proved.

Consider an N and an N-1 stage optimisation process with corresponding optimal control policies $m(N-j)$ and $m(\overline{N-1}-j)$. Let $\bar{m}(\overline{N-1}-j)$ be the truncated control policy, made up from the optimal policy $m(N-j)$. By the definition of optimality,

$$f_{\overline{N-1}-j}[x(j)] \leq \mathcal{J}_{\overline{N-1}-j}[x(j), \bar{m}(\overline{N-1}-j)]$$

and due to positive definite or semi-definite cost matrices

$$\mathcal{J}_{\overline{N-1}-j}[x(j), \bar{m}(\overline{N-1}-j)] \leq f_{N-j}[x(j)].$$

Hence $$f_{\overline{N-1}-j}[x(j)] \leq f_{N-j}[x(j)].$$

By assumed form of $f_{\overline{N-1}-j}[x(j)]$,

$$x'(j)\left[P(N-j) - P(N-\overline{j+1})\right] x(j) \geq 0$$

i.e. $[P(N-j) - P(N-j+1)]$ is positive semi-definite.

Matrix P(N-j) is therefore a member of a bounded increasing sequence of matrices, and it follows that the individual elements of the unaugmented E.D.R. matrix converge to their respective ultimate values for a sufficiently large optimisation interval. Nevertheless, nothing can be said about the specific elemental convergence properties, even for the case where the cost matrices are positive definite.

The E.D.R. matrix iteration procedure used in the calculation of the optimal control policy for an unaugmented system has therefore been proved to be theoretically convergent.

It was noted in Chapter 2 that the optimal control policy and the matrix iterative equation in the E.D.R. matrix for a system with zero mean additive noise was independent of the noise

characteristics. The existence of an optimal control policy and the convergence of the E.D.R. matrix will therefore follow from the proof for the deterministic system. The main difference between the stochastic and deterministic systems is that for the former systems the performance index does not have an upper limit due to the additional term $Z(N-j)$ that appears therein.

The difference equation governing $Z(N-j)$ was detailed in equation (2.87), which can be rearranged into

$$Z(N-j) - Z(N-\overline{j+1}) = E\Big[r'(j)\, P(N-\overline{j+1})\, r(j) + \sum_{is=1}^{n} r'(j,is)\, Q\, r(j,is)\Big] \qquad (3.9)$$

As $P(N-\overline{j+1})$ converges to its ultimate steady state value, the left hand side of equation (3.9) tends to a limit; thus the rate of change of $Z(N-j)$ and hence the rate of change of the performance index tends to a constant non-zero value with increasing number of sampling intervals, as will be seen in the following example.

Consider a system described by the state equation

$$\begin{bmatrix}\dot{x}_1(t)\\ \dot{x}_2(t)\end{bmatrix} = \begin{bmatrix}-0.5 & 1\\ 0 & -1\end{bmatrix}\begin{bmatrix}x_1(t)\\ x_2(t)\end{bmatrix} + \begin{bmatrix}0\\ 1\end{bmatrix} m(t) + \begin{bmatrix}n(t)\\ 0\end{bmatrix}$$

where the covariance of the independent noise input $n(t)$ is 0.0133 units^2 and the mean value of $n(t)$ is zero. (cf Appendix G).

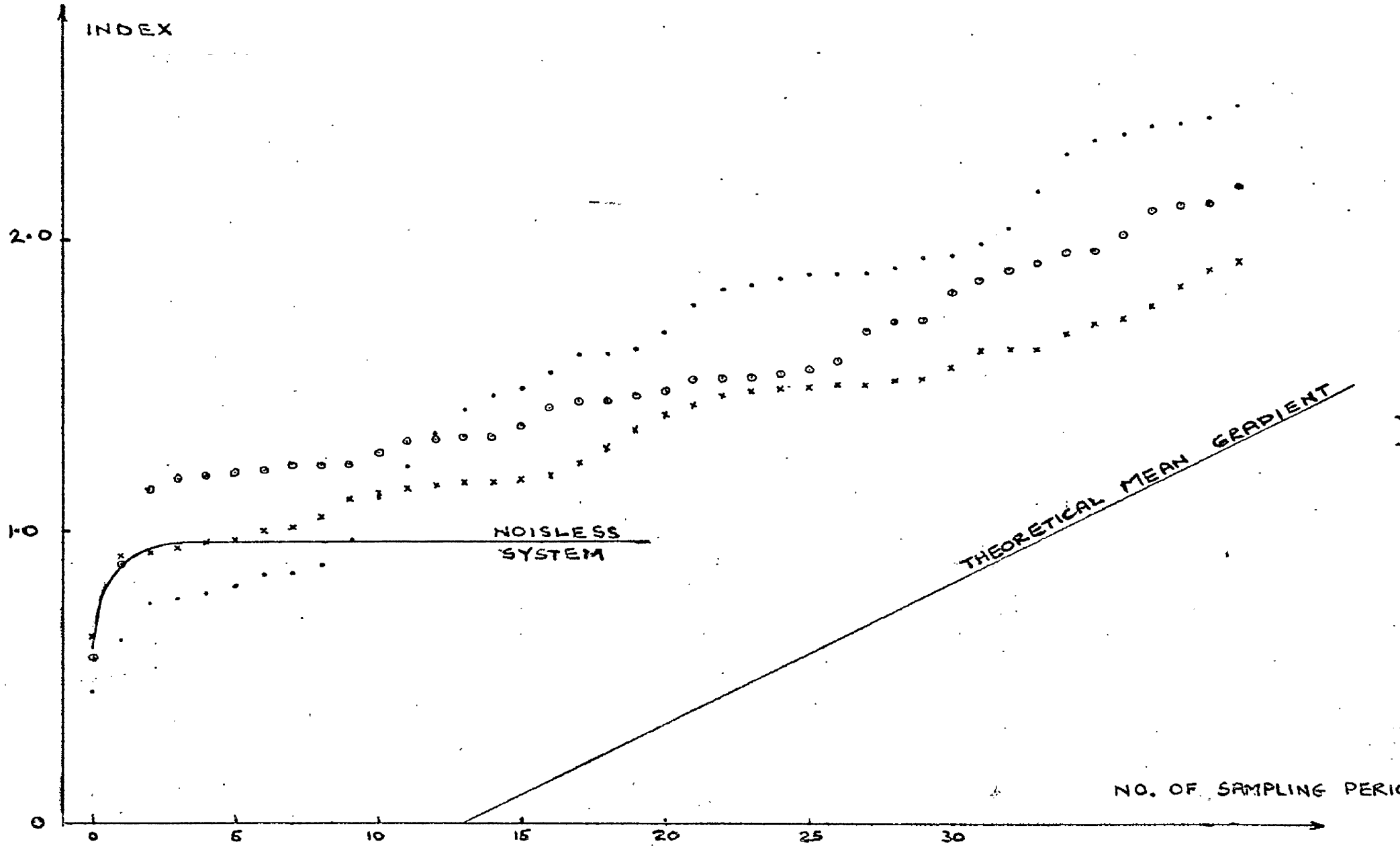


FIGURE 10 INDEX FOR NOISY SYSTEM

The intersample criterion is equivalent to $E \int_0^\infty [x_1^2(t) + x_2^2(t)]\, dt$.

Figure (10) shows the graph of the performance index for the noiseless and the noisy system. In the latter case, the actual value of the performance index is plotted rather than its expected value. From the three cases shown with different noise sequences, but with the same prescribed characteristics, the rate of change of the mean or expected value of the performance index approximates to the calculated value of the left hand side of equation (3.9), namely 4.94_{10}-2 units^2/ sampling period.

The effect of the independent additive noise with zero mean is therefore the introduction of an additional loss term in the performance index for every sampling period, the loss term tending to a constant non-zero value with an increasing number of samples.

3.2.3. - CONVERGENCE THEOREM FOR AUGMENTED STATE SYSTEMS.

In the preceding section, the iterative equation for the unaugmented E.D.R. matrix was proved to be convergent. This convergent property will now be used in this section to prove the stability of the additional difference equations found in augmented state systems. Hence it follows that an optimal control policy can be calculated; thus, there is no necessity to prove an existence theorem for augmented systems.

Rather than deal with the specific types of augmented state system

and the resulting difference equations, only the general structures of these equations need be considered, since it has been found that the difference equations have definite structures universal to all augmented systems. In Chapter 2 it was noted that the augmented state E.D.R. matrix was made up of several different matrices, each satisfying a particular type of difference equation. In particular, for a system with true state vector $x(j)$ and additional state vector $y(j)$, the performance index was found to be equal to

$$f_{N-j}\begin{bmatrix} x(j) \\ \hline y(j) \end{bmatrix} = \begin{bmatrix} x(j) \\ \hline y(j) \end{bmatrix}' \begin{bmatrix} P(N-j) & V(N-j) \\ \hline V'(N-j) & R(N-j) \end{bmatrix} \begin{bmatrix} x(j) \\ \hline y(j) \end{bmatrix}$$

The partitioned parts of the augmented E.D.R. matrix can be categorised by their relationship with respect to the state vectors $x(j)$ and $y(j)$ as,

(a) Matrix $P(N-j)$ which is only associated with the true state vector,

(b) Matrix $V(N-j)$ which is associated with the true and the additional state vectors,

(c) Matrix $R(N-j)$ which is only associated with the additional state vector.

These matrices will now be examined in detail, noting that variables $\phi(T)$, $G(T)$, Q and H refer to the unaugmented system, whereas quantities with bar above refer to completely augmented general system.

P(N-j)

The matrix difference equation governing P(N-j) has the form indicated below, assuming an intersample criterion,

$$P(N-j) = Q^{*} + \phi'(T)\, P(N-\overline{j+1})\, \phi(T) - [\phi'(T)\, P(N-\overline{j+1})\, G(T) + U][H^{*} + G'(T)\, P(N-\overline{j+1})\, G(T)]^{-1} \times [G'(T)\, P(N-\overline{j+1})\, \phi(T) + U']$$

Only the partitioned parts of the system matrices and the performance cost matrices associated with the true state are used in this equation. The latter is therefore identical to the difference equations, which would result if one considered the dynamics etc. associated with the true state to be a completely separate unaugmented system. It has already been proved that such a system results in a stable difference equation for P(N-j) subject to a partial controllability condition being obeyed. The partitioned part matrix P(N-j) of the augmented system will also converge to a steady state value, subject to an identical partial controllability condition.

V(N-j)

The matrix equation governing V(N-j) has been found by examination to have the general structure (cf equations (2.62) (2.75))

$$V(N-j) = g_1[P(N-\overline{j+1}), \bar{\phi}(T), \bar{G}(T), \bar{Q}^{*}, \bar{H}^{*}] + [\phi'(T) + B'(N-\overline{j+1})G'(T)] \times V(N-\overline{j+1})$$

where g_1 is a non-linear function of the matrix arguments, and $B(N-\overline{j+1})$ is the feedback coefficient matrix associated with the true state.

It is important to note the significance of matrix $[\phi'(T)+ B'(N-\overline{j+1})G'(T)]$ which is the transpose of the unaugmented dynamic state transition matrix of the controlled system. It has already been proved that matrix $P(N-j)$ converges; thus, the function g_1 and $[\phi'(T)+B'(N-\overline{j+1})G'(T)]$ will tend to their respective constant matrices α_1 and β giving

$$V(N-j) = \alpha_1 + \beta V(N-\overline{j+1}) \tag{3.10}$$

This is a constant coefficient difference equation, whose stability can be investigated by looking at the solution of the homogeneous part of equation (3.10), i.e.

$$V(N-j) = \beta^{N-j} = [\phi'(T) + B'G']^{N-j} \tag{3.11}$$

In all practical cases of interest, the controlled system will be stable, i.e. eigenvalues of $\phi + GB$ will be contained within the unit circle. This is the condition for the stability of equation (3.11).

The matrix $V(N-j)$ has thus been proved to converge for a sufficiently large optimisation interval. The fact that one was unable to determine the properties of $V(N-j)$ when it was governed by a time varying coefficient difference equation does not alter the fact that

the difference equation in V(N-j) has been proved to be eventually stable.

R(N-j)

The matrix difference equation governing R(N-j) has been found to comply with the general structure (cf equation (2.63))

$$R(N-j) = g_2 \left[P(N-\overline{j+1}),\ V(N-\overline{j+1}), \bar{\phi}(T),\ \bar{G}(T),\ \bar{Q}^*,\ \bar{H}^*\right] + R(N-\overline{j+1}) \qquad (3.12)$$

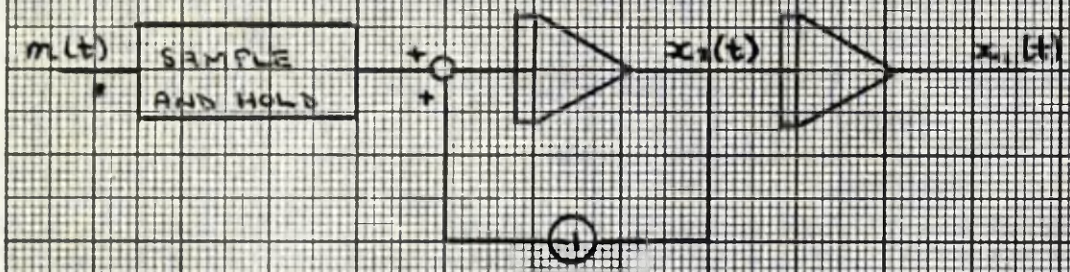
Let it again be assumed that sufficient time has elapsed for $P(N-\overline{j+1})$ and $V(N-\overline{j+1})$ to have attained their steady state values, then g_2 can be replaced by α_2. Rearranging equation (3.12) gives

$$R(N-j) - R(N-\overline{j+1}) = \alpha_2$$

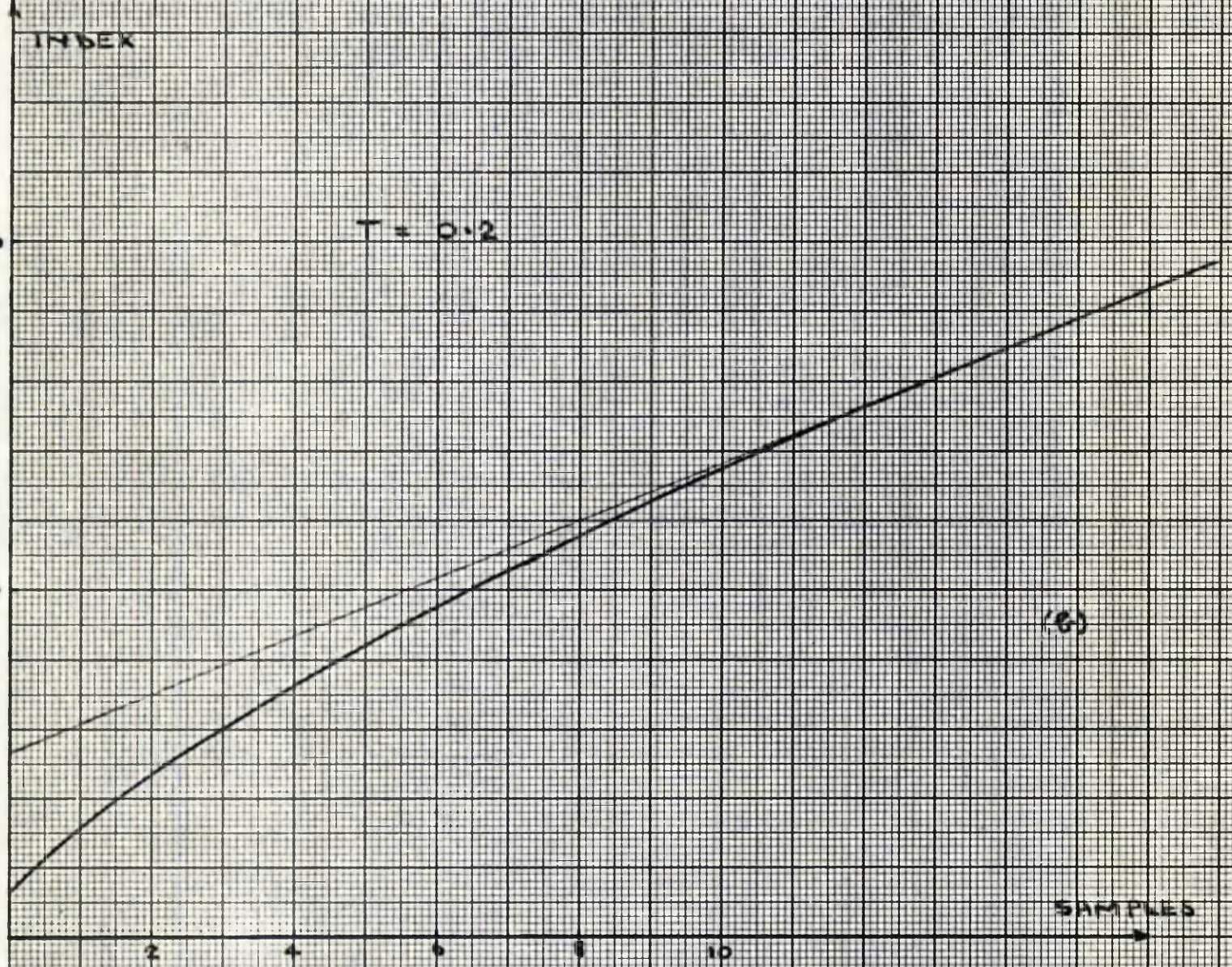
The rate of change of R(N-j) thus tends to a constant limit with increasing number of sampling instants. The optimal performance index has been shown to be

$$f_{N-j}\left[x(j)\right] = x'(j)\ P(N-j)\ x(j) + 2\ x'(j)\ V(N-j)\ y(j) + y'(j)\ R(N-j)\ y(j)$$

As a result, only when α_2 is a null matrix will the performance index converge to a steady state value (assuming that $\| y(j) \| \neq 0$). Otherwise it will adopt a constant rate of change with time for large values of time.

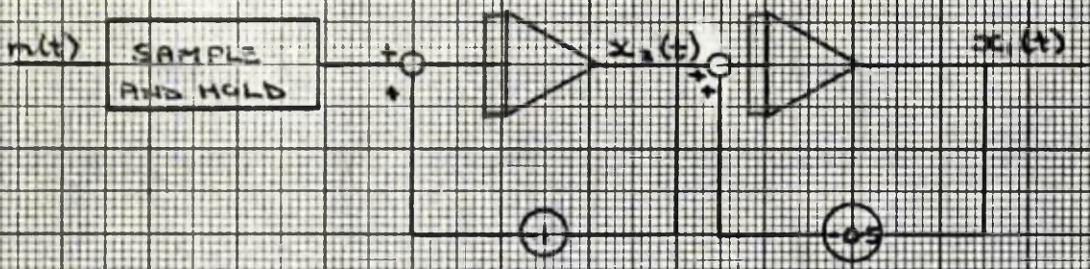


(a)



(b)

FIGURE 11 INCOMPATIBLE CRITERION 1



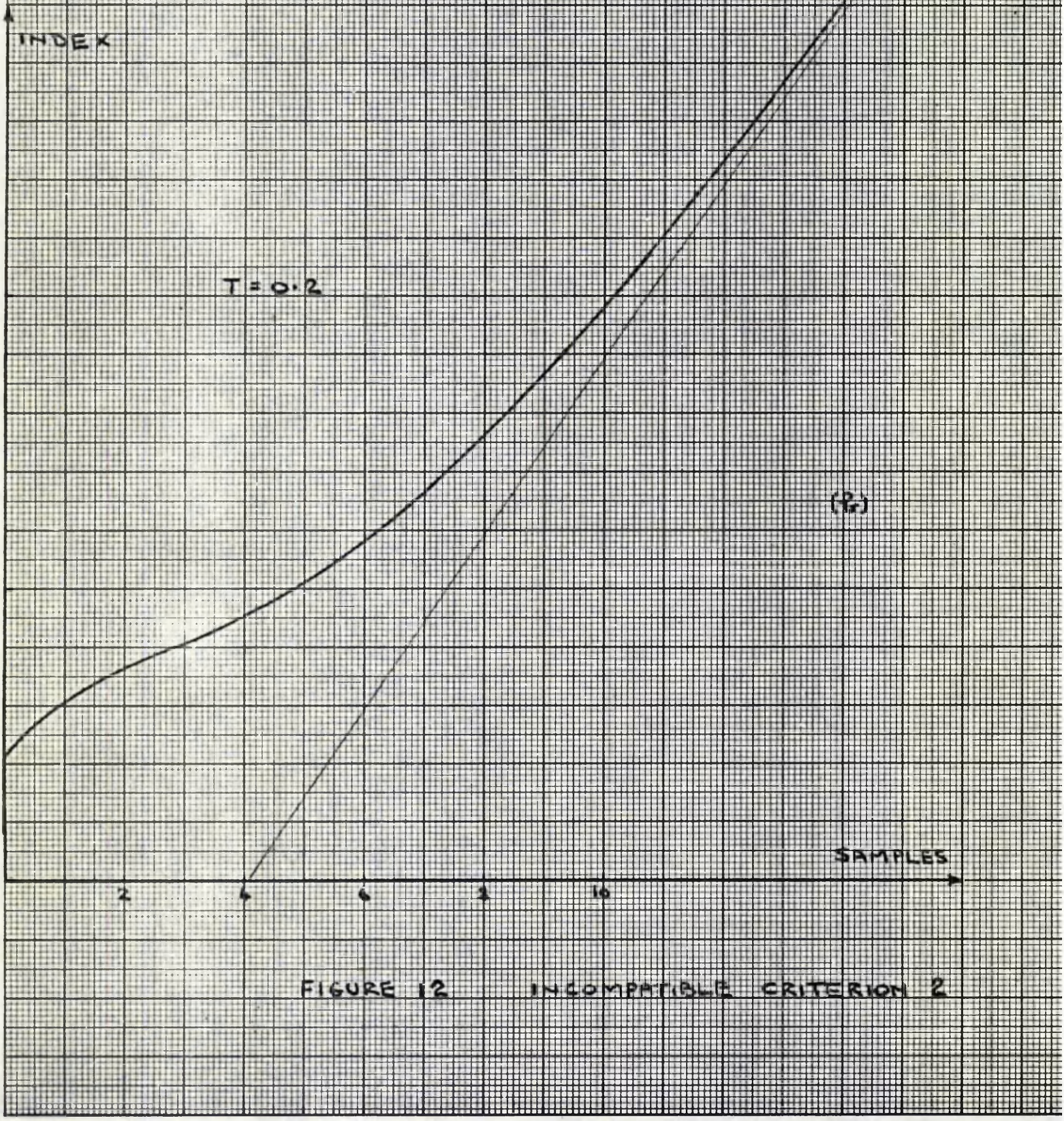


FIGURE 12 INCOMPATIBLE CRITERION 2

The matrix α_2 has been found to be directly related to the compatibility of the performance criterion and the systems dynamic equations. If the performance criterion is incompatible, it is found that matrix α_2 is non-null, whereas a compatible criterion gives a null matrix. To illustrate the relationship, three examples of systems with incompatible criteria are indicated below.

1. Consider the system represented by the diagram in Figure (11), with intersample criterion equivalent to $\int_0^\infty [x_1^2(t) + (x_2(t)-r_2)^2]\, dt$, where r_2 is a constant reference input, whose value $x_2(t)$ is desired to achieve. Input r_2 has value 2 units and the initial conditions are $x_1(0) = x_2(0) = 1$ unit.
   The criterion is obviously incompatible with the system dynamics due to the presence of the integrator. Figure (11) shows the resulting non-zero rate of change of index with large values of time.

2. Consider system represented by diagram in Figure (12) with intersample criterion equivalent to $\int_0^\infty [(x_1(t) - r_1)^2 + (x_2(t)-r_2)^2]\, dt$ where r_1 and r_2 are the constant reference inputs equal to 2 units. The initial conditions are $x_1(0) = x_2(0) = 1$ unit. The criterion is incompatible, since one is attempting to control two states with only one input. The index has again a non-zero constant gradient at large values of time, as can be seen in Figure (12).

   In this case $R(N-j) - R(N-j+1) \rightarrow \begin{bmatrix} 0.04 & -0.08 \\ -0.08 & 0.16 \end{bmatrix}$

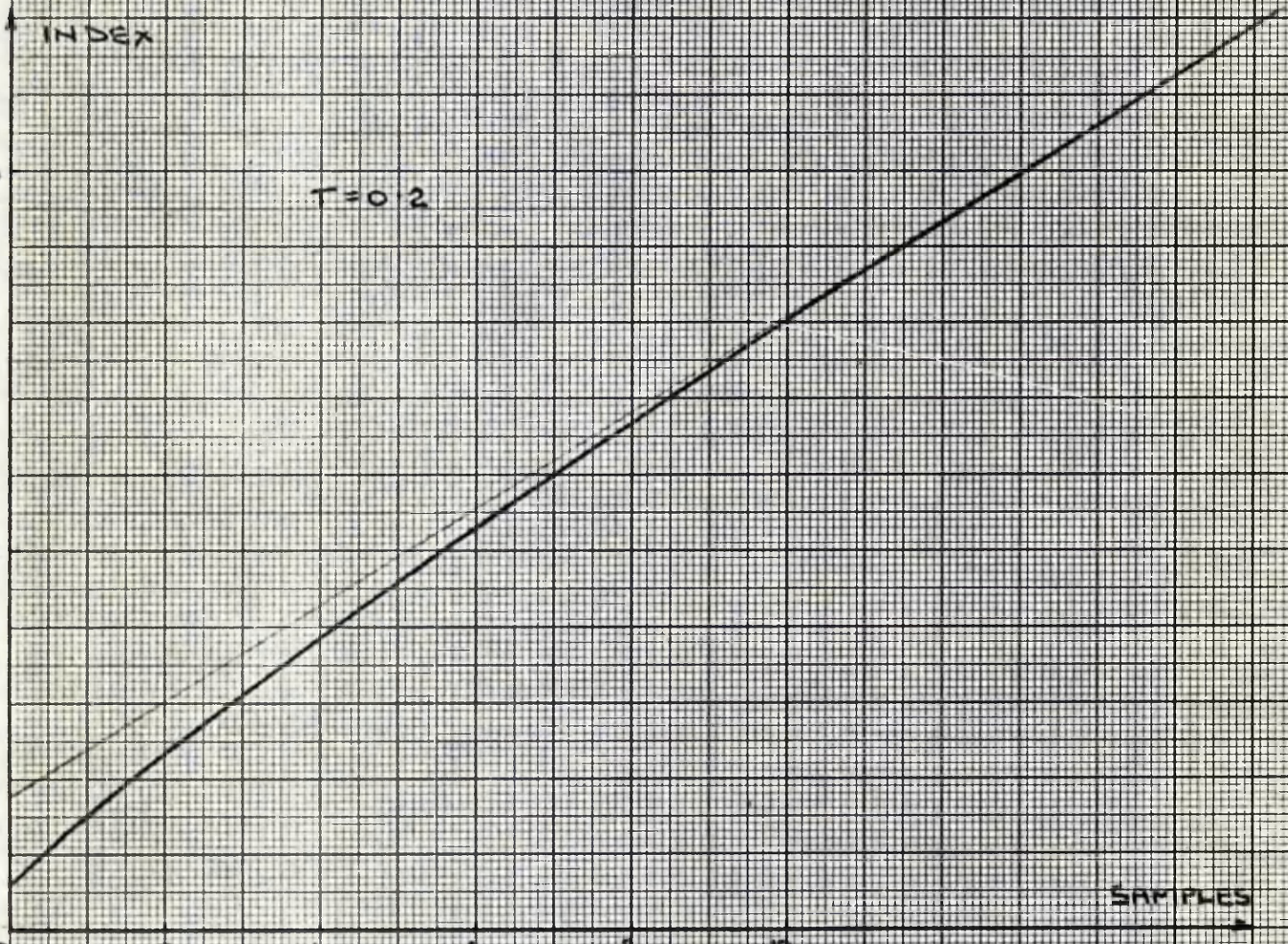


FIGURE 13 INCOMPATIBLE CRITERION 3

with non-zero elements associated with r_1 and r_2. If the situation should arise where, for example, only those elements associated with r_2 were non-zero, then this would indicate that $x_1(t)$ achieves the value r_1, whereas $x_2(t)$ does not achieve value r_2.

3. Consider system of example (2) with no reference inputs, but with a constant deterministic disturbance input h of unit value, added at state x_2, and a criterion
$\int_0^\infty x_1^2(t) + x_2^2(t) \quad dt$.
The criterion is incompatible, since to bring $x_1^2(t)$ to zero, the variable $x_2(t)$ must necessarily have a value equal and opposite to the disturbance input. The non-zero constant gradient of the performance index with large values of period is shown in Figure (13).

In practice only compatible criteria are of interest, but with complex systems the compatibility or otherwise of the criteria is not always as obvious as in the above examples. It is therefore of interest to note that the digital optimal control technique devised in Chapter 2 inherently carries out a compatibility check, the result of which can easily be determined by examining the elements of matrix R(N-j).

The augmented E.D.R. matrix as a whole has thus been shown in this section to converge to a steady state limit, if the unaugmented system is partially controllable, the controlled system is stable and the performance criterion is compatible with the augmented state dynamic equations.

3.2.4. - A LOWER BOUND ON THE NUMBER OF CONTROL VARIABLES.

If a system is not completely or partially controllable, then it would be advantageous to the designer to know whether he has sufficient control inputs. In this section, the dynamic equations of the system are manipulated into a form which enables the determination of the minimum number of control variables which are necessary, although not sufficient, to ensure that the system obeys the controllability conditions.

Consider a system described by the differential equation,

$$\dot{x}(t) = A\,x(t) + D\,m(t) \qquad (3.13)$$

It is well known that a non-singular transformation matrix S can be found (33), such that

$$\dot{x}(t) = S^{-1}AS\,x(t) + S^{-1}D\,m(t)$$

where $S^{-1}AS$ is the jordon cannonic form of matrix A.

Consequently, the difference equation form of equation (3.13)

$$x(j+1) = \phi\, x(j) + G\, m(t)$$

can be transformed into

$$x(j+1) = S_1^{-1}\, \phi\, S_1\, x(j) + S_1^{-1}\, G\, m(j)$$

where $S_1 = \exp(ST)$

The matrix $S_1^{-1} \phi S_1$, the discrete jordon cannonic form of matrix A, has

(a) Diagonal elements equal to $e^{\lambda_i \tau}$, where λ_i are eigenvalues of matrix A, the number of diagonal elements with same value corresponding to the total multiplicity of the eigenvalue,

(b) A jordon block associated with each eigenvalue of multiplicity greater than unity having non-zero elements above the diagonal, yet within the block,

(c) Zero elements elsewhere.

As in the continuous jordon cannonic form, it is possible to have more than one block associated with the same eigenvalue.

To recapitulate, the property of complete controllability requires that all the modes of the system be stimulated by the control

inputs. Hence, for a system described in discrete jordon cannonic form to be completely controllable,

(a) The rows of the matrix $S_1^{-1}G$ corresponding to each distinct (i.e. unit multiplicity) eigenvalue must have at least one non-zero element, otherwise the associated mode is uncontrollable.

(b) The rows of the matrix $S_1^{-1}G$ corresponding to each of the last rows of the discrete jordon blocks associated with the same eigenvalue must be linearly independent of each other, otherwise the associated modes cannot be controlled independently.

For requirement (b) to hold, it is necessary, although not sufficient, that there be at least as many control inputs as there are discrete jordon cannonic blocks associated with the same eigenvalue. A lower limit on the number of control inputs has thus been determined.

The above analysis is dependent on one's ability to calculate the transformation matrix S_1, which is not the simplest of tasks for systems of modest order with multiple eigenvalues. The complete rather than the partial controllability condition was considered above since, although the property of partial controllability was all that

was required to prove the existence of the optimal solution, it is important in practical systems to be able to control all the modes of the unaugmented system.

3.3. - THE AUGMENTED E.D.R. MATRIX ITERATION.

To implement the digital optimisation technique devised in Chapter 2, one is required to calculate the completely augmented E.D.R. matrix. If one is solely interested in the case where the optimisation interval is infinite, then this only necessitates the calculation of the steady state value of the E.D.R. matrix. Three possible methods come to mind.

(1) One could obtain the general solution of the time varying coefficient non-linear difference matrix equation governing the E.D.R. matrix and thence evaluate the solution at a sufficiently large N-j. The non-linear and time varying characteristics of this problem prohibit any attempt at such a solution, so one must resort to more direct methods.

(2) At sufficiently large N-j, it has been proved that the E.D.R. matrix converges to a constant matrix. Considering $P(N-j)$ and $P(N-\overline{j+1})$ as equal to the converged value of the matrix, the difference equation reduces to a non-linear algebraic matrix

equation. The author spent considerable time using ordinary and generalised matrix theory in an attempt to obtain a matrix solution, but no satisfactory results were obtained. A brute force method of determination would be the elemental expansion of the algebraic non-linear equations and their solution for the $\frac{1}{2}n(n+1)$ unknown elements of the symmetric E.D.R. matrix. For systems of even modest order, this approach is formidable.

(3) The final and, it is concluded, only feasible method is the successive substitution in the difference matrix equations to evaluate the sequence of matrices $P(N-j)$ until the latter converges to some steady state value.

3.3.1. - COMPUTATIONAL INSTABILITY.

Computational instability is said to exist in a numerical calculation, when the round off errors become predominant over the true result with repeated substitutions in an iterative equation. It was found that in many cases of interest, any attempt to calculate the steady state value of the E.D.R. matrix by successive substitution was balked by computational instability. Even on a floating-point computer with twelve significant figures, the round off errors in many cases could accumulate after only a few iterations to a value several thousand per cent greater than the true solution. The computational stability of the iteration will now be examined and

a computationally stable iteration procedure developed.

Before proceeding with the analysis, it is useful to derive a relationship between the eigenvectors of the continuous system transition matrix A and its discrete couterpart $\phi(T)$.

The Caley-Hamilton theory states that a matrix, A, satisfies its own characteristic equation, i.e.

$$A^p + \alpha_1 A^{p-1} + \alpha_2 A^{p-2} + \ldots + \alpha_p I = 0 \qquad (3.14)$$

where α_i are the polynomial coefficients
and p is the order of matrix A.

Equation (3.14) can be rearranged into,

$$\begin{aligned} A^p = {} & \beta_1 \left[(A - \lambda_2 I)(A - \lambda_3 I) \ldots (A - \lambda_p I)\right] + \ldots \\ & + \beta_i \left[(A - \lambda_1 I)(A - \lambda_2 I) \ldots (A - \lambda_{i-1} I)(A - \lambda_{i+1} I) \ldots (A - \lambda_p I)\right] \\ & + \ldots + \beta_p \left[(A - \lambda_1 I)(A - \lambda_2 I) \ldots (A - \lambda_{p-1} I)\right] \end{aligned} \qquad (3.15)$$

where β_i are suitable scalar coefficients
and λ_i are the eigenvalues of matrix A.

Post-multiplying both sides of equation (3.15) by y_i, the eigenvector of λ_i, one obtains

$$A^p y_i = \beta_i \left[(\lambda_i - \lambda_1)(\lambda_i - \lambda_2) \ldots (\lambda_i - \lambda_{i-1})(\lambda_i - \lambda_{i+1}) \ldots (\lambda_i - \lambda_p)\right]$$

$$= \lambda_i^p \, y_i$$

Similarly it could be proved that $A^r y_i = \lambda_i^r y_i$. The matrix $\phi(T)$ is a polynomial in A; thus the eigenvectors of A are also the eigenvectors of $\phi(T)$, independent of the sampling period.

Returning to the main theme of this section, let ideal matrix $P(N-\overline{j+1})$ be contaminated with an additive error matrix $\mathcal{E}(N-\overline{j+1})$, hence one can rewrite the difference equation (2.86) in P(N-j) as

$$\begin{aligned} P(N-j) + \mathcal{E}(N-j) = {} & Q^* + \phi'\left[P(N-\overline{j+1}) + \mathcal{E}(N-\overline{j+1})\right]\phi \\ & - \left\{\phi'\left[P(N-\overline{j+1}) + \mathcal{E}(N-\overline{j+1})\right]G + U\right\} \times \\ & \left\{H^* + G'\left[P(N-\overline{j+1}) + \mathcal{E}(N-\overline{j+1})\right]G\right\}^{-1} \times \\ & \left\{G'\left[P(N-\overline{j+1}) + \mathcal{E}(N-\overline{j+1})\right]\phi + U'\right\} \end{aligned} \tag{3.15}$$

where notation is that of equation (2.82) with all the matrices referring to the completely augmented system.

Assuming $\mathcal{E}(N-\overline{j+1}) \ll P(N-\overline{j+1})$ and ignoring terms with second order terms in the error matrix, the equation (3.15) can be rewritten as

$$\begin{aligned} \mathcal{E}(N-j) \doteq {} & \phi'\mathcal{E}(N-\overline{j+1})\phi \\ & - \left[\phi' P(N-\overline{j+1})G + U\right]\left[H^* + G'P(N-\overline{j+1})G\right]^{-1}\left[G'\mathcal{E}(N-\overline{j+1})\phi\right] \\ & - \left[\phi'\mathcal{E}(N-\overline{j+1})G\right]\left[H^* + G'P(N-\overline{j+1})G\right]^{-1}\left[G'P(N-\overline{j+1})\phi + U'\right] \end{aligned} \tag{3.16}$$

Pre and post multiplying equation (3.16) by y' and y, where y is the eigenvector of any eigenvector λ of matrix $\phi(T)$,

$$\zeta(N-j) = \lambda^2 \zeta(N-\overline{j+1}) - 2\lambda^2 y' \left[P(N-\overline{j+1}) + \sum_{i=0}^{\tilde{\infty}} QG(is)\lambda(is)/\lambda(T)\right] \times \left[H^* + G' P(N-\overline{j+1})G\right]^{-1} G' \varepsilon(N-\overline{j+1})y$$

where $y'\varepsilon(N-j)y = \zeta(N-j)$. (3.17)

To determine the exact stability conditions governing equation (3.17) would, if possible, be a difficult task. Since the objective of this analysis is to develop a computationally stable iteration, a good approximation to the stability condition would be just as useful.

If the last term of equation (3.17) is assumed to be zero, then the resulting equation will be stable, if and only if $|\lambda| < 1$. Since equation (3.17) holds for all the eigenvalues of $\phi(T)$, an approximate computational stability condition for complete equation (3.17) is that all the eigenvalues of $\phi(T)$ lie within the unit circle in the z-plane, i.e. that the uncontrolled system is stable.

Numerical experience indicated that this approximate stability condition was reasonable, since computational stability was only, but not always, found in uncontrolled unstable systems. A combination of computational experience and analytic calculation therefore leads to the conclusion that if the uncontrolled system is unstable, there is more likelihood of computational instability occuring in the E.D.R. iteration.

3.3.2. - PROGRESSIVE COMPUTATIONAL STABILITY.

If one were to assume some discrete state feedback for an otherwise uncontrolled unstable system, so that the resulting system was stable, and thence calculate the optimal state feedback, the optimal state feedback arrangement for the original unstable system could be found by simply adding the assumed and the calculated feedback coefficients. Since a stable system was used in the optimisation calculations, the chances of computational instability would be greatly reduced. All optimally controlled systems of practical importance must be stable; thus, if the assumed state feedback matrix was determined by making progressively better estimates of the overall optimal feedback matrix, one would eventually have a stable system in the optimisation calculations. The author has called this method of determining the E.D.R. matrix sequence and the optimal control policy, the progressive stability method. The details of the latter are given below.

Consider a system described by equation (2.80) and intersample criterion defined in equation (2.79).

Assuming that r approximations to the steady state value of the feedback coefficient matrix have been made, the non-optimal state feedback is defined by

$$\bar{m}(j) = \left[B_1 + B_2 + \ldots + B_r\right] x(j) = \bar{B}\, x(j) \tag{3.18}$$

Let m(j) denote the optimal control input calculated by the digital optimal control technique. The modified systems dynamic state difference equation is then represented by

$$\begin{aligned} x(j+1T) &= \left[\phi(T) + G(T)\bar{B}\right] x(jT) + G(T)\, m(jT) \\ &\triangleq \Phi(T) x(jT) + G(T)\, m(jT) \end{aligned} \tag{3.19}$$

The performance index also is modified and becomes

$$f_{N-j}[x(j)] = E \sum_{j=0}^{N} \left\{ \sum_{i=0}^{n} \left[x'(k)\, Q\, x(k)\right] + \left[\bar{m}(jT) + \tilde{m}(jT)\right]' H \left[\bar{m}(jT) + m(jT)\right] \right\}$$

where notation is that of equation (2.79).

Substituting for $\bar{m}(j)$ from equation (3.18),

$$\begin{aligned} f_{N-j}[x(j)] = E \sum_{j=0}^{N} \Big\{ & x'(jT)\, Q^{**}\, x(jT) + 2\, x'(jT)\, U^{*}\, m(jT) \\ & + m'(j)\, H^{*}\, m(j) \Big\} \end{aligned} \tag{3.20}$$

where $Q^{**} = \sum_{i=0}^{n} \Phi'(is)\, Q\, \Phi(is) + \bar{B}'\, H \bar{B}$

$$H^{**} = H + \sum_{i=1}^{n} G'(is)\, Q\, G(is)$$

and $$U^{*} = \sum_{i=1}^{n} \Phi'(is)\, Q\, G(is) + \bar{B}'\, H$$

The structure of state equations (3.19) and (2.81) and the performance index equations (3.20) and (2.82) are identical, thus the difference equations for P(N-j) and Z(N-j) could be obtained by the substitution of

$$\Phi(T) \quad \text{for} \quad \phi(T)$$
$$Q^{**} \quad \text{for} \quad Q^{*}$$
$$H^{**} \quad \text{for} \quad H^{*}$$
$$\text{and} \quad U^{*} \quad \text{for} \quad U$$

in equations (2.86) and (2.87) respectively. By expanding these equations in terms of the original variables $\phi(T)$, Q^{*}, H^{*} and U, and then performing some elementary matrix manipulations, it can be shown that the sequence of E.D.R. matrices calculated by the progressive stability technique is identical to the sequence that would be ideally generated by direct method, if there was no computational instability. In other words, the sequence of matrices P(N-j) including the initial condition matrix is independent of the estimated value of the feedback coefficient matrix, provided the resulting system is computationally stable. The overall optimal control policy could therefore either be calculated by the addition of the estimated and calculated modified optimal feedback coefficients or from (cf equation(2.85)).

$$B(N-\overline{j+1}) = -\left[H^{*} + G'(T)\ P(N-\overline{j+1})\ G(T)\right]^{-1} \times \left[G'(T)\ P(N-\overline{j+1})\,\phi(T) + U'\right]$$

Details of the algol procedure used to determine the optimal control policy by the progressive stability method are given in Appendix (C). The author has used this procedure in many extremely

computationally unstable systems, and in no case did the iteration fail to eventually converge to the true optimal solution.

3.3.3. - TEST FOR OPTIMALITY.

Since the non-linear equation governing the error matrix (cf equation (3.16)) has time dependent coefficients, in the form of matrix $P(N-\overline{j+1})$, it is possible that the error matrix may build up over part of the optimisation interval, but not diverge excessively. This could result in undetected errors in the E.D.R. matrix, and therefore a test for optimality is required.

Due to the mechanics of the solution of the digital optimisation problem, the substitution of the calculated values of $P(N-j)$ and $m(j)$ into the equation defining the first variation of the performance index with respect to the control input, will always give a null vector, even when the calculated values of $P(N-j)$ are in error. The only sure method of confirming optimality is therefore to perturb the elements of the feedback coefficient matrix about the calculated value and determine the resulting performance index. If the latter is greater than the calculated optimal index for small perturbations, the calculated optimal control policy is truly optimal.

3.4. - PERFORMANCE INDEX INSTABILITY

In section (3.2) the sufficient conditions for complete controllability were found by considering the discrete version of the system's matrix dynamic equation. This condition is therefore a function of the sampling period, with the result that the complete controllability condition could be violated for some periods, whereas it could hold for others. Since the complete controllability conditions are only sufficient, their violation does not necessarily mean that an optimal control policy does not exist. Nevertheless, one could suspect that if the condition was a good approximation to the necessary condition, the performance index would change rapidly in the vicinity of the sampling periods for which the condition was violated. Such a behaviour of the performance index has been called index instability. A particularly interesting situation arises when one is attempting to control a system described by either differential or difference equations and the system has at least one pair of complex roots.

3.4.1. - INSTABILITY SAMPLING PERIOD

In subsection (3.2.4.) a continuous system was considered in its transformed discrete jordon cannonic form. For a system with one control variable, consider the last row of two jordon blocks, associated with eigenvalues λ_k and λ_l of the continuous system's equation, are

$$x_k(i+1) = e^{\lambda_k T} x_k(i) + \bar{g}_k\, m(i) \tag{3.21}$$

$$x_l(i+1) = e^{\lambda_l T} x_l(i) + \bar{g}_l\, m(i) \tag{3.22}$$

where $\bar{g}_k$ and $\bar{g}_l$ are assumed non-zero.

Pre-multiplying equation (3.21) and equation (3.22) by $\bar{g}_l$ and $\bar{g}_k$ respectively and subtracting gives

$$\bar{g}_l\, x_k(i+1) - \bar{g}_k\, x_l(i+1) = \bar{g}_l\, e^{\lambda_k T} x_k(i) - g_k\, e^{\lambda_l T} x_l(i) \,. \tag{3.23}$$

Let $e^{\lambda_k T} = e^{(\alpha_k + j\beta_k)T}$

and $e^{\lambda_l T} = e^{(\alpha_l + j\beta_l)T}$

where $\alpha_k, \beta_k, \alpha_l, \beta_l$ are real quantities

and $j = (-1)^{\frac{1}{2}}$

Assuming that $\alpha_k = \alpha_l$, one can rewrite equation (3.23) as

$$y(i+1) = e^{(\alpha_k + j\beta_k)T} \left[\bar{g}_l\, x_k(i) - \bar{g}_k\, x_l(i) e^{j(\beta_l - \beta_k)T} \right]$$

where $y(i+1) = \bar{g}_l\, x_k(i+1) - \bar{g}_k\, x_l(i+1)$

If $(\beta_l - \beta_k)T = 2h\pi$ for $h = 0, 1, 2 \ldots$ then

$$y(i+1) = e^{\alpha_k T + j\beta_k T}\, y(i) \tag{3.24}$$

Equation (3.24) indicates that the value of x_k and x_l cannot be controlled independently and thus the system is uncontrollable in the complete sense.

The assumptions made concerning the eigenvalues were

$$\mathrm{Re}(\lambda_k) = \mathrm{Re}(\lambda_1) \tag{3.25}$$

$$\text{and } \left[\mathrm{Im}(\lambda_k) - \mathrm{Im}(\lambda_1)\right] T = 2h\pi \qquad h = 0, 1, 2 \ldots \tag{3.26}$$

where Re and Im refer to Real and Imaginary parts respectively.

These are approximately the conditions which Kalman stated, without proof in paper (35), must be violated to ensure that the complete controllability conditions derived for continuously controlled systems automatically hold for its discrete formulation with sampling period T. The difference between equations (3.25) and (3.26) and those of Kalman is that in the latter h is not permitted to have a zero value. If $h = 0$, then either $T = 0$ and/or $\left[\mathrm{Im}(\lambda_k) - \mathrm{Im}(\lambda_1)\right] = 0$ for conditions expressed in equation (3.26) to hold. When $T = 0$, there can be no control input, since $G(T) = 0$; thus system is uncontrollable in both complete and practical sense. It should be noted that making $T = 0$ is not equivalent to a continuous input, but making $T = \varepsilon$, for $\varepsilon > 0$ and $\varepsilon \to 0$, is the limiting case of the digitally controlled system. Nevertheless, as $T \to 0$, the system will tend to become uncontrollable, according to the sufficient conditions derived; thus, unless the controllability conditions are poor approximations to the necessary conditions, one would expect large deviations in the index as $T \to 0$. Numerical experience and intuition would lead one to conclude that as $T \to 0$, the performance of the system would improve with a corresponding reduction in performance

index and therefore the controllability conditions must be poor approximations to the necessary conditions as $T \rightarrow 0$.

Returning to the conditions of equations (3.25) and (3.26), there are only two cases likely to occur.

(a) Real equal eigenvalues with separate jordon blocks — Since $[Im(\lambda_k) - Im(\lambda_l)] = 0$, the equations (3.25) and (3.26) will hold for all T. The system is therefore uncontrollable (cf sub-section (3.24), condition (a) for uncontrollability).

(b) Complex pair of eigenvalues —
Since such eigenvalues are very common in systems, the following results are of considerable importance. For a complex pair with natural frequency w_n, the uncontrollability or index instability period T_s is given by equation (3.26), i.e.

$$T_s = h\pi/w_n = hT_n/2 \qquad \text{for } h = 1, 2, \ldots$$

where T_n is the natural period.

It is current design practice to sample at a period smaller than T_n, if the complex roots are the predominant roots. Nevertheless, as

is shown in following example, the side effect of an increased performance index also extends to periods considerably less than the critical period. Finally, if the system had several pairs of complex roots, then there may be several sets of periodic index instability critical periods.

The above analysis was carried out under the initial assumptions that there was only one control input variable and $\bar{g}_k$ and $\bar{g}_l$ are non-zero. When there are more than one control variable, then equations (3.25) and (3.26) must hold, and the rows of cannonical input matrix, corresponding to last rows of the jordon blocks associated with eigenvalues λ_k and λ_l, must also be linearly dependent for uncontrollability and, hence, index instability to occur(cf subsection 3.2.4.). Hence, a method of overcoming the index instability problem would be to increase the number of independent control inputs to at least the calculated lower bound value. It should be remembered that, as was noted in sub-section (3.2.4.), such a move does not necessarily result in a controllable system, but with a suitable choice of the systems input matrix controllability can be achieved.

Finally, it should be noted that the index instability is

not inevidence in all systems with complex poles. In some cases, notably stable uncontrolled systems, the performance index is not adversely affected in the vicinity of the theoretical instability period or multiples thereof. In such cases, the sufficient conditions for controllability are not good approximations to the undetermined necessary conditions for controllability.

3.4.2. - CONTROLLABILITY CHARACTERISTIC NUMBER.

The problem of index instability highlights the desirability of calculating whether a system obeys the controllability condition or not. The rank of the condition matrix C(p-1) can be calculated in several ways, the most common being to form an echelon matrix from the original matrix, and thence the rank is equal to the number of rows which contain at least one non-zero element. Unfortunately, the accuracy of such a method becomes questionable when the matrix tends to having linearly dependent columns.

It would be more suitable and informative to quantitatively represent the quality of the controllability of a system by a scalar quantity rather than the boolean rank condition.

Consider the controllability condition

$$\text{Rank } [C(p-1)] = p$$

where matrix C(p-1) is rectangular matrix of dimensions $p \times q$.

But $\quad \text{Rank } [C(p\text{-}1)] = \text{Rank } [C(p\text{-}1)\ C'(p\text{-}1)]$

where dimension of $C(p\text{-}1)\ C'(p\text{-}1)$ is $p \times p$, i.e. square.

Since the matrix is square, its rank can be determined by evaluating its determinant, which is zero if the rank is less than the order of the matrix. The evaluation of the determinant by standard procedures also becomes inaccurate when the matrix tends to having linearly dependent columns. The inaccuracy results from subtracting two almost equal numbers, the resulting number having an excessively large round-off error. When the latter number is used in further calculations, the error builds up with the associated reduction in accuracy. A second drawback of the direct determinant approach to controllability determination is that one arrives at a single number, upon which one must base a decision upon the quality of the linear dependence of the rows.

To overcome the inaccuracies and to provide some quantitative insight, the author has written a short but involved program, which evaluates the determinant of any square matrix by the fundamental expansion of the cofactors (cf Appendix D). In this program, the determinant calculation was performed in such a way that no numbers of opposite sign were added together and no numbers of equal sign were subtracted from each other. The result of the calculation is two numbers which represent the positive and negative part of the determinant, the latter being the sum of these two parts. The quality of the linear

dependence of the rows of the matrix can be gauged by examining by how much the two parts differ from each other in absolute value, compared to the sum of their absolute values. For complete dependence, the positive and negative parts would have equal absolute value.

Let α and $-\beta$ be the positive and negative parts respectively, then $|\ C(p-1)\ C'(p-1)\ | = \alpha - \beta$ and the characteristic controllability number $= (\alpha-\beta)/(\alpha+\beta)$. The characteristic number has a maximum value of unity and a minimum value of zero, the smaller the number the more dependent are the rows of the matrix. The number is indeterminant when $\alpha = \beta =$ 0, but in this case the determinant is zero anyway.

There are two drawbacks to the above method. Firstly, the calculation time can become excessive for systems of large order, since the number of multiplicative and additive operations is a function of the factorial of the order of the matrix. Secondly, the controllability conditions are only sufficient, thus the results so obtained must be viewed with caution.

3.4.3. - EXAMPLE

Consider a system described by the matrix state differential equation

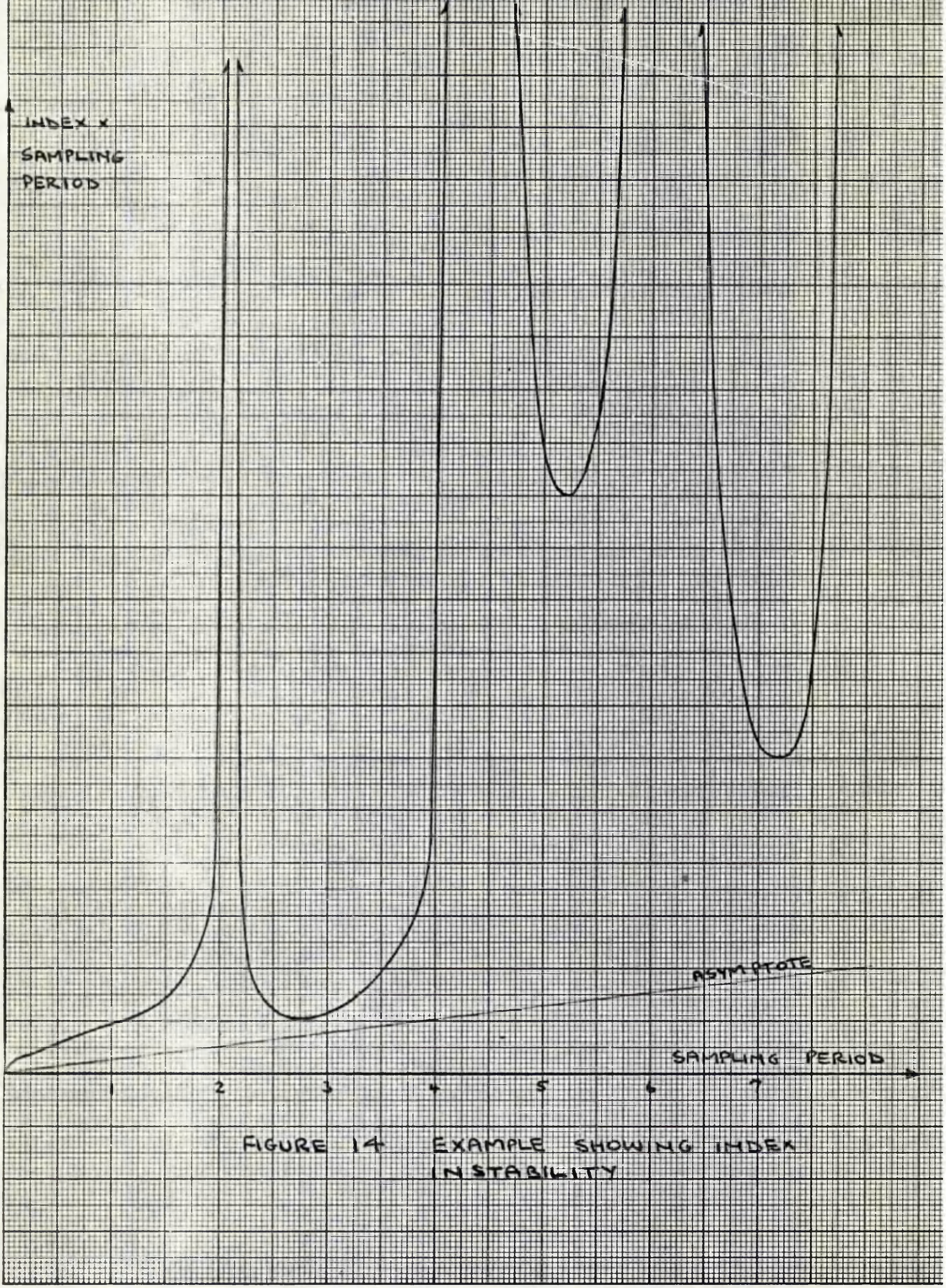


FIGURE 14 EXAMPLE SHOWING INDEX INSTABILITY

$$\begin{bmatrix}\dot{x}_1(t)\\ \dot{x}_2(t)\\ \dot{x}_3(t)\end{bmatrix} = \begin{bmatrix}-0.5 & 1 & 0\\ 0 & 0 & 1\\ -1 & -2 & 0\end{bmatrix}\begin{bmatrix}x_1(t)\\ x_2(t)\\ x_3(t)\end{bmatrix} + \begin{bmatrix}0\\ 0\\ 1\end{bmatrix} m_1(t)$$

Since the results of this example are to be reused in Chapter 5, the performance criterion chosen was the quadratic summation criterion

$$f_{N-j}[x(j)] = \sum_{j=0}^{\infty} x_1^2(j) + x_2^2(j) + x_3^2(j)$$

(It should be noted that similar results also hold for the intersample criterion).

The uncontrolled continuous system is unstable with eigenvalues

$$\lambda_{1,2} = 1.820408_{10} - 1 \pm j\, 1.510449_{10} + 0$$

and $\lambda_3 = -8.640816_{10} - 1.$

From the analysis of sub-section (3.4.1), the instability period has value

$$T_s = \pi/w_n = 2.0799_{10} + 0.$$

Index instability is therefore likely to, and in fact does, occur at all positive integer multiples of this period, as can be seen in Figure (14), which is the graph of index against sampling period for initial conditions $x_1(0) = x_2(0) = x_3(0) = 1$ unit. Although the values of index at multiples of the index instability period were excessively large, they were in fact finite quantities. An optimal

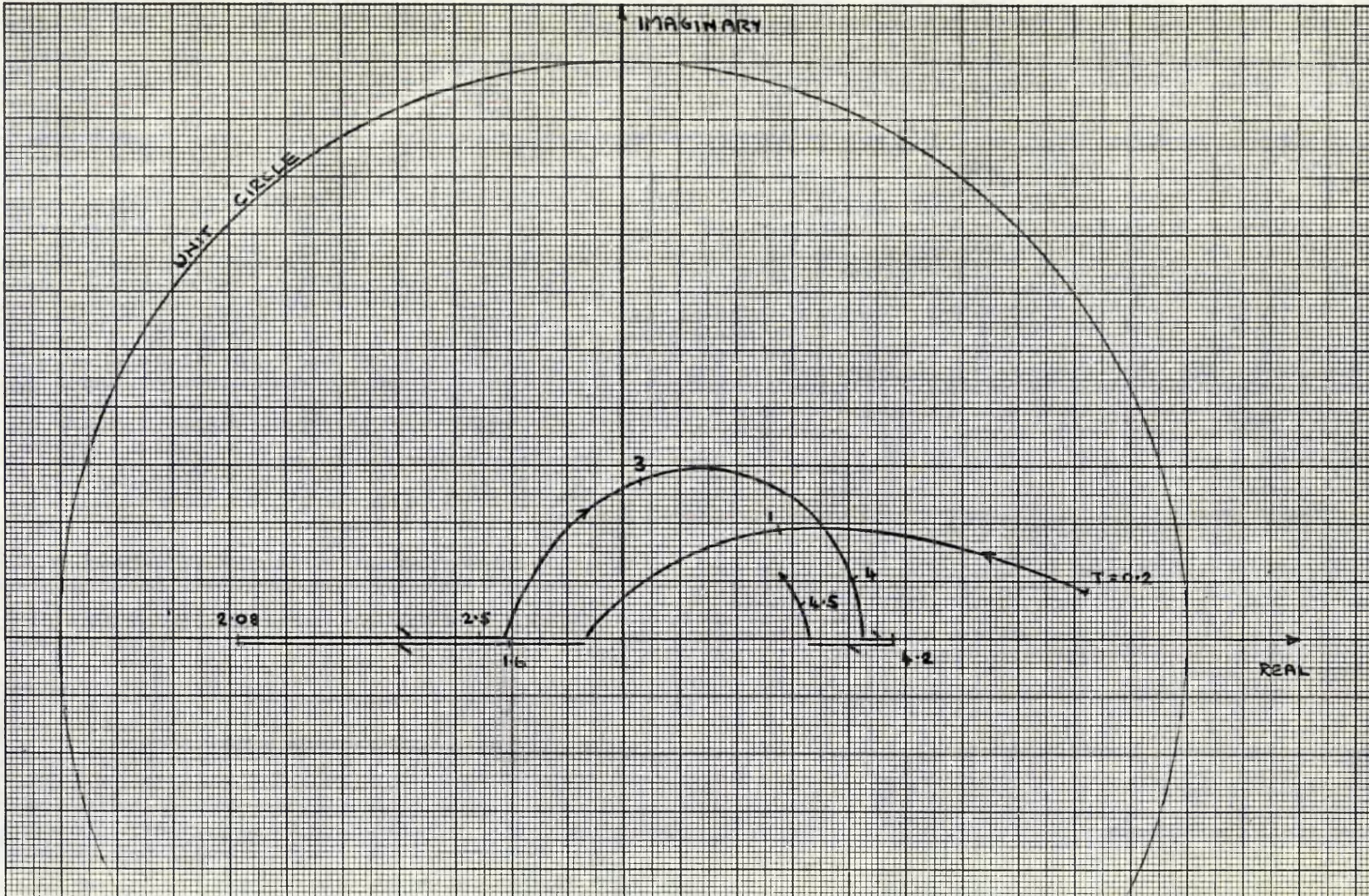


FIGURE 15 HALF Z-PLANE ROOTS

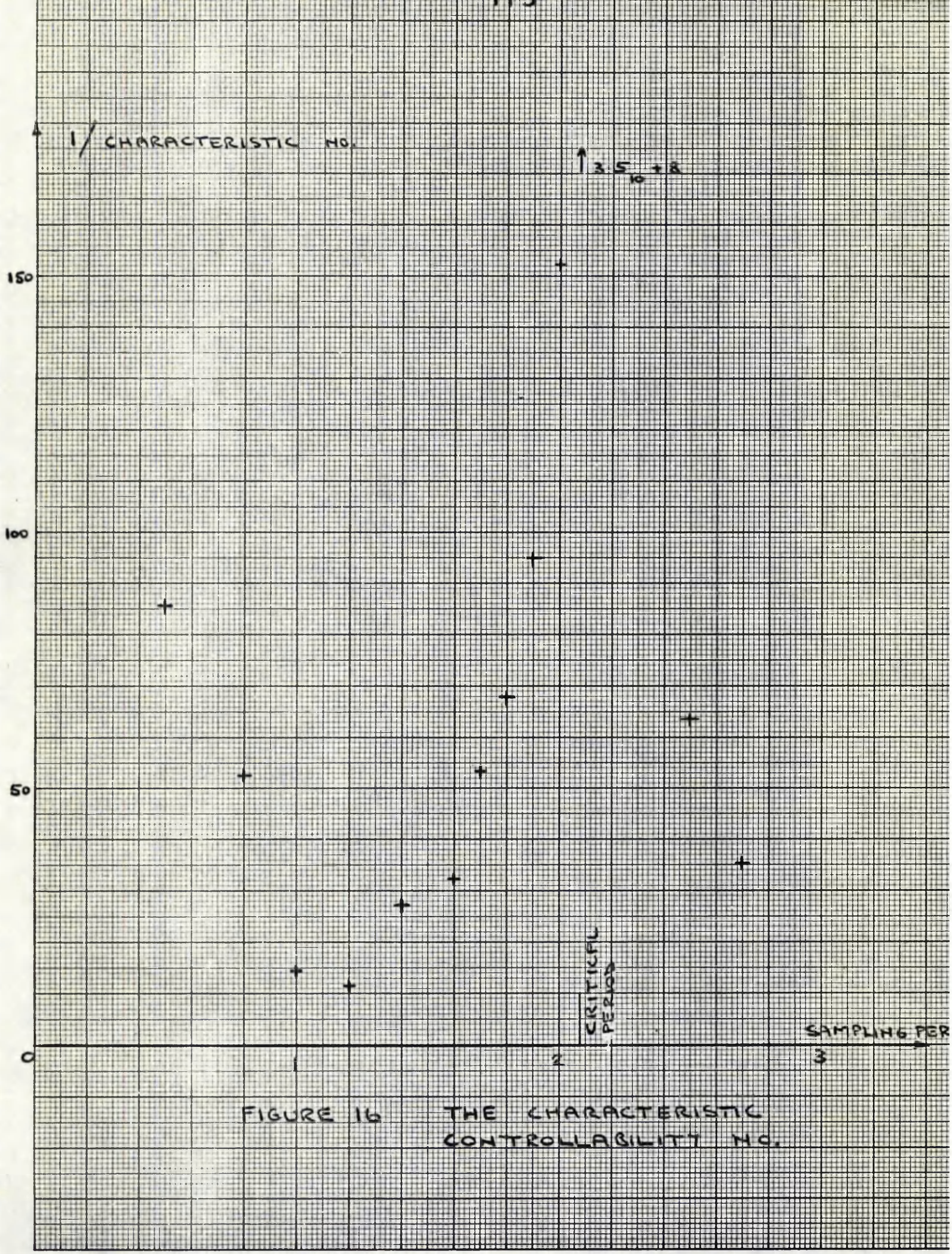


FIGURE 16 THE CHARACTERISTIC CONTROLLABILITY NO.

control policy was calculable, even although the complete controllability condition was violated, which indicates the sufficiency rather than the necessity of the controllability condition. The finite nature of the performance criterion can be seen by examining Figure (15), which is the locus of the complex pair of roots of the controlled system. The third root is zero for all sampling periods, which is later proved in Chapter 5 to be case for all single control input systems. The controlled system is therefore stable for all sampling periods except for the trivial value $T = 0$; thus the performance index is always finite.

The inverse of the controllability characteristic number is plotted against the sampling period in Figure (16). Due to the sufficiency of the condition, one cannot expect an exact correspondence between the performance index curve of Figure (14) and Figure (16). Nevertheless, it can be seen that there is reasonable agreement in the shape of the curves in the vicinity of the instability periods. Figure (16) shows that as $T \rightarrow 0$, the quality of controllability deteriorates as was discussed in sub-section (3.4.1).

The lower bound on the control inputs is two at the instability periods and one at the other periods. Increasing the number of independent control inputs to two, such that the system is described by state equation

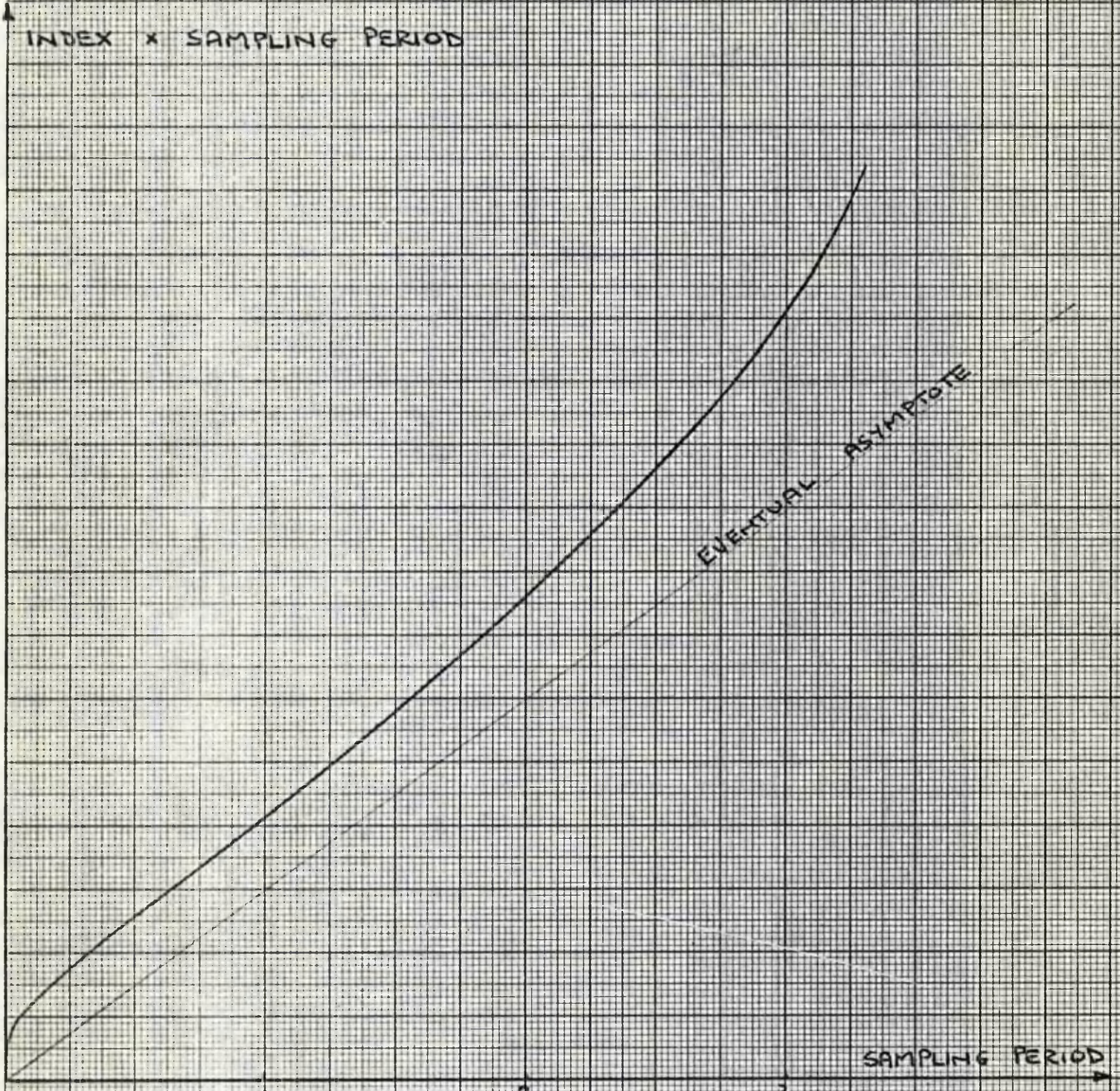


FIGURE 17 INDEX AGAINST SAMPLING PERIOD FOR TWO CONTROL VARIABLES

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \end{bmatrix} = \begin{bmatrix} -0.5 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -2 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} m_1(t) \\ m_2(t) \end{bmatrix}$$

will therefore result in an optimal system, which does not exhibit index instability. This can be seen from Figure (17), which is the graph of the performance index against sampling period. The appearance of divergence at high values of sampling period is due to the effect of sampling at a frequency approaching one of the uncontrolled system's natural frequencies. This phenomenon will be considered in detail in Chapter 5. The natural frequency, in the particular case above, is the natural frequency of the complex eigenvalues.

CHAPTER 4 - SENSITIVITY OF OPTIMAL DIGITAL SYSTEMS

4.1. - THE SENSITIVITY PROBLEM

The optimal control policy has been derived under the basic assumption that one has a complete and accurate knowledge of the dynamic equations governing the system. In reality, the designer is rarely, if ever, in this ideal position, since only approximate values of the system's parameters are usually available. In most physical systems, the parameters also change due to ageing, variations in working conditions etc. The designer must therefore produce a controlled system, whose response is insensitive to variations in system parameters, and thus compensate for his lack of information. The situation also arises, in which the designer wishes to design one controller, which is to be used with several systems with slightly different parameter values. Here again the controller must be designed to reduce the effect of different parameter values (36).

One of the properties of a well designed feedback system is the reduced sensitivity of the overall response due to changes in system parameters (13), (37). Since the implementation of the optimal digital control policy for linear systems results in an optimal feedback configuration, it would be hoped that the optimal system would be insensitive to parameter variations. Kalman, in his paper (38) on the inverse optimisation problem for continuous systems, proved that the absolute value of the return difference of classical control theory is

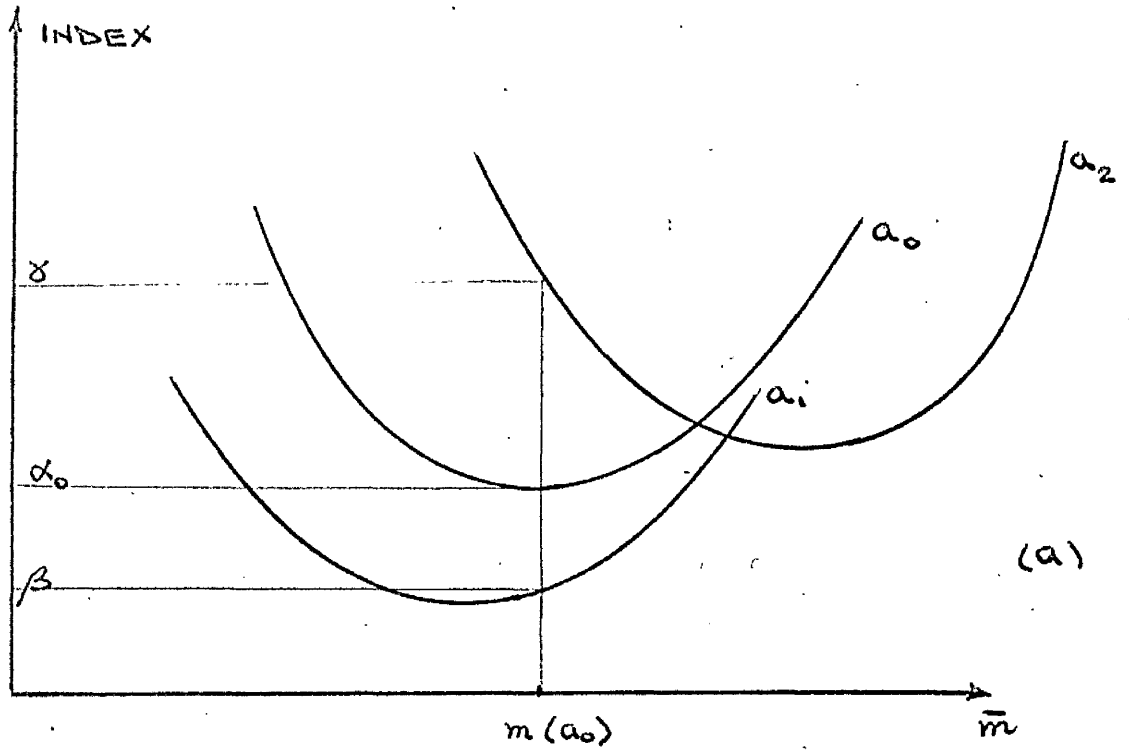
always less than unity for an optimal linear state feedback system. This means that sensitivity of the system to parameter perturbations in the open loop configuration is always reduced by the addition of the state feedback loops. Similar results can also be shown to hold for optimal digital systems. The degree of insensitivity is not specifically defined, and therefore it would be better to design the optimal system, subject to the constraint that the resulting system has minimum sensitivity with respect to the parameter variations. This chapter is devoted to the attempted solution of this problem. In particular, the sensitivity of the optimal system with respect to variations in the coefficients of the state equations and the feedback coefficients will be examined.

4.2 - LARGE PERTURBATION SENSITIVITY FUNCTIONS

Before the sensitivity of a system's response can be examined, it must be defined. Since the optimality of the system's response is determined by the performance criterion, it seems only logical to consider the variations in the performance criterion due to variations in parameters of the system, as a measure of sensitivity. (41),(39),(40). One of the assets of such a measure is that it is a scalar quantity.

The terminology used in this chapter concerning the performance index varies slightly from that previously used, so it is explained below.

(1) The performance index $F[m(a_i), a_j]$ is the index



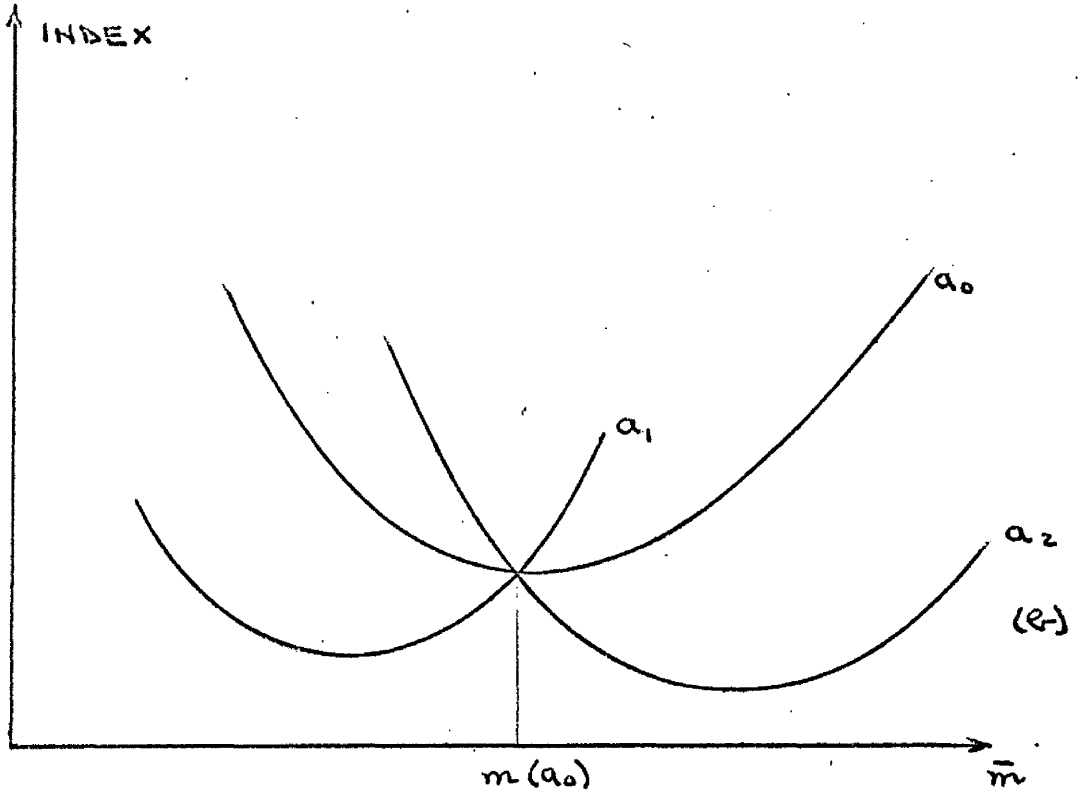


FIGURE 18 ABSOLUTE SENSITIVITY

which results when the optimal feedback coefficient calculated for a system with system parameters a_i is used in conjunction with a system with parameters a_j.

(2) The performance index $F[m(b_i), a_j]$ is the index which results when control feedback coefficient matrix with parameters b_i is used in conjunction with a system with parameters a_j.

4.2.1. - ABSOLUTE AND RELATIVE SENSITIVITY FUNCTIONS

The first sensitivity function to be considered is the absolute sensitivity function which, as the name suggests, is a measure of the absolute deviation of the performance of the perturbed system away from its optimal value calculated for the unperturbed system (39). In mathematical terms it is

$$\Delta F_{abs} = F[m(a_o), a] - F[m(a_o), a_o]$$

A graphical illustration of this sensitivity function is given in Figure (18a), where the performance index is assumed, for simplicity of argument, to be a function of a scalar control quantity $\bar{m}$ with $\bar{m} = m(a_o)$ being the optimal value of this control quantity for a system with parameters a_o. The curves for three sets of system parameters a_o, a_1 and a_2 are shown.

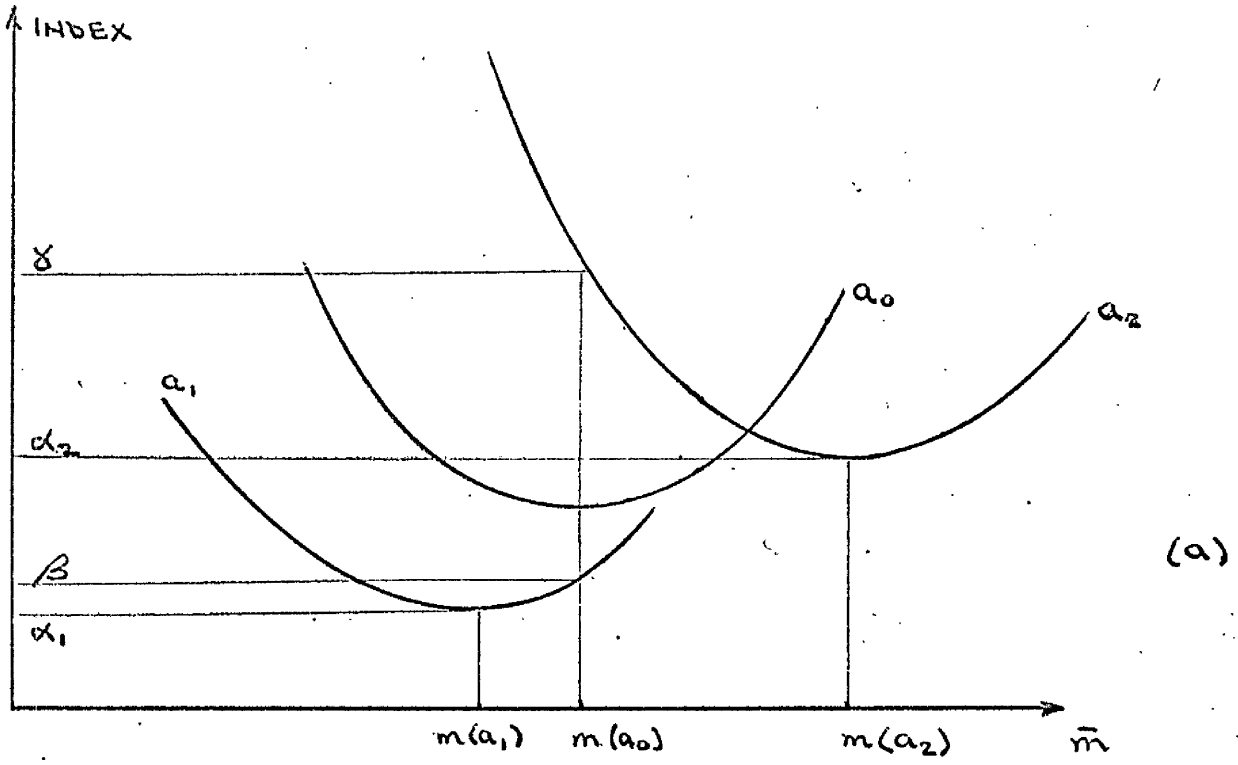
Assuming parameter values a_1 and a_2 represent perturbed values of the ideal parameter value a_o, the sensitivity functions are

$$\Delta F_{abs} = \gamma - \alpha_o.$$

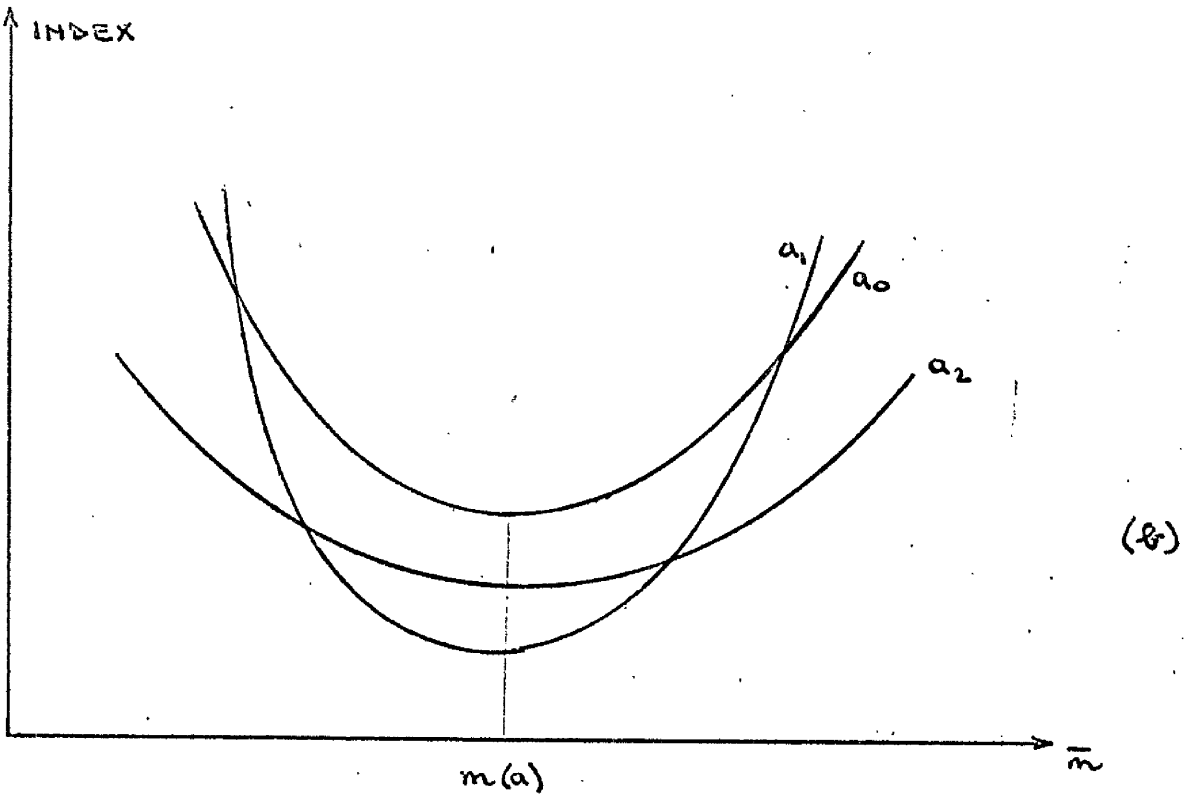
and $$\Delta F_{abs} = \beta - \alpha_o.$$

As can be seen from Figure (18a), it is possible for the absolute sensitivity function to be negative. This indicates that an alteration in the parameter values causes a reduction in the performance index and, consequently, a better system's response. Ideally, one would like to design a system such that any parameter perturbation would result in an increase in the performance index, when compared with the optimal performance index. It is assumed henceforth that the parameter values a_o have been chosen due to their having some significant value, e.g. centre of the range in which parameter values are known to lie.

Figure (18b) shows the situation where the sensitivity functions are zero in both cases of parameter deviations. Assuming a_o is fixed, designing a system so as to minimise the absolute sensitivity function would result in a system which is independent of any considerations about the optimal performance index, which could be achieved by using the optimal control policy of the perturbed system. As a means of overcoming this shortcoming of the absolute sensitivity function, a second sensitivity function is introduced. It is called the relative sensitivity function, as it is a measure of the performance of the perturbed system relative to the optimal performance that could be achieved, using the optimal control



(a)



(b)

FIGURE 19 RELATIVE SENSITIVITY

policy for the perturbed system. In mathematical terms it is

$$\Delta F_{rel} = F[m(a_o), a] - F[m(a), a]$$

Figure (19a) is a graphical illustration of the relative sensitivity function. The control quantities $m(a_1)$ and $m(a_2)$ are the optimal control scalar quantities associated with the systems with perturbed parameters a_1 and a_2 respectively.

Assuming that a_o are the unperturbed parameters, the sensitivity functions are

$$\Delta F_{rel} = \beta - \alpha_1$$

$$\Delta F_{rel} = \gamma - \alpha_2$$

Since $F[m(a), a]$ is always less than $F[m(a_o), a]$ by the definition of optimality, the relative sensitivity function is always non-negative.

Figure (19b) shows the ideal case where the sensitivity functions are both zero for perturbed values a_1 and a_2. This figure illustrates the shortcoming of a design procedure, which attempts to develop a system with minimal relative sensitivity. No consideration is given to the absolute value of the optimal performance index of the perturbed system relative to that of the unperturbed system. A measure of this deviation has already been defined as the absolute sensitivity function. Although neither sensitivity function is ideal from the design point of view, a combination of both would result in a reasonable design, since

the shortcomings of one are compensated by the other sensitivity function. Since the two functions are inter-related, it is necessary to minimise both functions simultaneously, rather than design a system which minimises one, then redesign to minimise the other.

4.2.2. - THE CONTROLLER SENSITIVITY FUNCTION

The ideal coefficient feedback matrix will be accurately known in practice, but parameter variations will occur when the controller is implemented. The corresponding controller sensitivity function has been defined as the difference between the performance index of the system with a perturbed feedback coefficient matrix and the optimal index for the unperturbed system, i.e.

$$\Delta F_{cont} = F\left[m(b),\ a_o\right] - F\left[m(b_o),\ a_o\right].$$

where b_o are the optimal coefficients for system with parameters a_o.

Due to the definition of optimality, the controller sensitivity function will always be greater than or equal to zero.

4.3 - SMALL PERTURBATION SENSITIVITY FUNCTIONS

Having developed some useful sensitivity functions for large parameter perturbations, one is faced with use of these functions in the design problem. The method of attack adopted was to consider the

corresponding sensitivity functions for infinitely small perturbations of the parameters and use these to impose sensitivity constraints on the optimal solution. This type of approach is not uncommon in traditional control theory, since it usually allows a general analytic solution to a problem, for which no general large perturbation solution can easily be implemented or derived.

4.3.1. - MINIMAL ABSOLUTE SENSITIVITY.

Assuming the state equation coefficients, i.e. elements of matrices A and D, are perturbed by an infinitely small quantity δa_o, the corresponding absolute sensitivity function can be redefined as

$$\delta F_{abs} = F[m(a_o),\ a_o + \delta a] - F[m(a_o),\ a_o] \tag{4.1}$$

Expanding $F[m(a_o),\ a_o + \delta a]$ in a Taylor series truncated at the second term, equation (4.1) becomes

$$\delta F_{abs} = \sum_i \partial F[m(a_o),\ a] / \partial a_i \Big|_{a=a_o} \delta a_i \tag{4.2}$$

where a_i represents the elements of the vector of all the perturbed variables.

In matrix notation equation (4.2) can be written

$$\delta F_{abs} = \partial F[m(a_o),\ a] / \partial a \Big|_{a = a_o} \delta a$$

It is required to design a system in which δF_{abs} is a minimum, which is equivalent to making $\partial F[m(a_o),\ a] / \partial a$ a minimum. In particular

consider a digital control system

$$\dot{x}(t) = A\,x(t) + D\,m(t) \tag{4.3}$$

which has to be controlled in such a way as to minimise the quadratic summation criterion

$$F[m(a_o),\ a] = \sum_{j=0}^{N} J[x(j),\ m(a_o)]$$

$$= \sum_{j=0}^{N} x'(j)\ Q\ x(j) + m'(j)\ H\ m(j) \tag{4.4}$$

But $\partial F[m(a_o),\ a]\ /\partial a = \sum_{j=0}^{N} \partial J[x(j), m(a_o)]\ /\partial x(j) \times \partial x(j)\ /\partial a$

By definition

$$J[x(j),\ m(a_o)] = x'(j)\ Q\ x(j) + m'(j)\ H\ m(j)$$

$$= x'(j)[Q + B'(N-j)\ H\ B(N-j)]\ x(j)$$

where B(N-j) is the coefficient feedback matrix.

Hence $\partial J[x(j),\ m(a_o)]\ /\partial x(j)\ \Big|_{a=a_o} = [2x'(j)Q + 2x'(j)\ B'(N-j)HB(N-j)]\Big|_{a=a_o}$

where defining $a = a_o$ means that x(j) is the state trajectory with parameters a_o.

Let $\partial x(j)/\partial a$ be denoted by y(j) and let it be assumed for the moment that the dynamic equations governing the trajectory of y(j) can be determined. Then

$$\partial F\,[x(j),\, m(a_o)]\,/\partial\, x(j)\;\Big|_{a=a_o} = \sum_{j=o}^{N} [2x'(j)\; Q\; y(j) + 2x'(j)\; B'(N-j) H\; B(N-j)\; y(j)\;] \qquad (4.5)$$

Optimal control requires the minimisation of equation (4.4), whereas minimal absolute sensitivity required the minimisation of equation (4.5). The similarity of the structures of the equations (4.4) and (4.5) suggests the use of the function $\partial F[x(j),\, m(a_o)]/\partial x(j)$ as a constraint on the system's response by including it in the performance criterion. The drawback, which balks the analysis, is that the coefficient feedback matrix $B(N-j)$ is not known at the outset of the calculation, and thus is not a permissible variable in the criterion. If the perturbed system is considered to have an open loop structure, with the control input that of the unperturbed system, the coefficient feedback matrix would not appear in equation (4.5). One is essentially considering an open loop rather than a closed loop or feedback system. Since the optimal closed loop system has been shown (38) to be less sensitive than the open loop structure, a design method based on the open loop configuration should also give satisfactory results for the actual closed loop system. Note that if the control cost matrix is zero, there is no problem, since $B(N-j)$ does not appear in equation (4.5).

Under the assumption of open loop control

$$\partial F\left[x(j),\ m(a_o)\right]/\partial x(j)\Big|_{a=a_o} = \sum_{j=o}^{N} 2x'(j)Q\ y(j)\Big|_{a=a_o}$$

A combined performance criterion is then

$$F_1\left[x(j),\ y(j),\ m(a_o)\right] = \sum_{j=o}^{N} x'(j)Q\ x(j) + m'(j)H\ m(j) + 2w_1\, x'(j)Q\ y(j) \qquad (4.6)$$

where w_1 is a weighting constant.

This criterion is sign indefinite, due to term $2w_1 x'(j)Q\ y(j)$; thus the optimisation methods already discussed are not applicable. To ensure at least sign semi-definiteness, a further term $w_2 y'(j)Q\ y(j)$ is required to be added. The variable $y(j)$ was defined as the rate of change of the trajectory with the perturbed parameters, and therefore represents a measure of system sensitivity, which has been called the state sensitivity function. The inclusion of $y(j)$ in the performance criterion in a suitably weighted quadratic form is therefore mathematically and practically justified.

Using the augmented state vector $\left[x(j)\ \vdots\ y(j)\right]'$ the final combined performance criterion becomes

$$F_2\left[x(j),\ y(j),\ m(a_o)\right] = \begin{bmatrix} x(j) \\ y(j) \end{bmatrix}' \begin{bmatrix} Q & w_1 Q \\ w_1 Q & w_2 Q \end{bmatrix} \begin{bmatrix} x(j) \\ y(j) \end{bmatrix} + m'(j)Hm(j) \qquad (4.7)$$

where w_1 and w_2 are the weighting factors for the absolute and state sensitivity functions.

The remaining problem is the development of the dynamic equations governing y(t). Expressing equation (4.3) in partial derivative form, since x(j) is a function of the perturbed parameter values, and the independent time variable t, one obtains

$$\partial x(a,t)/\partial t = A(a)\, x(a,t) + D(a)\, m(t) \tag{4.8}$$

But $\frac{\partial}{\partial t}\left[\partial x(a,t)/\partial a\right] = \frac{\partial}{\partial a}\left[A(a)\, x(a,t) + D(a)\, m(t)\right]$

$$= \frac{\partial\left[A(a)\, x(a,t)\right]}{\partial x(a,t)} \cdot \frac{\partial x(a,t)}{\partial a} + \frac{\partial}{\partial a}\left[A(a)\, x(a_o,t) + D(a)\, m(t)\right]$$

where it has been assumed that m(t) is the optimal open loop input of unperturbed system.

Therefore

$$\partial y(t)/\partial t \Big|_{a=a_o} = A(a_o)y(t) + \partial\left[A(a)x(a_o,t) + D(a)m(t)\right]/\partial a \Big|_{a=a_o} \tag{4.9}$$

where stipulation that $a = a_o$ determines that the equation describes the dynamics of y(t) for the unperturbed system.

Initially, the rate of change of x(0) with respect to the perturbed parameters, i.e. y(0), will be zero, since the initial condition of the state is independent of the perturbed parameters. The dynamic equations governing y(t) have thus been completely defined. Using

the augmented state vector $\left[x(t) \,\vdots\, y(t)\right]$ the dynamic equations became

$$\begin{bmatrix} \dot{x}(t) \\ \hdashline \dot{y}(t) \end{bmatrix} = \begin{bmatrix} A(a_o) & \vdots & 0 \\ \hdashline \alpha(a_o) & \vdots & A(a_o) \end{bmatrix} \begin{bmatrix} x(t) \\ \hdashline y(t) \end{bmatrix} + \begin{bmatrix} D \\ \hdashline \beta(a_o) \end{bmatrix} m(t) \qquad (4.10)$$

$$\text{where } \alpha(a_o)\, x(t) + \beta(a_o)\, m(t) = \partial[A(a)x(a,t)+D(a)m(t)]/\partial a \Big|_{a=a_o}$$

Using the augmented state equation (4.10) and the performance criterion of equation (4.7), the resulting optimally controlled system will have its sensitivity to parameter variations constrained in the absolute and state sensitivity senses.

The formulation of the performance criterion with sensitivity constraints for the intersample criterion can most easily be dealt with by considering the equivalent continuous integral criterion. The results exactly correspond to those for the quadratic summation criterion.

In this sub-section, by making some reasonable assumptions, a procedure has been developed, which, when implemented, results in an optimal system, whose combined absolute and state sensitivity is minimised. The introduction of the additional state vector $y(j)$ and the need for its value to be known at each sampling instant to implement the constrained system's control policy, are inherent disadvantages of the above design procedure. From the augmented state dynamic equations it can be seen that the state vector $x(j)$ is completely uncoupled from the state vector $y(j)$, if the control input is independent of $y(j)$. In Chapter 5, a method

of determining a control input subject to such an independence constraint will be illustrated. The net result would therefore be a design procedure, which allows one to determine a sub-optimal control policy for the unaugmented system, the sub-optimal system being less sensitive to parameter variations than the optimal system.

The above technique for including a sensitivity constraint, with or without the independence constraint, did not prove to be as useful in optimal system's design as had been originally hoped for. Although, by a suitable weighting of the sensitivity segments of the performance index, it was always possible to reduce the sensitivity of the optimal system to parameter variations, a very frequently occurring side effect detracted from the method's practical usefulness. For systems which did exhibit sensitivity to system parameters, the performance index, which resulted from using the sensitivity constrained criterion, was found, in all the cases analysed by the author, to be considerably greater than the unconstrained index plus the deviation due to the parameter variations. Hence, from a practical point of view, the optimal sensitive system is better than the optimal sensitivity constrained system. If the independence constraint is also imposed, there will be a further increase in the constrained performance index, since the sensitivity and independence constrained system is sub-optimal compared to the purely sensitivity constrained system. This further increase in performance index, discussed more fully in Chapter 5,

completely outweighs the benefit of the reduced sensitivity and complexity of the system.

The author examined many optimal systems and found that in the range of sampling period in which they were likely to be used (cf Chapter 5) the optimal systems were comparatively insensitive to variations of the order of 10% in the systems parameters. A typical value of the absolute sensitivity functions was 0.1% of the optimal unconstrained performance index. To, in fact, obtain system which were sensitive, such as the example at the end of this chapter, the author had to go outwith the range of likely sampling periods. Before drawing any conclusions, the two other proposed sensitivity functions will be examined.

4.3.2. - MINIMAL RELATIVE SENSITIVITY

If one assumes infinitesimally small parameter variations, the definition of the relative sensitivity function becomes

$$\delta F_{rel} = F[m(a_o),\ a_o + \delta a_o] - F[m(a_o + \delta a_o),\ (a_o + \delta a_o)] \tag{4.11}$$

Expanding $F[m(a_o),\ a_o + \delta a_o]$ as a Taylor series in $F[m(a_o+\delta a_o),(a_o+\delta a_o)]$ truncated at the third term, one obtains

$$\delta F_{rel} = \tfrac{1}{2} \sum_i \sum_j \frac{\partial^2 F\,[m(a),\ a]}{\partial m(a_i)\partial m(a_j)} \Bigg|_{a=a_o+\delta a_o} \delta m(a_i)\delta m(a_j) \,. \tag{4.12}$$

where $\delta m(a) = m(a_o + \delta a_o) - m(a_o)$.

It should be noted that the second term in the Taylor expansion does not appear, since it is zero, due to definition of optimality.

Rewriting equation (4.12) in matrix form

$$\delta F_{rel} = \tfrac{1}{2} \delta m'(a) \left[\partial^2 F \left[m(a),\ a \right] / \partial^2\ m(a) \right]_{a=a_o+\delta a_o} \delta m(a) \qquad (4.13)$$

As in the case of the large perturbation relative sensitivity function, the sensitivity function of (4.13) is always positive as long as the performance index satisfies the convexity requirements of a unique minimum.

Since $\delta m(a)$ is a fixed quantity depending on the perturbed parameters, minimising $\partial^2 F \left[m(a),\ a \right] / \partial^2\ m(a)$ is equivalent to minimising the sensitivity function of equation (4.13). Consider the system described by state equation (4.3) and the intersample performance index of equation (2.79), repeated below

$$F[m(a),\ a] = \sum_{j=0}^{N} \left\{ \sum_{i=0}^{n} \left[x'(k)\ Q\ x(k) \right] + m'(jT)\ H\ m(jT) \right\}$$

Hence from the results of section (2.5)

$$\partial^2 F \left[m(a),\ a \right] / \partial^2 m(a) = \partial^2 f_{N-j}[x(j)] / \partial^2 m(a)^2$$

$$= 2\left[H^* + G'(T)\ P(N-j)\ G(T) \right]$$

where H^* is defined in equation (2.82).

The design problem of minimising the relative sensitivity function has thus been reduced to minimising the matrix $[H^* + G'(T)\ P(N-j)\ G(T)]$

It has been shown in Chapter 2 that

$$F[m(a),\ a] = f_{N-j}\ [x(j)] = x'(0)\ P(N-j)\ x(0) \qquad (4.14)$$

which has a minimum value for all x(0). The E.D.R. matrix P(N-j) can therefore be considered as a minimal positive semi-definite matrix. The matrix $G'(T)\ P(N-j)\ G(T)$ would also be a minimal semi-definite matrix, this being particularly obvious in the case when there is only one control variable and G(T) is reduced to a vector, which could be considered as a particular initial condition vector x(0) in equation (4.14). Unless the dynamic equations governing the system or the performance index are altered, the positive definite matrix H^* is invariable. (Note that matrix H^* must be positive definite, otherwise the initial value P(0) could not be determined, and optimisation technique breaks down). The complete matrix $[H^* + G'(T)\ P(N-j)\ G(T)]$ can therefore be thought of as a minimal positive definite matrix, i.e.

$$\delta m'(a)\ [H^* + G'(T)\ P(N-j)\ G(T)]\ \delta m(a)$$

has a minimal value for any vector $\delta m(a)$.

The optimisation procedure therefore automatically results in a system

which has minimal relative sensitivity. Further reduction would only be possible by altering the system dynamics or the performance criterion.

4.3.3. - MINIMAL CONTROLLER SENSITIVITY.

For small perturbations the controller sensitivity function is defined as

$$\delta F_{cont} = F[m(b_o + \delta b),\ a_o] - F[m(b_o),\ a_o] \qquad (4.15)$$

where δb is an infinitesimally small valued vector.

Expanding $F[m(b_o + \delta b),\ a_o]$ in a Taylor series in $F[m(b_o), a_o]$ truncated at third term, one obtains

$$F_{cont} = \tfrac{1}{2} \sum_i \sum_j \frac{\partial^2 F[m(b),\ a_o]}{\partial b_i \partial b_j} \Bigg|_{b=b_o} \delta b_i\, \delta b_j \qquad (4.16)$$

where b_i and b_j are elements of controller parameter variations

The resemblance between equation (4.16) and (4.12) allows one to deduce that the small perturbation controller sensitivity function is also minimised by the digital optimisation procedure.

4.3.4. - CONCLUSIONS

It has been found that

(a) the proposed method of minimising the absolute and state sensitivity functions results, in general, in

a system which, although less sensitive, is sub-optimal compared to the original unconstrained optimal system, even with its sensitivity,

and (b) the relative and controller sensitivity functions are automatically minimised as a by-product of the digital optimisation procedure.

A point of interest is that, as was stated in the introduction to the sensitivity problem, all the interrelated sensitivity functions would have to be minimised in parallel rather than serially. Due to the relative and controller sensitivity functions being minimised automatically, the absolute - state sensitivity method is such a parallel method.

4.4. - EXAMPLE

To illustrate the design procedure proposed in sub-section

(4.3.1) and, at the same time, show the prevelant drawback of the method, an example will be illustrated.

Consider a system described by state equations

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -\alpha & \beta \\ 0 & 2\gamma \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} m_1(t) \qquad (4.17)$$

which is to be optimally controlled according to the performance criterion $\int_0^\infty [x_1^2(t) + x_2^2(t) + m_1^2(t)]\, dt$

Let it be assumed that α, β and γ have nominal values of unity, but for the purposes of this example they adopt perturbed values of -0.9, 1.1 and 1.1 units respectively. Let the initial conditions be $x_1(0) = x_2(0) = 1$ unit and the sampling period be 0.2 units.

The performance indeces of the perturbed and unperturbed systems were calculated without an absolute - state sensitivity constraint. The sub-optimal performance index was calculated for the perturbed system, but in this case the optimal feedback matrix for the unperturbed system was used. Similar calculations were carried out for the system with an absolute - state sensitivity constraint, using weighting factors $w_1 = 1$ and $w_2 = 2$. In addition, that part of the augmented state performance index which corresponds to the actual performance index of interest, namely, $\int_0^\infty [x_1^2(t) + x_2^2(t) + m^2(t)]\, dt$, was also determined for the constrained systems. These values will be used in determining

(a) System without Sensitivity Constraint and Unaugmented Criterion:

	Unperturbed	Perturbed
Optimal Index	5.166955	5.687209
Sub-optimal Index		5.756726

(b) System with Sensitivity Constraint and Augmented Criterion:

	Unperturbed	Perturbed
Optimal Index	9.399358	9.909170
Sub-optimal Index		9.981889

(c) System with Sensitivity Constraint but with Index Calculated from Unaugmented Criterion:

	Unperturbed	Perturbed
Partial Optimal Index	9.317275	9.821141
Partial Sub-optimal Index		9.875390

(d) Unperturbed System without Sensitivity Constraint

Optimal Index	5.166955
Index with 10% increase in coefficient feedback matrix	5.250724

TABLE 1 - SENSITIVITY RESULTS.

the absolute sensitivity functions. For the purposes of calculating the controller sensitivity function, the value of the performance index for the unperturbed unconstrained system was also determined when the elements of the coefficient matrix were perturbed by 10% of their optimal values. All these results are tabulated in table (1) where the index values have common arbitrary units.

The absolute sensitivity functions for the system of equation (4.17), with and without sensitivity constraints, was found to be 5.99% and 11.41% of their respective optimal values of the performance index for the corresponding unperturbed system. The absolute - state sensitivity technique has therefore approximately halved the sensitivity of the performance index to parameter variations. On the other hand, the value of the performance index for the sensitivity constrained system is considerably in excess of the value of the unconstrained system. The advantages derived from using the constraint technique are therefore completely outweighed by the excessive increase in the resulting performance index.

In conclusion, the relative and controller sensitivity functions for the unperturbed, unconstrained system were found to be 1.22% and 1.62% of the perturbed and unperturbed systems' indeces respectively. The small values of these functions is in agreement with the analysis already carried out, which indicated automatic minimisation.

CHAPTER 5 - DESIGN OF DIGITAL CONTROL SYSTEMS

5.1. - DESIGN PROBLEMS.

The optimal control policy has been shown to be a linear function of all the state variables of the system, even when some of those state variables may, in fact, be fictitious augmented state variables. In general, all the state variables are not measurable, even in a deterministic system; thus the implementation of the optimal control policy is difficult, if not impossible. Apart from some reported work on the simulation of all or part of the augmented state system and thence measuring the complete state vector (42), there has been little research activity in the elimination of the state measurement problem for deterministic systems. On the other hand, the estimation of state variables and system parameters of stochastic systems has received considerable attention by several authors, using a variety of techniques (43), (44), (45). In this chapter, sub-optimal methods for both deterministic and stochastic systems are discussed and developed.

In conclusion, such aspects of the design problem as the optimal system's stability conditions and a semi-quantitative method of choosing the sampling period will be analysed.

5.2. - STATE VARIABLE MEASUREMENT IN DETERMINISTIC SYSTEMS.

In the deterministic problem, it will be assumed that only

some of the state variables can be measured, but that those can be measured accurately. The problem is then to devise the best control policy, which is linearly dependent on only those variables which can be measured.

5.2.1. - SYSTEM EQUIVALENCE OR SINGULARITY.

Inasmuch as it was eventually decided that the quadratic summation criterion was not a suitable criterion for design purposes, this sub-section is a digression. Its direct connection with the state measurement problem nevertheless merits its inclusion in this thesis. It will be proved that for a linear deterministic system, with a single control input and no cost on that control input, the resulting optimal transition matrix has a rank less than its order. At least one of the state variables is then linearly related to the other n-1 state variables, where n is the order of the system. The state measurement problem is simplified in this particular case, since it is only necessary to measure n-1 of the state variables, in order to be able to calculate the other state variable, and hence have a complete knowledge of the state vector. The n dimensional optimal system is therefore equivalent to a n-1 dimensional system.

Consider a system described by the dynamic differential equations

$$\dot{x}(t) = A\,x(t) + D\,m(t)$$

The optimal digitally controlled system can be described by matrix

difference equation

$$x(\overline{j+1}\ T) = \left[\phi(T) + G(T)\ B(N-j)\right] x(jT) \qquad (5.1.)$$

where, assuming a quadratic summation criterion, the coefficient feedback matrix is defined by

$$B(N-j) = -\left[H + G'(T)\ P(N-j)\ G(T)\right]^{-1} G'P(N-j)\phi(T)$$

Dropping all arguments superfluous to an understanding of the problem, the optimal transition matrix Φ is

$$\Phi = \phi - G\left[H + G'PG\right]^{-1} G'P\,\phi$$

$$= \left\{I - \left[H + G'PG\right]^{-1} GG'P\right\}\phi \qquad \text{since G is a vector}$$

Since P is a symmetric matrix and G is a vector

$$G'PG = \text{trace } GG'P = \sum_{i=1}^{n} g_i \sum_{j=1}^{n} g_j\, p_{ji}$$

The elements $w_{row,col}$ of matrix $W = \left[I - \left[H + G'PG\right]^{-1} GG'P\right]$ can be expressed as

(1) for row = col:

$$w_{row,col} = \left(\sum_{k} g_k \sum_{j=1}^{n} g_j p_{jk} + h\right) / \left(\sum_{i=1}^{n} g_i \sum_{j=1}^{n} g_j p_{ji} + h\right) \qquad (5.2)$$

where k has all values between 1 and n except k = row

(2) for row $\neq$ col:

$$w_{row,col} = -(g_{row} \sum_{j=1}^{n} g_j p_{jcol})/(\sum_{i=1}^{n} g_i \sum_{j=1}^{n} g_j p_{ji} + h) \qquad (5.3)$$

where H = h, a scalar, for uniformity of notation.

Consider the matrix $\bar{W}$, which is formed by multiplying all the i th. rows of matrix W by factor

$$g_{col} \sum_{j=1}^{n} g_j p_{ji} \quad * \sum_{i=1}^{n} g_i \sum_{j=1}^{n} g_j p_{ji} + h$$

Note that the last part of multiplying factor is simply the denominator of equations (5.2) and (5.3).

The off-diagonal elements of matrix $\bar{W}$ can be expressed as

$$\begin{aligned} \bar{w}_{row,col} &= -(g_{row} \sum_{j=1}^{n} g_j p_{jcol}) \times (g_{col} \sum_{j=1}^{n} g_j p_{jrow}) \\ &= -(g_{row} \sum_{j=1}^{n} g_j p_{jrow}) \times (g_{col} \sum_{j=1}^{n} g_j p_{jcol}) \end{aligned} \qquad (5.4)$$

whereas the diagonal elements of $\bar{W}$ are

$$\bar{w}_{row,col} = (\sum_{g_k} g_k \sum_{j=1}^{n} g_j p_{jk} + h) \times (g_{col} \sum_{j=1}^{n} g_j p_{jcol}) \qquad (5.5)$$

where k has all values from 1 to n except k = row.

From equations (5.4) and (5.5) it is seen that matrix $\bar{W}$ has linearly dependent rows, if h = 0, since, in this case, the sum of the off diagonal elements equals the diagonal elements of each column. The determinant of $\bar{W}$ is therefore zero and hence determinant of W is also zero, i.e. rank $[W] < n$, where n is order of system.

Now Rank $[\Phi] \leq$ Min $\{$Rank $[W]$, Rank $[\phi]\}$

Now Rank $[\Phi] \leq$ Min $\{$Rank $[W]$, Rank $[\phi]\}$

If ϕ represents the transition matrix of a continuous system, then its rank is equal to the order of system; therefore

$$\text{Rank}\,[\Phi] \leq \text{Rank}\,[W] < n$$

The optimal system's transition matrix has linearly dependent rows; therefore, at least one of the state variables at the sampling instant is linearly related to the other n-1 state variables. There may, in fact, be i linearly dependent state variables, in which case only the n-i independent state variables need be measured for a complete knowledge of the state.

An attempt was made to extend the above analysis to systems with an intersample criterion, but subsequent numerical examples indicated that, in general, the optimal transition matrix is non-singular.

As an example, consider the example of sub-section (3.4.3). It has already been stated that one of the roots of the optimal transition matrix is zero for all sampling periods, indicating the singularity of the said transition matrix. For a sampling period of 2 units, it is found that the corresponding linear relationship between the state variables is

$$x_1(jT) \doteq 13.80\, x_2(jT) + 23.56\, x_3(jT) \qquad (5.6)$$

Thus being able to measure any two of the three state variables is

equivalent to a complete knowledge of the state. It should be noted that the linear dependence is exhibited throughout the sampling interval, since only the structural properties of the various matrices were used in the above analysis.

Using equation (5.6) the equivalent 2nd order system can be derived, whose dynamic difference equation is in terms of $x_2(jT)$ and $x_3(jT)$ only, i.e.

$$\begin{bmatrix} x_2(j+1T) \\ x_3(j+1T) \end{bmatrix} = \begin{bmatrix} -8.3491 & -12.1247 \\ 4.1484 & 6.4222 \end{bmatrix} \begin{bmatrix} x_2(jT) \\ x_3(jT) \end{bmatrix} + \begin{bmatrix} 0.8278 \\ -0.0714 \end{bmatrix} m(t)$$

The corresponding optimal feedback coefficient for sufficiently large (N-j) is

$$B(N-j) = \begin{bmatrix} 1.7288 & 1.5501 \end{bmatrix}$$

The response of this system is identical to that of the 3rd order system with the restriction that the initial conditions also obey equation (5.6).

5.2.2. - THE BEST LINEAR CONTROL POLICY.

It has been shown that the optimal control input is a linear function of, in general, all the state variables of the fully augmented system. If, and only if, all elements of the i th column of the feedback coefficient matrix is zero, will the optimal control input not be a function of the i th state variable. The system designer

would ideally like to specify which state variables should be used in the generation of the control input, and which should not. The development of such a control policy, which the author has called the best linear control policy, is the subject of this sub-section. Prior to the author's own research, an extremely involved, but illuminating paper by Anderson (46) tackled a similar problem in connection with the Merriam parametric expansion continuous optimisation procedure. In that paper, the state independence constraint was introduced into the problem's solution by the inclusion of an initially unspecified Lagrangian constraint in the performance criterion, and thence, assuming certain properties to be exhibited by the constraint, a satisfactory control policy could be generated. The method was so involved, and to some extent arbitrary, due to the inclusion of further unspecified Lagrange multiplier, that the significance of the technique and the final result was obscured. Using the same source of information as the author of this thesis, a recent paper (47) by Hosking has arrived at the continuous ricatti equation analogue of the result derived in this sub-section. Once again the significance and the sub-optimality of the result does not seem to have been appreciated. The author of this thesis originally devised the best linear control policy from an intuitive point of view, but, upon becoming aware of Anderson's paper, he has based the technique on a more mathematically rigourous foundation. After some

preliminary remarks, the general technique will be presented below.

Assuming that the control input is a linear function of all the state variables, if one specifies that the i th column of the coefficient feedback matrix is zero, then, in order to maintain the optimal value of the input vector, the other elements of the coefficient feedback matrix, by the laws of linear algebra, must be replaced by non-linear functions of the state variables, including the i th state variable. Thus, by imposing the state independence constraint, one sacrifices the simplicity of linear state feedback, even for an otherwise linear system. Any linear control policy, which results from a procedure in which one imposes an independence constraint, is therefore sub-optimal. Furthermore, for any given initial conditions and the resulting trajectory, there will be a linear control policy, which is the best linear approximation to the ideal optimal non-linear control policy. Since this approximation is dependent on the system's trajectory and the designer does not know beforehand exactly which trajectories the system is likely to encounter, it is necessary to derive the best linear approximation to the optimal control policy for all state trajectories. In summary, the designer would like to derive the least upper (sub-optimal) bound, over all system trajectories, of the best linear approximation to the ideal non-linear feedback optimal control policy.

Consider the deterministic version of the system, whose optimal control policy and associated difference equations were derived in

section (2.5). Let it be assumed that an unspecified Lagrangian constraint exists, which is a functional of the state and control vectors and is zero for some trajectories of the system. Denote the constraint functional by

$$\Lambda\,[x(jT),\ m(jT)] = 0. \tag{5.7}$$

The constrained system can then be analysed by minimising the intersample performance criterion (cf equation 2.79)

$$J_N\,[x(0)] = \sum_{j=0}^{N}\left\{\sum_{i=0}^{n} x'(k)\ Q\ x(k) \ + m'(jT)\ Hm(jT)+\Lambda[x(jT),m(jT)]\right\}$$

where notation is that of equation (2.79).

By Principle of Optimality,

$$f_{N-j}\,[x(j)] = \min_{m(j)}\left\{\sum_{i=0}^{n} x'(k)Q\ x(k) \ + m'(jT)\ Hm(jT) \right.$$
$$\left. +\Lambda[x(jT),\ m(jT)] + f_{N-\overline{j+1}}[x(j+1)]\right\} \tag{5.8}$$

Let it be assumed, remembering that it is a deterministic system, that

$$f_{N-\overline{j+1}}\,[x(j+1)] = x'(j+1)\ P(N-\overline{j+1})\ x(j+1) \tag{5.9}$$

where $P(N-j+1)$ is a $\underset{\text{positive}}{\text{symmetric/semi-definite}}$, as yet, unknown matrix.

Substituting this expression for $f_{N-\overline{j+1}}\,[x(j+1)]$ into equation (5.8) and following a similar procedure to that of section (2.5) to obtain

the minimum by taking first variation of $f_{N-j}[x(j)]$ with respect to $m(jT)$, one obtains the control policy defined by

$$m(jT) = -[H^* + G'(T)\ P(N-\overline{j+1})G]^{-1} \times$$

$$\left\{ [G'(T)\ P(N-\overline{j+1})\phi + U'(N-\overline{j+1})]\, x(jT) + \partial\Lambda[x(jT), m(jT)]/\partial m(jT) \right\} \quad (5.10)$$

where H^* and $U(N-\overline{j+1})$ are defined in equation (2.82)

Since $\Lambda[x(jT),\ m(jT)]$ is as yet not completely defined, it can be assumed that

$$\partial\Lambda[x(jT), m(jT)]\ /\partial m(jT) = R(N-\overline{j+1})\ x(jT)$$

where $R(N-\overline{j+1})$ is a matrix to be specified.

The coefficient feedback matrix is therefore

$$B(N-\overline{j+1}) = -[H^* + G'(T)\ P(N-\overline{j+1})\ G(T)]^{-1} \times$$

$$[G'(T)\ P(N-\overline{j+1})\phi(T) + U'(N-\overline{j+1}) + R(N-\overline{j+1})]$$

Since $[H^* + G'(T)\ P(N-\overline{j+1})G(T)]$ is assumed non-singular for the inverse and hence the control input to be calculable, its inverse will also be non-singular. The matrix $B(N-\overline{j+1})$ can then have zero column if, and only if the matrix $[G'(T)\ P(N-\overline{j+1})\phi(T) + U'(N-\overline{j+1}) + R(N-\overline{j+1})]$ has the elements of its corresponding column zero. Since matrix $R(N-\overline{j+1})$ is so far undefined, let it have columns of elements equal and opposite to

those column of $\left[G'(T)\ P(N-\overline{j+1})\phi(T) + U'(N-\overline{j+1})\right]$ corresponding to the states not to be included in the control policy and zero elements elsewhere. The matrix $B(N-\overline{j+1})$ will then have the desired zero columns and the control input will be linearly dependent only on the specified state variables, i.e.

$$B(N-\overline{j+1}) = -\left[H^* + G'(T)P(N-\overline{j+1})G(T)\right]^{-1} \times \left[G'(T)\ P(N-\overline{j+1})\phi(T) + U'(N-\overline{j+1})\right] K \qquad (5.11)$$

where K is matrix with unity on the diagonal if corresponding state variable is specified as appearing in the control policy and zero elsewhere.

As in section (2.5), to obtain the equations governing the E.D.R. matrix P(N-j), the control policy derived above is substituted into the expanded version of equation (5.8) giving

$$f_{N-j}\left[x(j)\right] = x'(jT)\left[Q^* + \phi'(T)P(N-\overline{j+1})\phi(T) - K\alpha - \alpha K + K\alpha K\right]x(jT) \qquad (5.12)$$

where $\alpha = \left[\phi'(T)\ P(N-\overline{j+1})G(T) + U(N-\overline{j+1})\right] \times \left[H^* + G'(T)P(N-\overline{j+1})G(T)\right]^{-1} \times \left[G'(T)\ P(N-\overline{j+1})\phi(T) + U'(N-\overline{j+1})\right]$

and Q^* is defined in equation (2.82).

By linear algebra, the matrix $K\alpha + \alpha K$ is symmetric and at least positive semi-definite, therefore assumed form of $f_{N-j}\left[x(j)\right]$ in

equation (5.9) is valid. If equations (5.9) and (5.12) hold for all x(jT), then

$$P(N-j) = Q^{*} + \phi'(T)P(N-j+1)\phi(T) - K\alpha - \alpha K + K\alpha K \qquad (5.13)$$

By similar argument to that used in section (2.5), the initial conditions for difference equation (5.13) can be found to be

$$P(0) = Q^{*} - K\beta - \beta K + K\beta K$$

where $\beta = U[H^{*}]^{-1} U'$

The above optimisation procedure will now be discussed. Firstly, a control input policy has been developed, which is linear, due to the linear dynamics of the system, the quadratic nature of the criterion and, most important, due to the assumed structure for $f_{N-j}[x(j)]$. Secondly, in the derivation of the difference equation for the E.D.R. matrix, one desired equations (5.9) and (5.12) to hold for all x(j). Theret result is therefore a control policy, which minimises the performance criterion subject to a state independence and a linearity constraint for all state trajectories. This is exactly the definition of the best linear control policy; therefore, the procedure detailed above is the desired method of including a state independence constraint in the digital optimisation solution.

When K is the unit matrix, the control policy derived is the optimal control policy, being a linear function of all the state variables.

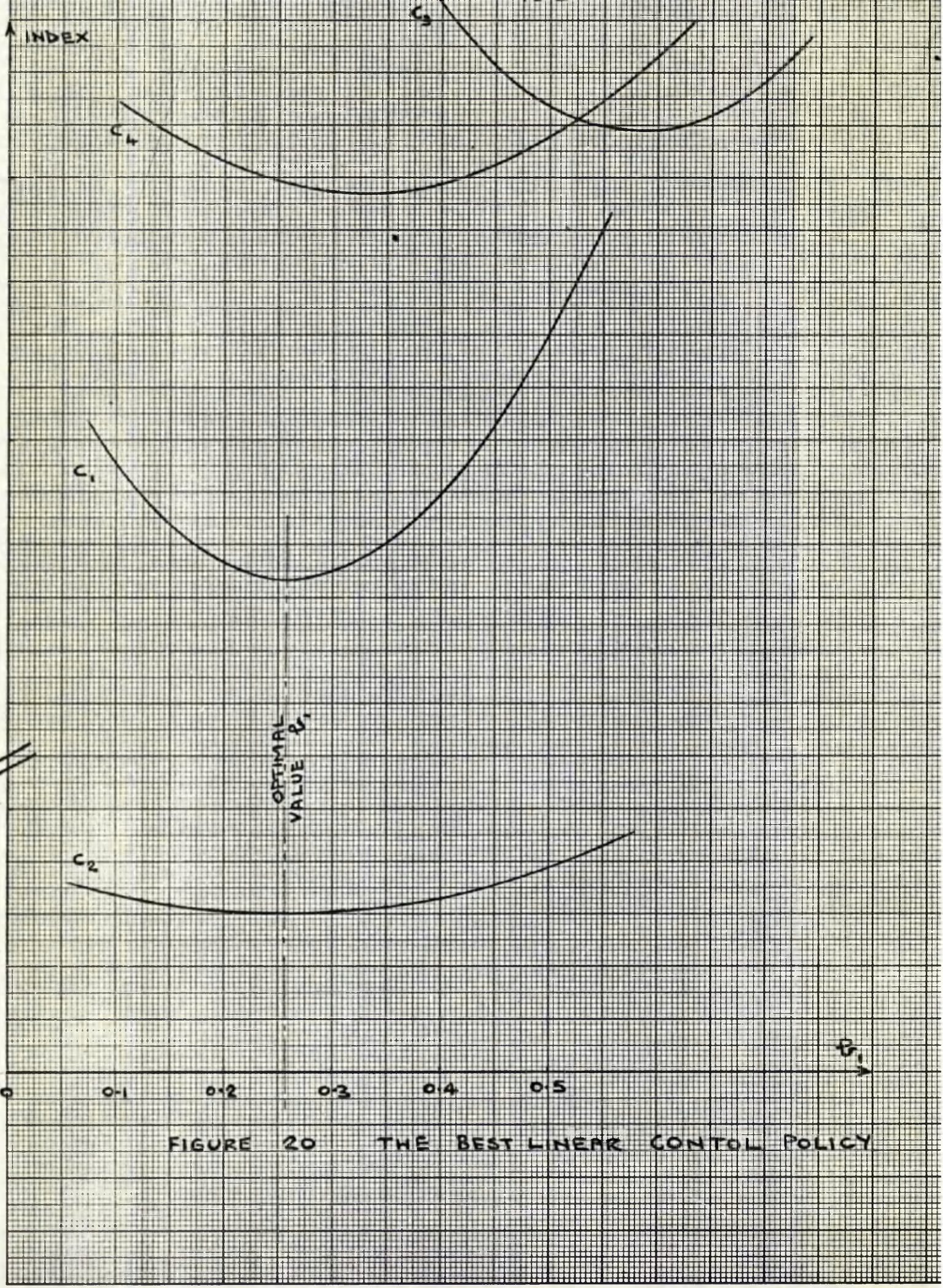


FIGURE 20 THE BEST LINEAR CONTOL POLICY

To illustrate the best linear control policy, consider the system described by the state equations

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -0.5 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} m(t)$$

which is to be minimised according to performance criterion equivalent to $\int_0^\infty [6x_1^2(t) + 6x_2^2(t) + 0.5m^2(t)]\,dt$ with sampling period of 0.5 units.

With no independence constraint, the converged value of the feedback coefficient matrix is found to be

$$B_{opt} = [-0.258 \quad -0.465].$$

The graphs of the performance index for two sets of initial conditions, namely $x_1(0) = x_2(0) = 1$ and $x_1(0) = 0$, $x_2(0) = 1$, against variations in the first element, b_1, of the feedback matrix are shown as curves c_1 and c_2 respectively in figure (20).

Imposing a constraint such that $x_1(t)$ only is specified as appearing in the control policy, results in a best linear feedback coefficient matrix for all state trajectories of

$$B_{sub\text{-}opt} = [-0.222 \quad 0].$$

The graphs of the performance index for both sets of initial conditions against variations in b_1 are also shown as curves c_3 and c_4 in figure (20).

The optimality of the unconstrained system's coefficient feedback matrix

B_{opt} and its independence of the initial conditions, i.e. its linearity, can be seen from figure (20) by the fact that the calculated value of B_{opt} corresponds to the minimum of the index curve for both sets of initial conditions. One is assuming that b_2 has its optimal value, which, in fact, it does. The essential non-linear nature of the constrained feedback system can be seen from the dependence of the best linear approximation to the optimal feedback matrix on the initial conditions. The best linear approximation for the specific initial conditions corresponds to the values of b_1 at the minimum of curves c_3 and c_4.

The performance index, which results from using the best linear feedback coefficient matrix for all state trajectories, i.e. $b_1 = -0.222$ in this case, will always be greater than or equal to the smallest value of the index obtainable by using the best linear approximation for any specified initial state. This is due to any specified set of initial conditions being a sub-set of all possible sets of initial conditions, for which the best linear feedback coefficient matrix $B_{sub-opt}$ was devised. The graphical evidence agrees with this conclusion.

Finally, since the best linear control policy is sub-optimal, the corresponding performance index is greater than its unconstrained optimal value. In the two particular cases illustrated, the increases due to imposing the constraint were 15% and 111% of their corresponding unconstrained optimal values for initial conditions $x_1(0) = x_2(0) = 1$ and $x_1(0) = 0, x_2(0) = 1$

respectively. In many cases, (cf Chapter 6), the increases can be excessively large, if not infinite, and thus the procedure devised above for including a state independence constraint must not be used with abandon.

5.2.3. - ASYMPTOTIC PROPERTIES OF STATE CONSTRAINED SYSTEM.

To establish the conditions for the existence and convergence of the optimal solution to the digital control problem without a state independence constraint, the property of partial controllability was used in sub-section (3.2.2) to ensure an upper bound on the index. This approach is not applicable to the independence constrained system, since the constrained system is only sub-optimal. It is not difficult to imagine a case where, upon stipulating that certain control variables should not appear in the optimal control policy, it is impossible to derive a control input, which does not result in a system which is unstable in some of the variables appearing in the criterion. The non-negative definiteness of the performance criterion is sufficient to prove that a best linear control policy exists, if an upper bound can be found for the linear state constrained system. The conditions for such an upper bound will now be found by considering the system described by the state difference equation

$$x(j+1) = \phi x(j) + Gm(j)$$

where $m(j) = \overline{B}\, x(j)$.

The matrix $\bar{B}$ is assumed to be an arbitrary matrix, whose columns corresponding to those state variables which have not to appear in the control policy, have zero elements throughout, i.e. matrix $\bar{B}$ satisfies the independence constraint.

Hence $x(j+n) = [\phi + G\bar{B}]^n x(j)$

In optimisation theory, one requires the state variables, which appear directly or indirectly in the performance criterion, ~~to~~ eventually attain a zero value. If vector $y(j)$ is vector of those state variables which appear in criterion, then, for suitable matrix M,

$$y(j) = M\, x(j).$$

The requirement for stability of $y(j)$ is then

$$\lim_{n \to \infty} M[\phi + G\bar{B}]^n \to 0. \qquad (5.14)$$

This condition is independent of the initial conditions of the system and therefore holds for all state trajectories. Thus, if this condition holds for the semi-arbitrary matrix $\bar{B}$, then the performance index for linear constrained system has a non-infinite upper bound, and the existence of a best linear control policy and the convergence of the iterative procedure for the E.D.R. matrix can be proved (cf sub-section (3.2.2.)).

The determination of whether equation (5.14) holds or not is not a

simple problem, since one must analyse the condition in an elemental rather than in a matrix form. Even for a system of modest order, an analytic analysis becomes tedious. One must therefore, in practice, resort to a form of hill-climbing procedure, which adjusts the allowable elements of $\bar{B}$ in order to minimise the resulting value of the performance index over some specified time interval. This procedure need only be continued until condition of equation (5.14) is obeyed. Note that the logical conclusion of such a hill-climber would be the determination of the best linear approximation to the optimal non-linear control policy for the specific initial conditions chosen, and not the determination of the best linear control policy for any initial conditions.

A sufficient condition for equation (5.14) to be obeyed is that the eigenvalues of $[\phi + G\bar{B}]$ lie within the unit circle in the Z-plane. If $\bar{B}$ is assumed to be a null matrix, then condition is obeyed if the uncontrolled system is stable. Hence the uncertainty about convergence only occurs when the uncontrolled system is unstable. The difficulty in determining an easily verifiable condition for unstable system suggests that the most direct method of determining convergence is to actually carry out the optimal iterative procedure. It was noted in sub-section (3.3.2) that computational instability was also most often in evidence with unstable systems. In agreement with the previously reported practical success of the progressive stability technique of calculating the E.D.R. Matrix sequence, it was found that if convergence was not

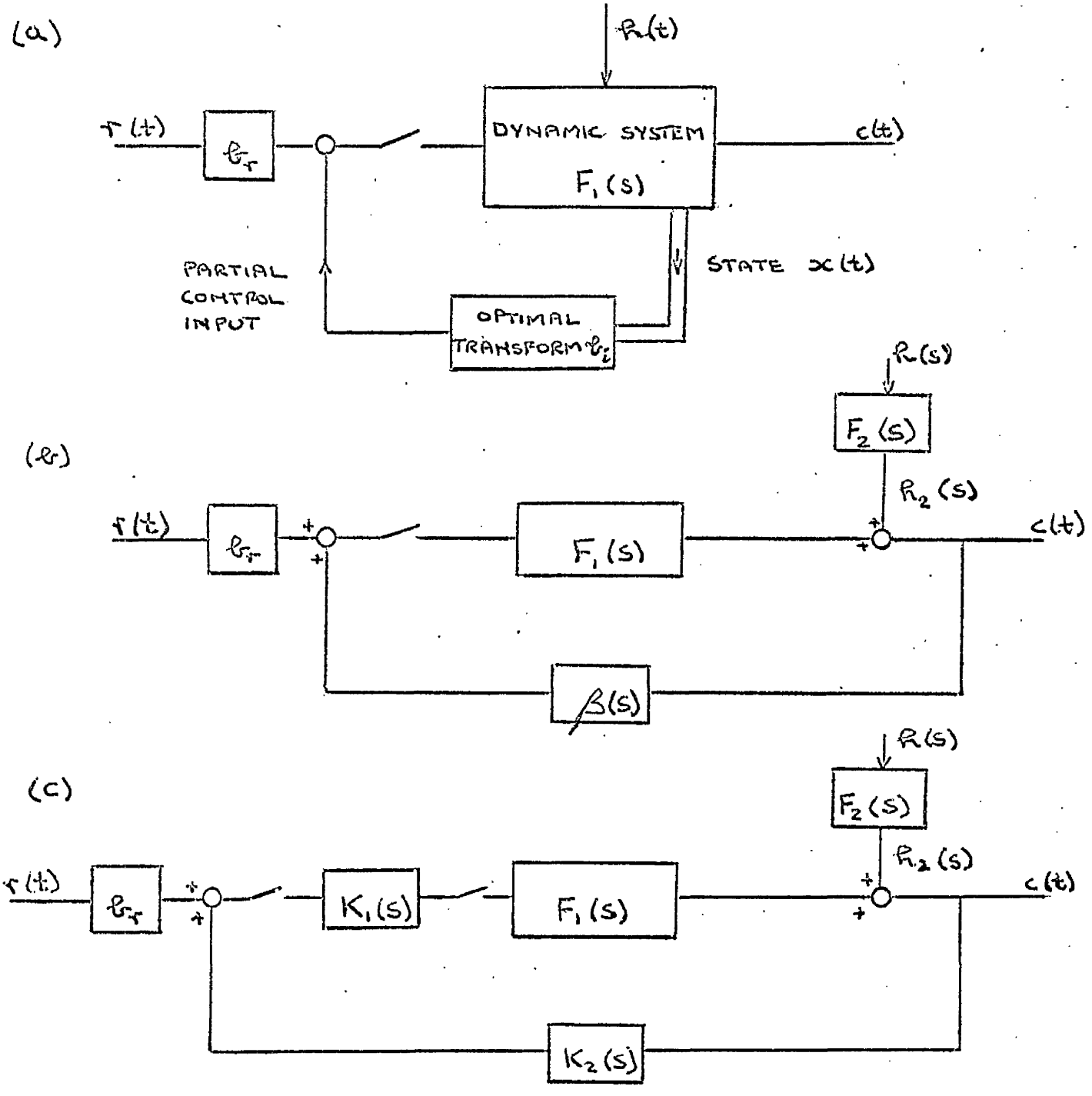


FIGURE 21 EQUIVALENT OPTIMAL SYSTEMS

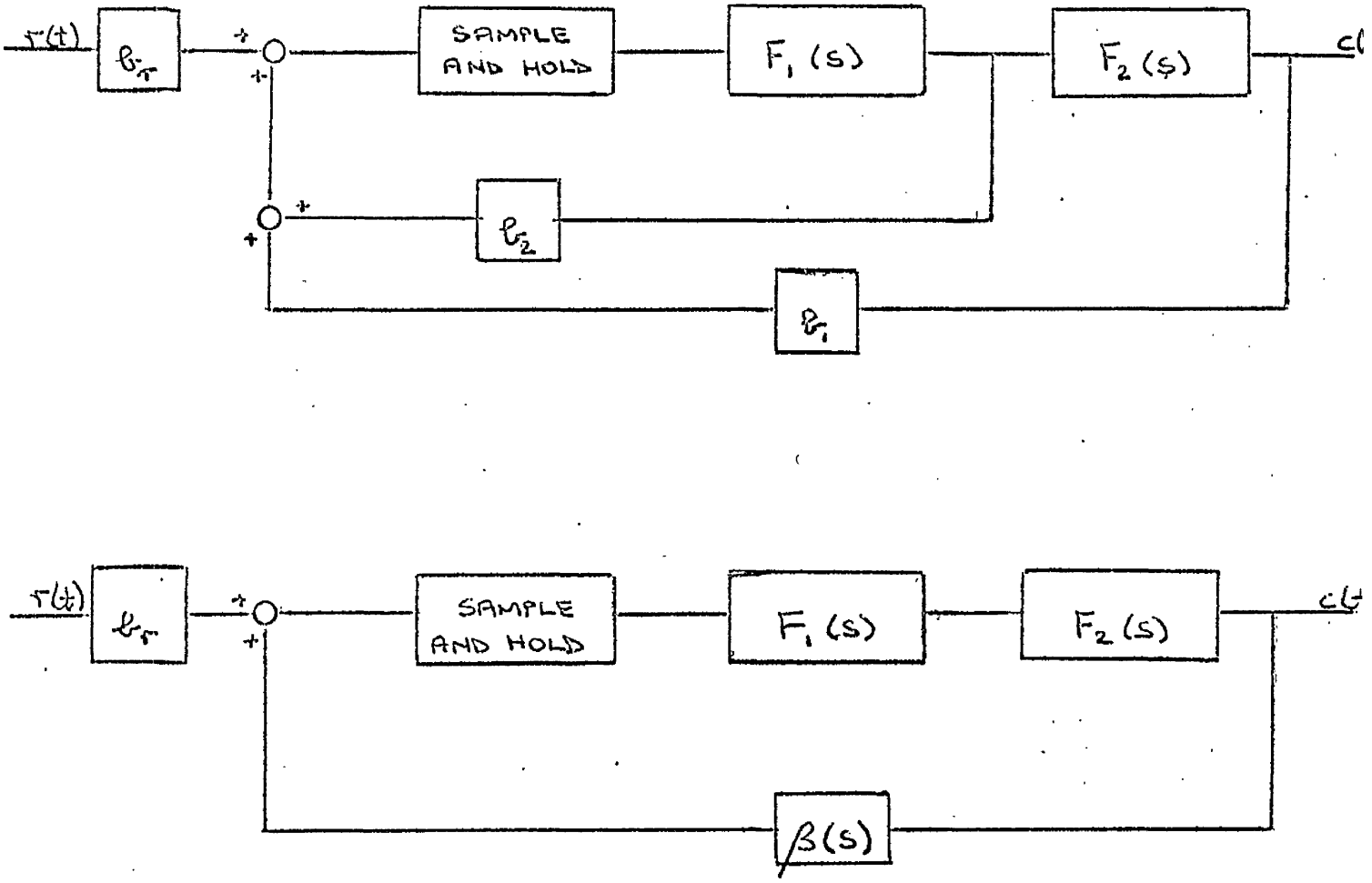


FIGURE 22 UNREALISABLE $\beta(s)$

eventually attained, it was due to condition of equation (5.14) being violated and not due to computational instability.

5.2.4. - Z-TRANSFORM IMPLEMENTATION OF THE CALCULATED CONTROL POLICY.

In classical sampled-data systems, controllers are not in general in the form of feedback from all the state variables, but rather in the form of dynamic filters in cascade with the system or in feedback paths from some of the state variables, usually the output variables. It is therefore of interest to examine the possibility of constructing a more conventional controller configuration from that of the optimal state feedback controller.

Consider an optimal single input single output reference system depicted in figure (21a) with disturbance input variable h(t). The system can be analytically rearranged into the structure of figure (21b) where $\beta(s)$ and $F_2(s)$ are found from the dynamic relationships between the variables. Direct implementation would suggest that filtering the output through a compensator $\beta(s)$ would result in an identical conventionally controlled system. The drawback is that $\beta(s)$ has, in general, more zeros than poles and, hence, is physically difficult to realise. This is not too obvious for figure (21b) but, if one considers the system depicted in figure (22), the transfer function $\beta(s)$ becomes

$$\beta(s) = b_1 + b_2/F_2(s)$$

$$= b_1U(s) + b_2V(s) \,/U(s)$$

where $F_2(s) = U(s)/V(s)$.

Since $F_2(s)$ is a physical system, order $[U(s)] <$ order $[V(s)]$, hence $\beta(s)$ will have a numerator of greater order than its denominator, and therefore will be physically difficult to implement. Equality of order can occur, but it is not common.

Consider now the discrete cascade plus continuous feedback controllers of system depicted in figure (21c), where $K_1(s)$ and $K_2(s)$ are the respective transfer functions of the controllers. The problem is to find what values of $K_1(s)$ and $K_2(s)$ will make this system and the optimal system equivalent. Little thought is required to see that making $K_1(s) = 1$ and $K_2(s) = \beta(s)$ would result in an equivalent, but physically unrealisable system, but the mathematical derivation of this result highlights certain fundamental aspects of the equivalence requirements of systems.

Using classical modified z transform theory (48) and assuming that the disturbance is zero, the output/reference input transfer function of the optimal system of figure (21b) is found to be

$$c(z,m)/b_r r(z) = F_1(z,m)/\left[1-F_1\beta(z)\right].$$

where $c(z,m)$ is the modified z transform of the output $c(t)$ etc.

The corresponding relationship for compensator system is

$$c(z,m)/b_r r(z) = K_1(z)F_1(z,m)/[1-K_1(z)F_1K_2(z)].$$

Equating these transfer functions gives relationship

$$1/[1-F_1\beta(z)] = K_1(z)/[1-K_1(z)F_1K_2(z)]. \qquad (5.15)$$

Thus by equating the optimal and compensation system's output/reference input transfer functions, an expression has been obtained in the two unknown compensators $K_1(s)$ and $K_2(s)$. Some semi-arbitrary choice of one of these compensators could then be made, and the other calculated from equation (5.15), so that both were physically realisable. Nevertheless, although the two systems have identical responses to the same inputs, assuming no disturbances, this is not a sufficient condition to dictate that the two systems are dynamically identical. One must also consider the output/disturbance (actual or fictitious) input transfer functions of the two system configurations, and also equate these before exact equivalence is assured.

In z transform theory, one is not always able to calculate the transfer function between two variables, due to the mechanics of the transform. In this case, the output response can only be calculated as a function of the disturbance input since, unless the disturbance is input through a sampler, $K_2h_2(z) \neq K_2(z)h_2(z)$ and $h_2(z) \neq \beta(z)h_2(z)$. For the optimal configuration, the output response, assuming zero reference input, is found to be

$$c(z,m) = h_2(z,m) + F_1(z,m)\beta h_2(z)/[1- F_1\beta(z)]$$

and for the compensator system

$$c(z,m) = h_2(z,m) + F_1(z,m)K_1(z)K_2h_2(z)/[1-K_1(z)K_2F_1(z)].$$

Equating these two output responses gives expression

$$K_1(z)K_2h_2(z)/[1-K_1(z)K_2F_1(z)] = \beta h_2(z)/[1-\beta F_1(z)] \qquad (5.16)$$

From expressions (5.15) and (5.16) one can obtain expression

$$K_2d_2(z) = \beta h_2(z) \qquad (5.17)$$

A sufficient condition for equation (5.17) to hold is for $K_2(s) = \beta(s)$. Substituting this value of $K_2(s)$ into equation (5.15) yields $K_1(z) = 1$. These values of $K_1(s)$ and $K_2(s)$ were the values of the compensators that were deduced by inspection.

Although by simply equating the output/reference input transfer functions of the system one would seem to be able to implement the optimal state feedback system with realisable compensators, the two systems are not truely identical. It is only when the disturbance output responses of the systems are also equated that true equivalence is obtained with the resulting unrealisability of the controllers.

Other compensator and system arrangements could be studied in the light of the above analysis, but due to the shortage of time, the author was not able to further investigate this avenue of approach. This type of analysis is tending towards methods of calculating

optimal compensator transfer functions to minimise some specified criteria, usually some form of integral error squared, which have been studied in the past with some success (49). An interesting area of research would be to attempt, with the aid of the recently developed optimisation techniques, to determine optimal transfer functions, or their equivalents, which have a prescribed position in the system. The development of the best linear control policy is the first step in this direction, but more time and effort would be required before a complete solution to the problem could be found.

5.3. - STATE VARIABLE MEASUREMENT IN NOISY SYSTEMS.

All systems are subject to random variations of various types to a greater or lesser extent. For the solution of the digital optimisation problem, the statistical characteristics were restricted to additive independent noise. The additive nature of the noise input is not too restrictive, but the independent or uncorrelated property is extremely restrictive. Correlated noise with a specified spectral density, which can be approximated to by a meromorphic function, is equivalent to passing white noise through a suitably defined filter (50). The white noise has the necessary independence property and the dynamic filter can be included in the overall dynamics of the system by the augmented state technique. The restriction of independent noise can therefore be replaced by the much less restrictive condition that the spectral density can be adequately represented by a meromorphic function.

The spectral density will not in practice be known to any high degree of accuracy, therefore the above restriction allows for a more general use of the digital optimisation technique in noisy systems.

The estimation of state variables in stochastic systems has been the subject of several papers. A particular method, to which several other proposed methods have been shown to be related, is based on the theory of orthogonal projections in multi-dimensional spaces. It has commonly been given the name of the Kalman Filter. This method results in the optimal estimate of the state vector, where the estimate is optimal, inasmuch as it minimises the covariance matrix of the error between the predicted value and the true value of the state for a noise input with independent gaussian characteristics. If the input is non-gaussian, then the estimate is only the best linear estimate, which minimises the covariance error matrix. It has been shown by Tou (19) that the optimal control and optimal estimation procedures are entirely separable and that if one uses the optimal estimate of the state in the generation of the optimal control policy, then the resulting combined system can be considered as optimal.

In this section, it is intended to develop the Kalman filter technique to cope with systems in which the measurement of the system's output variables are subject to pure time delays, a common feature of process control systems. Thence, a sub-optimal estimator suitable for optimal control purposes will be developed by considering the estimation

difference equations in the light of the generalised inverse matrix theory.

5.3.1. - THE KALMAN FILTER.

In order to present the development of the estimator with multiple delays in the measurement of the output variables, it is first necessary to rearrange the results obtained in references (19) (43) in a suitable form. The general details of the method will not be given, since these are well recorded in the before mentioned papers.

Consider the system described by the matrix state difference equation

$$x(j+1) = \phi(T)\, x(j) + G(T)m(j) + r(j) \qquad (5.19)$$

where $r(j)$ is an additive independent discrete noise input with zero mean.

Let the output vector $y(j)$ be a linear combination of the state variables plus an additive independent noise disturbance with zero mean, i.e.

$$y(j) = M\, x(j) + w(j)$$

Define $Y(j)$ to be the space made up of all linear combinations of the measurable output vector $y(k)$, $k \leq j$. The space $Y(j)$ then represents the space of knowledge upon which the optimal estimations will be based. Similarly, $X(j)$ is defined to be the space made up of all linear combinations

of the vectors $x(k)$, $k \leq j$. It is important to note that the dimension of the space $X(j)$ is greater than that of space $Y(j)$ if the number of independent output variables is less than the order of the system, which is always the case.

The orthogonal projection of matrix $x(j)$ onto space $Y(k)$ is denoted by $\hat{x}(j/k)$ and the corresponding normal of $x(j)$ on $Y(k)$ is denoted by $\tilde{x}(j/k)$, therefore one has the fundamental relationship

$$x(j) = \hat{x}(j/k) + \tilde{x}(j/k) \qquad \text{for } j \geq k. \tag{5.20}$$

The objective of the analysis is to derive an expression for $\hat{x}(j/j)$, the orthogonal projection of $x(j)$ on $Y(j)$, since it has been proved to be the optimal estimate of the state vector $x(j)$ in the sense of minimising the variance of the error matrix (43).

By using standard vector space methods (51), it can be shown (19) that a linear transition matrix exists between $\hat{x}(j+1/j)$ and $\hat{x}(j/j)$, namely,

$$\begin{aligned} \hat{x}(j+1/j) &= \left[\phi(T) + G(T)B(N-j)\right] \hat{x}(j/j) \\ &\triangleq \Phi(j)\, \hat{x}(j/j). \end{aligned} \tag{5.21}$$

This equation is the predictor equation, since the orthogonal projection of $x(j+1)$ is predicted from the knowledge contained in space $Y(j)$.

Since the value of $y(j+1)$ is known at time $(j+1)T$, it is desirable

that this knowledge be used in the evaluation of $\hat{x}(j+1/j+1)$. In order to do so, a further vector space $Z(j+1)$ is considered to be made up of vectors $V(j+1)\,\tilde{y}(j+1/j)$ where $V(j+1)$ is an, as yet, undefined matrix. Using the property that space $Z(j+1)$ is orthogonal to all vectors in space $Y(j)$, and hence that space $Y(j+1)$, by definition, is the sum of the two spaces $Y(j)$ and $Z(j+1)$, the optimal estimator equation is found to be (19).

$$\hat{x}(j+1/j+1) = \Phi(j)\,\hat{x}(j/j) - V(j+1)\left[M\Phi(j)\hat{x}(j/j) - y(j+1)\right] \qquad (5.22)$$

The optimal estimate of $x(j+1)$ knowing the output $y(j+1)$ is therefore seen to be a linear combination of the predicted value of $x(j+1)$ based on $Y(j)$ minus a correction factor due to the difference between the predicted value of the output vector and the actual output at time $(j+1)T$.

The pair of matrix difference equations governing the matrix $V(j+1)$ is determined by considering the covariance matrix, $J(j+1)$ of the error between the predicted value of the state $x(j+1)$ based on $Y(j)$ and its true value, i.e.

$$J(j+1) = E\left\{\tilde{x}(j+1/j)\,\tilde{x}'(j+1/j)\right\}$$

which upon using equations (5.19), (5.21) and (5.22) becomes

$$J(j+1) = \Phi(j)\left[I - V(j)M\right]J(j)\left[I - V(j)M\right]'\Phi'(j)$$

$$+ R(j) + \Phi(j)V(j)\,W(j)V'(j)\Phi'(j) \qquad (5.23)$$

where $R(j)$ and $W(j)$ are the covariance matrices of the disturbance input $r(j)$ and measurement input $w(j)$ respectively.

Using the orthogonal properties of the spaces $Y(j)$ and $Z(j+1)$, an expression for $V(j+1)$ can be derived (19), namely,

$$V(j+1) = J(j+1)\, M'\left[MJ(j+1)M' + W(j+1)\right]^{-1} \tag{5.24}$$

assuming that inverse exists.

Assuming that an initial error covariance matrix $J(0)$ and an initial best estimate of state $\hat{x}(-1/-1)$ are known, then one is in a position to calculate $V(j+1)$ and $\hat{x}(j+1/j+1)$ for $j \geq 0$, the latter being the optimal estimate of the state vector, the objective of the above analysis.

5.3.2. - ASYMPTOTIC PROPERTIES OF THE OPTIMAL ESTIMATOR.

In order that the estimation procedure derived above is to be of use, it must be proved that the covariance matrix of the error converges to a limiting value with an increasing number of samples of the output variables. It must also be decided exactly what the initial condition $J(0)$ of the sequence of covariance matrices is, since the ultimate value of the sequence is dependent on its initial conditions. Before discussing these points, the conditions governing a property of the controlled and uncontrolled system, commonly known as the observability property, will be derived.

A deterministic system is said to be observable if all the modes of the system are evident in the measurable output vector of the system. In

simpler terms a deterministic system is observable if it is possible to calculate the initial state of the system given the output vectors and control input vectors over some finite time interval.

To determine the conditions for observability, consider the system described by the state difference and output equations,

$$x(j+1) = \phi x(j) + Gm(j)$$

$$y(j) = Mx(j)$$

Hence, by successive substitution, assuming an initial condition of x(0),

$$y(n) = M\phi^{n}x(0) + M\sum_{i=0}^{n-1}\phi^{n-1-i}\, G\, m(i)$$

If $z(n) = y(n) - M\sum_{i=0}^{n-1}\phi^{n-1-i}\, Gm(i)$ then

$$z(n) = M\phi^{n}\, x(0)$$

Consider the augmented state matrix equation

$$\left[z(0) \mid z(1) \mid \ldots . z(n)\right]' = \bar{C}(n-1)\, x(0)$$

where $\bar{C}(n-1) = \left[M \mid M\phi \mid M\phi^{2} \ldots . M\phi^{n-1}\right]'$

In order for system to be observable, according to the definition, the above matrix equation must be solvable for x(0). For any solution to exist, the matrix algebraic equation must be compatible, a sufficient condition being

$$\text{Rank}\left[\bar{C}(n-1)\right] = p$$

where p is order of system, i.e. dimension of ϕ.

According to Caley-Hamilton theorem ϕ^n is linearly dependent on ϕ^{n-1} for $n \geq p$, therefore rank condition is reduced to

$$\text{Rank}\left[\bar{C}(p-1)\right] = p \tag{5.25}$$

Note that this sufficient condition for observability is independent of the method of generating the control vector and is therefore of use when the control vector is unknown. On the other hand, consider the case where the control policy is given by

$$m(j) = B(N-j)\, x(j)$$

In the remainder of the analysis it will be assumed, for simplicity that B(N-j) has converged to its ultimate value.(Similar results also hold for the time varying case).

Then $$y(n) = M\left[\phi + GB\right]^n x(0)$$

$$\triangleq M \Phi^n x(0)$$

The corresponding sufficient condition for observability is found to be

$$\text{Rank}\left[M \mid M\Phi \mid M\Phi^2 \;\ldots\; M\Phi^{p-1}\right] = p \tag{5.26}$$

The conditions for controllability will now be used in the analysis of

the convergence properties of the estimator.

By eliminating V(j+1) from equations (5.23) and (5.24) one obtains the expression

$$J(j+1) = R(j) + \Phi(j)\, J(j)\, \Phi'(j) - \Phi(j)\, J(j)\, M' \left[W(j) + MJ(j)M'\right]^{-1} MJ(j)\, \Phi'(j) \qquad (5.27)$$

Compare this equation with equation (2.57), repeated below, which is the E.D.R. matrix difference equation for a quadratic summation performance criterion.

$$P(N-j+1) = Q + \phi' P(N-j) \phi - \phi' P(N-j) G \left[H + G' P(N-j) G\right]^{-1} G' P(N-j) \phi . \qquad (5.28)$$

Equations (5.27) and (5.28) are duals of one another. This duality property was first noted by Kalman in reference (31), but due to complex notation, the essential simplicity of the argument was lost. The above presentation clearly indicates the following dual matrix pairs $\{\phi, \Phi'(j)\}$, $\{G, M'\}$, $\{Q, R(j)\}$, $\{H, W(j)\}$ and $\{P(N-j),\ J(j+1)\}$.

The difference equation (5.28) in augmented form was shown in section (2.3) to be convergent, subject to the following conditions

(a) Matrices Q and H are positive semi-definite.

(b) The augmented state system satisfies a partial controllability condition, the performance criterion and the augmented system's dynamics

are compatible and the controlled augmented system is stable.

Thus for optimal estimator to be proved convergent, the following dual conditions must be satisfied

(a) Matrices R(j) and W(j) are positive semi definite.

(b) There is no corresponding condition to the criterion compatability condition, therefore the dual of the conditions for complete controllability must be satisfied. Complete controllability is seldom satisfied in augmented state systems, but its dual fortunately is usually satisfied.

The sufficient complete controllability condition is

$$\text{Rank}\left[G \,\vdots\, \phi G \,\vdots\, \phi^2 G \;\; \ldots\ldots \phi^{p-1} G\right] = p$$

which has dual

$$\text{Rank}\left[M \,\vdots\, M\Phi \,\vdots\, M\Phi^2 \;\; \ldots\ldots \; M\Phi^{p-1}\right]' = p \qquad (5.29)$$

Equation (5.29) is identical with the sufficient condition for observability (cf equation (5.26)); therefore, controllability and

observability are duals of one another. Since one does not know at the outset of the optimisation calculation what the control input will be, the sufficient observability condition of equation (5.25) is a more convenient method of determining convergence, i.e.

$$\text{Rank}\left[M \;\vdots\; M\phi \;\vdots\; M\phi^2 \;\ldots\; M\phi^{p-1}\right] = p.$$

The duality property has therefore greatly facilitated the determination of the sufficient conditions for convergence. Three further aspects of the estimator will now be discussed.

Firstly, the initial value of the covariance error matrix, $J(0)$, was not specifically defined but, by analogy with the E.D.R. matrix iteration, the initial value should be equal to the covariance matrix of the discrete disturbance noise input, $R(0)$. With this initial value, one is effectively stipulating that prior to the first state vector estimation, there is no estimation problem, i.e. ideal measurement of the state vector in the presence of noise disturbances. Secondly, since matrix Φ is the transition matrix of the controlled system, it will have all its eigenvalues on or within the unit circle in the z-plane. According to the progressive stability theory, the probability of computational instability is therefore greatly reduced. Thirdly, the estimator could be subject to periodic instabilities in the covariance error matrix of a type similar to those found in the E.D.R. matrix (cf section 3.4). Excessive instabilities are usually only found in unstable systems with complex roots. Since the controlled system will always in practice be stable, the probability

of excessive periodic instabilities is greatly reduced.

5.3.3. - OPTIMAL ESTIMATION WITH PURE TIME DELAYS IN THE OUTPUT MEASUREMENT.

For the purposes of this sub-section, consider an optimally controlled system, described by the state and output equations

$$x(j+1) = \phi(T)\, x(jT) + G(T)\, B(N-j)\; x(jT) \tag{5.30}$$

and $$\bar{y}(jT) = \sum_{i=o}^{n_1} M_i\, x(jT - \delta_i) \tag{5.31}$$

where M_i are the output matrices

and δ_i are pure time delays.

The output vector $y(jT)$ is a linear combination of past and present values of the state vector. If one considers M_i to be null matrices for all but $i = 1$, then

$$\bar{y}(jT) = M_1\, x(jT - \delta_1) + w(jT)$$

One no longer has a knowledge of the output at the time one wishes to make an estimate of the state vector, and therefore the previously pure estimation problem is now a combined estimation and prediction problem. As will be shown later, this particular case can be separated into easily recognisable estimation and prediction parts, but in general this is not obvious.

The author at first tried to solve the problem by using a

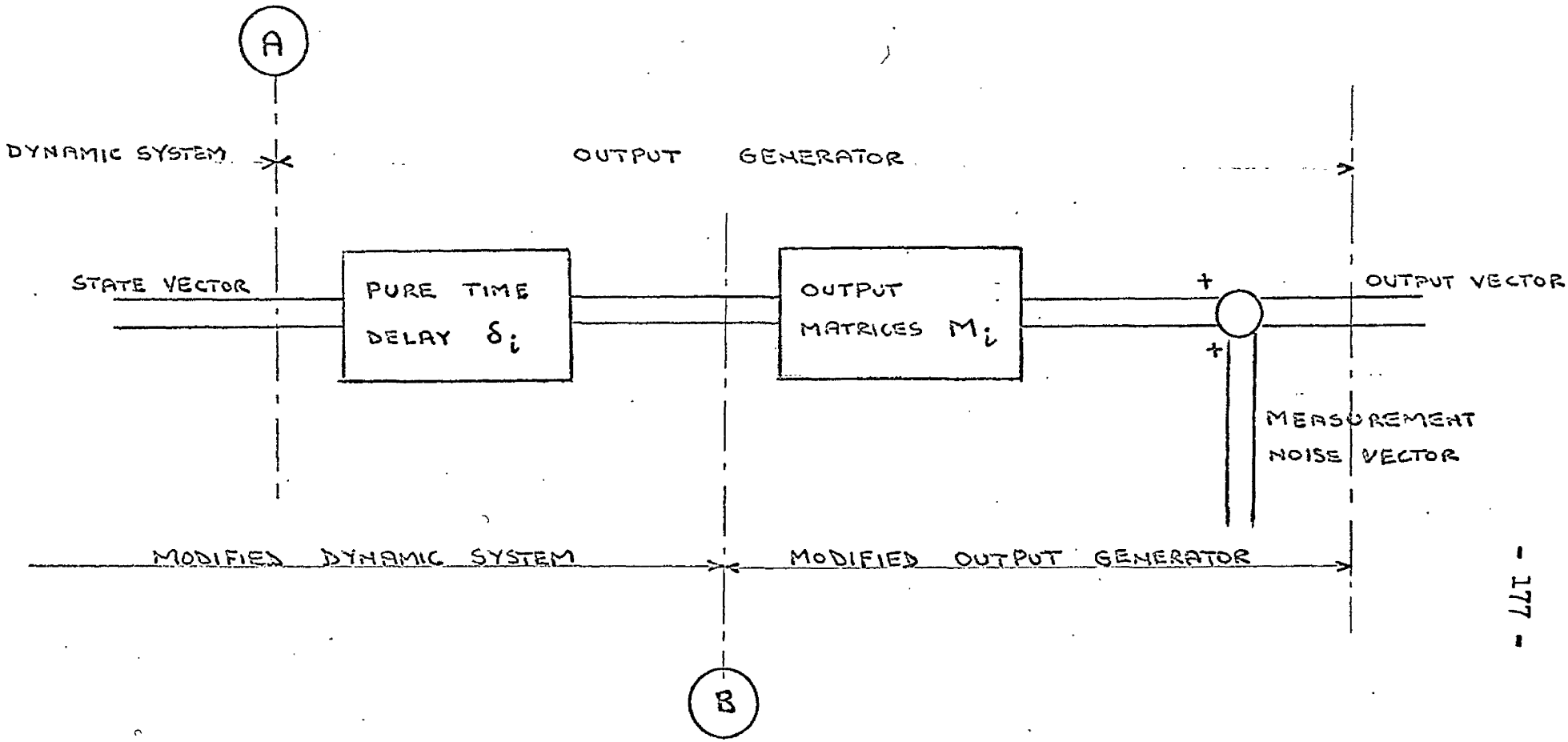


FIGURE 23 DELAYED MEASUREMENT SCHEME

more complicated orthogonal projection theory, but this approach did not meet with much success. A more elegant and conceptionally simple method was then developed, based on a reappraisal of the problem. Consider the graphical representation of the matrix output equation (5.31), illustrated in figure (23). The state vector of the system is that at position(A), but if one considers the delay elements δ_i as being included in the dynamics of the system, rather than in the output vector generator, the state vector would be that at position (B). The state vector would then be

$$\bar{x}(jT) = \left[x(jT) \,\vdots\, x(jT - \delta_1) \,\vdots\, \ldots\ldots \,\vdots\, x(jT - \delta_{n_1})\right]'$$

The state transition equation of the controlled system would then be, assuming for simplicity that $B(N-j) \rightarrow$ constant value B

$$\bar{\Phi} = \begin{bmatrix} \alpha(T) & \ldots & 0 \\ \beta(\delta_1) & & 0 \\ & & \\ \beta(\delta_{n_1}) & & 0 \end{bmatrix}$$

where $\alpha(T) = \phi(T) + G(T)B$

and $\beta(\delta_i) = \alpha(T-\sigma)\alpha(T)^{-k}$

with k = integer part of δ_i/T and $\sigma = \delta_i - kT$.

and the output matrix equation would be

$$\bar{y}(jT) = \left[M_o \,\vdots\, M_1 \,\vdots\, \ldots\ldots M_{n_1}\right] \bar{x}(jT) + w(jT)$$

Thus, by changing the formulation of the multiple measurement delay

problem, it has been reduced to a problem similar to that already solved in the derivation of the Kalman filter. The optimal estimate of the state vector can therefore be calculated by the substitution of the augmented state matrices into the difference equations (5.22), (5.23) and (5.24).

It is illustrative to consider the case where M_i is only non-null for i = 1 and the delay in measurement is equal to one sampling period. In this case, the controlled systems transition matrix can be considered as

$$\bar{\bar{\Phi}}(j) = \begin{bmatrix} \Phi(j) & 0 \\ 0 & \Phi(j-1) \end{bmatrix}$$

and the augmented output matrix is

$$M = [0 \,|\, M_1]$$

Assuming the augmented matrix $\bar{V}(j+1)$ of equation (5.24) to be suitably partitioned, i.e.

$$\bar{V}(j+1) = \begin{bmatrix} V_1(j+1) \\ V_2(j+1) \end{bmatrix}$$

and hence, substituting into equation (5.22), one obtains the optimal estimate of $\bar{x}(j+1)$, knowing $\bar{y}(j+1)$ as

$$\begin{bmatrix} \hat{x}(j+1/j+1) \\ \hat{x}(j/j+1) \end{bmatrix} = \begin{bmatrix} \Phi(j)\hat{x}(j/j)-V_1(j+1)\left[M_1\Phi(j-1)\hat{x}(j-1/j)-\bar{y}(j+1)\right] \\ \Phi(j-1)\hat{x}(j-1/j)-V_2(j+1)\left[M_1\Phi(j-1)\hat{x}(j-1/j)-\bar{y}(j+1)\right] \end{bmatrix} \quad (5.33)$$

The terminology is such that $\hat{x}(j/j+1)$ is the orthogonal projection of $x(j)$ on space $\bar{Y}(j+1)$, where the latter is space made up of all linear combinations of $\bar{y}(k)$, $k \leq j+1$. If one considers $y(j+1)$ to be the ideal measurement with no delay as used in derivation of Kalman filter, then $\bar{y}(j+1)$ is equivalent to $y(j)$ and $\hat{x}(j/j+1)$ and $\hat{x}(j+1/j+1)$ are equivalent to $\hat{x}(j/j)$ and $\hat{x}(j+1/j)$ respectively. The two algebraic equations of equation (5.33) become

$$\hat{x}(j+1/j) = \Phi(j)\,\hat{x}(j/j-1)-V_1(j+1)\left[M_1\Phi(j-1)\hat{x}(j-1/j-1)-y(j)\right] \quad (5.34)$$

and

$$\hat{x}(j/j) = \Phi(j-1)\hat{x}(j-1/j-1) - V_2(j+1)\left[M_1\Phi(j-1)\hat{x}(j-1/j-1)-y(j)\right] \quad (5.35)$$

Note that equations are interrelated through vector $\hat{x}(j-1/j-1)$. The equation (5.34) is in the form of a predictor, predicting $\hat{x}(j+1/j)$ from $\hat{x}(j/j-1)$ (c.f. equation (5.21)). In this case there is an additional correction term, due to the difference between the predicted value and the actual value of the measurable output $y(j)$. On the other hand, equation (5.35) is in the form of an estimator, estimating $\hat{x}(j/j)$ from $\hat{x}(j-1/j-1)$ (c.f. equation (5.22)). Again, there is a corresponding correction term.

The above particular case of the multiple measurement delay problem

has therefore been shown to be a combination of a predictor and an estimator, as was envisaged when the problem of multiple measurement delays was formulated at the beginning of this sub-section.

5.3.4. - A SUB-OPTIMAL ESTIMATION ALGORITHM.

The matrix coefficients of the difference equation governing the optimal estimate of the state vector in both the no-delay and multiple delay systems are time dependent, due to matrices $V(j+1)$ and $\Phi(j)$. If the control optimisation is to be carried out over a sufficiently long time interval, then $\Phi(j)$ will be time independent, a property that will be assumed to hold in the remainder of this sub-section. The problem of $V(j+1)$ remains. It would be of practical interest to develop a time invariant sub-optimal estimator and thus considerably reduce the complexity of the system. This will be carried out by investigating the generalised inverse solution of the difference equation in $V(j+1)$, namely

$$V(j+1) = J(j+1)M'\left[MJ(j+1)M' + W(j+1)\right]^{-1} \qquad (5.36)$$

In the above system $W(j+1)$ is the covariance matrix of the independent zero mean measurement noise. If, as is more likely, the measurement noise is correlated, a white noise representation must be used. In this case, the white noise is considered as a disturbance input associated with a filter, which augments the dynamic state equations.

The correlated measurement noise input is then one of the state variables of the system, and the output equation no longer contains an additive measurement noise vector w(j). In this more general case, or in the case where there is no measurement noise, the covariance matrix W(j+1) is zero. Hence, premultiplying equation (5.36) by M yields

$$MV(j+1) = I \tag{5.37}$$

The solution of the vector equivalent of equation (5.37) has been extensively studied by Penrose (52) for the case where M is a singular or rectangular matrix. The solution of equation (5.37) can be proved, by direct substitution, to be

$$V(j+1) = M^{+} + [I-M^{+}M]C.$$

where M^{+} is the generalised inverse of M, hence $MM^{+}M = M$ and C is an arbitrary matrix.

Since M is a constant matrix, so is its generalised inverse. The arbitrary matrix C would ideally be time varying, so that V(j+1) could adopt its optimal time dependent values. As a first approximation, nevertheless V(j+1) could be equated to M^{+} and the resulting sub-optimal estimation equation corresponding to equation (5.22) would then become

$$\hat{x}_s(j+1/j+1) = \Phi \hat{x}_s(j/j) - M^{+}[M\,\hat{x}_s(j/j) - y(j+1)]$$

where subscript s refers to sub-optimality.

In this time invariant coefficient equation, the correction term is

premultiplied by M^+ rather than $V(j+1)$. Equation (5.38) can be rearranged to give

$$\hat{x}_s(j+1/j+1) = M^+y(j+1) + [I-M^+M]\,\Phi\,\hat{x}_s(j/j) \qquad (5.38)$$

To add some justification for making the above approximation for $V(j+1)$, consider the optimal estimation equation (5.22) premultiplied by M, which gives

$$M\hat{x}(j+1/j+1) = y(j+1) \quad \text{since } M\,V(j+1) = I.$$

The solution of this equation can be proved by substitution to be

$$\hat{x}(j+1/j+1) = M^+y(j+1) + [I-M^+M]\,c \qquad (5.39)$$

where c is an arbitrary vector.

The structure of equation (5.38) and (5.39) are similar with arbitrary vector c in optimal equation replacing vector $\Phi\hat{x}_s(j/j)$ in sub-optimal equation. The matrix coefficients M^+ and $[I-M^+M]$ being identical in both cases adds some justification for making the simplifying assumption that $V(j+1)$ be replaced by M^+. This approach to reducing the complexity of the composite control-estimation system is applicable to both no-delay and multiple measurement delay systems.

So far the generalised inverse of M has been used without any indication of how it is calculated. As a result of the proposed relationship between $V(j+1)$ and M^+, a method of determining M^+ was devised

by considering equation (5.24), in which it is assumed that W(j+1) is a null matrix, i.e.

$$V(j+1) = J(j+1)M'\left[MJ(j+1)M'\right]^{-1}$$

In practice, the output matrix M has rank equal to its smaller dimension, otherwise there would be dependent outputs. To ensure the existence of the inverse, the matrix J(j+1) must be positive definite, so let J(j+1) be the unit matrix, then,

$$V^{*} = M'\left[MM'\right]^{-1}$$

The matrix couple V^{*} and M are such that

(a) $MV^{*}M = M$

(b) $V^{*}MV^{*} = V^{*}$

(c) $(MV^{*})' = MV^{*}$

(d) $(V^{*}M)' = V^{*}M$

If all these conditions hold, then V^{*} is by definition the generalised inverse of M (52). The generalised inverse of M can therefore be calculated from equation

$$M^{+} = M'\left[MM'\right]^{-1}.$$

In order that the sub-optimal estimator be of practical use, it must result in a convergent sequence of covariance error matrices. Making approximation that $V(j) = M^{+}$ for all j, results in an iterative

equation for J(j+1) corresponding to equation (5.23). Assuming a non-zero measurement covariance matrix W(j), although a null matrix was assumed in the justification for the approximation, one has

$$J_s(j+1) = \Phi\left[I-M^+M\right] J_s(j) \left[I-M^+M\right]' \Phi' + R(j) + \Phi M^+ W(j) M^{+\prime} \Phi'$$

where subscript s refers to sub-optimality.

By successive substitution and assuming that the statistics are stationary.

$$J_x(j+1) = \alpha^{j+1} J_s(0) \alpha'^{j+1} + \sum_{i=0}^{j} \alpha^i \left[R+\Phi M^+ W M^{+\prime} \Phi'\right] \alpha'^i$$

where $\alpha = \Phi\left[I-M^+M\right]$

If all the eigenvalues of matrix α lie within the unit circle in the z-plane, then $J_s(j+1)$ will tend to a limit defined by

$$\lim_{j \to \infty} J_s(j+1) = \lim_{j \to \infty} \sum_{i=0}^{j} \alpha^i \left[R + \Phi M^+ W M^{+\prime} \Phi'\right] \alpha'^i.$$

5.3.5. - EXAMPLES OF ESTIMATORS

To illustrate the results achieved by using the before-mentioned estimators, two examples will be detailed, in which

(a) the optimal and sub-optimal estimators are compared,

and (b) the optimal and delayed optimal estimators are compared.

(a) The system considered is

$$\begin{bmatrix}\dot{x}_1(t)\\ \dot{x}_2(t)\end{bmatrix} = \begin{bmatrix}-0.5 & 1\\ 0 & -1\end{bmatrix}\begin{bmatrix}x_1(t)\\ x_2(t)\end{bmatrix} + \begin{bmatrix}0\\ 1\end{bmatrix} m_1(t) + \begin{bmatrix}1\\ 0\end{bmatrix} u(t)$$

with intersample criterion equivalent to $\int_0^\infty [x_1^2(t) + x_2^2(t)]\, dt$.

The output equation is assumed to be

$$y_1(jT) = \begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}x_1(jT) & x_2(jT)\end{bmatrix}'$$

The sampling period is 0.5 secs. The disturbance noise input is an independent non-white sequence of delta functions with a fixed interval of 0.1 secs. between each impulse. The covariance of the continuous noise input u(t) is 0.133 units^2. The resulting discrete covariance matrix of disturbance noise input is calculated to be

$$R = \begin{bmatrix}5.24624_{10}\text{-}2 & 0\\ 0 & 0\end{bmatrix}$$

To examine the performance of the estimators devoid of control considerations one need only examine the error covariance matrix. The above example was chosen because it allows an illustrative comparison between the optimal and sub-optimal estimators since it represents the estimator equivalent of a dead-beat system, i.e. the error covariance matrix converges after one iteration to have a value equal to its initial value. The error covariance matrix is thus

$$J_{optimal} = \begin{bmatrix} 5.24624_{10}-2 & 0 \\ 0 & 0 \end{bmatrix} = R$$

With $V(j) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ for $j \geq 0$.

On the other hand, the sub-optimal estimator is not of the dead-beat type and takes seven iterations to converge to value

$$J_{sub\text{-}optimal} = \begin{bmatrix} 5.67433_{10} - 2 & -1.80912_{10}-3 \\ -1.80912_{10}-3 & 7.63840_{10}-4 \end{bmatrix}$$

with the constant generalised inverse of $\begin{bmatrix} 1 & 1 \end{bmatrix}$ being

$$V^* = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$

The reduction in optimality by using the generalised inverse approximation is seen to be two fold. Firstly, the error covariance associated with $x_1(t)$ is increased in the sub-optimal case. Secondly, whereas in the optimal estimator, the error auto and cross covariance associated with $x_1(t)$ was zero, the sub-optimal estimator no longer maintains these zero values. The overall increase in the covariance is nevertheless small for this particular system, which would indicate the acceptability of the considerably less complex sub-optimal estimator.

(b) For the second example, the system considered is

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -0.5 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} m_1(t) + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u(t)$$

The performance criterion, sampling period and noise input are assumed to be the same as those in example (a). In this case, the discrete covariance matrix of the disturbance noise input is calculated to be

$$R = \begin{bmatrix} 5.56722_{10}-2 & 9.51874_{10}-3 \\ 9.51874_{10}-3 & 4.21413_{10}-2 \end{bmatrix}$$

The optimal estimators for a no delay and a single period delay in the measurement output are to be compared to illustrate the deterioration due to a predictor having to be added to the estimator in the latter case. The output equations are

(i) No delay: $y_1(jT) = x_1(jT) + x_2(jT)$

(ii) Single delay: $y_1(jT) = x_1(jT) + x_2(\overline{j-1}\ T)$

For the no-delay measurement, the covariance error matrix was found to converge after seven iterations, to give

$$J_{\text{no-delay}} = \begin{bmatrix} 6.18680_{10}-2 & 6.90278_{10}-3 \\ 6.90278_{10}-3 & 4.32458_{10}-2 \end{bmatrix}$$

In the case of delayed measurement of output, the corresponding partitioned part of the error covariance matrix converged after sixteen iterations to give

$$J_{\text{delay}} = \begin{bmatrix} 6.98738_{10}-2 & -4.35125_{10}-3 \\ -4.35125_{10}-3 & 5.88950_{10}-2 \end{bmatrix}$$

The introduction of measurement delay therefore increases the number of iterations for the covariance error matrix to converge and also increases the auto-correlation error componenets, i.e. the diagonal elements of matrix J. The off diagonal or cross-correlation components are seen to be reduced, but this is of minor importance. The delayed measurement system, as expected, results in an overall deterioration in the estimator.

5.3.6. - COMPATIBILITY OF OPTIMAL DIGITAL CONTROL AND ESTIMATION ALGORITHMS.

The optimal digital control techniques developed in Chapter (2) required that the additive disturbance noise be independent. On the other hand, the optimal estimation procedures developed in this chapter required the disturbance inputs and also the measurement inputs be independent. The restriction on the noise inputs are therefore compatible for both optimal control and estimation.

To increase the class of systems to which the optimal control technique was applicable, it was found necessary to resort to the white noise representation already discussed. Since the variance of pure white noise is infinite, some of the elements of discrete covariance matrix will have infinite values. In practice, ideal white noise does not exist, there being only approximations to it (cf Appendix (G)), whose spectral density remains flat over a sufficiently large frequence range, but eventually tends to zero at high enough frequencies. The discrete

covariance matrix will therefore have some elements, whose values tend to be large valued but finite.

By the duality property between J(j+1) and P(N-j) one can write the norm inequality

$$\| J(0) \| \leq \| J(j+1) \| \qquad \text{for } j \geq 0.$$

where $\| J(0) \| = y' J(0)\, y$ for any arbitrary vector y.

Since J(0) = R(0),

$$\| R(0) \| \leq \| J(j+1) \| \qquad \text{for } j \geq 0.$$

Hence the covariance matrix of the error between the actual state and its predicted value, note not its best estimate, will contain some large valued elements, particularly on the diagonal. The large values of the norm of J(j+1) indicate that the estimate of the state vector is poor, although in fact optimal. The implementation of the optimal digital control policy, using an optimal estimator for white noise configuration would therefore result in a system's performance, which was considerably inferior to the performance with ideal measurement.

5.3.7. - SUB-OPTIMAL ESTIMATION IN A DETERMINISTIC SYSTEM.

In a deterministic system in which the number of the output variables measured at each sampling instant is less than the order, p, of the system, one need only measure the output at successive sampling

instants, until one has p independent output measurements to be able to calculate the initial and all subsequent state vectors (assuming that the system is completely observable). This approach suffers from a high sensitivity to the set of initial measurements of the output, a disadvantage common to feedforward systems. The sub-optimal estimation algorithm can be used to develop a less sensitive feedback structure. Initially, one does not know the complete state vector, so let the initial covariance error matrix have some reasonable assumed value, as if a random disturbance were present. The sub-optimal estimator is as before

$$\hat{x}_s\ (j+1/j+1) = M^{+}y(j+1) + \left[I-M^{+}M\right]\Phi\hat{x}(j/j)$$

which has a continually updated, i.e. feedback, structure and does not rely solely on one set of initial measurements; thus it is less sensitive to measurement errors and noise disturbances, which may exist but have not been taken into account.

Since the R(j) and W(j) are null matrices, the limiting value of the error covariance matrix would be

$$J(j+1) = \alpha^{j+1}\ J(0)\ \alpha^{j+1}$$

where $\alpha = \Phi\left[I-M^{+}M\right]$

which would tend to zero if all the eigen values of α were within the unit circle in the z-plane.

If one attempted to use the optimal estimator for a deterministic system, then immediately p independent output variables had been measured, the covariance matrix becomes zero and the inverse in the difference equation for V(j+1) no longer exists. Further values of the state vector could then only be calculated by the sensitive feedforward procedure.

The use of the above sub-optimal estimator in a deterministic system is an alternative to the previously proposed best linear control policy technique. In general, the state dynamic equations would have to be rearranged in terms of the output variables, before one could directly apply the best linear control technique. A drawback of the sub-optimal estimator is that it requires a finite memory, since the previous best estimate of the state vector is used in the calculation of the succeeding best estimate. The sub-optimal estimator for deterministic systems would therefore result in a considerably more complex system than the best linear control system.

5.4. - OPTIMAL SYSTEM STABILITY.

Although a system may be optimally controlled according to a prescribed performance criterion, the overall system is not automatically stable. The conditions, derived in Chapter 3, which ensure the existence of an optimal control input and hence the convergence of the performance index, are alternatively, the sufficient conditions required to prove that the state variables which appear directly or indirectly in the performance

criterion are stable. The state variables not contained in the performance criterion may, however, be unstable without adversely affecting the optimality of the system. If the coefficient feedback matrix is known, then the stability of the optimal transition matrix could be determined by an examination of the eigenvalues of that matrix. At the outset of the optimal calculations, the coefficient feedback matrix is unknown, therefore this simple approach to stability analysis is thwarted. It should also be remembered that the convergence of the optimal iterative equations for augmented systems and the quasi-analytic proof of computational stability depended on the controlled system being completely stable. It may not, therefore, always be possible to calculate the optimal feedback matrix and the direct approach to stability analysis may not be feasible. It is intended in this section to devise a complete stability condition based on the uncontrolled system's dynamic equations and the performance criterion.

5.4.1. - PERFORMANCE CRITERION VARIABLES.

The state variables can be included in a performance criterion in two ways, namely, directly through the state cost matrix Q, or indirectly through the control cost matrix H. In practice, the control input vector is usually weighted and is in general a function of all the state variables. The performance criterion therefore usually includes all the state variables of the system either directly or indirectly.

For a linear undisturbed system, the optimal steady state is reached when all the steady state variables appearing in the criterion become zero. Sufficient conditions have already been determined, which prove the existence of an optimal trajectory, and hence a null steady state vector for those state variables, which appear in the criterion. Thus, if the cost control matrix is non-null, the optimal system will in general be stable. If the state cost matrix were positive definite, then one could categorically state that the optimal system would be stable, but this matrix is seldom positive definite in practice.

Of more interest is the best linearly controlled undisturbed system, in which only some of the state variables are prescribed as being included in the control input. It will be assumed that the control input is, in fact, a linear function of all the prescribed variables, and that the control cost matrix is non-null. The unprescribed variables are therefore not included in the criterion unless they are included directly through the state cost matrix Q. Although the corresponding sufficient conditions for optimality (cf sub-section (5.2.3.)) ensure that variables appearing in criterion attain a null value, the trajectory of the unprescribed variables not appearing in the criterion may be unstable.

5.4.2. - STABILITY CONDITIONS.

The stability analysis to be used is based on the first method of Lyapunov (53) as applied to discrete systems. The stability theorem

can be stated as -

If a system can be described by the difference equation

$$x(\overline{j+1})\ T) = \Phi(T)\ x(jT)$$

and there exists a function $V\left[x(jT)\right]$ which is mathematically continuous and has the following properties

(a) $V\left[x(jT)\right]$ is positive definite for $\|x\| \neq 0$, where the norm is the euclidean norm,

(b) the first forward difference, $\Delta V\left[x(jT)\right]$, is negative semi-definite for $\|x\| \neq 0$,

and(c) $V\left[x(jT)\right] \to \infty$ for $\|x\| \to \infty$.

then the system is stable inasmuch as for every finite subspace of state vectors, $\varepsilon > 0$, there exists a corresponding subspace $\delta > 0$, such that whenever $\|x(jT)\| < \delta$, $\|x(rT)\|$ is contained in ε for $r > j$.

For the sake of simplicity, it is assumed that the coefficient feedback matrix has its converged best linear value. This is not a restriction on the analysis, since if a linear system can be proved to be stable over some time interval $t_o < t < t_1$, then it is also stable over the time interval $t_o < t < t_2$, where $t_2 < t_1$. The best linearly controlled system is therefore described by

$$x(\overline{j+1}\ T) = [\phi(T) + G(T)B]\, x(jT)$$

$$= \Phi(T)\, x(jT) \qquad (5.40)$$

As a possible Lyapunov function, consider the function

$$V\ x(jT) = \sum_{j=0}^{N} x'(jT)\, \bar{Q}\, x(jT) \qquad (5.41)$$

where $\bar{Q}$ is a matrix with unity on diagonal corresponding to all the state variables, which appear directly or indirectly in the performance criterion governing the system.

The function of equation (5.41) is mathematically continuous in the state $x(jT)$ and has properties,

(a) $V[x(jT)]$ is positive <u>semi</u>-definite, since it is a function of only some of the state variables,

(b) $\Delta V[x(jT)]$ is negative semi-definite, since matrix $\bar{Q}$ is positive semi-definite

and (c) $V[x(jT)] \to \infty$ if and only if $\|x(jT)\| \to \infty$, since only those variables which appear in performance criterion appear in $V[x(jT)]$ and hence it has an upper bound for finite $\|x(jT)\|$.

All the conditions for stability hold, except $V[x(jT)]$ is positive semi-definite, rather than positive definite for $\|x(jT)\| \neq 0$. It is therefore necessary to determine a further set of conditions which, when satisfied, ensure that $V[x(jT)]$ is, in fact, positive definite.

To this end consider a system of order p, which is described by the partitioned difference equation equivalent to equation (5.40), namely

$$\begin{bmatrix} y(j+1\ T) \\ \hline z(j+1\ T) \end{bmatrix} = \begin{bmatrix} \bar{\Phi}_{11} & \bar{\Phi}_{12} \\ \hline \bar{\Phi}_{21} & \bar{\Phi}_{22} \end{bmatrix} \cdot \begin{bmatrix} y(jT) \\ \hline z(jT) \end{bmatrix}$$

in which the state vector has been partitioned into a vector, $y(jT)$, of variables which appear in the Lyapunov function, and a vector, $z(jT)$, of variables which neither appear in criterion nor are specified to appear in best linear control policy. The partitioned parts $\bar{\Phi}_{12}$ and $\bar{\Phi}_{22}$ of the controlled system will therefore only contain the corresponding elements of the uncontrolled systems state transition matrix and therefore will be independent of the feedback coefficient matrix. The vectors $y(jT)$ and $z(jT)$ are defined to have dimension q and (p-q) respectively.

Let it be assumed that $y(jT)$ is a null vector, then for $y(\overline{j+1}\ T)$ also to be a null vector

$$\text{row}(i)\ z(jT) = 0 \quad \text{for } i = 1, z, \ldots q. \tag{5.42}$$

where row(i) is i th row of $\bar{\Phi}_{12}$.

If equation (5.42) is satisfied, then $z(j+1) = \bar{\Phi}_{22}\ z(jT)$ and condition for $y(\overline{j+2}\ T)$ to be zero is

$$\text{row}(i)\ \bar{\Phi}_{22}\ z(jT) = 0 \qquad \text{for } i = 1, z, \ldots q.$$

Similarly, $y(kT)$ is null if

$$\text{row}(i)\,\Phi_{22}^{n}\,z(jT) = 0 \qquad \text{for } i = 1, 2, \ldots q \text{ and } n = 0, 1, \ldots k-1$$

In matrix notation, the condition that $y(kT) = 0$ can be written

$$\begin{bmatrix} \text{row}(i) \\ \text{row}(i)\,\Phi_{22} \\ \cdots \\ \text{row}(i)\,\Phi_{22}^{k-1} \end{bmatrix} z(jT) = 0 \qquad \text{for } i = 1, 2 \ldots q. \tag{5.43}$$

By linear algebra, if the matrix in row(i) and Φ_{22} for k=p-q has rank equal to its dimensions, i.e. p-q, then, for equation (5.43) to be obeyed, the vector $z(jT)$ must be a null vector. Due to choice of the Lyapunov function, the only occasion when the latter is zero is when $y(jT)$ is a null vector. If $y(jT)$ is a null vector, then so also is $z(jT)$ for all future time if

$$\text{Rank}\left[\text{row}(i) \;\vdots\; \text{row}(i)\,\Phi_{22} \;\vdots\; \ldots\ \text{row}(i)\,\Phi_{22}^{p-q-1}\right] = p-q \tag{5.44}$$

Hence, if this additional condition is satisfied $V[x(jT)]$ is positive definite for $\| x(jT) \| \neq 0$. The undisturbed best linear control system can thus be proved stable, if condition of equation (5.44) is satisfied. To obtain this result, it must be remembered that it was assumed that the control cost matrix was non-null and that all the state variables which were prescribed as being included in the best linear

control policy were, in fact, included. Otherwise one could not derive a stability condition equivalent to equation (5.44), which was independent of the feedback coefficient matrix.

5.5. - DETERMINATION OF SAMPLING PERIOD

In classical sampled-data theory and more particularly in optimal digital control theory, the determination of the sampling frequency has not received much attention. In classical theory the sampling frequency seems to be usually chosen by either of two methods

(a) The empirical rule method, in which the sampling period is chosen to be 1/n times the predominant time constant of the system, where n varies from 2 to 20, depending on reference.

or (b) The ripple factor method, in which a factor, which is a measure of the intersample ripple, is minimised by reducing the sampling period until the factor satisfies the system's specifications (48).

In optimal digital control systems, the quality of the responseis determined by the resulting performance index; therefore, a quantitative method of determining the sampling frequency should be based on the variation of the optimal performance index with the sampling period. For reasons of reliability and economy, the smaller

the sampling frequency the better. The optimal sampling frequency could be defined as the smallest frequency which upon further reduction, does not significantly reduce the performance index of the system.

The ultimate performance index of the system is directly dependent on the initial state vector of the system since

$$f_N [x(0)] = x'(0)\ P(N)\ x(0)$$

where P(N) is E.D.R. matrix of dimensions $(n * n)$

It would be preferable to have a method which was independent of the initial state vector, but such a method would necessitate the examination of the $\frac{1}{2}n(n+1)$ independent elements of the symmetrix matrix P(N). This approach becomes exceedingly tedious for systems of even modest order so, as an alternative, the index-sampling period graph can be determined for a specific initial state, and hence an approximate range of possible sampling periods could be obtained. A particular sampling period could then be determined by a closer examination of the elements of the E.D.R. matrix in this range.

To illustrate the typical performance index sampling period curves and to show relationship between optimal sampling period method and the empirical rule method, two examples are detailed below.

(a) Second order damped oscillatory system:

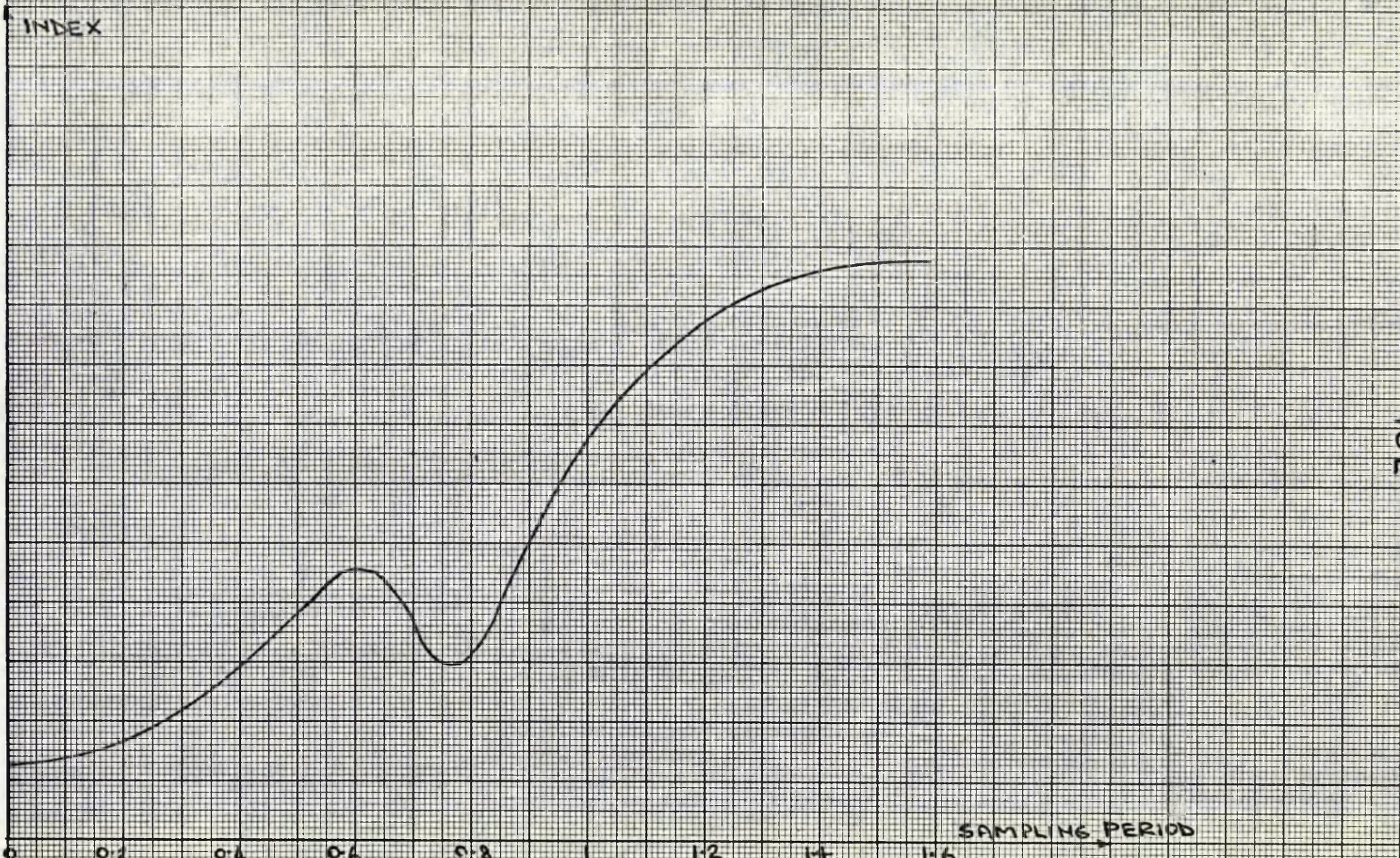


FIGURE 24 INDEX AGAINST SAMPLING PERIOD FOR OSCILATORY SYSTEM

The system of this example is described by the differential state equation

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -17.796 & -1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} m_1(t)$$

which is to be optimally controlled according to the intersample criterion equivalent to $\int_0^\infty [x_1^2(t) + x_2^2(t) + m_1^2(t)]\, dt$.

The graph of performance index against sampling period for the initial conditions $x_1(0) = x_2(0) = 1$ unit is shown in figure (24). Significant divergence of the index is not apparent until a sampling period of approximately 0.5 units. The method of determining the sampling period is not precise due to the qualitative term significant. The amount of trading of optimality for increased sampling period will depend on the particular problem being studied, so each case would have to be judged on its own merits.

Figure (24) also shows the relationship between the time constants of the uncontrolled system, the sampling period and the performance index. The eigenvalues of continuous system are

$$\lambda_{1,2} = -2 \pm j4.1887902$$

from which the exponential time constant is 0.5 units and the period of natural oscillations is 1.5 units. From Figure (24) the index begins to increase rapidly at approximately half to a third of these values.

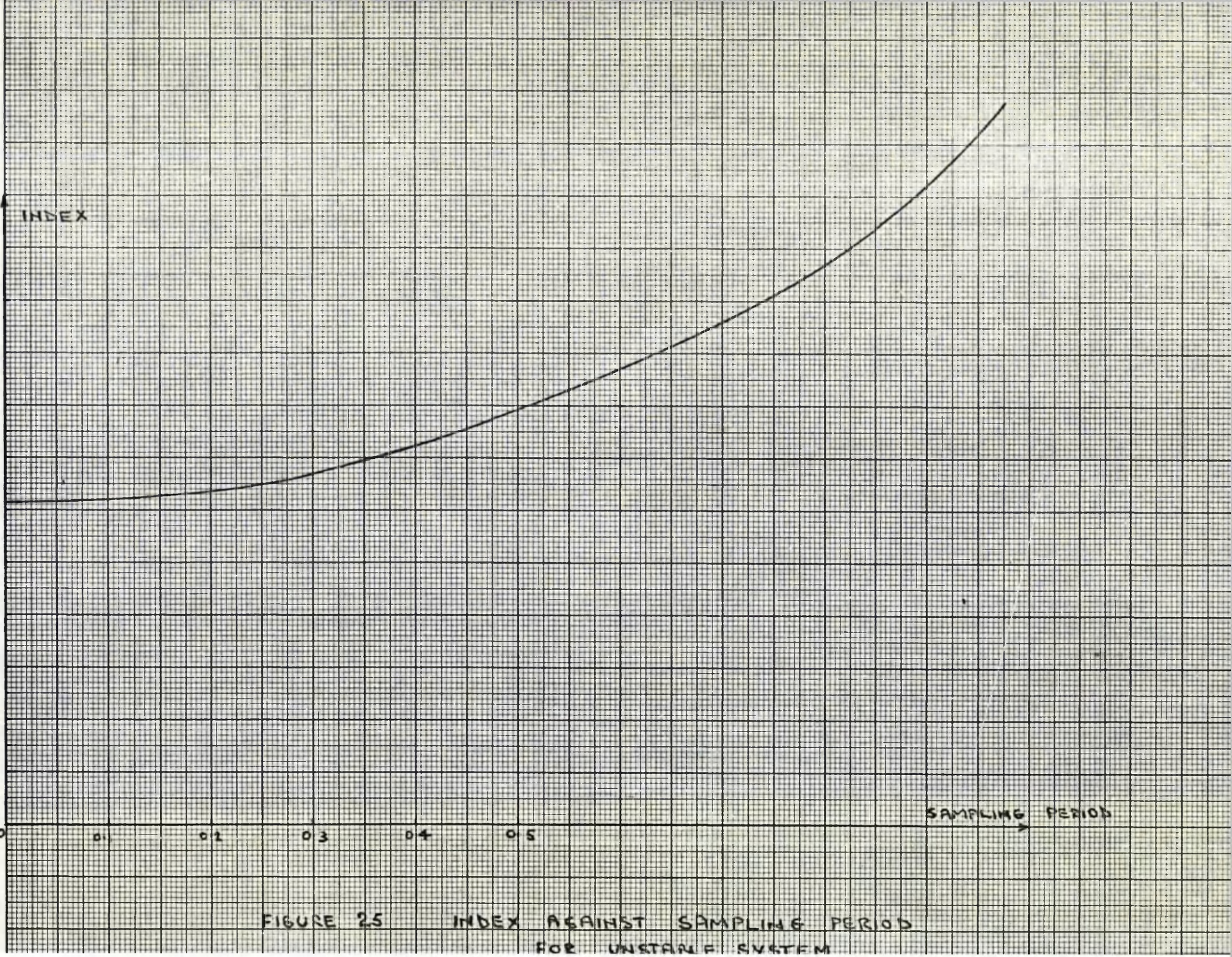


FIGURE 2.5 INDEX AGAINST SAMPLING PERIOD FOR UNSTABLE SYSTEM

The approximate relationship of a half to a third was found to be the case in many numerical examples analysed. The empirical rule that the predominant time constant and the sampling period are related in a ratio of 1 to 4 would give a reasonable sampling period, if anything a slightly overcautious estimate.

(b) Unstable second order system:

The system of this example is described by

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} m_1(t)$$

which is to be optimised according to the intersample criterion equivalent to $\int_0^\infty [x_1^2(t) + x_2^2(t) + m_1^2(t)]\, dt$.

The graph of the performance index against sampling period is shown in Figure (25) for the initial conditions $x_1(0) = x_2(0) = 1$ unit. This would indicate a sampling period in the approximate range of 0.1 to 0.2 units. In the case of unstable systems, the predominant root is the unstable root, which does not have a time constant in the conventional sense of the term. The empirical rule for determining the sampling period therefore breaks down for unstable systems, and one must resort to an examination of the index-sampling period graph.

CHAPTER 6 - TWO EXAMPLES

In this chapter, two systems, namely the output concentration control of a chemical reactor and the short period pitch control of an aircraft, will be examined. Both these systems are basically non-linear systems but, for the digital optimisation analysis which has been developed, it is necessary to linearise them. The two systems were chosen because the reactor was unstable if uncontrolled and, as a result, is difficult to control, whereas the aircraft was stable and is thus easily controlled. The details of the dynamics of the systems and the notation used in their description are given in Appendices E and F. The two systems will now be examined in turn.

6.1. - THE OUTPUT CONCENTRATION CONTROL OF A CHEMICAL REACTOR.

6.1.1. - THE PERFORMANCE CRITERION

Firstly, one must decide upon a performance criterion. Minimum deviation from the maximum steady state product concentration is desired, thus the linearised product concentration perturbed variable is naturally weighted. The reactor temperature and the control inputs are also assumed to be weighted. The absolute values of the elements of the performance criterion cost matrices do not influence the calculation of the coefficient feedback matrix; it is only the relative weighting between the elements of the matrices which matter. The relative weighting was determined by the rule of thumb that the cost of a

2% variation in the steady state values of the product concentration should equal the cost of a 5% variation in each of the other weighted variables. All the linearised perturbed variables were normalised by dividing by their corresponding steady state values (cf Appendix E), and thus the performance criterion was defined by the intersample criterion equivalent to

$$\int_0^\infty [x_4^2(t) + 0.16T^2(t) + 0.16\ f_2^2(t) + 0.16\ f_w^2(t)]\ dt \qquad . \quad (6.1)$$

where x_4 is the output/product concentration,

T is the reactor temperature,

f_2 is the control input flow of reactant 2,

f_w is the control input flow of cooling water

and all variables are the linearised perturbed variables.

6.1.2. - CHARACTERISTICS OF SYSTEM WITH NO RECYCLE DYNAMICS.

The linearised reactor configuration, where it is assumed that there are no recycle dynamics in the feedback path, is governed by the equation

$$\dot{y}(t) = A\ y(t) + D\ m(t) \qquad (6.2)$$

where $y_i(t)$ for $i = 1, 2, \ldots 5$ are the respective reactant concentrations in effluent flow of reactor (cf Appendix E),

$y_6(t)$ is the reactor temperature,

CHEMICAL REACTOR MODEL WITH NO RECYCLE DYNAMICS - FIGURE 26.

REACTOR TEMPERATURE 6.6200 10 +2

A

-3.71443 10 +1	-3.14986 10 +1	-1.90542 10 -1	0.00000	-5.91203 10 +0	-4.90007 10 +
-8.10950 10 +0	-2.57805 10 +1	-5.58273 10 +0	0.00000	-5.91203 10 +0	-2.48335 10 +
3.25129 10 +2	6.90159 10 +1	-3.26470 10 +2	-6.45202 10 +1	-5.91203 10 +0	-1.76285 10 +
-1.14993 10 +0	1.60854 10 +1	1.51160 10 +1	-1.42076 10 +1	-5.91203 10 +0	3.07207 10 +
-1.14993 10 +0	4.49576 10 +0	8777172 10 +0	0.00000	-1.48743 10 +1	2.03072 10 +
2.21573 10 +0	3.24576 10 +0	1.78190 10 +0	7.51867 10 -1	0.00000	6.39346 10 +

D

-8.12347 10 +0	0.00000
1.31905 10 +1	0.00000
-8.12347 10 +0	0.00000
-3.72569 10 +0	0.00000
-8.12347 10 +0	0.00000
-3.25718 10 -1	-1.39262 10 -1

$m_1(t)$ is the input flow of reactant 2,

$m_2(t)$ is the input flow of cooling water,

and all variables are linearised perturbed variables.

The matrices A and D are defined in Figure (26) for a reactor temperature of $662^{\circ}R$.

It is found that the eigenvalues of the system are

$$\lambda_{1,2} = +6.93 \pm j\,1.99$$
$$\lambda_{3,4} = -21.65 \pm j\,1.09$$
$$\lambda_5 = -8.96$$
$$\lambda_6 = -316.15$$

The system is therefore unstable and, in agreement with the approximate computational stability analysis, it was found that the iterative procedure for calculating the E.D.R. matrix sequence was extremely computationally unstable. The proposed progressive stability technique nevertheless enabled an accurate determination of the E.D.R. matrix sequence free from instabilities. A possibility also exists of there being periodic index instabilities due to the system having complex roots, but upon calculating the controllability numbers for the critical periods of 1.578 and 2.878 hrs., it was found that since the system had two control inputs, the controllability number was non-zero. In any case, the critical periods are considerably in excess of the chosen sampling period, as will be seen in the next sub-section.

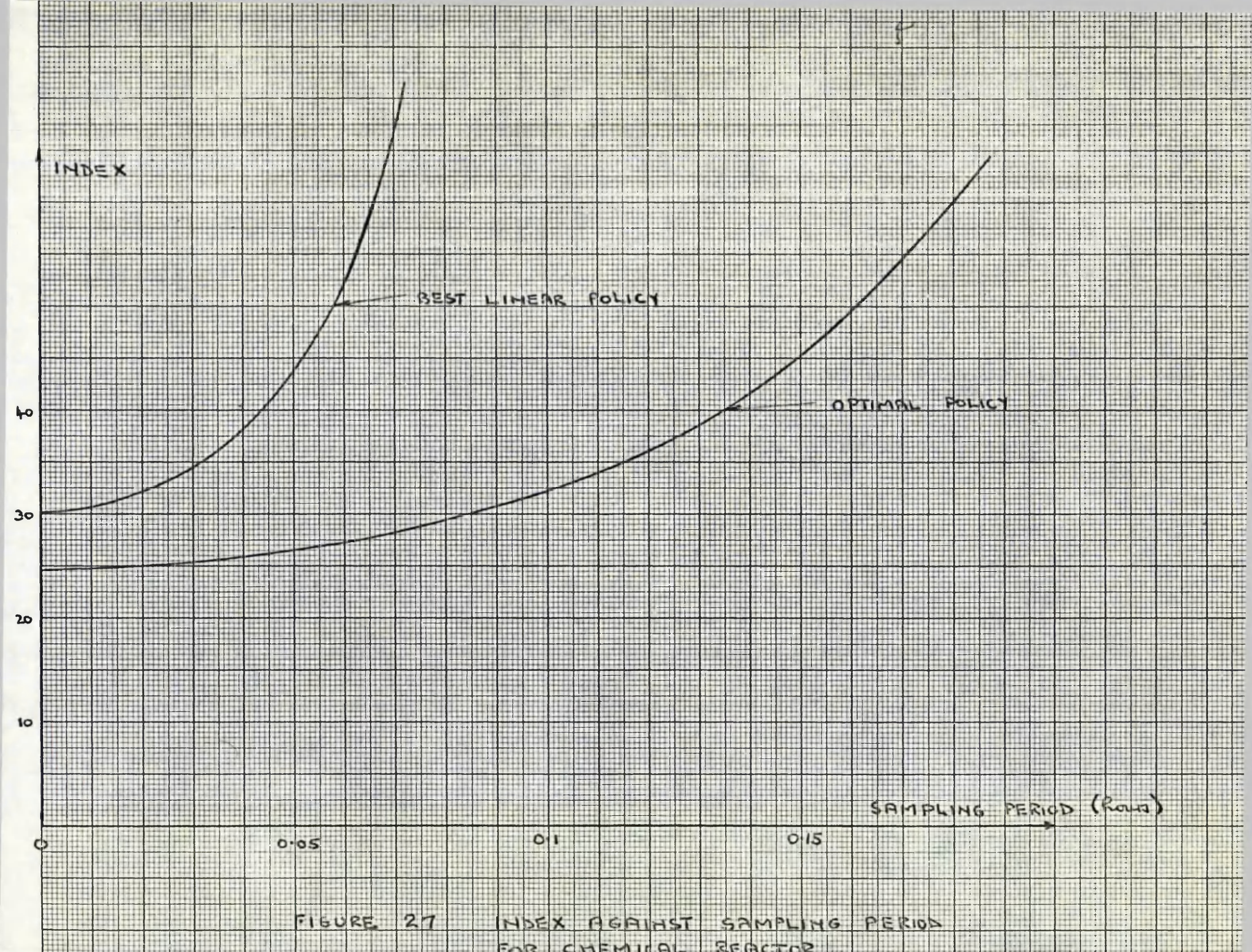


FIGURE 27 INDEX AGAINST SAMPLING PERIOD FOR CHEMICAL REACTOR

6.1.3. - SAMPLING PERIOD

To determine a suitable sampling period, the graph of the performance index against sampling period was constructed for initial conditions $y_4 = 0.01$ and $y_6 = 0.025$. It is illustrated in Figure (27). The monotonically divergent curve is typical of those for unstable systems. A sampling period of 0.05 hrs. was decided as being a compromise between optimality and a large sampling period. Examination of the E.D.R. matrix in the vicinity of this period indicated its suitability for all initial conditions.

The corresponding curve for the best linear control policy is also illustrated in Figure (27) for the case where the state variable $y_3(t)$ was not permitted to appear in the feedback control matrix. It is noticed that the divergence of the curve is considerably increased, and that a much lower sampling period is indicated. The reason for the increased divergence can possibly be explained by the results obtained in the analysis of the sensitivity of optimal systems. A change in sampling period is essentially the same as a change in the system's parameters. It was noted in Chapter 4 that the optimal system seemed to result in an insensitive system; thus one would expect a sub-optimal system, such as the best linear control policy, to be more sensitive to parameter variations, and thus to variations in the sampling frequency. The increased divergence for the best linear control system is a characteristic common to many systems, a second example being illustrated in the case of the aircraft pitch control system.

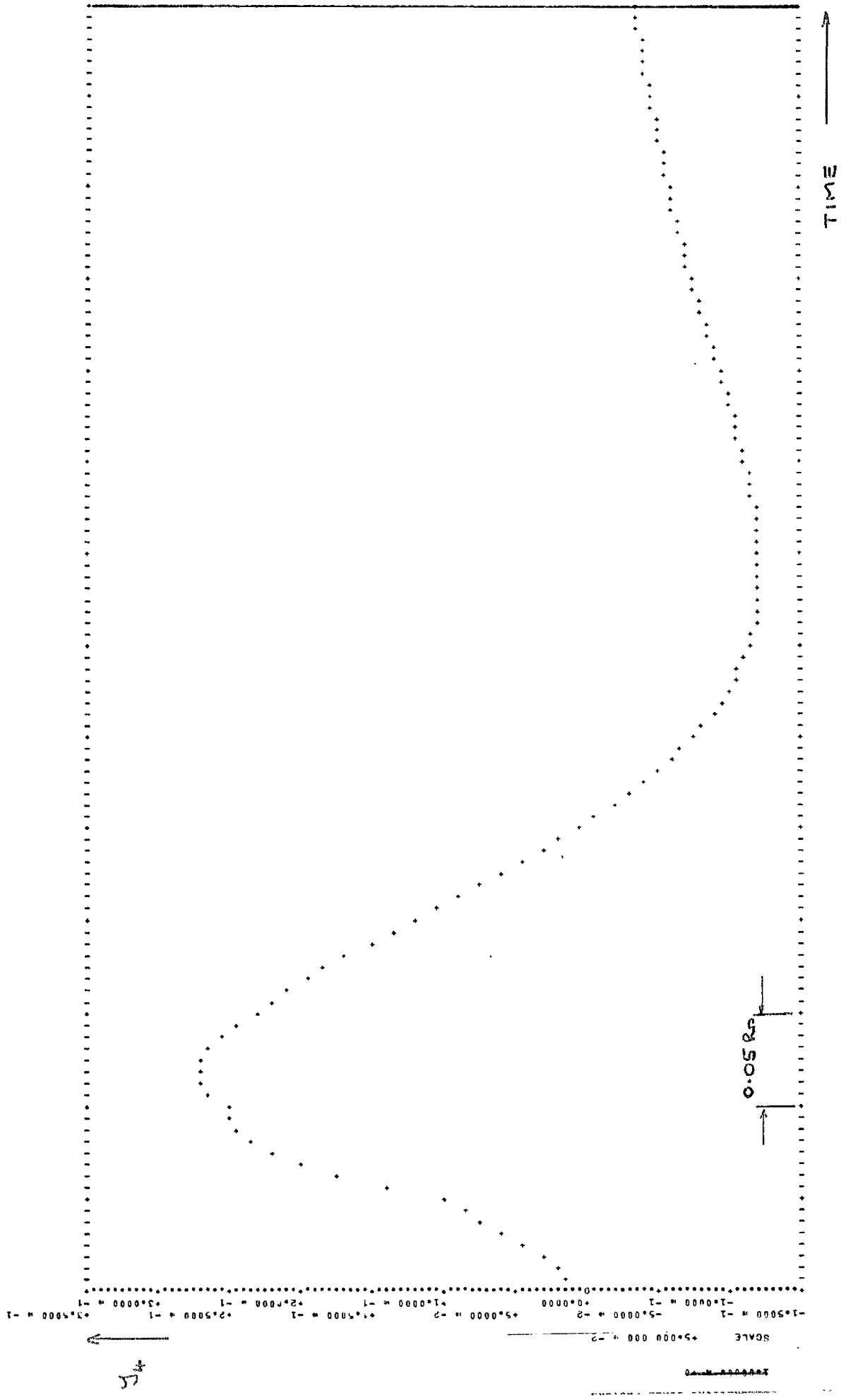


FIGURE 28. - PRODUCT CONCENTRATION TRAJECTORY.

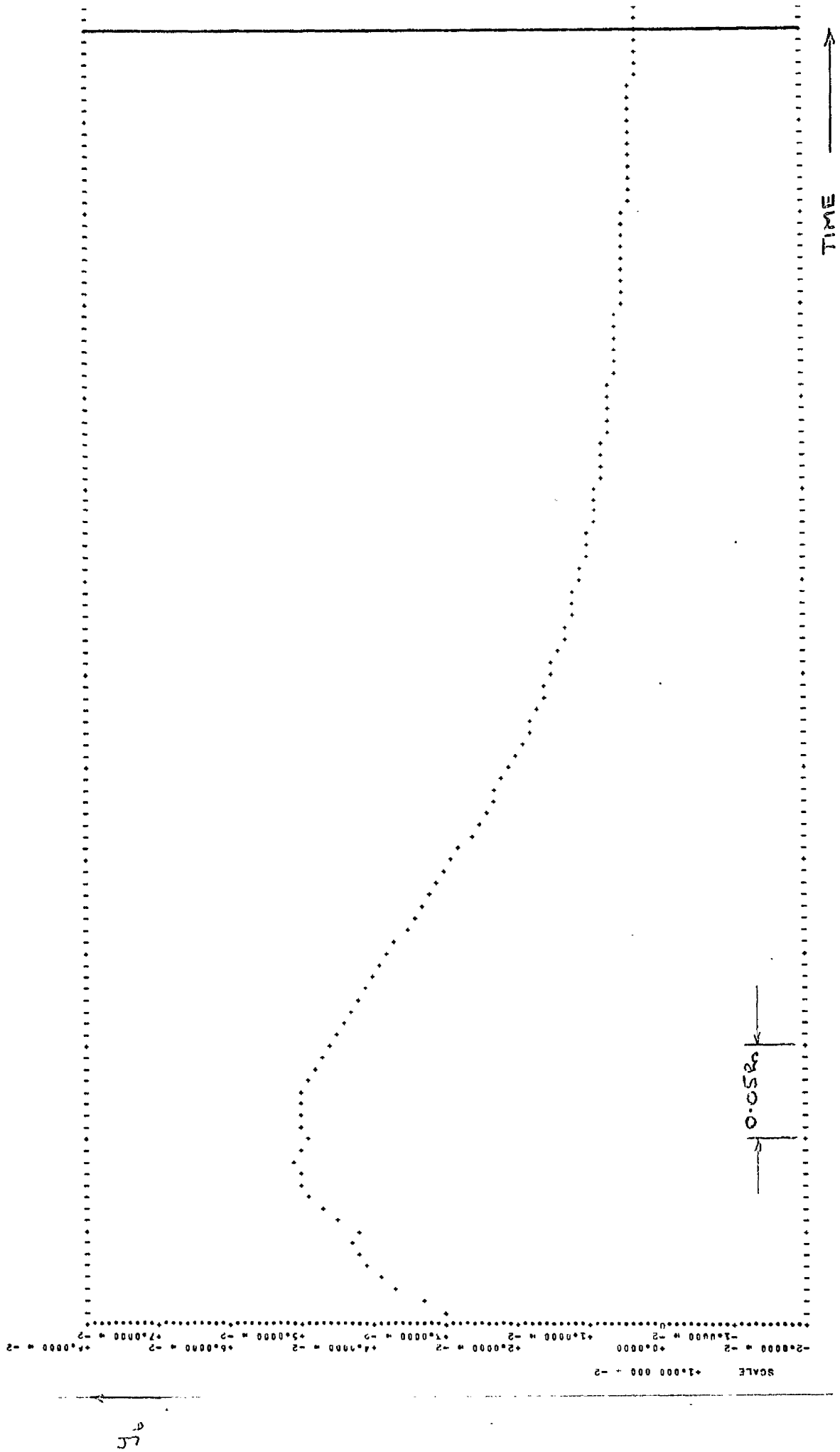


FIGURE 29 - REACTOR TEMPERATURE TRAJECTORY.

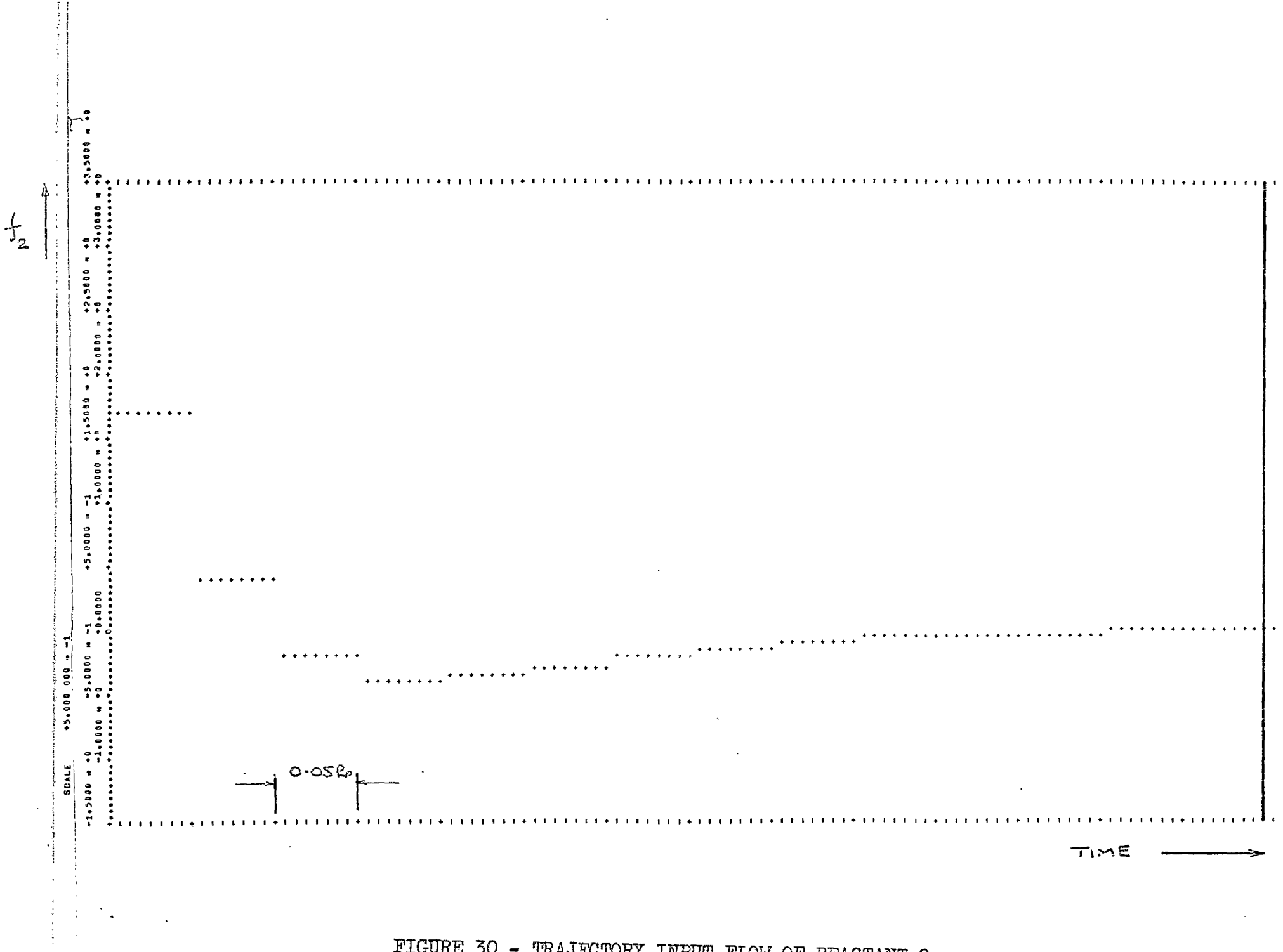


FIGURE 30 - TRAJECTORY INPUT FLOW OF REACTANT 2.

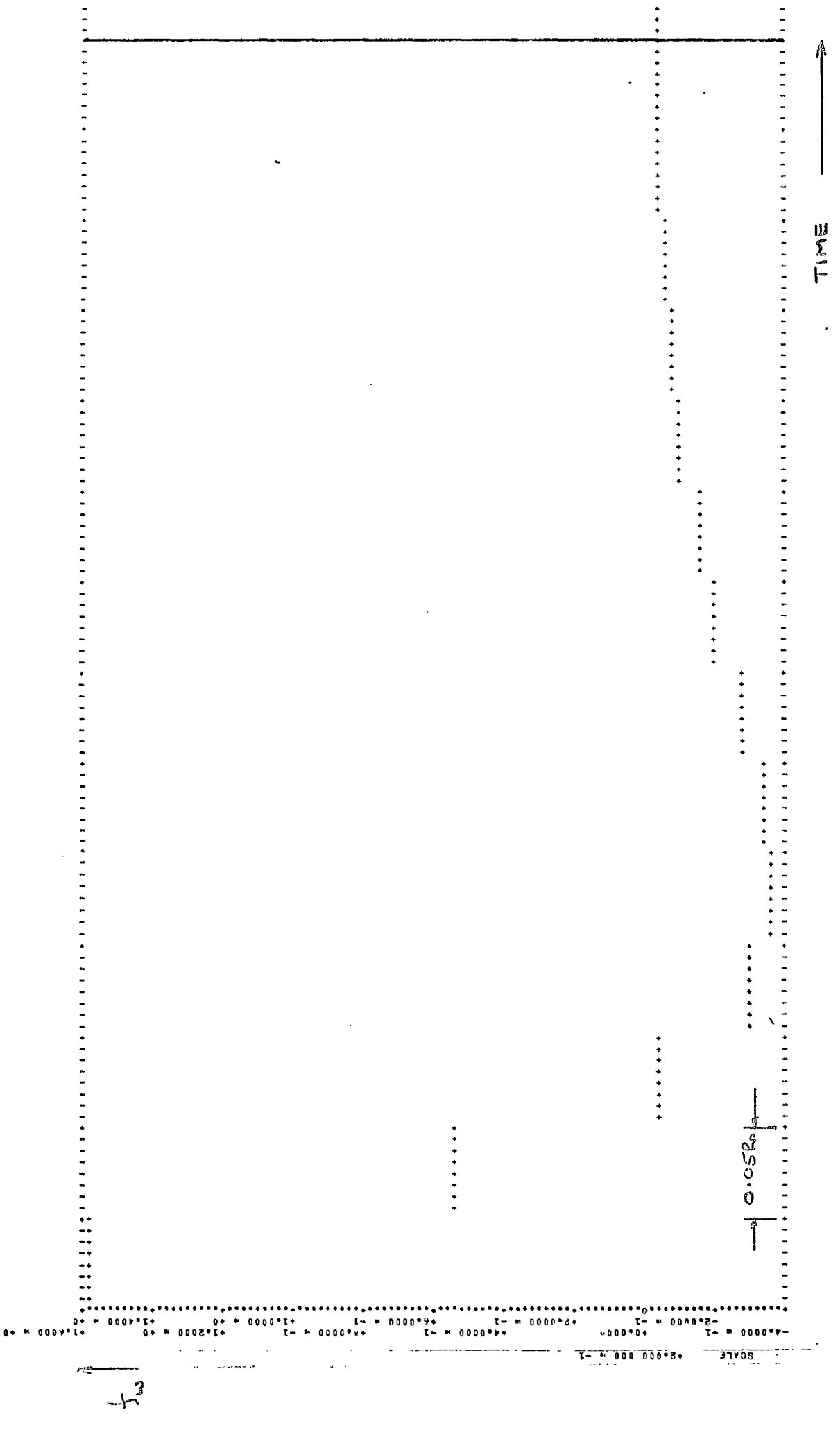


FIGURE 31 - TRAJECTORY OF INPUT FLOW OF COOLING WATER.

6.1.4. - OPTIMAL STATE TRAJECTORY.

The optimal state trajectories of the reactor concentrations and temperature were determined by digital computation for a sampling period of 0.05 hrs. and initial conditions of $y_4(0) = 0.01$ and $y_6(0) = 0.025$, i.e. 1% and 2.5% perturbations in the steady state values of the product concentration and reactor temperature respectively. The optimal feedback coefficient matrix was found to be

$$B = \begin{bmatrix} 4.469_{10}{+0} & 2.822_{10}{+0} & 3.124_{10}{-1} & 1.042_{10}{+0} & -1.573_{10}{+0} & 6.776_{10}{+1} \\ 4.105_{10}{+0} & 2.768_{10}{+0} & 2.857_{10}{-1} & 1.045_{10}{+0} & -1.584_{10}{+0} & 6.239_{10}{+1} \end{bmatrix}$$

Reduced copies of some of the graphical output of the author's simulation program are shown in Figures (28), (29), (30) and(31). In Figures (28), (30) and (31), it is seen that even for the very small initial perturbations, the trajectories exhibit excessively large perturbations in the product concentration and control flow inputs. Even larger perturbations occur in the unweighted variables, so much so that some of the perturbed variables have maximum values of greater than $\pm$ 100% of their steady state values. The linearised model of the chemical reactor is therefore completely inadequate for optimisation purposes with the proposed performance criterion. It might be possible that a performance index which weighted all the state and control variables could be derived on a trial and error basis, such that excessive perturbations were not evident. In this case, one is no longer optimally controlling the system

CHEMICAL REACTOR MODEL WITH RECYCLE DYNAMICS - (PADE FILTER) - FIGURE 32.

REACTOR TEMPERATURE 6.6200 10 +2

DELAY 2.0000 10 -2

A

-4.77133 10 +1	-2.70321 10 +1	0.00000	0.00000	0.00000	-4.90007 10 +
-6.95957 10 +0	-3.30330 10 +1	-5.39218 10 +0	0.00000	0.00000	-2.48335 10 +
3.26279 10 +2	7.34824 10 +1	-3.37998 10 +2	-6.45202 10 +1	0.00000	-1.76285 10 +
0.00000	2.05519 10 +1	1.53065 10 +1	-2.59266 10 +1	0.00000	3.07207 10 +
0.00000	8.96226 10 +0	8.96226 10 +0	0.00000	-2.06813 10 +1	2.03072 10 +
2.21573 10 +0	3.24576 10 +0	1.78190 10 +0	7.51867 10 -1	0.00000	6.39346 10 +
1.37901 10 +2	-1.10812 10 +1	-1.62592 10 +0	0.00000	-5.04483 10 +1	4.90007 10 +
-2.85293 10 +0	9.49197 10 +1	3.76626 10 +0	0.00000	-5.04483 10 +1	2.48335 10 +
-3.36092 10 +2	-1.11596 10 +2	4.36372 10 +2	6.45202 10 +1	-5.04483 10 +1	1.76285 10 +

1.17190 10 +1	0.00000	0.00000
0.00000	1.17190 10 +1	0.00000
0.00000	0.00000	1.17190 10 +1
-1.04541 10 +0	-4.06055 10 +0	-1.73224 10 -1
-2.27942 10 +0	-8.85361 10 +0	-3.77697 10 -1
0.00000	0.00000	0.00000
-1.11719 10 +2	0.00000	0.00000
0.00000	-1.11719 10 +2	0.00000
0.00000	0.00000	-1.11719 10 +2

D

-8.12347 10 +0	0.00000
1.31905 10 +1	0.00000
-8.12347 10 +0	0.00000
-3.72569 10 +0	0.00000
-8.12347 10 +0	0.00000
-3.25718 10 -1	-1.39262 10 -1
8.12347 10 +0	0.00000

to minimise the deviations in the product concentration, but one is simply altering the criterion for no reason other than to solve the large perturbation problem. The author found that even weighting all the elements equally did not result in an acceptable state trajectory. It must therefore be concluded that the idealised chemical reactor problem cannot be optimally controlled without taking into account its non-linear dynamics.

Similar results also hold for the best linearly controlled system.

6.1.5. - SECOND REACTOR CONFIGURATION.

The second, slightly more realistic, reactor configuration, as discussed in Appendix E, allowed for a first approximation Pade filter being included in the recycle path. It was assumed that the pure time delay to which the approximation was being made was 0.02 hrs. The resulting dynamic state equations are

$$\dot{z}(t) = A\, z(t) + D\, m(t) \qquad (6.3)$$

where $z_i(t)$ for $i = 1, 2 \ldots .6$ and $m_1(t)$ and $m_2(t)$ are defined in equation (6.2)

and $z_7(t)$ is the recycle concentration of reactant 1.

$z_8(t)$ " " " " " " 2.

$z_9(t)$ " " " " " " 3.

The matrices A and D are defined in Figure (32).

CHEMICAL REACTOR MODEL WITH IDEAL RECYCLE DELAY - FIGURE 33.

REACTOR TEMPERATURE 6.6200 10 +2

AO

-4.77133 10 +1	-2.70321 10 +1	0.00000	0.00000	0.00000	-4.90007 10
-6.95957 10 +0	-3.30330 10 +1	-5.39218 10 +0	0.00000	0.00000	-2.48335 10
3.26279 10 +2	7.34824 10 +1	-3.37998 10 +2	-6.45202 10 +1	0.00000	-1.76285 10
0.00000	2.05519 10 +1	1.53065 10 +1	-2.59266 10 +1	0.00000	3.07207 10
0.00000	8.96226 10 +0	8.96226 10 +0	0.00000	-2.06813 10 +1	2.03072 10
2.21573 10 +0	3.24576 10 +0	1.78190 10 +0	7.51867 10 -1	0.00000	6.39346 10

A1

1.05691 10 +1	-4.46649 10 +0	-1.90542 10 -1	0.00000	-5.91203 10 +0	0.00000
-1.14993 10 +0	7.25250 10 +0	-1.90542 10 -1	0.00000	-5.91203 10 +0	0.00000
-1.14993 10 +0	-4.46649 10 +0	1.15285 10 +1	0.00000	-5.91203 10 +0	0.00000
-1.14993 10 +0	-4.46649 10 +0	-1.90542 10 -1	1.17190 10 +1	-5.91203 10 +0	0.00000
-1.14993 10 +0	-4.46649 10 +0	-1.90542 10 -1	0.00000	5.80696 10 +0	0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

D

-8.12347 10 +0	0.00000
1.31905 10 +1	0.00000
-8.12347 10 +0	0.00000
-3.72569 10 +0	0.00000
-8.12347 10 +0	0.00000
-3.25718 10 -1	-1.39262 10 -1

Again it was found that the optimally and best linearly controlled system resulted in a system with excessively large deviations in the state trajectories.

6.1.6. - THIRD REACTOR CONFIGURATION.

The third reactor configuration, as discussed in Appendix E, allowed for a pure time delay of 0.02 hrs. to be included in the recycle dynamics. The system equations in this case are described by the difference differential equation

$$\dot{y}(t) = A_o y(t) + A_1 y(t-0.2) + Dm(t) \qquad (6.4)$$

where $y(t)$ and $m(t)$ are defined in equation (6.2).

The matrices A_o, A_1 and D are defined in Figure (33).

Using the methods proposed in Chapter 2, the state difference equation approximately equivalent to equation (6.4) was determined, but not without difficulty. Due to system being unstable, the convergence of the various infinite series encountered was slow, so the calculations required a larger amount of computer storage space than was available. The author had therefore to be satisfied with a slightly less accurate approximate state difference equation than he would have liked. As expected, this system arrangement once again resulted in excessively large state trajectory perturbations.

It must therefore be concluded that even the more realistic reactor configurations cannot be optimally controlled by considering the linearised form of their dynamics and thence deriving the small perturbation optimal control policy.

6.2 - THE PITCH CONTROL OF AN AIRCRAFT

6.2.1. - THE PERFORMANCE CRITERION.

As discussed in Appendix F, it is desired that the pitch of the aircraft to a step input on the joy-stick should follow an ideal pitch trajectory as closely as possible. One must also take into account that the elevator angle and pitch rate must also be weighted, so as not to result in a system with excessively large values of either, which would be practically unacceptable. The relative weightings were obtained by the rule of thumb that a cost of a 2^{o}error in(actual minus ideal) pitch should equal each of the costs of a 40^{o} elevator angle and a 4^{o}/sec. pitch rate. The resulting difference intersample criterion was thus equivalent to the integral criterion

$$\int_{o}^{\infty} \left\{ \left[x_4(t) - x_7(t) \right]^2 + 0.025\, x_1^2(t) + 0.25 x_3^2(t) \right\} dt \qquad (6.5)$$

where notation is that of Appendix F.

6.2.2. - CHARACTERISTICS OF SYSTEM

The completely augmented state system for an aircraft altitude

$$
\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \\ \dot{x}_4(t) \\ \dot{x}_5(t) \\ \dot{x}_6(t) \\ \dot{x}_7(t) \\ \dot{x}_8(t) \\ \dot{x}_9(t) \end{bmatrix} = \begin{bmatrix} -8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 57.95 & -2.60 & -50.48 & 0 & 42.26 & 0 & 0 & 0 & 0 \\ -44.43 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1.44 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.71 & -2.03 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -10 & -4.4 & 10 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \\ x_5(t) \\ x_6(t) \\ x_7(t) \\ x_8(t) \\ x_9(t) \end{bmatrix} + \begin{bmatrix} 5 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} m(t) + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} u(t
$$

FIGURE 34 - AIRCRAFT STATE EQUATIONS AT 30,000 ft. AND MACH 1.

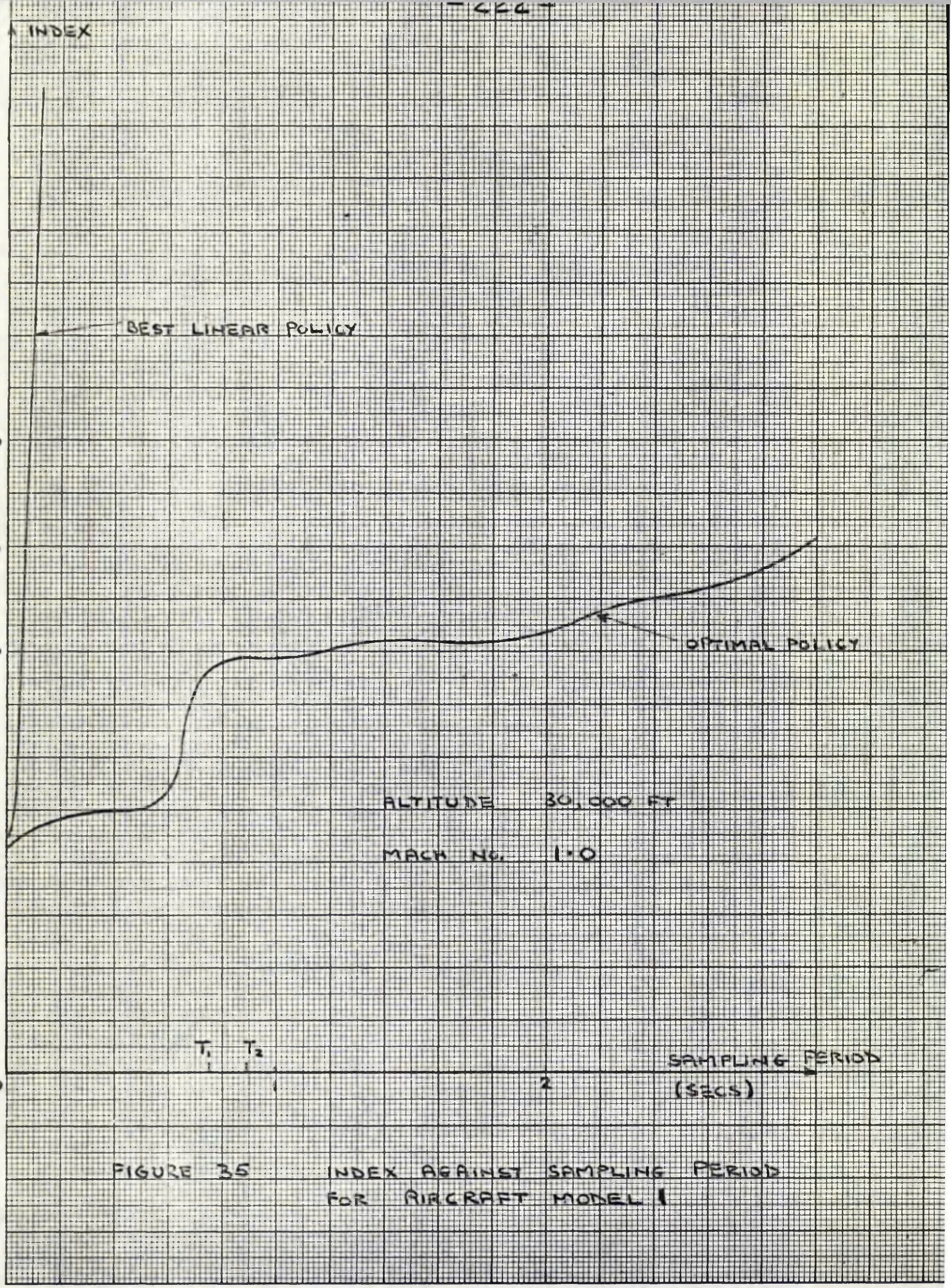


FIGURE 35 INDEX AGAINST SAMPLING PERIOD FOR AIRCRAFT MODEL 1

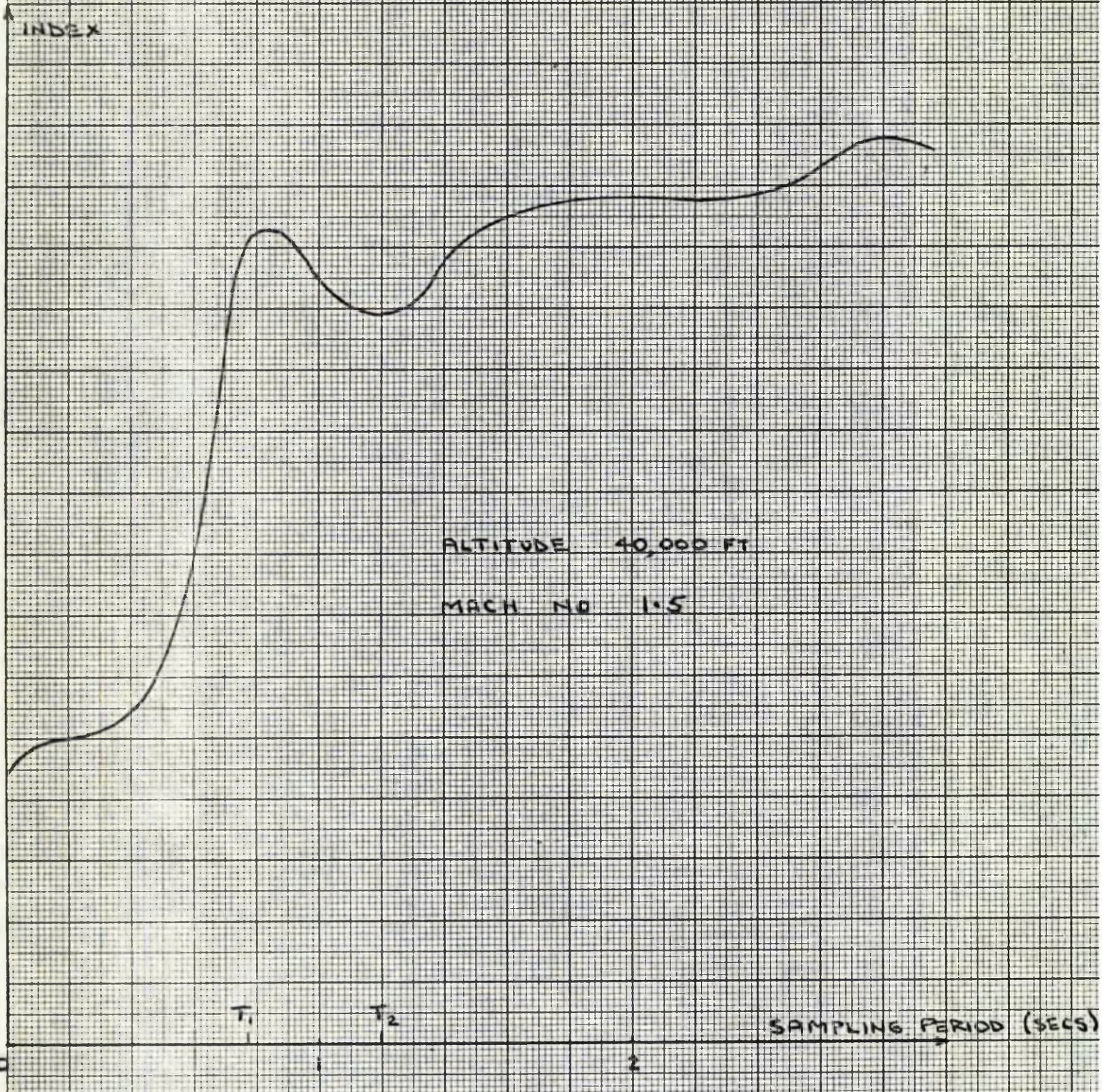


FIGURE 36 INDEX AGAINST SAMPLING PERIOD FOR AIRCRAFT MODEL 2

of 30,000ft. and a velocity of mach 1 is given in Figure (34). The state vector is fully augmented by the noise filter variables and the reference pitch variables. (cf Appendix F for notation).

The dynamics are stable throughout the flight envelope and, in the five particular combinations of altitude and velocity studied, it was found that there was no evidence of computational instability. This observation is in agreement with the computational stability analysis of Chapter 3. The system has a pair of complex roots, but due to the system being stable, there is no index instability at the corresponding critical periods.

6.2.3. - SAMPLING PERIOD

Typical graphs of the performance index against sampling period are shown in Figures (35) and (36) for two particular points in the flight envelope. In obtaining these graphs, it was assumed that there was no noise disturbance input and that the joy-stick step input was 10^{o}. These figures indicate a suitable sampling period of 0.2 secs. This is a low value of period and would indicate possible difficulties in implementing the optimal control policy, particularly if any calculations such as updating the feedback coefficient matrix, were to be performed between successive samples. The predominant roots of the aircraft in both cases are the complex pair, whose exponential time constant, T_1, and natural period of oscillation T_2, are shown on the time axis of each graph. It is seen that there is an abrupt rise in the performance index

as the sampling period approaches these values. This shows the approximate relationship between sampling period and the predominant time constants as noted in Chapter 5.

It was found that the performance index for a sampling period of 0.2 secs. was remarkably constant throughout the flight trajectory, as can be seen from the following table

Altitude (ft.)	Mach No.	Performance Index
10,000	0.5	27.4
40,000	0.5	29.8
10,000	1.5	24.5
40,000	1.5	29.9
30,000	1.0	27.8

This would indicate that the optimal performance of the aircraft, although it has widely differing stability characteristics, has been forced to respond in such a way as to be approximately independent of its uncontrolled dynamics. The associated optimal feedback coefficient matrices naturally do not carry through the approximate constancy of the performance indeces; thus, to obtain a response independent of the position in the flight envelope, it would be necessary to continually update the feedback coefficient matrix with the aircraft's position in its flight envelope.

The performance index-sampling period graph is also shown in Figure (35) for the best linear control policy for the case where the fictitious

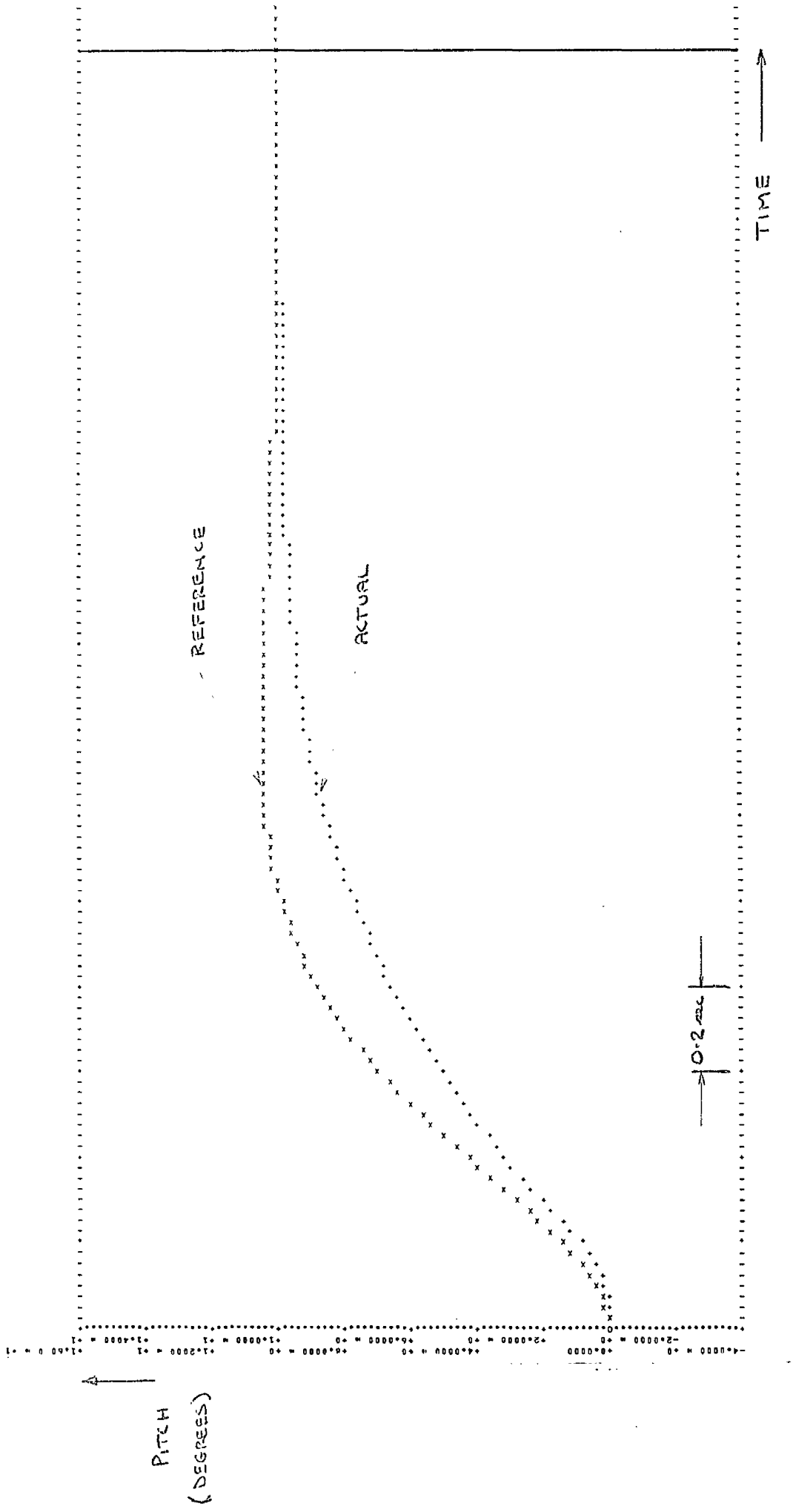


FIGURE 37 - ACTUAL AND REFERENCE PITCH TRAJECTORIES.

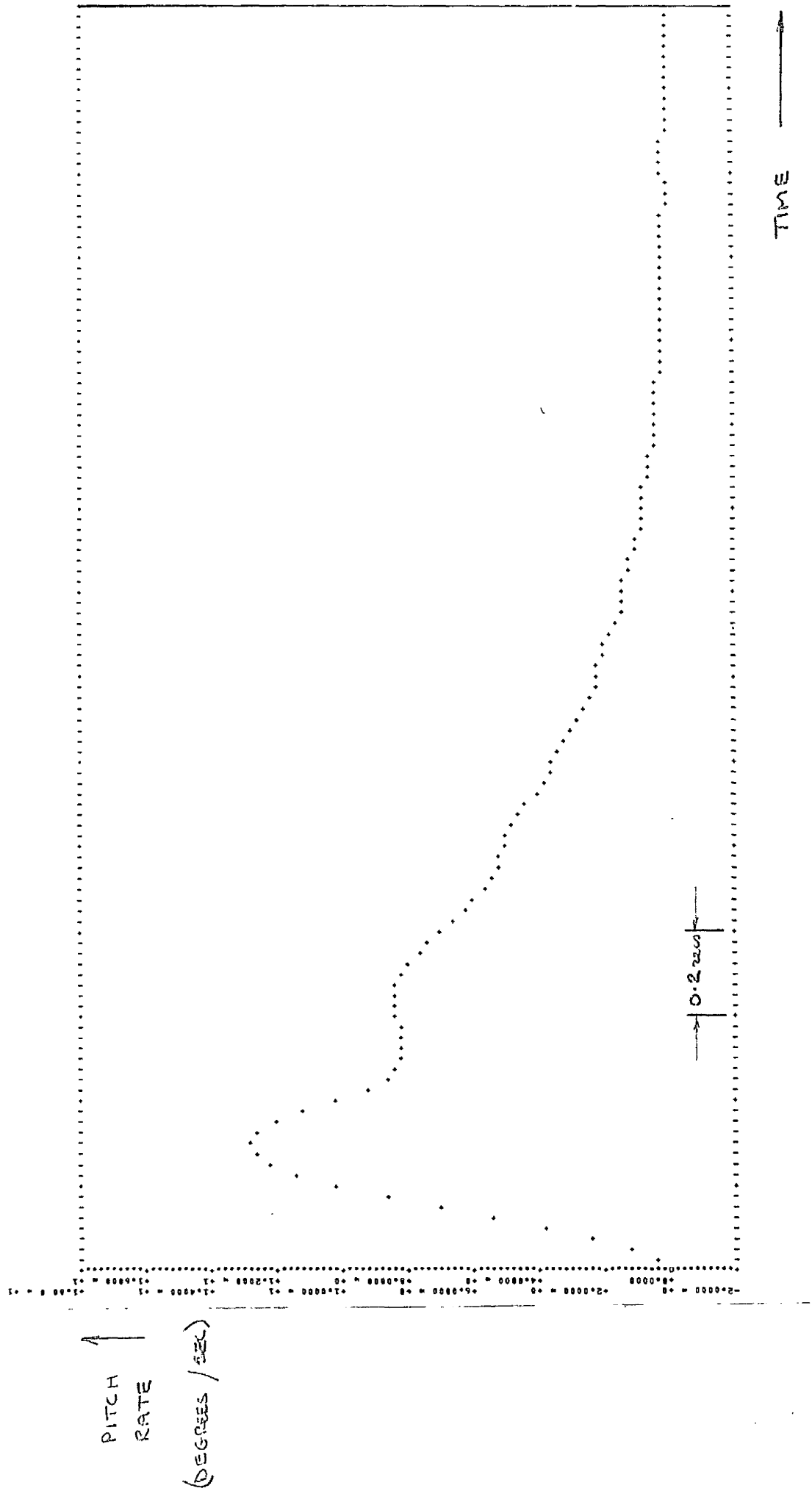


FIGURE 38 - PITCH RATE TRAJECTORY.

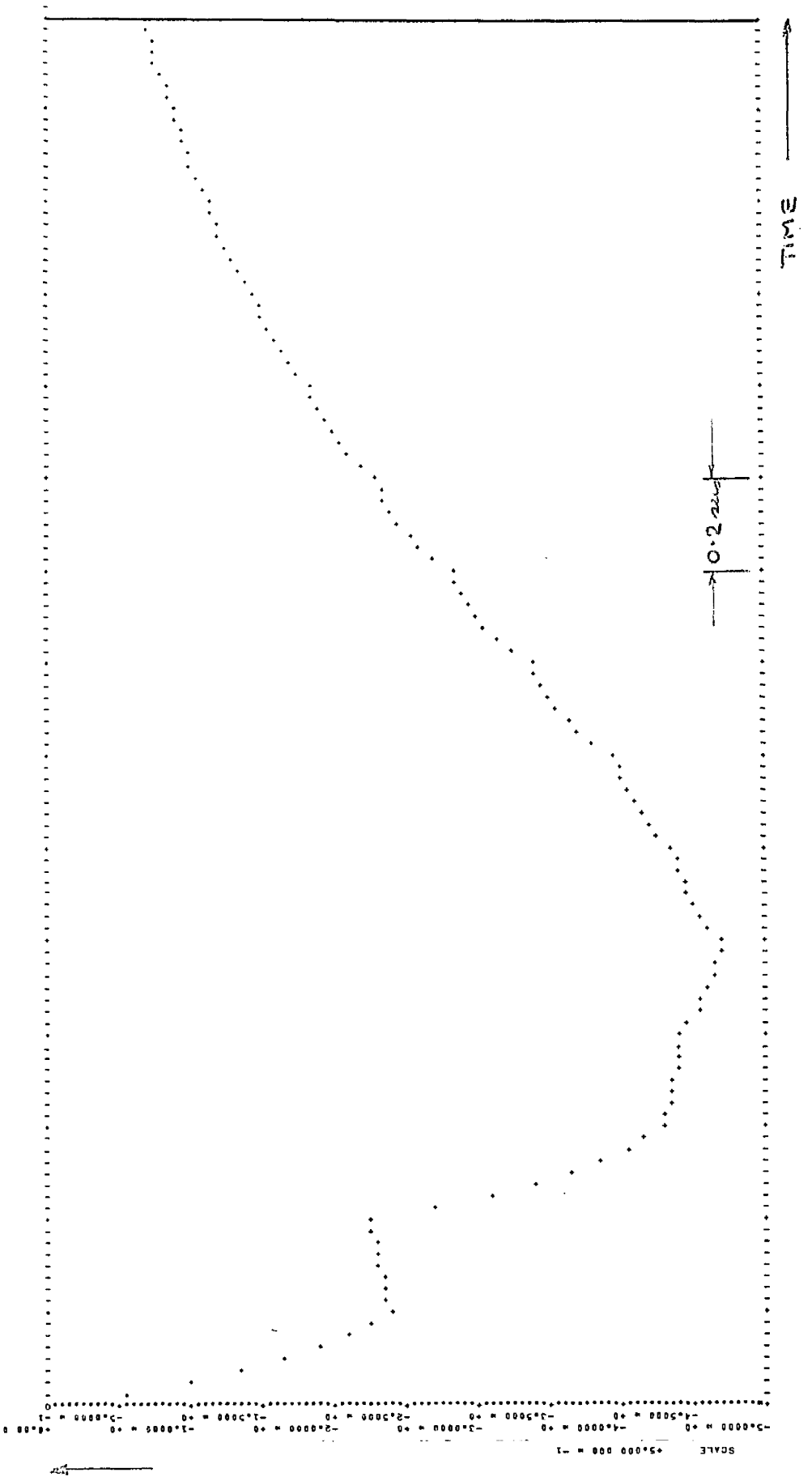


FIGURE 39 - ELEVATOR ANGLE TRAJECTORY.

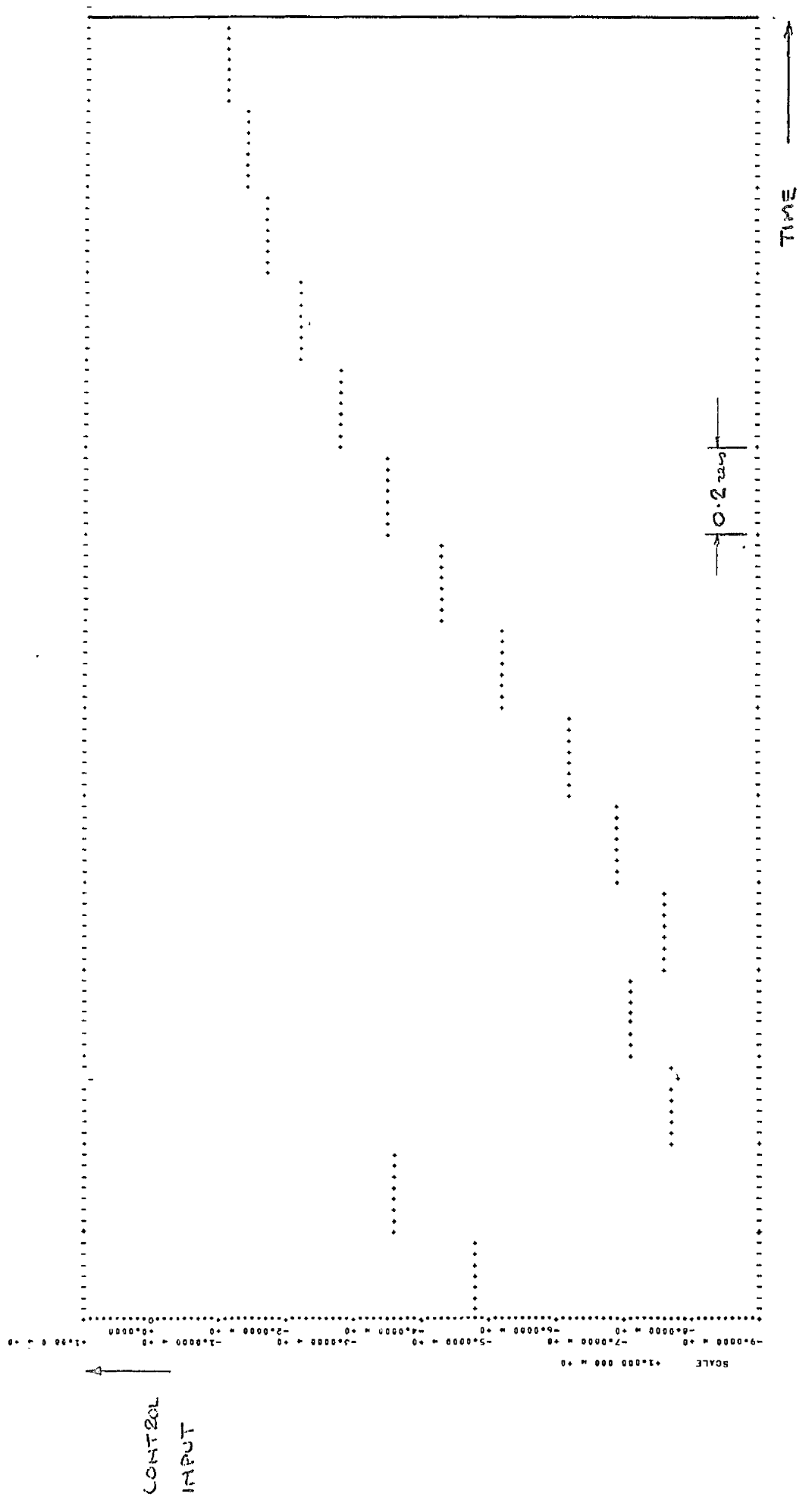


FIGURE 40 - TRAJECTORY OF CONTROL INPUT TO ELEVATOR ACTUATOR.

state variable $x_2(t)$ is not included in the control policy generator. There is a considerable increase in the divergence of the curve compared to that of the unconstrained control policy. This would indicate that feedback from state variable $x_2(t)$, although the latter is fictitious, is important and would have to be implemented in some other way, possibly by using a dynamic feedback filter.

6.2.4. - STATE TRAJECTORY.

The trajectory of the pitch, pitch rate, elevator and control input to the elevator actuator for a deterministic aircraft system (i.e. no noise disturbance) at an altitude of 30,000ft. and a velocity of Mach 1 is shown in Figures (37), (38), (39) and (40) for a joy-stick step input of 10°. The feedback coefficient matrix in this case is

$$B=\left[1.107_{10}{+0}\quad 5.189_{10}{-2}\quad 1.909_{10}{-1}\quad 7.577_{10}{-1}\quad -2.754_{10}{-1}\quad -7.240_{10}{-2}-4.822_{10}{-1}\right]$$

If no cost weighting was put on the elevator angle and the pitch rate, then for a period of 0.2 secs. the actual pitch could be made to follow the ideal pitch, so much so that one would not be able to distinguish between the curves if plotted on same scale as Figure (37). In this case, however, the maximum elevator angle and pitch rate are -5.82° and 17.65°/sec. respectively, whereas for weighted performance criterion the maximum values are reduced to -4.69° and 12.75°/sec. respectively.

R - DISCRETE COVARIANCE MATRIX.

0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.00000	2.22675 10 +3	1.84229 10 +2	1.01933 10 +1	6.07008 10 +2	-5.83444 10 -2
0.00000	1.84229 10 +2	1.71991 10 +1	1.01822 10 +0	4.51269 10 +1	3.48742 10 -6
0.00000	1.01933 10 +1	1.01822 10 +0	6.26922 10 -2	2.38914 10 +0	1.23965 10 -4
0.00000	6.07008 10 +2	4.51269 10 +1	2.38914 10 +0	2.00000 10 +2	-4.22898 10 -2
0.00000	-5.83444 10 -2	3.48742 10 -6	1.23965 10 -4	-4.22898 10 -2	1.99841 10 +2

J - ERROR COVARIANCE MATRIX.

1.05480 10 -1	1.17148 10 +1	1.13446 10 +0	9.76894 10 -2	1.87019 10 +0	-3.39546 10 -
1.17148 10 +1	4.02659 10 +3	3.80240 10 +2	2.65551 10 +1	9.30139 10 +2	-1.73783 10 +
1.13446 10 +0	3.80240 10 +2	3.93192 10 +1	2.85313 10 +0	8.15932 10 +1	-1.51785 10 +
9.76894 10 -2	2.65551 10 +1	2.85313 10 +0	2.15394 10 -1	5.41414 10 +0	-1.00346 10 -
1.87019 10 +0	9.30139 10 +2	8.15932 10 +1	5.41414 10 +0	2.60118 10 +2	-4.90187 10 +
-3.39546 10 -2	-1.73787 10 +1	-1.51785 10 +0	-1.00346 10 -1	-4.90187 10 +0	5.87248 10 +

FIGURE 41. COVARIANCE MATRICES.

6.2.5. - ESTIMATION OF STATE VECTOR WITH WHITE NOISE DISTURBANCE CONFIGURATION NOISE INPUT

Firstly, one must decide upon which of the state variables are measurable. Let it be assumed that there is negligible measurement noise and that the elevator angle, pitch rate and pitch are measurable. The reference system's variables are uncontaminated by noise, and therefore will be accurately measurable and have zero auto and cross-covariance elements associated with them. The output matrix equation can therefore be considered as

$$\begin{bmatrix} y_1(jT) \\ y_2(jT) \\ y_3(jT) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} x(jT) \qquad (6.6)$$

The best linear control policy will be considered where one does not permit the fictitious variables, used in the white noise filter, to be included in the control law. The corresponding feedback coefficient matrix will have the identical elements associated with the other state variables, as they had in the case of the deterministic system. This is due to the control input and the other state variables having no influence on the response of the noise filter.

Although the white noise input has a theoretically infinite variance, consider the approximate case where the variance is 10^{+3}. The corresponding discrete disturbance covariance noise matrix was calculated and is defined in Figure (41). Note that the elements associated with

the reference system are not shown, since these are all zero. The zero values associated with the elevator angle, i.e. $x_1(t)$, are due to the noise input not influencing the elevator angle.

If one attempted to use the output matrix of equation (6.6) and the above discrete covariance matrix in the determination of the best estimate of the state vector, one finds that the inverse appearing in equation (5.24) does not exist, and therefore the method breaks down. To overcome this, one must assume that the knowledge of the elevator angle is not used in the estimation. The output equation then becomes

$$\begin{bmatrix} y_2(jT) \\ y_3(jT) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

In this case, the error covariance matrix converges after 20 iterations to have a value defined in Figure (41).

The values of the elements of the error covariance matrix associated with the fictitious white noise filter variables are naturally large, but since these are not included in the control policy, this is unimportant. The error covariance associated with the fictitious state aircraft variable $x_2(t)$ is also large, i.e. $4.02658_{10}+3$, since the correlated noise input is considered additive to the actuator output. The auto-covariance elements associated with the other state variables are acceptably low.

It has already been shown, when considering the best linear control system of the deterministic aircraft, that the fictitious state variable $x_2(t)$

plays an important part in the optimal control policy. Since the optimal state estimation scheme results in an unacceptibly high error auto-covariance for $x_2(t)$, one has not satisfactorily overcome the lack of knowledge about $x_2(t)$. As the author indicated in Chapter 5, such difficulties indicate the need for an optimal control theory where, for linear systems, optimal transfer functions at specified points in the system's structure could be determined, rather than complete state feedback. The best linear control policy is the first step in such a theory based on modern control techniques. Its use in the above example is advantageous in the case of removing the fictitious variables associated with white noise filter, but when one attempts to remove the aircraft's fictitious variable, $x_2(t)$, one must reduce the sampling period to almost zero or sacrifice a small performance index (cf Figure (35)).

CHAPTER 7 - CONCLUSIONS

Firstly, the author would like to point out that a research project such as the one reported in this thesis requires an exceptionally large amount of computer programming to be carried out. The author was not fortunate enough to join an established research group, which was interested in digital simulation and control, nor was he associated with an active computer department. As a result, the author has had to build up a computer library from scratch. This took some considerable time and effort, particularly as the hardware and software available for the KDF 9 computer, which was installed in 1963, was always found to be one stage behind the requirements of the author. Originally the generalised simulation program was written in the author's algol matrix scheme but, due to the better storage facilities afforded by the official English Electric matrix scheme (55), which became available with compiler errors in late 1965, the simulation program was rewritten in this language. One of the major bugbears of the computational work was the small storage space, 8192 words, which was available for storing the compiled programs. As a result, the final optimal control-estimation program had to be split into 6 parts, each of approximately 8,000 words, which automatically fed from one into another.

In the remainder of this chapter, the author sums up and draws some conclusions from the work reported in this thesis.

1. It has been possible to extend the class of systems, to which the

dynamic programming digital control technique is applicable, to include systems whose dynamics are described by difference differential equations, subject to the condition that all the delays appearing in the equations are multiples of some basic delay. The approximate method developed necessitates the use of a large, theoretically infinite, state vector. The latter may in certain circumstances, such as unstable uncontrolled systems, prove to be so large as to forbid the solution of the optimisation problem by this method.

2. A discrete intersample quadratic performance criterion has been developed, which can be made equivalent to the continuous integral criterion, and thus is more acceptable from conventional design standards than the quadratic summation criterion. The corresponding digital optimisation technique was devised and shown to be a more general case of the quadratic summation criterion technique. A further generalisation of the optimisation technique was made possible by the development of the sub-optimal best linear control policy, in which case one could specify which state variables were allowed to appear in the state feedback policy.

The theoretical convergence properties of the various unaugmented iterative equations used in the calculation of the optimal policy were examined with the aid of the partial controllability property of the system, the latter being introduced in easily understandable terms. The unaugmented optimal solution was shown to be convergent if the sufficient partial controllability

condition was obeyed. In the case of augmented state systems, with non-zero mean noise inputs and difference performance criterion, it was found necessary to solve the optimisation problem by a direct method, which was then shown to be equivalent to an augmented state solution, before the corresponding convergence conditions could be determined. These conditiitions turned out to be that the corresponding unaugmented state system was partially controllable, the controlled system was stable and the performance criterion was compatible with the system's dynamics.

An approximate computational stability analysis of the augmented E.D.R. matrix iteration indicated that the probability of computational instability was greatly increased if the uncontrolled system was unstable. An alternative method of generating the E.D.R. matrix sequence was developed, which has since proved to be most successful in eradicating computational instability.

The stability and performance criterion compatibility characteristics of the uncontrolled system have thus been shown to affect the generation of the optimal control policy, although the digital optimisation technique would seem to indicate, on the surface, that the control policy could be generated independent of these properties.

3. In the analysis of the sensitivity of the optimal system's response to parameter changes, it was shown that two of the proposed sensitivity functions were automatically minimised as a by-product of the digital optimisation procedure. A method of minimising a third sensitivity function

was developed, but upon its application, it was noted that the performance index of the insensitive system was, in all the cases analysed by the author, greater than the corresponding index for the perturbed sensitive system. The sensitivity constrained system is therefore practically unacceptable.

4. The sub-optimal best linear control policy was developed as a first attempt at overcoming the problem of not being able to measure the complete state vector in deterministic systems. The resulting increase in the performance index, compared with that for the optimal policy, is found in many cases to be excessively large, and thus the best linear control policy is not a panacea to the measurement problem.

For stochastic systems, the convergence properties of the Kalman filter were examined in simple terms by considering a duality property that exists between the estimator and optimal controller for a quadratic summation criterion. The Kalman filter estimator was then developed to include measurement systems with multiple state delays, which commonly occur in process control systems. By considering the generalised inverse solution of the equations used in the Kalman filter, a time invariant sub-optimal estimator was proposed which, at least in the example considered, gave reasonable results. The estimators proposed above are practical in the case of independent additive noise inputs but, for a white noise system configuration, which allows a larger class of systems

to be considered, the resulting covariance error matrix can have very large valued elements and the estimate of the state vector is correspondingly poor. Again, the measurement problem has not been conclusively solved.

It is thought that it would therefore be advantageous to develop an optimisation technique based on modern control theory, which would determine optimal transfer functions at specific points in the system, making full use of the measurable state variables.

5. Finally, the empirical relationship between the sampling period and the predominant stable system's time constant, which has been used in classical sampled-data theory for the determination of a suitable sampling period, is particularly obvious in the case of optimal systems. An approximate ratio of 1/4 has been found to hold but, in the case when sub-optimal policies, such as the best linear control policy, are used, the ratio must be reduced.

APPENDIX A - THE PROGRAMMED SOLUTION OF THE GENERAL DIFFERENCE-DIFFERENTIAL EQUATION (2.11).

As stated in Chapter 2, the matrix functions appearing in the solution of the general difference-differential equation (cf equations (2.11) and (2.43)) are ideal for computer handling. Five ALGOL procedures (54), which determine the various matrices appearing in equation (2.43) are detailed below. These procedures are slightly simplified versions of those actually used, the latter being considerably longer, due to various tolerance conditions being included to save computer time. Since the matrix infinite series in the solution were shown to be convergent, but not necessarily elementally uniformly convergent, some care must be exercised in the application of tolerances. For example, if a tolerance condition is placed on ϕ_i, then it is necessary when applying the convergence test to consider groups of ϕ_i, rather than a single matrix. The group must contain at least as many consecutive ϕ_i as there are continuous transition matrices A_i. Similar remarks also hold for the other matrices determined by the procedures.

Wherever possible, the notation used in the derivation of the solution of the difference-differential equation has been carried over to the ALGOL procedures, but to clarify matters, a pair grouping is given below. In the case of array variables, only those parameters corresponding to the subscripts are shown, the other parameters corresponding to the row and column of the matrix elements are left blank.

Procedure	Result in notation of equation (2.43)	Procedure Notation
FIJ	F_{ij}	F[-, -, i, j]
PHII	ϕ_i	PHI[-, -, i]
S12J	$S_{1j}(D_k)$	S1[-, -, j, k]
	$S_{2j}(D_k)$	S2[-, -, j, k]
GIJ	G_{ik}	G[-, -, i, k]
GI	G_i	Gs[-, -, i]

The other general variable pair groupings are

Procedure Notation	Analysis Notation
A [-, -, i]	A_i
D [-, -, i]	D_i
plant	dimension of continuous transition matrix
control	no. of control variables
statedelay	r_1 of equation (2.11)
cndelay	r_2 of equation (2.11)
period	sampling period
delay	δ
interval	interval of time over which one wishes to determine the solution
uplimit1	upper limit on i in F_{ij}
uplimit2	upper limit on j in F_{ij}
ftol	tolerance condition information storage delay
tol1	convergence tolerance measure

The value of the variable interval is equal to the sampling period if one wishes to determine the system variables at one sampling period from those of past sampling periods. In the state equation (cf Figure 2) the state at intermediate time intervals is also required, thus substituting a suitable value for interval allows the corresponding matrices to be determined.

There now follows the five procedures.

```
rocedure FIJ(plant, statedelay, interval, i, uplimit1,
             uplimit2, tol1, ftol, A, F);
eal interval, tol1;
nteger plant, statedelay, i, uplimit1, uplimit2;
eal array A, F;
nteger array ftol;

    begin
    integer j, k, l, m, n;
    real array APHI[1: plant, 1: plant];

    if i = 0 then
          begin for j:= 0 step 1 until uplimit1 do
          for k:= 0 step 1 until uplimit2+1 do
          if j = 0 and k = 0 then ftol[j, k]:= 1 else ftol[j, k]:= 0;
          for j:= 1 step 1 until plant do
          for k:= 1 step 1 until plant do
          if j = k then F[j, k, 0, 0]:= 1 else F[j, k, 0, 0]:= 0;
          end else
          begin for j:= 1 step 1 until plant do
          for k:= 1 step 1 until plant do F [j, k, i, 0]:= 0;
          end;

    for j:= 0 step 1 until uplimit2 do
          begin for k:= 1 step 1 until plant do
          for l:= 1 step 1 until plant do F [k, l, i, j +1]:= 0;

          for k:= 0 step 1 until statedelay do
                begin if i-k < 0 then goto ZEROUT;
                if ftol[i-k, j] = 0 then goto R1;

                for l:= 1 step 1 until plant do
                for m:= 1 step 1 until plant do
                      begin APHI[l, m]:= 0;
                      for n:= 1 step 1 until plant do
                      APHI[l, m]:= APHI[l, m]+A[l, n, k]×
                                   F[n, m, i-k, j];
                      APHI[l, m]:= APHI[l, m]×interval/(j+1);
                      F[l, m, i, j+1]:= APHI[l, m]+F[l, m, i, j+1];
                      end;
                R1:
                end;

          ZEROUT:
          for k:= 1 step 1 until plant do
          for l:= 1 step 1 until plant do
          if abs (F[k, l, i, j+1]) > tol1 then ftol[i, j+1]:= 1;
          end;

    end;
```

```
rocedure PHII(plant, statedelay, interval, uplimit1,
              uplimit2, tol1, ftol, F, PHI);
eal interval, tol1;
nteger plant, statedelay, uplimit1, uplimit2;
eal array F, PHI;
nteger array ftol;

     begin
     integer i, j, k, l;

     for i:= 0 step 1 until uplimit1 do
          begin FIJ(plant, statedelay, interval, i, uplimit1,
                    uplimit2, tol1, ftol, A, F);
          for j:= 1 step 1 until plant do
          for k:= 1 step 1 until plant do PHI[j, k, i]:= 0;

          for j:= 0 step 1 until uplimit2 do
               begin if ftol[i, j] = 0 then goto R2;
               for k:= 1 step 1 until plant do
               for l:= 1 step 1 until plant do
               PHI[k, l, i]:= PHI[k, l, i]+F[k, l, i, j];
          R2:end;
          end;

     end;


rocedure S12J(plant, control, statedelay, cndelay, interval,
              period, delay, uplimit1, uplimit2,
              tol1, ftol, A, D, F, S1, S2);
eal interval, period, delay, tol1;
nteger plant, control, statedelay, cndelay, uplimit1, uplimit2;
eal array A, D, F, S1, S2;
nteger array ftol;

     begin
     real tau, sigma;
     integer i, j, k, l, m, n;
     real array AG[1: plant, 1: control];

     comment preliminary partial calculation of S2j;
     for i:= 0 step 1 until uplimit1 do
     for j:= 0 step 1 until cndelay do
          begin for k:= 1 step 1 until plant do
          for l:= 1 step 1 until control do S2[k, l, i, j]:= 0;
          for k:= 0 step 1 until uplimit2 do
               begin if ftol[i, k] = 0 then goto R3;
               for l:= 1 step 1 until plant do
               for m:= 1 step 1 until control do
```

```
                    begin AG[1, m]:= 0;
                    for n:= 1 step 1 until plant do
                    AG[1, m]:= AG[1, m]+F[1, n, i, k]×D[n, m, j];
                    AG[1, m]:= AG[1, m]×interval/(k+1);
                    S2[1, m, i, j]:= AG[1, m] + S2[1, m, i, j];
                    end;
          R3:end;
      end;

comment calculation of S1j if interval > sigmai;
for i:= 0 step 1 until uplimit1 do
      begin tau:= period×entier(i×delay/period);
      sigma:= i×delay-tau;
      if sigma > interval then
            begin for j:= 0 step 1 until cndelay do
            for l:= 1 step 1 until plant do
            for m:= 1 step 1 until control do
            S1[l, m, i, j]:= 0;
            goto R5;
            end;

      for j:= 0 step 1 until i do FIJ(plant, statedelay,
            interval-sigma, j, i, uplimit2, tol1, ftol, A, F);
      for j:= 0 step 1 until cndelay do
            begin for k:= 1 step 1 until plant do
            for l:= 1 step 1 until control do S1[k, l, i, j]:= 0;
            for k:= 0 step 1 until uplimit2 do
                  begin if ftol[i, k] = 0 then goto R4;
                  for l:= 1 step 1 until plant do
                  for m:= 1 step 1 until control do
                        begin AG[l, m]:= 0;
                        for n:= 1 step 1 until plant do
                        AG[l, m]:= AG[l, m]+F[l, n, i, k]×
                                    D[n, m, j];
                        AG[l, m]:= AG[l, m]×
                                    (interval-sigma)/(k+1);
                        S1[l, m, i, j]:= AG[l, m] +
                                          S1[l, m, i, j];
                        end;
                  R4:end;
            end;
   R5:end;

comment completion of calculations of S2j;
for i:= 0 step 1 until uplimit1 do
for j:= 0 step 1 until cndelay do
for k:= 1 step 1 until plant do
for l:= 1 step 1 until control do
S2[k, l, i, j]:= S2[k, l, i, j]-S1[k, l, i, j];

end;
```

```
procedure GIJ(plant, control, cndelay, period, delay, uplimit1,
              S1, S2, G);
real period, delay;
integer plant, control, cndelay, uplimit1;
real array S1, S2, G;

      begin
      integer i, j, k, l, m;
      integer array start[0: cndelay];

      for i:= 0 step 1 until entier (uplimit1×delay/period+1) do
      for j:= 0 step 1 until cndelay do
            begin  for k:= 1 step 1 until plant do
            for l:= 1 step 1 until control do G[k, l, i, j]:= 0;

            if i = 0 then start[j]:= j-1;
            k:= start[j];
            for k:= k+1 while entier (k×delay/period) = i do
                  begin for l:= 1 step 1 until plant do
                  for m:= 1 step 1 until control do
                  G[l, m, i, j]:= G[l, m, i, j]+
                  (if k-j < 0 or k-j > uplimit1 then 0
                   else S1[l, m, k-j, j])+
                  (if k-j-1 < 0 or k-j-1 > uplimit1 then 0
                   else S2[l, m, k-j-1, j]);
                  start[j]:= k;
                  end;
            end;

      end;


procedure GI(plant, control, cndelay, period, delay, uplimit1, G, Gs);
real period, delay;
integer plant, control, cndelay, uplimit1;
real array G, Gs;

      begin
      integer i, j, k, l;

      for i:= 0 step 1 until entier(uplimit1×delay/period+1) do
            begin for j:= 1 step 1 until plant do
            for k:= 1 step 1 until control do Gs[j, k, i]:= 0;

            for j:= 0 step 1 until cndelay do
            for k:= 1 step 1 until plant do
            for l:= 1 step 1until control do
            Gs[k, l, i]:= Gs[k, l, i]+G[k, l, i, j];
            end;

      end;
```

APPENDIX B - TWO PARTICULAR OPTIMAL SOLUTIONS

In this appendix, the two particular problems discussed in section (2.3) are solved, using a similar dynamic programming formulation and terminology as that used in section (2.2).

A.1. - NON-ZERO MEAN ADDITIVE NOISE

The optimal control problem will be solved for a system described by equation (2.59) with performance criterion defined by equation (2.45).

The equation (2.59) has the difference equation solution

$$x(j+1) = \phi x(j) + G\, m(j) + r_1(j) + r_2(j) \qquad (A.1)$$

where ϕ and G are defined in section (2.1)

$$r_1(j) = \int_{jT}^{\overline{j+1}\,T} \phi(\overline{j+1}\,T - q)\, K\, u_1(q)\, dq$$

and $$r_2(j) = \int_{jT}^{\overline{j+1}\,T} \phi(\overline{j+1}\,T - q)\, u_2(q)\, dq$$

The principle of optimality allows the multistage decision process to be redefined as a series of single stage decision processes, namely,

$$f_{N-j}\left[x(j)\right] = \min_{m(j)} E \left\{ x'(j) Q\, x(j) + m'(j)\, H\, m(j) + f_{N-\overline{j+1}}\left[x(j+1)\right] \right\} \qquad (A.2)$$

Let it be assumed that

$$f_{N-j}\left[x(j)\right] = x'(j)\, P(N-j)\, x(j) + 2x'(j)\, V(N-j)\, r_1(j) + r_1'(j)\, R(N-j)\, r_1(j) + Z(N-j) \qquad (A.3)$$

where $P(N-j)$ is a symmetric, as yet unknown matrix,

and $V(N-j)$ and $Z(N-j)$ are, as yet, unknown matrices.

The procedure is now to substitute the expression for $f_{N-j+1}\left[x(j+1)\right]$ from equation (A.3) and thence expression for $x(j+1)$ and $r_1(j+1)$ into equation (A.2).

So far, no expression has been developed or defined, which determines $r_1(j+1)$ from a knowledge of $r_1(j)$. In the case of random disturbances, one can usually make the assumption that the statistics are stationary and thus $r_1(j+1) = r_1(j)$. It was indicated in Chapter 2 that equation of type (A.1) also describes a system with deterministic inputs $u_1(t)$, a complete knowledge of whose dynamic characteristics is unlikely in a practical system. In order to proceed further with the solution, it is necessary to make some reasonable assumption about the dynamics. Taking a lead from the stationary statistics case, it will be assumed that $u_1(t)$ is constant over the time interval $0 \leq t \leq NT$, where N is number of stages in optimisation problem and T is the sampling period.

Equation (A.2) then becomes

$$f_{N-j}[x(j)] = \underset{m(j)}{\text{Min}}\, E \Big\{ x'(j)\,[Q + \phi' P(N-\overline{j+1})\,\phi]\,x(j)$$

$$+ 2\,x'(j)[\phi' P(N-\overline{j+1})G]\,m(j)$$

$$+ m'(j)\,[H + G' P(N-\overline{j+1})G]\,m(j)$$

$$+ 2\,x'(j)\,[\phi' P(N-\overline{j+1}) + \phi' V(N-\overline{j+1})]\,r_1(j)$$

$$+ 2\,m'(j)\,[G' P(N-\overline{j+1}) + G' V(N-\overline{j+1})]\,r_1(j)$$

$$+ r_1'(j)\,[R(N-\overline{j+1}) + P(N-\overline{j+1}) + 2V(N-\overline{j+1})]\,r_1(j)$$

$$+ r_2'(j)\,P(N-\overline{j+1})\,r_2(j) + Z(N-\overline{j+1}) \Big\} \qquad \text{(A.4)}$$

assuming that $u_2(t)$ is an independent random noise input.

To find minimum, the first variation of $f_{N-j}[x(j)]$ is made with respect to $m(j)$ and then equated to zero, giving,

$$m(j) = -[H + G' P(N-\overline{j+1})G]^{-1} \times$$

$$\Big\{ G' P(N-\overline{j+1})\phi\, x(j) + [G' P(N-\overline{j+1}) + G' V(N-\overline{j+1})]\, r_1(j) \Big\} \qquad \text{(A.5)}$$

The optimal control policy is therefore a combination of a linear feedback from the state variables, plus an additional contribution due to the non-zero mean value of the noise (or deterministic) disturbance inputs.

The assumption of equation (A.3) is shown to be valid by the substitution of the value of $m(j)$ from equation (A.5) into equation (A.4), which results in

$$f_{N-j}\left[x(j)\right] = x'(j)\left\{Q + \phi' P(N-\overline{j+1})\phi - \phi' P(N-\overline{j+1})G\left[H+G' P(N-\overline{j+1})G\right]^{-1} G' P(N-\overline{j+1})\phi\right\}x(j)$$

$$+ 2x'(j)\left\{-\phi' P(N-\overline{j+1})G\left[H+G' P(N-\overline{j+1})G\right]^{-1}\left[G' P(N-\overline{j+1}) + G' V(N-\overline{j+1})\right]\right.$$

$$\left. + \phi' P(N-\overline{j+1}) + \phi' V(N-\overline{j+1})\right\} r_1(j)$$

$$+ r_1'(j)\left\{2V(N-\overline{j+1}) + R(N-\overline{j+1}) + P(N-\overline{j+1})\right.$$

$$- \left[G' P(N-\overline{j+1}) + G' V(N-\overline{j+1})\right]'\left[H+G' P(N-\overline{j+1})G\right]^{-1} \times$$

$$\left.\left[G' P(N-\overline{j+1}) + G' V(N-\overline{j+1})\right]\right\} r_1(j)$$

$$+ E\left[r_2'(j)\, P(N-\overline{j+1})\, r_2(j)\right] + Z_1(N-\overline{j+1}) \qquad \text{(A.6)}$$

The right hand side of equation (A.6) is seen to have the assumed form; thus the assumption of equation (A.3) is valid. Since equation (A.6) holds for all $x(j)$, $r_1(j)$ and $r_2(j)$, it is possible to equate the corresponding parts of equation (A.3) and (A.6) to obtain the difference equations which are detailed in the main text of section (2.3) in equations (2.61) - (2.64). The initial conditions of these difference equations can be calculated by an exactly similar argument to that used in section (2.2), giving

$$P(0) = Q,$$

and $V(0)$, $R(0)$ and $Z(0)$ null matrices.

A.2. - THE DIFFERENCE CRITERION

Consider the system described by equation (2.70) with performance

criterion defined by equation (2.71). The difference equation solution of equation (2.70) is

$$x(j+1) = \phi\, x(j) + G\, m(j) \tag{A.7}$$

Once again the multistage decision process can be arranged as a series of single stage decision processes, i.e.

$$f_{N-j}\ x(j) = \underset{m(j)}{\text{Min}} \left\{ [x(j) - y]' Q\, [x(j) - y] + m'(j) H m(j) + f_{N-j+1}\, [x(j+1)] \right\} \tag{A.8}$$

Let it be assumed that

$$f_{N-j}\, [x(j)] = x'(j)\, P(N-j)\, x(j) + 2x'(j)\, V(N-j) y + y\, R(N-j) y \tag{A.9}$$

where $P(N-j)$ is a symmetric, as yet unknown matrix and $V(N-j)$ and $R(N-j)$ are, as yet, unknown matrices.

Substituting expression for $f_{N-j+1}[x(j+1)]$ from equation (A.9) and thence expression for $x(j+1)$ from equation (A.7), one obtains

$$\begin{aligned} f_{N-j}[x(j)] = \underset{m(j)}{\text{Min}} \Big\{ & x'(j) \left[Q + \phi' P(N-j+1) \phi \right] x(j) \\ & + 2x'(j) \left[\phi' P(N-j+1) G \right] m(j) \\ & + m'(j) \left[H + G' P(N-j+1) G \right] m(j) \\ & + 2x'(j) \left[-Q + \phi' V(N-j+1) \right] y \\ & + 2m'(j) \left[G' V(N-j+1) \right] y \end{aligned}$$

$$+ y'R(N-\overline{j+1})y \Big\} \qquad \text{(A.10)}$$

Taking first variation of $f_{N-j}[x(j)]$ with respect to $m(j)$ and equating to zero gives

$$m(j) = -\left[H+G'P(N-\overline{j+1})\,G\right]^{-1}\left[G'P(N-\overline{j+1})\,\phi\,x(j) + G'V(N-\overline{j+1})y\right]$$

The assumption of equation (A.9) is proved to be valid by back substitution of $m(j)$ into equation (A.10), which gives,

$$\begin{aligned} f_{N-j}[x(j)] = {} & x'(j)\Big\{Q + P(N-\overline{j+1}) \\ & + \phi'P(N-\overline{j+1})G\left[H+G'P(N-\overline{j+1})G\right]^{-1}G'P(N-j+1)\,\phi\Big\}x(j) \\ & + 2x'(j)\Big\{-Q + \phi'V(N-j+1) \\ & - \phi'P(N-\overline{j+1})G\left[H+G'P(N-\overline{j+1})G\right]^{-1}G'V(N-\overline{j+1})\Big\}\,y \\ & + y'\Big\{Q+R(N-\overline{j+1}) \\ & - V'(N-\overline{j+1})G\left[H+G'P(N-\overline{j+1})G\right]^{-1}G'V(N-\overline{j+1})\Big\}\,y \end{aligned} \qquad \text{(A.12)}$$

The right hand side of equation (A.12) has same form as equation (A.9); thus the latter is valid. Since equation (A.12) holds for all $x(j)$ and y, the corresponding parts of equations (A.9) and (A.12) can be equated to obtain the difference equation detailed in the main text of section (2.3) in equations (2.74) - (2.76).

The initial conditions are found to be

$$P(0) = Q$$

and $V(0)$ and $R(0)$ null matrices.

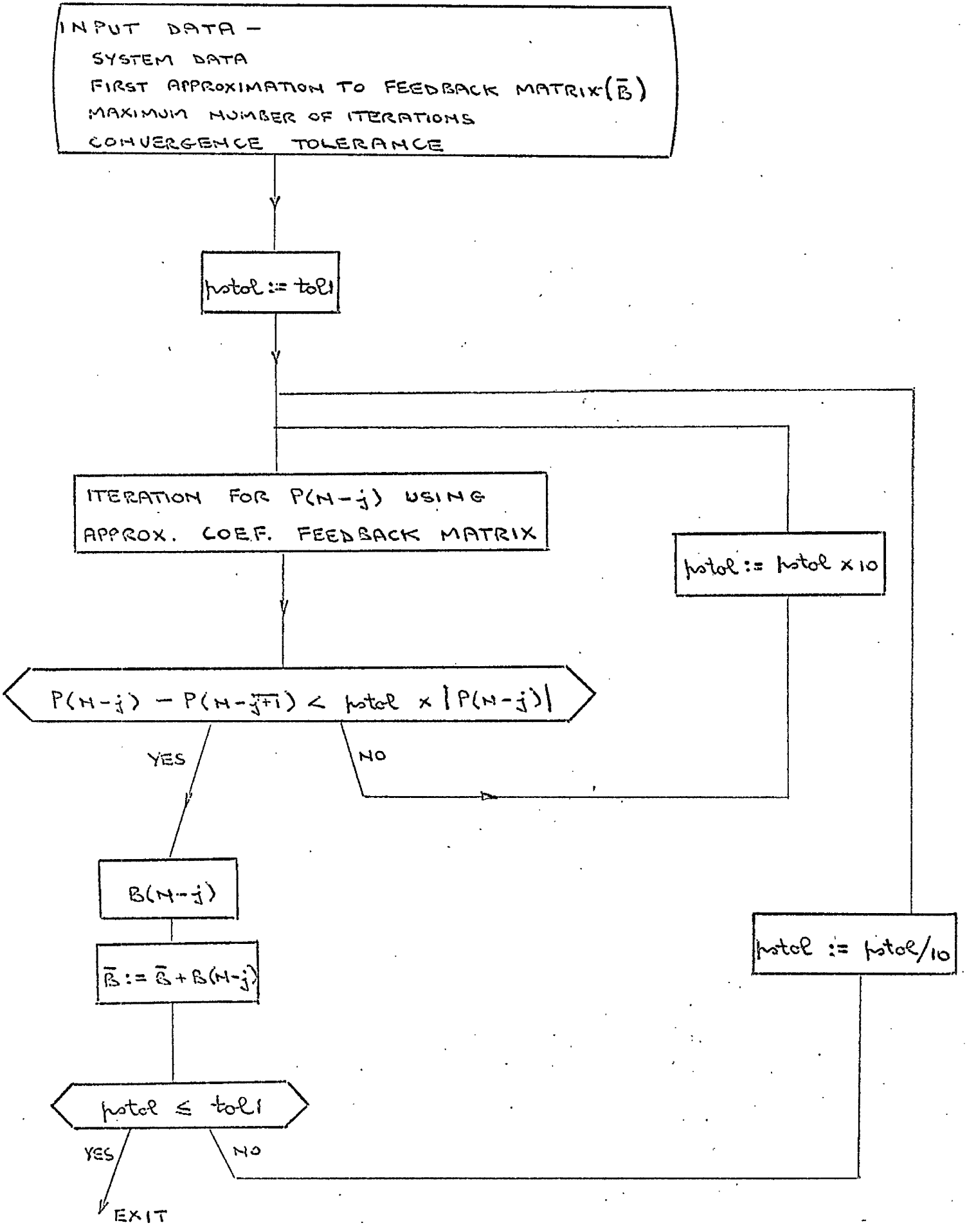


CHART 1 E.D.R. MATRIX SEQUENCE

APPENDIX C - PROCEDURE FOR CALCULATION OF E.D.R. MATRIX SEQUENCE

An ALGOL procedure (54), written in the ALGOL matrix scheme language (55), is detailed below for calculating the E.D.R. matrix sequence. The procedure is applicable to completely augmented state systems with either the quadratic summation or the weddle integral criterion of Chapter 2. It includes provision for the calculation of either the optimal or the best linear control policy, and also illustrates the progressive computational stability technique discussed in Chapter 3.

A flow chart of the progressive stability iteration is illustrated in Chart 1 opposite. The iteration for the steady state value of $P(N-j)$ is terminated when the difference between the elements of the matrices on successive iterations is less than some tolerance, toll, times the absolute value of the element in question. Computational instability in general results in a divergent sequence of matrices; thus the convergence test also acts as an approximate stability test. If convergence with initial tolerance, toll, is not achieved due to computational instability, the tolerance is increased until convergence is achieved, when an approximate value of the coefficient feedback matrix is calculated. The latter is added to the previous value of the coefficient feedback matrix, $\overline{B}$, and the tolerance is reduced by a factor of ten. The process is then repeated until the iteration for the E.D.R. matrix and consequently the feedback coefficient matrix

converges with respect to the initial tolerance, toll.

By the frequent use of comments, the operation of each section of the program is self-explanatory. The notation of the analysis of the optimisation problem has been carried through to the procedure wherever possible, with the following pair groupings.

Procedure Notation	Analysis Notation
PHI	$\phi(T)$
G	G(T)
intersample	integer subdivision of sampling period for weddle criterion thus must be multiple of 6.
INTERPHI	ϕ(T/intersample)
INTERG	G(T/intersample)
SELECTB	matrix with unit on diagonal if corresponding state variable to be included in control input calculation.
plant	dimension of state vector
control	dimension of control vector
weddle	Boolean variable, true if integral criterion, false if quadratic summation criterion.
nop	maximum number of iterations permitted in calculation of converged value of P(N-j).
tol3	convergence tolerance condition for E.D.R. matrix sequence.
P	P(N-j)
B	B(N-j)

```
rocedure EDR(PHI, G, INTERPHI, INTERG, Q, H, SELECTB, plant,
             control,weddle, intersample, nop, tol3, P, B);
alue PHI, G, INTERPHI, INTERG, SELECTB, plant, control, weddle,
     intersample, nop, tol3;
eal tol3;
nteger plant, control, intersample, nop,
       PHI, G, INTERPHI, INTERG, Q, H, SELECTB, P, B;
oolean weddle;

    begin
    real pstol;
    integer i, j, k, l, f1,
            INTERQ, INTERH, INTERGPHI,
            PSPHI, PSINTERPHI, HGPG, GPHI, HIPHI, PLPL, PLPL3, PLCN;

    pstol:= tol3; f1:= format([5sndddc]);

    comment preparation of cost matrices if integral criterion;
    if weddle then
         begin call(Q);  mult by(period/intersample);  equals(Q);
         call(H);  multby (period);  equals(H);
         end;

    comment calculation of P(0) for quadratic summation criterion;
    if not weddle then
         begin call (Q); dup;
         equals(PLPL3); equals (P);
         end;

NW2:null(control, plant);  equals (INTERGPHI);
    null(control, control);  equals(INTERH);
    null(plant, plant);  equals (INTERQ);

    comment calculation of preliminary transition matrix;
    call(PHI); call(G); call(B);
    mult; add; equals(PSPHI);

    comment calculation of additional terms if integral criterion;
    if not weddle then goto NW4;
    unit (plant); equals (PSINTERPHI);
    unit (plant); equals(PLPL);
    null (plant, control); equals(PLCN);

    for i:= 1 step 1 until intersample/6+0.5 do
    for j:= 0 step 1 until 6 do
         begin if j = 0 or j = 2 or j = 4 or j = 6 then k:= 1;
         if j = 1 or j = 5 then k:= 5;
         if j = 3 then k:= 6;
         if j = 0 then goto S1;
```

```
      call (PSINTERPHI); call (INTERPHI); mult;
      call (PSINTERPHI); call (INTERG); mult;
      call (PLCN); add; delete(PLCN);
      equals(PLCN); dup; delete(PSINTERPHI); equals(PSINTERPHI);
      call (PLCN); call (B); mult; add;
      delete(PLPL); equals (PLPL);

      S1:
      call (PLPL); trans; call (Q); call (PLPL); mult; mult;
      mult by (k×3/10); call (INTERQ); add;
      delete(INTERQ); equals(INTERQ);

      call (PLCN); trans; call (Q); call (PLCN); mult; mult;
      mult by (k×3/10); call(INTERH); add;
      delete(INTERH); equals(INTERH);

      call (PLCN); trans; call (Q); call (PLPL); mult; mult;
      mult by (k×3/10); call (INTERGPHI); add;
      delete(INTERGPHI); equals (INTERGPHI);

    end;
delete(PLPL); delete(PLCN); delete(PSINTERPHI);

comment calculation of additional computational stability terms
        for integral index;
call(INTERQ); call(B); trans; call(H); call(B);
mult; mult; add; delete(INTERQ); equals(INTERQ);

call(INTERGPHI); call(H); call(B);
mult; add; delete(INTERGPHI); equals(INTERGPHI);

call (INTERQ); call (Q); subtract; dup;
delete(INTERQ); equals (INTERQ);
call (Q); add; delete(Q); equals (Q);
call (H); call (INTERH); add; delete(H); equals (H);

comment calculation of P(0) for integral criterion;
if abs(pstol - tol3) > 10-3 × tol3 then goto NW5;
call (Q);
call (INTERGPHI); trans; call (H); invert;
mult; call (INTERGPHI); mult;
call (SELECTB); mult; dup; trans; dup; call (SELECTB); mult;
subtract; add; subtract; dup;
delete(PLPL3); delete(P); equals(PLPL3); equals (P);
goto NW5;
```

```
comment calculation of additional computational stability terms
        for quadratic summation criterion;
NW4:
call(Q); call(B); trans; call(H); call(B);
mult; mult; dup; delete(INTERQ); equals(INTERQ); add;
delete(Q); equals(Q);

call(H); call(B); mult; delete(INTERGPHI); equals(INTERGPHI);

comment calculation of EDR matrix sequence;
NW5:
call(PLPL3); equals(P);
for i:= 1 step 1 until nop do
     begin
     if i = nop then
          begin write text (70, [[2c5s]
                    NO * CONVERGENCE * OF * P[5s] ITERATION ]);
          write (70, f1, i);
          write matrix (70, 6, P, [P]);
          write matrix (70, 6, PLPL, [DIFFERENCE]);
          write text (70, [[5s]TOLERANCE]);
          write (70, f2, pstol);
          pstol:= pstol × 10;
          delete(P); delete(PLPL);
          goto NW5;
          end;

     comment calculation of intermediate matrices;
     call (G); trans; call(P); call(G); mult; mult;
     call(H); add; invert; equals (HGPG);

     call (G); trans; call (P); call (PSPHI); mult;
     mult; call (INTERGPHI); add; equals (GPHI);

     call (PSPHI); trans; call(P); call (PSPHI); mult;
     mult; equals (HIPHI);

     comment actual calculation of P(N-j);
     call(Q); call(HIPHI); add;
     call (GPHI); trans;
     call(HGPG); call(GPHI); mult; mult;
     call(SELECTB); mult; dup; trans; dup; call(SELECTB); mult;
     subtract; add; subtract;
     delete(P); equals (P);

     for j:= 1 step 1 until plant do
     for l:= 1 step 1 until plant do
     if abs(element(j, l, PLPL)) > pstol ×
       (if abs(element(j, l, P)) < 10-4 then 10-4 else
        abs(element(j, l, P))) then goto I3;
```

```
      write matrix (70, 6, P, [P]);
      write text (70, [[5s] NO. * OF * ITERATIONS]);
      write (70, f1, i);
      write matrix (70, 6, PLPL, [DIFFERENCE]);
      write text (70, [[5s]TOLERANCE]);
      write (70, f2, pstol);
      goto L1;

      I3:
      delete(HGPG); delete(GPHI); delete(HIPHI);
      end;

comment reasignment of state and control cost matrices;
L1:
call (Q); call (INTERQ); subtract; delete(Q); equals (Q);
call (H); call (INTERH); subtract; delete(H); equals (H);

comment calculation of B and test for tolerance;
call(HGPG); negate; call(GPHI); call(SELECTB);
mult; mult; call(B); add; delete(B); equals(B);
write matrix (70, 6, B, [B]);

delete(HGPG); delete(GPHI); delete(HIPHI);

if pstol - tol3 > 10-3 × tol3 then
      begin pstol:= pstol/10;
      goto NW2;
      end;

end;
```

APPENDIX D - COFACTOR EXPANSION

By definition, the determinant of a matrix, as a function of the cofactors of the elements of the first row is

$$|A| = \sum_{i=1}^{n} a_{1i} B_{1i}$$

where n is the dimension of matrix A with elements a_{ij} and B_{1i} is cofactor of a_{1i}

The cofactors B_{1i} are defined by

$$B_{1i} = (-1)^{1+i} \sum_{j=1}^{n-1} \bar{a}_{1j} \bar{B}_{1j}$$

where matrix $\bar{A}$, of which $\bar{a}_{1j}$ are elements of the first row with corresponding cofactors $\bar{B}_{1j}$, is the matrix A with the i th row and column of A removed, i.e. square matrix of dimension n-1.

The determination of the cofactor is thus similar to the calculation of the determinant, except that the dimension of the problem has been reduced by 1 to n-1. Proceeding in this way, as one does in manual calculation of small order matrices, one can eventually reduce the dimension of the problem to two, in which case the evaluation of the corresponding cofactors is trivial. Knowing the path by which one reduced the dimension of the problem, one can retrace one's steps and determine the cofactor for successively larger dimensions, until one finally obtains the determinant of the complete matrix A. Such a method of calculating the latter is performed by the ALGOL procedure

detailed below.

A recursive use is made of the procedure, i.e. it calls upon itself, and thus the operation of the procedure is rather involved. It should be noted that little storage space is required, since only elements of matrix A and none of the matrices of smaller dimensions made up from elements of A are required to be stored. The path taken to reduce the dimension of the problem is stored in the integer array route, which is a column vector of dimension n. The integer array selection contains the information on which elements of the matrix A are to be considered in the evaluation of the cofactors. The integer variables row and col are the subscripts of the various cofactors, whose values are stored and calculated in the form of a positive part, posdet, and a negative part, negdet, such that no two parts of the same sign are subtracted and no two parts of opposite sign are added. The real arrays posdet and negdet are matrices of dimension $(n \times n)$.

The procedure call

CODET (A, 1, 0, n, route, posdet, negdet)

with route, posdet and negdet having all zero elements prior to the call, would evaluate the determinant of the matrix A of dimension n in the form of a positive part, posdet $[1, 0]$, and a negative part, negdet $[1, 0]$.

```
procedure CODET(A, row, col, order, route, posdet, negdet);
value row, col;
integer row, col, order;
real array A, posdet, negdet;
integer array route;

    begin
    real term, dumpos;
    integer i, j, acol;
    integer array selection[1: order+1-row];

    j:= 1;
    for i:= 1 step 1 until order do if route[i] = 0 then
        begin selection[j]:= i;
        j:= j+1;
        end;

    for i:= row step 1 until order do
        begin acol:= selection[i+1-row];
        route[acol]:= 1;
        if row+1 > order then goto L1;

        CODET(A, row+1, i, order, route, posdet, negdet);

        L1:
        term:= (-1)↑(1+i-(row-1))×A[row, acol];

        if row+1 > order then
            begin posdet[row+1, i]:= 1;
            negdet[row+1, i]:= 0;
            end;
        if term ≥ 0 then
            begin posdet[row+1, i]:= posdet[row+1, i]×term;
            negdet[row+1, i]:= negdet[row+1, i]×term;
            end else
            begin dumpos:= posdet[row+1, i];
            posdet[row+1, i]:= negdet[row+1, i]×term;
            negdet[row+1, i]:= dumpos×term;
            end;

        route[acol]:= 0;
        end;

    for i:= row step 1 until order do
        begin posdet[row, col]:= posdet[row, col]+posdet[row+1, i];
        negdet[row, col]:= negdet[row, col]+negdet[row+1, i];
        posdet[row+1, i]:= 0;
        negdet[row+1, i]:= 0;
        end;
end;
```

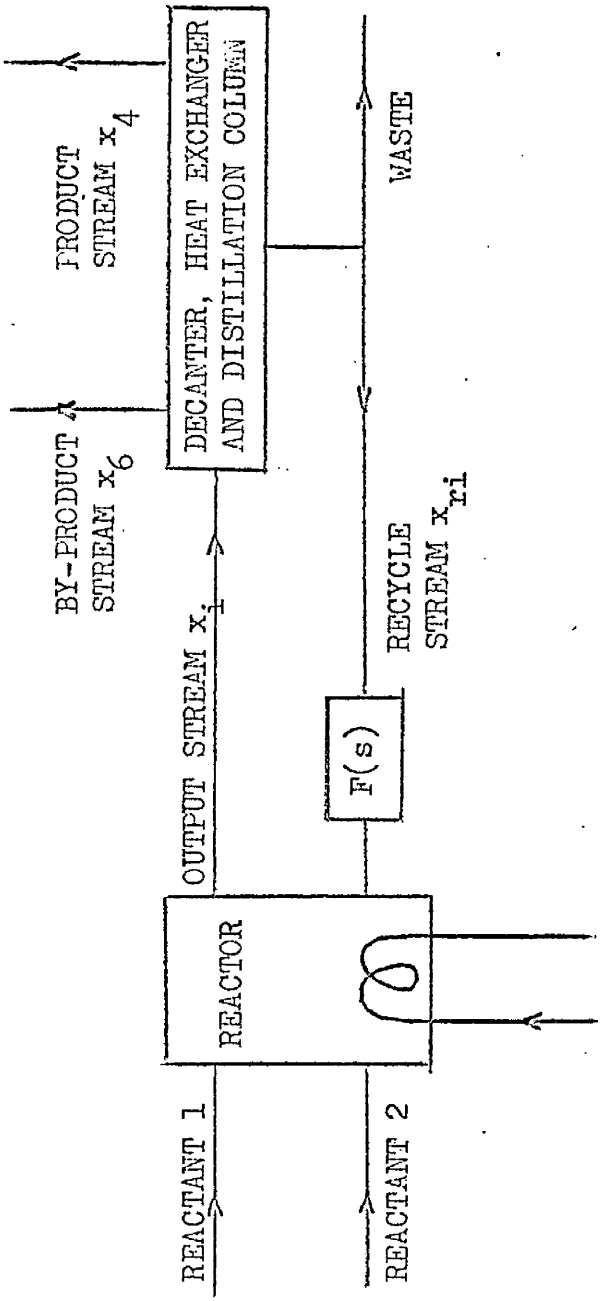


FIGURE 42 CHEMICAL REACTOR MODEL.

APPENDIX E - AN IDEALISED MODEL OF A CHEMICAL REACTOR

The chemical reactor model discussed in this appendix was proposed in a paper by Williams and Otto (56) for the express purpose of providing a model upon which comparisons could be made of the ability of various process digital computers, to solve the steady state optimisation problem. In the above mentioned paper full details, including structural details, are given of a chemical plant, which includes a reactor, heat-exchanger, decanter and distillation column. It is only proposed to use the reactor part of this model.

A schematic sketch of the operation of the reactor and associated recycle stream is given in Figure (42). The input reactant flows are assumed to be 100% pure, and only the flow of stream (2) is adjustable, i.e. can be used as a control input variable. The chemical reaction within the reactor is assumed to be three interconnected exothermic reactions, which obey the Arrhenius reaction rate equation (57). Three by-products with concentrations represented by x_3, x_5 and x_6 result from the reaction. The by-product with concentration x_6 is assumed harmful to the reaction, and therefore is decanted from the output stream before passing to other sections of the plant or entry to recycle stream. The product being manufactured has concentration x_4 and is assumed to form a 10% azeotropic mixture with the by-product x_5. All the product is distilled from the output stream, except the 10% azeotropic mixture, before being recycled. The concentrations in the recycle stream are denoted by x_r with an additional subscript denoting the reactant, product,

Variable	Physical Significance	Value at 662°
x_1	concentration of stream 1 (effluent)	0.0868
x_2	concentration of stream 2 (effluent)	0.3372
x_3	concentration of stream 3 (effluent)	0.0143
x_4	concentration of stream 4 (effluent)	0.0884
x_5	concentration of stream 5 (effluent)	0.4057
x_6	concentration of stream 6 (effluent)	0.0673
x_{r1}	concentration of recycle stream 1	0.0981
x_{r2}	concentration of recycle stream 2	0.3811
x_{r3}	concentration of recycle stream 3	0.01625
x_{r4}	concentration of recycle stream 4	0.0458
x_{r5}	concentration of recycle stream 5	0.4586
T	reactor temperature	662 °R
$T_{1,2}$	temperature of input streams 1 and 2	530 °R
T_r	recycle temperature	560 °R
T_w	input water temperature	520 °R
a_1	Arrhenius equation constants	$5.9755_{10}+9$
a_2	Arrhenius equation constants	$2.5962_{10}+12$
a_3	Arrhenius equation constants	$9.6283_{10}+15$
b_1	Arrhenius equation constants	12,000
b_2	Arrhenius equation constants	15,000
b_3	Arrhenius equation constants	20,000
h_1	heat of reaction 1	-125 BTU/hou
h_2	heat of reaction 2	-50 BTU/hour
h_3	heat of reaction 3	-143 BTU/hou
V	effective mass volume of reactor	4,640 lbs.
c	specific heat of reactor contents	0.4
U	heat transfer coefficient	5,000 BTU/R°h
f_1	flow rate of stream 1 (input)	14,500 lbs./h
f_2	flow rate of stream 2 (input)	33,350 lbs./h
f_w	water flow rate	3,671 lbs./ho
f_r	recycle flow rate	48,111 lbs./h
f	reactor effluent flow rate	95,961 lbs./h

TABLE (2)

etc.concerned. It is assumed that the reactor has ideal level control and that a fixed recycle flow is maintained.

It is a fairly straightforward matter to derive the differential equations governing the output concentrations of the reactor. Since this is adequately detailed in reference (56), only the resulting first order nonlinear equations will be listed below. The notation is given in detail in Table (2).

$$dx_1/dt = f_1/V + x_{r1}\, f_r/V - x_1 f/V - k_1 x_1 x_2$$

$$dx_2/dt = f_2/V + x_{r2}\, f_r/V - x_2 f/V - k_k x_1 x_2 - k_2 x_2 x_3$$

$$dx_3/dt = x_{r3}\, f_r/V - x_3 f/V + 2k_1 x_1 x_2 - 2k_2 x_2 x_3 - k_3 x_3 x_4$$

$$dx_4/dt = x_{r4}\, f_r/V - x_4 f/V + k_2 x_2 x_3 - 0.5\, k_3 x_3 x_4$$

$$dx_5/dt = x_{r5}\, f_r/V - x_5 f/V + 2k_2 x_2 x_3$$

$$dx_6/dt = -x_6\, f/V + 1.5\, k_3 x_3 x_4$$

where $k_i = a_i \exp(-b_i/T)$

In addition to these equations, there are the algebraic conditions governing the concentrations and flow rates of the various streams, namely,

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 1$$

$$x_{r1} + x_{r2} + x_{r3} + x_{r4} + x_{r5} = 1$$

$$x_{r4} = 0.1\, x_{r5}$$

and $f = f_1 + f_2 + f_r$.

The author will now manipulate this model into a form suitable for the digital optimisation problem.

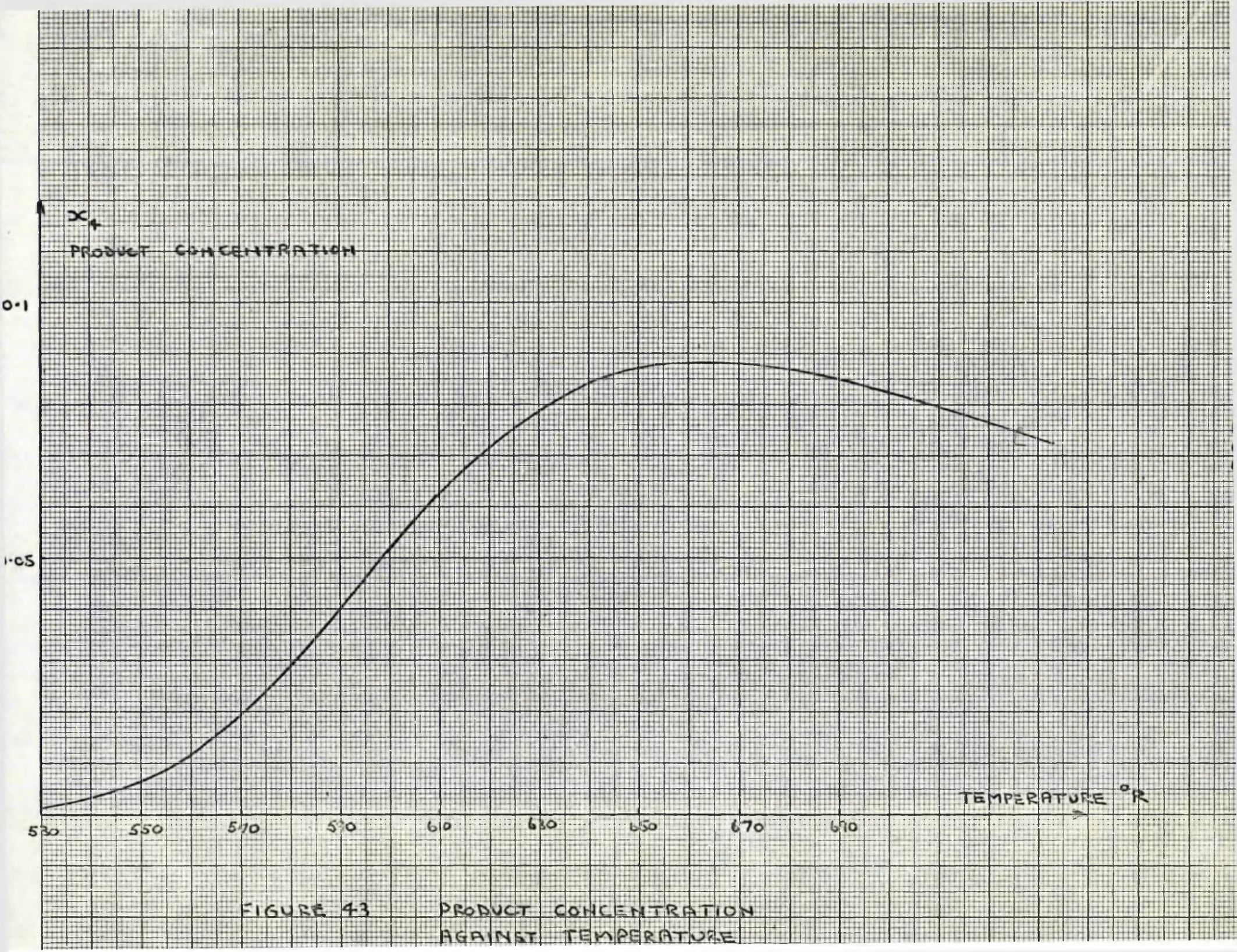


FIGURE 43 PRODUCT CONCENTRATION AGAINST TEMPERATURE

The concentrations in the recycle stream are related to the concentrations in the ouput stream, since the former is only the latter with the stream components fx_6 and $f(x_4 - 0.5x_5)$ removed by decanting and distillation respectively. The ideal recycle concentrations, assuming no dynamics in the recycle path, can therefore be determined from the output flow concentration by relationship

$$x_{ri} = x_i/(1 - x_6 - x_4 + 0.1x_5) \quad \text{for } i = 1, 2, 3, 5.$$

An algol program was written to obtain the steady state solution of the non-linear concentration equations. The solution was accomplished by expressing all the concentrations as a function of x_2 using the non-linear and algebraic equations, and thence, knowing that x_2 must lie between zero and unity, successive approximations were made for x_2 until the true solution was obtained. By this method, the graph of the steady state product concentration against temperature could be obtained and is shown in Figure (43). It is seen that the maximum concentration occurs at $662^{\circ}R$; therefore it is at this temperature that the linearisation of concentration equations was performed, since one was desirous of maintaining the output concentration at its maximum value. In table (2) the steady state values of variables are given at a reactor temperature of $662^{\circ}R$.

If there are dynamics in the recycle path, as will undoubtedly be the case, the recycle concentrations input to the reactor will be defined by

$$\bar{x}_{ri}(s) = F(s)\ x_{ri}(s)$$

where F(s) is the transfer function of the recycle dynamics.

Three cases were considered by the author

(1) No recycle dynamics, i.e. $F(s) = 1$

(2) First approximation Pade filter for pure delay

i.e. $F(s) = (1-0.5\delta s)/(1+0.5\delta s)$

and (3) Pure time delay i.e. $F(s) = \exp(-\delta s)$

It should be noted that the last type of recycle dynamics results in the system being described by difference-differential equations.

Since the reactions are all exothermic, it is necessary to have some form of cooling, this being provided in the model by passing water through coils in the reactor vessel. It is assumed that the water flow rate is adjustable, and therefore acts as the second control input to the system. By equating the rate of loss of enthalpy from the reactor vessel to the rate of gain of enthalpy of the cooling water, it is possible to develop an expression for the rate at which the enthalpy is extracted from the reactor in terms of the reactor temperature, T, and the water flow rate, f_w, namely

$$\text{Enthalpy/hr. lost} = 2Uf_w(T_w - T)/(U + 2f_w)$$

where the nomenclature is that of table (2), and it is assumed that the rate of removal of enthalpy by cooling water is U× (the mean

water temperature), where U is a heat transfer coefficient, whose value is given in table (2).

Upon making heat balance for reactor, i.e.

$$\begin{aligned} d\,[\text{reactor enthalpy}]/dt = {} & d\,[\text{input enthalpy}]/dt \\ & - d[\text{output enthalpy}]/dt \\ & + d[\text{chemical reaction enthalpy}]/dt \\ & - d[\text{enthalpy removed by cooling water}]/dt \end{aligned}$$

one can obtain the expression for the rate of change of the reactor temperature as

$$\begin{aligned} dT/dt = {} & [f_1T_1 + f_2T_2 + f_rT_r]/V - fT/V \\ & + [2k_1x_1x_2h_1/c + 3k_2x_2x_3h_2/c + 1.5k_3x_3x_4h_3/c] \\ & + 2Uf_w(T_w - T)/Vc(U + 2f_w) \end{aligned}$$

For complete details of the assumptions made in the derivation of the differential equations governing the system, the author recommends the reference (56). Since the author's main interest lies with the resulting model, he did not think it necessary to expand at length on the derivation of the model, except where the equations have been manipulated for the purposes of the digital optimisation problem.

Finally, for each of the three cases of the recycle dynamics configurations, a computer program was written to determine the linearised

equivalent of the non-linear reactor concentration and temperature equations in the linearly independent output concentrations (x_1, x_2, x_3, x_4 and x_5), the reactor temperature (T), the two control inputs (f_2 and f_w) and, for case of Pade filter system, in the three linearly independent recycle concentrations (x_{r1}, x_{r2} and x_{r3}). The linearisation was carried out by a multivariable Taylor expansion, truncated at the second term. In addition, the variables were normalised by dividing them by their steady state values, since there was a wide disparity in the numerical values of the various variables. The linearised equations were suitably altered. The results of these calculations are shown in matrix form in Figures (26), (32) and (33) and equations (6.2), (6.3) and (6.4) in the main text of Chapter 6.

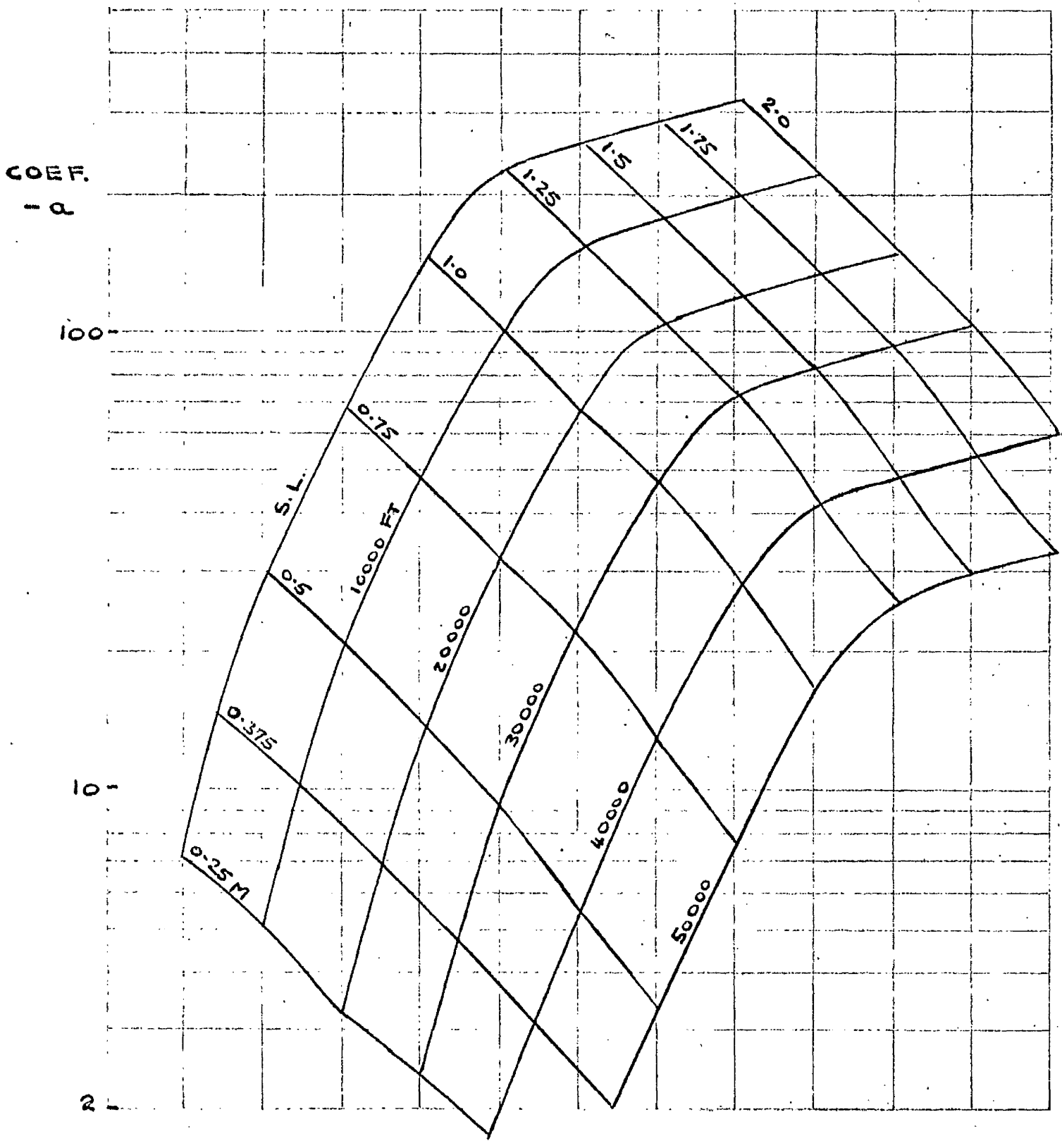


FIGURE 44

VARIATION OF "a" WITH ALTITUDE AND MACH No.

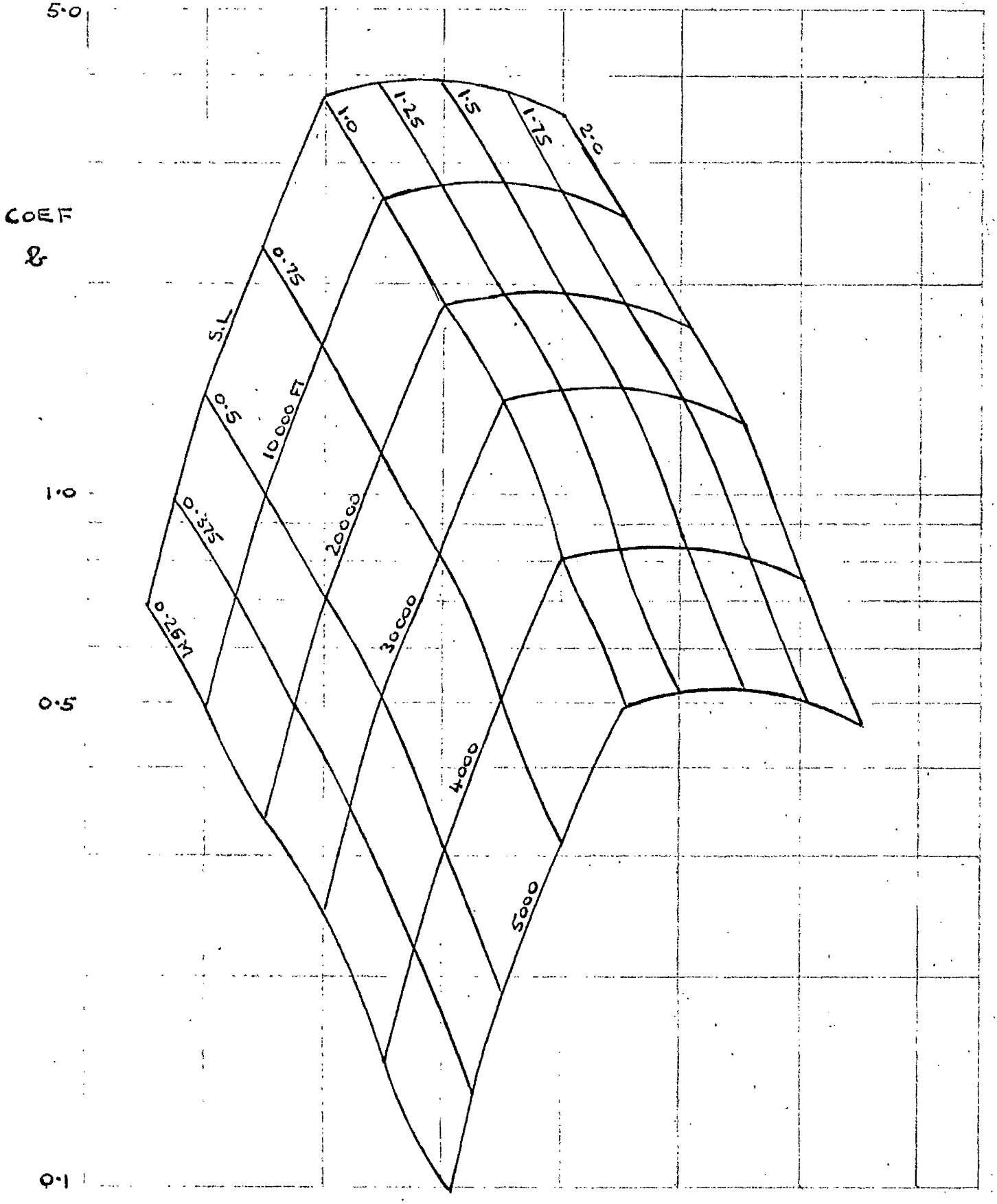


FIGURE 4-5

VARIATION OF 'δ' WITH ALTITUDE AND MACH NO.

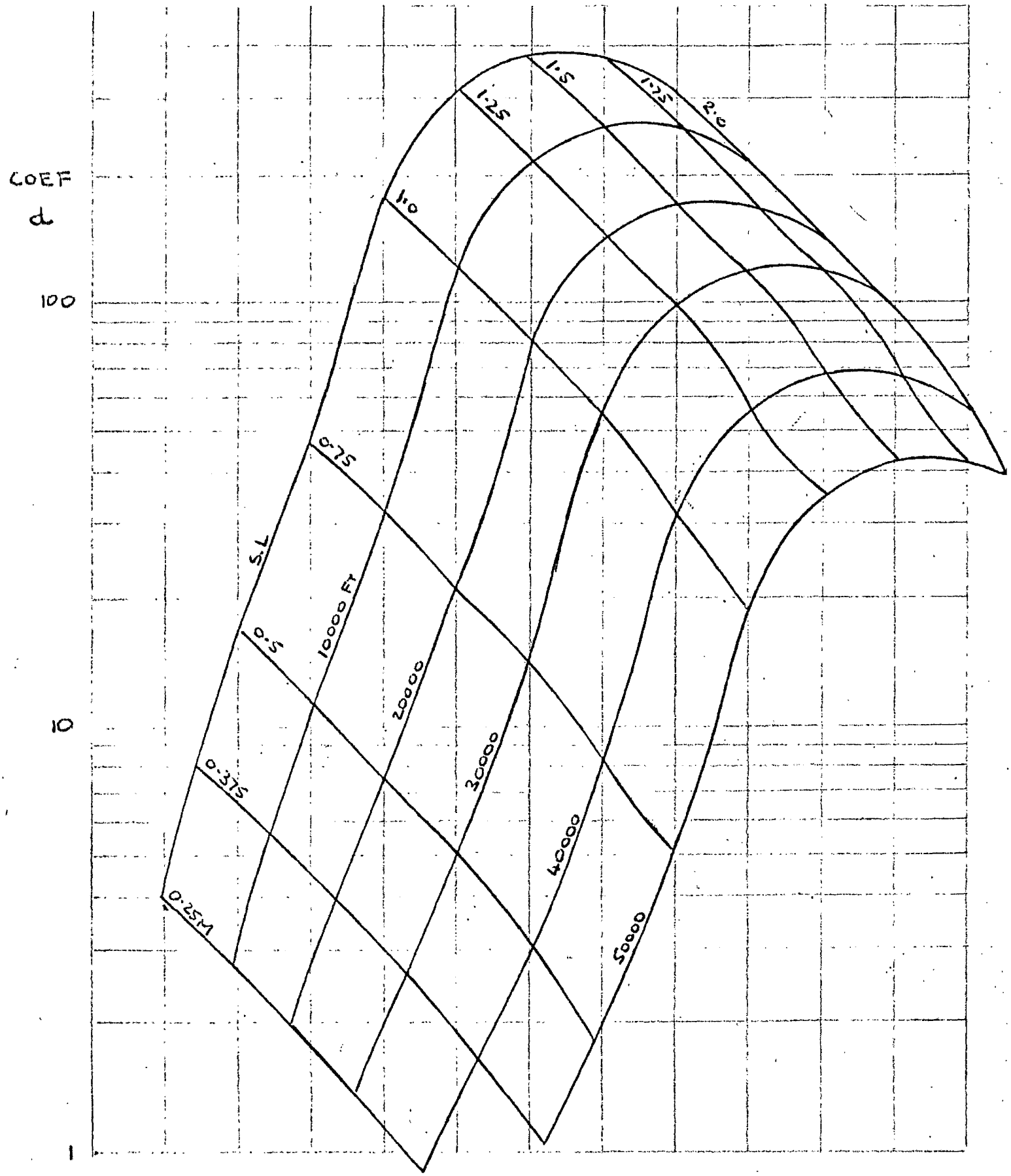


FIGURE 46

VARIATION OF "d" WITH ALTITUDE AND MACH No.

APPENDIX F - AN AIRCRAFT PITCH CONTROL MODEL

The fundamentals of the aircraft short period pitch control system, which is presented in this appendix, was obtained from Smiths Aviation Co. Ltd. in a private communication. It represents a hypothetical aircraft, upon which that Company were carrying out some investigations.

The model is non-linear, since the aerodynamic characteristics of the aircraft vary with its altitude and velocity (58). The model to be considered is the equivalent linearised model, whose transfer function between the pitch of the aircraft and the control input to the elevator actuator is given by

$$(s)/i(s) = a(s + b)/s(s + 8)(s^2 + 2b + d) \qquad (F.1)$$

The actuator has a transfer function of $1/(s+8)$, the remainder of the left hand side of equation (F.1) being due to the aircraft's aerodynamics. The constants a, b and d are the coefficients which vary with the aircraft's altitude and velocity, their values being obtained from the isometric graphs illustrated in Figures (44), (45) and (46). The stability characteristics vary widely over the flight envelope, since the roots of system are dependent on b and d. It was intended to examine the effect of the changing system parameters on the resulting optimally controlled system.

With the control input assumed to act through an amplifier of gain 5,

the aircraft pitch control system can be represented by the matrix state equation,

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \\ \dot{x}_4(t) \end{bmatrix} = \begin{bmatrix} -8 & 0 & 0 & 0 \\ -ab & -2b & -d & 0 \\ a & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{bmatrix} + \begin{bmatrix} 5 \\ 0 \\ 0 \\ 0 \end{bmatrix} m_1(t)$$

where $x_1(t)$ is the elevator angle in degrees,

$x_2(t)$ is a fictitious state variable,

$x_3(t)$ is the pitch rate in degrees/sec.,

$x_4(t)$ is the pitch in degrees,

and $m(t)$ is the control input to the elevator-accuator.

Pilot opinion and other studies associated with the human pilot (59) have indicated that the majority of pilots would prefer all aircraft to have a certain type of response in order to reduce fatigue etc. Such a proposed response is represented by a transfer function between the pitch output and the joy-stick input of

$$\Theta_{ideal}(s)/i_{joy\text{-}stick}(s) = 10/(s^2 + 4.4s + 10).$$

This system has the desirable eigenvalues of -2.2 ± 2.27, i.e. natural angular velocity of 2.27 radians/sec. and a coefficient of damping of 0.695. It is intended to optimally control the system so that the response of the aircraft follows the ideal response as closely as possible. Assuming that only step responses are to be considered, the

matrix state equation governing this reference response system is

$$\begin{bmatrix} \dot{x}_7(t) \\ \dot{x}_8(t) \\ \dot{x}_9(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ -10 & -4.4 & 10 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_7(t) \\ x_8(t) \\ x_9(t) \end{bmatrix}$$

where $x_7(t)$ is the ideal pitch,

$x_8(t)$ is a fictitious state variable,

and $x_9(t)$ is the joy-stick input.

An aircraft in flight experiences atmospheric turbulence, whose statistical properties have been the subject of several papers (58) and (60). It has been found, both theoretically and experimentally, that a fairly good representation of the spectral density of the gust velocity experienced by an aircraft travelling horizontally at a constant velocity of V ft./sec. is

$$G(w) = 12\,\sigma^2\, V/L \times (w^2 + V^2/3L^2)/(w^2 + V^2/L^2)^2 \qquad (F.2)$$

where w is the angular velocity in radians/sec.,

L is the scale of turbulence in feet (58) (60),

and σ^2 is the variance of the gust velocity in $(\text{ft./sec.})^2$.

It is unusual to have the spectral density of the gust velocity as a function of its variance. Since the spectral density and correlation function are fourier transform pairs and the variance equals the correlation function at zero delay, if one determines the inverse fourier transform of G(w), one finds that it does, in fact, obey the property

that it equals the variance of the gust velocity at zero delay. It should be remembered that the above approximate spectral density has been determined under the assumption that the aircraft was in level flight, a requirement which will be assumed to hold in the optimisation analysis. In order to have time invariant dynamics, it will also be assumed that the velocity of the aircraft is constant. A scale of disturbance of 750ft. and a variance of gust velocity of $100(\text{ft./sec.})^2$ were used for the model.

The spectral density of equation (F.2) is considered to be generated by passing white noise with unit, spectral density through a filter with transfer function

$$F(s) = \alpha (s + \beta)/(s + \gamma)^2$$

where $\alpha = (12\,\sigma^2 V/L)^{\frac{1}{2}}$

$\beta = V/\sqrt{3}\,L$

and $\gamma = V/L$

The filter can be represented by the matrix state equations

$$\begin{bmatrix} \dot{x}_5(t) \\ \dot{x}_6(t) \end{bmatrix} = \begin{bmatrix} 0 & \beta - 2\gamma \\ \dfrac{-\gamma^2}{\beta - 2\gamma} & -2\gamma \end{bmatrix} \begin{bmatrix} x_5(t) \\ x_6(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u_1(t)$$

where $x_5(t)$ is the output from the filter without gain α,

$x_6(t)$ is a fictitious state variable,

and $u_1(t)$ is the white noise input.

It is assumed that the random noise disturbance is additive to the output of the elevator actuator through a fictitious amplifier of gain α.

APPENDIX G - RANDOM SEQUENCE GENERATION

For simultation purposes, both sequences of independent non-white and white random disturbance inputs were required in the examples detailed in this thesis. These sequences were generated digitally by methods discussed below, but before going into details, the relevant statistical tests used in the analysis of the generators will be listed.

G.1. STATISTICAL TESTS.

(a) Mean and Variance Tests (61).

The theoretical probability frequency distribution of the sequence of the random numbers will be known, thus one can calculate the theoretical values of the mean and variance. These can be compared with the actual values determined by numerical calculation from the random sequence itself.

(b) Auto-correlation Function (61).

The correlation function of independent noise is a delta function at zero delay. It should be noted that such a correlation function is only a sufficient, but not a necessary condition, for independence. In numerical calculations, a normalised correlation function was used, the latter being defined by

$$\bar{R}(\tau) = \text{covariance}\,[u(t),\ u(t+\tau)]\,/\text{variance}\,[u(t)]$$

where $u(t)$ is random sequence

and τ is delay.

(c) Chi-squared Test (61).

The chi-squared test is a null hypothesis test. As an example, one can divide the range in which a random variable lies into k subranges and then calculate the chi-sequence value

$$\chi^2 = \sum_{i=1}^{k} (\text{actual}(i) - \text{hypothetical})^2 / \text{hypothetical}\ (i)$$

where actual (i) is the number of times the random sequence lies in i th subrange and hypothetical (i) is the corresponding number of times which can be theoretically calculated under the hypothesis about the shape of the probability frequency distribution curve. The chi-squared value is therefore a measure of the deviation of the actual from the hypothetical values. From the chi-squared value one can then determine the probability of chi-squared having this value. If the probability is small, one must assume either that one has examined an extremely rare case or that the original hypothesis was wrong. On the other hand, if the probability was not small, then one can only say that at least the hypothesis has not been proved wrong.

The hypotheses under test in the examination of the random sequences are the rectangular shape of the probability frequency distribution function of the random number generator and the uncorrelated property of the random sequence.

(d) Spectral Density.

Since one requires to generate a random signal with a prescribed spectral density by passing white noise through a suitable filter, it is necessary

to be able to examine the spectral density of a random signal. The spectral density and the auto-correlation function are Fourier transforms of one another. The definition of the Fourier transform and its inverse varies widely in the literature and often on consecutive pages of some texts. Before continuing, one must therefore decide upon one definition to be used throughout the following analysis and digital calculations. The related Laplace transform fortunately does not suffer from this ambiguity, and therefore by analogy the transform pair is defined by

$$R(\tau) = 1/2\pi \int_{-\infty}^{\infty} G(w)\, e^{jw\tau}\, dw$$

and $$G(w) = \int_{-\infty}^{\infty} R(\tau)\, e^{-jw\tau}\, d\tau$$

where $R(\tau)$ is auto-correlation function for delay τ,

$G(w)$ is spectral density for angular frequency w,

and $j = (-1)^{\frac{1}{2}}$.

Thus $$G(w) = 2 \int_{o}^{\infty} R(\tau) \cos(w\tau)\, d\tau$$

Knowing the auto-correlation function of a signal therefore allows one to calculate the spectral density of that signal. The upper infinite limit can be replaced by a finite limit, since the correlation function will tend to become asymptotic to zero with increasing delay, i.e. sufficiently far apart values of a random signal tend to be independent. The integration could be carried out by any of the standard techniques. In the spectral density program written by the author, the weddle formula was used (30).

G.2. - RANDOM NUMBER GENERATION

The recursive multiplicative method of generating random sequences has been used frequently, particularly in Monte Carlo type calculations. The numerical generating process can be described by the difference equation

$$u_1(j+1) = a\, u_1(j) + b$$

where $u_1(j)$ is a member of the random sequence with initial value $u_1(0)$, the variables a and b are chosen constants, and the arithmetic is carried out in some prescribed modulo, M. The sequence will be periodic, the exact value of the period being M or less. The relationship between the period and the constants a, b and M has been the subject of several papers with the net result (62) that choosing constant a prime to the modulo and constant b odd would give a sequence with period equal to the modulo. Further work (63) has also shown that to reduce the correlation between any number and the succeeding ten (approximately) numbers of the sequence, the relationship

$$b = M(0.5 + 3^{\frac{1}{2}}/6) \qquad \text{(G.1)}$$

should hold and neither a nor b should have a small value.

Since the computer to be used was a binary machine with overflow occurring at an approximate value of 2^{39}, the modulo of the generator was chosen to be 2^{34} with constant a having value $(2^4 + 1)$ and constant b, calculated from equation (H.1), having value 13549335641. The quasi random number generator then becomes

$$u_1(j+1) = (2^4+1)\, u_1(j) + 13549335641 \text{ (Modulo } 2^{34}) \qquad (G.2)$$

If the sequence u(j) is then substituted into equation

$$u(j) = (\text{upper}-\text{lower})\, u_1(j)/2^{34} + \text{lower} \qquad (G.3)$$

one would theoretically have a random sequence u(j) with a rectangular probability frequency distribution between the upper limit of upper and the lower limit of lower.

G.3. - TEST OF RANDOM NUMBER GENERATOR.

The random number generator under test was assumed to have an upper and lower limit of 0.5 and -0.5 units respectively. The mean and variance tests gave typical values of 0.0089 and 0.3388 for a sample of 2,000 numbers, whereas the theoretical values are 0 and $\frac{1}{3}$. By increasing the sample length to 10,000, one could improve on the agreement between the actual and the theoretical values by one further decimal place, but little improvement could be obtained by a still further increase in the sampling length.

To test the probability distribution, the range from -0.5 to 0.5 was divided into 25 equal sub-intervals. It is generally accepted that a χ^2 probability of calculating the actually determined chi-squared value of less than 0.05 indicates that the hypothesis being tested is of questionable validity. Typical results from the probability frequency distribution test were found to give a χ^2 probability of greater than 0.25 for sample lengths from 2,000 to 10,000 numbers; thus, at least

the rectangular hypothesis has not been proved invalid.

The correlation between the random numbers was examined by the two-dimensional chi-squared test (61). In order to get consistent results, the independence mesh was limited to 25 elements and the sample length was taken as 10,000 numbers. The results indicated that the independence hypothesis was questionable, since typical χ^2 probability values were less than 0.05. It was noticed that it was only the correlation between successive numbers that seemed to be large, so the random number generator was modified to output only every second random number of the sequence. The corresponding values of the χ^2 probability were found to be approximately independent of the delay and to have a value greater than 0.15. This would indicate a valid independence hypothesis. To add some weight to this conclusion, the normalised correlation function was calculated. It was found to be unity at zero delay, a property of the function, and to have typical values of $\bar{+}$ 0.01 for delays greater than zero.

G.4. - INDEPENDENT NON-WHITE AND WHITE NOISE GENERATORS.

Various approximate white noise generators have been proposed (49), (64) and (65), but the author was unable to find any results obtained from the implementation of these methods far less a comparative study of them. Due to the limited time available, the author was only able to derive and examine one white noise generator based on a method proposed in reference (64). It should be noted that since white noise is defined as noise with a constant spectral density for all frequencies,

it is impossible to generate pure white noise. If the spectral density is flat over the range of frequencies of interest in any specific problem, then this is assumed to be an adequate representation of the ideal white noise.

Consider the random process

$$u(t) = \sum_{i=0}^{\infty} a_i \, H(t-t_i) \qquad (G.4)$$

where a_i is a random independent delta function with zero mean distributed at random and independently over the time axis with a constant average density d or a mean value of time interval between impulses of $1/d$.

The probability distribution of there being n delta functions in the time interval T is governed by the poisson distribution (49).

$$\text{Prob}\,[n] = e^{-dT} \, (dT)^n / n!$$

Assuming H(t) is a stable filter, then using standard probability techniques (64) one can determine an expression for the unnormalised correlation function, namely,

$$R(\tau) = d \, \sigma_a^2 \int_{-\infty}^{\infty} H(t) \, H(t+\tau) \, dt$$

where σ_a^2 is variance of the random variable a_i.

Since the correlation function and the spectral density are fourier transforms of one another

$$G(w) = \int_{-\infty}^{\infty} R(\tau)\, e^{-jw\tau}\, d\tau$$

$$= d\, \sigma_a^2 \left| \int_{-\infty}^{\infty} H(t)\, e^{-jwt}\, dt \right|^2$$

Consider the case when filter H(t) is represented by

$$H(t) = (1/T)\, e^{-t/T}$$

then the spectral density becomes

$$G(w) = d\, \sigma_a^2/(1 + w^2T^2)$$

Hence, for the limiting case as $T \rightarrow 0$, i.e. H(t) tends to a unit delta function,

$$G(w) = d\, \sigma_a^2 \qquad (G.5)$$

The spectral density is constant and independent of angular frequency; therefore the random process described by equation (G.4), under the assumption that H(t) tends to a delta function, is a pure white noise generator. As T is allowed to tend to zero, the mean value of the time interval must also tend to zero for the above theory to be valid. An infinite average impulse density is impractical, thus one must make do with a sufficiently large average impulse density, which will give an adequate representation of white noise over the frequency range of interest.

The purpose of generating white noise was to be able to generate noise with a meromorphic type spectral density, by passing white noise

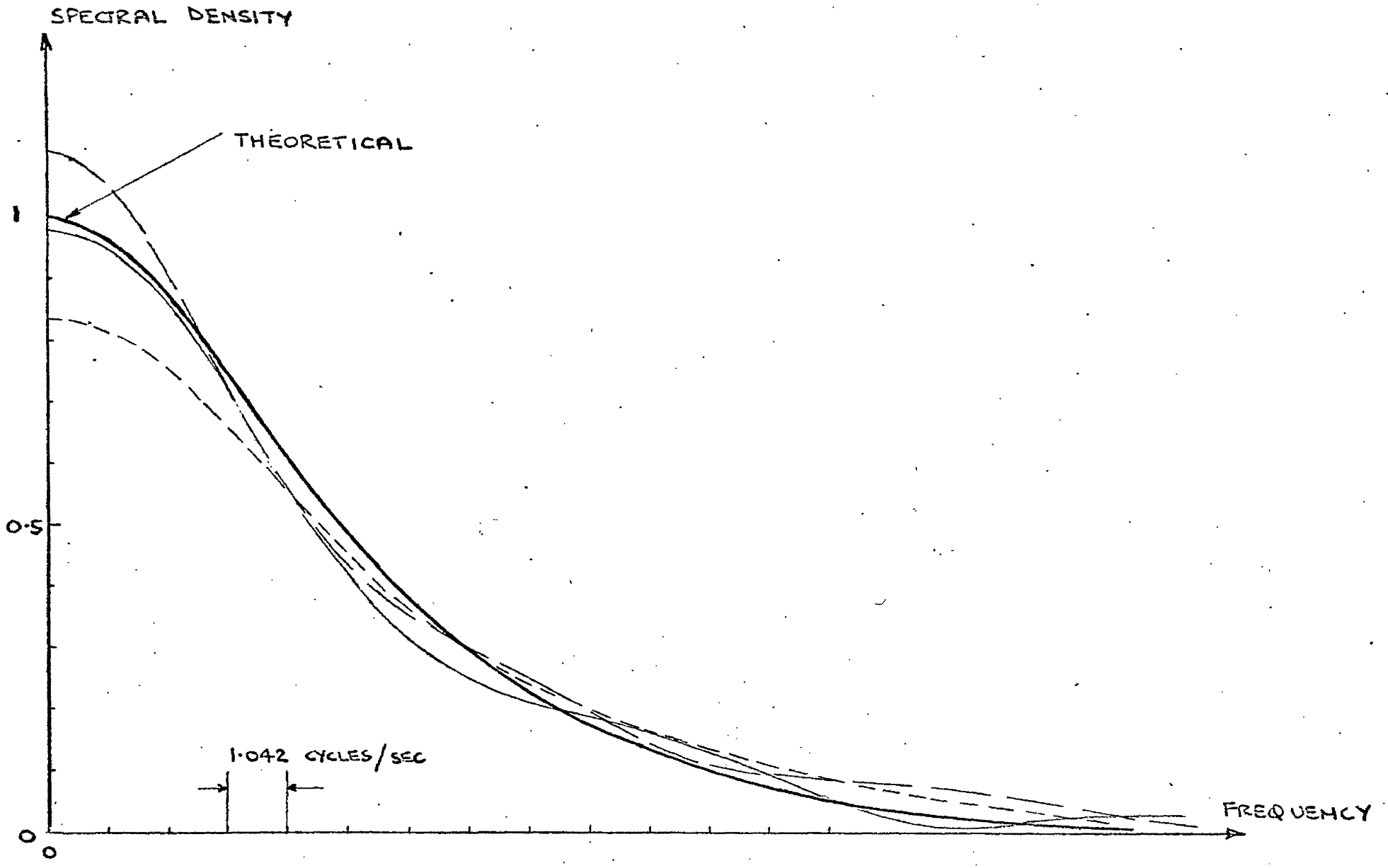


FIGURE 47 SPECTRAL DENSITY OF TEST SIGNAL

with unit spectral density through a suitable filter. The output from the latter will therefore be

$$v(t) = \sum_{i=o}^{\infty} a_i F(t-t_k)$$

where $F(t)$ is a suitable filter and a_i is random independent delta function with zero mean distributed at random and independently over the time axis with constant average density d, such that $d\,\sigma_a^2 = 1$.

The random number generator of equations (G.2) and (G.3) was used to generate the required random independent amplitude quantities a_i and the random independent time intervals between the delta functions. Care must be exercised that the starting numbers for time interval and amplitude generators are not closely sequentially related. To test the output of a filter, and hence indirectly test the white noise generator, the particular case when

$$F(s) = 1/(1 + s)^2$$

was examined. Firstly, the correlation function was calculated, and from that the spectral density was derived, as discussed in section (G.1). The curves in Figure (47) represent the worst upper and lower and also the best approximation to the ideal spectral density of $1/(1+w^2)^2$, which were obtained in twelve tests made with different sets of starting numbers for the random generators. The correlation function was determined for the output of the filter by sampling every 0.04 secs. for the 15,000 samples following the first 2,000 samples, in order to

nullify any generator start up errors, due to the absence of the tails of the responses that would have been applied prior to the start up of the generator. The lower limit for the time interval generator was chosen to be 1% of the period at which the spectral density of the ideal output of filter was 0.5% of its value at zero frequency. This gave a lower time interval limit of 0.016 secs., which, with an upper and lower amplitude limit of +0.25 and -0.25 units respectively, gave an upper limit on the time interval of 0.025 secs. (cf equation (G.5) with unit spectral density).

The agreement between the generated and theoretical spectral densities was not as good as it had been hoped; the major disparity in agreement was at low frequencies. The deviation due to using different sets of starting numbers would most probably indicate that some of the sequences of variables a_i and time intervals are not as independent as others. Since the theory relies on the independence property of these sequences, any violation will tend to give a deviation from ideal spectral density. Nevertheless, since the spectral density is in practice a function which is not usually known to a high degree of accuracy, the white noise generator proposed above was considered acceptable.

For simulation purposes, a non-white independent noise input is also required. Such an input can be generated, using the white noise generator, except that the time interval between successive impulses

is fixed at some constant value T_1.

G.5. - PROGRAMMED GENERATORS

The noise vector required for simulation of optimal controllers and estimators is the sampled data version of the continuous noise input, where the former is defined from the latter to be

$$r(j) = \int_{jT}^{\overline{j+1}\,T} \phi\,(\overline{j+1}\,T - q)\,u(q)\,dq \qquad (G.6)$$

where ϕ is uncontrolled system's transition matrix,

$u(q)$ is continuous noise input, calculated by white noise or independent generators

$r(j)$ is discrete noise input,

and T is sampling period.

It is required for optimisation purposes that $r(j)$ be independent of $r(k)$, $k \neq j$; thus as a final statistical test, the independence of the corresponding discrete output of filter $1/(1+s)^2$ was tested by the χ^2 independence test for non-white and white noise continuous inputs. The resulting χ^2 probabilities in both cases were of the order of 0.2 independent of the delay, and thus are considerably in excess of the significance level of 0.05.

Two procedures were written in ALGOL matrix scheme language for use in the general digital simulation program. These were a procedure RANOGEN, which is equivalent to equations (G.2) and (G.3) and a procedure TRANSITNU, which generated the discrete noise input vector defined by

equation (G.6). The former procedure requires no explanation, while the latter possibly does.

For a system with continuous transition matrix, A, and sampling period, period, the noise vector is calculated and stored in vector NU. Matrix DISFORM is the disturbance - state continuous transition matrix for the vector of noise inputs whose upper and lower bounds and last numbers used in the corresponding random number generators for amplitude and time intervals are stored in the vectors AUPPER, ALOWER AND ALASTNO and TUPPER, TLOWER AND TLASTNO respectively. The vector XSTIME stores the necessary information on what time remains of a random time interval at the end of each sampling period. A procedure call for TRANSITPHI is made in the procedure. This procedure calculates the discrete transition matrix $\phi(t)$ for any time interval t by infinite series method (cf Chapter 2.1). It is not thought necessary to give details of this simple procedure.

There follows a printout of the procedures RANOGEN and TRANSITNU.

```
procedure RANOGEN (upper, lower, lastno, rn);
value upper, lower;
real upper, lower, rn;
integer lastno;

      begin
      integer random, i;
      random:= lastno;
      for i:= 1 step 1 until 2 do
           begin random:= 17×random+13549335641;
           random:= random-entier(random/2↑34)×2↑34;
           end;

      rn:= random×(upper-lower)/2↑34+lower;
      lastno:= random;
      end;


procedure TRANSITNU(DISFORM, period, TUPPER, TLOWER, TLASTNO,
                    AUPPER, ALOWER, ALASTNO, A, XSTIME, NU);
value DISFORM, period, TUPPER, TLOWER, AUPPER, ALOWER, A;
real period;
integer DISFORM, TUPPER, TLOWER, TLASTNO, AUPPER, ALOWER,
        ALASTNO, A, XSTIME, NU;

        begin
        real rtime, tlapse, ramp, tup, tlow, aup, alow;
        integer i, j, tlast, alast,
                PHI, NU2, NU3;

        null(rows(A), 1);  equals(NU);
        for i:= 1 step 1 until cols(DISFORM) do
                begin tlapse:= 0;
                tup:= element(i, 1, TUPPER);
                tlow:= element(i, 1, TLOWER);
                aup:= element(i, 1, AUPPER);
                alow:= element(i, 1, ALOWER);

                null(rows(A), 1);  equals(NU3);

                null(rows(A), 1);
                for j:= 1 step 1 until rows(A) do
                put(element(j, i, DISFORM), j, 1);
                equals(NU2);
```

```
        R1:
        tlast:= element(i, 1, TLASTNO);
        if abs(tlapse) > 10-9 × period then
            RANOGEN(tup, tlow, tlast, rtime) else
            rtime:= element(1, i, XSTIME);
        call(TLASTNO);  put(tlast, i, 1);
        delete(TLASTNO);  equals(TLASTNO);

        tlapse:= rtime+tlapse;
        if tlapse > period then
              begin call(XSTIME); put(tlapse - period, 1, i);
              delete(XSTIME);  equals(XSTIME);
              goto R2;
              end;

        alast:= element(i, 1, ALASTNO);
        RANOGEN(aup, alow, alast, ramp);
        call(ALASTNO);  put(alast, i, 1);
        delete(ALASTNO);  equals(ALASTNO);

        TRANSITPHI(period-tlapse, A, PHI);
        call(PHI);  call(NU2);
        delete(PHI);
        mult by(ramp);  mult;
        call(NU3);  add;  delete(NU3);  equals(NU3);
        goto R1;

        R2:
        call(NU3); call(NU); add; delete(NU);  equals(NU);
        delete (NU2);  delete (NU3);
        end;

end;
```

APPENDIX H - BIBLIOGRAPHY

1. Bliss, G.A., "Lectures on the Calculus of Variations", Seventh Impression, University of Chicago Press, Chicago, 1963.

2. Pontryagin, L.S. et al., "The Mathematical Theory of Optimal Processes", Wiley Inc., New York, 1962.

3. Bellman, R.E., "Dynamic Programming", Princeton University Press, Princeton, 1957.

4. Bryson, A.E., Denham, W.F., "A Steepest-Ascent Method for Solving Optimum Programming Problems", Journal of Applied Mechanics, Series E., Vol.29, 1962.

5. Bryson, A.E., Denham, W.F., "The Solution of Optimal Programming Problems with Inequality Constraints", Raythean Report BP 2122, 1962.

6. Westcott, J.H., Florentin, J.J., Pearson, J.D., "Approximation Methods in Optimal and Adaptive Control", 2nd I.F.A.C. Congress, Butterworth, London, 1963.

7. Noton, A.R.M., "Optimal Control and the Two-point Boundary Problem", Proc.Inst. of Mech.Eng., Part E, Vol.179, 1964-65.

8. Levine, M.D., "A Steepest Descent Method of Synthesising Optimal Control Programmes", Proc.Inst. of Mech. Eng., Part E, Vol.179, 1964-65.

9. Pearson, J.D., "Studies in the Optimal Control of Dynamic Systems", Ph.D. Thesis, London University, 1963.

10. Davis, D.N., "Approximately Optimal Control of Non-Linear Systems by Means of an Analogue Computer", M.Sc. Thesis, London University, 1964.

11. Kalman, R.E., "Contributions to the Theory of Optimal Control", Boletin de la Sociedad Matematica Mexicana, Vol.5, 1960.

12. Merriam, C.W., "An Optimisation Theory for Feedback Control System Design", Information and Control, Vol.3, 1960.

13. Horowitz, I. M., "Synthesis of Feedback Systems", Academic Press, New York, 1963.

14. Miller, W.E., (editor), "Digital Computer Applications to Process Control", Plenium Press, New York, 1965.

15. Halkin, H., "Optimal Control of Systems Described by Difference Equations" in "Advances in Control Systems", edited by C. T. Leondes, Academic Press, New York, 1964.

16. Holtzman, J.M., "Convexity and the Maximum Principle for Discrete Systems", I.E.E.E. Trans. on Automatic Control Vol. AC(11) No. 1, 1966.

17. Kalman, R.E., Koepcke, R.W., "Optimal Synthesis of Linear Sampling Systems Using Generalised Performance Indeces," Trans., A.S.M.E., Vol.80, 1958.

18. Gunckel, T.L., Franklin, G.F., "The General Solution for Linear Sampled-Data Control", Trans. A.S.M.E., Series D., Vol.85, 1963.

19. Tou, J. T., "Optimum Design of Digital Control Systems", Academic Press, New York, 1963.

20. Joseph, P.D., "Optimum Design of Linear Multivariable Digital Control Systems", Ph.D. Thesis, Purdue University, Lafayette, 1962.

21. Deley, G.W., Franklin, G.F., "Optimal Bounded Control of Linear Sampled-Data Systems with Quadratic Loss", Trans. A.S.M.E. Series D., Vol.87, 1965.

22. Zadeh, L.A., Desoer, C.A., "An Introduction to State Space Techniques", Proc. J.A.C.C., June, 1962.

23. Zadeh, L.A., Desoer, C.A., "Linear Systems Theory", McGraw-Hill, New York, 1963.

24. Ince, E.L., "Ordinary Differential Equations", Dover, New York, 1956.

25. Gilbert, E.O., "A Method for the Symbolic Representation and Analysis of Linear Periodic Feedback Systems," Trans. A.I.E.E., Applications and Industry, 1960.

26. Kalman, R.E., Bertram, J.E., "A Unified Approach to the Theory of Sampling Systems", Jour. of Franklin Institute, Vol.267, 1959.

27. Koepcke, R.W., "On the Control of Linear Systems with Pure Time Delay", Trans. A.S.M.E., Series D, Vol.86, 1965.

28. Noton, A.R.M., "Application of Dynamic Programming to the Synthesis of High-Order Multivariable Control Systems, "Proc. I.E.E., Vol.111, 1964.

29. Bellman, R.E., "Adaptive Control Processes", Princeton University Press, Princeton, 1961.

30. Hartree, D.R., "Numerical Analysis", Oxford University Press, Oxford, 1952.

31. Kalman, R.E., "On the General Theory of Control Systems", 1st I.F.A.C. Congress, Butterworth, London, 1960.

32. Adorno, D.S., "Optimum Control of Certain Linear Systems with Quadratic Loss", Information and Control, Vol.5, 1962.

33. Gilbert, E.G., "Controllability and Observability in Multivariable Control Systems", Journal of S.I.A.M., Series A, Vol.1, 1963.

34. Sarachik, P.E., "Controllability and Observability of Linear Discrete -Time Systems, "International Journal of Control Vol.1, 1965.

35. Kalman, R.E., "Mathematical Description of Linear Dynamical Systems", Journal of S.I.A.M., Series A, Vol.1, 1963.

36. Tomovic, R., "Sensitivity Analysis of Dynamic Systems", McGraw-Hill, New York, 1963.

37. Cruz, J.B., Perkins, W.R., "Criteria for System Sensitivity to Parameter Variations", 3rd. I.F.A.C. Congress, 1966 (preprints).

38. Kalman, R.E., "When is a Linear Control System Optimal?", Trans. A.S.M.E., Series D, Vol.85, 1963.

39. Dorato, P., "On Sensitivity of Optimal Control Systems", Trans. I.E.E.E., Vol. AC 8, 1963.

40. Rohrer, R.A., "Sensitivity Considerations in Optimal System Design", Trans I.E.E.E., Vol. AC10, 1965.

41. Pagurek, B., "Sensitivity of the Performance of Optimal Linear Control Systems to Parameter Variations, " International Journal of Control, Vol.1, 1965.

42. Noton, A.R.M., "Application of Dynamic Programming to the Synthesis of High-order Multivariable Control Systems", Proc.I.E.E.,Vol.111, 1964.

43. Kalman, R.E., Bucy, R.S., "A New Approach to Linear Filtering and Prediction Problems", Trans. A.S.M.E., Series D., Vol.82, 1960.

44. Fagin, S.L., "Recursive Linear Regression Theory, Optimal Filtering Theory and Error Analysis of Optimal Systems", M.S. Thesis, New York University, New York, 1964.

45. Ho, Y., Lee, R.C.K., "Identification of Linear Dynamic Systems", Information and Control, Vol.8, 1965.

46. Anderson, G.B., "The Optimisation of Computer-controlled Systems Using Partial Knowledge of the Output State", 2nd. I.F.A.C. Congress, Butterworth, London, 1963.

47. Hosking, K.J.B., "Dynamic Programming and the Synthesis of Linear Optimal Control Systems", Proc. I.E.E., Vol.113, 1966.

48. Tou, T.T., "Digital and Sampled-Data Control Systems", McGraw Hill, New York, 1959.

49. Chang, S.S.L., "Synthesis of Optimal Control Systems", McGraw Hill, New York, 1961.

50. Solodovnikov, V.V., "Statistical Dynamics of Linear Automatic Control Systems", Van Nostrand, Princeton, 1965.

51. Halmos, P.R., "Finite Dimensional Vector Spaces", Van Nostrand, Princeton, 1958.

52. Penrose, R., "A Generalised Inverse for Matrices", Proc. Cambridge Philosophical Society, Vol.51, 1955.

53. Letov, A.M., "Stability in Non-linear Control Systems", Princeton University Press, Princeton, 1961.

54. English-Electric - Leo - Marconi", KDF 9 - Algol Programming Users Manual", Kidsgrove, 1962.

55. English-Electric - Leo - Marconi", KDF 9 - Algol Matrix Scheme Users Manual", Kidsgrove, 1965.

56. Williams, T. J., Otto, R.E., "A Generalised Chemical Processing Model for the Investigation of Computer Control", Trans. A.I.E.E., Part 1, Vol.79, 1960.

57. Perry, J.H., (Editor) "Chemical Engineers's Handbook", McGraw-Hill, New York, 1963.

58. Etkin, B., "Dynamics of Flight", Wiley Inc., New York, 1959.

59. Hall, I.A.M., "Study of the Human Pilot as a Servo-Element", Journal of Royal Aeronautical Society, Vol.67, 1963.

60. Press, H., Meadows, M.T., "A Re-evaluation of Gust Load Statistics for Application in Spectral Calculations", N.A.C.A. Tech.Report 1272, Washington, 1955.

61. Fry, T.C., "Probability and its Engineering Uses", Bell Telephone Series, Van Nostrand, Princeton, 1965.

62. Rotenberg, A., "A New Pseudo-Random Number Generator", Journal A.C.M., Vol.7, 1960.

63. Coveyou, R.R., "Serial Correlation in the Generation of Pseudo-Random Numbers", Journal A.C.M., Vol.7, 1960.

64. Langing, J.H., Batin, R.H., "Random Processes in Automatic Control", McGraw-Hill, New York, 1956.

65. Fuller, A.T., "Notes on the Random Telegraph Signal as an Approximation to Gaussian White Noise", Journal of Electronics and Control, Vol.14, 1963.