TIME - AND FREQUENCY - DOMAIN APPROXIMATION TECHNIQUES FOR OPTIMAL CONTROL

Shaw Choon Chuang

A Thesis submitted for the
degree of Doctor of Philosophy.

Department of Computing and Control,
Imperial College of Science and Technology,
University of London.

October 1971

ABSTRACT

Approximation techniques for optimising control functions using both time and frequency domain methods are of interest. The use of models of high-order systems with a low-order state-space subsystem in parallel with a convolution-described subsystem is considered. An optimisation algorithm for the parallel model, using the Dynamic Programming approach, is developed and computed results are given for an example. Algorithms for obtaining feedback gains for the parallel model with a quadratic cost is developed, using a variational approach. Similar algorithms are also presented for the same model when the quadratic cost includes a sensitivity function term. The convergence of these algorithms is proved.

A continued fraction expansion (CFE) technique is proposed for modelling multivariable transfer functions to give good approximations to the initial transient response while yielding the correct steady-state response for a class of inputs. Another approximation technique, called Modified Padé Approximation (MPA) which differs from the well-known Padé approximation, is also presented, for modelling multivariable transfer functions. It is shown that, under certain conditions, the CFE and MPA techniques yield the same model.

A technique is developed for obtaining approximations to the transfer function of a system such that the steady-state responses of the system and its low-order model are the same for a class of inputs. The technique is based on modelling the system's impulse response by a linear combination of

orthonormal functions. For multivariable systems, low-order models of the transfer function matrices can be obtained by applying the technique to each element of the system's weighting function matrix. A minimal realisation of the resulting low-order model of the transfer function matrix can then be simply obtained.

New lower bounds for the minimum cost for convolution-described linear dynamical systems with quadratic cost functions are presented. Besides its application as stopping conditions for iterative optimisation algorithms to decide when a control sufficiently close to the optimal control has been achieved, it can also be used to check the accuracy of a simplified model to a system.

The CFE and MPA approximation techniques developed before are used to approximate the optimal control for regulator problems formulated in the frequency domain. It is done by performing a sequence of simple spectral factorisations which each require considerably less effort than the complete spectral factorisation which is necessary for the determination of the optimal control in the frequency domain. It should be particularly useful for high-order linear multivariable systems. A frequency domain algorithm is proposed which iterates to give a suboptimal control with increasingly detailed structure.

To my family

ACKNOWLEDGMENTS

I should like to express my deepest appreciation and thanks to my supervisor, Dr. J.C. Allwright, for his encouragement and criticism during the course of this research. My thanks go also to Professor Westcott and Professor Mayne for their interest.

Discussions with the staff and students of the Control Section of the C.C.D. have been beneficial.

The work has been supported by the University of London Postgraduate Studentship during the 1969-70 and 1970-71 Sessions.

CONTENTS

CHAPTER 1

INTRODUCTION TO THE THESIS

1.	INTRODUCTION	10
2.	OUTLINE OF THE THESIS	13
3.	CONTRIBUTIONS OF THE THESIS	16
4.	REFERENCES	19

CHAPTER 2

A TIME DOMAIN APPROXIMATION TECHNIQUE FOR OPTIMAL CONTROL

1.	INTRODUCTION	23
2.	SOLUTION OF THE OPTIMISATION PROBLEM	27
3.	ACCURACY OF THE MODEL	38
4.	EXAMPLE	43
5.	REFERENCES	46

CHAPTER 3

OPTIMAL FEEDBACK GAINS

1.	INTRODUCTION	48
2.	OPTIMAL FEEDBACK GAIN MATRICES FOR THE PARALLEL MODEL	51
3.	SENSITIVITY REDUCTION CONTROLLER DESIGN	68
4.	REFERENCES	75

CHAPTER 4

FREQUENCY DOMAIN APPROXIMATION TECHNIQUE FOR MODELLING

1.	INTRODUCTION	77
2.	CONTINUED FRACTION EXPANSION TO GIVE GOOD APPROXIMATION TO BOTH THE INITIAL TRANSIENT AND STEADY-STATE RESPONSES	79
3.	PADÉ APPROXIMATION FOR SCALAR RATIONAL FUNCTIONS	83
4.	'MODIFIED' PADÉ APPROXIMATION	90
5.	LOW-ORDER APPROXIMATIONS FOR TRANSFER FUNCTION MATRICES	94
6.	EXAMPLES	99
7.	REFERENCES	110

CHAPTER 5

A TIME DOMAIN TECHNIQUE FOR MODELLING

1.	INTRODUCTION	112
2.	TRANSFER FUNCTION DETERMINATION TO GIVE EXACT STEADY-STATE RESPONSE	114
3.	EXAMPLE	118
4.	MINIMAL REALISATION OF THE LOW-ORDER MODEL	123
5.	NUMERICAL EXAMPLE	126
6.	REFERENCES	130

CHAPTER 6

LOWER BOUNDS WITH APPLICATION TO MODEL ACCURACY

1.	INTRODUCTION	132
2.	THE LOWER BOUNDS	134
3.	TEST OF A SYSTEM MODEL	138
4.	STOPPING CONDITIONS FOR ITERATIVE ALGORITHMS	139
5.	EXAMPLE	142
6.	REFERENCES	144

CHAPTER 7

FREQUENCY DOMAIN APPROXIMATION TECHNIQUE FOR OPTIMAL CONTROL

1.	INTRODUCTION	146
2.	PROBLEM FORMULATION	149
3.	SUBOPTIMAL CONTROL	155
4.	ITERATIVE ALGORITHM	157
5.	COMPARISON BETWEEN THE SDMA AND REDUCED MODEL APPROACH	159
6.	EXAMPLES	165
7.	REFERENCES	181

CHAPTER 8

CONCLUSIONS AND FURTHER RESEARCH 183

CHAPTER 1

INTRODUCTION

TO THE

THESIS

1. INTRODUCTION

This thesis presents new and more effective algorithms for choosing optimal controls and controllers for linear multivariable dynamical systems. Since system descriptions are needed before such control can be considered, the problem of determining suitable models is also considered.

Many researchers have contributed to the study of the problem of optimising a linear time-invariant control system with respect to a quadratic performance index. If the system is given in the state-space formulation, it is usually tackled by Kalman's theory of optimal control [1-2] while Wiener's optimal filtering theory [3] is often used to solve the problem formulated in the frequency domain. The optimal feedback control law obtained by Kalman's theory requires the solution to the non-linear Riccati equation; to obtain optimal control using Wiener's formulation necessitates the spectral factorisation of a positive definite matrix commonly known as the spectral density matrix. Kalman's approach requires that all the states of the system be accessible to physical measurement, in order to generate the optimal feedback control law. This difficulty can be overcome by using a observer for the deterministic case and by using a filter for the stochastic system. However, engineering experience indicates that only the 'dominant' states need to be fed back (or estimated and fed back). For convolution-described systems, Lee[4] has proposed iterative algorithms for finding the optimal feedback gain matrix linking the outputs and the inputs of the

system. A more serious drawback is that system design based on optimality becomes very complicated if the order of the system is high and hence expensive to implement in practice.

Though there have been several methods proposed by Youla[5], Davis[6], Anderson[7] and not so well-known, by Rozanov[8] for the spectral factorisation of the spectral density matrix, the procedure to obtain the optimal control in the frequency domain is often very tedious, especially for high-order multivariable systems. It is due to the fact that all these methods are not easily mechanised.

Therefore, in practice, one often has to be satisfied with an approximate solution (i.e. suboptimal control) provided it is easy to obtain and the performance of the resulting system is satisfactory. It is the aim of this thesis to present approximation techniques for optimal control in both time and frequency domains and also to provide a means of evaluating the merits of these techniques.

A popular approach to obtain an approximate solution to the optimal control problem is to represent the actual physical system, usually of high order, by a simplified model for which the optimal control is relatively easy to obtain. The success of this approach would obviously depend on the performance of the actual system using the optimal control of the model. A model may be regarded as an ideal one if it satisfies the following criteria: (1) it must be simple enough so that its optimal control can

be easily obtained; (2) accurate enough so that the final system has a satisfactory performance. It is not easy to find a model which satisfies these criteria simultaneously and to do so might well require deep physical insight of the physical system being considered. Several approximation techniques for obtaining simplified models have been proposed, e.g. Kitamori[9], Davison[10] in the time domain and Evans[11], Chen and Shieh[12] in the frequency domain. The techniques given in [10] and [11] are based on the concept that the dynamical behaviour of the system is determined mainly by the poles nearest to the imaginary axis, i.e. the dominant poles. In Kitamori's approach, the impulse response of the system is approximated by a linear combination of orthogonal functions while Chen's method yields a reduced model by applying the continued fraction expansion to system's transfer function. In this thesis, two modelling techniques in the frequency domain are proposed; one of them is based on the continued fraction expansion which differs from that by Chen while the other is of Padé type. We also extend the technique proposed by Kitamori so that the model obtained will have the correct steady-state response for specified inputs. Cumming[13] has recently suggested a model which consists of a state-space subsystem (to describe the slow modes) in parallel with a convolution-described subsystem (to describe the fast modes) operating only on the recent past history of the system. We obtain the optimal control law for this model with a quadratic performance index. Algorithms to generate the optimal feedback gain matrices

of the model are developed. Development of similar algorithms to reduce the sensitivity of convolution-described systems to parameter variations are also presented in this thesis.

Clearly related to the approximation techniques for the optimal control are the lower bound results for the optimal performance index. These can be used to estimate the difference between the optimal performance index V^o and the performance index V using the optimal control of the model, without actually knowing the optimal control of the system. Thus, it provides a simple means of testing the validity of the simplified model. Another important application is its use as stopping conditions for iterative optimisation algorithms. Allwright[14] has obtained a lower bound result for both state-space described systems and convolution-described systems. We improve these lower bounds for convolution-described systems so that they give a more accurate test of the validity of the model and also provide improved stopping conditions for iterative optimisation algorithms.

In this thesis, a new approximation technique for obtaining a suboptimal control in the frequency domain is also proposed; no simplified model is used in this approach.

2. OUTLINE OF THE THESIS

A brief outline of the contents of this thesis follows.

In Chapter 2, discrete-time systems are described by a parallel model which consists of a state-space

described subsystem in parallel with a convolution-described subsystem operating only on the recent past history of the input to the system. The optimal control law for the parallel model with the usual quadratic performance index is obtained, by using the Dynamic Programming approach. A numerical example is given to justify the case of the parallel model.

Iterative algorithms to generate the optimal feedback gain matrices for the parallel model are presented in Chapter 3. These algorithms are similar to those proposed by Mee [4] for pure convolution-described systems. Then, convergence is proved, using Polak's lemma [15]. For convolution-described systems, a sensitivity reduction design technique is presented. A sensitivity function term is added to the usual quadratic performance index to be minimised and the controller is assumed to be a linear combination of the current state and the sensitivity vector function, or of the current state alone, which is the more practical case. Algorithms are presented to find the optimal feedback matrices and are proved to converge.

A brief summary of the available modelling techniques in the frequency domain is given in Chapter 4. Using the time moments concept, a new result concerning the responses of the model obtained by Chen's method [12] to a class of inputs is presented. A continued fraction expansion technique denoted CFE2, which differs from Chen's method is proposed for model simplification. The relation between the power series expansion of the original transfer function and that of the simplified model obtained by CFE2 is investigated. An approximation technique, denoted MPA,

which differs from the well-known Padé approximation technique is proposed. It is shown that, under certain conditions, the simplified models, of the same order, obtained by CFE2 and MPA are the same. It is also indicated that the CFE2 proposed in this chapter can sometimes be used to provide a stable model while Chen's method gives an unstable model. Several examples are given to illustrate the modelling techniques presented in this chapter.

Chapter 5 is concerned with a modelling technique in the time domain. A well-known technique for approximating a function by a linear combination of linearly independent functions is described in the Introduction; the function to be modelled here is the impulse response of the system. A modelling technique, similar to the above technique, is proposed. It has a desirable feature, namely that the simplified model obtained by the technique proposed will have the correct steady-state response for some pre-specified inputs. A numerical example is presented to illustrate this feature. The minimal realisation of the simplified model is obtained by Gilbert's method [16] and an example is given.

In chapter 6, new lower bound results for the minimum performance index are presented, for convolution-described systems. This is followed by a discussion of their main applications: (i) a simple means for testing the validity of a simplified model to the system: (ii) stopping conditions for iterative optimisation algorithms. Two numerical examples are given to illustrate their usefulness.

Chapter 7 begins with a brief summary of the

available methods for the spectral factorisation of the spectral density matrix which is a necessary step to obtain the optimal control in the frequency domain. The CFE2 and MPA techniques described in chapter 4 are then applied to obtain a low-order approximation to the spectral density matrix and this results in a suboptimal control which can be obtained easily. An algorithm in the frequency domain is also given which interates to give a higher order approximation by 'updating' the previous lower order one. Important properties of the algorithm are indicated. Some interesting results are presented which relate this approach to the well-known simplified model approach. Two numerical examples are given; the second example shows that this approach can be superior to the simplified model approach.

3. CONTRIBUTIONS OF THE THESIS

In general, the techniques, the algorithms and the methods of proof presented in this thesis are original unless otherwise stated.

CHAPTER 2

A new algorithm, based on Dynamic Programming, for optimising the parallel model suggested by Cumming is developed which exploits the structure of the parallel model and is of conceptual interest. The bounds on the difference between the performance index for the optimal control for the parallel model and for the system when the parallel model is obtained from the system in a certain way are new.

CHAPTER 3

The generalisation of the algorithms, proposed originally by Mee for pure convolution-described systems, to exploit the structure of the parallel model when determining the optimal feedback gain matrices for the parallel model is believed to be new. The idea has been extended to cases where a sensitivity term occurs in the cost function, which is believed to be new. Perhaps, the most important contribution made in this chapter is the proof of convergence of the algorithms (including those by Mee). For the sensitivity reduction design techniques for state-space described systems suggested by other authors, no convergence proof for the algorithms proposed has, to the auther's knowledge, appeared in the literature.

CHAPTER 4

The proof that the simplified model obtained by Chen's method will give the correct steady-state response for a class of inputs is original. A continued fraction expansion technique (CFE2) has been applied to system modelling to attempt to obtain good approximation to both the transient and steady-state responses of the system. This use of continued fraction expansion is believed to be new. A new approximation technique of Padé kind (MPA) has been developed for the same purpose. The result on the relationship between the power series of the model obtained by CFE2 or MPA and that of the original transfer function is new. Also, a condition has been obtained for which models, of the same order,

obtained by the CFE2 and MPA are the same.

A letter by the author based on the CFE2 techniques has appeared in Electronics Letters [17].

CHAPTER 5

An important contribution has been made in extending the well-known technique for fitting a function by a linear combination of orthogonal functions so that the model so obtained has the correct steady-state response, for specified inputs.

The first three sections (1-3) of the chapter have appeared as a letter in Electronics Letters [18].

CHAPTER 6

For convolution-described systems, the lower bound results for the minimum performance index presented in this chapter are believed to be original.

CHAPTER 7

Apart from the methods for spectral factorisation given in the Introduction, results in this chapter are original.

4. REFERENCES

[1] R.E. Kalman, On the General Theory of Control Systems, Proc. of IFAC, 1960, Vol. I, pp.481-492.

[2] R.E. Kalman, Contribution to the Theory of Optimal Control, Bol. Soc. Mat. Mexicana, 1960.

[3] N. Wiener, The Extrapolation, Interpolation and Smoothing of Stationary Time Series, Wiley, New York, 1949.

[4] D.H. Mee, A Weighting Function Approach to Linear Control Systems Design, Ph.D. thesis, University of London, 1969.

[5] D.C. Youla, On the Factorisation of Rational Matrices, I.R.E. Trans. Information Theory, Vol.IT-7, July, 1961, pp. 172-189.

[6] M.C. Davies, Factoring the Spectral Matrix, IEEE Trans. Automatic Control, Vol. AC-8, October 1963, pp. 296-305.

[7] B.D.O. Anderson, An Analytic Solution to the Spectral Factorisation Problem, IEEE Trans. Automatic Control, Vol. AC-12, August 1967, pp. 410-414.

[8] Yu.A. Rozanov, Statsionarnye Sluchairye Protsessy Gosuderstvennoe Izdateestvo Fiziko-Matematicheskoi Literatury, Moscow 1963. Translation: Stationary Random Processes, Holden-Day, San Francisco, 1967.

[9] T. Kitamori, Application of Orthogonal Functions to the Determination of Process Dynamic Characteristic and to the Construction of Self-Optimising

Control Systems, Proc. of IFAC, 1960, Moscow, pp. 613-618.

[10] E.J. Davison, A New Method for Simplifying Large Linear Dynamic Systems, IEEE Trans. Automatic Control, AC-13, April 1968, pp.214-215.

[11] W.R. Evans, Control Systems Dynamics, McGraw Hill Book Co., 1954, p.46.

[12] C.F. Chen and L.S. Shieh, A Novel Approach to Linear Model Simplification, 1968 JACC Record, pp.454-461.

[13] S.D.G. Cumming, Observer Theory and Control Systems Design, Ph.D. Thesis, University of London, 1969.

[14] J.C. Allwright, Stopping Conditions for Iterative Control Function Optimisation, Electronics Letters, Vol. 6, No. 17, 1970, pp.545-547.

[15] E. Polak and G. Ribiere, Note Sur La Convergence De Methodes De Directions Conjuguees, to be published.

[16] E.G. Gilbert, Controllability and Obserability in Multivariable Control Systems, J. SIAM Control, ser A, 2, 1963, pp.128-151.

[17] S.C. Chuang, Application of Continued Fraction Expansion Method for Modelling Transfer Functions to Give More Accurate Initial Transient Response, Electronics Letters, Vol.6, No. 26, December 1970, pp.861-863.

[18] J.C. Allwright and S.C. Chuang, Obtaining Transfer

Functions with Zero Steady-State Error from Impulse Response, Electronics Letters, Vol.6, No.22, October, pp. 706-708.

CHAPTER 2

A TIME DOMAIN APPROXIMATION TECHNIQUE FOR OPTIMAL CONTROL

1. INTRODUCTION

There has been an interest in finding a new formulation to describe the dynamical behaviour of high order linear systems with wide range of eigenvalues; linear systems of this nature are often encountered in practice. It would seem that describing linear systems of this type for optimisation purposes by means of a state-space subsystem in parallel with a convolution-described subsystem has potential advantages over either a pure state-space description or a pure convolution description of the system. This is due to the fact that a pure state-space representation describing both the slowly varying and rapidly varying contributions to the system output can cause numerical difficulties when the system of interest has a wide range of eigenvalues and a pure convolution model requires excessive storage of past input history when it is used to describe the slowly varying contributions to the system outputs. Hence, a combination of two descriptions consisting of a state-space subsystem (to describe the slow modes) in parallel with a convolution-described subsystem (to describe the fast modes) operating only on the recent past input history of the system would appear to give a useful form of model for linear dynamical systems. This idea was first suggested by Cumming [1]. The optimal control law for such a parallel model with the usual quadratic performance index is therefore of interest. In this Chapter, a procedure similar to that of

Allwright [2] for pure convolution-described systems will be presented, for calculating the optimal control law for the model.

To enable the state-space and convolution-described subsystems to be synchronised together in the most simple way, a discrete-time formulation is used. The parallel model is therefore written in the form :

$$S_1 \quad : x_{k+1} = A_k x_k + B_k u_k \quad : x_1 = x_s$$

$$S_2 \quad : z_{k+1} = \sum_{i=1}^{p} W_{k+1}^{i} u_{k+1-i} \quad : u_o, u_{-1}, \ldots u_{1-p} \qquad (1)$$

$$y_{k+1} = C_{k+1} x_{k+1} + z_{k+1} \qquad \forall\, k \in [0,1,2,\ldots N]$$

where $x_k \in R^n$; $z_k, y_k \in R^r$, $\forall\, k \in [1,2,\ldots N+1]$

$u_k \in R^m$, $\forall\, k \in [1-p, 1-p+1, .., 0, 1, 2, \ldots N]$.

The final outputs y_k of the model consist of the outputs x_k generated by the state-space subsystem and the output z_k by the convolution-described subsystem. The integer p determines the length of the past input history of the system which is used by the convolution-described subsystem S_2. For any particular p, the matrices A_k, B_k, C_k and W_{k+1}^{i} of the parallel model can be chosen so that the model approximates (in some sense) the physical system that is to be controlled. By increasing p, the accuracy of the model can be improved, at the expense of extra computation. By reducing p, the computation required can be reduced, but the accuracy will also be reduced. Hence, varying p gives a simple way of adjusting the accuracy with which the parallel model approximates

the physical system to be controlled.

The state of a system is defined intuitively as that information about the past history of the system which, together with the present and future inputs, enables the present and future outputs of the system to be determined. In the case of the parallel model described by equation (1), it is clear that the state at time index k is s_k where s_k is the ordered tuplet

$$s_k = (x_k, u_{k-1}, u_{k-2}, \ldots u_{k-p}). \qquad (2)$$

The initial condition for the parallel model can therefore be considered to be $s_1 \triangleq (x_s, u_o, u_{-1}, \ldots u_{1-p})$.

The performance index function which is to be minimised with respect to $u_1, u_2, \ldots u_N$ is considered to be $V(s_1)$ defined by

$$V(s_1, u_1, u_2, \ldots u_N) = \sum_{k=1}^{N} L_k(y_k, u_k) + L_{N+1}(y_{N+1}) \qquad (3)$$

where

$$L_k(y_k, u_k) = \frac{1}{2} \langle (y_k - y_k^d), Q_k(y_k - y_k^d) \rangle + \frac{1}{2} \langle u_k, R_k u_k \rangle$$

$$L_{N+1}(y_{N+1}) = \frac{1}{2} \langle (y_{N+1} - y_{N+1}^d), Q_{N+1}(y_{N+1} - y_{N+1}^d) \rangle$$

and where y_i^d is the desired output value at time index i, Q_i is a r x r symmetric and positive-semi-definite matrix costing the departure from the desired output value and R_i is the m x m symmetric and positive definite matrix costing the control energy.

Cumming suggested, in [1], that the optimisation problem could be solved in the conventional way for

discrete time systems by extending the dimension of the state-space subsystem of equation (1) so that equation (1) can be entirely written in the difference equation form,

$$\tilde{x}_{k+1} = \tilde{A}_k \tilde{x}_k + \tilde{B}_k u_k \qquad : \tilde{x}_1 = \tilde{x}_s$$

$$\tilde{y}_{k+1} = \tilde{C}_{k+1} \tilde{x}_{k+1}$$

where

$$\tilde{x}_k = (x_k^T, u_{k-1}^T, u_{k-2}^T, \ldots\ldots u_{k-p}^T)^T \in R^{n+pm}$$

$$\tilde{x}_s = (x_s^T, u_o^T, u_{-1}^T, \ldots\ldots\ldots u_{1-p}^T)^T \in R^{n+pm}$$

$$\tilde{A}_k = \begin{bmatrix} A_k & & & & & & \\ 0 & 0 & & & & & \\ 0 & I_m & 0 & & & & \\ \cdot & 0 & I_m & 0 & & & \\ \cdot & & & \cdot & & & \\ \cdot & & & & \cdot & & \\ 0 & 0 & \cdot & \cdot & \cdot & I_m & 0 \end{bmatrix} \in R^{(n+pm)x(n+pm)}$$

$$\tilde{B}_k = \begin{bmatrix} B_k \\ I_m \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} \in R^{(n+pm)xm}$$

$$\tilde{C}_{k+1} = (C_{k+1}, W_{k+1}^1, W_{k+1}^2, \ldots\ldots, W_{k+1}^p) \in R^{rx(n+pm)}.$$

However, this formulation has a disadvantage in

that the conventional approach for solving the optimisation problem for discrete time systems does not take account of the sparsities of the matrices $\tilde{A}_k$, $\tilde{B}_k$ and $\tilde{C}_{k+1}$ exhibited here. The procedure to be described does take account of this.

2. SOLUTION OF THE OPTIMISATION PROBLEM

The results, which are proved at the end of this Section, are first stated and commented on.

The Optimal Control Law If the state at time index k for parallel model of equation (1) is $s_k = (x_k, u_{k-1}, u_{k-2}, \ldots u_{k-p})$, then the optimal control at time index k, denoted here by u_k^o is yielded by the optimal control law

$$u_k^o = \alpha_k + \beta_k x_x + \sum_{i=1}^{p-1} \gamma_k^i u_{k-i} \qquad (4)$$

where it is assumed that $u_{k+1}, u_{k+2}, \ldots\ldots u_N$ are chosen optimally given the state s_k. In equation (4) and in the following text, a superscript such as i in γ_k^i denotes an index, rather than a power. Further, in equation (4) : $\alpha_k \in R^m$; $\beta_k \in R^{mxn}$; $\gamma_k^i \in R^{mxm}$ where R^{mxn} denotes the space of mxn matrices etc. The terms α_k, β_k and γ_k^i are determined as follows :

$$\begin{aligned} \alpha_k &= - (H_k)^{-1} e_k \\ \beta_k &= - (H_k)^{-1} F_k \\ \gamma_k^i &= - (H_k)^{-1} G_k^i \end{aligned} \qquad (5)$$

where

$$H_k = R_k + (B_k)^T[\hat{V}^{k+1}_{x_{k+1}x_{k+1}}]B_k + \hat{V}^{k+1}_{u_k u_k}[\hat{V}^{k+1}_{u_k x_{k+1}}]B_k$$

$$+ (B_k)^T[\hat{V}^{k+1}_{u_k x_{k+1}}]^T \hat{V}^{k+1}_{u_k u_k}$$

$$e_k = (B_k)^T[\hat{V}^{k+1}_{x_{k+1}}] + \hat{V}^{k+1}_{u_k} \qquad (6)$$

$$F_k = [\hat{V}^{k+1}_{u_k x_{k+1}}]A_k + (B_k)^T[\hat{V}^{k+1}_{x_{k+1}x_{k+1}}]A_k$$

$$G^i_k = \hat{V}^{k+1}_{u_k u_{k-i}} + (B_k)^T[\hat{V}^{k+1}_{u_{k-i}x_{k+1}}]^T.$$

Here $\hat{V}^{k+1}_{x_{k+1}}$ etc., are determined by running the following recurrence relations in reverse time, by starting at $k = N$ and decreasing k in unit step until $k = 1$, for $i \in [1,2,....p]$:

$$\hat{V}^k_{x_k} = (A_k)^T[\hat{V}^{k+1}_{x_{k+1}}] - (F_k)^T(H_k)^{-1}e_k - (C_k)^T Q_k y^d_k$$

$$\hat{V}^k_{x_k x_k} = (A_k)^T[\hat{V}^{k+1}_{x_{k+1}x_{k+1}}]A_k + (C_k)^T Q_k C_k - (F_k)^T(H_k)^{-1}F_k$$

$$\hat{V}^k_{u_{k-i}} = [\hat{V}^{k+1}_{u_{k-i}} - (G^i_k)^T(H_k)^{-1}e_k](1-\delta_{ip}) - (W^i_k)^T Q_k y^d_k$$

$$\hat{V}^k_{u_{k-i}u_{k-j}} = [\hat{V}^{k+1}_{u_{k-i}u_{k-j}} - (G^i_k)^T(H_k)^{-1}G^j_k](1-\delta_{ip})(1-\delta_{jp})$$

$$+ (W^i_k)^T Q_k W^j_k \qquad (7)$$

$$\hat{V}^k_{u_{k-i}x_k} = [\hat{V}^{k+1}_{u_{k-i}x_{k+1}}A_k - (G^i_k)^T(H_k)^{-1}F_k](1-\delta_{ip})$$

$$+ (W^i_k)^T Q_k C_k$$

subject to the initial conditions, at $k = N+1$, of

$$\hat{V}^{N+1}_{x_{N+1}} = - (C_{N+1})^T(Q_{N+1})y^d_{N+1}$$

$$\hat{V}^{N+1}_{x_{N+1}x_{N+1}} = (C_{N+1})^T(Q_{N+1})(C_{N+1})$$

$$\hat{V}^{N+1}_{u_{N+1-i}} = -(W^i_{N+1})^T(Q_{N+1})y^d_{N+1} \qquad (8)$$

$$\hat{V}^{N+1}_{u_{N+1-i}u_{N+1-j}} = (W^i_{N+1})^T(Q_{N+1})(W^j_{N+1})$$

$$\hat{V}^{N+1}_{u_{N+1-i}x_{N+1}} = (W^i_{N+1})^T(Q_{N+1})(C_{N+1}).$$

In the above : $\hat{V}^k_{x_k} \in R^n$; $\hat{V}^k_{x_kx_k} \in R^{nxn}$; $\hat{V}^k_{u_{k-i}} \in R^m$;

$\hat{V}^k_{u_{k-i}u_{k-j}} \in R^{mxm}$; $\hat{V}^k_{u_{k-i}x_k} \in R^{mxn}$, $\forall\, k \in [1,2,\ldots.N+1]$; and all these can be interpreted as certain partial derivatives, as will be seen later. Also, δ is the Kronecker delta and terms are multiplied by $(1-\delta_{j,p})$ etc. to indicate that they do not apply when i = p etc.

Comment 1 Due the symmetry, $\hat{V}^k_{u_{k-i}u_{k-j}} = (\hat{V}^k_{u_{k-j}u_{k-i}})^T$; which reduces the amount of computation needed. A further reduction can be obtained by noting that $\hat{V}^k_{x_kx_k}$ is symmetric.

Comment 2 Using the optimal control law given in equation (4) in forward time for k increasing from 1 in unit steps enables the optimal controls $u^o_1, u^o_2, \ldots.u^o_N$ which minimise the performance index in equation (3) for the parallel model to be determined as follows :

$$u^o_1 = \alpha_1 + \beta_1 x_s + \sum_{i=1}^{p-1} \gamma^i_1 u_{1-i} \;;$$

$$u_2^o = \alpha_2 + \beta_2 x_2 + \gamma_2^1 u_1^o + \sum_{i=2}^{p-1} \gamma_2^i u_{2-i}$$

where $x_2 = A_1 x_s + B_1 u_1^o$;

$$u_3^o = \alpha_3 + \beta_3 x_3 + \gamma_3^2 u_1^o + \gamma_3^1 u_2^o + \sum_{i=3}^{p-1} \gamma_3^i u_{3-i}$$

where $x_3 = A_2 x_2 + B_2 u_2^o$; etc.

The Minimal Performance Index The minimal value of the performance index in equation (3) for any initial state $s_1 = (x_s, u_o, u_{-1}, \ldots u_{1-p})$ is :

$$V(s_1)^o \triangleq \min_{u_1, u_2, \ldots u_N} V(s_1, u_1, u_2, \ldots u_N)$$

$$= \frac{1}{2} \sum_{k=1}^{N+1} \langle y_k^d, Q_k y_k^d \rangle - \frac{1}{2} \sum_{k=1}^{N} \langle e_k, (H_k)^{-1} e_k \rangle$$

$$+ \langle \hat{V}^1_{x_1}, x_s \rangle + \frac{1}{2} \langle x_s, \hat{V}^1_{x_1 x_1} x_s \rangle$$

$$+ \sum_{i=1}^{p} \langle \hat{V}^1_{u_{1-i}}, u_{1-i} \rangle + \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \langle u_{1-i}, [\hat{V}_{u_{1-i} u_{1-j}}] u_{1-j} \rangle + \sum_{i=1}^{p} \langle u_{1-i}, [\hat{V}_{u_{1-i} x_1}] x_s \rangle \qquad (9)$$

Comment 3 Result (9) for the minimal performance index can be used to give a check on the calculated controls $u_1^o, u_2^o, \ldots u_N^o$. For if the performance index value calculated using equation (3), the calculated controls $u_1^c, u_2^c, \ldots u_N^c$ and the associated response of the parallel model described by equation (1) is not equal to $V(s_1)^o$ of equation (9), then errors must be occuring.

Proof of the Optimisation Results Let the 'cost to go' at time index k given the state s_k at k and the following controls $u_k, u_{k+1}, \ldots u_N$ be

$$V^k(s_k, u_k, u_{k+1}, \ldots u_N) = \sum_{i=k}^{N} L_i(y_i, u_i) + L_{N+1}(y_{N+1}),$$

$$k \in [1, 2, \ldots N] \qquad (10)$$

where at time index N+1, there are no following controls so that

$$V^{N+1}(s_{N+1}) = L_{N+1}(s_{N+1}). \qquad (11)$$

$$\text{Let } V^k(s_k)^o = \min_{u_k, u_{k+1}, \ldots u_N} V^k(s_k, u_k, u_{k+1}, \ldots u_N) \qquad (12)$$

with

$$V^{N+1}(s_{N+1})^o = V^{N+1}(s_{N+1}) = L_{N+1}(s_{N+1}). \qquad (13)$$

Using the Principle of Optimality with equation (10) thus gives

$$V^k(s_k)^o = \min_{u_k} L_k(y_k, u_k) + V^{k+1}(s_{k+1})^o \qquad (14)$$

and the required optimal control u_k^o is given by the minimising control u_k of equation (14). The result of equation (14) is used to establish the dynamic-programming based algorithm. For convenience, it will be rewritten as

$$V^k(s_k)^o = \min_{u_k} \bar{V}^k(s_k, u_k) \qquad (15)$$

where

$$\bar{V}^k(s_k, u_k) \triangleq L_k(y_k, u_k) + V^{k+1}(s_{k+1})^o. \qquad (16)$$

Suppose $V^{k+1}(s_{k+1})^o$ has the following form, which is essentially a second-order expansion of $V^{k+1}(s_{k+1})^o$

in $s_{k+1} = (x_{k+1}; u_k; u_{k-1}; \ldots\ldots; u_{k-p+1})$

$$V^{k+1}(s_{k+1})^{\circ} = \hat{V}^{k+1} + \langle[\hat{V}^{k+1}_{x_{k+1}}], x_{k+1}\rangle + \frac{1}{2}\langle x_{k+1}, [\hat{V}^{k+1}_{x_{k+1}x_{k+1}}]x_{k+1}\rangle + \sum_{i=1}^{p}\langle[\hat{V}^{k+1}_{u_{k+1-i}}], u_{k+1-i}\rangle + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\langle u_{k+1-i}, [\hat{V}^{k+1}_{u_{k+1-i}u_{k+1-j}}]u_{k+1-j}\rangle + \sum_{i=1}^{p}\langle u_{k+1-i}, [\hat{V}^{k+1}_{u_{k+1-i}x_{k+1}}]x_{k+1}\rangle \quad (17)$$

where the terms $\hat{V}^{k+1}_{x_{k+1}}$ etc can obviously be interpreted as terms of a truncated Taylor series expansion.

Now, from equation (1),

$$y_k = C_k x_k + \sum_{i=1}^{p} W_k^i u_{k-i}$$

$$x_{k+1} = A_k x_k + B_k u_k.$$

So, from equations (3),(16) and (17)

$$\bar{V}^k(s_k, u_k) = \tilde{V}^k + \langle[\bar{V}^k_{x_k}], x_k\rangle + \frac{1}{2}\langle x_k, [\bar{V}^k_{x_k x_k}]x_k\rangle + \sum_{i=1}^{p}\langle[\bar{V}^k_{u_{k-i}}], u_{k-i}\rangle + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\langle u_{k-i}, [\bar{V}^k_{u_{k-i}u_{k-j}}]u_{k-j}\rangle + \sum_{i=1}^{p}\langle u_{k-i}, [\bar{V}^k_{u_{k-i}x_k}]x_k\rangle + \langle u_k, [e_k + F_k x_k + \sum_{i=1}^{p-1} G_k^i u_{k-i}\rangle + \frac{1}{2}\langle u_k, H_k u_k\rangle \quad (18)$$

where e_k, F_k, G_k^i and H_k are as defined in equation (6) and where, for all $i \in [1,2,\ldots p]$

$$\tilde{V}^k = \hat{V}^k + \frac{1}{2}\langle y_k^d, Q_k y_k^d\rangle$$

$$\bar{V}^{k}_{x_k} = (A_k)^T[\hat{V}^{k+1}_{x_{k+1}}] - (C_k)^T Q_k y^d_k$$

$$\bar{V}^{k}_{x_k x_k} = (A_k)^T[\hat{V}^{k+1}_{x_{k+1}x_{k+1}}](A_k) - (C_k)^T Q_k (C_k)$$

$$\bar{V}^{k}_{u_{k-i}} = [\hat{V}^{k+1}_{u_{k-i}}](1 - \delta_{ip}) - (W^i_k)^T Q_k y^d_k$$

$$\bar{V}^{k}_{u_{k-i}u_{k-j}} = [\hat{V}^{k+1}_{u_{k-i}u_{k-j}}](1-\delta_{ip})(1-\delta_{jp}) + (W^i_k)^T Q_k (W^j_k)$$

$$\bar{V}^{k}_{u_{k-i}x_k} = [\hat{V}^{k+1}_{u_{k-i}x_{k+1}}](A_k)(1-\delta_{ip}) + (W^i_k)^T Q_k C_k \qquad (19)$$

and where δ is the Kronecker delta and some terms are multiplied by $(1-\delta_{ip})$ etc to indicate that such terms are ignored where i=p, etc.

It will be shown later that H_k is positive definite. It then follows from equations (16) and (18) that the optimal control u^o_k is

$$u^o_k = -(H_k)^{-1}[c_k + F_k x_k + \sum_{i=1}^{p-1} G^i_k u_{k-i}]$$

$$\triangleq \alpha_k + \beta_k x_k + \sum_{i=1}^{p-1} \gamma^i_k u_{k-i} \qquad (20)$$

where α_k, β_k and γ^i_k are as defined in equation (5).

Substituting u^o_k of equation (20) into $\bar{V}^k(s_k, u_k)$ of equation (18) and using equation (15) give

$$V^k(s_k)^o = \hat{V}^k + \langle[\hat{V}^k_{x_k}], x_k\rangle + \frac{1}{2}\langle x_k, [\hat{V}^k_{x_k x_k}]x_k\rangle$$

$$+ \sum_{i=1}^{p} \langle[\hat{V}^k_{u_{k-i}}], u_{k-i}\rangle + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} \langle u_{k-i},$$

$$[\hat{V}^{k}_{u_{k-i}u_{k-j}}]u_{k-j}> + \sum_{i=1}^{p} <u_{k-i}, [\hat{V}^{k}_{u_{k-i}x_k}]x_k> \qquad (21)$$

where the terms $\hat{V}^{k}_{x_k}$ etc., are as given in equation (7)

and where

$$\hat{V}^{k} = \hat{V}^{k+1} + \frac{1}{2} <y^{d}_{k}, Q_k y^{d}_{k}> - \frac{1}{2} <e_k, (H_k)^{-1} e_k>. \qquad (22)$$

Thus, if $V^{k+1}(s_{k+1})^{o}$ has a second-order expansion which is exact as in equation (17), then so does $V^{k}(s_k)^{o}$, as in equation (21). Therefore, if $V^{N+1}(s_{N+1})^{o}$ has a second-order expansion which is exact, the above may be used iteratively, starting with k=N with k being reduced in unit steps until k=1, to enable the terms of the control law of equation (20) for determining the optimal control u^{o}_{k} to be established for all $k \in [1,2,\ldots N]$.

Now, from equation (1), (3) and (13),

$$V^{N+1}(s_{N+1})^{o} = \hat{V}^{N+1} + <[\hat{V}^{N+1}_{x_{N+1}}], x_{N+1}> + \frac{1}{2} <x_{N+1}, [\hat{V}^{N+1}_{x_{N+1}x_{N+1}}]x_{N+1}> + \sum_{i=1}^{p} <[\hat{V}^{N+1}_{u_{N+1-i}}], u_{N+1-i}>$$

$$+ \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} <u_{N+1-i}, [\hat{V}^{N+1}_{u_{N+1-i}u_{N+1-j}}]u_{N+1-j}>$$

$$+ \sum_{i=1}^{p} <u_{N+1-i}, [\hat{V}^{N+1}_{u_{N+1-i}x_{N+1}}]x_{N+1}> \qquad (23)$$

where

$$\hat{V}^{N+1} = \frac{1}{2} <y^{d}_{N+1}, Q_{N+1} y^{d}_{N+1}> \qquad (24)$$

and where the terms $\hat{V}^{N+1}_{x_{N+1}}$ etc are as defined in equation

(8).

Thus, since $V^{N+1}(s_{N+1})^{o}$ has a second-order expansion which is exact, the above technique can be used to solve the optimisation problem provided H_k is positive definite.

Since the minimal value of the performance index $V[s_1, u_1, u_2, \ldots u_N]$ of equation (3) is obviously equal to $V^k(s_k)^{o}$ of equation (12) for k=1, we have from equations (21), (22) and (24) that the minimum performance index value is as shown in equation (9).

Finally, the positive-definiteness of H_k is now considered*. The response of the parallel model described by equation (1) can clearly be written as

$$y_j = \phi_j x_s + \sum_{i=1}^{j-1} \bar{W}_j^i u_{j-i} \qquad (25)$$

for some ϕ_j, $\bar{W}_j^i$.

Now, from equations (10) and (3),

$$\begin{aligned} V^{k+1}(s_{k+1}, u_{k+1}, u_{k+2}, \ldots u_N) &= \sum_{i=k+1}^{N} L_i(y_i, u_i) \\ &\quad + L_{N+1}(y_{N+1}) \\ &= \frac{1}{2} \sum_{i=k+1}^{N+1} \langle (y_i - y_i^d),\ Q_i(y_i - y_i^d) \rangle \\ &\quad + \frac{1}{2} \sum_{i=k+1}^{N} \langle u_i,\ R_i u_i \rangle . \end{aligned} \qquad (26)$$

Letting

$$Q = \text{block diag}[Q_{k+1}, Q_{k+2}, \ldots\ldots Q_{N+1}]$$

$$R = \text{block diag}[R_{k+1}, R_{k+2}, \ldots\ldots R_N]$$

* The proof was suggested by Allwright.

$$y = (y_{k+1}^T,\ y_{k+2}^T, \ldots\ldots\ldots y_{N+1}^T)^T$$

$$y^d = ((y_{k+1}^d)^T, (y_{k+2}^d)^T, \ldots\ldots, (y_{N+1}^d)^T)^T \qquad (27)$$

$$u = (u_{k+1}^T, u_{k+2}^T, \ldots\ldots\ldots u_N^T)^T$$

enables $V^{k+1}(s_{k+1}, u_{k+1}, u_{k+2}, \ldots u_N)$ to be written as

$$V^{k+1}(s_{k+1}, u_{k+1}, u_{k+2}, \ldots u_N) = \frac{1}{2} \langle (y-y^d),\ Q(y-y^d) \rangle + \frac{1}{2} \langle u,\ Ru \rangle. \qquad (28)$$

It is clear from equation (25) that y in equation (27) can be written as

$$y = \bar{y} + \bar{W}_1 u_k + \bar{W}_2 u \qquad (29)$$

where $\bar{y}$ is the contribution to y due to s_k.

Substituting equation (29) into (28) gives

$$V^{k+1}(s_{k+1}, u_{k+1}, u_{k+2}, \ldots u_N) = \frac{1}{2} \langle (\bar{y}-y^d),\ Q(\bar{y}-y^d) \rangle + \langle u_k,\ \bar{W}_1^* Q(\bar{y}-y^d) \rangle + \frac{1}{2} \langle u_k,\ \bar{W}_1^* Q \bar{W}_1 u_k \rangle + \langle u,\ [\bar{W}_2^* Q(y-y^d) + \bar{W}_2^* Q \bar{W}_1 u_k] \rangle + \frac{1}{2} \langle u,\ [R + \bar{W}_2^* Q \bar{W}_2] u \rangle \qquad (30)$$

where $[R + \bar{W}_2^* Q \bar{W}_2] > 0$ as $R > 0$ (because $R_{k+1}, R_{k+2}, \ldots\ldots R_N > 0$) and $Q \geqslant 0$ (because $Q_{k+1}, Q_{k+2}, \ldots\ldots Q_{N+1} \geqslant 0$). The matrices R and Q are also symmetric, so $[R + \bar{W}_2^* Q \bar{W}_2]$ is also symmetric.

From equations (12) and (30)

$$V^{k+1}(s_{k+1})^o = \min_{u_{k+1}, u_{k+2}, \ldots u_N} V^{k+1}(s_{k+1}, u_{k+1}, \ldots u_N)$$

and so $V^{k+1}(s_{k+1})^o$ can be found by minimising $V^{k+1}(s_{k+1}, u_{k+1}, u_{k+2}, \ldots u_N)$ of equation (30) with respect to u. This is possible since $[R + \bar{W}_2^* Q \bar{W}_2]$ is positive-definite and gives

$$V^{k+1}(s_{k+1})^o = \frac{1}{2} \langle (\bar{y}-y^d),\ Q(\bar{y}-y^d) \rangle + \langle u_k,\ \bar{W}_1^* Q(\bar{y}-y^d) \rangle$$
$$- \frac{1}{2} \langle [\bar{W}_2^* Q(y-y^d) + \bar{W}_2^* Q \bar{W}_1 u_k],\ [R+\bar{W}_2^* Q \bar{W}_2]^{-1}$$
$$[\bar{W}_2^* Q(y-y^d) + \bar{W}_2^* Q \bar{W}_1 u_k] \rangle + \frac{1}{2} \langle u_k,\ \bar{W}_1^* Q \bar{W}_1 u_k \rangle . \tag{31}$$

From equations (16) and (3),

$$\bar{V}^k(s_k, u_k) = \frac{1}{2} \langle (y_k - y_k^d),\ Q_k(y_k - y_k^d) \rangle + \frac{1}{2} \langle u_k,\ R_k u_k \rangle$$
$$+ V^{k+1}(s_{k+1})^o \tag{32}$$

and using the result of equation (31) in (32) and the fact that y_k does not depend on u_k gives

$$\frac{\partial^2 \bar{V}^k(s_k, u_k)}{\partial u_k \partial u_k} = R_k + \bar{W}_1^* Q \bar{W}_1 - \bar{W}_1^* Q \bar{W}_2 [R+\bar{W}_2^* Q \bar{W}_2]^{-1} \bar{W}_2^* Q \bar{W}_1 . \tag{33}$$

But, from equation (18)

$$\frac{\partial^2 \bar{V}^k(s_k, u_k)}{\partial u_k \partial u_k} = H_k . \tag{34}$$

Thus H_k is positive definite if the expression on the right hand side of equation (33) is positive definite. Since $R_k > 0$, the right hand side is certainly positive

definite if

$$\bar{W}_1^* Q \bar{W}_1 - \bar{W}_1^* Q \bar{W}_2 [R + \bar{W}_2^* Q \bar{W}_2]^{-1} \bar{W}_2^* Q \bar{W}_1 \geq 0.$$

Now, it can be easily seen that

$$Q^{\frac{1}{2}} \bar{W}_2 [R + \bar{W}_2^* Q \bar{W}_2]^{-1} \bar{W}^* Q^{\frac{1}{2}} = T(I + T)^{-1}$$

where

$$T = HR^{-1}H^*$$

and

$$H = Q^{\frac{1}{2}} \bar{W}_2.$$

Hence,

$$\begin{aligned} Z &\triangleq \bar{W}_1^* Q \bar{W}_1 - \bar{W}_1^* Q \bar{W}_2 [R + \bar{W}_2^* Q \bar{W}_2]^{-1} \bar{W}_2^* Q \bar{W}_1 \\ &= \bar{W}_1^* Q^{\frac{1}{2}} [I - T(I + T)^{-1}] Q^{\frac{1}{2}} \bar{W}_1 \\ &= \bar{W}_1^* Q^{\frac{1}{2}} (I + T)^{-1} Q^{\frac{1}{2}} \bar{W}_1. \end{aligned}$$

Since $R > 0$, $T \geq 0$ and $(I + T)^{-1} > 0$. Hence $Z \geq 0$ and from equations (33) and (34),

$$H_k > 0$$

as required.

3. ACCURACY OF THE MODEL

As has been mentioned in the Introduction, the accuracy with which the parallel model approximates the actual physical system depends on p which is the length of the input history to the system used by the convolution-described subsystem. By increasing p, the accuracy of the model will be improved but more computational effort will be required. In reverse, the computational effort can be reduced by decreasing p but the accuracy of approximating the system is also decreased. One possible

measure of the accuracy of the parallel model would be the difference between the optimal performance index V^o of the model and the performance index $\hat{V}^o$ of the system using the optimal control u^o of the model. The smaller the difference is, the more accurate is the model. In this Section, it is intended to investigate the effect of varying p on the accuracy of the model, based on this measure.

Following the notations used in the Introduction, a discrete time formulation will be used and the state-space subsystem S_1 will also be expressed in the convolution form i.e. the final output $\hat{y}_{k+1}$ of the parallel model is given by

$$\tilde{y}_{k+1} = \sum_{i=1}^{k} \hat{W}^i_{k+1} u_{k+1-i} + \sum_{i=1}^{p} W^i_{k+1} u_{k+1-i} \quad \forall\ k \in [0,1,..N],$$

where $\tilde{y}_{k+1} \in R^r$, $u_k \in R^m$, $\hat{W}^i_{k+1}$ and W^i_{k+1} are the r x m weighting function matrices of the state-space subsystem and the convolution-described subsystem respectively. The above equation can be written in a more compact form

$$\tilde{\underline{y}} = \begin{bmatrix} (\hat{M}_1 + M_1)u \\ (\hat{M}_2 + M_2)u \\ \vdots \\ (\hat{M}_p + M_p)u \\ (\hat{M}_{p+1} + \tilde{M}_{p+1})u \\ \vdots \\ (\hat{M}_{N+1} + \tilde{M}_{N+1})u \end{bmatrix}$$

$$\underline{y} = [y_1^T, y_2^T, \ldots\ldots y_{N+1}^T]^T, \quad u = [u_o^T, u_1^T, \ldots\ldots u_N^T]^T$$

$$\hat{M}_i = [\hat{W}_i^i, \hat{W}_i^{i-1}, \ldots\ldots \hat{W}_i^1, 0 \ldots 0] \in R^{r \times (N+1)m}$$

$$M_i = [W_i^i, W_i^{i-1}, \ldots\ldots W_i^1, 0 \ldots 0] \in R^{r \times (N+1)m}$$

$$\forall\ i \in [1, 2, \ldots\ \rho\ ],$$

$$\tilde{M} = [0_{j-p}, W_j^{j-p+1}, \ldots . W_j^j, 0 \ldots\ldots 0]$$

$$\forall\ j \in [p+1, \ldots N+1]$$

where 0_{j-p} is a null matrix of dimension r x (mj-mp). Similarly, the output $\hat{\underline{y}}$ of the actual system is given by

$$\hat{\underline{y}} = \begin{bmatrix} (\hat{M}_1 + M_1)u \\ \vdots \\ (\hat{M}_{p+1} + M_{p+1})u \\ \vdots \\ (\hat{M}_{N+1} + M_{N+1})u \end{bmatrix}$$

$$= \begin{bmatrix} (\hat{M}_1 + M_1)u \\ \vdots \\ (\hat{M}_p + M_p)u \\ (\hat{M}_{p+1} + \tilde{M}_{p+1} + E_{p+1})u \\ \vdots \\ (\hat{M}_{N+1} + \tilde{M}_{N+1} + E_{N+1})u \end{bmatrix}$$

where

$$E_j = [W_j^{j-p}, W_j^{j-p-1}, \ldots W_j^{1} \quad , 0, \ldots, 0] \in R^{r \times (N+1)m}$$

$$\forall\, j \in [p+1, \ldots N+1]$$

which is the part of the dynamics of the actual system ignored by the parallel model.

The performance index function for the parallel model given by equation (3) can now be written as

$$V = \frac{1}{2}[\langle u, Ru\rangle + \langle \underline{y}, Q\underline{y}\rangle - 2\langle \underline{y}, Q\underline{\tilde{y}}^d\rangle + \langle \underline{\tilde{y}}^d, Q\underline{\tilde{y}}^d\rangle] \quad (35)$$

where $\tilde{y}_k^d = y_k^d - A^k x_s$

$$R = \text{block diag}[R_1, \ldots\ldots R_N]$$

$$Q = \text{block diag}[Q_1, Q_2, \ldots\ldots Q_{N+1}] \quad (35a)$$

$$\underline{y}^d = [(y_1^d)^T, (y_2^d)^T, \ldots\ldots (y_{N+1}^d)^T]^T.$$

The optimal control u^o of the model which minimises equation (35) is found to be

$$\underline{u}^o = [R + \sum_{i=1}^{p} (\hat{M}_i + M_i)^T Q_i (\hat{M}_i + M_i) + \sum_{i=p+1}^{N+1} (\hat{M}_i + \tilde{M}_i)^T Q_i (\hat{M}_i + \tilde{M}_i)]^{-1} [\sum_{i=1}^{p} (\hat{M}_i + M_i)^T Q_i y_i^d + \sum_{i=p+1}^{N+1} (\hat{M}_i + \tilde{M}_i)^T Q_i y_i^d]. \quad (36)$$

The performance index function for the actual system is defined to be

$$\hat{V} = \frac{1}{2}[\langle u, Ru\rangle + \langle \underline{\hat{y}}, Q\underline{\hat{y}}\rangle - 2\langle \underline{\hat{y}}, Q\underline{y}^d\rangle + \langle \underline{y}^d, Q\underline{y}^d\rangle],$$

where R and Q are as defined in equation (35a); $\underline{\hat{y}}$ is the outputs of the actual system. It is easy to show that

$$\hat{V} = V + \sum_{i=p+1}^{N+1} [(E_i u)^T Q_i (E_i u) + 2(E_i u)^T Q_i (\hat{M}_i + \tilde{M}_i) u - 2(y_i^d)^T Q_i E_i u]$$

where V is the performance index of the model defined in equation (35). The value of the performance index $\hat{V}$ using the optimal control u^o of the model is then given by

$$\hat{V}^o = V^o + \sum_{i=p+1}^{N+1} [<u^o, (E_i^T Q_i E_i + 2E_i^T Q_i (\hat{M}_i + \tilde{M}_i) u^o> - 2<y_i^d, Q_i E_i u^o>],$$

V^o being the optimal performance index of the model. If the complete history of the input is applied to the convolution-described subsystem, i.e. $p = N+1$, we have $\hat{V}^o = V^o$ in the above equation as expected, since the dynamics of the parallel model are then exactly that of the actual system so that no approximation is made. It is easy to see that

$$|\hat{V}^o - V^o| \leq \sum_{i=p+1}^{N+1} [(\|E_i\|^2 \|Q_i\| + 2\|E_i\| \|Q_i\| \|\hat{M}_i + \tilde{M}_i\|) \|u^o\|^2 + 2\|y_i^d\| \|Q_i\| \|E_i\| \|u^o\|]. \quad (37a)$$

Also, from equation (36), we have

$$\|u^o\| \leq \|R^{-1}\| [\sum_{i=1}^{p} \|\hat{M}_i + M_i\| \|Q_i\| \|y_i^d\| + \sum_{i=p+1}^{N+1} \|\hat{M}_i + \tilde{M}_i\| \|Q_i\| \|y_i^d\|]. \quad (37b)$$

Combining the bounds given in equations (37a) and (37b), it is now possible to estimate the dependence of the difference in performance indices $\hat{V}^o - V^o$ on p, the length of the input history applied to the convolution-described subsystem. However, the norms of the matrices in equations (37a) and (37b) may not be easy to calculate in practice.

Another possible measure of the model's accuracy would be the departure of the optimal control u^o of the model from that $\hat{u}^o$ of the actual system, i.e. $\|\hat{u}^o - u^o\|$. This norm can be expressed as a function of p but again it could involve considerable amount of calculation. Much further research is, thus, required in finding a suitable measure of model accuracy and hopefully, a simpler expression for model accuracy as a function of p.

4. EXAMPLE

The actual system considered here is a single-input, single-output system of 8th order. The $\hat{A},\hat{B},\hat{C}$ matrices of the system, in the usual state-space notations, are

$$\hat{A} = \begin{bmatrix} -1 & & & & & & & \\ & -2 & & & & & & \\ & & -5 & & & & & \\ & & & -10 & & & & \\ & & & & -20 & & & \\ & & & & & -30 & & \\ & & & & & & -40 & \\ & & & & & & & -50 \end{bmatrix}$$

$$\hat{B} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\hat{C} = [2,-6,5,10,10,-15,20,-25].$$

The system has two relatively small eigenvalues -1,-2. They are, thus, taken to be the eigenvalues of the state-space subsystem which describes the slow modes of the actual system. The impulse response $W_1(t)$ of the state-space subsystem is obtained by using the modelling technique described later in Chapter 5 and its minimal realisation is, then, obtained by Gilbert's method [3]. The A_1, B_1, C_1 matrices of the 2nd order state-space subsystem are found to be

$$A_1 = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

$$C_1 = [-2.94775,\ 5.91042].$$

This subsystem and the actual system have the same steady state response to a step input. The impulse response of the convolution-described subsystem is then obtained from $\hat{W}(t) - W_1(t)$ which is the difference between the impulse response $\hat{W}(t)$ of the actual system and that $W_1(t)$ of the state-space subsystem.

To solve the optimisation problem described by equations (1) and (3), the systems are discretised with a sampling interval $\Delta = 0.05$ sec; the total number of sampling intervals of interest is 100, i.e. $N = 101$. Q_i, R_i and y_i^d in equation (3) are assumed to be constant.

$$Q_i = 5,\ R_i = 1,\ y_i^d = 1 \qquad \forall\ i \in [1,2,\ldots,101].$$

The optimisation problem is solved with different models by using different p; the range of p considered is [0,1,2,3]. The value of the performance index V of the actual system using the optimal control of different models is tabulated below as a function of p

p	value of the performance index V	computing time for finding the optimal control of the model, using CDC 6600 (sec)
0	1.0598	.5167
1	.86456	.5784
2	.72001	.7000
3	.74844	1.0667

When p = 0, only the 2nd order state-space subsystem is used to approximate the original 8th order system, the effect of the fast modes of the original system being totally neglected. By increasing p from 0 to 3, the accuracy of the model is increased by taking into account the effect of the fast modes of the original system. As can be seen from the above table, this results in the reduction in V from 1.0598 (when p = 0) to 0.74844 (when p = 3). This would therefore, justify the use of the parallel model, even though the computational effort needed to find the optimal control of the model is increased.

5. REFERENCES

[1] S.D.G. Cumming, Observer Theory and Control System Design, Ph.D. Thesis, Univ. of London, 1969.

[2] J.C. Allwright, Optimal Control Synthesis by Hill-Climbing in Function Space, Ph.D. Thesis, Univ. of London, 1969.

[3] E.G. Gilbert, Controllability and Observability in Multivariable Control Systems, J.SIAM Control, Ser A, 2, 1963, pp.128-151.

CHAPTER 3

OPTIMAL FEEDBACK GAINS

1. INTRODUCTION

The main topics we shall deal with in this Chapter are the development of algorithms to generate optimal feedback gain matrices and a design technique for the optimal feedback controller to reduce the sensitivity of the system to plant parameter variation. It is well known that the solution to the optimal linear regulator problem i.e. the control of a linear system $\dot{x}=Ax+Bu$, $y=Cx$ with respect to a quadratic performance index, is of feedback form and is generated by a linear combination of all the current states,

$$u^{o}(t) = -B^{T}R^{-1}P(t)x(t) = -L^{o}(t)\,x(t)$$

where $L^{o}(t)$ is the time-varying feedback gain matrix [1]; to obtain $L^{o}(t)$ requiresthe solution $P(t)$ to the usual Riccati equation. Of particular interest is the regulator problem over an infinite time interval where the optimal feedback gain matrix is constant. However, engineering experience indicates that, for many practical control systems, not all the states are available for feedback and in fact, it is often necessary to feedback only the dominant outputs.

Mee [2] has developed algorithms to generate the feedback gain matrix L for the control law of the form

$$u(t) = -Ly(t)$$

where $y(t)$ is the output vector of a convolution-described system

$$y(t) = y_{s}(t) + \int_{t_{s}}^{t} W(t,\tau)u(\tau)\,d\tau.$$

These algorithms generate a reduction in cost at each iteration but the convergence was not proved. In the present Chapter, algorithms similar to those by Mee are developed

for the parallel model described in Chapter 2; these algorithms lead to a reduction in cost in each iteration and will be shown to converge.

For state-space described systems, Cassidy [3] and Bradt [4] have proposed design techniques for an optimal feedback controller which reduces the first order sensitivity of the system to plant parameter variations. In both methods, a term dependent on the sensitivity function is added to the usual quadratic performance index to be minimised and the feedback controller is assumed to be a linear conbination of the current state as well as the sensitivity function. The design problem is thon solved by finding the feedback gain matrices which minimises the resulting performance index. The Maximum Principle is used to obtain the equations for the necessary conditions of optimality. Unfortunately, the problem formulated in this way does not reduce to that of state regulator feedback and associated Riccati formulation and it has to be solved iteratively. Bradt has suggested in [4] an iterative scheme, similar to the gradient technique, to solve these equations. However, the convergence of the algorithm was not proved.

For convolution-described systems, a sensitivity reduction design technique will be described later in the Chapter. The controller is assumed to be a linear conbination of the current state and the sensitivity vector function. Algorithms are developed to find the optimal feedback gain matrices; these algorithms have the similar structure and properties as those of the algorithms developed for the parallel model.

To develop the algorithms, a result is required which

can be found in [2]. However, for convenience, a brief discussion of it is given below. Consider, in general, two Hilbert spaces H_u and H_y and define ℓ as a space of Linear transformations from H_u to H_y i.e.

$$b = L\,a$$

where $a \in H_u$, $b \in H_y$ and $L \in \ell$. By writing

$$f = \langle y, Lu \rangle = \langle [y, u], L \rangle$$

where f is the usual inner product on the Hilbert space H_y, we can associate with any elements u, y, an element $[y,u]$ of the space of linear functionals on ℓ; these functionals can be interpreted as a generalised trace [5] for if H_u and H_y are finite dimensional spaces and L a matrix, we have

$$\langle y, Lu \rangle = y^T Lu = \text{trace } uy^T L$$

$$\triangleq \langle [y, u], L \rangle$$

If H_u and H_y are taken to be L_2 spaces where the inner product is defined as

$$\langle x, y \rangle = \int x^T(t)\, y(t)\, dt,$$

we have

$$\langle y, Lu \rangle = \int y^T(t)\, Lu(t)\, dt$$

$$\triangleq \text{trace } \left(\int u(t)\, y^T(t)\, dt \right) L$$

The main result is stated in the following Lemma.

Lemma: For any two constant matrices M_1, M_2 of appropriate dimensions,

$$\langle [M_1 y_1, y_2], M_2 \rangle = \langle M_1 [y_1, y_2], M_2 \rangle$$

and therefore

$$[M_1 y_1, y_2] = M_1 [y_1, y_2]$$

Proof:

$$\langle [M_1 y_1,\ y_2],\ M_2 \rangle \triangleq \langle M_1 y_1,\ M_2 y_2 \rangle$$

$$= \langle y_1,\ M_1^* M_2 y_2 \rangle$$

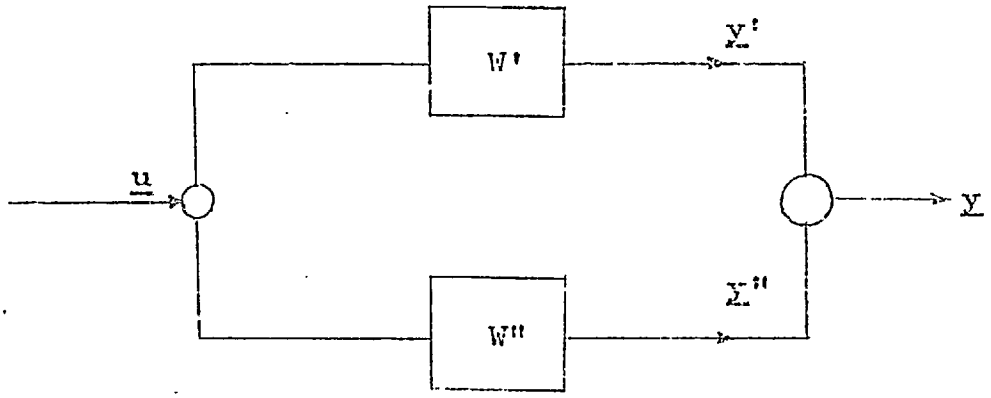
$$= \langle [y_1,\ y_2],\ M_1^* M_2 \rangle$$

$$= \langle M_1 [y_1,\ y_2],\ M_2 \rangle$$

and since M_1, M_2 are arbitrary constant matrices,

$$[M_1 y_1,\ y_2] = M_1 [y_1,\ y_2]$$

2. OPTIMAL FEEDBACK GAIN MATRICES FOR PARALLEL MODEL



The Parallel model of a control system is shown above which consists of a state-space described subsystem in parallel with a convolution-described subsystem. However, in the development of the algorithms, the state-space model will be expressed by the convolution description.

$$y'_k = y'_s + \sum_{i=1}^{k-1} W'_i u_{k-i}$$

Note that the full history of the control u is used in generating the output y', whereas y'' is generated by only using the recent past history of u.

Assume that the feedback control law has the form

$$u_k = - Py'_k - \sum_{i=1}^{q-1} S_i u_{k-i} \qquad (1)$$

where u_k is an m-dimensional vector and y'_k a l- dimensional vector at time instant k; P is a m x l constant matrix and S_i, i=1, 2, ... q-1 m x m constant matrices; q is a positive integer proportional to the length of the past history of the control used to generate the output $\underline{y}''$. Assuming that, in practice, the output y_k of the system can be measured, the control u_k of the form given by equation (1) can be easily implemented. For, knowing the recent past history of the control which is required in finding u_k in equation (1), the output y''_k of the convolution-described subsystem can be calculated and the output y'_k of the state-space subsystem is then given by

$$y'_k = y_k - y''_k.$$

The control u_k in equation (1) can now be found easily. The equation (1) can be written as

$$\underline{u} = - \underline{P}\underline{y}' - \underline{S}\,\underline{\bar{u}} \qquad (2)$$

where

$$\underline{u} = \begin{bmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ \cdot \\ u_n \end{bmatrix}, \quad \underline{\bar{u}} = \begin{bmatrix} 0 \\ u_1 \\ u_2 \\ \cdot \\ \cdot \\ \cdot \\ u_{n-1} \end{bmatrix}, \quad \underline{y}' = \begin{bmatrix} y'_1 \\ y'_2 \\ \cdot \\ \cdot \\ \cdot \\ y'_n \end{bmatrix}$$

$$\underline{P} = \begin{bmatrix} P & & & \\ & P & & 0 \\ & & \cdot & \\ & & & \cdot \\ & 0 & & \cdot \\ & & & & P \end{bmatrix}$$

$$\underline{S} = \begin{bmatrix} 0 & & & & & \\ 0 & S_1 & & & & \\ \cdot & S_2 & S_1 & & 0 & \\ \cdot & \cdot & \ddots & \ddots & & \\ \cdot & \cdot & & \ddots & & \\ \cdot & S_{q-1} & S_{q-2} & \cdots & S_1 & \\ \cdot & 0 & S_{q-1} & \ddots & \ddots & \\ \cdot & \vdots & 0 & \ddots & \ddots & \\ \cdot & \cdot & & \ddots & \ddots & \\ 0 & 0 & \cdots & 0 & S_{q-1} & \cdots & S_1 \end{bmatrix}$$

N being the total number of sampling points; $y'_1 = y'_{s1}$ the initial output. Using these notations, the dynamical equation of the convolution-described system can be, compactly, written as

$$\underline{y'} = \underline{y'_s} + \underline{W}\,\underline{u}$$

where

$$\underline{W} = \begin{bmatrix} 0 & & & & \\ W_o & & & & \\ W_1 & W_o & & 0 & \\ W_2 & W_1 & W_o & \ddots & \\ \cdot & \cdot & \ddots & \ddots & \\ \cdot & \cdot & & \ddots & \\ \cdot & \cdot & & & \\ W_{n-1} & W_{n-2} & & & W_o \end{bmatrix} \qquad \underline{y'_s} = \begin{bmatrix} y'_{s1} \\ y'_{s2} \\ \cdot \\ \cdot \\ \cdot \\ \vdots \\ \cdot \\ \cdot \\ y'_{sn} \end{bmatrix}$$

It is obvious that any change in output $\delta\underline{y}$ of the convolution-described system due to the change $\delta\underline{u}$ in input can be expressed as

$$\delta \underline{y} = \underline{W}\, \delta \underline{u}$$

Consider a quadratic performance index

$$J = \frac{1}{2} < \underline{u}, R\underline{u} > + \frac{1}{2} < (\underline{y}' + \underline{y}''), Q(\underline{y}' + \underline{y}'') >$$

For small perturbation $\delta \underline{P}$ and $\delta \underline{S}$ in feedback gain matrices $\underline{P}$ and $\underline{S}$, the corresponding first order change $\delta \underline{u}$ in the feedback control law $\underline{u}$ is given by

$$\delta \underline{u} = -\delta \underline{P} \underline{y}' - \underline{P}\, \delta \underline{y}' - \delta \underline{S}\, \overline{\underline{u}} - \underline{S}\, \delta \overline{\underline{u}} \qquad (3)$$

where $\delta \underline{y}'$ is the change in $\underline{y}'$ due to the changes $\delta \underline{P}$, $\delta \underline{S}$. Since $\underline{y}' = \underline{W}'\, \underline{u}$, we have

$$\delta \underline{u} + \underline{PW}' \delta \underline{u} + \underline{S}\, \delta \overline{\underline{u}} = - \delta \underline{P}\, \underline{y}' - \delta \underline{S}\, \overline{\underline{u}}$$

or

$$T\, \delta \underline{u} = - \delta \underline{P}\, \underline{y}' - \delta \underline{S}\, \overline{\underline{u}} \qquad (4)$$

where

$$T = \begin{bmatrix} I & & & & & \\ PW'_0 + S_1 & I & & & & \\ PW'_1 + S_2 & PW'_0 + S_1 & \ddots & & 0 & \\ \vdots & & & & & \\ PW'_{q-2} + S_{q-1} & & & & & \\ PW'_{q-1} & & \ddots & & & \\ \vdots & & & \ddots & & \\ PW'_{N-1} & PW'_{n-2} & \cdots & PW'_{q-2} + S_{q-1} & \cdots & I \end{bmatrix}$$

which, by inspection, is invertible.

The first order change in cost function δJ due to the arbitiary small change in control $\delta \underline{u}$ is given by

$$\delta J = \langle R\underline{u},\ \delta\underline{u} \rangle + \langle Q\ (\underline{y}' + \underline{y}''),\ (\delta\underline{y}' + \delta\underline{y}'') \rangle$$

Substitution of the equation (3) into the above equation yields

$$\delta J = \langle R\underline{u},\ (-\delta P\underline{y}' - \delta S\ \bar{\underline{u}}) \rangle + \langle R\underline{\acute{u}},\ (-P\delta\underline{y}' - \underline{S}\ \delta\bar{\underline{u}}) \rangle$$
$$+ \langle Q\ (\underline{y}' + \underline{y}''),\ \underline{W}\ \delta\bar{\underline{u}} \rangle$$

where $\underline{W} = \underline{W}' + \underline{W}''$ is the weighting function matrix of the composite model, $\underline{W}''$ being given by

$$\underline{W}'' = \begin{bmatrix} 0 & & & & \\ W_o'' & & & & \\ W_1'' & W_o'' & & 0 & \\ \cdot & \cdot & \cdot & & \\ \cdot & & \cdot & \cdot & \\ W_{q-1}'' & & & \cdot & \cdot \\ & W_{q-1}'' & & & \cdot \\ 0 & & \cdot & & \cdot \\ & & & W_{q-1}'' & \cdot \cdot \cdot W_o'' \end{bmatrix}$$

It is easy to see that

$$\underline{P}\ \delta\underline{y}' + \underline{S}\ \delta\bar{\underline{u}} = \underline{P}\ \underline{W}'\ \delta\bar{\underline{u}} + \underline{S}\ \delta\bar{\underline{u}}$$
$$= (\underline{P}\ \underline{W}' + \underline{S})\delta\bar{\underline{u}}$$

and therefore

$$\delta J = -\langle R\underline{u}, \delta P\underline{y}' + \delta\underline{S}\ \bar{\underline{u}} \rangle$$
$$-\langle (P\underline{W}' + \underline{S})^{*}\ R\underline{u} - \underline{W}^{*}Q\ (\underline{y}' + \underline{y}''),\ \delta\bar{\underline{u}} \rangle$$

By rearranging the equation (4), we have

$$T\,\delta\underline{\bar{u}} = -\,\delta\underline{P}\,\underline{\bar{y}}' - \delta\underline{S}\,\underline{\bar{\bar{u}}}$$

where

$$\underline{\bar{y}}' = \begin{bmatrix} 0 \\ y_1' \\ y_2' \\ \cdot \\ \cdot \\ \cdot \\ y_{n-2}' \end{bmatrix}, \quad \underline{\bar{\bar{u}}} = \begin{bmatrix} 0 \\ 0 \\ u_1 \\ u_2 \\ \cdot \\ \cdot \\ \cdot \\ u_{n-2} \end{bmatrix}.$$

Thus,

$$\delta\underline{\bar{u}} = -\,T^{-1}\,\delta\underline{P}\,\underline{\bar{y}}' - T^{-1}\,\delta\underline{S}\underline{\bar{u}}$$

and

$$\begin{aligned}\delta J &= \langle R(\underline{P}y' + \underline{S}\bar{u}),\ \delta\underline{P}\,\underline{y}' + \delta\underline{S}\,\underline{\bar{u}}\rangle \\ &\quad + \langle(\underline{P}\underline{W}' + \underline{S})^{*}\,R\underline{u} - \underline{W}^{*}Q(\underline{y}'+\underline{y}''),\ T^{-1}\delta\underline{P}\,\underline{\bar{y}}' + T^{-1}\delta\underline{S}\,\underline{\bar{\bar{u}}}\rangle \\ &= \langle R\underline{P}G_{y'} - [T^{-1}{}^{*}\{\underline{W}^{*}Q(y'+y'') - (\underline{P}\,\underline{W}'+\underline{S})^{*}R\underline{u}\},\underline{\bar{y}}'] + [R\underline{S}\bar{u},\underline{y}'],\delta\underline{P}\rangle \\ &\quad + \langle R\underline{S}\,G_u - [T^{-1}{}^{*}\{\underline{W}'{}^{*}Q(\underline{y}'+\underline{y}'') - (\underline{P}\,\underline{W}'+\underline{S})^{*}R\underline{u}\},\underline{\bar{\bar{u}}}] + [R\underline{P}\underline{y}',\underline{\bar{u}}],\delta\underline{S}\rangle \\ &= \langle R\underline{P}G_{y'} - \Phi_{y'},\ \delta\underline{P}\rangle + \langle R\underline{S}\,G_u - \Phi_u,\ \delta\underline{S}\rangle\end{aligned}$$

where

$$\Phi_{y'} = [T^{-1}{}^{*}\{W'{}^{*}Q(\underline{y}' + \underline{y}'') - (\underline{P}\,\underline{W}' + \underline{S})^{*}R\underline{u}\},\underline{\bar{y}}'] - [R\underline{S}\bar{u},\underline{y}']$$

$$\Phi_u = [T^{-1}{}^{*}\{\underline{W}'{}^{*}Q(\underline{y}'+\underline{y}'') - (\underline{P}\underline{W}'+\underline{S})^{*}R\underline{u}\},\ \underline{\bar{\bar{u}}}] - [R\underline{P}\underline{y}',\underline{\bar{u}}];$$

$G_{y'} = [\underline{y}',\ \underline{y}']$ and $G_u = [\underline{\bar{u}},\ \underline{\bar{u}}]$ are square matrices and are assumed to be invertible; this assumption is valid if the components of $\underline{y}'$ are independent and so are the components of $\underline{\bar{u}}$.

To achieve optimality, δJ must be zero for arbitrary allowable perturbations $\delta \underline{P}$ and $\delta \underline{S}$, i.e.

$$\underline{R}\,\underline{P}\,G_{y'} - \Phi_{y'} = 0$$

and (5)

$$\underline{RS}\,G_u - \Phi_u' = 0$$

To find the optimal feedback gain matrices $\underline{P}^o$ and $\underline{S}^o$ iteratively, the basic algorithm proposed is

ALGORITHM 1

$$\underline{P}^{k+1} = R^{-1}\,\Phi_{y'}^k\,(G_{y'}^{-1})^k \tag{6a}$$

$$\underline{S}^{k+1} = R^{-1}\,\Phi_u^k\,(G_u^{-1})^k \tag{6b}$$

Two more algorithms are obtained by introducing the parameters $0< \epsilon \leqslant 1$ and M a positive semi-definite matrix.

ALGORITHM 2

$$P^{k+1} = (1-\epsilon_1)\,\underline{P}^k + \epsilon_1 R^{-1}\,\Phi_{y'}^k (G_{y'}^{-1})^k \quad 0< \epsilon_1 \leqslant 1 \tag{7a}$$

$$\underline{S}^{k+1} = (1-\epsilon_2)\underline{S}^k + \epsilon_2\,R^{-1}\,\Phi_u^k\,(G_u^{-1})^k \quad 0< \epsilon_2 \leqslant 1 \tag{7b}$$

ALGORITHM 3

$$\underline{P}^{k+1} = (1-\epsilon_1)\underline{P}^k + \epsilon_1 (R+M)^{-1}[\Phi_{y'}^k (G_{y'}^{-1})^k + M\underline{P}^k]$$

$$0< \epsilon_1 \leqslant 1 \tag{8a}$$

and

$$\underline{S}^{k+1} = (1-\epsilon_2)\,\underline{S}^k + \epsilon_2 (R+M)^{-1}[\Phi_u^k (G_u^{-1})^k + M\underline{S}^k]$$

$$0< \epsilon_2 \leqslant 1 \tag{8b}$$

$M \geqslant 0$.

Note that if M=0, Algorithm 3 is reduced to Algorithm 2 and, if further $\epsilon_1 = \epsilon_2 = 1$, to the basic Algorithm 1. We shall now proceed to show that Algorithm 2 and 3 lead to a non-zero reduction in cost in each iteration. Since the proof

is fairly tedious, it will be accomplished in two parts presented as Lemma 1 and 2 below.

Lemma 1: The Algorithms 1, 2 and 3 give rise, to first order, a non-zero reduction in cost in each iteration.

Proof: Since the Algorithm 3 can be reduced to the Algorithm 2 and 1 by setting, respectively, M=0 and M=0, $\epsilon_1 = \epsilon_2 = 1$, we shall give only the proof for the Algorithm 3; for the Algorithm 2, proof follows immediately by setting M=0 and for the Algorithm 1, by setting M=0, $\epsilon_1 = \epsilon_2 = 1$.

By writing $\underline{P}^{k+1} - \underline{P}^k = \delta\underline{P}^k$, $\underline{S}^{k+1} - \underline{S}^k = \delta\underline{S}^k$, it is obvious that equation (8a) of Algorithm 3 can be written as

$$\begin{aligned}\underline{P}^{k+1} &= \underline{P}^k + \epsilon_1 (R+M)^{-1}[-(R+M)\underline{P}^k + \Phi^k_{y'}(G^{-1}_{y'})^k + M\underline{P}^k] \\ &= \underline{P}^k + \epsilon_1 (R+M)^{-1}[\Phi^k_{y'}(G^{-1}_{y'})^k - R\underline{P}^k] \\ &= \underline{P}^k + \epsilon_1 (R+M)^{-1}[\Phi^k_{y'} - R\underline{P}^k G^k_{y'}](G^{-1}_{y'})^k\end{aligned}$$

$$\text{i.e. } \delta\underline{P}^k = -\epsilon_1 (R+M)^{-1}[\Phi^k_{y'} - R\underline{P}^k G^k_{y'}](G^{-1}_{y'})^k$$

$$\Rightarrow R\underline{P}^k G^k_{y'} - \Phi^k_{y'} = -\frac{1}{\epsilon_1}(R+M)\,\delta\underline{P}^k G^k_{y'} \qquad (9a)$$

Similarly, we have, from equation (8b)

$$R\underline{S}^k G^k_u - \Phi^k_u = -\frac{1}{\epsilon_2}(R+M)\,\delta\underline{S}^k G^k_u \qquad (9b)$$

To the first order, the change in cost δJ^k at kth iteration of Algorithm 3 is given by

$$\begin{aligned}\delta J^k &= -\langle \delta\underline{P}^k, \frac{1}{\epsilon_1}(R+M)\,\delta\underline{P}^k G^k_{y'}\rangle \\ &\quad -\langle \delta\underline{S}^k, \frac{1}{\epsilon_2}(R+M)\delta\underline{S}^k G^k_u\rangle \\ &= -\frac{1}{\epsilon_1}\langle \delta\underline{P}^k{}_{y'}{}^k, (R+M)\,\delta\underline{P}^k{}_{y'}{}^k\rangle \\ &\quad -\frac{1}{\epsilon_2}\langle \delta\underline{S}^k \underline{u}^k, (R+M)\,\delta\underline{S}^k \underline{u}^k\rangle \qquad (9c)\end{aligned}$$

Hence,

$$\delta J^k \leqslant -\frac{1}{\epsilon_1}\beta_{min}\|\delta\underline{P}^k\underline{y}'^k\|^2 - \frac{1}{\epsilon_2}\beta_{min}\|\delta\underline{S}^{k-k}\underline{u}\|^2 < 0 \qquad (10)$$

where β_{min}= the smallest eigenvalue of the positive definite matrix (R + M).

In the case of Algorithm 2, we have

$$\delta J^k \leqslant -\frac{1}{\epsilon_1}\alpha_{min}\|\delta\underline{P}^k\underline{y}'^k\|^2 - \frac{1}{\epsilon_2}\alpha_{min}\|\delta\underline{S}^{k-k}\underline{u}\|^2 < 0 \qquad (10a)$$

where α_{min}= the smallest eigenvalue of the positive definite matrix R. In the case of Algorithm 1, we have

$$\delta J^k \leqslant -\alpha_{min}\|\delta\underline{P}^k\underline{y}'^k\|^2 - \alpha_{min}\|\delta\underline{S}^{k-k}\underline{u}\|^2 < 0. \qquad (10b)$$

Since the components of $\underline{y}'^k$ and $\underline{u}^k$ are assumed to be independent, an assumption required for the invertibility of the matrices G_u and $G_{y'}$, $\delta\underline{P}^k\underline{y}'^k$ and $\delta\underline{S}^{k-k}\underline{u}$ cannot vanish unless $\delta\underline{P}^k$ and $\delta\underline{S}^k$ vanish respectively. Therefore, the strict inquality sign in expressions (10), (10a) and (10b) holds, unless $\underline{P}$ and $\underline{S}$ satisfy the equation (5) simultaneously. This guarantees, to first order, a non-zero reduction in cost, a vital fact in proving the convergence of the Algorithm 2 and 3.

Lemma 2: For sufficiently small ϵ_1 and $\epsilon_2 > 0$, Algorithm 2 and 3 produce a non-zero reduction in cost in each iteration.

Proof: So far, in the development of the necessary condition for optimality of the algorithms and the proof of Lemma 1, only the first order change in control $\delta\underline{u}$ and first order change in cost function δJ have been considered. To prove that Algorithm 2 and 3 lead to a non-zero reduction in cost in each iteration for sufficiently small ϵ_1, $\epsilon_2 > 0$, the expressions giving the exact change in control $\Delta\underline{u}$ and in cost function ΔJ will have to be used.

For the control of the form $\underline{u} = -\underline{P}\,\underline{y}' - \underline{S}\,\underline{\bar{u}}$ as given by equation (2), the total change in control $\Delta\underline{u}$ due to the changes $\delta\underline{P}$ and $\delta\underline{S}$ in $\underline{P}$ and $\underline{S}$ is given by

$$\Delta\underline{u} = -\delta\underline{P}\underline{y}' - \underline{P}\Delta\underline{y}' - \delta\underline{S}\underline{\bar{u}} - \underline{S}\,\Delta\,\underline{\bar{u}} - \delta\underline{P}\Delta\underline{y}' - \delta\underline{S}\Delta\underline{\bar{u}}$$

or

$$\Delta\underline{u} + \underline{PW}'\Delta\underline{u} + \underline{S}\Delta\underline{\bar{u}} = -(\delta\underline{P}\underline{y}' + \delta\underline{S}\underline{\bar{u}}) - (\delta\underline{P}\Delta\underline{y}' + \delta\underline{S}\Delta\underline{\bar{u}}).$$

As for equation (4), the above equation can be written as

$$\Delta\underline{u} = -T^{-1}(\delta\underline{P}\,\underline{y}' + \delta\underline{S}\underline{\bar{u}}) - T^{-1}(\delta\underline{P}\,\Delta\underline{y}' + \delta\underline{S}\Delta\underline{\bar{u}})$$

$$T_1\Delta\underline{u} = -T^{-1}(\delta\underline{P}\underline{y}' + \delta\underline{S}\underline{\bar{u}})$$

Where T is the matrix as defined in equation (4) and

$$T_1 \triangleq I + T^{-1}(\delta\underline{PW}' + \delta\underline{S}).$$

Now, by rearranging the matrix T_1

$$T_1 = T^{-1}[I + (\underline{P}+\delta\underline{P})\,\underline{W}' + (\underline{S}+\delta\underline{S})],$$

it can be seen that $I+(\underline{P}+\delta\underline{P})\,\underline{W}' + (\underline{S}+\delta\underline{S})$ is a matrix of the same structure as T and hence is invertible. This immediately means that T_1 is invertible and the exact change in control $\Delta\underline{u}$ is given by

$$\Delta\underline{u} = -T_1^{-1}\,T^{-1}(\delta\underline{P}\,\underline{y}' + \delta\underline{S}\,\underline{\bar{u}}).$$

Expressing T_1^{-1} in terms of its Neumann series expansion, i.e.

$$T_1^{-1} = [I + T^{-1}(\delta\underline{P}\,\underline{W}' + \delta\underline{S})]^{-1}$$

$$= \sum_{i=o}^{\infty} (-1)^i\,(T^{-1}(\delta\underline{P}\,\underline{W}' + \delta\underline{S}))^i$$

and making use of equation (4), the total change in control $\Delta\underline{u}$ due to the changes $\delta\underline{P}$ and $\delta\underline{S}$ is given by

$$\Delta \underline{u} = -\left\{ \sum_{i=0}^{\infty} (-1)^i (T^{-1}(\delta \underline{PW}' + \delta \underline{S}))^i \right\} T^{-1}(\delta \underline{Py}' + \delta \underline{S\bar{u}})$$

$$= -\left\{ \sum_{i=1}^{\infty} (-1)^i (T^{-1}(\delta \underline{PW}' + \delta \underline{S}))^i \right\} T^{-1}(\delta \underline{Py}' + \delta \underline{S\bar{u}})$$

$$- T^{-1}(\delta \underline{Py}' + \delta \underline{S\bar{u}})$$

$$= -\underline{e} + \delta \underline{u} \qquad (10c)$$

where $\delta \underline{u}$ as in equation (4), is the first order change in control due to the changes $\delta \underline{P}$ and $\delta \underline{S}$ and

$$\underline{e} \triangleq \left\{ \sum_{i=1}^{\infty} (-1)^i (T^{-1}(\delta \underline{PW}' + \delta \underline{S}))^i \right\} T^{-1}(\delta \underline{Py}' + \delta \underline{S\bar{u}})$$

$$= \left\{ \sum_{i=1}^{\infty} (-1)^i (T^{-1}(\delta \underline{PW}' + \delta \underline{S}))^i \right\} \delta \underline{u} \qquad (10d)$$

It is obvious that $\underline{e}$ involves only the second and higher order changes in $\underline{P}$ and $\underline{S}$ and is, thus, the error incurred in replacing the exact change in control $\Delta \underline{u}$ by the first order change $\delta \underline{u}$. Note that ignoring the second and higher order changes in $\underline{P}$ and $\underline{S}$ does not affect the derivation of the optimality condition in equation (5) and the first order reduction in cost δJ depends only on the first order changes $\delta \underline{P}$ and $\delta \underline{S}$.

For any change $\Delta \underline{u}$ in the control, the resulting total change ΔJ in the value of the cost function J of equation (2a), which is quadratic in $\underline{u}$, is

$$\Delta J = \langle \Delta \underline{u}, \underline{g} \rangle + \tfrac{1}{2} \langle \Delta \underline{u}, A \Delta \underline{u} \rangle$$

where

$$g = R\underline{u} + (\underline{W}' + \underline{W}'')^{*}Q\,(\underline{y}' + \underline{y}'')$$

and $$A = R + (\underline{W}' + \underline{W}'')^{*}Q(\underline{W}' + \underline{W}'').$$

Substituting $\Delta\underline{u}$ of equation (10c) into the above equation for ΔJ gives

$$\Delta J = \langle(\delta\underline{u} - \underline{e}),\ g\rangle + \tfrac{1}{2}\langle \Delta\underline{u},\ A\Delta\underline{u}\rangle$$

$$= \langle\delta\underline{u},\ g\rangle - \langle\underline{e},\ g\rangle + \tfrac{1}{2}\langle\Delta\underline{u},\ A\Delta\underline{u}\rangle.$$

Of all the three terms on the right hand side of the above equation, only the term $\langle\delta\underline{u}, g\rangle$ involves the first order changes $\delta\underline{P}$ and $\delta\underline{S}$; the change in cost due to the second and higher order changes in $\underline{P}$ and $\underline{S}$ is accounted for by the other two terms. It, hence, follows that

$$\langle\delta\underline{u},\ g\rangle = \delta J$$

which is the change in cost due to the first order changes $\delta\underline{P}$ and $\delta\underline{S}$ in $\underline{P}$ and $\underline{S}$.

That Algorithm 2 and 3 always lead to a non-zero reduction in cost in each iteration can now be proved. As in the proof of Lemma 1, only Algorithm 3 will be considered. Following the notations in Lemma 1, we have

$$\delta\underline{P}^{k} = -\epsilon_1 (R+M)^{-1}[R\underline{P}^{k}G_{y'}^{k} - \Phi_{y'}^{k}]\ (G_{y'}^{-1})^{k}$$

and $$\delta\underline{S}^{k} = -\epsilon_2 (R+M)^{-1}[R\underline{S}^{k}G_{u}^{k} - \Phi_{u}^{k}]\ (G_{u}^{-1})^{k}$$

This indicates that the changes $\delta\underline{P}^{k}$ and $\delta\underline{S}^{k}$ are linear functions of the adjustable scalars $0 < \epsilon_1 \leq 1$, $0 < \epsilon_2 \leq 1$, i.e. $\delta\underline{P}^{k} = -\epsilon_1 \underline{D}_{p}^{k}$

and $\delta \underline{S}^k = - \epsilon_2 \underline{D}_s^k$

where $\underline{D}_p^k$ and $\underline{D}_s^k$ are matrices independent of ϵ_1 and ϵ_2.

Similarly, from equations (4) and (10d)

$$\delta \underline{u}^k = -T^{-1}(\delta \underline{Py}'^k + \delta \underline{S\bar{u}}^k)$$

$$= \epsilon_1 d_p^k + \epsilon_2 d_s^k$$

for some vectors d_p^k and d_s^k and

$$\underline{e} = - \left\{ \sum_{m=1}^{\infty} (-1)^m (\epsilon_1 \underline{E}_p^k + \epsilon_2 \underline{E}_s^k)^m \right\} (\epsilon_1 d_p^k + \epsilon_2 d_s^k)$$

$$= - \sum_{i=i}^{\infty} \sum_{j=1}^{\infty} \epsilon_1^i \epsilon_2^j f_{ij}^k$$

for some matrices $\underline{E}_p^k$, E_s^k and vectors f_{ij}^k ; thus, the total change in control $\Delta \underline{u}^k$ is

$$\Delta \underline{u}^k = \epsilon_1 d_p^k + \epsilon_2 d_s^k + \sum_{i=1}^{\infty} \sum_{j=i}^{\infty} \epsilon_1^i \epsilon_2^j f_{ij}^k.$$

Consider now the terms involved in the total change in cost function ΔJ^k

(i) $\langle \delta \underline{u}^k , g^k \rangle = \delta J^k$ is, from the result of Lemma 1, a non-zero negative number and is linearly dependent on the scalars ϵ_1 and ϵ_2, i.e.

$$\langle \delta \underline{u}^k, g^k \rangle = - (\epsilon_1 a_1^k + \epsilon_2 a_2^k)$$

where $a_i^k > 0$, $i=1,2$;

(ii) $-\langle \underline{e}^k, g^k \rangle + \frac{1}{2} \langle \Delta \underline{u}^k, A \Delta \underline{u}^k \rangle$ is, from the above results, dependent on the second and higher order power of ϵ_1 and ϵ_2, i.e.

$$-\langle \underline{e}^k,\ g^k\rangle + \langle \Delta\underline{u}^k,\ A\Delta\underline{u}^k\rangle = \sum_{i=1}^{\infty}\sum_{j=i}^{\infty} \epsilon_1^i\,\epsilon_2^j\, b_{ij}^k$$

$$+ \epsilon_1^2\, b_{20}^k + \epsilon_2^2\, b_{02}^k$$

$$= C^k, \text{ say.}$$

Thus, the total change in cost ΔJ^k can be expressed as

$$\Delta J^k = -(\epsilon_1 a_1^k + \epsilon_2 a_2^k) + C^k$$

Obviously, $\Delta J^k < 0$ if $C^k \leqslant 0$ and the Lemma is proved.

However, if $C^k > 0$, then since $0 < \epsilon_1 \leqslant 1$, $0 < \epsilon_2 \leqslant 1$,

$$0 < C^k \leqslant \epsilon_1\epsilon_2 \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} |b_{ij}^k| + \epsilon_1^2 |b_{20}^k| + \epsilon_2^2 |b_{02}^k| = \bar{C}^k, \text{ say,}$$

$$\Rightarrow \Delta J^k \leqslant -\epsilon_1[a_1^k - \epsilon_1 |b_{20}^k| - \epsilon_2 \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} |b_{ij}^k|] - \epsilon_2[a_2^k - \epsilon_2 |b_{02}^k|]$$

$$= -\epsilon_1\theta_1^k - \epsilon_2\theta_2^k \qquad (11)$$

where

$$\theta_1^k \triangleq a_1^k - \epsilon_1 |b_{20}^k| - \epsilon_2 \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} |b_{ij}^k|$$

$$\theta_2^k \triangleq a_2^k - \epsilon_2 |b_{02}^k|.$$

To ensure that $\Delta J^k < 0$, it is sufficient that $\theta_1 > 0$ and $\theta_2 \geqslant 0$. Now $\theta_2 > 0$ for any

$$\epsilon_2 < \frac{a_2^k}{|b_{20}^k|}$$

and $\theta_1 > 0$ for any

$$\epsilon_1 < \frac{a_1^k - \epsilon_2 \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} |b_{ij}^k|}{|b_{20}^k|}$$

provided $a_1^k - \epsilon_2 \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} |b_{ij}^k| > 0$ which holds if

$$\epsilon_2 < \frac{a_1^k}{\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} |b_{ij}^k|} .$$

Hence, it can be guaranteed that $\Delta J^k < 0$ for any ϵ_1, ϵ_2 where

$$0 < \epsilon_2 < \min \left(1, \frac{a_2^k}{|b_{02}^k|}, \frac{a_1^k}{\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} |b_{ij}^k|} \right)$$

and

$$0 < \epsilon_1 < \min \left(1, \frac{a_1^k - \epsilon_2 \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} |b_{ij}^k|}{|b_{20}^k|} \right)$$

This completes the proof of Lemma 2.

To show that Algorithm 2 and 3 converge, it is necessary to invoke the modified form of Polak's Lemma [6] . As it is not very well-known, the Lemma, in the modified form to suit our purpose, and its associated definitions are given below.

We first consider a general algorithm which minimises a continous function $f(\underline{P}, \underline{S})$, $\underline{P} \in R^{m \times n}$, $\underline{S} \in R^{m \times n}$. Starting from an arbitrary point (P_o, S_o), we form successively

$$\begin{aligned} \underline{P}_{k+1} &= a_p(\underline{P}_k) \\ \underline{S}_{k+1} &= a_s(\underline{S}_k) \end{aligned} \qquad (12)$$

such that $f(\underline{P}_{k+1}, \underline{S}_{k+1}) \leqslant f(\underline{P}_k, \underline{S}_k)$.

The point $(\underline{P}, \underline{S})$ is called a 'desirable' point of the minimisation problem if

$$f[\, a_p(\underline{P}), a_s(\underline{S}) \,] \geq f(\underline{P}, \underline{S}).$$

Note that $a_p(a_s)$ is an operation which generates P_{k+1} (S_{k+1}) from P_k (S_k), keeping S_k (P_k) constant.

MODIFIED POLAK'S LEMMA: If for all (P, S) not 'desirable', there exist constants depending on (P, S), $\varepsilon_p(P, S)$, $\varepsilon_s(P, S)$ and $\delta(P, S) > 0$ such that

$$\|P - P_k\| \leqslant \varepsilon_p(P,S) \text{ and } \|S - S_k\| \leqslant \varepsilon_s(P,S)$$

$$f(P_k,S_k) - f[a_p(P_k), a_s(S_k)] \geqslant \delta(P,S) > 0, \tag{12a}$$

then every convergent subsequence of the sequence defined by algorithm in equation (12) has for its point of accumulation, a 'desirable' point.

Proof: To prove the lemma, it is assumed that the sequence (P_k, S_k) is bounded which is not stated explicitly in the original version of Polak's lemma.

If the sequence is finite, the proof is trivial, for at certain kth iteration, $P_{k-1}=P_k, S_{k-1}=S_k$ with $f(P_{k-1},S_{k-1}) = f(P_k,S_k)$ and hence (P_k,S_k) is 'desirable'.

The case of infinite sequence is now considered. As the sequence consists of bounded elements in finite dimensional spaces $R^{m \times 1}$ and $R^{m \times m}$, there exists a convergent subsequence (P_j, S_j), $j \in J$ with the accumulation point (P^o, S^o), where J is a subset of the set of all positive integers, $P^o \in R^{m \times 1}$, $S^o \in R^{m \times m}$. The lemma is proved by contradiction. Suppose that (P^o, S^o) is not 'desirable'. Then, according to the hypothesis of the lemma, there exists an $\varepsilon_p^o > 0$, $\varepsilon_s^o > 0$ and $\delta^o > 0$ such that

$$\|P^o - P\| \leqslant \varepsilon_p^o \text{ and } \|S^o - S\| \leqslant \varepsilon_s^o$$

$$\Longrightarrow f(P,S) - f(a_p(P), a_s(S)) \geqslant \delta^o > 0. \tag{13}$$

As the subsequence $(\underline{P}_j, \underline{S}_j)$, $j \in J$ is convergent, there exists an integer m such that $j \in J$ and $j \geqslant m$ implies that $\|\underline{P}^o - \underline{P}\| \leqslant \varepsilon_p^o$ and $\|\underline{S}^o - \underline{S}\| \leqslant \varepsilon_s^o$. Consider now two consecutive elements of the convergent subsequence $(\underline{P}_j, \underline{S}_j)$ and $(\underline{P}_{j+i}, \underline{S}_{j+i})$ with j and $j+i \in J$, $j > m$. It is obvious that

$$f(\underline{P}_j, \underline{S}_j) - f(\underline{P}_{j+i}, \underline{S}_{j+i})$$

$$= [f(\underline{P}_j, \underline{S}_j) - f(\underline{P}_{j+1}, \underline{S}_{j+1})] + [f(\underline{P}_{j+1}, \underline{S}_{j+1}) - f(\underline{P}_{j+2}, \underline{S}_{j+2})] + \ldots + [f(\underline{P}_{j+i-1}, S_{j+i-1}) - f(\underline{P}_{j+i}, \underline{S}_{j+i})]. \qquad (14)$$

All the terms on the right hand side of equation (14) are positive as the algorithm guarantees a reduction in cost in each iteration. Hence, equations (13) and (14) yield

$$f(\underline{P}_j, \underline{S}_j) - f(\underline{P}_{j+i}, \underline{S}_{j+i}) \geqslant f(\underline{P}_j, \underline{S}_j) - f(\underline{P}_{j+1}, S_{j+1})$$

$$\geqslant \delta^o > 0. \qquad (15)$$

But, by the definition of continuity, $f(\underline{P}_j, \underline{S}_j) \rightarrow f(\underline{P}^o, \underline{S}^o)$ as $(\underline{P}_j, \underline{S}_j) \rightarrow (\underline{P}^o, \underline{S}^o)$, a contradiction to equation (15). Thus, the accumulation point $(\underline{P}^o, \underline{S}^o)$ of every convergent subsequence is 'desirable'.

To prove the convergence of Algorithms 2 and 3, we define that $(\underline{P}, \underline{S})$ is a stationary point if

$$\frac{\partial J}{\partial \underline{P}} \triangleq J'_p(\underline{P}, \underline{S}) = 0$$

$$\frac{\partial J}{\partial \underline{S}} \triangleq J'_s(\underline{P}, \underline{S}) = 0.$$

Theorem : Algorithm 2 and 3 satisfy the hypothesis of both Lemma 2 and Polak's lemma and hence every convergent

subsequence of the sequence generated by the algorithms converges to a stationary point.

Proof : It is easy to see that the quadratic performance index J is continous in both $\underline{P}$ and $\underline{S}$. From Lemma 1 and 2, Algorithm 2 and 3 always produce a non-zero reduction in cost unless $\underline{P}^{k+1} = \underline{P}^{k}$ and $\underline{S}^{k+1} = \underline{S}^{k}$. This means that Algorithm 2 and 3 satisfy Polak's definition of general algorithm (see equation (12)). We now proceed to show that the hypothesis of Polak's lemma is also satisfied. Again, only the case of Algorithm 3 is presented.

Consider any point $(\underline{P},\underline{S})$ which is not desirable and $\delta\underline{P} \triangleq \underline{P} - \underline{P}^{k}$ and $\delta\underline{S} \triangleq \underline{S} - \underline{S}^{k}$. If

$$\|\delta\underline{P}\| = \|\underline{P} - \underline{P}^{k}\| \leqslant \varepsilon_{p}(\underline{P},\underline{S})$$

and $$\|\delta\underline{S}\| = \|\underline{S} - \underline{S}^{k}\| \leqslant \varepsilon_{s}(\underline{P},\underline{S}),$$

then, from equation (11),

$$J(\underline{P}^{k},\underline{S}^{k}) - J(\underline{P}^{k+1},\underline{S}^{k+1}) \geqslant \delta(\underline{P},\underline{S}) > 0,$$

where $\delta(\underline{P},\underline{S}) = \epsilon_1\theta_1^{k} + \epsilon_2\theta_2^{k}$. Thus, Algorithm 3 satisfies the hypothesis of Polak's lemma and the accumulation point of every convergent subsequence of the sequence formed by the Algorithm is a desirable point. To prove that the desirable point is also a stationary point, suppose that $(\underline{P}^{k},\underline{S}^{k})$ is desirable. Then, by definition (10)

$$J(\underline{P}^{k},\underline{S}^{k}) - J(\underline{P}^{k+1},\underline{S}^{k+1}) \leqslant 0$$

$$\Longrightarrow \|\delta\underline{P}^{k}\underline{y}'^{k}\| = 0 \quad \text{and} \quad \|\delta\underline{S}^{k}\underline{u}^{k}\| = 0$$

$$\Longrightarrow \quad \delta \underline{P}^k \equiv 0 \quad \text{and} \quad \delta \underline{S}^k \equiv 0.$$

From equations (9a) and (9b), since $(R + M)$, $G_{y'}^k$ and G_u^k are invertible,

$$\delta \underline{P}^k \equiv 0 \Longrightarrow R\underline{P}^k G_{y'}^k - \Phi_{y'}^k \equiv 0$$

and $$\delta \underline{S} \equiv 0 \Longrightarrow R\underline{S}^k G_u^k - \Phi_u^k \equiv 0.$$

Thus, $(\underline{P}^k, \underline{S}^k)$ is a stationary point, noting that

$$\left.\frac{\partial J}{\partial \underline{P}}\right|_{\underline{P}^k} = R\underline{P}^k G_{y'}^k - \Phi_{y'}^k \quad \text{and} \quad \left.\frac{\partial J}{\partial \underline{S}}\right|_{\underline{S}^k} = R\underline{S}^k G_u^k - \Phi_u^k .$$

Now if $(\underline{P}^k, \underline{S}^k)$ is stationary, then by reverse argument it is easy to show that it is also desirable and hence the proof is complete.

3. SENSITIVITY REDUCTION CONTROLLER DESIGN

Since the early 1960s, effort has been concentrated on various sensitivity problems which are of importance in the practical design of control systems. Sensitivity reduction controller design techniques have been developed for state-space described systems [3-4]. To solve the problem, Bradt [4] has proposed an iterative scheme, similar to the gradient technique but the convergence of the method was not proved. In this Section, we shall consider a sensitivity reduction design technique, for convolution-described systems. Algorithms similar to those developed for the composite model will be presented and proved to converge.

Sensitivity Function

Consider a convolution-described system

$$y(t;v) = y_s(t) + W(t;v)u(v) \qquad (16)$$

where $y(t;v)$ is the l-dimensional output function; $y_s(t)$ is the disturbance resulting from initial condition or other external influences; $u(v)$ is the m-dimensional output function; $W(t;v)$ is the open-loop transfer operator linking input to output; v is the p-dimensional plant parameter function. u, W and hence y are assumed to be functions of v. The l-dimensional sensitivity function $\frac{\partial y}{\partial v_i}$ is given by

$$\frac{\partial y}{\partial v_i} = \frac{\partial W}{\partial v_i} u + W \frac{\partial u}{\partial v_i} \qquad i = 1,2,....p. \qquad (17)$$

Algorithms For Finding Optimal Feedback Gain Matrices

To reduce the sensitivity of the system to parameter variation, the controller is assumed to generate and feed back a first order approximation to the sensitivity functions as well as the system outputs. The algorithms are developed to determine the feedback gain matrices which minimises a quadratic performance index containing the the sensitivity functions; i.e. the controller is of the form

$$u = -K_1 y - \sum_{i=1}^{p} K_i' \frac{\partial y}{\partial v_i} \qquad (18)$$

where K_1 and K_i', $i = 1,2,....p$ are m x l dimensional feedback gain matrices and the quadratic performance

index to be minimised is

$$J = \frac{1}{2}\left\{\langle y,\ Qy\rangle + \langle u,\ Ru\rangle + \sum_{i=1}^{p} \langle \frac{\partial y}{\partial v_i},\ H_i \frac{\partial y}{\partial v_i}\rangle\right\} \qquad (19)$$

where $R > 0$, Q, $H_i \geq 0$, $i = 1,2,\ldots,p$. Letting

$$K_2 = [K_1', K_2', \ldots\ldots K_p'],$$

$$q = [\frac{\partial y}{\partial v_1}^T, \frac{\partial y}{\partial v_2}^T, \ldots\ldots \frac{\partial y}{\partial v_p}^T]^T$$

and

$$H = \begin{bmatrix} H_1 & & & O \\ & H_2 & & \\ & & \ddots & \\ O & & & H_p \end{bmatrix},$$

equations (18) and (19) become, respectively,

$$u = -K_1 y - K_2 q$$

$$J = \frac{1}{2}\left\{\langle y,\ Qy\rangle + \langle u,\ Ru\rangle + \langle q,\ Hq\rangle\right\}.$$

Also, from equation (17), we have

$$q = \frac{\partial W}{\partial v} u - WK_1 q$$

i.e. $$q = W_1 \frac{\partial W}{\partial v} u \qquad (21)$$

where

$$\frac{\partial W}{\partial v} = \begin{bmatrix} \frac{\partial W}{\partial v_1} \\ \frac{\partial W}{\partial v_2} \\ \vdots \\ \frac{\partial W}{\partial v_p} \end{bmatrix}, \quad W_1 = (I + WK_1)^{-1}; \qquad (21a)$$

W_1 exists, since $I + WK_1$ is an Volterra operator and is therefore invertible.

Consider arbitrary perturbations δK_1 and δK_2 in in feedback gain matrices K_1 and K_2; the first order change δJ in J is

$$\delta J = \langle \delta y, Qy \rangle + \langle \delta u, Ru \rangle + \langle \delta q, Hq \rangle \qquad (22)$$

and the corresponding first order change δu, δy and δq in u, y and q are given, respectively, by

$$\delta u = -(\delta K_1 y + \delta K_2 q) - (K_1 \delta y + K_2 \delta q) \qquad (23)$$

$$\delta y = W \delta u \qquad (24)$$

$$\delta q = \frac{\partial W}{\partial v} \delta u - W \delta K_1 q - W K_1 \delta q$$

$$= W_1 \frac{\partial W}{\partial v} \delta u - W_1 W \delta K_1 q \ , \qquad (25)$$

where W_1 is defined in equation (21a). Substitution of (23) into equation (22) yields

$$\delta J = \langle \delta y, Qy \rangle + \langle \delta q, Hq \rangle - \langle K_1 \delta y, Ru \rangle - \langle K_2 \delta q, Ru \rangle$$

$$- \langle \delta K_1 y, Ru \rangle - \langle \delta K_2 q, Ru \rangle$$

$$= \langle \delta y, Qy - K_1^* Ru \rangle + \langle \delta q, Hq - K_2^* Ru \rangle$$

$$- \langle \delta K_1 y, Ru \rangle - \langle \delta K_2 q, Ru \rangle$$

$$= \langle \delta y, z_1 \rangle + \langle \delta q, z_2 \rangle - \langle \delta K_1 y_1, Ru \rangle - \langle \delta K_2 q, Ru \rangle$$

where

$$z_1 = Qy - K_1^* Ru$$

$$z_2 = Hq - K_2^* Ru.$$

Using equations (24) and (25), the above equation can be written as

$$\delta J = \langle \delta u, W^* z_1 \rangle + \langle \delta u, (\frac{\partial W}{\partial v})^* W_1^* z_2 \rangle$$

$$- \langle \delta K_1 q, W^* W_1^* z_2 \rangle - \langle \delta K_1 y, Ru \rangle - \langle \delta K_2 q, Ru \rangle$$

$$= \langle \delta u,\ W^* z_1 + (\frac{\partial W^*}{\partial v}) W_1^* z_2 \rangle - \langle \delta K_1,\ [q,\ W^* W_1^* z_2] \rangle$$

$$- \langle \delta K_1,\ [y,\ Ru] \rangle - \langle \delta K_2,\ [q,\ Ru] \rangle .$$

By substituting equations (24) and (25) into equation (23), we have

$$\delta u = -(\delta K_1 y + \delta K_2 q) - (K_1 W \delta u + K_2 W_1 \frac{\partial W}{\partial v} \delta u - K_2 W_1 W \delta K_1 q)$$

i.e. $$\delta u = T(K_2 W_1 W \delta K_1 q - \delta K_1 y - \delta K_2 q)$$

where

$$T = (I + K_1 W + K_2 W_1 \frac{\partial W}{\partial v})^{-1} .$$

Therefore,

$$\delta J = \langle T(K_2 W_1 W \delta K_1 q - \delta K_1 y - \delta K_1 q),\ W^* z_1 + (\frac{\partial W}{\partial v})^* W_1^* z_2 \rangle$$

$$- \langle \delta K_1,\ [q,\ W^* W_1^* z_2] \rangle - \langle \delta K_1,\ [y,\ Ru] \rangle$$

$$- \langle \delta K_2,\ [q,\ Ru] \rangle$$

and, after some rearrangement of terms, we obtain

$$\delta J = \langle \delta K_1,\ [y,\ RK_1 y] \rangle + \langle \delta K_1,\ [q,\ (K_2 W_1 W)^* T^*$$

$$(W^* z_1 + (W_1 \frac{\partial W^*}{\partial v}) z_2) - (W_1 W)^* z_2] \rangle$$

$$- \langle \delta K_1,\ [y,\ T^*(W^* z_1 + (W_1 \frac{\partial W^*}{\partial v}) z_2) - RK_2 q] \rangle$$

$$+ \langle \delta K_2,\ [q,\ RK_2 q] \rangle - \langle \delta K_2,\ [q,\ T^*(W^* z_1 +$$

$$(W_1 \frac{\partial W^*}{\partial v}) z_2) - RK_1 y] \rangle$$

$$= \langle \delta K \cdot [y,\ RK\ y] - \Phi_y \rangle + \langle \delta K\ ,\ [q,\ RK\ q] - \Phi_q \rangle$$

where

$$\Phi_y \triangleq [y,\ T^*(W^* z_1 + (W_1 \frac{\partial W^*}{\partial v}) z_2) - RK_2 q]$$

$$- [q,\ (K_2 W_1 W)^* T^* (W^* z_1 + (W_1 \frac{\partial W^*}{\partial v}) z_2) - (W_1 W)^* z_2]$$

and $\Phi_q \triangleq [q,\ T^*(W^*z_1 + (W_1\frac{\partial W^*}{\partial v})z_2) - RK_1y]$.

To achieve optimality we must have $\delta J = 0$ for arbitrary changes δK_1 and δK_2 . This means that

$$[y,\ RK_1y] - \Phi_y = 0 \qquad (26a)$$

and $$[q,\ RK_2q] - \Phi_q = 0\ . \qquad (26b)$$

Since R, K_1 and K_2 are constant matrices, by using the result given in the Introduction, we have

$$[y,\ RK_1y] = RK_1[y,\ y] = RK_1G_y$$

and $$[q,\ RK_2q] = RK_2[q,\ q] = RK_2G_q$$

where

$$G_y \triangleq [y,\ y] \text{ and } G_q = [q,\ q].$$

Thus, equations (26a) and (26b) can be written as

$$RK_1G_y - \Phi_y = 0 \qquad (27a)$$

$$RK_2G_q - \Phi_q = 0\ . \qquad (27b)$$

It will be assumed that G_y and G_q are invertible which is valid provided the components of y and q are independent. Note the close resemblance of equations (27a), (27b) and equation (5) for the parallel model.

We now suggest the following algorithms to find the optimal K_1 and K_2 .

Algorithm 4 : The basic algorithm

$$K_1^i = R^{-1}\Phi_y^i(G_y^{-1})^i$$

$$K_2^i = R^{-1}\Phi_q^i(G_q^{-1})^i$$

Algorithm 5 :

$$K_1^{i+1} = (1 - \epsilon_1)K_1^i + \epsilon_1 R^{-1}\Phi_y^i(G_y^{-1})^i \qquad 0 < \epsilon_1 \leqslant 1$$

$$K_2^{i+1} = (1 - \epsilon_2)K_2^i + \epsilon_2 R^{-1}\Phi_q^i(G_q^{-1})^i \qquad 0 < \epsilon_2 \leqslant 1$$

Algorithm 6 :

$$K_1^{i+1} = (1-\epsilon_1)K_1^i + \epsilon_1(R+M)^{-1}[\Phi_y^i(G_y^{-1})^i + MK_1^i] \qquad 0 < \epsilon_1 \leqslant 1$$

$$K_2^{i+1} = (1-\epsilon_2)K_2^i + \epsilon_2(R+M)^{-1}[\Phi_q^i(G_q^{-1})^i + MK_2^i] \qquad 0 < \epsilon_2 \leqslant 1$$

where $M \geqslant 0$.

Note the similarity between Algorithm 2 and 5 and Algorithm 3 and 6. Again, Algorithm 6 is reduced to Algorithm 5 if $M = 0$ and, if in addition $\epsilon_1 = \epsilon_2 = 1$, reduced to the basic Algorithm 4. Therefore, following the arguments used in Lemma 1 and 2, it can be easily proved that Algorithm 5 and 6 always lead to a non-zero reduction in performance index in each iteration. Likewise, using the modified Polak's lemma presented earlier, Algorithm 5 and 6 can be proved to converge to a stationary point.

4. REFERENCES

[1] R.E. Kalman, Contribution to the Theory of Optimal Control, Bol. Soc. Mat. Mex., Vol. 5, 1960, pp. 102-119.

[2] D.H. Mee, A weighting Function Approach to Linear Control Systems Design, Ph.D. Thesis, Univ. of London, 1969.

[3] J.F. Cassidy, Jr and I. Lee, On the Optimal Feedback Control of a Large Launch Vechicle to Reduce Trajectory Sensitivity, Preprints, J.A.C.C. Philadelphia, Pa, June 1967, pp. 587-595.

[4] A.J. Bradt, Sensitivity Functions in the Design of Optimal Controllers, IEEE Tran. Automatic Control (Correspondence), Vol. AC-13, February 1968, pp. 110-113.

[5] N. Dunford and J.T. Schwartz, Linear Operators, Vol. II, Interscience, Wiley.

[6] E. Polak and G. Ribiere, Note Sur La Convergance De Mithodes De Directions Conjugnees, to be published.

CHAPTER 4

FREQUENCY DOMAIN APPROXIMATION TECHNIQUE FOR MODELLING

1. INTRODUCTION

It is often desirable to reduce the order of the transfer function of a complex system, for analysis and design purposes. In the classical approach where the problems are usually formulated in the frequency domain, there are several rules which have been well-known: (a) Chen[1] suggested that, in the Bode diagram, the corner frequency can be ignored if the gains are below -15db or above +15db; (b) Evans[2] suggested that only the dominant poles of the system are needed in estimating the transients of a closed loop transfer function. All these techniques are based on the concept that the dynamical behaviour of the system is determined mainly by the poles nearest to the imaginary axis i.e. dominant poles. However, since many practical control systems do not have dominant poles, the above methods cannot be used in general. In 1968, Chen and Shieh [3] suggested a method for reducing the order of transfer functions, using the technique of continued fraction expansion (CFE). Regarding the control system as a low pass filter, the approach arranges the numerator and denominator polynomials of the transfer function in the ascending order in the power of s and starts the CFE from the constant terms of the transfer function; this, in effect, is equivalent to Taylor series expansion about $s = 0$ and can be shown to be a special form of Padé approximation for scalar rational function[4]. The method assures that the model gives the correct steady-state response for constant input. However, the approximation

to the initial transient response may not be good. Furthermore, like many other approximation methods based on power series expansions, the stability of the model is not guaranteed, even if the original system is stable.

In this Chapter, a different form of CFE will be suggested for model simplification which yields good approximations to both the initial transient and steady-state responses. For convenience, we shall, in the sequel, write CFE1 and CFE2 to represent, respectively, the CFE suggested by Chen and that suggested by the author. It will be indicated that, by using the concept of time moments of a impulse response, the model obtained by CFE1 will, in general, give the correct steady-state response for a class of inputs, namely $\alpha_i t^i$, i= 0,1,2,....p where p is a positive integer depending on the order of the model. This will be followed by a proof that the power series expansions, about s= 0 and s=∞, of the low-order model obtained by CFE2 agree with the corresponding series, about s= 0 and s=∞, of the original transfer function up to certain terms. An approximation technique which differs from the ordinary Padé approximation will be proposed; this technique which will be called 'Modified'Pade Approximation (MPA) will ensure good approximation to both the initial transient and steady-state responses. And it will be shown that, under certain conditions, the approximations, of the same order, obtained by CFE2 and MPA are the same. Similar results for a transfer function matrix will be presented. This result should facilitate the computation of the low-order model, especially for

the multivariable systems. It is also shown by example (ii) of Section 6 that the CFE2 can sometimes be used to provide a stable, reasonably good model when CFE1 gives an unstable model.

2. CONTINUED FRACTION EXPANSION TO GIVE GOOD APPROXIMATION TO BOTH THE INITIAL TRANSIENT AND STEADY-STATE RESPONSES

(a) Scalar Case

We shall here consider the general transfer function C(s) of a single-input single-output system

$$C(s) = \frac{A_{21} + A_{22}s + \ldots\ldots\ldots + A_{2n}s^{n-1}}{A_{11} + A_{12}s + \ldots\ldots\ldots + A_{1,n+1}s^{n}} .$$

Expansions from the constant terms in the CFE1 are equivalent to Taylor series expansion about s= 0. Thus, C(s) is expanded in the following way

$$C(s) = \cfrac{1}{\cfrac{A_{11}}{A_{21}} + s\,\cfrac{A_{31} + A_{32}s + \ldots\ldots\ldots\ldots}{A_{21} + A_{22}s + \ldots\ldots\ldots\ldots}}$$

$$= \cfrac{1}{\cfrac{A_{11}}{A_{21}} + \cfrac{s}{\cfrac{A_{21}}{A_{31}} + s\,\cfrac{A_{41} + A_{42}s + \ldots\ldots}{A_{31} + A_{32}s + \ldots\ldots}}}$$

$$= \cfrac{1}{h_1 + \cfrac{s}{h_2 + \cfrac{s}{h_3 + \cfrac{s}{h_4 + \cfrac{s}{h_5 + \ddots}}}}} \qquad (1)$$

where

$$A_{31} = \frac{A_{21}A_{12} - A_{11}A_{22}}{A_{21}} \qquad A_{41} = \frac{A_{22}A_{31} - A_{32}A_{21}}{A_{31}}$$

$$A_{42} = \frac{A_{23}A_{31} - A_{33}A_{21}}{A_{31}}$$

$$A_{3,n-1} = \frac{A_{21}A_{1n} - A_{11}A_{2n}}{A_{21}} \qquad A_{4,n-1} = \frac{A_{2n}A_{31} - A_{3n}A_{21}}{A_{31}}$$

$$A_{3,n} = A_{1,n+1}$$

and $h_1 = \frac{A_{11}}{A_{21}}$, $h_2 = \frac{A_{21}}{A_{31}}$, $h_3 = \frac{A_{31}}{A_{41}}$ etc.

To obtain good approximations to both initial transient and steady-state responses, Taylor series expansions should be carried out about $s = 0$ and $s = \infty$. One way of achieving this is to have continued fraction expansion from the constant terms first and then the highest order terms yielding, say.

$$C(s) = \frac{1}{\frac{A_{11}}{A_{21}} + s\,\frac{A_{3n}s^{n-1} + A_{3,n-1}s^{n-2} + \ldots\ldots + A_{31}}{A_{2n}s^{n-1} + A_{2,n-1}s^{n-2} + \ldots\ldots + A_{21}}}$$

$$= \cfrac{1}{\frac{A_{11}}{A_{21}} + \cfrac{s}{\frac{A_{2n}}{A_{3n}} + \cfrac{A_{41} + A_{42}s \ldots + A_{4,n-1}s^{n-2}}{A_{31} + A_{32}s + \ldots\ldots + A_{3n}s^{n-1}}}}$$

where

$$A_{31} = \frac{A_{21}A_{12} - A_{11}A_{22}}{A_{21}} \qquad A_{41} = \frac{A_{21}A_{3n} - A_{31}A_{2n}}{A_{3n}}$$

$$A_{4,2} = \frac{A_{22}A_{3n} - A_{32}A_{2n}}{A_{3n}}$$

$$A_{3,n-1} = \frac{A_{21}A_{1n} - A_{11}A_{2n}}{A_{21}}$$

$$A_{4,n-1} = \frac{A_{2,n-1}A_{3n} - A_{3,n-1}A_{2n}}{A_{3n}}$$

$$A_{3n} = A_{1,n+1}$$

By repeating this sequence of expansions, C(s) can be expanded as

$$C(s) = \cfrac{1}{h_1 + \cfrac{s}{h_2 + \cfrac{1}{h_3 + \cfrac{s}{h_4 + \cfrac{1}{\ddots}}}}} \qquad (2)$$

where $h_1 = \frac{A_{11}}{A_{21}}$, $h_2 = \frac{A_{2n}}{A_{3n}}$, $h_3 = \frac{A_{31}}{A_{41}}$ etc.

If a p-order model is required, only the first 2p quotients h_i in either expansion (1) or (2) have to be known and the desired simplified model is obtained by applying the inverse procedure to the truncated continued fraction.

(b) Matrix Case

Chen has recently extended his method of model simplification using CFE to multivariable systems i.e. C(s) is a transfer function matrix. Let C(s) be a m x m transfer function matrix. Chen showed that, analogous to the scalar case, the CFE of C(s) is of the form

$$C(s) = [H_1 + s[H_2 + s[H_3 + s[H_4 + \ldots]^{-1}]^{-1}]^{-1}]^{-1} \qquad (3)$$

where H_1, H_2, ... in the above equation are m x m

constant matrices. It is obvious that equation (3) can be written as

$$C^{-1}(s) = H_1 + s[H_2 + s[H_3 + s[H_4 + \ldots]^{-1}]^{-1}]^{-1} \quad (4)$$

provided $C^{-1}(s)$ exists. H_1 can be found by equating the constant matrices on both sides of the above equation; H_2, H_3, ... can be found similarly by inverting and then equating the constant terms on both sides of anequation similar to equation (4). We shall now proceed to show how the CFE2 can also be extended to multivariable systems. It is easy to see that the transfer function matrix C(s) can be expanded, formally, into

$$C(s) = [H_1 + s[H_2 + [H_3 + s[H_4 + \ldots]^{-1}]^{-1}]^{-1}]^{-1}. \quad (5)$$

To obtain the correct steady-state response, the constant matrices H_1, H_3, H_5,... in equation (5) are obtained in the similar way as the corresponding matrices in equation (4); however, to ensure good initial transient response approximation, H_2, H_4, ... are obtained by inverting and then equating the highest order terms on both sides of an equation similar to equation (5). A low-order model can then be obtained by inverse procedure, after H_1, H_2, ... H_n, say, have been found.

However, the above procedure only works if the transfer function matrix C(s) is invertible. A necessary condition for the invertibility of C(s) is, obviously, that the number of inputs to the system and the number of outputs should be the same, which is often untrue in practice.

3. PADÉ APPROXIMATION FOR SCALAR RATIONAL FUNCTIONS

For convenience, a brief description of the ordinary Padé approximation for a function C(s) of the complex variable s is given below; for further information the reader is referred to Wall [4]. The function C(s) is assumed to be analytic about $s = s_a$, a point of interest; the points which we are most interested are s = 0 and $s = \infty$.

For C(s) analytic about s = 0, we can expand it into a power series about s = 0, i.e.

$$C(s) = c_o + c_1 s + c_2 s^2 + \ldots\ldots \qquad (6)$$

where c_o, c_1, c_2, ... are known. Suppose that an approximation to C(s) is to be obtained by Padé approximation. Then, it has been shown in [4] that, with each ordered pair (p,q) of nonnegative integers, there is associated a uniquely determined rational function approximation $C_{p,q}(s)$, whose numerator and denominator are of degree not exceeding q and p respectively and whose expansion in ascending power of s agrees term by term with the power series $\sum_{i=0}^{\infty} c_i s^i$ of C(s) for more terms than that of any other such rational functions. In fact, the power series of C(s) and $C_{p,q}(s)$ will agree up to the s^{p+q} term. This means that, if the rational function approximation $C_{p,q}(s)$ is expanded into a power series

$$C_{p,q}(s) = t_o + t_1 s + t_2 s^2 + \ldots\ldots$$

then, we have, in general, $t_o = c_o$, $t_1 = c_1$, ... $t_{p+q} = c_{p+q}$

$t_{p+q+1} \neq c_{p+q+1}$, ... By arranging these rational function approximations $C_{p,q}(s)$ to $C(s)$ in a table of double entry with $C_{p,q}(s)$ in the (p+1) row and (q+1) column of the table (p,q = 0,1,2, ...), we obtain the well-known Padé table.

It is trivial to prove that an approximation $C_1(s)$ to $C(s)$ obtained by CFE1 is, in general, a rational function with p-q = 1 and forms the subdiagonal sequence of the Padé table (theorem 96.1, p.380 in [4]); i.e.

$$C_1(s) = C_{p,p-1}(s)$$

$$= \frac{a_o + a_1 s + a_2 s^2 + \ldots\ldots + a_{p-1} s^{p-1}}{b_o + b_1 s + b_2 s^2 + \ldots\ldots\ldots + b_p s^p}$$

$$= c_o + c_1 s + \ldots + c_{2p-1} s^{2p-1} + t_{2p} s^{2p} + t_{2p+1} s^{2p+1} + \ldots\ldots \qquad (7)$$

If $C(s)$ is the transfer function of a linear time-invariant system having the impulse response $c(t)$, we have, by using the power series for e^{-st},

$$C(s) \triangleq \int_o^\infty c(t) e^{-st}\, dt$$

$$= \int_o^\infty c(t)dt - s\int_o^\infty t c(t)dt + \frac{s^2}{2}\int_o^\infty t^2 c(t)dt + \ldots\ldots$$

By identifying the above equation with equation (6), we have $c_i = [(-1)^i/i!]\int_o^\infty t^i c(t)dt$ which is called the ith time moment of the impulse response $c(t)$. Hence, the coefficient of the s^i term in the power series expansion, about s = 0, of $C(s)$ is the ith time moment of $c(t)$. Since the first (p+q) terms in the power series expansion, about s = 0,

of C(s) and $C_1(s)$ are the same, their first (p+q) time moments are the same. It is obvious that the ith time moment of c(t) ($c_1(t)$) is the steady state response of c(t) ($c_1(t)$) to the input $[(-1)^i/i!]t^i$. Therefore, we conclude that the steady state response of C(s) and $C_{p,p-1}(s) = C_1(s)$ to the inputs $\alpha_i t^i$, i=0,1,2,....2p-1, where α_i are constants, are the same.

We shall now proceed to show that the power series expansions, about s = 0 and s =∞, of the low order model obtained by the CFE2 technique expressed in equation (2) agree with the corresponding power series, about s = 0 and s =∞, of the original transfer function C(s) up to certain terms. The result will be illustrated in example (ii) of Section 6.

It is assumed that C(s) is a rational proper transfer function of finite order having a general CFE of the form

$$C(s) = \cfrac{a_1 s}{b_1 + \cfrac{a_2 s}{b_2 + \cfrac{a_3 s}{b_3 + \cdots + \cfrac{a_{2n} s}{b_{2n}}}}} \qquad (8)$$

where n is a finite positive integer; C(s) can also be expanded into the power series

$$C(s) = p_o + p_1 s + p_2 s^2 + \ldots \text{ expanded about } s = 0$$
$$= q_o + q_1 s^{-1} + q_2 s^{-2} + \ldots \text{ expanded about } s = \infty. \qquad (9)$$

Define

$$A_1 \triangleq a_1 s$$
$$B_1 \triangleq b_1$$
$$A_2 \triangleq b_2 a_1 s$$
$$B_2 \triangleq b_1 b_2 + a_2 s$$

or in general

$$\begin{aligned} A_p(s) &= b_p A_{p-1}(s) + a_p s A_{p-2}(s) \\ B_p(s) &= b_p B_{p-1}(s) + a_p s B_{p-2}(s) \end{aligned} \tag{10}$$

assuming that

$$A_{-1} = 1, \quad B_{-1} = 0, \quad A_o = 0, \quad B_o = 1.$$

Using equation (10) iteratively, we have

$$A_{2(p+1)} B_{2p} - A_{2p} B_{2(p+1)} = (-1)^{2p} a_1 a_2 \ldots a_{2p+1} b_{2(p+1)} s^{2p+1} \tag{11}$$

It is easy to see that equation (8) can be reduced to the CFE2 of equation (2) by letting

$$a_1 = a_3 = \ldots = a_{2n-1} = \frac{1}{s}$$
$$a_2 = a_4 = \ldots = a_{2n} = 1$$
$$b_1 = h_1, \; b_2 = h_2, \; \ldots \; b_{2n} = h_{2n}.$$

Substituting these a_i, b_i into equation (11) yields immediately

$$\begin{aligned} A_{2(p+1)} B_{2p} - A_{2p} B_{2(p+1)} &= a_2 a_4 \ldots a_{2p} b_{2(p+1)} s^p \\ &= b_{2(p+1)} s^p. \end{aligned}$$

Also, we have

$$\begin{aligned} \frac{1}{B_{2p} B_{2(p+1)}} &= \frac{1}{\theta_o + \theta_1 s + \theta_2 s^2 + \ldots} \\ &= \alpha_o + \alpha_1 s + \alpha_2 s^2 + \ldots \end{aligned}$$

Thus, the difference between the two approximations

$\frac{A_{2(p+1)}}{B_{2(p+1)}}$ and $\frac{A_{2p}}{B_{2p}}$ is given by

$$\frac{A_{2(p+1)}}{B_{2(p+1)}} - \frac{A_{2p}}{B_{2p}} = \frac{A_{2(p+1)}B_{2p} - A_{2p}B_{2(p+1)}}{B_{2p}B_{2(p+1)}}$$

$$= b_{2(p+1)}s^{p}(\alpha_o + \alpha_1 s + \alpha_2 s^2 + \ldots)$$

$$= \sum_{i=o}^{\infty} \beta_i s^{p+i} \qquad (12)$$

Therefore, it follows that the power series expansions, about s=0, of $\frac{A_{2(p+1)}}{B_{2(p+1)}}$ and $\frac{A_{2p}}{B_{2p}}$ obtained by the CFE2 technique agree up to the s^{p-1} term.

We shall, now, show that the power series expansion, about $s = \infty$, of $\frac{A_{2(p+1)}}{B_{2(p+1)}}$ and $\frac{A_{2p}}{B_{2p}}$ agree up to the s^{-p} term. Obviously, equation (2) can also be written as

$$C(s) = \cfrac{\frac{1}{s}}{\frac{h_1}{s} + \cfrac{1}{h_2 + \cfrac{\frac{1}{s}}{\frac{h_3}{s} + \cfrac{1}{h_4 + \ddots + \frac{1}{h_{2n}}}}}}$$

and equation (8) can be reduced to the above equation by putting

$$a_1 = a_3 = \ldots\ldots\ldots\ldots = a_{2n-1} = \frac{1}{s^2}$$

$$a_2 = a_4 = \ldots\ldots\ldots\ldots\ldots = a_{2n} = \frac{1}{s}$$

$$b_1 = \frac{h_1}{s}, \ldots\ldots\ldots\ldots, b_{2n-1} = \frac{h_{2n-1}}{s}$$

$$b_2 = h_2, \ldots\ldots\ldots\ldots\ldots, b_{2n} = h_{2n}.$$

For the above a_i, b_i, we have, from equation (11)

$$A_{2(p+1)}B_{2p} - A_{2p}B_{2(p+1)}$$

$$= (-1)^{2p}\left(\frac{1}{s^2}\right)^{p+1}\left(\frac{1}{s}\right)^{p} h_{2(p+1)}s^{2p+1}$$

$$= h_{2(p+1)}s^{-(p+1)}.$$

Also

$$\frac{1}{B_{2p}B_{2(p+1)}} = \frac{1}{\lambda_1 + \lambda_2 s^{-1} + \ldots\ldots\ldots}$$

$$= \gamma_o + \gamma_1 s^{-1} + \ldots\ldots\ldots$$

Hence,

$$\frac{A_{2(p+1)}}{B_{2(p+1)}} - \frac{A_{2p}}{B_{2p}} = \frac{A_{2(p+1)}B_{2p} - A_{2p}B_{2(p+1)}}{B_{2p}B_{2(p+1)}}$$

$$= h_{2(p+1)} s^{-(p+1)} (\gamma_o + \gamma_1 s^{-1} + \gamma_2 s^{-2} + \ldots)$$

$$= \sum_{i=o}^{\infty} h_{2(p+1)} \gamma_i s^{-(p+1+i)} \qquad (13)$$

This means that the power series expansion about $s=\infty$, of $\frac{A_{2(p+1)}}{B_{2(p+1)}}$ and $\frac{A_{2p}}{B_{2p}}$ agree up to the s^{-p} term.

Since, by definition, the original rational proper transfer function $C(s) \triangleq \frac{A_{2n}}{B_{2n}}$ (see equation (8)), we have, for $1 \leqslant p \leqslant n$,

$$C(s) - \frac{A_{2p}}{B_{2p}}$$

$$= \frac{A_{2n}}{B_{2n}} - \frac{A_{2p}}{B_{2p}}$$

$$= \left(\frac{A_{2n}}{B_{2n}} - \frac{A_{2(n-1)}}{B_{2(n-1)}}\right) + \left(\frac{A_{2(n-1)}}{B_{2(n-1)}} - \frac{A_{2(n-2)}}{B_{2(n-2)}}\right)$$

$$+ \ldots\ldots + \left(\frac{A_{2(p+1)}}{B_{2(p+1)}} - \frac{A_{2p}}{B_{2p}}\right)$$

$$= \sum_{i=o}^{\infty} (\beta_{2n})_i s^{(n-1+i)} + \sum_{i=o}^{\infty} \left(\beta_{2(n-1)}\right)_i s^{(n-2+i)}$$

$$+ \ldots\ldots + \sum_{i=o}^{\infty} \left(\beta_{2(p+1)}\right)_i s^{p+i)}$$

$$= \sum_{i=o}^{\infty} \delta_i s^{p+i}$$

by using equation (12). Similarly, by using equation (13) we have

$$C(s) - \frac{A_{2p}}{B_{2p}}$$

$$= \frac{A_{2n}}{B_{2n}} - \frac{A_{2p}}{B_{2p}}$$

$$= \sum_{i=o}^{\infty} \xi_i s^{-(p+1+i)} .$$

Hence, the power series expansion, about s=o (about $s=\infty$), of the low order approximation $\frac{A_{2p}}{B_{2p}}$ obtained by CFE2 of equation (2) agree with the corresponding series of the original rational proper transfer function C(s) up to the s^{p-1} term (s^{-p} term).

4. MODIFIED PADÉ APPROXIMATION

From the discussions given in Section 3, it is obvious that, like the CFE1, the PA is, essentially, an approximation about one point i.e. s=0 or $s=\infty$, if desired. We shall now propose an approximation technique which is based on the power series expansion about s=0 and $s=\infty$ and will therefore be called the 'Modified' Padé approximation (MPA). This method will ensure good approximations to both the initial transient and steady state responses. It will also be proved that, under certain condition, the approximations, of the same order, obtained by CFE2 and MPA are the same.

MPA may be defined as a approximation method which associates, with each ordered pair (p,q) of nonnegative integers, a rational function approximation $C_1(s)$ whose series expansions in both ascending and descending power

of s agree with the corresponding power series of the given function C(s) up to certain term. The scalar rational transfer function C(s) of the system is assumed to be proper and has the form

$$C(s) = \frac{a_o + a_1 s + \ldots\ldots + a_{n-1} s^{n-1}}{b_o + b_1 s + \ldots\ldots\ldots + b_n s^n} \quad ;$$

it is assumed to have the following power series expansions

$$C(s) = p_o + p_1 s + p_2 s^2 + \ldots\ldots \text{ expanded about } s=0$$

$$= q_o + q_1 s^{-1} + q_2 s^{-2} + \ldots\ldots \text{ expanded about } s=\infty$$

Suppose that C(s) is to be approximated by a low-order model $C_1(s)$ of the form

$$C_1(s) = \frac{r_o + r_1 s + \ldots + r_{m-1} s^{m-1}}{d_o + d_1 s + \ldots + d_m s^m}$$

where $m<n$, by using the MPA defined above. Without loss of generality, we can put $d_o = 1$ so that there remain 2m coefficients r_i, d_i to be found. One of the models that could be obtained by the MPA technique will have the power series expansion, about s=o (about $s=\infty$), agree with the corresponding power series, about s=0, (about $s=\infty$) of C(s) up to the s^{m-1} (s^{-m}) term. This can be accomplished by having the coefficients r_i, d_i of the model C(s) to satisfy the following set of 2m linear equations

$$
\begin{aligned}
r_o &= p_o \\
r_1 &= p_1 + d_1 p_o \\
&\vdots \\
r_{m-1} &= p_{m-1} + d_1 p_{m-2} + \cdots + d_{m-1} p_o \\
r_{m-1} &= d_m q_1 \\
r_{m-2} &= d_m q_2 + d_{m-1} q_1 \\
&\vdots \\
r_o &= d_m q_m + d_{m-1} q_{m-1} + \cdots + d_1 q_1
\end{aligned}
\qquad (14)
$$

This model will have equal emphasis on its approximations to the initial transient and steady state responses of C(s). It is easy to see that models with 'biased' emphasis can also be obtained by using the MPA, though it is not possible with the ordinary PA. For example, a model $C_1(s)$ to have more emphasis on the steady state response approximation than on the initial transient response approximation can be obtained by finding its coefficients r_i, d_i in the following set of 2m linear equations

$$
\begin{aligned}
r_o &= p_o \\
r_1 &= p_1 + d_1 p_o \\
&\vdots \\
r_{m-1} &= p_{m-1} + d_1 p_{m-2} + \cdots + d_{m-1} p_o \\
0 &= p_m + d_1 p_{m-1} + \cdots + d_{m-1} p_1 \\
&\vdots \\
0 &= p_k + d_1 p_{k-1} + \cdots + d_{m-1} p_{k-(m-1)}
\end{aligned}
\qquad (15)
$$

$$r_{m-1} = d_m q_1$$

$$r_{m-2} = d_m q_2 + d_{m-1} q_1$$

$$\vdots$$

$$r_{-m+k+1} = d_m q_{2m-(k+1)} + \cdots + d_{-m+k+2} q_1$$

Where $m < k$. By 'fitting' more coefficients in the power series expansion about $s=0$, the $C_1(s)$ given by equation (15) would, usually, give a better approximation to the steady state response than $C_1(s)$ given by equation (14), though its initial transient response approximation may not be as good.

It is obvious that the d_i, $i=1,2, \ldots m$ in equation (14) can be found by solving the following equation

$$\underline{P} = T\,\underline{d} \qquad (16)$$

where

$$\underline{d} = (d_1, d_2, \ldots d_m)'$$

$$\underline{P} = (p_o, p_1, \ldots p_{m-1})'$$

$$T = \begin{bmatrix} q_1 & q_2 & \cdots & q_{m-1} & q_m \\ -p_o & q_1 & & & q_{m-1} \\ -p_1 & -p_o & q_1 & & q_{m-2} \\ \vdots & & \ddots & & \vdots \\ -p_{m-2} & -p_{m-3} & \cdots & -p_o & q_1 \end{bmatrix}$$

If the matrix T has an inverse, $\underline{d}$ is uniquely determined and so is $\underline{r}$ where $\underline{r} = (r_0, r_1, \ldots r_{m-1})'$ is given by equation (14). Then the model $C_1(s)$ with these r_i and d_i as its numerator and denominator coefficients respectively is thus unique.

This result together with the relation between the CFE2 and the power series expansion, about s=0 an s=∞, of the function C(s), proved in Section 3 leadsto the conclusion that, provided the matrix T in equation (16) is invertible, the low-order approximation obtained by the CFE2 can also be obtained by the MPA technique of equation (14) and vice versa. This fact will be illustrated in a numerical example of Section 6.

However, if the matrix $\underline{T}$ is not invertible, there exist other models of the same order, besides the model obtained by the CFE2, whose coeficients a_i, b_i satisfy the equation (14). Among these models, some may be stable whereas others including the one obtained by the CFE2 may not be stable.

5. LOW-ORDER APPROXIMATIONS FOR TRANSFER FUNCTION MATRICES.

It is easy to see that the PA and MPA described in Section 3 and 4, for a scalar transfer function, can be easily extended to a transfer function matrix; the scalar coefficients in the power series expansions, about s=0 and s=∞, will now be replaced by constant matrices of appropriate dimension and a low-order multivariable model is obtained by 'fitting' these constant matrices up to certain power of s. Obviously,

the PA and MPA techniques can be applied to non-square transfer function matrices whereas the CFE1 and CFE2 cannot. Furthermore, as no matrix inversion is involved in using the PA or MPA, they are, computationally, more attractive than the CFE1 or CFE2, especially for systems with large numbers of inputs and outputs.

We shall now proceed to show that the relation shown in Section 3 for the scalar case between the power series expansions, about $s=0$ and $s=\infty$, of the original transfer function and that of the low-order model obtained by the CFE1 or CFE2 can be easily extended to a rational proper transfer function matrix $G(s)$; $G(s)$ is assumed to be invertible and has a CFE1 and CFE2.

Analogous to equation (8), the rational proper transfer function matrix $G(s)$ is assumed to have a general CFE of the form

$$G(s) = A_1 s\,[B_1 + A_2 s[B_2 + \ldots + A_{2n} s[B_{2n}]^{-1}..]^{-1}]^{-1} \tag{17}$$

and its power series expansions exist

$$\begin{aligned} G(s) &= P_o + P_1 s + P_2 s^2 + \ldots \text{ expanded about } s=0 \\ &= Q_o + Q_1 s^{-1} + Q_2 s^{-2} + \ldots\ldots \text{ expanded about } s=\infty. \end{aligned} \tag{18}$$

Define

$$\begin{aligned} E_1 &\triangleq A_1 s \\ F_1 &\triangleq B_1 \\ E_2 &\triangleq B_2 A_1 s \\ F_2 &\triangleq B_1 B_2 + A_2 s \\ &\vdots \end{aligned}$$

or in general

$$E_p(s) = B_p E_{p-1}(s) + A_p s E_{p-2}(s)$$

$$F_p(s) = B_p F_{p-1}(s) + A_p s F_{p-1}(s) \qquad (19)$$

assuming that

$$E_{-1} = I,\ F_{-1} = 0,\ E_0 = 0,\ F_0 = I$$

It is easy to see that the equation (17) can be reduced to equation (3) by setting

$$A_1 = \frac{I}{s} I,\ A_2 = A_3 = \ldots = A_{2n} = I,$$

$$B_1 = H_1,\ B_2 = H_2,\ \ldots\ldots\ B_{2n} = H_{2n}.$$

With the F_p, E_p defined in equation (19), and F_p assumed to be invertible, a low-order model M_p can be obtained where

$$M_p = F_p^{-1} E_p.$$

For p = 1,2, ... 2n, we have

$$M_1 = F_1^{-1} E_1 = H_1^{-1}$$

$$M_2 = F_2^{-1} E_2 = [H_2H_1 + s\,I]^{-1} H_2$$

$$= [H_1 + sH_2^{-1}]^{-1}$$

$$M_3 = F_3^{-1} E_3$$

$$= [H_3\,[H_2H_1 + sI] + sH_1]^{-1}\,[H_3H_2 + sI]$$

$$= [\,[H_3H_2 + sI]H_1 + sH_3]^{-1}\,[H_3H_2 + sI]$$

$$= [H_1 + s\,[H_3H_2 + sI]^{-1} H_3]^{-1}$$

$$= [H_1 + s\,[H_2 + sH_3^{-1}]^{-1}\,]^{-1}$$

$$\vdots$$

$$M_{2n} = F_{2n}^{-1} E_{2n}$$

$$= [H_1 + s[H_2 + s[H_3 + \ldots + s[H_{2n}]^{-1} \ldots]^{-1}]^{-1}]^{-1}$$

$$= G(s)$$

These low-order models can also be obtained directly by the CFE1 by retaining only the first few constant matrices H_1, H_2, ... in equation (3). A low-order model M_{p+1} can be written as

$$M_{p+1} = F_{p+1}^{-1} E_{p+1}$$

$$= [H_{p+1}F_p + sF_{p-1}]^{-1}[H_{p+1} E_p + sE_{p-1}]$$

$$= [H_{p+1} F_p + sF_{p-1}]^{-1}[H_{p+1}F_pF_p^{-1}E_p + sF_{p-1}$$

$$F_p^{-1}E_p - sF_{p-1}F_p^{-1}E_p + sE_{p-1}]$$

$$= [H_{p+1}F_p + sF_{p-1}]^{-1} \left\{[H_{p+1}F_p + sF_{p-1}]\right.$$

$$\left. F_p^{-1} E_p - sF_{p-1}[F_p^{-1}E_p - F_{p-1}^{-1}E_{p-1}]\right\}$$

$$= F_p^{-1}E_p - sF_{p+1}^{-1} F_{p-1} [F_p^{-1}E_p - F_{p-1}^{-1} E_{p-1}]$$

$$= M_p - sF_{p+1}^{-1} F_{p-1} [M_p - M_{p-1}]$$

Thus, the difference between the two successive models M_{p+1} and M_p is equal to

$$M_{p+1} - M_p$$

$$= - sF_{p+1}^{-1} F_{p-1} (M_p - M_{p-1})$$

By repeating the same procedure on $M_p - M_{p-1}$ and so on, we have

$$M_{p+1} - M_p$$

$$= (-1)^2 F_{p+1}^{-1} F_{p-1} F_p^{-1} F_{p-2} (M_{p-1} - M_{p-2}) s^2$$

$$\vdots$$

$$= (-1)^p F_{p+1}^{-1} F_{p-1} F_p^{-1} F_{p-2} \cdots F_2^{-1} F_o F_1^{-1} s^p$$

It is trivial to show that the product $F_{p+1}^{-1} F_{p-1} F_p^{-1} F_{p-2} \cdots F_o F_1^{-1}$ where the F_i, $i = 0,1, \ldots p+1$ are defined in the equation (19) can be expressed as

$$F_{p+1}^{-1} F_{p-1} \cdots F_o F_1^{-1} = (-1)^p \sum_{i=o}^{\infty} L_i s^i$$

L_i being constant matrices. Hence, we have

$$M_{p+1} - M_p$$

$$= \sum_{i=o}^{\infty} L_i s^{i+p}.$$

Now, by definition, the original transfer function matrix $G(s) \triangleq M_{2n} = F_{2n}^{-1} E_{2n}$, the difference between $G(s)$ and a low-order model, say, M_p is equal to

$$G(s) - M_p$$

$$= M_{2n} - M_p$$

$$= (M_{2n} - M_{2n-1}) + (M_{2n-1} - M_{2n-2}) + \ldots + (M_{p+1} - M_p)$$

$$= \sum_{i=o}^{\infty} (L_{2n-1})_i s^{i+2n-1} + \sum_{i=o}^{\infty} (L_{2n-2})_i s^{i+2n-2} + \cdots$$

$$+ \sum_{i=o}^{\infty} (L_p)_i s^{i+p}$$

$$= \sum_{i=o}^{\infty} K_i s^{p+i}$$

Thus, we can conclude that the power series expansions, about $s=0$, of the low- order model M_p obtained by the

CFE1 of the equation (3) agrees with the corresponding power series expansion of the original transfer function matrix G(s) up to the s^{p-1} term. This fact will be illustrated in example (IV) of Section 6.

By following a similar procedure, we can also conclude that the power series expansions, about s=0 and s=∞, of the low-order model M_p obtained by the CFE2 of equation (5) agree with the corresponding power series expansions, about s=0 and s=∞, of the original transfer function matrix G(s) up to certain terms.

6. EXAMPLES

(i) The original system, taken from [6], is of 8th order and has transfer function

$$C(s) = \frac{a(s)}{b(s)}$$

where

$$a(s) = 18s^7 + 514s^6 + 5982s^5 + 363823s^4 + 122664s^3 + 222088s^2 + 185760s + 40320$$

$$b(s) = (s+1)\ (s+2)\ (s+3)\ (s+4)\ (s+5)\ (s+6)\ (s+7)\ (s+8).$$

Suppose that a 2nd order model is required. The CFE2 technique described in this chapter yields a model having the transfer function

$$C_2(s) = \frac{18s + 5.436}{s^2+7.7s+5436}$$

While the CFE1 introduced by Chen gives a model having the transfer function

$$C_1(s) = \frac{15.16s + 4.838}{s^2+6.022s+4.838}$$

Fig. 1 shows the transient portion of the step response of the original system (curve a), the model $C_1(s)$ (curve b) and the model $C_2(s)$ (curve c), which finally settle down to the same steady state. It can be seen that $C_2(s)$ gives a more accurate approximation to the initial transient than $C_1(s)$.

(ii) Given a control system having the transfer function

$$C(s) = \frac{2 + 6s + 8s^2}{(s+2)(s+1)^2}$$

the models obtained by using the CFE1 and CFE2 having the transfer functions

$$C_1(s) = \frac{1.776s + 0.222}{(s+0.1239)(-s+1.791)}$$

$$C_2(s) = \frac{8s + 7.6}{(s+2.1)^2 + (1.786)^2}$$

respectively. Fig. 2 shows the step response of C(s) (curve a), $C_1(s)$ (curve b) and $C_2(s)$ (curve c). The model $C_2(s)$ again gives good approximation to the initial transient and yields the correct steady-state response. Although its behaviour after the initial transiant is not a very good approximation, it is at least stable, whereas model $C_1(s)$, given by the CFE1, is unstable.

The original system C(s) in the example (ii) above has the following power series expansions

$$C(s) = \frac{8s^2+6s+2}{(s+2)(s+1)^2}$$

$$= 1 + \frac{s}{2} + \frac{3s^2}{4} + \ldots\ldots \quad \text{expanded about } s=0$$

$= 8s^{-1} - 26s^{-2} + 66s^{-3} + \ldots\ldots$ expanded about $s=\infty$,

and the corresponding power series expansions of $C_2(s)$ obtained by the CFE2 are

$$C_2(s) = \frac{8s + 7.6}{(s+2.1)^2 + (1.786)^2}$$

$$= 1 + \frac{s}{2} - \frac{3.1}{7.6}s^2 + \ldots\ldots \quad \text{expanded about } s=0$$

$$= 8s^{-1} - 26s^{-2} + 48.4s^{-3} + \ldots\ldots\ldots \quad \text{expanded about } s=\infty$$

It is easy to see that the power series expansions, about $s=0$ ($s=\infty$), of the 2nd order model $C_2(s)$ obtained by the CFE2 agree with the corresponding power series expansion, about $s=0$ (about $s=\infty$), of the original system $C(s)$ up to the s (s^{-2}) term, as shown in Section 3.

To obtain a 2nd order model, say, $C_2(s)$ by the MPA technique of equation (14) in Section 4, the 2 x 2 matrix T in equation (16) is found to be

$$T = \begin{bmatrix} 8 & -26 \\ -1 & 8 \end{bmatrix}$$

and its determinant $|T| = 8 \times 8 - (26) = 38 \neq 0$. Thus, T is invertible and the resulting 2nd order model $C_3(s)$ is unique. The coefficients d_i in the denominator of $C_3(s)$ can be found from equation (16) i.e.

$$\begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} 8 & -26 \\ -1 & 8 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ \frac{1}{2} \end{bmatrix}$$

$$= \frac{1}{38} \begin{bmatrix} 8 & 26 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} 1 \\ \frac{1}{2} \end{bmatrix}$$

$$= \frac{1}{38} \begin{bmatrix} 21 \\ 5 \end{bmatrix}$$

assuming that $d_o = 1$; the coefficients r_i in the numerator can now be obtained from equation (14) i.e.

$$r_o = p_o = 1$$

$$r_1 = p_1 + d_1 p_o$$

$$= \frac{1}{2} + \frac{21}{38} \times 1 = \frac{40}{38}.$$

Thus, the 2nd order model $C_3(s)$ obtained by the MPA is found to be

$$C_3(s) = \frac{1 + \frac{40}{38}s}{1 + \frac{21}{38}s + \frac{5}{38}s^2}$$

$$= \frac{8s + 7.6}{s^2 + 4.2s + 7.6}$$

which is equal to the 2nd order model $C_2(s)$ obtained by the CFE2 technique, as shown in Section 4.

(IV) a 2-input, 2-output system, taken from [7], is now considered. The transfer function matrix C(s) of the system is given by

$$C(s) = \frac{\begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}}{G(s)} \quad .6250$$

where

$$C_{11} = 0.01496s^2 + 1.525s + 2.54$$
$$C_{12} = 95.15s^2 + 1130s + 1810$$
$$C_{21} = 0.0852s^2 + 8.6\,s + 12.3$$
$$C_{22} = 124s^2 + 1491s + 2520$$
$$G(s) = 6.251s^4 + 703.2s^3 + 8416s^2 + 21834s + 15862$$

C(s) is square and has a CFE1 expansion given by equation (3) where

$$H_1 = \frac{1}{6250} \begin{bmatrix} 2520 & -1810 \\ -12.3 & 2.54 \end{bmatrix}$$

$$H_2 = \frac{6250}{7444} \begin{bmatrix} 1.525 & 1130 \\ 8.6 & 1491 \end{bmatrix}$$

⋮

; its power series expansion, about s=0, is given by

$$C(s) = P_o + P_1 s + P_2 s^2 + \ldots\ldots\ldots$$

where

$$P_o = \begin{bmatrix} 2.54 & 1810 \\ 12.3 & 2520 \end{bmatrix} \frac{1}{15862}$$

$$P_1 = \begin{bmatrix} 1.971 & 1362 \\ 8.33 & 1978 \end{bmatrix} \frac{1}{15862}$$

$$P_2 = \begin{bmatrix} 1.3799 & 1009.8 \\ 5.030 & 1510 \end{bmatrix} \frac{1}{15862}$$

$$\vdots$$

A low-order model $M_2(s) = [H_1 + sH_2^{-1}]^{-1}$ obtained by CFE1 is

$$M_2(s) = \frac{\begin{bmatrix} 1.525s+2.54 & 1130s+1810 \\ 8.6s+12.3 & 1491s+2580 \end{bmatrix}}{7444s^2+21834s+15862} .6250$$

Whose power series expansion, about s=0 is given by

$$M_{21}(s) = \bar{P}_o + \bar{P}_1 s + \bar{P}_2 s^2 + \cdots\cdots$$

Where

$$\bar{P}_o = \begin{bmatrix} 2.54 & 1810 \\ 12.3 & 2520 \end{bmatrix} \frac{1}{15862}$$

$$\bar{P}_1 = \begin{bmatrix} 1.971 & 1362 \\ 8.33 & 1978 \end{bmatrix} \frac{1}{15862}$$

$$\bar{P}_2 = \begin{bmatrix} 1.521 & 1025.6 \\ 5.698 & 1541 \end{bmatrix} \frac{1}{15862}$$

$$\vdots$$

Thus as shown theoretically in Section 5, the power series expansion, about s=0, of the low-order model $M_2(s) = [H_1 + sH_2^{-1}]^{-1}$ obtained by the CFE1 technique agrees with the corresponding power series expansion, about s=0, of the original system C(s) up to the s term.

The two output step responses of the original system C(s) and its model $M_2(s)$ are shown in Fig. 3 and 4. It can be seen that the step responses of the model's outputs are practically identical to that of the original system. (in fact, coincident in the Figs)

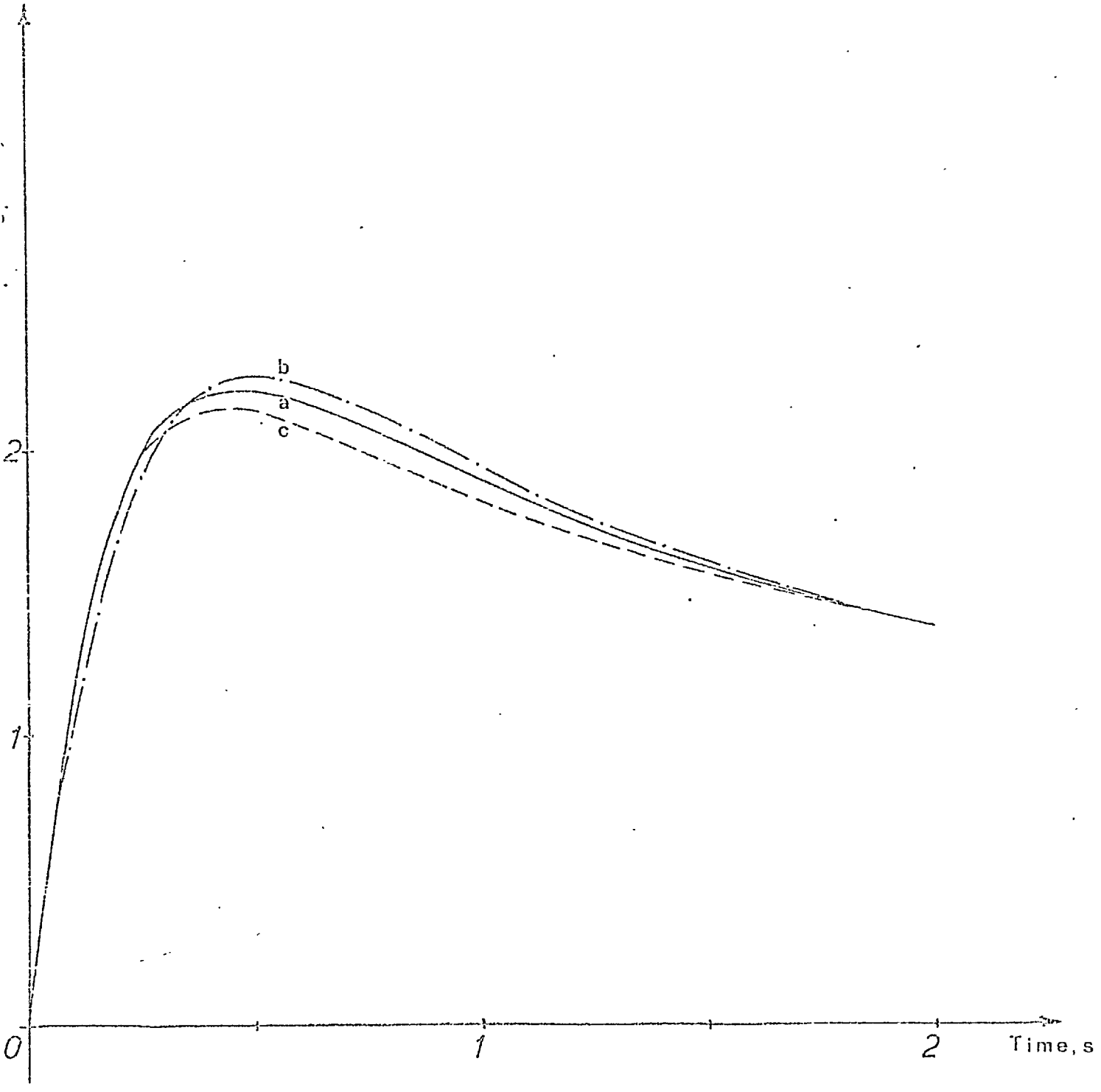


Fig 1

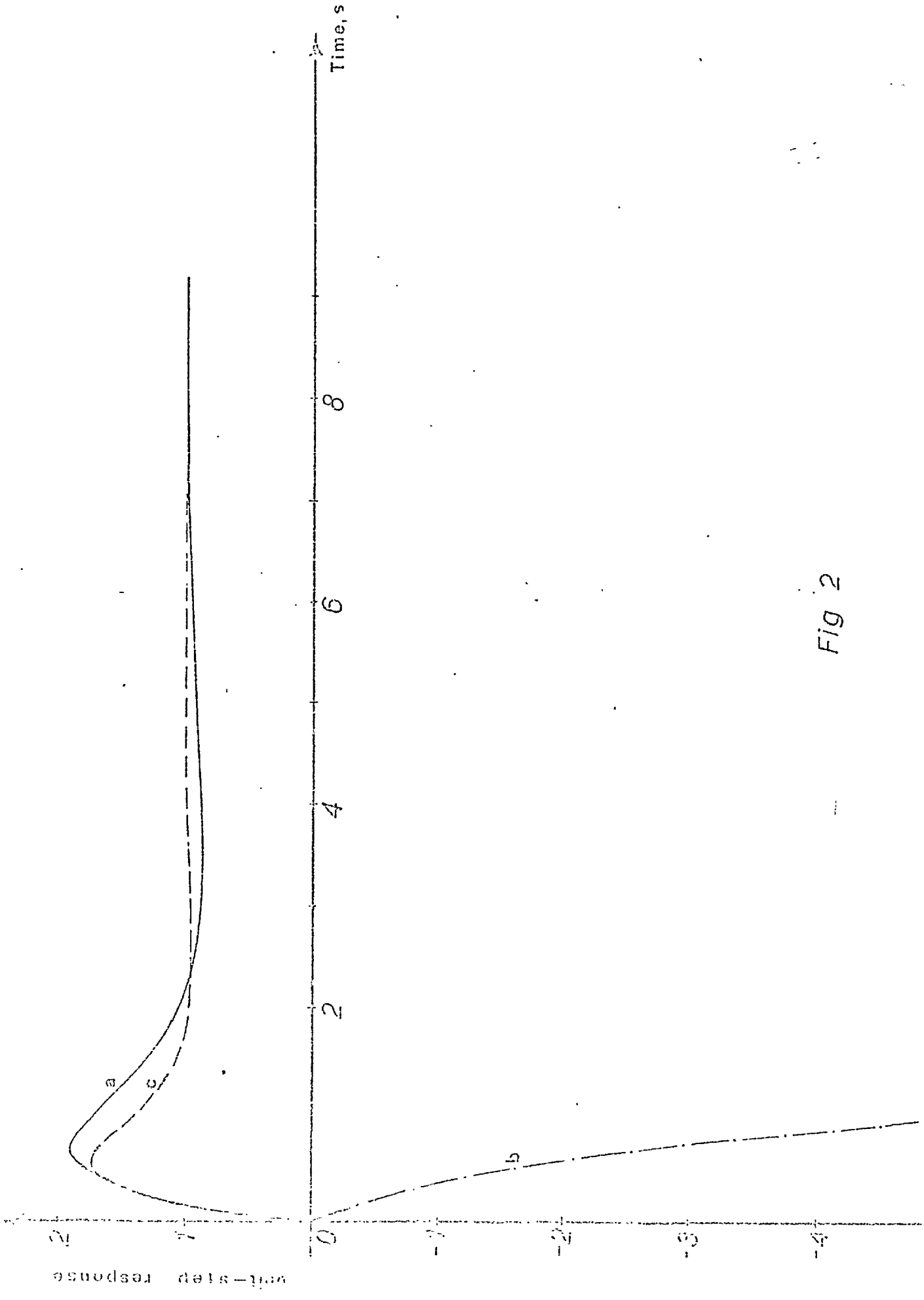


Fig 2

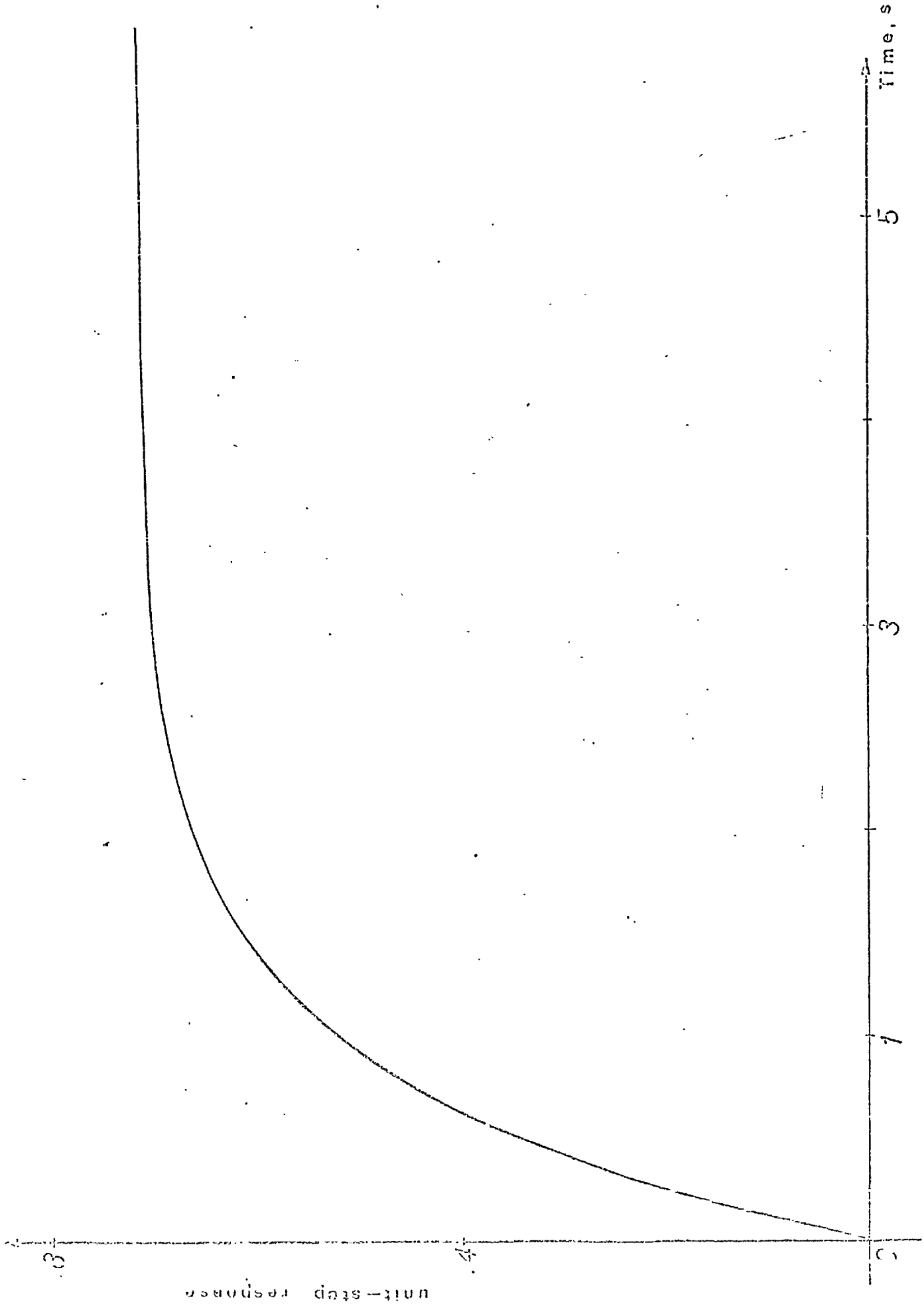


Fig 3

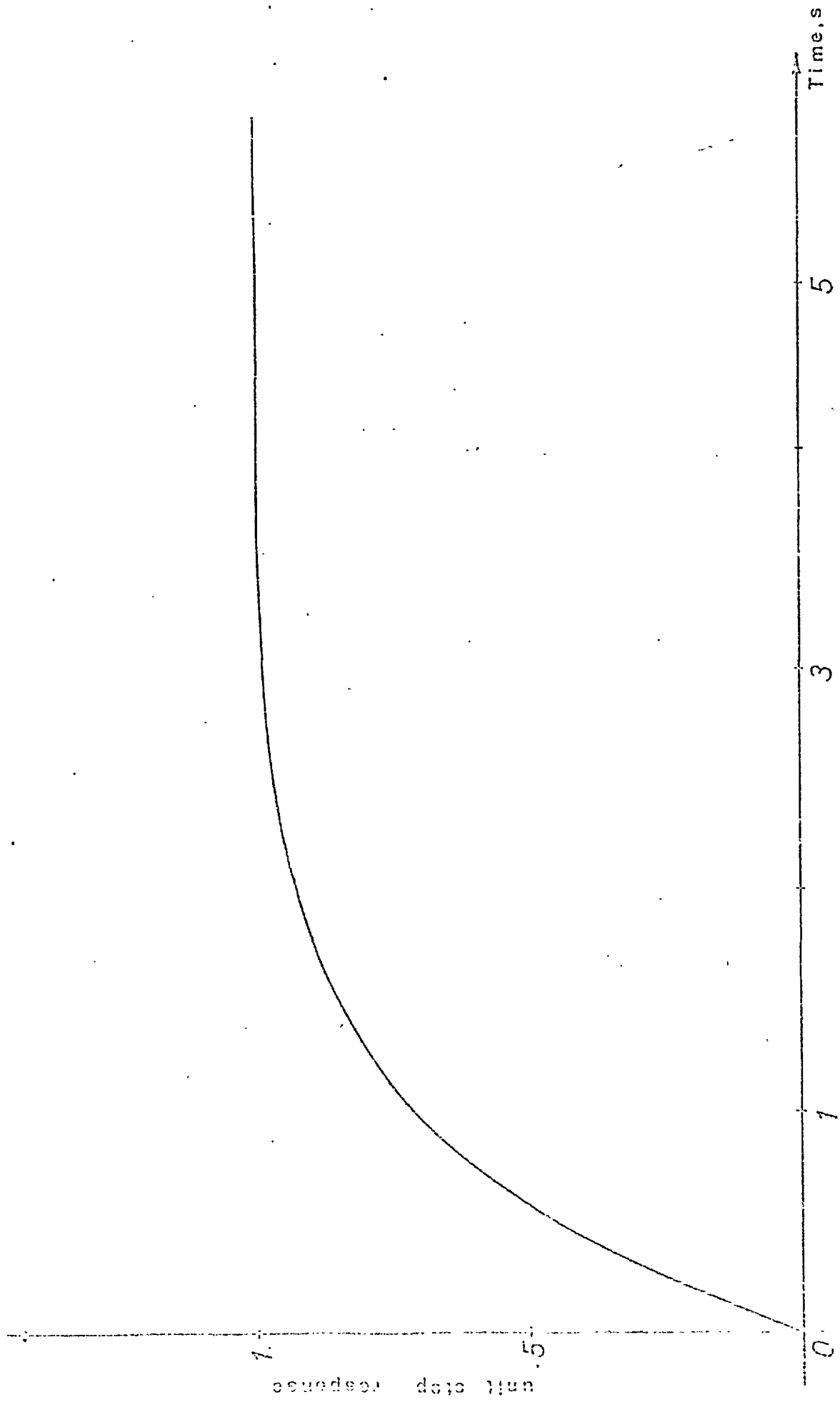


Fig 4

7. REFERENCES

(1) K. Chen, 'A Quick Method for Estimating Closed Loop Poles of Control Systems' AIEE Tran. Vol. 76, Pt. II; May, 1957, pp, 80-87.

(2) W. R. Evans, 'Control Systems Dynamics' McGraw Hill Book Co., 1954, p. 46.

(3) C. F. Chen and L. S. Shieh, 'A Novel Approach to Linear Model Simplification' 1968, JACC Record, pp. 454-461.

(4) H. S. Wall, 'Continued Fractions' Van Nostrand Co., 1948, p. 377.

(5) C. F. Chen, 'A Matrix Cauer Model in Multivariable Systems Analysis and Design' 4th UKAC Control Convention, Multivariable Control System Design and Applications, September 1971, pp. 66-70.

(6) M. R. Chidambara, 'Two Simple Techniques for the Simplification of Large Dynamic System 1969 Preprints of JACC.

(7) H. H. Rosenbrock and P. D. McMorran, 'State-space Analysis of Cascaded Controller' Proc. IEE, Vol. 117, No. 5, May, 1970, pp. 1026-1030.

CHAPTER 5

A TIME DOMAIN TECHNIQUE FOR MODELLING

1. INTRODUCTION

The problem of finding a transfer function model of a linear stable time-invariant plant described by an impulse response is considered in this Chapter. Two possible situations when this might arise are as follows: first, when the impulse response of the plant has been found using correlation techniques, but a transfer function description is needed so that conventional control-system design techniques can be employed; secondly, when it is desired to find a low order model of a system with known transfer function and impulse response. A solution to the problem when the same situation arises has been presented in Chapter 4, where the low order model is obtained by using either continued fraction expansion or Padé approximation techniques; however, as yet, the stability of the resulting model is not guaranteed even when the original system is stable.

Suppose that the impulse response is denoted by W. The output y and the input u of the plant are then related by

$$y(t) = \int_{-\infty}^{t} W(t-\tau)u(\tau)d\tau. \qquad (1)$$

A well known technique for fitting to a function involves fitting to a function a model consisting of a linear combination of linearly independent functions[1-2]. Since the transfer function of the plant considered, denoted by G(s), is the Laplace transform of the impulse response, an approximation to G(s) can be obtained by fitting to W a model consisting of a linear combination

of functions with known Laplace transform. We shall only consider the case when u and y are scalar-valued functions of time since other cases can be treated by applying the approach developed here to each element of the corresponding impulse response matrix.

Suppose that the functions from which the model of W is to be constructed are $f_1, f_2, \ldots . f_p$; then the model of the impulse response W is of the form

$$\tilde{W}(\theta)_{\underline{\alpha}} = \alpha_1 f_1(\theta) + \alpha_2 f_2(\theta) + \ldots . + \alpha_p f_p(\theta) \qquad (2)$$

where $\underline{\alpha} \triangleq (\alpha_1, \alpha_2, \ldots . \alpha_p)'$ and the scalar components α_i can be chosen to minimise the error functon

$$E(\underline{\alpha}) = \int_o^\infty (W(\theta) - \tilde{W}(\theta)_{\underline{\alpha}})^2 \rho(\theta) d\theta \qquad (3)$$

where $\rho(\theta)$ is a positive scalar.

The problem of choosing the components α_i to minimise E of equation (3) is greatly simplified if the functions f_i are orthonormal, in that

$$\int_o^\infty f_i(\theta) f_j(\theta) \rho(\theta) d\theta = \delta_{ij} \quad \text{(Kronecker Delta)}$$

since it is then easy to show that

$$E(\underline{\alpha}) = \int_o^\infty W(\theta)^2 \rho(\theta) d\theta - 2 \sum_{i=1}^{p} \alpha_i \int_o^\infty W(\theta) f_i(\theta) \rho(\theta) d\theta + \sum_{i=1}^{p} \alpha_i^2$$

so that the minimising components α_i are given by

$$\alpha_i^* = \int_o^\infty W(\theta) f_i(\theta) \rho(\theta) d\theta . \qquad (4)$$

The optimised model of the impulse response is then given by

$$\tilde{W}(\theta)_{\underline{\alpha}^*} = \sum_{i=1}^{p} \alpha_i^* f_i(\theta)$$

and the associated approximation to the transfer function G(s) is

$$\tilde{G}(s) = \sum_{i=1}^{p} \alpha_i^* F_i(s) \qquad (5)$$

where $F_i(s)$ is the Laplace transform of $f_i(\theta)$.

It is often desirable for an approximation to a plant transfer function to have the same steady state response as the plant for some pre-specified inputs, e.g. step input, ramp input etc; this feature is not guaranteed for the above modelling procedure. In this Chapter, it will be shown how essentially the same technique can be used to choose an approximation to the transfer function which has the same steady state response as the plant, for specified inputs. This method has a great advantage over the continued fraction expansion or other power series approximation techniques in that it is easy to guarantee at least the open loop stability of the model. It will also be described how a minimal realisation of the impulse response $\tilde{W}(t)$ of the model can be obtained, using Gilbert's method [5]. The eigenvalues of the model are assumed to be simple. The influence of the locations of the eigenvalues of $\tilde{W}(t)$ on the order of the minimal realisation is described.

2. TRANSFER FUNCTION DETERMINATION TO GIVE EXACT STEADY STATE RESPONSE

Since the transfer function $\tilde{G}(s)$ associated with a model $\tilde{W}$ of the impulse response W is just the Laplace transform of $\tilde{W}$, for $\tilde{G}(s)$ to give the same steady state

response as the system with impulse response W, it is necessary only that the model $\tilde{W}$ of W yields the same steady state response as W when used in equation (1). Suppose that the model $\tilde{W}$ is required to give the same steady state response as W for all the q pre-specified input functions $u_1(t), u_2(t), \ldots u_q(t)$. Let the steady state response of W to the inputs $u_j(t)$, $j = 1,2,\ldots q$ be W_j, i.e.

$$W_i = \int_{-\infty}^{o} W(-\tau)u_j(\tau)d\tau \qquad j = 1,2,\ldots q,$$

and the steady output for the model $\tilde{W}$ of equation (2) to the inputs u_j be

$$\begin{aligned}\tilde{y}_i &= \int_{-\infty}^{o} \sum_{i=1}^{p} \alpha_i f_i(-\tau)u_j(\tau)d\tau \\ &= \sum_{i=1}^{p} \alpha_i \int_{-\infty}^{o} f_i(-\tau)u_j(\tau)d\tau \\ &= \sum_{i=1}^{p} \alpha_i \beta_{ij}\end{aligned}$$

where

$$\beta_{ij} = \int_{-\infty}^{o} f_i(-\tau)u_j(\tau)d\tau \qquad \begin{matrix} i = 1,2,\ldots.p \\ j = 1,2,\ldots.q. \end{matrix}$$

Thus, the problem of choosing the components α_i of the model $\tilde{W}$ of equation (2) so as to give the best possible fit to W (i.e. so as to minimise $E(\underline{\alpha})$ of equation (3)) while ensuring that the model gives the same steady state response as the plant can be formulated as the problem of minimising $E(\underline{\alpha})$ with respect to $\alpha_1, \alpha_2, \ldots \alpha_p$ subject to a set of linear equality constraints

$$W_j = \sum_{i=1}^{p} \alpha_i \beta_{ij} \qquad j = 1,2,\ldots q; \qquad (6)$$

for the problem to be well-defined, $q \leqslant p$. We shall

now show that the optimal α_i which satisfy the constraints are given by

$$\tilde{\alpha}_i = \sum_{j=1}^{q} \beta_{ij}\theta_j + \alpha_i^* \qquad i = 1,2,\ldots.p \qquad (7)$$

where α_i^* is as given by equation (4); the scalars $\theta_1, \theta_2, \ldots$ θ_q satisfy the following set of linear simultaneous algebraic equations

$$W_k - \gamma_k = \sum_{j=1}^{q} \xi_{kj}\theta_j \qquad k = 1,2,\ldots.q$$

$$\gamma_k \triangleq \sum_{i=1}^{p} \alpha_i^* \beta_{ik} \qquad (8a)$$

and

$$\xi_{kj} \triangleq \sum_{i=1}^{p} \beta_{ik}\beta_{ij}. \qquad (8b)$$

Substituting α_i of equation (7) into equation (6) yields that

$$\begin{aligned}\sum_{i=1}^{p} \tilde{\alpha}_i \beta_{ij} &= \sum_{i=1}^{p} \Big(\sum_{k=1}^{q} \beta_{ik}\theta_k + \alpha_i^* \Big)\beta_{ij} \\ &= \sum_{k=1}^{q} \Big(\sum_{i=1}^{p} \beta_{ij}\beta_{ik} \Big)\theta_k + \sum_{i=1}^{p} \alpha_i^* \beta_{ij} \\ &= \sum_{k=1}^{q} \xi_{jk}\theta_k + \gamma_j \\ &= W_j,\end{aligned}$$

by making use of equations (8a) and (8b). Hence, $\tilde{\alpha}_i$ satisfy the constraints. From equation (3), it is easy to show that

$$E(\underline{\tilde{\alpha}}+\delta\underline{\alpha}) = E(\underline{\tilde{\alpha}}) + 2 \sum_{i=1}^{p} (\tilde{\alpha}_i - \alpha_i^*)\delta\alpha_i + \sum_{i=1}^{p} \delta\alpha_i^2$$

$$= E(\tilde{\underline{\alpha}}) + 2 \sum_{i=1}^{p} \sum_{j=1}^{q} \beta_{ij}\theta_j \delta\alpha_i + \sum_{i=1}^{p} \delta\alpha_i^2$$

$$= E(\tilde{\underline{\alpha}}) + 2 \sum_{j=1}^{q} \left(\sum_{i=1}^{p} \beta_{ij}\delta\alpha_i \right)\theta_j + \sum_{i=1}^{p} \delta\alpha_i^2 \qquad (9)$$

where

$$\delta\underline{\alpha} = (\delta\alpha_1, \delta\alpha_2, \ldots\ldots \delta\alpha_p)'.$$

Now $\tilde{\underline{\alpha}}$ minimises $E(\tilde{\underline{\alpha}})$ subject to the constraints defined by equation (6) only if $E(\tilde{\underline{\alpha}}+\delta\underline{\alpha}) \geqslant E(\tilde{\underline{\alpha}})$ for all $\tilde{\underline{\alpha}} + \delta\underline{\alpha}$ such that the constraints are satisfied; i.e. for all $\tilde{\underline{\alpha}} + \delta\underline{\alpha}$ such that

$$W_j = \sum_{i=1}^{p} (\alpha_i + \delta\alpha_i)\beta_{ij} \qquad j = 1,2,\ldots\ldots q.$$

Now since $\tilde{\underline{\alpha}}$ has been shown to satisfy the constraints, the only changes

$$\delta\underline{\alpha} = (\delta\alpha_1, \delta\alpha_2, \ldots\ldots \delta\alpha_p)'$$

which has to be considered are those for which

$$\sum_{i=1}^{p} \beta_{ij}\delta\alpha_i = 0.$$

For such changes $\delta\underline{\alpha}$, it is obvious from equation (9) that $E(\tilde{\underline{\alpha}} + \delta\underline{\alpha}) \geqslant E(\tilde{\underline{\alpha}})$, which proves that $\tilde{\alpha}_i$ given by equation (7) minimises $E(\underline{\alpha})$ subject to the constraints defined by equation (6). The required approximation to the transfer function G(s) which has the same steady state response as the plant is therefore $\tilde{G}(s)$ of equation (5) with α_i^* replaced by $\tilde{\alpha}_i$.

Note that, if $u_j(t) = t^{j-1}$, $j = 1,2,\ldots\ldots$,

$(-1)^{j-1}W_j$ = (j-1)th time moment of the impulse response W, where $W_j = \int_{-\infty}^{o} W(-\tau)u_j(\tau)d\tau$; the 0th time moment corresponds to the area under the W(t) curve and is equal to the steady state response to a unit step input. Thus, the model $\tilde{W}$ which satisfies the constraints defined by equation (6) has its first (q-1) time moments equal to those of impulse response W. For a model $\tilde{W}$ to satisfy only one constraint $u_1(t) = t^o = 1$, i.e. $\tilde{W}$ and W have the same steady state response to a constant input, the expression for the optimal scalars $\tilde{\alpha}_i$ reduces to

$$\tilde{\alpha}_i = \beta_i\theta + \alpha_i^* \qquad i = 1,2,.....p \qquad (10)$$

where α_i^* is as given by equation (4) and where

$$\theta = (W_s - \sum_{i=1}^{p} \alpha_i^*\beta_i)/\gamma \qquad (10a)$$

$$W_s = \int_o^\infty W(\tau)d\tau \qquad (10b)$$

$$\beta_i = \int_o^\infty f_i(\tau)d\tau,$$

and

$$\gamma = \sum_{i=1}^{p} \beta_i^2 \neq 0.$$

Equations (10) willbe used in the following numerical example where the model $\tilde{W}$ is required to have the same steady state response as W, for constant input.

3. EXAMPLE

The single-input single-output plant considered here, taken from [3], is of 8th order, has transfer

function

$$G(s) = a(s)/b(s)$$

where

$$a(s) = 18s^7+514s^6+5982s^5+363823s^4+122664s^3 +222088s^2+185760s+40320$$

$$b(s) = (s+1)(s+2)(s+3)(s+4)(s+5)(s+6)(s+7)(s+8)$$

and has a impulse response

$$W(t) = -3e^{-t}+4e^{-2t}-6e^{-3t}+8e^{-4t}+5e^{-5t}-12e^{-6t} +14e^{-7t}+8e^{-8t}.$$

The orthonormal functions used for impulse response approximation purposes are

$$f_1(\theta) = \sqrt{2\lambda_1}e^{-\lambda_1\theta}$$

$$f_2(\theta) = \frac{\sqrt{2\lambda_2}}{(\lambda_1-\lambda_2)} (2\lambda_1 e^{-\lambda_1\theta}-(\lambda_1+\lambda_2)e^{-\lambda_2\theta})$$

for which $\rho(\theta) = 1 \quad \forall\ \theta$. Such functions are described in [4] and have Laplace transforms

$$F_1(s) = \sqrt{2\lambda_1}\frac{1}{s+\lambda_1}$$

$$F_2(s) = \sqrt{2\lambda_2}\ \frac{s-\lambda_1}{(s+\lambda_1)(s+\lambda_2)}.$$

Initially the problem of fitting 1st- and 2nd-order models of W (the 1st-order model being of the form $W = \alpha_1 f_1$ and the 2nd-order model being of the form $W = \alpha_1 f_1 + \alpha_2 f_2$, with $\lambda_1=-1$ and $\lambda_2=-4$) was considered without attempting to satisfy the constraint that the steady state response

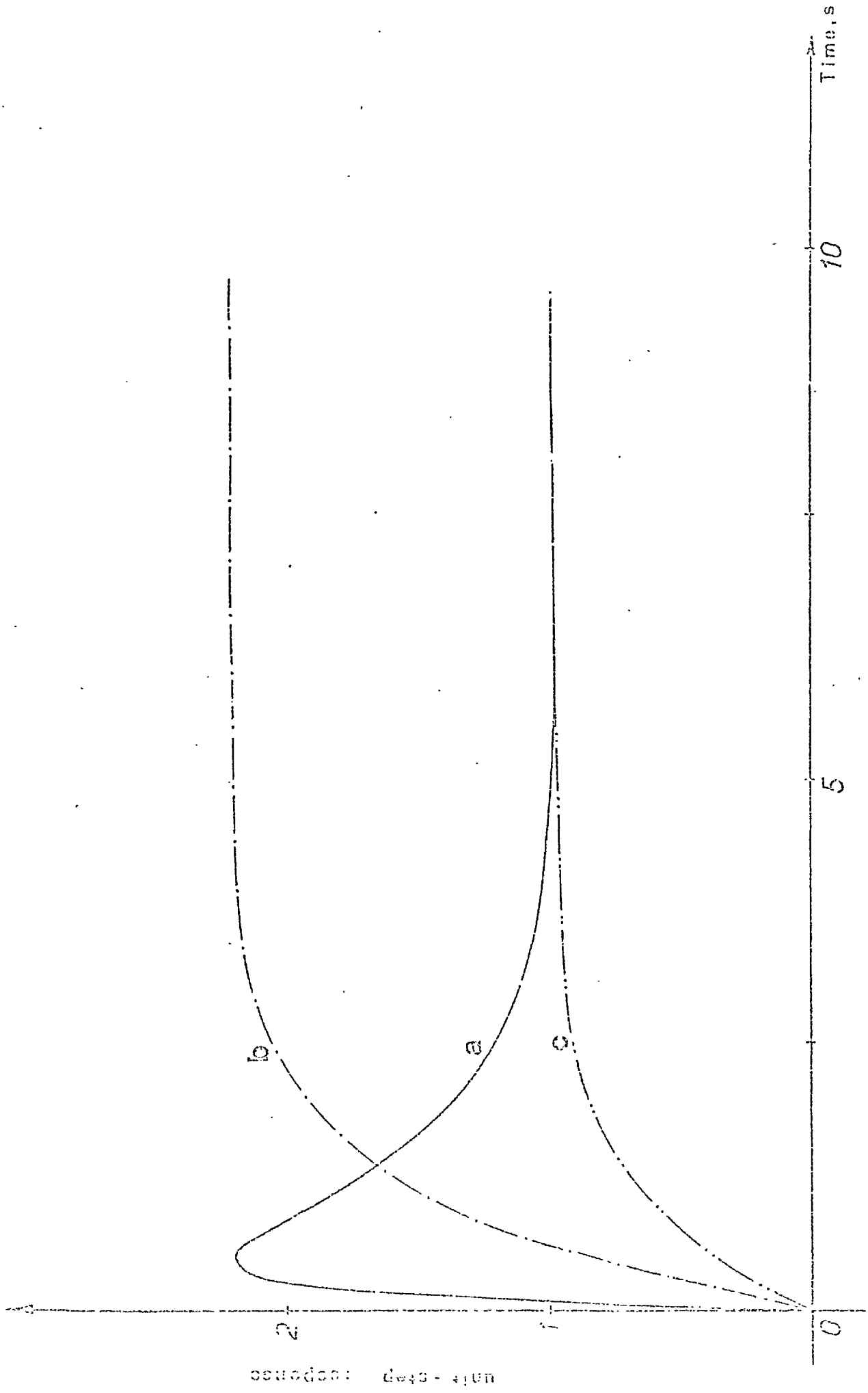


FIG 1

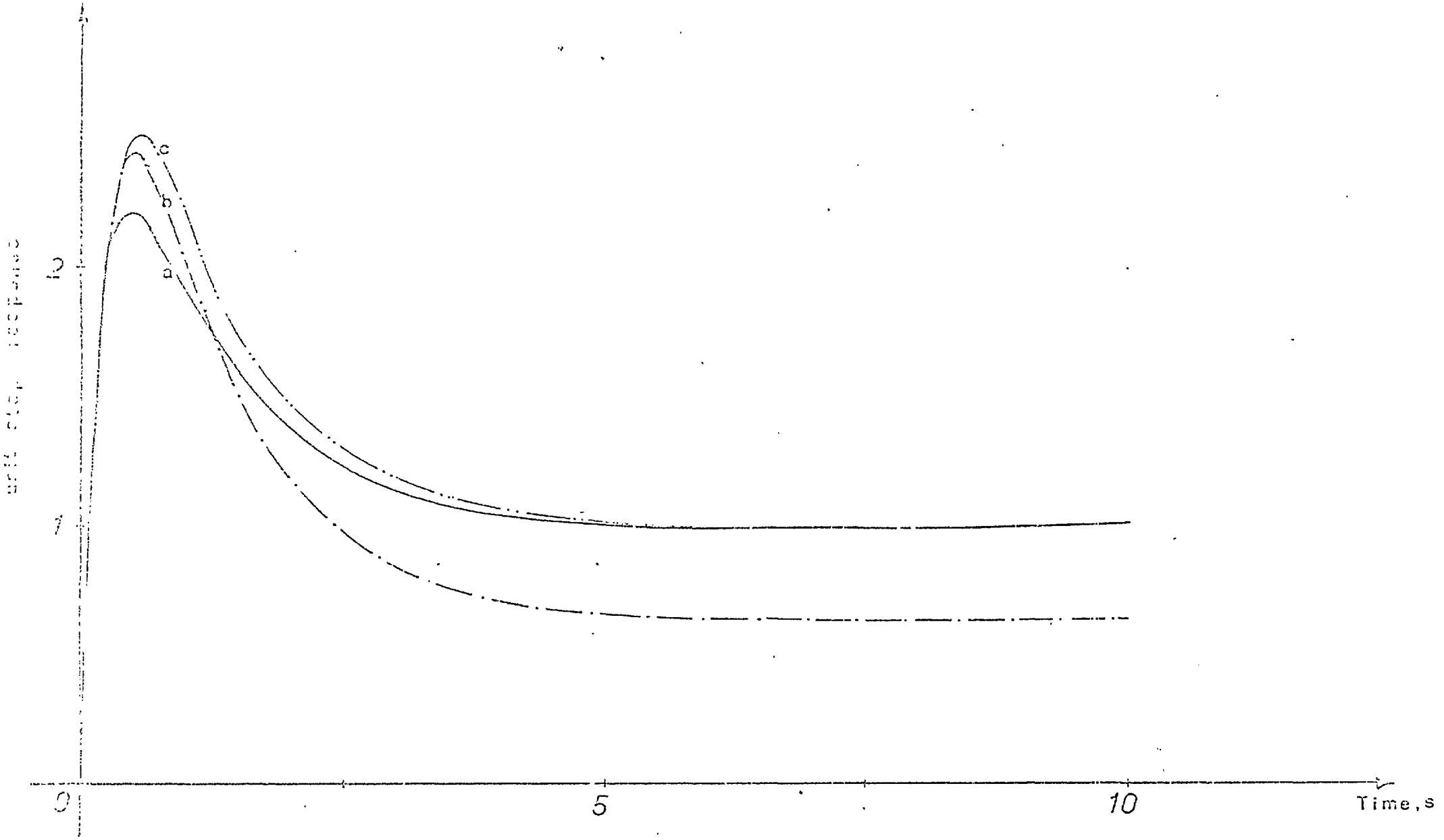


Fig 2

of the plant and the model be identical, for step input, using the method outlined in the Introduction. The optimised transfer function for the 1st-order model was $G(s) = 2.25/(s+1)$, and that for the 2nd-order model was $G(s) = (14.43s+2.52)/(s+1)(s+4)$. The responses of the actual plant and of the 1st-order model for a unit step in the input (starting at time $t = 0$, with the input zero for $t < 0$) are shown in Figures 1a and b respectively. Similar responses for the optimised 2nd-order model are shown in Figure 2. As would be expected, the 2nd-order model gives a response closer to the plant response than does the 1st-order model. However, the steady state responses for both models are not equal to the steady state response of the plant. Next, the technique for obtaining transfer functions presented in Section 2 to give the same steady state response as the plant was applied to the choice of 1st and 2nd-order models. The resulting transfer functions were $G(s) = 1/(s+1)$ and $G(s) = (14.2s+4)/(s+1)(s+4)$, respectively. The associated response for a unit step input for the 1st-order model is shown in Figure 1c and that for the 2nd-order model in Figure 2c. This time, the steady state response for both models is the same as that of the system, and the model responses, seem, overall, to be more acceptable approximations to the sysyem response than the responses given by the models chosen without regard to steady state response.

4. MINIMAL REALISATION OF THE LOW ORDER MODEL

Each element $\tilde{\omega}_{ij}(t)$ of the impulse response matrix $\tilde{W}(t)$ of the model is assumed to be obtained as a linear combination of orthonormal functions, using the technique described in Section 2. The orthonormal functions used here are the Laguerre functions

$$f_1(t) = \sqrt{2\lambda_1}\, e^{-\lambda_1 t}$$

$$f_2(t) = \frac{\sqrt{2\lambda_2}}{(\lambda_1-\lambda_2)} (2\lambda_1 e^{-\lambda_1 t} - (\lambda_1+\lambda_2)e^{-\lambda_2 t}),$$

described in detail in [4]. Thus, each element $\tilde{\omega}_{ij}(t)$ of the p×q matrix $\tilde{W}(t)$ can be expressed as

$$\mathcal{L}[\tilde{\omega}_{ij}(t)] = \sum_{k=1}^{N_{ij}} \frac{\alpha_k}{(s+\lambda_k)} \qquad \begin{matrix} i = 1,2,.....p \\ j = 1,2,.....q. \end{matrix}$$

It is now easy to see that the Laplace transform $\tilde{G}(s)$ of $\tilde{W}(t)$ can be expressed as

$$\tilde{G}(s) = \sum_{i=1}^{m} \frac{H_i}{(s+\lambda_i)}$$

where the p×q matrix

$$H_i = \lim_{s \to -\lambda_i} (s+\lambda_i)\tilde{G}(s).$$

m is the number of simple eigenvalues $-\lambda_i$, each of rank r_i. Then by Gilbert's method [4], a minimal realisation of the model of order $n = \sum_{i=1}^{m} r_i$ can be expressed as

$$\dot{\underline{x}} = A\underline{x} + B\underline{u}$$

$$\underline{y} = C\underline{x}$$

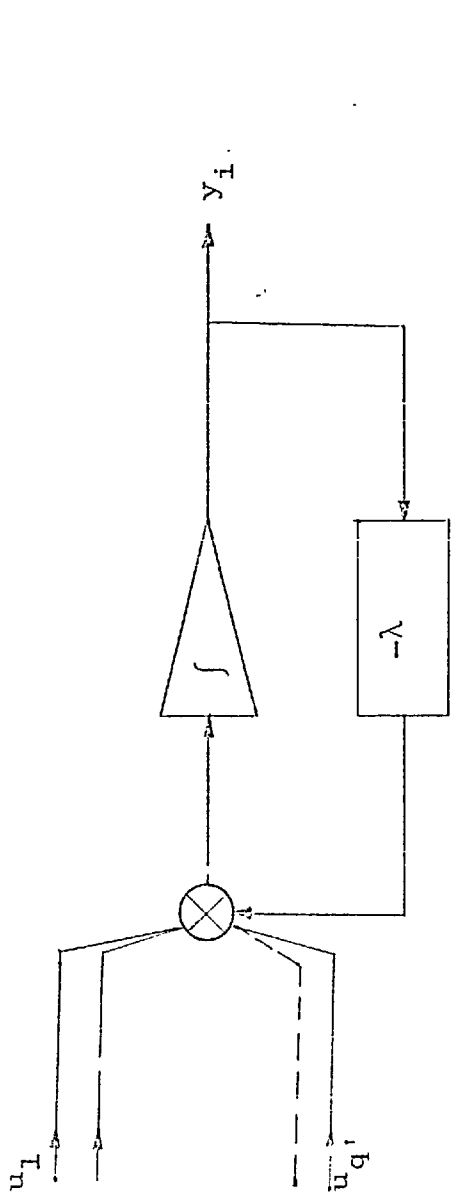
where

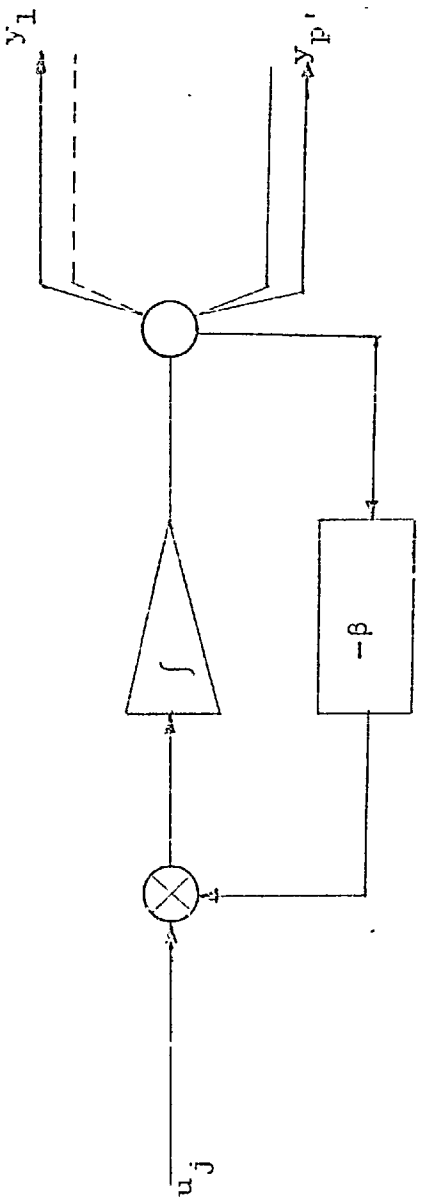
$$A = \begin{bmatrix} -\lambda_1 I_1 & 0 & \cdot & \cdot & 0 \\ 0 & -\lambda_2 I_2 & & & \cdot \\ \cdot & \cdot & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ \cdot & & & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & -\lambda_m I_m \end{bmatrix}, B = \begin{bmatrix} F_1 \\ F_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ F_m \end{bmatrix}$$

$$C = [\, E_1, E_2, \ldots\ldots\ldots E_m \,]$$

and I_i is a unit matrix of order r_i. E_i is a $p \times r_i$ matrix consisting of the linearly independent columns of H_i of rank r_i. F_i is a $r_i \times q$ matrix given by $F_i = (E_i' E_i)^{-1} E_i' H_i$.

It is clear that if the same eigenvalue $-\lambda$ has been located at any q' ($q' \leqslant q$) elements in any row of the transfer function matrix $\tilde{G}(s)$ of the model, instead of distinct eigenvalues, the total number of distinct eigenvalues will be $m-(q'-1)$. And since the rank of H_i remains unchanged, the number of integrators required to realise $\tilde{G}(s)$ will be $(q'-1)$ less than that required when all eigenvalues are distinct. Similarly, the number of integrators required is decreased by $p'-1$ ($p' \leqslant p$) if any p' elements of any column of $\tilde{G}(s)$ has the same eigenvalue $-\beta$, say. These effects are illustrated below.


y_i
$\int$
$-\lambda$
u_1
$u_{q'}$


y_1
$y_{p'}$
$\int$
$-\beta$
u_j

5. NUMERICAL EXAMPLE

The complex system considered is a two-input, two-output linear stationary system of order 14. The 2 x 2 transfer function matrix G(s) of the system is given by

$$G(s) = \sum_{i=1}^{12} \frac{H_i}{(s+\lambda_i)} ,$$

where

$$H_1 = \frac{1}{167}\begin{bmatrix} 1.01 & 0 \\ 0.866 & -1.012 \end{bmatrix}, \quad H_2 = \frac{1}{83}\begin{bmatrix} 0 & -0.872 \\ 0 & 0 \end{bmatrix}$$

$$H_3 = \frac{1}{1.1}\begin{bmatrix} 0 & 0.1256 \\ 0 & 0.0736 \end{bmatrix}, \quad H_4 = -\begin{bmatrix} 0.00898 & 0.103 \\ 0.142 & .0602 \end{bmatrix}$$

$$H_5 = \frac{1}{0.7}\begin{bmatrix} 0 & 0 \\ 0.358 & 0 \end{bmatrix} \qquad H_6 = -\frac{1}{0.6}\begin{bmatrix} 0 & 0 \\ 0.0585 & 0 \end{bmatrix}$$

$$H_7 = \frac{1}{0.5}\begin{bmatrix} 0 & 0 \\ 0.016 & 0 \end{bmatrix} \qquad H_8 = -\frac{1}{0.4}\begin{bmatrix} 0 & 0 \\ 0.0145 & 0 \end{bmatrix}$$

$$H_9 = \frac{1}{0.11}\begin{bmatrix} 0.0368 & 0 \\ 0 & 0 \end{bmatrix} \qquad H_{10} = -\frac{1}{0.1}\begin{bmatrix} 0.0333 & 0 \\ 0 & 0 \end{bmatrix}$$

$$H_{11} = \frac{1}{0.09}\begin{bmatrix} 0.0195 & 0 \\ 0 & 0 \end{bmatrix} \qquad H_{12} = -\frac{1}{0.08}\begin{bmatrix} 0.00404 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\lambda_1 = \frac{1}{167}, \quad \lambda_2 = \frac{1}{83}, \quad \lambda_3 = \frac{1}{1.1}, \quad \lambda_4 = 1, \quad \lambda_5 = \frac{1}{0.7},$$

$$\lambda_6 = \frac{1}{0.6}, \quad \lambda_7 = \frac{1}{0.5}, \quad \lambda_8 = \frac{1}{0.4}, \quad \lambda_9 = \frac{1}{0.11}, \quad \lambda_{10} = \frac{1}{0.1},$$

$$\lambda_{11} = \frac{1}{0.09}, \quad \lambda_{12} = \frac{1}{0.08}.$$

By using the modelling method described in [1], a model $\tilde{G}(s)$ of the transfer function $G(s)$ is found to be

$$\tilde{G}(s) = \begin{bmatrix} \frac{.00793}{s+\frac{1}{167}} - \frac{0.327}{s+1} & -\frac{0.0117}{s+\frac{1}{83}} + \frac{0.123}{s+1} \\ \frac{.00942}{s+\frac{1}{167}} - \frac{.555}{s+1} & -\frac{0.00603}{s+\frac{1}{167}} + \frac{.0105}{s+1} \end{bmatrix}.$$

It is obvious that each element of $\tilde{G}(s)$ matrix can be realised by a linear system of order 2. Thus, a direct realisation of $\tilde{G}(s)$ has order $4 \times 2 = 8$. Now we express $\tilde{G}(s)$ as

$$\tilde{G}(s) = \sum_{i=1}^{3} \frac{\tilde{H}_i}{(s+\lambda_i)}$$

where

$$\tilde{H}_1 = \begin{bmatrix} 0.00793 & 0 \\ 0.00942 & -0.00603 \end{bmatrix}, \quad \tilde{H}_2 = \begin{bmatrix} 0 & -0.0117 \\ 0 & 0 \end{bmatrix}$$

$$\tilde{H}_3 = \begin{bmatrix} -0.327 & 0.123 \\ -0.555 & 0.0105 \end{bmatrix}$$

$\lambda_1 = \frac{1}{167}$, $\lambda_2 = \frac{1}{83}$ and $\lambda_3 = 1$. Note that $\tilde{H}_1$ and $\tilde{H}_3$ are of rank 2 and $\tilde{H}_2$ is of rank 1. Hence, by Gilbert's method of minimal realisation described above, $\tilde{G}(s)$ can be represented by the following linear, vector differential equation of order 5.

$$\dot{x} = Ax + Bu$$

$$y = Cx$$

where

$$A = \begin{bmatrix} -\frac{1}{167} & 0 & \cdots & \cdots & 0 \\ 0 & -\frac{1}{167} & & & \vdots \\ \vdots & & -\frac{1}{83} & & \vdots \\ \vdots & & & -1 & \vdots \\ 0 & \cdots & \cdots & \cdots & -1 \end{bmatrix}$$

$$B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and}$$

$$C = \begin{bmatrix} 0.00805 & 0. & -0.0117 & -0.327 & 0.123 \\ 0.00946 & -0.00604 & 0 & -0.555 & 0.0105 \end{bmatrix} .$$

Hence, for this example, minimal realisation of $\tilde{G}(s)$ requires 3 less integraters than the direct realisation.

The unit-step response of the system output y_1 and the corresponding output of the model are shown in Fig. 3. (curve (a) and (b) respectively).

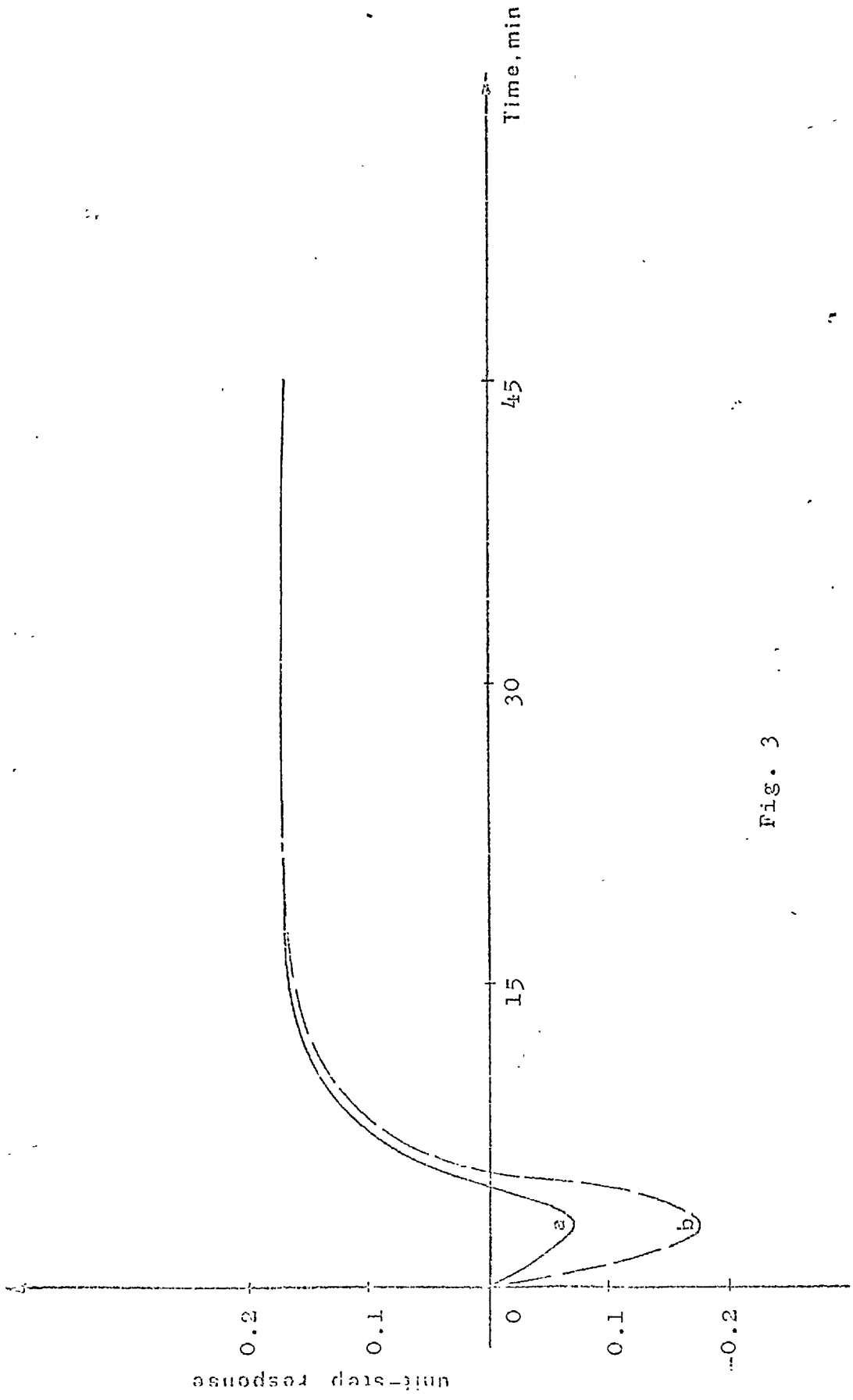


Fig. 3

6. REFERENCES

[1] T. Kitamori, Applications of Orthogonal Functions to the Deyermination of Process Dynamic Characteristics and to the Construction of Self-Optimising Control Systems, Proc. of IFAC, 1960, Moscow, pp.613-618.

[2] P.D. Roberts, The Application of Orthogonal Functions to Self-Optimising Control Systems Containing a Digital Compensator, Proc. of IFAC, 1966, London, paper 36.c.

[3] M.R. Chidambara, Two Simple Techniques for the Simplification of Large Dynamic Systems, Proc. of JACC, 1969.

[4] W.H. Kautz, Transient Synthesis in the Time Domain, I.R.E. Trans. 1954, CT-1, pp.29-39.

[5] E.G. Gilbert, Controllability and Observability in Multivariable Control Systems, J.SIAM. Control, Vol.2, No.1, 1963, pp.128-151.

CHAPTER 6

LOWER BOUNDS WITH APPLICATION TO MODEL ACCURACY

1. INTRODUCTION

Consider the optimisation problem:
minimise with respect to the control function u the performance index

$$V[u] = \frac{1}{2}\int_T [\langle (y(t)-y_d(t)),\ Q(t)(y(t)-y_d(t))\rangle + \langle u(t),\ R(t)u(t)\rangle]dt \qquad (1)$$

for the linear dynamical system

$$y(t) = y_s(t) + \int_{t_s}^{t} W(t,\tau)u(\tau)d\tau \qquad (2)$$

where y is the output function; y_s is the contribution to the output due the initial conditions or other external influances; y_d is the desired system response; u is the control function; W is the open-loop transfer function operator linking input to output; $Q(t) \geqslant 0$; $R(t) > 0$; $u(t) \in R^m$ and $y(t),\ y_s(t),\ y_d(t) \in R^r\ \forall\, t \in T = [t_s, t_f]$, $t_f - t_s < \infty$. Using Hilbert space notation, equations (1) and (2) can be expressed as

$$V[u] = \frac{1}{2}\langle [y-y_d],\ Q[y-y_d]\rangle + \frac{1}{2}\langle u,\ Ru\rangle \qquad (3)$$

and

$$y = y_s + Wu$$

where

$$\langle [y-y_d],\ Q[y-y_d]\rangle \triangleq \int_T \langle [y(t)-y_d(t)],\ Q[y(t)-y_d(t)]\rangle dt,$$

$$[Ru](t) \triangleq R(t)u(t),$$

$$[Wu](t) \triangleq \int_{t_s}^{t} W(t,\tau)u(\tau)d\tau$$

and where u, y etc are taken to be members of appropriate Hilbert spaces. It can be shown that the optimal control function u^o for the above problem is given by

$$u^o = - (R + W^*QW)^{-1} W^*Q(y_s - y_d) \qquad (4)$$

where the operator $W^* \triangleq$ the adjoint of W. The direct solution of equation (4) is often avoided, for $(R + W^*QW)$ is an infinite-dimensional compact operator and is therefore difficult to invert numerically. The problem is usually solved by iterative algorithms e.g. conjugate gradient, contraction mapping algorithms. A suboptimal solution to the problem is sometimes sought by specifying the structure of the control function u [1]. Also, it may be necessary, computationally, to approximate the impulse response of the system by a simple model. Whichever approximation approach is used, it is often desirable to know how close is the approximate solution to the optimal one without actually knowing the optimal solution. In this Chapter, we shall develop several lower bounds for the minimum performance index V^o and discuss their application as a simple test on the validity of a system model. Also, their use as a stopping condition for iterative optimisation algorithms is discussed. These lower bounds are closer than those obtained by Allwright [2], at the expense of extra computation, for the linear convolution-described systems. The further computational effort is about that which is required to perform one extra optimisation iteration.

2. LOWER BOUNDS

The lower bound results are stated below

i) For $R > 0$ and $Q \geqslant 0$,

$$V^o \geqslant V(u) - \frac{1}{2}\langle g, R^{-1}g\rangle = \tilde{V}_1(u)$$

ii) For $R > 0$ and $Q > 0$,

$$V^o \geqslant V(u) - \frac{1}{2}\langle g, R^{-1}g\rangle + \frac{1}{2}\frac{\|WR^{-1}g\|^4}{\langle WR^{-1}g, (Q^{-1}+WR^{-1}W^*)WR^{-1}g\rangle}$$

$$= \tilde{V}_2(u)$$

iii) For $R > 0$ and $Q \geqslant 0$,

$$V^o \geqslant V(u) - \frac{1}{2}\langle g, R^{-1}g\rangle + \frac{1}{2}\frac{\|Q^{\frac{1}{2}}WR^{-1}g\|^4}{\langle Q^{\frac{1}{2}}WR^{-1}g, (I+Q^{\frac{1}{2}}WR^{-1}W^*Q^{\frac{1}{2}})Q^{\frac{1}{2}}WR^{-1}g\rangle}$$

$$= \tilde{V}_3(u)$$

iv) For $R > 0$ and $Q > (\text{or } \geqslant)\ 0$,

$$V^o \geqslant V(u) - \frac{1}{2}\langle g, (R^{-1}-R^{-1}W^*QWR^{-1}+R^{-1}W^*QWR^{-1}W^*QWR^{-1})g\rangle$$

$$= \tilde{V}_4(u)$$

where V^o is the minimal performance index; $V(u)$ is the performance index and g is the gradient of $V(u)$ with respect to u, for the control function u. The superscript $*$ denotes the adjoint of an operator and inequalities > 0 and $\geqslant 0$ denote, respectively, positive definiteness and positive semi-definiteness of a symmetric operator.

The proofs of the above results are facilitated by the following simple lemmas:

Lemma 1 For linear bounded operators Q, R, S and T on appropriate Banach spaces, we have

$$(R+S^*QT)^{-1}=R^{-1}-R^{-1}S^*Q^{\frac{1}{2}}(I+Q^{\frac{1}{2}}TR^{-1}S^*Q^{\frac{1}{2}})^{-1}Q^{\frac{1}{2}}TR^{-1}$$

provided $R > 0$ and $Q \geqslant 0$.

Proof : The following constructive proof was suggested by Allwright. By writing

$$A = R + S^*QT$$

we have

$$A = R + S^*Q^{\frac{1}{2}}Q^{\frac{1}{2}}T.$$

Pre-multiplying both sides by A^{-1} and then post-multiplying by R^{-1} yields

$$A^{-1} = R^{-1} - A^{-1}S^*Q^{\frac{1}{2}}Q^{\frac{1}{2}}TR^{-1}.$$

Post-multiplying this by $S^*Q^{\frac{1}{2}}$ and rearranging gives

$$A^{-1}S^*Q^{\frac{1}{2}} = R^{-1}S^*Q^{\frac{1}{2}}(I+Q^{\frac{1}{2}}TR^{-1}S^*Q^{\frac{1}{2}})^{-1}.$$

Substituting for $A^{-1}S^*Q^{\frac{1}{2}}$ in the above expression for A^{-1} then yields the required result.

Lemma 2 For the same operators and hypothesis as in Lemma 1, save that $Q > 0$,

$$(R+S^*QT)^{-1} = R^{-1} - R^{-1}S^*(Q^{-1} + TR^{-1}S^*)^{-1}TR^{-1}.$$

Proof : The result of Lemma 2 follows from that of Lemma 1 by noting that, when $Q > 0$,

$$Q^{\frac{1}{2}}(I + Q^{\frac{1}{2}}TR^{-1}S^*Q^{\frac{1}{2}})^{-1}Q^{\frac{1}{2}}=(Q^{-1} + TR^{-1}S^*)^{-1}.$$

Lemma 3 For any non-zero vectors x, y and any linear bounded symmetric operator $P > 0$,

$$\langle y, P^{-1}y\rangle \geqslant \frac{\langle x, y\rangle^2}{\langle x, Px\rangle}.$$

Proof : Since $P^{\frac{1}{2}}$ and $P^{-\frac{1}{2}}$ exist and are positive-definite, we have

$$\langle x,y\rangle^2 = \langle x,\ P^{\frac{1}{2}}P^{-\frac{1}{2}}y\rangle^2 = \langle P^{\frac{1}{2}}x,\ P^{-\frac{1}{2}}y\rangle^2$$

$$\leqslant \|P^{\frac{1}{2}}x\|^2 \|P^{-\frac{1}{2}}y\|^2 = \langle x,\ Px\rangle < y,\ P^{-1}y\rangle.$$

As

$$\langle x,\ Px\rangle > 0 \qquad \forall\ x \neq 0,$$

it follows that

$$\langle y,\ P^{-1}y\rangle \geqslant \frac{\langle x,y\rangle^2}{\langle x,\ Px\rangle}.$$

Using the above Lemmas, we can now come to the

Proof of the Lower Bound Results

i) By substituting equation (2) into equation (1), we have

$$V(u) = \frac{1}{2}\langle u,\ (R+W^*QW)u\rangle + \langle u,\ W^*Q(y_s-y_d)\rangle$$

$$+ \frac{1}{2}\langle (y_s-y_d),\ Q(y_s-y_d)\rangle.$$

It is obvious that V has a unique minimum and differentiating V with respect to u yields the gradient function g :

$$g = Ru + W^*Q(y - y_d)$$

$$= (R + W^*QW)u + W^*Q(y_s - y_d). \qquad (4)$$

At the optimum, g = 0 and the unique optimal control function u^o is given by

$$u^o = - R^{-1}W^*Q(y - y_d)$$

and the minimun performance index is

$$V^o = - \frac{1}{2}\langle W^*Q(y_s-y_d),\ (R+W^*QW)^{-1}W^*Q(y_s-y_d)\rangle$$

$$+ V(u)$$

where V(u) is the performance index for the control function u. It is now trivial to show , using equation

(3), that

$$V^o = V(u) - \frac{1}{2}\langle g,\ (R+W^*QW)^{-1}g\rangle$$

where g is the gradient function for the control u.

Since $W^*QW \geqslant 0$,

$$\langle x,\ R^{-1}x\rangle \geqslant \langle x,\ (R+W^*QW)^{-1}x\rangle \qquad \forall\, x \neq 0.$$

The result of (i) follows immediately.

ii) From equation (4) and Lemma 2,

$$V^o = V(u) - \frac{1}{2}\langle g, R^{-1}g\rangle + \frac{1}{2}\langle g, R^{-1}W^*(Q^{-1}+WR^{-1}W^*)^{-1}WR^{-1}g\rangle$$

$$= V(u) - \frac{1}{2}\langle g, R^{-1}g\rangle + \frac{1}{2}\langle WR^{-1}g, (Q^{-1}+WR^{-1}W^*)^{-1}WR^{-1}g\rangle.$$

Then, Lemma 3 gives

$$\langle WR^{-1}g,\ (Q^{-1}+WR^{-1}W^*)^{-1}WR^{-1}g\rangle$$

$$\geqslant \frac{\|WR^{-1}g\|^4}{\langle WR^{-1}g,\ (Q^{-1}+WR^{-1}W^*)WR^{-1}g\rangle}\ .$$

Thus,

$$V^o \geqslant V(u) - \frac{1}{2}\langle g, R^{-1}g\rangle + \frac{1}{2}\,\frac{\|WR^{-1}g\|^4}{\langle WR^{-1}g, (Q^{-1}+WR^{-1}W^*)WR^{-1}g\rangle}\ ,$$

as required.

iii) The proof here is similar to that of (ii), save that Lemma 1 is now used.

iv) Applying Lemma 2 to $(Q^{-1}+WR^{-1}W^*)^{-1}$ reveals that

$$\langle WR^{-1}g,\ (Q^{-1}+WR^{-1}W^*)^{-1}WR^{-1}g\rangle$$

$$= \langle WR^{-1}g,\ (Q - QW(R+W^*QW)^{-1}W^*Q)WR^{-1}g\rangle.$$

Hence,

$$V^o = V(u) - \frac{1}{2}\langle g,\ R^{-1}g\rangle$$

$$+ \frac{1}{2}\langle WR^{-1}g, (Q-QW(R+W^*QW)^{-1}W^*Q)WR^{-1}g\rangle$$

$$\geqslant V(u) - \frac{1}{2}\langle g, (R^{-1}-R^{-1}W^*QWR^{-1}+R^{-1}W^*QWR^{-1}W^*QWR^{-1})g\rangle,$$

as $W^*QW \geqslant 0$.

It is easy to see that the same result is obtained if $Q > 0$.

3. TEST OF A SYSTEM MODEL

As has been explained in the previous Chapters, it is often desirable to reduce the order of the control system for analysis, design or optimisation purposes. Several techniques for model simplification have been presented in Chapter 4 for systems described in the frequency domain and in Chapter 5 for systems described in the time domain. So far, the only test that has been made on the validity of the low order model is by comparing the response of the system and its model to some specified inputs e.g. α, αt,... etc. However, for the optimisation problem using a model, it would be desirable to compare the system response to the optimal control u^o of the system and to the optimal control $\tilde{u}$ of the model; usually, it is done by finding the difference between the optimal performance index V^o and the performance index $V[\tilde{u}]$ using $\tilde{u}$. Furthermore, it is hoped that one comparison can be made without knowing the optimal control u^o of the system. The results of the lower bound on the optimal performance index derived in Section 2 may be used for this purpose; for the lower bounds $\tilde{V}_i[\tilde{u}]$, $i=1,2,3,4$ give an estimate of V^o. This can be demonstrated by using the example in Chapter 1. The state-space subsystem is obtained by using the modelling technique presented in Chapter 5 and the accuracy of the model is altered by changing the length

p of the recent past input history applied to the convolution-described subsystem. The lower bound $\tilde{V}_1$ and one of the closer lower bounds $\tilde{V}_2$ of Section 2 were calculated for p = 0,1,2,3 and the results are presented in the following table, where $\tilde{u}$ is the optimal control of the model and $V[\tilde{u}]$ is the performance index using $\tilde{u}$

p	$\tilde{V}_1[\tilde{u}]$	$\tilde{V}_2[\tilde{u}]$	$V[\tilde{u}]$
0	-.56009	.52109	1.0598
1	.00936	.53214	.86456
2	.080131	.54472	.78001
3	.15818	.54984	.74844

To compare the results, V^o is also calculated; V^o=.56478. It can be seen that , as p increases, the value of the lower bound $\tilde{V}_1$ and $\tilde{V}_2$ increases, so getting nearer to V^o and hence a better estimation of V^o is obtained. At p = 3, the error in using $\tilde{V}_2$ to estimate V^o is 0.56478 - 0.54984 = 0.01494. It can also be seen that, for each p, the closer lower bound $\tilde{V}_2$ gives a much better estimation than $\tilde{V}_1$, especially at p=0. This should offset the extra computation required in calculating $\tilde{V}_2$.

4. STOPPING CONDITIONS FOR ITERATIVE ALGORITHMS

When iterative optimisation algorithms are used to optimise the control function u for the problem described in the Introduction, it is desirable to know when u_j, the control function obtained at iteration j

of such an algorithm, is an adequate approximation to u^o, the optimal control function which minimises $V[u]$. For once an adequate approximation has been obtained, iteration can cease. Allwright has obtained stopping conditions for iterative control function optimisation using the following types of approximation to the optimal control function. For real pre-specified scalar $\epsilon > 0$, u_j is said to be an ϵ(absolute) approximation to u^o if $|V[u_j]-V[u^o]| \leqslant \epsilon$; similarly, u_j is an ϵ(relative) approximation to u^o if $|(V[u_j]-V[u^o])/V[u^o]| \leqslant \epsilon$. It is assumed here that $V[u^o] > 0$, which will occur whenever the free response of the system is not equal to the desired response.

Let $g[u_j]$ be the gradient function for the control function u_j. Then, as shown in Section 2, a closer lower bound for the minimum performance index $V[u^o]$ is

$$\tilde{V}_2[u_j] = V[u_j] - \frac{1}{2}\langle g[u_j], R^{-1}g[u_j]\rangle + \frac{1}{2}\frac{\|WR^{-1}g[u_j]\|^4}{\langle WR^{-1}g[u_j], (Q^{-1}+WR^{-1}W^*)WR^{-1}g[u_j]\rangle} \qquad (5)$$

which implies that $\tilde{V}_2[u_j] \leqslant V[u^o]$.

<u>ϵ(ABSOLUTE) APPROXIMATION</u>

It is easy to see that $|V[u_j]-V[u^o]| \leqslant |V[u_j] - \tilde{V}_2[u_j]| \triangleq \epsilon_j^a$ so that u_j is an ϵ(absolute) approximation to u^o whenever $\epsilon > \epsilon_j^a$.

<u>ϵ(RELATIVE) APPROXIMATION</u>

Since $|(V[u_j]- V[u^o])/V[u^o]| \leqslant |(V[u_j]- \tilde{V}_2[u_j])/\tilde{V}_2[u_j]| \triangleq \epsilon_j^r$, if $\tilde{V}_2[u_j] > 0$, u_j is an ϵ(relative)

approximation to u^o if $\epsilon \geqslant \epsilon_j^r$.

The gradient function $g[u_j]$ for u_j is usually required by iterative optimisation algorithms, so the lower bound $\tilde{V}_2[u_j]$ on $V[u^o]$ can be calculated by using it in equation (5). It is then easy to decide whether iteration can cease, by checking whether u_j is an ϵ(absolute) approximation or ϵ(relative) approximation for a suitable ϵ as required. In [2], Allwright's lower bound $\tilde{V}_1[u_j]$ is

$$\tilde{V}_1[u_j] = V[u_j] - \frac{1}{2} \langle g[u_j], R^{-1} g[u_j] \rangle$$

and is therefore related to the new lower bound $\tilde{V}_2[u_j]$ by

$$\tilde{V}_2[u_j] = \tilde{V}_1[u_j] + \frac{1}{2} \frac{\| WR^{-1} g[u_j] \|^4}{\langle WR^{-1} g[u_j], (Q^{-1} + WR^{-1} W^*) WR^{-1} g[u_j] \rangle} \qquad (6)$$

Since the second term is always positive when the gradient $g[u_j]$ is non-zero, the new lower bound $\tilde{V}_2[u_j]$ is more positive and therefore closer to $V[u^o]$. This will lead to smaller $\tilde{\epsilon}^a$ and $\tilde{\epsilon}^r$ than the corresponding ϵ_j^a and ϵ_j^r obtained by using $\tilde{V}_1[u_j]$. Hence, if u_j is actually an ϵ approximation to u^o, the new lower bound $\tilde{V}_2[u_j]$ is more likely to reveal that it is than $\tilde{V}_1[u_j]$. Furthermore, since the check on whether u_j is an

ϵ(relative) approximation is only applicable if $\tilde{V}_2[u_j] > 0$ (or $\tilde{V}_1[u_j] > 0$) and $\tilde{V}_2[u_j]$ (or $\tilde{V}_1[u_j]$) could be negative with $|\tilde{V}_2[u_j]| > V[u^o]$ (or $|\tilde{V}_1[u_j]| > V[u^o]$), it would be likely to be applicable when the new lower bound $\tilde{V}_2[u_j]$ is used instead of $\tilde{V}_1[u_j]$. The additional term involved in the new lower bound in equation (6) is easy to calculate, since Q and R are operators which are easy to invert.

For the example presented in Section 3, we have, at p=3, $\tilde{\epsilon}^a = |.74844 - .54984| = .19860$ and $\tilde{\epsilon}^r = |(.74844 - .54984)/.54984| = .36118$. Thus, the optimal control u of the parallel model with p=3, may be regarded as an ϵ(absolute, $\epsilon = .19860$) approximation or an ϵ(relative, $\epsilon = .36118$) approximation to the optimal control u^o of the system.

An numerical example is given below to demonstrate the use of the lower bound results as a stopping condition for iterative optimisation algorithms.

5. EXAMPLE

For comparison with Allwright's results, the system treated is that used as an example in [3], where $y_s = (1, -1)'$, $y_d(t) = (1, 0)'$; $R(t) = 0.1$; $T = [0, 1]$;

$$A(t) = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \qquad B(t) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$C(t) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad Q(t) = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}.$$

j	$\tilde{\epsilon}_j^a$	$\tilde{\epsilon}_j^r$	ϵ_j^a	ϵ_j^r
0	3.6874939	-	51.221298	-
1	0.8181761	23.6149712	1.5880249	-
2	0.0631328	0.1332017	0.1124313	0.2647528
3	0.0044296	0.0091227	0.0070402	0.0145776
4	0.0002036	0.0004191	0.0002801	0.0005767
5	0.0000069	0.0000142	0.0000213	0.0000438
6	0.0000056	0.0000115	0.0000081	0.0000166
7	0.0000001	0.0000002	0.0000001	0.0000002

The convolution description of the system used here is derived in [3]. Simpson's rule with a step length of 0.01 was used for numerical integrations and the conjugate-gradient algorithm was used to perform the iterative optimisation with an initial control function u_o identically zero on T.

The values of $\tilde{\epsilon}_j^a$ and $\tilde{\epsilon}_j^r$ as well as those of ϵ_j^a and ϵ_j^r in [2] at iteration j are given in the table above, for $j = 0, 1, \ldots 7$.

It can be seen from the table that $\tilde{\epsilon}_j^a < \epsilon_j^a$ and $\tilde{\epsilon}_j^r < \epsilon_j^r$ and that $\tilde{\epsilon}_1^r$ can be calculated when ϵ_1^r cannot. The new lower bound results show that u_5 is actually an ϵ(absolute) approximation with $\epsilon = 0.00001$, even though Allwright's lower bound $V_1[u]$ does not reveal that a ϵ(absolute) approximation has been obtained until u_6 has been calculated. Though additional computational effort is required to calculate $\tilde{\epsilon}_j^a$ and $\tilde{\epsilon}_j^r$, the possible saving in the number of iterations required could offset this.

6. REFERENCES

[1] E.A. Freeman and A. Todd, Sub-optimal Control for Multivariable Systems Based on an Orthogonal Decomposition Technique, Int. J. Control, 1971, Vol. 13, No.4, pp.737-761.

[2] J.C. Allwright, Stopping Conditions for Iterative Control Function Optimisation, Electronics Letters, Vol.6, No.17, 1970, pp.545-547.

[3] E.A. Freeman and G.L. Brown, Digital Calculation of Optimum Controls for Time-Varying Linear Multivariable System - Realisation of Contraction Mapping Algorithm, Proc. IEE, 1970, 117,pp.449-458.

CHAPTER 7

FREQUENCY DOMAIN APPROXIMATION TECHNIQUE FOR OPTIMAL CONTROL

1. INTRODUCTION

Extensive study has been made on the problem of optimising a linear time-invariant control system with respect to a quadratic performance index. This problem can be solved by using either Kalman's theory of optimal control [1-2] in the time domain or Wiener's optimal filtering theory in the frequency domain [3]. Kalman's theory leads to the solution of the nonlinear Riccati equation while Wiener's formulation requires the spectral factorisation of a positive definite matrix often known as the spectral density matrix (SDM). The most common complaint in using Kalman's approach is that not all the states of the system are accessible to physical measurement and therefore not available for feedback control scheme. A more serious difficulty is that systems designs based on optimality become very complicated if the order of the system is high and therefore expensive to implement. As have been discussed in the previous Chapters, a well-known approach to overcome this difficulty is to find a suboptimal solution to the problem, using a suitable low-order model to the system. The main theme of this Chapter will be the linear regulator problem formulated in the frequency domain.

Several methods for the spectral factorisation of a rational regular continous time SDM have been suggested. Before discussing these methods, we shall explain some terminology: let $A(s)$ be a $m \times n$ rational matrix function of the complex variable s, then the matrix $A(s)$ is said to be paraconjugate hermitian if $A^*(s) = A(-\bar{s})$ (where $(\cdot)^*$ denotes the conjugate transpose matrix operation), and is said to be

paraconjugate unitary if either $A^*(-\bar{s})\ A(s) = I_n$ or $A(s)\ A^*(-\bar{s}) = I_m$ or both. A polynomial matrix is said to be of normal rank r if (i) there exists at least one subminor of order r which does not vanish identically, and (ii) all minors of order greater than r vanish identically. Lastly, we term a matrix elementary if and only if it has a constant determinant.

The most complete and elegant solution to the factorisation problem was given by Youla [4]. Central to Youla's method of proof is the use of the Smith-McMillan canonical form for rational matrices and the technique for factoring a positive elementary polynomial paraconjugate hermitian matrix due to Oono and Yasurra [7]. Youla proves, among other things, a main theorem required in the spectral factorisation of rational matrices.

Theorem: If $P(s) = P^*(-\bar{s})$ is a rational n x n paraconjugate hermitian matrix of normal rank r which is non-negative for $s=jw$ (w real) then there exists an r x n rational matrix G(s) with the following properties:

(1) $P(s) = G^*(-\bar{s})\ G(s)$.

(2) G(s) and $G^{-1}(s)$ (the right inverse of G(s)) are both analytic in Re (s) > 0.

(3) G(s) is unique up to within a constant unitary r x r matrix multiplier on the left; i.e. $G_1(s)$ also satisfies 1 and 2, $G_1(s) = TG(s)$ where T is rxr constant matrix and satisfies $T^*T = I_r$.

(4) If P(s) is real (i.e. $\overline{P(s)} = P(\bar{s})$), G(s) is real and T is real.

Anderson [6] proposed a method which is based on the minimal realisation theory of rational matrices and requires

the solution of a matrix quadratic equation. In 1963, Davies [5] put forward a method which avoids the calculation of the Smith-McMillan form. However, though elegant in form, these methods are not computationally attractive, since they are not easily mechanised. The number of operations involved in these methods increases with the degree of the characteristic polynomial of the SDM which, in turn, depends on the order of the system. Thus, these methods becomes very tedious for high order multivariable control systems.

We shall demonstrate, in this Chapter, that the continued faction expansion CFE2 and the 'Modified' Padé approximation techniques MPA which have been shown useful in model reduction in Chapter 4 can also be applied to the SDM P to obtain a low-order approximation. The spectral factorisation of the low-order approximation can then be accomplished with only a few operations and a suboptimal control to the system can be obtained easily. No low-order model to the system is required in this approach. In fact, it will be shown that this approach can be related to and also proved to be more flexible than the popular approach using the reduced model, if the model is obtained by the CFE2 or MPA technique described in Chapter 4. An iterative algorithm, in the frequency domain, will also be presented. The algorithm generates a new approximation, say P_{i+1}, to P by updating the former approximation P_i such that the spectral factorisation of P_i will form part of that of P_{i+1}; this feature of the algorithm would greatly reduce the amount of effort required to obtain the spectral factorisation of the new approximation P_{i+1}. Furthermore, it will be showed that, when expanded into power series about $s=0$ ($z=\infty$), the new

approximation P_{i+1} generated by the algorithm will agree with P up to a higher order term in s (s^{-1}) than P_i. The spectral factorisation method proposed by Davies [5] will be used in the example presented in Section 6 of this Chapter.

2. PROBLEM FORMULATION

To find the control to a linear constant system which minimises a quadratic performance index requires, in the frequency domain, the solution of a matrix Wiener-Hopf equation. For convenience, a short, simple but not rigorous derivation of the equation is given below.

Consider a time-invariant linear control system with a r x m transfer function matrix W(s), a r- dimensional output vector y(s) and a m- dimensional input victor u(s) i.e.

$$y(s) = y_o(s) + W(s)\, u(s) \qquad (1)$$

where $y_o(s)$ is a r- dimensional vector due to initial conditions or other external influences. A control $u^o(s)$ is to be found which will minimise the quadratic performance index

$$J = \frac{1}{2\pi j} \int_{-j\infty}^{+j\infty} [(y^*-y_d^*)Q(y-y_d) + u^*Ru]\, ds. \qquad (2)$$

By Parseval's theorem,

$$\frac{1}{2\pi j} \int_{-j\infty}^{+j\infty} [(y^*-y_d^*)Q(y-y_d) + u^*Ru] ds$$

$$= \int_o^\infty [(y(t)-y_d(t))^t Q(y(t)-y_d(t)) + u^T(t)Ru(t)]dt$$

Which is the well-known continue quadratic performance index in optimal control theory. In equation(2), y_d is the desired output vector, $y^*=y^T(-s)$ = the transponse of the vector y(-s), $Q \geq 0$ and $R>0$. Substituting equation (1) into (2) gives

$$J = \frac{1}{2\pi j} \int_{-j\infty}^{+j\infty} [(u^*W^* + z^*)Q(Wu + z) + u^*Ru]ds$$

Where $z(s) = y_o(s) - y_d(s)$. We now apply the variational method suggested by Chang [8] to derive the Wiener-Hopf equation. Let $h(s)$ be a vector and

$$J = \frac{1}{2\pi j} \int_{-j\infty}^{+j\infty} V(h^*,h)ds \qquad (2a)$$

Where $h^* = h^T(-s)$. Then, the Wiener-Hopf equation which defines the vector $h(s)$ minimising equation (2a) is the vector equation

$$\frac{\partial V}{\partial h^*} = g(s)$$

Where $g(s)$ is a vector required to give rise to function of time which exists only for negative time. Comparing equations (2) and (2a), we have

$$V = (y^* - y_d^*)\ Q\ (y - y_d) + u^*R\ u$$

and $h(s) = u(s)$. Thus,

$$\frac{\partial V}{\partial h^*} = \frac{\partial V}{\partial u^*} = W^*Q(Wu+z) + Ru = g(s)$$

or

$$(R + W^*QW)u + W^*Qz = g(s) \qquad (3)$$

which is the frequency response equivalent of Wiener-Hopf equation in matrix form. Note that $(R+W^*QW) \triangleq P(s)$ is a rational regular m x m real paraconjugate hermitian matrix of normal rank m which is positive on the real-frequency axis p=jw. Hence, by the theorem given in the Introduction, due to Youla [4], there exists an m x m rational regular real matrix G(s) such that

$$(R+W^*QW) \triangleq P(s)$$

$$= G^T(-s)\ G(s);$$

for real matrix G(s), $G^*(-\bar{s}) = G^T(-s)$. Also G(s) and $G^{-1}(s)$ could both represent the transfer function of stable linear systems. It can be shown that the optimal control $u^o(s)$, which is the solution of the equation (3), is given by

$$u^o = -G^{-1}(s)\ [(G^T(-s))^{-1}\ W^T(-s)\ Qz]^+ \qquad (3a)$$

where the operator $[\quad]^+$ indicates that only the portion of the transform relating to positive time is considered.

By studying the expression

$$P \triangleq R + W^*QW = R + W^T(-s)QW(s)$$

$$= G^T(-s)\ G(s) \qquad (4)$$

carefully, it can be seen that, in some cases, the amount of work involved in obtaining the spectral factorisation of the SDM P i.e. finding the matrices G(s) and $G^T(-s)$ could be considerably reduced. The transfer function matrix W(s) of the system is assumed to be a square matrix of the form

$$W(s) = \frac{\sum_{i=0}^{m} A_j s^i}{\sum_{i=0}^{n} b_i s^i}$$

where A_j are constant matrices, b_i constant scalars and $m < n$. The SDM P is then given by

$$P = R + \frac{\sum_{i=0}^{2n} \tilde{A}_i s^i}{\sum_{i=0}^{2n} \tilde{b}_i s^i}$$

$$= \frac{\sum_{i=0}^{2n} \tilde{B}_i s^i}{\sum_{i=0}^{2n} \tilde{b}_i s^i}.$$

Note that P is, in general, a rational regular matrix i.e. $\lim_{s\to\infty} P = K$, a non-zero constant matrix. The number of operations required to obtain the spectral factorisation of P increases with higest degree of the polynomial matrix $\sum_{i=0}^{2n} \tilde{B}_i s^i$ which, in this case, is 2n where n is the order of the system.

From equation (4), we have

$$P = W^T(-s)[(W^T(-s))^{-1} R(W(s))^{-1} + Q] W(s) \qquad (4a)$$

The matrix $(W^T(-s))^{-1} R(W(s))^{-1} + Q$ is, in general, a positive definite rational matrix. The degree of its matrix polynomial numerator has to be higher than that of its scalar polynomial denominator so that pre-multiplication by $W^T(-s)$ and post-multiplication by $W(s)$ which are rational proper matrices give rise to the rational regular matrix P. By the theorem due to Youla [4], this matrix has a spectral factorisation i.e. there exists a matrix M(s) such that

$$(W^T(-s))^{-1} R(W(s))^{-1} + Q = M^T(-s) M(s) \qquad (5)$$

Thus, the matrices G(s) and $G^T(-s)$ are obtained as soon as the spectral factorisation of $(W^T(-s)^{-1} R(W(s))^{-1} + Q$ is found, for, from equation (4a)

$$P = W^T(-s) M^T(-s) M(s) W(s)$$
$$= (M(-s) W(-s))^T (M(s) W(s))$$

i.e.

$$G(s) = M(s) W(s).$$

In some cases, it is possible that $(W^T(-s))^{-1} R(W(s))^{-1} + Q$ is just a polynomial matrix of the form $\frac{1}{\alpha} \sum_{i=0}^{2(n-m)} \theta_i s^i$ where α is a constant scalar; i.e.

$$(W^T(-s)^{-1}R(W(s))^{-1} + Q = \sum_{i=0}^{2(n-m)} \frac{1}{\alpha}\theta_i s^i.$$

Note that the highest degree of the polynomial matrix has to be 2(n-m) where n and m are the highest degree of the denominator and the numerator of W(s) respectively, in order to ensure that the matrix P in equation (4a) will be a rational regular matrix. Hence the highest degree of the polynomial matrix $\frac{1}{\alpha}\sum_{i=0}^{2(n-m)} \theta_i s^i$ is less than that of the SDM $P = \frac{\sum_{i=0}^{2n} \tilde{B}_i s^i}{\sum_{i=0}^{2n} \tilde{b}_i s^i}$, And it follows immediately that the number of operations required to find the matrix M(s) in equation (5) is less than that required to find the G(s) in equation (4). Thus, by transforming the SDM P from its original form in equation (4) to the form in equation (5), the amount of effort required to obtain the spectral factorisation of P could be considerably reduced, in some cases. Consider a practical example given in [10] where W(s) is a 2 x 2 transfer function matrix of a 4th order system

$$W(s) = \frac{\begin{bmatrix} W_{11}(s) & W_{12}(s) \\ W_{21}(s) & W_{22}(s) \end{bmatrix}}{C(s)}$$

where

$$w_{11}(s) = 0.01496s^2 + 1.525s + 2.54$$

$$w_{12}(s) = 95.15s^2 + 1130s + 1810$$

$$w_{21}(s) = 0.0852s^2 + 8.6s + 12.3$$

$$w_{22}(s) = 124s^2 + 1491s + 2520$$

$$C(s) = 6.2517s^4 + 703.1606s^3 + 8416.228s^2 + 21834.8s + 15862.2$$

W(s) is the transfer function matrix for a 2-spool gas turbine

based on experimental data. For convenience, take

$$Q = \begin{bmatrix} 10^2 & 0 \\ 0 & 10^2 \end{bmatrix}, \quad R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and the SDM P of the optimisation problem is found to be

$$P = \frac{\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}}{C(-s)\ C(s)}$$

where

$$p_{11} = .39084 \times 10^2 s^8 - .38920 \times 10^6 s^6 + .43245 \times 10^8 s^4$$
$$- .29160 \times 10^{11} s^2 + .61869 \times 10^{11}$$
$$p_{12} = .46829 \times 10^{10} s^4 - .41702 \times 10^{12} s^3 - .48973 \times 10^{13} s^2$$
$$- .12589 \times 10^{13} s + .13903 \times 10^{14}$$
$$p_{21} = .46829 \times 10^{11} s^4 + .41702 \times 10^{12} s^3 - .48973 \times 10^{13} s^2$$
$$+ .12589 \times 10^{13} s + .13903 \times 10^{14}$$
$$p_{22} = .39084 \times 10^2 s^8 - .8920 \times 10^6 s^6 + .40323 \times 10^8 s^4$$
$$- .20976 \times 10^9 s^2 + .25161 \times 10^9.$$

The matrix P is a rational regular matrix where both the matrix polynomial numerator and the scalar polynomial denominator are of degree 8 in s. The inverse transfer function $W^{-1}(s)$ of the system is

$$W^{-1}(s) = \frac{\begin{bmatrix} w_{22}(s) & -w_{12}(s) \\ -w_{21}(s) & w_{11}(s) \end{bmatrix}}{6250}$$

which is a polynomial matrix of degree 2 in s with a constant scalar denominator 6250. Also, since

$$Q = \begin{bmatrix} 10^2 & 0 \\ 0 & 10^2 \end{bmatrix},$$

the corresponding positive definite matrix

$$(W^T(-s))^{-1}R\,(W(s))^{-1} + Q = \frac{\begin{bmatrix} \bar{w}_{11}(s) & \bar{w}_{12}(s) \\ \bar{w}_{21}(s) & \bar{w}_{22}(s) \end{bmatrix}}{6250^2}$$

where

$$\bar{w}_{11}(s) = w_{22}(-s)w_{22}(s) + w_{21}(-s)w_{21}(s) + 62500^2$$

$$\bar{w}_{12}(s) = -w_{22}(-s)w_{12}(s) - w_{21}(-s)w_{11}(s)$$

$$\bar{w}_{21}(s) = \bar{w}_{12}(-s)$$

$$\bar{w}_{22}(s) = w_{12}(-s)w_{12}(s) + w_{11}(-s)w_{11}(s) + 62500^2$$

is a polynomial matrix of degree 4 in s. Therefore, the amount of effort required to obtain the spectral factorisation of $(W^T(-s))^{-1}R(W(s))^{-1} + Q$ will be considerably less than that required to obtain the spectral factorisation of $R + W^T(-s)QW(s)$. It is, hence, worth checking if the inverse transfer function matrix $W^{-1}(s)$ of the system is a polynomial matrix, even though it is not often true.

3. SUBOPTIMAL CONTROL

As discussed in the Introduction, the spectral factorisation of the SDM P(s) for a high order multivariable system could be a very tedious process, as it is not easily

mechanised. And it is not always true that the inverse transfer function matrix of the system will be a polynomial matrix so that the argument given in Section 2 can be used to reduce the amount of effort required. The continued fraction expansion CFE2 and the 'Modified' Padé approximation MPA techniques described in Chapter 4 will now be applied to obtain an low-order approximation, say $\bar{P}$, to the SDM P. The spectral factorisation of $\bar{P}$ can then be easily obtained and a suboptimal control $\bar{u}$ to the system can be found by replacing P in equation (3) by $\bar{P}$. We shall name this technique for obtaining suboptimal control, the spectral density matrix approximation technique (SDMA).

The matrix P is a rational regular matrix and is assumed to have a general CFE of the form

$$P = A_1 s[B_1 + A_2 s[B_2 + \ldots + A_{2n} s[B_{2n}]^{-1} \ldots]^{-1}]^{-1}$$

and the following power series expansions

$$P = C_o + C_1 s + C_2 s^2 + \ldots\ldots \quad \text{expanded about } s=0$$

$$= D_o + D_1 s^{-1} + D_2 s^{-2} + \ldots. \quad \text{expanded about } s=\infty$$

Define

$$E_1 \triangleq A_1 s$$

$$F_1 \triangleq B_1$$

$$E_2 \triangleq B_2 A_1 s$$

$$F_2 \triangleq B_1 B_2 + A_2 s$$

$$\vdots$$

or in general

$$E_p \triangleq B_p E_{p-1} + A_p s E_{p-2}$$

assuming that

$$E_{-1} = I,\ F_{-1} = 0,\ E_0 = 0,\ F_0 = I.$$

As have been shown in Section 5 of Chapter 4, a low-order approximation $\tilde{P}_i$ to P obtained by the CFE technique, is given by

$$\tilde{P}_i = F_i^{-1} E_i.$$

A low-order approximation $\tilde{P}_i$ to P can also be obtained by using the MPA technique. Since no matrix inversion is involved in obtaining $\tilde{P}_i$, the MPA technique is, computationally, more efficient than the CFE technique, especially for high order multivariable systems and will be used to obtain the low-order approximations in example (ii) of Section 6.

4. ITERATIVE ALGORITHM

We shall now develop an iterative algorithm which will prove to be particularly useful for multivariable systems. The algorithm will generate a sequence of approximations $P_1, P_2, \ldots P_i, P_{i+1}, \ldots$ such that the degree of the polynomial of the elements in P_{i+1} is, in general, higher than that of the polynomial of the corresponding elements in P_i.

Suppose that the ith approximation to the SDM P is P_i which has the spectral factorisation $P_i(s) = G_i^T(-s)\, G_i(s)$ where $G_i(s)$ and $G_i^T(-s)$ are both invertible. Then, it is obvious that

$$P = P_i + (P - P_i)$$

$P - P_i$ being the error incurred by using the approximation

P_i instead of P. $P - P_i$ is a rational matrix and can be written as

$$\begin{aligned} P-P_i &= P - G_i^T(-s)\, G_i(s) \\ &= G_i^T(-s)\, [(G^T{}_i(-s))^{-1} P\, (G_i(s))^{-1} - I]\, G_i(s) \\ &= G^T{}_i(-s)\, V_i\, G_i(s) \end{aligned}$$

where

$$V_i = (G_i^T(-s)^{-1}\, P(G_i(s))^{-1} - I. \qquad (6a)$$

Hence, the SDM P can be written as

$$P = P_i + G_i^T(-s)\, V_i G_i(s). \qquad (6b)$$

By using either the CFE2 or MPA technique, a low-order approximation $\bar{V}_i$ to the rational matrix V_i can be found and we have

$$\begin{aligned} P &= P_i + G^T{}_i(-s)\, (V_i - \bar{V}_i + \bar{V}_i) G_i(s) \\ &= P_i + G^T{}_i(-s)\bar{V}_i G_i(s) + G^T{}_i(-s)(V_i - \bar{V}_i) G_i(s) \qquad (7) \end{aligned}$$

where the rational matrix $G^T{}_i(-s)(V_i - \bar{V}_i)\, G_i(s)$ is the error when the error matrix $P-P_i$ is replaced by its low-order approximation $G_i^T(-s)\, \bar{V}_i\, G_i(s)$. By writing $P_{i+1} = P_i + G_i^T(-s)\, \bar{V}_i G_i(s)$, we obtain a new approximation (i.e. the (i+1) th approximation) P_{i+1} to P. Note that the new approximation P_{i+1} consists of the former approximation P_i plus a low-order approximation $G_i^T(-s)\, \bar{V}_i\, G_i(s)$ to the error matrix $P-P_i$.

Also, since

$$\begin{aligned} P_{i+1} &= P_i + G_i^T(-s)\bar{V}_i G_i(s) \\ &= G_i^T(-s)\, (I+\bar{V}_i)\, G_i(s) \qquad (8) \end{aligned}$$

and thus

$$P - P_{i+1} = G_i^T(-s)\, (V_i - \bar{V}_i) G_i(s) \qquad (8a)$$

where $\bar{V}_i$ is a rational matrix, the degree of the polynomical elements of P_{i+1} is, in general, higher than that of the corresponding polynomial elements of P_i. Intuitively, the new higher order approximation P_{i+1} would, therefore, be a closer approximation to P than P_i, as the error matrix $P-P_i$ is accounted for by a low-order approximation in P_{i+1}. Equation (8) reveals an useful feature of the algorithm which is that to find the spectral factorisation of the new approximation P_{i+1} requires only that of the low-order matrix $I+\bar{V}_i$, for $G_i(s)$ and $G_i^T(-s)$ are already found in the previous approximation $P_i = G_i^T(-s)G_i(s)$. Thus if

$$I + \bar{V}_i = L_i^T(-s)L_i(s),$$

we have

$$P_{i+1} = G_{i+1}^T(-s)\, G_{i+1}(s)$$

where

$$G_{i+1}(s) \triangleq L_i(s)\, G_i(s) \qquad (8b)$$

and

$$G_{i+1}^T(-s) \triangleq G_i^T(-s)\, L_i^T(-s)$$

Analogous to equation (6), we can again write

$$P = P_{i+1} + (P - P_{i+1})$$

and a new approximation P_{i+2} may be found in terms of P_{i+1} by the algorithm described above, if necessary.

5. COMPARISON BETWEEN THE SDMA AND REDUCED MODEL APPROACH

Difficulties arise when the theory of optimal control is applied to high order multivariable control systems,

In the time domain where Kalman's theory of optimal control is used, high order of the system often leads to excessive computer storage and hence the optimal solution is expensive to implement. When Wiener's filtering theory is used in the frequency domain, the amount of work required to obtain the spectral factorisation of the SDM of a high order multivariable system could be tremendous. So far, the most popular approach used to tackle this problem is the reduced model approach [7-9]. In this approach, a low-order model to the system's dynamic (expressed either in the state-space form or in the transfer function form) is obtained, using various techniques; for example, Chen [9] uses the CFE1 technique described in Chapter 4. The optimal control theory is then applied to the simplified model and the resulting optimal control of the model is taken as a suboptimal control to the original system.

We shall now show that the SDMA approach suggested in this Chapter for obtaining a suboptimal control can be related to the reduced model approach described above, provided the reduced model is obtained by using the CFE2 technique. Consider the optimisation problem described by equation (1) and (2). The transfer function matrix $W(s)$ is assumed to be a rational proper matrix and has a low-order approximation, say $\bar{W}(s)$, obtained by the CFE2 technique which ensures good initial transient approximation and correct steady-state response. Then, as shown in Section 3 of Chapter 4, the power series expansion, about $s=0$ (about $s=\infty$), of $\bar{W}(s)$ agree with the corresponding power series expansion, about $s=0$ (about $s=\infty$) of $W(s)$ up to, say, the s^{p-1} ($s^{-(p-1)}$) term i.e. if

$$W(s) = A_o + A_1 s + \ldots + A_{p-1}s^{p-1} + C_p s^p + \ldots \text{ expanded about } s=0$$

$$= B_o + B_1 s^{-1} + \ldots + B_{p-1}s^{-(p-1)} + D_p s^{-p} + \ldots,$$

expanded about $s=\infty$

it follows immediately that the difference between $W(s)$ and $\bar{W}(s)$ can be expressed as

$$W(s) - \bar{W}(s) = \sum_{i=p}^{\infty} (A_i - C_i)s^i \qquad (9a)$$

$$= \sum_{i=p}^{\infty} (B_i - D_i)s^{-i}. \qquad (9b)$$

If

$$P \triangleq R + W^T(-s)QW(s)$$

and

$$\bar{P} \triangleq R + \bar{W}^T(-s)\, Q\bar{W}(s),$$

it is easy to show that

$$P - \bar{P} = W^T(-s)\, QW(s) - \bar{W}^T(-s)\, Q\bar{W}(s)$$

$$= (W(-s) - \bar{W}(-s))^T Q(W(s) - \bar{W}(s)) + (W(-s) - \bar{W}(-s))^T Q\bar{W}(s) + \bar{W}^T(-s)Q(W(s) - \bar{W}(s))$$

When expressed in terms of the power series expansion, about $s=0$, of $W(s)$ and $\bar{W}(s)$, we have

$$P - \bar{P} = \sum_{i=p}^{\infty} (-1)^i (A_i - C_i)^* s^i\, Q \sum_{j=p}^{\infty} (A_j - C_j)s^j + \sum_{i=p}^{\infty} (-1)^i (A_i - C_i)^* s^i Q$$

$$\sum_{j=0}^{\infty} A_j s^j + \sum_{i=0}^{\infty} (-1)^i A_i^* s^i Q \sum_{j=p}^{\infty} (A_j - C_j)s^j$$

$$= \sum_{i=p}^{\infty} \alpha_i s^i,$$

Similarly if expressed in terms of the power series expansion, about $s=\infty$, of $W(s)$ and $\bar{W}(s)$, we have

$$P - \bar{P} = \sum_{i=p}^{\infty} \beta_i s^{-1}$$

If the SDMA approach is used where a low-order approximation to P is obtained by the CFE2 or MPA technique, then from the results presented in Section 5 of Chapter 4, there exists a low-order approximation, say $\tilde{P}$, such that

$$P - \tilde{P} = \sum_{i=p}^{\infty} \gamma_i s^{i}$$

$$= \sum_{i=p}^{\infty} \theta_i s^{-i}.$$

In other words, the power series expansion, about $s=0$ and $s=\infty$, of the approximation P obtained by the SDMA technique agree with the corresponding power series expansions, about $s=0$ and $s=\infty$, of the approximation $\bar{P}$ obtained by using the reduced model approach up to certain term.

However, it should be noted that the iterative algorithm developed in Section 4 cannot be used in the reduced model approach. Unlike the SDMA approach where a higher order approximation to P can be obtained by using the iterative algorithm to update the previous approximation, no information on the pervious approximation could be used in the reduced model approach. Thus, if a low-order model used in the reduced model approach does not yield a satisfactory suboptimal control and a higher order model has to be used, the spectral factorisation of the entire higher order SDM approximation, say p_n, has to be found. This would require a tremendous amount of

computational effort, especially since the spectral factorisation procedure is not easily mechanised.

We shall now proceed to show that, when expanded into power series about $s=0$ ($s=\infty$), the new approximation P_{i+1} generated by the iterative algorithm agree with P up to the higher order term in $s(s^{-1})$ than P_i. If the approximation P_i is obtained by CFE2 or computationally more efficient by MPA, then as shown in Section 5 of Chapter 4, there exist positive integers n and n' such that the difference $P-P_i$ between P and its ith approximation P_i can be expressed as

$$P-P_i = \sum_{i=n}^{\infty} \gamma_i s^i \qquad (10a)$$

$$= \sum_{i=n'}^{\infty} \theta_i s^{-i} \qquad (10b)$$

In general, the matrices $G_i(s)$ and $G_i^T(-s)$ in the expression $P_i = G_i^T(-s)\, G_i(s)$ are rational regular matrices and can therefore be expressed as

$$G_i(s) = \sum_{i=0}^{\infty} \alpha_i s^i \qquad (11a)$$

$$= \sum_{i=0}^{\infty} \beta_i s^{-i} \qquad (11b)$$

also

$$G_i^{-1}(s) = \sum_{i=0}^{\infty} \xi_i s^i \qquad (12a)$$

$$= \sum_{i=0}^{\infty} \delta_i s^{-i} \qquad (12b)$$

Since, from equation (6a),

$$P-P_i = G_i^T(-s)\, V_i\, G_i(s)$$

$$\Rightarrow V_i = (G_i^T(-s))^{-1}\, (P-P_i)\, (G_i(s))^{-1}, \qquad (13)$$

We have, by substituting equation (10a) and (12a) into equation (13),

$$V_i = \sum_{i=0}^{\infty} (-1)^i \xi_i T_s^i \sum_{i=n}^{\infty} \gamma_i s^i \sum_{i=0}^{\infty} \xi_i s^i$$

$$= \sum_{i=n}^{\infty} \xi'_i s^i$$

and, by substituting equation (10b) and (12b) into equation (13)

$$V_i = \sum_{i=0}^{\infty} (-1)^i \delta_i T_s^{-i} \sum_{i=n'}^{\infty} \theta_i s^{-o} \sum_{i=0}^{\infty} \delta_i s^i$$

$$= \sum_{i=n'}^{\infty} \delta'_j s^{-i}$$

To obtain a new approximation P_{i+1} by using the algorithm presented in Section 4, a low-order approximation $\bar{V}_i$ is obtained by the MPA technique i.e. by 'fitting', say, the first m and m' matrix coefficients of the power series expansions of V_i, about s=0 and s=∞ respectively. Therefore, by expressing the approximation $\bar{V}_i$ also in terms of its power series expansions about s=0 and s=∞, we have

$$V_i - \bar{V}_i = \sum_{i=m+m}^{\infty} \xi''_i s^i \qquad (14a)$$

$$= \sum_{i=n'+m'}^{\infty} \delta''_i \, s^{-i} \qquad (14b)$$

respectively.. In consequence, by using equation (8a) and (14a), the difference between the SDM P and the new approximation P_{i+1} given by equation (8) is, when expressed in terms of power series expansion about s=0

$$P-P_{i+1} = \sum_{i=o}^{\infty} (-1)^i \alpha_i T s^i \sum_{i=n+m}^{\infty} \xi'' s^i \sum_{i=o}^{\infty} \alpha_i s^i$$

$$= \sum_{i=n+m}^{\infty} \gamma'_i \, s^i \; ;$$

similarly, the difference $P-P_{i+1}$, when expressed in terms of their power series expansions, about $s=\infty$, is of the form

$$P-P_{i+1} = \sum_{i=n'+m'}^{\infty} \theta'_i \, s^{-i}$$

Thus, we have proved that, when expressed in terms of their power series expansion about s=0 ($s=\infty$), the new approximation P_{i+1} generated by the iterative algorithm agrees with the SDM P up to a higher power term in $s(s^{-1})$ than P_i.

6. EXAMPLES

(i) Consider a scalar system, of 10th order, having the transfer function W(s) where

$$W(s) = \frac{a(s)}{b(s)}$$

$$a(s) = 5.057s^9+469.7s^8+15707s^7+264590s^6+2554700s^5$$
$$+14913000s^4+53105000s^3+111660000s^2+125800000s$$
$$+57600000$$

$$b(s) = s^{10}+109s^9+4363s^8+88875s^7+1051404s^6+7674576s^5$$
$$+35302112s^4+101347520s^3+173235200s^2+158112000s$$
$$+57600000.$$

Consider an optimisation problem of equation (1) and (2) where

$$y_o(s) = 0, \quad y_d(s) = \frac{1}{s}$$

$$Q = q = 10, \quad R = r = 1.$$

The SDM P is then given by

$$P = \frac{c(s)}{d(s)}$$

where

$$c(s) = s^{20} - .34097 \times 10^4 s^{18} + .23851 \times 10^6 s^{16}$$
$$- .56591 \times 10^9 s^{14} + .59891 \times 10^{11} s^{12}$$
$$- .31936 \times 10^{13} s^{10} + .90333 \times 10^{14} s^8$$
$$- .135767 \times 10^{15} s^6 + .103025 \times 10^{17} s^4$$
$$- .346947 \times 10^{17} s^2 + .364958 \times 10^{17}$$

and

$$d(s) = s^{20} - .3155 \times 10^4 s^{18} + .17638 \times 10^7 s^{16} - .32667$$
$$\times 10^9 s^{14} + .276 \times 10^{11} s^{12} - .11974 \times 10^{13} s^{10}$$
$$+ .2778 \times 10^{14} s^8 - .33987 \times 10^{15} s^6 + .20382$$
$$\times 10^{16} s^4 - .50427 \times 10^{16} s^2 + .33178 \times 10^{16}.$$

To ensure that the suboptimal control will have the correct steady-state value and good transient approximation to the

optimal control, P is expanded using the CFE2 technique described in Chapter 4. An low-order approximation P_1 is obtained where

$$P_1 = \frac{-s^2 + 4.19136^2}{-s^2 + 1.26374^2}$$

For s=jw,

$$P_1(jw) = \frac{w^2 + 4.19136^2}{w^2 + 1.26374^2} > 0 \quad \forall \text{ real } w$$

Thus, $P_1(s)$ satisfies the hypothesis of the theorem by Youla [4] given in the Introduction and has a spectral factorisation

$$P_1(s) = G_1^T(-s)\, G_1(s)$$

where

$$G_1(s) = \frac{s+4.19136}{s+1.26374}$$

Therefore,

$$[(G_1^T(-s))^{-1} W^T(-s)\, Qz]^+$$

$$= \frac{3.0151}{s}$$

and, from equation (5a), the suboptimal control u^o is given by

$$\bar{u}^o = G_1^{-1}(s)\, [(G_1^T(-s)^{-1} W^T(-s) Qz]^+$$

$$= 3.0151 \quad \frac{s + 1.26374}{s(s+4.19136)} .$$

For comparison, the optimal control u^o of the system was calculated using the Differential Dynamic Programming technique. Fig. 1 shows the trajectories of the optimal control u^o (curve a) and the suboptimal control $\bar{u}^o$ (curve b).

It can be seen that the suboptimal control $\bar{u}^o$ gives a good transient approximation and has the correct steady-state value.

COMMENTS

It should be noted that, for $y_o=0$ and $y_d = \frac{a}{s}$ where a is a real constant, the initial and steady-state value of the control given by equation (5) depends on the ratio between the constant terms in the SDM while the transient behaviour is controlled by the difference in degree of the highest order terms. Thus, to achieve a good suboptimal control, the CFE2 proposed by the author in Chapter 4 has to be used instead of Chen's method. Also, Chen's method will result in a control which is unbounded

. It is due to the fact that the degree of the numerator is, generally, less than that of the denominator in a low-order approximation to P obtained by Chen's method. For instance, in the above example, the approximation, say $\bar{P}_1$, to P by Chen's method will be of the form

$$\bar{P}_1 = \frac{a_1^2}{-b_2^2 s^2 + b_1^2}$$

$$= \bar{G}_1^T(-s)\, \bar{G}_1(s)$$

where

$$\bar{G}_1(s) = \frac{a_1}{b_2 s + b_1}$$

Hence

$$\bar{G}_1^{-1}(s) = \frac{b_2 s}{a_1} + \frac{b_1}{a_1}$$

and the resulting suboptimal control $\bar{u}_1$ will be of the form

$$\bar{u}_1 = c_1 + \frac{c_2}{s}.$$

The constant term c_1 in the above equation represents a linear change in the magnitude of $\bar{u}_1$ with time.

(ii) It will be demonstrated in this example that the SDMA approach together with the iterative algorithm proposed in this chapter could yield a closer approximation to the optimal control than the popular reduced model approach.

The system, taken from [10], is a two-input two-output system of 4th order having the transfer function matrix

$$W(s) = \frac{\begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}}{C(s)} \quad .6250 \qquad (10)$$

where

$$w_{11} = .01496s^2+1.525s+2.54$$
$$w_{12} = 95.15s^2+1130s+1810$$
$$w_{21} = .0852s^2+8.6s+12.3$$
$$w_{22} = 124s^2+1491s+2520$$
$$C(s) = 6.25174s^4+703.16064s^3+8416.228s^2+21834.86s+15862.2.$$

This system has been used in Chapter 4 to demonstrate the results in model approximation. To use the reduced model approach, a low-order model $\tilde{W}(s)$ to $W(s)$ is obtained by CFEI technique proposed by Chen [9] where

$$W(s) = \frac{\begin{bmatrix} \tilde{w}_{11} & \tilde{w}_{12} \\ \tilde{w}_{21} & \tilde{w}_{22} \end{bmatrix}}{\tilde{c}(s)} .6250$$

where

$$\tilde{w}_{11} = 1.525s+2.54$$
$$\tilde{w}_{12} = 1130s+1810$$
$$\tilde{w}_{21} = 8.6s+12.3$$
$$\tilde{w}_{22} = 1491s+2520$$
$$\tilde{c}(s) = 7444s^2+21834s+15862$$

The output step responses of W and $\tilde{W}$ are practically the same, impart actually coincident ~~as shown~~ in Figs 2 and 3, where y_1 and y_2, respectively, are shown.

Consider the optimisation problem described by equation (1) and (2) where W is given by equation (10) of this example,

$$Q = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}, \quad R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$y_o(s) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad y_d(s) = \frac{1}{s} \begin{bmatrix} .7157 \times 10^3 \\ .10044 \times 10^4 \end{bmatrix}$$

Then, the SDM' P is given by

$$P = \frac{\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}}{C(-s)C(s)} \triangleq R + W^T(-s)Q\,W(s)$$

Where

$$p_{11} = .39084 \times 10^2 s^8 - .38920 \times 10^6 s^6 + .43245 \times 10^8 s^4 - .29160 \times 10^{11} s^2 + .61869 \times 10^{11}$$

$$P_{12} = .40829 \times 10^{10}s^4 - .41702 \times 10^{12}s^3 - .48973 \times 10^{13}s^2 - .12589 \times 10^{13}s + .13903 \times 10^{14}$$

$$P_{21} = .46829 \times 10^{10}s^4 + .41702 \times 10^{12}s^3 - .48973 \times 10^{13}s^2 + .12589 \times 10^{13}s + .13903 \times 10^{14}$$

$$P_{22} = .39084 \times 10^{2}s^8 - .38920 \times 10^{6}s^6 + .95428 \times 10^{13}s^4 - .98851 \times 10^{15}s^2 + .37603 \times 10^{16}$$

$$C(-s)C(s) = .39084 \times 10^{2}s^8 - .38920 \times 10^{6}s^6 - .40323 \times 10^{8}s^4 - .20976 \times 10^{9}s^2 + .25161 \times 10^{9}$$

The SDMA approach is now applied to obtain an approximation to the matrix P and hence an suboptimal control, say $\tilde{u}^o$, to the system. In order to obtain a good approximation to the transient behaviour of the optimal control u^o and achive the correct steady-state value, approximation in the frequency domain has to be made around $s=\infty$ and $s=0$. In this example, the first step is to obtain an approximation, say P_1, by the MPA technique about $s=\infty$ which yields a good approximation to the transient behaviour of the optimal control, but has the incorrect steady-state values. The approximation P_1 which is rational regular is found to be

$$P_1 = \frac{\begin{bmatrix} (p_1)_{11} & (p_1)_{12} \\ (p_1)_{21} & (p_1)_{22} \end{bmatrix}}{C_1(-s)\ C_1(s)}$$

where

$$(p_1)_{11} = s^4 - .995807859 \times 10^{4}s^2 + .110647152 \times 10^{7}$$

$$(p_1)_{12} = .119815717 \times 10^{9}$$

$(p_1)_{21} = .119815717 \times 10^9$

$(p_1)_{22} = s^4 - .995807859 \times 10^4 s^2 + .24416029 \times 10^{12}$

$C_1(-s)\ C_1(s) = s^4 - .995807859 \times 10^4 s^2 + .103168475 \times 10^7$

It is easy to check that the matrix $P_1(jw) > o$ $\forall$ real finite w and hence by the theorem due to Youla [4], it has a spectral factorisation,

$$P_1(s) = G_1^T(-s)\ G_1(s).$$

The method suggested by Davies [5] is used to find the matrices $G_1(s)$ and $G_1^T(-s)$ where

$$G_1(s) = \frac{\begin{bmatrix} (g_1)_{11} & (g_1)_{12} \\ (g_1)_{21} & (g_1)_{22} \end{bmatrix}}{C_1(s)}$$

where

$(g_1)_{11} = s^2 + .109568551 \times 10^3 s + .102369032 \times 10^4$

$(g_1)_{12} = .435938555s + .241653644 \times 10^3$

$(g_1)_{21} = .435938555s + .241653644 \times 10^3$

$(g_1)_{22} = s^2 + .999104363 \times 10^3 s + .494125725 \times 10^6$

$C_1(s) = s^2 + .10949665 \times 10^3 s + .101571884 \times 10^4$

The iterative algorithm will now be used to obtain an higher order approximation P_2 by applying the MPA technique about s=0 to the matrix V_1 in equation (6a). This will ensure that P_2 has the correct steady-state value and since $P_2 = P_1 + G_1^T(-s)\bar{V}_1 G_1(s)$ from equation (8) where $\bar{V}_1$ is an low-order approximation to V_1, P_2 will also retain the characteristics of P_1 which yields good transient approximation. Again, by Davies' method, the matrix L_1 in equation (8b) is found to be

$$L_1 = \frac{\begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix}}{m(s)}$$

where

$$L_{11} = 8.91454575s + 15.1068989$$
$$L_{12} = 1.05423892s + 7.67109367$$
$$L_{21} = -6.83607924s$$
$$L_{22} = 1.1102444435 + 3.8403921$$
$$m(s) = s + 1.08901938$$

and the higher order approximation P_2 generated by the iterative algorithm is then given by

$$P_2 = G_2^T(-s)\, G_2(s)$$

where

$$G_2(s) = L_1(s)\, G_1(s);$$

to find the suboptimal control $\tilde{u}^o$ with $G_2(s)$ in place of $G(s)$ in equation (3a), $G_2^{-1}(s)$ is required and found to be

$$G_2(s) = \frac{\begin{bmatrix} (g_2)_{11} & (g_2)_{12} \\ (g_2)_{21} & (g_2)_{22} \end{bmatrix}}{c_2(s)}$$

where

$$(g_2)_{11} = 1.11024443s^6 + .12328874 \times 10^4 s^5 + .674807638 \times 10^6 s^4 + .64067725 \times 10^8 s^3 + .83619689 \times 10^9 s^2 + .276296107 \times 10^{10} s + .209904692 \times 10^{10}$$

$$(g_2)_{12} = -1.0542389s^6 + .117366332 \times 10^4 s^5 + .644524354 \times 10^6 s^4 + .632033574 \times 10^8 s^3 + .101741731 \times 10^{10} s^2 + .488021978 \times 10^{10} s + .397433242 \times 10^{10}$$

$$(g_2)_{21} = 6.83607924s^6 + .15045077 \times 10^4 s^5 + .972640319 \times 10^5 s^4$$
$$+ .160022027 \times 10^7 s^3 + .836145166 \times 10^7 s^2$$
$$+ .638886247 \times 10^7 s - .102654509 \times 10^7$$

$$(g_2)_{22} = 8.91454575s^6 + .197814135 \times 10^4 s^5 + .130893167 \times 10^6 s^4$$
$$+ .237418998 \times 10^7 s^3 + .15541769 \times 10^8 s^2$$
$$+ .318666322 \times 10^8 s + .191566798 \times 10^8$$

$$C_2(s) = 17.1041855(s+5.42258322)\ (s+.625521538)$$
$$(s+10.3124537)(s+99.2558909)((s+.4993522275 \times 10^3)^2$$
$$+ .494543643^2 \times 10^6)$$

With the approximation $G_2^{-1}(s)$ in place of $G^{-1}(s)$ in equation (3a), we have

$$[(G_2^T(-s))^{-1} W^T(-s)Qz]^+ = -\frac{1}{s}\begin{bmatrix} .3467029124 \times 10^4 \\ .185648506 \times 10^4 \end{bmatrix}$$

and the resulting suboptimal control $\tilde{u}^o$ given by

$$\tilde{u}^o = \begin{bmatrix} u_1\ (s) \\ u_2\ (s) \end{bmatrix} = -G_2^{-1}(s)\ [(G_2^T(-s))^{-1} W^T(-s)\ Qz]^+$$

is then found to be

$$\tilde{u}_1^o(s) = -\frac{3.4363}{s} + \frac{2.304822}{s+.625521538} + \frac{110.862779}{s+5.42258322}$$
$$+ \frac{.693215285}{s+10.3124537} - \frac{.904520441 \times 10^{-2}}{s+99.2558909}$$
$$+ \frac{.20488(s+.4995522275 \times 10^3) - .13265 \times .494543643 \times 10^3}{(s+.4995522275 \times 10^3)^2 + .494543643^2 \times 10^6}$$

$$\tilde{u}_2^o(s) = \frac{.9907154}{s} - \frac{.58986126 \times 10^{-1}}{s+.625521538} + \frac{1.65901906}{s+5.42258322}$$
$$- \frac{.875765578 \times 10^{-3}}{s+10.3124537} + \frac{.41603634 \times 10^{-5}}{s+99.2558909}$$

$$+\frac{.235067078 \times 10^{4}(s+.4995522275 \times 10^{3})-.188218929\times 10^{4}\times .4945364\times 10^{3}}{(s+.4995522275 \times 10^{3})^{2}+.494543643^{2}\times 10^{6}}$$

By applying inverse Laplace transform to $\tilde{u}^{o}_{1}(s)$ and $\tilde{u}^{o}_{2}(s)$, the suboptimal control $\tilde{u}^{o}$ as a function of time can be obtained.

For comparison, the trajectories of the optimal control u^{o} calculated by Differential Dynamic Programming technique (curve a), the suboptimal control $\bar{u}^{o}$ by the reduced model approach (curve c) and the suboptimal control $\hat{u}^{o}$ by the SDMA approach (curve b) are shown in Fig. 4 and 5. Component 1 is shown in Fig 4, and component 2 in Fig 5, for u^{o}, $\bar{u}^{o}$ and $\tilde{u}^{o}$. It can be seen that the suboptimal control $\tilde{u}^{o}$ by the SDMA approach has a better transsient approximation to u^{o} than $\bar{u}^{o}$ obtained by the reduced model approach, especially in Fig. 4. Also, both $\hat{u}^{o}$ and $\bar{u}^{o}$ have the correct steady-state values.

That the reduced model approach yields poor transient approximation is due to the fact that the transient behaviour of the control is mainly determined by the behaviour of G(s) around $s=\infty$ whereas the reduced model approach using a model obtained by the CFE1 technique suggested by Chen is usually an approximation technique about s=0. By using the iterative algorithm, it is possible to obtain good approximation at both s=0 and $s=\infty$ by the SDMA approach, as shown in this example .

Control Trajectories

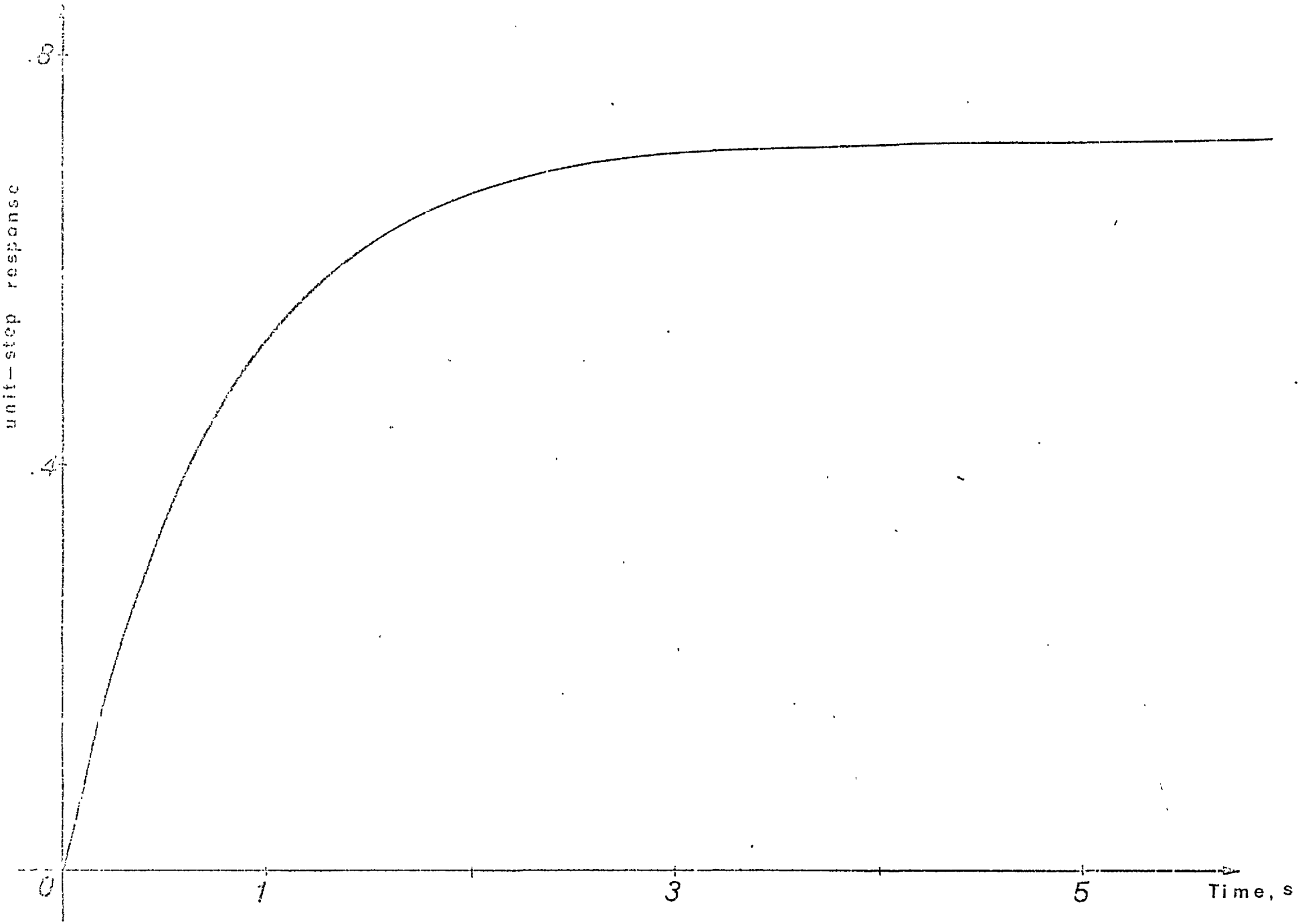


Fig 2

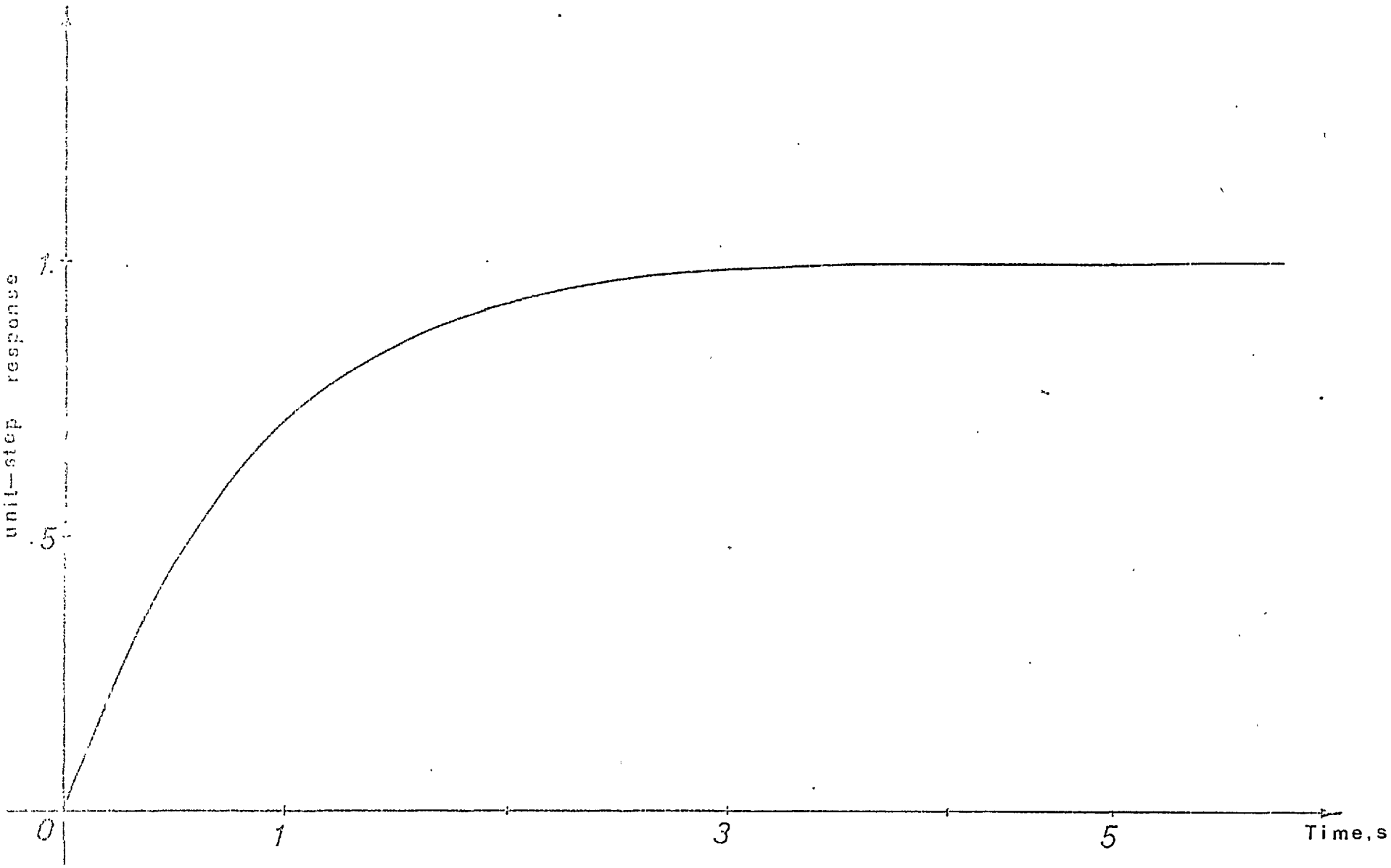


Fig 3

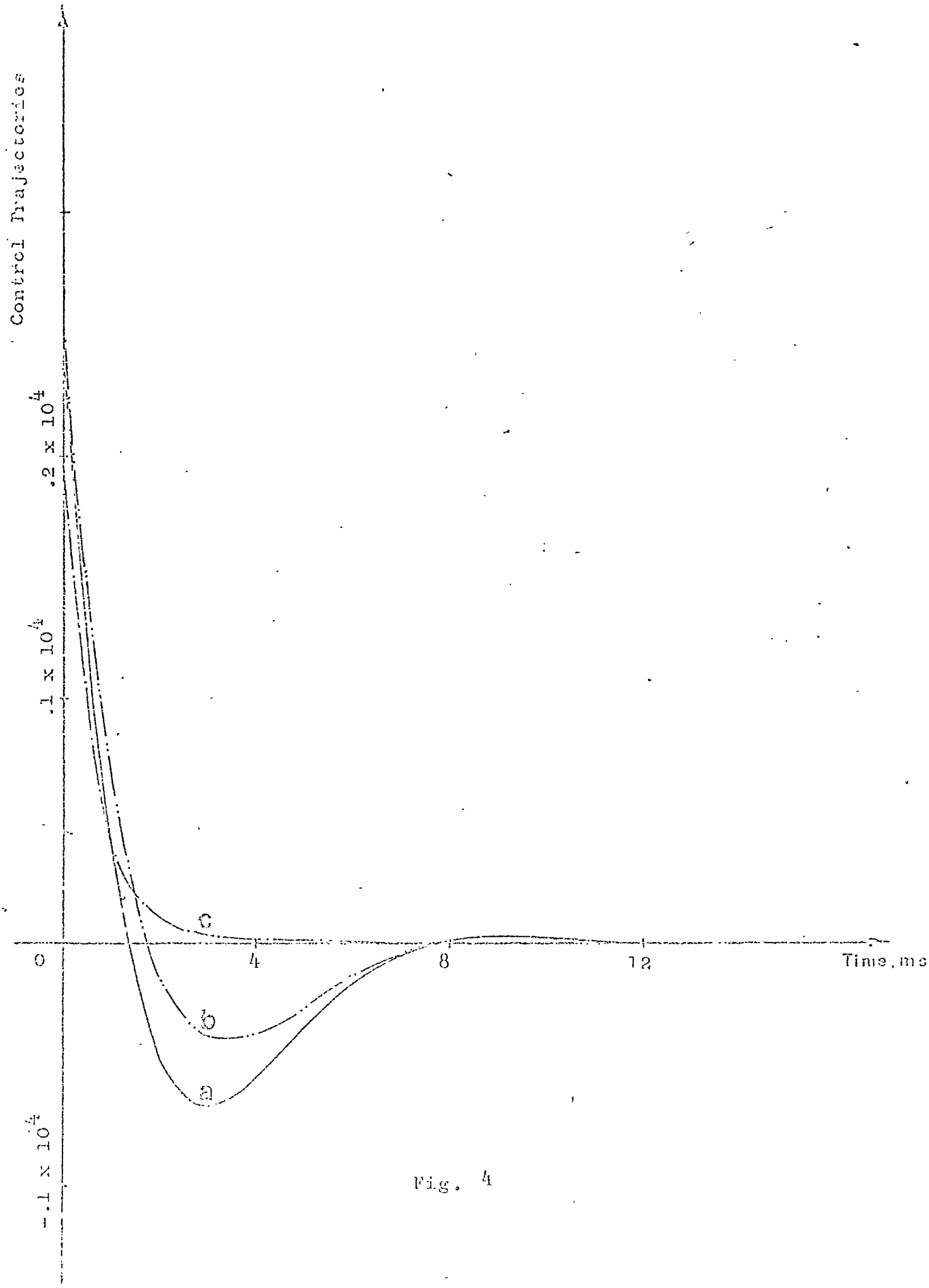


Fig. 4

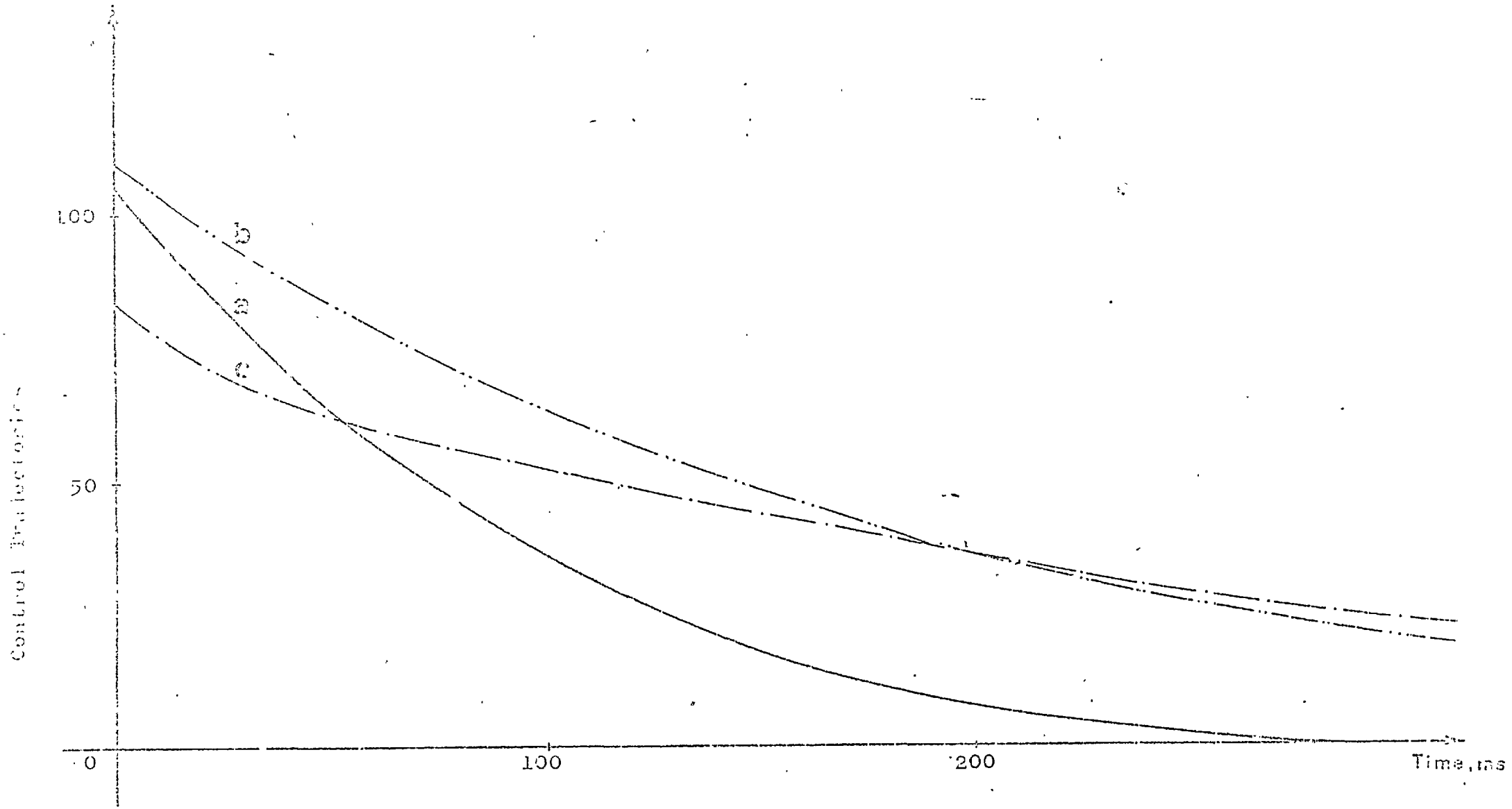


Fig. 5

180.

7. REFERENCES

[1] N. Wiener, The Extrapolation, Interpolation and Smoothing of Stationary Time Series, Wiley, New York, 1949.

[2] R.E. Kalman, On the General Theory of Control Systems, Proc. of IFAC, 1960, Vol. I. pp. 481-492.

[3] R.E. Kalmen, Contribution to the Theory of Optimal Control, Bol. Soc. Mat. Mexicana, 1960.

[4] D.C. Youla, On the factorisation of Rational Matrices, I.R.E. Trans. Information Theory, Vol. IT-7, July, 1961, pp. 172-189.

[5] M.C. Davies, Factoring the Spectral Matrix, IEEE Trans. Automatic Control, Vol. AC-8, October, 1963, pp. 296-305.

[6] B.D.O. Anderson, An Algebraic Solution to the Spectral Factorisation Problem, IEEE Trans. Automatic Control, Vol. AC-12, August 1967, pp.410-414.

[7] H. Nicholson, Dynamic Optimisation of a Boiler, Proc. IEE, Vol. 111, 1964, pp. 1479-1499.

[8] E.J. Davison, A New Method for Simplifying Large Linear Dynamic Systems, IEEE Trans. Automatic Control, Vol. AC-13, April 1968, pp. 296-305.

[9] C.F. Chen, A Matrix Cauer Model in Multivariable Systems Analysis and Design, 4th UKAC Multivariable Control System Design and Application Conference, September 1971, pp. 66-70.

[10] H.H. Rosenbrock and P.K. McMorran, State-Space Analysis of Cascaded Controller, Proc. IEE, Vol. 117, No. 5, May 1970, pp. 1026-1030.

CHAPTER 8

CONCLUSIONS AND FURTHER RESEARCH

CONCLUSIONS AND EXTENSIONS

The details of the Contributions of the thesis have been given in Section 3 of Chapter 1. This thesis contains several approximation techniques for optimal control in both time and frequency domains. Two different approaches have been considered. The first approach which is popular, makes use of a simplified model to the actual physical system being considered. This leads naturally to the problem of model simplification which has been studied in both time and frequency domains. The other approach, which has been considered only in the frequency domain, does not require a simplified model for the system dynamics but involves using a low-order approximation to the spectral density matrix of the linear regulator problem.

The success of the popular simplified model approach will obviously depend on the performance of the final system using the optimal control of the model. It would be expected that the performance of the system being controlled would improve with improvements in the accuracy of the model used. That the lower bound results on the minimum performance index can be used to test the accuracy of the model has been suggested and a numerical example has been given. Many model simplification techniques in the frequency domain are based on the concept of power series expansions and the numerical results have shown that reasonably good models can be obtained by those techniques presented in this thesis. However, the stability of the model obtained by these techniques is, by no means,

guaranteed. It should thus be useful to devote some effort to investigate the necessary and, if possible, sufficient conditions to ensure the stability of the model. One possible way of getting a stable model is to use the technique proposed in Chapter 5, which is applicable in both time and frequency domains, with a suitably chosen set of orthonormal functions, e.g. Laguarre functions with positive λ_i. But there is still no simple way of choosing an optimal set of eigenvalues λ_i. The parallel model appears to be useful and numerical experience indicates that a satisfactory control can be obtained by applying only a short recent past history of the input to the convolution-described subsystem; of course, the p required to achive a good control will also depend on the modes described by the state-space subsystem. Thus, an useful but not necessarily easy extension is to find a new measure of model accuracy which will lead to a simple expression for the model accuracy as a function of p. Further study is also required on the stability of the closed system resulting when the optimal 'feedback' control of the parallel model is applied to the actual system.

An interesting fact about the second approach, treated only in the frequency domain, is that no simplified model is needed so that the question of instability does not arise. Numerical results showed that this approach could yield a more satisfactory control than the popular simplified model approach and is thus

potentially useful.

The suboptimal control obtained by using either approach has but the first time moment equal to that of the optimal control. It is, therefore, of interest to investigate the existence of approximation techniques which yield suboptimal control with time moments equal to those of the optimal control, up to certain order.