

IMPERIAL COLLEGE OF SCIENCE AND TECHNOLOGY
(UNIVERSITY OF LONDON)

DEPARTMENT OF COMPUTING AND CONTROL
(CONTROL SYSTEMS SECTION)

A VARIATIONAL APPROACH TO LINEAR CONTROL STRUCTURE PROBLEMS

BY

MICHAEL ARTHUR JOHNSON

A THESIS SUBMITTED FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

MARCH, 1978

Abstract

Large and often complex simulations are usually employed in any serious industrial study. In the first stage of the study, the designer frequently needs to solve at least two important structure problems. The first is the selection of the smallest possible set of controlled and measured variables necessary for a satisfactory design. In practice, a surprisingly small subset may suffice, and the inclusion of additional variables may represent a considerable installation investment, and maintenance liability. The second is the "pairing" of the input and output variables to obtain a satisfactory sequence of simple feedback/feedforward designs hopefully of a single input/single output nature. Once the "pairing" is decided, and provided only low dimensional subdesigns are involved, modern control theory provides powerful design tools whereas little aid seems available to solve the two structure problems. The use of variational methods in constructing practical design aids for the structures question represents the main theme of this thesis.

A general form of the linear dynamic quadratic cost problem is utilized in obtaining a formal problem statement. This form is used in various ways to study structure questions and the detection of important control theoretic properties. Asymptotic properties of the cost function and a related characterization of the system pole-zero patterns turn out to be of central interest.

Straightforward solution of the structures problem is easily shown to pose significant combinatorial and computational problems. Therefore considerable attention is devoted to the use of approximation

procedures, in particular to the development of cost bounds.

Subsequently a viable algorithm is constructed for problems of non-trivial dimensionality, and an industrial example is used to illustrate the algorithm's capabilities.

The structures problem is of interest in a far wider field than control dynamic design, and several of these areas are briefly discussed.

To my Mother, and Father,

and to my wife, Margaret.

Acknowledgements

I wish to sincerely thank my supervisor Dr. Greyham Bryant (Project Director, Industrial Automation Group) for suggesting the theme of this thesis as a topic for extended research. His continued enthusiasm as the project developed, and his suggestions on the finished manuscript were gratefully appreciated.

Thanks are also extended to Dr. Peter Spooner, and John Caley for technical assistance and stimulating discussions on the use of the five stand rolling mill model as a trial problem for the structures algorithm.

It is a particular pleasure to thank Mrs. Christine Carter for typing such a difficult manuscript with exceptional efficiency.

Contents and Page Index

Title Page	Page No. 1
Abstract	2
(Dedication)	4
Acknowledgements	5
Contents and Page Index	6
Tables	15
Figures	17
Notation	18

Introduction

I.1 Structures and Controller Design

I.1.0 Introduction

I.1.1 Controller Structures in the Literature of Linear Systems

I.1.2 Two Industrial Examples

I.1.3 Conclusions

I.2 The Problem of Isolating Controller Structures

I.2.0 Introduction: Structures in the Design Process

I.2.1 Basic Problem Formulation

I.2.2 The Input Structure Diagnostics

Part 1: Input Structures

Chapter 1 Asymptotic Optimal Control Theory for Linear Systems

- 1.0 Introduction
- 1.1 The Problem Basics and the Optimal Return Difference Relation
- 1.2 Generalization of a Result of Kalman's
- 1.3 The Infinite Zeros and their Butterworth Patterns
- 1.4 Summary of Asymptotic Optimal Control Theory

Chapter 2 The Control Properties of the Input Structure Diagnostics

- 2.0 Introduction
- 2.1 Asymptotic Properties for the Input Structure Diagnostics
- 2.2 Basic Relations for the Input Structure Assessment
- 2.3 Input Deletion and the Preservation of Controllability
- 2.4 Input Deletion and the Loss of Controllability
- 2.5 Summary of the Control Properties of Input Structure Diagnostics

Chapter 3 Bounds for the Input Structure Diagnostics

- 3.0 Introduction
- 3.1 A Hilbert Space Canonical Form for Diagnostic, ΔJ_0 and the Formulation of a Reduced Order Optimization Problem
- 3.2 Analytical Bound Formulae for the Diagnostic,
 $\Delta J_0(r=0,1,\dots,4)$
- 3.3 Analytical Bound Formulae for the Diagnostic,
 $IESQ(r=0,1,\dots,4)$
- 3.4 Properties of the Moments and the Bounds
- 3.5 Asymptotic Properties of the Bound Formulae.

3.6 A Finite Time Domain Example

3.7 Summary

Chapter 4 Aspects of Bound Computation for Structures

4.0 Introduction

4.1 Calculation of the Structure Moment Sequence

4.2 The Decomposition of Moments

4.2.0 Decomposition of Moment, m_0

4.2.1 Decomposition of Moment, m_1

4.2.2 Decomposition of Moment, m_2

4.2.3 Decomposition of Moment, m_3

4.2.4 Decomposition of Moment, m_4

4.3 The Formation of Structure Moment Sequences

4.4 The Problem of an Upper Bound for the Maximum Eigenvalue

$$\lambda_M \text{ of Operator } D = Q^{\frac{1}{2}}WR^{-1}W^*Q^{\frac{1}{2}}$$

4.4.0 Introduction

4.4.1 Eigenvalue Bounds for Single Input-Single Output Systems

4.4.2 Eigenvalue Bounds for Multivariable Systems

4.4.3 Implementation of the Bounds on the Maximum Eigenvalue for the Assessment of Input Structures

4.5 An Algorithm for Input Structure Assessment

Chapter 5 An Industrial Example

5.0 Introduction

5.1 Plant, Disturbance and Problem Description

5.2 The Logistics of the Calculation of Input Structure Diagnostic Bounds

- 5.3 Control Structure Interpretation of the Calculated Bounds
 - 5.3.1 Bounds for a 10% Step Disturbance on the Exit Gauge of Stand 2 and their Structural Implications.
 - 5.3.2 Bounds for a 10% Step Disturbance on Interstand Tension, T_2 and their Structural Implications
- 5.4 Conclusions

Chapter 6 Extensions and Applications in the Infinite Time Domain

- 6.0 Introduction
- 6.1 Input Structure Assessment and the Infinite Time Domain
- 6.2 Removal of Initial Condition Dependence of Input Structure Assessment for State Space Systems
- 6.3 Diagonal Dominance and Structures
 - 6.3.0 Introduction
 - 6.3.1 On Pseudo-Diagonalisation
 - 6.3.2 On Diagonal Dominance
- 6.4 Least Squares Parameter Estimation Problems and an Application of the Structures Procedure.

Part 2: Output Structures

II.0 Introduction

Chapter 7 Primal and Dual System Concepts

- 7.0 Introduction
- 7.1 Primal and Lagrange Dual Systems
- 7.2 Conclusions

Chapter 8 Optimal Output Structures for Linear Stochastic Systems

8.0 Introduction

8.1 Minimum Variance Estimators

8.2 Output Structures for Linear Stochastic Systems

8.2.0 Introduction

8.2.1 Basic Theory

8.2.2 An Output Structures Problem for Linear Stochastic Systems

8.2.3 Controllers of Reduced Dimension

8.3 Conclusions

III.0 Future Research

IV.0 References

Appendix A: A Theorem for the Smith Form of $\phi^T(-S) \phi(S)$ and Other Results

Appendix B: Some Linear Space Concepts

Appendix C: Norm Concepts; Lemmas and Theorems Relating to the Problem of Bounding the Maximum Eigenvalue of VV^* .

Appendix D: On the Modified Dominance Cost Function.

Page Index For Contents

Introduction

I.1 Structures and Controller Design

<u>Section</u>		<u>Page No.</u>	
I.1.0		I-1-1	
I.1.1		I-1-8	
I.1.2		I-1-16	
I.1.3		I-1-18	

I.2 The Problem of Isolating Controller Structures

<u>Section</u>		<u>Page No.</u>	
I.2.0		I-2-1	
I.2.1		I-2-3	
I.2.2		I-2-5	

Part 1 Input Structures

Chapter 1 Asymptotic Optimal Control Theory for Linear Systems

<u>Section</u>		<u>Page No.</u>	
1.0		1-1	
1.1		1-2	
1.2		1-6	
1.3		1-9	
1.4		1-18	

Chapter 2 The Control Properties of Input Structure Diagnostics

<u>Section</u>		<u>Page No.</u>	
2.0		2-1	
2.1		2-4	
2.2		2-17	
2.3		2-38	
2.4		2-34	
2.5		2-46	

Chapter 3 Bounds for the Input Structure Diagnostics

<u>Section</u>		<u>Page No.</u>	
3.0		3-1	
3.1		3-3	
3.2		3-11	
3.3		3-31	
3.4		3-35	
3.5		3-44	
3.6		3-54	
3.7		3-58	

Chapter 4 Aspects of Bound Computation for Structures

<u>Section</u>		<u>Page No.</u>	
4.0		4-1	
4.1		4-3	
4.2			
4.2.0		4-5	
4.2.1		4-5	
4.2.2		4-5	
4.2.3		4-6	
4.2.4		4-7	
4.3		4-9	
4.4			
4.4.0		4-11	
4.4.1		4-15	
4.4.2		4-19	
4.4.3		4-21	
4.5		4-25	

Chapter 5 An Industrial Example

<u>Section</u>		<u>Page No.</u>	
5.0		5-1	
5.2		5-6	

<u>Section</u>	5.3	<u>Page No.</u>	
	5.3.1		5-10
	5.3.2		5-11
	5.4		5-13

Chapter 6 Extensions and Applications in the Infinite Time Domain

<u>Section</u>	6.0	<u>Page No.</u>	6-1
	6.1		6-2
	6.2		6-9
	6.3		
	6.3.0		6-16
	6.3.1		6-17
	6.3.2		6-29
	6.4		6-40

Part 2 : Output Structures

II.0	<u>Introduction</u>	II-0
------	---------------------	------

Chapter 7 Primal and Dual System Concepts

<u>Section</u>	7.0	<u>Page No.</u>	7-1
	7.1		7-2
	7.2		7-6

Chapter 8 Optimal Output Structures for Linear Stochastic Systems

<u>Section</u>	8.0	<u>Page No.</u>	8-1
	8.1		8-2
	8.2		
	8.2.0		8-5
	8.2.1		8-6
	8.2.3		8-11
	8.3		8-17

III.O Future Research

III-O

IV.O References

IV-O

Appendix A

Pages A-0 to A-9

Appendix B

Pages B-0 to B-7

Appendix C

Pages C-0 to C-14

Appendix D

Pages D-1 to D-5

Tables

	<u>Page No.</u>
3.6.1 Table of Bounds: Using Analytical Moment Values	3-54
3.6.2 Variation of Upper Bounds on ΔJ_0 with respect to λ_M	3-55
3.6.3 Variation of Lower Bounds on ΔJ_0 with respect to λ_M	3-55
3.6.4 Moment and Moment Ratio Sequences for Different Approximation Intervals	3-56
3.6.5 Fourth Order Bounds for ΔJ_0 obtained using Different Approximation Intervals	3-57
3.6.6 Fourth Order Bounds for IESQ obtained using Different Approximation Interval Sizes	3-57
4.2.1 Computer Storage Requirements for Moment Decomposition	4-8
4.3.1 Total Number of Structures for some Values of m	4-9
4.4.1 Convergence of a Structure Moment Sequence	4-11
4.4.2 Generation of $\Omega(t) = \int_0^{T_f} \left \int_0^{\min(t,\tau)} v(t-s)v(\tau-s) ds \right dt$	4-18
4.4.3 Summary of Eigenvalue Bounds for Single Input-Single Output Systems	4-18
5.2.1 The Computing Time used to Calculate the Bounds for a Complete Set of Structures	5-7
5.2.2 The Percentage Error in the Four Bounds for Twelve Selected Structures and Details of the Power Method for these Structures	5-8
5.3.1 10% Step Disturbance to Exit Gauge, Stand 2 Cost Functional Weights : Set I	5-14
5.3.2 10% Step Disturbance to Exit Gauge, Stand 2 Cost Functional Weights : Set II	5-16
5.3.3 Bounds of Table 5.3.1 Recalculated Using Exact Eigenvalues	5-15

5.3.4	10% Step Disturbance to Interstand Tension, T_2	5-17
	Cost Functional Weights : Set I	
5.3.5	Bounds of Table 5.3.4 Recalculated Using Exact Eigenvalues	5-17
5.3.6	10% Step Disturbance to Interstand Tension, T_2	5-18
	Cost Functional Weights : Set II	
5.3.7	Bounds of Table 5.3.6 Recalculated Using Exact Eigenvalues	5-18
6.1.1	Error in $m_o(\infty)$ versus Approximation Interval, T_f	6-7
6.1.2	Variation of Numerical Results for Different	6-7
	Approximation Intervals and Integration Intervals, T .	
6.4.1	Data : Draper and Smith (1966) Example	6-42
6.4.2	Details of Complete/Structure Least Squares Solutions ($\rho=0$)	6-43
6.4.3	Optimal Costs and Parameters for the Modified Cost	6-43
	where $Q=I_{13}$, and $R_o=I_5$ ($0 \leq \rho \leq 0.1$)	
6.4.4	Three Parameter Solutions obtained for $J_M(\rho=10^{-5})$	6-44
	Ranked by $\Delta J_o^S(\rho)$	
6.4.5	Three Parameter Solutions Ranked by the Lower Bound	6-44
	Values on Diagnostic $\Delta J_o^S(\rho)$	
6.4.6	Optimal Cost and Parameters for the Modified Cost Problem	6-45
	where $Q=I_{13}$, using R_o (modified) with $0 \leq \rho \leq 1$	
6.4.7	Three Parameter Solutions obtained for $J_M(\rho=10^{-5})$,	6-46
	Ranked by $\Delta J_o^S(\rho)$	
6.4.8	Three Parameter Structures Ranked by Lower Bound Values	6-46
	on the Diagnostic $\Delta J_o^S(\rho)$	
6.4.9	Variation of the Lower Bound Values for ΔJ_o^S	6-47
	(Using the Modified Coefficient Weight) for $10^{-5} \leq \rho \leq 1.0$	

<u>Figures</u>	<u>Page No.</u>
2.2.1 A Geometric Analogy for the Decomposition	2-25
$H_r = \overline{R(W_s)} \oplus X \oplus N(W)$ with $\overline{R(W)} = \overline{R(W_s)} \oplus X$	
2.2.2 Relationship between the Optimal Control Action $\hat{Z}(\rho)$ and the Optimal Output $\hat{Y}(\rho)$	2-25
2.2.3 The Geometric Relationship between the Structure Optimal Control Action $\hat{Z}_s(\rho)$ and the Optimal Structure Output, $\hat{Y}_s(\rho)$	2-26
2.4.1 Open Loop Structured System	2-42
2.5.1 Decomposed Structured System	2-49
3.6.1 The Asymptotic Behaviour of the Diagnostic Bounds for ΔJ_0	3-48
4.4.1 Typical Behaviour of Estimates of the Maximum Eigenvalue λ_M , For Various Structures	4-12
4.5.1 An Algorithm for Input Structure Assessment	4-26
5.1.1 Diagram Showing Mill Variables	5-3
5.1.2 Normalized Mill Step Responses Four Stand Cold Rolling Mill	5-3a,b
5.1.3 Decoupling Network for Stand 2 Exit Gauge	5-3
5.1.4 Decoupling Network for Interstand Tension, T_2	5-3
6.3.1.1 The Geometric Configuration of $q \in Q_{D_j}$ such that $\ q\ = \mu$	6-20
6.3.1.2 Geometric Configuration for Optimization Problems	6-22
6.3.2.1 Desired Dominance Trajectories and the Dominance achieved for Hawkins Example	6-39
7.1.1 Primal/Dual System Configuration	7-3
8.2.1.1 Linear Stochastic System Configuration	8-6

Notation

General Symbols:

$\approx$	approximately equal		
$\triangleq$	equal by definition		
$\equiv$	equivalent		
$\Rightarrow$	implies	$\Leftrightarrow$	implies and is implied by
$\forall$	for all	$\exists$	there exists
s.t.	such that		
$ \cdot $	modulus	$x + j y$	complex number representation
$\text{Re}[\cdot]$ and $\text{Im}[\cdot]$	real and imaginary parts of a complex number		
$\lim$	denotes deleted limit		
$E[\cdot]$	statistical expectation operator		

'o' Landau symbol: Sequence $X = \{x_n\}$ is of lower order of magnitude than non-zero sequence $Y = \{y_n\}$ written

$$X = o(Y) \text{ or } x_n = o(y_n) \text{ for which } \lim_{n \rightarrow \infty} \left(\frac{|x_n|}{|y_n|} \right) = 0.$$

Set Theoretic Notation:

$x \in X$ denotes that element x is a member of set X

$\cap$ denotes set intersection

$\subset$ denotes set containment

ϕ denotes the null set

Linear Space Notation:

A transformation W , mapping the space X into space Y is written as $W: X \rightarrow Y$.

Hence $\mathcal{R}(W)$ and $\mathcal{N}(W)$ denote the range and nullity respectively of linear transformation W .

The mapping of element x into element y by the transformation W is written $y = Wx$ for which a frequent definition is given the form $y(.) \triangleq (Wx)(.)$.

W^* denotes the adjoint, a particular example is the matrix transpose denoted W^T .

The inner product of vectors x , and y on the Hilbert space H_n is denoted $\langle x, y \rangle_{H_n}$.

$\mathbb{R}^n$ denotes the space of all ordered n -tuples of real numbers, an element of which is written

$$x \triangleq (x_1, x_2, \dots, x_n).$$

E_n the Euclidean space of all ordered n -tuples of real numbers

with inner product $\langle x, y \rangle_{E_n} \triangleq \sum_{i=1}^n x_i y_i$

and norm $\|x\| = (\langle x, x \rangle_{E_n})^{\frac{1}{2}}$

$L_p [0, T_f]$ the space of real Lebesgue integrable functions on the interval $0 \leq t \leq T_f$ with norm

$$||x||_p \triangleq \left(\int_0^{T_f} |x(t)|^p dt \right)^{1/p}$$

H_n the Hilbert space usually taken to be $L_2 [0, T_f]$

so that $\langle x, y \rangle_{H_n} \triangleq \int_0^{T_f} \langle x(t), y(t) \rangle_{E_n} dt$

and $||x||_{H_n}^2 = ||x||_2^2 \triangleq \int_0^{T_f} \langle x(t), x(t) \rangle_{E_n} dt.$

$\oplus$ denotes direct sum $^\perp$ denotes orthogonal complement

I denotes identity operator I_r denotes $r \times r$ unit matrix

$\det [.]$ denotes determinant.

Matrix $A: \mathbb{R}^m \rightarrow \mathbb{R}^n$ is written in sub-script form as

$A = [\alpha_{ij}]_{i=1}^n \substack{n \\ m} \substack{j=1}^m$, in column form as

$A = [\alpha_{\cdot 1}, \dots, \alpha_{\cdot m}]$ where $\alpha_{\cdot j} \in \mathbb{R}^n, j = 1, 2, \dots, m$

or in row form as $A = \begin{bmatrix} \alpha_{1\cdot} \\ \alpha_{2\cdot} \\ \vdots \\ \alpha_{n\cdot} \end{bmatrix}$ where $\alpha_{i\cdot}^T \in \mathbb{R}^m, i=1, \dots, n.$

Other Symbols

The state space system definition employs a system triple notation

$$\mathcal{S}(A,B,C) \triangleq \begin{cases} \dot{x}(t) = Ax(t) + Bu(t) & x_0 = x(t_0) \\ y(t) = Cx(t) & t_0 \leq t \leq T_f \end{cases}$$

where A, B and C are real constant matrices of dimensions $(n \times n)$, $(n \times m)$ and $(r \times n)$ respectively. The one-sided laplace transform is denoted $\mathcal{L}_1[.]$ whilst the double sided transform is written $\mathcal{L}_2[.]$. In both cases the laplace transform variable is denoted by s , and frequencies on the imaginary axis $s = j\omega$ are often of interest. $P_+x \triangleq \{x(s)\}_+$ defines a projection isolating that part of $x(s)$ which is analytic in the closed right half plane.

The two diagnostics of the thesis are denoted

$\Delta J_0 \sim$ optimal cost improvement and

$IESQ \sim$ the integral error squared value. Optimality of a vector is usually denoted by the ' $\wedge$ ' notation and frequent use is made of

subscripts ' s ' and ' f ' where ' s ' denotes a subset of controls

(a structure) and ' f ' denotes the full or complete set of controls

available. Dependence of the diagnostics on an asymptotic

parameter ρ is usually written $\Delta J_0(\rho)$ or $IESQ(\rho)$ whilst

dependence on a control structure $\hat{u}_s$ may be written $\Delta J_0(\hat{u}_s)$ or ΔJ_0^s ,

similarly $IESQ(\hat{u}_s)$ or $IESQ^s$.

Introduction

I.1 Structures and Controller Design

I.1.0 Introduction

I.1.1 Controller Structures in the Literature of Linear Systems

I.1.2 Two Industrial Examples

I.1.3 Conclusions

I.2 The Problem of Isolating Controller Structures

I.2.0 Introduction: Structures in the Design Process

I.2.1 The Problem Basics

I.2.2 The Input Structure Diagnostics

I.1 Structures and Controller Design

I.1.0 Introduction

The decision to adopt a particular control scheme for implementation on a complex process is almost always taken after a thorough investigation of the economics of installation, and maintenance, and after quantifying the benefits which will accrue from its operation. The techniques used to model the physical aspects of the process usually involve advanced mathematical theory, and probably several types of computer simulations but ultimately the designer knows his control scheme should be simple and yet effective. Ideally only a few inputs and selected measurements will figure in the design, and frequently through a judicious choice such a small selection is found to suffice. This problem of choosing an input-output structure on which to base the control design is the theme of this thesis. The concept of structures as described here is not only relevant to the problems of controller design but find a much wider application in quite diverse fields. To help clarify the concept, several of these applications are detailed further.

In linear regression analysis, the problem arises of fitting a linear combination of n independent variables to a set of N given observations.

The fit is usually described
$$Y_i = \sum_{j=1}^n \beta_j X_{ij} + e_i, \quad i = 1, \dots, N$$

where $\{Y_i\}_{i=1}^N$ form the given observations

$\{X_{ij}\}_{i=1}^N \quad j=1, \dots, n$ are the independent variables

$\{\beta_j\}_{j=1}^n$ and $\{e_i\}_{i=1}^N$ are the regression coefficients and residual errors respectively.

These N equations may be succinctly written as:

$$\underline{Y} = W \underline{\beta} + \underline{e} \quad \text{where } \underline{Y} \in \mathbb{R}^N, \underline{e} \in \mathbb{R}^N, \underline{\beta} \in \mathbb{R}^n \text{ and } W: \mathbb{R}^n \rightarrow \mathbb{R}^N.$$

The optimal set of regression coefficients is chosen to minimize the least squares cost function $J = \frac{1}{2} \sum_{i=1}^N e_i^2 \triangleq \frac{1}{2} \langle \underline{e}, \underline{e} \rangle_{E_N}$.

The subsequent regression analysis determines which of the n regression coefficients β_j should be included in the linear relation to explain the given data. This represents a simple form of the class of structure problems discussed in this thesis. Namely to decide which of the β_j 's may be zeroed without appreciable degradation in "optimal cost". In regression analysis, the degradation is measured using F ratio tests in an ANOVA table. Each of the structures may be written $\beta_s = \{\beta_{i_1}, \beta_{i_2}, \dots, \beta_{i_s}\}$, $i_s \leq n$, and the linear equation as $\underline{Y} = W_s \beta_s + \underline{e}$, where the matrix $W_s: \mathbb{R}^s \rightarrow \mathbb{R}^N$ is obtained from the matrix W by deletion of those columns associated with the β_j 's considered to be zero. The least squares structures problem is then
$$\min_{\beta_s \in \mathbb{R}^s} J = \frac{1}{2} \langle \underline{e}, \underline{e} \rangle_{E_N}$$

subject to $\underline{Y} = W_s \beta_s + \underline{e} \in \mathbb{R}^N$.

To consider all the different regression structures, a total of $(2^n - 1)$ candidates would have to be assessed. To overcome this combinatorial aspect of regression structure assessment various statistical techniques have been devised. Backward elimination is one. This begins from the full optimum where all the β_j , $j = 1, \dots, n$ have been fitted and utilizes an F-test to discard variables one by one until no further variables may be removed. In contrast forward selection operates in the opposite direction; building up the regression set from single variables to pairs and so on until no further variables can be included. Draper and Smith (1966) have given examples where these two techniques provide different regression structure for the same set of data.

Even this simple problem exhibits all the essential ingredients of the structure problems of interest in this thesis:

- 1) The problem has a combinatorial aspect, which for a large problem may well preclude the simple enumeration of all combinations.
- 2) The selection of best coefficient subsets is not obtained by setting to zero those coefficients close to zero, unless the problem is well conditioned.
- 3) If sequential techniques are used to isolate a best parameter structure the problem arises of ensuring no better structure exists.

An application of the techniques of this thesis to a linear least squares problem (Draper and Smith, 1966) is given in the sequel.

Consider now a controller design problem. Taken from linear

systems theory and receiving considerable research attention is the design of post compensators to "square" a system prior to the application of a linear multivariable design procedure. Assuming a state space system description $\dot{S}(A,B,C)$ where there are more outputs than inputs ($r > m$), a proposed solution to the "squaring" problem (Kouvaritakis, 1972) is the application of a matrix compensator $C_1: \mathbb{R}^r \rightarrow \mathbb{R}^m$ to give a new output set $y_{SQ} = C_1 y$. Although such a proposal is mathematically tidy, it is not necessarily good engineering design. An important concern here, as remarked earlier, is the minimization of the numbers of measured and manipulated physical variables. Since every measured, and manipulated variable represents a major cost in transducers, actuators, expensive cabling, and long term maintenance liability, and the minimization of these overheads is of dominant interest. Hence a seemingly innocuous matrix manipulation can sweep in a vast array of these overheads in the first stage of the design. It is instructive to note that all the multivariable design techniques (Inverse Nyquist Array, Sequential Return Difference, Characteristic Loci) tend to ignore the basic engineering realities in this manner..

For example, the selection of diagonal dominance pre-compensators prior to the application of the Nyquist Array techniques is one area where structure assessment might be profitably performed. The definition of diagonal dominance provides the "cost function" from which optimal pre-compensators can be calculated, and through

which sub-optimal compensators can be compared. Assuming a constant precompensator matrix, and utilizing the technique of pseudo-diagonalisation, it is shown in this thesis how a quadratic cost of precompensator gains can be utilized as the basis of a structural assessment. Subsequently a least squares type of optimization formulation is used to determine whether or not pre-compensator gains may be usefully zeroed.

These introductory examples have demonstrated other salient features of structures assessment. The problems usually specify a cost function, or diagnostic from which to determine the optimal solution, and against which to measure the degradation caused by a sub-optimal structure. The combinatorial aspect of the assessment makes it imperative to avoid the direct evaluation of the many optimization problems; since for n inputs whether regression coefficients, compensator gains or control inputs there are $(2^n - 1)$ possible combinations or structure to be assessed.

A recurrent theme in the literature of industrial control applications is the availability of only a few states for measurement and the use of 'simple' control loop structures to obtain satisfactory control. Such schemes are easily tuned and economical to install and maintain. Sometimes they are designed on the basis of known physical interactions within the process and occasionally not even on this basis but as a series of single loop designs for which the total interaction when all loops are closed simultaneously, is not too severe. The main difficulty in bringing some rationale to this method of design for large scale processes is to provide answers of a

qualitative nature to fundamental questions such as: "Which input subset can be used to control a particular process variable or set of process variables satisfactorily?" or "Which measurement or set of measurements should be used to generate the necessary control action?" Although attention has previously been drawn to the problem of answering these questions (Rosenbrock, 1962; Foss, 1973) the research response has been surprisingly disappointing so that the only methods available involve either the direct simulation of several controller structures, or the structural simplification of a complex but satisfactory controller. The difficulty with the first approach is that it represents a compromise answer to the brutal solution of simulating all possible structures. Clearly with this compromise, it is possible for the designer to miss the "good" structure. The method of structural simplification unfortunately lacks reliable design aids to be considered as a viable technique.

In this thesis, a diagnostic or variational index is utilized to provide the qualitative answers to some controller structure questions. An important ingredient of the theory, the diagnostic accommodates the relative costs of the control objectives, and indicates the control properties of the various structures. Asymptotic optimal control theory has been used to show how the index reflects the system control properties, and how the control properties of structures can also be inferred from this asymptotic behaviour.

The combinatorial problem of structures remains to be solved. Here bound estimates of the cost, which can be computed with far less work than direct evaluation of the optimal solution, are developed.

It will be seen that the formulation of the structures problem using Hilbert space notation lends itself to re-interpretation as a finite dimensional structures problem. A feature which permits applications of the technique in such diverse areas as the choosing of constant precompensator matrices in diagonal dominance, or in some least squares parameter estimation problems.

I.1.1 Controller Structures in the Literature of Linear Systems

Begin with the state space description, and the associated optimal control synthesis technique (Kalman, 1964). The requirement of complete state feedback and the difficulties of its implementation led to the formulation of a specific structure problem, known as the optimal fixed structure problem. It is essentially a non-linear optimization problem, probably first formulated by Rekasius (1967) for which Jameson (1970) appears to give a comprehensive treatment. Extensions of this approach to controller structures have been attempted (Kosut, 1970 for example) but their subsequent development has been somewhat restricted by the nonlinearity of the formulation.

A related idea, that of predicting or bounding the value of the performance index for a sub-optimal control without resort to full scale computation or simulation has also received some attention. The motivation for this is to compare a simple but suboptimal controller against a complex but optimal controller. However none of the available techniques have been designed for the automatic assessment of the many controller structures which would exist for a large scale process. Invariably they are only suitable (and appear only to be used) for the comparison of one or two fixed structure feedback matrices on low order state space models. (Mc Clamroch, 1969; Canales, 1970; Pritchard and Parker, 1974). The work in this thesis makes a major contribution to this area. The methodology for producing increasingly accurate bounds on the optimal cost value

without direct simulation of the optimal solution is established. Further it is shown that these bounds are readily calculated from process simulation data and are easily automated to effect the structural assessment of a large scale process.

A branch of linear systems theory which has undergone several advances recently is that of decoupling theory. Loosely a multi-variable system is called decoupled if each system output can be independently controlled by a corresponding input (subset). These developments have occurred in parallel with the growth of geometric analysis pioneered by A.S. Morse and W.M. Wonham. On the face of it, it would appear that decoupling theory could give useful data on the controller structure problem. However the theory is based on precise algebraic relationships, for example identities between finite dimensional subspaces, and for the moment appears to lack any element of approximation.

With a distinguished lineage from classical control is the linear transform system description. Rosenbrock, Mayne and MacFarlane have pioneered several multivariable controller design techniques using the complex frequency domain description. For these techniques, the system structure is an a-priori assumption for which the controller is designed en-bloc in an algorithmic fashion. They are basically quantitative in nature and are therefore most usefully applied once the decomposition of the system structure into weakly interactive subsystems has been achieved. Furthermore, their use is restricted to this role

by their inability to cope successfully with large scale systems.

At this point, it is of interest to recall the concept of diagonal dominance, a prerequisite to the application of Rosenbrock's Inverse Nyquist Array technique. Very little has been written of the attainment of this quality, although there has been more interest shown recently (Hannah, 1975; Owen, 1976; Mee, 1976). In the sequel, the work of Hawkins(1972) is extended, and re-interpreted as a finite dimensional structures problem to which the techniques of this thesis may be applied.

Recalling the notion of using an index, or diagnostic to compare the control properties of one structure with those of another. The question arises as to the nature of the changes in the control properties (poles and zeros, for example) of a system if only a subset of the available inputs or outputs are utilized. More fundamental are the changes caused by a radical re-selection of the system input and output variables.

In a definitive treatise on non-minimum phase effects in multi-variable linear systems, Rosenbrock (1966) observed that for a state space system $\mathcal{S} (A,B,C)$ "the structural design problem concerns the choice of B and C ". Furthermore the possible control difficulties were shown to be directly dependent on this choice, so that once made any inherent control difficulty such as non-minimum phase is immutable by controller design.

Davison and Goldberg (1968) followed this work by devising a theory of structures designed to isolate the best measurement subset. For system $\mathcal{S}(A,B,C)$ the output matrix C was chosen to comprise of eigenvectors so that the dominant modes of the system could be eliminated by a diagonal constant feedback matrix. This measurement matrix was used with an "average steady state" estimate to assess the contribution from each state to each output. If a particular state was seen to contribute insignificantly to all the outputs, then it was deleted from set of states to be measured. The technique was applied with some success to the selection of a measurement subset for a boiler system and a nuclear reactor (ninth and eleventh order systems respectively). Although this paper is one of the very few attempts to provide qualitative answers to questions of controller structures its application is limited by the state space formulation (which for example cannot accommodate time delay systems), and the assumption that the outputs will comprise of only linear combinations of accessible states.

The earlier work of Rosenbrock was translated into the formalism of geometric analysis by Bengtsson (1973) who defined the following extensions of control, and measurement. System $\mathcal{S}(A,B,C_e)$ was said to be an output extension of $\mathcal{S}(A,B,C)$ if $\mathcal{N}(C_e) \subset \mathcal{N}(C)$. Similarly system $\mathcal{S}(A,B_e,C)$ was said to be an input extension of $\mathcal{S}(A,B,C)$ if $\mathcal{R}(B_e) \supset \mathcal{R}(B)$. Further

analysis showed that suitable extensions might be chosen to avoid any non-minimum phase difficulty of the original system. These concepts of system structure were applied by Bengtsson (1973) to the state space model of a drum boiler. Here it was shown that a particular choice of inputs, and outputs did not cause non-minimum phase behaviour and had a satisfactory left half plane zero pattern before proceeding with the controller design.

Motivated by a desire to apply the Characteristic Loci controller design technique to non-square systems A.J. MacFarlane and B. Kouvaritakis researched the problem of "squaring" such systems. Without loss of generality (Kouvaritakis, 1972) attention was focused on a reduction in the number of outputs for a system description with more outputs than inputs ($r > m$) colloquially known as "squaring down". For the state space system $\dot{S}(A,B,C)$ this was achieved using a post compensator real constant matrix L to obtain a new state-to-output matrix given by $C_{sq} = LC$. The control consequences of this reduction was that the system's original zero pattern was preserved, and that new invariant zeros were introduced. An algorithm was developed to choose the compensator L so that as many of the new zeros as possible were placed in desirable positions. This work, and other related topics are reported in Kouvaritakis (1972 ; 1974) and in Kouvaritakis and MacFarlane (1976 a,b). As remarked earlier, the choice of L which utilizes all the outputs may incur heavy installation and maintenance liabilities unnecessarily.

The afore-mentioned research of Rosenbrock through to Kouvaritakis and MacFarlane , appears to constitute the bulk of in depth research applied to the problem of choosing a suitable system structure. The important feature of this analysis is the relationship between the system structure and its associated control properties. However the basic motivation for their research is to produce a system structure with few inherent control difficulties prior to the design of a controller en bloc.

In contrast, Industry has often found it desirable to employ single input-single output process control loops. Behind such industrial practice are the expediences of simplicity and economy. Single loops are simple to design and tune, and are economical to install. Unfortunately, for the more complex processes simultaneous application of all the control loops was frequently found to interact adversely. It is in this situation that the need for systematic procedures to isolate simple control networks, or structures arises.

Bristol (1966) gave a static measure of non-interaction which is also exhibited in Rosenbrock (1974) as a measure of non-minimum phase. The essentials of the measure can also be translated to an LQP setting although as yet no scheme has been developed to use the measure.

Two differing non-interaction indices have been proposed by Davison (1969) and, Davison and Man (1970). They purport to measure the degradation caused by the simultaneous application of controllers each known to give satisfactory control over

individual (groups of) outputs. A more constructive approach of course, would be to indicate which inputs and outputs can be controlled non-interactively prior to the design of the controllers. (The reader is referred to Hannah (1975) for more comments on the use of these indices, especially with reference to the relationship between non-minimum phase behaviour and the index in Davison (1969)). Niederlinski (1971) adopts this approach by searching through all the possible paired structures of inputs and outputs, rejecting those structures which are unstable when integrating control (namely I, PI or PID) is applied and finally uses two heuristics to isolate the 'best' pair structure for control purposes. As the name "heuristics" implies the technique lacks a theoretical basis and the algorithm proposed is only suitable for transfers of low dimension ($n \leq 3$). In the realm of stochastic analysis H. Kwakernaak, and R. Sivan (1972b) have proposed a controller design philosophy utilizing the Kalman filter and the deterministic LQP problem. Invoking the separation theorem, the Kalman filter uses measureable outputs to create a state estimate, and the optimal LQP gains form the feedback law. With this structure for the system-controller configuration, a non-linear optimization is posed to reduce the dynamic order of the controller. They remark that considerable simplifications result if structural constraints are placed on the controller. These may be specific in the sense of zeroing a local feedback coefficient or more general in the sense of deleting inputs, or outputs, however no guidance on this

choice was given. In the second part of this thesis, an output structures problem for stochastic systems is considered. In its present form it appears possible to determine which outputs are necessary for the satisfactory estimation of particular states. When all the states have been considered, it should be possible to determine which outputs are necessary for the satisfactory estimation of the state vector. This output structural information, when united with the input structure information obtained using the first part of this thesis could well provide the structural guidance Kwakernaak and Sivan's design algorithm lacked.

Of related interest is the so-called measurement structure problem (Athans, 1972; Herring and Melsa, 1974; Mellefont and Sargent, 1977). Again the control problem is set in a noisy environment for which the question under consideration is when to measure (as a function of time) to maintain satisfactory control. Whether this formulation can be extended to indicate what to measure remains as yet unreported.

In a study of this nature it is desirable to include some examples of realistic industrial problems; in particular examples for which a structures problem was defined and resolved. Clearly this type of example is not found among the small problems which pervade much of the academic research literature. The two examples chosen are recent industrial applications of structures assessment.

I.1.2 Two Industrial Examples

Taken from the Steel Industry, the first example is the multivariable control of a five stand cold rolling steel mill. The main control objectives for a cold mill are to reduce deviations of strip thickness from a nominal schedule of values, and to constrain the interstand tensions to prevent looping or tearing.

Utilizing a linearised rolling model having some 33 states, 8 outputs and 9 inputs, Tanuma (1974) developed a multivariable control scheme of considerable complexity using the theory of the LQP problem. The complexity of the control techniques (implemented by computer algorithms) permitted only the broadest of conclusions to be drawn on the physical mechanism at work in the control scheme.

In a definitive study of the control of cold rolling mills, Bryant et al (1973, 1975) developed non-interactive networks for control based on an extensive and exhaustive research programme, the main thrust of which was directed at modelling the process physics. The structures obtained have their basis in the theory of the process with their simplicity being an aesthetic bonus. The important role of simulations in this study (some 6000 runs in all) serves only to indicate the scarcity of design tools for isolating the significant controller structures that might exist to provide satisfactory control.

The second example is the control of a fluidic catalytic cracker, a process from the petroleum refining industry, reported by Schuldts and Smith (1971). The process was simulated in steady state

conditions using a linearised model. Nine different controller structures were chosen for assessment, where the gains were calculated by minimizing a quadratic cost function. The effectiveness of each controller (simulated in the presence of large random disturbances) was measured using a quadratic loss index. The results of these trials isolated new controller structures showing that certain measurements could be discarded without serious degradation in control and under certain operating conditions a particular control input could be omitted.

I.1.3 Conclusion

One aspect of the so-called 'gap' between main-stream linear systems theory (as reported in the control literature) and control practice arises directly from the constraints placed on the solution of the control problem. This is especially true of industry where economy of design and implementation have to be balanced against the possible increased benefit which might accrue from a more complex design.

The reduction of a system structure to the smallest possible input/output subsets and the subsequent decomposition into a series of weakly interacting control structures is therefore seen as an alternative design philosophy to that of en bloc controller design. It is felt that such a structural decomposition will depend strongly on the underlying physical nature of the system giving the procedure a satisfying intuitive appeal.

As the survey of the literature of linear systems theory shows, the importance of the choice of system structure for control has long been recognized but there is a marked absence of systematic computational aids to decompose the structure into simple (weakly non-interactive) control networks. This is the basic structure problem in controller design.

The sequel reports the development of such an aid; a diagnostic which can be used to identify simple controller structures having good control properties. The technique is shown to have a theoretical basis in terms of pole-zero patterns and the related non-minimum phase difficulty when applied to the state space system $\dot{S}(A,B,C)$. To assess the many possible structures arising from a large scale process, a set of bounds on the

diagnostic value are used to indicate the potentially good structures. For this purpose a computer algorithm has been developed and validated. The computation of bounds is many orders faster than a full optimization and although there is some imperfection in the estimate of optimal cost, many more combinations can be assessed.

Utilizing a digital simulation of a five stand cold rolling mill (developed by the Industrial Automation Group), the diagnostic bounds have successfully isolated several of the simple controller structures proposed by the Industrial Automation Group as the basis for non-interactive control networks recently implemented on several mills in this country.

I.2 The Problem of Isolating Controller Structures

I.2.0 Introduction: Structures in the Design Process

There are two basic approaches to the design of controllers, depending very much on the nature of the process and its constraints. Given the most generous conditions it may be possible to synthesize a controller using a multivariable design package which could be followed by a search for structural simplifications (especially if the synthesis technique used is based on some costing which provides a ready made ranking). The second approach is the reverse of the above, namely simple control structures are found, and the controllers are designed using the framework of these structures. In practice, the practical constraints of realistic systems usually ensure the second approach is adopted.

The selection of controller structure is therefore deciding which input, measured output subsets to use to attain specified control objectives. The indications are that these simple structures frequently exist but are obscured by the complexity of the large scale process under consideration. The benefits to be obtained are simplicity in the subsequent controller design and economy in implementation.

To be useful, a diagnostic for the selection of external structures in large scale systems would have a sound theoretical basis for the comparison of the control properties of different structures. Following this it should be easily evaluated and readily assimilated so that even if the combinatorial aspect of structure comparison cannot be defeated it is still feasible to examine the

smaller sized structures. The diagnostics of this thesis and its associated bounds for comparison are thought to go a long way towards satisfying these stringent criteria.

I.2.1 Basic Problem Formulation

Using Hilbert space notation the linear system description is

$$y = Wu + d \quad \text{I.2.1.1}$$

where the output vector $y \in H_r$, the input vector $u \in H_m$, the disturbance vector $d \in H_r$ and the system operator W satisfies $W: H_m \rightarrow H_r$. Set in the time domain a causal weighting function description is assumed. This incorporates those processes with irrational transfer functions, although more important is the prospect of utilizing process simulation models (digital, analogue or hybrid) directly to provide the weighting function description.

To facilitate a valid comparison of the different controller structure, a measure, diagnostic or index is required. It will be important for this index to accommodate the objectives of the design, factors like actuator cost, transducer costs as well as the rather broader control aims like "tight output" control.

In the framework of these design objectives the index must also reflect the control properties of the various structures.

The variational index chosen is the quadratic cost performance index

$$J = \frac{1}{2} \langle y, Qy \rangle_{H_r} + \frac{1}{2} \langle u, Ru \rangle_{H_m} \quad \text{I.2.1.2}$$

where the output weighting Q (assumed self adjoint and semi-positive definite, $Q \geq 0$) and the input weighting R (assumed self adjoint and positive definite, $R > 0$) can be used to determine the relative merits of the individual outputs and inputs, while the balance between $\|Q\|$ and $\|R\|$ can be used to specify the overall control objectives. This ratio between

$||Q||$ and $||R||$ is extended further when it become necessary to relate the value of the optimal cost to the control properties of the structure. An asymptotic analysis is developed by setting $R = \rho R_0, \rho > 0$ and examining in control terms the effect of $\rho \rightarrow 0$.

It is convenient at this point to introduce the concepts of "input structure" and "output structure". For given control objectives the former specifies the inputs required and the latter the outputs required. In an optimal control setting, it turns out that the input structure problem is naturally handled via the deterministic LQP formulation whilst the output structure problem has so far been most usefully exploited in the stochastic environment of the LQG problem. However it is instructive to note that the output structure problem has been recast as a deterministic input structure problem with an increased computational requirement. For this reason, the work on the output structure problem remains largely academic rather than practical and the part of the thesis devoted to this problem is mainly of an expository nature only.

I.2.2 The Input Structure Diagnostics

By deleting some of the inputs of a given linear system description, one of the many different input structures is obtained as the structured system:

$$y = W_S u_S + d \quad \text{I.2.2.1}$$

where y , and $d \in H_r$, structured control input $u_S \in H_S$ ($s \in \{1, 2, \dots, m\}$) and the structured system operator W_S is obtained from W , by deletion so that $W_S: H_S \rightarrow H_m$. (This notation simply relabels the situation where certain inputs, removed from the input set are constrained to be zero.)

Assuming that the full input and output sets have been assigned suitable cost weights Q , and R , it is clear that input deletion will give the structure an associated cost functional J_S , where

$$J_S = \frac{1}{2} \langle y, Q y \rangle_{H_r} + \frac{1}{2} \langle u_S, R_S u_S \rangle_{H_S} \quad \text{I.2.2.2.}$$

The comparison of the optimal values of the costs J , and J_S forms the basis of the structures assessment technique. Since the lowest optimal cost will be achieved by the use of all inputs, the structures are examined to ascertain which structure has an optimal cost value sufficiently close to this overall optimum.

The optimal cost value is obtained as follows:

$$\text{Using } J = \frac{1}{2} \langle y, Q y \rangle + \frac{1}{2} \langle u, R u \rangle \quad (\text{I.2.1.2})$$

substituting for output y using the linear constraint I.2.1.1

$$J(u) = \frac{1}{2} \langle u, (W^* Q W + R) u \rangle + \langle u, W^* Q d \rangle + \frac{1}{2} \langle d, Q d \rangle$$

where W^* is the operator adjoint of W .

A gradient on H_m may be defined as:

$$g(u) \triangleq \frac{\partial J}{\partial u} = Au + b$$

where $A \triangleq W^* QW + R$ and $b \triangleq W^* Qd$.

Under suitable conditions (Riesz and Sz-Nagy, 1965)

A^{-1} exists, and there is a unique optimal control $\hat{u} = -A^{-1}b$ for which $g(\hat{u}) = 0$, and $J(\hat{u}) = \frac{1}{2} \langle d, Qd \rangle - \frac{1}{2} \langle b, A^{-1}b \rangle$.

The two costs for comparison are therefore the full optimal cost

$J(\hat{u}) = \frac{1}{2} \langle d, Qd \rangle - \frac{1}{2} \langle b, A^{-1}b \rangle$ and the structured optimal cost $J_S(\hat{u}_S) = \frac{1}{2} \langle d, Qd \rangle - \frac{1}{2} \langle b_S, A_S^{-1} b_S \rangle$

These cost values motivate the introduction of two input

structure diagnostics ΔJ_0 , and IESQ. The first of these is based

on the improvement in cost $\Delta J(u) \triangleq J(u) - J(\hat{u})$ for any $u \in H_m$.

Clearly this is a measure across constant cost contours of the

available cost improvement from cost $J(u)$ at u , to the optimal

cost $J(\hat{u})$ at $\hat{u}$. Of particular interest is the improvement in

cost available in applying optimal control compared to applying

no control at all, viz $\Delta J_0 = J(0) - J(\hat{u}) = \frac{1}{2} \langle b, A^{-1}b \rangle$

or in structured terms $\Delta J_0^S = \frac{1}{2} \langle b_S, A_S^{-1} b_S \rangle$

The second of these is the integral error squared diagnostic

$IESQ \triangleq \frac{1}{2} \langle y(\hat{u}), Qy(\hat{u}) \rangle$ (its name coming from the

related quantity in the optimal servo problem:

$\frac{1}{2} \langle e(\hat{u}), Qe(\hat{u}) \rangle$ where $e(\hat{u}) = r - y(\hat{u})$ which measures

the success of output regulation through applying optimal control.

In structured terms this is written $IESQ_S = \frac{1}{2} \langle y(\hat{u}_S), \hat{Q}y(u_S) \rangle_{H_T}$.

Several important issues must be resolved the most basic being whether the proposed diagnostics can be related to the control characteristics of the structures. That this is the case is established using new results on the asymptotic behaviour of the pole-zero patterns of optimal control systems. These results are rederived using a linear space analysis to give a completely fresh "geometric" or "least squares" interpretation of the proposed structure assessment technique (Chapters 1 and 2).

It is computationally inefficient to evaluate the diagnostic of each different structure by solving the related optimization problem. Consequently following the previous researches of G.F. Bryant (1973) additional information contained in a set of moment equations was utilized to obtain a set of diagnostic bounds. Here an infinite dimensional linear programming theorem was developed to reduce an infinite dimensional optimization problem to one of finite dimension.

A set of four finite dimensional problems were solved to obtain a set of bounds with the possibility of using numerical techniques to obtain a further set. These bounds are thought to be original to this work.

Various properties of the bound formulae, and the moment sequences are developed and the computation for a simple finite time-domain example given. With this important computational problem surmounted, structure assessment is now a feasible proposition.

A new algorithm was developed to decompose the moment quantities, so that the moment sequences for the different structures may be readily reconstituted. Also required in the algorithm to perform structure assessment was an (low) upper bound for the maximum eigenvalue of the operator $Q^{\frac{1}{2}} W^* R^{-1} W Q^{\frac{1}{2}}$ and its structured relation $Q^{\frac{1}{2}} W_S^* R_S^{-1} W_S Q^{\frac{1}{2}}$. The results of Bryant (1973) are recast in a Hilbert space setting, and several new extensions given. The problems are discussed in Chapter 4 where a set of bounds of varied accuracy and complexity are seen to be available. The validation of the algorithm, and a successful trial of the structures assessment package (written in FORTRAN IV) on the model of a five-stand cold rolling mill developed by the Industrial Automation Group is detailed in Chapter 5. Various extensions and consequences of the technique are explored in Chapter 6. Perhaps the most significant of these are a completely new re-appraisal of pseudo-diagonalisation (Hawkins, 1972) and an application of the structures procedure to a least squares problem (Draper and Smith, 1966). The notion of applying structure procedures to the constant pre-compensators of the pseudo-diagonalisation technique is new, whilst the problem of least square parameter structures is well-known. The output structures problem is briefly discussed in Chapter 7 and 8. In Chapter 7 is a discussion of the implication of output deletion in the quadratic (primal) cost index where it is shown that this deletion procedure does not appear to yield useful information on the output structures problem. Whilst in Chapter 8, the output problem is usefully set in a stochastic environment.

Finally the thesis concludes with a brief set of topics and recommendations for further research.

Chapter 1 Asymptotic Linear Systems Theory for Optimal Controllers

1.0 Introduction

1.1 The Problem Basics and the Optimal Return Difference Relation

1.2 Generalization of a Result of Kalman

1.3 The Infinite Zeros and their Butterworth Patterns

1.4 Summary

1.0 Introduction

A considerable amount of work has been expended on detailing the asymptotic behaviour of the closed loop poles of an optimal linear system. (Kalman, 1964; Tyler and Tuteur, 1966; Kwakernaak and Sivan, 1971; and Kwakernaak, 1976.) Although the results for the finite zeros and the first order infinite zeros are available, the details for the higher order infinite zeros remain unknown. In this chapter Kalman's result for the single-input/multi output system is successfully given a multivariable treatment and several results relating to all orders of infinite zeros given.

1.1 The Problem Basics and the Optimal Return Difference Relation

The performance index considered is the integral of a quadratic cost subject to a linear system constraint described by the time invariant system triple $\dot{S}(A,B,C)$, namely

$$J = \frac{1}{2} \int_0^{\infty} \{ \langle y(t), Qy(t) \rangle + \langle u(t), Ru(t) \rangle_{E_m} \} dt \quad 1.1.1$$

with $\dot{S}(A,B,C) \triangleq \dot{x}(t) = A x(t) + B u(t) \quad x(0) = x_0.$

$$y(t) = C x(t) \quad 1.1.2$$

Two additional assumptions are made:

- i) Symmetric, constant matrices R and Q are positive and positive semi-definite respectively. ($R > 0$, $Q \geq 0$.)
- ii) The matrix Q is capable of a factorization $Q = D^T D$ so that the system triple $\dot{S}(A,B,DC)$ is completely controllable and observable. These assumptions ensure the optimal closed loop system is asymptotically stable. (Anderson and Moore, 1971.) They also ensure no pole-zero cancellations occur in forming the transfer matrix $G(s) = Q^{\frac{1}{2}} C (sI - A)^{-1} B R^{-\frac{1}{2}}.$

The optimal return difference determinantal relation is obtained succinctly using the properties of the matrix $\bar{A} = \begin{bmatrix} A & -BR^{-1}B^T \\ -C^TQC & -A^T \end{bmatrix}$

These include:

- i) Matrix $\bar{A}$ has no purely imaginary eigenvalues. If λ is an eigenvalue, then $(-\lambda)$ is also one.

- ii) For $\bar{A}$ with distinct eigenvalues, and associated eigenvector

$$\text{matrix } V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}, \quad \text{then } V^{-1} = \begin{bmatrix} V_{22}^T & -V_{12}^T \\ -V_{21}^T & V_{11}^T \end{bmatrix}$$

iii) The symmetric positive definite matrix $P_\infty = \begin{bmatrix} V_{12}^T \\ V_{22} \end{bmatrix}^{-1} V_{22}^T$ is a solution of the steady-state matrix Riccati equation with cost, equation 1.1.1 and system described by equation 1.1.2.

iv) The optimal control law is $\hat{u}(t) = -R^{-1}B^T P_\infty x(t)$ so that the closed loop system matrix is $A_{CL} = A - BR^{-1}B^T P_\infty$ and the optimal closed loop poles are $(-\lambda)$, $R(\lambda > 0)$.

Introduce the closed loop characteristic polynomial as

$$\det [sI - A_{CL}] = \det [sI - A] = \Delta_c(s)$$

$$(\text{using } A_{CL} = A - BR^{-1}B^T P_\infty = V_{12} \begin{bmatrix} -\lambda \end{bmatrix} V_{12}^T)$$

and the open loop characteristic polynomial as

$$\det [sI - A] \triangleq \Delta_o(s).$$

Matrix $\bar{A}$ has the canonical form

$$\bar{A} = \begin{bmatrix} A & -BR^{-1}B^T \\ -C^TQC & -A^T \end{bmatrix} = V \begin{bmatrix} \lambda & 0 \\ 0 & -\lambda \end{bmatrix} V^{-1}$$

$$\text{Hence } \det [sI - \bar{A}] = \det [sI - A] \det [sI + A]$$

$$\text{giving } (-1)^n \Delta_c(s) \Delta_c(-s) = \det \begin{bmatrix} sI - A & BR^{-1}B^T \\ C^TQC & sI + A^T \end{bmatrix}$$

Using Schur's Identity obtain:

$$\begin{aligned} (-1)^n \Delta_c(s) \Delta_c(-s) &= \det [sI - A] \det [sI + A^T - C^TQC(sI - A)^{-1}BR^{-1}B^T] \\ &= \det [sI - A] \det \left[-(-sI - A^T) \begin{Bmatrix} I - \\ (sI + A^T)^{-1}C^TQC(sI - A)^{-1}BR^{-1}B^T \end{Bmatrix} \right] \\ &= (-1)^n \det [sI - A] \det [-sI - A^T] \det \left[I + B^T(-sI - A^T)^{-1}C^TQC(sI - A)^{-1}BR^{-1} \right] \end{aligned}$$

Recalling $\Delta_o(s) = \det [sI - A]$ and defining $\tilde{G}(s) = C(sI - A)^{-1}B$

gives $\Delta_c(s) \Delta_c(-s) = \Delta_o(s) \Delta_o(-s) \det [I_m + \tilde{G}^T(-s) Q \tilde{G}(s) R^{-1}]$

A modification of notation is made: it is assumed that the input weighting matrix R is written $R = \rho R_o$, where scalar ρ satisfies $\rho > 0$, and $R_o > 0$ enables R_o to be factorized as $R_o^{\frac{1}{2}} R_o^{\frac{1}{2}}$, similarly the output weighting matrix is factorized as $Q = D^T D$ so that writing $G(s) = \tilde{D} \tilde{G}(s) R_o^{-\frac{1}{2}}$, the optimal return difference determinantal relation is obtained as:

$$\Delta_c(s) \Delta_c(-s) = \Delta_o(s) \Delta_o(-s) \det [I_m + \mu G^T(-s) G(s)] \quad 1.1.3$$

where $\mu = 1/\rho$ and $\rho > 0$.

The optimal closed loop pole positions have been specified through the canonical form of the matrix $\bar{A}$,

$$\bar{A} = \begin{bmatrix} A & -\mu B R_o^{-1} B^T \\ -C^T Q C & -A^T \end{bmatrix} = V \begin{bmatrix} \Lambda(\mu) & 0 \\ 0 & -\Lambda(\mu) \end{bmatrix} V^{-1}$$

and through the zeros of the optimal return difference determinant

$$\begin{aligned} \Delta_c(s) \Delta_c(-s) &= \prod_{i=1}^n (s + \lambda_i(\mu))(-s + \lambda_i(\mu)) \quad R(\lambda_i > 0) \\ &= \Delta_o(s) \Delta_o(-s) \det [I_m + \mu G^T(-s) G(s)] \end{aligned}$$

The asymptotic behaviour of the poles of the optimal closed loop system arises as the scalar ρ , which weights the input cost, tends to zero giving $\mu = 1/\rho \rightarrow \infty$. More formally if $\{\mu_j\}_{j=1}^{\infty}$ is an unbounded sequence of positive real numbers, with $\lambda(\mu_j)$ and $v(\mu_j)$ as the corresponding sequence of eigenvalue/vector pairs satisfying

$$\bar{A}(\mu_j) v(\mu_j) = -\lambda(\mu_j) v(\mu_j) \quad \Re(\lambda) > 0$$

then a suitable subsequence may be extracted so that

$$\lim_{j \rightarrow \infty} v(\mu_j) = v_\infty \text{ and } ||v||_\infty = k > 0.$$

The related eigenvalue sequence has either a finite limit or is unbounded. The equivalence between the canonical form of $\bar{A}$, and the relation 1.1.3 enables the asymptotic properties of the optimal closed loop poles to be investigated as asymptotic properties of the zeros of the latter, that is through the asymptotic roots of

$$\prod_{i=1}^n (s + \lambda_i(\mu_j))(-s + \lambda_i(\mu_j)) = \Delta_o(s)\Delta_o(-s)\det [I_m + \mu G^T(-s)G(s)] = 0.$$

Anticipating the results, some of the closed loop poles attain finite positions in the open left half plane related to the system zeros whilst the remaining closed loop poles tend to infinity in Butterworth patterns of varying orders and radii. These latter poles are said to be attracted by infinite zeros whose order is that of the Butterworth pattern.

1.2 Generalization of a Result of Kalman's

In an early paper on the LQP problem, Kalman detailed the asymptotic behaviour of the optimal closed loop poles for a single input/multi output system using a transfer function analysis. Here this approach is extended to specify the behaviour of those closed loop poles attaining finite positions for multi-variable systems.

The principal ingredient of this extension is the relationship between the Smith-McMillan form of $G(s)$ and $G^T(-s)G(s)$. It is shown in Appendix A, that if $G(s)$ has the Smith-McMillan form $P(s)\Gamma(s)Q(s)$ where both $P(s)$ and $Q(s)$ are unimodular matrices of appropriate dimension, then $G^T(-s)G(s) = T_1(s) \Gamma^T(-s)\Gamma(s)T_2(s)$ where $T_1(s)$ and $T_2(s)$ are also appropriate unimodular matrices. The matrix $\Gamma(s)$ has a leading diagonal whose rational elements in product yield the transmission zeros and the open loop system poles (all here assumed controllable and observable).

In the optimal return difference determinantal relation

$$\begin{aligned} \det [I_m + \mu G^T(-s)G(s)] &= \det [I_m + \mu T_1(s)\Gamma^T(-s)\Gamma(s)T_2(s)] \\ &= \det [I_m + \mu \Gamma^T(-s)\Gamma(s)T_{21}(s)] \end{aligned}$$

where $T_{21}(s) = T_2(s)T_1(s)$ is $m \times m$ unimodular.

Hence the closed loop poles for the sequence $\{\mu_j\}_{j=1}^{\infty}$ are given

$$\text{by the roots of } \Delta_0(s)\Delta_0(-s)\det [I_m + \mu_j \Gamma^T(-s)\Gamma(s)T_{21}(s)] = 0.$$

Thus for all $\mu_j > 0$, have that

$$\Delta_0(s)\Delta_0(-s)\det [I_m/\mu_j + \Gamma^T(-s)\Gamma(s)T_{21}(s)] = 0.$$

so that as $\mu_j \rightarrow \infty$, $\det [I_m/\mu_j + \Gamma^T(-s)\Gamma(s)T_{21}(s)] \rightarrow \det [\Gamma^T(-s)\Gamma(s)T_{21}(s)]$

$$\text{giving } \frac{\Delta_0(s) \Delta_0(-s) \prod_{i=1}^{n_Z} (s-Z_i)(-s-Z_i) \det[T_{21}(s)]}{\Delta_0(s) \Delta_0(-s)} = 0$$

Hence n_Z asymptotic roots or equivalently optimal closed loop poles limit at positions associated with the system transmission zeros. Noting that the optimal closed loop is always asymptotically stable, those closed loop poles limiting at finite positions do so at either the left half plane zero positions or at the left half plane mirror image position of right half plane zeros.

The behaviour of the remaining closed loop poles is indicated by the analysis following, however a more detailed exposition is given in the next section. Consider here $s = jw$; hence matrix $G^T(-jw) G(jw)$ is real symmetric and positive semi-definite (real hermitian). In the optimal return difference determinantal relation $\Delta_c(jw) \Delta_c(-jw) = |\Delta_c(jw)|^2 = |\Delta_0(jw)|^2 \det[I_m + \mu G^T(-jw) G(jw)]$

Using Minkowski's Inequality

$$\{\det[I_m + \mu G^T(-jw) G(jw)]\}^{1/m} \geq \{\det[I_m]\}^{1/m} + \{\det[\mu G^T(-jw) G(jw)]\}^{1/m}$$

$$\text{hence } \det[I_m + \mu G^T(-jw) G(jw)] \geq \mu^m \det[G^T(-jw) G(jw)]$$

$$\text{to give } |\Delta_c(jw)|^2 \geq |\Delta_0(jw)|^2 \mu^m \det[G^T(-jw) G(jw)]$$

Using the theorem of Appendix A, namely that

$$G^T(-s) G(s) = T_1(s) \Gamma^T(-s) \Gamma(s) T_2(s)$$

$$\text{obtain } |\Delta_c(jw)|^2 \geq |\Delta_0(jw)|^2 \mu^m \det[\Gamma^T(-jw) \Gamma(jw) T_{21}(jw)]$$

$$\text{and subsequently } |\Delta_c(jw)|^2 \geq |\Delta_0(jw)|^2 \mu^m \frac{|\Delta_Z(jw)|^2 k_{21}}{|\Delta_0(jw)|^2}$$

$$\text{where } \Delta_Z(s) \triangleq \prod_{i=1}^{n_Z} (s-Z_i) \quad (\text{a real coefficient polynomial})$$

and $k_{21} = \det[T_{21}(s)]$

It has already been shown, that as $\mu \rightarrow \infty$, some of the closed loop poles tend to finite positions associated with the system zeros.

Partition $|\Delta_c(j\omega)|^2$ accordingly so that

$$|\Delta_c(j\omega)|^2 = |\Delta_z(j\omega, \mu)|^2 |\Delta_R(j\omega, \mu)|^2$$

where $\Delta_z(j\omega, \mu)$ is that part of $\Delta_c(j\omega)$ associated with the closed loop poles limiting at finite positions (related to the system zeros)

giving $\frac{|\Delta_z(j\omega)|^2}{|\Delta_z(j\omega, \mu)|^2} \rightarrow 1$ as $\mu \rightarrow \infty$; and $\Delta_R(j\omega, \mu)$ is the

$$|\Delta_z(j\omega, \mu)|^2$$

remaining part of $\Delta_c(j\omega)$.

$$\text{Then } |\Delta_R(j\omega, \mu)|^2 \geq k_{21} \mu^m \frac{|\Delta_z(j\omega)|^2}{|\Delta_z(j\omega, \mu)|^2}$$

Clearly as $\mu \rightarrow \infty$, the excess of closed loop poles over the number of system zeros tend to infinity, subject to the constraint of asymptotic stability which ensures that these loci are in the left half plane. In the following section a more comprehensive analysis is given to describe this asymptotic closed loop poles behaviour.

1.3 The Infinite Zeros and their Associated Butterworth Patterns

Begin by recalling the concept of the Butterworth Pattern which arose in the early investigations of network theory (Butterworth, 1930). They are simply the roots of the polynomial equation

$$s^{2n} = (-1)^{n+1} \alpha \text{ where } \alpha > 0 \text{ and } n = 1, 2, \dots$$

For example:

$$n = 1: \text{ 1st Order Pattern } s = \pm \alpha^{\frac{1}{2}}$$

$$n = 2: \text{ 2nd Order Pattern}$$

$$s_k = \alpha^{\frac{1}{4}} (\cos + i \sin) ((2k + 1) \pi / 4) \quad k = 0, 1, 2, 3$$

$$n = 3: \text{ 3rd Order Pattern}$$

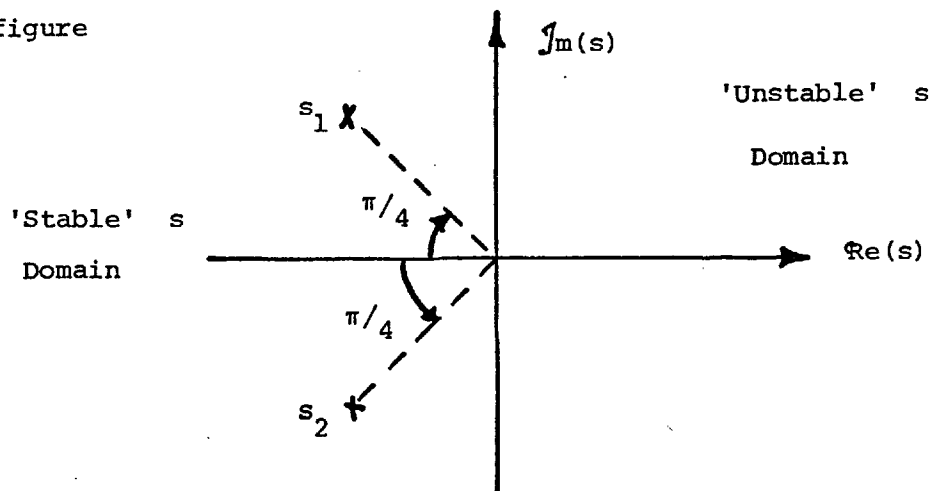
$$s_k = \alpha^{\frac{1}{6}} (\cos + i \sin) (k\pi / 6) \quad k = 0, 1, \dots, 5$$

Each pattern is seen to contain $2n$ roots. Those roots in the left half plane will play an important role when interpreted as closed loop system poles. For this purpose, introduce an n th order stable Butterworth polynomial which is formed from the left half plane roots. As an example consider $n = 2$, then the 2nd order stable Butterworth polynomial is given

$$f_2(s) = \left(s - \frac{\alpha^{\frac{1}{4}}(-1+i)}{\sqrt{2}} \right) \left(s - \frac{\alpha^{\frac{1}{4}}(-1-i)}{\sqrt{2}} \right)$$

$$f_2(s) = s^2 + \sqrt{2}\alpha^{\frac{1}{4}}s + \alpha^{\frac{1}{2}}$$

The 2nd order stable Butterworth pattern is given in the following figure



The unit step responses of a single input/single output system having no open loop zeros with a transfer function denominator of varying orders of Butterworth polynomials have been depicted in Kwakernaak and Sivan, 1972b. Their percentage overshoot increases from 0% for $n = 1$ to a limit of about 23% as $n \rightarrow \infty$ (Kalman, 1964).

The linear system is assumed to be defined through the matrix triple $\mathcal{S}(A, B, C)$ where the cost function weightings Q and R_0 have been absorbed into the matrices C and B respectively in the manner of the previous section. The basic assumption on the system dimensions is that $n \gg \max(r, m)$ where there are n states, r outputs and m inputs. The input/output dimensions r , and m give rise to three possible cases

i) $r > m$, ii) $r = m$ and iii) $r < m$ for which three analogous treatments are possible. Here the details of the case i) are given, where $m < r \ll n$. Kou varitakis and

MacFarlane (1976b) have shown that the maximum number of zeros defined through the system matrix $P(s) = \begin{bmatrix} sI - A & B \\ -C & 0 \end{bmatrix}$

is given by $n - \max(r, m)$. In the case above this will be $(n - r)$ consequently the behaviour of at least r infinite closed loop poles is sought. For this reason the optimal return difference determinantal relation is used in the form

$$\Delta_c(s)\Delta_c(-s) = \Delta_o(s)\Delta_o(-s) \det [I_r + \mu G(s)G^T(-s)]$$

Let $\{\eta_i\}_{i=1}^n$ be the n open loop pole positions, and

define $\bar{\eta} = \max_i \{|\eta_i|\}$ The behaviour of the zeros of

$\Delta_c(s)\Delta_c(-s)$ is to be examined for $|s| \gg \bar{\eta}$, and asymptotically

as $\mu \rightarrow \infty$. The open loop characteristic polynomial $\Delta_o(s)$ is defined $\Delta_o(s) = \prod_{i=1}^n (s - \eta_i)$, therefore in the region $|s| \gg \bar{\eta}$ $\Delta_o(s) \neq 0$.

Hence, recalling $\Delta_c(s)\Delta_c(-s) = \Delta_o(s)\Delta_o(-s)\det [I_r + \mu G(s)G^T(s)]$

the zeros of $\Delta_c(s)\Delta_c(-s)$ in this region will be given by the roots of $\det [I_r + \mu G(s)G^T(-s)] = 0$. Thus the problem of the asymptotic behaviour of the optimal closed loop poles reduces to a study of the roots of $\det [I_r + \mu G(s)G^T(-s)] = 0$ under the conditions $|s| \gg \bar{\eta}$ and $\mu \rightarrow \infty$.

Under such conditions the system resolvent $(sI - A)^{-1}$ is capable of a series expansion:

$$(sI - A)^{-1} = 1/s (I - 1/s A)^{-1} = 1/s (I_n + A 1/s + A^2 1/s^2 + \dots)$$

$$\text{to give } G(s) = C(sI - A)^{-1}B = 1/s (M_0 + M_1 1/s + M_2 1/s^2 + \dots)$$

where the k th Markov parameter is defined $M_k = CA^k B, k=0,1,\dots$

Hence the asymptotic closed loop pole behaviour is given by the

$$\text{roots of } \det [I_r - \mu/s^2 \{M_0 M_0^T + (M_1 M_0^T - M_0 M_1^T) 1/s + \dots\}] = 0 \quad 1.3.1$$

where $|s| \gg \bar{\eta}$ and $\mu \rightarrow \infty$. Since the relationship between rank, and the number of non-zero eigenvalues for real definite hermitian matrices plays an important role in the sequel a minor

digression is made to give several relevant results.

Theorem (Lancaster, 1969 Pg 76) Every hermitian matrix is simple.

Theorem (Lancaster, 1969, Pg 94) The $n \times n$ matrix A is a definite matrix of rank $r \leq n$ iff A is real, symmetric and has r positive eigenvalues and $(n-r)$ eigenvalues equal to zero.

Theorem (Noble, 1969, Pg 139) If A is a real matrix then A and $A^T A$ have the same rank.

The behaviour of the asymptotic roots of 1.3.1 is given by an inductive type of analysis. Based on the optimal return difference determinant introduce the j th order reduced optimal return difference relation

$$RD_j(s, \mu) \triangleq \det \left[I_{q_{j-1}} + (-1)^j (\mu/s^{2j}) \{ A_{j-1} A_{j-1}^T + (A_j A_{j-1}^T - A_{j-1} A_j^T) 1/s + (A_{j+1} A_{j-1}^T - A_j A_j^T + A_{j-1} A_{j+1}^T) 1/s^2 + \dots \} \right]$$

whose relationship with the return difference expression (1.3.1) should become clear by setting $j=1$. The subscript q_{j-1} defines the dimensions of the (square) matrix terms in the series.

The rank of the leading term $A_{j-1} A_{j-1}^T$ is defined as d_{j-1} , which by the preliminary theorems is also the rank of A_{j-1} . These relations are used as the basis of an inductive type argument to detail the asymptotic roots of 1.3.1.

If $|s| \rightarrow \infty$ and $\mu \rightarrow \infty$ such that $\mu = o(|s|^k)$ then the asymptotic zeros of the j th order reduced optimal return difference determinant

$$RD_j(s, \mu) = \det \left[I_{q_{j-1}} + (-1)^j \mu/s^{2j} \{ A_{j-1} A_{j-1}^T + (A_j A_{j-1}^T - A_{j-1} A_j^T) 1/s + (A_{j+1} A_{j-1}^T - A_j A_j^T + A_{j-1} A_{j+1}^T) 1/s^2 + \dots \} \right]$$

$$j = 1, 2, \dots, \quad 1.3.2$$

are i) non existent for the rate $\mu = o(|s|^k)$ $k < 2j+1$

but ii) exist for the rate $\mu = o(|s|^k)$ $k = 2j+1$

and for all rates $\mu = o(|s|^k)$ $k > 2j+1$ are given by the

(j+1)th reduced optimal return difference expression $RD_{j+1}(s, \mu)$ whose terms are defined from those of $RD_j(s, \mu)$.

In particular for the rate $k = 2j+1$, if d_{j-1} is the rank of A_{j-1} , then the asymptotic zeros occupy d_{j-1} sets of jth order Butterworth patterns whose radii are related to the positive eigenvalues of $A_{j-1} A_{j-1}^T$.

Proof i) For the rates $\mu = o(|s|^k)$ $k < 2j+1$

Thus for all rates $\mu = o(|s|^k)$ $k < 2j+1$ have that

$\mu / |s|^{2j} \rightarrow 0$ giving $RD_j(s, \mu) = 1$ and no asymptotic zeros exist.

ii) For the rate $\mu = o(|s|^{2j+1})$

clearly $\mu / |s|^{2j} \rightarrow \text{constant}$, and $\mu / |s|^k \rightarrow 0$ $k \geq 2j+1$

so that the zeros of

$$RD_j(s, \mu) = \det \left[I_{q_{j-1}} + (-1)^j A_{j-1} A_{j-1}^T \mu / s^{2j} \right] \quad \text{are required.} \quad 1.3.3$$

Using the spectral properties of real hermitian matrices have

$$\begin{bmatrix} T_{j-1}^T(1) \\ T_{j-1}^T(2) \end{bmatrix} [A_{j-1} A_{j-1}^T] \begin{bmatrix} T_{j-1}(1) & T_{j-1}(2) \end{bmatrix} = \begin{bmatrix} \lambda_{d_{j-1}} & \\ & 0_{q_j} \end{bmatrix}$$

where $\text{rank}[A_{j-1} A_{j-1}^T] = \text{rank}[A_{j-1}] \triangleq d_{j-1}$

$$q_j \triangleq q_{j-1} - d_{j-1} \quad \text{and} \quad T_{j-1}^T(2) A_{j-1} A_{j-1}^T T_{j-1}(2) = 0_{q_j}$$

gives $A_{j-1}^T T_{j-1}(2) = 0$

1.3.4

Substituting into 1.3.3 gives on equating to zero

$$RD_j(s, \mu) = \prod_{i=1}^{d_{j-1}} (1 + (-1)^j \lambda(i) \mu / s^{2j}) = 0$$

hence $s^{2j} = (-1)^{j+1} \lambda(i) \mu \quad i = 1, 2, \dots, d_{j-1} \quad 1.3.5$

where $\{\lambda(i)\}_1^{d_{j-1}}$ are the positive eigenvalues of $A_{j-1} A_{j-1}^T$.

This proves the second part of the assertion since the above 1.3.5

clearly define d_{j-1} different sets of j th order Butterworth Pattern.

For any rate $\mu = o(|s|^k)$, $k > 2j+1$ $\mu/|s|^{2j} \rightarrow \infty$ so that the first term of $RD_j(s, \mu)$ (equation 1.3.2) becomes unbounded and obscures the zero behaviour of the remaining terms. This unbounded term associated with rate $\mu = o(|s|^{2j+1})$ is factored out using Schur's Identity to give the $(j+1)$ th reduced expression. The first two properties i) and ii) may then be re-applied to isolate more asymptotic zero behaviour. Use the orthogonal matrix

$[T_{j-1}(1) \ T_{j-1}(2)]$ to produce

$$\begin{bmatrix} T_{j-1}^T(1) \\ T_{j-1}^T(2) \end{bmatrix} RD_j(s, \mu) \begin{bmatrix} T_{j-1}(1) & T_{j-1}(2) \end{bmatrix} = \det \begin{bmatrix} B_{11}(s) & B_{12}(s) \\ B_{12}^T(-s) & B_{22}(s) \end{bmatrix}$$

where the relation 1.3.4 has been used to obtain simplifications

to give $B_{11}(s) = I_{d_{j-1}} + (-1)^{j+1} \mu/s^{2j} \{ \Lambda_{d_{j-1}} + B_{11}(1)/s + B_{11}(2)/s^2 + \dots \}$

with $B_{11}(1) = T_{j-1}^T(1) (A_j A_{j-1}^T - A_{j-1} A_j^T) T_{j-1}(1)$

$B_{11}(2) = T_{j-1}^T(1) (A_{j+1} A_{j-1}^T - A_j A_j^T + A_{j-1} A_{j+1}^T) T_{j-1}(1)$

and $B_{12}(s) = (-1)^{j+1} \mu/s^{2j+1} \{ B_{12}(0) + B_{12}(1)/s + B_{12}(2)/s^2 + \dots \}$

with $B_{12}(0) = T_{j-1}^T(1) A_{j-1} A_j^T T_{j-1}(2)$

$B_{12}(1) = T_{j-1}^T(1) (A_j A_j^T - A_{j-1} A_{j+1}^T) T_{j-1}(2)$

$B_{12}(2) = T_{j-1}^T(1) (A_{j+1} A_j^T - A_j A_{j+1}^T + A_{j-1} A_{j+1}^T) T_{j-1}(2)$

$$\text{and } B_{22}(s) = I_{q_j} + (-1)^{j+1} \mu / s^{2j+2} \{ B_{22}(0) + B_{22}(1) \frac{1}{s} + B_{22}(2) \frac{1}{s^2} + \dots \}$$

$$\text{with } B_{22}(0) = T_{j-1}^T(2) A_j A_j^T T_{j-1}(2)$$

$$B_{22}(1) = T_{j-1}^T(2) (A_{j+1} A_j^T - A_j A_{j+1}^T) T_{j-1}(2)$$

$$B_{22}(2) = T_{j-1}^T(2) (A_{j+1} A_j^T - A_{j+1} A_{j+1}^T + A_j A_{j+2}^T) T_{j-1}(2)$$

$$B_{22}(3) = T_{j-1}^T(2) (A_{j+3} A_j^T - A_{j+2} A_{j+1}^T + A_{j+1} A_{j+2}^T - A_j A_{j+3}^T) T_{j-1}(2)$$

Consequently using Schur's Identity to factor out $|B_{11}(s)|$

$$RD_j(s, \mu) = |B_{11}(s)| |B_{22}(s) - B_{12}^T(-s) B_{11}^{-1}(s) B_{12}(s)|$$

$$\text{and define } RD_{j+1}(s, \mu) \triangleq \det [B_{22}(s) - B_{12}^T(-s) B_{11}^{-1}(s) B_{12}(s)]$$

$$\text{Define } \hat{B}_{11}(s) \triangleq B_{11}^{-1}(s) = \frac{(-1)^j s^{2j}}{\mu} \left[\hat{B}_{11}(0) + \hat{B}_{11}(1) \frac{1}{s} + \dots \right]$$

where the details of this inversion are in Appendix A.

$$\text{Thus } B_{12}^T(-s) \hat{B}_{11}(s) B_{12}(s) = \frac{(-1)^{j+1} \mu}{s^{2j+2}} \{ B_{12}^T(0) \hat{B}_{11}(0) B_{12}(0)$$

$$+ (B_{12}^T(0) \hat{B}_{11}(1) B_{12}(0) + B_{12}^T(0) \hat{B}_{11}(0) B_{12}(1) - B_{12}^T(1) \hat{B}_{11}(0) B_{12}(0)) \frac{1}{s}$$

$$+ (B_{12}^T(0) \hat{B}_{11}(2) B_{12}(0) + B_{12}^T(0) \hat{B}_{11}(1) B_{12}(1) + B_{12}^T(0) \hat{B}_{11}(0) B_{12}(2)$$

$$- B_{12}^T(1) \hat{B}_{11}(1) B_{12}(0) - B_{12}^T(1) \hat{B}_{11}(0) B_{12}(1) + B_{12}^T(2) B_{11}(0) B_{12}(0)) \frac{1}{s^2} + \dots \}$$

Substituting this term into $RD_{j+1}(s, \mu)$ and utilizing the equivalences

for the terms of $B_{12}(s)$, $B_{22}(s)$ and $\hat{B}_{11}(s)$ it is possible after

much tedious manipulation (given in Appendix A) to produce the

correct form for the first two terms of the $(j+1)$ th reduced return

difference determinant formula as:

$$RD_{j+1}(s, \mu) = \det \left[I_{q_j} + \frac{(-1)^{j+1} \mu}{s^{2(j+1)}} \left\{ \tilde{A}_j \tilde{A}_j^T + (\tilde{A}_{j+1} \tilde{A}_j^T - \tilde{A}_j \tilde{A}_{j+1}^T)^{1/s} \right. \right. \\ \left. \left. + \dots \right\} \right] \quad 1.3.6.$$

$$\text{where } \tilde{A}_j = T_{j-1}^T (2) A_j P_{j-1} \quad P_{j-1} = I - A_{j-1}^T T_{j-1} (1) \Lambda_{d_{j-1}}^{-1} T_{j-1}^T A_{j-1}$$

$$\text{and } \tilde{A}_{j+1} = T_{j-1}^T (2) (A_{j+1} - A_j A_{j-1}^T T_{j-1} (1) \Lambda_{d_{j-1}}^{-1} T_{j-1}^T (1) A_j) P_0$$

This new reduced return difference relation may then be used to examine the asymptotic behaviour of the zeros for the rates

$$\mu = o(|s|^k) \quad k > 2j+1.$$

Beginning from $j=1$, where $RD_1(s, \mu)$ can be identified with the optimal return difference expression it can be seen that the above analysis forms a recursive procedure for describing the behaviour of all the asymptotic zeros of expression 1.3.1. Using the constraint of asymptotic stability for the optimal closed loop, the asymptotic behaviour of the closed loop poles can then be identified with a stable Butterworth polynomial of asymptotic zeros. Such a recursive procedure would terminate at that order for which $\text{rank}(A_j) = q_j$. The above analysis would also form the basis as an inductive proof for a theorem if all the terms of the series for $RD_{j+1}(s, \mu)$ could be identified. An examination of Appendix A will indicate that this does not appear feasible. In the absence of such a theorem the following corollary must suffice.

Corollary 1.3.1 Setting $j=1$ in the above recursive procedure

obtain for the case where $r = m$:

$$i) \quad d_0 \triangleq \text{rank}(CB \triangleq M_0)$$

If $d_0 = r$, then there are r 1st order Butterworth Patterns and exactly $(n - r)$ finite zeros.

ii) If $d_0 < r$, define $q_1 \triangleq r - d_0$

$d_1 \triangleq \text{rank} (T_{O2}^T \text{CAB } P_O)$ where $P_O = I_m - M_O^T T_{O1}^{-1} \Lambda_{d_0}^{-1} T_{O1}^T M_O$

$$\text{and } \begin{bmatrix} T_{O1}^T \\ T_{O2}^T \end{bmatrix} M_O M_O^T \begin{bmatrix} T_{O1} & T_{O2} \end{bmatrix} = \begin{bmatrix} \Lambda_{d_0} & \\ & O_{q_1} \end{bmatrix}$$

If $d_1 = q_1$ then there are d_0 1st Order and d_1 2nd Order stable Butterworth Patterns giving exactly $n - (d_0 + 2d_1)$ finite zeros.

iii) If $d_1 < q_1$ then clearly a further step may be taken.

The numbers of infinite zeros, and zeros for these cases can be shown to be equivalent to results presented in Kou varitakis and Shaked, 1976 where the asymptotic properties of a scalar feedback of the form kI , $k \rightarrow \infty$ were investigated. (Appendix A contains results demonstrating the correspondences for the corollary above.)

In toto, a recursive procedure has been developed for describing the behaviour of those closed loop poles tending to infinity. Explicit results have been obtained for the 1st and 2nd order Stable Butterworth patterns.

Footnote: Since this investigation was completed, the author's attention has been drawn to the report Kou varitakis, 1977 in which the asymptotic root locus theory developed in Kou varitakis and Shaked, 1976 is applied to the optimal linear system configuration to obtain the same results.

1.4 Summary of Asymptotic Optimal Control Theory

The asymptotic behaviour of the optimal closed loop poles has been described for the quadratic cost functional

$$J = \frac{1}{2} \int_0^{\infty} \{ \langle y(t), Qy(t) \rangle + \rho \langle u(t), R_0 u(t) \rangle \} dt$$

with time invariant linear system

$$S(A,B,C) \triangleq \begin{cases} \dot{x}(t) = A x(t) + Bu(t) & x(0) = x. \\ y(t) = C x(t) \end{cases}$$

where there are n states, r outputs and m inputs.

As $\rho \rightarrow 0$, the optimal closed loop poles behave according to the following general rules:

- 1) The finite system zeros attract (according to their multiplicity) the same number of closed loop poles to various left half plane positions. The right half plane zeros attract to their mirror image position in the left half plane and the left half plane zeros attract to their actual left half plane locations.
- 2) The remaining optimal closed loop poles tend to infinity in various orders of differing radii stable Butterworth patterns. A method of deriving the various patterns was presented in some detail in Section 1.3.

The asymptotic theory is the framework in which the control properties of the input structure diagnostics is to be viewed. It relates directly how the closed loop poles are behaving as the cost parameter ρ is being modified. The next chapter relates (using two kinds of analyses) the value

of the optimal cost (through which the diagnostic is defined)
to the limiting process $\rho \rightarrow 0$, and the underlying system pole-
zero configuration. With these relationships established the
control-theoretic basis of structure assessment is complete.

Chapter 2 The Control Properties of Input Structure Diagnostics

2.0 Introduction

2.1 Asymptotic Properties for Input Structure Diagnostics

2.2 Basic Relations for Input Structure Assessment

2.3 Input Deletion and the Preservation of Controllability

2.4 Input Deletion and the Loss of Controllability

2.5 Summary of Control Properties of Input Structure Diagnostics

2.0 Introduction

The input structure diagnostics can be described in the following general notational scheme. The quadratic cost functional is

$$J = \frac{1}{2} \int_0^{T_f} (\langle e(t), Qe(t) \rangle + \langle u(t), Ru(t) \rangle) dt,$$

with process equations $e(t) = r(t) - y(t)$

$$\text{and } y(t) = (Wu)(t) + d(t). \quad 0 \leq t \leq T_f.$$

By setting $d(t) \equiv 0$, and $r(t)$ to the desired system response, the optimal servo problem is obtained. For the regulation of an initial condition disturbance, set $r(t) \equiv 0$ and $d(t)$ to the initial condition disturbance.

To be consistent with the control interpretations obtained by specializing to the state space system description $\dot{S}(A,B,C)$, this chapter is written in terms of the notation for the regulation of an initial condition disturbance. The diagnostics are then written as:

$$i) \quad \Delta J_0 = J(u=0) - J(u=\hat{u})$$

representing the total cost improvement obtained by applying optimal control, and

ii) $IESQ = \frac{1}{2} \int_0^{T_f} \langle \hat{y}(t), Q\hat{y}(t) \rangle dt$ which is simply the output inner product term of the cost functional evaluated with the optimal output. This quantity is a measure of how well the desired system response is being achieved by the use of optimal control.

Further simplification is achieved by the formal use of Hilbert space notation, and by making some assumptions on the cost functional weighting operators which facilitate a rescaling of the original problem statement.

Assume, therefore that operators Q and R are self-adjoint and positive definite. The control weighting operator R is taken to be of the form $R = \rho R_0$, where scalar $\rho > 0$ is utilized in the discussion of the asymptotic behaviour and R_0 has a diagonal structure. This latter quality enables deletions of inputs to be related to the deletions of diagonal elements of R_0 .

Using square root factorizations for R_0 , and Q , the following rescalings are obtained:

$$\begin{aligned} J &= \frac{1}{2} \langle y, Q y \rangle_{H_r} + \frac{\rho}{2} \langle u, R_0 u \rangle_{H_m} \\ &= \frac{1}{2} \langle Q^{\frac{1}{2}} y, Q^{\frac{1}{2}} y \rangle_{H_r} + \frac{\rho}{2} \langle R_0^{-\frac{1}{2}} u, R_0^{-\frac{1}{2}} u \rangle_{H_m} \end{aligned}$$

Redefine $Q^{\frac{1}{2}} y$ as y , and $R_0^{-\frac{1}{2}} u$ as u to give

$$J = \frac{1}{2} \langle y, y \rangle_{H_r} + \frac{\rho}{2} \langle u, u \rangle_{H_m}$$

Similarly the process equation $y = Wu + d$ becomes

$$Q^{\frac{1}{2}} y = Q^{\frac{1}{2}} W R_0^{-\frac{1}{2}} u + Q^{\frac{1}{2}} d$$

to give $y = Wu + d$, where $Q^{\frac{1}{2}} W R_0^{-\frac{1}{2}}$ has been replaced by W , and $Q^{\frac{1}{2}} d$ by d . The full problem may be stated in rescaled form as:

$$\begin{aligned} \text{Min } J &= \frac{1}{2} \langle y, y \rangle_{H_r} + \frac{1}{2} \rho \langle u, u \rangle_{H_m} \\ u &\in H_m \\ \text{subject to } y &= Wu + d, \rho > 0. \end{aligned}$$

In the context of this rescaled problem, the control properties of the diagnostics are to be investigated. In the early part of the chapter linear space concepts are used to describe the asymptotic behaviour of the diagnostics. Using the pole-zero results of the first chapter, control interpretations are given for the state space system description.

Using the same linear space concepts the discussion moves onto the use of the diagnostics for structure assessment. In the process of proving consistency for ΔJ_0 , and asymptotic consistency for IESQ much insight is gained as to the significance of structure assessment.

Finally the consequences of deletion are related to the multivariable concepts of controllability and various types of pole-zero patterns. Later it will be shown how sets of bounds can be generated for the diagnostics enabling a rapid assessment of input structures.

Notation Frequent use is made of the sub/super script 's', and 'f'; these indicate dependence on a control subset (a structure) or the full control set respectively.

2.1 Asymptotic Properties for Input Structure Diagnostics

Recall the optimization problem as:

$$\text{Min}_{u \in H_m} J = \frac{1}{2} \langle y, y \rangle_{H_r} + \frac{\rho}{2} \langle u, u \rangle_{H_m} \quad \text{subject to } y = Wu + d$$

Substituting for y obtain

$$J = \frac{1}{2} \langle u, (W^*W + \rho I) u \rangle_{H_m} + \langle u, W^*d \rangle_{H_m} + \frac{1}{2} \langle d, d \rangle_{H_r}$$

for which a gradient on H_m may be defined

$$g(u) = \frac{\partial J}{\partial u} = (W^*W + \rho I) u + W^*d$$

to give the optimal control as $\hat{u} = -(W^*W + \rho I)^{-1} W^*d$

It follows from this that the optimal cost is

$$\hat{J}(\rho) = \frac{1}{2} \langle d, d \rangle_{H_r} - \frac{1}{2} \langle W^*d, (W^*W + \rho I)^{-1} W^*d \rangle_{H_m}$$

the optimal cost improvement is

$$\Delta J_0(\rho) = \frac{1}{2} \langle W^*d, (W^*W + \rho I)^{-1} W^*d \rangle_{H_m}$$

and the IESQ diagnostic is $\text{IESQ}(\rho) = \frac{1}{2} \langle \hat{y}, \hat{y} \rangle_{H_r}$ where $\hat{y} = W\hat{u} + d$

The asymptotic behaviour of these three quantities is obtained

using Householder's relation for operators (Chuang, 1970) and the

eigen-structure of compact self adjoint operators on Hilbert spaces.

Begin by recalling Householder's matrix equality which can be extended

for use with suitably defined Hilbert space operators (Chuang, 1970), as

$$(B + CDE)^{-1} = B^{-1} - B^{-1}CD(D + DEB^{-1}CD)^{-1}DEB^{-1} \quad 2.1.1$$

where B , and D are assumed positive definite. This enables the following

commutativity to be obtained.

Lemma 2.1.1 If $F: H_r \rightarrow H_m$ and $G: H_m \rightarrow H_r$

then $(I+GF)^{-1}G = G(I+FG)^{-1}$ (assuming the inverses exist).

Proof Using 2.1.1 on the left hand side:

$$\begin{aligned}(I+GF)^{-1}G &= (I-G(I+FG)^{-1}F)G \\ &= G(I - (I+FG)^{-1}FG)\end{aligned}$$

Using 2.1.1 with $B=I, C=I, D=I, E=FG$ to give

$$(I+GF)^{-1}G = G(I+FG)^{-1}$$

◻

Define operator $D: H_r \rightarrow H_r$ where $D=WW^*$, then the assumption of compactness introduced on D enables the eigen-structures of compact self adjoint operators between Hilbert spaces to be used to further the analysis. Hsieh (1965) has justified the use of such an assumption for a large class of linear systems including those described by a state space triple $S(A, B, C)$.

Compact operator $D=WW^*$ is positive semi definite and self adjoint hence it has a denumerable set of positive eigen values $\{\lambda_i\}_{i=1}^{\infty}$ and a corresponding set of orthonormal eigenfunctions $\{e_i\}_{i=1}^{\infty}$. Further one of the eigenvalues attains the value of the norm of operator D , and the only possible point of accumulation for $\{\lambda_i\}_{i=1}^{\infty}$ is zero.

Denote $\lambda_M = \|D\| = \sup_{x \in H_r} \{ \langle x, Dx \rangle / \langle x, x \rangle \}$

and $\lambda_1 \in (0, \lambda_M]$. The space H_r is therefore capable of a direct sum decomposition $H_r = H_1 \oplus H_2$ where $H_1 = \text{span} \{e_i\}_{i=1}^{\infty}$ and H_2 is the null space of operator D (denoted by $N(D)$). The relationship between the nullities of D , and W^* is expressed in the following;

Lemma 2.1.2 If $D = WW^*$ where $W^*: H_r \rightarrow H_m$, and $D: H_r \rightarrow H_r$ then $N(D) \equiv N(W^*)$

Proof (i) Suppose $x \in N(D)$ implies $Dx = 0 \in H_r$
then $\langle x, Dx \rangle_{H_r} = \langle W^*x, W^*x \rangle_{H_m} = 0$ (by assumption)
so that $W^*x = 0$ and $x \in N(W^*)$. $N(D) \subseteq N(W^*)$
(ii) Suppose $x \in N(W^*)$ then $W^*x = 0 \in H_m$
consequently $WW^*x = Dx = 0$ giving $N(W^*) \subseteq N(D)$
(i) and (ii) together yield the result.

Since H_1 is the orthogonal complement to $H_2 \triangleq N(D)$, have through the identification in Lemma 2.1.2 and Luenberger, (1969) that

$H_1 = N(W^*)^\perp = \overline{R(W)}$ which is a closed linear subspace of a complete Hilbert space, consequently (Naylor and Sell, 1971) it too is complete and may therefore be treated as a complete Hilbert space.

Finally since $H_1 = \overline{R(W)}$, this implies by definition (Naylor and Sell, 1971) that $R(W)$ is dense in H_1 . Intuitively then, the elements of the range of the system operator can approach arbitrary close to any element in $H_1 = \text{sp} \{e_i\}_{i=1}^{\infty}$.

With this characterization decompose the disturbance d , as

$$d = d_1 + d_2 \quad \text{where} \quad d_1 \in H_1, \quad \text{and} \quad d_2 \in H_2.$$

Then $\exists$ scalars $\{\alpha_i\}_{i=1}^{\infty}$ such that $d_1 = \sum_{i=1}^{\infty} \alpha_i e_i$, and further $Dd = Dd_1 = \sum_{i=1}^{\infty} \alpha_i \lambda_i e_i$ where $\{\lambda_i\}_{i=1}^{\infty}$ are the eigenvalues of D .

Theorem 2.1.1 Asymptotic Diagnostic Values

If the disturbance d is decomposed as $d = d_1 + d_2$ where

$d_1 \in H_1 = \overline{\mathcal{R}(W)}$ and $d_2 \in H_2 = \mathcal{N}(W^*)$ then

$$(i) \quad \lim_{\rho \rightarrow 0} \hat{J}(\rho) = \frac{1}{2} \langle d_2, d_2 \rangle_{H_r}$$

$$(ii) \quad \lim_{\rho \rightarrow 0} \Delta J_0(\rho) = \frac{1}{2} \langle d_1, d_1 \rangle_{H_r}$$

$$\text{and (iii)} \quad \lim_{\rho \rightarrow 0} \text{IESQ}(\rho) = \frac{1}{2} \langle d_2, d_2 \rangle_{H_r}$$

Proof (i) Recall $\hat{J}(\rho) = \frac{1}{2} \langle d, d \rangle_{H_r} - \frac{1}{2} \langle W^* d, (W^* W + \rho I)^{-1} W^* d \rangle_{H_m}$

Use Lemma 2.1.1 with $G = W$ and $F = W^*$ to give

$$\hat{J}(\rho) = \frac{1}{2} \langle d, d \rangle_{H_r} - \frac{1}{2} \langle d, (\rho I + W W^*)^{-1} W W^* d \rangle_{H_r}$$

Substituting for $D = W W^*$, $d = d_1 + d_2$ and using orthogonality

$(d_1 \perp d_2)$, and $d_2 \in \mathcal{N}(D)$ obtain

$$\hat{J}(\rho) = \frac{1}{2} \langle d_1, d_1 \rangle_{H_r} + \frac{1}{2} \langle d_2, d_2 \rangle_{H_r} - \frac{1}{2} \langle d, (\rho I + D)^{-1} D d_1 \rangle_{H_r}$$

Using the eigenstructure of D , have $D d_1 = \sum_{i=1}^{\infty} \alpha_i \lambda_i e_i$,

$$\text{also } (\rho I + D) e_i = \rho e_i + D e_i = (\rho + \lambda_i) e_i$$

$$\Rightarrow (\rho + \lambda_i)^{-1} e_i = (\rho I + D)^{-1} e_i$$

$$\text{hence } (\rho I + D)^{-1} D d_1 = \sum_{i=1}^{\infty} \alpha_i (\rho + \lambda_i)^{-1} \lambda_i e_i$$

$$\text{Giving } \hat{J}(\rho) = \frac{1}{2} \langle d_1, d_1 \rangle_{H_r} + \frac{1}{2} \langle d_2, d_2 \rangle_{H_r} - \frac{1}{2} \sum_{i=1}^{\infty} \alpha_i^2 \lambda_i / (\rho + \lambda_i)$$

$$\text{Then } \lim_{\rho \rightarrow 0} \hat{J}(\rho) = \frac{1}{2} \langle d_1, d_1 \rangle_{H_r} + \frac{1}{2} \langle d_2, d_2 \rangle_{H_r} - \frac{1}{2} \sum_{i=1}^{\infty} \alpha_i^2$$

$$\text{But } \frac{1}{2} \sum_{i=1}^{\infty} \alpha_i^2 = \frac{1}{2} \langle d_1, d_1 \rangle \Rightarrow \lim_{\rho \rightarrow 0} \hat{J}(\rho) = \frac{1}{2} \langle d_2, d_2 \rangle, \quad d_2 \in \mathcal{N}(W^*)$$

(ii) The improvement in cost obtained by applying optimal control is

$$\text{given as } \Delta J_0(\rho) = J(u=0) - \hat{J}(\rho)$$

$$\Rightarrow \lim_{\rho \rightarrow 0} \Delta J_0(\rho) = J(u=0) - \lim_{\rho \rightarrow 0} \hat{J}(\rho)$$

$$= \frac{1}{2} \langle d, d \rangle_{H_r} - \frac{1}{2} \langle d_2, d_2 \rangle_{H_r}$$

$$\Rightarrow \lim_{\rho \rightarrow 0} \Delta J_0(\rho) = \frac{1}{2} \langle d_1, d_1 \rangle, \quad d_1 \in \overline{\mathcal{R}(W)}$$

$$\begin{aligned} \text{(iii) } \text{IESQ}(\rho) &= \frac{1}{2} \langle \hat{y}(\rho), \hat{y}(\rho) \rangle_{H_r} \\ &= \frac{1}{2} \langle W \hat{u}(\rho) + d, W \hat{u}(\rho) + d \rangle_{H_r} \end{aligned}$$

$$\text{where } \hat{u}(\rho) = -(W^*W + \rho I)^{-1} W^*d \in H_m$$

Expanding obtain

$$\begin{aligned} \text{IESQ}(\rho) &= \frac{1}{2} \langle d, d \rangle - \langle d, W(W^*W + \rho I)^{-1} W^*d \rangle_{H_r} \\ &\quad + \frac{1}{2} \langle W(W^*W + \rho I)^{-1} W^*d, W(W^*W + \rho I)^{-1} W^*d \rangle_{H_r} \end{aligned}$$

Using Lemma 2.1.1 rearrange to obtain:

$$\begin{aligned} \text{IESQ}(\rho) &= \frac{1}{2} \langle d, d \rangle_{H_r} - \langle d, (\rho I + D)^{-1} Dd \rangle_{H_r} \\ &\quad + \frac{1}{2} \langle (\rho I + D)^{-1} Dd, (\rho I + D)^{-1} Dd \rangle_{H_r} \end{aligned}$$

Using the eigenstructure for D, obtain

$$\text{IESQ}(\rho) = \frac{1}{2} \langle d_1, d_1 \rangle + \frac{1}{2} \langle d_2, d_2 \rangle - \frac{1}{2} \sum_{i=1}^{\infty} \frac{\alpha_i^2 \lambda_i (\lambda_i + 2\rho)}{(\lambda_i + \rho)^2}$$

$$\lim_{\rho \rightarrow 0} \text{IESQ}(\rho) = \frac{1}{2} \langle d_1, d_1 \rangle + \frac{1}{2} \langle d_2, d_2 \rangle - \frac{1}{2} \sum_{i=1}^{\infty} \alpha_i^2$$

Using $\frac{1}{2} \langle d_1, d_1 \rangle = \frac{1}{2} \sum_{i=1}^{\infty} \alpha_i^2$ gives the result

$\lim_{\rho \rightarrow 0} \text{IESQ}(\rho) = \frac{1}{2} \langle d_2, d_2 \rangle$, $d_2 \in H_m$. which concludes the proof.

An immediate consequence of this theorem is the asymptotic behaviour of the difference

$$\hat{J}(\rho) - \text{IESQ}(\rho) = \frac{\rho}{2} \langle \hat{u}(\rho), \hat{u}(\rho) \rangle_{H_m}$$

which is given in the following corollary.

Corollary 2.1.1 $\lim_{\rho \rightarrow 0} \frac{\rho}{2} \langle \hat{u}(\rho), \hat{u}(\rho) \rangle_{H_m} = 0$

Proof Recall $\hat{J}(\rho) = \frac{1}{2} \langle \hat{y}(\rho), \hat{y}(\rho) \rangle_{H_r} + \frac{\rho}{2} \langle \hat{u}(\rho), \hat{u}(\rho) \rangle_{H_m}$

that is $\hat{J}(\rho) = \text{IESQ}(\rho) + \frac{\rho}{2} \langle \hat{u}(\rho), \hat{u}(\rho) \rangle_{H_m}$

From theorem 2.1.1 $\lim_{\rho \rightarrow 0} \hat{J}(\rho) = \lim_{\rho \rightarrow 0} \text{IESQ}(\rho)$

but $\lim_{\rho \rightarrow 0} \hat{J}(\rho) = \lim_{\rho \rightarrow 0} \text{IESQ}(\rho) + \lim_{\rho \rightarrow 0} \frac{\rho}{2} \langle \hat{u}(\rho), \hat{u}(\rho) \rangle_{H_m}$

Hence the result $\lim_{\rho \rightarrow 0} \frac{\rho}{2} \langle \hat{u}(\rho), \hat{u}(\rho) \rangle_{H_m} = 0$

If d is typical of a class of disturbances, the asymptotic value of the optimal cost, or the IESQ diagnostic isolates that part of the disturbance which the optimal control cannot remove. This motivates the concept of systems of maximal accuracy (Kwakernaak and Sivan, 1972a). These are systems for which the optimal cost can always be reduced to zero as input amplitudes are permitted to increase indefinitely. The conditions under which this occurs for state space systems were formulated by Davison and Chow, (1974) as the Control Inequality Principle. It states that if a linear system has more outputs than inputs (namely $r > m$) it is not possible to achieve maximal accuracy. Two necessary conditions for maximal accuracy are:

- 1) more or the same number of inputs to outputs, and
- 2) no right half plane zeros. (The conditions are also sufficient for the case of square systems, $r = m$).

Formally, a system is maximally accurate with respect to a class of disturbances $\mathcal{C} \subseteq H_r$ if $\forall d \in \mathcal{C} \subseteq H_r$

$$\lim_{\rho \rightarrow 0} \text{IESQ}(\rho) = 0$$

To obtain necessary and sufficient conditions for maximal accuracy partition the disturbance d , as $d = d_1 + d_2$ where $d_1 \in \overline{\mathcal{R}(W)} = H_1$ and $d_2 \in \mathcal{N}(W^*) = H_2$. From theorem 2.1.1

$$\lim_{\rho \rightarrow 0} \text{IESQ}(\rho) = \frac{1}{2} \langle d_2, d_2 \rangle_{H_r} \quad \text{hence} \quad \frac{1}{2} \langle d_2, d_2 \rangle = 0 \quad \forall d \in \mathcal{C} \subseteq H_r$$

requires $d_2 = 0 \quad \forall d \in \mathcal{C}$ giving the necessary and sufficient condition for maximal accuracy as $d \in \overline{\mathcal{R}(W)} \quad \forall d \in \mathcal{C} \subseteq H_r$.

Consider a stable state space system description $\mathcal{S}(A, B, C)$, for which $(Wu)(t) \triangleq \int_0^t C e^{A(t-\tau)} B u(\tau) d\tau$, subject to a class of initial condition disturbances $\mathcal{C} \triangleq \{ d: d(t) = C e^{At} x_0, 0 \leq t < \infty, x_0 \in \mathbb{R}^n \}$

The conditions for maximal accuracy are those for which $d \in \overline{\mathcal{R}(W)}$.

If $d \in \mathcal{R}(W)$, then $d \in \overline{\mathcal{R}(W)}$ and if d is a limit point of $\overline{\mathcal{R}(W)}$,

then for any given $\eta > 0$, $\exists d_1 \in \mathcal{R}(W)$ such that $\|d - d_1\| < \eta$

Examine therefore the conditions for the simpler case $d \in \mathcal{R}(W)$.

On the infinite time interval $0 \leq t < \infty$: $d(t) = (Wu)(t)$.

Taking laplace transforms $C(sI - A)^{-1} x_0 = C(sI - A)^{-1} B \bar{u}$.

Assume (i) complete controllability and observability,

(ii) $n \gg \max(r, m)$ (iii) $\text{rank}(C) = r$, $\text{rank}(B) = m$ and

$\text{rank}(C(sI - A)^{-1} B) = \min(r, m)$.

Case (i) $r \leq m$ clearly $\text{rank } C(sI - A)^{-1}B = r$

The rank decomposition for

$$G(s) = C(sI - A)^{-1}B = P(s) \begin{bmatrix} \Gamma(s) \\ 0_{r \times m-r} \end{bmatrix} Q(s)$$

where $P(s)$ and $Q(s)$ are $(r \times r)$ and $(m \times m)$ unimodular matrices and $\Gamma(s)$ is a $(r \times r)$ diagonal matrix whose diagonal entries expose the system transmission zeros and open loop poles.

$$\text{Then } C(sI - A)^{-1}x_0 = P(s) \begin{bmatrix} \Gamma(s) \\ 0_{r \times m-r} \end{bmatrix} Q(s) u$$

$$\text{implying } P(s)^{-1} C(sI - A)^{-1}x_0 = \Gamma(s) Z_r$$

$$\text{where } \begin{bmatrix} Z_r \\ Z_{m-r} \end{bmatrix} = Q(s) \bar{u}$$

$$\text{Hence } Z_r = \Gamma(s)^{-1} P(s)^{-1} C(sI - A)^{-1}x_0$$

$$\text{giving } \bar{u} = Q(s)^{-1} \begin{bmatrix} \Gamma(s)^{-1} P(s)^{-1} C(sI - A)^{-1}x_0 \\ Z_{m-r} \end{bmatrix}$$

where Z_{m-r} is arbitrary (subject to $u \in H_m$).

Therefore under the assumption $r \leq m$, the control $\bar{u}$ has been

exhibited for which $d \in \overline{\mathcal{R}(W)}$, $\forall d \in \mathcal{Z}$. Note that Z_{m-r} is

arbitrary and does not appear in the argument for the case $r = m$.

Further it is also necessary to restrict the original system to have no right half plane zeros otherwise $\bar{u}$ is not a stable control.

(see Kwakernaak and Sivan, 1972a).

Case (ii) $r > m$ $\text{rank } C(sI - A)^{-1}B = m$.

$$\text{Here } C(sI - A)^{-1}x_0 = P(s) \begin{bmatrix} \Gamma(s) \\ 0_{r-m \times m} \end{bmatrix} Q(s) \bar{u}$$

$$\text{Giving } P(s)^{-1} [C(sI - A)^{-1}x_0] = \begin{bmatrix} \Gamma(s) Q(s) \bar{u} \\ 0_{r-m} \end{bmatrix}$$

Where the equality is to hold for all $x_0 \in \mathbb{R}^n$.

But $P(s)$ and $(sI - A)^{-1}$ are non-singular and C (by assumption) has rank r where $m < r \ll n$ hence $P(s)^{-1} C (sI - A)^{-1}$ provides r linearly independent columns for which x_0 can be chosen to contradict the zero vector 0_{r-m} of the left hand side.

The necessary conditions for maximal accuracy have therefore been demonstrated, namely the need for more or the same number of inputs to outputs and for a minimum phase system. That these conditions are also sufficient in the case of square systems is easily seen since the only the analysis of case(i) would be required and the required control would simplify to

$$\bar{u} = Q^{-1}(s) \Gamma^{-1}(s) P^{-1}(s) \cdot C (sI - A)^{-1} x_0$$

hence the necessary and sufficient condition of a minimum phase system results.

Using the pole-zero concepts obtained previously in the first chapter the implications of attaining maximal accuracy are explored further. Recall, for the system triple $\mathcal{S}(A, B, C)$ and the usual quadratic cost functional, the optimal closed loop system as $\dot{\hat{x}}(t) = A_{CL} \hat{x}(t)$; $\hat{x}(t_0) = x_0$ where $A_{CL} = A - B R^{-1} B^T P_\infty$ and P_∞ is the solution to the associated steady state matrix Riccati equation. It has been shown that $A_{CL} = T[-\Lambda]T^{-1}$ where $[-\Lambda]$ is a matrix of stable eigenvalues and T is a matrix of the corresponding eigenvectors.

Hence $Z(t) = e^{-\Lambda(t-t_0)} Z_0$ where $Z(t) = T^{-1} \hat{x}(t)$ and $Z_0 = T^{-1} x_0$.

The optimal output is $\hat{y}(t) = C \hat{x}(t)$ which may be written as

$$\hat{y}(t) = C T e^{-\Lambda(t-t_0)} Z_0 . \text{ This gives}$$

$$IESQ = \frac{1}{2} \int_0^{\infty} \langle e^{-\Lambda t} Z_0, (CT)^T Q CT e^{-\Lambda t} Z_0 \rangle_{E_n} dt.$$

The asymptotic analysis of Chapter 1 has shown that those left half plane closed loop poles tending to infinity do so in a Butterworth pattern, whilst the remainder of the closed loop poles are attracted by the finite zeros. In the case where maximal accuracy is achievable the following theorem is relevant.

Theorem 2.1.2 If $\lim_{\rho \rightarrow 0} IESQ(\rho) = 0$, then the following must hold:

- i) those closed loop poles attracted to the left half plane poles become (as $\rho \rightarrow 0$) unobservable,
- ii) there are no right half plane zeros, and
- iii) the remaining closed loop poles become arbitrary fast.

Proof Recall the steady-state matrix Riccati equation as

$$P A + A^T P - P B R^{-1} B^T P + C^T Q C = 0 \quad 2.1.2$$

$$\text{then } \lim_{\rho \rightarrow 0} IESQ(\rho) = \lim_{\rho \rightarrow 0} \hat{J}(\rho) = \lim_{\rho \rightarrow 0} \frac{1}{2} \langle x_0, P x_0 \rangle = 0$$

giving $\lim_{\rho \rightarrow 0} P = 0_{n \times n}$ since $x_0 \neq 0$ is arbitrary.

i) Rearrange 2.1.2 to give

$$P A_{CL} + A_{CL}^T P = -C^T Q C \quad 2.1.3$$

Using $A_{CL} = T[-\Lambda]T^{-1}$, the above 2.1.3 becomes

$$P T[-\Lambda]T^{-1} + A^T P = -C^T Q C \quad 2.1.4$$

Suppose columns t_{k_1} and t_{k_2} are associated with a pair of complex conjugate closed loop poles which are attracted by a pair of complex conjugate zeros as $\rho \rightarrow 0$. Denote $T_{k_1 k_2} = [t_{k_1} t_{k_2}] \in \mathbb{R}^{n \times 2}$

then pre-multiplying 2.1.4 by $T_{k_1 k_2}^T$, post-multiplying by $T_{k_1 k_2}$

and using the properties of T obtain

$$T_{k_1 k_2}^T P T [-\Lambda]_{k_1 k_2} + T_{k_1 k_2}^T A^T P T_{k_1 k_2} = -[C T_{k_1 k_2}]^T Q C T_{k_1 k_2}$$

where $[-\Lambda]_{k_1 k_2}$ is a $(n \times 2)$ matrix containing two columns of $[-\Lambda]$ associated with the limiting positions of those closed loop poles attracted to the left half plane zeros.

Then as $\rho \rightarrow 0$

- a) $P \rightarrow 0$
- b) $[-\Lambda]_{k_1 k_2} \rightarrow$ finite left hand plane zero positions
- c) T remains finite since it is an eigenvector matrix.

$$\text{Hence } \lim_{\rho \rightarrow 0} T_{k_1 k_2}^T P T [-\Lambda]_{k_1 k_2} + T_{k_1 k_2}^T A^T P T_{k_1 k_2} = 0$$

$$\text{implying } \lim_{\rho \rightarrow 0} [C T_{k_1 k_2}]^T Q C T_{k_1 k_2} = 0 \quad (Q > 0)$$

hence $C T_{k_1 k_2} = 0$ and the closed loop poles attracted by the left half plane zeros become unobservable in the output as $\rho \rightarrow 0$.

ii) Assume that there are a pair of complex conjugate right half plane zeros, and associate with them an eigenvector matrix

$$T_{k_1 k_2} = [t_{k_1}, t_{k_2}] \quad (\text{as with the pair of left half plane zeros}).$$

Using the same analysis as above conclude that if

$$\lim_{\rho \rightarrow 0} \text{IESQ}(\rho) = 0 \quad \text{then} \quad \lim_{\rho \rightarrow 0} P = 0 \quad (\forall x_0) \quad \text{and} \quad C T_{k_1 k_2} = 0$$

That is, the modes which tend to the limiting positions associated with the right half plane zeros become decoupled from the output in the limiting position. This however would imply the existence of a pair of zeros in the limiting left half plane position. If such a pair did exist, they themselves would attract two closed loop poles which would still leave the closed loop poles attracted by the right half plane zeros undecoupled giving a contradiction, so there are no right half plane zeros.

iii) For the remaining part, knowledge of the effects of the zeros is used. Assuming no right half plane zeros, partition the output

$$\hat{y}(t) = \hat{y}_1(t) + \hat{y}_2(t) \quad \text{where} \quad \hat{y}_1(t) = C T_1 [e^{-\Lambda t} z_0]_1$$

contains the modes of the Butterworth patterns and

$$\hat{y}_2(t) = C T_2 [e^{-\lambda^t z_0}]_2$$

contains the modes associated with the left half plane zeros. For the latter, as $\rho \rightarrow 0$ the decoupling effect of the zero implies

$\lim_{\rho \rightarrow 0} y_2(t) = 0 \quad \forall t \in [0, \infty)$, hence the quantity $IESQ(\rho)$ is wholly

explained by the responses of the modes attracted by the infinite zeros:

that is $\lim_{\rho \rightarrow 0} IESQ(\rho) = \lim_{\rho \rightarrow 0} \frac{1}{2} < y_1, Q y_1 >_{H_T}$. Under maximal accuracy

$\lim_{\rho \rightarrow 0} \frac{1}{2} < y_1, Q y_1 >_{H_T} = 0$ consequently the Butterworth pattern modes become arbitrary fast.

If maximal accuracy is not achieved then the limiting value of $IESQ(\rho)$ may be interpreted as quantifying the control difficulties not overcome. For those systems with more or the same number of inputs as outputs, the degrees of freedom available for control lie most favourably with the inputs consequently the loss of maximal accuracy is very likely to be due to the presence of a right half plane zero. In the case of systems with more outputs than inputs the loss may be attributed to the following factors:

- a) Right half plane zeros: A right half plane zero will attract (as $\rho \rightarrow 0$) a closed loop pole to a finite position in the left half plane. The response of this mode will (given a suitable excitation) appear in the output since complete observability is assumed. However the actual magnitude of this type of term will depend on the numbers and positions of the right half plane zeros (since a far-away zero will give rise to a mode response which decays quickly and a zero close to the origin will generate a slowly decaying mode response) and the size of the initial condition excitation.
- b) Left half plane zeros: The loss of maximal accuracy reflects the loss of asymptotic unobservability associated with those poles attracted by the left half plane zeros. Hence in the limit these modes will generate a non-zero output term whose magnitude will depend on the numbers and positions of the left half plane zeros and upon the size of the excitation.

c) The Butterworth pattern poles. If the system has no zeros then the loss of maximal accuracy implies that the Butterworth pattern poles do not become arbitrary fast (otherwise a contradiction with $\lim_{\rho \rightarrow 0} \text{IESQ}(\rho) > 0$ follows). Although not arbitrary fast these modes attracted by infinite zeros tend to decay rapidly and unless excited by an unusually large initial condition are thought to form a less significant part of the limiting value of $\text{IESQ}(\rho)$.

So far maximal accuracy has been discussed in terms of the diagnostic $\text{IESQ}(\rho)$. However maximal accuracy is also quantified by the optimal control improvement diagnostic $\Delta J_0(\rho)$. Here the asymptotic optimal control improvement is

$$\lim_{\rho \rightarrow 0} \Delta J_0(\rho) = \frac{1}{2} \langle d_1, d_1 \rangle_{H_r} \leq \frac{1}{2} \langle d, d \rangle_{H_r},$$

which is clearly maximized when $d_1 = d \in \overline{R(W)}$. The difference between the asymptotic value of $\Delta J_0(\rho)$ and the possible maximal value, $m_0 = \frac{1}{2} \langle d, d \rangle_{H_r}$, is naturally interpreted as a measure of the control difficulties which the application of optimal control has failed to remove. These are the inherent system difficulties like right half plane zeros or insufficient numbers of inputs.

The use of the diagnostics for structure assessment is based on their quantification of system control difficulties. Input deletion used to create structures may cause control difficulties such as the loss of degrees of freedom to cope with left half plane zeros, or the introduction of new zeros. Since the system excitation has been related to the physical environment of the process, these new difficulties will be reflected in the value of the diagnostic in a manner which enables a valid comparison to be made between the control properties of a structure and those of the full optimal system.

2.2 Basic Relations for the Input Structure Assessment

It is intuitively felt that the full control set should be as effective as, if not more effective than, a control structure comprising of a subset of the available inputs. This intuitive notion is formalised as the property of consistency, which is investigated in this section. The optimal cost improvement diagnostic is shown to have consistency whatever the value of ρ , whilst diagnostic $\text{IESQ}(\rho)$ is shown to have asymptotic consistency. In the course of the investigation further insight is obtained into the composition of the two diagnostics.

As structures created by the deletion of inputs are produced, the assumption of complete observability remains unchanged. This ensures all the possible effects of deletion (for example an uncontrolled subsystem) are monitored by the diagnostics. The property of consistency for diagnostic ΔJ_0 is shown next.

Consider the primal problem:

$$\text{Min}_{u_f \in H_m} J_f = \frac{1}{2} \langle y_f, y_f \rangle_{H_r} + \frac{\rho}{2} \langle u_f, u_f \rangle_{H_m}$$

$$\text{subject to } y_f = W u_f + d; \quad y_f \in H_r.$$

and a problem obtained from this by input deletion:

$$\text{Min}_{u_s \in H_m} J_s = \frac{1}{2} \langle y_s, y_s \rangle_{H_r} + \frac{\rho}{2} \langle u_s, u_s \rangle_{H_s}$$

$$\text{subject to } y_s = W_s u_s + d; \quad y_s \in H_r$$

Set $u_f = \begin{pmatrix} \hat{u}_s \\ 0 \end{pmatrix} + u$ where $u_f, u \in H_m$ and $\hat{u}_s$ is the optimal control for

the structures problem above.

$$\text{Then } y_f = W u_f + d = W_s \hat{u}_s + W u + d = \hat{y}_s + W u.$$

Substituting in the full cost, J_f gives

$$J_f(y_f, u_f) = J_s(\hat{y}_s, \hat{u}_s) + \frac{1}{2} \langle u, (W^*W + \rho I) u \rangle_{H_m} + \langle u, g_s \rangle_{H_m}$$

$$\text{where } g_s = W^* \hat{y}_s + \rho \begin{pmatrix} \hat{u}_s \\ 0 \end{pmatrix}.$$

Performing the minimization gives $\tilde{u} = -(W^*W + \rho I)^{-1} g_s$ and

$$\hat{J}_f = \hat{J}_s - \frac{1}{2} \langle \tilde{u}, (W^*W + \rho I) \tilde{u} \rangle_{H_m} \quad 2.2.1$$

$$\text{or } \Delta J_o^f - \Delta J_o^s = \frac{1}{2} \langle \tilde{u}, (W^*W + \rho I) \tilde{u} \rangle_{H_m} \quad 2.2.2$$

This basic relation for ΔJ_o shows the consistency to be independent of scalar ρ , and shows that the difference $\Delta J_o^f - \Delta J_o^s$ comprises of a term measuring (in the same control metric) the extra control energy required to reduce the optimal structure cost to the full optimal cost. Further, the positive-definiteness of this term ensures it is only zero if $\tilde{u} = 0$, as for example will occur if the structure is exactly optimal.

The investigation of the expression 2.2.2 is pursued further. Initially the effectiveness of $\hat{u}_s(\rho)$ and $\hat{u}_f(\rho)$ are compared by decomposing the disturbance, and eventually asymptotic considerations of the expression 2.2.2 are discussed.

Partition the system operator W as $[W_s, W_{m-s}]$ where operators W_s and W_{m-s} are associated with the structured and deleted systems respectively. In the previous section, the output space H_r was given the orthogonal decomposition $H_r = \overline{\mathcal{R}(W)} \oplus \mathcal{N}(W^*)$ which was also related to the eigenstructure of the compact self adjoint operator $D = WW^*$. The linear subspace $\overline{\mathcal{R}(W)}$ is a closed subspace in a complete

metric space, hence (Naylor and Sell, 1971) it too may be treated as an individual Hilbert space. Clearly the closed linear subspace associated with $\mathcal{R}(W_S)$, satisfies the following inclusion :

$\overline{\mathcal{R}(W_S)} \subseteq \overline{\mathcal{R}(W)}$. Using an orthogonal decomposition (Naylor and Sell, 1971)

obtain : $\overline{\mathcal{R}(W)} = \overline{\mathcal{R}(W_S)} \oplus X$, where X is the orthogonal complement

to $\overline{\mathcal{R}(W_S)}$ restricted to the subspace $\overline{\mathcal{R}(W)}$. Thus the output space now

has the orthogonal decomposition $H_r = \overline{\mathcal{R}(W_S)} \oplus X \oplus N(W^*)$ 2.2.3

where $\overline{\mathcal{R}(W)} = \overline{\mathcal{R}(W_S)} \oplus X$ (X is closed) 2.2.4

and $N(W_S^*) = X \oplus N(W^*)$ (by observation) 2.2.5

With such a decomposition for H_r , partition $d \in H_r$ as

$d = d_1 + d_2 + d_3$ where $d_1 \in \overline{\mathcal{R}(W_S)}$, $d_2 \in X$ and $d_3 \in N(W^*)$ 2.2.6

Lemma 2.2.1 Define $\hat{Z}_S(\rho) = W_S \hat{u}_S(\rho) \in \mathcal{R}(W_S) \subseteq \overline{\mathcal{R}(W_S)}$

and with $d = d_1 + d_2 + d_3$ (as specified in 2.2.6) then

$\lim_{\rho \rightarrow 0} \hat{Z}_S(\rho) = -d_1$, giving $\lim_{\rho \rightarrow 0} \hat{y}_S(\rho) = d_2 + d_3$

Proof $\hat{Z}_S(\rho) = W_S \hat{u}_S(\rho) = -W_S (W_S^* W_S + \rho I)^{-1} W_S^* d$
 $= -(\rho I + W_S W_S^*)^{-1} W_S W_S^* d$ (lemma 2.1.1)

Recall $d = d_1 + d_2 + d_3$ where $d_1 \in \overline{\mathcal{R}(W_S)}$ and $d_2 + d_3 \in N(W_S^*)$

$\Rightarrow \hat{Z}_S(\rho) = -(\rho I + W_S W_S^*)^{-1} W_S W_S^* d_1$

Using the eigenstructure of the compact self adjoint operator

$D_S = W_S W_S^*$, $\overline{\mathcal{R}(W_S)}$ has an ortho-normal eigen basis $\{e_i\}_{i=1}^{\infty}$

and associated positive eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$. Write d_1 as

$d_1 = \sum_{i=1}^{\infty} \alpha_i e_i$, then $\hat{Z}_S(\rho) = -\sum_{i=1}^{\infty} \alpha_i \left(\frac{\lambda_i}{\lambda_i + \rho} \right) e_i$

giving $\lim_{\rho \rightarrow 0} \hat{Z}_S(\rho) = -\sum_{i=1}^{\infty} \alpha_i e_i = -d_1$

Further $\hat{y}_s(\rho) = W_s \hat{u}_s(\rho) + d$

$$= \hat{Z}_s(\rho) + d_1 + d_2 + d_3$$

Hence $\lim_{\rho \rightarrow 0} \hat{y}_s(\rho) = \lim_{\rho \rightarrow 0} (\hat{Z}_s(\rho) + d_1) + d_2 + d_3 = d_2 + d_3$

This lemma indicates that the structure optimal control asymptotically removes from the full disturbance that part in the range of the structure operator. Similarly for the full structure, the full optimal control will remove from disturbance d , that part in the range of the full operator W . For the full structure, an equivalent lemma would be :

Lemma 2.2.2 Define $\hat{Z}(\rho) = W \hat{u}(\rho) \in R(W) \subseteq \overline{R(W)}$ and

$d = d_1 + d_2 + d_3$ (as specified in 2.2.6) then

$$\lim_{\rho \rightarrow 0} \hat{Z}(\rho) = -(d_1 + d_2) \quad \text{giving} \quad \lim_{\rho \rightarrow 0} \hat{y}(\rho) = d_3.$$

Proof As for Lemma 2.2.1.

A comparison of the two limiting outputs, namely

$\lim_{\rho \rightarrow 0} \hat{y}_s(\rho) = d_2 + d_3$ and $\lim_{\rho \rightarrow 0} \hat{y}(\rho) = d_3$, show that the structure optimal control to be less efficient in reducing the disturbance.

The portion of the disturbance d_2 , unreached by the structure optimal control, is solely related to the imposition of structure. It is an additional portion of the disturbance that the structure optimal control cannot remove, over and above that part d_3 which cannot be removed by the full optimal control.

The assessment of structures through $\Delta J_o^f - \Delta J_o^s$ is now shown asymptotically to quantify this structure dependent part of the disturbance. The most direct route (as taken here) to this result is to apply the asymptotic value theorem (2.1.1) of the previous section, however in Appendix 2 an alternative proof is given

beginning from $\Delta J_0^f - \Delta J_0^s = \frac{1}{2} < \tilde{u}, (W^*W + \rho I) \tilde{u} >$.

Theorem 2.2.1 Asymptotic Structure Assessment with ΔJ_0

If the output space H_r , is orthogonally decomposed as

$$H_r = \overline{R(W_S)} \oplus X \oplus N(W^*) \quad \text{with} \quad \overline{R(W)} = \overline{R(W_S)} \oplus X$$

and the disturbance d , partitioned as $d = d_1 + d_2 + d_3$

where $d_1 \in \overline{R(W_S)}$, $d_2 \in X$ and $d_3 \in N(W^*)$

then $\lim_{\rho \rightarrow 0} \Delta J_0^f(\rho) - \Delta J_0^s(\rho) = \frac{1}{2} < d_2, d_2 >_{H_r}$.

Proof (i) Full problem : define $\bar{d}_1 \triangleq d_1 + d_2 \in \overline{R(W)}$

$$\text{and } \bar{d}_2 = d_3 \in N(W^*)$$

Apply theorem 2.1.1, part (ii) to obtain

$$\lim_{\rho \rightarrow 0} \Delta J_0^f(\rho) = \frac{1}{2} < \bar{d}_1, \bar{d}_1 >_{H_r} = \frac{1}{2} < d_1, d_1 >_{H_r} + \frac{1}{2} < d_2, d_2 >_{H_r}$$

(ii) Structure problem : define $\bar{d}_1 \triangleq d_1 \in \overline{R(W_S)}$

$$\text{and } \bar{d}_2 \triangleq d_2 + d_3 \in N(W_S^*)$$

Apply theorem 2.1.1, part (ii) to obtain

$$\lim_{\rho \rightarrow 0} \Delta J_0^s(\rho) = \frac{1}{2} < \bar{d}_1, \bar{d}_1 >_{H_r} = \frac{1}{2} < d_1, d_1 >_{H_r}$$

$$\begin{aligned} \text{(iii)} \quad \lim_{\rho \rightarrow 0} (\Delta J_0^f(\rho) - \Delta J_0^s(\rho)) &= \lim_{\rho \rightarrow 0} \Delta J_0^f(\rho) - \lim_{\rho \rightarrow 0} \Delta J_0^s(\rho) \\ &= \frac{1}{2} < d_1, d_1 > + \frac{1}{2} < d_2, d_2 > - \frac{1}{2} < d_1, d_1 > \end{aligned}$$

$$\text{Hence } \lim_{\rho \rightarrow 0} (\Delta J_0^f(\rho) - \Delta J_0^s(\rho)) = \frac{1}{2} < d_2, d_2 >$$

◻

The assessment of structures is therefore based on the comparison of the magnitude of this term which may be associated with the extra control difficulties introduced by the deletion of inputs. The effect of deletion in terms of control properties is discussed later in the chapter.

The conditions for which a structure is asymptotically as accurate as the full system may be investigated. The necessary condition for this is $d_2=0$, which may occur for a particular disturbance choice. However, a sufficient condition is $\mathcal{R}(W_s)=\mathcal{R}(W)$ since this implies $X=\{0\}$ giving $d_2=0 \forall d \in H_r$. Although neither of these cases cause any difficulties in interpretation they would probably merit further investigation in physical terms, for example linearly dependent inputs.

In contrast to the optimal cost improvement diagnostic, consistency for the IESQ diagnostic is an asymptotic property.

Theorem 2.2.2 Asymptotic Structure Assessment with IESQ

If the output space H_r , is orthogonally decomposed as

$$H_r = \overline{\mathcal{R}(W_s)} \oplus X \oplus \mathcal{N}(W^*) \quad \text{with} \quad \overline{\mathcal{R}(W)} = \overline{\mathcal{R}(W_s)} \oplus X$$

and the disturbance d , partitioned as $d = d_1 + d_2 + d_3$

where $d_1 \in \overline{\mathcal{R}(W_s)}$, $d_2 \in X$, and $d_3 \in \mathcal{N}(W^*)$

$$\text{then} \quad \lim_{\rho \rightarrow 0} (\text{IESQ}_s(\rho) - \text{IESQ}_f(\rho)) = \frac{1}{2} \langle d_2, d_2 \rangle_{H_r}$$

Proof (i) Full problem : define $\bar{d}_1 \triangleq d_1 + d_2 \in \overline{\mathcal{R}(W)}$ and $\bar{d}_2 = d_3 \in \mathcal{N}(W^*)$

Apply theorem 2.1.1, part (iii) to obtain

$$\lim_{\rho \rightarrow 0} \text{IESQ}_f(\rho) = \frac{1}{2} \langle \bar{d}_2, \bar{d}_2 \rangle_{H_r} = \frac{1}{2} \langle d_3, d_3 \rangle_{H_r}$$

(ii) Structure problem : define $\bar{d}_1 = d_1 \in \overline{\mathcal{R}(W_s)}$ and $\bar{d}_2 = d_2 + d_3 \in \mathcal{N}(W_s^*)$

Apply theorem 2.1.1, part (iii) to obtain

$$\lim_{\rho \rightarrow 0} \text{IESQ}_s(\rho) = \frac{1}{2} \langle \bar{d}_2, \bar{d}_2 \rangle_{H_r} = \frac{1}{2} \langle d_2, d_2 \rangle_{H_r} + \frac{1}{2} \langle d_3, d_3 \rangle_{H_r}$$

$$(iii) \quad \lim_{\rho \rightarrow 0} (\text{IESQ}_s(\rho) - \text{IESQ}_f(\rho)) = \lim_{\rho \rightarrow 0} \text{IESQ}_s(\rho) - \lim_{\rho \rightarrow 0} \text{IESQ}_f(\rho)$$

$$\text{Hence} \quad \lim_{\rho \rightarrow 0} (\text{IESQ}_s(\rho) - \text{IESQ}_f(\rho)) = \frac{1}{2} \langle d_2, d_2 \rangle_{H_r}$$

4

Immediately observe that asymptotically, structure assessment by either diagnostic is based on the same quantity, namely $\frac{1}{2} \langle d_2, d_2 \rangle_{H_r}$. Also that the situation of a structure being asymptotically as accurate as the full system either through disturbance choice, or more fundamentally through the sufficiency condition $R(W) = R(W_S)$ will be indicated. For those structures where $d_2 \neq 0$ asymptotic consistency is shown as follows :

$$\text{If } \lim_{\rho \rightarrow 0} \text{IESQ}_S(\rho) - \text{IESQ}_f(\rho) = \frac{1}{2} \langle d_2, d_2 \rangle_{H_r} > 0$$

then by definition for any given $\eta > 0$ it is possible to find

$$\rho_{\max}(\eta) > 0 \quad \text{such that whenever } 0 < \rho \leq \rho_{\max}$$

$$\frac{1}{2} \langle d_2, d_2 \rangle_{H_r} - \eta < (\text{IESQ}_S(\rho) - \text{IESQ}_f(\rho)) < \frac{1}{2} \langle d_2, d_2 \rangle_{H_r} + \eta$$

Choose $\eta = \frac{1}{2} \langle d_2, d_2 \rangle_{H_r}$ then $\exists \rho_{\max}(\eta)$ such that

$$\text{IESQ}_S(\rho) - \text{IESQ}_f(\rho) > 0 \quad \text{whenever } \rho < \rho_{\max}.$$

One of the appealing features of the Hilbert space formalism used here is the geometrical analogy which can be drawn for the results. The salient features of this geometry are the orthogonal decomposition of the output space H_r , recalled as

$$H_r = \overline{R(W_S)} \oplus X \oplus N(W^*) \quad \text{where} \quad \overline{R(W)} = \overline{R(W_S)} \oplus X,$$

which enables a disturbance d to be represented as $d = d_1 + d_2 + d_3$

with $d_1 \in \overline{R(W_S)}$, $d_2 \in X$ and $d_3 \in N(W^*)$, and the two limits

describing the effect of applying optimal control, namely

$$\lim_{\rho \rightarrow 0} \hat{Z}(\rho) = -(d_1 + d_2) \quad \text{and} \quad \lim_{\rho \rightarrow 0} \hat{Z}_S(\rho) = -d_1$$

where $\hat{Z}(\rho) \triangleq W \hat{u}(\rho)$ and $\hat{Z}_S(\rho) = W_S \hat{u}_S(\rho)$. Using the linear

space E^3 to represent the Hilbert space of output H_r , the mutually orthogonal axes x, y , and z may be associated with the orthogonal

subspaces $\overline{\mathcal{R}(W_S)}$, X , and $\mathcal{N}(W^*)$ respectively. So that

$\overline{\mathcal{R}(W)} = \overline{\mathcal{R}(W_S)} \oplus X$ is represented by the entire (x,y) plane and the disturbance decomposition is represented by the three vectors d_1, d_2, d_3 as shown in Figure 2.2.1. The trajectories of $\hat{Z}(\rho)$ and $\hat{Z}_S(\rho)$ as functions of ρ are also depicted : $\hat{Z}(\rho) \in \mathcal{R}(W)$ hence its trajectory is seen as lying in the (x,y) plane limiting, as $\rho \rightarrow 0$, at

$-\bar{d}_1 = -(d_1 + d_2)$ and $\hat{Z}_S(\rho) \in \mathcal{R}(W_S)$ so that it is seen as lying in the x axis limiting, as $\rho \rightarrow 0$, at $-d_1$. The relationship between the control action $\hat{Z}(\rho)$ and the resulting output $\hat{y}(\rho) = Z(\rho) + d$ is shown in Figure 2.2.2. In Figure 2.2.3 a similar relationship for a structure control effect $\hat{Z}_S(\rho)$ and the resulting output $\hat{y}_S(\rho) = \hat{Z}_S(\rho) + d$ is shown. Also on this figure is the limiting value of the full optimal output, namely $\lim_{\rho \rightarrow 0} \hat{y}(\rho) = d_3$, which

is the best output vector obtainable. It is clear, for sufficiently small ρ , that the limiting values of $\hat{y}(\rho)$ and $\hat{y}_S(\rho)$ will only be close if the component d_2 is small. This then is the basis of the structures assessment, a structure is sought for which the introduced structure dependent component is small so that the optimal structure output compares favourably with the full optimal output.

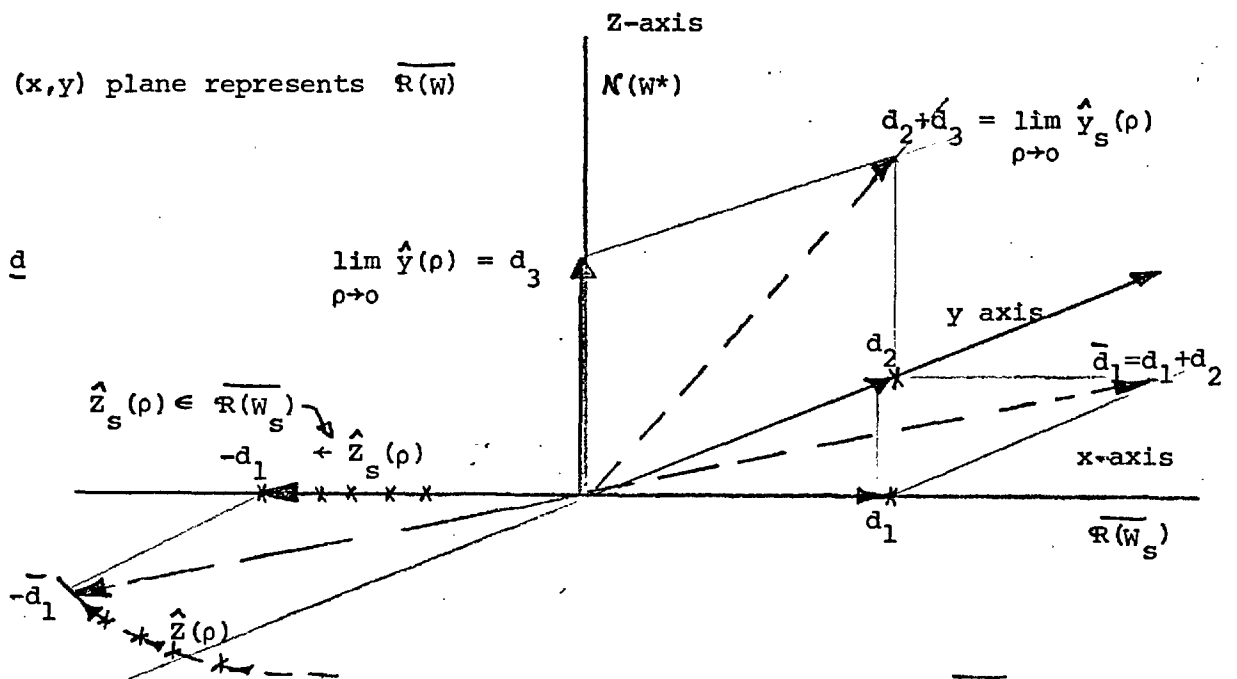


Fig 2.2.1 A Geometric Analogy for the Decomposition $H_r = \overline{\mathcal{R}(W_s)} \oplus X \oplus \mathcal{N}(W)$

with $\overline{\mathcal{R}(W)} = \overline{\mathcal{R}(W_s)} \oplus X$

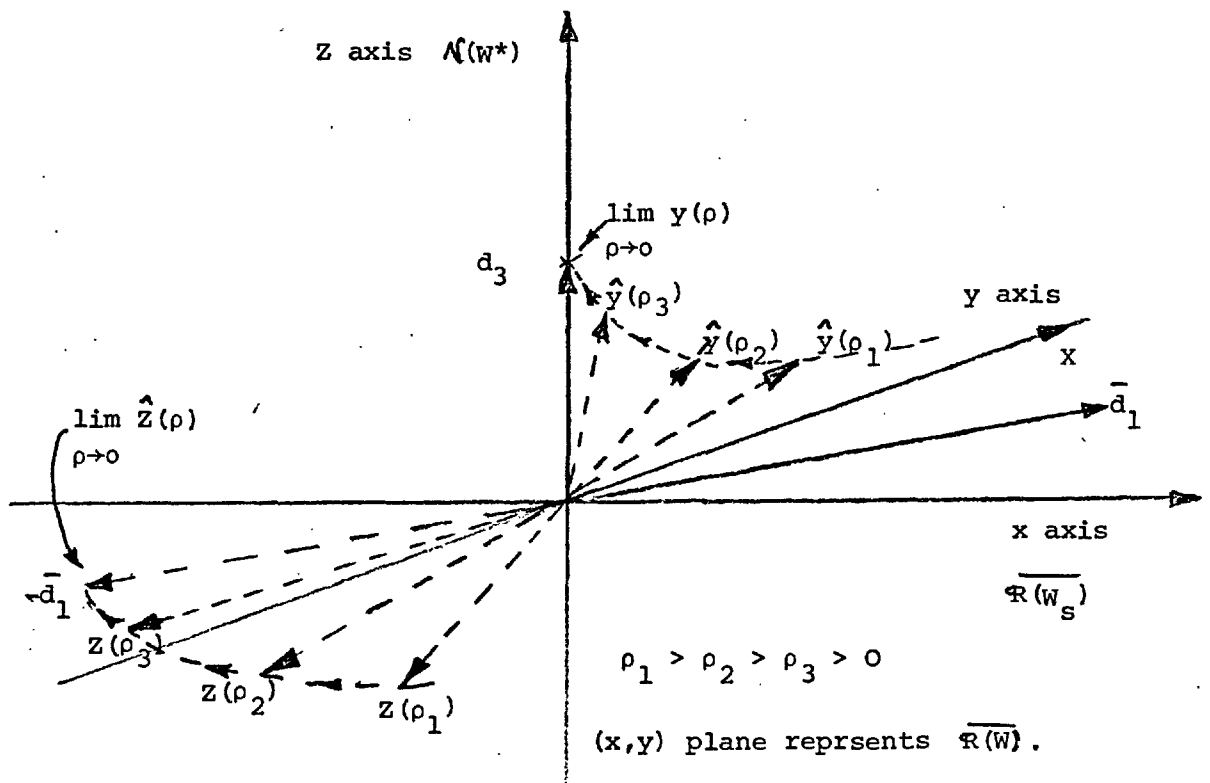
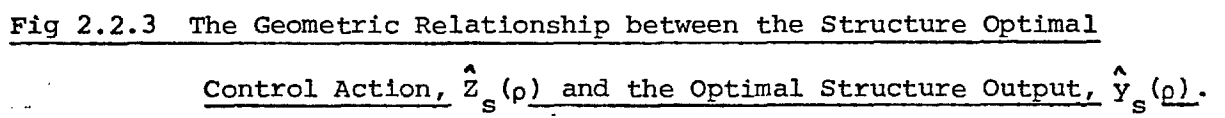


Fig 2.2.2 Relation between Optimal Control Action $\hat{z}(\rho)$ and

Optimal Output $\hat{y}(\rho)$



Having shown in this section that the diagnostics satisfy a consistency property and in the previous section that their values are indicative of the control properties of the system, it remains therefore to give some control interpretation to the quantity isolated by the assessment of structures. This is the quantity arising from the two limits :

$$\lim_{\rho \rightarrow 0} (\Delta J_O^f(\rho) - \Delta J_O^s(\rho)) = \lim_{\rho \rightarrow 0} (IESQ_s(\rho) - IESQ_f(\rho)) = \frac{1}{2} \langle d_2, d_2 \rangle_{H_r} \quad \text{where } d_2 \in X, \text{ and } \overline{R(W)} = \overline{R(W_s)} \oplus X. \text{ Its relationship with } \overline{R(W_s)}, X \text{ and } \overline{R(W)}$$

indicates that it represents an additional part of the disturbance vector which cannot be removed by the structure optimal control and is therefore a newly introduced obstacle to good control. (Its size giving some indication of its importance.) What then are the new control difficulties introduced by input deletion ? To answer this question a state space system description is utilized which is initially assumed completely controllable and observable. Input deletion is shown to effect the complete controllability of the system and to change the system zero pattern, both giving rise to features which are reflected in the diagnostics as new control difficulties.

2.3 Input Deletion and the Preservation of Controllability.

For a given system triple $\mathcal{S}(A,B,C)$ representing the time-invariant state space system, input deletions effectively nullify the columns of the input matrix B . Consequently if the full system is assumed completely controllable, the process of nullifying the columns of B will reach a point at which complete controllability is lost.

(Trivially the complete observability of the system remains unchanged.)

Input structures arise quite naturally since there are many differing combinations of columns which can be nulled. Dressler and Larson (1968) called systems which are not completely controllable, partially controllable systems. This situation is discussed at length in the next section, whilst here concern is focused on the conditions *sufficient* for the preservation of complete controllability and their interpretation in terms of the pole-zero concepts and the diagnostic values.

The first results concerning the conditions *sufficient* for the preservation of complete controllability are established in the more general setting of "squaring" a rectangular system as a prerequisite for controller design using the techniques of Rosenbrock or MacFarlane.

Kouvaritakis (1972) has discussed this problem and using a rank argument has prescribed that 1) in the case of more outputs than inputs ($r = \text{no. of outputs} > m = \text{no. of inputs}$) an output compensator should

be used to give tight control over the m new outputs by the m inputs, and ii) in the case of more inputs than outputs ($m > r$) an input precompensator should be used. In later work (Kouvaritakis and MacFarlane, 1976b) it was recognized that the process of squaring was effecting the system zero pattern to the extent of introducing new system zeros, and so algorithms were devised to place as many of the new zeros as possible. In the case of reduction of inputs (squaring down) this was achieved through a modifying matrix $\bar{B}$, so that $B_s = B\bar{B}$ is the new input matrix of dimensions (n, s) . Input deletion is a special case of squaring down with a deletion matrix $\bar{B} = [e_{j_1}, e_{j_2}, \dots, e_{j_s}]$ where $j_1 < j_2 < \dots < j_s$ and $e_{j_i} \in \mathbb{R}^m$. Consequently the first theorem, and its corollary use a general structure matrix B_s which could have been obtained by either process.

Theorem 2.3.1 If (A, B) form a completely controllable pair, then (A, B_s) are completely controllable if $\mathcal{R}(B_s) \equiv \mathcal{R}(B)$.

Proof) Show $\mathcal{R}(B_s) \equiv \mathcal{R}(B) \Rightarrow (A, B_s)$ completely controllable.

Denote $\mathcal{P} = \mathcal{R}(B)$ and $\mathcal{P}_s = \mathcal{R}(B_s)$

then $\mathcal{R}(B) \equiv \mathcal{R}(B_s) \Rightarrow \mathcal{P} = \mathcal{P}_s, A^i \mathcal{P} = A^i \mathcal{P}_s \quad i=1, \dots, n-1$

(A, B) completely controllable $\Rightarrow \text{rank} [B, AB, \dots, A^{n-1} B] = n$.

Hence $\forall Z \in \mathbb{R}^n \exists \gamma \in \mathbb{R}^{nm}$ such that

$$Z = B\gamma_0 + AB\gamma_1 + \dots + A^{n-1}B\gamma_{n-1} \quad \gamma_i \in \mathbb{R}^m \quad i=0, 1, \dots, n-1$$

But $\mathcal{R}(B) = \mathcal{R}(B_s) \Rightarrow \exists \delta_i \in \mathbb{R}^s \text{ s.t. } B\gamma_i = B_s \delta_i$

$\forall \gamma_i$, and $i=0, 1, \dots, n-1$

$$\Rightarrow Z = B_s \delta_0 + AB_s \delta_1 + \dots + A^{n-1}B_s \delta_{n-1}$$

Since this is true $\forall Z \in \mathbb{R}^n \Rightarrow \text{rank} [B_s, AB_s, \dots, A^{n-1}B_s] = n$

$\Rightarrow (A, B_S)$ completely controllable. △

Squaring down or input deletion always leads to the condition $\mathcal{R}(B_S) \subseteq \mathcal{R}(B)$. As already seen, complete controllability is preserved under equality of the range spaces $\mathcal{R}(B)$ and $\mathcal{R}(B_S)$. However, the alternative condition $\mathcal{R}(B_S) = \mathcal{R}(B)$ does not always imply a loss of complete controllability. For the case of input deletion this can be more readily appreciated by rewriting the controllability matrix to emphasize the cyclic subspaces generated by the columns of the input matrix:

$$[B, AB, \dots, A^{n-1}b] = [b_1, Ab_1, A^2b_1, \dots, A^{n-1}b_1, b_2, Ab_2, \dots, A^{n-1}b_2, \dots, b_m, Ab_m, \dots, A^{n-1}b_m].$$

Loss of controllability (for deletion) is therefore dependant on whether the columns of the structure input matrix B_S can generate a set of cyclic subspaces which span the whole space, $\mathbb{R}^n$.

The following corollary details the changes to the zero pattern under the assumption $\mathcal{R}(B) = \mathcal{R}(B_S)$.

Corollary 2.3.1 If a system's input matrix B is modified to B_S whilst preserving range spaces, then no new system zeros are introduced.

Proof If the range spaces are preserved, namely $\mathcal{R}(B) = \mathcal{R}(B_S)$, then the input matrices B and B_S have the following rank decompositions :

$$B = P \begin{bmatrix} I_{\bar{m}} & 0_{\bar{m} \times m - \bar{m}} \\ \hline 0 & 0 \end{bmatrix} Q_m \Rightarrow B_s = P \begin{bmatrix} I_{\bar{m}} & 0_{\bar{m} \times (s - \bar{m})} \\ \hline 0 & 0 \end{bmatrix} Q_s$$

where P , Q_m and Q_s are the usual non-singular matrices, and $\bar{m} \leq m$ is the common rank. This result enables the statement on the system zeros to be proved.

1) Matrix B is given the rank decomposition

$$B = P \begin{bmatrix} I_{\bar{m}} & 0_{\bar{m} \times (m - \bar{m})} \\ \hline 0 & 0 \end{bmatrix} Q_m. \quad \text{Write } B_s = B\bar{B} \text{ where matrix } \bar{B} \text{ is}$$

$(m \times s)$ and has rank $s \geq m$. Give $\bar{B}$ the rank decomposition

$$\bar{B} = \bar{Q}_m \begin{bmatrix} I_s \\ \hline 0 \end{bmatrix} \bar{P}_s \text{ and form } B_s :$$

$$B_s = B\bar{B} = P \begin{bmatrix} I_{\bar{m}} & 0_{\bar{m} \times (m - \bar{m})} \\ \hline 0 & 0 \end{bmatrix} Q_m \bar{Q}_m \begin{bmatrix} I_s \\ \hline 0 \end{bmatrix} \bar{P}_s$$

Under the assumption of preservation of ~~range spaces~~,

$\text{rank}(B_s) = \text{rank}(B) = \bar{m}$, and since P , and $\bar{P}_s$ are non-singular

$$\begin{bmatrix} I_{\bar{m}} & 0_{\bar{m} \times m - \bar{m}} \\ \hline 0 & 0 \end{bmatrix} Q_m \bar{Q}_m \begin{bmatrix} I_s \\ \hline 0 \end{bmatrix} \text{ has rank } \bar{m}.$$

$$\text{Multiplying out obtain } \begin{bmatrix} I_{\bar{m}} & 0_{\bar{m} \times (m - \bar{m})} \\ \hline 0 & 0 \end{bmatrix} Q_m \bar{Q}_m \begin{bmatrix} I_s \\ \hline 0 \end{bmatrix} = \begin{bmatrix} R \\ \hline 0 \end{bmatrix}$$

where R is $(\bar{m} \times s)$ and is of rank $\bar{m}$. Since $s \geq \bar{m}$, column

$$\text{manipulations on } \begin{bmatrix} R \\ \hline 0 \end{bmatrix} \text{ can be used to produce } \begin{bmatrix} R \\ \hline 0 \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} \\ \hline 0 & 0 \end{bmatrix} R_s$$

where R_{11} is non-singular and R_s is the non-singular matrix manipulating columns.

$$\text{Finally } \begin{bmatrix} R \\ \hline 0 \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} \\ \hline 0 & 0 \end{bmatrix} R_s = \begin{bmatrix} I_{\bar{m}} & 0_{\bar{m} \times (s - \bar{m})} \\ \hline 0 & 0 \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ \hline 0 & I \end{bmatrix} R_s$$

gives $B_s = P \begin{bmatrix} I_{\bar{m}} & 0_{\bar{m}} \times (s-\bar{m}) \\ 0 & 0 \end{bmatrix} Q_s$ where P and Q_s are non-singular.

ii) The rank decompositions for B and B_s can be used to prove the result on system zeros.

Recall $G(s) = C (sI - A)^{-1} B$ as the system transfer matrix and

$$\bar{G}(s) = G(s) \bar{B} = C (sI - A)^{-1} B \bar{B} = C (sI - A)^{-1} B_s$$

as the structured system transfer function matrix. Writing the kernel

of rank decompositions as $\Gamma_{\bar{m}} \triangleq \begin{bmatrix} I_{\bar{m}} & 0_{\bar{m}} \times (m-\bar{m}) \\ 0 & 0 \end{bmatrix}$

then $B = P \Gamma_{\bar{m}} Q$ and $B_s = P \Gamma_{\bar{m}} Q_s$. Substituting into the respective system transfers obtain

$$G(s) = C (sI - A)^{-1} P \Gamma_{\bar{m}} Q \quad \text{and} \quad C (sI - A)^{-1} P \Gamma_{\bar{m}} Q_s$$

Writing $T_1(s) \psi(s) T_2(s) = C (sI - A)^{-1} P \Gamma_{\bar{m}}$, where $\psi(s)$ is the usual Smith McMillan kernel, and $T_1(s)$ and $T_2(s)$ are unimodular matrices.

Noting that matrices Q and Q_s are both nonsingular and that complete controllability has been preserved, $G(s) = T_1(s) \psi(s) T_2(s) Q$ and $\bar{G}(s) = T_1(s) \psi(s) T_2(s) Q_s$ have the same system zeros, that is the original system zeros. This proves the theorem.

The above treatment deals with modifications to the input matrix B , although duality ensures a similar result holds for modifications of the output matrix C . The immediate consequence for the problem of squaring a rectangular system is that ~~if~~ rank preserving compensators ^{will not} are used decoupling zeros ^{appear}, a fact which does not appear to have been stated in the literature of the squaring problem.

Returning to the consideration of input deletion, if complete controllability has been preserved, and consequently no new zeros appear it is expected asymptotically that control sets u_s and u_f will be equally efficient, since no new control difficulties have been introduced by the structure.

In fact it is quite straightforward to justify this since

$\mathcal{R}(B_s) \equiv \mathcal{R}(B)$ implies for the state space system operator
 $(Wu)(t) \triangleq \int_0^t C e^{A(t-\tau)} B u(\tau) d\tau$ that $\mathcal{R}(W) \equiv \mathcal{R}(W_s)$

which has already been noted as a sufficient condition for

$$\lim_{\rho \rightarrow 0} \Delta J_O^f(\rho) - \Delta J_O^s(\rho) = 0 \quad \text{and} \quad \lim_{\rho \rightarrow 0} IESQ_s(\rho) - IESQ_f(\rho) = 0.$$

2.4 Input Deletion and the Loss of Controllability

In the previous section, the general results (Theorem 2.3.1, and Corollary 2.3.1) indicated that if the process of squaring down a system was achieved through the use of a reduced rank input matrix (that is $\text{rank}(B_S = B\bar{B}) < \text{rank}(B)$), then complete controllability of the system ^{may be} lost. (A dual of this result holds for squaring up a system using an output compensator of the form $C_S = \bar{C}C$.)

This section is concerned with the effect of this loss of complete controllability. The initial discussion deals with the more general case of squaring down a rectangular system. It is then shown how input deletion is a far more radical procedure, and finally Kalman's Canonical form is used to demonstrate the registering of these effects in the structure diagnostics.

The assumption is made that system $\mathcal{S}(A, B, C)$ has the following qualities. The numbers of outputs, inputs and states satisfy $r < m < n$. The output matrix is assumed to attain its row rank r , and the input matrix its column rank, m . Input matrix B is then squared down using an input precompensator $\bar{B}$ such that $\text{rank}(B_S = B\bar{B}) = \bar{r}$ where $r < \bar{r} < m$. Using the Smith-McMillan form of the original system, the corresponding form is to be produced for the squared down system. Such a procedure demonstrates the modifications being made on the original zero pattern.

System $\mathcal{S}(A, B, C)$, assumed completely controllable and observable, may be written in transfer function form as $G(s) = C(sI - A)^{-1}B = N(s)/d(s)$, where $N(s)$ is a polynomial matrix. The corresponding Smith form for $N(s)$ is given as $N(s) = P(s)\Gamma(s)Q(s)$ where $P(s)$ and $Q(s)$ are unimodular matrices. From this $G(s)$ may be written in Smith-McMillan form as $G(s) = \frac{P(s)\Gamma(s)Q(s)}{d(s)} = P(s)\tilde{\Gamma}(s)Q(s)$

$$\text{where } \tilde{\Gamma}(s) = \begin{bmatrix} \sigma_1(s)/\psi_1(s) & & & \\ 0 & \sigma_2(s)/\psi_2(s) & & \\ & \ddots & \ddots & \\ & & \sigma_r(s)/\psi_r(s) & \\ & & & 0_{r \times (m-r)} \end{bmatrix}$$

Since complete controllability and observability are assumed, the system zeros are given by $\prod_{i=1}^r \sigma_i(s) = 0$, and the system poles by

$$\prod_{i=1}^r \psi_i(s) = 0. \quad (\text{MacFarlane and Karcanas 1976}).$$

(note that $\sigma_1(s) | \sigma_2(s) | \dots | \sigma_r(s)$ and $\psi_r(s) | \psi_{r-1}(s) | \dots | \psi_1(s)$.)

The transfer of the squared down system triple $\mathcal{S}(A, B_s, C)$ is

$$\bar{G}(s) = C(sI - A)^{-1}B\bar{B} = N(s)\bar{B}/d(s) = \bar{N}(s)/d(s).$$

To be able to exhibit the new zeros, the Smith form for the numerator $\bar{N}(s)$ is found :

i) The invariants of $\bar{N}(s) = P(s)\Gamma(s)Q(s)\bar{B}$ are the same as those of

$$\bar{N}_1(s) = \Gamma(s)Q(s)\bar{B}. \quad (P(s) \text{ is unimodular})$$

ii) Assume the full rank decomposition for $\bar{B}$,

$$\bar{B} = L \begin{bmatrix} I_r \\ 0 \end{bmatrix} M \text{ where matrices } L, \text{ and } M \text{ are nonsingular.}$$

Hence $\bar{N}_1(s) = \Gamma(s)Q(s)L \begin{bmatrix} I_{\bar{r}} \\ 0 \end{bmatrix} M$, and the invariants of $N_1(s)$

are the same as those of $\bar{N}_2(s) = \Gamma(s)Q(s)L \begin{bmatrix} I_{\bar{r}} \\ 0 \end{bmatrix}$.

iii) Write $\bar{Q}(s) = Q(s)L$, and partition $\bar{Q}(s)$ to obtain

$$\bar{N}_2(s) = \Gamma(s)\bar{Q} \begin{bmatrix} I_{\bar{r}} \\ 0 \end{bmatrix} = \begin{bmatrix} \Gamma_r(s) & 0 \\ & \vdots \end{bmatrix} \begin{bmatrix} \bar{Q}_{11}(s) & \bar{Q}_{12}(s) \\ \bar{Q}_{21}(s) & \bar{Q}_{22}(s) \end{bmatrix} \begin{bmatrix} I_{\bar{r}} \\ 0 \end{bmatrix} = \Gamma_r(s)\bar{Q}_{11}(s)$$

iv) The invariants of $N_2(s)$ are found using the determinantal divisors, and by exploiting the structures of $\bar{N}_2(s)$, namely

$$\begin{aligned} [\bar{N}_2(s)]_{ij} &= \Gamma_i(s) [\bar{Q}_{11}(s)]_{ij} \quad i=1, \dots, r \\ &\quad j=1, \dots, \bar{r} \end{aligned}$$

kth order determinantal divisors are then given for $k=1, \dots, r$ as:-

2.4.1

$\Delta_1 \triangleq$ greatest common divisor of all minors of order 1

$= \Gamma_1(s) \cdot f_1(s)$, since $\Gamma_1(s) | \Gamma_2(s) | \dots | \Gamma_r(s)$ (property of invariants) and all the elements of $[\bar{Q}_{11}(s)]_{ij}$ may have a common factor.

$\Delta_2 \triangleq$ greatest common divisor of all minors of order 2

$= \Gamma_1(s)\Gamma_2(s)f_1^2(s)f_2(s)$, where $f_2(s)$ is a common factor to all second order minors of $\bar{N}_2(s)$ after $\Gamma_1(s)\Gamma_2(s)f_1^2(s)$ has been removed

Continuing up to Δ_r :

$\Delta_r \triangleq$ greatest common divisor of all minors of order r .

$= \Gamma_1(s)\Gamma_2(s)\dots\Gamma_r(s)f_1^r(s)f_2^{r-1}(s)\dots f_r(s)$, where $f_r(s)$

is a common factor to all r th order minors of $\bar{N}_2(s)$ after

$\Gamma_1(s)\dots\Gamma_r(s)f_1^r(s)\dots f_{r-1}^2(s)$ has been removed.

The invariants are therefore found as:

$$\bar{\Gamma}_1(s) = \Delta_1(s) = \Gamma_1(s)f_1(s)$$

$$\bar{\Gamma}_2(s) = \Delta_2(s)/\Delta_1(s) = \Gamma_2(s)f_1(s)f_2(s)$$

$$\bar{\Gamma}_r(s) = \Delta_r(s)/\Delta_{r-1}(s) = \Gamma_r(s)f_1(s)\dots f_r(s)$$

The new Smith form is therefore

$$\bar{N}(s) = \bar{P}(s) \begin{bmatrix} \Gamma_1(s)f_1(s) & & \\ & \ddots & \\ & & \Gamma_r(s)f_1(s)\dots f_r(s) \end{bmatrix} \bar{Q}(s)$$

from which the squared down Smith McMillan form is

$$\bar{G}(s) = \bar{P}(s) \begin{bmatrix} \sigma_1(s)f_1(s)/\psi_1(s) & & \\ & \ddots & \\ & & \sigma_r(s)f_1(s)\dots f_r(s)/\psi_r(s) \end{bmatrix} \begin{matrix} \\ \\ \\ \end{matrix} \bar{Q}(s)$$

Clearly three observations may be made :

- 1) The initial assumption of $\text{rank}(B_s) < \text{rank}(B) \Rightarrow$ possible loss of controllability. Hence some of the factors $f_1(s), \dots, f_r(s)$ would supply input decoupling zeros to cancel with some of the system poles contained in the $\psi_i(s)$, $i = 1, \dots, r$.
- 2) The original system zeros contained in the terms $\sigma_i(s)$, $i=1, \dots, r$ are preserved
- and 3) Terms $f_i(s)$, $\dots$, $f_r(s)$ may well supply new transmission zeros, although the analysis given does not provide explicit proof whether or not such new zeros appear.

The deletion of inputs to form structures is a more radical process than squaring down a rectangular system, since eventually the normal rank of the transfer will be reduced. The effect is to remove whole terms from the kernel of Smith McMillan form of the deleted system. This is illustrated by the previous analysis, for if s is the size of the structure index set, determinantal divisors (2.4.1) will then be obtained for $k=1, \dots, r$ when $s \geq r$ and for $k=1, 2, \dots, s$ where $s < r$. This latter case is where $\bar{G}(s)$ has lost normal rank

(from r down to s). Giving a transfer of the form

$$\bar{G}(s) = \bar{P}(s) \begin{bmatrix} \alpha_1(s)f_1(s)/\psi_1(s) \\ \vdots \\ \alpha_s(s)f_1(s)\dots f_s(s)/\psi_s(s) \\ \vdots \\ 0_{(r-s), s} \end{bmatrix} \bar{Q}(s) \quad (s < r)$$

from which it can be seen that some system zeros and poles have been completely removed. The details are encapsulated in the following

Theorem 2.4.1 Zeros of a Deleted System.

For completely controllable and observable system triple $\mathcal{S}(A, B, C)$

assume $r, \bar{r}$, and m (the number of outputs, the rank of the input matrix B and the number of inputs respectively) satisfy $r < \bar{r} < m$.

If s is the structure index size then:

- i) $r \leq s < \bar{r} \Rightarrow$ a) ^{Possible} Loss of controllability and introduction of decoupling zeros.
 - b) The transmission zeros of the system are preserved
 - c) New transmission zeros are possibly introduced
- ii) $s < r \Rightarrow$ a) ^{Possible} Loss of controllability and introduction of decoupling zeros.
 - b) Loss of normal rank of transfer $\Rightarrow$ loss of some original and new transmission zeros.

Proof Recall that the zeros of the original system $\mathcal{S}(A, B, C)$

are defined as the zeros of $\phi(s)$, the monic greatest divisor of all

$$\text{minors: } m(i_1, \dots, i_r) = \begin{vmatrix} sI - A & B_r \\ -C & 0 \end{vmatrix}$$

$$\text{for all } B_r = [b_{i_1}, b_{i_2}, \dots, b_{i_r}] \subset B = [b_1, b_2, \dots, b_m]$$

(read: the columns of B_r are a subset of those in matrix B).

- i) $r \leq s < \bar{r} \Rightarrow$ a) $R(B_s) \neq R(B) \Rightarrow$ ^{Possible} Loss of controllability (by Theorem 2.3.1).

Appeal to Rosenbrock, 1970 for the number of decoupling zeros introduced as the rank defect of the controllability matrix, $[B_s, AB_s, \dots, A^{n-1}B_s]$.

The zeros of the structured system $S'_s(A, B, C)$ are defined as the zeros of $\phi_s(s)$, the monic greatest divisor of all minors of

$$P_s(s) = \begin{bmatrix} sI - A & B_s(i_1, \dots, i_r) \\ -C & 0 \end{bmatrix} \quad \text{for all}$$

$$B_s(i_1, i_2, \dots, i_r) \subset B_s = [b_{j_1}, b_{j_2}, \dots, b_{j_s}]$$

a) Let $s=Z$ be a transmission zero of the original system

$$\Rightarrow (s-Z) \mid \det [C(sI-A)^{-1}B_r] \quad \forall B_r \subset B \text{ in particular}$$

$$(s-Z) \mid \det [C(sI-A)^{-1}B_r] \quad \forall B_r \subset B_s$$

$\Rightarrow s=Z$ is a transmission zero of $S'_s(A, B_s, C)$.

b) Let $s=Z$ be a transmission zero of a structured system

$$S'_s(A, B_s, C) \Rightarrow (s-Z) \mid \det [C(sI-A)^{-1}B_r] \quad \forall B_r \subset B_s$$

Since $B_s \subset B$, this does not necessarily imply

$$(s-Z) \mid \det [C(sI-A)^{-1}B_r] \quad \forall B_r \subset B;$$

consequently new transmission zeros may appear.

ii) $s < r$. a) The ^{possible} loss of controllability occurs as before because

$\mathcal{R}(B_s) \subset \mathcal{R}(B)$ but $\mathcal{R}(B_s) \neq \mathcal{R}(B)$, consequently input decoupling zeros may be introduced.

b) The transfer function matrix is given as $G(s) = C(sI-A)^{-1}B$, for

which the normal rank was assumed to be the row rank, r . When input deletion occurs so that $s < r \Rightarrow$ normal rank $= s$. This reduction in normal rank, removes transmission zeros of the original system, and new transmission zeros depending on their order.

The zeros of the reduced rank system are given as the zeros of $\phi_s(s)$

the monic greatest divisor of all the minors of $P_s(s) = \begin{bmatrix} sI-A & B_s \\ C_s & 0 \end{bmatrix}$

for all

$$C_s = \begin{bmatrix} C_{i_1} \\ \vdots \\ C_{i_s} \end{bmatrix} \subset C = \begin{bmatrix} C_1 \\ \vdots \\ C_r \end{bmatrix} \quad (s \leq r)$$

Let $(s-Z)$ be a divisor of all system matrices

$$P_s(s) = \begin{bmatrix} sI-A & B_s \\ C_s & 0 \end{bmatrix} \quad \forall B_s \subset B_r \text{ and } C_s \subset C$$

then, $(s-Z) \mid \det [C_s (sI-A)^{-1} B_s] \quad \forall C_s \subset C$ and $s=Z$ is a transmission zero of the reduced rank structure.

Also $(s-Z) \mid \det [C(sI-A)^{-1} B_r]$ for a given B_r , hence $(s-Z)$ is a transmission zero of the system $\mathcal{S}(A, B_r, C)$.

Suppose $(s-Z_1)$ is a divisor of all system matrices

$$P_{s+1}(s) = \begin{bmatrix} sI-A & B_{s+1} \\ C_{s+1} & 0 \end{bmatrix} \quad \forall B_{s+1} \subset B_r \text{ and } C_{s+1} \subset C$$

but not a divisor of $P_s(s)$, then $(s-Z_1)$ is not a transmission

zero of $\mathcal{S}(A, B_s, C)$ and yet $(s-Z_1) \mid \det [C(sI-A)^{-1} B_r]$ giving that

$s=Z_1$ is a transmission zero of $\mathcal{S}(A, B_r, C)$. The situation is there-

fore, when $s=r$, and the system is square it will have a set of trans-

mission zeros comprising of the original system zeros and new zeros

introduced by the squaring down. But as the systems normal rank is reduced

$s < r$, some of these zeros are deleted depending on their order. This

concludes the proof of the theorem.

A fundamental feature of both squaring a rectangular system, or deletion whilst preserving the normal rank of the transfer function is the preservation of the original system transmission zeros. Consequently if the initial zero distribution is poor (containing a right half plane zero for example) this will remain unaffected by squaring or deletion. The only cure is, therefore, the selection of new input, output connections having a better zero pattern (Rosenbrock, 1966).

The appearance of input decoupling zeros or equivalently the loss of complete controllability enables recourse to the Kalman Canonical Form to give an interpretation of the effect caused in the input structure diagnostics IESQ and ΔJ_0 . Using a suitable transformation the loss of complete controllability (but the preservation of complete observability) may be exhibited in the structured system as:

$$\begin{bmatrix} \dot{z}_1(t) \\ \dot{z}_2(t) \end{bmatrix} = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ 0 & \bar{A}_{22} \end{bmatrix} \begin{bmatrix} z_1(t) \\ z_2(t) \end{bmatrix} + \begin{bmatrix} \bar{B}_1 \\ 0 \end{bmatrix} u_s ; \begin{bmatrix} z_1(0) \\ z_2(0) \end{bmatrix} \text{ given}$$

$$y(t) = \begin{bmatrix} \bar{C}_1 & \bar{C}_2 \end{bmatrix} \begin{bmatrix} z_1(t) \\ z_2(t) \end{bmatrix} = \bar{C}_1 z_1(t) + \bar{C}_2 z_2(t) = y_1(t) + y_2(t)$$

where $z_1(t) \in \mathbb{R}^{n_1}$ and $z_2(t) \in \mathbb{R}^{n_2}$, $n_1 + n_2 = n$.

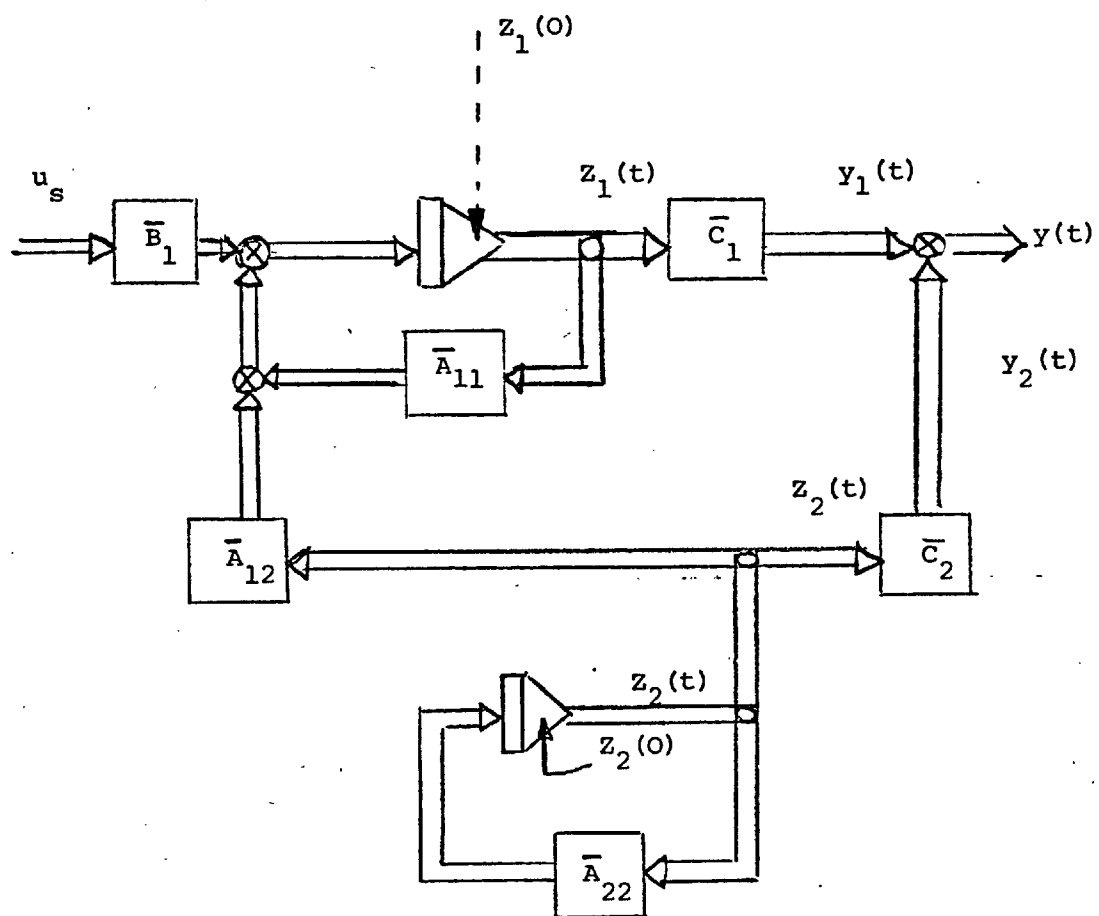


Fig 2.4.1 Open Loop Structured System

The infinite time interval optimal closed loop is specified in the usual manner through the steady state matrix Riccati equation, details of which follow:

The optimal feedback law is $\hat{u}_s = -R_s^{-1} [\bar{B}_1^T, 0] P \begin{bmatrix} Z_1(t) \\ Z_2(t) \end{bmatrix}$

where P is the symmetric, positive definite solution of the equation $\bar{A}^T P + P \bar{A} - P \bar{B}_s R_s^{-1} \bar{B}_s^T P - \bar{C}^T Q \bar{C} = 0$.

Partition P as $P = \begin{bmatrix} P_{11} & P_2 \\ P_2^T & P_3 \end{bmatrix}$ to give the optimal

closed loop system as:

$$\begin{bmatrix} \dot{Z}_1(t) \\ \dot{Z}_2(t) \end{bmatrix} = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} Z_1(t) \\ Z_2(t) \end{bmatrix} + \begin{bmatrix} \bar{B}_1 \\ 0 \end{bmatrix} [-R_s^{-1} [\bar{B}_1^T, 0]] \begin{bmatrix} P_{11} & P_2 \\ P_2^T & P_3 \end{bmatrix} \begin{bmatrix} Z_1(t) \\ Z_2(t) \end{bmatrix}$$

namely, $\begin{bmatrix} \dot{Z}_1(t) \\ \dot{Z}_2(t) \end{bmatrix} = \begin{bmatrix} \bar{A}_{11} - \bar{B}_1 R_s^{-1} \bar{B}_1^T P_{11} & \bar{A}_{12} - \bar{B}_1 R_s^{-1} \bar{B}_1^T P_2 \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} Z_1(t) \\ Z_2(t) \end{bmatrix}$

The closed loop poles are given by

$$\begin{aligned} \det[sI - \bar{A}_{CL}] &= \det[sI - (\bar{A}_{11} - \bar{B}_1 R_s^{-1} \bar{B}_1^T P_{11})] \det[sI - A_{22}] \\ &= \det[sI - \Lambda_{CL}] \det[sI - \Lambda_{22}] \end{aligned}$$

where Λ_{CL} contains the closed loop pole positions associated with the closed loop matrix $(\bar{A}_{11} - \bar{B}_1 R_s^{-1} \bar{B}_1^T P_{11})$ and Λ_{22} contains the pole positions of the uncontrolled subsystem.

Furthermore, it can be shown that P_1 satisfies a steady state

matrix Riccati equation: $\bar{A}_{11}^T P_1 + P_1 \bar{A}_{11} - P_1 \bar{B}_1 R_s^{-1} \bar{B}_1^T P_1 + \bar{C}_1^T Q \bar{C}_1 = 0$

This has the important implication that the behaviour of the closed looppoles contained in Λ_{CL} is determined by the asymptotic optimal control theory developed in Chapter 1. Presuppose for the closed loop a decomposition of the form

$$T^{-1}A_{CL}T = \begin{bmatrix} \Lambda_{CL}(\rho) & 0 \\ 0 & \Lambda_{22} \end{bmatrix}, \text{ then } \begin{bmatrix} \dot{Z}_1(t) \\ \dot{Z}_2(t) \end{bmatrix} = A_{CL} \begin{bmatrix} Z_1(t) \\ Z_2(t) \end{bmatrix}$$

becomes

$$\begin{bmatrix} \dot{w}_1(t) \\ \dot{w}_2(t) \end{bmatrix} = \begin{bmatrix} \Lambda_{CL}(\rho) & 0 \\ 0 & \Lambda_{22} \end{bmatrix} \begin{bmatrix} w_1(t) \\ w_2(t) \end{bmatrix}; \begin{bmatrix} w_1(0) \\ w_2(0) \end{bmatrix} = T^{-1} \begin{bmatrix} Z_1(0) \\ Z_2(0) \end{bmatrix}$$

$$\text{and } y(t) = \begin{bmatrix} \bar{C}_1 & \bar{C}_2 \end{bmatrix} \begin{bmatrix} Z_1(t) \\ Z_2(t) \end{bmatrix} = \begin{bmatrix} \tilde{C}_1 & \tilde{C}_2 \end{bmatrix} \begin{bmatrix} w_1(t) \\ w_2(t) \end{bmatrix} = \tilde{y}_1(t) + \tilde{y}_2(t)$$

where $\tilde{y}_1(t) = \tilde{C}_1 w_1(t)$ and $\tilde{y}_2(t) = \tilde{C}_2 w_2(t)$. Substitute into the diagnostic $IESQ_S(\rho)$ to obtain

$$IESQ_S(\rho) = \langle \hat{y}, Q\hat{y} \rangle_{H_T} = \frac{1}{2} \langle \tilde{y}_1, Q\tilde{y}_1 \rangle_{H_T} + \langle \tilde{y}_1, Q\tilde{y}_2 \rangle_{H_T} + \frac{1}{2} \langle \tilde{y}_2, Q\tilde{y}_2 \rangle_{H_T}.$$

Clearly the three terms may be related to three system features:

$\frac{1}{2} \langle \tilde{y}_1, Q\tilde{y}_1 \rangle \sim$ to the response of the controlled optimal subsystem,

$\frac{1}{2} \langle \tilde{y}_1, Q\tilde{y}_2 \rangle \sim$ to the interaction between the optimal subsystem and the uncontrolled subsystem,

and $\frac{1}{2} \langle \tilde{y}_2, Q\tilde{y}_2 \rangle \sim$ to the uncontrolled subsystem. The relationship

in corollary 2.1.1 between $\hat{J}_S(\rho)$ and $IESQ_S(\rho)$, namely

$$\hat{J}_S(\rho) = IESQ_S(\rho) + \frac{1}{2} \rho \langle \hat{u}_S(\rho), \hat{u}_S(\rho) \rangle_{H_M}$$

and the $\lim_{\rho \rightarrow 0} \frac{1}{2} \rho \langle \hat{u}_S(\rho), \hat{u}_S(\rho) \rangle_{H_M} = 0$ shows asymptotically that

these system features also dominate the value of $\hat{J}_s(\rho)$.

In section 2.2, it was shown that for sufficiently small $\rho > 0$, structure assessment based on the quantities $\Delta J_0^f - \Delta J_0^s$ or $IESQ_s - IESQ_f$ isolated the additional part of the disturbance that the structure optimal control was unable to remove. (Over and above that part of the disturbance which the full optimal control failed to remove.) In this and the previous section it has been detailed how such additional obstacles to control arise. They result from the introduction of new zeros: decoupling zeros signifying uncontrollable subsystems and possibly new transmission zeros which can restrict the effectiveness of the control action. The breakdown of the diagnostic $IESQ_s(\rho)$ (using Kalman's Canonical form) has also been seen to reflect these new control difficulties.

2.5 Summary of the Control Properties of Input Structure

Diagnostics

This chapter has been concerned with the asymptotic properties of the input structure diagnostics and the notion of consistency. Linear space concepts have been utilized to provide direct proofs and a geometrical insight to the results. Equivalent interpretation of these properties have been obtained through the use of a state space system description and an initial condition disturbance. The later parts of the chapter contain a detailed investigation of the effects of input deletion using the familiar concepts of controllability and the system zeros. A summary of the various results follow:

Asymptotic Diagnostic Values

By partitioning the output space as $H_r = \overline{R(W)} \oplus N(W^*)$ and the disturbance as $d = d_1 + d_2$ where $d_1 \in \overline{R(W)}$ and $d_2 \in N(W^*)$, the following limits were obtained:

$$\lim_{\rho \rightarrow 0} \hat{J}(\rho) = \lim_{\rho \rightarrow 0} \text{IESQ}(\rho) = \frac{1}{2} \langle d_2, d_2 \rangle_{H_r}, \quad \lim_{\rho \rightarrow 0} \Delta J_0(\rho) = \frac{1}{2} \langle d_1, d_1 \rangle_{H_r},$$

$$\text{for which it can also be shown } \lim_{\rho \rightarrow 0} (Wu(\rho) + d_1) = 0.$$

Asymptotically, the optimal control removed from the disturbance that part associated with $\overline{R(W)}$, so that the diagnostic $\text{IESQ}(\rho)$ represents (in the limit) that part of the cost $J(u=0)$ that the optimal control cannot remove, whilst $\Delta J_0(\rho)$ represents that part of the cost $J(u=0)$ that the optimal control has removed. Clearly $d_2 \in N(W^*) = \overline{R(W)}^\perp$ represents an obstacle to perfect control, or maximal accuracy.

A system was described as being maximally accurate with respect to a class of disturbances $\mathcal{C}$ if

$$\lim_{\rho \rightarrow 0} \text{IESQ}(\rho) = 0 \quad \forall d \in \mathcal{C} \subseteq H_r$$

Using a state space description and the class of initial condition disturbances, conditions for maximal accuracy were pursued.

These were found to be i) fewer outputs than controls and ii) no right half plane zeros. Finally the asymptotic theory of optimal control developed in Chapter 1 was used to give a control interpretation of this condition, namely that maximal accuracy implies i) asymptotic unobservability for those closed loop poles attracted by the left half plane zeros, ii) no right half plane zeros and iii) the remaining closed loop poles (over and above those attracted by zeros) become arbitrary fast.

Basic Relations for Input Structure Assessment

For the optimal cost improvement index, the notion of consistency was proved: $\Delta J_O^f - \Delta J_O^s = \frac{1}{2} \langle \tilde{u}, (W^*W + \rho I) \tilde{u} \rangle_{H_m}$, and asymptotically for the IESQ diagnostic as $\exists \rho_{\max}$ such that $\forall \rho < \rho_{\max} : \text{IESQ}^s(\rho) - \text{IESQ}^f(\rho) > 0$.

Again using the linear space concepts partition the output space as $H_r = \overline{R(W_s)} \oplus X \oplus N(W^*)$ where $\overline{R(W)} = \overline{R(W_s)} \oplus X$ and the disturbance as $d = d_1 + d_2 + d_3$ with $d_1 \in \overline{R(W_s)}$, $d_2 \in X$ and $d_3 \in N(W^*)$.

$$\text{Then } \lim_{\rho \rightarrow 0} (\Delta J_O^f(\rho) - \Delta J_O^s(\rho)) = \frac{1}{2} \langle d_2, d_2 \rangle = \lim_{\rho \rightarrow 0} (\text{IESQ}^s(\rho) - \text{IESQ}^f(\rho))$$

and $\lim_{\rho \rightarrow 0} (W_s \hat{u}(\rho) + d_1) = 0$. The component of the disturbance

$(d_2 + d_3)$ represents the obstacles to control for the deleted system W_s (since $(d_2 + d_3) \in \mathcal{N}(W_s^*)$) but component $d_3 \in \mathcal{N}(W^*)$ and may therefore be associated with the inherent control difficulties of the full system to leave the component d_2 to represent the additional control difficulties introduced by the deletion procedure. This then is the basis for input structure assessment: the comparison of additional control restrictions quantified by the diagnostic differences $(\Delta J_O^f - \Delta J_O^s)$ or $(IESQ_s(\rho) - IESQ_f(\rho))$ and the choosing of a satisfactory compromise between simplicity (few inputs) and suboptimality (small diagnostic difference).

Using a state space system description the remaining parts of the chapter were devoted to the control difficulties introduced by deletion.

Input Deletion and the Preservation of Controllability

A sufficient condition for the preservation of controllability was shown to be $\mathcal{R}(B_s) \equiv \mathcal{R}(B)$. It was also shown that if squaring down occurred so that $\mathcal{R}(B_s = \bar{B}\bar{B}) \equiv \mathcal{R}(B)$ then the original system zeros were preserved, and no new transmission zeros appeared. These quite general results were then specialized to the case of input deletion where it was shown

$\lim_{\rho \rightarrow 0} (\Delta J_O^f(\rho) - \Delta J_O^s(\rho)) = \lim_{\rho \rightarrow 0} (J_O^f(\rho) - J_O^s(\rho)) = 0$. From which it is obvious that under such conditions, asymptotically no new control difficulties arise.

Input Deletion and the Loss of Controllability

The discussion of deletion and the loss of complete controllability introduced a new feature, namely the loss of the normal rank of

the system matrix transfer function. This was caused by the consideration of two cases a) $r \leq s < \bar{r}$ in which there is a *possible* — loss of complete controllability and b) $s < r$ in which there is an additional loss of normal rank.

The full details are:

a) $r \leq s < \bar{r}$ i) loss of complete controllability *possible*
 ii) preservation of the original system zeros
 iii) introduction of new transmission zeros
possible

and b) $s < r$ i) loss of normal rank causing the loss of original system zeros and new transmission zeros originating from minors of order $> s$.
 ii) loss of complete controllability or equivalently the appearance of input decoupling zeros *possible*

The Kalman canonical form was utilized to demonstrate how the system structure was reflected in the diagnostic $IESQ(\rho)$. In the presence of an uncontrolled subsystem, the diagnostic comprised of three terms quantifying the parameters of the controlled subsystem (by appeal to the asymptotic optimal control theory), the parameters of the uncontrolled subsystem and their interaction. The decomposed system structure is shown in the figure below.

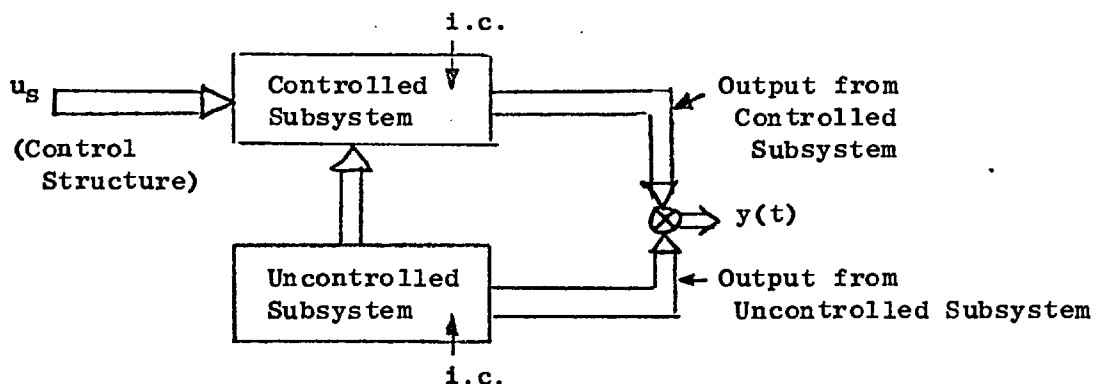


Fig. 2.5.1 Decomposed Structured System

Finally, two general observations made in the chapter. The first concerns the preservation of the original system zeros in the process of squaring down the number of input m to r . Here a poor initial system zero distribution (say right half plane zero) can only be modified by the radical move of taking new input/output sets, namely rechoosing matrices B , and C . Secondly if a system is squared down and the input matrix rank is reduced then an uncontrolled subsystem appears.

Chapter 3 Bounds for the Input Structure Diagnostics

3.0 Introduction

3.1 A Hilbert Space Canonical Form for Diagnostic, ΔJ_0 and the Formulation of a Reduced Order Optimization Problem.

3.2 Analytical Bound Formulae for the Diagnostic, ΔJ_0 (Cases $r=0,1,\dots,4$)

3.3 Analytical Bound Formulae for the Diagnostics, IESQ (Cases $r=0,1,\dots,4$).

3.4 Properties of the Moments and Bounds.

3.5 Asymptotic Properties of the Bound Formulae

3.6 A Finite Time Domain Example

3.7 Summary

Chapter 3 Bounds for the Input Structure Diagnostics

3.0 Introduction

The preceding two chapters have detailed the theoretical basis for the utility of the quantities ΔJ_0 , and IESQ as diagnostics of control structures obtained by input deletion. For the many possible structures it is clearly not possible to calculate all the corresponding diagnostic values (each one resulting from a Hilbert space optimization problem) so some form of simplified procedure is required. Following the work of Bryant (1973), the emphasis will be entirely on the calculation of bounds on the diagnostic values.

Using finite dimensional approximations for infinite dimensional operators, and linear programming theory, Bryant reduced the bound problem to that of solving a set of finite dimensional optimization problems. Here the spectral expansions for the diagnostics, and the moment constraints are interpreted as an infinite dimensional linear program. The reduction in dimensionality is obtained by a result similar to the fundamental theorem of linear programming (Luenberger, 1973). The subsequent development is concerned with providing explicit solutions to the resulting set of finite dimensional optimization problems. These are obtained using a technique of Hadley's (1970). Because of the similarities of the lengthy bound derivation for both diagnostics, only one is given in detail. Section 3.1 contains the theory for reduction in dimensionality of diagnostic ΔJ_0 , and the following section (3.2) the detailed derivation of its bounds for the finite dimensional problems of order $r=0,1,2,3$ and 4.

The important features of a similar analysis for the IESQ diagnostic are examined in Section 3.3, where the formulae for the low order cases ($r=0$, and 1) are given. Various properties of the bounds, and the moments are considered in Section 3.4. Permissible moment sequences are established, the behaviour of the moment ratio sequence examined and as far as possible the nesting of the bound formulae ascertained.

A discussion of the asymptotic properties of the bound formulae (Section 3.5) and some numerical examples (Section 3.6) for a single input, single state system conclude the chapter.

3.1 A Hilbert Space Canonical Form for ΔJ_0 , and the Formulation of a Reduced Order Optimization Problem.

The time-domain optimal control problem is posed as

$$\text{Min}_{u \in H_m} J = \frac{1}{2} \int_0^{T_f} \{ \langle y(t), Qy(t) \rangle_{E_r} + \langle u(t), Ru(t) \rangle_{E_m} \} dt$$

$$\text{with system equation } y(t) = \int_0^t W(t-\tau)u(\tau)d\tau + d(t), \quad 0 \leq t \leq T_f.$$

As before this may be written (using Hilbert space formalism) as

$$\text{Min}_{u \in H_m} J = \frac{1}{2} \langle y, Qy \rangle_{H_r} + \frac{1}{2} \langle u, Ru \rangle_{H_m}$$

with system equation $y = Wu + d$

where y , and $d \in H_r$; $u \in H_m$; $W: H_m \rightarrow H_r$, $R: H_m \rightarrow H_m$

and $Q: H_r \rightarrow H_r$. Assume operators Q , and R are both self adjoint

and semi-positive and positive definite respectively, namely

$$Q^* = Q, Q \geq 0 \text{ and } R^* = R, R > 0.$$

Then the input structure diagnostic ΔJ_0 is given as

$$\begin{aligned} \Delta J_0 &= J(u=0) - J(u=\hat{u}) \\ &= J(0) - (J(0) - \frac{1}{2} \langle b, A^{-1}b \rangle_{H_m}) \\ &= \frac{1}{2} \langle b, A^{-1}b \rangle_{H_m} \end{aligned}$$

where $b \triangleq W^*Qd$, and $A \triangleq W^*QW + R$.

Householder's matrix equality can be extended for use with suitably defined Hilbert space operators (Chuang, 1970) so that if $B > 0$ and $D > 0$ then

$$(B+CDE)^{-1} = B^{-1} - B^{-1}CD(D+DEB^{-1}CD)^{-1}DEB.$$

Consequently ΔJ_0 may be expanded as:

$$\begin{aligned}\Delta J_0 &= \frac{1}{2} \langle b, A^{-1}b \rangle_{H_m} = \frac{1}{2} \langle b, (R+W*Q^{\frac{1}{2}} \cdot Q^{\frac{1}{2}}W)^{-1}b \rangle_{H_m} \\ &= \frac{1}{2} \langle b, R^{-1}R^{-1}W*Q^{\frac{1}{2}}(I+Q^{\frac{1}{2}}WR^{-1}W*Q^{\frac{1}{2}})^{-1}Q^{\frac{1}{2}}WR^{-1}b \rangle_{H_m}\end{aligned}$$

Recall $b = W*Qd$, write $Z=Q^{\frac{1}{2}}d$, and $D=Q^{\frac{1}{2}}WR^{-1}W*Q^{\frac{1}{2}}$

to obtain $\Delta J_0 = \frac{1}{2} \langle Z, [D-D(I+D)^{-1}D]Z \rangle_{H_r}$.

3.1.1

Compact operator $D=Q^{\frac{1}{2}}WR^{-1}W*Q^{\frac{1}{2}}$ is positive semi-definite and self

adjoint hence it has a denumerable set of positive eigenvalues

$\{\lambda_i\}_{i=1}^{\infty}$ and a corresponding set of ortho-normal eigen-functions

$\{w_i\}_{i=1}^{\infty}$. Further one of the eigenvalues attains the value of the

norm of operator D , and their only possible point of accumulation is

$\{0\}$.

Denote $\lambda_M = \|D\| \triangleq \sup_{x \in H_r} \{\langle x, Dx \rangle_{H_r} / \langle x, x \rangle_{H_r}\}$

then $\lambda_1 \in (0, \lambda_M]$. The space H_r is complete (being $L^2_r [0, T_f]$)

and decomposes as $H_r = H_1 \oplus H_2$ where H_1 is spanned by $\{w_i\}_{i=1}^{\infty}$, and

H_2 is the null space of operator D (denoted $N(D)$). Any $y \in H_r$, has

therefore the unique decomposition $y = \sum_{i=1}^{\infty} a_i w_i + y_0$, where $a_i \in \mathbb{R}^1$,

$i=1, 2, \dots$ and $y_0 \in H_2 \triangleq N(D)$. In particular for $Z=Q^{\frac{1}{2}}d$ obtain the

decomposition $Z = \alpha_0 w_0 + \sum_{i=1}^{\infty} \alpha_i w_i$, where $\alpha_0 w_0$ denotes the null space component of Z .

The various eigenvalue/vector properties include:

$$Dw_i = \lambda_i w_i, \quad (I+D)^{-1}w_i = (1+\lambda_i)^{-1}w_i \quad i=1, 2, 3, \dots$$

where $\{0 < \lambda_1 < \lambda_M\}_{i=1}^{\infty}$, and further

$$\langle w_i, w_j \rangle = \delta_{ij} \quad i, j=0, 1, 2, \dots \text{(ortho-normality)}$$

$$DZ = \sum_{i=1}^{\infty} \alpha_i \lambda_i w_i \quad (I+D)^{-1}DZ = \sum_{i=1}^{\infty} \alpha_i \lambda_i (1+\lambda_i)^{-1} w_i$$

Utilizing these relationships in 3.1.1 obtain

$$\Delta J_0 = \frac{1}{2} < \sum_{i=0}^{\infty} \alpha_i w_i , \quad \sum_{i=1}^{\infty} \alpha_i (\lambda_i - \lambda_i^2 / (1 + \lambda_i)) w_i >_{H_r}$$

$$\text{so that } \Delta J_0 = \frac{1}{2} \sum_{i=1}^{\infty} \alpha_i^2 \lambda_i / (1 + \lambda_i) \quad \{ 0 < \lambda_i \leq \lambda_M \}_{i=1}^{\infty}$$

This series clearly converges since $\frac{1}{2} \sum_{i=0}^{\infty} \alpha_i^2 = m_0$, and

$$0 \leq \lambda_i / (1 + \lambda_i) < 1.$$

The expression 3.1.1 for ΔJ_0 may be further expanded to give

$$\Delta J_0 = \frac{1}{2} < Z, DZ > - \frac{1}{2} < Z, D^2 Z > + \frac{1}{2} < Z, D^{3/2} (I + D)^{-1} D^{3/2} Z >$$

$$\text{Defining } m_j = \frac{1}{2} < Z, D^j Z > , \text{ and } R_{j+1} = \frac{1}{2} < Z, D^{\frac{j+1}{2}} (I + D)^{-1} D^{\frac{j+1}{2}} Z >$$

a more general expansion for ΔJ_0 is :

$$\Delta J_0 = \sum_{j=1}^r (-1)^{j+1} m_j + R_{r+1} , \quad r = 1, 2, \dots$$

$$\text{The terms } m_j \stackrel{\Delta}{=} \frac{1}{2} < Z, D^j Z >_{H_r} = \frac{1}{2} \sum_{i=0}^{\infty} \alpha_i^2 (\lambda_i)^j \quad (j = 0, 1, 2, \dots)$$

where $\lambda_0 \stackrel{\Delta}{=} 0$, and $\{ 0 < \lambda_i \leq \lambda_M \}_{i=1}^{\infty}$ are given the interpretation

" jth order moments " (Bryant, 1973) since they have the same

formulae as the jth order moments for masses $\{ \alpha_i^2 \}_{i=0}^{\infty}$ set at

distances $\{ \lambda_i \}_{i=0}^{\infty}$ from the origin. The term R_{j+1} is taken to

be the remainder term of the expansion. In general the series cannot

be used to calculate ΔJ_0 since it diverges if for any i, $\lambda_i > 1$

and $\alpha_i \neq 0$. However the moments m_j , $j = 0, 1, 2, \dots$ give indirect

information about the distribution of the numbers $\{ \lambda_i \}_{i=1}^{\infty}$ and

$\{ \alpha_1^2 \}_{i=0}^{\infty}$. Furthermore the first few moments are readily obtained by system simulation since

$$m_0 = \frac{1}{2} \langle Z, Z \rangle_{H_R} = \frac{1}{2} \langle d, Q d \rangle_{H_R}$$

$$m_1 = \frac{1}{2} \langle Z, D Z \rangle_{H_R} = \frac{1}{2} \langle W^* Q d, R^{-1} W^* Q d \rangle_{H_m} = \frac{1}{2} \langle b, R^{-1} b \rangle_{H_m}$$

$$\text{and } m_2 = \frac{1}{2} \langle Z, D^2 Z \rangle_{H_R} = \frac{1}{2} \langle W R^{-1} b, Q W R^{-1} b \rangle_{H_R}$$

The problem arises therefore of bounding the Hilbert space canonical

form $\Delta J_0 = \frac{1}{2} \sum_{i=1}^{\infty} \alpha_1^2 \lambda_i / (1 + \lambda_i)$ given the additional constraints

$$\frac{1}{2} \sum_{i=0}^{\infty} \alpha_1^2 = m_0, \quad \frac{1}{2} \sum_{i=1}^{\infty} \alpha_1^2 (\lambda_i)^j = m_j : j = 1, 2, \dots, r$$

with $\{ \alpha_1^2 \geq 0 \}_{i=0}^{\infty}$ and $\{ 0 < \lambda_i \leq \lambda_M \}_{i=1}^{\infty}$.

Using the definition $\{ a_1^2 \triangleq \frac{1}{2} \alpha_1^2 \}_{i=1}^{\infty}$, the following relabelled problem is obtained.

$$\text{Extremise } \Delta J_0 = \sum_{i=1}^{\infty} \frac{a_1^2 \lambda_i}{(1 + \lambda_i)} \quad 3.1.2$$

$$\left. \begin{aligned} \text{subject to } a_0^2 + \sum_{i=1}^{\infty} a_1^2 &= m_0 \\ \sum_{i=1}^{\infty} a_1^2 (\lambda_i)^j &= m_j, \quad j = 1, 2, \dots, r \end{aligned} \right\} \quad 3.1.3$$

$$\text{with } \{ a_1^2 \geq 0 \}_{i=0}^{\infty} \quad \text{and} \quad \{ 0 < \lambda_i \leq \lambda_M \}_{i=1}^{\infty} \quad 3.1.4$$

To obtain the reduction in dimensionality, these equations 3.1.2 to 3.1.4 can be re-written in the form of an infinite dimensional linear program

$$\Delta J_0 = \langle c(\lambda), x \rangle \triangleq \sum_{i=0}^{\infty} c_i x_i$$

$$\text{subject to } A(\lambda) x = b \in \mathbb{R}^{r+1} \quad \text{and} \quad x \geq 0$$

where the following identifications are made :

$$c_0 \triangleq 0, \quad \{ 0 < c_1 \triangleq \lambda_1 / (1 + \lambda_1) < 1 \}_{i=1}^{\infty}, \quad \{ x_i \triangleq a_i^2 \}_{i=0}^{\infty}$$

$$\text{with } A(\lambda) = \begin{bmatrix} 1 & 1 & 1 & - & - \\ 0 & \lambda_1 & \lambda_2 & & \\ & \lambda_1^2 & \lambda_2^2 & & \\ & \vdots & \vdots & & \\ & \lambda_1^r & \lambda_2^r & - & - \\ 0 & \lambda_1^r & \lambda_2^r & - & - \end{bmatrix} \quad \text{and } b = \begin{bmatrix} m_0 \\ m_1 \\ \vdots \\ m_r \end{bmatrix} \in \mathbb{R}^{r+1}$$

Define as any feasible solution any pair (λ, x) such that

$$(i) \quad A(\lambda) x = b \quad (ii) \quad \{ x_i \geq 0 \}_{i=0}^{\infty} \quad \text{and} \quad (iii) \quad \{ 0 < \lambda_i \leq \lambda_M \}_{i=1}^{\infty}$$

The desired reduction in dimensionality is given as :

Theorem 3.1.1 For a linear program of the above form (i) if there is

a feasible solution, there is a basic feasible solution and

(ii) if there is an optimal feasible solution there is an optimal basic feasible solution.

Proof (a) If x has only a finite number of non-zero elements, then the linear program may be written as one on $\mathbb{R}^N$. Consequently results (i) and (ii) follow from the fundamental theorem of linear programming (Lueubeger, 1973).

(b) If x has a denumerable number of elements, then to prove

(i) Pair (λ, x) are feasible, hence $A(\lambda) x = b$, so that for any $\rho > 0$,

$\exists N(\rho)$ for which the constraints may be partitioned as

$$\sum_{i=0}^N x_i = m_0 - \sum_{i=N+1}^{\infty} x_i, \quad \text{and} \quad \sum_{i=1}^N x_i (\lambda_i)^j = m_j - \sum_{i=N+1}^{\infty} x_i (\lambda_i)^j$$

$j = 1, 2, \dots, r$. Or in vector form as $A(N) x(N) = b(N)$ satisfies

$$\| A(N) x(N) - b \| \leq \rho$$

Choose $\rho = 1/n$, $n = 1, 2, \dots$ and applying the fundamental theorem of linear programming obtain a denumerable sequence of matrices $A(n) \in \mathbb{R}^{r+1 \times r+1}$ and basic feasible solutions $x(n) \in \mathbb{R}^{r+1}$ such that

$$A(n) x(n) = A(N(n)) x(N(n)) = b(N(n)).$$

The elements of $A(n)$ satisfy $\{ 0 \leq A(n)_{ij} \leq \lambda_M^{r+1} \}_{i=1, j=1}^{r+1, r+1}$

whilst the m_0 constraint implies that the elements of $x(n)$ satisfy $\{ 0 \leq x(n)_i \leq m_0 \}_{i=1}^{r+1}$. Consequently (Heine-Borel Theorem) a

subsequence may be selected so that $A(n_k), x(n_k)$ converge element-wise to $\bar{A} \in \mathbb{R}^{r+1 \times r+1}$ and $\bar{x} \in \mathbb{R}^{r+1}$.

Hence $0 \leq \| A(n_k) x(n_k) - b \| \leq 1/n_k$

as $n_k \rightarrow \infty$ $0 \leq \| \bar{A} \bar{x} - b \| \leq 0$

So that the feasible solution is basic.

To prove (ii) partition the optimal feasible solution $(\hat{\lambda}, \hat{x})$ as before so that $\| A(N) \hat{x}(N) - b \| \leq \rho$

Choosing $\| \cdot \|$ to be the sup norm on $\mathbb{R}^{r+1}$, have that the m_0 constraint implies $\sum_{i=N+1}^{\infty} \hat{x}_i \leq \rho$. Recalling that $c_i \in [0, 1]$

$\forall i = N+1, N+2, \dots$ gives $0 \leq \sum_{i=N+1}^{\infty} c_i \hat{x}_i \leq \rho$

Partition the optimal cost at $N(\rho)$ to obtain

$$\hat{\Delta J}_0 = \hat{\Delta J}_0(N) + \sum_{i=N+1}^{\infty} c_i \hat{x}_i \quad \left(\hat{\Delta J}_0(N) = \sum_{i=0}^N c_i \hat{x}_i \right)$$

giving $0 \leq \hat{\Delta J}_0 - \hat{\Delta J}_0(N) \leq \rho$

For the given $\rho > 0$, $\exists N(\rho)$ such that

$$(i) \quad A(N) \hat{x}(N) = b(N) \quad (ii) \quad 0 \leq \| A(N) \hat{x}(N) - b \| \leq \rho$$

and (iii) $0 \leq \hat{\Delta J}_0 - \hat{\Delta J}_0(N) \leq \rho$

Noting that $\hat{x}(N)$ is optimal for the truncated cost $\Delta J_0(N)$ (otherwise the optimality of $\hat{\Delta J}_0$ could be contradicted), choose $\rho = 1/n$,

$n = 1, 2, \dots$ and apply the fundamental theorem of linear programming to obtain a denumerable sequence of matrices $A(n) \in \mathbb{R}^{(r+1) \times (r+1)}$, and optimal basic feasible solutions $\hat{x}(n) \in \mathbb{R}^{r+1}$ such that

$$(i) \quad A(n) \hat{x}(n) = A(N(n)) \hat{x}(N(n)) = b(N(n)) \quad \text{and}$$

$$(ii) \quad \hat{\Delta J}_0(\hat{x}(n)) = \hat{\Delta J}_0(\hat{x}(N(n)))$$

As before a subsequence may be selected so that $A(n_k)$, and $x(n_k)$ converge elementwise to $\bar{A} \in \mathbb{R}^{(r+1) \times (r+1)}$ and $\hat{x}_{r+1} \in \mathbb{R}^{r+1}$ and

also that $\hat{\Delta J}(\hat{x}(n_k))$ converges to $\hat{\Delta J}(\hat{x}_{r+1})$.

Consequently as $n_k \rightarrow \infty$ $\bar{A} \hat{x}_{r+1} = b$

$$\text{and } 0 \leq \hat{\Delta J}_0 - \hat{\Delta J}_0(\hat{x}(n_k)) \leq 1/n_k$$

implies $\hat{\Delta J}_0 = \hat{\Delta J}_0(\hat{x}_{r+1})$

so that to every optimal feasible solution there is an optimal basic feasible solution.

Corollary 3.1.1 The infinite dimensional optimization of 3.1.2 to 3.1.4 can be replaced by an equivalent finite dimensional problem in $\mathbb{R}^{2(r+1)}$.

Proof From theorem 3.1.1, optimization over all feasible solutions may be reduced to optimizing over all basic feasible solutions of $A(\lambda) x = b$. That is optimizing over a compact set in $\mathbb{R}^{2(r+1)}$ on which the constraints are satisfied and for which the extrema are attained. Finally the optimal feasible solutions are the optimal basic feasible solutions.

The constraints 3.1.3 and 3.1.4 suggest that the infinite dimensional optimization will be basic in two ways depending on whether or not $a_0^2 = 0$; namely

a) Extremise $C^r(\alpha, e) = \sum_{i=1}^r \alpha_i e_i / (1 + e_i)$, subject to

$$\alpha_0 + \sum_{i=1}^r \alpha_i = m_0, \quad \sum_{i=1}^r \alpha_i (e_i)^j = m_j; \quad j = 1, 2, \dots, r$$

$$\{\alpha_i \geq 0\}_{i=0}^r \quad \text{and} \quad \{0 < e_i \leq \lambda_M\}_{i=1}^r$$

or b) Extremise $C^r(\alpha, e) = \sum_{i=1}^{r+1} \alpha_i e_i / (1 + e_i)$

$$\sum_{i=1}^{r+1} \alpha_i = m_0, \quad \sum_{i=1}^{r+1} \alpha_i (e_i)^j = m_j; \quad j = 1, 2, \dots, r$$

$$\{\alpha_i \geq 0\}_{i=1}^{r+1} \quad \text{and} \quad \{0 < e_i \leq \lambda_M\}_{i=1}^{r+1}$$

It is readily seen that using the second type of formulation in conjunction with a new "eigenvalue" constraint

$\{0 \leq e_i \leq \lambda_M\}_{i=1}^{r+1}$ subsumes both cases above. (Since if e_i activates the zero constraint, then the effect on problem (type b) is to give it the structure of problem type a)). The solutions to this new finite dimensional problem are obtained in the next section.

3.2 Analytical Bound Formulae for the Diagnostic, ΔJ_0 ($r=0,1,2,\dots,4$)

The problem of bounding ΔJ_0 reduces to finding solutions to the following type of non-linear optimization problem:

$$\text{rth Order Problem} \quad \text{Extremise} \quad C^r(\alpha, e) = \sum_{i=1}^{r+1} \alpha_i e_i / (1 + e_i)$$

$$\text{subject to} \quad \sum_{i=1}^{r+1} \alpha_i = m_0$$

$$\sum_{i=1}^{r+1} \alpha_i (e_i)^j = m_j \quad j = 1, 2, \dots, r$$

$$\text{with} \quad \alpha_i \geq 0 \quad \text{and} \quad 0 \leq e_i \leq \lambda_M \quad i = 1, 2, \dots, r+1$$

Bryant (1973) considered some low order bounds namely the cases $r = 0, 1$ and produced a univariate optimization problem for the case $r = 2$. A point estimator was also constructed for this case. Here a technique of Hadley (1970) is used to find analytical solutions to the above problem. The procedure is described as follows:

i) Compute the free optima ignoring the inequality constraints. If the point (or points) so obtained also satisfy the inequality constraints, then the global optima to the problem which includes the inequality constraints have been found.

ii) If one or more of the inequality constraints are not satisfied select one (not necessarily a constraint that was not satisfied at the first stage) and treat it as though active.

Repeat the process of finding the free optima with this constraint active. If the point (or points) yielding the global optima to this problem satisfy the inequality constraints not included then the global optima to the problem which includes all the inequality constraints has been found. If not, the process is continued until all combinations of including just

one inequality constraint is exhausted.

iii) If the optima are not obtained after stage ii) then the whole process is repeated activating combinations of two of the inequality constraints, and so on.

This procedure, although laborious, was successful in producing results for all the cases $r = 0, 1, \dots, 4$; the case $r = 5$ not appearing to be tractable. The details follow.

$$\text{Case } r = 0 \quad 0 \leq \Delta J_0 \leq m_0 \lambda_M / (1 + \lambda_M) \quad 3.2.1$$

$$\text{Proof} \quad \text{Extremise } C^0(\alpha, e) = \alpha_1 e_1 / (1 + e_1)$$

$$\text{subject to } \alpha_1 = m_0, \alpha_1 \geq 0, 0 \leq e_1 \leq \lambda_M$$

The inequality $0 \leq e_1 / (1 + e_1) \leq \lambda_M / (1 + \lambda_M)$ gives the result.

$$\text{Case } r = 1 \quad 0 \leq m_1 / (1 + \lambda_M) \leq \Delta J_0 \leq m_0 m_1 / (m_0 + m_1) \quad 3.2.2$$

$$\text{Proof} \quad \text{Extremise } C^1(\alpha, e) = \frac{\alpha_1 e_1}{(1 + e_1)} + \frac{\alpha_2 e_2}{(1 + e_2)}$$

$$\text{subject to } \alpha_1 + \alpha_2 = m_0 \text{ and } \alpha_1 e_1 + \alpha_2 e_2 = m_1$$

$$\text{with } \alpha_i \geq 0 \text{ and } 0 \leq e_i \leq \lambda_M, i = 1, 2.$$

1) Adjoin equality constraints and search for free optima:

$$C^1(\alpha, e, \mu) = \frac{\alpha_1 e_1}{(1 + e_1)} + \frac{\alpha_2 e_2}{(1 + e_2)} + \mu_1 (\alpha_1 + \alpha_2 - m_0) + \mu_2 (\alpha_1 e_1 + \alpha_2 e_2 - m_1)$$

The condition of optimality $\nabla C = 0 \in \mathbb{R}^6$ gives

$$e_1 \left(\mu_2 + \frac{1}{(1 + e_1)} \right) + \mu_1 = 0 \quad 3.2.3$$

$$e_2 \left(\mu_2 + \frac{1}{(1 + e_2)} \right) + \mu_1 = 0 \quad 3.2.4$$

$$\alpha_1 \left(\mu_2 + \frac{1}{(1 + e_1)^2} \right) = 0 \quad 3.2.5$$

$$\alpha_2 \left(\mu_2 + \frac{1}{(1 + e_2)^2} \right) = 0 \quad 3.2.6$$

$$\alpha_1 + \alpha_2 = m_0 \quad 3.2.7$$

$$\text{and } \alpha_1 e_1 + \alpha_2 e_2 = m_1 \quad 3.2.8$$

Equations 3.2.5 and 3.2.6 have three possible types of solution

i) a symmetric pair of the form

$$\alpha_1 = 0; \mu_2 = - (1 + e_2)^{-2}$$

$$\alpha_2 = 0; \mu_2 = - (1 + e_1)^{-2}$$

ii) $\alpha_1 = \alpha_2 = 0$

$$\text{iii) } \mu_2 = - (1 + e_1)^{-2} = - (1 + e_2)^{-2}$$

The first two cases activate an inequality constraint, whilst the third produces a problem insufficiently determined in α_1 , and α_2 . Conclude there are no optima without activating an inequality constraint.

2) Activate one inequality constraint at a time

Ignoring the symmetries which arise, three types of problems are possible

i) $\alpha_1 = 0$, ii) $e_1 = 0$ and iii) $e_1 = \lambda_M$.

In detail i) $\alpha_1 = 0$: $C^1(0, \alpha_2, e_1, e_2) = \alpha_2 e_2 / (1 + e_2)$

subject to $\alpha_2 = m_0$ and $\alpha_2 e_2 = m_1$. This problem is independent of e_1 , therefore setting $e_1 \in [0, \lambda_M]$ have

$$\hat{\alpha}_2 = m_0, \quad \hat{e}_2 = m_1/m_0 < \lambda_M \quad \text{and} \quad \hat{C} = m_0 m_1 / (m_0 + m_1)$$

$$\text{ii)} \quad e_1 = 0 : C^1(\alpha_1, \alpha_2, 0, e_2) = \alpha_2 e_2 / (1 + e_2)$$

$$\text{subject to } \alpha_1 + \alpha_2 = m_0 \quad \text{and} \quad \alpha_2 e_2 = m_1.$$

$$\text{Eliminating } \alpha_2 \text{ and } e_2 : \alpha_2 = m_0 - \alpha_1, \quad \text{and} \quad e_2 = m_1 / (m_0 - \alpha_1)$$

$$\text{Obtain } C^1(\alpha_1) = m_1 (m_0 - \alpha_1) / (m_0 + m_1 - \alpha_1)$$

$$\text{hence } \frac{dC^1}{d\alpha_1} = \frac{-m_1^2}{(m_0 + m_1 - \alpha_1)^2} \neq 0 \quad \alpha_1 \in [0, \infty)$$

$$\text{iii)} \quad e_1 = \lambda_M : C^1(\alpha_1, \alpha_2, \lambda_M, e_2) = \frac{\alpha_1 \lambda_M}{(1 + \lambda_M)} + \frac{\alpha_2 e_2}{(1 + e_2)}$$

$$\text{subject to } \alpha_1 + \alpha_2 = m_0 \quad \text{and} \quad \alpha_1 \lambda_M + \alpha_2 e_2 = m_1$$

Using Lagrange multipliers form an unconstrained cost

$$C_{ad}^1 = \frac{\alpha_1 \lambda_M}{(1 + \lambda_M)} + \frac{\alpha_2 e_2}{(1 + e_2)} + \mu_1 (\alpha_1 + \alpha_2 - m_0) + \mu_2 (\alpha_1 \lambda_M + \alpha_2 e_2 - m_1)$$

The condition of optimality $\nabla C_{ad}^1 = 0 \in \mathbb{R}^5$ provides the equations

$$\lambda_M \left(\mu_2 + \frac{1}{(1 + \lambda_M)} \right) + \mu_1 = 0 \quad 3.2.9$$

$$e_2 \left(\mu_2 + \frac{1}{(1 + e_2)} \right) + \mu_1 = 0 \quad 3.2.10$$

$$\alpha_2 \left(\mu_2 + \frac{1}{(1 + e_2)^2} \right) = 0 \quad 3.2.11$$

$$\alpha_1 + \alpha_2 = m_0 \quad 3.2.12$$

$$\alpha_1 \lambda_M + \alpha_2 e_2 = m_1 \quad 3.2.13$$

Equation 3.2.11 has two types of solution

a) $\alpha_2 = 0$ gives a contradiction between equations 3.2.12 and 3.2.13.

b) $\mu_2 = -(1 + e_2)^{-2}$ Substitution in equation 3.2.10 gives

$$\mu_1 = -e_2^2 / (1 + e_2)^2 \quad \text{and subsequently in 3.2.9 to obtain } e_2 = \lambda_M.$$

Obtain from equations 3.3.12 and 3.3.13 the contradiction

$\alpha_1 + \alpha_2 = m_0$ and $\alpha_1 + \alpha_2 = m_1/\lambda_M$. No new optima have been found, so a further step in the Hadley procedure is taken.

3) Activate two inequality constraints at a time

Ignoring symmetries eight types of cases arise:

- | | | | |
|---------------------|-------------------|-------------------------|-------------------|
| i) $\alpha_1 = 0$ | $\alpha_2 = 0$ | v) $\alpha_1 = 0$ | $e_2 = \lambda_M$ |
| ii) $\alpha_1 = 0$ | $e_1 = 0$ | vi) $e_1 = 0$ | $e_2 = 0$ |
| iii) $\alpha_1 = 0$ | $e_2 = 0$ | vii) $e_1 = 0$ | $e_2 = \lambda_M$ |
| iv) $\alpha_1 = 0$ | $e_1 = \lambda_M$ | viii) $e_1 = \lambda_M$ | $e_2 = \lambda_M$ |

In detail i) $\alpha_1 = \alpha_2 = 0$: Contradicts $\alpha_1 + \alpha_2 = m_0 > 0$.

ii) $\alpha_1 = 0, e_1 = 0$: Cost obtained in (2) (i).

iii) $\alpha_1 = 0, e_2 = 0$: $C^1(0, \alpha_2, e_1, 0) = 0$

subject to $\alpha_2 = m_0$ and $m_1 = 0$: Contradiction for m_1 .

iv) $\alpha_1 = 0, e_1 = \lambda_M$: Cost obtained in (2) (i).

v) $\alpha_1 = 0, e_2 = \lambda_M$: In the equality constraints

obtain $\alpha_2 = m_0, \alpha_2 \lambda_M = m_1$: Contradiction for α_2 .

vi) $e_1 = e_2 = 0$: $C^1(\alpha_1, \alpha_2, 0, 0) = 0$,

subject to $\alpha_1 + \alpha_2 = m_0$ and $0 = m_1$: Contradiction for m_1 .

vii) $e_1 = 0, e_2 = \lambda_M$: $C^1(\alpha_1, \alpha_2, 0, \lambda_M) = \alpha_2 \lambda_M / (1 + \lambda_M)$

subject to $\alpha_1 + \alpha_2 = m_0$ and $\alpha_2 \lambda_M = m_1$. Consequently

$$C^1(\hat{\alpha}_1, \hat{\alpha}_2, 0, \lambda_M) = m_1 / (1 + \lambda_M) \text{ where } \hat{\alpha}_1 = \frac{m_0}{\lambda_M} \left(\lambda_M - \frac{m_1}{m_0} \right) > 0$$

and $\hat{\alpha}_2 = m_1 / \lambda_M > 0$

$$\text{viii) } e_1 = e_2 = \lambda_M : C^1(\alpha_1, \alpha_2, \lambda_M, \lambda_M) = \frac{(\alpha_1 + \alpha_2) \lambda_M}{(1 + \lambda_M)}$$

subject to $\alpha_1 + \alpha_2 = m_0$ and $(\alpha_1 + \alpha_2) \lambda_M = m_1$. The problem is insufficiently determined in the pair α_1 , and α_2 .

The process of Hadley is terminated here since further activation of inequality constraints produces either the results obtained so far, or inconsistent cases. The two optima obtained are:

$$C^1(0, \alpha_2, e_1, e_2) = m_0 m_1 / (m_0 + m_1)$$

$$\text{with } \hat{\alpha}_2 = m_0, \hat{e}_2 = m_1/m_0 < \lambda_M \text{ and } \hat{e}_1 \in [0, \lambda_M]$$

$$\text{and } C^1(\hat{\alpha}_1, \hat{\alpha}_2, 0, \lambda_M) = m_1 / (1 + \lambda_M) \text{ where } \alpha_1 = \frac{m_0}{\lambda_M} \left(\lambda_M - \frac{m_1}{m_0} \right) > 0$$

and $\hat{\alpha}_2 = m_1/\lambda_M > 0$. (The property $\lambda_M > m_1/m_0$ is proved in Section 3.4.)

This concludes the analysis for the pair of bounds to be derived from the optimization problem of order $r = 1$. As can be seen the procedure is straightforward but tedious and will be more so for the higher order cases $r = 2, 3, 4$ and 5 .

A re-examination of the analysis for the above case indicates how much of the process can be ignored. The following points are noted:

- i) Expect no free optima to the problem when all the inequality constraints are disregarded. (The solutions to the optimality equations 3.2.3 to 3.2.8 required the activation of an inequality in α_1 , the only other solutions producing an insufficiently determined problems)
- ii) Sufficient inequality constraints must be activated to produce a feasible solution from the equality constraints that is the moment equations. Activating too few or too many inequality constraints produces insufficiently determined or over-determined problems respectively.

In the remaining cases these guidelines are used to avoid much tedious analysis:

$$\text{Case } r = 2 \quad 0 < \frac{m_1^2}{(m_1+m_2)} < \Delta J_0 < \frac{-m_0 m_1 \lambda_M^2 + (m_2 - m_1) m_0 \lambda_M + m_1^2}{(1 + \lambda_M)(m_2 + m_1 - (m_1 + m_0) \lambda_M)}$$

$$\text{Proof} \quad \text{Extremise } C^2(\alpha_1, \alpha_2, \alpha_3, e_1, e_2, e_3) = \sum_{i=1}^3 \frac{\alpha_i e_i}{1+e_i}$$

$$\text{subject to } \alpha_1 + \alpha_2 + \alpha_3 = m_0$$

$$\alpha_1 e_1 + \alpha_2 e_2 + \alpha_3 e_3 = m_1$$

$$\alpha_1 e_1^2 + \alpha_2 e_2^2 + \alpha_3 e_3^2 = m_2$$

with $\alpha_i \geq 0$ and $0 \leq e_i \leq \lambda_M$, $i=1,2,3$.

The three equality constraints form a set of equations in six variables $\alpha_1, \alpha_2, \alpha_3, e_1, e_2, e_3$, consequently three variables are to be removed by activating constraints. (The various symmetries will be ignored.)

Set $\alpha_1 = 0$, $e_1 \in [0, \lambda_M]$ to obtain

$$\alpha_2 + \alpha_3 = m_0, \quad \alpha_2 e_2 + \alpha_3 e_3 = m_1 \quad \text{and} \quad \alpha_2 e_2^2 + \alpha_3 e_3^2 = m_2.$$

A further variable is to be removed. This can be achieved in one of three ways:

$$\text{i) } \alpha_2 = 0 \quad \text{ii) } e_2 = 0 \quad \text{iii) } e_2 = \lambda_M$$

In detail i) $\alpha_2 = 0$ requires the additional specification of

$$e_2 \in [0, \lambda_M] \quad \text{and leads to the contradiction } \alpha_3 = m_0, \quad \text{and } \alpha_3 = m_1^2 / m_2$$

ii) $e_2 = 0$ yields the following set of equations

$$\alpha_2 + \alpha_3 = m_0, \quad \alpha_3 e_3 = m_1 \quad \text{and} \quad \alpha_3 e_3^2 = m_2.$$

$$\text{hence } \hat{e}_3 = m_2 / m_1, \quad \hat{\alpha}_3 = m_1^2 / m_2 \quad \text{and} \quad \hat{\alpha}_2 = m_0 - m_1^2 / m_2$$

Using the properties $0 < \frac{m_1}{m_0} < \frac{m_2}{m_1} < \lambda_M$

3.2.14

(these are proved in Section 3.4) gives $\hat{e}_3 \in [0, \lambda_M]$, $\hat{\alpha}_3 > 0$

and $\hat{\alpha}_2 = m_1 \left(\frac{m_0}{m_1} - \frac{m_1}{m_2} \right) > 0$. The value of the cost is

$$C^2(0, \hat{\alpha}_2, \hat{\alpha}_3, 0, 0, \hat{e}_3) = \frac{\hat{\alpha}_3 \hat{e}_3}{(1 + \hat{e}_3)} = \frac{m_1^2}{(m_1 + m_2)}$$

iii) $e_2 = \lambda_M$ yields the following equations :

$$\alpha_2 + \alpha_3 = m_0 \quad \alpha_2 \lambda_M + \alpha_3 e_3 = m_1 \quad \text{and} \quad \alpha_2 \lambda_M^2 + \alpha_3 e_3^2 = m_2.$$

The first equation gives $\hat{\alpha}_2 = m_0 - \hat{\alpha}_3$, substitute into the remaining two equations to obtain

$$\hat{e}_3 + \lambda_M = \frac{m_2 - m_0 \lambda_M^2}{m_1 - m_0 \lambda_M} ; \text{ consequently}$$

$$\hat{e}_3 = (m_2 - m_1 \lambda_M) / (m_1 - m_0 \lambda_M) ,$$

$$\hat{\alpha}_3 = (m_1 - m_0 \lambda_M)^2 / (m_0 \lambda_M^2 - 2m_1 \lambda_M + m_2) \quad \text{and}$$

$$\hat{\alpha}_2 = (m_0 m_2 - m_1^2) / (m_0 \lambda_M^2 - 2m_1 \lambda_M + m_2) .$$

The moment properties 3.3.14 are again used to show that $\hat{\alpha}_2$, $\hat{\alpha}_3$

and $\hat{e}_3$ are feasible. Namely that 3.3.14 giving $(m_2/m_1 - m_1/m_0) > 0$

hence

$$m_0 \lambda_M^2 - 2m_1 \lambda_M + m_2 = m_0 \left[\left(\lambda_M - \frac{m_1}{m_0} \right)^2 + \frac{m_1}{m_0} \left(\frac{m_2}{m_1} - \frac{m_1}{m_0} \right) \right] > 0$$

to give $\hat{\alpha}_2 > 0$ and $\hat{\alpha}_3 > 0$; and

$$\frac{m_2 - m_1 \lambda_M}{m_1 - m_0 \lambda_M} = \frac{m_1}{m_0} \left(\frac{\lambda_M - m_2/m_1}{\lambda_M - m_1/m_0} \right) \quad \text{to give} \quad 0 < \hat{e}_3 < \lambda_M.$$

It remains therefore to calculate the cost

$$C^2(0, \hat{\alpha}_2, \hat{\alpha}_3, 0, \lambda_M, \hat{e}_3) = \frac{\hat{\alpha}_2 \lambda_M}{(1+\lambda_M)} + \frac{\hat{\alpha}_3 \hat{e}_3}{(1+\hat{e}_3)}$$

$$\begin{aligned} C^2(\text{optimal}) &= \frac{m_0 \lambda_M}{(1+\lambda_M)} + \hat{\alpha}_3 \left[\frac{\hat{e}_3}{(1+\hat{e}_3)} - \frac{\lambda_M}{(1+\lambda_M)} \right] \\ &= \frac{m_0 \lambda_M}{(1+\lambda_M)} + \frac{\hat{\alpha}_3 (\hat{e}_3 - \lambda_M)}{(1+\hat{e}_3)(1+\lambda_M)} = \frac{1}{(1+\lambda_M)} \left[m_0 \lambda_M + \frac{(m_1 - m_0 \lambda_M)}{(1+\hat{e}_3)} \right] \\ &= \frac{1}{(1+\lambda_M)} \left[m_0 \lambda_M + \frac{(m_1 - m_0 \lambda_M)^2}{(m_1 + m_2 - (m_0 + m_1) \lambda_M)} \right] \end{aligned}$$

$$C^2(0, \hat{\alpha}_2, \hat{\alpha}_3, 0, \lambda_M, \hat{e}_3) = \frac{-m_0 m_1 \lambda_M^2 + m_0 (m_2 - m_1) \lambda_M + m_1^2}{(1+\lambda_M)(m_1 + m_2 - (m_1 + m_0) \lambda_M)}$$

This value is to be ordered with $C^2(0, \hat{\alpha}_2, \hat{\alpha}_3, 0, 0, \hat{e}_3) = m_1^2 / (m_1 + m_2)$.

Set $\lambda_M = m_2 / m_1 + \delta$, $\delta > 0$ then

$$C^2(\delta) = \frac{-m_0 m_1 (m_2/m_1 + \delta)^2 - m_2 (m_2 - m_1) (m_2/m_1 + \delta) + m_1^2}{(1 + m_2/m_1 + \delta)(m_1 + m_2 - (m_1 + m_0)(m_2/m_1 + \delta))}$$

where $C^2(0) = m_1^2 / (m_1 + m_2)$ and after much manipulation

$$\frac{\partial C^2(\delta)}{\partial \delta} = \frac{a_1 \delta^2 + a_2 \delta + a_3}{(1 + m_2/m_1 + \delta)^2 (m_1 + m_2 - (m_0 + m_1)(m_2/m_1 + \delta))^2} > 0, \quad \delta \geq 0$$

where $a_1 = m_0 m_1 \left(\frac{m_2}{m_1} - \frac{m_1}{m_0} \right) > 0$ $a_2 = 2m_0^2 m_1 \left(\frac{m_2}{m_1} - \frac{m_1}{m_0} \right)^2 > 0$

and $a_3 = m_0^2 m_2 \left(\frac{m_2}{m_1} - \frac{m_1}{m_0} \right)^2 > 0$.

Hence the result as given at the beginning of this case is obtained.

Before continuing with the cases $r = 3$, and 4 , introduce some new notation and summarise the bounds obtained so far. Denote the r th order upper and lower bounds as $C_{\max}^r(\dots)$ and $C_{\min}^r(\dots)$ where the argument list contains the moments, and possibly λ_M upon which the bound is dependent. Then the formulae obtained so far are:

$$C_{\min}^0 = 0$$

$$C_{\max}^0(m_0, \lambda_M) = \frac{m_0 \lambda_M}{(1 + \lambda_M)}$$

$$C_{\min}^1(m_1, \lambda_M) = \frac{m_1}{(1 + \lambda_M)}$$

$$C_{\max}^1(m_0, m_1) = \frac{m_0 m_1}{(m_0 + m_1)}$$

$$C_{\min}^2(m_1, m_2) = \frac{m_1^2}{(m_1 + m_2)}$$

$$C_{\max}^2(m_0, m_1, m_2, \lambda_M) = -\frac{m_0 m_1 \lambda_M^2 + (m_2 - m_1) m_0 \lambda_M + m_1^2}{(1 + \lambda_M)(m_2 + m_1 - (m_1 + m_0) \lambda_M)}$$

The finding of the optimal costs for the cases $r = 3$, and $r = 4$ is straightforward but the complexity of the formulae obtained prohibits further analysis to determine the ordering of the bounds.

Instead numerical tests and the observation that the bound formulae alternate in being a function of λ_M is relied upon to settle the question of ordering.

$$\text{Case } r = 3 \quad C_{\min}^3(m_1, m_2, m_3, \lambda_M) < \Delta J_0 < C_{\max}^3(m_0, m_1, m_2, m_3)$$

Where for the minimum the optimal values are:

$$\hat{e}_4 = \frac{m_2 \lambda_M - m_3}{m_1 \lambda_M - m_2}, \quad \hat{\alpha}_3 = \frac{m_1 \hat{e}_4 - m_2}{\lambda_M (\hat{e}_4 - \lambda_M)}, \quad \hat{\alpha}_4 = \frac{m_2 - m_1 \lambda_M}{\hat{e}_4 (\hat{e}_4 - \lambda_M)}$$

$$\text{to give } C_{\min}^3(m_1, m_2, m_3, \lambda_M) = \frac{\hat{\alpha}_3 \lambda_M}{(1 + \lambda_M)} + \frac{\hat{\alpha}_4 \hat{e}_4}{(1 + \hat{e}_4)}$$

These optimal values are shown to be feasible in the sequel but numerical trials were used to show the associated cost to be a minimum.

For the maximum: Set $k_1 = (m_1/m_0)^2 - m_2/m_0$

$k_2 = (m_1/m_0)^3 - m_3/m_0$ and $k_3 = m_1/m_0$

Solve the associated quadratic:

$$x^2 + (3k_3 - k_2/k_1) x + k_1 = 0$$

for the roots $\hat{x}$ and $\hat{y}$

Compute $\hat{e}_3 = \hat{x} + m_1/m_0$, $\hat{e}_4 = \hat{y} + m_1/m_0$

$$\hat{\alpha}_3 = m_0 \hat{y} / (\hat{y} - \hat{x}) \quad \text{and} \quad \hat{\alpha}_4 = m_0 \hat{x} / (\hat{x} - \hat{y})$$

Obtaining the cost as

$$C_{\max}^3(m_0, m_1, m_2, m_3) = \frac{\hat{\alpha}_3 \hat{e}_3}{(1 + \hat{e}_3)} + \frac{\hat{\alpha}_4 \hat{e}_4}{(1 + \hat{e}_4)}$$

Numerical tests should be included to check feasibility.

Proof Extremise $C^3(\alpha, e) = \sum_{i=1}^4 \frac{\alpha_i e_i}{(1 + e_i)} \quad \alpha \text{ and } e \in \mathbb{R}^4$

subject to $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = m_0$

$$\alpha_1 e_1 + \alpha_2 e_2 + \alpha_3 e_3 + \alpha_4 e_4 = m_1$$

$$\alpha_1 e_1^2 + \alpha_2 e_2^2 + \alpha_3 e_3^2 + \alpha_4 e_4^2 = m_2$$

$$\alpha_1 e_1^3 + \alpha_2 e_2^3 + \alpha_3 e_3^2 + \alpha_4 e_4^2 = m_3$$

with $\alpha_i \geq 0$ and $0 \leq e_i \leq \lambda_M \quad i = 1, \dots, 4$.

The four equality constraints form a set of equations in eight variables, consequent four are to be determined by activating the inequality constraints.

Set $\alpha_1 = 0$, $e_1 \in [0, \lambda_M]$ to obtain

$$\alpha_2 + \alpha_3 + \alpha_4 = m_0$$

$$\alpha_2 e_2 + \alpha_3 e_3 + \alpha_4 e_4 = m_1$$

$$\alpha_2 e_2^2 + \alpha_3 e_3^2 + \alpha_4 e_4^2 = m_2$$

$$\alpha_2 e_2^3 + \alpha_3 e_3^3 + \alpha_4 e_4^3 = m_3$$

Symmetries aside seven cases arise:-

$$i) \alpha_2 = 0; e_2 \in [0, \lambda_M]$$

$$v) e_2 = e_3 = 0$$

$$ii) \alpha_2 = \alpha_3 = 0$$

$$vi) e_2 = 0; e_3 = \lambda_M$$

$$iii) \alpha_2 = 0; e_3 = 0$$

$$vii) e_2 = e_3 = \lambda_M$$

$$iv) \alpha_2 = 0; e_3 = \lambda_M$$

In detail i) $\alpha_2 = 0; e_2 \in [0, \lambda_M]$ the constraints reduce to

$$\alpha_3 + \alpha_4 = m_0$$

$$\alpha_3 e_3^2 + \alpha_4 e_4^2 = m_2$$

$$\alpha_3 e_3 + \alpha_4 e_4 = m_1$$

$$\alpha_3 e_3^3 + \alpha_4 e_4^3 = m_3$$

Assume $e_3 \neq e_4$ then

$$\begin{bmatrix} 1 & 1 \\ e_3 & e_4 \end{bmatrix} \begin{bmatrix} \alpha_3 \\ \alpha_4 \end{bmatrix} = \begin{bmatrix} m_0 \\ m_1 \end{bmatrix}$$

$$\text{gives } \alpha_3 = \frac{m_0 e_4 - m_1}{e_4 - e_3} \quad \text{and} \quad \alpha_4 = \frac{m_1 - m_0 e_3}{(e_4 - e_3)}$$

Substitution in the remaining constraints gives

$$m_0 e_3 e_4 - m_1 (e_3 + e_4) + m_2 = 0$$

$$m_0 e_3 e_4 (e_3 + e_4) - m_1 (e_3^2 + e_3 e_4 + e_4^2) + m_3 = 0$$

These equations may be factorized as

$$(e_3 - m_1/m_0)(e_4 - m_1/m_0) = (m_1/m_0)^2 - m_2/m_0$$

$$\text{and } (e_3 - m_1/m_0)(e_4 - m_1/m_0)(e_3 + e_4 + m_1/m_0) = (m_1/m_0)^3 - m_3/m_0$$

$$\text{Set } x = e_3 - m_1/m_0; y = e_4 - m_1/m_0$$

$$k_1 = (m_1/m_0)^2 - m_2/m_0, \quad k_2 = (m_1/m_0)^3 - m_3/m_0$$

$$\text{and } k_3 = m_1/m_0$$

$$\text{Solve } xy = k_1 \text{ with } xy(x+y+3k_3) = k_2$$

to obtain $y = k_1/x$ with x satisfying quadratic

$$x^2 + (3k_3 - k_2/k_1)x + k_1 = 0.$$

The roots of this quadratic are $\hat{x}$, and $k_1/\hat{x} = \hat{y}$ which give directly

$$\hat{e}_3 = \frac{m_1}{m_0} + \hat{x} \text{ and } \hat{e}_4 = \frac{m_1}{m_0} + \hat{y}, \text{ then}$$

$$\text{using } \hat{\alpha}_3 = \frac{m_0 \hat{e}_4 - m_1}{(\hat{e}_4 - \hat{e}_3)} = \frac{m_0 \hat{y}}{\hat{y} - \hat{x}} \quad \text{and}$$

$$\text{similarly } \hat{\alpha}_4 = \frac{m_1 - m_0 \hat{e}_3}{(\hat{e}_4 - \hat{e}_3)} = \frac{m_0 \hat{x}}{\hat{x} - \hat{y}} \quad \text{to obtain}$$

$$C_{\max}^3(m_0, m_1, m_2, m_3) = \frac{\hat{\alpha}_3 \hat{e}_3}{(1 + \hat{e}_3)} + \frac{\hat{\alpha}_4 \hat{e}_4}{(1 + \hat{e}_4)}$$

ii) $\alpha_2 = \alpha_3 = 0$. The equality constraints reduce to

$$\alpha_4 = m_0, \alpha_4 e_4 = m_1, \alpha_4 e_4^2 = m_2 \text{ and } \alpha_4 e_4^3 = m_3$$

with e_2 , and e_3 undetermined. Pathological case requiring e_2 and

$$e_3 \in [0, \lambda_M] \text{ and } m_1/m_0 = m_2/m_1 = m_3/m_2.$$

iii) $\alpha_2 = 0$ and $e_3 = 0$. The equality constraints reduce to

$$\alpha_3 + \alpha_4 = m_0, \alpha_4 e_4 = m_1, \alpha_4 e_4^2 = m_2, \alpha_4 e_4^3 = m_3$$

which again is a pathological case requiring specification of e_2 ,

$$\text{and } m_2/m_1 = m_3/m_2.$$

iv) $\alpha_2 = 0, e_3 = \lambda_M$. The constraints reduce to

$$\alpha_3 + \alpha_4 = m_0, \quad \alpha_3 \lambda_M^2 + \alpha_4 e_4^2 = m_2$$

$$\alpha_3 \lambda_M + \alpha_4 e_4 = m_1, \quad \alpha_3 \lambda_M^3 + \alpha_4 e_4^3 = m_3$$

The cost is independent of e_2 which has to be assumed feasible.

If it is assumed that $\hat{e}_4 \neq \lambda_M$ it can be shown (by elimination of α_3 and α_4) that $\hat{e}_4$ is a zero of

$e_4^2 + (\lambda_M - m_3/m_2)e_4 + \lambda_M(\lambda_M - m_3/m_2)$. This has only imaginary zeros (use $0 < m_3/m_2 < \lambda_M$ which is proved in Section 3.4)

consequently the case is dismissed.

v) $e_2 = e_3 = 0$. The equality constraints reduce to

$$\alpha_2 + \alpha_3 + \alpha_4 = m_0, \quad \alpha_4 e_4 = m_1, \quad \alpha_4 e_4^2 = m_2 \quad \text{and} \quad \alpha_4 e_4^3 = m_3.$$

Clearly of the type (iii) above with additional indeterminacy in α_2 .

vi) $e_2 = 0$, and $e_3 = \lambda_M$. This gives

$$\alpha_2 + \alpha_3 + \alpha_4 = m_0 \qquad \alpha_3 \lambda_M^2 + \alpha_4 e_4^2 = m_2$$

$$\alpha_3 \lambda_M + \alpha_4 e_4 = m_1 \qquad \alpha_3 \lambda_M^3 + \alpha_4 e_4^3 = m_3$$

$$\text{with cost } C^3(\alpha, e) = \frac{\alpha_3 \lambda_M}{(1+\lambda_M)} + \frac{\alpha_4 e_4}{(1+e_4)}$$

The cost is independent of α_2 , and α_2 does not appear in last

three constraints hence solve for $\hat{\alpha}_2 = m_0 - \hat{\alpha}_3 - \hat{\alpha}_4$ and use

remaining equations to solve for $\hat{\alpha}_3$, $\hat{\alpha}_4$ and $\hat{e}_3$.

$$\text{Assume } e_4 \neq \lambda_M \text{ solve } \begin{bmatrix} \lambda_M & e_4 \\ \lambda_M^2 & e_4^2 \end{bmatrix} \begin{bmatrix} \alpha_3 \\ \alpha_4 \end{bmatrix} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}$$

$$\text{to obtain } \hat{\alpha}_3 = \frac{m_1 e_4 - m_2}{\lambda_M (e_4 - \lambda_M)} \quad \text{and} \quad \hat{\alpha}_4 = \frac{m_2 - m_1 \lambda_M}{e_4 (e_4 - \lambda_M)}$$

Substitution in the remaining constraint gives

$$0 < \hat{e}_4 = \frac{(m_2 \lambda_M - m_3)}{(m_1 \lambda_M - m_2)} = \frac{m_2 (\lambda_M - m_3/m_2)}{m_1 (\lambda_M - m_2/m_1)} < \lambda_M$$

$$\text{implies } \hat{\alpha}_3 = \frac{m_1 (m_2/m_1 - \hat{e}_4)}{\lambda_M (\lambda_M - \hat{e}_4)} > 0 \quad \text{and} \quad \hat{\alpha}_4 = \frac{m_1 (\lambda_M - m_2/m_1)}{\hat{e}_4 (\lambda_M - \hat{e}_4)} > 0$$

vii) $e_2 = e_3 = \lambda_M$. This yields a problem insufficiently determined in α_2 and α_3 .

This concludes the analysis for the case $r = 3$.

Case $r = 4$ $C_{\min}^4(m_1, \dots, m_4) < \Delta J_O < C_{\max}^4(m_O, m_1, \dots, m_4, \lambda_M)$

For the minimum : set $k_1 = m_2/m_1$, $k_2 = (m_2/m_1)^2 - m_3/m_1$

and $k_3 = (m_2/m_1)^3 - m_4/m_1$. Solve the quadratic equation

$$x^2 + (3k_1 - k_3/k_2) x + k_2 = 0 \quad \text{for the roots } \hat{x}, \text{ and } \hat{y}.$$

$$\text{Set } \hat{e}_4 = \hat{x} + m_2/m_1 \quad \text{and} \quad \hat{e}_5 = \hat{y} + m_2/m_1$$

$$\hat{\alpha}_4 = \frac{\hat{e}_5 m_1 - m_2}{\hat{e}_4 (\hat{e}_5 - \hat{e}_4)} = \frac{m_1 \hat{y}}{\hat{e}_4 (\hat{y} - \hat{x})}$$

$$\hat{\alpha}_5 = \frac{m_2 - m_1 \hat{e}_4}{\hat{e}_5 (\hat{e}_5 - \hat{e}_4)} = \frac{m_1 \hat{x}}{\hat{e}_5 (\hat{x} - \hat{y})}$$

$$\text{Calculate } C_{\min}^4(m_1, \dots, m_4) = \frac{\hat{\alpha}_4 \hat{e}_4}{(1 + \hat{e}_4)} + \frac{\hat{\alpha}_5 \hat{e}_5}{(1 + \hat{e}_5)}$$

For the maximum :

$$\text{Define } M_O = m_O \lambda_M - m_1 \quad M_1 = m_1 \lambda_M - m_2$$

$$M_2 = m_2 \lambda_M - m_3 \quad M_3 = m_3 \lambda_M - m_4$$

$$\text{Also } K_1 = M_1/M_O, \quad K_2 = (M_1/M_O)^2 - M_2/M_O$$

$$K_3 = (M_1/M_0)^3 - M_3/M_0$$

Solve the quadratic $x^2 + \left(3K_1 - \frac{K_3}{K_2}\right)x + K_2 = 0$

for $\hat{x}$, and $\hat{y}$; form $\hat{e}_4 = \hat{x} + M_1/M_0$ and $\hat{e}_5 = \hat{y} + M_1/M_0$

Compute the α 's from

$$\hat{\alpha}_3 = (m_0 \hat{e}_4 \hat{e}_5 - m_1 (\hat{e}_4 + \hat{e}_5) + m_2) / (\hat{e}_4 - \lambda_M)(\hat{e}_5 - \lambda_M)$$

$$\hat{\alpha}_4 = (-m_0 \hat{e}_5 \lambda_M + m_1 (\hat{e}_5 + \lambda_M) - m_2) / (\hat{e}_5 - \hat{e}_4)(\hat{e}_4 - \lambda_M)$$

$$\hat{\alpha}_5 = (m_0 \lambda_M \hat{e}_4 - m_1 (\hat{e}_4 + \lambda_M) + m_2) / (\hat{e}_5 - \hat{e}_4)(\hat{e}_5 - \lambda_M)$$

Compute the optimal cost as

$$C_{\max}^4(m_0, \dots, m_4, \lambda_M) = \frac{\hat{\alpha}_3 \lambda_M}{(1+\lambda_M)} + \frac{\hat{\alpha}_4 \hat{e}_4}{(1+\hat{e}_4)} + \frac{\hat{\alpha}_5 \hat{e}_5}{(1+\hat{e}_5)}$$

Proof Extremise $C^4(\alpha, e) = \sum_{i=1}^5 \frac{\alpha_i e_i}{(1+e_i)} \quad \alpha \text{ and } e \in \mathbb{R}^5$

subject to $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5 = m_0$

$$\alpha_1 e_1 + \alpha_2 e_2 + \alpha_3 e_3 + \alpha_4 e_4 + \alpha_5 e_5 = m_1$$

$$\alpha_1 e_1^2 + \alpha_2 e_2^2 + \alpha_3 e_3^2 + \alpha_4 e_4^2 + \alpha_5 e_5^2 = m_2$$

$$\alpha_1 e_1^3 + \alpha_2 e_2^3 + \alpha_3 e_3^3 + \alpha_4 e_4^3 + \alpha_5 e_5^3 = m_3$$

$$\alpha_1 e_1^4 + \alpha_2 e_2^4 + \alpha_3 e_3^4 + \alpha_4 e_4^4 + \alpha_5 e_5^4 = m_4$$

with $\alpha_i \geq 0$ and $0 \leq e_i \leq \lambda_M \quad i = 1, 2, \dots, 5$

The five constraints form a set of equations in ten variables. Four variables are removed by (ignoring symmetries) setting

$$\alpha_1 = 0, \quad e_1 \in [0, \lambda_M] \quad \text{and} \quad \alpha_2 = 0, \quad e_2 \in [0, \lambda_M]$$

to leave

$$\alpha_3 + \alpha_4 + \alpha_5 = m_0$$

$$\alpha_3 e_3 + \alpha_4 e_4 + \alpha_5 e_5 = m_1$$

$$\alpha_3 e_3^2 + \alpha_4 e_4^2 + \alpha_5 e_5^2 = m_2$$

$$\alpha_3 e_3^3 + \alpha_4 e_4^3 + \alpha_5 e_5^3 = m_3$$

$$\alpha_3 e_3^4 + \alpha_4 e_4^4 + \alpha_5 e_5^4 = m_4$$

Three cases arise in removing one more variable

$$\text{i) } \alpha_3 = 0 \quad \text{ii) } e_3 = 0 \quad \text{iii) } e_3 = \lambda_M$$

In detail i) $\alpha_3 = 0$ causes the cost to be independent of e_3 which must be assumed to be feasible, and also leaves five equations in four variables. Under the assumption $e_4 \neq e_5$, the equations can be reduced to the system :

$$-m_0 e_4 e_5 + m_1 (e_4 + e_5) = m_2$$

$$-m_1 e_4 e_5 + m_2 (e_4 + e_5) = m_3$$

$$-m_2 e_4 e_5 + m_3 (e_4 + e_5) = m_4$$

thus for a unique solution of pair (e_4, e_5) require a degeneracy in the moments $m_0, \dots, m_4$.

ii) $e_3 = 0$. The equations reduce to

$$\alpha_3 + \alpha_4 + \alpha_5 = m_0$$

$$\alpha_4 e_4 + \alpha_5 e_5 = m_1$$

$$\alpha_4 e_4^2 + \alpha_5 e_5^2 = m_2$$

$$\alpha_4 e_4^3 + \alpha_5 e_5^3 = m_3$$

$$\alpha_4 e_4^4 + \alpha_5 e_5^4 = m_4$$

Solve for α_4 and α_5 using second and third equations, assuming

$$e_4 \neq e_5. \text{ That is solve } \begin{bmatrix} e_4 & e_5 \\ e_4^2 & e_5^2 \end{bmatrix} \begin{bmatrix} \alpha_4 \\ \alpha_5 \end{bmatrix} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}$$

to obtain $\alpha_4 = \frac{e_4 m_1 - m_2}{e_4(e_5 - e_4)}$ and $\alpha_5 = \frac{m_2 - m_1 e_4}{e_5(e_5 - e_4)}$ Substitution in the

remaining equations yields the pair $m_1 e_4 e_5 - m_2(e_4 + e_5) + m_3 = 0$

and $m_1 e_4 e_5(e_4 + e_5) - m_2(e_4^2 + e_4 e_5 + e_5^2) + m_4 = 0$

Factorize into $(e_4 - m_2/m_1)(e_5 - m_2/m_1) = (m_2/m_1)^2 - m_3/m_1$

and $(e_4 - m_2/m_1)(e_5 - m_2/m_1)(e_4 + e_5 + m_2/m_1) = (m_2/m_1)^2 - m_4/m_1$

Let $k_1 = m_2/m_1$, $k_2 = (m_2/m_1)^2 - m_3/m_1$, $k_3 = (m_2/m_1)^3 - m_4/m_1$

with $x = e_4 - m_2/m_1$ and $y = e_5 - m_2/m_1$

Form the equations $xy = k_2$ and $xy(x + y + 3k_1) = k_3$

Eliminate y to obtain the quadratic

$x^2 + (3k_1 - k_3/k_2)x + k_2 = 0$ whose roots are $\hat{x}$ and $k_2/\hat{x} = \hat{y}$.

From here the process is reversed to find the optimal values $\hat{\alpha}_4$, $\hat{\alpha}_5$

$\hat{e}_4$ and $\hat{e}_5$, consequently the optimal cost $C_{\min}^4(m_1, \dots, m_4)$.

iii) $e_3 = \lambda_M$. The equality constraints become

$$\alpha_3 + \alpha_4 + \alpha_5 = m_0$$

$$\alpha_3 \lambda_M + \alpha_4 e_4 + \alpha_5 e_5 = m_1$$

$$\alpha_3 \lambda_M^2 + \alpha_4 e_4^2 + \alpha_5 e_5^2 = m_2$$

$$\alpha_3 \lambda_M^3 + \alpha_4 e_4^3 + \alpha_5 e_5^3 = m_3$$

$$\alpha_3 \lambda_M^4 + \alpha_4 e_4^4 + \alpha_5 e_5^4 = m_4$$

Solve the first three equations as

$$\begin{bmatrix} 1 & 1 & 1 \\ \lambda_M & e_4 & e_5 \\ \lambda_M^2 & e_4^2 & e_5^2 \end{bmatrix} \begin{bmatrix} \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{bmatrix} = \begin{bmatrix} m_0 \\ m_1 \\ m_2 \end{bmatrix}$$

$$\text{to obtain } \alpha_3 = \frac{m_0 e_4 e_5 - m_1 (e_4 + e_5) + m_2}{(e_4 - \lambda_M)(e_5 - \lambda_M)}$$

$$\alpha_4 = \frac{-m_0 \lambda_M e_5 + m_1 (e_5 + \lambda_M) - m_2}{(e_5 - e_4)(e_4 - \lambda_M)}$$

$$\alpha_5 = \frac{m_0 \lambda_M e_4 - m_1 (e_4 + \lambda_M) - m_2}{(e_5 - e_4)(e_5 - \lambda_M)}$$

Substitute these results for α_3, α_4 and α_5 in the remaining two equality constraints to obtain the pair

$$e_4 e_5 \lambda_M m_0 - m_1 (e_4 e_5 + (e_4 + e_5) \lambda_M) + m_2 (e_4 + e_5 + \lambda_M) = m_3.$$

$$e_4 e_5 \lambda_M (e_4 + e_5 + \lambda_M) - m_1 (e_4 + e_5) (e_4 + \lambda_M) (e_5 + \lambda_M)$$

$$+ m_2 (e_4^2 + e_5^2 + \lambda_M^2 + e_4 e_5 + \lambda_M (e_4 + e_5)) = m_4.$$

The first of these may be written

$$(m_0 \lambda_M - m_1) e_4 e_5 - (m_1 \lambda_M - m_2) (e_4 + e_5) + (m_2 \lambda_M - m_3) = 0$$

$$\text{Define } M_0 = m_0 \lambda_M - m_1 \quad M_2 = m_2 \lambda_M - m_3$$

$$M_1 = m_1 \lambda_M - m_2 \quad M_4 = m_3 \lambda_M - m_4$$

$$\text{Clearly } M_0 e_4 e_5 - M_1 (e_4 + e_5) + M_2 = 0$$

$$\text{or } (e_4 - M_1/M_0) (e_5 - M_1/M_0) = (M_1/M_0)^2 - M_2/M_0$$

and after some manipulation the second may be written as

$$(e_4 - M_1/M_0) (e_5 - M_1/M_0) (e_4 + e_5 + M_1/M_0) = (M_1/M_0)^3 - M_3/M_0.$$

$$\text{Define } K_1 = M_1/M_0, K_2 = (M_1/M_0)^2 - M_2/M_0, K_3 = (M_1/M_0)^3 - M_3/M_0$$

$$\text{with } x = e_4 - M_1/M_0 \text{ and } y = e_5 - M_1/M_0.$$

$$\text{As before form the quadratic } x^2 + (3K_1 - K_3/K_2) x + K_2 = 0$$

whose roots are $\hat{x}$, and $K_2/\hat{x} = \hat{y}$, from here the process of back

substitution proceeds to find $C_{\max}^4(m_0, \dots, m_4, \lambda_M)$.

This concludes the analysis for all the cases $r = 0, \dots, 4$,
the case $r = 5$ has been attempted but does not appear tractable.

It will also be shown later that the computation involved in
decomposing the moment m_5 is formidable for a large system .

The only result which has been used but not proved here is one
concerning the behaviour of the sequence of moment ratios namely

$$0 < \frac{m_1}{m_0} < \frac{m_1}{m_0} < \frac{m_2}{m_1} < \dots < \frac{m_4}{m_3} < \lambda_M.$$

The proof of this result is to be found in section 3.4, along with
other moment and bound formulae properties.

3.3 Analytical Bound Formulae for the Diagnostic IESQ

The development for the analysis of the IESQ diagnostic follows that of ΔJ_0 . Hilbert space canonical forms are derived for IESQ from which a finite dimensional optimization problem is deduced. The solution of this is found to be closely related that of the same order problem for ΔJ_0 , thus enabling the calculation of IESQ bounds at almost no extra computational expense.

$$\text{Recall } \text{IESQ} = \frac{1}{2} \langle \hat{y}, Q \hat{y} \rangle_{H_r}$$

$$\text{where } \hat{y} = W \hat{u} + d, \quad \hat{u} = -A^{-1}b, \quad A = W^* Q W + R$$

$$\text{and } b = W^* Q d.$$

$$\text{Hence } \text{IESQ} = \frac{1}{2} \langle W A^{-1} b, Q W A^{-1} b \rangle_{H_r} - \langle W A^{-1} b, Q d \rangle_{H_r} + \frac{1}{2} \langle d, Q d \rangle_{H_r}$$

Using Householder's equality write

$$A^{-1} = (W^* Q W + R)^{-1} = R^{-\frac{1}{2}} (I + R^{-\frac{1}{2}} W^* Q^{-\frac{1}{2}} Q^{\frac{1}{2}} W R^{-\frac{1}{2}})^{-1} R^{-\frac{1}{2}}$$

$$\text{Define } V = Q^{\frac{1}{2}} W R^{-\frac{1}{2}}$$

$$\text{then } A^{-1} = R^{-\frac{1}{2}} (I + V^* V)^{-1} R^{-\frac{1}{2}} = R^{-\frac{1}{2}} (I - V^* (I + V V^*)^{-1} V) R^{-\frac{1}{2}}$$

$$\text{As in section 3.1, define } D \triangleq V V^* = Q^{\frac{1}{2}} W R^{-1} W^* Q^{\frac{1}{2}}$$

$$\text{and } Z = Q^{\frac{1}{2}} d. \quad \text{Obtain then for the terms of IESQ,}$$

$$\begin{aligned} \text{i) } \langle A^{-1} b, W^* Q d \rangle_{H_m} &= \langle Q^{\frac{1}{2}} d, Q^{\frac{1}{2}} W A^{-1} W^* Q^{\frac{1}{2}} Q^{\frac{1}{2}} d \rangle_{H_r} \\ &= \langle Z, V (I - V^* (I + V V^*)^{-1} V) V^* Z \rangle_{H_r} \\ &= \langle Z, (D - D(I + D)^{-1} D) Z \rangle_{H_r} \end{aligned}$$

$$\begin{aligned} \text{and ii) } \langle W A^{-1} b, Q W A^{-1} b \rangle_{H_r} &= \langle Q^{\frac{1}{2}} d, Q^{\frac{1}{2}} W A^{-1} W^* Q W A^{-1} W^* Q^{\frac{1}{2}} Q^{\frac{1}{2}} d \rangle_{H_r} \\ &= \langle Z, V (I - V^* (I + V V^*)^{-1} V) V^* V (I - V^* (I + V V^*)^{-1} V) V^* Z \rangle \\ &= \langle Z, (D - D(I + D)^{-1} D)^2 Z \rangle_{H_r} \end{aligned}$$

Recall from section 3.1 the various eigenvalue/vector properties of compact, self-adjoint, semipositive definite Hilbert space operator $D = Q^{\frac{1}{2}} W R^{-1} W^* Q^{\frac{1}{2}}$. Using these properties and with the same notation write:

$$Z = Q^{\frac{1}{2}} d = \alpha_o w_o + \sum_{i=1}^{\infty} \alpha_i w_i = \sum_{i=0}^{\infty} \alpha_i w_i$$

where $\{w_i\}_{i=1}^{\infty}$ form a complete set of orthonormal eigenfunctions and $\alpha_o w_o$ denotes the possible null space component of Z .

$$\text{Thus } [D - D(I+D)^{-1} D] Z = \sum_{i=1}^{\infty} \frac{\alpha_i \lambda_i w_i}{(1+\lambda_i)}$$

$$\text{and } [D - D(I+D)^{-1} D]^2 Z = \sum_{i=1}^{\infty} \alpha_i \lambda_i^2 w_i / (1+\lambda_i)^2$$

$$\begin{aligned} \text{Hence } \text{IESQ} &= \frac{1}{2} \sum_{i=1}^{\infty} \frac{\alpha_i^2 \lambda_i^2}{(1+\lambda_i)^2} - \sum_{i=1}^{\infty} \frac{\alpha_i^2 \lambda_i}{(1+\lambda_i)} + \frac{1}{2} \sum_{i=0}^{\infty} \alpha_i^2 \\ &= \frac{1}{2} \alpha_o^2 + \frac{1}{2} \sum_{i=1}^{\infty} \frac{\alpha_i^2}{(1+\lambda_i)^2} \end{aligned}$$

The related moment expansions are $\frac{1}{2} \sum_{i=0}^{\infty} \alpha_i^2 = m_o$

and $\frac{1}{2} \sum_{i=1}^{\infty} \alpha_i^2 (\lambda_i)^j = m_j$; $j=1, 2, \dots, r$ with $\{\alpha_i^2 \geq 0\}_{i=1}^{\infty} = 0$

and $\{0 < \lambda_i \leq \lambda_M\}_{i=1}^{\infty}$. The problem of bounding IESQ subject to the

above moment constraints is considered. A reduction in dimensionality is obtained using reasoning completely analogous to that used for Theorem 3.1.1. (In particular it is noted that the cost IESQ has a similar linear form if the coefficient c_i are defined as $c_o = 1$ and $c_i(\lambda) = 1/(1+\lambda_i)^2$ $i=1, 2, \dots, \infty$.)

As before the finite dimensional problem can be basic in one of two ways

either a) $R^r(\alpha, e) = \alpha_0 + \sum_{i=1}^r \alpha_i / (1+e_i)^2$

subject to $\sum_{i=0}^r \alpha_i = m_0, \quad \sum_{i=1}^r \alpha_i (e_i)^j = m_j, \quad j=1,2,\dots,r$

with $\{\alpha_i \geq 0\}_{i=0}^r$, and $\{0 < e_i \leq \lambda_M\}_{i=1}^r$

or b) $R^r(\alpha, e) = \sum_{i=1}^{r+1} \alpha_i / (1+e_i)^2$ subject to $\sum_{i=1}^{r+1} \alpha_i = m_0,$

and $\sum_{i=1}^{r+1} \alpha_i (e_i)^j = m_j; \quad j=1,2,\dots,r$ with $\{\alpha_i \geq 0\}_{i=1}^{r+1}$

and $\{0 < e_i \leq \lambda_M\}_{i=1}^{r+1}$. Again with the modified constraint

$\{0 \leq e_i \leq \lambda_M\}_{i=1}^{r+1}$ the case b) subsumes that of a) to

give the finite dimensional optimization problems associated with IESQ as :

rth Order Problem Extremise $R^r(\alpha, e) = \sum_{i=1}^{r+1} \alpha_i / (1+e_i)^2$

subject to $\sum_{i=1}^{r+1} \alpha_i = m_0$

$\sum_{i=1}^{r+1} \alpha_i (e_i)^j = m_j; \quad j=1,2,\dots,r$

with $\{\alpha_i \geq 0 \text{ and } 0 \leq e_i \leq \lambda_M\}_{i=1}^{r+1}$

The optimal solutions of this set of problems are intimately related to those of the corresponding finite dimensional problems posed for ΔJ_0 , since order for order the optimal feasible solutions are identical. This is a consequence of both problems having the same constraints, hence for the higher order problems there will be considerable saving in

computation time. For the low order cases, $r=0$, and 1 explicit bound formulae are given, but for the higher order cases application of the feasible optimal points from the analysis given for ΔJ_0 (Section 3.2) is assumed.

$$\text{Case } r = 0 \quad \text{Extremise } R^0(\alpha, e) = \frac{\alpha_1}{(1+e_1)^2}$$

$$\text{subject to } \alpha_1 = m_0, \quad \text{and} \quad 0 \leq e_1 \leq \lambda_M$$

$$\text{Clearly } 0 \leq R_{\min}^0(m_0, \lambda_M) = \frac{m_0}{(1+\lambda_M)^2} \leq \text{IESQ} \leq R_{\max}^0(m_0) = m_0.$$

$$\text{Case } r = 1 \quad R_{\min}^1(m_0, m_1) \leq \text{IESQ} \leq R_{\max}^1(m_0, m_1, \lambda_M)$$

$$\text{where } R_{\min}^1(m_0, m_1) = m_0^3 / (m_0 + m_1)^2$$

$$\text{and } R_{\max}^1(m_0, m_1, \lambda_M) = m_0 - m_1 (\lambda_M + 2) / (1 + \lambda_M)^2$$

It has been shown how a similar analysis to that used for the ΔJ_0 diagnostic, also produces explicit bound formulae for the diagnostic IESQ. A most important feature being the use of identical feasible optimal solutions, the consequence being a formidable saving in computation.

3.4 Properties of the Moments and Bounds

The moment sequences, and bounds have some interesting theoretical features which may be used to ensure the consistency of their numerical values obtained by approximation techniques. First the moment sequences are shown to appear in only four possible variants.

Recall the moment definitions as :

$$m_k = \frac{1}{2} \langle Z, D^k Z \rangle_{H_r} \quad k = 0, 1, \dots,$$

$$\text{where } Z = Q^{\frac{1}{2}} d \in H_r \quad \text{and} \quad D = Q^{\frac{1}{2}} W R^{-1} W^* Q^{\frac{1}{2}}.$$

The compactness of D enables a direct sum decomposition for H_r as

$$H_r = H_1 \oplus H_2, \quad \text{where } H_1 \text{ is spanned by the complete set of}$$

ortho-normal eigenfunctions $\{w_i\}_{i=1}^{\infty}$, and H_2 is the nullity of

operator D. A spectral expansion for $Z = Q^{\frac{1}{2}} d \in H_r$ is obtained as

$$Z = \alpha_0 w_0 + \sum_{i=1}^{\infty} \alpha_i w_i, \quad \text{where } \alpha_0 w_0 \text{ is the null space component}$$

($\alpha_0 w_0 \in H_2$) of Z. The moments have the canonical forms :

$$m_0 = \frac{1}{2} \alpha_0^2 + \frac{1}{2} \sum_{i=1}^{\infty} \alpha_i^2 \triangleq \sum_{i=0}^{\infty} a_i^2$$

$$\text{and } m_j = \frac{1}{2} \sum_{i=1}^{\infty} \alpha_i^2 (\lambda_i)^j \triangleq \sum_{i=1}^{\infty} a_i^2 (\lambda_i)^j \quad j = 1, 2, \dots$$

Assume the eigenvalues are ordered so that

$$\lambda_M = \lambda_1 > \lambda_2 > \lambda_3 > \lambda_4 > \dots > \lambda_N > \lambda_{N+1} > \dots > 0 = \lambda_0$$

then these may be partitioned to form three sets :

$$\{ \lambda_i : \lambda_i \in (1, \lambda_M], \text{ associated indices } i = 1, 2, \dots, N-1 \}$$

$$\{ \lambda_i : \lambda_i = 1, \text{ associated index } i = N \}$$

and $\{ \lambda_i : \lambda_i \in (0,1) , \text{ associated indices } i = N+1, N+2, \dots \}$

Using these sets, moments m_k , and m_{k+1} are written as

$$m_k = \sum_{i=1}^{N-1} a_i^2 \lambda_i^k + a_N^2 \lambda_N^k + \sum_{i=N+1}^{\infty} a_i^2 (\lambda_i)^k$$

$$m_{k+1} = \sum_{i=1}^{N-1} a_i^2 \lambda_i^{k+1} + a_N^2 \lambda_N^{k+1} + \sum_{i=N+1}^{\infty} a_i^2 (\lambda_i)^{k+1} \quad k = 1, 2, \dots$$

the moment m_0 may be adjoined as :

$$m_0 = a_0^2 + \sum_{i=1}^{N-1} a_i^2 + a_N^2 + \sum_{i=N+1}^{\infty} a_i^2.$$

Theorem 3.4.1 The moment sequence $\{ m_0, m_1, \dots \}$ has only four

behavioural variants :

- (i) $m_1 > m_2 > m_3 > \dots > m_k > m_{k+1} > \dots$
- (ii) $m_0 \geq m_1 = m_2 = m_3 \dots = m_k = m_{k+1} = \dots$
- (iii) $m_1 < m_2 < m_3 < m_4 \dots < m_k < m_{k+1} < \dots$
- (iv) $m_1 > m_2 > m_3 > \dots > m_k \leq m_{k+1} < m_{k+2} < \dots$

Proof

(i) Consider $a_1 = 0$; $i = 1, 2, \dots, N-1$ hence the moments reduce to

$$m_0 = a_0^2 + \sum_{i=N}^{\infty} a_i^2, \quad m_k = \sum_{i=N}^{\infty} a_i^2 (\lambda_i)^k \quad \text{and} \quad m_{k+1} = \sum_{i=N}^{\infty} a_i^2 (\lambda_i)^{k+1}$$

where $a_0^2 \geq 0$, $\lambda_N = 1$, $\lambda_i \in (0,1)$ $i=N+1, N+2, \dots$ and $k=1, 2, 3, \dots$

Hence $\lambda_i^{k+1} < \lambda_i^k$, $k = 1, 2, \dots$ and $i = N+1, N+2, \dots$

implying $m_k > m_{k+1}$, $k = 1, 2, \dots$

For $k = 0$, and $k = 1$ $m_0 = a_0^2 + \sum_{i=N}^{\infty} a_i^2$ and $m_1 = \sum_{i=N}^{\infty} a_i^2 \lambda_i$

from which no ordering conclusion can be drawn.

(ii) Consider $a_i = 0$, $i = 1, 2, 3, \dots, N-1, N+1, \dots$

so that $m_0 = a_0^2 + a_N^2$, $m_1 = a_N^2$, $m_k = a_N^2 = m_{k+1}$ $k=1, 2, \dots$

to give $m_0 \geq m_1 = m_2 = \dots$ as required.

(iii) Consider $a_i = 0$, $i = N+1, N+2, \dots$ Here the moments are

$$m_0 = a_0^2 + \sum_{i=1}^N a_i^2, \quad m_k = \sum_{i=1}^N a_i^2 (\lambda_i)^k, \quad \text{and} \quad m_{k+1} = \sum_{i=1}^N a_i^2 (\lambda_i)^{k+1}$$

where $a_0^2 \geq 0$, $\lambda_i \in (1, \lambda_M]$, $i = 1, 2, \dots, N-1$; $\lambda_N = 1$

and $k = 1, 2, \dots$

Hence $(\lambda_i)^k < (\lambda_i)^{k+1}$ $i = 1, 2, \dots, N-1$; $k = 1, 2, \dots$

so that $m_k < m_{k+1}$ $k = 1, 2, \dots$ (as required)

For $k = 0$, and $k = 1$ no conclusion may be drawn as to the ordering of moments m_0 , and m_1 .

(iv) The remaining case to be considered is that where λ_i takes values in all three sets. Let $I_1 \triangleq \{i : i \in 1, 2, \dots, N-1\}$, and

$I_{N+1} \triangleq \{i : i \in N+1, N+2, \dots\}$, then assuming $a_i \neq 0$ for some

$i \in I_1$, $i = N$, and some $i \in I_{N+1}$, the moments take the form

$$m_0 = \sum_{i=0}^{\infty} a_i^2, \quad m_k = \sum_{i \in I_1} a_i^2 (\lambda_i)^k + a_N^2 \lambda_N^k + \sum_{i \in I_{N+1}} a_i^2 (\lambda_i)^k,$$

where $k = 1, 2, \dots$

For the moments m_0 and m_1 , no ordering conclusion may be drawn.

First show that $m_k < m_{k+1}$ implies $m_{k+1} < m_{k+2}$ $k = 1, 2, 3, \dots$

Inductively assume $m_{k+1} > m_k$

$$\text{then } m_{k+1} - m_k = \sum_{i \in I_1} a_i^2 \lambda_i^k (\lambda_i - 1) + \sum_{i \in I_{N+1}} a_i^2 \lambda_i^k (\lambda_i - 1) > 0$$

$$\text{hence } \sum_{i \in I_1} a_i^2 \lambda_i^k (\lambda_i - 1) > \sum_{i \in I_{N+1}} a_i^2 \lambda_i^k (1 - \lambda_i) > 0$$

Using the properties of λ_i according to whether index $i \in I_1$

or I_{N+1} obtain

$$\begin{aligned} \sum_{i \in I_1} a_i^2 \lambda_i^{k+1} (\lambda_i - 1) &> \sum_{i \in I_1} a_i^2 \lambda_i^k (\lambda_i - 1) > \sum_{i \in I_{N+1}} a_i^2 \lambda_i^k (1 - \lambda_i) \\ &> \sum_{i \in I_{N+1}} a_i^2 \lambda_i^{k+1} (1 - \lambda_i) > 0 \end{aligned}$$

$$\text{so that } m_{k+2} - m_{k+1} = \sum_{i \in I_1} a_i^2 \lambda_i^{k+1} (\lambda_i - 1) + \sum_{i \in I_{N+1}} a_i^2 \lambda_i^{k+1} (\lambda_i - 1) > 0$$

The converse is not necessarily true. In fact for the distribution of linear coefficients assumed for this case if $m_k > m_{k+1}$ for the first few values of k ($= 1, 2, \dots$) the sequence will eventually reverse to become monotonic increasing. This can be seen from the expression

$$m_{k+1} - m_k = \sum_{i \in I_1} a_i^2 \lambda_i^k (\lambda_i - 1) + \sum_{i \in I_{N+1}} a_i^2 \lambda_i^k (\lambda_i - 1).$$

If the right hand side is initially negative (so that $m_k > m_{k+1}$)

this implies the second of the two terms dominates, but as k increases

this second term can only become smaller in magnitude (since

$\lambda_i \in (0, 1) \quad \forall i \in I_{N+1}$) until an index is reached at which the first

term dominates giving $m_{k+1} > m_k$ and consequently from the inductive

argument all subsequent moments obey this type of inequality. This concludes the proof.

That the moment sequences have only four behavioural configurations can be used to ensure a sufficiently accurate numerical approximation to the convolution operators has been obtained.

The second result of interest concerns the limit of the sequence of moment ratios which is shown to be the largest eigenvalue λ_M .

Theorem 3.4.2 Consecutive moment ratios satisfy

$$0 < \frac{m_1}{m_0} \leq \frac{m_2}{m_1} \leq \dots \leq \frac{m_{k+1}}{m_k} < \dots < \lambda_M$$

and $\lim_{k \rightarrow \infty} m_{k+1}/m_k = \lambda_M$

Proof Recall $Z = \sum_{i=0}^{\infty} \alpha_i w_i$, where it is assumed that $\alpha_1 \neq 0$

(in all numerical implementations of the scheme rounding error usually ensures this is the case).

Then $m_k = \sum_{i=1}^{\infty} a_i^2 (\lambda_i)^k$ and $m_{k+1} = \sum_{i=1}^{\infty} a_i^2 (\lambda_i)^{k+1}$, $k = 1, 2, \dots$

$$\begin{aligned} \text{so that } \frac{m_{k+1}}{m_k} &= \frac{a_1^2 \lambda_1^{k+1} + \sum_{i=2}^{\infty} a_i^2 (\lambda_i)^{k+1}}{a_1^2 \lambda_1^k + \sum_{i=2}^{\infty} a_i^2 (\lambda_i)^k} \\ &= \lambda_M \frac{\left[1 + \sum_{i=2}^{\infty} (a_i/a_1)^2 (\lambda_i/\lambda_M)^{k+1} \right]}{\left[1 + \sum_{i=2}^{\infty} (a_i/a_1)^2 (\lambda_i/\lambda_M)^k \right]} \end{aligned}$$

Define $e_i = a_i/a_1$, and $v_i = \lambda_i/\lambda_M$ $i = 2, 3, \dots$

to write $d_k = \sum_{i=2}^{\infty} e_i^2 v_i^k$

Hence $m_{k+1}/m_k = \lambda_M (1 + d_{k+1}) / (1 + d_k)$

$$\begin{aligned} \text{so that } \Delta = \frac{m_{k+1}}{m_k} - \frac{m_k}{m_{k-1}} &= \lambda_M \left\{ \frac{(1+d_{k+1})}{(1+d_k)} - \frac{(1+d_k)}{(1+d_{k-1})} \right\} \\ &= \frac{\lambda_M (d_{k+1} - 2d_k + d_{k-1} + d_{k+1}d_{k-1} - d_k^2)}{(1+d_k)(1+d_{k-1})} \end{aligned}$$

In detail for the numerator

$$d_{k+1} - 2d_k + d_{k-1} + d_{k-1}d_{k+1} - d_k^2 = \sum_{i=2}^{\infty} e_i^2 v_i^{k-1} (v_i - 1)^2 +$$

$$\left(\sum_{i=2}^{\infty} e_i^2 v_i^{k+1} \right) \left(\sum_{i=2}^{\infty} e_i^2 v_i^{k-1} \right) - \left(\sum_{i=2}^{\infty} e_i^2 v_i^k \right)^2$$

Which on collating the terms of the series gives :

$$= \sum_{i=2}^{\infty} e_i^2 v_i^{k-1} (v_i - 1)^2 + \sum_{i=2}^{\infty} e_i^2 v_i^{k-1} \sum_{j=i+1}^{\infty} e_j^2 v_j^{k-1} (v_i - v_j)^2 > 0 \quad k=2,3,$$

Noting that the above analysis is also valid for $k = 1$, using

$$\bar{m}_0 = m_0 - a_0^2 = \sum_{i=1}^{\infty} a_i^2 \quad \text{enables the result to be extended to include } m_0.$$

$$\text{Thus } \frac{m_{k+1}}{m_k} - \frac{m_k}{m_{k-1}} > 0 \quad k = 1, 2, \dots$$

The limit is obtained directly from $v_i \in (0,1)$ so that as $k \rightarrow \infty$,

$v_i^k \rightarrow 0$, and consequently $d_k \rightarrow 0$ hence

$$\lim_{k \rightarrow \infty} \frac{m_{k+1}}{m_k} = \lim_{k \rightarrow \infty} \lambda_M \frac{(1 + d_{k+1})}{(1 + d_k)} = \lambda_M$$

△

The behaviour of successive moment ratios was frequently used in the proofs of the bound formulae, not only to obtain the bounds but to verify their ordering. The limit property of the sequence was also investigated when the problem of obtaining an upper bound to the largest eigenvalue was considered.

Finally some properties of the bound formulae are demonstrated for the input diagnostic ΔJ_0 . Naturally only the low order formulae are amenable to manipulation, but the properties have been shown numerically in the high order bound formulae. Such numerical examples appear in the Section 3.6. Although the details have been omitted, similar properties are exhibited by the bound formulae of the IESQ diagnostic.

The two properties are (i) the bounds nested according to increasing order, and (ii) the r th order bound which is a function of λ_M , limits as $\lambda_M \rightarrow \infty$ at the respective $(r-1)$ th order bound which will be independent of λ_M . Examining the cases in order :

$$\text{Case } r = 0 \quad C_{\min}^0 = 0 < \Delta J_0 < C_{\max}^0(m_0, \lambda_M) = \frac{m_0 \lambda_M}{1 + \lambda_M}$$

$$\text{Set } \lambda \geq \lambda_M, \text{ then } C_{\max}^0(m_0, \lambda) \geq C_{\max}^0(m_0, \lambda_M)$$

$$\lim_{\lambda \rightarrow \infty} C_{\max}^0(m_0, \lambda) = m_0$$

$$\text{Hence } 0 < \Delta J_0 < \frac{m_0 \lambda_M}{(1 + \lambda_M)} \leq \frac{m_0 \lambda}{(1 + \lambda)} < m_0$$

$$\text{Case } r = 1 \quad C_{\min}^1(m_1, \lambda_M) < \Delta J_0 < C_{\max}^1(m_0, m_1)$$

$$\text{where } C_{\min}^1(m_1, \lambda_M) = \frac{m_1}{(1 + \lambda_M)} \text{ and } C_{\max}^1(m_0, m_1) = \frac{m_0 m_1}{(m_0 + m_1)}$$

$$\text{Thus } \lambda \geq \lambda_M \quad 0 < C_{\min}^1(m_1, \lambda) \leq C_{\min}^1(m_1, \lambda_M)$$

$$\lim_{\lambda \rightarrow \infty} C_{\min}^1(m_1, \lambda) = 0 = C_{\min}^0$$

$$\text{Also } C_{\max}^0(m_0, \lambda_M) = \frac{m_0 \lambda_M}{(1 + \lambda_M)} \text{ and } 0 < \frac{m_1}{m_0} < \lambda_M$$

$$\text{Thus } \infty > 1/(m_1/m_0) > 1/\lambda_M > 0 \Rightarrow 0 < \frac{m_1}{(m_0 + m_1)} < \frac{\lambda_M}{(1 + \lambda_M)} < 1$$

$$\text{Hence } C_{\max}^1(m_0, m_1) < C_{\max}^0(m_0, \lambda_M)$$

Thus between cases $r = 0$, and $r = 1$ exist the following relationship:

$$0 = C_{\min}^0 < C_{\min}^1(m_1, \lambda_M) < \Delta J_0 < C_{\max}^1(m_0, m_1) < C_{\max}^0(m_0, \lambda_M) < m_0$$

$$\text{with } \lim_{\lambda_M \rightarrow \infty} C_{\min}^1(m_1, \lambda_M) = C_{\min}^0 \quad \lim_{\lambda_M \rightarrow \infty} C_{\max}^0(m_0, \lambda_M) = m_0$$

Case $r = 2$ $C_{\min}^2(m_1, m_2) < \Delta J_0 < C_{\max}^2(m_0, m_1, m_2, \lambda_M)$

For the minimum recall $C_{\min}^1(m_1, \lambda_M) = \frac{m_1}{(1+\lambda_M)}$

also that $0 < m_2/m_1 < \lambda_M$. Using an argument similar to that used for $C_{\max}^1$ and $C_{\max}^0$ it is straightforward to show

$$C_{\min}^1(m_1, \lambda_M) < C_{\min}^2(m_1, m_2)$$

For the maximum recall $C_{\max}^2$

$$\begin{aligned} C_{\max}^2(m_0, \dots, m_2, \lambda_M) &= \frac{-m_0 m_1 \lambda_M^2 + (m_2 - m_1) m_0 \lambda_M + m_1^2}{(1+\lambda_M)(m_2 + m_1 - (m_1 + m_0) \lambda_M)} \\ &= \left(\frac{m_0 m_1}{m_0 + m_1} \right) \frac{(\lambda_M^2 + (1 - m_2/m_1) \lambda_M - m_1/m_0)}{(1+\lambda_M)(\lambda_M - (m_2 + m_1)/(m_1 + m_0))} \end{aligned}$$

$$C_{\max}^2(m_0, \dots, m_2, \lambda_M) = \left(\frac{m_0 m_1}{m_0 + m_1} \right) \left(1 - f(m_0, \dots, m_2, \lambda_M) \right)$$

$$\text{where } f(m_0, \dots, m_2, \lambda_M) = \frac{\left(\frac{m_2}{m_1} - \left(\frac{m_1 + m_2}{m_0 + m_1} \right) \right) \lambda_M + \frac{m_1}{m_0} - \frac{m_1 + m_2}{m_0 + m_1}}{(1+\lambda_M) \left(\lambda_M - \left(\frac{m_1 + m_2}{m_0 + m_1} \right) \right)}$$

Rationalize the numerator to obtain

$$f(m_0, \dots, m_2, \lambda) = \frac{\frac{m_0}{(m_0 + m_1)} \left(\frac{m_2}{m_1} - \frac{m_1}{m_0} \right) \left(\lambda - \frac{m_1}{m_0} \right)}{(1+\lambda) \left(\lambda - \left(\frac{m_1 + m_2}{m_0 + m_1} \right) \right)}$$

$$\text{Using the identity } 0 < \frac{m_1}{m_0} < \frac{m_1 + m_2}{m_0 + m_1} < \frac{m_2}{m_1} < \lambda_M \leq \lambda$$

$$\text{to give } f(m_0, \dots, m_2, \lambda_M) \geq f(m_0, \dots, m_2, \lambda) \geq 0$$

$$\text{and } \lim_{\lambda \rightarrow \infty} f(m_0, \dots, m_2, \lambda) = 0$$

$$\text{Hence } C_{\max}^2(m_0, \dots, m_2, \lambda_M) < \frac{m_0 m_1}{(m_0 + m_1)} = C_{\max}^1(m_0, m_1)$$

$$\text{and } \lim_{\lambda_M \rightarrow \infty} C_{\max}^2(m_0, \dots, m_2, \lambda_M) = C_{\max}^1(m_0, m_1)$$

Consequently between the cases $r = 1$ and $r = 2$ exist the following relationships between the bounds :

$$0 < C_{\min}^1(m_1, \lambda_M) < C_{\min}^2(m_1, m_2) < \Delta J_0 < C_{\max}^2(m_0, \dots, m_2, \lambda_M) < C_{\max}^1(m_0, m_1)$$

$$\text{with } \lim_{\lambda_M \rightarrow \infty} C_{\max}^2(m_0, \dots, m_2, \lambda_M) = C_{\max}^1(m_0, m_1)$$

Using the explicit bound formulae for the higher order cases to demonstrate these properties proves too complicated. As already mentioned Section 3.6 contains numerical examples for these higher order cases. Clearly it should be feasible to isolate these properties from the general problem statement; this should provide a fruitful topic for further research. The importance of these properties should not be underestimated since the first guarantees that a closer set of bounds will result from a higher order problem, whilst the second quantifies the worst possible degradation obtainable by using a poor upper bound on the largest eigenvalue λ_M .

3.5 Asymptotic Properties of the Bound Formulae

The properties of the asymptotic optimal control problem, described in Chapter 1, were used to relate various multivariable control concepts to the values of the input diagnostics. This quantification is particularly important if the diagnostics are to be indicative of system control restrictions (for example, non-minimum phase).

In the proposed structure assessment procedure, the diagnostic values will be indicated by a set of bounds consequently the asymptotic behaviour of the bounds under the same limiting process (namely input cost reduced to zero) used to expose the control restrictions is of importance.

The limiting behaviour of the bound formulae derived for ΔJ_0 , and IESQ from the finite dimensional optimization problems can be ascertained using two different analyses.

The first is to incorporate the asymptotic cost parameter into the moments $m_0, m_1, \dots$, and the maximum eigenvalue λ_M and use bound formulae directly to obtain their asymptotic behaviour. Since only low order cases can be examined (cases $r = 0, 1$, and 2) the results can be interpreted geometrically using the concepts of Bryant (1973).

A more fundamental approach, also detailed here, is to return to the Hilbert space formalism and obtain the asymptotic form of the infinite dimensional canonical form. As a consequence the results of the low order cases can then be identified in a more general setting.

For the initial analysis, recall the moment formulae as:

$$m_k = \frac{1}{2} \langle Z, D^k Z \rangle_{H_r} \quad k = 0, 1, \dots,$$

$$\text{where } Z = Q^{\frac{1}{2}} d \text{ and } D = Q^{\frac{1}{2}} W R^{-1} W^* Q^{\frac{1}{2}}$$

The asymptotic behaviour is given by setting

$$R = \rho R_0 \text{ and letting } \rho \rightarrow 0.$$

$$\text{Hence } m_k = \rho^{-k} \tilde{m}_k \text{ where } \tilde{m}_k = \frac{1}{2} \langle Z, D_0^k Z \rangle_{H_r}$$

$$\text{and } D_0 = Q^{\frac{1}{2}} W R_0^{-1} W^* Q^{\frac{1}{2}}. \text{ The limiting behaviour of the maximum}$$

eigenvalue is therefore obtained as

$$\lambda_M = \lim_{k \rightarrow \infty} \frac{m_{k+1}}{m_k} = \lim_{k \rightarrow \infty} \frac{\rho^{-(k+1)} \tilde{m}_{k+1}}{\rho^{-k} \tilde{m}_k} = \rho^{-1} \tilde{\lambda}_M$$

$$\text{where } \tilde{\lambda}_M = \lim_{k \rightarrow \infty} \tilde{m}_{k+1} / \tilde{m}_k$$

Dealing first with the low order bounds, that is $r = 0, 1$, and 2 , for the optimal cost improvement diagnostic, ΔJ_0 .

$$\text{Case } r = 0 \quad C_{\min}^0 = 0 \quad \text{and} \quad C_{\max}^0(m_0, \lambda_M) = \frac{m_0 \lambda_M}{(1 + \lambda_M)}$$

$$\text{Clearly } C_{\max}^0(m_0, \lambda_M) = \frac{m_0 \tilde{\lambda}_M}{(\tilde{\lambda}_M + \rho)} \rightarrow m_0 \text{ as } \rho \rightarrow 0$$

$$\text{Hence } 0 \leq \Delta J_0(\rho) \leq \frac{m_0 \tilde{\lambda}_M}{(\tilde{\lambda}_M + \rho)} \leq m_0$$

$$\text{Case } r = 1 \quad C_{\min}^1(m_1, \lambda_M) = \frac{m_1}{(1+\lambda_M)} ; \quad C_{\max}^1(m_0, m_1) = \frac{m_0 m_1}{(m_0 + m_1)}$$

$$\text{Clearly } C_{\min}^1(m_1, \lambda_M) = \frac{\tilde{m}_1}{(\tilde{\lambda}_M + \rho)} \rightarrow \frac{\tilde{m}_1}{\tilde{\lambda}_M} \text{ as } \rho \rightarrow 0$$

$$\text{and } C_{\max}^1(m_0, m_1) = \frac{m_0 \tilde{m}_1}{(\tilde{m}_1 + \rho m_0)} \rightarrow m_0 \text{ as } \rho \rightarrow 0$$

$$\text{Hence } 0 < \frac{\tilde{m}_1}{\tilde{\lambda}_M} < \frac{\tilde{m}_1}{(\tilde{\lambda}_M + \rho)} < \Delta J_0(\rho) < \frac{m_0 \tilde{m}_1}{(\tilde{m}_1 + \rho m_0)} < m_0 \quad (\rho > 0)$$

$$\text{Case } r = 2 \quad C_{\min}^2(m_1, m_2) = \frac{m_1^2}{(m_1 + m_2)} \text{ and}$$

$$C_{\max}^2(m_0, m_1, \dots, \lambda_M) = \frac{-m_0 m_1 \lambda_M^2 + (m_2 - m_1) m_0 \lambda_M + m_1^2}{(1+\lambda_M)(m_2 + m_1 - (m_1 + m_0) \lambda_M)}$$

$$\text{Here } C_{\min}^2(m_1, m_2) = \frac{\tilde{m}_1^2}{(\tilde{m}_2 + \rho \tilde{m}_1)} \rightarrow \frac{\tilde{m}_1^2}{\tilde{m}_2} \text{ as } \rho \rightarrow 0$$

$$\text{and } C_{\max}^2(m_0, m_1, \dots, \lambda_M) = \frac{-m_0 \tilde{m}_1 \lambda_M^2 + (\tilde{m}_2 - \rho \tilde{m}_1) m_0 \lambda_M + \rho \tilde{m}_1^2}{(\tilde{\lambda}_M + \rho)(\tilde{m}_2 + \rho \tilde{m}_1 - (\tilde{m}_1 + \rho \tilde{m}_0) \tilde{\lambda}_M)} \\ \rightarrow m_0 \text{ as } \rho \rightarrow 0$$

$$0 < \frac{\tilde{m}_1}{\tilde{\lambda}_M} < \frac{\tilde{m}_1^2}{\tilde{m}_2} < \frac{\tilde{m}_1^2}{(\tilde{m}_2 + \rho \tilde{m}_1)} < \Delta J_0(\rho) < C_{\max}^2(m_0, \tilde{m}_1, \dots, \tilde{\lambda}_M, \rho) < m_0$$

On the basis of these first few cases the finding is therefore that the lower bounds for $\Delta J_0(\rho)$ tend to a set of nested limit values whilst the upper bounds collapse into the upper bound m_0 . The graphical interpretation of Bryant (1973) shows succinctly this collapse. Consider then the case for $r = 1$ after normalization with respect to m_0 has occurred.

Extremise $\frac{\Delta J_o}{m_o} = \sum_{i=1}^2 \frac{\bar{\alpha}_i e_i}{(1+e_i)}$ subject to $\sum_{i=1}^2 \bar{\alpha}_i = 1$

and the linear constraint $\sum_{i=1}^2 \bar{\alpha}_i e_i = m_1/m_o$ where $\bar{\alpha}_i \geq 0$

and $0 \leq e_i \leq \lambda_M$. The points $(e_i/(1+e_i), e_i)$ lie on the concave curve $y = x/(1+x)$, and $\Delta J_o/m_o$ is a linear combination of the abscissa of these points whilst m_1/m_o is the same convex combination of the ordinates. It therefore follows that the point $(\Delta J_o/m_o, m_1/m_o)$ lies in the convex hull of the curve $y = x/(1+x)$, $x \in [0, \lambda_M]$, as shown in Figure 3.6.1. Clearly the bounds are given by $x/(1+x)$ evaluated at m_1/m_o , namely $m_1/(m_o + m_1)$ and (using ratios) by $m_1/m_o(1+\lambda_M)$. Introducing the asymptotic parameter, ρ then :

$$\lambda_M(\rho) = \rho^{-1} \tilde{\lambda}_M, \quad \frac{m_1}{m_o}(\rho) = \rho^{-1} \frac{\tilde{m}_1}{m_o}$$

and $\left(\frac{\Delta J_o}{m_o}\right)(\rho_1) \leq \frac{\Delta J_o}{m_o}(\rho_2) \leq \dots \leq \frac{\Delta J_o}{m_o} < \infty \quad \rho_1 > \rho_2 > \dots \geq 0$

The upper bound $B_u = \frac{m_1}{(m_o + m_1)}$ becomes $B_u(\rho) = \frac{1}{(1 + \rho m_o/\tilde{m}_1)}$

and the lower bound $B_L = \frac{m_1}{m_o(1+\lambda_M)}$ becomes $B_L(\rho) = \frac{\tilde{m}_1}{m_o(\tilde{\lambda}_M + \rho)}$

The figure 3.6.1 shows the trajectories of $B_u(\rho)$ and $B_L(\rho)$ and the collapse of $B_u(\rho)$ into the normalized value of unity.

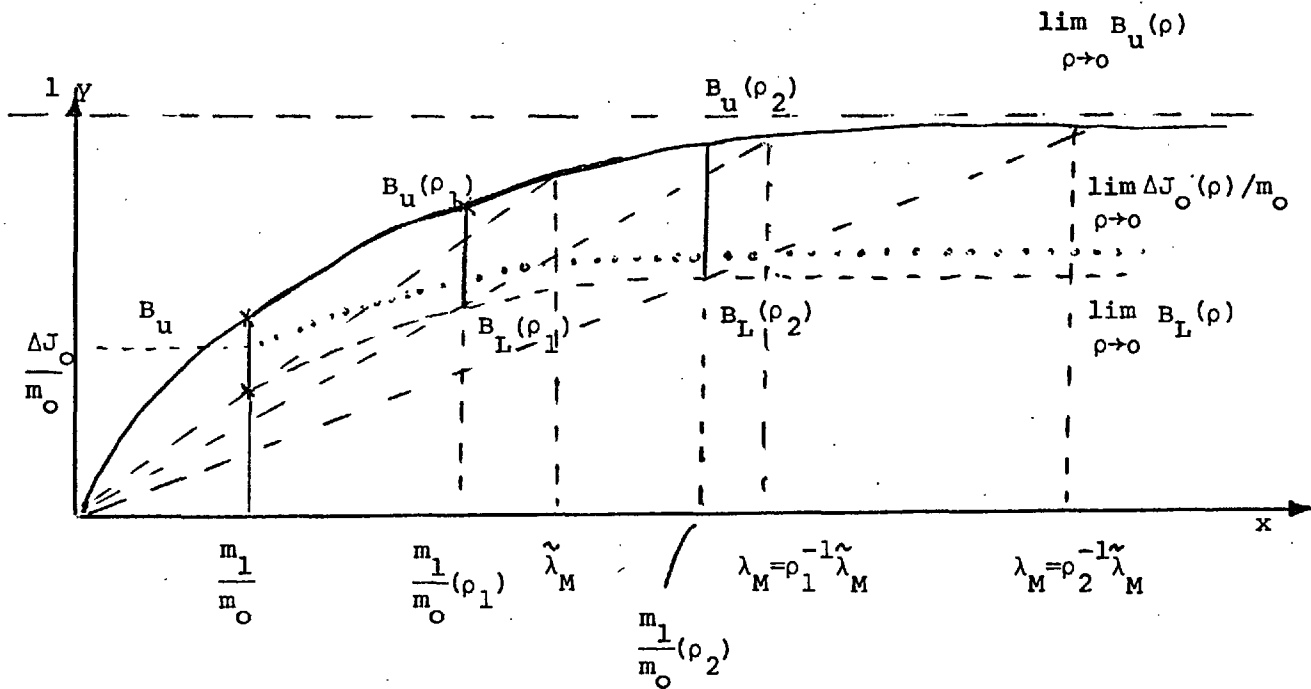


Figure 3.6.1 The Asymptotic Behaviour of the Diagnostic Bounds for ΔJ_0

As might be expected the bounds for the diagnostic, $IESQ(\rho)$ also exhibit a similar behaviour under the limiting operation $\rho \rightarrow 0$, here, the lower bound is seen to collapse into zero. The two low order cases $r = 0$, and $r = 1$ are examined in detail :

Case $r = 0$ $R_{\min}^0(m_0, \lambda_M) = \frac{m_0}{(1+\lambda_M)^2}$ and $R_{\max}^0(m_0) = m_0$

Clearly $R_{\min}^0(m_0, \lambda_M) = \frac{m_0}{(1 + \rho^{-1}\tilde{\lambda}_M)^2} = \frac{\rho^2 m_0}{(\tilde{\lambda}_M + \rho)^2}$

then $R_{\min}^0(m_0, \lambda_M) \rightarrow 0$ as $\rho \rightarrow 0$

Hence $0 \leq \frac{\rho^2 m_0}{(\tilde{\lambda}_M + \rho)^2} < IESQ(\rho) < m_0$

$$\text{Case } r = 1 \quad R_{\min}^1(m_0, m_1) = \frac{m_0^3}{(m_0 + m_1)^2}$$

$$\text{and } R_{\max}^1(m_0, m_1, \lambda_M) = m_0 - m_1(\lambda_M + 2)/(1 + \lambda_M)^2$$

$$\text{Then } R_{\min}^1(m_0, m_1) = \frac{m_0^3}{(m_0 + m_1)^2} = \frac{\rho^2 m_0^3}{(\tilde{m}_1 + \rho m_0)} \rightarrow 0 \text{ as } \rho \rightarrow 0$$

$$\text{and } R_{\max}^1(m_0, m_1, \lambda_M) = m_0 - \frac{m_1(\lambda_M + 2)}{(1 + \lambda_M)^2} = m_0 - \frac{\tilde{m}_1(\tilde{\lambda}_M + 2\rho)}{(\tilde{\lambda}_M + \rho)^2}$$

$$\therefore R_{\max}^1(m_0, m_1, \lambda_M) \rightarrow m_0 - \tilde{m}_1/\tilde{\lambda}_M \text{ as } \rho \rightarrow 0$$

$$\text{Hence } 0 \leq \frac{\rho^2 m_0^3}{(\lambda_M + \rho)^2} < \text{IESQ}(\rho) < m_0 - \frac{\tilde{m}_1}{\tilde{\lambda}_M} < m_0$$

A similar analysis can be presented for the case $r = 2$, which exhibits similar features.

The asymptotic behaviour of the bounds is now pursued in the more general setting of the Hilbert space canonical form. In the notation of Section 3.1, the input diagnostic ΔJ_0 is given by

$$\Delta J_0 = \frac{1}{2} \langle b, A^{-1} b \rangle_H = \frac{1}{2} \langle W^* Q d, (W^* Q W + R)^{-1} W^* Q d \rangle_{H_m}$$

Incorporate the asymptotic parameter $\rho > 0$ through $R \triangleq \rho R_0$

and set $Z = Q^{\frac{1}{2}} d$ and $D_0 = Q^{\frac{1}{2}} W^* R_0^{-1} W Q^{\frac{1}{2}}$ to obtain

$$\Delta J_0(\rho) = \frac{1}{2} \langle Z, [D_0 - D_0(\rho I + D_0)^{-1} D_0] Z \rangle_{H_r}$$

Operator D_0 is positive semi-definite, self adjoint and compact so that as before a spectral decomposition may be utilized to give

$\Delta J_O(\rho) = \frac{1}{2} \sum_{i=1}^{\infty} \alpha_i^2 \lambda_i / (\lambda_i + \rho)$ and similarly for IESQ to give

$\text{IESQ}(\rho) = \frac{1}{2} \alpha_O^2 + \frac{1}{2} \sum_{i=1}^{\infty} \alpha_i^2 \rho^2 / (\lambda_i + \rho)^2$ for which the moment

equations are $\frac{1}{2} \alpha_O^2 + \frac{1}{2} \sum_{i=1}^{\infty} \alpha_i^2 = m_O$ and $\frac{1}{2} \sum_{i=1}^{\infty} \alpha_i^2 (\lambda_i)^j = m_j$

$j = 1, 2, \dots, r$ where $\{\lambda_i\}_{i=1}^{\infty}$ are the denumerable eigenvalues

of operator D_O and $\{\alpha_i\}_{i=0}^{\infty}$ are the co-ordinate scalars of

$z = Q^{\frac{1}{2}} d$ with respect to $\{w_O, \{w_i\}_{i=1}^{\infty}\}$ where $w_O \in \mathcal{N}(D_O)$

and $\{w_i\}_{i=1}^{\infty}$ are the ortho-normal eigenfunctions of D_O ; hence

$\{\alpha_i^2 \geq 0\}_{i=0}^{\infty}$ and $\{0 < \lambda_i \leq \lambda_M\}_{i=1}^{\infty}$

The limiting values of the diagnostics are therefore

$\lim_{\rho \rightarrow 0} \Delta J_O(\rho) = \frac{1}{2} \sum_{i=1}^{\infty} \alpha_i^2$ and $\lim_{\rho \rightarrow 0} \text{IESQ}(\rho) = \frac{1}{2} \alpha_O^2$,

so that the two diagnostics complement each other as determined

in Chapter 2. Define $\Delta J_O(0) \triangleq \lim_{\rho \rightarrow 0} \Delta J_O(\rho) = \frac{1}{2} \sum_{i=1}^{\infty} \alpha_i^2$

After relabelling to remove the factor $\frac{1}{2}$, the bounding of $\Delta J_O(0)$

may be summarised as:

$$\text{Extremise } \Delta J_O(0) = \sum_{i=1}^{\infty} a_i^2 \quad 3.5.1$$

$$\text{subject to } a_O^2 + \sum_{i=1}^{\infty} a_i^2 = m_O \quad 3.5.2$$

$$\sum_{i=1}^{\infty} a_i^2 (\lambda_i)^j = m_j \quad ; j = 1, 2, \dots, r \quad 3.5.3$$

$$\text{where } \{a_i^2 \geq 0\}_{i=1}^{\infty} \quad \text{and} \quad \{0 < \lambda_i \leq \lambda_M\}_{i=1}^{\infty} \quad 3.5.4$$

The asymptotic cost $\Delta J_0(0)$ and the constraints 3.5.2 to 3.5.4 can be written as an infinite dimensional linear program so that reasoning analogous to that in Theorem 3.1.1 and the subsequent Corollary 3.1.1 can be used to obtain a reduction to finite dimensionality. Again the moment constraints suggest that the finite dimensional problems are basic in two forms:

$$a) \text{ Extremise } C^r(\alpha) = \sum_{i=1}^r \alpha_i \text{ subject to } \alpha_0 + \sum_{i=1}^r \alpha_i = m_0$$

$$\text{and } \sum_{i=1}^r \alpha_i (e_i)^j = m_j; j=1,2,\dots,r \text{ with } \{\alpha_i \geq 0\}$$

$$\text{and } \{0 < e_i \leq \lambda_M\}$$

$$\text{and b) Extremise } C^r(\alpha) = \sum_{i=1}^{r+1} \alpha_i \text{ subject to } \sum_{i=1}^{r+1} \alpha_i = m_0$$

$$\text{and } \sum_{i=1}^{r+1} \alpha_i (e_i)^j = m_j; j=1,2,\dots,r \text{ with } \{\alpha_i > 0\}$$

$$\text{and } \{0 < e_i \leq \lambda_M\}$$

Clearly all feasible solutions to the constraints of the form

$$b) \text{ will satisfy the } m_0 \text{ constraint, hence the cost } C^r(\alpha) = \sum_{i=1}^{r+1} \alpha_i$$

takes the constant value m_0 over all feasible solutions. Any

feasible solution to the constraints of the form a) will satisfy the m_0 constraint whence

$$C^r(\alpha) = \sum_{i=1}^r \alpha_i = m_0 - \alpha_0 \leq m_0.$$

Then if $\alpha_0 = 0$, the maximum value $\hat{C}^r = m_0$ is obtained and if

$\alpha_0 > 0$, a lower bound for $C^r(\alpha)$ is found, so that the form a) will provide a minimum. In summary the extrema are obtained as

$$\text{Max } C^r(\alpha \in \mathbb{R}^{r+1}) = m_0$$

and

$$\text{Min } C^r(\alpha) = \sum_{i=1}^r \alpha_i \quad r=1,2,\dots \quad 3.5.5.$$

$$\text{subject to } \alpha_0 + \sum_{i=1}^r \alpha_i = m_0 \quad 3.5.6.$$

$$\sum_{i=1}^r \alpha_i (\lambda_i)^j = m_j \quad 3.5.7.$$

$$\text{with } \alpha_0 > 0, \text{ and } \{\alpha_i \geq 0, 0 < \lambda_i \leq \lambda_M\}_{i=1}^r \quad 3.5.8.$$

The collapse of the upper bounds on $\Delta J_0(\rho)$ and the lower bounds on $\text{IESQ}(\rho)$ as $\rho \rightarrow 0$ has therefore been demonstrated in the more general setting of the infinite dimensional canonical form. Although an unfortunate theoretical property it is redeemed to some extent by the facts that the bounds retain their nesting, with the highest order bounds being the most accurate. By choosing $\rho > 0$ only sufficiently small enough to be satisfied that the desired output is achieved and that the consistency of $\text{IESQ}_s - \text{IESQ}_f$ is obtained for all structures, the more accurate higher order bounds are usually sufficiently disjoint to be able to eliminate the poorer structures and isolate candidates for further detailed investigation. In some cases the dispersion of the bounds is quite remarkable as will be seen in the case of a tension control scheme for a cold rolling steel mill exhibited in Chapter 5.

An alternative treatment to the optimizing of the infinite dimensional Hilbert space canonical forms subject to the given constraints is possible by permitting randomized solutions and then noting that the problem structure is almost identical

to that of calculating Tchebichev bounds (Whittle, 1973).

In fact, virtually all the results given in this chapter concerned with the bound formulae can be rederived using the schema of the Tchebichev problem. To the present, this new approach has yielded no new results apart from re-interpretation of existing results. One topic for future research is to discover whether there exists any relationship between the minimum of the asymptotic problem 3.5.5 to 3.5.8 as a function of its order (namely r) and the limiting value $\overline{\Delta J}_0 = \lim_{\rho \rightarrow 0} \Delta J_0(\rho)$. For

example it has been noted for the linear least squares problem, that if the input dimension is two, and four moment equations are used then the lower bound is identical to the actual optimal improvement.

3.6 A Finite Time Domain Example

The example discussed is a single input-single state system for which it is trivial enough to obtain analytical results for the moment values. Having compared analytical values, some results of a numerical trial are given; these raise several problems which will become more pertinent in the multivariable context.

The system is
$$x(t) = \int_0^t W(t-\tau) u(\tau) d\tau + d(t); t \in [0, T_f]$$

where the weighting function, impulse response or operator kernel is $W(t) = e^t$ $0 < t \leq T_f$ and the initial condition disturbance $d(t) = e^t x(0)$ $0 \leq t \leq T_f$.

Choose T_f (final time) = 1.0 and initial condition $x(0) = 1.0$.

The quadratic cost function is $J = \int_0^{T_f} (x^2(t) + u^2(t)) dt$.

The optimal improvement in control $\Delta J_0 = 1.505029658$ and the output regulation IESQ = 0.93467058.

The actual moment and moment ratio sequences are

$$\begin{aligned} m_0 &= 3.194528049 \\ m_1 &= 3.005240705 \\ m_2 &= 3.000220675 \\ m_3 &= 3.000010077 \\ m_4 &= 3.000000471 \end{aligned} \quad \begin{aligned} r_1 &= m_1/m_0 = 0.9407463820 \\ r_2 &= m_2/m_1 = 0.9983295747 \\ r_3 &= m_3/m_2 = 0.9999298060 \\ r_4 &= m_4/m_3 = 0.9999967978 \end{aligned}$$

The maximum eigenvalue λ_M is the limit of the moment ratio sequence so clearly $\lambda_M = 1.0$. Using this value of λ_M the following table was obtained:

Table 3.6.1 Table of Bounds: Using Analytical Moment Values

r	$\Delta J_0 = 1.505029658$		IESQ = 0.93467058	
0	0.0	1.597264025*	0.7986320124	3.194528049
1	1.502620352*	1.548497389	0.8481431041	0.940597521*
2	1.503876409	1.505065520*	0.934547774*	0.938084883
3	1.505029308*	1.505047134	0.9346019891	0.9346721215*
4	1.505029486	1.505029660*	0.934670569*	0.9346714247

'*' indicates that the bound formula is a function of λ_M .

The bounds for the quantity ΔJ_0 were subjected to variations in λ_M to obtain the following tables:

Table 3.6.2 Variation of the Upper Boundson ΔJ_0 with respect to λ_M .

r	$\lambda_M = 1.0$	$\lambda_M = 2.0$	$\lambda_M = 6.0$	$\lambda_M = 10.0$
0	1.597264025*	2.129685366	2.7381669	2.904116409
1	1.548497389	-	-	1.548497389
2	1.505065520*	1.532767928	1.541897417	1.544307669
3	1.505047134	-	-	1.505047135
4	1.505029660*	1.505041160	1.505044574	1.505045505

Table 3.6.3 Variation of the Lower Bounds on ΔJ_0 with respect to λ_M .

r	$\lambda_M = 1.0$	$\lambda_M = 2.0$	$\lambda_M = 6.0$	$\lambda_M = 10.0$
0	0.0	-	-	-
1	1.502620352*	1.001746902	0.4293201007	0.2732037004
2	1.503876409	-	-	1.503876409
3	1.505029308*	1.504277485	1.504048189	1.503985716
4	1.505029486	-	-	1.505029486

'-' indicates no change in the bound value.

These two tables clearly demonstrate the limiting behaviour of the bounds with respect to the value of parameter λ_M . The practical significance being that even if the least upper bound of λ_M is greatly in error, the r th order bound produced as a function λ_M cannot be any poorer than the (r-1)th order bound which is independent of λ_M .

The question arises as to how well these results will be emulated when they are produced using numerical approximations. In this context it is reasonable to suppose that a system impulse response will be obtained numerically from a measured system step response. For the following results the system step response was provided analytically as

$w_{\text{step}}(t) = e^t - 1.0 \quad 0 \leq t \leq T_f = 1.0$, and this was numerically differentiated to provide an impulse response matrix.

The convolutions that are required were performed using the trapezoidal rule of integration. The size of the integration

interval (denoted as T) used in the approximations enters as a parameter chosen to obtain a sensible balance between the accuracy of the approximation and the cost of work involved in computational manipulations. (This cost involves quantities like computing time, and computer storage requirements.) The moment and ratio values for this system using the above numerical procedure are given next.

Table 3.6.4 Moment and Moment Ratio Sequences for Different Approximation Interval Sizes

r	Moments			Moment Ratios		
	T=0.1	T=0.05	T=0.01	T=0.1	T=0.05	T=0.01
0	3.2052	3.1972	3.1946	-		
1	2.8471	2.9626	3.0034	0.8883	0.9266	0.9401
2	2.6869	2.9118	2.9963	0.9437	0.9829	0.9976
3	2.5386	2.8663	2.9940	0.9448	0.9844	0.9992
4	2.3986	2.8217	2.9919	0.9449	0.9845	0.9993

Before the calculation of the bounds can proceed an upper bound for λ_M must be chosen. In the case of the analytical operator the value of λ_M was seen to be unity, but when using a numerical approximation the value of the maximum eigenvalue or an upper bound on it has to be determined (in the absence of any a-priori formula) from the first few terms of the moment ratio sequence. The problem of finding a-priori bounds on λ_M assumes a greater importance in the multivariable context, since here the maximum eigenvalue λ_M becomes structure dependent. To return to the above numerical example: unity appears to be an upper bound on λ_M and was used to calculate the bounds. The fourth order bounds are given for the three approximation intervals along with the values for ΔJ_0 and IESQ obtained from equivalent numerical optimization solutions (; the posed optimization problems were solved using a conjugate gradient algorithm).

Table 3.6.5 Fourth Order Bounds for ΔJ_o obtained for Different Approximation Interval Sizes

T	ΔJ_o (Anal.) = 1.505029658		ΔJ_o (Numerical)
0.1	1.465441	1.465455	1.458883
0.05	1.495157	1.495169	1.493274
0.01	1.504646	1.5046549	1.504564

Table 3.6.6 Fourth Order Bounds for IESQ obtained for Different Approximation Interval Sizes

T	IESQ (Anal.) = 0.93467058		IESQ (Numerical)
0.1	0.984700	0.984764	0.991318
0.05	0.946376	0.946427	0.948311
0.01	0.935053	0.935091	0.935173

The two features of importance are that as T decreases in size, the results more closely resemble the analytical values and the bounds indicate optimal values of the same order as those obtained by solving the optimization problem numerically.

In this simple single-input, single output problem no structure complications occur; the application of the bound formulae was straightforward. The assessment of structures implies structure dependent moments, a maximum eigenvalue which is structure dependent and the computation of structure bounds, these are the problems to be solved for a successful implementation of the bound formulae.

3.7 Summary

The infeasibility of calculating the values of the diagnostics ΔJ_0 , and IESQ for all the input structure possibilities demands the development of a schema for displaying in some approximate form their respective values. Following the work of Bryant (1973) the emphasis was entirely on finding bounds for the diagnostic value. These were to be sufficiently disjoint to enable a structural assessment to be made.

Utilizing a Hilbert space canonical form for ΔJ_0 and IESQ, the bounds were constructed as the extrema of an infinite dimensional linear programming type problem. To assert the existence of basic feasible solutions a theorem similar to the Fundamental Theorem of Linear Programming was required. The resulting finite-dimensional optimization problems were subsequently solved using a technique of Hadley's (1970).

Following the derivation of explicit bound formulae various properties of the moment sequences, and the bounds were investigated. Perhaps the most significant of these was the asymptotic behaviour of the bounds which was seen to be theoretically detrimental to the structure assessment procedure. This area of work should provide several fruitful research topics and is at present the subject of on-going research.

Finally a simple finite time domain problem was used to demonstrate several of the properties isolated, and motivate several new problems particularly related to the algorithmic procedure of structures assessment to be discussed in the sequel.

Chapter 4 Aspects of Bound Computation for Structures

4.0 Introduction

4.1 Calculation of the Structure Moment Sequences

4.2 The Decomposition of the Moments

4.2.0 Decomposition of Moment, m_0

4.2.1 Decomposition of Moment, m_1

4.2.2 Decomposition of Moment, m_2

4.2.3 Decomposition of Moment, m_3

4.2.4 Decomposition of Moment, m_4

4.3 The Formation of Structure Moment Sequences

4.4 The Problem of an Upper Bound for the Maximum Eigenvalue, λ_M .

Introduction

4.4.1 Eigenvalue Bounds for Single Input-Single Output Systems

4.4.2 Eigenvalue Bounds for Multivariable Systems.

4.4.3 Implementation of the Bounds on the Maximum Eigenvalue of the Multivariable Operator, $D=Q^{\frac{1}{2}} W R^{-1} W^* Q^{\frac{1}{2}}$ for the Assessment of Input Structures.

4.5 An Algorithm for Input Structures Assessment.

Chapter 4 Aspects of Bound Computation for Structures

4.0 Introduction

Two problems arise when the diagnostic bound formulae are to be used for the assessment of structures. The first is to produce the moment sequence corresponding to each structure, and the second is to find either an upper bound or the actual value of the now structurally dependent maximum eigenvalue, λ_M^S .

The first of these is largely a matter of methodology, and with some not unreasonable problem size restrictions a moment decomposing algorithm is developed.

Finding an upper-bound or the value of the structurally dependent maximum eigenvalue, λ_M^S presents a formidable task. Discussion of the use of the moment ratio sequence is presented. The results of Bryant (1973) are re-derived in a function space setting and some useful extensions are given. Computer implementation of these results is considered. The chapter concludes with the complete input structure assessment algorithm.

Notation

It is assumed that there are m controls, and r outputs. As far as possible the letters 'i' and 'j' are reserved as indices for outputs and controls respectively. (Examples $i, i_1, \dots, j, j_1, j_3$ and so on). The controls $\{u_1, u_2, \dots, u_j, \dots, u_m\}$ are associated with the integer set $I_m = \{1, 2, \dots, j, \dots, m\}$ and a control structure $\{u_{j_1}, u_{j_2}, \dots, u_{j_s}\}$ associated with $\Omega_s = \{j_1, \dots, j_s\} \subseteq I_m$. Finally operators W, V are assumed to be the stationary convolution operators (whose adjoint is denoted W^*) typically given by:

$$Wu \triangleq (Wu)(t) = \int_0^t W(t-\tau) u(\tau) d\tau, \quad W: H_m \rightarrow H_r$$

and the adjoint by

$$W^*y \triangleq (W^*y)(t) = \int_t^{T_f} W^T(\tau - t) y(\tau) d\tau, \quad W^*: H_r \rightarrow H_m$$
$$0 < t \leq T_f$$

Related to these are the column and row operators

$$W_{\cdot j} u \triangleq (W_{\cdot j} u)(t) = \begin{bmatrix} \int_0^t W_{1j}(t-\tau) u(\tau) d\tau, \dots \\ \int_0^t W_{rj}(t-\tau) u(\tau) d\tau \end{bmatrix}^T$$

$$\text{and } W_{i\cdot} u \triangleq (W_{i\cdot} u)(t) = \sum_{j=1}^m \int_0^t W_{ij}(t-\tau) u_j(\tau) d\tau$$

hence $W_{\cdot j} : H_1 \rightarrow H_r$ and $W_{i\cdot} : H_m \rightarrow H_1$

The following adjoint operations hold for these operators

$$(W_{\cdot j})^* \equiv (W^*)_{\cdot j} \text{ written as } W^*_{\cdot j}$$

$$\text{and } (W_{i\cdot})^* \equiv (W^*)_{\cdot i} \text{ written } W^*_{\cdot i}$$

when R is assumed diagonal ' r_j ' denotes the (j,j) th element of the operator.

4.1 Calculation of the Structure Moment Sequences

An input structure is imposed on a system by the deletion of a subset of the inputs, to form a structured system, W_s . In the associated quadratic cost optimization problem, the effect is to produce a new input cost matrix R_s which is part of the full input cost matrix R . The calculation of ΔJ_O^S or $IESQ^S$ follows in the usual manner in which the operators are structurally dependent. Consequently the moment sequence, and the maximum eigenvalue are structurally dependent. In particular the structure moment sequence will be:

$$\begin{aligned} m_O^S &= m_O = \langle d, Q d \rangle_{H_r} \\ m_1^S &= \langle W_s^* Q d, R_s^{-1} W_s^* Q d \rangle_{H_s} = \langle g_O^S, R_s^{-1} g_O^S \rangle_{H_s} \\ m_2^S &= \langle W_s R_s^{-1} g_O^S, Q W_s R_s^{-1} g_O^S \rangle_{H_r} = \langle Z_O^S, Q Z_O^S \rangle_{H_r} \\ m_3^S &= \langle W_s^* Q Z_O^S, R_s^{-1} W_s^* Q Z_O^S \rangle_{H_s} = \langle \tilde{g}_O^S, R_s^{-1} \tilde{g}_O^S \rangle_{H_s} \\ m_4^S &= \langle W_s R_s^{-1} \tilde{g}_O^S, Q W_s R_s^{-1} \tilde{g}_O^S \rangle_{H_r} = \langle \tilde{Z}_O^S, Q \tilde{Z}_O^S \rangle_{H_r} \end{aligned}$$

(Remark: The normalizing factor of $\frac{1}{2}$ has been omitted for notational simplicity), which, when given some value for λ_M^S , can be used to calculate bounds for ΔJ_O^S and $IESQ^S$. For any given multivariable system, assessment of all the input structure will require the generation of all the corresponding structure moment sequences.

Hence the method of generation should minimize the repetition of computation common to the various moment sequences. Three methods are considered:

1. The first involves the formation of the system operator for each structure, producing the responses $g_O^S, \dots, \tilde{Z}_O^S$ and directly calculating the moment sequence $m_1^S, \dots, m_4^S$. Such a method has the advantages of simplicity and low computer storage requirements. If the input weight operator R is not

restricted to diagonal form this may be the only method available to obtain the moment sequences. However given the restrictions that Q be self-adjoint and R be of diagonal form then the method is grossly inefficient. The remaining two methods require these restrictions to apply, namely

- (i) Q is self adjoint
- (ii) R is of diagonal form.

2. The repetition of calculating structured system responses can be removed by decomposing the full system vectors $g_o = W^*Qd$, $\dots$, $\tilde{z}_o = W R^{-1} \tilde{g}_o$ into a number of basic responses from which the structure responses can be reconstituted. This does not eliminate the repetition of certain inner product calculations, furthermore the storage requirements may be prohibitive since the basic responses would be stored in sampled form.
3. The storage of basic responses can be eliminated by evaluating the inner products of these responses to form moment parts which sum to form the moments. This leads to the successful technique of decomposing the full system moments into a number of moment parts from which the structure moments can be reconstituted. This technique is ideal for both hybrid and digital computing facilities although experience to date is limited to digital computing. Computer storage requirements can present a problem as will be seen later but first the decomposition of the moments is detailed.

4.2 The Decomposition of the Moments

The full system moments are decomposed in order:

4.2.0 Decomposition of Moment, m_0

Moment $m_0 = \langle d, Qd \rangle_{H_r}$ is independent of input structures, hence it can be calculated, and stored for use.

4.2.1 Decomposition of Moment, m_1

Moment $m_1 \triangleq \langle g_0, R^{-1} g_0 \rangle_{H_m}$ where $g_0 = W^* Qd$

whence $m_1 = \sum_{j=1}^m \langle g_{0j}, r_j^{-1} g_{0j} \rangle_{H_1} = \sum_{j=1}^m m_1(j)$

where $m_1(j) \triangleq \langle g_{0j}, r_j^{-1} g_{0j} \rangle_{H_1}$

and $g_{0j} = W_j^* Qd$; $W_j^* : H_r \rightarrow H_1$ $j = 1, 2, \dots, m$.

The complete decomposition of moment, m_1 is into m inner product parts. For a given structure specified by the index set Ω_s , the structure moment $m_1^s = \sum_{j \in \Omega_s} m_1(j)$.

4.2.2 Decomposition of Moment, m_2

Moment $m_2 \triangleq \langle Z_0, Q Z_0 \rangle_{H_r}$ where $Z_0 = WR^{-1} g_0$

Recall $g_0 = \begin{bmatrix} W_{1.}^* Qd \\ W_{2.}^* Qd \\ \vdots \\ W_{m.}^* Qd \end{bmatrix}$ then $R^{-1} g_0 = \begin{bmatrix} r_1^{-1} W_{1.}^* Qd \\ r_2^{-1} W_{2.}^* Qd \\ \vdots \\ r_m^{-1} W_{m.}^* Qd \end{bmatrix}$

Then $Z_0 = WR^{-1} g_0 = W_{1.} r_1^{-1} W_{1.}^* Qd + \dots + W_{m.} r_m^{-1} W_{m.}^* Qd$

Defining $Z_{0j} = W_{.j} r_j^{-1} W_{j.}^* Qd \in H_r$ $j = 1, 2, \dots, m$

gives $m_2 = \langle \sum_{j=1}^m Z_{0j}, Q \sum_{j=1}^m Z_{0j} \rangle_{H_r}$
 $= \sum_{j=1}^m \langle Z_{0j}, Q Z_{0j} \rangle_{H_r} + 2 \sum_{j=1}^{m-1} \sum_{j_1=j+1}^m \langle Z_{0j}, Q Z_{0j_1} \rangle_{H_r}$

writing $m_2(j, j_1) = \langle Z_{0j}, Q Z_{0j_1} \rangle_{H_r}$ $(j, j_1) \in I_m \times I_m$.

and using the symmetry of Q : $m_2(j, j_1) = m_2(j_1, j)$

$$\text{Obtain } m_2 = \sum_{j=1}^m m_2(j,j) + 2 \sum_{j=1}^{m-1} \sum_{j_1=j+1}^m m_2(j,j_1) .$$

Hence a complete decomposition of moment, m_2 produces $\frac{1}{2}m(m+1)$ parts for storage, and the following formula for the generation of a 2nd order structure moment associated with index set

$$\Omega_s : m_2^s \sum_{(j,j_1) \in \Omega_s \times \Omega_s} m_2(j,j_1)$$

4.2.3 Decomposition of Moment, m_3

$$\text{Moment } m_3 = \langle \tilde{g}_o, R^{-1} \tilde{g}_o \rangle_{H_m} \quad \text{with} \quad \tilde{g}_o = W^* Q Z_o$$

$$\text{Recall from the decomposition of moment } m_2 \quad Z_o = \sum_{j=1}^m Z_{oj}$$

$$\text{where } Z_{oj} = W_{.j} r_j^{-1} W_j^* Q d \in H_r \quad j=1,2,\dots,m.$$

$$\text{Then } \tilde{g}_o = W^* Q Z_o = \begin{bmatrix} W_{1.}^* Q Z_o \\ \vdots \\ W_{m.}^* Q Z_o \end{bmatrix}$$

$$m_3 = \langle \tilde{g}_o, R^{-1} \tilde{g}_o \rangle_{H_m} = \sum_{j=1}^m \langle W_j^* Q Z_o, r_j^{-1} W_j^* Q Z_o \rangle_{H_1}$$

$$= \sum_{j=1}^m \langle W_j^* Q \sum_{j_1=1}^m Z_{oj_1}, r_j^{-1} W_j^* Q \sum_{j_2=1}^m Z_{oj_2} \rangle_{H_1}$$

$$= \sum_{j=1}^m \left\{ \sum_{j_1=1}^m \langle W_j^* Q Z_{oj_1}, r_j^{-1} W_j^* Q Z_{oj_1} \rangle_{H_1} \right.$$

$$\left. + \sum_{j_1=1}^{m-1} \sum_{j_2=j_1+1}^m \langle W_j^* Q Z_{oj_1}, r_j^{-1} W_j^* Q Z_{oj_2} \rangle_{H_1} \right\}$$

$$\text{Define } m_3(j,j_1,j_2) = \langle W_j^* Q Z_{oj_1}, r_j^{-1} W_j^* Q Z_{oj_2} \rangle_{H_1}$$

as the decomposed parts of m_3 then may write

$$m_3 = \sum_{j=1}^m \left\{ \sum_{j_1=1}^m m_3(j, j_1, j_1) + \sum_{j_1=1}^{m-1} \sum_{j_2=j_1+1}^m m_3(j, j_1, j_2) \right\}$$

Hence m_3 decomposes into $m \times (\frac{1}{2}m(m+1)) = \frac{1}{2}m^2(m+1)$ parts and a structure moment may be reconstituted as

$$m_3^S = \sum \sum \sum m_3(j, j_1, j_2) \epsilon_{\Omega_S} x_{\Omega_S} x_{\Omega_S}$$

using the symmetry $m_3(j, j_1, j_2) = m(j, j_2, j_1)$ $j_2 > j_1$

4.2.4 Decomposition of Moment, m_4

Moment $m_4 = \langle \tilde{Z}_O, Q \tilde{Z}_O \rangle_{H_R}$ with $\tilde{Z}_O = W R^{-1} \tilde{g}_O$

Using the decomposition of $\tilde{g}_O$ namely $\tilde{g}_O = \begin{bmatrix} W_{1.}^* Q Z_O \\ \vdots \\ W_{m.}^* Q Z_O \end{bmatrix}$

may write $\tilde{Z}_O = \sum_{j=1}^m W_{.j} r_j^{-1} W_{j.}^* Q Z_O$

Using the expansion for Z_O , namely $Z_O = \sum_{j=1}^m Z_{Oj}$ obtain

$$\tilde{Z}_O = \sum_{j=1}^m \sum_{j_1=1}^m W_{.j} r_j^{-1} W_{j.}^* Q Z_{Oj_1}$$

giving $\tilde{Z}_O = \sum_{j=1}^m \sum_{j_1=1}^m Z_{Ojj_1}$ where $Z_{Ojj_1} = W_{.j} r_j^{-1} W_{j.}^* Q Z_{Oj_1}$

Clearly $\tilde{Z}_O$ decomposes into m^2 distinct responses, thus forming the moment $m_4 = \langle \tilde{Z}_O, Q \tilde{Z}_O \rangle$ produces a quadratic form in these responses. Consequently m_4 decomposes into $\frac{1}{2}m^2(m^2 + 1)$ parts if the inner product symmetry is exploited, namely

$$\langle Z_{Ojj_1}, Q Z_{Oj_2j_3} \rangle = \langle Z_{Oj_2j_3}, Q Z_{Ojj_1} \rangle$$

Define $m_4(j, j_1, j_2, j_3) = \langle Z_{Ojj_1}, Q Z_{Oj_2j_3} \rangle_{H_R}$ as

the decomposed moment part then a structure moment may be reconstituted as: $m_4^S = \sum \sum \sum \sum m_4(j, j_1, j_2, j_3)$

$$(j, j_1, j_2, j_3) \in \Omega_S \times \Omega_S \times \Omega_S \times \Omega_S.$$

where the symmetry $m_4(j, j_1, j_2, j_3) = m_4(j_2, j_3, j, j_1)$ is used to reduce the storage requirements.

The two important features of the decomposition of the moments are the use of inner product symmetries to reduce the number of constituent parts of a given moment (hence reducing the computer storage requirements) and the fact: that the number of parts to be stored depends only on the number of controls or inputs. The following table serves to illustrate this dependency.

Table 4.2.1 Computer Storage Requirements for Moment Decomposition

No of Controls	No of Parts into Which Moment Decomposes				
	m_0	m_1	m_2	m_3	m_4
3	1	3	6	18	45
5	1	5	15	75	325
10	1	10	55	550	5,050
20	1	20	210	4,100	80,200
50	1	50	1,275	63,750	3,126,250
m	1	m	$\frac{1}{2}m(m+1)$	$\frac{1}{2}m^2(m+1)$	$\frac{1}{2}m^2(m^2+1)$

4.3 The Formation of Structure Moment Sequences

The reconstitution of decomposed moment parts to form a structure moment sequence simply requires the identification of an input structure with the subset Ω_s and the use of the structure moment formulae given in the previous section. The association between the input set $\{u_1, u_2, \dots, u_m\}$ and the integer set $I_m = \{1, 2, \dots, m\}$ enables the input structure $\{u_{j_1}, u_{j_2}, \dots, u_{j_s}\}$ to be associated with the integer subset $\Omega_s = \{j_1, j_2, \dots, j_s\} \subseteq I_m$. To effect a complete assessment of all input structure obtainable by input deletion these integer subsets (or their logical equivalent) must all be created a-priori or produced sequentially to enable the generation of the structure moment sequences. The expression for the total number of structures which can be generated is given as

$$\text{Total} = \sum_{j=1}^s i C_m \quad \text{where} \quad i C_m = m! / i! (m-i)!$$

and s is the maximum size of the input subsets forming a structure.

Some typical values are given in the table below:

Table 4.3.1 Total Number of Structures for some values of m

Total No of Inputs, m	3	5	10	20	50
Total No of Structures	7	31	1023	10485	69757

In the cases where m is large, it is probable that interest will be centred on low order structure causing a corresponding reduction in the number of structures to be examined. For example in the case where $m = 50$, if structures of up to 10 inputs are examined then there would be 13, 421 structures, a reduction of more than 56,000 combinations.

For each of the structures designated through Ω_s , the moment sequence $m_1^s, m_2^s, \dots, m_4^s$ is calculated using the following formulae:

$$m_1^s = \sum_{j \in \Omega_s} m_1(j)$$

$$m_2^s = \sum_{j \in \Omega_s} m_2(j, j) + 2 \sum_{(j, j_1) \in \Omega_s \times \Omega_s} m_2(j, j_1)$$

$$j < j_1$$

$$m_3^s = \sum_{j \in \Omega_s} \{ m_3(j, j, j) + 2 \sum_{(j_1, j_2) \in \Omega_s \times \Omega_s} m_3(j, j_1, j_2) \}$$

$$j \neq j_1 \neq j_2 \quad j_1 \leq j_2$$

$$m_4^s = \sum_{(j, j_1) \in \Omega_s \times \Omega_s} m_4(j, j_1, j, j_1) + 2 \sum_{(j, j_1) \in \Omega_s \times \Omega_s} m_4(j, j_1, j_2, j_3)$$

$$(j_1, j_2) \in \Omega_s \times \Omega_s$$

$$(j, j_1) < (j_2, j_3)$$

The moment, m_0 , is produced only once and can be adjoined to give the structure moment sequence as $m_0, m_1^s, \dots, m_4^s$.

The properties exhibited in Section 3.5 will apply to this sequence and so can be used as a check of consistency. A structure moment ratio sequence can be formed to aid in the calculation of a bound on the maximum eigenvalue, λ_M^s . This remains the only uncertain quantity in the process of calculating bounds for the structure diagnostics ΔJ_0^s and $IESQ^s$.

4.4 The Problem of an Upper Bound for the Maximum Eigenvalue, λ_M

Introduction

The preceding theory of bound calculation relied on the assumption that either the value or an upper bound was available for the maximum eigenvalue (λ_M) of the operator

$$D = VV^* = Q^{\frac{1}{2}}WR^{-1}W^*Q^{\frac{1}{2}}. \text{ The consideration of structures implies}$$

that the operators V , and D become structure dependent, viz

$$D_s = V_s V_s^* = Q_s^{\frac{1}{2}}W_s R_s^{-1}W_s^*Q_s^{\frac{1}{2}}, \text{ so to } \overset{o}{\text{do}} \text{ does the maximum eigenvalue}$$

denoted λ_M^s . In Section 3.4 it was shown that the moment

ratios form a sequence of lower bounds for λ_M , hence if this

sequence is sufficiently well converged an upper bound for

λ_M might be obtained almost by inspection. This also applies to the sequences formed for the structures. However experience has shown that while the sequences have converged sufficiently for some structures this is not necessarily the case for others.

Furthermore, it is not impossible that many more moments would have to be generated before any semblance of convergence

appears. To illustrate this Fig.4.4.1 shows the convergence

of the moment ratio sequences for several structures taken

from the industrial example presented in Chapter 5, whereas

Table 4.4.1 shows the development of the moment ratio

sequence for another structure from the same example.

Table 4.4.1 Convergence of a Structure Moment Ratio Sequence

A Five Stand Cold Rolling Mill : Run from Series IV

r1	r2	r4	r8	r12	r16	r20	r40	$r_i = \frac{m_i}{m_{i-1}}$
0.1738	0.6968	0.8259	2.494	2.613	2.641	2.662	2.722	

The conclusion to be drawn from these examples is that applying extrapolation techniques or other convergence

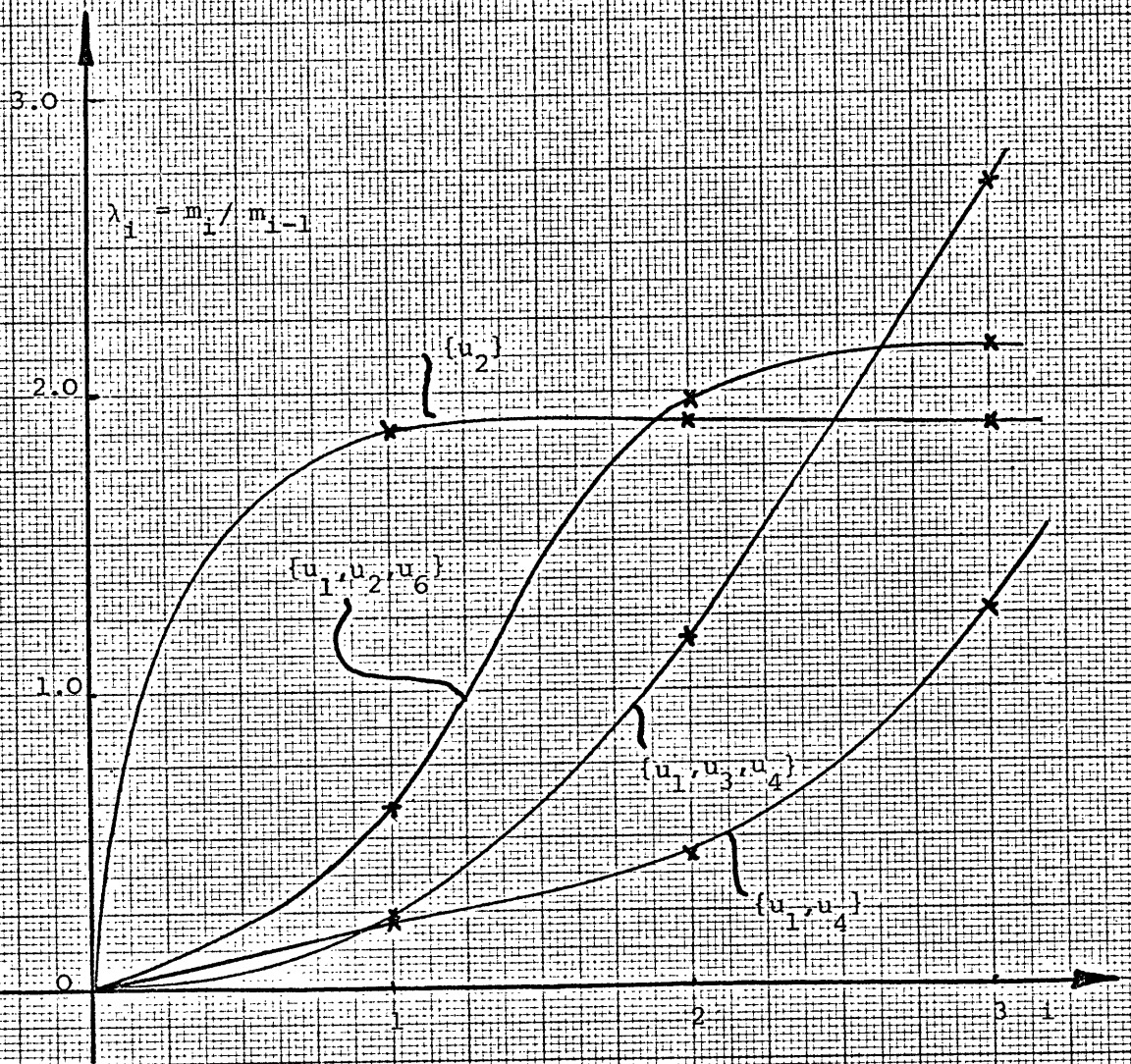


Figure 4.4.1 Typical Behaviour of Estimates for Maximum Eigenvalue λ_M
For Various Structures

Controls $u_1, \dots, u_4$ $s_1, \dots, s_4$

$u_5, \dots, u_7$ $w_1, \dots, w_3$

Step Disturbance on Interstand Gauge h_1

Cost Function Weights are $Q_1 = 20.0$ $i=1, \dots, 7$

$R_1 = 1.0$ $i=1, \dots, 4$

$R_1 = 5.0$ $i=5, \dots, 7$

accelerating techniques will not be successful because of the inherent poor conditioning of (possibly) some of the moment ratio sequences. The moment ratio sequence is used as an iterative technique, known as the "power" method to find the maximum eigenvalue of a self adjoint operator. The extensive literature on this topic (see, for example, Noble 1969 or Faddeev and Faddeeva, 1963) has been searched for references on the problem of finding or bounding the limit of the moment ratio sequence given the restricted information of the first few moments but to no avail.

Bryant (1973) addressed this problem through the use of operator norms. Using finite dimensional approximations for infinite dimensional operators, applying the properties of matrix norms and observing that the various limiting operations then taken are valid under a set of very general conditions. Bryant was able to produce a set of upperbounds for the maximum eigenvalue, λ_M . The new bounds subsumed the isolated result of Freeman (1969) and were capable of some extensions. The approach taken here is to rederive all the results in a function space setting, beginning with one of the important properties of the induced norms. This concerns the norm of a product of linear operators. (An elementary preamble on the topic of norms along with most of the proofs for the lemmas and theorems of this section are to be found in Appendix C). Suppose X , Y and Z are normed linear spaces, and the linear operators T_1 , and T_2 are such that

$$T_1: X \rightarrow Y, \quad ||T_1 x||_Y \leq ||T_1||_X ||x||_X$$

and $T_2: Z \rightarrow X, \quad ||T_2 z||_X \leq ||T_2||_Z ||z||_Z$

For the operator product $T_1 T_2$ obtain

$$\begin{aligned} \|T_1 T_2\|_Y &= \sup_z \|T_1 T_2 z\|_Y \leq \|T_1\|_X \sup_z \|T_2 z\|_X \\ &\quad \|z\|_Z = 1 \\ &\leq \|T_1\|_X \cdot \|T_2\|_Z. \quad 4.4.1 \end{aligned}$$

This relationship between the norms shows the operator product is also continuous. A compatible norm on a space of linear transformations $\gamma : \{T \in \gamma \Rightarrow T : X \rightarrow X\}$ satisfies the axioms of the norm (Appendix C) together with :

$$a) \quad \|Tx\|_X \leq \|T\|_{c(X)} \|x\|_X \quad \forall x \in X, T \in \gamma$$

$$\text{and b) } \|T_1 T_2\|_{c(X)} \leq \|T_1\|_{c(X)} \|T_2\|_{c(X)} \quad \forall T_1, T_2 \in \gamma$$

4.4.2

The notation $\|\cdot\|_{c(X)}$ denotes that the norm is compatible with the vector norm $\|\cdot\|_X$. The relationship between the induced operator norm, and a norm compatible with the same vector norm is given as $\|T\|_X \leq \|T\|_{c(X)} \quad \forall T \in \gamma$

4.4.3

Proof Suppose not, then $\|Tx\|_X \leq \|T\|_{c(X)} \|x\|_X \leq \|T\|_X \|x\|_X$

$$\Rightarrow \sup_{x \neq 0} \frac{\|Tx\|_X}{\|x\|_X} \leq \|T\|_{c(X)} < \|T\|_X$$

Contradicts the definition of the induced operator norm $\|\cdot\|_X$.

The discussion of the bounds on the maximum eigenvalue of the operator $D = Q^{\frac{1}{2}} W R^{-1} W^* Q^{\frac{1}{2}} = V V^*$ where $V = Q^{\frac{1}{2}} W R^{-\frac{1}{2}}$ begins by making several assumptions on the component operators, W , Q and R , and then specializes to the single-input single output case. The results for this case are then utilized in extensions to the multivariable situation.

Assume operators Q , R and W are such that 1) $Q^* = Q$, $Q \geq 0$

2) $R^* = R$, $R > 0$ and diagonal and 3) W is a time-invariant causal convolution operator.

Define a weighted impulse response

$$v_{ij}(t) = \left(\sum_{s=1}^r \tilde{q}_{is}(t) W_{sj}(t) \right) / r_j^{1/2}(t) \quad 0 \leq t \leq T_f$$

where $i, s = 1, 2, \dots, r$ and $j = 1, 2, \dots, m$

where $\tilde{q}_{is}(\cdot)$ are the elements of $Q^{1/2}$

$W_{sj}(\cdot)$ are the elements of the system impulse response matrix $W(\cdot)$.

and $r_j(\cdot)$ are the diagonal elements of operator R .

4.4.1 Eigenvalue Bounds for Single Input-Single Output Systems

The operator $D = Q^{1/2} W R^{-1} W^* Q^{1/2} = V V^*$ comprises of the

two convolution operators:

$$(Vu)(t) \triangleq \int_0^t v(t-\tau) u(\tau) d\tau \quad 0 \leq t \leq T_f$$

$$\text{and } (V^*y)(t) = \int_0^{T_f} v(\tau-t) y(\tau) d\tau \quad 0 \leq t \leq T_f$$

Two lemmas (proofs given in Appendix C) are introduced, these form the basis for the eigenvalue bounds.

Lemma 4.4.1 If $v \in L^1 [0, T_f]$, then $Vu \in L^p [0, T_f]$

whenever $u \in L^p [0, T_f]$. Similarly $Vy^* \in L^p [0, T_f]$ whenever $y \in L^p [0, T_f]$. $1 \leq p \leq \infty$.

Lemma 4.4.2 If $x \in L^2 [0, T_f] \Rightarrow x \in L^1 [0, T_f]$

The bounds on the maximum eigenvalue are produced first in the form of norm relationships, and then the particular norm formulae are evaluated.

If the kernel of operator V belongs to $L^1 [0, T_f]$ and the

domain of V^* is $L^2 [0, T_f]$, Lemma 4.4.1 gives

$$V^* : L^2 [0, T_f] \rightarrow L^2 [0, T_f] \quad \text{and} \quad V : L^2 [0, T_f] \rightarrow L^2 [0, T_f]$$

further $VV^* : L^2 [0, T_f] \rightarrow L^2 [0, T_f]$

$$\|VV^*\|_2 = \sup_{x \neq 0} \frac{\|VV^*x\|_2}{\|x\|_2} = \sup_{x \neq 0} \left\{ \frac{\langle VV^*x, VV^*x \rangle}{\langle x, x \rangle} \right\}^{\frac{1}{2}}$$

Since VV^* is self-adjoint and compact (Hsieh, 1965) on the Hilbert space $H=L^2[0, T_f]$, the theory utilized in Chapter 3 gives

$\|VV^*\|_2 = \lambda_M$, where λ_M is the maximum eigenvalue of VV^* . (For this reason this norm is known as the spectral norm.) Invoking the results for compatible norms, (namely inequalities 4.4.2 and 4.4.3)

$$\text{obtain } \lambda_M = \|VV^*\|_2 \leq \|VV^*\|_{C(2)} \leq \|V\|_{C(2)} \cdot \|V^*\|_{C(2)}$$

where suffix 'C(2)' denotes an operator norm compatible with the vector norm $\|x\|_2, x \in L^2[0, T_f]$.

Using the Hilbert space theory, the eigenvector corresponding to λ_M , denoted by e_M , satisfies $VV^*e_M = \lambda_M e_M$ where $\lambda_M > 0$ and (Lemma 4.4.2) $e_M \in L^2[0, T_f] \subset L^1[0, T_f]$.

$$\text{Hence } \|\lambda_M e_M\|_1 = \lambda_M \|e_M\|_1 = \|VV^*e_M\|_1 \leq \|VV^*\|_1 \|e_M\|_1 \\ \Rightarrow \lambda_M \leq \|VV^*\|_1 \leq \|V\|_1 \|V^*\|_1.$$

The formulae for these operator norms are given by the following

Theorem 4.4.1 If the operator kernel v is s.t. $v \in L^1[0, T_f]$

then

$$\text{a) } \|V\|_1 = \|V^*\|_1 = \int_0^{T_f} |V(s)| ds \\ \text{and b) } \|VV^*\|_1 \leq \max_{0 \leq t \leq T_f} \int_0^{T_f} \left| \int_0^{\min(t, \tau)} v(t-s)v(\tau-s) ds \right| d\tau \\ \leq \|V\|_1 \|V^*\|_1$$

If the operator kernel v is such that $v \in L^2[0, T_f]$

then

$$\text{c) } \|VV^*\|_E = \int_0^{T_f} \int_0^{T_f} \left| \int_0^{\min(t, s)} v(t-\tau)v(s-\tau) d\tau \right|^2 ds dt$$

where $||VV^*||_E \leq ||V||_E^2$ and $||V||_E^2 = ||V^*||_E^2 = \int_0^{T_f} \int_0^t v^2(t-\tau) d\tau dt$

and $||\cdot||_E$ is a norm compatible to the norm $||\cdot||_2$.

The proof of this theorem is found in Appendix C.

These formulae were applied to the unstable single input - single output system used in Section 3.6. The following details are

pertinent: $v(t) = q^{\frac{1}{2}} W(t) / r^{\frac{1}{2}} = e^t \quad 0 \leq t \leq 1.0$ and $\lambda_M = 1.0$.

The bound formulae gave :

$$i) \quad ||V||_1 = \int_0^{1.0} |v(s)| ds = e-1 \Rightarrow (||V||_1)^2 = 2.9525$$

$$ii) \quad ||V||_E^2 = \int_0^{1.0} \int_0^t v^2(t-\tau) d\tau dt = (e^2-3)/4 = 1.09726$$

$$iii) \quad ||VV^*||_E = 1.0072, \text{ using numerical methods.}$$

$$iv) \quad ||VV^*||_1 = \max_t \int_0^{T_f} \left| \int_0^{\min(t,\tau)} v(t-s) v(\tau-s) ds \right| d\tau$$

The details of this calculation for $||VV^*||_1$ are given here to

illustrate the complexity. Proceed by setting

$$f(t,\tau) = \int_0^{\min(t,\tau)} v(t-s) v(\tau-s) ds \text{ where } (t,\tau) \in [0, T_f] \times [0, T_f]$$

Advantage is taken of the symmetry of $f(t, \tau)$ about the

line $t = \tau$ to obtain the lower triangular half of an array

representing the function $f(t, \tau)$. The full function is

obtained by reflection about the diagonal. This is followed

by numerical integration along the τ axis to obtain the

function $\Omega(t) = \int_0^{T_f} |f(t, \tau)| d\tau$, of which the maximum is obtained by inspection. Table 4.4.2 contains the values for

this example.

Table 4.4.2 Generation of $\Omega(t) = \int_0^T \left| \int_0^{\min(t,\tau)} v(t-s) v(\tau-s) ds \right| d\tau$

τ Axis $\rightarrow$		0.0	0.25	0.50	0.75	1.00	$\Omega(t)$
t Axis $\downarrow$	0.0	0.0	Reflect Array to				
	0.25	0.0	0.3244	Obtain $f(t,\tau)$			
	0.50	0.0	0.4165	0.8591			
	0.75	0.0	0.5348	1.1032	1.7408		
	1.00	0.0	0.6867	1.4165	2.2353	3.1945	
Integrate Across to Obtain $\Omega(t) \rightarrow \Omega(t)$							

The column for $\Omega(t)$ obtained by integrating across the τ axis will clearly show a maximum at $t = 1.0$. This enables an analytical expression to be deduced for $||VV^*||_1$
namely $||VV^*||_1 = \int_0^1 \int_0^{\tau} (e^{1-s} e^{\tau-s}) ds = e(\cosh 1 - 1)$,
subsequently $||VV^*||_1 = 1.4762$.

In summary the four bounds found on $\lambda_M = 1.0$ are

- i) $\lambda_M \leq ||v||_1 = 2.9525$
- ii) $\lambda_M \leq ||VV^*||_1 = 1.4762 \leq ||v||_1^2$
- iii) $\lambda_M \leq ||v||_E^2 = 1.0973$
- iv) $\lambda_M \leq ||VV^*||_E^2 = 1.0072 \leq ||v||_E^2$

It should be mentioned that the ordering of the $||\cdot||_1$ and $||\cdot||_E$ bounds differs from problem to problem. The results for a single input, single output systems are summarized in the table below.

Table 4.4.3 Summary of Eigenvalue Bounds for Single Input-

Single Output Systems	
Norm Bounds	$0 < \lambda_M \leq VV^* _1 \leq v _1^2$ ($ v^* _1 = v _1$)
	$0 < \lambda_M = VV^* _2 \leq VV^* _E \leq v _E^2$
Bound Formulae	$ v _1 = v^* _1 = \int_0^T v(s) ds$

$$||VV^*||_1 \leq \max_t \int_0^{T_f} \left| \int_0^{\min(t-\tau)} v(t-s) v(\tau-s) ds \right| d\tau$$

$$||VV^*||_E = \int_0^{T_f} \int_0^{T_f} \left(\int_0^{\min(t,s)} v(t-\tau) v(s-\tau) d\tau \right)^2 ds dt$$

$$||V||_E = ||V^*||_E = \int_0^{T_f} \int_0^t v^2(t-\tau) d\tau dt$$

4.4.2 Eigenvalue Bounds for Multivariable Systems

The theory for the bounds on λ_M unfolds in much the same fashion as that for the single input, single output systems except that there is an added complexity arising from the operator kernel being a matrix valued function. However trivial extensions of the previous work are not proved. The operators under consideration are :

$$(Vu)(t) \triangleq \int_0^t V(t-\tau) u(\tau) d\tau, \text{ and } (V^*y)(t) = \int_0^{T_f} V^T(\tau-t) y(\tau) d\tau$$

where $V(\cdot) \in \mathbb{R}^{r \times m}$ and $t \in [0, T_f]$

Lemma 4.4.3 If $V(t) \in \mathbb{R}^{r \times m}$, $t \in [0, T_f]$ and the elements

$v_{ij}(t) \in L^1[0, T_f]$ $i = 1, 2, \dots, r$; $j = 1, 2, \dots, m$ then

$Vu \in L_r^p[0, T_f]$ whenever $u \in L_m^p[0, T_f]$ similarly

$V^*y \in L_m^p[0, T_f]$ whenever $y \in L_r^p[0, T_f]$.

Lemma 4.4.4 If $x \in L_n^2[0, T_f]$ then $x \in L_n^1[0, T_f]$.

The eigenvalue bound derivations follows as before using the above lemmas, norm axioms, norm property 4.4.1 and the definition of norm compatibility 4.4.2 to give

$$\lambda_M \leq \|VV^*\|_1 \leq \|V\|_1 \|V^*\|_1 \quad \text{and}$$

$$\lambda_M = \|VV^*\|_2 \leq \|VV^*\|_{C(2)} \leq \|V\|_{C(2)} \|V^*\|_{C(2)}$$

The norm formulae for multivariable systems follow as:

Theorem 4.4.2 Under the conditions of Lemma 4.4.3 obtain

$$a) \quad \|V\|_1 = \max_j \sum_{i=1}^r \int_0^{T_f} |v_{ij}(t)| dt \quad (\text{max. column sum})$$

$$b) \quad \|V^*\|_1 = \max_i \sum_{j=1}^m \int_0^{T_f} |v_{ij}(t)| dt \quad (\text{max. row sum})$$

$$\begin{aligned} \text{and c) } \|VV^*\|_1 &< \max_{i_0, s} \sum_{i=1}^r \int_0^{T_f} \left| \sum_{j=1}^m \int_0^{\min(t,s)} v_{ij}(t-\tau) v_{i_0 j}(s-\tau) d\tau \right| dt \\ &\leq \|V\|_1 \|V^*\|_1 \end{aligned}$$

Restricting further the kernel of operator V so that $v_{ij} \in L^2[0, T_f]$

$$\text{then d) } \|V\|_E^2 = \|V^*\|_E^2 = \int_0^{T_f} \int_0^t \sum_{i=1}^r \sum_{j=1}^m v_{ij}^2(t-\tau) d\tau dt$$

$$\begin{aligned} \text{and e) } \|VV\|_E &= \sum_{i=1}^r \sum_{i_0=1}^r \int_0^{T_f} dt \int_0^{T_f} ds \left| \sum_{j=1}^m \int_0^{\min(t,s)} v_{ij}(t-\tau) v_{i_0 j}(s-\tau) d\tau \right|^2 \\ &\leq \|V\|_E^2 \end{aligned}$$

This theorem is proved in Appendix C, and the implementation of the formulae is considered in the following sub-section.

4.4.3 Implementation of the Bounds on the Maximum Eigenvalue of the Multivariable Operator, $VV^* = Q^{\frac{1}{2}} W R^{-1} W^* Q^{\frac{1}{2}}$ for the Assessment of Input Structure.

The implementation of the bound formulae is discussed individually for the expressions in Theorem 4.4.2. Account must be taken of the method of computation, that is whether digital or hybrid, since some of the formulae are not suited to one of the forms of computation. Ease of calculation of the bounds for structure assessment is the remaining factor to be considered.

$$\text{Bounds } ||V||_1 = \max_j \sum_{i=1}^r \int_0^T |v_{ij}(s)| ds, \quad \text{and}$$

$$||V^*||_1 = \max_i \sum_{j=1}^m \int_0^T |v_{ij}(s)| ds \quad \text{are easily calculable by either}$$

digital or hybrid means, since the term $\int_0^T |v_{ij}(s)| ds$

$i = 1, 2, \dots, r$; $j = 1, 2, \dots, m$ is the area under the absolute value of the (i, j) th impulse response. For structure assessment :

$$\text{define } S_j = \sum_{i=1}^r \int_0^T |v_{ij}(s)| ds \quad (\text{column sums}) \quad \text{then}$$

$$||V_s||_1 = \max_{j \in \Omega_s} S_j \quad \text{similarly define } S_{ij}^* = \int_0^T |v_{ij}(s)| ds$$

$$\text{then } ||V_s^*||_1 = \max_i \sum_{j \in \Omega_s} S_{ij}^* . \quad \text{Clearly storage for } m(r+1) \text{ variables}$$

is required, and each bound calculation comprises of one summation, two searches for a maximum and one multiplication.

$$\text{The bound } ||VV^*||_1 \leq \max_{i, t} \sum_{i=1}^r \int_0^T \left| \sum_{j=1}^m \int_0^{\min(t, \tau)} v_{ij}(t-s) v_{ij}(\tau-s) ds \right| d\tau$$

does not have the simplicity of the previous two formulae.

The inner summation with respect to index j will become structure

dependent requiring an impossible amount of storage connected with the

terms $\int_0^{\min(t,\tau)} v_{ij}(t-s) v_{i_o j}(\tau-s) ds$. By making a slight modification

obtain:
$$\|v v^*\|_1 \leq \max_{i,t} \sum_{i_o=1}^r \int_0^{T_f} \sum_{j=1}^m \left| \int_0^{\min(t,\tau)} v_{ij}(t-s) v_{i_o j}(\tau-s) ds \right| d\tau$$

$$\leq \|v\|_1 \|v^*\|_1.$$

Note that compatibility is still preserved, and equality with the previous form is attained for wholly positive impulse responses. This modification permits the re-ordering of the two summations (with respect to i_o , and j) and the integration with respect to τ .

Define
$$S_{ij}(t) = \sum_{i_o=1}^n \int_0^{T_f} \left| \int_0^{\min(t,\tau)} v_{ij}(t-s) v_{i_o j}(\tau-s) ds \right|$$

$$\|v v^*\|_1 \leq \max_{i,t} \sum_{j=1}^m S_{ij}(t)$$

This circumvents the storage problems since $S_{ij}(t)$ for $i=1,2,\dots,r$;

$j=1,2,\dots,m$ and N sample points for $t \in [0, T_f]$ will require only

$r \times m \times N$ locations. A formula of this type also facilitates the

computation of bounds for structures since if Ω_s is a structure index

set then $\|v_s v_s^*\|_1 \leq \max_{i,t} \sum_{j \in \Omega_s} S_{ij}(t)$. The computation of the $S_{ij}(t)$

can be reduced by the exploitation of several symmetries :

i) that of $f_{ii_o j}(t,\tau) = \int_0^{\min(t,\tau)} v_{ij}(t-s) v_{i_o j}(\tau-s) ds$ with respect to

t , and τ , which can be used to reduce the computation of the integral

$$\int_0^{T_f} |f_{ii_o j}(t,\tau)| d\tau$$

and ii) that of $f_{ii_o j}(t,\tau)$ with respect to i , and i_o which reduces

the repetition of calculations when summing over i_o . Digital or hybrid techniques may be used to produce the sampled data form of $S_{ij}(t)$ from which the structure bounds may be generated using only one summation and simple search for each structure.

The simple Euclidean bound has a form which is easily evaluated and amenable for structures assessment :

$$||v||_E^2 = \sum_{i=1}^r \sum_{j=1}^m \int_0^{T_f} \int_0^t v_{ij}^2(t-\tau) d\tau$$

Reversing the summations, and defining

$$S_j = \int_0^{T_f} \int_0^t \sum_{i=1}^r v_{ij}^2(t-\tau) d\tau \quad \text{obtain}$$

$$||v||_E^2 = \sum_{j=1}^m S_j \quad \text{and for structures} \quad ||v_s||_E^2 = \sum_{j \in \Omega_s} S_j.$$

The more complex $||vv^*||_E^2$ bound poses formidable computational problems even though its form can be suitably decomposed for structure assessment.

$$\text{Recall } ||vv^*||_E^2 = \sum_{i=1}^r \sum_{i_o=1}^r \int_0^{T_f} dt \int_0^{T_f} d\tau \left(\sum_{j=1}^m \int_0^{\min(t,\tau)} v_{ij}(t-s) v_{i_o j}(\tau-s) ds \right)^2$$

$$\text{Define } f_{ii_o j}(t,\tau) = \int_0^{\min(t,\tau)} v_{ij}(t-s) v_{i_o j}(\tau-s) ds$$

$$\text{then } \left(\sum_{j=1}^m f_{ii_o j}(t,\tau) \right)^2 = \sum_{j=1}^m f_{ii_o j}^2(t,\tau) + 2 \sum_{j=1}^{m-1} \sum_{j_1=j+1}^m f_{ii_o j}(t,\tau) f_{ii_o j_1}(t,\tau)$$

$$\begin{aligned} \text{thus } & \sum_{i=1}^r \sum_{i_o=1}^r \int_0^{T_f} dt \int_0^{T_f} d\tau \left(\sum_{j=1}^m f_{ii_o j}(t,\tau) \right)^2 \\ &= \sum_{j=1}^m \left(\sum_{i=1}^r \sum_{i_o=1}^r \int_0^{T_f} dt \int_0^{T_f} f_{ii_o j}^2(t,\tau) d\tau \right) + \\ & 2 \sum_{j=1}^{m-1} \sum_{j_1=j+1}^m \left(\sum_{i=1}^r \sum_{i_o=1}^r \int_0^{T_f} dt \int_0^{T_f} f_{ii_o j}(t,\tau) f_{ii_o j_1}(t,\tau) d\tau \right) \end{aligned}$$

$$\text{Define } Q_{jj_1} = \sum_{i=1}^r \sum_{i_0=1}^r \int_0^{T_f} dt \int_0^{T_f} f_{ii_0j}(t, \tau) f_{ii_0j_1}(t, \tau) d\tau$$

$$\text{then } ||V V^*||_E^2 = \sum_{j=1}^m Q_{jj} + 2 \sum_{j=1}^{m-1} \sum_{j_1=j+1}^m Q_{jj_1}$$

and for the structure index set Ω_s

$$||V_s V_s^*||_E^2 = \sum_{j \in \Omega_s} Q_{jj} + 2 \sum_{j, j_1 \in \Omega_s} Q_{jj_1}$$

The problems arise in computing the components of the sum, Q_{jj_1} , $j=1, 2, \dots, m$; $j_1=j, j+1, \dots, m$. Even after exploiting all the symmetries and interchanging summations and integrations it does not seem to be a feasible computation.

Consideration of the details of computing the bounds on the maximum eigenvalue reveals that the two more complicated bounds

$$||V V^*||_1 \text{ and } ||V V^*||_E^2 \text{ are not realistic computational tasks for}$$

structures assessment. However a modified formula for $||V V^*||_1$

has been introduced which is computationally realistic, and compatible

with $||V||_1 ||V^*||_1$. The simpler bounds $||V||_1$, $||V^*||_1$, and

$||V||_E^2$ are easily evaluated either digitally or by hybrid techniques and are readily amenable for structure assessment.

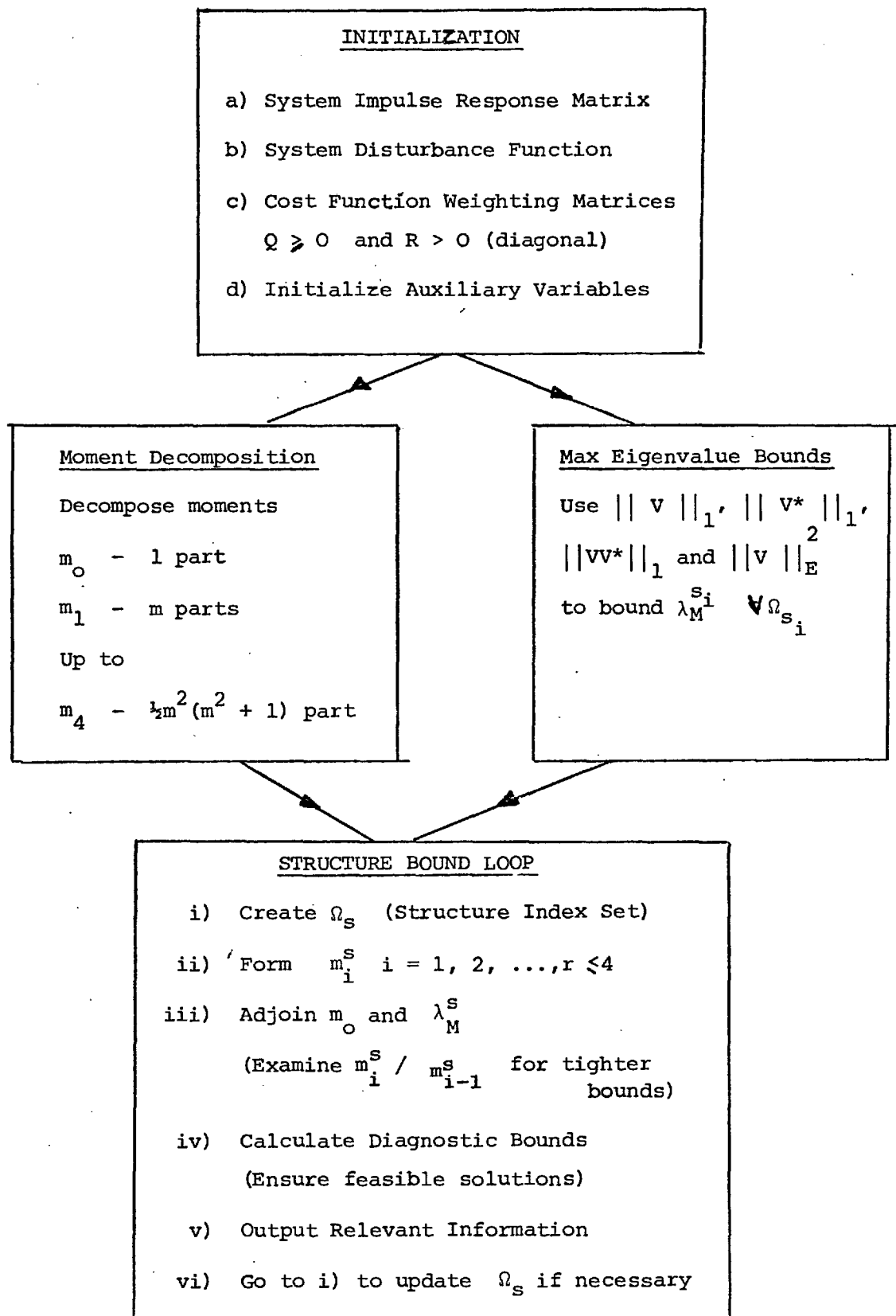
4.5 An Algorithm for Input Structures Assessment

The various parts of an algorithm for input structure assessment have been considered separately in order to give an adequate depth to their theory. This also reflects the off-line nature of certain parts, so that moment decomposition, or the generation of the structure dependent maximum eigenvalue bounds can be performed by separate computer programs, sometimes to a level of detail which is not necessary for the main purpose of producing bounds on the structure optima.

The algorithm is given in Fig. 4.5.1; its steps are discussed in detail here.

Initialization: There are four parts to this section: a) A system impulse response matrix or equivalent hybrid simulation must be obtained. b) The system disturbance must be determined. (Usually comparable in form to those the plant actually experiences). c) The choice of weight matrices Q (symmetric, semi positive definite) and R (diagonal, positive definite) must be made, reflecting the appropriate cost of system input and output variables; and d) The values of auxiliary variables should be settled.

Fig. 4.5.1 An Algorithm for Input Structure Assessment



These will include the maximum size of structure subsets, the order to which the moments should be formed, the interval size of system responses in sampled data form, and others.

Moment Decomposition: In accordance with the data provided by the above initialization, and the theory of Section 4.2, the moments are decomposed up to the appropriate order. This process may be performed on-line or off line depending chiefly on the number of inputs, and the order to which moments are to be created. If, as is usual for a large problem, this is an off line process then the decomposed moments must be stored in a form convenient for input to a program computing structure bounds.

Maximum Eigenvalue Bounds: This again is a process which may be performed on or off line. Using the process impulse response, the chosen weight matrices, and some of the auxiliary variables the bounds on λ_M^S are calculated using the theory of Section 4.4. If these bounds for the structures are created off-line, convenient data storage forms must be used for input to a program computing structure bounds.

Structure Bounds: This part of the algorithm requires as input only the decomposed moments, the maximum eigenvalue bounds, and those auxiliary variables necessary to be able to set up the structure index set Ω_s , (or its logical equivalent). The operation is a simple loop: create the structure index set, Ω_s ; form (using the theory of Section 4.3) the structure moment sequence; adjoin the maximum eigenvalue bound, λ_M^S (examining the moment ratio sequence for a possible tighter bound); calculate the bounds on the desired diagnostic (ΔJ_O^S , or $IESQ^S$, using the theory of Chapter 3) ensuring that the optimal solution is feasible; output the relevant information and return to create a new structure index.

These are the four parts of the input structures assessment algorithm, they have been implemented as a Fortran Library known as DECOMP. This library is a suite of subroutines which perform all the operations necessary for the above algorithm. The next Chapter contains details of the performance of the algorithm on a significant industrial example.

Chapter 5 - An Industrial Example

5.0 Introduction

5.1 Plant, Disturbance and Problem Description.

5.2 The Logistics of the Calculation Input Structure Diagnostic Bounds.

5.3 Control Structure Interpretation of the Calculated Bounds.

5.3.1 Bounds for a 10% Step Disturbance on Exit Gauge of Stand 2 and their Structural Implications .

5.3.2 Bounds for a 10% Step Disturbance on Interstand Tension, T_2 and their Structural Implications.

5.4 Conclusions.

5.0 Introduction

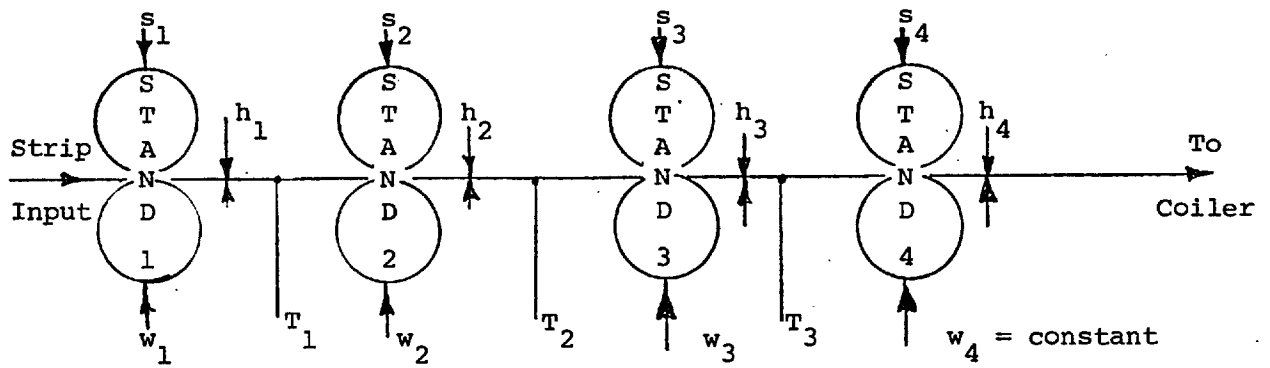
The automation of cold rolling steel mills has been a major subject of research for the Industrial Automation Group, Imperial College since April, 1968. This research generated large scale digital and hybrid mill simulations which were used to verify and test control structures devised using algebraic methods and physical reasoning for specific types of disturbance control. (Bryant et al, 1973). The control structures obtained were highly successful; their performance comparing favourably with a full treatment of the mill as a formal optimal control problem. (Tanuma, 1974).

Here the techniques of input structure analysis are applied to two specific disturbances, one gauge and one tension, on a four stand tandem mill having hydraulic actuators. For this medium size plant (seven input and seven outputs) it is known (Bryant et al, 1973, 1975) that these two disturbance types are controlled effectively by two particular input structures. Consequently this provides an ideal example to assess the feasibility of the computation involved and the clarity with which the input analysis indicates the known structures.

5.1 Plant, Disturbance and Problem Description

The modelling of a tandem cold rolling mill with hydraulic actuators (Edwards, 1968, Edwards et al, 1969, Caley and Spooner, 1972) resulted in a digital simulation (Bryant et al, 1969, Caley and Spooner, 1973) known by the Fortran name HYDMIL. This simulation provided a set of plant step responses and consequently the plant impulse responses for the system convolution operator, W . The simulation was run as a four stand mill giving seven output variables; namely four exit gauges $h_1, \dots, h_4$ and three interstand tensions $T_1, \dots, T_3$ and seven input variables; namely four actuator reference positions $s_1, \dots, s_4$ and three motor speed reference positions $w_1, \dots, w_3$. The plant, and the plant variables are shown diagrammatically in Fig. 5.1.1 and the associated normalized step responses in Fig. 5.1.2.

From the many types of disturbances which effect the rolling process (Bryant et al, 1973) two were chosen for which non-trivial control action is required. These were a 10% step disturbance to the exit gauge of Stand 2 (h_2) and a similar step disturbance to the tension between Stands 2 and 3 (T_2).



Nomenclature

s_1, s_2, s_3, s_4 :	actuator references (screws)	} controls/inputs
w_1, w_2, w_3 :	motor speed references (motors)	
h_1, h_2, h_3, h_4 :	exit gauges	} outputs
T_1, T_2, T_3 :	Interstand Tensions	

Fig. 5.1.1 Diagram showing Mill Variables

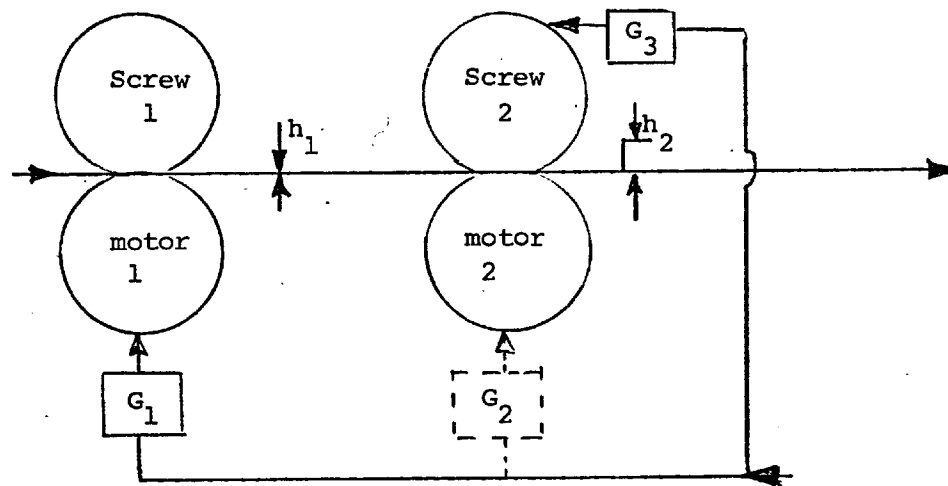


Fig. 5.1.3 Decoupling Network for Stand 2, Exit Gauge

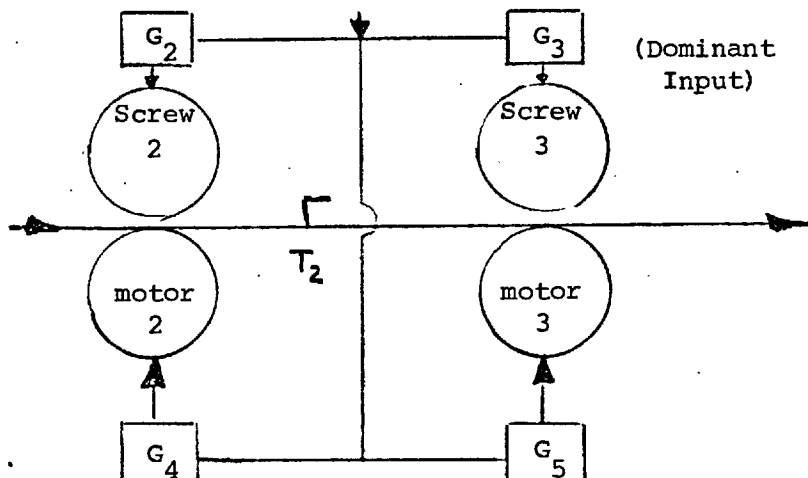


Fig. 5.1.4 Decoupling Network for Interstand Tension, T_2

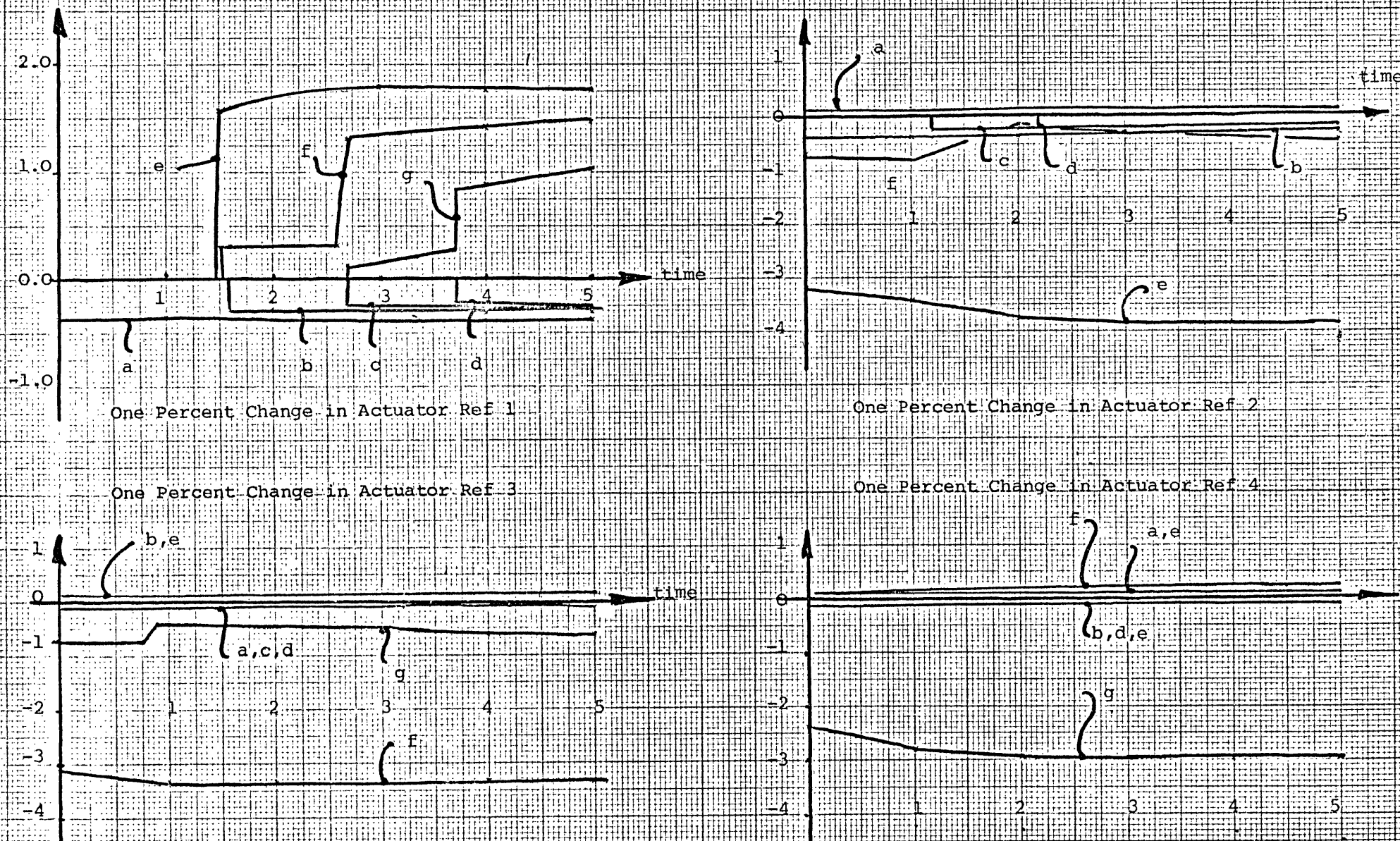


Figure 5.1.2 Normalized Mill Step Responses : 4 Stand Cold Rolling Mill

Open Loop at 1000 ft/min : Strip 36" Wide : 50 Samples at 0.1 Sec Intervals

5-3b

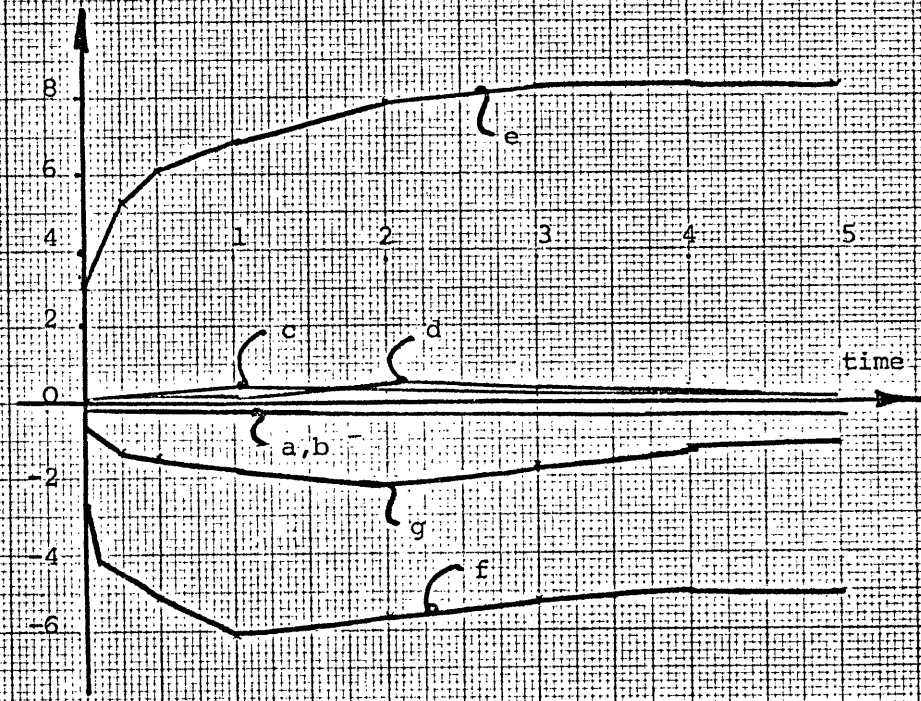
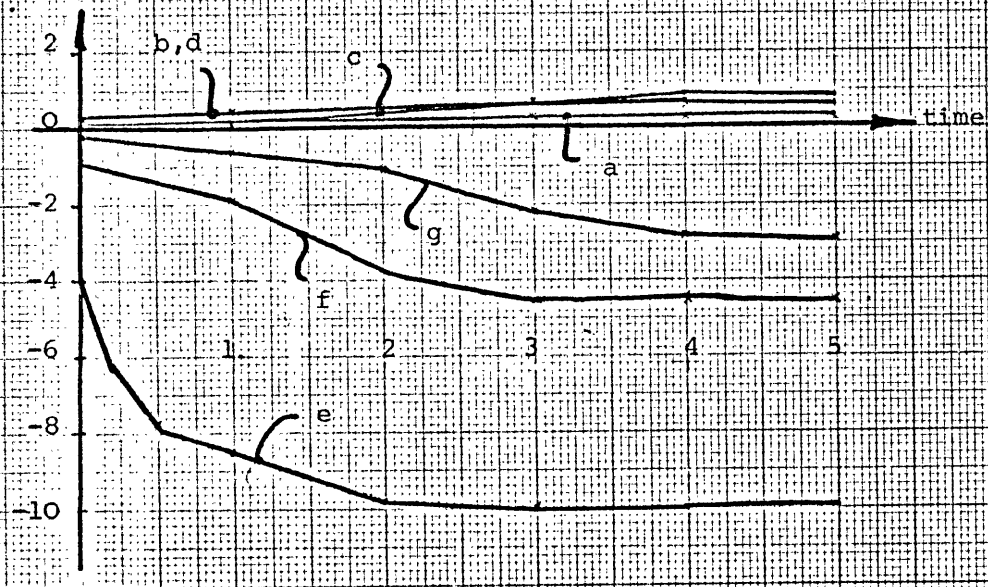
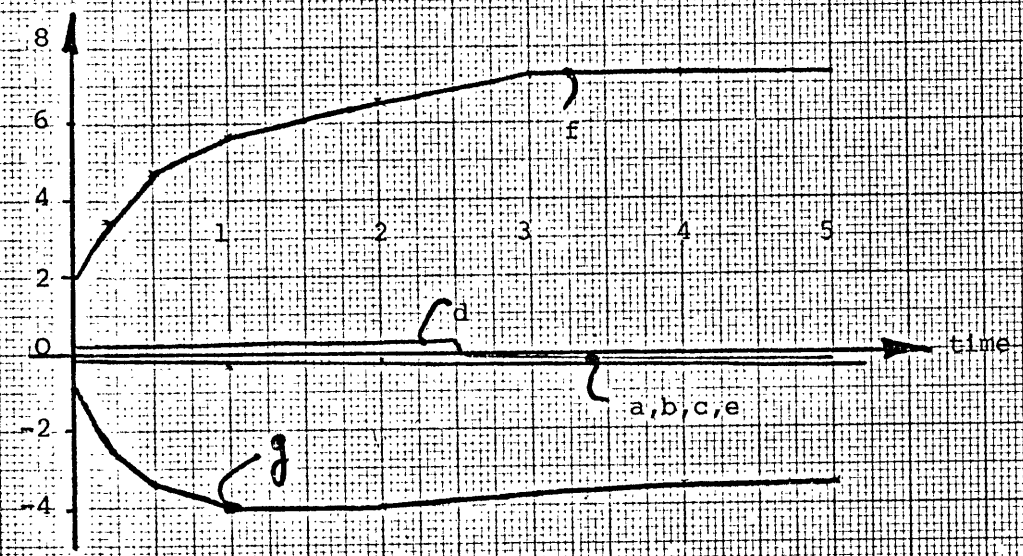


Figure 5.1.2 Continued

- Key
- | | |
|------------------|--------------------------|
| a = Exit Gauge 1 | d = InterStand Tension 1 |
| b = Exit Gauge 2 | e = InterStand Tension 2 |
| c = Exit Gauge 3 | f = InterStand Tension 3 |
| d = Exit Gauge 4 | |



The problem in both cases is to find out whether any input structure can remove the disturbances without upsetting the other mill outputs and with near optimal performance.

In the case of the disturbance to exit gauge h_2 , the Group have deduced (using algebraic methods and physical reasoning) that an effective decoupling network would utilize the inputs s_2 , w_1 , and s_2 to remove such a disturbance.

Further analysis and simulation runs have shown that the connection to motor 2 is of less importance so that in practical terms the structure should contain inputs s_2 , and w_1 . (Bryant et al, 1975). This network is shown along with the redundant arm (in dotted lines) in Fig. 5.1.3. The control of interstand tension was originally obtained on many mills wholly by the use of motor structures, however the Groups early analysis and simulation trials (Bryant et al, 1973) established the decoupling network of Fig. 5.1.4 which used the four inputs $\{s_2, s_3, w_2, w_3\}$. More recently an examination of the contribution from the arms of such a network has shown the forward screw arm to be dominant. (Bryant et al, 1975). Consequently tension is controlled by the forward screw alone, s_3 with little deterioration in performance when compared with the full network. The use of these two input structures formed a major advance in the automatic control of cold rolling tandem mills, and were used as the basis for control schemes on several mills in this country.

The input structure analysis was performed for the disturbances with two sets of cost weightings. One set assumed a blanket weight on all outputs such that in comparison with the input weight $\|Q\| \gg \|R\|$. This reflects the notion that control is cheap but output deviations from nominal are not tolerable. The second set, whilst adhering to this principle, costs the tension deviations less than the gauge changes to reflect the engineering consideration that a 1% tension change is

not so disastrous as a 1% gauge change. In detail the weight matrices were time-invariant and diagonal such that

$$\text{Set I} \quad Q_1 = \dots = Q_7 = 100.0 \quad R_1 = \dots = R_7 = 1.0$$

$$\text{Set II} \quad Q_1 = \dots = Q_4 = 100.0 \quad R_1 = \dots = R_7 = 1.0$$

$$Q_5 = \dots = Q_7 = 10.0$$

The logistics of producing the input structure diagnostics for these two disturbances are detailed next.

5.2 The Logistics of the Calculation of Input Structure

Diagnostic Bounds

The programs used to obtain the diagnostic bounds for all the structures were written in FORTRAN IV and run on the CDC 6400/CYBER 7314 machines installed at the Imperial College Computer Centre. A modifiable card image library (UPDATE) of programs was written to reflect the structure of the algorithm previously outlined in section 4.5. Hence the individual programs of the library are (in general) concerned with one distinct feature of the algorithm and data transfers between programs are achieved by the use of file stores.

The first group of programs are concerned with the production of an array of structure index sets, and the creation of the impulse response matrix. The step responses are read in card form (since this is the output medium used by the digital mill simulation) and a three point numerical differentiation formula is used to obtain the impulse response matrix for the plant convolution operator. In the case of the mill problem this resulted in a (7×7) input/output matrix sampled at 51 time instances hence an array of $(7 \times 7 \times 51)$ is required to store the mill responses. The array of structure index sets is a logical array (127×7) whose rows specify individual input structures, of which there are 127 for the mill problem.

The production of the bounds requires data specifying bounds on the maximum eigenvalue of the structured operators $D_s = V_s V_s^* = Q^{\frac{1}{2}} W_{s s s}^{-1} W_{s s s}^* Q^{\frac{1}{2}}$ and to decomposition of the moments.

Four formulae are available for the computation of the bounds on the maximum eigenvalue of the structured operators, D_s . These are given as:

$$B_s^1 = \{ \max_{j \in \Omega_s} \sum_{i=1}^r \int_0^{T_f} |v_{ij}(t)| dt \} \times \{ \max_i \sum_{j \in \Omega_s} \int_0^{T_f} |v_{ij}(s)| ds \}$$

$$B_s^2 = \max_{i,t} \sum_{j \in \Omega_s} \int_0^{T_f} \int_0^{\min(t,\tau)} |v_{ij}(t-s)| \sum_{i_o=1}^r |v_{i_o j}(\tau-s)| ds d\tau$$

$$B_s^3 = \max_{i,t} \sum_{j \in \Omega_s} \int_0^{T_f} \sum_{i_o=1}^r \left| \int_0^{\min(t,\tau)} v_{ij}(t-s) v_{i_o j}(\tau-s) ds \right| d\tau$$

$$\text{and } B_s^4 = \sum_{j \in \Omega_s} \int_0^{T_f} \int_0^t \sum_{i=1}^r v_{ij}^2(t-\tau) d\tau dt$$

where $B_s^1 \geq B_s^2 \geq B_s^3$ are the 1-norms and B_s^4 is a

Euclidean norm, (Section 4.4). The choice of which of the bounds

B_s^1 , B_s^2 and B_s^3 should be computed is largely dictated by computer

storage requirements and computing time. For the mill problem it

was possible to compute the four bounds for all the structures

enabling a comparison to be made between the bound formulae, and

against the exact value of the maximum eigenvalues. The following

tables document such a comparison for the two sets of cost functional

weightings.

Table 5.2.1 The Computing Time used to Calculate each Bound for a Complete Set of Structures (127).

Weight Sets	Computing Time (secs) for 127 Structures			
	B_s^1	B_s^2	B_s^3	B_s^4
I	0.103	44.34	336.0	0.133
II	0.098	47.03	354.8	0.142

Table 5.2.2 The Percentage Error in the Four Bounds for Twelve
Selected Structures, and Details of the Power Method for these
Structures. (Cost Functional Weights Set II)

Structure Subset	Percentage Error in Bound Formulae				No. of Moments Required to achieve $ \text{Est}_2 - \text{Est}_1 < 0.1$		
	B_s^1	B_s^2	B_s^3	B_s^4			
s_1	153.88	78.68	52.49	1788.61	9	6.4	Computing Time in Secs. to achieve this accuracy
s_2	166.23	120.73	107.17	3666.28	9	6.3	
s_3	97.52	101.01	92.35	3469.52	12	8.8	
s_4	56.01	68.71	63.37	3226.47	13	8.8	
w_1	114.40	45.57	44.52	942.50	25	15.8	
w_2	113.48	80.33	75.70	1207.57	> 50	> 30.0	
w_3	101.88	79.81	69.94	802.71	25	16.3	
s_1, s_2	242.45	157.31	135.92	3879.79	23	26.6	
s_1, s_3	188.47	153.26	133.58	3920.17	31	34.0	
s_2, w_1	178.45	61.93	59.00	1375.19	31	34.2	
s_2, w_6	200.05	103.00	96.46	1761.17	44	46.9	
$s_1, \dots$ $w_1, \dots, s_4$ w_3	284.16	139.91	122.43	2588.95	39	135.0	

These two tables indicate that the relatively minor increase in accuracy obtained through using bound formula B_s^3 does not warrant the almost eightfold increase in computing time. Notice that the theoretical inequalities relating to formulae B_s^1 , B_s^2 , and B_s^3 are not always observed when working with discrete operator approximation (structures s_3 , and s_4). The infeasibility of using the power method to obtain eigenvalue estimates for all structure is also demonstrated.

The decomposition of the first four moments for the mill problem was undertaken using the algorithms outlined in Sections 4.2 and 4.3. Such a decomposition was produced for each disturbance and each set of cost functional weights. For seven inputs, the moments were decomposed as follows: m_1 into 7 parts, m_2 into 28 parts, m_3 into 196 parts and m_4 into 1225 parts. Trapezoidal numerical integration was used to

produce the required plant, and adjoint plant responses and the integral inner products which constitute the decomposed moment parts. About 3500 seconds of computing time was used to produce the data file corresponding to one set of cost functional weights and one particular disturbance.

Subsequent to the production of the eigenvalue bounds for all the structures, and the decomposition of the moments the diagnostic bounds may be calculated. Since moments up to m_4 have been decomposed, five levels of bound formulae are available. For the mill problem about 930 seconds of computing time are required to produce the highest level bounds (that is using $m_0, \dots, m_4$) for all 127 possible structures, with the bulk of this time being used to produce structure moment sequences. For example, the calculation of the moment sequence $m_0^s, m_1^s, \dots, m_4^s$ associated with the full structure $(s_1, \dots, s_4, w_1, \dots, w_3)$ uses 53 seconds of computing time whereas calculating the related diagnostic bounds uses less than 0.005 seconds.

The diagnostic bounds for the posed mill problem, and a discussion of their control interpretation are given in the next section.

5.3 Control Structure Interpretation of the Calculated Bounds

The bounds are tabulated to show the diagnostic bounds for all the single input structures, and then the three highest ΔJ_0 (min) bounds, the three highest ΔJ_0 (max) bounds, the three lowest IESQ (min) and the three lowest IESQ (max) bounds for differing sizes of input structure subset. In this way the 127 structures of the mill problem are reduced to about 20 for closer scrutiny. The dispersion of the bounds is examined to isolate those structures which have bounds clearly separated from those of the other structures. Often secondary features must be assessed to obtain a conclusion on the ranking of structures. Perhaps the most important of these is the relationship between the bounds obtained for the complete input set and the bounds obtained for an input structure set of a specified size. In other words if an input subset is near optimal in performance, the bounds for the subset and the full set should be consistent.

5.3.1 Bounds for 10% Step Disturbance on Exit Gauge of Stand 2 and their Structural Implications

The results for cost functional weights sets I and II are contained in Table 5.3.1 and Table 5.3.2 respectively.

Table 5.3.1 An immediate conclusion is that among the single inputs screw, s_4 is a useless control. It may also be noted that screw, s_1 is more effective than screws s_3 and s_4 , and probably better than motor, w_3 , although over-all no single control has bounds close to the full optimal bounds. Motor, w_3 is also probably better than screw s_3 , although this is not conclusive. Examining the pairs of inputs, the potential to produce a well controlled output resides with structure $\{s_2, w_1\}$, further the four bounds for this structure lie closest to those of the full input set. This pair also occurs in the larger subsets which have bounds close to the full input set, for example $\{s_1, s_2, w_1\}$, $\{s_2, w_1, w_2\}$, $\{s_1, s_2, w_1, w_2\}$ and $\{s_2, w_1, w_2, w_3\}$. These higher order structures

appear to be the most significant, suitable for further examination by direct simulation . The bounds for the low order structure were recalculated using the exact values for the maximum eigenvalues of the structure operators. These new bounds, which are given in Table 5.3.3, simply confirm the earlier conclusion drawn above.

Table 5.3.2 The second set of cost functional weights were chosen to penalize tension deviations much less, but maintain tight control over gauge errors. The resulting bounds give conclusions similar to those deduced for the previous table, except that they appear more distinctive. Among the single inputs, screw s_4 is useless, screw s_1 is impressive being better than screws s_2, s_3, s_4 , and motor, w_3 . Motor, w_1 is also powerful being rated above screws, s_3 and s_4 and motor, w_3 although again no single control has bounds close to those of the full input set. In the pairs of inputs, $\{s_2, w_1\}$ is prominent in having the potential to give the best output control and having bounds closest to the full structure bounds. This pair also feature among the significant larger input subsets, namely $\{s_2, w_1, w_2\}$, $\{s_1, s_2, w_1\}$, $\{s_1, s_2, w_1, w_2\}$, $\{s_2, w_1, w_2, w_3\}$ and $\{s_2, s_3, w_1, w_2\}$. Again the use of exact eigenvalues in the bound calculations do not alter these conclusions significantly.

For the case of a step disturbance to exit gauge, stand 2, the input pair $\{s_2, w_1\}$ were identified as being a near optimal structure. Among the larger sized subsets isolated for further investigation were the structures $\{s_2, w_1, w_2\}$ and $\{s_1, s_2, w_1, w_2\}$ which have been shown to be important by other techniques.

5.3.2 Bounds for a 10% Step Disturbance Interstand Tension, T_2 and Their Structural Implications

The results for cost functional weights sets I and II are contained in Table 5.3.4 and Table 5.3.6 respectively.

Table 5.3.4 The immediate conclusion to be drawn here is that screw, s_3 above has near optimal bounds, and is unquestionably better than the

remaining screws s_1 , s_2 , and s_4 and the motor w_1 . Refinement of the bounds obtained by using the actual eigenvalues of the structures (Table 5.3.5) shows further that motor w_2 may also be added to this list. Examining the pairs of inputs when all those containing screw, s_3 have been eliminated the best of the remaining structures is a pair of motors $\{w_1, w_2\}$. Since the single input, s_3 is so outstandingly good no larger sized subsets were considered above the pairs of inputs.

Table 5.3.6 With the second set of cost functional weights the results are even more distinct. The single input screw, s_3 is alone better than all the other input except motor w_3 . Refinement of these bounds (Table 5.3.7) removes this exception. As with the pairs in Table 5.3.4 once those containing the input s_3 have been removed the structures containing motor inputs become significant.

In conclusion, for the control of interstand tension, T_2 screw, s_3 is advocated unequivocally as the near optimal input structure. When considering pairs of inputs and after eliminating those containing the input s_3 the motor based input structures become significant. These findings compare most favourably with those obtained previously.

5.4 Conclusions

This example, a four stand tandem mill presented an ideal opportunity to assess the practicality of the input structure calculation and the accuracy of the conclusions drawn from the diagnostic bounds.

The computation of the bounds on the largest eigenvalue of the structure operators was not a formidable computing problem, and the indistinctness introduced into the diagnostic bounds was not enough to destroy the accuracy of the important structural conclusions to be drawn. Moment decomposition, and reconstitution would probably be the first areas to cause computing difficulties when the technique is applied to plants with larger numbers of inputs. This is because moment decomposition is a direct function of the number of inputs whilst moment reconstitution is a function of the maximum size of input structure subsets which means that the more inputs there are, the more subsets there are of maximum size. The reconstitution of moment sequences for larger input subset sizes requires proportionately more computing time. The actual calculation of the diagnostic bounds is not likely to give rise to computing problems but for large numbers of structures some form of ordering algorithm is required to facilitate assimilation of the results.

The conclusions drawn from the diagnostic bounds show very favourable agreement with established results for the disturbances considered. The place of the technique in the design process is also highlighted, since some of the indistinctness can only be removed by direct simulation. But then this is the purpose of the technique to isolate the most profitable structures for further investigation. It does not preclude the need for physical reasoning to justify the effectiveness of isolated structures. In fact this very justification should produce a better understanding of the system physics and behaviour which in turn may cause a modification of the cost functional weights and so on ~ the innovative process of design.

Table 5.3.1 10% Step Disturbance to Exit Gauge, Stand 2
Cost Functional Weights: Set I
 $Q_1 = \dots = Q_7 = 100.0$; $R_1 = \dots = R_7 = 1.0$
Minimum Structure Eigenvalue Bounds Used.

Structure	ΔJ_o (min)	ΔJ_o (max)	IESQ(min)	IESQ(max)
s_1	680.43	5641.77	39852.46	49317.97
s_2	96.11	1139.41	47836.78	49903.85
s_3	18.40	210.48	49597.67	49981.59
s_4	0.05	0.58	49998.89	49999.95
w_1	190.31	11058.07	30441.71	49809.67
w_2	69.08	5159.97	40266.16	49930.91
w_3	18.27	981.19	48073.70	49981.73
s_1, s_2	2923.73	43873.75	752.47	47057.61
s_1, s_3	830.16	37584.67	3091.03	49167.28
s_2, w_1	3445.23	47331.87	142.70	46542.77
s_2, w_2	547.03	44228.45	666.89	49452.42
s_1, s_2, s_3	5768.69	43886.28	749.22	44147.84
s_2, s_3, w_3	4169.62	45851.13	344.48	45800.13
s_2, s_4, w_2	3940.22	47353.59	140.38	46043.98
s_1, s_2, w_1	3566.54	47658.58	109.92	46422.33
s_2, w_1, w_2	3154.73	47605.11	114.97	46836.28
s_1, s_2, s_3, s_4	7002.23	44002.53	720.87	42874.50
s_1, s_2, s_3, w_2	9348.05	47141.40	163.53	40548.15
s_2, s_3, s_4, w_1	4468.89	47337.96	142.05	45510.39
s_1, s_2, s_4, w_1	4038.34	47677.47	108.15	45947.30
s_1, s_2, w_1, w_2	3274.48	47891.87	89.10	46717.09
s_2, w_1, w_2, w_3	1879.69	47682.50	107.67	48117.49
$s_1, \dots, s_4, w_1, \dots, w_3$	2162.29	47935.44	85.47	47833.76

Table 5.3.3 Bounds Above Recalculated using Exact Eigen Values

Structure	ΔJ_o (min)	ΔJ_o (max)	IESQ(min)	IESQ(max)
s_1	680.43	1891.00	46912.72	49317.97
s_2	96.11	651.59	48797.64	49903.85
s_3	18.40	136.07	49746.18	49981.59
s_4	0.05	0.32	49999.41	49999.95
w_1	190.31	3774.47	42897.51	49809.67
w_2	69.08	2314.50	45539.93	49930.91
w_3	18.27	385.48	49249.66	49981.73
s_1, s_2	2923.73	42071.02	1269.20	47057.61
s_1, s_3	830.16	18230.40	20446.84	49167.28
s_2, w_1	3445.23	47156.48	162.17	46542.77
s_2, w_2	547.03	40829.11	1687.71	49452.42

Table 5.3.2

10% Step Disturbance on Exit Gauge, Stand 2

Cost Functional Weights: Set II

 $Q_1 = \dots = Q_4 = 100.0$ $Q_5 = \dots = Q_7 = 10.0$ $R_1 = \dots = R_7 = 10.0$

Minimum Structure Eigenvalue Bounds Used.

Structure	ΔJ_o (min)	ΔJ_o (max)	IESQ(min)	IESQ(max)
s_1	3875.84	7718.03	38524.95	46073.28
s_2	933.00	1983.09	46962.96	49062.40
s_3	181.30	371.31	49433.66	49817.73
s_4	0.52	1.03	49998.43	49999.48
w_1	1743.29	13236.20	27972.67	48255.28
w_2	670.96	6024.26	39182.48	49328.36
w_3	177.14	1156.47	47875.92	49822.56
s_1, s_2	9709.41	40222.13	2041.78	40093.02
s_2, w_1	16126.57	47163.06	164.60	33606.38
s_2, w_2	4834.28	44242.94	669.34	45122.79
s_1, w_2	3424.68	40441.36	1855.62	46545.25
s_2, s_3, w_2	20143.55	45809.82	353.36	29149.98
s_2, s_4, w_1	17095.99	47185.06	162.03	32602.10
s_2, w_1, w_2	16642.90	47514.04	126.40	33105.04
s_1, s_2, w_2	16030.22	47474.84	130.69	33740.51
s_1, s_2, s_3, w_2	23927.18	47059.05	174.25	25391.43
s_2, s_3, s_4, w_2	20671.10	45816.02	352.31	28584.39
s_2, s_3, w_1, w_3	17647.94	47123.87	169.26	32019.40
s_1, s_2, w_1, w_2	16542.58	47780.41	100.93	33239.32
s_2, w_1, w_2, w_3	13015.94	47587.17	119.22	36844.92
s_2, s_3, w_1, w_2	15205.51	47557.30	122.07	34592.60
$s_1, \dots, s_4, w_1, \dots, w_3$	13823.55	47829.92	96.62	36036.08

Table 5.3.4 10% Step Disturbance to Interstand Tension, T_2
 Cost Functional Weights: Set I
 $Q_1 = \dots Q_7 = 100.0$; $R_1 = \dots R_7 = 1.0$
 Minimum Structure Eigenvalue Bounds Used

Structure	ΔJ_o (min)	ΔJ_o (max)	IESQ(min)	IESQ(max)
s_1	8764.44	22620.83	18068.41	41217.17
s_2	951.06	13227.85	27548.26	49048.43
s_3	48890.95	49966.71	0.04	1082.39
s_4	122.97	1428.81	47294.45	49876.92
w_1	6363.20	46328.63	307.38	43636.23
w_2	16207.62	49094.38	23.82	33790.70
w_3	37334.50	49902.12	0.60	12659.08
s_1, s_3	48561.04	49967.49	0.040	1413.74
s_2, s_3	48632.75	49969.68	0.025	1341.64
s_3, s_4	48966.16	49972.16	0.016	1008.15
s_3, w_3	46754.60	49986.09	0.009	3237.08
w_1, w_2	43059.86	49986.07	0.005	6932.03
$s_1, \dots, w_3$	49170.58	49995.23	0.00051	825.01

Table 5.3.5 Bounds of Table 5.3.4 Recalculated using
 Exact Eigenvalues

Structure	ΔJ_o (min)	ΔJ_o (max)	IESQ(min)	IESQ(max)
s_1	8764.44	16340.66	27380.18	41217.17
s_2	951.06	7639.70	36569.22	49048.43
s_3	48890.95	49965.22	0.06	1082.39
s_4	122.97	859.48	48411.58	49876.92
w_1	6363.20	41680.31	1582.62	43636.23
w_2	16207.62	47527.48	179.56	33790.70
w_3	37334.50	49702.32	6.45	12659.08
$s_1, \dots, w_3$	49170.58	49995.16	0.00054	825.01

Table 5.3.6 10% Step Disturbance to Interstand Tension, T_2
Cost Functional Weights: Set II
 $Q_1 = \dots = Q_4 = 100.0$, $Q_5 = \dots = Q_7 = 10.0$
 $R_1 = \dots = R_7 = 1.0$
Minimum Structure Eigenvalues Bounds Used.

Structure	ΔJo (min)	ΔJo (max)	IESQ (min)	IESQ (max)
s_1	506.26	710.18	4077.08	4487.57
s_2	91.96	245.23	4602.79	4907.57
s_3	4817.06	4963.38	0.67	157.04
s_4	12.03	25.06	4961.35	4987.86
w_1	580.66	2871.43	1020.56	4418.86
w_2	1579.06	4303.45	139.06	3419.36
w_3	3627.22	4899.65	5.93	1366.72
s_1, s_3	4790.78	4966.17	0.434	184.29
s_2, s_3	4791.92	4968.24	0.306	183.29
s_3, s_4	4942.78	4971.15	0.193	29.53
s_3, w_3	4642.46	4985.69	0.091	349.153
w_1, w_2	3975.84	4983.54	0.074	1017.23
w_1, w_3	3704.06	4982.24	0.092	1290.56
$s_1, \dots, w_3$	4761.47	4994.60	0.008	234.37

Table 5.3.7 Bounds of Table 5.3.6 Recalculated using
Exact Eigenvalues

Structure	ΔJo (min)	ΔJo (max)	IESQ (min)	IESQ (max)
s_1	506.26	620.22	4254.55	4487.57
s_2	91.96	167.62	4755.65	4907.57
s_3	4817.06	4960.72	0.99	157.04
s_4	12.03	19.12	4973.39	4987.86
w_1	580.66	1770.05	2356.03	4418.86
w_2	1579.06	3510.94	643.14	3419.36
w_3	3627.22	4713.37	55.67	1366.72
$s_1, \dots, w_3$	4761.47	4994.26	0.01	234.37

Chapter 6 Extensions and Applications in the Infinite Time Domain

6.0 Introduction

6.1 Input Structure Assessment and the Infinite Time Domain

6.2 Removal of Initial Condition Dependence of Input Structure Assessment for State Space Systems.

6.3 Diagonal Dominance and Structures

6.3.0 Introduction

6.3.1 On Pseudo-Diagonalisation

6.3.2 On Diagonal Dominance

6.4 Least Squares Parameter Estimation, and an Application of the Structures Procedure.

6.0 Introduction

The application of the structures assessment technique to problems defined on the infinite time interval are discussed. In particular, specialization to a state space description is made and the possibility of using Parseval's Theorem to evaluate the moment sequence explored. It is demonstrated that in the absence of suitable recurrent formulae, the calculation in the Laplace s -domain is a complicated procedure to automate. However sufficiently accurate finite time domain approximations are found to be adequate.

Next the problem of removing the initial condition dependence of the state-space regulator formulation is considered. The technique of Levine and Athans (1970) is shown to lead to the definition of an equivalent disturbance and also to an n -fold increase in the computational requirements.

The conclusion drawn is that the initial condition dependence cannot be removed (practically) so the choice of initial conditions should be well justified.

Some space is devoted to an application of structures concepts to a problem arising in the process of creating diagonal dominance as a pre-requisite to using the multivariable design techniques of Rosenbrock (1974).

Finally the application of the structures procedure to the problem of least squares parameter estimation is discussed. An example taken from Draper and Smith (1966) is analysed in some detail.

6.1 Input Structures Assessment, and the Infinite Time Domain

The Hilbert space formalism utilized in the earlier chapters of this thesis unified the problem statement and the subsequent analysis of the diagnostics. One of the appealing features of this operator description is that it accommodates the idea of obtaining the kernel of the system convolution directly from a process simulation. Consequently simple finite time domain operations can be used to calculate the moment sequences.

Here input structures for a problem posed in the infinite time domain are considered. The quadratic cost functional is

$$J = \frac{1}{2} \int_0^{\infty} \{ \langle y, Qy \rangle_{E_r} + \langle u, Ru \rangle_{E_m} \} dt$$

subject to linear system constraint $y(t) = \int_0^t W(t-\tau) u(\tau) d\tau + d(t)$

where $0 \leq t < \infty$. Lemma 4.4.3 (Chapter 4) may be extended to the infinite time-domain so that whenever the elements of the convolution kernel satisfy $w_{ij}(\cdot) \in L^1[0, \infty)$ and the input functions satisfy

$u \in L_m^2[0, \infty)$ then $z = W u$ satisfies $z \in L_r^2[0, \infty)$, where

$$z(t) \triangleq \int_0^t W(t-\tau) u(\tau) d\tau \in \mathbb{R}^r, \quad 0 \leq t < \infty.$$

Assuming $d \in L_r^2[0, \infty)$ gives $y = z + d \in L_r^2[0, \infty)$. Set therefore

the basic problem in Hilbert space $H_r \triangleq L_r^2[0, \infty)$, with the permissible

set for inputs $u \in H_m = L_m^2[0, \infty)$. In this manner the cost becomes

$$J = \frac{1}{2} \langle y, Qy \rangle_{H_r} + \frac{1}{2} \langle u, Ru \rangle_{H_m} \quad \text{subject to} \quad y = W u + d \in H_r, \text{ and } u \in H_m.$$

Clearly the asymptotic theory of Chapter 2, and the approximation procedures of Chapter 3 apply unchanged.

If the system description is restricted to have compatible state space form $\dot{S}(A,B,C)$, the evaluation of the moments may be developed in terms of real rational transfer matrices.

The usual assumption made to ensure an asymptotically stable control law is the complete observability of pair (A,DC) , where $D^T D = Q$.

The disturbance restriction $d \in L_r^2[0, \infty)$ will, for any arbitrary choice of initial condition imply that the system matrix is asymptotically stable, namely A will have wholly negative eigenvalues. This in turn ensures that the elements of the operator kernel satisfy $w_{ij}(\cdot) \in L^1[0, \infty)$

since $W(t) = C e^{At} B = C T^{-1} e^{At} T B$ giving $|w_{ij}(t)| \leq m_{ij} e^{-\alpha_{ij}(t)}$

some $m_{ij} > 0$ and $\alpha_{ij} > 0$ hence $w_{ij} \in L^1[0, \infty)$. With these

restrictions and forms consider the problem of computing the moment

sequence $m_0 = \frac{1}{2} \langle d, Qd \rangle_{H_r}$, $m_1 = \frac{1}{2} \langle W^* Qd, R^{-1} W^* Qd \rangle_{H_m}, \dots$

using Parseval's Theorem. The two operators to be transformed are the system convolution operator Wu , and the adjoint convolution operator W^*y .

Operator $(Wu)(t) \triangleq \int_0^t W(t-\tau) u(\tau) d\tau$ is causal so that its Laplace

transform is given as : $\mathcal{L}_1\{Wu\} \triangleq G(s) u(s)$ where $G(s) \triangleq \mathcal{L}_1\{W(\cdot)\}$.

Adjoint operator $(W^*y)(t) = \int_t^\infty W^T(\tau-t) y(\tau) d\tau$ is not causal since

it is defined for $-\infty < t < \infty$ so that the bilateral Laplace transform

is required to give $\mathcal{L}_2\{W^*y\} \triangleq G^T(-s) y(s)$ which is analytic only

on a strip which includes the imaginary axis. Take as an example the

calculation of the moments $m_1 = \langle W^* Qd, R^{-1} W^* Qd \rangle_{H_m}$

and $m_2 = \langle WR^{-1} W^* Qd, QWR^{-1} W^* Qd \rangle_{H_r}$. In the moment m_1 , the function

$(W^* Qd)(t)$ is only required on the interval $[0, \infty)$ similarly in moment m_2 , where $W^* Qd$ serves as the input to system WR^{-1} . For this reason it is important when inverting $G^T(-s)y(s)$ to be able to obtain only that part of the function defined on $[0, \infty)$. Using the inverse Laplace transform this is $(W^* y)(t) H(t) = \frac{1}{2\pi j} \int_{-j\omega}^{j\omega} G^T(-s) y(s) e^{st} ds$

with the region of convergence being the closed right half plane. When real rational transfers are used, this can be achieved in the s-domain using a projection $P_+ \triangleq \{\cdot\}_+$ to isolate the open left half plane transfers (analytic in the closed right half plane : hence the "+" notation) by partial fraction expansions. Thus for the adjoint convolution : $\mathcal{L}_1\{ (W^* y)(t) \cdot H(t) \} = P_+ \mathcal{L}_2\{ W^* y(t) \} = P_+ \{ G^T(-s)y(s) \} = \{ G^T(-s)y(s) \}_+$

With this modification, Parseval's Theorem is used to evaluate the moment inner products.

Parseval's Theorem Let $x(t)$ and $y(t)$ have unilateral Laplace transforms $x(s)$ and $y(s)$. If all the elements of $x(s)$ are analytic in the closed right half plane (that is the region of convergence includes the imaginary axis) then

$$\int_0^\infty \langle y(t), x(t) \rangle dt = (2\pi j)^{-1} \int_{-j\infty}^{j\infty} \langle y(-s), x(s) \rangle ds$$

where equality holds for all those pairs (x, y) for which the time domain integral exists.

The restriction of the system to state space form $\mathcal{S}(A, B, C)$ implies that $G(s)$ and $d(s)$ have real rational entries consequently the

above contour ^eintegral may be evaluated by appeal to Cauchy's Residue Theorem. The generation of the moment sequence takes the following pattern :

$$i = 0 \quad m_0 = (2\pi j)^{-1} \int_{-j\infty}^{j\infty} \langle d(-s), Qd(s) \rangle_{E_r} ds$$

$i = 1$ Form $W^T(-s)Qd(s)$ and subsequently

$$f_1(s) = \{ W^T(-s)Qd(s) \}_+$$

$$\text{Evaluate } m_1 = (2\pi j)^{-1} \int_{-j\infty}^{j\infty} \langle f_1(-s), R^{-1}f(s) \rangle_{E_m} ds$$

$i = 2$ Form $f_2(s) = W(s)R^{-1}f_1(s)$

$$\text{Evaluate } m_2 = (2\pi j)^{-1} \int_{-j\infty}^{j\infty} \langle f_2(-s), Qf_2(s) \rangle_{E_r} ds$$

and so on.

Clearly for diagonal R , the structure of the finite time domain moments has been preserved so that the decomposition algorithm of Chapter 4 will still apply. However the actual evaluation of moments now requires the following operations

- 1) manipulations of matrices, and vectors whose entries are now real rational transfers,
- 2) a projection on these entries to isolate the part analytic in the closed right half plane, and
- 3) residue evaluations to obtain the moment values.

The consideration of even a simple example will show that the automation of such operations is not trivial and are frequently a practice unwarranted since a finite time domain approximation of sufficient accuracy can be used. For example, recall $m_0 = \frac{1}{2} \langle d, Qd \rangle_{H_m}$

where $d(t) = C e^{At} x_0$, $0 \leq t < \infty$ and system matrix A has wholly negative eigenvalues. The latter condition ensures that

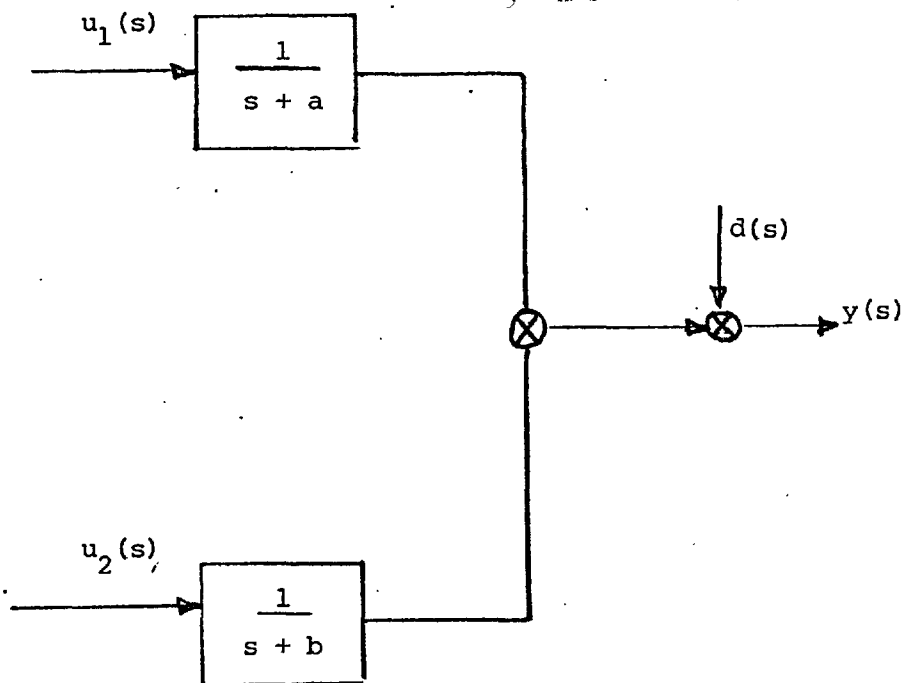
$$M(T_f) = \int_0^{T_f} e^{AT} C^T Q C e^{At} dt \rightarrow M(\infty) \quad \text{as } T_f \rightarrow \infty$$

and $\|M(\infty)\| < \infty$. Hence for any $\rho > 0 \exists \tilde{T}_f$ such that

$$\forall T_f > \tilde{T}_f \quad 0 < |m_o(T_f) - m_o(\infty)| < \rho.$$

Dependent therefore on the modes of $\mathcal{S}(A,B,C)$, a sufficiently accurate finite time domain approximation may often be utilized. To illustrate these assertions the pertinent details of an example are given :

System



$$d(s) = \frac{x_1(o)}{(s+a)} + \frac{x_2(o)}{(s+b)} \quad Q = 1, R = I_2$$

$$a = 1.618 \quad b = 0.618 \quad x_1(o) = x_2(o) = 1$$

$$m_o(T_f) = \frac{1 - e^{-aT_f}}{2a} + \frac{2(1 - e^{-(a+b)T_f})}{(a+b)} + \frac{(1 - e^{-bT_f})}{2b}$$

The value of T_f required to obtain a predetermined percentage error in $m_o(\infty)$ are given in Table 6.1.1

Table 6.1.1 Error in $m_o(\infty)$ versus Approximation Interval, T_f .

% Error in $m_o(\infty)$	20	10	5	1	0.1	10^{-4}
Approximation Interval T_f	0.9	1.34	1.82	3.03	4.8	10.0

Another comparison was made between the length of the approximation interval T_f , and the integration interval, T . These results are given in Table 6.1.2.

Table 6.1.2 Variation of Numerical Results for Different Approximation Intervals, T_f and Integration Intervals, T .

Analytical Results		Calculated using m_o, m_1, m_2, λ_M	
Structure	ΔJ_o^S	Lower Bound	Upper Bound
$\{u_1\}$	0.25314201	0.251607226	0.2723069
$\{u_2\}$	0.427210505	0.422411386	0.628971447
$\{u_1, u_2\}$	0.5818624	0.519317	0.763347

Numerical Results

<u>Lower Bound</u>	(*)		(**)	
Structure	$T_f=10.0 \ T=0.1,$	$T_f=10.0 \ T=0.02,$	$T_f=50.0 \ T=0.1,$	$T_f=50.0 \ T=0.02$
$\{u_1\}$	0.25527686	0.25175216	0.25527816	0.25175845
$\{u_2\}$	0.42597642	0.42262345	0.42590304	0.42255068
$\{u_1, u_2\}$	0.52479320	0.51959885	0.52472639	0.51953270
<u>Upper Bound</u>				
$\{u_1\}$	0.27576714	0.27244339	0.27576909	0.27244506
$\{u_2\}$	0.63259861	0.62908920	0.63262694	0.62911751
$\{u_1, u_2\}$	0.76849358	0.76352980	0.76851769	0.76355393

The two cases (*) and (**) required the same computer storage requirements and algorithms would have performed the same number of operations but the increase in integration accuracy far outweighs the effect of extending the approximation interval, T_f .

It is therefore concluded that a judicious choice of approximation interval, and integration step size, T , will enable infinite time domain problems (state space descriptions) to be replaced by a satisfactory finite time domain approximation. The main advantage here being the preservation of the moment calculated using straightforward system and adjoint system convolutions. Stiff systems with very lightly damped eigenvalues will (as usual) need careful treatment.

6.2 Removal of the Initial Condition Dependence of Input Structure Assessment for State Space Systems.

The approach followed here is that of Levine and Athans (1970) who were concerned with the elimination of initial condition dependence when finding constant feedback gains.

The system description is restricted to an asymptotically stable compatible matrix triple $\mathcal{S}(A,B,C)$ subject to an initial condition $x(0) = x_0$. In terms of the state-space process write

$$y(t) = \int_0^t W(t-\tau) u(\tau) d\tau + d(t) \quad 0 \leq t < \infty$$

$$\text{where } W(t) = C e^{At} B \quad \text{and} \quad d(t) = C e^{At} x_0. \quad x_0 \in \mathbb{R}^n$$

$$\text{The cost is } J = \frac{1}{2} \langle y, Qy \rangle_{H_r} + \frac{1}{2} \langle u, Ru \rangle_{H_m}.$$

Defining $T \triangleq \{ t : 0 \leq t \leq T_f \}$, let $A : T \rightarrow \mathbb{R}^{m \times n}$

and $B : T \rightarrow \mathbb{R}^{m \times n}$, and define a matrix valued quantity

$$[A,B] \triangleq \int_0^{T_f} A^T(t) B(t) dt \quad \text{where } [A,B] : \mathbb{R}^n \rightarrow \mathbb{R}^n. \quad \text{Also extend}$$

the definition of the convolution operator to accept matrix valued inputs namely $(WU)(t) \triangleq \int_0^t W(t-\tau) U(\tau) d\tau$ where $U(t) : \mathbb{R}^l \rightarrow \mathbb{R}^m$, $t \in [0, T_f]$

This enables expressions of the form $[A,WU]$ to be defined as

$$\begin{aligned} [A,WU] &\triangleq \int_0^{T_f} A^T(t) \int_0^t W(t-\tau) U(\tau) d\tau dt \\ &= \int_0^{T_f} \begin{bmatrix} a_{\cdot 1}^T(t) \\ \vdots \\ a_{\cdot \ell_1}^T(t) \end{bmatrix} \left[\int_0^t W(t-\tau) U_{\cdot 1}(\tau) d\tau, \dots, \int_0^t W(t-\tau) U_{\cdot \ell_2}(\tau) d\tau \right] dt \end{aligned}$$

$$\begin{aligned}
&= \begin{bmatrix} \langle a_{.1}, Wu_{.1} \rangle_{H_r} , \langle a_{.1}, Wu_{.2} \rangle , \dots \langle a_{.1}, Wu_{.l_2} \rangle \\ \vdots \\ \langle a_{.l_1}, Wu_{.1} \rangle , \dots \dots \dots \langle a_{.l_1}, Wu_{.l_2} \rangle \end{bmatrix} \\
&= \begin{bmatrix} \langle W^* a_{1.}, u_{.1} \rangle_{H_m} \dots \dots \dots \langle W^* a_{1.}, u_{.l_2} \rangle_{H_m} \\ \vdots \\ \langle W^* a_{l_1.}, u_{.1} \rangle_{H_m} \dots \dots \dots \langle W^* a_{l_1.}, u_{.l_2} \rangle_{H_m} \end{bmatrix} \\
&= \int_0^{T_f} \left[\int_t^{T_f} W^T(\tau-t) A(\tau) d\tau \right]^T U(t) dt \\
&\triangleq [W^* A, U]
\end{aligned}$$

Returning to the problem as given, recall

$$\Delta J_0 = \frac{1}{2} \langle b, A^{-1} b \rangle_{H_m} \quad \text{where} \quad b = W^* Q d \quad \text{and} \quad A = W^* Q W + R. \quad \text{The}$$

disturbance is given as $d(t) = C e^{At} x_0 = D(t) x_0$ where $D(t) \triangleq C e^{At}$

Define operator $D : \mathbb{R}^n \rightarrow H_r$ such that $(Dx_0)(t) \triangleq D(t)x_0$.

$$\begin{aligned}
\text{Then } \Delta J_0 &= \frac{1}{2} \langle x_0, [W^* Q D, A^{-1} W^* Q D] x_0 \rangle_{E_n} \\
&= \frac{1}{2} \text{trace} \{ [W^* Q D, A^{-1} W^* Q D] x_0 x_0^T \}
\end{aligned}$$

Following Athans and Levine (1970) make the assumption that the initial condition x_0 is a vector random variable distributed uniformly over the surface of an n-dimensional unit sphere, and that the cost is averaged over all such initial conditions. Hence $E[x_0 x_0^T] = I_n/n$

$$\text{giving } E[\Delta J_0] = 1/2n \text{ trace } [W^* QD, A^{-1} W^* QD]$$

It is now shown how previously developed theory can be used to find bounds for this new problem.

$$\begin{aligned} \text{Let } \overline{\Delta J}_0 &= n \times E[\Delta J_0] = \frac{1}{2} \text{ trace } [W^* QD, A^{-1} W^* QD] \\ &= \frac{1}{2} \text{ trace } [D, QWA^{-1} W^* QD] \end{aligned}$$

$$\text{Recall that } (D)(t) \triangleq D(t) = [d_1(t), \dots, d_n(t)] \triangleq [d_1, \dots, d_n](t)$$

$$\overline{\Delta J}_0 = \frac{1}{2} \text{ trace } [[d_1, \dots, d_n], QWA^{-1} W^* Q [d_1, \dots, d_n]]$$

$$\begin{aligned} \overline{\Delta J}_0 &= \frac{1}{2} \text{ trace } \begin{bmatrix} \langle d_1, QWA^{-1} W^* Q d_1 \rangle, \dots, \langle d_1, QWA^{-1} W^* Q d_n \rangle \\ \vdots \\ \langle d_n, QWA^{-1} W^* Q d_1 \rangle, \dots, \langle d_n, QWA^{-1} W^* Q d_n \rangle \end{bmatrix} \\ &= \frac{1}{2} \sum_{i=1}^n \langle d_i, QWA^{-1} W^* Q d_i \rangle_{H_n} \end{aligned}$$

Using the Householder equality of Chapter 3, and the subsequent eigen-analysis for the estimation of a term of the form

$$\langle W^* Qd, A^{-1} W^* Qd \rangle_{H_m} \text{ obtain}$$

$$\begin{aligned} \overline{\Delta J}_0 &= \frac{1}{2} \sum_{i=1}^n \langle W^* Qd_i, A^{-1} W^* Qd_i \rangle_{H_m} = \frac{1}{2} \sum_{i=1}^n \left(\sum_{j=1}^{\infty} \frac{v_{ij}^2 \lambda_j}{(1+\lambda_j)} \right) \\ &= \sum_{j=1}^{\infty} \frac{v_j^2 \lambda_j}{(1+\lambda_j)} \quad \text{where } v_j^2 = \frac{1}{2} \sum_{i=1}^n v_{ij}^2 \end{aligned}$$

Here $\lambda_j \in (0, \lambda_M]$ are the denumerable eigenvalues of the compact,

positive semi-definite, self adjoint operator $\Phi = Q^{\frac{1}{2}} W R^{-1} W^* Q^{\frac{1}{2}}$ and

the coefficients $\{v_{ij}\}_{j=1}^{\infty}$ are the eigenvector co-ordinates of the

disturbance $Z_i = Q^{\frac{1}{2}} d_i$ with respect to the ortho-normal eigenbasis

$\{w_j\}_{j=1}^{\infty}$ of the operator $Q^{1/2} W R^{-1} W^* Q^{1/2}$. As before introduce the j_1 th order moment $m_{j_1} \triangleq \frac{1}{2} \text{trace} [Q^{1/2} D, \phi^{j_1} Q^{1/2} D]$ $j_1 = 0, 1, \dots$

where $m_0 = \frac{1}{2} \text{trace} [Q^{1/2} D, Q^{1/2} D]$

$$= \frac{1}{2} \sum_{i=1}^n \langle Q^{1/2} d_i, Q^{1/2} d_i \rangle_{H_r}$$

Setting $z_i = Q^{1/2} d_i = \sum_{j=1}^{\infty} v_{ij} w_j + v_{i0} w_0$

where $v_{i0} w_0$ is a possible component from $N(\phi = Q^{1/2} W R^{-1} W^* Q^{1/2})$

$$= \frac{1}{2} \sum_{i=1}^n \left(\sum_{j=0}^{\infty} v_{ij}^2 \right)$$

$$m_0 = \sum_{j=0}^{\infty} v_j^2 \quad v_j^2 = \frac{1}{2} \sum_{i=1}^n v_{ij}^2, \quad j = 0, 1, 2, \dots$$

and $m_{j_1} = \frac{1}{2} \text{trace} [Q^{1/2} D, \phi^{j_1} Q^{1/2} D]$ $j_1 = 1, 2, \dots$

$$= \frac{1}{2} \sum_{i=1}^n \langle Q^{1/2} d_i, \phi^{j_1} Q^{1/2} d_i \rangle_{H_r}$$

Setting $z = Q^{1/2} d_i = \sum_{j=1}^{\infty} v_{ij} w_j + v_{i0} w_0$ $w_0 \in N(\phi)$

$$\Rightarrow m_j = \sum_{i=1}^n \sum_{j=1}^{\infty} v_{ij}^2 (\lambda_j)^{j_1} = \sum_{j=1}^{\infty} v_j^2 (\lambda_j)^{j_1} \quad j_1 = 1, 2, \dots$$

In summary, the resulting problem is :

$$\text{Extremise } \overline{\Delta J}_0 = \sum_{j=1}^{\infty} v_j^2 \lambda_j / (1 + \lambda_j) \quad \{v_j^2 \geq 0\}_{j=0}^{\infty}$$

$$\{0 < \lambda_j \leq \lambda_M\}_{j=1}^{\infty}$$

$$\text{with } m_0 = \sum_{j=0}^{\infty} v_j^2$$

$$\text{and } m_{j_1} = \sum_{j=1}^{\infty} v_j^2 (\lambda_j)^{j_1} \quad j_1 = 1, 2, \dots$$

The immediate conclusion is that it is possible to remove the initial condition dependence using the device of Levine and Athans, and the new problem has the form of the one discussed and solved in Chapter 3. (See in particular equations 3.1.2 to 3.1.4.) The actual form of the cost improvement, that is

$$\overline{\Delta J}_0 = \sum_{i=1}^n \langle W^* Q d_i, A^{-1} W^* Q d_i \rangle,$$

is the sum of n cost improvements resulting from the application of n individual initial conditions of the form $x(0)_i^T =$

$(0, 0, \dots, 1, 0, \dots, 0)$ where the unit is in the i th position.

Such a systematic set of initial conditions ensures that all the system modes are excited whether uncontrollable or not, so that any uncontrolled subsystem occurring with a particular input structure is not omitted from the structure assessment by a freak choice of initial condition. Although, the problem has the same form as that given previously it will now be shown that an n fold increase in computational requirements has occurred to implement the decomposition algorithm. Consider the first three moments m_0 , m_1 and m_2 :

$$\text{moment } m_0 = \text{trace} [D, QD]$$

$$= \sum_{i=1}^n \langle d_i, Q d_i \rangle_{H_r}$$

This moment, although independent of control structures requires the summation of n inner product terms on H_r .

$$\text{moment } m_1 = \text{trace} [W^* QD, R^{-1} W^* QD]$$

$$= \sum_{i=1}^n \langle W^* Q d_i, R^{-1} W^* Q d_i \rangle_{H_m}$$

$$\begin{aligned}
&= \sum_{i=1}^n \sum_{j=1}^m \langle W_j^* Qd_i, R_j^{-1} W_j^* Qd_i \rangle_{H_m} \\
&= \sum_{i=1}^n \sum_{j=1}^m m_{1j}^i = \sum_{j=1}^m \left(\sum_{i=1}^n m_{1j}^i \right)
\end{aligned}$$

Again each of the inner products $\langle W_j^* Qd_i, R_j^{-1} W_j^* Qd_i \rangle_{H_m}$ decomposes

into m parts (as in the original scheme) but now there are n different disturbances for which this decomposition is required.

$$\text{moment } m_2 = \text{trace} [WR^{-1} W^* QD, QWR^{-1} W^* QD]$$

$$= \sum_{i=1}^n \langle WR^{-1} W^* Qd_i, QWR^{-1} W^* Qd_i \rangle_{H_r}$$

Following the previous pattern each of the n terms

$$\langle WR^{-1} W^* Qd_i, QWR^{-1} W^* Qd_i \rangle_{H_r} \text{ can be decomposed into } m(m+1)/2$$

moment parts to give

$$\langle WR^{-1} W^* Qd_i, QWR^{-1} W^* Qd_i \rangle_{H_r} = \sum_{j=1}^m m_{2jj}^i + 2 \sum_{j=1}^{m-1} \sum_{k=j+1}^m m_{2jk}^i$$

Substituting into the expression for m_2 to obtain :

$$\begin{aligned}
m_2 &= \sum_{i=1}^n \left(\sum_{j=1}^m m_{2jj}^i + 2 \sum_{j=1}^{m-1} \sum_{k=j+1}^m m_{2jk}^i \right) \\
&= \sum_{j=1}^m \left(\sum_{i=1}^n m_{2jj}^i \right) + 2 \sum_{j=1}^{m-1} \sum_{k=j+1}^m \left(\sum_{i=1}^n m_{2jk}^i \right)
\end{aligned}$$

In total, for moment, m_2 , $m(m+1)/2$ control structure dependent parts are obtained but each of these parts requires inner product terms arising from n different disturbances.

The usual assumption placed on n , the number of states is $n \gg \max(r, m)$ which if satisfied would imply an impossibly large increase in the computational requirement for the decomposition technique. Note that the estimation of an upper bound for the maximum eigenvalue, λ_M

remains unaffected by this treatment, (this is because the kernel operator $Q^{\frac{1}{2}}WR^{-1}W^*Q^{\frac{1}{2}}$ is unchanged.)

Using the eigenbasis of the operator $Q^{\frac{1}{2}}WR^{-1}W^*Q^{\frac{1}{2}}$ it is possible to formally construct an equivalent disturbance and eliminate the n different disturbances, but such a technique is computationally infeasible.

In conclusion this analysis has not been without value because it highlights the need for a sensible choice of initial condition and if n is small gives a technique for control structures in the regulator problem.

6.3 Diagonal Dominance and Structures

6.3.0 Introduction

The concept of diagonal dominance (Rosenbrock, 1974) has arisen as a pre-requisite to the use of two multivariable design procedures, namely the Nyquist and Inverse Nyquist array techniques. Although much has been written of the multivariable design techniques very little guidance has been forthcoming on making a system transfer diagonally dominant. Hawkins (1972) developed the technique of pseudo-diagonalisation and Hannah (1975) extended this work as a non-linear optimization problem with some success. Hannah also gave some guidelines for choosing constant pre-compensators based on the systems impulse response matrix, but these results were given as observations and without theoretical justification.

The treatment given here is a reappraisal of the concept of pseudo-diagonalisation which leads to an alternative problem statement. A natural extension of this problem is proposed to achieve diagonal dominance, an extension which has a time domain interpretation, and permits the assessment of precompensator structures.

Dominance (Rosenbrock 1974) is defined as follows:

Let $g_{ij}(s)$ denote the (i,j) th element of the square system transfer matrix $G(s)$ and let $g_{ii}(s)$, $i=1,2,\dots,m$ map the usual Nyquist contour D into $\Gamma_i(s)$ in the complex plane then

$$\text{if } |g_{ii}(s)| > \sum_{\substack{j=1 \\ j \neq i}}^m |g_{ij}(s)| \quad i=1,2,\dots,m; \quad \forall s \text{ on } D \text{ the system}$$

is diagonally row dominant,

if $|g_{ii}(s)| > \sum_{\substack{j=1 \\ j \neq i}}^m |g_{ji}(s)|$, $i=1,2,\dots,m$ $\forall s$ on D the system

is diagonally column dominant and if the system is diagonally row dominant $\forall s$ on D or is diagonally column dominant $\forall s$ on contour D it is termed as diagonally dominant. Pseudo-diagonal dominance (Hawkins, 1972) is dominance at a particular frequency, usually $s_0 = iw_0$. This concept was introduced as a method for selecting constant pre-compensators to create diagonal dominance. The motivation for such a technique was mainly empirical since a precompensator pseudo-diagonalising at some frequency w_0 was often found to pseudo-diagonalise over a range of frequencies and frequently to produce diagonal dominance per se.

6.3.1 On Pseudo-Diagonalisation

Pose the problem of choosing a constant precompensator matrix K to make a given transfer $G(s)$ dominant at the particular frequency $s = iw$.

Hence $Q(s) = G(s)K$ at $s = iw$ giving $Q(iw) = G(iw)K$

This can be reduced to m independent subproblems by choosing to examine column dominance (note that if a post compensator is used, namely $Q(s) = KG(s)$, then the consideration of row dominance yields a similar decomposition), in which case $q_{\bullet j} = G(iw)k_{\bullet j}$ $j=1,2,\dots,m$ and the associated dominance cost is

$$J_{Dj}(k_{\bullet j}) = |q_{jj}(iw)| - \sum_{\substack{i=1 \\ i \neq j}}^m |q_{ij}(iw)| \quad j=1,2,\dots,m$$

The j th subproblem is to choose constant precompensator column $k_{\bullet j}$ so that $J_{Dj}(k_{\bullet j}) > 0$, or possibly improve the dominance which already exists at $s = iw$.

Write $G(i\omega) = \alpha + i\beta$ where $\operatorname{Re}[G(i\omega)] = \alpha$ and

$\operatorname{Im}[G(i\omega)] = \beta$. (Arrays α , and β are $m \times m$).

Hence $q_{\cdot j}(i\omega) = G(i\omega)k_{\cdot j} = [\alpha + i\beta] k_{\cdot j}$

or expressing $q_{\cdot j}(i\omega)$ in real and imaginary parts obtain

$$q_1^j + iq_2^j = \alpha k_{\cdot j} + i\beta k_{\cdot j} \quad q_1^j, q_2^j \in \mathbb{R}^m.$$

$$\text{giving } q \triangleq \begin{bmatrix} q_1^j \\ q_2^j \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} k_{\cdot j} = W k_{\cdot j}$$

where $q \in \mathbb{R}^{2m}$, $k_{\cdot j} \in \mathbb{R}^m$ and $W: \mathbb{R}^m \rightarrow \mathbb{R}^{2m}$.

The cost function for the j th dominance subproblem may be written

$$J_{D_j}(q) = (q^2(j) + q^2(j+m))^{\frac{1}{2}} - \sum_{\substack{i=1 \\ i \neq j}}^m (q^2(i) + q^2(i+m))^{\frac{1}{2}}$$

and the set of outputs q which are dominant for the j th column

may be specified as $Q_{D_j} \triangleq \{q \in \mathbb{R}^{2m} : J_{D_j}(q) > 0\}$

Observe that Q_{D_j} has the following properties :

- 1) If $q \in Q_{D_j} \Rightarrow \lambda q \in Q_{D_j} \quad \forall \lambda \in \mathbb{R} - \{0\}$
- 2) Let P_j be a projection $P_j : \mathbb{R}^{2m} \rightarrow \operatorname{span}\{e_j, e_{j+m}\}$. If for given $q \in \mathbb{R}^{2m}$, $P_j q = 0$ then $q \notin Q_{D_j}$. (Here e_j and e_{j+m} are the j th and $(j+m)$ th unit basis vectors for $\mathbb{R}^{2m}$.)
- 3) Simple counter-examples may be constructed to show that Q_{D_j} is not a subspace nor is it convex. Similar properties are true of the complement of Q_{D_j} .
- 4) The maximum of $J_{D_j}(q)$ over the set $\|q\|_2 = \mu$, $q \in Q_{D_j}$ is attained by $q \in \operatorname{sp}(e_j, e_{j+m})$ normalized so that $\|q\|_2 = \mu$.

Varied over this set $\|q\|_2 = \mu$, $q \in Q_{D_j}$; $J_{D_j}(q)$ satisfies the inequality $0 < J_{D_j}(q) \leq \mu$

Incorporating the linear constraint $q = W k_{\cdot j}$ observe that

- 5) If $\mathcal{R}(w) \cap Q_{D_j} = \emptyset$, then there exists no $k_{\cdot j}$ to make the j th column dominant at frequency $s = iw_0$. This in turn implies that there is no constant precompensator column $k_{\cdot j}$ to make the j th column dominant over any frequency range which contains w_0 .

Observation 1) is important since it indicates that if any q is column dominant, then $\lambda q = W(\lambda k_{\cdot j})$ is also column dominant where the dominance costs are related by $J_{D_j}(\lambda q) = \lambda J_{D_j}(q)$, so that the

dominance cost can be made arbitrary large by suitable scaling. By specifying a norm on the dominance output, observation 4) specifies the maximum value of the dominance cost over this set, and the form of output q to attain it. Observation 5) gives the necessary and sufficient condition for the existence of dominant output q , namely $\mathcal{R}(w) \cap Q_{D_j} \neq \emptyset$. However conditions on W so that this is true have not to date been forthcoming. It is therefore assumed that this feasible set of dominant outputs for the subsequent minimizations is not empty.

The square-root operator within the summand of the cost function $J_{D_j}(q)$ does not make it a very attractive function for further analysis. Consideration of the geometric situation enables a new optimization problem to be posed. Impose a length constraint

on the output q , $||q||_2 = \mu$, then Fig. 6.3.1.1 gives a geometric interpretation where q is depicted in relation to the subspace $\text{span}(e_j, e_{j+m})$. This is the subspace of maximally dominant outputs q , in the sense that if $q \in \text{span}(e_j, e_{j+m})$ then $J_{D_j}(q)$ attains its maximum value μ . In pursuit of this ideal $||q||_2 = \mu$

recall the projection operator $P_j : \mathbb{R}^{2m} \rightarrow \text{span}(e_j, e_{j+m})$ where

$$P_j \cong \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{is an } (2m \times 2m) \text{ idempotent matrix with}$$

unity in the j th and $(j+m)$ th diagonal entries.

For any $q \in \mathbb{R}^{2m}$, $q = (I - P_j + P_j)q = (I - P_j)q + P_j q$

where $(I - P_j)q \perp P_j q$, and $P_j q \in \text{span}(e_j, e_{j+m})$. On Fig. 6.3.1.1,

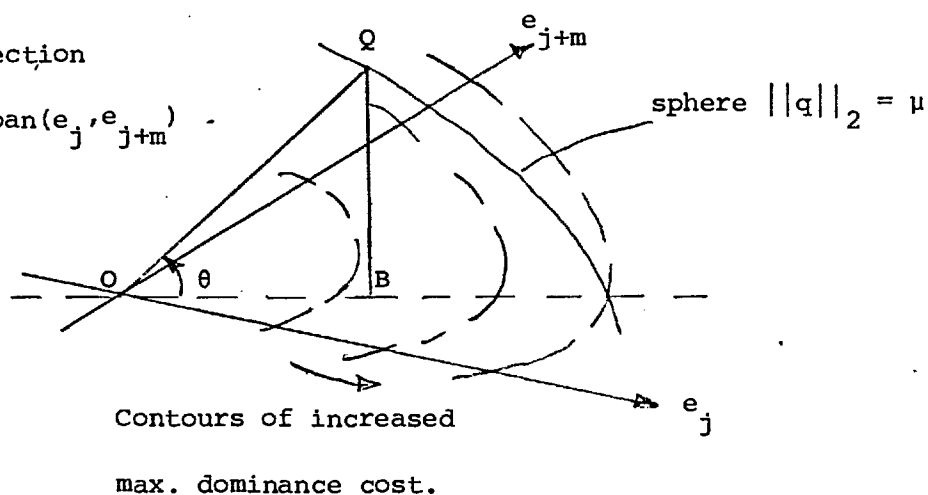
$\vec{OQ}$ represents q , $\vec{OB}$ represents $P_j q$ and BQ represents length

$$||(I - P_j)q||$$

Fig. 6.3.1.1 The Geometric Configuration of $q \in Q_{D_j}$ s.t. $||q||_2 = \mu$

Define a projection

$$P_j : \mathbb{R}^{2m} \rightarrow \text{span}(e_j, e_{j+m})$$



From the geometrical configuration several optimization problems may be proposed all related to reducing the size of angle QOB labelled θ , so that q is as close as possible to attaining maximal dominance, μ .

$$\tan \theta = \frac{QB}{OB} = \frac{|| (I-P_j)q ||_2}{|| P_j q ||} = \left[\sum_{\substack{i=1 \\ i \neq j}}^m q^2(i) + q^2(i+m) \right]^{\frac{1}{2}} / (q^2(j) + q^2(j+m))^{\frac{1}{2}}$$

$$\text{Hence } \tan^2 \theta = \frac{\sum_{\substack{i=1 \\ i \neq j}}^m \frac{q^2(i) + q^2(i+m)}{q^2(j) + q^2(j+m)}}{\left[\sum_{\substack{i=1 \\ i \neq j}}^m (q^2(i) + q^2(i+m))^{\frac{1}{2}} \right]^2} \leq \left[\sum_{\substack{i=1 \\ i \neq j}}^m \frac{(q^2(i) + q^2(i+m))^{\frac{1}{2}}}{(q^2(j) + q^2(j+m))^{\frac{1}{2}}} \right]^2$$

An equivalent expression for $\tan^2 \theta$ is obtained using a scale factor

$k^2 > 0$ chosen so that $k^2 (q^2(j) + q^2(j+m)) = \mu^2$ then

$$\tan^2 \theta = \sum_{\substack{i=1 \\ i \neq j}}^m \frac{(kq(i))^2 + (kq(i+m))^2}{(kq(j))^2 + (kq(j+m))^2} = \sum_{\substack{i=1 \\ i \neq j}}^m \frac{(kq(i))^2 + (kq(i+m))^2}{\mu^2}$$

Define $\tilde{q} = kq$ where $q \in \mathcal{R}(w) \cap \mathcal{Q}_{D_j}$ so that $\tilde{q} \in \mathcal{R}(w) \cap \mathcal{Q}_{D_j}$

to give the following problem equivalence :

$$\min_{\substack{||q||_2 = \mu \\ q \in \mathcal{R}(w)}} \left\{ \tan^2 \theta = \sum_{\substack{i=1 \\ i \neq j}}^m \frac{(q^2(i) + q^2(i+m))}{q^2(j) + q^2(j+m)} \right\}$$

$$\equiv \min_{\substack{||P_j q||_2 = \mu \\ q \in \mathcal{R}(w)}} \left\{ \sum_{\substack{i=1 \\ i \neq j}}^m (q^2(i) + q^2(i+m)) / \mu^2 \right\} \quad 6.3.1.1$$

In terms of the projection operator define

$$T_j(q) \triangleq \sum_{\substack{i=1 \\ i \neq j}}^m (q^2(i) + q^2(i+m)) = || (I-P_j)q ||_2^2.$$

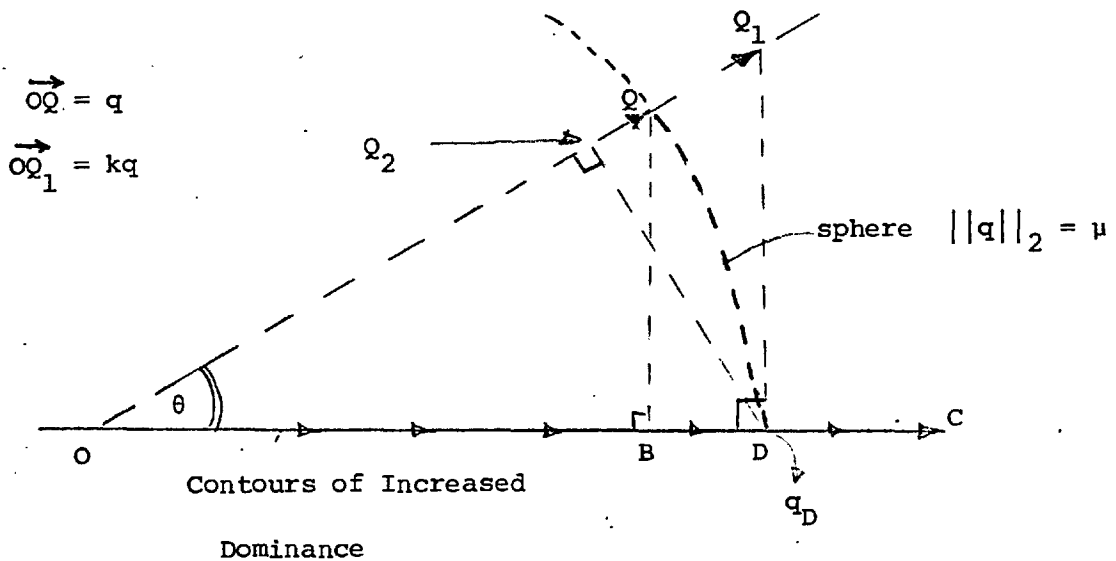
In the left hand side of the equivalence 6.3.1.1. there is no loss of generality in restating the minimization as :

$$\begin{aligned} \text{Min}_{\substack{\|P_j q\|_2 = \mu \\ q \in \mathcal{R}(w)}} T_j(q) &= \|(I - P_j)q\|_2^2 \end{aligned}$$

6.3.1.2.

This can be given geometrical significance by isolating triangle QOB from figure 6.3.1.1. as shown below in Fig 6.3.1.2. Clearly the scalar factor k , chosen to produce the equivalence relation 6.3.1.1., magnifies vector $\vec{OQ}$ representing q , to $\vec{OQ}_1$ representing kq . Under these circumstances $\|P_j(kq)\|_2 = \mu$ and the problem 6.3.1.2 seen to be the minimization of length Q_1D which incidently also minimizes $\tan^2 \theta$.

Fig 6.3.1.2. Geometric Configuration for Optimization Problems .



Two other related optimization problems may also be given directly from the geometrical configuration above :

$$\begin{aligned} 1) \quad \text{Min}_{\substack{\|q\|_2 = \mu \\ q \in \mathcal{R}(w)}} \|(I - P_j)q\|_2^2 \quad \text{and} \quad 2) \quad \text{Max}_{\substack{\|q\|_2 = \mu \\ q \in \mathcal{R}(w)}} \|P_j q\|_2^2 \end{aligned} \quad 6.3.1.3$$

which are immediately seen to be the minimization of length QB and the maximization of length OB subject to norm, linear and dominance constraints.

To return to the minimization of length $Q_1 D$, as given in 6.3.1.2, this may be written in terms of the original problem as :

$$\text{Min } \left\{ T_j(q) = \sum_{\substack{i=1 \\ i \neq j}}^m |q_{ij}(iw)|^2 \right\}$$

$$\text{subject to } |q_{jj}(iw)| = \mu \quad \text{and} \quad q(iw) = G(iw)k_{\cdot j}$$

which clearly corresponds to the minimization of the off-diagonal terms of the j th column subject to a constraint preventing the modulus of the j th diagonal element from disappearing, and a linear system constraint. This is precisely the modified problem proposed by Rosenbrock (1975) to overcome the deficiencies of the Hawkins (1972) formulation. The solution to this constrained minimization problem is obtained as that of an equivalent eigenvalue problem of the form $Ax = \lambda Bx$, where the eigenvector corresponding to the smallest non-negative eigenvalue is required. The related optimization problems of 6.3.1.3 can also be reduced to similar eigenvalue/vector problems.

The relationship between minimizing a function of angle θ (Fig. 6.3.1.2) and maximizing $J_{D_j}(q)$ both subject to the norm and linear constraints is very difficult to relate analytically even though intuitively it is thought that minimizing $\tan \theta$ (or some similar function for example $T_j(q)$) encourages $q \in Q_{D_j}$ and further increases $J_{D_j}(q)$, that is encourages column dominance at $s = iw_0$. The following theorem gives a sufficient condition for iterates in one of the above optimization procedures (6.3.1.1, 6.3.1.2 or 6.3.1.3) to give improved column dominance.

Theorem 6.3.1.1 Let q_1 , and $q_2 \in Q_{D_j}$ where $\|q_1\|_2 = \|q_2\|_2 = \mu$. Suppose $\tan \theta(q_1) < \tan \theta(q_2)$ where q_1 and $q_2 \in \mathbb{R}^{2m}$.

If for the j th column
$$a_{1j}^2 \geq \frac{a_{2j}^2}{(m-1)^{1/2}} + \mu^2 \left(1 - \frac{1}{(m-1)^{1/2}} \right)$$

then $J_{D_j}(q_1) > J_{D_j}(q_2)$; where $m > 1$

and $a_{ki} \triangleq (q_k^2(i) + q_k^2(i+m))^{1/2}$ $k = 1, 2, \dots$; and $i = 1, 2, \dots, m$.

Proof (See Appendix D).

In terms of the original optimization procedures, suppose the initial output $q_o \in Q_{D_j}$, where $a_{oj}^2 = q_o^2(j) + q_o^2(j+m) = \frac{1}{2}\mu^2 + \epsilon$

then if subsequent iterates satisfy

$$a_{kj}^2 \geq \mu^2(1 - 0.5(m-1)^{-k/2}) + \epsilon (m-1)^{-k/2} \quad k = 1, 2, 3, \dots$$

the dominance cost is successively increased. For example

$$m = 4, \quad a_{oj}^2 = \frac{1}{2}\mu^2 + \epsilon, \quad a_{1j}^2 \geq .71 \mu^2 + 0.58 \epsilon,$$

$$a_{2j}^2 \geq 0.833 \mu^3 + \epsilon/3 \text{ and so on.}$$

The problem description specifies $q \in \mathcal{R}(W)$ so that the theorem will now apply to

iterates q_k produced in the range space of the operator W .

Hence sufficient conditions have been exhibited to show that

minimizing $\tan \theta(q)$ (or its equivalent) can improve the

dominance cost $J_{D_j}(q)$, however it is expected that the

process will terminate at some sub-optimal cost value.

The geometrical situation of the modified Rosenbrock formulation

is that of minimizing the length of a vector between the

two subspaces $\mathcal{R}(W)$, and $\text{span} \{e_j, e_{j+m}\}$. In particular the

vector is assumed to be orthogonal to the latter. An alternative

formulation is to assume the required vector is orthogonal

to the subspace $\mathcal{R}(W)$, a situation which is now pursued.

Set $q_D = \alpha_j e_j + \alpha_{j+m} e_{j+m}$, where $\|q_D\|_2 = \mu$ (shown by the vector $\vec{OD}$) and pose the minimization problem $\text{Min}_{q = Wk_{\cdot j}} \|q - q_D\|_2^2$

which is the minimization of the length of $Q_2 D$ for a given q_D .

Considering all possible q_D yields a least squares minimization problem

$$\text{as } \text{Min}_{q = Wk_{\cdot j}} \|q - q_D\|_2^2 \quad 6.5.1.4.$$

$$q_D = \alpha_j e_j + \alpha_{j+m} e_{j+m}, \quad \|q_D\|_2 = \mu$$

Perhaps the most interesting feature of this problem is its essential equivalence to the modified Rosenbrock problem 6.3.1.3(2) and the simplified computational problem which it yields.

Theorem 6.3.1.2 The solution of Rosenbrock's modified problem (1975)

$$\text{namely } \text{Min}_{\|P_j q\|_2 = \mu, q \in \mathcal{R}(w) \cap Q_{D_j}} \|(I - P_j)q\|_2^2 \quad \text{lies on the same ray as the solution}$$

of the least squares formulation

$$\text{Min}_{q_D = \alpha_j e_j + \alpha_{j+m} e_{j+m}, \|q_D\|_2 = \mu, q \in \mathcal{R}(w) \cap Q_{D_j}} \|q - q_D\|_2^2$$

Proof First some geometrical concepts which motivate the result are

given : (i) If $\hat{q}_R \in \mathbb{R}^{2m}$ solves Rosenbrock's modified problem

$$\text{then } \langle P_j \hat{q}_R, (I - P_j) \hat{q}_R \rangle = 0 \quad \text{where } P_j \hat{q}_R = \text{span}(e_j, e_{j+m})$$

and (ii) If $\hat{q}_L \in \mathbb{R}^{2m}$ solves the least squares formulation then

$$\langle \hat{q}_L, \hat{q}_L - q_D \rangle = 0 \quad \text{where } \hat{q}_L = W(W^* W)^{-1} W^* q_D. \quad \text{These orthogonality}$$

conditions form the basis of the intuitive geometric setting in

Figure 6.3.1.2.

The remainder of the proof is in two parts under the assumption that

a dominant output $q = Wk_{\cdot j}$ exists. (That is $\mathcal{Q}(W) \cap \mathcal{Q}_{D_j} \neq \emptyset$).

(a) Adjoin the constraints in Rosenbrock's modified formulation to obtain $J_R = ||(I-P_j)q||_2^2 + \lambda(\mu^2 - ||P_j q||_2^2)$. Substitute for the

linear constraint $q = Wk_{\cdot j}$:

$$J_R(k_{\cdot j}, \lambda) = \langle k_{\cdot j}, W^*(I-P_j)Wk_{\cdot j} \rangle + \lambda(\mu^2 - \langle k_{\cdot j}, W^*P_jWk_{\cdot j} \rangle)$$

$$\frac{1}{2} \frac{\partial J_R}{\partial k_{\cdot j}} = W^*(I-P_j)Wk_{\cdot j} - \lambda W^*P_jWk_{\cdot j}$$

$$\frac{\partial J_R}{\partial \lambda} = \mu^2 - \langle k_{\cdot j}, W^*P_jWk_{\cdot j} \rangle$$

The condition of optimality gives $\hat{J}_k = \lambda\mu^2 \geq 0$, seek therefore the smallest non-negative eigenvalue, $\lambda_{\min}$, satisfying

$$W^*(I-P_j)\hat{Wk}_{\cdot j} = \lambda_{\min} W^*P_j\hat{Wk}_{\cdot j} \quad 6.3.1.5$$

from which the corresponding normalized eigenvector is the required pre-compensator column.

The eigenvalue/vector relationship may be rewritten as:

$$W^*\hat{Wk}_{\cdot j} = (1 + \lambda_{\min}) W^*P_j\hat{Wk}_{\cdot j} \quad 6.3.1.6$$

or assuming $W^*W > 0$ as $\hat{k}_{\cdot j} = (1+\lambda_{\min}) (W^*W)^{-1} W^*P_j\hat{Wk}_{\cdot j}$

and using $\hat{q}_R = \hat{Wk}_{\cdot j}$ as $\hat{q}_R = (1+\lambda_{\min}) W(W^*W)^{-1} W^*P_j\hat{q}_R$.

(b) Recall the least squares formulation as :

$$\text{Min } ||q - q_D||_2^2 \quad \text{subject to } q = Wk_{\cdot j}, \quad q_D = \alpha_j e_j + \alpha_{j+m} e_{j+m}$$

and $||q_D||_2 = \mu$. Write $E_j = (e_j, e_{j+m})$ then $q_D = E_j \alpha$, $\alpha \in \mathbb{R}^2$.

Adjoin the constraints to obtain :

$$J_L = ||q - q_D||_2^2 + \xi(\mu^2 - \langle \alpha, E_j^* E_j \alpha \rangle)$$

$$= \langle q - E_j \alpha, q - E_j \alpha \rangle + \xi(\mu^2 - \langle \alpha, \alpha \rangle)$$

Substitute for the linear constraint $q = Wk_{\cdot j}$, then

$$J_L(k_{\cdot j}, \alpha, \xi) = \langle k_{\cdot j}, W^* W k_{\cdot j} \rangle - 2 \langle k_{\cdot j}, W^* E_j \alpha \rangle + (1-\xi) \langle \alpha, \alpha \rangle + \xi \mu^2$$

$$\text{Giving } \frac{1}{2} \frac{\partial J_L}{\partial k_{\cdot j}} = W^* W k_{\cdot j} - W^* E_j \alpha$$

$$\frac{1}{2} \frac{\partial J_L}{\partial \alpha} = -E_j^* W k_{\cdot j} + (1-\xi) \alpha$$

$$\frac{\partial J_L}{\partial \xi} = \mu^2 - \langle \alpha, \alpha \rangle$$

Under the condition of optimality $\nabla J_L = 0$ obtain

$$(i) \quad W^* \hat{W} k_{\cdot j} = W^* E_j \hat{\alpha}, \quad (ii) \quad E_j^* \hat{W} k_{\cdot j} = (1-\xi) \hat{\alpha} \quad \text{and} \quad (iii) \quad \mu^2 = \langle \alpha, \alpha \rangle$$

with the optimal cost as $\hat{J}_L = \xi \mu^2 \geq 0$.

Clearly (i) and (ii) together give

$$W^* W k_{\cdot j} = (1-\xi)^{-1} W^* E_j E_j^* W k_{\cdot j} \quad (E_j E_j^* = P_j)$$

The minimum optimal cost is attained for the smallest non-negative number ξ , denoted $\xi_{\min}$. Assuming that $W^* W > 0$ implies $\xi_{\min}$

satisfies $(1-\xi_{\min}) \hat{\alpha} = E_j^* W (W^* W)^{-1} W^* E_j \hat{\alpha}$. Thus for the least squares formulation $\hat{k}_{\cdot j}$ satisfies

$$W^* \hat{W} k_{\cdot j} = (1-\xi_{\min})^{-1} W^* P_j W k_{\cdot j} \quad 6.3.1.7$$

A comparison of 6.3.1.6, and 6.3.1.7 shows the pre-compensator column $k_{\cdot j}$ to satisfy

$$W^* \hat{W} k_{\cdot j} = \bar{\lambda} W^* P_j W k_{\cdot j}$$

where in both cases $\bar{\lambda}$ is the smallest eigenvalue satisfying the inequality $\bar{\lambda} \geq 1$. Since $k_{\cdot j}$ is an eigenvector it can be scaled

to satisfy the constraints of either problem, and also gives

$$\hat{q}_L = \eta \hat{q}_R \quad \text{for some scalar, } \eta.$$

4

The details of the proof also indicate the technique for the solution of the least squares problem. The relevant equations are (assuming $W^*W > 0$) :

$$\hat{k}_{\cdot j} = (W^*W)^{-1} W^* q_D, \quad q_D = E_j \alpha; \quad \alpha \in \mathbb{R}^2 \quad \text{and}$$

$$(1 - \xi_{\min}) \hat{\alpha} = E_j^* W (W^*W)^{-1} W^* E_j \hat{\alpha}. \quad \text{The latter can be rewritten as}$$

$$E_j^* (I - W(W^*W)^{-1} W^*) E_j \hat{\alpha} = \xi_{\min} \hat{\alpha}, \quad \text{which is a trivial } (2 \times 2)$$

eigenvalue/vector problem. Consequently the main computing burden lies in the calculation of the inverse of W^*W , the remaining operations being rather trivial. This compares favourably with the computing load associated with the eigenvalue problem of the modified Rosenbrock problem.

The problem readily admits the incorporation of precompensator gain costs as a quadratic term in $k_{\cdot j}$, vis:

$$\text{Min } J_Q(q, \alpha_1, \alpha_2) = \frac{1}{2} \langle q - q_D, q - q_D \rangle_{E_{2m}} + \frac{1}{2} \langle k_{\cdot j}, R k_{\cdot j} \rangle_{E_m}$$

$$q, \alpha_1, \alpha_2 \quad q_D = \alpha_1 e_j + \alpha_2 e_{j+m}$$

$$\alpha_1^2 + \alpha_2^2 = \mu^2 \quad q \in \mathbb{R}(W)$$

The cost matrix R can be given the interpretation "control"

cost and the normalizing parameter μ scales the solution to the maximal achievable dominance at frequency $s = iw$. However the incorporation of a gain cost will require the verification $\hat{q} \in Q_{D_j}$. Clearly in this form the structures procedure of Chapter 3 can be applied to find gain structures with almost optimal dominance.

In transferring from pseudo-diagonalisation to diagonal dominance proper, the important features of the above formulation are its frequency dependence and the simplicity of the eigenvalue/vector calculation. These enable, through a suitable choice for $\mu(w)$, a desired dominance trajectory $q_D(w)$ to be found which becomes the basis of a least squares approach for choosing $k_{\bullet j}$.

6.3.2 On Diagonal Dominance

The precompensator is chosen to be a constant matrix. The technique of pseudo-diagonalisation as described above enables the constant precompensator to be calculated at a specific frequency by using the following optimization problem:

At the frequency $s = iw$ the desired dominance is

$$q_D(w) \in Q_{D_j}, \quad q_D(w) = \alpha_j(w)e_j + \alpha_{j+m}(w)e_{j+m} \text{ with } \alpha_j^2(w) + \alpha_{j+m}^2(w) = \mu^2(w)$$

where procedures for the selection of $\mu(w)$ are given below. Choose

$$q(w) = W(w)k_{\bullet j}(w) \text{ to minimize}$$

$$J(w) = \frac{1}{2} \| q_D(w) - q(w) \|^2, \quad q_D(w) - q(w) \in E_{2m}$$

$$\text{where } W(w) : \mathbb{R}^m \rightarrow \mathbb{R}^{2m} \text{ and } k_{\bullet j} \in \mathbb{R}^m \quad j = 1, 2, \dots, m.$$

For a sequence of frequencies representing the frequency range of interest, the solution of this problem is straightforward since only one (m, m) inversion and m minimum eigenvalue/vector pairs for a 2×2 problem are required at each frequency point giving a desired dominance trajectory $q_D(w)$ and an optimal output $q(w)$ obtained by the optimal gain $k_{\bullet j}(w)$. It is this desired dominance disturbance which forms the basis of the problem for dominance.

Restrict the precompensator gain to be constant over the frequency range $(-\infty, \infty)$ and write $e(w) = q_D(w) - W(w)k_{\bullet j}$. To seek to minimize an integral quadratic cost of error squared requires some restrictions to ensure the existence of such integrals defined on the infinite domain $-\infty < w < \infty$. It is sufficient to demand i) $\|e(w)\|_2$ is bounded everywhere and ii) $\lim_{w \rightarrow \pm\infty} \|e(w)\|_2 = 0$. The conditions on the error function

$$w \rightarrow \pm\infty$$

produce restrictions on the plant $G(s)$ and the disturbance $q_D(s)$. Noting an equivalence between $\|e(w)\|_2$, $e \in \mathbb{R}^{2m}$ and $\|e(iw)\|_2$, $e \in \mathbb{C}^m$ use the form $e(iw) = q_D(iw) - G(iw)k_{\bullet j}$. Hence $\|e(iw)\|_2 \leq \|q_D(iw)\|_2 + \|G(iw)\|_2 \|k_{\bullet j}\|_2$

Then if i) $\|q_D(iw)\|_2$ and $\|G(iw)\|_2$ are bounded everywhere

$$\Rightarrow \|e(iw)\|_2 < \infty \Rightarrow \|e(w)\|_2 < \infty, \quad -\infty < w < \infty.$$

$$\text{and ii) } \lim_{w \rightarrow \pm\infty} \|q(iw)\|_2 = 0 \quad \text{and} \quad \lim_{w \rightarrow \pm\infty} \|G(iw)\|_2 = 0$$

$$\Rightarrow \lim_{w \rightarrow \pm\infty} \|e(iw)\|_2 = \lim_{w \rightarrow \pm\infty} \|e(w)\|_2 = 0$$

These conditions are sufficient to ensure $\|e(w)\|_2$ behaves as required. Those on the plant are, therefore

$$\text{i) } \|G(iw)\|_2 < \infty \quad \text{and} \quad \text{ii) } \lim_{w \rightarrow \pm\infty} \|G(iw)\|_2 = 0, \quad \text{and these } w \in (-\infty, \infty)$$

are only satisfied by a plant whose transfer has elements of the type

$$g_{ij}(s) = \frac{a_0 s^m + a_1 s^{m-1} + \dots + a_m}{s^n + a_1 s^{n-1} + \dots + b_n} \quad n > m \quad \text{and} \quad b_n \neq 0.$$

Such a plant also has the property $\operatorname{Re}(G(-iw)) = \operatorname{Re}(G(iw))$ and

$$\int_m(G(-iw)) = -\int_m(G(iw)) \quad \text{so that} \quad W(-w) = \begin{bmatrix} W_1(w) \\ -W_2(w) \end{bmatrix}$$

The disturbance $q_D(iw)$ has to satisfy i) $\|q_D(iw)\|_2 < \infty \forall w$ and ii) $\lim_{w \rightarrow i\infty} \|q(iw)\|_2$ and to be physically meaningful

if $q_D(w) = \begin{bmatrix} q_1(w) \\ q_2(w) \end{bmatrix}$ then set $q_D(-w) = \begin{bmatrix} q_1(w) \\ -q_2(w) \end{bmatrix}$. Since

$\mu(w) = \|q_D(iw)\|_2$, the conditions (i) and (ii) are properties on $\mu(w)$, whereas the latter property can always be obtained from the eigenvector nature of $q_D = \alpha_j e_j + \alpha_{j+m} e_{j+m}$.

The parameter $\mu(w)$ has already been represented as the maximum dominance obtainable at frequency w . Suppose, therefore such dominance could be achieved by $q_D(iw) = G(iw)k_{\bullet j}$, then

$$\begin{aligned} \mu(w) = |q_{Djj}(iw)| &= \left| \sum_{\ell=1}^m G_{j\ell}(iw) k_{\ell j} \right| \\ &\leq \sum_{\ell=1}^m |G_{j\ell}(iw)| |k_{\ell j}| \\ &\leq \max_{\ell} |G_{j\ell}(iw)| \sum_{\ell=1}^m |k_{\ell j}| \end{aligned}$$

since any $k_{\bullet j}$ producing maximum dominance may be normalized so that $\|k_{\bullet j}\|_1 = 1$ suggesting the constraint

$$\mu(w) \leq \max_{\ell} |G_{j\ell}(iw)| \quad -\infty < w < \infty$$

Choose as maximum achievable dominance $\mu(w) = \max_{\ell} |G_{j\ell}(iw)|$

on $w \in (-\infty, \infty)$. Although this measure is simple computational and appealing conceptually it does not preclude other choices for example

$$\mu(w) = \sum_{\ell=1}^m |G_{j\ell}(iw)| \quad \text{which is obtained as follows}$$

$$\mu(w) = |q_{Djj}(iw)| \leq \sum_{\ell=1}^m |G_{j\ell}(iw)| |k_{\ell j}|$$

$$\leq \sum_{\ell=1}^m |G_{j\ell}(iw)| \max_{\ell} |k_{\ell j}|.$$

Either of these formulae for $\mu(w)$, $-\infty < w < \infty$ produce a disturbance whose general properties depend on those of the system hence under the restrictions already placed on the system $q_D(w)$ will satisfy

$$\|q_D(w)\|_2 < \infty, \quad -\infty < w < \infty \quad \text{and} \quad \lim_{w \rightarrow \pm\infty} \|q_D(w)\|_2 = 0.$$

The consistency property, that is if $q_D(w) = \begin{bmatrix} q_1(w) \\ q_2(w) \end{bmatrix}$

then $q_D(-w) = \begin{bmatrix} q_1(w) \\ -q_2(w) \end{bmatrix}$ is shown to be obtained from the eigenvector

nature of $q_D(w)$. At each frequency w , $q_D(w)$ is related to a vector $v \in \mathbb{R}^2$ where $q_D(w) = v_1 e_j + v_2 e_{j+m}$ and v is an eigenvector of the matrix $A(w) : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ where

$$A(w)_{11} = [W(w) [W^T(w)W(w)]^{-1} W^T(w)]_{jj}$$

$$A(w)_{22} = [W(w) [W^T(w)W(w)]^{-1} W^T(w)]_{j+m,j+m} \quad \text{and}$$

$$A(w)_{12} = A(w)_{21} = [W(w) [W^T(w)W(w)]^{-1} W^T(w)]_{j,j+m}. \quad \text{Under the}$$

conditions already imposed on the system, have that

$$\text{if } W(w) = \begin{bmatrix} \alpha(w) \\ \beta(w) \end{bmatrix}, \text{ then } W(-w) = \begin{bmatrix} \alpha(w) \\ -\beta(w) \end{bmatrix}$$

$$\text{Hence } W^T(-w)W(w) = W^T(w)W(w) = \Phi(w) \quad \text{and}$$

$$W(-w) [W^T(-w)W(-w)]^{-1} W^T(-w) = \begin{bmatrix} \alpha(w) \Phi^{-1}(w) \beta(w) & -\alpha(w) \Phi^{-1}(w) \beta^T(w) \\ -\beta(w) \Phi^{-1}(w) \alpha^T(w) & \beta(w) \Phi^{-1}(w) \beta^T(w) \end{bmatrix}$$

$$\text{giving } A(-w)_{11} = A(w)_{11} \quad A(-w)_{22} = A(w)_{22}$$

$$\text{and } A(-w)_{12} = A(-w)_{21} = -A(w)_{12} = -A(w)_{21}$$

Consequently if $A(w)v(w) = \lambda(w)v(w)$ then

$$A(-w) \begin{bmatrix} v_1(w) \\ -v_2(w) \end{bmatrix} = \lambda(w) \begin{bmatrix} v_1(w) \\ -v_2(w) \end{bmatrix} \quad \text{giving the correspondence}$$

$$v(-w) = \begin{bmatrix} v_1(w) \\ -v_2(w) \end{bmatrix} \quad \text{and} \quad \lambda(-w) = \lambda(w)$$

$$\text{Hence the consistency property } q_D(-w) = \begin{bmatrix} q_1(w) \\ -q_2(w) \end{bmatrix} \text{ follows as required.}$$

Assuming that a desired dominance trajectory $q_D(w)$ is acquired having the above properties then an infinite frequency domain optimization problem is motivated by the desire to find a frequency independent precompensator gain. For the j th column have

$$e(w) = q_D(w) - W(w)k_{\bullet j} = \begin{bmatrix} q_1(w) \\ q_2(w) \end{bmatrix} - \begin{bmatrix} \alpha(w) \\ \beta(w) \end{bmatrix} k_{\bullet j} = \begin{bmatrix} e_1(w) \\ e_2(w) \end{bmatrix}$$

$$\text{then } e(-w) = \begin{bmatrix} q_1(w) \\ -q_2(w) \end{bmatrix} - \begin{bmatrix} \alpha(w) \\ -\beta(w) \end{bmatrix} k_{\bullet j} = \begin{bmatrix} e_1(w) \\ -e_2(w) \end{bmatrix}$$

$$\text{so that } \langle e(w), e(w) \rangle_{E_{2m}} = \langle e(-w), \bar{Q}e(w) \rangle_{E_{2m}}, \quad \bar{Q} = \begin{bmatrix} I_m & \\ & -I_m \end{bmatrix}$$

A least squares optimization is posed :

$$\begin{array}{l} \text{Min} \quad J = \frac{1}{2} \int_{-\infty}^{\infty} \langle e(-w), \bar{Q}e(w) \rangle_{E_{2m}} dw \\ \text{w.r.t. } k_{\bullet j} \end{array}$$

$$\text{such that } e(w) = q_D(w) - W(w)k_{\bullet j}$$

Such a problem formulation cannot guarantee dominance, since there are systems for which no constant precompensator can create dominance. It has however been posed to maximize dominance if it exists. From the problem as it stands it is possible to develop a computational frequency domain version and a complex frequency domain version. This latter problem has the virtue that it is possible to apply Parseval's Theorem to obtain a time-domain interpretation.

Utilizing the various system symmetries a computational frequency domain version is obtained as

$$\text{Min } J(k_{\cdot j}) = \langle k_{\cdot j}, A k_{\cdot j} \rangle_{E_m} - 2 \langle b, k_{\cdot j} \rangle_{E_m} + c$$

w.r.t. $k_{\cdot j}$

$$\text{where } A = \int_0^\infty W^T(w) W(w) dw \in \mathbb{R}^{m \times m}$$

$$b = \int_0^\infty W^T(w) q_D(w) dw \in \mathbb{R}^m \text{ and } c = \int_0^\infty \langle q_D(w), q_D(w) \rangle dw$$

Clearly the optimal pre-compensator column is $\hat{k}_{\cdot j} = A^{-1}b$.

Observe that the matrix A requires inverting only once since the same matrix occurs in each of the $j, j = 1, \dots, m$ problems. If

the matrix $W(w)$ is retained in sampled form from the pseudo-diagonalisation it is relatively straight-forward to calculate

A. The vector $b = \int_0^\infty W^T(w) q_D(w) dw$ will differ for each of the j subproblems since $q_D(w)$ is dependent on the particular $k_{\cdot j}$ column of the pre-compensator being obtained. The fact that

$q_D(w) = \alpha_j(w) e_j + \alpha_{j+m}(w) e_{j+m}$ can be exploited to give b as

$$b = \int_0^\infty (\alpha_j(w) W_{\cdot j}^T(w) + \alpha_{j+m}(w) W_{\cdot j+m}^T(w)) dw$$

Again utilizing various system symmetries an equivalent complex frequency domain formulation may be obtained as:

$$\text{Min } J = \frac{1}{2} \langle k_{\cdot j}, \bar{A} k_{\cdot j} \rangle_{E_m} - \langle \bar{b}, k_{\cdot j} \rangle_{E_m} + \frac{1}{2} \bar{c}$$

w.r.t. $k_{\cdot j} \in \mathbb{R}^m$

$$\text{where } \bar{A} = \int_{-\infty}^\infty G^T(-iw) G(iw) dw, \bar{b} = \int_{-\infty}^\infty G^T(-iw) q_D(iw) dw$$

and $\bar{c} = \int_{-\infty}^\infty \langle q_D(-iw), q_D(iw) \rangle_{E_m} dw$, for which the optimal

solution is $k_{\cdot j} = [\bar{A}]^{-1} \bar{b}$. The restrictions already placed on

the plant enable Parseval's Theorem to be applied to obtain:

$$\bar{A} = \int_{-\infty}^{\infty} G^T(-iw) G(iw) dw = 2\pi \int_0^{\infty} \Phi^T(t) \Phi(t) dt$$

$$\bar{b} = \int_{-\infty}^{\infty} G^T(-iw) q_D(iw) dw = 2\pi \int_0^{\infty} \Phi^T(t) q_D(t) dt$$

$$\text{and } \bar{c} = \int_{-\infty}^{\infty} \langle q_D(-iw), q_D(iw) \rangle_D dw = 2\pi \int_0^{\infty} \langle q_D(t), q_D(t) \rangle_{E_m} dt$$

where $\Phi(t)$ is the system impulse response defined on $t \in [0, \infty)$ and $q_D(t)$ is the time function associated with frequency response $q_D(jw)$. These results illuminate the naive examples given by Hannah (1975) and provide the correct function of the impulse response matrix to use. Further they indicate the need for a correct disturbance, q_D . At present, this disturbance is found from the pseudo-diagonalisation procedure as described in the previous section, and in practise has shown to be highly successful. The above link appears to indicate that a priori disturbance design in the time domain would be of great practical utility.

The assessment of precompensator structures can be introduced by the device of a compensator cost in the formulation of the least squares optimization problem, viz

$$\text{Min } J = \frac{1}{2} \int_{-\infty}^{\infty} \langle e(-w), \bar{Q} e(w) \rangle_{E_{2m}} dw + \langle k_{\bullet j}, R k_{\bullet j} \rangle_{E_m}$$

w.r.t. $k_{\bullet j}$

$$\text{subject to } e(w) = q_D(w) - W(w) k_{\bullet j}$$

from which a computational frequency domain form may be

$$\text{developed as } \text{Min } J = \langle k_{\bullet j}, (A + R) k_{\bullet j} \rangle_{E_m} - 2 \langle b, k_{\bullet j} \rangle_{E_m} + c$$

w.r.t. $k_{\bullet j}$

$$\text{with } A + R = \int_0^{\infty} W^T(w) W(w) dw + R, \quad b = \int_0^{\infty} W^T(w) q_D(w) dw$$

$$\text{and } c = \int_0^{\infty} \langle q_D(w), q_D(w) \rangle_{E_m} dw. \quad \text{Perform a Cholesky}$$

decomposition for A as $A = L^T L$, then the effect of deleting a particular element of $k_{\cdot j}$ can be obtained by deleting a column of L . Further write $b = L^T z$ where $z = (L^T)^{-1} b$, then removal of an element of b corresponds to the deletion of a column of L . Consequently may write

$$\begin{aligned} \text{Min } J &= \langle e, e \rangle_{E_m} + \langle k_{\cdot j}, R k_{\cdot j} \rangle_{E_m} + \bar{c} \\ \text{w.r.t. } k_{\cdot j} \end{aligned}$$

where $e = L k_{\cdot j} - z$ $z = (L^T)^{-1} b$ and $L^T L = A$
which is a suitable form for structures assessment.

The overall algorithm has three parts, the first is to produce the desired dominance trajectory using the pseudo-diagonalisation technique, the second is to find the complete least squares precompensator. At this point the normalized dominance curves $J_{Dj} = \sum_{\ell=1}^m |q_{\ell j}(i\omega)| / |q_{jj}(i\omega)|$ should be examined to see if the progression to structures, the third and final part, is sensible. This latter check must be made in case a constant precompensator only marginally creates dominance.

An example to illustrate the algorithm is that of an industrial boiler taken from Hawkins, 1972.

$$G(s) = \begin{bmatrix} 1.0/(1+4s) & 0.7/(1+5s) & 0.3/(1+5s) & 0.2/(1+5s) \\ 0.6/(1+5s) & 1.0/(1+4s) & 0.4/(1+5s) & 0.35/(1+5s) \\ 0.35/(1+5s) & 0.4/(1+5s) & 1.0/(1+4s) & 0.35/(1+5s) \\ 0.2/(1+5s) & 0.3/(1+5s) & 0.7/(1+5s) & 0.6/(1+4s) \end{bmatrix}$$

In producing column dominance, it can be seen that the plant has a symmetry which implies columns 1 and 4, and columns 2 and 3 may be paired. The desired dominance trajectories obtained from the pseudo-diagonalisation procedure, and the dominance achieved by a constant precompensator are displayed in Figure 6.3.2.1. The dominance achieved is so good that structure assessment may prove profitable.

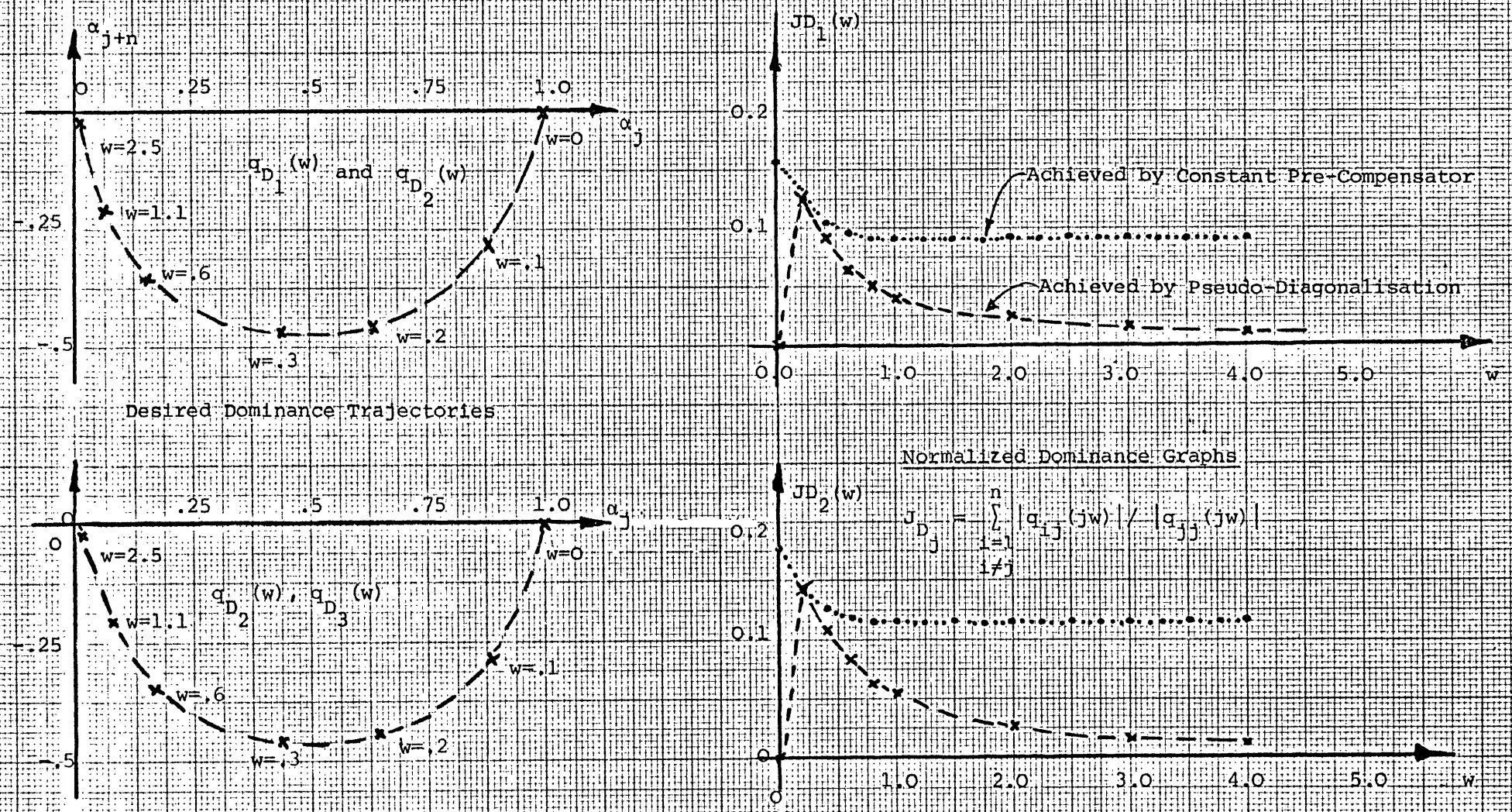


Figure 6.3.2.1 Desired Dominance Trajectories and the
Dominance Achieved for Hawkin's Example.

6.4 Least Squares Parameter Estimation, and an Application of the Structures Procedure

In the introductory parts of this thesis, the main features of the concept of structures were explained using the problem of least squares parameter estimation. Here the least squares analysis is briefly restated, and the details of an application of the structures procedure to an example taken from Draper and Smith, 1966 presented.

The problem arises of fitting a linear combination of n (independent) variables to a set of N given observations. The linear model is usually written

$$y(i) = \sum_{j=1}^n \beta_j x_j(i) + e(i) \quad : \quad i=1, 2, \dots, N$$

where $\{y(i)\}_{i=1}^N$ are the given observations

$\{x_j(i)\}_{i=1}^N$ are the independent variables

$\{\beta_j\}_{j=1}^n$ and $\{e(i)\}_{i=1}^N$ are the regression coefficients, and residual errors respectively.

In matrix form: $y = W \beta + e$, where y and $e \in \mathbb{R}^N$

$\beta \in \mathbb{R}^n$ and $W: \mathbb{R}^n \rightarrow \mathbb{R}^N$. The rows of matrix W , comprise of the independent variables at the i th instance. The optimal set of regression coefficients is chosen to minimize the quadratic cost function $J = \frac{1}{2} \langle e, Q e \rangle_{E_N}$, where Q is some positive-definite matrix. The complete problem statement is

Min $J = \frac{1}{2} \langle e, Q e \rangle_{E_N}$
subject to $e = y - W\beta \in \mathbb{R}^N$

The optimal solution (assuming that $W^T Q W > 0$) is given as

$\hat{\beta} = (W^T Q W)^{-1} W^T Q y$, for which the optimal cost is

$$\hat{J} = \frac{1}{2} \langle y, Q y \rangle_{E_N} - \frac{1}{2} \langle W^T Q y, (W^T Q W)^{-1} W^T Q y \rangle_{E_n}$$

The associated structures problem is to determine which of the β_j can be zeroed without appreciable degradation in the optimal cost. Each structure may be written $\beta_s^T = (\beta_{i_1}, \beta_{i_2}, \dots, \beta_{i_s})$, $i_s \leq n$ and the related linear equation as $y = W_s \beta_s + e$, where the matrix $W_s: \mathbb{R}^s \rightarrow \mathbb{R}^N$ is obtained from matrix W by deleting those columns associated with the β_j 's constrained to zero. The least squares structures problem is

$$\begin{aligned} \text{Min } J_s &= \frac{1}{2} \langle e, Q e \rangle_{E_N} \\ \text{subject to } y &= W_s \beta_s + e \in \mathbb{R}^N. \end{aligned}$$

The computation required to assess all $(2^n - 1)$ regression structures by direct evaluation is clearly prohibitive for all but small n . The structures diagnostic is shown by example to offer an alternative method of assessing the regression structures.

This is achieved by the addition to the cost function of a quadratic term costing the regression coefficients, and the use of an asymptotic parameter ρ , to give a modified cost function $J_M = \frac{1}{2} \langle e, Q e \rangle_{E_N} + \rho/2 \langle \beta, R_0 \beta \rangle_{E_n}$. The term costing regression coefficients, namely

$\langle \beta, R_0 \beta \rangle$ is inserted to give the least squares problem the same form as that used in the structures analysis of the earlier chapters, and the parameter ρ is used to ensure that this extra term does not upset the cost contours of the least squares problem unduly. The parameter ρ will clearly be utilized in an asymptotic sense, so that as $\rho \rightarrow 0$, $J_M(\beta, \rho)$ replicates the behaviour of the least squares cost $J(\beta)$.

The interpretations of using an asymptotic parameter in the structure diagnostic has been fully discussed in second chapter

of this thesis. In particular, the descriptions and theorems using the linear space setting will be directly relevant. The effect of the asymptotic parameter ρ , on the bound calculation is also well documented in Chapter 3. It will be recalled from that particular chapter, that as $\rho \rightarrow 0$ the lower bounds limit at a set of nested values whilst the upper bounds collapse into the value $m_0 = \frac{1}{2} \langle y, Qy \rangle$. Consequently, the information regarding structures resides in the lower bound, and the full optimal cost (which is assumed to be known).

The application of the structures procedure to the example from Draper and Smith (1966) is primarily made to demonstrate the features described above.

Table 6.4.1 Data: Draper and Smith (1966) Example

i	y(i)	$x_1(i)$	$x_2(i)$	$x_3(i)$	$x_4(i)$
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

Error Weight
 Q^{-1}_{13}

Table 6.4.2 Details of Complete/Structure Least Squares
Solutions ($\rho = 0$)

$$m_o = 60544.045 \quad \Delta J_o^f = 60520.113$$

$$k^{AT} \text{ (full set)} = (62.405, 1.551, 0.51, 0.102, -0.144)$$

Ranking Structures with Three Parameters Using ΔJ_o^s ($\rho = 0$)

Rank	Parameter Values			ΔJ_o (Structure)
1	$\beta_o = 52.577$	$\beta_1 = 1.468$	$\beta_2 = 0.662$	60515.093
2	$\beta_o = 103.097$	$\beta_1 = 1.44$	$\beta_4 = -0.614$	60506.664
3	$\beta_o = 131.282$	$\beta_3 = -1.2$	$\beta_4 = -0.725$	60456.176
4	$\beta_1 = 1.528$	$\beta_2 = 1.349$	$\beta_4 = 0.631$	60451.654

From the above table of structures with three parameters it can be seen that the optimal cost improvement ΔJ_o^s differs from the full optimal value ΔJ_o^f by at most .113%. Clearly very sensitive bounds will be required to isolate and order these top three parameter structures.

Assessment 1 Assuming a blanket coefficient cost of $R=I_5$. The full modified problem was solved for a range of $\rho > 0$ values to establish a suitable approximation.

Table 6.4.3 Optimal Costs and Parameters for the Modified
Cost where $Q = I_{13}$, and $R_o = I_5$ ($0 \leq \rho \leq 0.1$)

ρ	β_o	β_1	β_2	β_3	β_4	$\Delta J_o^f(\rho)$
10^{-1}	0.798	2.183	1.145	0.749	0.478	60517.425
10^{-3}	34.291	1.84	0.8	0.394	0.139	60519.042
10^{-5}	61.897	1.556	0.515	0.107	-0.139	60520.094
0	62.405	1.551	0.51	0.102	-0.144	60520.113

Noting that there is fair agreement between the elements of the optimal parameter vector β ($\rho = 10^{-5}$) and those of the full least squares parameter vector β ($\rho = 0$), and that the difference in the optimal cost improvements is minute, this value of ρ was chosen for the parameter structure assessment.

Table 6.4.4 Three Parameter Solutions obtained for $J_M(\rho = 10^{-5})$ Ranked by $\Delta J_O^S(\rho)$

Rank	Parameter Values	Lower Bound	$\Delta J_O^S(\rho)$
1	$\beta_0 = 52.577, \beta_1 = 1.468, \beta_2 = 0.662$	58995.12	60515.08
2	$\beta_0 = 103.097, \beta_1 = 1.44, \beta_4 = -0.614$	51791.48	60506.61
3	$\beta_0 = 131.282, \beta_3 = -1.2, \beta_4 = -0.724$	46430.10	60456.09
4	$\beta_1 = 1.528, \beta_2 = 1.349, \beta_4 = 0.631$	60170.11	60451.65

Comparing tables 6.4.2, and 6.4.4 reveals only minor discrepancies, so that the ranking for three parameter structures obtained using the modified cost $J_M = \frac{1}{2} \langle e, Qe \rangle + \rho/2 \langle \beta, R_0 \beta \rangle$ has not changed from that of the proper least squares approach.

Table 6.4.5 Three Parameter Solutions Ranked by the Lower Bound Values on Diagnostic $\Delta J_O^S(\rho)$

Rank	Structure	Lower Bound	$\Delta J_O^S(\rho)$	Rank by $\Delta J_O^S(\rho)$
1	$\beta_1 \beta_2 \beta_4$	60170.11	60451.65	4
2	$\beta_0 \beta_2 \beta_4$	59989.56	60109.56	) All below) 4th Position)
3	$\beta_1 \beta_2 \beta_3$	59910.28	60114.61	
4	$\beta_2 \beta_3 \beta_4$	59900.73	60108.24	

Although the bounds isolate a structure whose optimal cost improvement is within .113% of the full improvement possible, the lower bound ranking does not follow that of the actual optimal cost improvements $\Delta J_O^S(\rho)$.

One of a-priori assumptions made for this investigation was that the optimal least squares parameter vector $\beta(\rho=0)$ had been calculated. Further, a uniform weighting was chosen for all the least squares parameters.

This assumption meant that the least square parameters contributed unequally to the parameter cost $\langle \beta, R_o \beta \rangle_{E_n}$ at the full optimum. In a structure assessment the emphasis is on determining which parameters are important to producing a small error, $e = y - W\beta$, given that any selection may be used. That is individual parameter costs at the optimum must be relatively equal. The actual solution of the least squares problem can be used to construct a new coefficient weight, so that at the full optimum all the parameter contributions are similar. Again the parameter ρ was used to ensure only a small perturbation to the original least squares problem.

Assessment 2 For $\rho=0$, the optimal parameter vector was evaluated as $\beta^T(\rho=0) = (62.405, 1.551, 0.51, 0.102, 0.144)$. The coefficient weights were set at $R_o(i) = [\hat{\beta}(i)]^{-2}$ to give $R_o = \text{diag. } (0.0003, 0.4157, 3.8447, 96.1146, 48.2252)$.

Table 6.4.6 Optimal Cost and Parameters for the Modified Cost Problem where $Q = I_{13}$, using R_o (modified) with $0 \leq \rho \leq 1$

ρ	β_o	β_1	β_2	β_3	β_4	$\Delta J_o^f(\rho)$
1.0	63.287	1.513	0.509	0.061	-0.146	60517.814
10^{-1}	63.811	1.533	0.497	0.083	-0.157	60519.869
10^{-3}	62.441	1.551	0.510	0.101	-0.144	60520.111
10^{-5}	62.406	1.551	0.510	0.102	-0.144	60520.113
0	62.405	1.551	0.510	0.102	-0.144	60520.113

It is immediately clear how much more robust the parameter estimates are compared to the variation in ρ . This conclusion is re-inforced by the details of the three parameter structures obtained for $\rho = 10^{-5}$.

Table 6.4.7 Three Parameter Solutions obtained for
 $\frac{J}{M} (\rho=10^{-5})$, Ranked by $\Delta J_o^S(\rho)$

Rank	Parameter Values	Lower Bound	$\Delta J_o^S(\rho)$
1	$\beta_o = 52.577, \beta_1 = 1.468, \beta_2 = 0.662$	60500.7	60515.093
2	$\beta_o = 103.097, \beta_1 = 1.44, \beta_4 = -0.614$	59987.2	60506.664
3	$\beta_o = 131.282, \beta_3 = -1.2, \beta_4 = 0.725$	60153.51	60456.175
4	$\beta_1 = 1.528, \beta_2 = 1.349, \beta_4 = 0.631$	60184.88	60451.653

Comparing tables 6.4.2, 6.4.4 and the above table 6.4.7 indicate that the structures of this second assessment reflect more closely the actual least square structures both in parameter and optimal cost improvement values.

Table 6.4.8 Three Parameter Structures Ranked by Lower Bound
Values on the Diagnostic $\Delta J_o^S(\rho)$

Rank	Structure	Lower Bound	$\Delta J_o^S(\rho)$	Rank by $\Delta J_o^S(\rho)$
1	$\beta_o, \beta_1, \beta_2$	60500.722	60515.093	1
2	$\beta_1, \beta_2, \beta_4$	60184.877	60451.653	4
3	$\beta_o, \beta_3, \beta_4$	60153.510	60456.175	3
4	$\beta_o, \beta_2, \beta_3$	60094.490	60336.323	-

The additional information incorporated through the coefficient weight matrix has produced results which reflect the actual structure ranking much more closely, especially when it is realized that three out of the top four least square structures of three parameters have been isolated. Further this top structure $\{\beta_o, \beta_1, \beta_2\}$ attains all but 0.008% of the full

optimal cost improvement available.

Another feature of the lower bound behaviour is also exhibited by the case $\rho = 10^{-5}$. This is the property that as $\rho \rightarrow 0$, the lower bounds attain a limiting value as is shown in the table below:

Table 6.4.9 Variation of the Lower Bound Values for ΔJ_0^S
(using the Modified Coefficient Weight) for $10^{-5} \leq \rho \leq 1$.

Structure	$\rho = 1.0$	$\rho = 10^{-1}$	$\rho = 10^{-3}$	$\rho = 10^{-5}$
$\beta_0 \quad \beta_1 \quad \beta_2$	60499.13	60500.56	60500.72	60500.72
$\beta_1 \quad \beta_2 \quad \beta_4$	60174.75	60183.86	60184.87	60184.88
$\beta_0 \quad \beta_3 \quad \beta_4$	60137.68	60151.91	60153.49	60153.51
$\beta_0 \quad \beta_2 \quad \beta_4$	60092.86	60094.33	60094.49	60094.49

This behaviour is symptomatic of the information limit determined by the number of moments available, a barrier which can only be broken by the generation of additional moments to yield yet tighter bounds. However, the finer differentiation of more structure bounds has to be balanced against the additional computation expense which would be incurred.

In conclusion, this application of the structures procedure to the least squares parameter estimation problem has usefully demonstrated some of the asymptotic features of the diagnostic bound calculations. Although a more thorough theoretical investigation is required, it has been shown that, used with sensitivity, the structures procedure is capable of isolating the more profitable sub-optimal structures for further investigation.

Part II Output Structures

Introduction

Chapter 7 Primal and Dual System Concepts

Chapter 8 Optimal Output Structures for Linear Stochastic
Systems.

II.0 Introduction

When posing the input structure problem, the question being asked is whether system performance is appreciably diminished by imposing the constraints $u_{i_1} = 0, u_{i_2} = 0, \dots, u_{i_s} = 0$ for some subset of the complete control set $\{u_1, u_2, \dots, u_m\}$. The assumed linearity of the system equations is exploited directly in the subsequent analysis of the problem.

The output structures problem is best explained by assuming an output feedback control law, namely $u = Fy$. Output structures are then seen to be the imposition of the constraints $y_{i_1} = 0, y_{i_2} = 0, \dots, y_{i_s} = 0$ (for some subset of the complete output set $\{y_1, \dots, y_r\}$) in the feedback law, and the subsequent monitoring of the degradation in system performance.

Since the input structure problem is set in the mathematical framework of a quadratic cost functional subject to a deterministic system constraint, it was initially thought that the Lagrange dual problem might provide the correct setting for the output structures analysis. The preliminary investigation of Chapter 7 only serves to emphasise the fact that although y_{i_s} is regarded as a zero input to the feedback law, any successful output structures diagnostic must also monitor the actual system output y_{i_s} which may be significantly non-zero.

Attempts to implement this feature in the primal problem, using an operator formulation also failed, but this time the cause was the complexity and subsequent intractability of the expressions obtained.

However, by setting the output structure problem in a stochastic environment an approach is developed which is both conceptually pleasing and computationally feasible. The resulting formulation finds its affirmation in the design philosophy of Kwakernaak and Sivan, 1972b for which it provides a powerful computational tool.

Chapter 7 Primal and Dual System Concepts

7.0 Introduction

7.1 Primal and Lagrange Dual Systems

7.2 Conclusions

7.0 Introduction

The motivation for the output structures problem is to assess which outputs can be removed from a feedback scheme with an acceptable degradation in the system performance. In the spirit of Part I, the system performance is measured by a quadratic cost functional and the system is considered to be linear. With this framework, a feedback description is sought which has the same mathematical structure of the input deletion problem so that the structures assessment technique can be applied.

The method of Lagrange duality gives such a feedback description. It is unfortunate that a straightforward output deletion procedure applied to this formulation does not have a very useful structures interpretation. A reformulation utilizing a stochastic setting will be seen to have a more satisfying interpretation.

The analysis in the following formally uses Hilbert space notation with specialization to a state space description (system triple $\mathcal{S}(A, B, C)$) for control interpretations.

7.1 Primal and Lagrange Dual Systems

Introduce the primal problem as

$$\text{Min } J = \frac{1}{2} \langle y, Qy \rangle_{H_r} + \frac{1}{2} \langle u, Ru \rangle_{H_m}$$

$u \in H_m$

$$\text{subject to } y = Wu + d \in H_r, \quad (Q > 0 \text{ and } R > 0) \quad 1.7.1$$

Adjoining the constraint using a Lagrange multiplier, v obtain :

$$J_a = \frac{1}{2} \langle y, Qy \rangle_{H_r} + \frac{1}{2} \langle u, Ru \rangle_{H_m} - \langle v, y - Wu - d \rangle_{H_r}$$

for which the equations of optimality are :

$$Qy - v = 0, \quad Ru + W^*v = 0, \quad \text{and} \quad y = Wu + d \quad 1.7.2$$

Eliminating y and u from J_a obtain

$$J_a = -\frac{1}{2} \langle v, Q^{-1}v \rangle_{H_r} - \frac{1}{2} \langle v, WR^{-1}W^*v \rangle_{H_r} + \langle v, d \rangle_{H_r}$$

$$\text{Define } Z \triangleq W^*v, \quad \text{and noting that } \frac{\partial^2 J_a}{\partial v^2} = -(Q^{-1} + WR^{-1}W^*) < 0$$

the Lagrange dual optimization problem is

$$\text{Max } J_a = -\frac{1}{2} \langle v, Q^{-1}v \rangle_{H_r} - \frac{1}{2} \langle Z, R^{-1}Z \rangle_{H_m} + \langle v, d \rangle_{H_r}$$

$v \in H_r$

$$\text{subject to } Z = W^*v \in H_m. \quad 1.7.3$$

Using state space representations for the Hilbert space operators enables the optimality conditions 1.7.2 to be given a state space interpretation :

$$Qy = v \Rightarrow Qy(t) = v(t); \quad Z = W^*v \Rightarrow Z(t) = \int_t^{T_f} B^T e^{A^T(\tau-t)} C^T v(\tau) d\tau$$

$$\text{Define } \lambda(t) \triangleq - \int_t^{T_f} e^{A^T(\tau-t)} C^T v(\tau) d\tau \quad \text{and using Leibniz rule}$$

$$\text{for differentiation of an integral obtain } \dot{\lambda}(t) = -A^T \lambda(t) + C^T v(t)$$

$$\text{for which } \lambda(T_f) = 0 \text{ gives the above solution. Finally } Ru + W^* \lambda = 0$$

$$\text{gives } u(t) = R^{-1} B^T \lambda(t). \text{ Consequently the complete Lagrange dual}$$

$$\text{system } \dot{\lambda}(-A^T, C^T, B^T) \triangleq \dot{\lambda}(t) = -A^T \lambda(t) + C^T v \quad \lambda(T_f) = 0$$

$$\tilde{z}(t) = B^T \lambda(t)$$

Using the relations $u(t) = R^{-1} \tilde{z}(t)$, $v = Q^{-1} y(t)$ and choosing

$\lambda(0)$ so that $\lambda(T_f) = 0$ (Euler-Lagrange Transversality condition)

the familiar primal/dual system configuration may be drawn, as in Fig 7.1.1

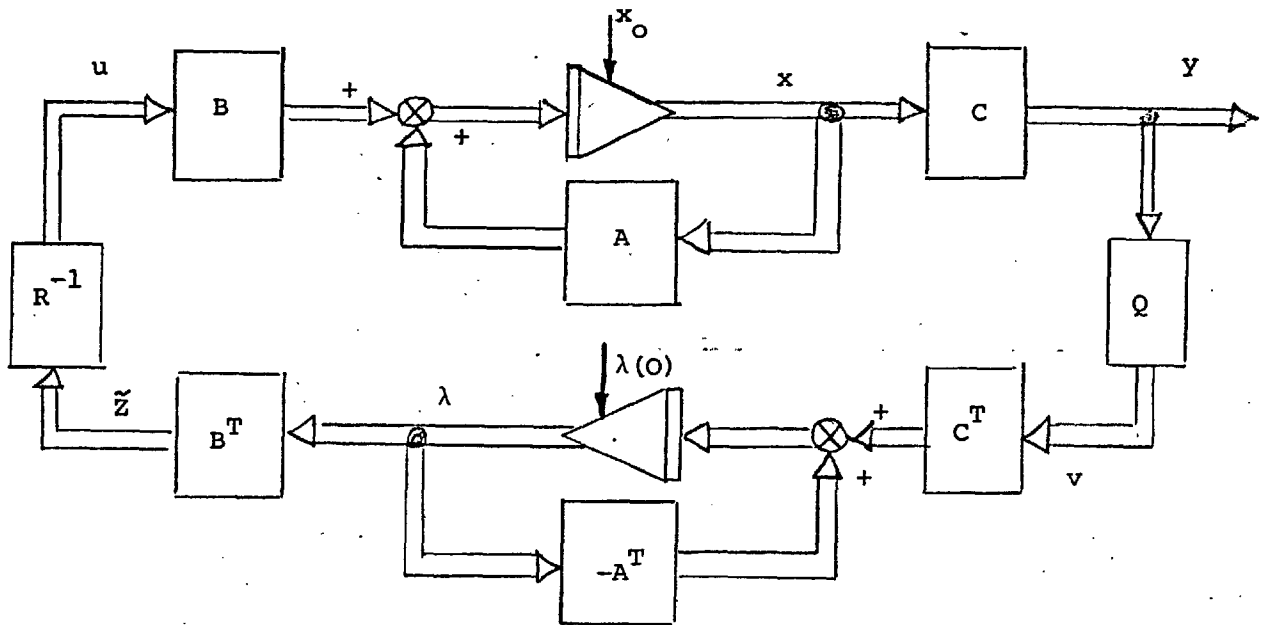


Figure 7.1.1 Primal/Dual System Configuration

If cost weighting matrices Q and R are assumed diagonal there is a direct relationship between the entries of $\tilde{z}$ and u , and the entries of v and y . Hence the Lagrange dual system has the primal system outputs as inputs, and the primal system inputs as outputs. Clearly if $\lambda(0)$ is chosen so that $\lambda(T_f) = 0$ the Lagrange dual system may be written in weighting function form as $Z = \tilde{W}v + \tilde{d}$ where

$$(\tilde{W}v)(t) \triangleq \int_0^t B^T e^{-A^T(t-\tau)} C^T v(\tau) d\tau \text{ and } \tilde{d}(t) \triangleq B^T e^{-A^T t} \lambda(0).$$

This is the description of the ideal feedback operator, and output structure are considered by setting $v_s = 0 \Leftrightarrow y_s = 0$ for different structures and assessing the degradation in the optimal cost.

First, the effect of imposing such a constraint in the Lagrange Dual

optimization problem is considered. Suppose v_s is obtained from v by deletion, then the structured dual optimization problem is :

$$\text{Max}_{v_s \in H_s} J_d^s = -\frac{1}{2} \langle v_s, Q_s^{-1} v_s \rangle_{H_s} - \frac{1}{2} \langle Z, R^{-1} Z \rangle_{H_m} + \langle v_s, d_s \rangle_{H_r}$$

$$\text{subject to } Z = W_s^* v_s \in H_m$$

$$\text{That is } \text{Max}_{v_s \in H_s} J_d^s = -\frac{1}{2} \langle v_s, (Q_s^{-1} + W_s R^{-1} W_s^*) v_s \rangle_{H_s} + \langle v_s, d_s \rangle_{H_r}$$

$$\Rightarrow \hat{v}_s = (Q_s^{-1} + W_s R^{-1} W_s^*)^{-1} d_s \Rightarrow \hat{J}_d^s = \frac{1}{2} \langle d_s, (Q_s^{-1} + W_s R^{-1} W_s^*)^{-1} d_s \rangle_{H_r}$$

An application of Householder's inequality gives

$$\hat{J}_d^s = \frac{1}{2} \langle d_s, Q d_s \rangle_{H_s} - \frac{1}{2} \langle W_s^* Q d_s, (R + W_s^* Q W_s)^{-1} W_s^* Q d_s \rangle_{H_m}$$

which is clearly the optimal cost of the primal problem with y_s obtained from y by the same deletions as v_s . This structured primal problem is stated as

$$\text{Min}_{u \in H_m} J^s = \frac{1}{2} \langle y_s, Q y_s \rangle_{H_s} + \frac{1}{2} \langle u, R u \rangle_{H_m}$$

$$\text{subject to } y_s = W_s u + d_s \in H_s$$

Denoting the full optimal cost by $\hat{J}(\hat{y}, \hat{u})$, and the structure optimal cost as $J^s(\tilde{y}_s, \tilde{u})$ the following analysis provides the interpretations which can be made by examining the quantity $\delta = \hat{J}(\hat{y}, \hat{u}) - J^s(\tilde{y}_s, \tilde{u})$.

$$\text{Partition the output equation as } y_s = W_s u + d_s$$

$$y_{r-s} = W_{r-s} u + d_{r-s}$$

$$\text{Then } \hat{J}(\hat{y}, \hat{u}) = \frac{1}{2} \langle \hat{y}, Q \hat{y} \rangle_{H_r} + \frac{1}{2} \langle \hat{u}, R \hat{u} \rangle_{H_m}$$

$$= \frac{1}{2} \langle \hat{y}_s, Q_s \hat{y}_s \rangle_{H_s} + \frac{1}{2} \langle \hat{u}, R \hat{u} \rangle_{H_m} + \frac{1}{2} \langle \hat{y}_{r-s}, Q_{r-s} \hat{y}_{r-s} \rangle_{H_{r-s}}$$

$$= J^s(\hat{y}_s, \hat{u}) + \frac{1}{2} \langle \hat{y}_{r-s}, Q_{r-s} \hat{y}_{r-s} \rangle_{H_{r-s}}$$

$$\text{Hence } \delta = \hat{J}(\hat{y}, \hat{u}) - J^S(\tilde{y}_s, \tilde{u})$$

$$= J^S(\hat{y}_s, \hat{u}) - J^S(\tilde{y}_s, \tilde{u}) + \frac{1}{2} \langle \hat{y}_{r-s}, Q_{r-s} \hat{y}_{r-s} \rangle_{H_{r-s}}$$

$$\text{Clearly } i) \quad J^S(\hat{y}_s, \hat{u}) - J^S(\tilde{y}_s, \tilde{u}) \geq 0$$

Otherwise $(\hat{y}_s, \hat{u})$ contradicts the optimality of $(\tilde{y}_s, \tilde{u})$

$$\text{and } ii) \quad \frac{1}{2} \langle \hat{y}_{r-s}, Q_{r-s} \hat{y}_{r-s} \rangle_{H_{r-s}} \geq 0 \quad \text{since } Q_{r-s} > 0$$

$$\text{Implying } \delta \geq J^S(\hat{y}_s, \hat{u}) - J^S(\tilde{y}_s, \tilde{u}) \geq 0 \quad 7.1.4$$

$$\text{and } \delta \geq \frac{1}{2} \langle \hat{y}_{r-s}, Q_{r-s} \hat{y}_{r-s} \rangle_{H_{r-s}} \geq 0 \quad 7.1.5$$

The interpretation of these inequalities depends on the smallness of the parameter δ . Given δ sufficiently small relation 7.1.4 shows that the pairs $(\tilde{y}_s, \tilde{u})$ and $(\hat{y}_s, \hat{u})$ produce similar cost values, and the second relation 7.1.5 gives the smallness of the outputs $\hat{y}_{r-s}$. However for an unambiguous conclusion to be drawn the restrictions on the size of δ would be quite severe. Furthermore the effect of the control $\tilde{u}$ on the output y_{r-s} is not monitored. These two reasons lead to the conclusion that output deletion in the primal problem or equivalently input deletion in the Lagrange dual problem does not answer the question of the importance of particular outputs in the feedback structure.

Attempts to incorporate the monitoring of the effect of the sub-optimal control $\tilde{u}$, on the outputs y_{r-s} using an operator formulation in the primal problem also failed. The cause here was the complexity and apparent intractability of the expressions obtained.

7.2 Conclusions

A preliminary discussion into the use of the Lagrange dual formulation as a setting for the output structures problem has been given.

Although no useful computational procedures have been devised, the control interpretations of the dual system setting, and the output structures problem seem to have an affinity which must merit further investigation.

Chapter 8 Optimal Output Structures for Linear Stochastic Systems

8.0 Introduction

8.1 Minimum Variance Estimators

8.2 Output Structures for Linear Stochastic Systems

8.2.0 Introduction

8.2.1 Basic Theory

8.2.2 An Output Structures Problem for Linear Stochastic Systems

8.2.3 Controllers of Reduced Dimensions

8.3 Conclusions

8.0 Introduction

The first part of this chapter is concerned with an application in linear regression theory, in particular to the minimum variance estimator. The development initially follows that of Luenberger, 1969 until a useful identification between a minimum variance problem and optimization problem is made. The optimization problem is seen to have the mathematical form to which a structures procedure may be applied.

This more general formulation is particularized to the problem of the construction of a Kalman-Bucy Filter where the application of deletions is made physically meaningful as a type of output structure problem. It is shown (section 8.2.1) that the state estimation problem can be formulated as n independent optimization problems, each problem corresponding to the estimation of a specific state. In fact each state optimization problem fits the format assumed earlier in the thesis, and hence the structures procedures can be applied. Structures are here concerned with the removal of output measurements in constructing a state estimator. In this context the well-known separation theorem enables the input structure procedure for the deterministic description of the first part of the thesis to be united with this stochastic output structure problem to form an overall controller design philosophy. This philosophy has previously been formulated by Kwakernaak and Sivan, 1972b but the computational tools developed here to determine the controller structure should greatly strengthen the design procedure.

8.1 Minimum Variance Estimators

Consider the estimation of vector β from a set of measurements of the form $y = W\beta + \epsilon$ where both β and ϵ are considered to be random vectors. The criterion for optimality is the minimization of $E \left[\|\hat{\beta} - \beta\|_2^2 \right]$ where $\hat{\beta}$ is a linear estimate for β . In this general setting, have Theorem 8.1.1. (Luenberger, 1969) Let y and β be random vectors (not necessarily of the same dimension). Assume $\{E[yy^T]\}^{-1}$ exists. The linear estimate $\hat{\beta}$ of β , based on y , minimizing $E \left[\|\hat{\beta} - \beta\|_2^2 \right]$ is $\hat{\beta} = E[\beta y^T] \{E[yy^T]\}^{-1} y$, with corresponding error covariance matrix

$$\begin{aligned} E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] &= E[\beta\beta^T] - E[\hat{\beta}\hat{\beta}^T] \\ &= E[\beta\beta^T] - E[\beta y^T] E[yy^T]^{-1} E[y\beta^T]. \end{aligned}$$

(This is the minimum variance estimator, which is unbiased if both y and β have zero means.) The theorem is applied to a specific case to obtain the following corollary.

Corollary 8.1.1 (Luenberger, 1969) Suppose $y = W\beta + \epsilon$ where y is a known m dimensional data vector, β is an n -dimensional unknown random vector, ϵ is an unknown m -dimensional random error vector, W is a known $(m \times n)$ constant matrix and $E[\epsilon\epsilon^T] = Q$, $E[\beta\beta^T] = R$, $E[\epsilon\beta^T] = 0_{m \times n}$.

Assuming that R and Q are positive-semidefinite matrices (of appropriate size) and that $WRW^* + Q$ is non singular. Then the linear estimate $\hat{\beta}$, of β minimizing $E \left[\|\hat{\beta} - \beta\|_2^2 \right]$ is $\hat{\beta} = RW^*(WRW^* + Q)^{-1}y$ with error covariance $E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = R - RW^*(WRW^* + Q)^{-1}WR$.

With the additional restriction that R and Q are positive definite and invertible, the Householder matrix equality may be used to give $E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = (Q^{-1} + W^*R^{-1}W)^{-1}$.

The theorem and its corollary are used to motivate the following development which will conclude with an identification between a minimum variable problem and a finite dimensional optimization problem.

Recall the cost to be minimized is

$$J = E \left[\left\| \hat{\beta} - \beta \right\|_2^2 \right] = E \left[\sum_{i=1}^n (\hat{\beta}_i - \beta_i)^2 \right] = \sum_{i=1}^n J_i$$

where $J_i = E \left[(\hat{\beta}_i - \beta_i)^2 \right]$ Each of these subproblems is independent so that the i th problem may be isolated as:

$$\text{Min } J_i = E \left[(\hat{\beta}_i - \beta_i)^2 \right] \quad i=1, 2, \dots, n$$

$$\text{subject to } y = W\beta + \epsilon, \quad E[\epsilon\epsilon^T] = Q \quad E[\epsilon\beta^T] = 0 \\ E[\beta\beta^T] = R$$

The corollary suggests a linear estimator of the form $\hat{\beta}_i = k^T y$

$$\text{Hence } J_i = E \left[(k^T y - \beta_i)^2 \right] = E \left[(k^T y - e_i^T \beta)^2 \right]$$

where $e_i^T = (0, 0, \dots, 1, 0, \dots, 0)$, unity in the i th entry.

$$\text{Giving } J_i = E \left[k^T y y^T k - 2k^T y \beta e_i + e_i^T \beta \beta^T e_i \right]$$

Have, therefore $yy^T = (W\beta + \epsilon)(W\beta + \epsilon)^T$ to obtain

$$E[yy^T] = E \left[W\beta\beta^T W^T + W\beta\epsilon^T + \epsilon\beta^T W^T + \epsilon\epsilon^T \right] = WRW^T + Q$$

$$\text{and } E[y\beta^T] = E \left[W\beta\beta^T + \epsilon\beta^T \right] = WR$$

$$\text{Min } E \left[(\hat{\beta}_i - \beta_i)^2 \right] = \text{Min}_{k \in \mathbb{R}^m} J_i = \underset{E_m}{< k, (WRW^* + Q) k >} - 2 \underset{E_m}{< k, WR e_i >} + \underset{E_n}{< e_i, Re_i >}$$

Clearly the optimal value for k is $\hat{k} = (WRW^* + Q)^{-1} WR e_i$ which should

be compared with the result in corollary 8.1.1. The finite dimensional optimization problem can be stated in a more familiar form as

$$\text{Min}_{k \in \mathbb{R}^m} J_i = \underset{E_n}{< z, R z >} + \underset{E_m}{< k, Q k >}$$

$$\text{subject to } z = W^* k - e_i \quad i=1, 2, \dots, n$$

An input deletion procedure applied to this problem is a method of determining which elements in the vector $k \in \mathbb{R}^m$ can be zeroed without a large degradation in the optimal cost. On a more practical level such a structure analysis could indicate which measurements are not required for the estimation of a particular parameter. Furthermore by considering all n subproblems en bloc the effect of the removal of a measurement from all the parameters can be assessed.

Finally a straightforward extension to theorem 8.1.1 is derived.

Theorem 8.1.2 Let y , and β be random vectors (not necessarily of the same dimension) such that $E[y] = y_0$, $E[\beta] = \beta_0$,

and with finite co-variance matrices. Then the minimum variance estimate $\hat{\beta}$, of β in the form $\hat{\beta} = K y + b$ is

$$\hat{\beta} = \beta_0 + E[(\beta - \beta_0)(y - y_0)^T] \{E[(y - y_0)(y - y_0)^T]\}^{-1}(y - y_0)$$

Proof Recall $y = W\beta + \varepsilon$

$$\text{then } E[y] \triangleq y_0 = WE[\beta] + E[\varepsilon] = W\beta_0$$

$$\Rightarrow y - y_0 = W(\beta - \beta_0) + \varepsilon$$

Define $\tilde{y} = y - y_0$ and $\tilde{\beta} = \beta - \beta_0$ where $\tilde{y} = W\tilde{\beta} + \varepsilon$

$$\Rightarrow E[\tilde{y}] = 0 \quad \text{and} \quad E[\tilde{\beta}] = 0$$

Apply theorem 8.1.1 to pair $(\tilde{y}, \tilde{\beta})$ to obtain

$$\hat{\tilde{\beta}} = \hat{\beta} - \beta_0 = E[\tilde{\beta}\tilde{y}^T] \{E[\tilde{y}\tilde{y}^T]\}^{-1}\tilde{y}$$

$$\Rightarrow \hat{\beta} = \beta_0 + E[(\beta - \beta_0)(y - y_0)^T] \{E[(y - y_0)(y - y_0)^T]\}^{-1}(y - y_0) \quad \triangle$$

It is easily seen that this estimator $\hat{\beta}$ is unbiased since

$$E[\hat{\beta}] = \beta_0. \quad \text{This more general result concludes the discussion on}$$

minimum variance estimators an application of which follows.

8.2 Output Structures for Linear Stochastic Systems

8.2.0 Introduction

The analysis to follow is of a formal nature which can be justified under more mathematically rigorous conditions. The tract of Bucy and Joseph, 1968, is cited as a source for a rigorous exposition of many of the operations performed in the sequel.

For the problem of determining the optimal output feedback regulator for the control of a linear stochastic system, the separation principle states that the problem decouples into i) the determination of an estimate of the state (using a filter) and ii) using the optimal feedback law (operating on this state estimate) for the deterministic quadratic regulator. The estimate of the state is to be obtained from the system output which contains noise, and the known system input. An output structures problem is developed similar in outline to the structures problem for minimum variance estimators. In fact the same question is being asked, namely whether an output need be measured to estimate a particular state or more generally whether an output need be measured at all for the purposes of estimation.

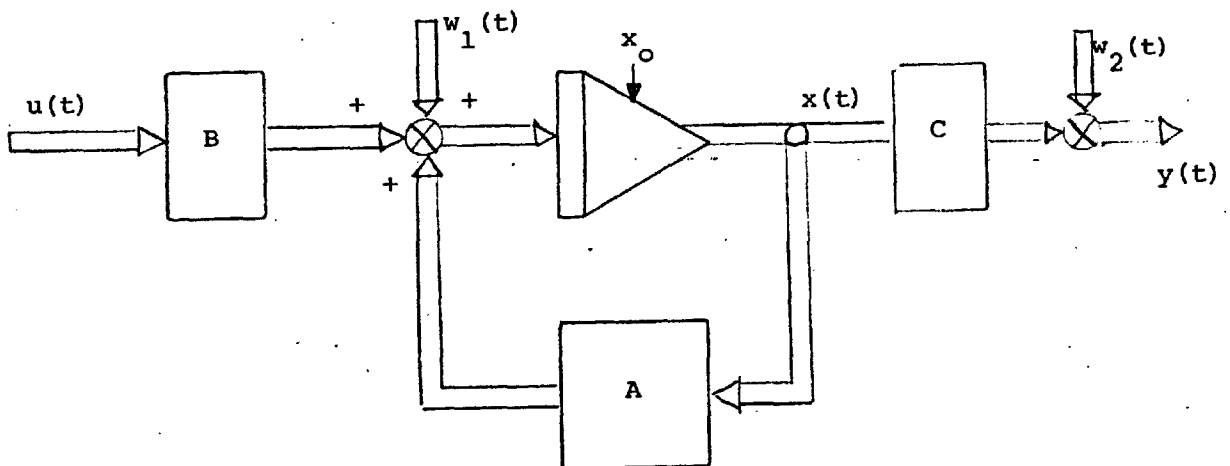
8.2.1 Basic Theory

The system is assumed to be a linear differential system subject to additive white noise disturbances on the input and output. The system configuration is shown in Fig 8.2.1.1 below.

$$\text{System Description} \quad \dot{\mathbf{x}}(t) = \mathbf{A} \mathbf{x}(t) + \mathbf{B} \mathbf{u}(t) + \mathbf{w}_1(t) \quad 8.2.1$$

$$\mathbf{y}(t) = \mathbf{C} \mathbf{x}(t) + \mathbf{w}_2(t) \quad 8.2.2$$

Fig. 8.2.1.1 Linear Stochastic System Configuration



The additive white noise disturbances have the following statistics :

$$E[\mathbf{w}_1(t)\mathbf{w}_2^T(\tau)] = \mathbf{Q}(t) \delta(t-\tau) \quad E[\mathbf{w}_1(t)] = 0$$

$$E[\mathbf{w}_2(t)\mathbf{w}_2^T(\tau)] = \mathbf{R}(t) \delta(t-\tau) \quad E[\mathbf{w}_2(t)] = 0$$

$$\text{and } E[\mathbf{w}_1(t)\mathbf{w}_2^T(\tau)] = \mathbf{0}_{n \times r} \quad \forall (t, \tau) \text{ pairs} \quad 8.2.3$$

The initial condition $\mathbf{x}_0$ is assumed to occur at $t = t_0$

$$\text{for which } E[\mathbf{x}_0] = \bar{\mathbf{x}}_0 \quad \text{and} \quad E[(\mathbf{x}_0 - \bar{\mathbf{x}})(\mathbf{x}_0 - \bar{\mathbf{x}})^T] = \mathbf{P}_0 \quad 8.2.4$$

and furthermore $\mathbf{x}_0$ is independent of both white noise processes

$$\mathbf{w}_1(t) \quad \text{and} \quad \mathbf{w}_2(t) \quad \forall t.$$

Define $\Phi(t) = e^{At}$, $t > 0$ for which $\dot{\Phi}(t) = A \Phi(t)$

Then from equation 8.2.1 obtain the state at time t_1 as :

$$x(t_1) = \int_{t_0}^{t_1} \Phi(t_1 - \tau) B u(\tau) d\tau + \Phi(t_1 - t_0) x_0 + \int_{t_0}^{t_1} \Phi(t_1 - \tau) w_1(\tau) d\tau.$$

The control input $u(\tau)$ is assumed known on $[t_0, t_1)$, consequently

$$\text{under expectation } E[x(t_1)] = \Phi(t_1 - t_0) \bar{x}_0 + \int_{t_0}^{t_1} \Phi(t_1 - \tau) B u(\tau) d\tau. \quad 8.2.5$$

Recall 8.2.2 and again take expectations to obtain

$$E[y(t_1)] = C E[x(t_1)] \quad 8.2.6$$

Using expressions 8.2.5, and 8.2.6 the system equations are to be transformed after the manner of Theorem 8.1.2.

$$\text{Define } \tilde{x}(t_1) = x(t_1) - E[x(t_1)] \quad 8.2.7$$

$$\text{then } \frac{d\tilde{x}}{dt_1}(t_1) = \dot{x}(t_1) - \frac{dE[x(t_1)]}{dt_1}$$

$$= Ax(t_1) + Bu(t_1) + w_1(t_1) - \{ A \Phi(t_1 - t_0) \bar{x}_0$$

$$+ \int_{t_0}^{t_1} A \Phi(t_1 - \tau) B u(\tau) d\tau + Bu(t_1) \}$$

$$\text{gives } \dot{\tilde{x}}(t_1) = A \tilde{x}(t_1) + w_1(t_1)$$

$$\text{Define } \tilde{y}(t_1) = y(t_1) - E[y(t_1)] = C x(t_1) + w_2(t_1) - C E[x(t_1)]$$

$$\Rightarrow \tilde{y}(t_1) = C \tilde{x}(t_1) + w_2(t_1) \quad 8.2.8$$

With this reformulation, consider the estimation of the i th element of $\tilde{x}_i(t_1)$ by a linear estimator of the form

$$\hat{\tilde{x}}_i(t_1) = \int_{t_0}^{t_1} v^T(\tau) \tilde{y}(\tau) d\tau$$

$$\text{to minimize } J_i = E \left[\left(\tilde{x}_i(t_1) - \int_{t_0}^{t_1} v^T(\tau) \tilde{y}(\tau) d\tau \right)^2 \right]$$

Introduce $e_1^T \triangleq (0, 0, \dots, 1, 0, \dots, 0)$ unity in the i th position and the differential equation $\dot{\lambda}(t) = -A^T \lambda(t) + C^T v(t)$ with final condition $\lambda(t_1) = e_1$.

Observe that $\frac{d}{dt} \lambda^T(t) \tilde{x}(t) = v^T(t) \tilde{y}(t) - v^T(t) w_2(t) + \lambda^T(t) w_1(t)$

which may be integrated on $[t_0, t_1]$ to obtain

$$\begin{aligned} \lambda^T(t_1) \tilde{x}(t_1) - \lambda^T(t_0) \tilde{x}(t_0) &= \int_{t_0}^{t_1} v^T(t) \tilde{y}(t) dt - \int_{t_0}^{t_1} v^T(t) w_2(t) dt \\ &+ \int_{t_0}^{t_1} \lambda^T(t) w_1(t) dt \end{aligned}$$

Using the initial condition $\lambda(t_1) = e_1$ the above may be rearranged to give

$$\tilde{x}_1(t_1) - \int_{t_0}^{t_1} v^T(t) \tilde{y}(t) dt = \lambda^T(t_0) \tilde{x}(t_0) - \int_{t_0}^{t_1} v^T(t) w_2(t) dt + \int_{t_0}^{t_1} \lambda^T(t) w_1(t) dt$$

Squaring and taking expectations gives when all the independences are considered.

$$\begin{aligned} E \left[\left(\tilde{x}_1(t_1) - \int_{t_0}^{t_1} v^T(t) \tilde{y}(t) dt \right)^2 \right] &= \langle \lambda(t_0), P_0 \lambda(t_0) \rangle_{E_n} \\ &+ \int_{t_0}^{t_1} \{ v(t), R(t) v(t) \}_{E_r} + \langle \lambda(t), Q(t) \lambda(t) \rangle_{E_n} dt \end{aligned}$$

This enables the following equivalence with respect to minimization to be stated.

$$\begin{aligned} \text{Min } E \left[(\tilde{x}_1(t_1) - \hat{x}_1(t_1))^2 \right] &= \text{Min } \langle \lambda(t_0), P_0 \lambda(t_0) \rangle_{E_r} \\ &\quad v \in H_r \\ &+ \int_{t_0}^{t_1} \{ \langle v(t), R(t) v(t) \rangle_{E_r} + \langle \lambda(t), Q(t) \lambda(t) \rangle_{E_n} \} dt \end{aligned}$$

subject to $\dot{\lambda}(t) = -A^T \lambda(t) + C^T v(t), \lambda(t_1) = e_1$ 8.2.9

for which the minimum variance estimator is

$$\hat{x}_i(t_1) = \int_{t_0}^{t_1} v^T(t) \tilde{y}(t) dt \quad 8.2.10$$

As is well known it is possible to realize the function $v(\cdot)$ as :

$$v(t) = R^{-1}(t) C P(t) \lambda(t) \quad t_0 \leq t \leq t_1 \quad 8.2.11$$

where $P(\cdot)$ satisfies a matrix riccati equation of the form

$$\frac{dP}{dt} = P(t)A^T + AP(t) - P(t)C^T R(t)^{-1} C P(t) + Q(t) \quad P(t_0) = P_0 \quad 8.2.12$$

Substituting the expression for $v(\cdot)$ in to the adjoint system

$$\text{equation obtain } \lambda(t) = \phi_F^T(t_1, t) e_i \quad t < t_1$$

where $\phi_F(t_1, t)$ is the transition matrix associated with

$$\dot{\lambda}(t) = - [A - P(t)C^T R(t)^{-1} C] \lambda(t)$$

With this form for $\lambda(\cdot)$ and $v(\cdot)$ substitute in to 8.2.10 to obtain

$$\hat{x}_i(t_1) = \int_{t_0}^{t_1} e_i^T \phi_F(t_1, t) P(t) C^T R^{-1}(t) \tilde{y}(t) dt$$

Differentiate to give

$$\frac{d\hat{x}_i(t_1)}{dt_1} = e_i^T \{ F(t_1) \hat{x}(t_1) - K(t_1) \tilde{y}(t_1) \} \quad \hat{x}(t_0) = 0$$

where $F(t) = A + K(t)C$ and $K(t) = -P(t) C^T R(t)^{-1}$

or equivalently

$$\frac{d\hat{x}_i(t_1)}{dt_1} = e_i^T \{ A \hat{x}(t_1) + K(t_1) [C \hat{x}(t_1) - \tilde{y}(t_1)] \}$$

For all the indices $i = 1, 2, \dots, n$ the optimal filter can be realized

$$\text{as } \dot{\hat{x}}(t) = A \hat{x}(t) + K(t) [C \hat{x}(t) - \tilde{y}(t)] \quad \hat{x}(t_0) = 0$$

where it should be recalled that $K(t) = -P(t) C^T R(t)^{-1}$ and

$\hat{x}$ estimates $\tilde{x}$ which along with $\tilde{y}$ are transformed variables.

(See expressions 8.2.7 and 8.2.8)

A final result to conclude this discussion of the basic theory is to obtain the actual value of the minimum of the deterministic problem described by equations 8.2.9. A reversal of the time variable establishes an equivalent minimization problem as (Anderson and Moore, 1971)

$$\text{Min}_{\mathbf{v} \in H_r} J_i = \langle \hat{\lambda}(\tau_0), P_0 \hat{\lambda}(\tau_0) \rangle + \int_{\tau_1}^{\tau_0} \{ \langle \hat{v}(\tau), \hat{R}(\tau) \hat{v}(\tau) \rangle + \langle \hat{\lambda}(\tau), Q(\tau) \hat{\lambda}(\tau) \rangle \} d\tau$$

$$\text{subject to } \frac{d\hat{\lambda}(\tau)}{d\tau} = A^T \hat{\lambda}(\tau) - C^T \hat{v}(\tau) \quad \hat{\lambda}(\tau_1) = e_i \quad \tau_1 < \tau_0$$

for which the minimum is clearly

$$J_i = \langle e_i, \hat{P}(\tau_1) e_i \rangle_{E_n} = \langle e_i, P(t_1) e_i \rangle_{E_n}$$

where $P(\cdot)$ is defined by the equation 8.2.12. An examination of the given analysis should indicate that $\lambda(t_1)$ previously set to e_i

could be set to any arbitrary vector $z \in \mathbb{R}^n$ from which it is deduced that $E[(\tilde{x}(t_1) - \hat{x}(t_1)) (\tilde{x}(t_1) - \hat{x}(t_1))^T] = P(t_1)$

thus defining the role of $P(t_1)$ as the covariance of the error $\tilde{x}(t_1) - \hat{x}(t_1)$.

8.2.2 An Output Structures Problem for Linear Stochastic Systems

The major result of the previous subsection was the equivalence between the estimation formulation and a deterministic minimization problem recalled here as

$$\begin{aligned} \text{Min } E \left[(\tilde{x}_i(t_1) - \hat{x}_i(t_1))^2 \right] &\equiv \text{Min}_{v \in H_r} \langle \lambda(t_0), P_0 \lambda(t_0) \rangle_{E_n} \\ &+ \int_{t_0}^{t_1} \{ \langle v(t), R(t) v(t) \rangle_{E_r} + \langle \lambda(t), Q(t) \lambda(t) \rangle_{E_n} \} dt \end{aligned}$$

$$\text{subject to } \dot{\lambda}(t) = -A^T \lambda(t) + C^T v(t) \quad \lambda(t_1) = e_i$$

$$\text{where } \hat{x}_i(t_1) = \int_{t_0}^{t_1} v^T(t) \tilde{y}(t) dt.$$

It has already been assumed that the system is time-invariant (although this was an unnecessary restriction on the preceeding analysis) and now further restrictions on the noise statistics are imposed to enable the development of a time invariant optimal estimator.

Assuming that matrices Q and R are constant consider limiting behaviour as $t_0 \rightarrow -\infty$. The matrix riccati equation 8.2.12 will under mild restrictions converge to a steady state solution P which is independent of the initial value P_0 . This steady state behaviour may be arrived at by taking a limiting boundary condition of the form $\lim_{t_0 \rightarrow -\infty} P(t_0) = 0$, and evaluating $P(t)$ for any t according to

equation 8.2.12. Effecting such a limit operation on the equivalent minimization the steady state estimation problem is given as

$$\lim_{t_0 \rightarrow \infty} E[(\tilde{x}_i(t_1) - \hat{x}_i(t_1))^2] = \min_v \int_{-\infty}^{t_1} \{ \langle \lambda(t), Q\lambda(t) \rangle_{E_n} + \langle v(t), Rv(t) \rangle_{E_r} \} dt$$

$$\text{subject to } \dot{\lambda}(t) = -A^T \lambda(t) + C^T v(t)$$

$$\lambda(t_1) = e_i$$

which is realized by the form

$$\frac{d\hat{x}_i(t_1)}{dt_1} = e_i^T \{ A \hat{x}(t_1) + K [C \hat{x}(t_1) - \tilde{y}(t_1)] \}$$

$$\text{where } K = -P(t_1) C^T R^{-1}.$$

The introduction of structures into this estimation problem is now straightforward since it has the mathematical form of the input structures problem. Introducing the constraint $v_s = 0$ for some subset of the inputs v has the effect of removing a subset of the output measurements on which estimation is based. Consequently the degradation on the optimal estimation cost can be assessed for the removal of these measurements. A situation may arise where the estimation of a particular state is of interest rather than the whole system state. The above structure procedure can be used to decide the minimum number of measurements required to estimate the individual state satisfactorily.

Performing a structures assessment for all the n states enables a decision to be taken on the minimum number of measurements required to estimate the whole state vector satisfactorily. This larger problem is considered to be the output structure problem for a linear stochastic system since recalling the estimator as:

$$\dot{\hat{x}}(t) = A \hat{x}(t) + K(t) [C \hat{x}(t) - \tilde{y}(t)]$$

the resulting structured estimator has the form

$$\dot{\hat{x}}(t) = A \hat{x}(t) + K_s(t) [C_s \hat{x}(t) - \tilde{y}_s(t)]$$

where the number of measurements has been reduced from r (the dimension of the full output vector) to s the dimension of $\tilde{y}_s(t)$, and consequently $y_s(t)$.

A number of observations are in order at this point. The first of which concerns the effect of input deletions in the adjoint state equation

$$\dot{\lambda}(t) = -A^T \lambda(t) + C^T v(t) \quad \lambda(t_1) = e_i$$

$$\text{and } -\infty < t < t_1$$

Since the equation is being integrated backwards in time the process is numerically stable. Controllability of the pair $(-A^T, C^T)$ corresponds to the observability of the pair (A, C) , hence the assessment of observability structure is under analysis here.

Without persisting with unnecessary detail most of the controllability theorems of Chapter 2 have a counterpart here for the pair $(-A^T, C^T)$. The same observation will apply to the asymptotic analysis of Chapter 1 except that the interpretation will in terms of the speed of reconstruction of the state since this will be determined by the balance of Q to R . Suppose

$R = \rho R_0$, then as $\rho \rightarrow 0$ the speed of reconstruction becomes faster (associated with faster closed loop estimator poles) unless of course the pair $(-A^T, C^T)$ has either a right half plane zero or an input decoupling zero in which case the speed of reconstruction will be limited. In the absence of exact values for the noise statistics Q and R , Kwakernaak and Sivan(1972b) suggest the use of the parameter ρ to obtain a satisfactory design.

It is also clear that computational requirements for a complete output structure analysis will depend on the parameter r (as the number of inputs) and the fact that there are n subproblems of the form:

$$\text{Min } J_i = \int_{-\infty}^{t_1} \{ \langle \lambda(t), Q \lambda(t) \rangle_{E_n} + \langle v(t), R v(t) \rangle_{E_r} \} dt$$

$$\text{subject to } \dot{\lambda}(t) = -A^T \lambda(t) + C^T v(t) \quad \lambda(t_1) = e_i$$

$$i=1,2,\dots,n.$$

Under the usual assumption $n \gg r$, the computational requirements are the same as those discussed for the removal of the initial condition in the deterministic problem (Chapter 6). Obviously there will be cases for which the procedure is not feasible but for those cases for which it is; a possible reduction in the number of measurements to be taken for estimation and control purposes seem such justifiable aims.

8.2.3 Controllers of Reduced Dimensions.

Kwakernaak and Sivan(1972b) formulated a design philosophy motivated by the structure of the optimal output feedback regulator for a linear stochastic system. Here the separation principle decoupled the problem into a linear estimator to reconstruct the state and an optimal feedback law operating on this reconstruction.

From this the problem of controllers of reduced dimension was defined for the system $\dot{x} = Ax + Bu + w_1$ $x(t_0) = x_0$

$$y = Cx + w_2$$

as choosing a controller of the form $\dot{q} = Fq + Gy$ $q(t_0) = q_0$

$$u = -Hq$$

to minimize $E \left[\int_{t_0}^{t_1} \{ \langle x, Qx \rangle + \langle u, Ru \rangle \} dt \right]$

The solution is obtained using a numerical non-linear minimization algorithm. However it is observed that the value of the criterion is determined only by its external representation

$H(sI-F)^{-1}G$ since it is well known that a given external representation has many internal representations. Kwakernaak and Sivan further remark that a considerable reduction in the number of free parameters can be achieved by the blocking of certain feedback paths of minor significance.

This is the imposition of structural constraints, a process for which algorithms of the input structure problem of deterministic systems have been developed. Applied to the optimal control problem the

number of system inputs can be reduced and applied to the linear stochastic system the number of measurements for estimation can be reduced. Using this reduced structure the order of the controller (that is dimension of q) can be obtained from the solution of the problem of controllers of reduced dimension to conclude the design process.

8.3 Conclusions

The general analysis for minimum variance estimators has been applied to the problem of estimating the state of a linear stochastic system. As in the general case an identification between a statement for estimation and a minimization of a quadratic cost functional is obtained. It is for the latter that a physically meaningful structures problem is posed. This concerns the choice of measurements to be used as the input to a state estimator. The separation principle decouples the estimation and control problems in this formulation so that the input structures technique may be used to decide the optimal control structure, and the optimal measurement structure which may then be taken together to give the external structure for a dynamic controller. Finally the algorithm of Kwakernaak and Siven (1972b) may be used to decide the appropriate controller order.

III.O Future Research

A major aim of this thesis was to present a satisfactory theoretical setting for a new aid to be utilized in the design of linear system controllers. Some practical experience with a realistic problem was also presented. However, all projects of this nature usually conclude with a number of research avenues as yet unexplored, and this project is no exception.

The asymptotic optimal control theory which opened this thesis still lacks a complete solution. One possibility here is that Geometric Analysis of Wonham and Morse may provide the necessary unifying framework for the known results.

The technique of Hadley used to obtain the bound formulae of Chapter 3 is certainly trustworthy but is very tedious especially for the higher order bounds. The investigation of alternative methodologies to obtain the bounds could bring dividends of even higher order bound formulae.

The asymptotic behaviour of the bounds still requires some clarification, this appears to be a most promising area for future research. Experience with the finite dimensional examples indicates some form of relationship between the order of the limiting lower bound value, and the limit value of the optimal cost, $J(\rho)$. What form this relation takes generally remains to be researched. The actual derivation of the asymptotic behaviour of the bounds is given in terms of particular examples

for which the analysis is tractable. It is felt that a more unified theory should be possible based on the finite dimensional optimization problem from which all the bounds are derived.

The application of the structures concept to other areas seems capable of considerable exploitation. The preliminary foray consisting of an example from the field of least squares fitting and some developments in pseudo-diagonalization serve to indicate the direction for future progress.

Similar remarks apply to the exposition on the output structures problem (Part II), although here further clarification of the problem statement is probably advisable before embarking on a prolonged research programme.

IV. REFERENCES

- Anderson, B.D.O. and Moore J.B., 1971,
Linear Optimal Control, Prentice-Hall, Inc., New Jersey
- Athans, M., 1972,
Automatica, Vol. 8. pgs 397-412.
- Bengtsson, G., 1973,
Ph.D. Thesis, Division of Automatic Control, Lund Institute,
Sweden.
- Bristol, E.H., 1966,
I.E.E.E. Trans on A.C. Vol AC-11, Pgs 133-134
- Bryant, G.F., 1973,
Ph.D. Thesis, Imperial College of Science and Technology,
London S.W.7.
- Bryant et al, 1969,
Industrial Automation Group Report 3/69,
Imperial College, London S.W.7.
- Bryant et al, 1973,
Automation of Tandem Mills,
The Iron and Steel Institute, London, S.W.1.
- Bryant et al., 1975,
Preprints 6th IFAC World Congress, Boston, Massachusetts, U.S.A.
Part IIC, Paper No. 53.3
- Bucy, R.S. and Joseph, P.D. (1968)
Filtering for Stochastic Processes with Applications to
Guidance, New York:Wiley 1968

Butterworth, S., 1930,

Wireless Engineer, Vol 7, Pg 536.

Caley, J.M. and Spooner, P.D. 1972,

Industrial Automation Group Report 11/72,

Imperial College, London, S.W.7.

Caley, J.M. and Spooner, P.D. 1973

Industrial Automation Group Report 4/73

Imperial College, London, S.W.7.

Canales, R. 1970,

I.E.E.E. Trans on A.C. Vol. AC-15, No.4,

Pgs 409-415

Chuang, S.C.

Control and Automation Section Report

No. 70/7, Dept. of Computing and Control,

Imperial College of Science and Technology,

London, S.W.7.

Davison, E.J., 1969,

Automatica Vol 5. pgs 791-799.

Davison, E.J. and Chow S.G. 1974,

8th Annual Princeton Conference on Information

Science and Systems, Pgs. 73-9.

Davison E.J. and Goldberg R.W. 1968.

IFAC Sympo. on Multivariable Control Systems

Pages 1-16, Dusseldorf.

Davison, E.J. and Man, F.T. 1970

Proc. IEE, Vol 117, No.2, Pgs 459-462.

Desoer, C.A., and Vidyasagar, M., 1975.

Feedback Systems: Input-Output Properties

Academic Press, New York, 1975.

Draper, M.R. and Smith, H. 1966.

Applied Regression Analysis : Wiley

Dressler, R.M. and Larson, R.E. 1968

Preprints of the 9th J.A.C. Conference, Pg 711.

Edwards, W.J. 1968

Industrial Automation Group Report 5/68,

Imperial College, London S.W.7.

Edwards et al., 1969,

Industrial Automation Group Report 11/69,

Imperial College, London S.W.7.

Faddeev, D.K., and Faddeeva, V.N. 1963

Computational Methods of Linear Algebra, W.H. Freeman and Co.

Foss, A.C. 1973,

IEEE Trans. A.C. Vol A.C.-18, Pg 646

Freeman, E.A. 1969.

Journal of Optimization Theory and Applications,

Vol 3, No.6, June 1969.

Hadley, G., 1970,

Nonlinear and Dynamic Programming,

Addison-Wesley Publ. Co.

Hannah, R., 1975

Ph.D. Thesis, School of Electrical Engineering.

University of New South Wales, New South Wales, Australia.

Hawkins, D.J., 1972

Proc. IEE. Vol 119, No.3., Pgs 337-342.

Herring, K.D. and Melsa, J.L. 1974

IEEE Trans on A.C. Vol AC-19,
No.6, Pgs 264-266.

Hsieh, H.C., 1965

Advances in Control, Vol 2 Pgs 117-208.

Jameson, A. 1970

Int. J of Control, Vol 11, No.3., Pgs 409-421

Kalman, R.E., 1964,

J. Basic Eng. Trans. A.S.M.E. Ser.D. Vol 86, Pg 51.

Kosut, R.L., 1970.

IEEE Trans on A.C. Vol AC-15, No.5,
Pgs 557-563

Kouvaritakis, B.A. 1972,

Design of Controllers for a Non Square Systems, M.Sc. Thesis,
U.M.I.S.T.

Kouvaritakis, B.A., 1974,

Ph.D. Thesis, University of Manchester

Institute of Science and Technology

Kouvaritakis, B.A., and MacFarlane, A.G.J., 1976a,

Int. J. Control, Vol 23, No.2., Pgs 149-166.

Kouvaritakis, B.A., and MacFarlane, A.G.J., 1976b,

Int.J. Control, Vol 23, No.2. Pgs 167-181

Kouvaritakis, B.A., and Shaked, U., 1976,

Int. J. Control, Vol. 23, No.2., Pg 297.

Krall, A.M. 1973.

Linear Methods of Applied Analysis, Addison-Wesley Pub. Co. Ltd.
Massachusetts.

Kwakernaak, H. 1976,

IEEE Trans. Auto. Control., Vol AC-21,
No. 3, Pgs 378-382.

Kwakernaak, H. and Sivan, R., 1971,

Proc. of 3rd Annual South-Eastern Symposium on Systems Theory,
Atlanta, U.S.A.

Kwakernaak, H. and Sivan, R., 1972a

IEEE Trans. A.C. Vol AC-17, No. 1, pg 79.

Kwakernaak, H. and Sivan, R., 1972b

Linear Optimal Control Systems,
Wiley-Interscience.

Lancaster, P., 1969,

Theory of Matrices, Academic Press, London.

Levine W.S. and Athans M. 1970

IEEE Transactions on Automatic Control AC-15, No.1,
Feb, (44-48), 1970.

Liusternik, L.A. and Sobolev, V.J. 1961

Elements of Functional Analysis
F. Ungar Publ. Co., New York.

Luenberger, D.G., 1969

Optimization of Vector Space Methods
John Wiley and Sons, Inc.,

Luenberger, D.G. 1973,

Introduction to Linear and Non-Linear Programming,
Addison-Wesley Publ. Co. Ltd.,

MacFarlane, A.G.J., 1973,

Notes on the Vector Frequency Response Approach to Multi-variable Feedback Systems, Dept. of Electrical Engineering and Electronics Rpt., UMIST.

MacFarlane, A.G.J., and Karcnias, N., 1976,

Poles and Zeros of Multivariable System, University of Cambridge Research Report No. CUED/ F-Control/TR 105

McClamroch, N.H. 1969,

IEEE Transactions on Automatic Control AC-14,
June, Pgs 282-285, 1969.

Mee, D.H., 1976,

Joint Automatic Control Conference

Paper No. Thry 9-5, Pgs 627-633.

Mellefont, D.J., and Sargent, R.W.H., 1977

Internal Report: 'Optimal Measurement Policies for Control',
Dept. of Chemical Engineering and Chemical Technology,
Imperial College.

Morse, A.S., and Wonham, W.M., 1971,

IEEE Trans on A.C., Vol AC-16, No.6, Pgs 568-581.

Naylor, A.W., and Sell, G.R., 1971,

Linear Operator Theory in Engineering and Science, Holt,
Rinehart and Winston, Inc., New York.

Niederlinski, A., 1971,

Automatica, Vol 7, pgs. 691-701.

Noble, B. 1969,

Applied Linear Algebra,

Prentice-Hall, Inc., New Jersey.

Owen, D.H. 1976,
 Research Report No. 50, Control Dept.,
 University of Sheffield, Sheffield, U.K.

Pritchard, A.J. and Parker, K.T.,
 J. Inst. Maths. Applics. 1974
 Vol 13 Pgs 97-106.

Rekasius, Z.V., 1967
 IEEE Trans. A.C. Vol AC-12, Pg 296

Rosenbrock, H.H., 1962
 Chem. Eng.-Progr., Vol 58, No.9, Pg 43.

Rosenbrock, H.H., 1966,
 Proc. 3rd I.F.A.C. Congress, London, Vol 1. Book 1, Paper 1A.

Rosenbrock, H.H., 1970.
 State Space and Multivariable Theory, Nelson-Wiley, London.

Rosenbrock, H.H., 1974,
 Computer-Aided Control System Design,
 published by Academic Press.

Schuldt, S.B. and Smith, F.B. 1971,
 Proc. Joint Automatic Control Conference,
 Paper No. 3-E4, Pgs 270-276.

Tanuma, M. 1974,
 Preprints 4th IFAC/IFIP International Conference on
 Digital Computer Application to Process Control,
 Zurich, Switzerland.

Tyler, J.S. and Tuteur, F.B. 1966,

IEEE Trans. AC Vol AC-11,

No. 1, Pg. 84.

Whittle, P. 1971

Optimization Under Constraints: Theory and Applications of
Non-Linear Programming. Wiley.

Riesz and Sz-Nagy , 1965

Functional Analysis,

Frederic Ungar Publishing Co.; 1965.

Appendix A A Theorem for the Smith Form of $\Phi^T(-s)\Phi(s)$ and Other Results.

Theorem If the polynomial matrix $\Phi(s)$ is $(r \times m)$ with normal

rank $r < m$ and Smith canonical form $\Phi(s) = P(s) \Gamma(s) Q(s)$

then the Smith canonical form for $\Phi^T(-s) \Phi(s) = P_1(s) \Gamma^T(-s) \Gamma(s) Q_1(s)$.

(Matrices $P(s), Q(s), P_1(s)$ and $Q_1(s)$ are all unimodular with compatible dimensions.)

Proof 1) $\Phi^T(-s) \Phi(s) = Q^T(-s) \Gamma^T(-s) P^T(-s) P(s) \Gamma(s) Q(s)$

2) The invariants of $\Phi^T(-s) \Phi(s)$ are the same as those of $\Gamma^T(-s) P^T(-s) P(s) \Gamma(s)$ since $Q(s)$ is unimodular.

3) Denote $D(s) = P^T(-s) P(s)$ where $P(s)$ is a $(r \times r)$ unimodular matrix then (a) $D(s)$ is unimodular and

$$(b) \quad d_{ij}(s) = d_{ji}(-s) \quad \forall (i, j) \quad i = 1, \dots, r; j = 1, \dots, r.$$

$$4) \quad \Gamma^T(-s) D(s) \Gamma(s) = \begin{bmatrix} E(s) & 0 \\ 0 & 0 \end{bmatrix}$$

$$\text{where } E_{ij}(s) = \Gamma_i(-s) d_{ij}(s) \Gamma_j(s) \quad i=1, \dots, r; j=1, \dots, r$$

and $\Gamma_i(s)$ are the invariant polynomials of $\Phi(s)$

$$\text{hence } \Gamma_i(s) / \Gamma_{i+1}(s) \quad i=1, \dots, r-1.$$

The invariants of the block matrix $E(s)$ are now constructed:

5) Define δ_k to be the greatest common divisor of all

k rowed co-factors of $E(s)$, $k=1, 2, \dots, \bar{r} \leq r$. The invariants are given by the ratios $\delta_1, \delta_2/\delta_1, \dots, \delta_{\bar{r}}/\delta_{\bar{r}-1}$.

$$k = 1 \quad \delta_1 = \Gamma_1(-s) \Gamma_1(s)$$

$$\text{Since a) } \Gamma_1(s) \mid \Gamma_2(s) \mid \dots \mid \Gamma_r(s)$$

$$b) \quad \Gamma_1(-s) \mid \Gamma_2(-s) \mid \dots \mid \Gamma_r(s)$$

c) If there exists a common factor in all the elements

of $D(s) \Rightarrow \det[D(s)]$ is not independent of s and therefore is not unimodular, contradicting property 3).

$$k = 2 \quad \delta_2 = \Gamma_1(s) \Gamma_2(s) \Gamma_1(-s) \Gamma_2(-s)$$

Since a) as for a) and b) in $k = 1$ above

$$b) \quad \Gamma_2(s) \mid \Gamma_3(s) \mid \dots \mid \Gamma_r(s) \quad \text{for } s \text{ and } (-s)$$

c) Each two rowed cofactor is of the form

$$\begin{vmatrix} E_{i_1 j_1}(s) & E_{i_1 j_2}(s) \\ E_{i_2 j_1}(s) & E_{i_2 j_2}(s) \end{vmatrix} = \begin{vmatrix} \Gamma_{i_1}(-s) d_{i_1 j_1} \Gamma_{j_1}(s) & \Gamma_{i_1}(-s) d_{i_1 j_2} \Gamma_{j_2}(s) \\ \Gamma_{i_2}(-s) d_{i_2 j_1} \Gamma_{j_1}(s) & \Gamma_{i_2}(-s) d_{i_2 j_2} \Gamma_{j_2}(s) \end{vmatrix} \\ = \Gamma_{i_1}(-s) \Gamma_{i_2}(-s) \Gamma_{j_1}(s) \Gamma_{j_2}(s) \begin{vmatrix} d_{i_1 j_1} & d_{i_1 j_2} \\ d_{i_2 j_1} & d_{i_2 j_2} \end{vmatrix}$$

Hence if every two rowed co-factor contributes a common factor of s to δ_2 , this will contradict the unimodularity of $D(s)$.

The arguments for $k=2,3,\dots,r$ follow similar lines giving the i th

$$\text{determinantal divisor as: } \delta_i = \prod_{j=1}^i \Gamma_j(-s) \Gamma_j(s).$$

This enables the invariants of $E(s)$ to be calculated as :

$$I_1(s) = \delta_1 = \Gamma_1(-s) \Gamma_1(s)$$

$$I_2(s) = \delta_2 / \delta_1 = \Gamma_2(-s) \Gamma_2(s)$$

$$\vdots$$

$$I_r(s) = \delta_r / \delta_{r-1} = \Gamma_r(-s) \Gamma_r(s)$$

Hence the rank of $\Phi^T(-s) \Phi(s)$ is r , and its invariants are

$$\{ \Gamma_1(-s) \Gamma_1(s), \Gamma_2(-s) \Gamma_2(s), \dots, \Gamma_r(-s) \Gamma_r(s) \}$$

giving a Smith canonical form $\Phi^T(-s) \Phi(s) = P_1(s) \Gamma^T(-s) \Gamma(s) Q_1(s)$

This theorem finds application in the asymptotic optimal control theory of Chapter 1 where it is used in analysis based on the optimal return

difference relation. In particular it is used for the re-writing of $G^T(-s) G(s)$. Suppose the system transfer function has Smith form :

$$G(s) = \frac{P(s)}{d(s)} \Gamma(s) \frac{Q(s)}{d(s)} = \frac{\phi(s)}{d(s)} \quad \text{then}$$

$$\begin{aligned} G^T(-s) G(s) &= \frac{\phi^T(-s)}{d(-s)d(s)} \phi(s) = \frac{P_1(s)}{d(-s)d(s)} \Gamma^T(-s) \Gamma(s) \frac{Q_1(s)}{d(-s)d(s)} \\ &= P_1(s) \psi^T(-s) \psi(s) Q_1(s) \end{aligned}$$

where $\psi(s)$ is the Smith McMillan kernel of $G(s)$.

For the case of square systems it is possible to give a more general form of the theorem proved above.

Theorem (For Square Matrices) If the polynomial matrix $\phi(s)$ is $(r \times r)$, with normal rank r and Smith canonical form

$$\phi(s) = P(s) \Gamma(s) Q(s), \text{ then}$$

$$\left[\phi^T(-s) \phi(s) \right]^{\ell} = P_1(s) \left[\Gamma^T(-s) \Gamma(s) \right]^{\ell} Q_1(s), \quad \ell = 1, 2, \dots,$$

Matrices $P(s)$, $Q(s)$, $P_1(s)$ and $Q_1(s)$ are all square and unimodular.

The proof of this theorem uses an inductive argument for which the actual detail is very similar to that given in the theorem for the non-square matrices above. For this reason the proof is omitted.

To facilitate the assimilation of the recursive procedure to describe the asymptotic behaviour of the optimal closed loop poles much of the detail was omitted from Section 1.3. The derivation of the $(j+1)$ th reduced return difference determinantal relation is given here in its entirety.

$$\text{Recall } RD_j(s, \mu) \triangleq \det \begin{bmatrix} B_{11}(s) & B_{12}(s) \\ B_{12}^T(-s) & B_{22}(s) \end{bmatrix}$$

where (i) $B_{11}(s) = I_{d_{j-1}} + \frac{(-1)^j \mu}{s^{2j}} \{ \Lambda_{d_{j-1}} + B_{11}(1)1/s + B_{11}(2)1/s^2 + \dots \}$

with $B_{11}(1) = T_{j-1}^T(1) (A_j A_{j-1}^T - A_{j-1} A_j^T) T_{j-1}(1)$

and $B_{11}(2) = T_{j-1}^T(1) (A_{j+1} A_{j-1}^T - A_j A_j^T + A_{j-1} A_{j+1}^T) T_{j-1}(1)$

(ii) $B_{12}(s) = \frac{(-1)^{j+1} \mu}{s^{2j+1}} \{ B_{12}(0) + B_{12}(1)1/s + B_{12}(2)1/s^2 + \dots \}$

with $B_{12}(0) = T_{j-1}^T(1) A_{j-1} A_j^T T_{j-1}(2)$

$B_{12}(1) = T_{j-1}^T(1) (A_j A_j^T - A_{j-1} A_{j+1}^T) T_{j-1}(2)$

and $B_{12}(2) = T_{j-1}^T(1) (A_{j+1} A_j^T - A_j A_{j+1}^T + A_{j-1} A_{j+1}^T) T_{j-1}(2)$

(iii) $B_{22}(s) = I_{q_j} + \frac{(-1)^{j+1} \mu}{s^{2j+2}} \{ B_{22}(0) + B_{22}(1)1/s + B_{22}(2)1/s^2 + \dots \}$

with $B_{22}(0) = T_{j-1}^T(2) A_j A_j^T T_{j-1}(2)$

$B_{22}(1) = T_{j-1}^T(2) (A_{j+1} A_j^T - A_j A_{j+1}^T) T_{j-1}(2)$

and $B_{22}(2) = T_{j-1}^T(2) (A_{j+1} A_j^T - A_{j+1} A_{j+1}^T + A_j A_{j+2}^T) T_{j-1}(2)$

Consequently using Schur's Identity to factor out $|B_{11}(s)|$

$RD_{j+1}(s, \mu) \triangleq \det [B_{22}(s) - B_{12}^T(-s) B_{11}^{-1}(s) B_{12}(s)]$

Define $\hat{B}_{11}(s) \triangleq B_{11}^{-1}(s) = \frac{(-1)^j \mu s^{2j}}{\mu} \left[\hat{B}_{11}(0) + \hat{B}_{11}(1) \frac{1}{s} + \hat{B}_{11}(2) \frac{1}{s^2} + \dots \right]$

Recall $B_{11}(s) = I_{d_{j-1}} + \frac{(-1)^j \mu}{s^{2j}} \{ \Lambda_{d_j} + B_{11}(1) \frac{1}{s} + B_{11}(2) \frac{1}{s^2} + \dots \}$

$= I_{d_{j-1}} + \frac{(-1)^j \mu}{s^{2j}} X(s) = \frac{(-1)^j \mu X(s)}{s^{2j}} \left[I_{d_{j-1}} + \frac{(-1)^j \mu s^{2j}}{\mu} X^{-1}(s) \right]$

where $X(s) = \Lambda_{d_j} + B_{11}(1) \frac{1}{s} + \dots$ is convergent since $|s| \rightarrow \infty$.

$$B_{11}^{-1}(s) = \left[\left(\frac{(-1)^j \mu}{s^{2j}} \right) X(s) \left(I_{d_{j-1}} + \frac{(-1)^j s^{2j}}{\mu} X^{-1}(s) \right) \right]^{-1}$$

$$= \frac{(-1)^j s^{2j}}{\mu} \left[I_{d_j} + \frac{(-1)^j s^{2j}}{\mu} X^{-1}(s) \right]^{-1} \cdot X^{-1}(s)$$

Hence for all rates $\mu = o(|s|^k)$ $k > 2j+1$ a series expansion is valid to give

$$B_{11}^{-1}(s) = \frac{(-1)^j s^{2j}}{\mu} X(s)^{-1} - \frac{(-1)^{2j} s^{4j}}{\mu^2} X(s)^{-2} + \dots$$

The inversion of $X(s) = \Lambda_{d_{j-1}} + B_{11}(1) \frac{1}{s} + B_{11}(2) \frac{1}{s^2} + \dots$

is obtained next, to give the leading term of $\hat{B}_{11}(s)$.

$$X(s) = \Lambda_{d_{j-1}} \left[I_{d_{j-1}} + \Lambda_{d_{j-1}}^{-1} Z(s) 1/s \right]$$

where $Z(s) = B_{11}(1) + B_{11}(2) 1/s + B_{11}(3) 1/s^2 + \dots$

$$\begin{aligned} \text{Hence } X(s)^{-1} &= \left[I_{d_{j-1}} + \Lambda_{d_{j-1}}^{-1} Z(s) 1/s \right]^{-1} \Lambda_{d_{j-1}}^{-1} \\ &= \Lambda_{d_{j-1}}^{-1} - \Lambda_{d_{j-1}}^{-1} Z(s) \Lambda_{d_{j-1}}^{-1} 1/s \\ &\quad + \Lambda_{d_{j-1}}^{-1} Z(s) \Lambda_{d_{j-1}}^{-1} Z(s) \Lambda_{d_{j-1}}^{-1} 1/s^2 - \dots \\ &= \Lambda_{d_{j-1}}^{-1} - \Lambda_{d_{j-1}}^{-1} B_{11}(1) \Lambda_{d_{j-1}}^{-1} 1/s \\ &\quad - \Lambda_{d_{j-1}}^{-1} (B_{11}(2) - B_{11}(1) \Lambda_{d_{j-1}}^{-1} B_{11}(1)) \Lambda_{d_{j-1}}^{-1} 1/s^2 + \dots \end{aligned}$$

$$\text{Comparing } \hat{B}_{11}(s) = \frac{(-1)^j s^{2j}}{\mu} \{ \hat{B}_{11}(0) + \hat{B}_{11}(1) 1/s + \hat{B}_{11}(2) 1/s^2 + \dots \}$$

$$\text{with } \hat{B}_{11}(s) = \frac{(-1)^j s^{2j}}{\mu} X(s)^{-1} - \dots$$

$$= \frac{(-1)^j s^{2j}}{\mu} \{ \Lambda_{d_{j-1}}^{-1} - \Lambda_{d_{j-1}}^{-1} B_{11}(1) \Lambda_{d_{j-1}}^{-1} 1/s + \dots \}$$

leads to the following identifications

$$\hat{B}_{11}(0) = \Lambda_{d_{j-1}}^{-1}, \quad \hat{B}_{11}(1) = -\Lambda_{d_{j-1}} B_{11}(1) \Lambda_{d_{j-1}}^{-1}$$

$$\text{and } \hat{B}_{11}(2) = -\Lambda_{d_{j-1}}^{-1} (B_{11}(2) - B_{11}(1) \Lambda_{d_{j-1}}^{-1} B_{11}(1)) \Lambda_{d_{j-1}}^{-1}.$$

With these terms identified, recall from section 1.3 the expression

for $B_{12}^T(-s) \hat{B}_{11}(s) B_{12}(s)$ to obtain

$$\begin{aligned} B_{22}(s) - B_{12}^T(-s) \hat{B}_{11}(s) B_{12}(s) &= I_{q_j} + \frac{(-1)^{j+1} \mu}{s^{j+2}} \{ B_{22}(0) - B_{12}^T(0) \hat{B}_{11}(0) B_{12}(0) \\ &+ (B_{22}(1) + B_{12}^T(1) \hat{B}_{11}(0) B_{12}(0) - B_{12}^T(0) \hat{B}_{11}(0) B_{12}(1) - B_{12}^T(0) \hat{B}_{11}(1) B_{12}(0)) \frac{1}{s} \\ &+ (B_{22}(2) - B_{12}^T(0) \hat{B}_{11}(2) B_{12}(0) - B_{12}^T(0) \hat{B}_{11}(1) B_{12}(1) - B_{12}^T(0) \hat{B}_{11}(0) B_{12}(2) \\ &+ B_{12}^T(1) \hat{B}_{11}(1) B_{12}(0) + B_{12}^T(1) \hat{B}_{11}(0) B_{12}(1) - B_{12}^T(2) B_{11}(0) B_{12}(0)) \frac{1}{s^2} + \dots \} \end{aligned}$$

Using the equivalences for the terms of $B_{12}(s)$, $B_{11}(s)$, $\hat{B}_{11}(s)$ and

$B_{22}(s)$ this expression is simplified as follows :

$$\begin{aligned} B_{22}(0) - B_{12}^T(0) \hat{B}_{11}(0) B_{12}(0) &= T_{j-1}^T(2) A_j A_j^T T_{j-1}(2) \\ &- T_{j-1}^T(2) A_j A_j^T T_{j-1}(1) \Lambda_{d_{j-1}}^{-1} T_{j-1}^T(1) A_{j-1} A_{j-1}^T T_{j-1}(2) \\ &= T_{j-1}^T(2) A_j \left[I - A_{j-1}^T T_{j-1}(1) \Lambda_{d_{j-1}}^{-1} T_{j-1}^T(1) A_{j-1} \right] A_j^T T_{j-1}(2) \end{aligned}$$

$$\text{Define } P_{j-1} \triangleq I - A_{j-1}^T T_{j-1}(1) \Lambda_{d_{j-1}}^{-1} T_{j-1}^T(1) A_{j-1}$$

where clearly P_{j-1} is idempotent, and define

$$\tilde{A}_j = T_{j-1}^T(2) A_j P_{j-1} \quad \text{then } B_{22}(0) - B_{12}^T(0) \hat{B}_{11}(0) B_{12}(0) = \tilde{A}_j \tilde{A}_j^T.$$

Taking the second term $B_{22}(1) + B_{12}^T(1) \hat{B}_{11}(0) B_{12}(0) -$

$$B_{12}^T(0) \hat{B}_{11}(0) B_{12}(1) - B_{12}^T(0) \hat{B}_{11}(1) B_{12}(0) = T_{j-1}^T(2) (A_{j+1} A_j^T - A_j A_{j+1}^T) T_{j-1}(2)$$

$$\begin{aligned}
& + T_{j-1}^T (2) (A_j A_j^T - A_{j+1} A_{j-1}^T) T_{j-1} (1) \Lambda_{d_{j-1}}^{-1} T_{j-1}^T (1) A_{j-1} A_j^T T_{j-1} (2) \\
& - T_{j-1}^T (2) A_j A_{j-1}^T T_{j-1} (1) \Lambda_{d_{j-1}}^{-1} T_{j-1}^T (1) (A_j A_j^T - A_{j-1} A_{j+1}^T) T_{j-1} (2) \\
& + T_{j-1}^T (2) A_j A_j^T T_{j-1} (1) \Lambda_{d_{j-1}}^{-1} T_{j-1}^T (1) (A_j A_{j-1}^T - A_{j-1} A_j^T) T_{j-1} (1) \Lambda_{d_{j-1}}^{-1} T_{j-1}^T (1) \\
& \quad A_{j-1} A_j^T T_{j-1} (2) \\
& = (T_{j-1}^T (2) A_{j+1} - T_{j-1}^T (2) A_j A_{j-1}^T T_{j-1} (1) \Lambda_{d_{j-1}}^{-1} T_{j-1}^T (1) A_j) P_{j-1} A_j^T T_{j-1} (2) \\
& \quad - T_{j-1}^T (2) A_j P_{j-1} (-A_j^T T_{j-1} (1) \Lambda_{d_{j-1}}^{-1} T_{j-1}^T (1) A_{j-1} A_j^T T_{j-1} (2) + A_{j+1}^T T_{j-1}^T (2)) \\
\end{aligned}$$

Define $\tilde{A}_{j+1} = (T_{j-1}^T (2) A_{j+1} - T_{j-1}^T (2) A_j A_{j-1}^T T_{j-1} (1) \Lambda_{d_{j-1}}^{-1} T_{j-1}^T (1) A_j) P_{j-1}$

then 2nd term = $\tilde{A}_{j+1} \tilde{A}_j^T - \tilde{A}_j \tilde{A}_{j+1}^T$.

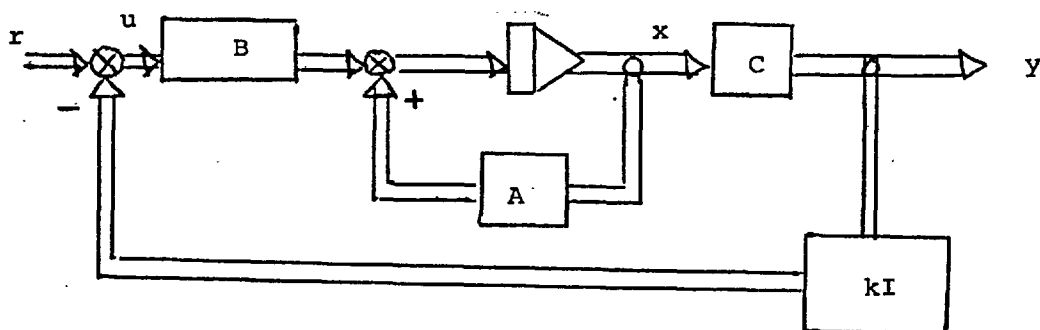
Simplification of the third and succeeding terms appears infeasible.

Recalling once more the (j+1)th reduced optimal return difference determinantal expression the above simplifications may be introduced to give :

$$RD_{j+1}(s, \mu) = \det \left[I_{q_j} + \frac{(-1)^{j+1} \mu}{s^{2j+2}} \{ \tilde{A}_j \tilde{A}_j^T + (\tilde{A}_{j+1} \tilde{A}_j^T - \tilde{A}_j \tilde{A}_{j+1}^T) \frac{1}{s} + \dots \} \right]$$

It remains to discuss the particular results given by the Corollary 1.3 which were obtained by setting $j=1$, and performing the first step of the recursive procedure.

Kou varitakis and Shaked (1976) described the asymptotic root loci of square linear multivariable systems using a feedback of the form $-kI$, where $k > 0$ is a scalar parameter. The configuration is depicted below :



Using the parameter k , the analysis is based on the asymptotic behaviour of the zeros of the closed loop characteristic polynomial, and the related return difference expression. In particular the 1st and 2nd order infinite zeros are given as follows : Assume the first Markov parameter $M_0 = CB$ has simple structure, with rank deficiency q_1 . Then CB has a spectral decomposition

$$CB = \begin{bmatrix} U^{(1)} & M^{(1)} \end{bmatrix} \begin{bmatrix} \Lambda_{d_0} & 0 \\ 0 & 0_{q_1} \end{bmatrix} \begin{bmatrix} V^{(1)} \\ N^{(1)} \end{bmatrix}$$

where $U^{(1)}$, Λ_{d_0} and $V^{(1)}$ are the eigenvector, eigenvalue and

reciprocal eigenvector matrices associated with the non-zero eigenvalues of CB . Matrices $M^{(1)}$ and $N^{(1)}$ complete the dual frames and satisfy

$CBM^{(1)} = 0$, and $N^{(1)}CB = 0$. The first order patterns are given by the determinantal identity $\det[I_{d_0} + k/s \Lambda_{d_0}] = 0$, thus :

i) If $d_0 = r$ ($=m$), then there are r 1st order patterns and exactly $(n-r)$ finite zeros.

ii) If $d_0 < r$, then there are d_0 1st order patterns and higher order patterns may be defined. Let $q_1 = (r-d)$ and define

$d_1 = \text{rank} [N^{(1)}_{CABM}^{(1)}]$. Now if $d_1 = q_1$ then the 2nd order

patterns are defined via the zeros of $\det[I_{d_1} + k/s^2 N^{(1)}_{CABM}^{(1)}]$

If $d_1 < q_1$ then there will be d_1 2nd order patterns and further steps are required to isolate the 3rd and higher order patterns.

The salient features of the above analysis are that

$$(i) \quad d_0 \triangleq \text{rank}[CB] \quad \text{and} \quad (ii) \quad d_1 \triangleq \text{rank}[N^{(1)} CABM^{(1)}]$$

give the numbers of the 1st and 2nd order infinite zeros. In the asymptotic optimal closed loop behaviour, the results were

$$(i) \quad d_0 = \text{rank}[M_0 \triangleq CB] \quad \text{1st Order Stable Butterworth patterns and}$$

$$(ii) \quad d_1 = \text{rank}[P_0 M_1^T T_{o2}] \quad (\text{where } M_1 \triangleq CAB) \quad \text{2nd order stable Butterworth patterns. That the same number of infinite zeros are specified by both approaches results from the following equivalence.}$$

Theorem Using the above notation for $N^{(1)}$, $M^{(1)}$, P_0 , M_1

$$\text{and } P_0 = I - M_0^T T_{o1} \Lambda_d^{-1} T_{o1}^T M_0$$

$$\text{rank}[N^{(1)} CAB M^{(1)}] \equiv \text{rank}[T_{o2}^T M_1 P_0]$$

Proof For the case of square systems $N^{(1)}$ is $d_1 \times r$, and satisfies

$$N^{(1)} CB = 0 \quad \text{with} \quad \text{rank}[N^{(1)}] = d_1 < r$$

$$\text{Define } Z = [N^{(1)}]^T \quad \text{then} \quad [CB]^T Z = 0$$

Recall also $M_0^T T_{o2} = 0$ where $M_0 \triangleq CB$, and the columns of T_{o2} are ortho-normal. A column of Z , denoted $Z(i)$, may be written in terms of the columns of T_{o2} , denoted $t_{o2}(j)$, so that

$$Z(i) = \sum_{j=1}^{d_1} \alpha_{ij} t_{o2}(j) = T_{o2} \alpha_i$$

and $Z = T_{o2} \alpha$ where T_{o2} is $r \times d_1$ and α is $d_1 \times d_1$.

Thus $\text{rank}[Z] = d_1 = \text{rank}[T_{O2} \alpha] < \min \{ \text{rank}[T_{O2}] \text{rank}[\alpha] \}$

hence $\text{rank}[\alpha] = d_1$ and α is non singular.

The columns of $M^{(1)} \in \mathcal{N}(M_O = CB)$, and are also linearly independent

by definition. Projection operator $P_O = I_r - M_O^T T_{O1} \Lambda_d^{-1} T_{O1}^T M_O$

is such that $\mathcal{R}(P_O) = \mathcal{N}(M_O)$ so that $P_O m_O(i) = m_O(i)$ where $i=1, \dots, d_1$.

But $\mathbb{R}^r = \mathcal{N}(M_O) \oplus \mathcal{N}(M_O)^\perp$, consequently there exists $\xi(i)$ $i=1, \dots, r-d_1$

such that $[M^{(1)} \xi]$ is non-singular and $P_O \xi(i) = 0$, $i=1, \dots, r-d_1$.

Noting that the rank of a matrix remains unchanged if it is pre- or post-multiplied by a nonsingular matrix

$$\begin{aligned} \text{rank}[T_{O2}^T M_1 P_O] &= \text{rank} \left[[\alpha^T]^{-1} N^{(1)} M_1 P_O [M^{(1)} \xi] \right] \\ &= \text{rank} \left[N^{(1)} M_1 [P_O M^{(1)}, 0] \right] \\ &= \text{rank} \left[N^{(1)} M_1 M^{(1)} \right] \\ &= \text{rank} \left[N^{(1)} CAB M^{(1)} \right] \end{aligned}$$

This concludes the proof.

It is worth remarking that although for the same system parameters the number of infinite zeros are identical, the actual Butterworth patterns are not. The patterns associated with the use of an LQP problem are always stable, but those for the diagonal feedback $[-kI]$ exhibit the properties of the classical root locus technique of Evans, namely that unstable branches are possible.

Appendix B Some Linear Space Concepts Relevant to Chapter Two

In this appendix several linear space concepts relevant to Chapter 2 are discussed and an alternative derivation for the theorem on the asymptotic behaviour of $\Delta J_0^f(\rho) - \Delta J_0^s(\rho)$ is also given. The major assumption of the analysis of Chapter 2 is that the operator $D = WW^*$, comprising of the system, and adjoint system operator is compact. This enables the eigenstructure of compact self adjoint operators between Hilbert spaces to be used to decompose the output space H_r , and relate part of a disturbance, $d \in H_r$, to $\overline{\mathcal{R}(W)}$ and part to the orthogonal complement $\mathcal{N}(W^*)$.

Compactness for the linear operator $L : X \rightarrow Y$ where X , and Y are Banach (complete, normed) spaces is the condition that the sequence $\{Lx_n\}$, $Lx_n \in Y$, where $\{x_n\}$, $x_n \in X$ is a bounded sequence, contains a convergent subsequence $\{Lx_{n_k}\}$. Hsieh (1965) detailed the use of this assumption in his function space approach to the synthesis of adaptive control systems. The operator $D = WW^*$ given by:

$$(Dy)(t) = \int_0^t W(t-\tau) \int_t^{T_f} W^T(s-\tau) y(s) ds d\tau \quad 0 \leq t \leq T_f$$

is recast as the integral operator

$$(Dy)(t) = \int_0^{T_f} K(t,s) y(s) ds$$

by suitably defining the operator kernel. Standard results in functional analysis (Liusternik and Sobolev, 1961) are then utilized to give compactness for the operator D if

$$\int_0^{T_f} \int_0^{T_f} \|K(t,s)\|_{E_m}^2 dt ds < \infty. \text{ Consequently the system impulse}$$

response matrix $W(\cdot)$ may be used to infer the compactness of the operator D .

After Hsieh (1965), the eigenstructure of compact self adjoint operators is used to partition the output space as $H_T = H_1 \oplus H_2$ where $H_2 \triangleq N(D)$ and $H_1 = H_2^\perp$. It is then straightforward to use the identification $N(D) = N(W^*)$ and the relationship $N(W^*)^\perp = \overline{R(W)}$ (Luenberger, 1969) to obtain $H_T = \overline{R(W)} \oplus N(W^*)$; a decomposition which forms the basis of the asymptotic analysis for $\hat{J}(\rho)$, $\Delta J_o(\rho)$ and $IESQ(\rho)$.

The asymptotic analysis for the structure assessment procedure is based on a further decomposition of the output space, in particular $\overline{R(W)}$. This decomposition is underpinned by some fundamental concepts from Hilbert space theory.

Definition 2.1 (Luenberger, 1969) Let P be a subset of a normed space X , then a point $x \in X$ is said to be a closure point of the subset P if given $\eta > 0$ there is a point $p \in P$ satisfying $\|x - p\| < \eta$.

The collection of all closure points of P is called the closure of P and is denoted $\bar{P}$. It is clear $P \subset \bar{P}$.

Definition 2.2 (Luenberger, 1969) A set P is said to be dense in a normed space X if for each $x \in X$ and each $\eta > 0$ there exists $p \in P$ with $\|x - p\| < \eta$.

The relationship between closure, and denseness is best expressed by the following alternative definition:

Definition 2.3 (Naylor and Sell, 1971) A set P in normed space X is said to be dense in X if the closure of P is X , that is $\bar{P} = X$. Finally the relationship between closure and completeness is given by:

Lemma 2.1 (Naylor and Sell, 1971) Let (X, ρ) be a complete metric space, and let (Y, ρ) be a subspace of (X, ρ) . Then the subspace (Y, ρ) is complete iff Y is a closed set in (X, ρ) . This naturally implies that (Y, ρ) can then be treated as a complete metric space.

This lemma is applied to the closed linear subspace $\overline{\mathcal{R}(W)} \subseteq H_r$ to be able to assert that it may be treated as an individual Hilbert space.

Then applying definition 2.3 it is clear that $\mathcal{R}(W)$ is dense in its closure, a fact used to prevent some of the analysis in part of section 2.1 from becoming unduly cumbersome.

The partition of the system operator W , as $W = [W_s, W_{m-s}]$ naturally arises in the discussion of structure assessment. An initial investigation for the decomposition of $\mathcal{R}(W)$ pursued the relationships between $\mathcal{R}(W)$, $\mathcal{R}(W_s)$ and $\mathcal{R}(W_{m-s})$. The following were obtained:

Theorem 2.1 $\mathcal{R}(W) \equiv \mathcal{R}(W_s) + \mathcal{R}(W_{m-s})$

Theorem 2.2 $\overline{\mathcal{R}(W)} \equiv \overline{\mathcal{R}(W_s)} + \overline{\mathcal{R}(W_{m-s})}$

Although a decompositions for $\overline{\mathcal{R}(W)}$ in these terms is obtained, it is lacking in orthogonality, and is complicated by the effects of the concept of closeness. In particular by the fact that a sum of two closed linear subspaces of a Hilbert space is not necessarily closed (Naylor and Sell, 1971). A requirement which removed this drawback is that the two subspaces are orthogonal.

Introduce three lemmas which establish an orthogonal decomposition for $\overline{\mathcal{R}(W)}$:

Lemma 2.2 (Naylor and Sell, 1971) Let M be a set in an inner product space X . Then $M^\perp$ is a closed linear subspace of X .

Lemma 2.3 (Naylor and Sell, 1971) Let M be any closed linear subspace of a Hilbert space H . Then $H = M \oplus M^\perp$. Moreover each $x \in H$ can be expressed uniquely as $x = m + n$, where $m \in M$ and $n \in M^\perp$, and

$$||x||^2 = ||m||^2 + ||n||^2$$

Lemma 2.4 $\overline{\mathcal{R}(W_s)} \subseteq \overline{\mathcal{R}(W)}$

Proof Let $y \in \overline{\mathcal{R}(W_s)}$ then $y \in H_r$, and for given $\eta > 0$

$$\exists p \in \mathcal{R}(W_s) \text{ such that } ||y - p|| < \eta$$

But $p \in \mathcal{R}(W_s)$ implies $p \in \mathcal{R}(W)$ hence $y \in \overline{\mathcal{R}(W)}$.

Thus $\overline{\mathcal{R}(W_S)}$ is a closed linear subspace of $\overline{\mathcal{R}(W)}$ where the latter may be regarded as a Hilbert space. By lemma 2.3, there is an orthogonal decomposition for $\overline{\mathcal{R}(W)}$ of the form $\overline{\mathcal{R}(W)} = \overline{\mathcal{R}(W_S)} \oplus X$ where X is interpreted as the orthogonal complement of $\overline{\mathcal{R}(W_S)}$ restricted to $\overline{\mathcal{R}(W)}$. Further, by lemma 2.2 linear subspace X is closed and finally $H_T = \overline{\mathcal{R}(W_S)} \oplus X \oplus N(W^*)$ from which $\overline{\mathcal{R}(W_S)}^\perp$ may be identified as $N(W_S^*) = X \oplus N(W^*)$. With this decomposition for H_T , the analysis for the structure assessment procedure is straightforward and has an appealing feature of a geometric analogy given in section 2.2

The final part of this appendix is taken up by an alternative proof of the theorem for asymptotic structure assessment using ΔJ_0 . Initially several properties and inequalities are recalled:

Lemma 2.5 If $\hat{u}(\rho)$ is the optimal control as a function of the asymptotic variable $\rho > 0$, then $\lim_{\rho \rightarrow 0} \rho \langle \hat{u}(\rho), \hat{u}(\rho) \rangle_{H_m} = 0$.

Proof See Corollary 2.1.1 of Chapter 2, Section 2.1

Lemma 2.6 If operator A is self adjoint and positive-semidefinite then $\exists m$, and M such that $0 \leq m \leq \frac{\langle x, Ax \rangle}{\langle x, x \rangle} \leq M$.

Proof Liusternik and Sobolev, 1961

Lemma 2.7 If operator $A = W(\rho I + W^*W)^{-1}W^*$ then

$$\frac{\langle x, Ax \rangle}{\langle x, x \rangle} \leq M_1(\rho) \leq 1 \quad \text{with} \quad \lim_{\rho \rightarrow 0} M_1(\rho) = 1$$

Proof Use Lemma 2.1.1 Chapter 2 Section 2.1 to obtain

$$\frac{\langle x, W(\rho I + W^*W)^{-1}W^*x \rangle}{\langle x, x \rangle} = \frac{\langle x, (\rho I + WW^*)^{-1}WW^*x \rangle}{\langle x, x \rangle}$$

Assume WW^* is a compact self adjoint positive semi definite operator so that $\{\lambda_i\}_{i=1}^{\infty}$ are its denumerable eigenvalues, $\{e_i\}_{i=1}^{\infty}$ are the associated ortho-normal eigenfunctions and the related decomposition for H_r is $H_r = H_1 \oplus H_2$ where $H_1 = \text{span}\{e_i\}_{i=1}^{\infty}$ and $H_2 = H_1^{\perp} = N(D)$.

Then any $x \in H_r$ may be written $x = \sum_{i=1}^{\infty} \alpha_i e_i + h_0$, $h_0 \in N(D)$.

$$\text{Giving } M_1(\rho) = \sup_{x \in H_m} \frac{\langle x, W(\rho I + W^*W)^{-1} W^* x \rangle}{\langle x, x \rangle}$$

$$= \sup_{\left\{ \begin{array}{l} \{\alpha_i\} \\ h_0 \end{array} \right\}} \left\{ \frac{\sum_{i=1}^{\infty} \alpha_i^2 \lambda_i / (\rho + \lambda_i)}{\sum_{i=1}^{\infty} \alpha_i^2 + \langle h_0, h_0 \rangle} \right\}$$

$$\text{Thus } \forall \rho > 0 \quad \left(\frac{\lambda_i}{\rho + \lambda_i} \right) < 1 \text{ implying } \sum_{i=1}^{\infty} \alpha_i^2 \left(\frac{\lambda_i}{\rho + \lambda_i} \right) < \sum_{i=1}^{\infty} \alpha_i^2$$

$$\therefore 0 < M_1(\rho) \leq 1 \text{ and } \lim_{\rho \rightarrow 0} M_1(\rho) = \sup_{\left\{ \begin{array}{l} \{\alpha_i\} \\ h_0 \end{array} \right\}} \left\{ \frac{\sum_{i=1}^{\infty} \alpha_i^2}{\sum_{i=1}^{\infty} \alpha_i^2 + \langle h_0, h_0 \rangle} \right\} = 1$$

Schwarz Inequality $|\langle x, y \rangle| \leq \|x\| \|y\|$

Lemma 2.8 If the output space H_r , is decomposed as $H_r = \overline{R(W)} \oplus N(W^*)$ and disturbance d as $d = d_1 + d_2$ where $d_1 \in \overline{R(W)}$ and $d_2 \in N(W^*)$. With $\hat{Z}(\rho) \triangleq W \hat{u}(\rho)$ have that $\lim_{\rho \rightarrow 0} (\hat{Z}(\rho) + d_1) = 0$

Proof See Lemma 2.2.2 of Chapter 2, Section 2.2.

Lemma 2.9 $\Delta J_o^f(\rho) - \Delta J_o^s(\rho) = \frac{1}{2} \langle g_s, (W^*W + \rho I)^{-1} g_s \rangle$ where $g_s = W^* \hat{y}_s(\rho) + \rho \begin{pmatrix} \hat{u}_s(\rho) \\ 0 \end{pmatrix}$ and $\hat{y}_s(\rho) = W_s \hat{u}_s + d$

Proof See Chapter 2, Section 2.2

Theorem 2.3 Asymptotic Structure Assessment with ΔJ_0

If the output space H_r , is orthogonally decomposed as

$$H_r = \overline{R(W_s)} \oplus X + N(W^*) \quad \text{with} \quad \overline{R(W)} = \overline{R(W_s)} \oplus X$$

and disturbance d partitioned accordingly as $d = d_1 + d_2 + d_3$

$d_1 \in \overline{R(W_s)}$, $d_2 \in X$ and $d_3 \in N(W^*)$ then

$$\lim_{\rho \rightarrow 0} \Delta J_0^f(\rho) - \Delta J_0^s(\rho) = \lim_{\rho \rightarrow 0} \frac{1}{2} \langle W^* \hat{y}_s(\rho), (W^*W + \rho I)^{-1} W^* \hat{y}_s(\rho) \rangle_{H_m}$$

$$= \frac{1}{2} \langle d_2, d_2 \rangle$$

Proof $\Delta J_0^f(\rho) - \Delta J_0^s(\rho) = \frac{1}{2} \langle g_s, (W^*W + \rho I)^{-1} g_s \rangle \quad (\text{Lemma 2.9})$

Substituting for $g_s = W^* \hat{y}_s(\rho) + \rho \begin{pmatrix} \hat{u}_s(\rho) \\ 0 \end{pmatrix}$ gives

$$\Delta J_0^f(\rho) - \Delta J_0^s(\rho) = \frac{1}{2} \langle W^* \hat{y}_s(\rho), (W^*W + \rho I)^{-1} W^* \hat{y}_s(\rho) \rangle + t_1(\rho) + t_2(\rho)$$

where $t_1(\rho) = \rho \langle W^* \hat{y}_s(\rho), (W^*W + \rho I)^{-1} \begin{pmatrix} \hat{u}_s(\rho) \\ 0 \end{pmatrix} \rangle$

and $t_2(\rho) = \frac{\rho^2}{2} \langle \begin{pmatrix} \hat{u}_s(\rho) \\ 0 \end{pmatrix}, (W^*W + \rho I)^{-1} \begin{pmatrix} \hat{u}_s(\rho) \\ 0 \end{pmatrix} \rangle$

$$t_1(\rho) = \rho \left\langle \begin{pmatrix} \hat{u}_s(\rho) \\ 0 \end{pmatrix}, (W^*W + \rho I)^{-1} W^* \left(W \begin{pmatrix} \hat{u}_s(\rho) \\ 0 \end{pmatrix} + d \right) \right\rangle$$

$$= t_{11}(\rho) - t_{12}(\rho)$$

where $t_{11}(\rho) = \rho \left\langle \begin{pmatrix} \hat{u}_s(\rho) \\ 0 \end{pmatrix}, (W^*W + \rho I)^{-1} W^* W \begin{pmatrix} \hat{u}_s(\rho) \\ 0 \end{pmatrix} \right\rangle$

and $t_{12}(\rho) = \rho \left\langle \begin{pmatrix} \hat{u}_s(\rho) \\ 0 \end{pmatrix}, \hat{u}(\rho) \right\rangle$

Using $(I + GF)^{-1} G = G (I + FG)^{-1} \quad (\text{Lemma 2.1.1 Chapter 2,}$

Section 2.1)

$$t_{11}(\rho) = \rho \left\langle \begin{pmatrix} \hat{u}_S(\rho) \\ 0 \end{pmatrix}, W^* (\rho I + WW^*)^{-1} W \begin{pmatrix} \hat{u}_S(\rho) \\ 0 \end{pmatrix} \right\rangle_{H_m}$$

Using Lemma 2.7 $0 < t_{11}(\rho) < M_1(\rho) \cdot \rho \left\langle \hat{u}_S(\rho), \hat{u}_S(\rho) \right\rangle_{H_m}$

Hence Lemma 2.5, and 2.7 gives $\lim_{\rho \rightarrow 0} t_{11}(\rho) = 0$

For $t_{12}(\rho) = \rho \left\langle \begin{pmatrix} \hat{u}_S(\rho) \\ 0 \end{pmatrix}, \hat{u}(\rho) \right\rangle$ use Schwarz Inequality

$$|t_{12}(\rho)| \leq \left\| \rho^{\frac{1}{2}} \hat{u}_S(\rho) \right\| \cdot \left\| \rho^{\frac{1}{2}} \hat{u}(\rho) \right\|$$

Using Lemma 2.5 $\lim_{\rho \rightarrow 0} |t_{12}(\rho)| = 0$ implies $\lim_{\rho \rightarrow 0} t_{12}(\rho) = 0$.

$\therefore t_1(\rho) = t_{11}(\rho) - t_{12}(\rho)$ implies $\lim_{\rho \rightarrow 0} t_1(\rho) = 0$

$$\begin{aligned} \text{For } t_2(\rho) &= \frac{\rho^2}{2} \left\langle \begin{pmatrix} \hat{u}_S(\rho) \\ 0 \end{pmatrix}, (W^*W + \rho I)^{-1} \begin{pmatrix} \hat{u}_S(\rho) \\ 0 \end{pmatrix} \right\rangle \\ &= \frac{1}{2} \left\langle \begin{pmatrix} \rho^{\frac{1}{2}} \hat{u}_S(\rho) \\ 0 \end{pmatrix}, \rho (W^*W + \rho I)^{-1} \begin{pmatrix} \rho^{\frac{1}{2}} \hat{u}_S(\rho) \\ 0 \end{pmatrix} \right\rangle \\ &\leq \frac{1}{2} M_2(\rho) \rho \left\langle \hat{u}_S(\rho), \hat{u}_S(\rho) \right\rangle \end{aligned}$$

$$M_2(\rho) = \sup_x \frac{\langle x, \rho (W^*W + \rho I)^{-1} x \rangle}{\langle x, x \rangle}$$

Use Householder's Equality then

$$M_2(\rho) = \sup_x \frac{\langle x, x \rangle - \langle x, W^* (\rho I + WW^*) W x \rangle}{\langle x, x \rangle}$$

$$= 1 - M_1(\rho) < 1 \quad \text{and} \quad \lim_{\rho \rightarrow 0} M_2(\rho) = 0$$

Hence $\lim_{\rho \rightarrow 0} t_2(\rho) = 0$ with $\lim_{\rho \rightarrow 0} t_1(\rho) = 0$ giving

$$\lim_{\rho \rightarrow 0} \Delta J_O^f(\rho) - \Delta J_O^s(\rho) = \lim_{\rho \rightarrow 0} \frac{1}{2} \langle W^* \hat{y}_S(\rho), (W^*W + \rho I)^{-1} W^* \hat{y}_S(\rho) \rangle$$

where $\hat{y}_S(\rho) = W_S \hat{u}_S(\rho) + d$

Decompose the output space as $H_r = \overline{R(W_s)} \oplus X + N(W^*)$ and the disturbance accordingly as $d = d_1 + d_2 + d_3$ where $d_1 \in \overline{R(W_s)}$, $d_2 \in X$ and $d_3 \in N(W^*)$. Write $\hat{Z}_s = W_s \hat{u}_s(\rho)$ then $\hat{Z}_s(\rho) \in \overline{R(W_s)}$, and further $\lim_{\rho \rightarrow 0} \hat{Z}_s(\rho) + d_1 = 0$. (See Lemma 2.2.1, Chapter 2, Section 2.2). Then

$$\begin{aligned} \text{(i)} \quad & \lim_{\rho_1 \rightarrow 0} \lim_{\rho_2 \rightarrow 0} \frac{1}{2} \langle W^* \hat{y}_s(\rho_1), (W^*W + \rho_2 I)^{-1} W^* \hat{y}_s(\rho_2) \rangle \\ &= \lim_{\rho_1 \rightarrow 0} \frac{1}{2} \langle \hat{Z}_s(\rho_1) + d_1 + d_2, \hat{Z}_s(\rho_1) + d_1 + d_2 \rangle \\ &= \frac{1}{2} \langle d_2, d_2 \rangle \end{aligned}$$

$$\begin{aligned} \text{and (ii)} \quad & \lim_{\rho_2 \rightarrow 0} \lim_{\rho_1 \rightarrow 0} \frac{1}{2} \langle W^* \hat{y}_s(\rho_1), (W^*W + \rho_2 I)^{-1} W^* \hat{y}_s(\rho_1) \rangle \\ &= \lim_{\rho_2 \rightarrow 0} \frac{1}{2} \langle W^* (d_2 + d_3), (W^*W + \rho_2 I)^{-1} W^* (d_2 + d_3) \rangle \\ &= \frac{1}{2} \langle d_2, d_2 \rangle \end{aligned}$$

$$\begin{aligned} \text{Hence (i) and (ii) imply} \quad & \lim_{\rho \rightarrow 0} \frac{1}{2} \langle W^* \hat{y}_s(\rho), (W^*W + \rho I)^{-1} W^* \hat{y}_s(\rho) \rangle \\ &= \frac{1}{2} \langle d_2, d_2 \rangle. \end{aligned}$$

Appendix C Norm Concepts, Lemmas and Theorems Relating to
the Problem of Bounding the Maximum Eigenvalue of Operator VV^*

The concepts of norm, and induced norm are introduced here.

A norm on a linear space X over a field F is a real valued function whose value at $x \in X$ is denoted $||x||$. This function satisfies the following axioms:

- a) $||x|| \geq 0$, the equality being attained iff $x = 0$.
- b) $||\alpha x|| = |\alpha| \cdot ||x|| \quad \forall \alpha \in F \text{ and } x \in X$
- c) $||x_1 + x_2|| \leq ||x_1|| + ||x_2|| \quad \forall \text{ pairs } x_1, x_2 \in X$

A normed linear space is a linear space on which a norm is defined, this becomes a metric space if the metric is defined via $d(x, y) = ||x - y||$; $x, y \in X$. If X , and Y are two normed spaces, and T is a linear operator on X into Y , then T is continuous on X iff $\exists$ a constant M such that

$$||Tx||_Y \leq M ||x||_X \quad \forall x \in X.$$

The smallest admissible value of the constant M is called the norm of operator T denoted $||T||_X$. The suffix ' X ' denotes that the operator norm is "induced" by or "subordinate" to the vector norm on the linear space X . This inequality motivates the "sup" definition of the induced operator norm:

$$||T||_X \triangleq \sup_{\substack{x \neq 0 \\ ||x||_X = 1}} \frac{||Tx||_Y}{||x||_X} = \sup_{||x||_X = 1} ||Tx||_Y$$

The two remain norm concepts used in Section 4.4 are the norm of a product of linear operators, and the concept of the compatible norm both explained in the main text.

The bounds on the maximum eigenvalue of VV^* depended on a number of results from the theory of Lebesgue integration. Preceding the main theorem of subsection 4.4.1, which gave the bound formulae for the case of a single input, single output system, were two technical lemmas; these are proved first. Recall the convolution operators under consideration as

$$(Vu)(t) = \int_0^t v(t-\tau)u(\tau)d\tau \text{ and } (V^*y)(t) = \int_t^{T_f} v(\tau-t)y(\tau)d\tau$$

$$0 \leq t \leq T_f < \infty$$

Lemma 4.4.1 If $v \in L^1[0, T_f]$ and $u \in L^p[0, T_f]$ then $(Vu) \in L^p[0, T_f]$ $\forall p \in [1, \infty]$. Similarly $V^*y \in L^p[0, T_f]$ whenever $y \in L^p[0, T_f]$.

Proof (Desoer and Vidyasagar, 1975)

i) $1 < p < \infty$: Assume $u \in L^p[0, T_f]$

$$|(Vu)(t)| = \left| \int_0^t v(t-\tau)u(\tau)d\tau \right| \leq \int_0^t |v(t-\tau)| |u(\tau)| d\tau$$

Choose q so that $1/p + 1/q = 1$ then

$$|(Vu)(t)| = \int_0^t |v(t-\tau)|^{1/p} |u(\tau)| \cdot |v(t-\tau)|^{1/q} d\tau$$

Define $z(t, \tau) \triangleq |v(t-\tau)|^{1/p} |u(\tau)|$ $0 \leq \tau \leq t \leq T_f < \infty$

$$\text{then } \|z\|_p^p = \int_0^{T_f} dt \int_0^{T_f} |v(t-\tau)| |u(\tau)|^p d\tau \leq \|v\|_1 \|u\|_p^p$$

$$\Rightarrow z \in L^p[0, T_f] \quad (\text{By Tonelli's Theorem})$$

Similarly $x(s) \triangleq |v(s)|^{1/q}$ $0 \leq s \leq T_f$

$$\|x\|_q^q = \int_0^{T_f} |v(s)| ds = \|v\|_1 \Rightarrow x \in L^q[0, T_f]$$

Using Holder's Inequality

$$|(Vu)(t)| \leq \left[\int_0^t |v(t-\tau)| |u(\tau)|^p d\tau \right]^{1/p} \left[\int_0^t |v(t-\tau)| d\tau \right]^{1/q}$$

The second term bounded by $\|v\|_1^{1/q} \quad \forall t \in [0, T_f]$

Forming $\|\cdot\|_p$ on both sides gives

$$\begin{aligned} \left[\int_0^{T_f} |(Vu)(t)| dt \right]^{1/p} &\leq \|v\|_1^{1/q} \left[\int_0^{T_f} dt \int_0^t |v(t-\tau)| |u(\tau)|^p d\tau \right]^{1/p} \\ &\leq \|v\|_1^{1/q} \|v\|_1^{1/p} \|u\|_p \end{aligned}$$

$$\text{Hence} \quad \|Vu\|_p \leq \|v\|_1 \|u\|_p < \infty$$

$$\Rightarrow Vu \in L^p[0, T_f]. \text{ whenever } u \in L^p[0, T_f], \quad 1 < p < \infty$$

ii) Assume $p=1$, hence $u \in L^1[0, T_f]$

$$\begin{aligned} \|Vu\|_1 &= \int_0^{T_f} dt \left| \int_0^t v(t-\tau) u(\tau) d\tau \right| \leq \int_0^{T_f} dt \int_0^t |v(t-\tau)| |u(\tau)| d\tau \\ &= \int_0^{T_f} u(\tau) \int_0^{T_f} |v(t-\tau)| dt d\tau \leq \int_0^{T_f} |u(\tau)| d\tau \int_0^{T_f} |v(s)| ds \end{aligned}$$

$$\Rightarrow \|Vu\|_1 \leq \|v\|_1 \|u\|_1 \Rightarrow Vu \in L^1[0, T_f]$$

iii) Assume $p=\infty$, hence $u \in L^\infty[0, T_f]$

$$\begin{aligned} \|Vu\|_\infty &= \text{ess sup}_{0 \leq t \leq T_f} |(Vu)(t)| = \text{ess sup}_{0 \leq t \leq T_f} \left| \int_0^t v(t-\tau) u(\tau) d\tau \right| \\ &\leq \text{ess sup}_{0 \leq t \leq T_f} \int_0^t |v(t-\tau)| |u(\tau)| d\tau \\ &\leq \text{ess sup}_{0 \leq \tau \leq T_f} u(\tau) \int_0^{T_f} |v(s)| ds \end{aligned}$$

$$\text{hence} \quad \|Vu\|_\infty \leq \|v\|_1 \|u\|_\infty \Rightarrow Vu \in L^\infty[0, T_f]$$

iv) The proof concerning $(V^*y)(t) \triangleq \int_t^{T_f} v(\tau-t) y(\tau) d\tau$ is a direct analogy of the arguments used for the operator V . This concludes the proof of the Lemma 4.4.1.

This lemma was important in the specification of the nature of the mapping of the composite operator VV^* . The next lemma establishes the relationship between the spaces $L^2[0, T_f]$ and $L^1[0, T_f]$, this was used to find $||\cdot||_1$ bounds on the maximum eigenvalue of VV^* .

Lemma 4.4.2 If $x \in L^2[0, T_f]$ then $x \in L^1[0, T_f]$

Proof If $x \in L^2[0, T_f] \Rightarrow \int_0^{T_f} |x(t)|^2 dt = B_x < \infty$

Partition the interval $[0, T_f]$ as $P_1 \triangleq \{t : |x(t)| \leq 1\}$

and $P_2 \triangleq \{t : t \in [0, T_f] ; t \notin P_1\}$

Clearly $[0, T_f] = P_1 + P_2$

Then i) $\int_{P_1} |x(t)| dt < T_f < \infty$

ii) $\int_{P_2} |x(t)| dt < \int_{P_2} |x(t)|^2 dt < B_x < \infty$

$\Rightarrow \int_{P_1 + P_2} |x(t)| dt < T_f + B_x < \infty \Rightarrow x \in L^1[0, T_f]$

This concludes the proof of Lemma 4.4.2.

Together the lemmas enable the bounds on λ_M to be obtained as

$$\lambda_M = ||VV^*||_2 \leq ||VV^*||_{C(2)} \leq ||V||_{C(2)} ||V^*||_{C(2)}$$

$$\text{and } \lambda_M \leq ||VV^*||_1 \leq ||V||_1 ||V^*||_1$$

The formulae for these bounds in the case of a single input, single output system formed the subject of the following theorem :

Theorem 4.4.1 If operator kernel $v \in L^1[0, T_f]$

then a) $\|v\|_1 = \|v^*\|_1 = \int_0^{T_f} |v(s)| ds$

b) $\|vv^*\|_1 \leq \max_{0 \leq t \leq T_f} \int_0^{T_f} \left| \int_0^{\min(t, \tau)} v(t-s) v(\tau-s) ds \right| d\tau$
 $\leq \|v\|_1 \|v^*\|_1$

If operator kernel $v \in L^2[0, T_f]$

then c) $\|vv^*\|_E = \int_0^{T_f} \int_0^{T_f} \left| \int_0^{\min(t, s)} v(t-\tau) v(s-\tau) d\tau \right|^2 ds dt$

where $\|vv^*\|_E \leq \|v\|_E^2$ and $\|v\|_E^2 = \|v^*\|_E^2 = \int_0^{T_f} \int_0^t v^2(t-\tau) d\tau dt$.

Proof Recall $(Vu)(t) = \int_0^t v(t-\tau) u(\tau) d\tau \quad 0 \leq t \leq T_f$

For convenience an associated kernel is defined :

$$\tilde{v}(s) = \begin{cases} 0 & s < 0, \quad s > T_f \\ v(s) & 0 \leq s \leq T_f \end{cases}$$

This is occasionally used in complicated iterated integrals.

a) $\|Vu\|_1 = \int_0^{T_f} |(Vu)(t)| dt = \int_0^{T_f} \left| \int_0^t v(t-\tau) u(\tau) d\tau \right| dt$
 $\leq \int_0^{T_f} \int_0^t |v(t-\tau)| |u(\tau)| d\tau dt \leq \int_0^{T_f} |u(\tau)| \int_0^{T_f} |\tilde{v}(t-\tau)| dt d\tau$
 $\leq \int_0^{T_f} |v(s)| ds \cdot \int_0^{T_f} |u(\tau)| d\tau$

Hence $\|v\|_1 \leq \int_0^{T_f} |v(s)| ds$

Introduce a one parameter function e_α which is an approximate

identity (see Krall, 1973) for $L^1(-\infty, \infty)$ where e_α satisfies

- i) $e_\alpha(t) \geq 0 \quad \forall \alpha, \text{ and } t$
- ii) $\int_{-\infty}^{\infty} e_\alpha(t) dt = \|e_\alpha\|_1 = 1$
- iii) $\forall \epsilon > 0 \quad \lim_{\alpha \rightarrow \infty} \int_{|t| > \epsilon} e_\alpha(t) dt = 0$

Theorem (Krall, 1973) If e_α is an approximate identity for

$L^1(-\infty, \infty)$ then $\lim_{\alpha \rightarrow \infty} \int_{-\infty}^{\infty} e_\alpha(s) x(t-s) ds = x(t) \quad \text{a.e.} \quad \forall x \in L^1(-\infty, \infty).$

Using a zero extension $\forall x \in L^1[0, T_f]$, the problem may be set in the extended interval $(-\infty, \infty)$. Equality of the norm is shown in this extension since if $x \in L^1[0, T_f]$ and its extension in $L^1(-\infty, \infty)$ is $\tilde{x}$

$$\tilde{x}(t) = \begin{cases} 0 & t < 0, \quad t > T_f \\ x(t) & 0 \leq t \leq T_f \end{cases}$$

Then $\|\tilde{x}\|_1 = \int_{-\infty}^{\infty} |\tilde{x}(t)| dt = \int_0^{T_f} |x(t)| dt = \|x\|_1$

Using the input $u(\tau) = \tilde{e}_\alpha(\tau)$ where $e_\alpha \in L^1[0, T_f]$

clearly $\|u\|_1 = 1$ (by property ii) of the approximate identity.)

and $(Vu)(t) = \int_{-\infty}^{\infty} \tilde{v}(t-\tau) e_\alpha(\tau) d\tau, \quad \lim_{\alpha \rightarrow \infty} (Vu)(t) = \tilde{v}(t) \quad t \in (-\infty, \infty)$

then $\|Vu\|_1 = \int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} v(t-\tau) e_\alpha(\tau) d\tau \right| dt \leq \|v\|_1 \cdot 1 \leq \int_0^{T_f} |v(t)| dt$

and $\alpha \rightarrow \infty \Rightarrow \|v\|_1 = \int_0^{T_f} |v(t)| dt$

A similar technique can be used to show $\|v^*\|_1 = \int_0^{T_f} |v(s)| ds$

b) Choosing $u \in L^1[0, T_f]$ and with $v \in L^1[0, T_f]$

Lemma 4.4.1 $\Rightarrow v^* : L^1[0, T_f] \rightarrow L^1[0, T_f]$

and $v : L^1[0, T_f] \rightarrow L^1[0, T_f]$. Forming the composition vv^*

have $VV^* : L^1[0, T_f] \rightarrow L^1[0, T_f]$ hence

$$\begin{aligned} \|VV^*y\|_1 &= \int_0^{T_f} \left| \int_0^\tau v(\tau-s) (V^*y)(s) ds \right| d\tau \\ &= \int_0^{T_f} \left| \int_0^\tau v(\tau-s) \int_s^{T_f} v(t-s)y(t) dt ds \right| d\tau \end{aligned}$$

In terms of $\tilde{v} : \|VV^*y\|_1 = \int_0^{T_f} \left| \int_0^{T_f} \tilde{v}(\tau-s) \int_0^{T_f} \tilde{v}(t-s)y(t) dt ds \right| d\tau$

$$= \int_0^{T_f} \left| \int_0^{T_f} y(t) \int_0^{T_f} \tilde{v}(t-s)\tilde{v}(\tau-s) ds dt \right| d\tau$$

$$\leq \int_0^{T_f} \int_0^{T_f} |y(t)| \left| \int_0^{T_f} \tilde{v}(t-s)\tilde{v}(\tau-s) ds \right| dt d\tau$$

$$= \int_0^{T_f} |y(t)| \int_0^{T_f} \left| \int_0^{T_f} \tilde{v}(t-s)\tilde{v}(\tau-s) ds \right| d\tau dt$$

$$\leq \max_{0 \leq t \leq T_f} \int_0^{T_f} \left| \int_0^{T_f} \tilde{v}(t-s)\tilde{v}(\tau-s) ds \right| d\tau \cdot \int_0^{T_f} |y(t)| dt$$

Giving $\|VV^*\|_1 \leq \max_{0 \leq t \leq T_f} \int_0^{T_f} \left| \int_0^{\min(t, \tau)} v(t-s)v(\tau-s) ds \right| d\tau$

In the absence of a proof of equality, it is shown that

$$\|VV^*\|_1 \leq \max_{0 \leq t \leq T_f} \int_0^{T_f} \left| \int_0^{\min(t, \tau)} v(t-s)v(\tau-s) ds \right| d\tau \leq \|v\|_1 \|v^*\|_1$$

$$\max_{0 \leq t \leq T_f} \int_0^{T_f} \left| \int_0^{\min(t, \tau)} v(t-s)v(\tau-s) ds \right| d\tau$$

$$= \max_{0 \leq t \leq T_f} \int_0^{T_f} \left| \int_0^{T_f} \tilde{v}(t-s)\tilde{v}(\tau-s) ds \right| d\tau$$

$$\leq \max_{0 \leq t \leq T_f} \int_0^{T_f} \int_0^{T_f} |\tilde{v}(t-s)| |\tilde{v}(\tau-s)| ds d\tau$$

$$= \max_{0 \leq t \leq T_f} \int_0^{T_f} |\tilde{v}(t-s)| \int_0^{T_f} |\tilde{v}(\tau-s)| d\tau ds$$

$$\leq \max_{0 \leq t \leq T_f} \int_0^{T_f} |\tilde{v}(t-s)| ds \int_0^{T_f} |\tilde{v}(\tau)| d\tau$$

$$\Rightarrow \|vv^*\|_1 \leq \max_{0 \leq t \leq T_f} \int_0^{T_f} \left| \int_0^{\min(t,\tau)} v(t-s)v(\tau-s) ds \right| d\tau$$

$$\leq \left\{ \int_0^{T_f} |v(s)| ds \right\}^2 = \|v\|_1 \|v^*\|_1$$

c) The norms compatible with $\|x\|_2$, $x \in L^2[0, T_f]$ satisfy the

inequality $\lambda_M = \|vv^*\|_2 \leq \|vv^*\|_E \leq \|v\|_E \|v^*\|_E$. The

derivation of the compatible bound formulae require an additional

restriction on the operator kernel, namely $v \in L^2[0, T_f] \Rightarrow \int_0^{T_f} v(t)^2 dt < \infty$

Under this assumption first show $\|v\|_E^2 = \|v^*\|_E^2 = \int_0^{T_f} \int_0^t v^2(t-\tau) d\tau dt$

$$\|Vu\|_2^2 = \int_0^{T_f} \left(\int_0^t v(t-\tau)u(\tau) d\tau \right)^2 dt$$

Since v , and $u \in L^2[0, T_f]$, have using Schwarz Inequality

$$\begin{aligned} \|Vu\|_2^2 &\leq \int_0^{T_f} \left\{ \int_0^t v^2(t-\tau) d\tau \cdot \int_0^t u^2(\tau) d\tau \right\} dt \\ &\leq \int_0^{T_f} \int_0^t v^2(t-\tau) d\tau \cdot \|u\|_2^2 \end{aligned}$$

Identify $\|v\|_E^2 \triangleq \int_0^{T_f} \int_0^t v^2(t-\tau) d\tau dt$.

A similar analysis for v^* gives $\|v^*\|_E^2 = \|v\|_E^2$

The derivation for $\|vv^*\|_E$ and its relation to $\|v\|_E^2$ follows :

$$\begin{aligned}
||vv^*y||_2^2 &= \int_0^{T_f} \left(\int_0^t v(t-\tau) \int_\tau^{T_f} v(s-\tau)y(s) ds d\tau \right)^2 dt \\
&= \int_0^{T_f} \left(\int_0^{T_f} \tilde{v}(t-\tau) \int_0^{T_f} \tilde{v}(s-\tau)y(s) ds d\tau \right)^2 dt \\
&= \int_0^{T_f} \left(\int_0^{T_f} y(s) \int_0^{T_f} \tilde{v}(t-\tau)\tilde{v}(s-\tau) d\tau ds \right)^2 dt
\end{aligned}$$

Again using Schwarz Inequality

$$\leq \int_0^{T_f} \left[\int_0^{T_f} \left(\int_0^{T_f} \tilde{v}(t-\tau)\tilde{v}(s-\tau) d\tau \right)^2 ds \cdot \int_0^{T_f} y^2(s) ds \right] dt$$

and identify $||vv^*||_E \triangleq \int_0^{T_f} \int_0^{T_f} \left(\int_0^{\min(t,s)} v(t-\tau)v(s-\tau) d\tau \right)^2 ds dt$

$$\begin{aligned}
\text{Hence } ||vv^*||_E &= \int_0^{T_f} \int_0^{T_f} \left(\int_0^{T_f} \tilde{v}(t-\tau)\tilde{v}(s-\tau) d\tau \right)^2 ds dt \\
&\leq \int_0^{T_f} \int_0^{T_f} \left(\int_0^{T_f} \tilde{v}^2(t-\tau) d\tau \cdot \int_0^{T_f} \tilde{v}^2(s-\tau) d\tau \right) ds dt \\
&= \int_0^{T_f} \int_0^t v^2(t-\tau) d\tau dt \int_0^{T_f} \int_0^s v^2(s-\tau) d\tau ds
\end{aligned}$$

$$\Rightarrow ||vv^*||_E \leq ||v||_E^2 \quad \text{as required.}$$

This concludes the results for the case of single input, single output systems. The development for the multivariable situation follows much the same pattern. Multivariable versions of Lemmas 4.4.1 and 4.4.2 are presented to enable the norm bounds to be stated followed by the derivation of the norm bound formulae, namely :

Theorem 4.4.2 If elements of the operator kernel, $v_{ij} \in L^1[0, T_f]$ then

$$\text{a) } ||v||_1 = \max_j \sum_{i=1}^r \int_0^{T_f} |v_{ij}(t)| dt \quad (\text{max. column sum})$$

$$\text{b) } ||v^*||_1 = \max_i \sum_{j=1}^m \int_0^{T_f} |v_{ij}(t)| dt \quad (\text{max. row sum})$$

$$\text{c) } ||vv^*||_1 \leq \max_{i_0 s} \sum_{i=1}^r \int_0^{T_f} \left| \sum_{j=1}^m \int_0^{\min(t,s)} v_{ij}(t-\tau)v_{i_0 j}(s-\tau) d\tau \right| dt$$

$$\leq \|v\|_1 \|v^*\|_1$$

Restricting further the kernel of operator V , so that $v_{ij} \in L^2[0, T_f]$

$$\text{then d) } \|v\|_E^2 = \|v^*\|_E^2 = \int_0^{T_f} \int_0^t \sum_{i=1}^r \sum_{j=1}^m v_{ij}^2(t-\tau) d\tau dt$$

$$\begin{aligned} \text{and e) } \|vv^*\|_E &= \sum_{i=1}^r \sum_{j=1}^m \int_0^{T_f} dt \int_0^{T_f} ds \left| \sum_{j=1}^m \int_0^{\min(t,s)} v_{ij}(t-\tau) v_{i_0j}(s-\tau) d\tau \right|^2 \\ &\leq \|v\|_E^2. \end{aligned}$$

Proof a) Recall $(Vu)_i(t) = \int_0^t \sum_{j=1}^m v_{ij}(t-\tau) u_j(\tau) d\tau$

$$\begin{aligned} \text{Then } \|Vu\|_1 &= \int_0^{T_f} \sum_{i=1}^r |(Vu)_i(t)| dt = \sum_{i=1}^r \int_0^{T_f} \left| \sum_{j=1}^m \int_0^t v_{ij}(t-\tau) u_j(\tau) d\tau \right| dt \\ &\leq \sum_{i=1}^r \sum_{j=1}^m \int_0^{T_f} \int_0^t |v_{ij}(t-\tau)| |u_j(\tau)| d\tau dt \\ &\leq \sum_{i=1}^r \sum_{j=1}^m \int_0^{T_f} |v_{ij}(s)| ds \int_0^{T_f} |u_j(\tau)| d\tau \\ &\leq \left\{ \max_j \sum_{i=1}^r \int_0^{T_f} |v_{ij}(s)| ds \right\} \cdot \|u\|_1 \end{aligned}$$

As in Theorem 4.4.1 equality is shown using an approximate identity e_α for $L^1(-\infty, \infty)$. Let j_1 be the value of j for which the above expression attains its maximum, and define

$$u(\tau) = \begin{pmatrix} 0 \\ \vdots \\ e_\alpha(\tau) \\ \vdots \\ 0 \end{pmatrix} \quad \text{where } e_\alpha(\tau) \text{ appears in the } j_1\text{th entry,} \quad \text{and } \|u\|_1 = 1$$

$$\text{Then } (Vu)(t) = \begin{pmatrix} \int_{-\infty}^{\infty} v_{1j_1}(t-\tau) e_\alpha(\tau) d\tau \\ \vdots \\ \int_{-\infty}^{\infty} v_{rj_1}(t-\tau) e_\alpha(\tau) d\tau \end{pmatrix} \quad \text{and } \lim_{\alpha \rightarrow \infty} (Vu)(t) = \begin{pmatrix} v_{1j_1}(t) \\ \vdots \\ v_{rj_1}(t) \end{pmatrix}$$

Hence $\|Vu\|_1 = \int_{-\infty}^{\infty} \sum_{i=1}^r \left| \int_{-\infty}^{\infty} v_{ij_1}(t-\tau) e_{\alpha}(\tau) d\tau \right| dt \leq \|v\|_1 \cdot 1$

and as $\alpha \rightarrow \infty \Rightarrow \|v\|_1 = \max_j \sum_{i=1}^r \int_0^{T_f} |v_{ij}(s)| ds.$

b) The analysis for $\|v^*\|_1$ is almost identical to that for $\|v\|_1$

except that the kernel matrix is transposed giving

$$\|v^*\|_1 = \max_i \sum_{j=1}^m \int_0^{T_f} |v_{ij}(s)| ds.$$

c) Recall $(Vv^*y)(t) = \int_0^t v(t-\tau) \int_{\tau}^{T_f} v^T(s-\tau)y(s) ds d\tau$

where $Vv^* : L_r^1[0, T_f] \rightarrow L_r^1[0, T_f]$

and $V(.) : \mathbb{R}^m \rightarrow \mathbb{R}^r \quad 0 \leq t \leq T_f$

Then $(Vv^*y)_i(t) = \int_0^t \int_{\tau}^{T_f} \sum_{i_0=1}^r \left[v(t-\tau)v^T(s-\tau) \right]_{ii_0} y_{i_0}(s) ds d\tau dt$

where $\left[v(t-\tau)v^T(s-\tau) \right]_{ii_0} = \sum_{j=1}^m v_{ij}(t-\tau)v_{i_0j}(s-\tau)$

$$\begin{aligned} \text{Hence } \|Vv^*y\|_1 &= \int_0^{T_f} \sum_{i=1}^r \left| \int_0^t \int_{\tau}^{T_f} \sum_{i_0=1}^r \sum_{j=1}^m v_{ij}(t-\tau)v_{i_0j}(s-\tau)y_{i_0}(s) ds d\tau \right| dt \\ &= \int_0^{T_f} \sum_{i=1}^r \left| \int_0^{T_f} \int_0^{T_f} \sum_{i_0=1}^r \sum_{j=1}^m \tilde{v}_{ij}(t-\tau)\tilde{v}_{i_0j}(s-\tau)y_{i_0}(s) ds d\tau \right| dt \end{aligned}$$

where $\tilde{v}_{ij}(s) = \begin{cases} 0 & \text{if } s < 0 \text{ or } s > T_f \\ v_{ij}(s) & 0 \leq s \leq T_f \end{cases}$

$$\|Vv^*y\|_1 = \int_0^{T_f} \sum_{i=1}^r \left| \int_0^{T_f} \sum_{i_0=1}^r y_{i_0}(s) \sum_{j=1}^m \int_0^{T_f} \tilde{v}_{ij}(t-\tau)\tilde{v}_{i_0j}(s-\tau) d\tau ds \right| dt$$

$$\leq \int_0^{T_f} \sum_{i=1}^r \int_0^{T_f} \sum_{i_0=1}^r |y_{i_0}(s)| \left| \sum_{j=1}^m \int_0^{T_f} \tilde{v}_{ij}(t-\tau)\tilde{v}_{i_0j}(s-\tau) d\tau \right| ds dt$$

$$\leq \max_{i_0, s} \left\{ \int_0^{T_f} \sum_{i=1}^r \left| \sum_{j=1}^m \int_0^{\min(t, \tau)} v_{ij}(t-\tau)v_{i_0j}(s-\tau) d\tau \right| dt \right\} \cdot \int_0^{T_f} \sum_{i_0=1}^r |y_{i_0}(s)| ds$$

Hence $||VV^*||_1 \leq \max_{i_0, s} \sum_{i=1}^r \int_0^{T_f} \left| \sum_{j=1}^m \int_0^{\min(t,s)} v_{ij}(t-\tau) v_{i_0 j}(s-\tau) d\tau \right| dt$

In the absence of a proof for equality, compatibility with

$||V||_1 ||V^*||_1$ is shown :

$$\begin{aligned}
 ||VV^*|| &\leq \max_{i_0, s} \sum_{i=1}^r \int_0^{T_f} \left| \sum_{j=1}^m \int_0^{\min(t,s)} v_{ij}(t-\tau) v_{i_0 j}(s-\tau) d\tau \right| dt \\
 &\leq \max_{i_0, s} \sum_{i=1}^r \int_0^{T_f} \sum_{j=1}^m \int_0^{T_f} |\tilde{v}_{ij}(t-\tau)| |\tilde{v}_{i_0 j}(s-\tau)| d\tau dt \\
 &= \max_{i_0, s} \sum_{i=1}^r \sum_{j=1}^m \int_0^{T_f} |\tilde{v}_{i_0 j}(s-\tau)| \int_0^{T_f} |\tilde{v}_{ij}(t-\tau)| dt d\tau \\
 &\leq \max_{i_0} \sum_{j=1}^m \int_0^{T_f} |v_{i_0 j}(s)| ds \sum_{i=1}^r \int_0^{T_f} |v_{ij}(t)| dt \\
 &\leq \max_{i_0} \sum_{j=1}^m \int_0^{T_f} |v_{i_0 j}(s)| ds \cdot \max_j \sum_{i=1}^r \int_0^{T_f} |v_{ij}(t)| dt \\
 &\Rightarrow ||VV^*|| \leq \max_{i_0, s} \sum_{i=1}^r \int_0^{T_f} \left| \sum_{j=1}^m \int_0^{\min(t,s)} v_{ij}(t-\tau) v_{i_0 j}(s-\tau) d\tau \right| dt \\
 &\leq ||V||_1 ||V^*||_1
 \end{aligned}$$

d) The norms compatible with $||x||_2$, $x \in L^2[0, T_f]$ satisfy the

inequality $\lambda_M = ||VV^*||_2 \leq ||VV^*||_E \leq ||V||_E ||V^*||_E$. An additional restriction on the elements of the operator kernel is assumed, namely $v_{ij}(\cdot) \in L^2[0, T_f]$ $i=1, 2, \dots, r$ and $j=1, 2, \dots, m$.

$$(Vu)_i(t) = \int_0^t \sum_{j=1}^m v_{ij}(t-\tau) u_j(\tau) d\tau$$

$$\begin{aligned}
 \text{Hence } ||Vu||_2^2 &= \int_0^{T_f} \sum_{i=1}^r \left| \int_0^t \sum_{j=1}^m v_{ij}(t-\tau) u_j(\tau) d\tau \right|^2 dt \\
 &= \int_0^{T_f} \sum_{i=1}^r \left| \sum_{j=1}^m \int_0^t v_{ij}(t-\tau) u_j(\tau) d\tau \right|^2 dt
 \end{aligned}$$

$$\begin{aligned}
&\leq \int_0^{T_f} \sum_{i=1}^r \left(\sum_{j=1}^m \int_0^t |v_{ij}(t-\tau) u_j(\tau)| d\tau \right)^2 dt \\
&\leq \int_0^{T_f} \sum_{i=1}^r \left(\sum_{j=1}^m \left[\int_0^t v_{ij}^2(t-\tau) d\tau \right]^{\frac{1}{2}} \left[\int_0^t u_j^2(\tau) d\tau \right]^{\frac{1}{2}} \right)^2 dt \\
&\leq \int_0^{T_f} \sum_{i=1}^r \sum_{j=1}^m \int_0^t v_{ij}^2(t-\tau) d\tau \cdot \sum_{j=1}^m \int_0^t u_j^2(\tau) d\tau dt
\end{aligned}$$

Use of Schwarz Inequality for integrals and finite sums has been made to give :

$$||Vu||_2^2 \leq \int_0^{T_f} \int_0^t \sum_{i=1}^r \sum_{j=1}^m v_{ij}^2(t-\tau) d\tau dt \quad ||u||_2^2.$$

Identify therefore $||V||_E^2 \triangleq \int_0^{T_f} \int_0^t \sum_{i=1}^r \sum_{j=1}^m v_{ij}^2(t-\tau) d\tau dt,$

A similar analysis for V^* yields $||V^*||_E^2 = ||V||_E^2$. (This result was obtained by Freeman, 1969.)

e) Recall $(V V^* y)_i(t) = \int_0^t \int_0^{T_f} \sum_{i_o=1}^r \sum_{j=1}^m v_{ij}(t-\tau) v_{i_o j}(s-\tau) y_{i_o}(s) ds d\tau$

$$\begin{aligned}
\text{Hence } ||V V^* y||_2^2 &= \int_0^{T_f} \sum_{i=1}^r |(V V^* y)_i(t)|^2 dt \\
&= \int_0^{T_f} \sum_{i=1}^r \left| \int_0^t \int_0^{T_f} \sum_{i_o=1}^r \sum_{j=1}^m \tilde{v}_{ij}(t-\tau) \tilde{v}_{i_o j}(s-\tau) y_{i_o}(s) ds d\tau \right|^2 dt
\end{aligned}$$

$$\begin{aligned}
\text{where } \tilde{v}_{ij}(s) &= 0 & s < 0 \\
&= v_{ij}(s) & 0 \leq s \leq T_f \\
&= 0 & s > T_f
\end{aligned}$$

Rearrange the order of integration to obtain

$$\begin{aligned}
||V V^* y||_2^2 &= \int_0^{T_f} \sum_{i=1}^r \left| \sum_{i_o=1}^r \int_0^{T_f} y_{i_o}(s) \sum_{j=1}^m \int_0^{T_f} \tilde{v}_{ij}(t-\tau) \tilde{v}_{i_o j}(s-\tau) d\tau ds \right|^2 dt \\
&\leq \int_0^{T_f} \sum_{i=1}^r \left(\sum_{i_o=1}^r \int_0^{T_f} |y_{i_o}(s)| \sum_{j=1}^m \int_0^{T_f} \tilde{v}_{ij}(t-\tau) \tilde{v}_{i_o j}(s-\tau) d\tau ds \right)^2 dt
\end{aligned}$$

Use Schwarz Inequality for integrals

$$\leq \int_0^{T_f} \sum_{i=1}^r \left(\sum_{i_0=1}^r \left[\int_0^{T_f} |y_{i_0}(s)|^2 ds \right]^{\frac{1}{2}} \cdot \left[\int_0^{T_f} \sum_{j=1}^m \int_0^{T_f} \tilde{v}_{ij}(t-\tau) \tilde{v}_{i_0j}(s-\tau) d\tau \right]^{\frac{1}{2}} ds \right)^2 dt$$

Using Schwarz Inequality for finite sums

$$||v v^*||_2^2 \leq \sum_{i=1}^r \sum_{i_0=1}^r \int_0^{T_f} dt \int_0^{T_f} ds \left(\sum_{j=1}^m \int_0^{\min(t,s)} v_{ij}(t-\tau) v_{i_0j}(s-\tau) d\tau \right)^2 \cdot ||u||_2^2$$

$$\text{Identify } ||v v^*||_E^2 = \sum_{i=1}^r \sum_{i_0=1}^r \int_0^{T_f} dt \int_0^{T_f} ds \left(\sum_{j=1}^m \int_0^{\min(t,s)} v_{ij}(t-\tau) v_{i_0j}(s-\tau) d\tau \right)^2$$

Finally compatibility with $||v||_E^2$ is shown.

$$\begin{aligned} ||v v^*||_E^2 &= \sum_{i=1}^r \sum_{i_0=1}^r \int_0^{T_f} dt \int_0^{T_f} ds \left| \sum_{j=1}^m \int_0^{T_f} \tilde{v}_{ij}(t-\tau) \tilde{v}_{i_0j}(s-\tau) d\tau \right|^2 \\ &\leq \sum_{i=1}^r \sum_{i_0=1}^r \int_0^{T_f} dt \int_0^{T_f} ds \left(\sum_{j=1}^m \int_0^{T_f} |\tilde{v}_{ij}(t-\tau) \tilde{v}_{i_0j}(s-\tau)| d\tau \right)^2 \\ &\leq \sum_{i=1}^r \sum_{i_0=1}^r \int_0^{T_f} dt \int_0^{T_f} ds \left(\sum_{j=1}^m \left[\int_0^{T_f} |\tilde{v}_{ij}^2(t-\tau)| d\tau \right]^{\frac{1}{2}} \cdot \left[\int_0^{T_f} |\tilde{v}_{i_0j}^2(s-\tau)| d\tau \right]^{\frac{1}{2}} \right)^2 \\ &\leq \sum_{i=1}^r \sum_{i_0=1}^r \int_0^{T_f} dt \int_0^{T_f} ds \sum_{j=1}^m \left\{ \int_0^{T_f} |\tilde{v}_{ij}^2(t-\tau)| d\tau \cdot \int_0^{T_f} |\tilde{v}_{i_0j}^2(s-\tau)| d\tau \right\} \\ &= \sum_{i=1}^r \sum_{j=1}^m \int_0^{T_f} dt \int_0^t v_{ij}^2(t-\tau) d\tau \cdot \sum_{i_0=1}^r \sum_{j=1}^m \int_0^{T_f} ds \int_0^s v_{i_0j}^2(s-\tau) d\tau \\ &\Rightarrow ||v v^*||_E \leq ||v||_E \cdot ||v^*||_E. \end{aligned}$$

This concludes the proof of the theorem giving the bound formulae. The computational importance of the formulae obtained is given in the same

chapter and some discussion on implementation is also given in Chapter 5 which details a significant industrial example.

Appendix D. On the Modified Dominance Cost Function

The cost function for dominance at the frequency $s=iw$ for the j th column $q_{\cdot j} = G(iw) k_{\cdot j}$ $j=1,2,\dots,m$ is given as

$$J_{D_j}(k_{\cdot j}) = |q_{jj}(iw)| - \sum_{\substack{i=1 \\ i \neq j}}^m |q_{ij}(iw)| \quad j=1,2,\dots,m.$$

Writing the matrix $G(iw)$ in real and imaginary parts as

$G(iw) = [\alpha + i\beta]$, and similarly for the vector $q_{\cdot j}$, obtain

$$q \triangleq \begin{bmatrix} q_1^j \\ q_2^j \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} k_{\cdot j} = W k_{\cdot j}$$

where $q_{\cdot j} \triangleq q_1^j + iq_2^j$, $q \in \mathbb{R}^{2m}$, $k_{\cdot j} \in \mathbb{R}^m$ and $W : \mathbb{R}^m \rightarrow \mathbb{R}^{2m}$

The dominance cost may therefore be written as

$$J_{D_j}(q) = (q^2(j) + q^2(j+m))^{\frac{1}{2}} - \sum_{\substack{i=1 \\ i \neq j}}^m (q^2(i) + q^2(i+m))^{\frac{1}{2}} \quad j=1,2,\dots,m.$$

The square root operator within the summand of the cost function

$J_{D_j}(q)$ does not make it a very attractive function for further

analysis. In Chapter 6, the geometric configuration of dominance costs are used to motivate the use of modified cost functions. The most useful of these is given as :

$$\tan^2 \theta(q) = \sum_{\substack{i=1 \\ i \neq j}}^m \frac{(q^2(i) + q^2(i+m))}{q^2(j) + q^2(j+m)} < \left[\sum_{\substack{i=1 \\ i \neq j}}^m \frac{(q^2(i) + q^2(i+m))^{\frac{1}{2}}}{(q^2(j) + q^2(j+m))^{\frac{1}{2}}} \right]^{\frac{1}{2}}$$

It is necessary to ascertain, whether iterates from a minimization of $\tan^2 \theta(q)$ are producing useful iterates when measured with the correct dominance cost. Sufficient conditions were obtained and are given in the following theorem.

Theorem D.1 Let q_1 and $q_2 \in Q_{D_j}$ where $||q_1||_2 = ||q_2||_2 = \mu$,

and $\tan \theta(q_1) < \tan \theta(q_2)$ with $q_k \in \mathbb{R}^{2m}$ $k=1,2$.

Define $\{ a_{ki} \triangleq (q_k^2(i) + q_k^2(i+m))^{1/2} \}_{k=1}^2 \{ i=1}^m$

If for the j th column $a_{1j}^2 \geq \frac{a_{2j}^2}{(m-1)^{1/2}} + \mu^2 \left(1 - \frac{1}{(m-1)^{1/2}} \right)$

then $J_{D_j}(q_1) > J_{D_j}(q_2)$ where $m > 1$.

Proof In the notation of the theorem statement

$$J_{D_j}(q_1) - J_{D_j}(q_2) = a_{1j} - a_{2j} + \sum_{\substack{i=1 \\ i \neq j}}^m (a_{2i} - a_{1i}). \text{ Based on this}$$

expression the proof is in two parts. The first part shows how the conditions of the theorem imply $(a_{1j} - a_{2j}) > 0$, and the second

part proves the sufficiency condition to ensure $\sum_{\substack{i=1 \\ i \neq j}}^m (a_{2i} - a_{1i}) > 0$.

The conditions of the theorem are :

$$q_1 \in Q_{D_j} \Rightarrow a_{1j} > \sum_{\substack{i=1 \\ i \neq j}}^m a_{1i} \quad q_2 \in Q_{D_j} \Rightarrow a_{2j} > \sum_{\substack{i=1 \\ i \neq j}}^m a_{2i} \quad 6.3.1.4$$

$$||q_1||_2^2 = a_{1j}^2 + \sum_{\substack{i=1 \\ i \neq j}}^m a_{1i}^2 = ||q_2||_2^2 = a_{2j}^2 + \sum_{\substack{i=1 \\ i \neq j}}^m a_{2i}^2 = \mu^2. \quad 6.3.1.5$$

$$\tan^2 \theta(q_1) = \sum_{\substack{i=1 \\ i \neq j}}^m a_{1i}^2 / a_{1j}^2 < \sum_{\substack{i=1 \\ i \neq j}}^m a_{2i}^2 / a_{2j}^2 = \tan^2 \theta(q_2) \quad 6.3.1.6$$

i) To show that 6.3.1.4, 6.3.1.5 and 6.3.1.6 implies $a_{1j} - a_{2j} > 0$.

$$\text{Relation 6.3.1.4 implies } a_{kj}^2 > \left(\sum_{\substack{i=1 \\ i \neq j}}^m a_{ki} \right)^2 > \sum_{\substack{i=1 \\ i \neq j}}^m a_{ki}^2 > 0 \quad k=1,2 \quad 6.3.1.7$$

Relations 6.3.1.7 and 6.3.1.6 give

$$\sum_{\substack{i=1 \\ i \neq j}}^m \frac{a_{1i}^2}{a_{1j}^2} < \sum_{\substack{i=1 \\ i \neq j}}^m \frac{a_{2i}^2}{a_{2j}^2} < 1 \quad \text{and hence} \quad \frac{a_{2j}^2}{a_{1j}^2} < \frac{\sum_{\substack{i=1 \\ i \neq j}}^m a_{2i}^2}{\sum_{\substack{i=1 \\ i \neq j}}^m a_{1i}^2} \\ < a_{2j}^2 / \sum_{\substack{i=1 \\ i \neq j}}^m a_{1i}^2$$

Using (6.3.1.5) obtain

$$\frac{a_{2j}^2}{a_{1j}^2} = 1 + \sum_{\substack{i=1 \\ i \neq j}}^m \left(\frac{a_{1i}^2 - a_{2i}^2}{a_{1j}^2} \right) < \frac{\sum_{\substack{i=1 \\ i \neq j}}^m a_{2i}^2}{\sum_{\substack{i=1 \\ i \neq j}}^m a_{1i}^2}$$

$$\text{Rearrange as} \quad \left(\frac{1}{a_{1j}^2} + \frac{1}{\sum_{\substack{i=1 \\ i \neq j}}^m a_{1i}^2} \right) \left(\sum_{\substack{i=1 \\ i \neq j}}^m (a_{1i}^2 - a_{2i}^2) \right) < 0$$

$$\text{Therefore} \quad \sum_{\substack{i=1 \\ i \neq j}}^m (a_{1i}^2 - a_{2i}^2) < 0 \quad \text{giving} \quad a_{1j}^2 > a_{2j}^2$$

A sufficient condition for $J_{D_j}(q_1) - J_{D_j}(q_2) > 0$ is

$$\sum_{\substack{i=1 \\ i \neq j}}^m (a_{2i} - a_{1i}) > 0 ; \quad \text{clearly such a condition is not necessary.}$$

This condition is found by bounding $f_k = \sum_{\substack{i=1 \\ i \neq j}}^m a_{ki}$ subject to

$$\sum_{\substack{i=1 \\ i \neq j}}^m a_{ki}^2 = \mu^2 - a_{kj}^2 = \alpha_k^2 > 0, \quad k=1,2, \quad \text{and then using the bounds to}$$

ensure $f_2 - f_1 > 0$.

Without loss of generality assume $j=m$, and write

$$a_{k1} = \alpha_k \cos \theta_1$$

$$a_{ki} \geq 0 \quad \forall i=1,2,\dots,m-1$$

$$a_{k2} = \alpha_k \sin \theta_1 \cos \theta_2$$

$$a_{k3} = \alpha_k \sin \theta_1 \sin \theta_2 \cos \theta_3$$

$$\text{hence } 0 \leq \theta_i \leq \pi/2 \quad i=1,\dots,m-1.$$

⋮

$$a_{km-1} = \alpha_k \sin \theta_1 \sin \theta_2 \dots \sin \theta_{m-2}$$

$$\text{Thus } f_k = \alpha_k [\cos \theta_1 + \sin \theta_1 \cos \theta_2 + \dots + \sin \theta_1 \sin \theta_2 \dots \sin \theta_{m-2}]$$

$$\text{where } \sum_{i=1}^{m-1} a_{ki}^2 = \alpha_k^2 \quad \text{and} \quad 0 \leq \theta_i \leq \pi/2$$

Search for unconstrained extrema

$$\frac{\partial f_k}{\partial \theta_1} = \alpha_k [-\sin \theta_1 + \cos \theta_1 \{ \cos \theta_2 + \sin \theta_2 \cos \theta_3 + \dots + \sin \theta_2 \dots \sin \theta_{m-2} \}]$$

$$\frac{\partial f_k}{\partial \theta_2} = \alpha_k \sin \theta_1 [-\sin \theta_2 + \cos \theta_2 \{ \cos \theta_3 + \dots + \sin \theta_3 \dots \sin \theta_{m-2} \}]$$

⋮

$$\frac{\partial f_k}{\partial \theta_{m-2}} = \alpha_k \sin \theta_1 \sin \theta_2 \dots \sin \theta_{m-3} [-\sin \theta_{m-2} + \cos \theta_{m-2}]$$

The condition of optimality assuming $\theta_1 \neq \theta_2 \neq \theta_3 \dots \neq \theta_{m-2} \neq 0$

$$\text{giving } \sin \theta_{m-2} = \cos \theta_{m-2} \quad \text{and hence } \hat{\theta}_{m-2} = \pi/4$$

Back substitution yields $\sin \theta_i = (m-1-i)^{1/2} \cos \theta_i$ hence

$$\tan \theta_i = (m-1-i)^{1/2} \quad i=m-2, m-3, \dots, 1$$

$$\text{Giving } \sin \theta_i = (m-1-i)^{1/2} / (m-i)^{1/2} \quad \text{and} \quad \cos \theta_i = (m-i)^{-1/2}$$

$$\text{so that } f_k(\hat{\theta}_1, \dots, \hat{\theta}_{m-1}) = (m-1)^{1/2} \alpha_k.$$

An examination of the low order cases shows this to be a maximum.

The minimum is obtained by activating the boundary constraints either

$\theta_i = 0$ or $\theta_i = \pi/2$, $i \in 1, 2, \dots, m-1$. The minimum function value is

$$f_k(\min) = \alpha_k. \quad \text{Have therefore}$$

$$\alpha_k \leq f_k = \sum_{\substack{i=1 \\ i \neq j}}^m a_{ki} \leq (m-1)^{\frac{1}{2}} \alpha_k \quad k=1,2.$$

Consequently a sufficient condition to ensure $f_2 - f_1 \geq 0$ is

$$(m-1)^{\frac{1}{2}} \alpha_1 < \alpha_2. \quad \text{Substituting for } \alpha_k = \mu^2 - \alpha_{kj}^2, \quad k=1,2$$

obtain the sufficiency condition as :

$$a_{1j}^2 \geq \frac{a_{2j}^2}{(m-1)^{\frac{1}{2}}} + \mu^2 \left(1 - \frac{1}{(m-1)^{\frac{1}{2}}} \right)$$

Hence sufficient conditions have been given to indicate when iterates from the minimization of $\tan^2 \theta(q)$ are maximizing the actual dominance cost.