<table>
<thead>
<tr>
<th>Title</th>
<th>Authors/Editors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improperly posed boundary value problems</td>
<td>A Carasso and A P Stone</td>
</tr>
<tr>
<td>Lie algebras generated by finite dimensional ideals</td>
<td>J N Stewart</td>
</tr>
<tr>
<td>Bifurcation problems in nonlinear elasticity</td>
<td>R W Dickey</td>
</tr>
<tr>
<td>Partial differential equations in the complex domain</td>
<td>D L Colton</td>
</tr>
<tr>
<td>Quasilinear hyperbolic systems and waves</td>
<td>A Jeffrey</td>
</tr>
<tr>
<td>Solution of boundary value problems by the method of integral operators</td>
<td>D L Colton</td>
</tr>
<tr>
<td>Taylor expansions and catastrophes</td>
<td>T Poston and I N Stewart</td>
</tr>
<tr>
<td>Function theoretic method in differential equations</td>
<td>R P Gilbert and R J Weinacht</td>
</tr>
<tr>
<td>Differential topology with a view to application</td>
<td>D R J Chillingworth</td>
</tr>
<tr>
<td>Characteristic classes of foliations</td>
<td>H V Pittie</td>
</tr>
<tr>
<td>Stochastic integration and generalized martingales</td>
<td>A U Kussmaul</td>
</tr>
<tr>
<td>Zeta-functions: A introduction of algebraic geometry</td>
<td>A D Thomas</td>
</tr>
<tr>
<td>Explicit a priori inequalities with applications to boundary value problems</td>
<td>V G Sigillito</td>
</tr>
<tr>
<td>Nonlinear diffusion</td>
<td>W E Fitzgibbon III and H F Walker</td>
</tr>
<tr>
<td>Unsolved problems concerning lattice points</td>
<td>J Hammer</td>
</tr>
<tr>
<td>Edge-colourings of graphs</td>
<td>S Fiorini and R J Wilson</td>
</tr>
<tr>
<td>Nonlinear analysis and mechanics: Heriot-Watt Symposium Volume I</td>
<td>R J Knops</td>
</tr>
<tr>
<td>Actions of fine abelian groups</td>
<td>C Kosniowski</td>
</tr>
<tr>
<td>Closed graph theorems and webbed spaces</td>
<td>M De Wilde</td>
</tr>
<tr>
<td>Singular perturbation techniques applied to integro-differential equations</td>
<td>H Grabmüller</td>
</tr>
<tr>
<td>Retarded functional differential equations: A global point of view</td>
<td>S E A Mohammed</td>
</tr>
<tr>
<td>Multiparameter spectral theory in Hilbert space</td>
<td>B D Sleeman</td>
</tr>
<tr>
<td>Mathematical modelling techniques</td>
<td>R Aris</td>
</tr>
<tr>
<td>Singular points of smooth mappings</td>
<td>C G Gibson</td>
</tr>
<tr>
<td>Nonlinear evolution equations solvable by the spectral transform</td>
<td>F Calogero</td>
</tr>
<tr>
<td>Nonlinear analysis and mechanics: Heriot-Watt Symposium Volume II</td>
<td>R J Knops</td>
</tr>
<tr>
<td>Constructive functional analysis</td>
<td>D S Bridges</td>
</tr>
<tr>
<td>Elongational flows: Aspects of the behaviour of model elasticoviscous fluids</td>
<td>C J S Petrie</td>
</tr>
<tr>
<td>Nonlinear analysis and mechanics: Heriot-Watt Symposium Volume III</td>
<td>R J Knops</td>
</tr>
<tr>
<td>Fractional calculus and integral transforms of generalized functions</td>
<td>A C McBride</td>
</tr>
<tr>
<td>Complex manifold techniques in theoretical physics</td>
<td>D E Lerner and P D Sommers</td>
</tr>
<tr>
<td>Hilbert’s third problem: scissors congruence</td>
<td>C H Sah</td>
</tr>
<tr>
<td>Graph theory and combinatorics</td>
<td>R J Wilson</td>
</tr>
<tr>
<td>The Tricomi equation with applications to the theory of plane transonic flow</td>
<td>A R Manwell</td>
</tr>
<tr>
<td>Abstract differential equations</td>
<td>S D Zaidman</td>
</tr>
<tr>
<td>Advances in twistor theory</td>
<td>L P Hughston and R S Ward</td>
</tr>
<tr>
<td>Operator theory and functional analysis</td>
<td>I Erdely</td>
</tr>
<tr>
<td>Nonlinear analysis and mechanics: Heriot-Watt Symposium Volume IV</td>
<td>R J Knops</td>
</tr>
<tr>
<td>Singular systems of differential equations</td>
<td>S L Campbell</td>
</tr>
<tr>
<td>N-dimensional crystallography</td>
<td>R L E Schwarzenberger</td>
</tr>
<tr>
<td>Nonlinear partial differential equations in physical problems</td>
<td>D Graffi</td>
</tr>
<tr>
<td>Shifts and periodicity for right invertible operators</td>
<td>D Przeworska-Rolewicz</td>
</tr>
<tr>
<td>Rings with chain conditions</td>
<td>A W Chatters and C R Hajarnavis</td>
</tr>
<tr>
<td>Moduli, deformations and classifications of complex manifolds</td>
<td>D Sundararaman</td>
</tr>
<tr>
<td>Nonlinear problems of analysis in geometry and mechanics</td>
<td>M Atteia, D Bancel and I Gumowski</td>
</tr>
<tr>
<td>Algorithmic methods in optimal control</td>
<td>W A Gruver and E Sachs</td>
</tr>
<tr>
<td>Abstract Cauchy problems and functional differential equations</td>
<td>F Kappel and W Schappacher</td>
</tr>
<tr>
<td>Sequence spaces</td>
<td>W H Ruckle</td>
</tr>
<tr>
<td>Recent contributions to nonlinear partial differential equations</td>
<td>H Berestycki and H Brezis</td>
</tr>
</tbody>
</table>

See inside back cover for further titles
Singular systems of differential equations II
S L Campbell

Singular systems of differential equations II

π

Pitman Advanced Publishing Program
SAN FRANCISCO · LONDON · MELBOURNE
Singular systems of differential equations.

(Remark notes in mathematics; 40, 61)
Includes indexes.
I. Title. II. Series: Research notes in mathematics; 40, [etc.]
QA371.C192 515.3'.5 79-20908
ISBN 0-8224-8438-2 (v. 1) AACR1
ISBN 0-273-08516-6 (v. 2)

British Library Cataloguing in Publication Data
Campbell, S. L.
Singular systems of differential equations II.
- (Research notes in mathematics; v. 61)
I. Title. II. Series
515.3'52 QA371

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system, or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording and/or
otherwise without the prior written permission of the publishers.
The paperback edition of this book may not be lent, resold, hired out
or otherwise disposed of by way of trade in any form of binding or
cover other than that in which it is published, without the prior
consent of the publishers.

Reproduced and printed by photolithography
in Great Britain by Biddles Ltd, Guildford
To Gail with love.
## Contents

0. Background Material

1. Introduction 1
   2. Matrix Theory Concepts and Notations 3

1. Singular Linear Control Theory

1. Introduction 13
   2. Unbounded Admissible Controls 13
   3. Controllability and Observability 20
   4. Stabilizability 22
   5. Time Optimal Control 27
   6. Cheap Control and Nonsymmetric Riccati Equations 31

2. Impulsive Behavior of Linear Systems

1. Introduction 46
   2. Distributions 47
   3. Distributions as Solutions 52
   4. Singular Perturbations 56
   5. Feedback Control of Impulses 64

3. Circuits and Circuit Models

1. Introduction 73
   2. Basic Laws 74
3. Linear Circuits 76
4. Nonlinear Circuits 85

4. Numerical Solution of Autonomous Linear Systems
1. Introduction 90
2. First Order Methods 91
3. Higher Order Methods 98
4. Starting Values and Variable Time Step Methods 101
5. Numerical Examples 107
6. Comment on $Ax = b$ 111

5. Linear Time Varying Systems
1. Introduction 114
2. The Index of a Linear Time Varying System 115
3. The Index One Case 119
4. Index Greater Than One 124
5. Numerical Solution of Linear Time Varying Systems 134

6. Nonlinear Singular Systems
1. Introduction 147
2. General Concepts 148
3. Index One Systems 153
4. Higher Order Systems 154
5. Linear Subsystems 164
6. Hysteresis and Impulses 178
7. Infinite Dimensional Linear Systems

1. Introduction 190
2. Systems with Infinite Matrices 191
3. Resolvent Techniques 205
4. A Geometric Approach 210

REFERENCES 220

INDEX 229

LIST OF SYMBOLS 232
0 Background material

1. INTRODUCTION

This book is primarily concerned with the differential equation

\[ \dot{A}x + B(x) = f \]  

(1.1)

where \( x \) is a vector valued function of the real variable \( t \), \( A \) is a matrix, and \( B \) is a vector valued function. We are particularly interested in what happens if \( A \) is, in some sense, singular. Systems of this type are referred to by a variety of names. Among them are: singular, semi-state, degenerate, descriptor, constrained, and differential-algebraic.

This book is a continuation of [8]. In [8] we studied the linear version of (1);

\[ \dot{A}x + Bx = f \]  

(1.2)

where \( A, B \) were matrices, independent of \( t \), but sometimes dependent on a small parameter \( \varepsilon \). The emphasis in [8] was on obtaining explicit solutions and applying them to various problems. The emphasis in several parts of this book will be quite different. As to be expected from the fact that (1) is nonlinear, there will be more concern with analytic and numerical procedures. The development will be independent of [8] in the sense that results needed from [8] will be carefully restated. However, proofs, examples, discussion, and motivation will not be repeated. The two books thus complement each other.
In this book, (1) is viewed as a dynamical system and the results are usually derived in terms of some type of state space. Chapter 2 is a partial exception. There are other approaches to (2) which have proved fruitful. The work of Verghese [110] involving transfer functions is a good example as is the work of Gohberg, et. al., [56] on matrix polynomials. However, these approaches do not extend as readily to (1) as the state variable, dynamical system approach does.

The remainder of this chapter is a quick review of the notation from [8]. Chapter 1 deals with the extension of several control theory concepts to linear control problems in the form (2). Besides their intrinsic interest, these results serve as a good way to introduce several important concepts and prepare the reader, who has not read [8], for what is to follow. The results of Chapter 1 do not appear in [8].

In [8], distributional solutions were derived for (2). Chapter 2 discusses in what sense these distributions are to be interpreted as solutions. This interpretation involves singular perturbations.

Chapter 3 presents a quick coverage of some basic concepts from circuit theory. They will be used in both applications and to motivate some of the results.

In order to understand the numerical solutions of (1), it is helpful to first discuss the numerical solution of (2). This is done in Chapter 4. Chapters 5 and 6 cover linear time varying and nonlinear systems respectively.

Chapter 7 deals with both infinite systems in the form (2) and with (2) when A,B are linear operators. These two topics are closely related but not identical.
Formulas are numbered by section. Theorems, lemmas, etc., are labeled by chapter and section. If referred to within the section in which it appears, just the last number is used. If it appears in a different section of the same chapter, the section number is included. If it is in a different chapter, the chapter number is also given. Thus XXXXX 2.3.4 refers to the fourth XXXXX in Section 3 of Chapter 2 whereas XXXXX 3.4 refers to the fourth XXXXX in Section 3 of the current chapter.

The author would like to acknowledge the contributions of several people to this volume. N. J. Rose, R. Newcomb, B. Cobb, A. Favini, L. Petzold, E. Yip, and B. Francis all influenced the author either through their published work or in discussions. The author is especially grateful to R. Newcomb, who through extensive correspondence, was instrumental in the author's introduction to non-linear circuits. M. D. Luu, N. J. Rose, R. Newcomb, and K. Clark read and commented on portions of the original manuscript. Most of the numerical results were programmed and run by K. Clark. Sharon Jones typed the final manuscript.

Finally, some of the research for this volume, particularly that in Chapters 5 and 6 was sponsored by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant No. AFOSR-81-0052. This material may be found in [14], [15], [16], [17], [18], [19], [20], where it is available for United States Government reproduction and distribution of reprints.

2. MATRIX THEORY CONCEPTS AND NOTATIONS

This section is an abbreviated form of Chapter 1 in [8]. We include it in order to make the presentation in this book more self-contained.

The mathematical prerequisites for reading this book are a working knowledge of linear algebra and matrix theory, real analysis, and a familiarity
with the basic theory of systems of ordinary differential equations.

Except for Chapter 7, the notation is consistent throughout the book.

The set of complex numbers is denoted by \( \mathbb{C} \); the set of real numbers by \( \mathbb{R} \). The set of \( m \times n \) matrices over \( \mathbb{C} \) is denoted by \( \mathbb{C}^{m \times n} \); the \( m \times n \) matrices over \( \mathbb{R} \) by \( \mathbb{R}^{m \times n} \). Unless stated otherwise, all matrices will be in \( \mathbb{C}^{m \times n} \).

The column vectors in the vector space \( \mathbb{C}^n \) will be denoted by \( u, x \) etc.

If \( A \in \mathbb{C}^{m \times n} \), \( A^* \) is the conjugate transpose of \( A \), \( A^T \) is the transpose of \( A \).

For vectors \( x, y \in \mathbb{C}^n \) the inner product of \( x \) and \( y \) is \( (x, y) = y^*x \). The norm of a vector \( x \in \mathbb{C}^n \) is the euclidean norm, \( ||x|| = (x, x)^{1/2} \). For matrices \( A \in \mathbb{C}^{m \times n} \), unless stated otherwise, the operator norm, \( ||A|| = \sup \{||Ax|| : ||x|| = 1\} \), is used.

If \( M \) is a subspace of \( \mathbb{C}^n \), (subspace always means linear subspace) \( \dim M \) denotes the dimension of \( M \). If \( A \in \mathbb{C}^{m \times n} \), the range (column space) of \( A \) is denoted by \( R(A) \) and the null space of \( A \), \( \{x : Ax = 0\} \), by \( N(A) \). Recall that \( \dim N(A) + \dim R(A) = n \).

Let \( M_1, \ldots, M_s \) be subspaces of \( \mathbb{C}^n \). The sum of these subspaces is the subspace, \( M_1 + \ldots + M_s = \{z = x_1 + \ldots + x_s : x_i \in M_i\} \). If \( M_i \cap M_j = \{0\} \) for \( i \neq j \), the subspaces are said to be independent; the sum is then called a direct sum and we write \( M_1 \oplus \ldots \oplus M_s \). Recall that \( \dim (M_1 \oplus \ldots \oplus M_s) = \dim M_1 + \ldots + \dim M_s \) and if \( x \in M_1 \oplus \ldots \oplus M_s \), then there exist unique \( x_i \in M_i \) such that \( x = x_1 + \ldots + x_s \).

A projection is a matrix \( P \in \mathbb{C}^{n \times n} \) such that \( P^2 = P \). It is easily seen that \( R(P) \oplus N(P) = \mathbb{C}^n \). Conversely if \( \mathbb{C}^n = M \oplus N \), there exists a unique projection \( P \) such that \( R(P) = M \) and \( N(P) = N \). This projection is denoted \( P_{M,N} \), and called the projection onto \( M \) along \( N \). If \( M \) is a subspace of \( \mathbb{C}^n \), the orthogonal complement of \( M \) is \( M^\perp = \{x \in \mathbb{C}^n : (x, y) = 0 \text{ for all} \}

\( y \in \mathbb{M} \). \( M^\perp \) is a subspace and \( M \oplus M^\perp = \mathbb{R}^n \). \( P_M, M^\perp \) is denoted by \( P_M \) and called the orthogonal projection onto \( M \).

We shall often make use of block matrices; that is, a matrix whose entries are matrices. In particular, if \( A \in \mathbb{R}^{n \times n} \) is block diagonal, that is \( A \) has blocks \( A_1, \ldots, A_s \) along the main diagonal and zero blocks elsewhere, we write \( A = \text{diag}(A_1, \ldots, A_s) \).

The determinant of a matrix is denoted \( \det A \). The eigenvalues of \( A \in \mathbb{C}^{n \times n} \) are the roots of the polynomial \( \det(\lambda I - A) \). The spectrum of \( A \) is the set of eigenvalues of \( A \) and is denoted by \( \sigma(A) \). The spectral radius of \( A \) is \( \rho(A) = \sup \{ |\lambda| : \lambda \in \sigma(A) \} \).

If \( B^2 = A \), \( B \) is called a square root of \( A \) and is denoted \( A^{1/2} \). A matrix \( A \in \mathbb{C}^{n \times n} \) is positive semi-definite if \( (Ax, x) \geq 0 \) for all \( x \in \mathbb{C}^n \). If \( A \) is positive semi-definite, it has a unique positive semi-definite square root.

If \( A, B \in \mathbb{C}^{n \times n} \), \( A \) is similar to \( B \) in case there exists a nonsingular matrix \( T \) such that \( A = TBT^{-1} \). Similar matrices represent the same linear transformation or operator on \( \mathbb{C}^n \) but with respect to different bases.

Frequent use will be made of the fact that every \( A \in \mathbb{C}^{n \times n} \) is similar to a matrix in Jordan Canonical form, that is \( A \) is similar to

\[
\begin{bmatrix}
\lambda_1 & 1 & 0 & \cdots \\
0 & \lambda_2 & 1 & \cdots \\
\vdots & \ddots & \ddots & \ddots \\
0 & \cdots & 0 & \lambda_s
\end{bmatrix}
\]

where \( J_i = \text{diag}(J_1, \ldots, J_s) \) where \( J_i = 
\begin{bmatrix}
\lambda_i & 1 & 0 & \cdots \\
0 & \lambda_i & 1 & \cdots \\
\vdots & \ddots & \ddots & \ddots \\
0 & \cdots & 0 & \lambda_i
\end{bmatrix}
\]

such block corresponding to the eigenvalue \( \lambda_i \).

If \( A \in \mathbb{C}^{n \times n} \), the index of \( A \), written \( \text{Ind}(A) \), is the least nonnegative integer \( \nu \) such that \( N(A^\nu) = N(A^{\nu+1}) \) (by convention \( A^0 = I \)). Since \( N(A^0) = \{0\} \subset N(A) \subset N(A^2) \cdots \subset N(A^n) \subset \mathbb{C}^n \) it is clear that \( \nu = \text{Ind}(A) \).
exists and $0 \leq u \leq n$. If $A$ is nonsingular, $\text{Ind}(A) = 0$, while if $A$ is nilpotent of index $u(A^u = 0, A^{u-1} \neq 0)$, then $\text{Ind}(A) = u$. The index could also be defined as the least nonnegative integer such that $R(A^u) = R(A^{u+1})$.

If $A \in \mathbb{C}^{n \times n}$, $\text{Ind}(A) = u$, $\dim R(A^u) = s$, $\dim N(A^u) = t (s + t = n)$, then there exists a nonsingular matrix $T$ such that

$$A = T \begin{bmatrix} C & 0 \\ 0 & N \end{bmatrix} T^{-1}$$

where $C$ is an $s \times s$ nonsingular matrix and $N$ is a $t \times t$ nilpotent matrix with $u = \text{Ind}(N)$. The form (1) follows immediately from the Jordan Canonical form. It is also easily proved directly, see [8, p. 7].

The Drazin inverse is convenient for summarizing certain spectral information. If $A$ is given by (1), then the Drazin inverse of $A$, is defined by

$$A^D = T \begin{bmatrix} C^{-1} & 0 \\ 0 & 0 \end{bmatrix} T^{-1}.$$

Note that if $A$ is nonsingular, the $N$ block is missing in (1) and $A^D = A^{-1}$ while if $A$ is nilpotent, the $C$ block is missing in (1) and $A^D = 0$. It can be shown that the Drazin inverse is unique. The following properties are useful.

If $A \in \mathbb{C}^{n \times n}$ and $\text{Ind}(A) = u$, then

$$AA^D = A^DA, A^DAA^D = A^D, A^{k+1}A^D = A^k \quad \text{for } k \geq \text{Ind}(A)$$

$$A^k(I - AA^D) = 0, k \geq \text{Ind}(A); A^k(I - AA^D) \neq 0, k < \text{Ind}(A)$$

$$AA^D = P_{R(A^u), N(A^u)}, I - AA^D = P_{N(A^u), R(A^u)}.$$

The properties (2) completely characterize $A^D$ and are often used as the definition of $A^D$. 

6
Note also that \( N(A^U) = N(A^D) \) and \( R(A^U) = R(A^D) \). In addition \( x \in N(A^U) \) if and only if \( AA^D x = 0 \) which holds if and only if \( x = (I - AA^D)x \).

In case \( \text{Ind}(A) = 0 \) or \( 1 \), the Drazin inverse of \( A \) is often called the group inverse and denoted by \( A^\# \). The group inverse has the additional property that \( \overrightarrow{A^\# A} = A \).

For \( A \in \mathbb{C}^{n \times n} \), let its characteristic polynomial be \( c(\lambda) = (\lambda - \lambda_1)^{m_1}(\lambda - \lambda_2)^{m_2} \cdots (\lambda - \lambda_s)^{m_s} \) where the eigenvalues \( \lambda_i \) are distinct and \( m_1 + m_2 + \cdots + m_s = n \). Let \( v_i = \text{Ind}(\lambda_i I - A) \) and \( N_i = N((\lambda_i I - A)^{\nu}) \).

Then \( N_i \) is an invariant subspace for \( A \), \( \dim N_i = m_i \), and \( E_i = I - (\lambda_i I - A)(\lambda_i I - A)^D \) is a projection on \( N_i \) such that \( E_i E_j = E_j E_i \).

If \( f \) is an analytic function of \( \lambda \in \mathbb{C} \) for \( |\lambda| < R \) and \( p(A) < R \), then \( f(A) \) exists and

\[
f(A) = \sum_{i=1}^{s} \sum_{k=0}^{\nu_i-1} \frac{f^{(k)}(\lambda_i)}{k!} (A - \lambda_i I)^k E_i. \tag{2.5}
\]

If \( f(\lambda) = \sum_{n=0}^{\infty} a_n \lambda^n \) for \( |\lambda| < R \), then (5) coincides with \( f(A) = \sum_{n=0}^{\infty} a_n A^n \).

The most important matrix function for our purposes is the matrix exponential. Since \( e^A = \sum_{n=0}^{\infty} A^n/n! \) converges for all \( \lambda \), \( e^A \) exists for every square matrix and \( e^A = \sum_{i=1}^{s} \sum_{k=0}^{\nu_i-1} \frac{e^{\lambda_i}}{k!} (A - \lambda_i I)^k E_i. \) \( e^A \lambda = 1 \) implies that \( e^{-A} = 1. \)

Thus \( e^A \) is invertible and \((e^A)^{-1} = e^{-A} \). For \( t \in \mathbb{R} \),

\[
e^{At} = \sum_{i=1}^{s} \sum_{k=0}^{\nu_i-1} \frac{e^{\lambda_i t}}{k!} (A - \lambda_i I)^k E_i. \tag{2.6}
\]

Sometimes it is helpful to use representations of matrix functions by contour integrals. Recall that if \( f(z) \) is analytic in and on a simple closed rectifiable curve or contour, \( C \), then \( \int_C f(z)dz = 0. \) Also if \( z \) is in the interior of \( C \), then
\[ f(z) = \frac{1}{2\pi i} \int_{C} \frac{f(\lambda)}{\lambda - z} \, d\lambda \quad \text{and} \quad f^{(k)}(z) = \frac{k!}{2\pi i} \int_{C} \frac{f(\lambda)}{(\lambda - z)^{k+1}} \, d\lambda. \quad (2.7) \]

Similar representations hold for functions of a matrix. If \( A \in \mathbb{F}^{n \times n} \), the matrix function \( (\lambda I - A)^{-1} \) is called the \textit{resolvent} of \( A \). It is analytic for \( \lambda \notin \sigma(A) \).

If \( A \in \mathbb{F}^{n \times n} \) and \( f \) is analytic in an open set \( \Omega \) containing \( \sigma(A) \), then
\[ f(A) = \frac{1}{2\pi i} \int_{C} f(\lambda)(\lambda I - A)^{-1} \, d\lambda \] where \( C \) is a contour lying in \( \Omega \) and enclosing all the eigenvalues of \( A \). In particular, a square root can be defined for any index one matrix.

If \( \sigma(A) = \sigma_1 \cup \sigma_2 \) where \( \sigma_1, \sigma_2 \) are disjoint sets of eigenvalues and \( C_1 \) is a contour containing \( \sigma_1 \) in its interior and \( \sigma_2 \) in its exterior, and \( f \) is analytic on a connected open set containing \( \sigma_1 \), then
\[ \frac{1}{2\pi i} \int_{C_1} f(\lambda)(\lambda I - A)^{-1} \, d\lambda = f(A)P = f(AP)P \] where \( P \) is the projection
\[ P = \sum_{\lambda \in \sigma_1} (I - (\lambda I - A)(\lambda I - A)^{D}) = \frac{1}{2\pi i} \int_{C_1} (\lambda I - A)^{-1} \, d\lambda. \]

Occasionally we shall talk about the impulsive behavior of systems. A convenient way to do so is with \textit{delta functions}.

The delta function \( \delta(t) \) and its \( i^{th} \) derivatives, \( \delta^{(i)}(t) \), are not actually functions, but rather objects in a larger space to which the operations of differentiation, integration, and taking Laplace transforms may be extended. It is not our intention at this point to carefully define delta functions but rather just state the properties they have. Section 2.2 has some additional information.

\( \delta(t) \) has the following properties. For \( t \neq 0 \), it is a "function" and is identically zero. \( \delta(0) \) is undefined. However,
\[ \int_{a}^{b} f(t) \delta(t) \, dt = f(0) \]  

(2.8)

for any function \( f \) that is continuous at zero and integrable on the interval \([a,b]\) with \( 0 \in [a,b] \). In particular,

\[
\int_{a}^{t} \delta(s) \, ds = h(t) = \begin{cases} 
0 & \text{if } t < 0 \\
1 & \text{if } t \geq 0
\end{cases}
\]  

(2.9)

Thus \( \delta(t) \) can be viewed as the derivative of the \( h(t) \) in (9).

The Laplace transform of \( \delta \) is \( L[\delta] = \int_{0}^{\infty} e^{-st} \delta(t) \, dt = 1 \). For integers \( i \geq 1 \), \( \delta^{(i)}(t) \) is the \( i \)-th derivative of \( \delta(t) \). It is also a function for \( t \neq 0 \) that is identically zero. In addition if \( f(t) \) is \( i \)-times continuously differentiable at 0, and \( 0 \in [a,b] \), then

\[
\int_{a}^{b} f(t) \delta^{(i)}(t) \, dt = (-1)^{i} f^{(i)}(0).
\]  

(2.10)

Similarly, \( L[\delta^{(i)}] = \int_{0}^{\infty} e^{-st} \delta^{(i)}(t) \, dt = s^{i} \).

Integration by parts holds with the proper interpretation. Also if \( c \) is a real number, \( f \) is \( i \)-times differentiable at \( c \), and \( c \in [a,b] \), then

\[
\int_{a}^{b} f(t) \delta^{(i)}(t - c) \, dt = (-1)^{i} f^{(i)}(c).
\]

Intuitively, one thinks of \( \delta(t - c) \) as a "sudden impulse" at time \( c \). It can be shown that, in the appropriate sense, \( \delta(t) \) is actually the limit of functions \( f_{n} \) such that \( \int_{-\infty}^{\infty} f_{n}(t) \, dt = 1, \ f_{n}(t) = 0 \text{ if } |t| > 1/n \).

For example, we could take \( f_{n}(t) = 0 \text{ if } t \notin [0,1/n] \) and \( f_{n}(t) = n \text{ for } 0 \leq t \leq 1/n \). For \( n \) large, these \( f_{n} \) look very much like what one would envision a "sudden impulse" to be. (See Figure 2.2.1.)
The remainder of this section reviews the basic results from [8] needed for later chapters. These results concern differentiable solutions of

\[ A\dot{x} + Bx = f. \]  

(2.11)

**Definition 0.1.1** For \( A, B \in \mathbb{R}^{n \times n} \) and \( t_0 \in \mathbb{R} \), the vector \( \xi \in \mathbb{R}^n \) is said to be a **consistent initial vector** associated with \( t_0 \) for (2) if (2) possesses at least one solution. Equation (1) is said to be **solvable at the point** \( t_0 \) if the initial value problem (2) has a unique solution for each consistent initial vector, \( \xi \), associated with \( t_0 \).

The term tractable was used in [8] instead of solvable.

**Theorem 0.1.1** For \( A, B \in \mathbb{R}^{n \times n} \), the homogeneous differential equation

\[ A\dot{x}(t) + Bx(t) = 0 \]

is solvable if and only if there exists a scalar \( \lambda \in \mathbb{R} \) such that \((\lambda A + B)^{-1}\) exists.

A useful technical theorem is

**Theorem 0.1.2** Suppose that \( A, B \in \mathbb{R}^{n \times n} \) are such that there exists a \( \lambda \in \mathbb{R} \) so that \((\lambda A + B)^{-1}\) exists. Let \( \hat{A}_\lambda = (\lambda A + B)^{-1}A \), \( \hat{B}_\lambda = (\lambda A + B)^{-1}B \), and \( \hat{f}_\lambda = (\lambda A + B)^{-1}f \) for \( f \in \mathbb{R}^n \). For all \( \alpha, \mu \in \mathbb{R} \) for which \((\alpha A + B)^{-1}\) and \((\mu A + B)^{-1}\) exist, the following statements are true.

\[ \hat{A}_\alpha \hat{A}_\mu = \hat{A}_{\alpha + \mu} \]  

\[ \hat{A}_\alpha \hat{D}_\mu = \hat{A}_{\alpha} \hat{D}_{\mu} \]  

and \( \hat{A}_\alpha \hat{B}_\mu = \hat{A}_{\alpha} \hat{B}_{\mu} \).

* Ind(\( \hat{A}_\alpha \)) = Ind(\( \hat{A}_\mu \)) and R(\( \hat{A}_\alpha \)) = R(\( \hat{A}_\mu \)).
\[ \hat{A}_a D_{\alpha}^e = \hat{A}_\mu D_{\mu}^e. \]
\[ \hat{B}_a D_{\alpha}^e = \hat{B}_\mu D_{\mu}^e. \]

In view of the preceding theorem, we can drop the subscript \( \lambda \) whenever the terms \( \hat{A}_\lambda A_{\lambda}^D, \hat{A}_\lambda B_{\lambda}^D, R(A), \) Ind(\( \hat{A}_\lambda \)), \( \hat{A}_\lambda D_{\lambda}^e \), and \( \hat{B}_\lambda D_{\lambda}^e \) appear. We shall do so.

The key theorem from [8] is

**Theorem 0.1.3** Suppose \( A \dot{x}(t) + Bx(t) = 0 \) is solvable. Then the general solution is given by

\[ x(t) = e^{-A^D(t-t_0)} \int_{\lambda} A A^\lambda q, q \in \mathbb{C}^n. \]

A vector \( c \in \mathbb{C}^n \) is a consistent initial vector for the homogeneous equation if and only if \( c = \hat{A} A^D c \) \( (c \in R(\hat{A}) = R(A^D) \) where \( k = \text{Ind}(\hat{A}) \).

Suppose that \( f(t) \) is \( k \)-times continuously differentiable around \( t_0 \). Then the nonhomogeneous equation \( \dot{x}(t) + Bx(t) = f(t) \) always possesses solutions and a particular solution is given by

\[ x(t) = e^{-A^D(t-t_0)} \int_{t_0}^t e^{A^D(s)} f(s) + (I - \hat{A} A^D) \sum_{i=0}^{k-1} (-1)^i [A B^D] f^i(s)(t). \]

(2.12)

The general solution is given by

\[ x(t) = e^{-A^D(t-t_0)} \int_{t_0}^t e^{A^D(s)} f(s) ds \]
\[ + (I - \hat{A} A^D) \sum_{i=0}^{k-1} (-1)^i [A B^D] f^i(s)(t), \]

11
where \( q \in \mathbb{C}^n \). Let \( \hat{w} = (I - \hat{A}^D)^k \sum_{i=0}^{k-1} (-1)^i \hat{A}^D i \hat{B}^D \hat{f}(1)(t) \). Then \( \hat{w} \) is independent of \( \lambda \). A vector \( c \in \mathbb{C}^n \) is a consistent initial vector associated with \( t_0 \in \mathbb{R} \) for the inhomogeneous equation if and only if \( c \) solves
\[
(I - \hat{A}^D)(c - \hat{w}(0)) = 0.
\]
Furthermore, the inhomogeneous equation is solvable at \( t_0 \) and the unique solution of the initial value problem with \( x(t_0) = c \), \( c \) a consistent initial vector associated with \( t_0 \), is given by (12) with \( q = c \).

In Chapter 4, we shall need the following discrete analogue of Theorem 1 [8, p. 69].

**Theorem 0.1.4** Suppose that \( A, B \in \mathbb{C}^{n \times n} \) and there exists \( \lambda \) such that
\( \lambda A + B \) is nonsingular. Set \( \hat{A} = (\lambda A + B)^{-1}A \), \( \hat{B} = (\lambda A + B)^{-1}B \), and
\( \hat{f} = (\lambda A + B)^{-1}f \). Then all solutions of
\[
Ax_{i+1} + Bx_i = \hat{f}_i, \quad i = 0, \ldots, N - 1
\]
are given by \( \hat{A}^D Ax_0 \) arbitrary,
\[
\hat{A}^D Ax_i = (-\hat{A}^D)^i Ax_0 + \sum_{\ell=0}^{i-1} (-\hat{A}^D)^\ell \hat{A}^D \hat{f}_{i-\ell-1}
\]
for \( i \geq 1 \) and
\[
(I - \hat{A}^D)x_i = (-\hat{A}^D)^{N-i}(I - \hat{A}^D)x_N + (I - \hat{A}^D)^{N-i-1} \sum_{\ell=0}^{N-i-1} (-\hat{A}^D)^\ell \hat{B}^D \hat{f}_{i+\ell}
\]
for \( i \geq 0 \). Note that \( (I - \hat{A}^D)\hat{A} \) is nilpotent of index equal to \( k = \text{Ind}(\hat{A}) \). Thus the sum in (14) has only a finite number of nonzero terms and the first term in (14) is zero if \( N - i > \text{Ind}(\hat{A}) \).
1 Singular linear control theory

1. INTRODUCTION

This chapter will be concerned with

\[ \dot{x}(t) + Bx(t) = Cu(t), \quad t \in I \]  

(1.1)

where \( A, B \in \mathbb{C}^{n \times n} \), \( u(t) \in \mathbb{C}^m \), \( \xi \in \mathbb{C}^{n \times n} \), and \( I \) is an interval of the form \([0,T]\) or \([0,\infty)\). Equation (1) will be referred to as the process, \( x \) the state vector, and \( u \) the control. Since \( A \) is allowed to be singular, there are restrictions on \( x \) imposed by (1). For this reason some authors refer to \( x \) as a semi-state [41], [85], [86] or a descriptor variable [73], [74].

In a control problem, the control is to be chosen from an admissible class of controls \( \Omega(I) \) in such a manner as to achieve some desired behavior. The desired behavior can range from keeping \( x \) bounded to trying to hit a moving target set with a minimal amount of "cost." In [8] we considered several such optimal control problems. This chapter will consider qualitative questions like reachability, controllability, and observability as well as time optimal control.

Of course, there are many variations on every control problem depending on the choice of \( I, \Omega(I) \), the process, and the performance criteria. We shall not attempt to treat all of them. Rather, certain basic problems will be examined.

2. UNBOUNDED ADMISSIBLE CONTROLS

A basic question to ask, even if one is doing an optimal control problem, is whether it is even possible to get \( x \) to have the desired behavior. As
with most control questions, the approach, and the answers, differ dramatically depending on $\Omega(I)$. One of the biggest differences is whether or not $u$ is required to be bounded in some sense. For singular systems one not only needs to make smoothness and boundedness decisions but other decisions as well. The general classes of functions on an interval we shall consider:

$C^k(I)$; The $k$-times continuously differentiable functions on $I$.

$PC^k(I)$; The piecewise $k$-times continuously differentiable functions on $I$.

$E(I)$; Those measurable functions on $I$ bounded by one is absolute value.

The statement $g \in G(I)$ will be taken to mean that each component of $g$ is in $G(I)$. Thus if $g = [g_1(t), \ldots, g_r(t)]$, then $g \in E(I)$ says that $|g_i(t)| \leq 1$, for almost every $t \in I$ and all $i = 1, \ldots, r$.

This section will use $C^k(I)$, $PC^k(I)$ where $k$ will be defined shortly.

Take as the process,

$$\dot{x} + Bx = Cu, \quad t \geq 0. \quad (2.1)$$

Assume that $A, B$ are constant matrices and $C$ has full column rank.

Furthermore, assume that there is a $\lambda$ such that $(\lambda A + B)^{-1}$ is invertible.

For a given initial condition $x_0$, the system (1) may or may not be consistent for a particular control $u$. For a given $x_0$ and class of controls, let the set of admissible controls, $A\Omega(x_0, I)$, be those $u$ such that (1) with $x(0) = x_0$ is consistent. From Theorem 0.1.3, if the controls are from $C^k(I)$;

$$A\Omega(x_0, I) = C^k(I) \cap \{u \mid (I - \hat{A}^D) x_0$$

$$= (I - \hat{A}^D) \sum_{i=0}^{k-1} (-\hat{A}^D) i \hat{B} (Cu)^{(i)}(0)\}, \quad (2.2)$$

Here $\hat{A} = (\lambda A + B)^{-1} A$, $\hat{B} = (\lambda A + B)^{-1} B$, $\hat{C} = (\lambda A + B)^{-1} C$, and $k = \text{Ind}(\hat{A})$.

14
If $PC^k(I)$ is used instead of $C^k(I)$ then $u \in A\Omega(x_0, I)$, if $u \in PC^k(I)$ and

$$
(I - \hat{A}^D \hat{A})x_0 = (I - \hat{A}^D \hat{A}) \sum_{i=0}^{k-1} (-\hat{A}^D^i \hat{A}^D^i B^D C \hat{A})(I)(0)
$$

(2.3)

and

$$
(I - \hat{A}^D \hat{A}) \sum_{i=0}^{k-1} (-\hat{A}^D^i \hat{A}^D^i B^D C[u(i) (t^+_o) - u(i) (t^-_o)] = 0
$$

(2.4)

for any $t_o$ at which there is a discontinuity of $u$ or its first $k-1$ derivatives.

Note that for a given $x_o$, there may not be any admissible controls at all. For any sequence of vectors in $q^n$: $\{u_o, u_{k-1}, v_o, ..., v_{k-1}\}$ and any $t_o, t_1$, there exists a $u \in C^k([t_o, t_1])$ such that $u(i)(t_o) = u_1$, $u(i)(t_1) = v_1$. Thus from (1) we have the following:

**Theorem 1.2.1** For $C^k(I)$, or $PC^k(I)$, there exist admissible controls if and only if

$$
(I - \hat{A}^D \hat{A})x_o \in R([(I - \hat{A}^D \hat{A})^k \hat{A}^D B^D C, ..., (I - \hat{A}^D \hat{A})(-\hat{A}^D^k \hat{A}^D B^D C)]).
$$

(2.5)

In particular, all $x_o$ admit controls if and only if

$$
R(I - \hat{A}^D \hat{A}) = R([(I - \hat{A}^D \hat{A})^k \hat{A}^D B^D C, ..., (I - \hat{A}^D \hat{A})(-\hat{A}^D^k \hat{A}^D B^D C)]).
$$

(2.6)

The reachable set from $x_o$, denoted $R(x_o)$, consists of all those $x_1$ for which there exists a time $t_1 \in I$ and a control $u \in A\Omega(x_o, I)$ such that the solution $x$ of (1) satisfies $x(0) = x_o$, $x(t_1) = x_1$.

The easiest way to analyze $R(x_o)$ is as follows. Multiply (1) by $(\lambda A + B)^{-1}$ to get $\hat{A}\hat{x} + B\hat{x} = \hat{C}u$. Since $A, B$ commute, it is possible to perform a similarity so that (1) becomes
\[
\begin{bmatrix}
C & 0 \\
0 & N
\end{bmatrix}
\begin{bmatrix}
\dot{z}_1 \\
\dot{z}_2
\end{bmatrix} +
\begin{bmatrix}
B_1 & 0 \\
0 & B_2
\end{bmatrix}
\begin{bmatrix}
z_1 \\
z_2
\end{bmatrix} =
\begin{bmatrix}
C_1 \\
C_2
\end{bmatrix}u
\]  
(2.7)

where \(C\) is invertible, \(N\) nilpotent of index \(k = \text{Ind}(\hat{A})\), \(B_2\) is invertible and \(CB_1 = B_1 C\), \(NB_2 = B_2 N\). Multiplying the first equation in (7) by \(C^{-1}\) and the second by \(B_2^{-1}\) gives

\[
\dot{z}_1 + B_1 z_1 = C_1 u,
\]  
(2.8a)

\[
N \dot{z}_2 + z_2 = C_2 u.
\]  
(2.8b)

In [113], (8) is called "the standard canonical form of the descriptor system." With the above machinery in place, we are now ready to characterize \(R(x_0)\). The characterization, and part of its proof are from [115]. A similar development can be found in [93]. Related ideas appear in the work of Cobb [31].

**Theorem 1.2.2** Given the assumptions of this section \(R(0) = R_1 \oplus R_2\) where

\[
R_1 = R((\hat{A}^{D_1} \hat{C}, (\hat{A}B)^{D_2} \hat{C}, \ldots, (\hat{A}B)^{n-1} \hat{C})),
\]  
(2.9a)

\[
R_2 = R(((I - \hat{A}^{D} \hat{C}, (AB)^{D} (I - \hat{A}A) \hat{C}, \ldots, (\hat{A}B)^{k-1} (I - \hat{A}A) \hat{C})).
\]  
(2.9b)

**Proof** In terms of the canonical form (8), this amounts to showing

\[
R(0) = \tilde{R}_1 \oplus \tilde{R}_2
\]

where

\[
\tilde{R}_1 = R((\tilde{C}_1, B_1 \tilde{C}_1, \ldots, B_1^{n-1} \tilde{C}_1)),
\]  
(2.10a)

\[
\tilde{R}_2 = R((\tilde{C}_2, N \tilde{C}_2, \ldots, N^{k-1} \tilde{C}_2))
\]  
(2.10b)

The set \(\tilde{R}_1\) is just the usual reachable set for the ordinary linear control problem (8a). The difficulty in the proof is that (8a) and (8b) use the same control \(u\).
If \( z_1, z_2 \) is a solution of (8), with \( x_0 = 0 \),

\[
\begin{align*}
\dot{z}_1 &= e^{-B_1 t} \int_0^t e^{B_1 s} C_1 u(s) ds = \int_0^t e^{B_1 (s-t)} C_1 u(s) ds, \\
\dot{z}_2 &= -\sum_{i=0}^{k-1} N^i C_2 u_1^i (t).
\end{align*}
\]  

(2.11)  

(2.12)

Now, \( e^{B_1 t} \) is expressible as a polynomial in \( B_1 \) of degree at most \( n \). Hence

\[
e^{B_1 t} = \sum_{i=0}^n \psi_i(s) B_1^i
\]

for functions \( \psi_0(s), \ldots, \psi_n(s) \) analytic in \( s \).

But then (11), (12) give that \( R(0) \subset \tilde{R}_1 \odot \tilde{R}_2 \). To see the converse, suppose that \( a \odot b \in \tilde{R}_1 \odot \tilde{R}_2 \). Note that this implies that \( A\Omega(x_0, I) \) is non-empty by Theorem 1.2.1. Pick a \( t_1 \in I \).

We shall now construct a \( u \) for which \( x(0) = \hat{x}_0, x(t_1) = a + b \). That is, \( \hat{z}(0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \hat{z}(t_1) = \begin{bmatrix} a \\ b \end{bmatrix} \)

Following [115], this is done as follows. Take \( u = u_1 + u_2 \). Let

\[
\hat{v}_1 = t^{k}(t_1 - t)^{k} \chi(t) \text{ where } \chi(t) \text{ is yet to be specified. Note that } \hat{v}_1^{(i)}(0) = \hat{v}_1^{(i)}(t_1) = 0 \text{ for } i = 0, \ldots, k - 1. \text{ Thus } \hat{v}_1 \text{ makes no contribution to } \hat{z}_2(t) \text{ at } t = 0 \text{ or } t = t_1.
\]

Since \( \hat{b} \in \tilde{R}_2 \), there exists \( \hat{b}_1, i = 0, \ldots, k - 1 \), such that

\[
\hat{b} = -\sum_{i=0}^{k-1} N^i C_2 \hat{b}_1.
\]

Let \( \hat{u}_2(t) \) be such that \( \hat{u}_2^{(i)}(0) = 0, \hat{u}_2^{(i)}(t_1) = \hat{b}_1 \) for \( i = 0, \ldots, m - 1 \).

Then (12) is satisfied if \( \hat{u} = \hat{u}_1 + \hat{u}_2 \). There remains to show how to pick \( \chi(t) \) so that (11) holds with \( \hat{z}_1(t_1) = a, \hat{z}_1(0) = 0 \). If (11) is to hold, then it suffices to find a \( \chi(t) \) so that
\[
\tilde{z}_1 = a - \int_0^{t_1} e^{(t_1-s)B_1^-}C_1^{-1}u_2(s)ds = \int_0^{t_1} e^{(t_1-s)B_1^-}C_1^{-1}u_1(s)ds.
\] (2.13)

Let
\[
u_1(s) = s^{2n}(s - t_1)^{2n}e^{(t_1-s)B_1^-}C_1^{-1}e^{(t_1-s)B_1^-}q
\]
for a yet to be specified \(q\).

Then (13) becomes
\[
\tilde{z}_1 = \int_0^{t_1} g(s)e^{(t_1-s)B_1^-}C_1^{-1}e^{(t_1-s)B_1^-}g(s)ds q
\]
where \(g(s) = s^n(s - t_1)^n\). It remains only to show that it is possible to pick \(q\). Since \(a \in \tilde{R}_1\), we have \(\tilde{z}_1 \in \tilde{R}_1\). Hence it suffices to show for any polynomial \(g(s)\) that
\[
\tilde{R}_1 = R(\int_0^{t_1} g(s)e^{(t_1-s)B_1^-}C_1^{-1}e^{(t_1-s)B_1^-}g(s)ds).
\] (2.14)

Clearly the right hand side is contained in \(\tilde{R}_1\). Suppose then there exists a vector \(w \in \tilde{R}_1\) such that \(w\) is perpendicular to the right hand side of (14).

But \(\int_0^{t_1} g(s)e^{(t_1-s)B_1^-}C_1^{-1}e^{(t_1-s)B_1^-}g(s)ds\) is a Hermitian matrix. Thus
\[
\int_0^{t_1} g(s)^2 w e^{(t_1-s)B_1^-}C_1^{-1}e^{(t_1-s)B_1^-}w ds = 0.
\]
Hence
\[
T e^{(t_1-s)B_1^-}C_1^{-1}e^{(t_1-s)B_1^-}w = 0\text{ for all } 0 < s < t_1,
\]
so that
\[
0 = w e^{-(s)B_1^-}C_1 = \sum_{m=0}^{\infty} \frac{(-s)^m}{m!} w^{-B_1m}C_1.
\] (2.15)

From (15), \(w^{-B_1m}C_1 = 0\) for all \(m \geq 0\) and \(w \perp \tilde{R}_1\) which is a contradiction. \(\square\)
Now suppose $x_0 \neq 0$. If $u$, $v$ are two admissible controls in $A\Omega(x_0, I)$, then $u - v \in A\Omega(0, I)$. Hence if $d \in R(x_0)$, then $R(x_0) = d + R(0)$.

**Theorem 1.2.3** The set of initial conditions for which $A\Omega(x_0, I)$ is non-empty is

$$AI = R(\hat{A}^n) \oplus R_2$$

where $R_2$ is given by (9b).

**Proof** Theorem 3 follows immediately from (12). □

**Theorem 1.2.4** If $x_0 \in AI$, then

$$R(x_0) = \nu_o + R(0)$$

where $\nu_o = e^{\hat{A}t}B^\dagger A\hat{x}_0$, $t \in I$.

**Proof** Suppose $x_0 \in AI$. By constructing $u$ so that $u \in A\Omega(x_0, I)$ and $u(i)(t) = 0$, $i = 0, \ldots, k - 1$ for some $t > 0$, we have that there is a vector of the form

$$e^{\hat{A}t}B^\dagger A\hat{x}_0 + q_1$$

in $R(x_0)$ where $q_1 \in R_1 \subset R(0)$. □

**Corollary 1.2.1** The complete set of reachable sets, $\bigcup_{x_0 \in AI} R(x_0)$, is the same as $AI$.

It is worth noting that in Theorem 2,

$$R_1 = R(\hat{A}^n[I, \hat{A}^nB, \ldots, (\hat{A}^nB)^{n-1}]C),$$

(2.16)
and

\[ R_2 = R((I - \hat{A}^D)^k)[I, \hat{A}^D, \ldots, (\hat{A}^D)^{k-1}]C. \]  \hspace{1cm} (2.17)

3. CONTROLLABILITY AND OBSERVABILITY

There are many variations on the ideas of controllability and observability of a process. The ones we shall use are quite natural and are from [115]. See also [93].

**Definition 1.3.1** A system is **completely controllable** (C-controllable), if and only if one can reach any state from any admissible state. The system

\[ \dot{x} + bx = cu \]  \hspace{1cm} (3.1)

is R-controllable if and only if one can reach any reachable state from any admissible state.

Both types of controllability are easy to characterize using the results of Section 2.

**Theorem 1.3.1** The system (1) is completely controllable if and only if \( R(0) = \mathbb{C}^n \), or equivalently, both

\[ R_1 = R(\hat{A}D^\hat{A}[I, \ldots, (\hat{A}^D)^{n-1}]C) = R(\hat{A}D^\hat{A}) \]  \hspace{1cm} (3.2)

and

\[ R_2 = R((I - \hat{A}^D)^k)[I, \ldots, (\hat{A}^D)^{k-1}]C) = R(I - \hat{A}^D). \]  \hspace{1cm} (3.3)

Before proving Theorem 1, it is worth noting that to verify (2), (3) it suffices to show that the matrix products on either side have the same rank.
Proof To be completely controllable, says that $R(x_0) = \mathbb{C}^n$ for any $x_0 \in AI$. In particular $R_1 \oplus R_2 = \mathbb{C}^n$. But $R_1 \subseteq R(\hat{A}^D)$, $R_2 \subseteq R(I - \hat{A}^D)$ and $R(\hat{A}^D) \oplus R(I - \hat{A}^D) = \mathbb{C}^n$. \( \square \)

Theorem 1.3.2 The system (1) is R-controllable if and only if (2) holds.

Proof To be R-controllable means that $R(0) = R(x_0)$ for $x_0 \in AI$. But then from Theorem 1.2, $R(0) = AI$. This happens if and only if (2) holds. \( \square \)

Sometimes one cannot observe all components of the state vector. Observability deals with whether it is possible to determine $\dot{x}(t)$ given information on the system and the control.

Definition 1.3.1 The system

$$\dot{Ax} + Bx = Cu \quad (3.4)$$

$$y = Ex \quad (3.5)$$

is observable if for each $\hat{t} \geq 0$, there is a $t_0 \geq \hat{t}$ such that it is possible to compute $x(t)$ from $A, B, C, E, y(t)$ and $u(t)$ for $\hat{t} \leq t \leq t_0$.

Note that from (2.12), $(I - \hat{A}^D)x$ is computable from $A, B, C$ and the function $u(t)$ (but not just the value of $u$ at $t$). Thus (4), (5) will be observable if and only if (2.11) is observable. Standard results are now available, for example from [2, p. 211]. The system (4), (5) is observable if and only if

$$R([E^T, (\hat{A}^D)^{T}E^T, \ldots, ((\hat{A}^D)^{n-1}E^T)^{T}, (\hat{A}^D)^{n}]) = R(\hat{A}^D). \quad (3.6)$$

It is also possible to arrive at (6) by considering
\[
\begin{bmatrix}
A \\
0
\end{bmatrix} \dot{x} + \begin{bmatrix}
B \\
E
\end{bmatrix} x = \begin{bmatrix}
C u \\
Y
\end{bmatrix}.
\] (3.7)

Then (4), (5) will be observable provided that (7) has a uniquely determined solution \( x \) for those \( u, y \) for which it is consistent. The characterization of consistent initial conditions and description of the solution of (7) is given in Theorem 0.1.3.

4. STABILIZABILITY

As noted earlier, some control problems may be phrased in terms of steering \( x_0 \) to 0. A weaker, but sometimes sufficient criterion is to make \( x \) go to zero exponentially fast. The simplest way to accomplish this would be by an autonomous linear state feedback of the form \( u = Fx \).

**Definition 1.4.1** The system

\[
\dot{A} x + B x = Cu
\] (4.1)

is stabilizable if there exists a matrix \( F \), and scalars \( \alpha > 0, \beta > 0 \) so that all solutions \( z \) of

\[
\dot{A} z + B z = C F z
\] (4.2)

satisfy

\[
\| z \| \leq \alpha e^{-\beta t} \| z(0) \|. \] (4.3)

There is a large body of literature on the stability of nonsingular systems and it is not our intention to repeat it. Rather we intend to show that the stabilizability of (1) is equivalent to the stabilizability of a simpler system.
The term compensation is sometimes used in place of stabilization.

Suppose that (1) is solvable and let \( \lambda \) be such that \( \lambda A + B \) is invertible. Multiply (1) by \((\lambda A + B)^{-1}\) to get

\[
\dot{A}_x + \dot{B}_x = \dot{C}_u. \tag{4.4}
\]

Perform a similarity, as in (2.7), to get

\[
\dot{C}_w_1 = B_1 \dot{w}_1 + C_1 u,
\]

\[
\dot{N}_2 = B_2 \dot{w}_2 + C_2 u,
\]

where \( CB_1 = B_1 C, NB_2 = B_2 N, C \) and \( B_2 \) are invertible. Multiplying the first equation by \( C^{-1} \) and the second by \( B_2^{-1} \) allows us to assume that (4) is in the form

\[
\begin{align*}
\dot{w}_1 &= B_1 \dot{w}_1 + C_1 u, \tag{4.5a} \\
\dot{w}_2 &= \dot{w}_2 + C_2 u, \tag{4.5b}
\end{align*}
\]

where \( B_1, N, C_1, C_2 \) have been appropriately changed.

**Theorem 1.4.1** The system (1) is stabilizable if and only if (1) is stabilizable with \( \tilde{A}^T \tilde{A}_x \) feedback, that is, with a control of the form

\[ u = \tilde{A}^T \tilde{A}_x. \]

Equivalently, the system (1) is stabilizable if and only if (5a) is stabilizable with \( w_1 \) feedback. That is, it is stabilizable with a control of the form \( u = F_1 w_1 \).

**Proof** The equivalence is clear. Suppose (5a) is stabilizable with feedback \( u = F_1 w_1 \). Then the entries of \( w_1 \) are of the form \( t e^{-\beta_j t} \), \( \beta_j > 0 \), and (5b) is
\[ \dot{\omega}_2 = \omega_2 + C_2 F_1 \omega_1. \]  

But from (2.12), the solution \( \omega_2 \) of (6) is given in terms of \( C_2 F_1 \omega_1 \) and its derivatives. Hence the entries of \( \omega_2 \) are also of the form \( t^j e^{-\beta_j t} \), \( \beta_j > 0 \). Thus (5) is stabilizable.

The converse is somewhat more difficult. Suppose that (5) is stabilizable but that (5a) is not stabilizable with \( \omega_1 \) feedback.

Now if (5a) is not stabilizable, a famous theorem of Wonham says (5a) is not controllable. (The controllability of (5a) is equivalent to being able to arbitrarily assign eigenvalues of the system matrix using state feedback). See [63] for a simple proof of this fact. But the non-controllability of (5a), as noted in Section 3, is equivalent to the fact that 

\[ \text{rank}[C_1, B_1 C_1, \ldots, B_1^m C_1] < \text{rank}(A^D A) = \text{rank} C. \]  

Let \( M = \mathbb{R}([C_1, \ldots, B_1^m C_1]) \). Since \( \text{dim} M < \text{rank}(A^D A) \), there is at least one non-zero vector in \( M \). Decompose the system (5a) with respect to \( M \perp M \). Then (5a) takes the form

\[ \dot{\upsilon}_1 = C_{11} \upsilon_1, \]  
\[ \dot{\upsilon}_2 = C_{12} \upsilon_1 + C_{22} \upsilon_2 + \tilde{C} \upsilon, \]  

Furthermore,

\[ \dot{\upsilon}_2 = C_{22} \upsilon_2 + \tilde{C} \upsilon \]  

is controllable and hence stabilizable by \( \upsilon_2 \) feedback.

Now with respect to this new basis, the system (5) is

\[ \dot{\upsilon}_1 = C_{11} \upsilon_1, \]  
\[ \dot{\upsilon}_2 = C_{12} \upsilon_1 + C_{22} \upsilon_2 + \tilde{C} \upsilon, \]
\[ \dot{w}_2 = w_2 + C_2 u. \quad (4.9c) \]

Note that only the initial conditions of \( w_2 \) are restricted. Since \( v_1 = e^{C_{11}t} \tilde{v}_1(0) \), and (9) is stabilizable, it must be that Re\( \lambda < 0 \) for all \( \lambda \in \sigma(C_{11}) \). But (9b) is controllable. Let \( u = Fv_2 \) be such that Re\( \lambda < 0 \) for all \( \lambda \in \sigma(C_{22} + \tilde{C}F) \) and \( \sigma(C_{22} + \tilde{C}F) \cap \sigma(C_{11}) = \emptyset \). Then all solutions of (7) or equivalently (5a), go to zero exponentially fast which contradicts the assumption that (5a) was not \( \dot{w}_1 \) stabilizable.

Note that the control law \( u = Fv_2 \) which stabilizes (7a) actually stabilizes the whole system (1).

We shall return to the question of stabilization in Section 2.2 where impulsive controls and solutions are considered.

One way to understand the usefulness of stabilization is the following.

Consider first the system

\[ \dot{x} = Ax + Cu \quad (4.10) \]

For a given \( u \), and \( x_0 \), the solution is

\[ x = x_p + x_h \]

where \( x_p \) is a particular solution of (10) and \( x_h = e^{At}x(0) \) is the homogeneous solution. If \( A \) is not stable, then choosing \( u \) to regulate (10) is made difficult by the fact that the \( x_h \) term does not necessarily decay to zero. Thus, if there is a perturbation in \( x \), at time \( t_0 \), that is, we are given new initial conditions \( x(t_0) \), but \( u \) is not changed, then the new solution \( x \) could be radically different from the originally desired \( x \).

On the other hand, if \( A \) is stable, then \( x_h \to 0 \) as \( t \to \infty \). Thus even if there are perturbations in our original system at a time \( t_0 \), producing a
new solution $\overline{x}$, we still have $\overline{x} - x \to 0$ as $t \to \infty$, where $x$ is the original state response to $u$ with no perturbation.

**Example 1.4.1** Consider $\dot{x} = x + u$, $x(0) = 2$. We want to hold $x$ at 2 so that the control is taken as $u = -2$. Suppose at time $t = 3$, the value of $x$ is perturbed to 1. The system is now $\dot{x} = x - 2$, $x(3) = 1$, so that $x(t) = 2 - e^{(t-3)}$, and $x(t)$ diverges from the desired behavior as $t \to \infty$.

**Example 1.4.2** Consider $\dot{x} = -x + u$, $x(0) = 2$. We want to hold $x$ at 2 so that the control is taken as $u = 2$. Suppose at time 3, the value of $x$ is perturbed to 1. The system is now $\dot{x} = -x + 2$, $x(3) = 1$, so that $x(t) = 2 - e^{-t+3}$. Note that as $t \to \infty$, $x(t)$ returns to the desired behavior since the term $e^{-t+3}$ from the homogeneous equation $\dot{x} = -x$ is transient.

Since real systems are subject to perturbations, the preceding discussion suggests that the mathematical results will often agree better with the physical reality if $A$ is stable. Consider then a control law of the form

$$u = Kx + y$$  \hspace{1cm} (4.11)

where $y$ is a new control variable. Then (10) becomes

$$\dot{x} = (A + CK)x + Cy.$$  \hspace{1cm} (4.12)

The matrix $K$ is chosen so as to make $A + CK$ stable if that is possible.

Similarly, one can apply controls of the form (11) to (1) in order to stabilize the process.
5. TIME OPTIMAL CONTROL

As mentioned earlier, whether or not controls are uniformly bounded greatly affects the analysis. This section will briefly consider the problem of steering \( x(t) \), governed by the control law

\[
\dot{x} + Bx = Cu
\]  

(5.1)

from a given \( x_0 \) to a given \( x_1 \) in the minimum amount of time. This problem has no solution unless the controls \( u \) are bounded in some sense for Section 3 shows that if it is possible to go from \( x_0 \) to \( x_1 \), and controls are unbounded, the trip can be made in an arbitrarily short time period.

At this stage of the theory's development the results attained do not merit extensive coverage. We shall content ourselves with a discussion of some of the difficulties involved. The problem has been considered in [87].

The usual assumptions are that if \( u = [u_1, \ldots, u_l]^T \), then \( |u_i(t)| \leq a_i \) for some \( a_i > 0 \). Alternatively, there exists a closed convex set \( \Sigma \subseteq \mathbb{R}^n \) and it is assumed that \( u(t) \in \Sigma \) for almost all \( t \).

The optimal control \( u \) in nonsingular time optimal problems tends to be discontinuous, but piecewise constant. Also, the values of \( u \) are often extreme points of \( \Sigma \).

An intuitive idea can be gotten by considering the problem of flying from point a to point b with a rocket of variable thrust \( u \), \( |u| \leq K \). Ignoring such things as other gravitational fields, the quickest way to get from a to b and not be moving at b, is to go wide open (\( u = k \)) for a fixed period \([0, t_0]\), flip the rocket over, and come in with the engines still wide open (\( u = -k \)). If you don't care how fast you are going at b, then \( u \equiv k \).

There are many variations on these problems. For example, the target \( x_1 \) may vary with time, or may be a set instead of a point. Our discussion will
consider the case when \( x_1 = 0 \). While this, of course, is a special case, it is a problem of some interest.

To see why this is to, suppose that \( x \) is a solution of the differential equation,

\[
\dot{A}x + Bx = 0, \quad t \geq 0 \tag{5.2}
\]

and that \( x \) represents the desired behavior. Suppose at time \( t_0 \) there is a perturbation, or error \( z_0 \), in the state \( x \). Let \( w \) be the state value for \( t \geq t_0 \). You want to control the process with a control \( u \) to eliminate the perturbed behavior. The control law is to be

\[
\dot{A}w + Bw = Cu, \quad t \geq t_0, \quad w(t_0) = x(t_0) + z(t_0).
\]

Let \( z = w - x \), so that \( z \) measures the difference between the actual state \( w \) and the desired state \( x \). Then \( z \) satisfies the differential equation

\[
\dot{A}z + Bz = Cu, \quad t \geq t_0, \quad z(t_0) = z_0, \tag{5.3}
\]

and the goal is to steer \( z \) to 0 as quickly as possible.

As in (8) of Section 2, system (1) is rewritten in the form,

\[
\dot{z}_1 = B_1 z_1 + C_1 u, \tag{5.4a}
\]

\[
N \dot{z}_2 = z_2 + C_2 u, \tag{5.4b}
\]

where \( N \) is nilpotent of index \( k \). (4a) is sometimes called the slow system and (4b) the fast system, [29].

The solution of (4b) involves \( C_2 u \) and its derivatives. The first question then is, are the (semi-) state variables \( z_1, z_2 \) to be continuous? In this section only continuous solutions will be considered. Discontinuous solutions will be considered in Chapter 2.
Let $E$ be a compact convex polyhedron containing the origin.

**Definition 1.5.1** A function $u(t)$ is admissible for the time optimal control problem if $u(t) \in E$ for almost all $t$, $u$ is measurable, and $N_i^iC_u$ is $i$-times continuously differentiable. Equivalently, $(I - \hat{A}^D)\hat{A}^i\hat{C}_u$ is $i$-times continuously differentiable where $\hat{A} = (\lambda A + B)^{-1}A$, $\hat{C} = (\lambda A + B)^{-1}C$. The set of admissible controls is denoted $\mathcal{T}_\Omega$.

Note that an initial condition $x(0)$ may be a consistent initial condition for the differential equation (3) but not be consistent for any $u \in \mathcal{T}_\Omega$.

**Definition 1.5.2** An initial condition $x_0$ is consistent for the time optimal problem, $x_0 \in \mathcal{T}_\Omega C$, if there is a $u \in \mathcal{T}_\Omega$ for which $x_0$ is a consistent initial condition of (3). Equivalently, in (4)

$$z_2(0) = - \sum_{i=0}^{k-1} \left( N_i^iC_u \right) (i)(0), \quad (5.5)$$

or

$$(I - \hat{A}^D\hat{A})x_0 = - \sum_{i=0}^{k-1} ((I - \hat{A}^D\hat{A})\hat{A}^i\hat{C}_u)(i)(0). \quad (5.6)$$

A sufficient condition for (5) is that $z_2(0) \in \text{Interior of CE} + \text{R(NE)}$.

If the original system is totally singular, that is $\hat{A}$ is nilpotent, then (4a) is not present. In this case, there never exists a minimal time. If $x_0 \in \mathcal{T}_\Omega C$, and $\hat{A}$ is nilpotent, then $x_0$ can be steered to 0 in $\varepsilon$ time units where $\varepsilon$ is any number greater than 0. What is of interest then, is how the two systems (4a), (4b) interact since they share the control law $u$.

First note that even for the nonsingular system (4a) not all initial conditions for (4a) need admit optimal controls.
Example 1.5.1 Let \( \dot{z} = z + u, \ |u| \leq 1. \) Then if \( z(0) > 1, \) it is impossible to get \( \dot{z}(t) = 0, \ t > 0. \)

In fact, it is easy to show that if it is possible for (4a) to go from any \( z_1(0) \) to \( 0 \) in finite time, then \( \text{Re} \lambda < 0 \) for all eigenvalues \( \lambda \) of \( B_1. \)

Example 1.5.2 Let

\[
\begin{align*}
\dot{z}_1 &= z_1 + u, \quad |u| \leq 1, \quad z_1(0) = 1/2, \quad (5.7a) \\
0 &= z_2 - u, \quad z_2(0) = 1/2. \quad (5.7b)
\end{align*}
\]

The problem (7a) has a unique optimal solution \( u = -1 \) on \( [0, \ln 2]. \) However, (7b) requires that \( u(0) = 1/2. \) For any \( \tilde{t} > \ln 2 \) it is possible to construct a \( u \) so that \( u(0) = 1/2, \ u(\tilde{t}) = 0, \ |u| \leq 1, \) and \( z_1(\tilde{t}) = 0. \) Thus the full problem (7) does not have minimal time solution, but the nonsingular one (7a) does. Notice that the minimal time, \( \ln 2, \) of (7a) is a lower bound for the possible time \( \tilde{t} \) for the full system, (7).

Of course, in actual applications, it is not necessary, nor usually possible to be exactly optimal. Even if the optimum of the mathematical problem is found, the mathematical problem is an approximation of the real world problem.

There are two ways in which a control for our problem could be nearly optimal. One is that \( u \) is defined on \( [0, t_o], \ z(t_o) = 0 \) and \( t_o \) is almost as small as possible. That is, the process has the desired behavior, it arrives at zero, but the cost, which is time, is only close to the minimum possible. The other way would be that the control operates for the minimum amount of time but it leaves \( z(t) \) only close to 0.

We shall consider the second kind of nearly optimal.
Theorem 1.5.1 Suppose that (4a) has a piecewise-continuous time optimal control \( \tilde{u} \) on \([0, \tilde{t}]\) for \( z_1(0) \). If \[
\begin{bmatrix}
Z_1(0) \\
Z_2(0)
\end{bmatrix}
\]
\( \in \text{TIC} \) for (4), then for every \( \varepsilon > 0 \) there is a control \( u_\varepsilon(t) \) on \([0, t]\) such that \( u_\varepsilon \in T\Omega \) and the solution \( z_2(t) \) of (3) for \( u_\varepsilon(t) \) satisfies \( ||z_2(t)|| < \varepsilon \).

Proof Let \( \tilde{u}(t) \in T\Omega \) be such that \( z(0) \) is admissible for it. Take \( \varepsilon > 0 \). Let \( K = ||C||\{\max\|e^{-t} \| : t \in [0, \tilde{t}]\} \{\max\|a\| : a \in E\} \). Suppose that \( \tilde{u} \) has \( m \) discontinuities. Define \( u_\varepsilon(t) \) as follows. If \( \tilde{u}(t) \) is not piecewise \( k \)-times differentiable, approximate it uniformly by a piecewise \( k \)-times differentiable control \( \tilde{u}(t) \) such that \( ||\tilde{u}(t) - \tilde{u}(t)|| < \varepsilon / 2K \tilde{t} \) for \( 0 < t < \tilde{t} \).

Let \( u_\varepsilon(t) = \tilde{u}(t) \) except in small intervals around \( 0, \tilde{t} \) and each discontinuity of \( \tilde{u} \) or its first \( k \) derivatives. Let the intervals about \( 0, \tilde{t} \) be \( [0, \delta) \), \( (\delta, \tilde{t}] \) respectively. Take the sum of the lengths of these intervals and those around the discontinuities to be less than \( \varepsilon / 2K \). Now define \( u_\varepsilon(t) \) in a smooth way inside these intervals so that \( u_\varepsilon(t) = \tilde{u}(t) \) on the interval \([0, \delta / 2] \), \( u_\varepsilon(t) = 0 \) on \([\delta - \delta / 2, \tilde{t}] \), \( u_\varepsilon(t) \) is \( k \)-times continuously differentiable on \([0, \tilde{t}] \), and \( u_\varepsilon(t) \in E \) for all \( t \). Then \( u_\varepsilon(t) \) is the desired control. \( \square \)

6. CHEAP CONTROL AND NONSYMMETRIC RICCATI EQUATIONS

The important role that the matrix Riccati equation
\[
\hat{X} + XA + AX^* - KBK^{-1}B^*X + Q = 0
\]
plays in the solution of the usual quadratic regulator problem with process,
\[
\dot{x} = Ax + Bu, \quad t_0 \leq t \leq t_1, \quad x(t_0) = x_0
\]
and cost functional,
\[ J[x,u] = \frac{1}{2} x(t_1)^* F x(t_1) + \frac{1}{2} \int_{t_0}^{t_1} x(s)^* Q x(s) + u(s)^* R u(s) ds \] (6.2b)

is well known. Here \( x_0, x(t) \in \mathbb{R}^n, u(t) \in \mathbb{R}^m \) and is differentiable, \( F, Q, R \) are positive semi-definite and \( t_0, t_1, x_0 \) are given. In deriving (1), it is assumed that \( R \) is invertible. In this section we shall consider the so-called "cheap control" problem [89], [90] where \( R \) is singular and derive nonsymmetric versions of (1) that play the same role for the cheap control problem that (1) does for the classical nonsingular problem. The problem (6.2) is often studied as the reduced order model in a singular perturbation problem [88], [89].

Let \( \lambda(t) \in \mathbb{R}^n \) be the costate variable. Then necessary conditions for an optimal solution are the Euler-Lagrange equations

\[ \dot{\lambda} + A^* \lambda + Q x = 0, \] (6.3a)

\[ \dot{x} - A x - B u = 0, \] (6.3b)

\[ B^* \lambda + R u = 0, \] (6.3c)

with boundary conditions

\[ x(t_0) = x_0, \] (6.3d)

\[ \lambda(t_1) = F x(t_1). \] (6.3e)

It has been known for some time [96], that one may process (2), (3) through a series of coordinate transformations and reductions until the final problem is totally singular (has \( R = 0 \)). The singular problem has been discussed from this and related viewpoints in [4], [28].

32
Riccati equations have been used with singular quadratic problems before, see [32], [94] for example. The purpose of this section is to suggest that it is possible to consider (2), (3) directly with $R \geq 0$ without going through any reduction process. Our approach will be to work directly with (3) and the explicit solutions of (3) given in Chapter 0. Riccati equations have also been used to analyze singularly perturbed cheap control problem [39], [57].

Cobb [32] has recently considered the problem (2) but instead of allowing $R$ to be singular, he replaces (2a) with a descriptor system in the form

$$\dot{Ex} = Ax + Bu$$

and obtains a reduced order Riccati equation. Cobb allows inconsistent $x_0$, that is impulsive solutions, but impulses have infinite cost so that they are not present on optimal trajectories. His results complement those of this section. The system (3a) - (3c) can be rewritten in the form,

$$\dot{Ax} + Bx = 0,$$

by writing

$$
\begin{bmatrix}
I & 0 & 0 \\
0 & I & 0 \\
0 & 0 & I
\end{bmatrix}
\begin{bmatrix}
\dot{\lambda} \\
\dot{x} \\
\dot{u}
\end{bmatrix}
+
\begin{bmatrix}
A^* & 0 & 0 \\
0 & -A & -B \\
B^* & 0 & R
\end{bmatrix}
\begin{bmatrix}
\lambda \\
x \\
u
\end{bmatrix}
=
\begin{bmatrix}
0 \\
0 \\
0
\end{bmatrix}.
$$

The system (4) may be solved using the approach of Chapter 0 under the mild assumption that there exists a scalar $\mu$ such that $\mu A + B$ is invertible.

That is, (4) is a solvable system. This was done in [8] to solve the cheap control problem when $F = 0$ and $x(t_1)$ is specified. We shall now show how this approach can be modified to produce matrix Riccati equations.
If (3a) - (3b) is solvable, then the proof of Theorem 1 in [9] [see also 8] may be modified slightly to show that (3) is not only necessary but sufficient for \( x, \lambda, u \) to be optimal and thus the optimum when it exists is uniquely determined by \( x_o, \lambda_o \).

As in [8], [9], [21], let \( N_\mu, M_\mu \) be defined by

\[
\begin{bmatrix}
\mu + A^* & Q & 0 \\
0 & \mu - A & -B \\
B^* & 0 & R
\end{bmatrix}^{-1}
\begin{bmatrix}
I & 0 & 0 \\
0 & I & 0 \\
0 & 0 & 0
\end{bmatrix}
= \begin{bmatrix}
N_\mu \\
0 \\
M_\mu \\
0
\end{bmatrix}
\tag{6.6}
\]

where \( u \) is such that \( uA + B \) is invertible.

Suppose that \( \lambda, x \) are from an optimal trajectory. Then

\[
\begin{bmatrix}
\lambda \\
x
\end{bmatrix}
= e^{-N_\mu (I-uN_\mu) (t-t_o)} N_\mu D_\mu \begin{bmatrix}
\lambda_o \\
x_o
\end{bmatrix}
\tag{6.7}
\]

where \( \lambda_o = \lambda(t_o) \), \( x(t_o) = x_o \), and the control \( u \) is given by the feedback law

\[
u = M_\mu N_\mu D_\mu \begin{bmatrix}
\lambda \\
x
\end{bmatrix}.
\tag{6.8}
\]

Define the \( n \times n \) matrices \( \theta_i ; i = 1, 2 \) by

\[
M_\mu N_\mu D_\mu + [0_1, \theta_2].
\tag{6.9}
\]

The classical approach First consider what happens if the classical approach is followed. Substituting (8) into (3a), (3b) gives that an optimal trajectory \( \lambda, x \) must satisfy

\[
\dot{\lambda} = -A^* \lambda - Qx
\tag{6.10a}
\]

and
\[ \dot{x} = Ax + B_u = Ax + B_{\theta_1}^{\lambda} + B_{\theta_2}x, \]

that is,

\[ \dot{x} = (A + B_{\theta_2})x + B_{\theta_1}^{\lambda} \quad (6.10b) \]

Let \( K_R(t) \) be the solution on \([t_0, t_1]\) of the matrix Riccati equation

\[ \dot{K} + K(B_{\theta_1})K + K(A + B_{\theta_2}) + A^*K + Q = 0, \quad (6.11) \]

such that \( K(t_1) = F \) if the solution exists. When discussing only one Riccati equation we shall drop the subscript.

The matrix \( K_R(t) \) in (11) has the usual properties, that is;

**Proposition 1.6.1** If \( \lambda, x \) satisfies (3), then \( \lambda = Kx \) where \( K \) is the solution of (11) if \( K \) exists.

**Proof** \( \lambda - Kx \) = \( \dot{x} - \dot{Kx} - K\dot{x} \). From (10) this is

\[ -A^*\lambda - Qx + [K(B_{\theta_1})K + K(A + B_{\theta_2}) + A^*K + Q]x - K[(A + B_{\theta_2})x + B_{\theta_1}^\lambda] \]

\[ = -A^*\lambda + K(B_{\theta_1})Kx + A^*Kx - KB_{\theta_1}^\lambda \]

\[ = -A^*(\lambda - Kx) + KB_{\theta_1}(Kx - \lambda). \]

Let \( h = \lambda - Kx \). Then \( h \) satisfies

\[ \dot{h} = -A^*h + KB_{\theta_1}h, \quad h(t_1) = 0. \quad (6.12) \]

But (12) is a first order linear, homogeneous differential equation for a given \( K \) so \( h \equiv 0 \) and \( \lambda = Kx \) as desired.

While (11) is the most natural and classical of the Riccati equations we derive, attempts to compute with it suggest that it is not always useful.
An example will be given later. Accordingly two other Riccati equations will be developed.

A different equation. A different Riccati equation can be derived by differentiating (7) to observe that

\[
\begin{pmatrix}
\dot{\lambda} \\
\dot{x}
\end{pmatrix} = -N^\mu D(I - \mu N^\mu) \begin{pmatrix}
\lambda \\
x
\end{pmatrix}.
\]

(6.13)

Let

\[-N^\mu D(I - \mu N^\mu) = \Sigma = \begin{bmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{bmatrix}
\]

where each \(\Sigma_{ij}\) is \(n \times n\). Then

\[
\dot{\lambda} = \Sigma_{11}\lambda + \Sigma_{12}x,
\]

(6.15a)

\[
\dot{x} = \Sigma_{21}\lambda + \Sigma_{22}x,
\]

(6.15b)

for optimal trajectories \((\lambda, x)\) and the \(\Sigma_{ij}\) are independent of \(\mu\). In this case the Riccati equation is

\[
\dot{\Sigma} + K\Sigma_{21}K + K\Sigma_{22} - \Sigma_{11}K - \Sigma_{22} = 0, \quad K(t_1) = \tilde{F}
\]

(6.16)

where \(\tilde{F}\) is a matrix, to be described later so that \(\tilde{F}q = Fq\) for \(q\) which are terminal values of optimal trajectories. To see this note that

\[
(\lambda - Kx)^\cdot = \dot{\lambda} - \dot{K}x - \dot{x}
\]

\[
= \Sigma_{11}\lambda + \Sigma_{12}x + (K\Sigma_{21}K + K\Sigma_{22} - \Sigma_{11}K - \Sigma_{12})x - K(\Sigma_{21}\lambda + \Sigma_{22}x)
\]

\[= \Sigma_{11}(\lambda - Kx) + K\Sigma_{21}(Kx - \lambda).
\]

36
Again this gives that $\lambda = Kx$. The solution of (16) will be denoted $K_x$.

The systems (11) and (15) have the optimal trajectories in common but produce different solutions for initial conditions that do not admit an optimal trajectory.

It is not obvious from the definitions (9), (11), that (11) (or (16)) reduce to the usual Riccati equation (1) when $R$ is invertible. This is most easily seen by utilizing the fact that (8) is correct when $R$ is invertible and thus $M = \begin{bmatrix} \dot{\lambda} \\ \dot{x} \end{bmatrix} = -R^{-1}B\lambda$ for any $\lambda, x$ solving (5). But $\lambda(t_o), x(t_o)$ are arbitrary (for any $x_o$, there is an $x_1$ so that $x_o, x_1$ would be used) so that

$$\theta_1 = -R^{-1} B \text{ and } \theta_2 = 0.$$  

The $K_R$ system will be discussed first. Note that (10) and (13) are linear but involve boundary conditions whereas (11) is nonlinear but involves only a terminal condition.

Not all $x_o$ will admit an optimal control. However, $K_R$ is independent of $x_o$ and the control law when it exists is given by

$$u = [\theta_1 K + \theta_2] x$$  

(6.17)

(where $K$ is not only $K_R$ but any of the $K$'s we develop).

The computation of $\theta_1, \theta_2$ is independent of $x_o$ and need be done only once.

There remains, of course, the problem of solutions to (11) or (15) existing on $[t_0, t_1]$. Of course, solutions exist on some $[\hat{t}, t_1]$ where $t_o < \hat{t} < t_1$.

Let $\Lambda, X$ be $(n \times n$ matrix) solutions of (10) such that $\Lambda(t_1) = F$, $X(t_1) = I$. Let $\tilde{K}(t) = \Lambda(t)X^{-1}(t)$. Then $\tilde{K}$ satisfies (11). Since $\Lambda, X$ are well defined on $[t_o, t_1]$, $\tilde{K}$ can only fail to exist if $X(t)$ is singular.

Now $(\Lambda, x)$ is an optimal trajectory for some $x_o$ if and only if
\[
\begin{bmatrix}
\lambda(t) \\
x(t)
\end{bmatrix}
\in \mathcal{R}(N_{\mu}D_{N_{\mu}}) \cap \mathcal{R}
\begin{bmatrix}
F \\
I
\end{bmatrix}
= M.
\] (6.18)

or equivalently
\[
\begin{bmatrix}
I - N_{\mu}D_{N_{\mu}} \\
I
\end{bmatrix}
\begin{bmatrix}
\lambda(t) \\
x(t)
\end{bmatrix}
= \begin{bmatrix}
0 \\
0
\end{bmatrix}.
\] (6.19)

Note that \(I - N_{\mu}D_{N_{\mu}}\) is \(2n \times 2n\), \(F\) is \(n \times n\), \(I\) is \(n \times n\). If \(\lambda, x\) satisfy (19), then (7) shows that there is a pair \(\lambda, x\) satisfying (7) so that \(\lambda(t) = \lambda, x(t) = x\) and \(x(0) = x_0\) for this \(x\) would be an initial condition admitting an optimal solution.

We note in passing that the orthogonal projection onto the subspace \(M\) in (18) is given by
\[
P_F(P_\mu + P_F)^+P
\] (6.20)
where
\[
P_\mu = (N_{\mu}D_{N_{\mu}})(N_{\mu}D_{N_{\mu}})^+, \quad P_F = \begin{bmatrix} F \\ I \end{bmatrix} \begin{bmatrix} F \\ I \end{bmatrix}^+
\] (6.21)
and \((\cdot)^+\) denotes the Moore Penrose inverse [21]. In particular, all vectors in (18) are of the form
\[
\lambda = FZq, \quad x = Zq,
\]
g arbitrary, and
\[
Z = \begin{bmatrix} F \\ I \end{bmatrix}^+ (P_\mu + P_F)^+ P_\mu.
\] (6.22)

We will return to these formulas shortly. The preceding discussion may be rephrased as
Proposition 1.6.2 A vector $x_1$ is the terminal value for a solution of (3) if and only if there is a $\lambda_1$ so that $(\lambda_1, x_1)$ satisfies (18). In this case the trajectory $x$ is given by

$$\dot{x} = (A + B\theta_2 + B\theta_1 K)x, \quad x(t_1) = x_1$$

where $K$ is the solution of (11) if $K$ exists. (Again any of the $K$'s we develop in this section may be used in (23).)

We will now develop a different Riccati equation. By construction this equation has solutions on $[t_0, t_1]$. It is more complicated than the $K^*_\Sigma$ or $K^*_R$ equations but provides more information and does not suffer existence problems. A related equation was constructed in a simpler case by [94].

Let $C_1 = A + B\theta_2, \ C_2 = B\theta_1$. If $(\lambda, x)$ is an optimal trajectory, from (10)

$$\dot{\lambda} = -A^*\lambda - Qx,$$

$$\dot{x} = C_1 x + C_2 \lambda,$$

If $M$ in (18) is $s$-dimensional, let $X_1$ be an $n \times s$ matrix of full column rank such that $M = R\left[\begin{bmatrix} FX_1 \\ X_1 \end{bmatrix}\right]$. Let $(\lambda_1, x_1), i = 1, \ldots, s$ be solutions of (24) such that $x_1(t_1)$ is the $i^{th}$ column of $X_1$ and $\lambda_1(t_1) = FX_1(t_1)$.

Let $X = [x_1, \ldots, x_s], \ \Lambda = [\lambda_1, \ldots, \lambda_s]$. Then

$$\dot{\Lambda} = -A^*\Lambda - QX, \quad \Lambda(t_1) = FX$$

$$\dot{x} = C_2 \Lambda + C_1 X, \quad X(t_1) = X_1.$$
Proof Suppose that rank $X(t) < s$; $\hat{t} < t_1$ and rank $X(t) = s$ on $[\hat{t}, t_1]$. Let $h \in \mathbb{R}^s$ be such that $X(t)h = 0$, $X(t)h \neq 0$ for $t > \hat{t}$. Then $\Lambda(t)h$, $X(t)h$ give an optimal trajectory on $[\hat{t}, t_1]$ from $(\Lambda(t_1)h, 0)$ to $(\Lambda(t_1)h, X(t_1)h)$. But $x(t) = 0$, $\Lambda(t) = 0$ is an optimal trajectory on $[\hat{t}, t_1]$ such that $x(\hat{t}) = 0$. By uniqueness, $X(t) = 0$ which is a contradiction. \[ \square \]

Since $X$ has constant rank, $X^\dagger$ is differentiable if $X$ is. Let

$$K = \Lambda X^\dagger, \quad P = XX^\dagger.$$  \hfill (6.25)

$P$ is the orthogonal projection onto $R(X)$. Thus at any time $t$, $P(t)$ gives all possible initial values for optimal trajectories at time $t$. Note that by definition

$$KP = P, \quad P^2 = P, \quad P = P^*.$$  \hfill (6.26)

These properties will be used in what follows. The formula

$$(X^\dagger)^* = -X^\dagger X (XX^\dagger) + X^\dagger (XX^\dagger)^* (I - XX^\dagger)$$  \hfill (6.27)

from [21] will be needed. Now

$$\dot{\lambda} = -A^* K - AP + \Lambda(-X^\dagger (C_1 P + C_2 K) + X^\dagger (C_1 P + C_2 K)^* (I - P))$$


And

$$\dot{P} = \dot{XX}^\dagger + X(X^\dagger)^*$$

$$= (C_1 X + C_2 A)X^\dagger + X(-X^\dagger (\dot{XX}^\dagger) + X^\dagger (\dot{XX}^\dagger)^* (I - XX^\dagger))$$

$$= C_1 P + C_2 K - P(C_1 P + C_2 K) + P(C_1 P + C_2 K)^* (I - P).$$
Thus we have the system

\[ \dot{K} = -A^*K - QP - KC_1^*P - KC_2^*K + KC_1^*(I - P) + KK^*C_2^*(I - P), \quad (6.28a) \]

\[ \dot{p} = (I - P)(C_1^*P + C_2^*K) + (C_1^*P + C_2^*K)^*(I - P). \quad (6.28b) \]

The solution of (28) will be denoted \((K_S, P)\). The preceding may be summarized as follows.

**Theorem 1.6.1** Let \(P_1 = ZZ^*\) where \(Z\) is given by (21). Then the solution of (28) such that \(P(t_1) = P_1, K(t_1) = FP_1\) exists on \([t_0, t_1]\). \(P(t)\) gives the possible state values at time \(t\). For any vector \(q\), \(x = Pq, \lambda = Kq\) is an optimal solution on \([t_0, t_1]\).

Other variants on (28) are possible. For example, (28b) may be replaced by

\[ \dot{p} = [(I - P)(C_1^*P + C_2^*K)] + [(C_1^*P + C_2^*K)(I - P)]^*. \quad (6.28c) \]

Equation (28c) is equivalent analytically to (28b) on optimal trajectories but computationally is slightly less work and maintains the symmetry of \(P\).

Another variation would be to use (13) instead of (24) in deriving (28).

**An example** The preceding three equations \(K_R, K_L, \) and \((K_S, P)\) were tried on several low dimensional examples using an explicit Euler's. In almost every case \(K_L\) and \((K_S, P)\) worked well and \(K_R\) blew up. The one exception is the example in [9] which since it decomposes into a nonsingular and a singular problem in an especially nice way, produces a good \(K_R\). As a typical example consider;
\[
A = \begin{bmatrix}
0 & 0 & 1 \\
2 & 0 & 0 \\
0 & 0 & 3 \\
1 & 0 & 0
\end{bmatrix}, \quad
B = \begin{bmatrix}
1 \\
0 \\
0 \\
0
\end{bmatrix}, \quad
R = 0, \quad
Q = \begin{bmatrix}
1 & 1 & 0 \\
0 & 0 & 2
\end{bmatrix}
\]

Then
\[
N^\mu = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
.5 & 0 & 0 & 0 & .25 & .167 \\
0 & 0.333 & 0 & 0 & .167 & .111 \\
0 & 0 & 0 & 0 & -.5 & 0 \\
0 & 0 & 0 & 0 & 0 & -.333 \\
0 & 0 & -.5 & 0 & 0 & 0
\end{bmatrix}
\]

\[
N^D = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 \\
2 & -2 & 3 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 2 \\
-4 & 0 & -6 & 0 & -2 & 0 \\
0 & 4 & 0 & 0 & 2 & 0 \\
-6 & 0 & 0 & 0 & -3 & 0
\end{bmatrix}
\]

The entries of \( N^D \) have a certain symmetry. This will be discussed later.

Finally,
\[
M^\mu = \begin{bmatrix}
0 & 0 & -.5 & -1 & 0 & 0
\end{bmatrix}
\]

*Since we were working on a small problem, not developing a code, and have \((\cdot)^+\) and \((\cdot)^D\) functions, we utilized (20) and got
\[
P(t_1) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad (6.33)
\]

\[
K(t_1) = FP(t_1) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.
\quad (6.34)
\]

Starting at \( t = 2 \), with a step of \(.001\), the \( K_R \) system blew up by around 1.6 for \( K_R(2) = I \) as well as \( K_R(2) = P(2) \). At \( t = 1.5 \), to four decimal places (which exceeds the expected accuracy with Eulers at \( h = .001 \))

\[
K_S(1.5) = \begin{bmatrix} .0001 & .0003 & -.0001 \\ -.2432 & .1044 & -.0023 \\ -.2975 & .1279 & -.0064 \end{bmatrix},
\]

\[
P(1.5) = \begin{bmatrix} .84734 & -.3605 & -.0008 \\ -.3606 & .1525 & .0055 \\ -.0027 & .0026 & -.0093 \end{bmatrix}.
\]

As a quick check note that \( P(1.5) = P(1.5)^* \) and trace \( P(1.5) = 1 \) to the expected two place accuracy for the hermitian projection \( P \). On the other hand, if \( K_L(2) = FP(2) \),

\[
K_L(1.5) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & .6686 & .8168 \\ 0 & .8168 & 1.2313 \end{bmatrix}.
\]

The matrices \( K_L \) and \( K_S \) should agree on \( R(P) \). That is, \((K_L - K_S)P = 0\). At \( t = 1.5 \),
\[(K_L - K_S)P = \begin{bmatrix} .0000 & .0000 & .0000 \\ .0004 & .0005 & -.0047 \\ .0006 & .0010 & -.0027 \end{bmatrix} \]

Now consider the \(K_R\) system at \(h = .001, t = 1.7\), with \(K_R(2) = P(2)\).

\[
K_R = \begin{bmatrix} 1.049 & 1.943 & 1.493 \\ 1.136 & 2.564 & 2.362 \\ .637 & 1.922 & 2.254 \end{bmatrix}
\]

(6.35)

while at \(t = 1.7, h = .001\),

\[
K_S(1.7) = \begin{bmatrix} .000 & .000 & .000 \\ -.309 & .084 & .100 \\ -.543 & .148 & .173 \end{bmatrix}
\]

(6.36)

\[
P(1.7) = \begin{bmatrix} .848 & -.229 & -.279 \\ -.229 & .061 & .078 \\ -.281 & .077 & .086 \end{bmatrix}
\]

Then

\[
K_R(1.7)P(1.7) = \begin{bmatrix} -.038 & -.007 & -.013 \\ -.288 & .078 & .086 \\ -.533 & .129 & .166 \end{bmatrix}
\]

(6.37)

\(K_R(1.7)\) was computed with \(h = .01\) since \(K_R(1.6)\) was larger with \(h = .001\) then \(.01\) and overflow occurred by \(t = 1.5\).

Comparing (36) to (37), we see that if \(K_R(2) = FP(2)\), then \(K_R\) and \(K_S\) agree on \(R(P)\) also. The differences between the different Riccati systems is now more apparent. All agree on \(R(P)\). However \(K_S(I - P) = 0\) while
\( K_\Sigma(I - P) \neq 0, \ K_\Sigma(I - P) \neq 0. \) At \( h = .01, t = 1.6, \) (35), (37) show that
\( K_\Sigma(1.7)(I - P(1.7)) \) is an order of magnitude greater than \( K_\Sigma(1.7)P(1.7). \)

For \( K_\Sigma \) we do not have \( K_\Sigma(I - P) \) dominating \( K_\Sigma P. \)

The preceding example does not prove anything, of course. The behavior, however, is typical of the several examples considered.

In summary, three Riccati systems have been presented which generalize the usual Riccati equation. The obvious one, the \( K_\Sigma \) equation (11), exhibits numerical difficulties. The two alternatives, the \( K_\Sigma \) and \( (K_\Sigma P) \) equations perform better. Existence of solutions is assured for the \( (K_\Sigma P) \) equation. \( K_\Sigma, K_\Sigma, K_\Sigma \) all agree on \( R(P) \) when they exist.

Several questions remain. The \( K_\Sigma \) system needs further study. In the examples tested, \( K_\Sigma \) was symmetric if \( K_\Sigma(2) \) was. The needed symmetry properties of \( \Sigma \) seem to be almost present in \( N_\mu \) but we have been unable to prove that \( K_\Sigma \) is symmetric except in special cases.

For larger problems, or ones where the equations (19) may be ill-conditioned, some care must be exercised in computing \( P(t_1). \)
2 Impulsive behavior of linear systems

1. INTRODUCTION

In [8], distributional solutions of

\[ \dot{A}x + Bx = f, \]  

with \( f \) \( k \)-times continuously differentiable, were formally developed using

Laplace transforms. The solution arrived at was

\[ x = x_c + x_i, \]  

where

\[ x_c = e^{-A^D(t)}A^Dx(0) + A^D \int_0^t e^{A^D(t-s)}\hat{f}(s) \, ds \]  

\[ + (I - A\hat{A})^D \sum_{m=0}^{k-1} (-A\hat{B})^m \hat{f}(m)(t), \]

\[ x_i = (I - A\hat{A})^D \sum_{m=0}^{k-1} (-A\hat{B})^m [\delta(m)(t)A\hat{x}(0) \]  

\[ + \sum_{i=0}^{m-1} \delta(m-i-1)(t)\hat{f}(i)(0)] \]

with \( \hat{A} = (\lambda A + B)^{-1}A, \hat{B} = (\lambda A + B)^{-1}B, \hat{f} = (\lambda A + B)^{-1}f, k = \text{Ind}(\hat{A}). \) There

are certain problems with (2) as a solution of (1) however. Subsequent to

[8], Cobb and Francis have put results like (2) into the correct mathematical

perspective.

Section 2 will provide a brief outline of distributional theory. Section

3 will give Cobb's [30] justification of (2) as a solution for (1). Section
4 will outline Francis' [51]. Section 5 will present Cobb's results on elimination of impulsive behavior by state feedback [31].

2. DISTRIBUTIONS

Obviously, the theory of distributions cannot be fully developed in a few pages. The most we can hope for is to give the reader unfamiliar with distributions a rough intuitive idea of what distributions are.

One reason distributions are introduced is to be able to 'differentiate' functions that are not differentiable. What is desired is a class of objects \( \mathcal{D} \) called distributions, on which there is an operation called differentiation. We want \( \mathcal{D} \) to include ordinary differentiable functions and distributional differentiation to correspond with ordinary differentiation.

This can be done as follows. Let \( \mathcal{T} \) be the space of test functions. \( \mathcal{T} \) consists of infinitely differentiable functions \( f \), defined on \( (-\infty, \infty) \), which have compact support. That is there is a compact set (closed, bounded) \( K \subseteq (-\infty, \infty) \) such that \( f(x) = 0 \) on \( (-\infty, \infty) \setminus K \).

Let \( \mathcal{S} \) denote the space of testing functions of rapid descent. That is, \( f \in \mathcal{S} \) if and only if \( f \) is indefinitely differentiable and

\[
|t^mf_n^{(k)}(t)| \leq C_{mk}, \quad -\infty < t < \infty
\]

where \( C_{mk} \) is a constant depending on \( f \) and \( m,k \) vary through all nonnegative integers.

If \( \{f_n\} \subseteq \mathcal{T} \), \( f \in \mathcal{T} \) we say \( f_n \to f \) if there is a compact set \( K \) containing their supports and \( f_n^{(k)} \) converges uniformly to \( f^{(k)} \) on \( K \) for each \( k \geq 0 \).

If \( \{f_n\} \subseteq \mathcal{S} \), \( f \in \mathcal{S} \), then \( f_n \to f \) if \( |t^mf_n^{(k)}(t)| \leq C_{mk}, \quad -\infty < t < \infty \), \( C_{mk} \) does not depend on \( n \) and for each \( k \) and on each bounded \( K \), \( f_n^{(k)} \) converges uniformly to \( f^{(k)} \).
Note that $T \subset S$ and that $f_n + f$ implies $f_n + f$. Furthermore $T$ is dense in $S$. That is, for every $g \in S$, there is $(f_n) \subset T$ such that $f_n \to g$.

**Definition 2.2.1** A distribution is a continuous linear functional on $T$. The space of distributions is denoted $\mathcal{D}$. If $f \in T$ and $d \in \mathcal{D}$, the value of $d$ at $f$, $d(f)$ is denoted $\langle f, d \rangle$. A tempered distribution (or a distribution of slow growth) is a continuous linear functional on $S$. We shall denote these by $S\mathcal{D}$. Again if $f \in S$ and $d \in S\mathcal{D}$, $d(f) = \langle f, d \rangle$.

Our notation is slightly different from the standard notation where our $T$ is denoted $\mathcal{D}$ and our $\mathcal{D}$ is denoted $\mathcal{D}'$, but it is easier to remember.

Definition 1 may seem odd if one has not seen it before. There are alternative definitions in terms of equivalence classes of sequences of functions.

If $g$ is a piecewise continuous function on $(-\infty, \infty)$, then for $f \in T$

$$d(f) = \int_{-\infty}^{\infty} f(t)g(t)dt$$  \hspace{1cm} (2.1)

defines a continuous linear functional on $T$. Thus we can view piecewise continuous functions as distributions. Because of (1) it is suggestive to write

$$\langle f, g \rangle = \int_{-\infty}^{\infty} f(t)g(t)dt$$  \hspace{1cm} (2.2)

if $f \in T$, $g \in \mathcal{D}$ where if $fg$ is not an integrable function, the right side of (2) is defined to be the left side.

For an arbitrary piecewise continuous function $g$, (1) may not be defined for all $f \in S$. An example is $g(t) = e^{t^2}$. However if for some $N$
\[ \lim_{t \to \pm \infty} |t|^{-N}|g(t)| = 0, \quad (2.3) \]

g is called a function of slow growth. In this case, (1) does define a continuous linear functional. Thus functions of slow growth can be viewed as tempered distributions.

If \( \{g_n\} \subseteq D, (SD) g \in D, (SD) \) we say \( g_n \to g \) in \( D(SD) \) if \( \langle f, g_n \rangle \to \langle f, g \rangle \) for all \( f \in T(S) \). Note that ordinary differentiation is a continuous linear transformation of \( T, S \), into themselves.

**Definition 2.2.2** The distributional derivative, \( \dot{g} \), of \( g \in D \) (SD) is defined by

\[ \langle f, \dot{g} \rangle = -\langle f, g \rangle \quad \text{for all } f \in T, \quad (f \in S). \quad (2.4) \]

Since the right hand side of (4) is a continuous functional of \( f \), (4) defines \( \dot{g} \) as one. Alternatively, if \( f \in T \), and \( g \) is a differentiable function, (4) is

\[ \langle f, \dot{g} \rangle = \int_{-\infty}^{\infty} f(t) \dot{g}(t) dt = \lim_{T \to \infty} \int_{-T}^{T} f(t) g(t) dt \]

\[ = -\int_{-\infty}^{\infty} \dot{f}(t) g(t) dt = -\langle \dot{f}, g \rangle \]

since \( f(t) \) is zero for large enough \( |t| \). Thus the distributional and ordinary derivative agree for differentiable functions. A similar argument works for \( f \in S \) and \( g \in SD \).

The evaluation map \( d(f) = (-1)^k f^{(k)}(0) \) is a continuous linear functional on \( T \) or \( S \). This (tempered) distribution is denoted as \( \delta^{(k)} \) and \( \delta^{(0)} \) is denoted \( \delta \). This gives the famous formula
\[ \int_{-\infty}^{\infty} \delta^{(k)} f(t) \, dt = (-1)^k f^{(k)}(0). \] 

(2.5)

It is not hard to see that distributions are infinitely differentiable and that \( \delta^{(k)} \) is the \( k \)-th derivative of \( \delta \).

The distribution \( \delta \) is the limit, in the distributional sense, of functions of fixed area and decreasing support. That is, it is the limit of functions which are progressively more and more "impulse-like." For example, let \( h_n(t) = n \) if \( |t| \leq 1/2n \), \( h_n(t) = 0 \) if \( |t| > 1/2n \), as shown in Figure 2.2.1.

![Figure 2.2.1](image)

Then for any \( f \in \mathcal{F} \), it is not hard to show that

\[ \lim_{n \to \infty} \langle f, h_n \rangle = \lim_{n \to \infty} \int_{-\infty}^{\infty} f(t) h_n(t) \, dt = f(0) = \langle f, \delta \rangle. \]

Thus \( h_n \to \delta \) in \( \mathcal{D} \).
A vector or matrix will be said to be in $\mathcal{T}, \mathcal{S}, \mathcal{D}$, or $\mathcal{SD}$ if each entry is.

In the context of ordinary differential equations, one usually takes the Laplace transform of a function by

$$L[f](s) = \int_0^\infty e^{-st}f(t)dt.$$  \hspace{1cm} (2.6)

We wish to take the Laplace transform of a distribution. To do so it is necessary to work with distributions with support in $[0, \infty)$.

Let $\mathcal{D}_o \cap (\mathcal{SD}_o)$ consists of those $g \in \mathcal{D}$ ($g \in \mathcal{SD}$) such that $\langle f, g \rangle = 0$ for all $f \in \mathcal{T}$ ($f \in \mathcal{S}$) such that $f(t) = 0$ for $t \geq 0$.

**Definition 2.2.3** If $f \in \mathcal{D}_o$ and $e^{-ct}f \in \mathcal{SD}$ for some real $c$, then

$$L[f](s) = \langle \lambda(t)e^{-(s-c)t}, e^{-ct}f(t) \rangle$$ \hspace{1cm} (2.7)

where $\lambda$ is infinitely differentiable, $\lambda = 1$ for $t \geq 0$ and $\lambda(t) = 0$ for $t < \varepsilon$, some $\varepsilon < 0$.

The $\lambda$ is introduced so that $\lambda e^{-(s-c)t} \in \mathcal{S}$. Then $e^{-ct}f \in \mathcal{SD}$ and (7) makes sense. The usual formulas hold for the Laplace transform and $L$ is continuous as a transformation on $\mathcal{D}_o \cap (\mathcal{SD}_o)$.

As an example, consider $\delta \in \mathcal{SD}_o$. We may take $c = 0$. Then (7) becomes

$$L[\delta](s) = \langle \lambda(t)e^{-st}, \delta \rangle = \lambda(0)1 = 1.$$

Similarly,

$$L[\dot{\delta}](s) = \langle \lambda(t)e^{-st}, \dot{\delta} \rangle = \dot{\lambda}e^{-st} - s\lambda(t)e^{-st} \bigg|_{t=0} = \dot{\lambda}(0) - s = -s.$$
The distributional derivative is the same as the ordinary derivative for functions in $T$. For a function $f$ on $(-\infty, \infty)$, let $f'(t)$ be the ordinary derivative for $t \neq 0$ and the right hand derivative for $t = 0$. Suppose that $f(t) = 0$ for $t < 0$, $f$ and $f'$ are continuous on $[0, \infty)$, $f \in SD$, and both $f$ and $f'$ are of exponential growth. Then

$$L[f'] = sf[f] - f(0).\tag{2.8}$$

Let $\dot{f}$ be the distributional derivative of $f$. Since $f \in SD_0$, $c$ in (7) can be taken to be zero. Thus

$$L[\dot{f}] = <\lambda e^{-st}, \dot{f}> = -<\lambda e^{-st} - s\lambda e^{-st}, f>$$

$$= -<\lambda e^{-st}, f> + s<\lambda e^{-st}, f>.$$

But $\lambda e^{-st}$ is zero for $t \geq 0$ and $f \in SD_0$ so that $<\lambda e^{-st}, f> = 0$. Thus

$$L[\dot{f}] = sf[f].\tag{2.9}$$

The two types of derivatives are related by the formula

$$\dot{f} = f' + \delta f(0).\tag{2.10}$$

Repeated application of (10) gives

$$f^{(m)} = f[m] + \delta f^{[m-1]}(0) + \ldots + \delta^{(m-1)}f(0)\tag{2.11}$$

where round brackets are distributional derivatives and square brackets use the right hand derivative at zero.

3. DISTRIBUTIONS AS SOLUTIONS

As in Chapter 1, it will be assumed that the system $A\dot{x} + Bx = f$ has been written as
\[ \dot{z}_1 = B_1 z_1 + f_1, \quad (3.1a) \]
\[ \dot{z}_2 = A_2 z_2 + f_2. \quad (3.1b) \]

We are going to formally apply Laplace transforms to (1) which is viewed as an initial value problem with \( t_0 = 0 \). If \( \dot{z}_1 \) in (1) is considered as the ordinary derivative for \( t > 0 \) and the right hand derivative for \( t = 0 \), the system (1) becomes, in the notation of Section 2,

\[ \dot{z}_1' = B_1 z_1 + f_1, \quad (3.2a) \]
\[ \dot{z}_2' = A_2 z_2 + f_2. \quad (3.2b) \]

The distributional equivalent of (2b) is

\[ \dot{\mathcal{N}z}_2 = \mathcal{N}z_2 + \delta \mathcal{N}z_2(0) + f_2 \quad (3.3) \]

since \( \mathcal{N}z_2' = \mathcal{N}z_2 - \delta \mathcal{N}z_2(0) \). Assume also that \( z_1(t) = f_1(t) = 0 \) for \( t < 0 \) and consider (1) to be a distributional equation. If the Laplace transform is applied to (2) we are implicitly assuming that \( \mathcal{N}z_2' \) exists on \([0, \infty)\), \( \mathcal{N}z_2 \) has a Laplace transform, and \( f \) grows slowly enough. Then from (2b)

\[ N(s\mathcal{N}z_2 - z_2(0)) = \mathcal{Z}_2(s) + \mathcal{F}_2(s). \]

Thus

\[
\mathcal{Z}_2(s) = (sN - I)^{-1}Nz_2(0) + (sN - I)^{-1}f_2(s)
\]

\[ = \sum_{m=0}^{k-1} s^m N^{m+1} z_2(0) - \sum_{m=0}^{k-1} s^m N^{m} f_2(s), \]

so that

\[ z_2(t) = \sum_{m=0}^{k-2} \delta(m) N^{m+1} z_2(0) - \sum_{m=0}^{k-1} N^{m} f_2(m)(t) \quad (3.4) \]

53
which is the solution given in Section 1 and [8]. Here $f_2^{(m)}$ is the $m$th
distributional derivative of $f$. There is, unfortunately, some confusion on
this point in [8, p. 84]. The formula on page 84 requires as an assumption
that $x(t)$ is differentiable on $(-\infty, \infty)$.

Formula (3.4) may be rewritten as

$$
x_2(t) = -\sum_{m=0}^{k-1} N^m f_{22}^{(m)}(t) - \sum_{m=0}^{k-2} \delta(r) N^{r+1} x_2(0)
+ \sum_{m=0}^{k-1} N^m f_{2}^{(m-1)}(0).
$$

Thus $x_2$ continuous on $[0, \infty)$ implies

$$
x_2(0) = -\sum_{m=0}^{k-1} N^m f_{22}^{(m)}(0) \tag{3.6}
$$
as in [8] and (4) reduces to $x_2(t) = -\sum_{m=0}^{k-1} N^m f_{22}^{(m)}(t)$ as in [8]. Suppose
however $x_2(0)$ does not satisfy (6). Then (4) provides a solution of (3) or
(2) and not (1). In fact, (4) will only be a solution of (1) for very
special $f_2$. What is the "correct" solution then for inconsistent initial
conditions? Is (4) the correct answer and thus (2) the correct system or is
(1) preferable?

Vergehez has arguments for (2) on physical grounds in his study of the
reassembly of subsystems [108]. See also [37]. We shall give here Cobb's
reasoning since it involves the idea of model reduction and singular
perturbations.

For any $z_0 \in \mathbb{R}^n$, $f \in \mathcal{D}$, define

54
\[ \Lambda(z_0, f) = -\sum_{m=0}^{k-2} \delta(m) N^{m+1} z_0 \sum_{i=0}^{k-1} N^{-m}(m). \]  

(3.7)

so that \( \Lambda(z_0, f) \) is the formal solution (4).

Following Cobb define:

**Definition 2.3.1** \( z \in \mathcal{D} \) is a limiting solution of \( N \dot{z} = z + f \) with initial condition \( z_0 \) if there exist sequences \( \{N_j\} \subseteq \mathbb{R}^{m \times n}, \{f_j\} \subseteq \mathcal{D}, \{z_{0j}\} \subseteq \mathbb{R}^n \), with \( N_j \) invertible, such that \( N_j \rightarrow N, \ z_{0j} \rightarrow z_0, \ f_j \rightarrow f \) and the solution \( z_j \) of \( N_j \dot{z}_j = z_j + f_j, \ z_j(0) = z_{0j} \), converges to \( z \).

**Theorem 2.3.1** (Cobb [30]) For every \( z_0 \in \mathbb{R}^n, f \in \mathcal{D} \), there exists a limiting solution of \( N \dot{z} = z + f \) with initial condition \( z_0 \). The limiting solution is unique and it is \( \Lambda(z_0, f) \).

**Proof** (sketch) The first step is to take the special sequence \( N_j = N - (\frac{1}{j})I \), use the fact that \( \int_0^\infty \|N_j e^{-tN_j} \|\, dt \) is bounded for all \( j \), for some \( k \), and get that \( \Lambda(z_0, f) \) is a limiting solution. The next step is to show that the limiting solution is unique. Details can be found in [30].

One way to interpret Theorem 1 is the following. Given any sequence of systems \( N_j \dot{z}_j = z_j + f_j, \ z_j(0) = z_{0j} \), where the "equations converge" to \( N \dot{z} = z + f, \ z(0) = z_0 \), then if the \( z_j \) converge to anything in the distributional sense, they converge to (7). Furthermore, there are sequences \( \{z_j\} \) such that the \( z_j \) do converge to (7).

Of course, in many applications the sequence \( \{z_j\} \) is given and it is necessary to determine whether \( \{z_j\} \) converges. If we knew that the \( \{z_j\} \) converged, Theorem 1 would tell us what it converged to.
In the next section, we shall give a result due to Francis which addresses this problem. It gives sufficient conditions for convergence in a distributional sense of the full problem to the reduced problem.

4. SINGULAR PERTURBATIONS

The singularly perturbed system

\[ \dot{x} = A(\epsilon)x(t) + B(\epsilon)y(t), \]  
\[ \dot{y} = C(\epsilon)x(t) + E(\epsilon)y(t), \]  
\[ x(0) = x_0, \quad y(0) = y_0 \]

with \( A(\epsilon), B(\epsilon), C(\epsilon), E(\epsilon) \) continuous at \( \epsilon = 0 \) was studied in [8]. The reduced system associated with (1) is

\[ \dot{x} = Ax + By, \]  
\[ 0 = Cx + Ey, \]  
\[ x(0) = x_0. \]

One of the fundamental questions is the relationship between (1) and (2). The solution of (2) is given by Theorem 0.1.3 provided that

\[ \det \begin{bmatrix} sI - A & -B \\ -C & -E \end{bmatrix} = r(s) \neq 0 \]

for some scalar \( s \). Under this assumption, Francis [51] has proven that

**Theorem 2.4.1** (Francis [51]). If \( A(\epsilon), B(\epsilon), C(\epsilon), D(\epsilon) \) are continuous on \([0, \epsilon_0]\) some \( \epsilon_0 > 0 \), and

\[
\begin{bmatrix}
A(\epsilon) & B(\epsilon) \\
\frac{1}{\epsilon} C(\epsilon) & \frac{1}{\epsilon} E(\epsilon)
\end{bmatrix}
\]
has eigenvalues in a half-plane \( \text{Re}(s) \leq \sigma_o \), for \( 0 < \varepsilon < \varepsilon_o \), then for
\[ \sigma_1 > \sigma_o, \]
the solutions \( x(\varepsilon, t), y(\varepsilon, t) \) of (1) converge to the (possibly
distributional) solution \( \tilde{x}, \tilde{y} \) of (2) developed in Section 3 in the sense
that for all \( \sigma > \sigma_o \)
\[ e^{-\sigma t} \tilde{x}(\varepsilon, t) + e^{-\sigma t} \tilde{y}(\varepsilon, t) = e^{-\sigma t} x(\varepsilon, t) + e^{-\sigma t} y(\varepsilon, t), \]
in the space of tempered distributions as \( \varepsilon \to 0^+ \).

The applicability of Theorem 1 depends in part on being able to determine
the location of the eigenvalues of (3). We shall now address this problem.
For a matrix valued function \( Q(\varepsilon) \), \( 0 < \varepsilon < \varepsilon_o \), let \( \text{Re}(\sigma(Q(\varepsilon))) =
\{ \text{Re} : \lambda \in \sigma(Q(\varepsilon)) \} \). We wish to determine conditions on (3) which will
insure that

\[ \text{there exists } \sigma_o > 0 \text{ such that } \text{Re}(\sigma(Q(\varepsilon))) \leq \sigma_o, \ 0 < \varepsilon < \varepsilon_o. \quad (4.4) \]

Assume that \( A(\varepsilon), B(\varepsilon), C(\varepsilon), E(\varepsilon) \) are differentiable functions of \( \varepsilon \). Let
\( C(\varepsilon) = C(0) + \varepsilon \tilde{C}(\varepsilon), E(\varepsilon) = E(0) + \varepsilon \tilde{E}(\varepsilon) \) where \( \tilde{C}(\varepsilon), \tilde{E}(\varepsilon) \) are continuous in
\( \varepsilon \). Then

\[
\begin{bmatrix}
A(\varepsilon) & B(\varepsilon) \\
\frac{1}{\varepsilon} C(\varepsilon) & \frac{1}{\varepsilon} E(\varepsilon)
\end{bmatrix}
= \begin{bmatrix}
A(\varepsilon) & B(\varepsilon) \\
\tilde{C}(\varepsilon) & \tilde{E}(\varepsilon)
\end{bmatrix}
+ \frac{1}{\varepsilon}
\begin{bmatrix}
0 & 0 \\
C & E
\end{bmatrix}
= G(\varepsilon) + \frac{1}{\varepsilon} H. \quad (4.5)
\]

Let \( \text{Rm}(\sigma(X)) = \max_{\lambda \in \sigma(X)} \text{Re}(\lambda) \).

**Proposition 2.4.1** If (4) holds for (5), then \( \text{Re}(\lambda) \leq 0 \) for all \( \lambda \in \sigma(E) \).

**Proof** If (4) holds, \( \text{Rm}(\sigma(G(\varepsilon) + \frac{1}{\varepsilon} H)) \leq \sigma_o \). Thus
\[ 0 \geq \lim_{\varepsilon \to 0^+} \varepsilon \text{Rm}(\sigma(G(\varepsilon) + \frac{1}{\varepsilon} H)) = \lim_{\varepsilon \to 0^+} \text{Rm}(\sigma(\varepsilon G(\varepsilon) + H)) = \text{Rm}(\sigma(H)). \]
But \( \sigma(H) = \sigma(E) \cup \{0\} \). \( \square \)
Proposition 2.4.2 A sufficient condition for (4) to hold for (5) is that \( \text{Re}\lambda \leq 0 \) for all \( \lambda \in \sigma(E) \) and \( \text{Ind}(\lambda I - H) = 1 \) for all \( \lambda \in \sigma(H) \) such that \( \text{Re}\lambda = 0 \).

Proof If \( \text{Ind}(\lambda I - H) = 1 \) for all \( \lambda \in \sigma(H) \) such that \( \text{Re}\lambda = 0 \), then \( H \) may be written as

\[
H = T \begin{bmatrix} 
\lambda_1 & \cdots & \lambda_r \\
& \ddots & \circ \\
& & \lambda_r \\
\circ & & \\
& & & H_1
\end{bmatrix} T^{-1} \text{ where } \text{Re}\lambda_i = 0, \text{Rm}(\sigma(H_i)) < 0, T
\]

an invertible matrix. Now \( G(\varepsilon) \) is bounded. Partition \( T^{-1}GT = [G_{ij}] \) conformally with \( T^{-1}HT \). Then by a block Greschgorin theorem \[48\] there are \( r \) circles centered at the \( \lambda_i/\varepsilon \) of bounded radius and the set

\[\{ z : \| (z - H_1/\varepsilon - G_{r+1,r+1}(\varepsilon))^{-1} \|^{-1} \leq K \} = \Omega(\varepsilon) \text{ whose union contains the spectrum of } G(\varepsilon) + H/\varepsilon. \]

If we can show \( \text{Re} \Omega(\varepsilon) \) is bounded above we are finished. Suppose that \( \text{Re} \Omega(\varepsilon) \) is not bounded above. Then there exists \( \varepsilon(z), \text{Re} \varepsilon(z) \to \infty \) as \( \varepsilon \to 0^+ \), \( z(\varepsilon) \in \Omega(\varepsilon) \). But then

\[
\| (\varepsilon z(\varepsilon) - H_1 - \varepsilon G_{r+1,r+1}(\varepsilon))^{-1} \| \geq 1/\varepsilon K.
\]

Now \( H_1 - \varepsilon G_{r+1,r+1}(\varepsilon) \to H_1 \) so that \( \sigma(H_1 - \varepsilon G_{r+1,r+1}(\varepsilon)) \to \sigma(H_1) \). But \( \text{Re} \sigma(H_1) < 0 \) and \( \text{Re} \varepsilon(z(\varepsilon)) \geq 0 \). Thus the left hand side of (6) is bounded which contradicts (6) since \( 1/\varepsilon K \to \infty \) as \( \varepsilon \to 0^+ \).

If \( \text{Re}\lambda \leq 0 \) for all \( \lambda \in \sigma(E) \) and \( \text{Ind}(\lambda I - H) > 1 \) for some \( \lambda \in \sigma(E) \), then whether (4) holds depends not only on \( H \) but also the particular \( G(\varepsilon) \).
Example 2.4.1 Let \( A(\varepsilon) = E(\varepsilon) = 0, B(\varepsilon) = C(\varepsilon) = 1 \). Then \( H = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \)
so that \( \sigma(H) = \sigma(E) = \{0\} \). But \( \sigma \left( \begin{bmatrix} 0 & 1 \\ 1/\varepsilon & 0 \end{bmatrix} \right) = \{\pm 1/\sqrt{\varepsilon}\} \). On the other hand, if \( B(\varepsilon) = A(\varepsilon) = E(\varepsilon) = 0, C(\varepsilon) = 1 \), we again have \( H = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \),
\( \sigma(H) = \sigma(E) = \{0\} \), but \( \sigma \left( \begin{bmatrix} 0 & 0 \\ 1/\varepsilon & 0 \end{bmatrix} \right) = \{0\} \) which is bounded.

One might still hope that if

\[
THT^{-1} = \begin{bmatrix} H_1 & 0 \\ 0 & H_2 \end{bmatrix}, \quad \text{Re}(\sigma(H_1)) = 0, \quad \text{Re}(\sigma(H_2)) < 0, \quad (4.7)
\]

and

\[
T G(\varepsilon) T^{-1} = \begin{bmatrix} G_{11}(\varepsilon) & G_{12}(\varepsilon) \\ G_{21}(\varepsilon) & G_{22}(\varepsilon) \end{bmatrix},
\]

that it would be possible by an argument similar to the proof of Proposition 2 to reduce the checking of (4) to considering the smaller matrices \( G_{11}(\varepsilon) + H_1/\varepsilon \) and \( G_{22}(\varepsilon) + H_2/\varepsilon \). The next two examples show that this is not possible and that Proposition 2 is about as good a general result as one could hope for.

The first example shows that the diagonal blocks in (7) satisfying (4) is not sufficient for (4) to hold for \( G(\varepsilon) + H/\varepsilon \).

Example 2.4.2 Let

\[
H = \begin{bmatrix} H_1 & 0 \\ 0 & H_2 \end{bmatrix} = \begin{bmatrix} 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{bmatrix},
\]

59
and $G(\varepsilon) = \begin{bmatrix} G_{11}(\varepsilon) & G_{12}(\varepsilon) \\ G_{21}(\varepsilon) & G_{22}(\varepsilon) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\varepsilon & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$.

The characteristic polynomial of $G_{11}(\varepsilon) + H_1/\varepsilon$ is $\lambda^4 + 1$ so that it satisfies (4) and $\sigma(H_1) = \{0\}$. On the other hand $\sigma(H_2) = \{-1\}$. The characteristic polynomial of $G(\varepsilon) + H/\varepsilon$ is $(\varepsilon^2(\lambda^5 + \lambda) + \varepsilon(\lambda^4 + 1) - \lambda)/\varepsilon^2$. Setting this to zero and solving for the larger value of $\varepsilon$ gives

$$\varepsilon(\lambda) = \frac{-((\lambda^4 + 1) + [(\lambda^4 + 1)^2 + 4\lambda(\lambda^5 + \lambda)])^{1/2}}{2(\lambda^5 + \lambda)}.$$

Now $\varepsilon(\lambda) \to 0^+$ as $\lambda \to \infty$. Thus for every $\varepsilon > 0$, $G(\varepsilon) + H/\varepsilon$ has a positive eigenvalue $\lambda(\varepsilon)$ and $\lambda(\varepsilon) \to +\infty$ as $\varepsilon \to 0^+$. Hence $G(\varepsilon) + H/\varepsilon$ does not satisfy (4).

The next example shows that it is not necessary for the diagonal blocks to satisfy (4) even if the whole matrix $G(\varepsilon) + H/\varepsilon$ does.

**Example 2.4.3** Let

$$H = \begin{bmatrix} H_1 & 0 \\ 0 & H_2 \end{bmatrix} = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}, \text{ and}$$

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \varepsilon & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$
Then the characteristic polynomial of $G_{11}(\varepsilon) + H_1/\varepsilon$ is $\lambda^3 - 1/\varepsilon$ so that it does not satisfy (4). Also note that $\sigma(H_2) = \{-1\}$. The characteristic polynomial of $G(\varepsilon) + H/\varepsilon$ is $\lambda^4 + \lambda^3/\varepsilon - \lambda/\varepsilon = \lambda(\lambda^3 + \lambda^2/\varepsilon - 1/\varepsilon)$. Let $p(\lambda) = \lambda^2 + \lambda^2/\varepsilon - 1/\varepsilon$. For small $\varepsilon > 0$, $p(-2) > 0$, $p(0) < 0$, $p(1) > 0$, and $G(\varepsilon) + H/\varepsilon$ has four real eigenvalues. Two are negative, one is zero, and one is between zero and one. Hence $G(\varepsilon) + H/\varepsilon$ satisfies (4). This example exhibits the behavior claimed for Example 3 in [13]. That example is incorrect.

Recall from [8] that a matrix $A$ is semi-stable if $\Re \lambda < 0$ for all $\lambda \in \sigma(A) \setminus \{0\}$ and $\text{Ind}(A) \leq 1$. Under the assumption of semi-stability of $\begin{bmatrix} 0 & 0 \\ C & E \end{bmatrix}$, it was shown in [8] that solutions of (1) converge uniformly on compact subsets of $(0, \infty)$ as $\varepsilon \to 0^+$. In this setting the results of this section can be viewed as a statement about boundary layer convergence, that is, convergence on $[0, t_0]$, $t_0 > 0$. However, (4) is implied by the semi-stability of $\begin{bmatrix} 0 & 0 \\ C & E \end{bmatrix}$ but not conversely. Thus the results of this section are providing convergence, in some sense, for cases not covered in [8].

We conclude this section with an instructive example from [51].

**Example 2.4.5** Consider the singularly perturbed system

$$
\dot{x} = y, \quad x(0) = x_0, \tag{4.9a}
$$

$$
\varepsilon \dot{y} = -x, \quad y(0) = y_0. \tag{4.9b}
$$

For $\varepsilon > 0$, the solution of (9) is,

$$
x(t) = x_0 \cos(t/\sqrt{\varepsilon}) + \sqrt{\varepsilon} y_0 \sin(t/\sqrt{\varepsilon}), \tag{4.10a}
$$

$$
y(t) = \frac{x_0}{\sqrt{\varepsilon}} \sin(t/\sqrt{\varepsilon}) + y_0 \cos(t/\sqrt{\varepsilon}). \tag{4.10b}
$$
Applying Laplace transforms to (9) gives

\[
\begin{bmatrix}
    s & -1 \\
    1 & \epsilon s
\end{bmatrix}
\begin{bmatrix}
    \hat{x}_\epsilon \\
    \hat{y}_\epsilon
\end{bmatrix}
= \begin{bmatrix}
    x_0 \\
    \epsilon y_0
\end{bmatrix}.
\] (4.11)

Thus

\[
\begin{bmatrix}
    \hat{x}_\epsilon \\
    \hat{y}_\epsilon
\end{bmatrix} = \begin{bmatrix}
    s & -1 \\
    1 & \epsilon s
\end{bmatrix}^{-1} \begin{bmatrix}
    x_0 \\
    \epsilon y_0
\end{bmatrix} = \frac{1}{\epsilon s^2 + 1} \begin{bmatrix}
    \epsilon s & 1 \\
    -1 & s
\end{bmatrix} \begin{bmatrix}
    x_0 \\
    \epsilon y_0
\end{bmatrix}
\]

\[
= \frac{1}{\epsilon s^2 + 1} \begin{bmatrix}
    \epsilon sx_0 + \epsilon y_0 \\
    -x_0 + \epsilon sy_0
\end{bmatrix}.\]

Then \( \lim_{\epsilon \to 0^+} \begin{bmatrix}
    \hat{x}_\epsilon(s) \\
    \hat{y}_\epsilon(s)
\end{bmatrix} = \begin{bmatrix}
    0 \\
    -x_0
\end{bmatrix}.

Thus the solution should be

\[
x(t) = 0, \quad y(t) = -x_0 \delta(t). \tag{4.12}
\]

Setting \( \epsilon = 0 \) in (9) gives

\[
x = y, \tag{4.13a}
\]

\[
0 = -x \quad \text{for } t > 0, \quad x(0) = x_0. \tag{4.13b}
\]

Given (13b), from (13a) we have \( y = -x(0) \delta \) as in (12). On the other hand, two functions that agree almost everywhere define the same distribution. Thus (13b) and \( x \equiv 0 \) from (12) agree as distributions. Since \( x(t) \equiv 0 \) does not satisfy (13b) as a function, care must be exercised when talking of distributional solutions as noted earlier.

Now look at (10). According to (12) and Theorem 1;

\[
\cos(t/\sqrt{\epsilon}) \to 0, \tag{4.14a}
\]

\[
\frac{\sin(t/\sqrt{\epsilon})}{\sqrt{\epsilon}} \to \delta. \tag{4.14b}
\]
in the sense of tempered distributions. This is not obvious. While we will not prove (14), a simple calculation should help to put (14) in perspective.

Consider \( \cos(t/\sqrt[3]{\varepsilon}) \), \( \sin(t/\sqrt[3]{\varepsilon}) \) as being zero for \( t < 0 \). Then from Theorem 1 we are to show for each \( \delta > 0 \)

\[
e^{-\delta t} \cos(t/\sqrt[3]{\varepsilon}) \to 0,
\]

\[
e^{-\delta t} \sin(t/\sqrt[3]{\varepsilon}) \to \delta,
\]

as tempered distributions. That is, for each \( f \in S \), as \( \varepsilon \to 0^+ \)

\[
<f, e^{-\delta t} \cos(t/\sqrt[3]{\varepsilon})> \to <f, 0> = 0,
\]

\[
<f, e^{-\delta t} \sin(t/\sqrt[3]{\varepsilon})> \to <f, \delta> = f(0).
\]

We shall verify (16) for a "typical" function in \( S \). For any \( r > 0 \), let \( f(t) \) be an infinitely differentiable function on \((-\infty, \infty)\) such that \( f(t) = e^{-rt}, \)

\( t \geq 0 \), and \( f(t) = 0 \) for \( t < -1 \). Then

\[
<f, e^{-\sigma t} \cos(t/\sqrt[3]{\varepsilon})> = \int_0^\infty e^{-(\sigma+r)t} \cos(t/\sqrt[3]{\varepsilon}) dt
\]

\[
= \lim_{t \to \infty} \frac{e^{-(\sigma+r)t}}{(\sigma + r)^2 + 1/\varepsilon} \left[ -(\sigma + r) \cos(t/\sqrt[3]{\varepsilon}) + \frac{\sin(t/\sqrt[3]{\varepsilon})}{\sqrt[3]{\varepsilon}} \right]_0^t
\]

\[
= \frac{\sigma + r}{(\sigma + r)^2 + 1/\varepsilon}.
\]

Letting \( \varepsilon \to 0^+ \), (16a) follows.

On the other hand,

\[
<f, e^{-\sigma t} \sin(t/\sqrt[3]{\varepsilon})> = \left( \frac{1}{\sqrt[3]{\varepsilon}} \right) \int_0^\infty e^{-(\sigma+r)t} \sin(t/\sqrt[3]{\varepsilon}) dt
\]
\[ e^{-(\sigma+r)t} \left( \frac{1}{(\sigma + r)^2 + 1/\varepsilon} \right) \left. \left. \frac{1}{\sqrt{\varepsilon}} \right| \right| \begin{vmatrix} - (\sigma + r) \sin(t/\sqrt{\varepsilon}) - \frac{\cos(t/\sqrt{\varepsilon})}{\sqrt{\varepsilon}} \end{vmatrix}^T \right|_0^T \]

Thus (16b) holds since \( f(0) = 1 \).

5. FEEDBACK CONTROL OF IMPULSES

In Chapter 1 the problem of stabilization using state feedback was considered.

In this section, we shall discuss the problem of eliminating impulses by state feedback. Take the system

\[ \dot{A}x + Bx = Cu \]  

(5.1)

which is assumed solvable, \( x \in \mathbb{R}^n \), \( u \in \mathbb{R}^m \). Let \( \sigma(A,B) = \{\lambda_1, \ldots, \lambda_r\} \) be those \( \lambda \) for which \( \lambda A + B \) is singular. The \( \lambda_i \) are sometimes called the eigenvalues of (1). If \( \det(\lambda A + B) = \sigma \prod_{i=1}^r (\lambda - \lambda_i)^{m_i} \), then \( \Sigma m_i < n \). When \( A \) is singular, \( m = n - \Sigma m_i > 0 \) and \( \lambda A + B \) is said to have \( m \) infinite eigenvalues.

Again write (1) in the form

\[ \dot{z}_1 = B_1 z_1 + C_1 u \]  

(5.2a)

\[ \dot{z}_2 = z_2 + C_2 u, \]  

(5.2b)

Following Cobb, we shall refer to (2a) as the **slow subsystem** and (2b) as the **fast subsystem**. An arbitrary linear state feedback law would be of the form

\[ u = Kx + v = K_1 z_1 + K_2 z_2 + v. \]  

(5.3)

Substituting (3) into (1) gives

\[ \dot{A}x + (B - CK)x = Cy. \]  

(5.4)
Assume that
\[ \lambda A + B = CK \] is invertible for some \( \lambda \). \hfill (5.5)

Note that since \( \lambda A + B \) is invertible for some \( \lambda \), that (5) holds for all sufficiently small \( K \) and hence on an open dense subset of \( \mathbb{C}^{m \times n} \). Then the impulses will not be present in (4) for any initial condition and any control \( Cv \) if and only if
\[ \text{Ind}[(\lambda A + B - CK)^{-1}A] \leq 1. \hfill (5.6) \]

Also, the number of "infinite eigenvalues" of (4) will then be
\[ \text{Im}(N(\lambda A + B - CK)^{-1}A)) = \dim N(A) = \dim N(N) \], where \( N \) is from (2b).

Thus if it is possible to eliminate impulses by state feedback of the form \( u = K_2 z_2 \), one cannot reduce the number of infinite eigenvalues any further by considering the more general feedback (3). We shall consider then the effect of two types of feedback;

"slow feedback," \( u = K_1 z_1 + v \), \hfill (5.7)

and

"fast feedback," \( u = K_2 z_2 + v \). \hfill (5.8)

In terms of (1), (7) is \( u = K_1\hat{A}^D x + v \), (8) is \( u = K_2 (I - \hat{A}^D\hat{A}) x + v \) where
\[ \hat{A} = (\lambda A + B)^{-1}A. \]

**Slow feedback**

If slow feedback (7) is applied to (2), under the assumption (5), system (2) becomes.
\[
\begin{bmatrix}
I & 0
\end{bmatrix}
\begin{bmatrix}
\dot{z}_1 \\
\dot{z}_2
\end{bmatrix} =
\begin{bmatrix}
B_1 + C_1 K_1 & 0 \\
C_2 K_1 & I
\end{bmatrix}
\begin{bmatrix}
\dot{z}_1 \\
\dot{z}_2
\end{bmatrix} +
\begin{bmatrix}
C_1 \\
C_2
\end{bmatrix} v.
\]

(5.9)

One could now argue using the Laplace transform of (9). We shall continue our state variable approach which is equivalent. Multiplying (9) by

\[
\begin{bmatrix}
\lambda I - B_1 - C_1 K_1 & 0 \\
-C_2 K_1 & \lambda N - I
\end{bmatrix}^{-1}
\begin{bmatrix}
(\lambda I - B_1 - C_1 K_1)^{-1} & 0 \\
(\lambda N - I)^{-1} C_2 K_1 (\lambda I - B_1 - C_1 K_1)^{-1} & (\lambda N - I)^{-1}
\end{bmatrix}
\begin{bmatrix}
G_1 \\
(\lambda N - I)^{-1} C_2 K_1 (\lambda I - B_1 - C_1 K_1)^{-1} & 0
\end{bmatrix}
\]

gives

\[
\begin{bmatrix}
G_1 \\
(\lambda N - I)^{-1} C_2 K_1 G_1 & N(\lambda N - I)^{-1}
\end{bmatrix}
\begin{bmatrix}
\dot{z}_1 \\
\dot{z}_2
\end{bmatrix} +
\begin{bmatrix}
\ddot{C}_1 \\
\ddot{C}_2
\end{bmatrix} v,
\]

(5.10)

\[\ddot{C}_2 = (\lambda N - I)^{-1} C_2 + (\lambda N - I)^{-1} C_2 K_1 G_1 C_1, \quad \ddot{C}_1 = G_1 C_1, \quad G_1 \text{ is invertible and}
\]

\[N(\lambda N - I)^{-1} \text{ is nilpotent. Thus a similarity of the form}
\begin{bmatrix}
I & 0 \\
L & I
\end{bmatrix}
\begin{bmatrix}
\dot{z}_1 \\
\dot{z}_2
\end{bmatrix} =
\]

\[
\begin{bmatrix}
I & 0 \\
L & I
\end{bmatrix}
\begin{bmatrix}
\dot{w}_1 \\
\dot{w}_2
\end{bmatrix},
\]

gives

\[G_1 \dot{w}_1 = (\lambda G_1 - I) \dot{w}_1 + \ddot{C}_1 v,
\]

(5.11a)

\[(\lambda N - I)^{-1} \dot{w}_2 = (\lambda N - I)^{-1} \dot{w}_2 + \ddot{C}_2 v,
\]

(5.11b)
\[ \dot{c}_2 = -Lc_1 + \hat{c}_2 \] where \( L \) is the solution of \(-LC_1 + (\lambda N - I)^{-1}C_2 K_1 G_1 + N(\lambda N - I)^{-1}L = 0\) which exists since \( \sigma(G_1) \cap \sigma(N(\lambda N - I)^{-1}) \) is empty [98].

Multiplying (11a) by \( C_1^{-1} \), and (11b) by \((\lambda N - I)\), and noting \( z_1 = w_1 \) we have

\[
\dot{z}_1 = (B_1 + C_1 K_1)z_1 + C_1 u \tag{5.12a}
\]

\[
N\dot{z}_2 = w_2 + C_2 u \tag{5.12b}
\]

where the \( N, C_1 \) in (11b) are the same as those in (2). Thus slow feedback essentially leaves the structure of the fast system unchanged and the results of Section 1.4 are applicable.

**Fast feedback**

Applying fast feedback (8) to (2) gives

\[
\dot{z}_1 = B_1 z_1 + C_1 K_2 z_2 + C_1 v \tag{5.13a}
\]

\[
N\dot{z}_2 = (I + C_2 K_2)z_2 + C_2 v \tag{5.13b}
\]

Before considering (13) we need to analyze (13b).

**Theorem 2.5.1** (Cobb [31]) There exists a \( K_2 \) satisfying (5) so that (13b) does not admit impulsive behavior if and only if

\[
\text{Ind}[(\lambda N - I - C_2 K)^{-1}N] = 1 \tag{5.14}
\]

or equivalently,

\[
\text{dim}[R(N) + N(N) + R(C_2)] = r \tag{5.15}
\]

where \( z_2 \in \Phi^r \).

67
Proof From Chapter 0 (or Section 3), impulses will not be present for any \( x(0) \), \( y(0) \) smooth, if and only if (14) holds. Now (14) is equivalent to

\[
\{0\} = R((\lambda N - I - C_2 K_2)^{-1} N) \cap N((\lambda N - I - C_2 K_2)^{-1} N)
\]

\[
= R((\lambda N - I - C_2 K_2)^{-1} N) \cap N(N). \text{ That is,}
\]

\[
\{0\} = R(N) \cap \{(\lambda N - I - C_2 K_2)N(N)\} = R(N) \cap \{(I + C_2 K_2)N(N)\}. \quad (5.16)
\]

Suppose that (16) holds. Since \( N - I - C_2 K_2 \) is invertible and \( \dim(R(N)) + \dim(N(N)) = r \), we have \( \mathcal{K} = R(N) \oplus (\lambda N - I - C_2 K_2)N(N) \subseteq R(N) + N(N) + C_2 K_2 N(N) \subseteq R(N) + N(N) + R(C_2) \subseteq \mathcal{K} \). Thus (14) implies (15). Conversely, suppose that (15) holds. Since \( N \) is nilpotent \( R(N) + N(N) \neq \mathcal{K} \). Thus there is a subspace \( M \subseteq R(C_2) \) so that \( [R(N) + N(N)] \oplus M = \mathcal{K} \). Let \( N(N) = N_2 \oplus \{R(N) \cap N(N)\} \), \( R(N) = R_1 \oplus \{R(N) \cap N(N)\} \). Then \( R(N) + N(N) = N_1 \oplus R_1 \oplus \{R(N) \cap N(N)\} \). But \( r = \dim R(N) + \dim N(N) = \dim N_1 + \dim R_1 + 2 \dim(R(N) \cap N(N)) \) and \( r = \dim N_1 + \dim R_1 + \dim M + \dim(R(N) \cap N(N)) \). Thus \( \dim M = \dim(R(N) \cap N(N)) \). Define \( K_2 \mid N_1 = 0 \) and \( C_2 K_2 \) to map \( R(N) \cap N(N) \) onto \( M \). To complete the definition of \( K_2 \) let \( K_2 \mid R_1 = 0 \). Then (14) holds. \( \square \)

There are many ways to define \( K_2 \) on \( R(N) \cap N(N) \). For example, if \( \{m_1, \ldots, m_s\} \) is a basis for \( M \), and \( \{p_1, \ldots, p_s\} \) a basis for \( R(N) \cap N(N) \), let \( K_2 p_1 = C_2^+ m_1 \). Then \( C_2 K_2 p_1 = C_2 C_2^+ m_1 = m_1 \) since \( m_1 \in R(C_2) \).

An argument similar to that used for slow feedback shows that (13) may be rewritten as

\[
\dot{\tilde{y}}_1 = B_1 v_1 + C_1 v,
\]

\[
N \dot{y}_2 = (I + C_2 K_2) y_2 + \tilde{C}_2 v.
\]

\[5.17a\]

\[5.17b\]
for an open dense set of $K_2$'s, namely those for which (5) holds and for which
\[
\sigma((\lambda - B_1)^{-1}) \cap \sigma((\lambda N - I + C_2K_2)^{-1}N) = \phi. \tag{5.18}
\]
Thus most fast feedback does not alter the eigenvalues of the slow system, \(\sigma(B_1)\). However, it can sometimes introduce new finite eigenvalues into the system (17b). Can they be made stable? That is, \(\text{Range}(N,I + C_2K_2) < 0\)?

Suppose that $K_2$ is such that assumptions (5) and (18) hold and that $K_2$ is constructed as in the proof of Theorem 1. Then $\delta K_2$ also works for any $\delta \neq 0$ such that (5) and (18) hold for $\delta K_2$. Conditions (5) and (18) exclude at most a finite number of $\delta$'s for each $\lambda$. Suppose that $||C_2K_2|| < 1$ and let \(\gamma_1, \ldots, \gamma_m\) be the finite number of scalars for which $\gamma N - I - C_2K_2$ is singular. Note that $\gamma N - I - C_2K$ is singular if and only if $I + (I - \gamma N)^{-1}C_2K_2$ is. Thus by taking $K_2$ small enough to start with, $\gamma_1 \notin \sigma(B_1)$ for all $i$. Now note that if $N_\lambda = (\lambda N - I - C_2K_2)^{-1}N$, then $N_\lambda$ has index one for $\lambda$ in a neighborhood of zero by the choice of $K_2$ and the proof of Theorem 1. Thus by analytic continuation (5), (6), (18) hold for all except a finite number of $\lambda$. Repeat the process used in deriving (2) to get
\[
\dot{w}_2 = Bw_2 + C_3v \tag{5.19a}
\]
\[0 = w_3 + C_4v \tag{5.19b}
\]
in place of (17b). Here \(\begin{bmatrix} w_2 \\ w_3 \end{bmatrix}\) are coordinates of $x_2$ with respect to the decomposition $R(N_\lambda^D N_\lambda^{-1}) \oplus R(I - N_\lambda^D N_\lambda^{-1})$, which is independent of $\lambda$, and \(\begin{bmatrix} B & 0 \\ 0 & 0 \end{bmatrix}\) is the matrix, with respect to this same decomposition, of
\[ \tilde{N}^D \tilde{N} (\tilde{N}N - I) \tilde{N} \tilde{N} = \lambda \tilde{N}^D \tilde{N} - \tilde{N}^D = \tilde{N}^D (\lambda - \tilde{N}^D) \tilde{N} \tilde{N}. \]

Note that if \( \gamma \neq 0 \)
\[ \gamma - \tilde{N}_\lambda = (\lambda \tilde{N} - I - C_2 K_2)^{-1} \gamma (\lambda \tilde{N} - I - C_2 K_2 - \frac{1}{\gamma} N). \]

Thus \( \gamma \neq 0, \gamma \in \sigma(\tilde{N}^D_\lambda) \) if and only if \( \gamma = \lambda - \gamma_1. \) Thus \( \sigma(\tilde{N}) = \{\gamma_1, \ldots, \gamma_r\}. \)

The question is if it is possible to choose \( K_2 \) so that (5), (14) hold and \( Rm\sigma(\tilde{N}) < 0? \) (Incidently, we could have gotten to this point more quickly if our development had been in terms of pole placement, etc.) The answer is, yes. A proof is given in [31] based on the fact that the eigenvalues of (19a) are controllable and thus can be placed arbitrarily by additional state feedback. Details may be found in [31].

The preceding discussion, together with [31] pretty well covers the complete elimination of impulses by fast state feedback. What of the worst possible case? Namely, when no change at all is possible. Also, if a change is possible, do there exist easy to compute matrices that will do the job?

**Proposition 2.3.1** The system
\[ \dot{N}x = x + Cu, \]  
(5.20)

\( N \) nilpotent, does not have a finite eigenvalue for any feedback of the form \( u = Kx, \) with
\[ \lambda N - I - CK \text{ invertible some } \lambda \]

if and only if \( NC = 0. \)
Proof Suppose that (20) does not have a finite eigenvalue for any $K$ for which (21) holds. Take $H$ to be an arbitrary matrix. For $\varepsilon$ sufficiently small $K = \varepsilon H$ satisfies (19). Let $u = \varepsilon Hx$. Then (20) becomes $N\dot{x} = x + \varepsilon CHx$.

Let $\varepsilon$ be small enough that $I + \varepsilon CH$ is invertible. Then by assumption, $(I + \varepsilon CH)^{-1}N$ is still nilpotent. Thus $0 = \text{trace}[(I + \varepsilon CH)^{-1}N]$

$$= \sum_{m=0}^{\infty} (-\varepsilon)^m \text{trace}[(CH)^m N] = 0 \text{ for some neighborhood of zero. Hence trace(CHN) = 0. Let } H = C^* N^* . \text{ Then } 0 = \text{trace}(CC^* N N) = \text{trace}(C^* N NC) = \text{trace}[(NC)^* NC] \text{ which implies } NC = 0 \text{ as desired.}$$

Suppose on the other hand, that $NC = 0$. Let $K$ be any matrix satisfying assumption (21). Then $(\lambda N - I)C = -C$. Hence $C = -(\lambda N - I)^{-1}C$. But then $(\lambda N - I - CK) = (\lambda N - I)[I - (\lambda N - I)^{-1}CK] = (\lambda N - I)^{-1}(I + CK)$. Since $(\lambda N - I)$ is invertible for all scalars $\lambda$ we have $I + CK$ is invertible and thus $\lambda N - I - CK$ is invertible for all finite $\lambda$. That is, $\lambda N - I - CK$ does not have any finite eigenvalues.

As a consequence of the proof of Proposition 1, we have

**Corollary 2.5.1** If $NC \neq 0$, then $u = \varepsilon Kx$ produces a finite eigenvalue for all but a finite number of $\varepsilon$ if $\text{trace}(CK) \neq 0$.

**Proof** Since $I + \varepsilon CK$ is invertible for $\varepsilon = 0$, it is invertible for all but a finite number of $\varepsilon$. Let $u = \varepsilon Kx$ where $(I + \varepsilon CK)$ is invertible. Then (20) becomes

$$(I + \varepsilon CK)^{-1}N\dot{x} = x + (I + \varepsilon CK)^{-1}Cy$$

(5.22)

Now if $\text{Trace}((I + \varepsilon CK)^{-1}N) \neq 0$, then $(I + \varepsilon CK)^{-1}N$ is not nilpotent and (22) will have finite eigenvalues. But $\text{trace}((I + \varepsilon CK)^{-1}N)$
\[ = -\varepsilon \sum_{m=0}^{\infty} (-\varepsilon)^m \text{trace}((\mathbf{C}\mathbf{K})^m \mathbf{N}) \] for small \( \varepsilon \). If \( \text{trace}(\mathbf{C}\mathbf{K}\mathbf{N}) \neq 0 \), then \( \text{trace}(\mathbf{I} + \varepsilon \mathbf{C}\mathbf{K})^{-1} \mathbf{N} \neq 0 \) for small \( \varepsilon \) and hence for all but a finite number of \( \varepsilon \). \[\square\]

The condition \( \text{trace}(\mathbf{C}\mathbf{K}\mathbf{N}) \neq 0 \) is sufficient but not necessary.

**Example 2.5.1** Let \( \mathbf{N} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{C} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \), so that \( \mathbf{N}\mathbf{C} \neq 0 \). Then \( \mathbf{u} = \varepsilon \mathbf{C} \mathbf{N} \mathbf{x} \) has the desired behavior.

\[
(I + \varepsilon \mathbf{C}\mathbf{N})^{-1} \mathbf{N} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \varepsilon & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -\varepsilon \end{bmatrix}.
\]

The new eigenvalues are \( \infty, \infty, -1/\varepsilon \). Note that \(-1/\varepsilon\) may be placed arbitrarily except for 0 using this feedback.

On the other hand if \( \mathbf{u} = \varepsilon \mathbf{K}\mathbf{x} = \varepsilon [1, 0, 0] \mathbf{x} \), \( \text{trace}(\mathbf{C}\mathbf{K}\mathbf{N}) = 0 \) and

\[
(I + \varepsilon \mathbf{C}\mathbf{K})^{-1} \mathbf{N} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \varepsilon & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -\varepsilon & 0 \end{bmatrix}
\]

which has eigenvalues of \( \{\infty, \pm i/\sqrt{\varepsilon} \} \).
3 Circuits and circuit models

1. INTRODUCTION

In Chapter 5 we shall begin considering non-linear singular systems of differential equations. As with the nonsingular case, it will be necessary to make various assumptions. Rather than just making those assumptions that permit the proving of theorems, we prefer to make assumptions motivated by, or compatible with, various applications. As in [8], various control problems will be considered. However, our principal applications will be in terms of electrical circuits, or more properly, state-variable models for circuits.

Of course, many physical processes can be modeled in a manner similar to circuits. This was the idea behind the analog computer. Much of the current interest in singular systems is among control theorists, systems theorists, and electrical engineers. Also readability is improved by working within the framework of a particular application. For these reasons we shall primarily discuss singular non-linear systems within the context of circuit and control theory. The remainder of this chapter is a quick introduction to certain aspects of circuit theory. It is aimed at those readers whose knowledge of circuits is limited to simple linear RLC circuits or, perhaps, to no knowledge at all. The presentation will be informal, non-rigorous and intuitive. For the reader who has a more sophisticated knowledge of circuits, a glance over this chapter should help to standardize notation and terminology.
2. BASIC LAWS

We shall consider only lumped parameter circuits. The word circuit means that behavior is determined solely by positions along the path. Thus field effects are ignored. Lumped parameter means that the effects of the various components may be considered to be concentrated at a point. The converse of lumped parameter is distributed parameter. Antennas are an example of a distributed parameter system.

We will not worry about units such as ohms, but rather work with dimensionless variables. For our purposes the exact scale is usually not important. In practice, of course, scale is quite important. The idea of small or parasitic elements is closely related to that of reduced order model and singular perturbations. Both these approaches often lead to singular systems.

At each point in the circuit there are two quantities of interest. There is voltage (or potential) and current (or flow of charge). Current is, by convention [105, p. 9], the net flow of positive charge. A branch is part of a circuit with two terminals to which connections can be made, a node is the point where two or more branches come together, and a loop is a closed path formed by connecting branches" [105, p. 29].

Kirchoff’s two laws are that:

The algebraic sum of the currents into a node at any instant is zero, \( (KC) \)

The algebraic sum of the voltage drops around a loop at any instant is zero. \( (KC) \)

To set up the circuit equations a current variable is assigned to each branch. One can talk either of the potentials (voltages) at the nodes or of potential drops along the branches. Kirchoff’s current law may then be applied at each node, the voltage law to each loop. Of course some of these
equations will be redundant. For large scale linear systems the determination of a minimal set of equations, and thus a minimum number of variables, can be computationally nontrivial, but is straightforward using graph theory [27].

Consider the following two terminal device (or one-port).

![Diagram of a two terminal device](Image)

For our purposes, the behavior of the device is completely determined if we know \( v = v_o - v_i \) and \( i \) at any time \( t \). The relationship between \( v \) and \( i \) is called the \textit{v-i characteristic} of the device. There are three types of two terminal devices which we shall consider.

There is a loss or gain of power in a \textit{resistor}. (Some authors use the word resistance.) The resistors studied at the introductory level are often \textit{dissipative elements}. That is, there is an irreversible loss of power in the resistor. The amount of power loss (or gain) is determined by the current. If the voltage drop across the resistor is uniquely determined by the current, \( v = f(i,t) \), the resistor is said to be \textit{current controlled}. We shall primarily consider autonomous current controlled resistors, that is, \( v = f(i) \). If \( v = Ri \), \( R \) a function only of \( t \), the resistor is said to be \textit{linear}, and \( R \) is called the \textit{resistance}. If a resistor is not linear, it is called nonlinear. We shall use the following symbol for a resistor; 

![Symbol for a resistor](Image)

Nonlinear resistors are often
A capacitor stores energy in the form of a charge \( q \) or an electric field. For autonomous linear capacitors \( q = C v \) or equivalently, \( i = C \frac{dv}{dt} \), \( C \) is constant and the symbol is \( \square \). One could consider non-linear capacitors \( \frac{dv}{dt} = f(i) \), for example, or non-autonomous capacitors, \( C \) varies with \( t \), but we shall not do so. Capacitors are sometimes specified by giving their \( q - v \) characteristic.

An inductor stores energy in the form of a magnetic field. The voltage-current relationship for a linear inductor is \( v = L \frac{di}{dt} \), the symbol is \( \bullet \bullet \bullet \bullet \bullet \). Inductors are characterized by their \( \phi - v \) characteristic where \( \phi \) is the flux.

In addition, there are short circuits, denoted \( \square \square \square \square \square \) where \( v = 0 \) for any \( i \). A short circuit is just a linear resistor with zero resistance. There are open circuits, denoted \( \square \square \square \) where \( i = 0 \) for any \( v \). There are voltage sources; \( \square \ ) where \( v = v \) for any \( i \), and current sources; \( \square \square \ ) where \( i = i \), for any \( v \). A circuit consisting of only these devices and linear resistors, inductors and capacitors is called a linear RLC-circuit. In the next section we shall briefly discuss linear RLC circuits. The section following that will consider nonlinear circuits and certain three terminal devices such as transistors.

The voltage at a node will be considered as identically zero if the node is grounded. This is denoted \( \square \).

3. LINEAR CIRCUITS

Many devices can be modeled by a combination of current or voltage sources, (nonlinear) resistors, and linear capacitors. Since over a narrow enough operational range a nonlinear resistor can usually be approximated by a
linear one, many models consist only of linear resistors and capacitors. For this reason RC-circuits are often discussed.

Consider a linear RLC circuit with N nodes and B branches. For each branch one node is denoted as a +, the other as a −. (This is the same as assigning the current variables in each branch.) The current is drawn as + to −). The B-dimensional vector of branch currents is \( i_B \). At each node there is a nodal voltage. The N-vector of nodal voltages is \( v_N \). The network incidence matrix \( A \) has a 1 in row \( i \), column \( j \) if node \( j \) is the + node for branch \( i \). All other entries are zero. Thus a row \( i \) of \( A \) has only two nonzero entries corresponding to the nodes defining branch \( i \). The \( j^{th} \) column of \( A \) has entries corresponding to all the branches that interact at node \( j \). Kirchoff's current law is thus \( A^T i_B = 0 \). The branch voltage vector \( v_B \) is given by \( A v_N = v_B \). Since this includes all nodes and hence all loops, \( A v_N = v_B \) is Kirchoff's voltage law.

The branches may be grouped as voltage sources, capacitors, resistors, inductors, and current sources so that

\[
v_B = \begin{bmatrix}
v_E \\
v_C \\
v_R \\
v_I \\
v_S \\
\end{bmatrix}, \quad i_B = \begin{bmatrix}
i_E \\
i_C \\
i_R \\
i_I \\
i_S \\
\end{bmatrix}
\]

The appropriate v-i characteristics for linear circuits are \( v_E = v(t) \), \( C v_C = i_C \), \( v_R = R i_R \), \( v_I = L i_I \), \( i_S = i(t) \) where \( C \), \( R \), are diagonal nonsingular matrices. If the inductors are uncoupled, then \( L \) is also diagonal. If the inductors are "mutually coupled," then \( L \) would be positive semi-definite.

Repartition \( A \) as
\[ A = \begin{bmatrix} A_E \\ A_C \\ A_R \\ A_I \\ A_S \end{bmatrix} \]

The circuit equations are thus in the form

\[ \begin{bmatrix} v_E \\ \vdots \\ v_C \\ \vdots \\ v_R \\ \vdots \\ v_I \\ \vdots \\ v_S \end{bmatrix} + \begin{bmatrix} 0 & \cdots & -I & 0 & 0 & 0 & 0 & 0 & A_E \\ 0 & \cdots & 0 & \cdots & 0 & 0 & 0 & 0 & A_C \\ 0 & \cdots & 0 & \cdots & 0 & 0 & 0 & 0 & A_R \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & A_I \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & A_S \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{v}_E \\ \vdots \\ \dot{v}_C \\ \vdots \\ \dot{v}_R \\ \vdots \\ \dot{v}_I \\ \vdots \\ \dot{v}_S \end{bmatrix} = \begin{bmatrix} \dot{e}(t) \\ \vdots \\ \dot{v}_C \\ \vdots \\ \dot{v}_R \\ \vdots \\ \dot{v}_I \\ \vdots \\ \dot{v}_S \end{bmatrix} \tag{3.1} \]

where \( H \) has \( C \) in the 2,2-position, \( L \) in the \( 8,8 \)-position and all other entries zero. The system (1) is in the form

\[ H \ddot{x} + Gx = \ddot{f} \tag{3.2} \]

with \( H \) singular. The matrices \( R, I \) in \( G \) and \( C \) in \( H \) are diagonal. Thus \( H, G \) are in general quite sparse. For large circuits it may be desirable to try and maintain this sparsity as much as possible.
The system (2) is said to have a state variable form if there are variables $u = P \bar{x}$ so that (2) is equivalent to

$$\dot{u} = g(u,t), \quad u(0) \text{ arbitrary.} \quad (3.3)$$

Reduction to state variable form can destroy sparseness.

**Index** One way traditionally used to solve (2) is by the shuffle algorithm [73], [74]. Mathematically this is the same as Wilkinson's procedure [8, p. 154-158]. The number of shuffles, when $\lambda G + H$ is invertible for some $\lambda$, is closely related to the index of $(\lambda G + H)^{-1}H$. While the theory we have been developing in [8] and this volume has involved matrices of arbitrarily large index, it is not necessarily obvious that such matrices ever arise in dealing with practical circuits. To apply most of our theory it is also necessary that $\lambda H + G$ be invertible. The remainder of this section will examine to what extent the invertibility of $\lambda H + G$ is a natural circuit assumption and give examples to show that higher index systems naturally arise.

In [103, Chapter 4] it is shown that for any connected RC-circuit with independent voltage sources, the system (2) is solvable. That is, $\lambda G + H$ is invertible.

Recall that the solution of (2) involves up to $k - 1$ derivatives of $\bar{f}$ where $k = \text{Ind}((\lambda G + H)^{-1}H)$. Thus the index of $(\lambda G + H)^{-1}H$ in (2) has to be at least one greater than the number of times that $e(t)$ is differentiated in the solution of (1).
Operational amplifiers

The operational amplifier is a three terminal device denoted by

\[ i \rightarrow v_1 \quad + \quad v_3 \quad \rightarrow \infty \quad \quad \quad v_2 \quad - \]

(3.4)

Terminals 1, 2 are called input terminals and form the input branch. Terminal (node) 3 is the output terminal. For an ideal operational amplifier, no voltage can drop across the input branch, no current can flow through the input branch, and the amplifier is said to have infinite gain. This may seem rather odd if one is unfamiliar with it.

Consider an operational amplifier (4) with gain \( A \). Typically \( A \approx 10^5 \).

then \( i \approx 0 \) and \( v_3 = A(v_1 - v_2) \).

To see how the infinite gain of an ideal operational is used, now consider the circuit.

(3.5)

Here node 4 is grounded so that \( v_4 = 0 \). The current law at node 2 is

\[
\frac{v_1 - v_2}{R_1} + \frac{v_3 - v_2}{R_2} = 0
\]

(3.6)

since the input current is negligible. But \( v_3 = Av_2 \), or \( v_2 = v_3/A \). Thus
\[
\frac{v_1 - v_3/A}{R_1} + \frac{v_3 - v_3/A}{R_2} = 0, \text{ so that }
\]
\[
\frac{v_1}{R_1} - \frac{v_3}{AR_1} + \frac{v_3}{R_2} - \frac{v_3}{R_2A} = 0. \tag{3.7}
\]

If \( A \) is large, (7) is approximately
\[
\frac{v_1}{R_1} + \frac{v_3}{R_2} = 0 \tag{3.8}
\]
or
\[
\frac{v_3}{R_2} = -\frac{R_2}{R_1} v_1. \tag{3.9}
\]

Thus the output voltage \( v_3 \) is an amplification of \( v_1 \). Notice that the relationship (9) would also result from (6) by setting \( v_2 = 0 \). That is \( v_2 = v_4 = 0 \) since there is no voltage drop across the input branch of an ideal operational amplifier. The ideal operational amplifier is thus a limiting case of what happens as the gain goes to infinity.

Operational amplifiers are widely used in modern circuit design. A simple differential amplifier is given by

\[
(3.10)
\]
Now $v_4 = 0$ in (10) since it is grounded. Thus $v_1 = e$. For an ideal operational amplifier $v_2 = 0$. Since $i = 0$, the current in the capacitor $C(v_1 - v_2)$ is the same as the current in the resistor $(v_2 - v_3)/R$. Thus $-v_3/R = C \dot{e}$, or

$$v_3 = -CRe.$$ (3.11)

Note that the solution of (1) for this circuit would involve a derivative of $e$ and thus the index of $(\lambda G + H)^{-1}H$ in (2) would be at least two for this circuit.

By connecting the differential amplifier in cascade one can get derivatives of arbitrary orders. For example in

![Diagram](Image)

we have from the preceding example that $v_1 = -C_1R_1 \dot{e}$. Thus $v_2 = -C_2R_2 \dot{v}_1 = C_2R_2C_1R_1 \ddot{e}$. Similarly, $v_3 = -C_3R_3 \dddot{v}_2 = -C_3R_3C_2R_2C_1 \dddot{e}$. For this circuit the index of (1) would be at least four.

The presence of operational amplifiers is interesting in another way. For linear RC circuits, solvability is determined by the network topology, namely whether it is connected. If the circuit possesses operational amplifiers, then whether it is solvable, and the number of initial conditions that can be specified, can depend on specific values of the various resistors, capacitors, etc. The two examples we present are from [103].

Consider first
For (12), the system (2) has $G = 0$. Thus $H$ must be invertible for (12) to be solvable. $H$ will be invertible if and only if $\det H = \frac{1}{R_2 R_3} - \frac{1}{R_1 R_4} \neq 0$ so that solvability depends on the values of $R_1$, $R_2$, $R_3$, $R_4$. Next consider

Let $G_1 = 1/R_1$. The nodal equations are

$$
\begin{align*}
\begin{bmatrix}
C & 0 & 0 & 0 & 0 & \mathbf{v}_1 \\
0 & 0 & 0 & 0 & 0 & \mathbf{v}_2 \\
0 & 0 & 0 & 0 & 0 & \mathbf{v}_3 \\
0 & 0 & 0 & 0 & 0 & \mathbf{v}_4 \\
0 & 0 & 0 & 0 & 0 & \mathbf{i}
\end{bmatrix} + \begin{bmatrix}
G_1 & -G_1 & 0 & 0 & 0 & \mathbf{v}_1 \\
0 & -C_1 & G_1 + G_2 & -G_2 & 0 & \mathbf{v}_2 \\
0 & -C_2 & C_2 + C_4 & -G_4 & 1 & \mathbf{v}_3 \\
0 & 0 & -C_4 & G_3 + G_4 & 0 & \mathbf{v}_4 \\
0 & 0 & 0 & 1 & 0 & -1 & \mathbf{i}
\end{bmatrix} &= \begin{bmatrix}
0 \\
0 \\
0 \\
0 \\
0
\end{bmatrix}.
\end{align*}
$$

(3.13)
If (13) is viewed as $G_{\dot{x}} + Hx = 0$, we have

$$\det(\lambda G + H) = \lambda C(G_1 G_4 - G_2 G_3) - G_1 G_2 G_3. \quad (3.14)$$

Assuming that $R_i \neq 0$ for all $i$, $\det(\lambda G + H) \neq 0$ for small $\lambda$. To complete our analysis of (13), we need the following proposition.

Proposition 3.3.1 If $\lambda G + H$ is invertible for some $\lambda$, then the degree of $\det(\lambda G + H)$ is the same as the rank of $[(\lambda G + H)^{-1} G]D[(\lambda G + H)^{-1} H] = \text{rank}(G^D G)$.

Proof Let $(\delta G + H)^{-1} G = T \begin{bmatrix} C & 0 \\ 0 & N \end{bmatrix}^{-1}, C$ invertible, $N$ nilpotent, for some $\delta$ such that $\delta G + H$ is invertible. Then

$$\det(\lambda G + H) = \det((\delta G + H))\det((\delta G + H)^{-1}(\lambda G + H))$$

$$= \det(\delta G + H)\det\left(\lambda T \begin{bmatrix} C & 0 \\ 0 & N \end{bmatrix}^{-1} + I - \delta T \begin{bmatrix} C & 0 \\ 0 & N \end{bmatrix}^{-1}\right)$$

$$= \det(\delta G + H)\det\begin{bmatrix} \lambda C + I - \delta C & 0 \\ 0 & \lambda N + I - \delta N \end{bmatrix}$$

$$= \det(\delta G + H)\det(\lambda C + I - \delta C).$$

But $\det(\delta G + H)$ is constant and the degree of $\det(\lambda C + I - \delta C) = \text{rank}(C)$ since $C$ is invertible.

An immediate consequence of Proposition 1 is that the number of possible state variables for (13), equivalently the maximum possible number of independent initial conditions, is zero if $G_1 G_4 - G_2 G_3 = 0$ and one if $G_1 G_4 - G_2 G_3 \neq 0$.

These examples show that, in general, the questions of solvability, index, core-rank (state variable dimension) will depend on the particular
v-i characteristics of the circuit elements and not just on their general type.

4. NONLINEAR CIRCUITS

Most electrical devices are not really linear. However if they are only considered over a narrow operating range they may appear linear. Some components, such as transistors, tend to lead to circuit models which are often nonlinear. For our purposes a transistor may be represented as a three-terminal device. An example is the bipolar transistor:

![Bipolar Transistor Diagram](image)

For the bipolar transistor there is a nonlinear functional relationship between $i_c$ and the voltage drop across nodes $C$ and $E$, $v_{CE}$. For purposes of modeling and computer simulation and design it is desirable to replace the transistor by a circuit model. In doing so the transistor is viewed as a 2-port.

There are several ways to do this for the bipolar transistor. One is the common-base configuration.
Another is the common-emitter configuration

Transistors are of particular interest to us since they can be modeled by linear capacitors and nonlinear resistors. The bipolar transistor can be modeled by using linear capacitors and nonlinear current controlled resistors. The results and procedures we develop in subsequent chapters will be applicable to circuits made up of nonlinear current controlled resistors and linear capacitors.

There are other types of transistors. For example in a MOSFET (metal oxide semi-conductor field effect transistor), there is still a nonlinear functional relationship between $i_C$ and $v_{CR}$ but now it is determined by the voltage across notes B and E, $v_{BE}$. For a given $v_{BE}$, a typical curve for a MOSFET is

A circuit model for a MOSFET would utilize voltage controlled elements.

Circuit models involving linear capacitors, nonlinear resistors, and independent current or voltage sources need not admit a state variable representation. The next example is from [27] and will be discussed more carefully in Chapter 6.
Example 3.4.1 Consider

\[
\begin{align*}
\dot{x} + \begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & -3 \end{pmatrix} x + \begin{pmatrix} 0 \\ 0 \\ x_3 \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \\
\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} &= \begin{pmatrix} y \\ -v_R \\ i_R \end{pmatrix}
\end{align*}
\]

where \( v_R = h(i_R) = -3i_R + i_R^3 \). This circuit has the equation

\[
(4.5)
\]

Due to the fact that \( h \) is not one to one it is impossible to globally solve for \( i_R \) in terms of \( v_R \). As a consequence, (6) does not have a state variable representation. Newcomb [41], [68], [85], [86] refers to (6) as a semi-state representation.

Unicursal elements The preceding discussion shows that many circuit models are in the form

\[
\dot{\mathbf{x}} + \mathbf{B}(\mathbf{x}) = \mathbf{f}(t)
\]

(4.7)

where \( \mathbf{B} \) is nonlinear. These types of singular systems will be discussed in Chapter 6. There is another, more general class of circuits which lead to equations of the form

\[
\mathbf{A}(\mathbf{x})\dot{x} + \mathbf{B}(\mathbf{x}) = \mathbf{f}.
\]

(4.8)
The systems (8) will not be discussed in this volume other than to show how they arise.

Following [26] a unicursal curve is a subset \( C \) of \( \mathbb{R}^2 \) where
\[
C = \{(w(\tau), z(\tau)) \mid \alpha \leq \tau \leq \beta\}.
\]
Here \( w, z \) are continuous on \([\alpha, \beta]\) and of bounded variation on all finite intervals \([\alpha, \beta']\), \(\beta' < \beta\). A unicursal element is one which can be characterized by a unicursal curve. Thus a unicursal resistance would be of the form
\[
v(\tau), \quad i(\tau). \tag{4.9}
\]
That is, its \(v\)-\(i\) characteristic is any continuous curve which is rectifiable over bounded domains. Similarly unicursal inductances are parameterized by \((\phi\) is the flux)
\[
\phi = \phi(\tau), \quad i = i(\tau), \quad v = \frac{d\phi}{d\tau} \quad \frac{dt}{dt} \tag{4.10}
\]
and unicursal capacitances by
\[
q = q(\tau), \quad v = v(\tau), \quad i = \frac{dq}{d\tau} \quad \frac{dt}{dt} \tag{4.11}
\]

Example 3.4.2 This example is also from [26, p. 478]. The circuit is a single loop with no sources. The resistance has \(v\)-\(i\) characteristic a four-cusped hypocycloid with parametric representation
\[
v_R = \cos^3(\tau_1), \quad i_R = \sin^3(\tau_1).
\]
The \(\phi\)-\(i\) characteristic of the inductance \(L\) is the ellipse
\[
\phi_L = \cos(\tau_2), \quad i_L = 2 \sin(\tau_2).
\]
The \(q\)-\(v\) characteristic of the capacitance is the cycloid.
\[ q_C = r_3 - \sin r_3, \quad v_C = 1 - \cos(r_3). \]

The resulting equations are

\[
\begin{bmatrix}
\sin(r_2) & 0 & 0 \\
1 - \cos(r_3) & 0 & 0 \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
\dot{r}_2 \\
\dot{r}_3 \\
\dot{r}_1
\end{bmatrix}
= \begin{bmatrix}
\cos^3(r_1) - \cos(r_3) + 1 \\
2 \sin(r_2) \\
\sin^3(r_1) - 2 \sin(r_2)
\end{bmatrix}
\]

which is in the form (8). Here \( \dot{r}_i = \frac{d\tau_i}{dt} \).
4 Numerical solution of autonomous linear systems

1. INTRODUCTION

The emphasis in [8] and this book, up to this point, has been predominantly on analytic closed form solutions. This chapter begins a shift in emphasis to the development of numerical procedures. This approach will dominate the later chapters on nonlinear systems where it will be seen that even "theoretical" questions such as existence of solutions and determination of consistent initial conditions are closely intertwined with the behavior of various difference schemes.

This chapter will be concerned with the numerical solution of

\[ \dot{\mathbf{x}} + \mathbf{Bx} = \mathbf{f} \]  

(1.1)

under our standard assumptions that \( \mathbf{A} \) is singular and \( \lambda \mathbf{A} + \mathbf{B} \) is invertible for some scalar \( \lambda \) and \( \mathbf{f} \) is \( \text{Ind}((\lambda \mathbf{A} + \mathbf{B})^{-1} \mathbf{A}) = \text{Ind}(\hat{\mathbf{A}}) \) times continuously differentiable.

One approach would be to try to simultaneously upper triangularize \( \mathbf{A}, \mathbf{B} \) and then solve (1) by back substitution. A procedure for doing this is given in [113], [114]. See also [84], [111], [112] who discuss the QZ algorithm and the generalized eigenvalue problem

\[ \mathbf{Ax} = \lambda \mathbf{Bx} \]  

(1.2)

which involves similar operations. An earlier version is [40]. This volume will consider a different approach.
From Chapter 0 we know that the solution of (1) will involve derivatives of \( f \) if \( \text{Ind}(\hat{A}) > 1 \). This has two implications.

If the differential operator \( L \) is defined by \( L(x) = \dot{A}x + Bx \), its "inverse" will also involve differentiations. This makes the use of iterative arguments involving the contraction mapping theorem very difficult to develop. In fact, we have been unable to do so in a space containing enough interesting functions. Accordingly, this chapter will concern itself entirely with difference approximations to (1).

Secondly, numerical differentiation is known to be difficult. This raises the possibility that error would quickly overwhelm the results. However, as we shall see shortly, this error tends to be transient or local in nature and it is possible to get reasonably good results.

The basic ideas in this chapter were developed independently by us and the group then at Boeing [103]. The reading of [103] and Petzold's paper [95] have influenced the writing of this chapter and several of their ideas, particularly those on the higher order methods, have been incorporated. That variable time step methods may not work well for high index systems was first noted by Petzold [95].

All the numerical examples given in this book were run in APL on an IBM 360 using the standard APL double precision, \( \epsilon \approx 10^{-17} \). Linear equations of the form \( Ax = b \) were solved using the APL "domino" operator, \( x = \div \times A \).

2. FIRST ORDER METHODS

The notation introduced in this chapter will be kept throughout the remainder of this book.

Let \( h \) be a "small" time step, \( t_0 \) some initial time, and \( t^\ell = t_0 + \ell h \) for integers \( \ell \). For a function \( g(t) \) define \( \bar{g}^\ell = g(t^\ell) \). A function of \( h, p(h) \)
is $O(h^r)$ if $\lim_{h \to 0^+} h^{-r} p(h)$ is bounded. If $\lim_{h \to 0^+} h^{-r} p(h) = 0$ we will sometimes write $p(h) = o(h^r)$. Assume that $x(t)$ is at least twice continuously differentiable. (This would happen if $f$ in (1.1) were at least $k + 1$ times differentiable where $k = \text{Ind}(\hat{A})$.) From Taylor's Theorem

$$\bar{x}_{k+1} = \bar{x}_k + \dot{x}(t_k)h + O(h^2). \quad (2.1)$$

Thus

$$\dot{x}(t_k) = \frac{\bar{x}_{k+1} - \bar{x}_k}{h} + O(h). \quad (2.2)$$

Substituting (2.1) into (1.1) gives

$$A\bar{x}_{k+1} + (hB - A)\bar{x}_k = hf_k + O(h^2).$$

Thus suggests making the following definition. The first order forward difference approximation to (1.1) is

$$A\bar{x}_{k+1} + (hB - A)\bar{x}_k = hf_k. \quad (2.3)$$

The difference scheme (3) is also referred to as a forward Euler's.

From Taylor's theorem, instead of (1) we can also get

$$\bar{x}_{k-1} = \bar{x}_k + \dot{x}(t_k)(-h) + O(h^2) \quad (2.4)$$

so that

$$\dot{x}(t_k) = \frac{\bar{x}_k - \bar{x}_{k-1}}{h} + O(h). \quad (2.5)$$

Using (5) in (1.1) gives

$$(A + hB)\bar{x}_k - A\bar{x}_{k-1} = hf_k + O(h^2). \quad (2.6)$$
The first order backwards difference approximation (backwards Euler's) to (1.1) is then

\[(A + hB)x_{k+1} - Ax_k = h\tilde{f}_{k+1}.\]  (2.7)

Do solutions of (3), (7), converge as \(h \to 0^+\) to solutions of (1.1) for consistent initial conditions? If \(x(t_0)\) is not consistent for (1.1), is an approximation of impulsive behavior present?

In the next section we shall give proofs and error estimates for several different backwards difference schemes. But first we shall briefly consider forward differences.

Forward differences
Forward differences are known to be generally less desirable for nonsingular problems than backwards differences, particularly for solutions involving rapidly decaying terms such as might occur in a singular perturbations problem. However, for completeness we shall briefly consider forward differences.

Notice that the forward difference equation is still a singular system.

We shall assume that \(h\) is chosen so that \(hB - A\) is invertible. Let

\[\tilde{A} = (hB - A)^{-1}A, \quad \tilde{f}_k = (hB - A)^{-1}\tilde{f}_k.\]

Then (3) is equivalent to

\[\tilde{A}x_{k+1} + x_k = h\tilde{f}_k, \quad k = 0, \ldots.\]  (2.8)

From Theorem 0.1.4, (8) is consistent if and only if

\[(I - \tilde{A}^DA)x_0 = (I - \tilde{A}DA)^{k-1} \sum_{i=0}^{k-1} (-\tilde{A})^ih\tilde{f}_i,\]  (2.9)

where \(k = \text{Ind}(A)\), and in this case for \(k \geq 1\).
\[ x_k = (-A)^{\ell} A A x_0 + \sum_{i=0}^{\ell-1} (-A)^{\ell-i-1} h_f i \]

\[ + (I - AA) \sum_{i=0}^{k-1} (-A)^i h_f i. \]

Note that \(^{-}D\) is independent of \(h\).

If \(f \equiv 0\), then (10) says \(x_0\) is consistent for the forward difference equation if and only if \(x_0\) is consistent for the differential equation.

The next proposition shows that solutions of the homogeneous discrete equation (3) converge to solutions of the homogeneous differential equation.

**Proposition 4.2.1** Let \(\hat{A} = (hB - A)^{-1}A\), \(\hat{\lambda} = (\lambda A + B)^{-1}A\), \(\hat{B} = (\lambda A + B)^{-1}B\).

For a fixed \(t\), let \(h = t/\ell\). Then

\[ \lim_{\ell \to \infty} (-A)^{\ell} A A = e^{-A B t} A A. \]

**Proof** Now let \(C\) be a contour around the nonzero eigenvalues of \(A\), but not including the zero eigenvalues. Note that in general \((TX)^D = T(XT)^D\).

Then

\[ \lim_{\ell \to \infty} (-A)^{\ell} A A \]

\[ = \lim_{\ell \to \infty} (-[(hB - A)^{-1}A]^D)^{\ell}[(hB - A)^{-1}A][(hB - A)^{-1}A] A D \]

\[ = \lim_{\ell \to \infty} (-[(h\hat{B} - \hat{A})^{-1}\hat{A}]^D)^{\ell}[(h\hat{B} - \hat{A})^{-1}\hat{A}][(h\hat{B} - \hat{A})^{-1}\hat{A}] A D \]

\[ = \lim_{\ell \to \infty} (-1)^{\ell} (h\hat{B} - \hat{A})^{\ell} [\hat{A} D] A A \]

since \(\hat{A} \hat{B} = \hat{B} \hat{A}\). But \(\hat{\lambda} A + \hat{B} = I\) so that (12) is
\[
\lim_{\ell \to \infty} \left( \hat{A} - h(I - \lambda \hat{A}) \right)^{\ell} A^{\hat{D}} A = \lim_{\ell \to \infty} \frac{1}{2\pi i} \int_{\mathcal{C}} (s - h + h\lambda s)^{-\ell} s^{-\ell}(\lambda - \hat{A})^{-1} d\lambda.
\]

Now \( \lim_{\ell \to \infty} (s - h + h\lambda s)^{-\ell} s^{-\ell} = \lim_{\ell \to \infty} (1 - \frac{t}{\ell s} + \frac{t}{\ell} \lambda)^{-\ell} = e^{-t/s + t\lambda} \) for \( s \in \mathcal{C} \).

Thus (2.12) is

\[
\frac{1}{2\pi i} \int_{\mathcal{C}} e^{-t/s + t\lambda} (\lambda - \hat{A})^{-1} d\lambda = e^{t(-A^{\hat{D}} + \lambda I) A A^{\hat{D}}}
\]

\[
= e^{t(-A^{\hat{D}} + \lambda A^{\hat{D}}) A A^{\hat{D}}} = e^{t A^{\hat{D}}(-I + \lambda A) A A^{\hat{D}}} = e^{t A^{\hat{D}} B A A^{\hat{D}}}
\]
as desired. \( \Box \)

Note also that for \( 0 \leq t \leq r_o \), some \( r_o > 0 \), the above limit is uniform.

Proposition 1 was proven separately since the proof is interesting and useful in evaluating other limits that (11). However, to analyze (3) in general, it is easier to proceed as follows. Multiplying (3) by \((\lambda A + B)^{-1}\) for some \( \lambda \), we may assume \( A, B \) commute. Next a similarity may be performed so that \( T A T^{-1} = \begin{bmatrix} C & 0 \\ 0 & N \end{bmatrix} \), \( T B T^{-1} = \begin{bmatrix} B_1 & 0 \\ 0 & B_2 \end{bmatrix} \) where \( C \) is invertible, \( N \) is nilpotent, \( B_2 \) is invertible and \( C B_1 = B_1 C \), \( N B_2 = B_2 N \). Thus it suffices to consider the two special cases when \( A \) is invertible and \( A \) is nilpotent. If \( A \) is invertible, it is known that both the forward and backward difference approximations converge as \( h \to 0^+ \) to the solution of the differential equation and standard error estimates may be used to give the approximation is uniformly \( O(h) \) on an interval \([t_o, t_1]\). It is not hard to show that the discrete solution converges to a continuous solution and that the consistent initial condition \( x_0(h) \) of the discrete problem, while it may not be a consistent initial condition for the continuous problem, does converge to a consistent initial condition of the continuous problem (1.1). Since we shall
be concerned solely with backwards differences we leave this as an exercise for the interested reader.

**Backward differences**

For nonsingular autonomous problems the backward difference procedure is often preferred since it tends to be more stable in computing the exponential types of solutions involved.

In this section we shall see that for singular autonomous systems there are other advantages of backward differences. For one, they can be used to estimate impulsive solutions.

Secondly, note that for small enough time step $A + hB$ is invertible and (7) is a nonsingular system. In particular, unlike with the forward difference scheme, all initial conditions $x_0$ are consistent for the difference equation (7).

Again for purposes of analysis, it suffices to assume that $B = I$, and $A = N$ is nilpotent.

Then (7) is

$$x_{\ell+1}^i - (N + hI)^{-1}Nx_\ell^i = h(N + hI)^{-1}x_{\ell+1}^i.$$  \hspace{1cm} (2.13)

The solution of (13) is

$$x_\ell^i = [(N + hI)^{-1}N]^i x_0^i + \sum_{i=0}^{\ell-1} [(N + hI)^{-1}N]^{i+1}h^{\ell-i-1}.$$ \hspace{1cm} (2.14)

The first thing to notice is that if $\ell \geq k = \text{Ind}(N)$, then

$$x_\ell^i = \sum_{i=0}^{k-2} [(N + hI)^{-1}N]^i h^{\ell-i-1}.$$ \hspace{1cm} (2.15)

Thus $x_\ell^i$, for $\ell \geq k$, is independent of the initial condition provided round off error is ignored. Secondly, $h(N + hI)^{-1} = O(h^{1-k})$. Thus it is not at
all obvious that the solutions of (13) converge as $h \rightarrow 0^+$. Shortly we shall show that, in fact, (15) converges to the solution of

$$\mathbf{N} \dot{x} + x = f$$

(2.16)

uniformly on compact subsets of $(0, \infty)$. Note that (15) then says that if impulsive behavior is present, it lasts for at most $\text{Ind}(\mathbb{N})$ times steps.

Equation (15) also implies that, in principle, the error introduced at step $l$ in the singular subsystem is transient. However, in evaluating (15) we must consider terms like $(f_k - f_{k-1}) h^{-1}$ since the solution of the differential equation involves derivatives of $f$. However, if these terms are being numerically implemented, there is round off error since $f_k, f_{k-1}$ will not be entered exactly. Thus one really has

$$\frac{f_k - f_{k-1}}{h} = O(h) + \frac{\varepsilon}{h}$$

(2.17)

where $\varepsilon$ depends on the particular implementation being used. In double precision on many machines, $\varepsilon \approx 10^{-15}$. In general, there are $k - 1$ differentiations in the solution of (1.1). Thus if round off error is considered, the error estimate for a backwards Euler's is not $O(h)$ but rather

$$O(h) + \frac{\varepsilon}{h^{k-1}}.$$  

(2.18)

On the other hand, as the system decomposition shows, the $e h^{1-k}$ error only occurs in the singular subsystem. But error in these terms is transient, that is, it is a local error that is not propagated for more than $k - 1$ steps. Thus, in double precision, with moderate indices (say $k = 3$ or 4), it is possible to achieve accuracy to several decimal places.

The term $e h^{1-k}$ in (18) is only amenable to control to the extent that it is related to the chosen machine precision. If the $O(h)$ could be replaced
by \( O(h^r) \), \( r > 1 \), so that one had say \( O(h^3) + o h^{-k+1} \), then greater accuracy could, in principle, possibly be attained with larger \( h \). This is the purpose of the higher order methods discussed in the next section.

In the later sections we shall need to modify (7) slightly to

\[
(A + hB)x_{k+1} - Ax = \frac{h f_k}{2}.
\]  

(Note: we have used \( f_k \) instead of \( f_{k+1} \).)

The preceding analysis also works on (19). To see why, let \( g_h(t) = f(t - h) \). Then for a given \( h \), the solution \( x_{k} \) is within \( O(h) \) of the solution of

\[
Ax + Bx = g_h.
\]  

On the other hand, \( |(g_h)^{(1)}(t) - f^{(1)}(t)| = O(h) \) also for \( 0 \leq i \leq k \), \( 0 \leq t \leq T \), so that the solution of (20) is within \( O(h) \) of the solution of (1.1).

3. HIGHER ORDER METHODS

The preceding section used the fact that

\[
\dot{x}(t_0 + 2h) = \left[ x_k - x_{k-1} \right]^{-1} + O(h).
\]  

(3.1)

With higher order schemes we seek a better approximation to \( \dot{x}(t_0 + 2h) \) in terms of \( \left[ x_j : j \leq k \right] \). One way to develop such a formula is as follows. From Taylor’s theorem, if \( x(t) \) is a differentiable function,

\[
-x_{k-1} = x_k \cdot \dot{x}(t_k) (h) + \ddot{x}(t_k) \frac{h}{2} - \dddot{x}(t_k) \frac{h^2}{6} + O(h^4).
\]  

(3.2)

Also,
\[ \ddot{x}_{k-2} = \ddot{x}_k + \dddot{x}(t_k)(-2h) + \dddot{x}(t_k) \frac{(-2h)^2}{2} + \dddot{x}(t_k) \left( -\frac{2h}{6} \right)^3 + o(h^4). \] (3.3)

Then (2), (3) give
\[ \dot{x}(t_k) = (\ddot{x}_k - \ddot{x}_{k-1})h^{-1} + \dddot{x}(t_k) \frac{h}{2} - \dddot{x}(t_k) \frac{h^2}{6} + o(h^3), \] (3.4)
\[ \dot{x}(t_k) = (\ddot{x}_k - \ddot{x}_{k-2})(2h)^{-1} + \dddot{x}(t_k)h - \dddot{x}(t_k) \frac{2h}{3} + o(h^3). \] (3.5)

Let \( \alpha = 2, \beta = -1 \). Note that \( \alpha + \beta = 1 \) and \( \alpha h/2 + \beta h = 0 \). Thus if \( \alpha \) times (4) is added to \( \beta \) times (5) the result is
\[ \dot{x}(t_k) = \left[ \frac{3}{2} \ddot{x}_k - \ddot{x}_{k-1} + \frac{1}{2} \ddot{x}_{k-2} \right]h^{-1} + o(h^2). \] (3.6)

Similar formulas may be derived involving an error of \( o(h^3), o(h^4), \) etc.

Note that if we defined the operator \( D \) acting on sequences by
\[ D\{x_k\} = \left[ \frac{3}{2} \ddot{x}_k - \ddot{x}_{k-1} + \frac{1}{2} \ddot{x}_{k-2} \right]h^{-1}, \] (3.7)
then (6) could be written as
\[ \dot{x}(t_k) = Dx_k + o(h^2). \] (3.8)

Let \( D\{x_k\} = \left[ \sum_{i=0}^{s} \alpha_i x_{k-i} \right]h^{-1} \). Then \( D \) is a difference operator of order \( r \) if
\[ \dot{x}(t_k) = Dx_k + o(h^r) \] (3.9)
for \( r + 1 \) times continuously differentiable functions \( x(t) \). The constant in the \( o(h^r) \) is to be determined by the maximum of the \( \{ |x^{(i)}| : 0 \leq i \leq r \} \) on an interval including \( t_k \). If \( C \) is a matrix, \( C \) can be viewed as an operator on sequences where \( C\{x_k\} = \{Cx_k\} \). Note that \( DC = CD \) for any matrix...
C and difference operator D. We are now ready to state and prove the following result.

**Theorem 4.3.1** Suppose that \( D \) is a difference operator of order \( r \). Let \( f \) be \( k + r \) times continuously differentiable on \([t_o, t_1]\). Let \( x(t) \) be a solution of (1.1) on the interval \([t_o, t_1]\). Let \( \tilde{x}_{k} \) be the solution of

\[
AD\tilde{x}_{k} + B\tilde{x}_{k} = \tilde{f}_{k},
\]

(3.10)

and \( \tilde{x}_{i} = x(t + hi) \) for \( 0 \leq i \leq \frac{(t_1 - t_o)}{h} \). Then \( ||x_{k} - \tilde{x}_{k}|| = O(h^r) \) on \([t_o + kh, t_1]\), \( k = \text{Ind}(\hat{A}) \).

**Proof** By performing the usual decomposition (1.1) involving the constant matrix \((A + B)^{-1}\) and a similarity we may write (1.1) as

\[
\dot{x} + B_{1}x = \hat{g},
\]

(3.11)

\[
N\dot{x} + x = \hat{f},
\]

(3.12)

where \( N \) is nilpotent of index \( k \). For (11) it is known that the local error is \( O(h^{r+1}) \) and the global error is \( O(h^r) \). Consider then (12). The discretized version for \( l > k \) is

\[
ND\tilde{x}_{k} + \tilde{x}_{k} = \hat{f}
\]

(3.13)

or

\[
(I + ND)\tilde{x}_{k} = \hat{f}_{k}.
\]

(3.14)

Now \( ND = DN \) so that \( ND \) is a nilpotent operator of index \( k \) on sequences. Thus (14) can be solved as
\[ x_{\hat{n}} = (I + ND)^{-1}\hat{f} = \sum_{i=0}^{k-1} (-N)^i D^i\hat{f}_{\hat{n}}. \]  

(3.15)

But

\[ D f_{\hat{n}} = \hat{f}_{\hat{n}} + O(h^r). \]  

(3.16)

Thus

\[ D^2 f_{\hat{n}} = D(\hat{f}_{\hat{n}} + O(h^r)) = \hat{f}_{\hat{n}} + DO(h^r) \]

\[ = \hat{f}_{\hat{n}} + O(h^r) + O(h^r) = \hat{f}_{\hat{n}} + O(h^r). \]  

(3.17)

That \( DO(h^r) = O(h^r) \) follows from considering the Taylor series for \( \hat{f}_{\hat{n}} \) in which case the \( O(h^r) \) term is of the form \( h^r m(t, h) \) and is \( s \) times differentiable with respect to \( t \). Thus (15) becomes

\[ x_{\hat{n}} = \sum_{i=0}^{k-1} (-N)^i \hat{f}_{\hat{n}}^{(i)} + O(h^r). \]  

(3.18)

But \( \sum_{i=0}^{k-1} (-N)^i \hat{f}_{\hat{n}}^{(i)} \) is the solution of (12). \( \square \)

As noted before, for the subsystem (12), the local and global error are both \( O(h^r) \).

There are certain problems attendant to using higher order difference schemes. Many of them require the generation of additional starting values. Also some of them exhibit instability. That is, round off error can grow exponentially with the number of steps instead of linearly.

4. **STARTING VALUES AND VARIABLE TIME STEP METHODS**

The proof of Theorem 3.1 used (3.15). But this formula is only valid for \( x_{\hat{n}} \) with \( \Delta \geq k \). In this section we wish to discuss what is happening during
the first $k$ steps and the implications this has for variable time step methods.

As noted earlier, to understand the numerical behavior of (1.1) it suffices, for our purposes, to consider

$$\dot{N} + x = f$$

(4.1)

where $N$ is a simple Jordan nilpotent block. To illustrate what happens suppose that

$$N = \begin{bmatrix}
0 & -1 & 0 & 0 \\
0 & 0 & -1 & 0 \\
0 & 0 & 0 & -1 \\
0 & 0 & 0 & 0
\end{bmatrix}, \quad f = \begin{bmatrix}
0 \\
0 \\
g
\end{bmatrix}$$

(4.2)

and $x_0$ is a consistent initial condition so that

$$x_0^T = [\ddot{g}(t_0), \dddot{g}(t_0), \dddot{g}(t_0), g(t_0)]^T.$$  

Then applying a backwards Euler's gives

$$x_1 = (N + hI)^{-1}[N x_0 + hf_1]$$

$$= \frac{1}{h} \begin{bmatrix}
1 & h^{-1} & h^{-2} & h^{-3} \\
0 & 1 & h^{-1} & h^{-2} \\
0 & 0 & 1 & h^{-1} \\
0 & 0 & 0 & 1
\end{bmatrix} \begin{bmatrix}
\dddot{g}(t_0) \\
\dddot{g}(t_0) \\
\dddot{g}(t_0) \\
hg(t_1)
\end{bmatrix}$$

$$= \begin{bmatrix}
(-\dddot{g}(t_0) - h^{-1} \dddot{g}(t_0) - h^{-2} \dddot{g}(t_0) + h^{-2} \dddot{g}(t_1))h^{-1} \\
(-\dddot{g}(t_0) - h^{-1} \dddot{g}(t_0) + h^{-1} \dddot{g}(t_1))h^{-1} \\
(-g(t_0) + g(t_1))h^{-1} \\
g(t_1)
\end{bmatrix}.$$  

(4.3)
Thus
\[ x_{14} = g(t_1) \]  \hspace{1cm} (4.4)

which is exactly correct. Also
\[ x_{13} = (-g(t_0) + g(t_1))h^{-1} = (-g(t_1) + h\dot{g}(t_1) + o(h^2) + g(t_1))h^{-1} \]
\[ = \ddot{g}(t_1) + o(h) \]

so that \( x_{13} \) differs from the exact solution by an \( O(h) \) term as to be expected.

However,
\[ x_{12} = (-\dot{g}(t_0) - h^{-1}g(t_0) + h^{-1}g(t_1))h^{-1} \]
\[ = (-[\dot{g}(t_1) - hg(t_1) + o(h^2)] - h^{-1}[g(t_1) - \dot{g}(t_1) + \frac{h^2}{2} \ddot{g}(t_1)] \]
\[ + o(h^3)] + h^{-1}g(t_1))h^{-1} = (\ddot{g}(t_1) - \frac{h}{2} \dddot{g}(t_1) + o(h^2))h^{-1} \]
\[ = \frac{\ddot{g}(t_1)}{2} + o(h). \]  \hspace{1cm} (4.5)

But the exact solution is \( \dddot{g}(t_1) \). Thus the second component of \( x_1 \) will always be off by approximately \( \dddot{g}(t_1)/2 \) regardless of the step size \( h \).

The situation for \( x_{11} \) is even worse.
\[ x_{11} = h^{-3}(-h^2g(t_0) - \dddot{g}(t_0) - g(t_0) + g(t_1)) \]
\[ = h^{-3}(-h^2[\dddot{g}(t_1) + o(h)] - h[\dddot{g}(t_1) - \dddot{g}(t_1) + o(h^2)] \]
\[ - [g(t_1) - h\dddot{g}(t_1) + \frac{h^2}{2} \dddot{g}(t_1) + o(h^3)] + g(t_1)] \]
\[ = h^{-3}[-\frac{h^2}{2} \dddot{g}(t_1) + o(h^3)] = -\frac{\dddot{g}(t_1)}{2h} + o(1). \]  \hspace{1cm} (4.6)
while the exact value of $x_{11}$ is $\tilde{g}(t_1)$.

This is somewhat surprising. According to Theorem 3.1, solutions of the discrete equation approximate uniformly on $[t_0 + kh, t_1]$. Yet (5), (6) say that if $x_1$ is exactly correct, then $x_{i+1}$ will differ from the exact solution by an amount independent of $h$ if the index is three and by $O(h^{-1})$ if the index is four. The following theorem explains what is happening.

**Theorem 4.4.1** Given any consistent initial value $x_0$ for the differential equation (1.1) and an $r$th order difference scheme, the solution of the difference equation (3.10) will approximate the functional solution uniformly $O(h^r)$ on $[t_0 + \varepsilon, t_1]$ for any $\varepsilon < kh$. For each step size $h$, there is an initial value $x_0(x_0, h, r)$ (values if $r > 1$) such that the solution of the difference scheme (3.10) starting at $x_0$ approximates the solution of (1.1) uniformly $O(h^r)$ on $[t_0, t_1]$.

Furthermore $|x_h - x_0| = O(h^r)$.

**Proof** Given $x_0, h, r$, let $x(t)$ be the solution of (1.1) such that $x(t_0) = x_0$. Start the difference scheme (3.10) at $x(t_0) = (k + r)h$ to get an approximation $z_k$. Let $x_h = z_k + z_{k+1} + \cdots + z_{k+r-1}$ if $r > 1$. The rest of Theorem 1 now follows from Theorem 3.1.

One consequence of Theorem 1 is that if $x_0$ is not close to being a consistent initial condition for the differential equation, then the first $k$ values of the difference approximation may be expected to have larger fluctuations in value than if $x_0$ were a consistent initial condition. Thus the solution of the difference equation for inconsistent initial conditions not close to consistent initial conditions should exhibit an approximation of the distributional solutions. This is illustrated in the next section.
Variable step methods One way that increased accuracy is often obtained in numerically solving ordinary differential equations is by varying the step size. This does not work very well with singular systems if the index is 3 or higher.

We shall take a canonical example of index 3 to show that happens.

Consider again $N\dot{x} + x = f$. Assume $N^3 = 0$. Let $h_k$ be the step size of the $k$-th step. Then using a backwards Eulers we get that

$$x_k = (N + h_k)^{-1} [N x_{k-1} + h_k f_{k-1}]$$

$$= (N + h_k)^{-1} h_k f_{k-1} + (N + h_k)^{-1} N x_{k-1}$$

$$= (N + h_k)^{-1} h_k f_{k-1} + (N + h_k)^{-1} N [(N + h_{k-1})^{-1} g_{k-1} f_{k-1}$$

$$+ (N + h_{k-1})^{-1} N x_{k-2}]$$

$$= (N + h_k)^{-1} h_k f_{k-1} + (N + h_k)^{-1} (N + h_{k-1})^{-1} h_{k-1} N f_{k-1}$$

$$+ (N + h_k)^{-1} N (N + h_{k-1})^{-1} h_{k-1} N^2 f_{k-2}. \quad (4.7)$$

Let

$$N = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix}, \quad f = \begin{bmatrix} 0 \\ 0 \\ g \end{bmatrix}.$$ 

Then

$$x_{k+1} = \frac{g_k}{h_k} - \frac{1}{h_k} \left\{ \frac{g_{k-1}}{h_{k-1}} + \frac{g_{k-1}}{h_k} \right\} + \frac{g_{k-2}}{h_k h_{k-1}}$$

$$= \frac{1}{h_k} \left\{ \frac{g_k - g_{k-1}}{h_k} \right\} + \frac{1}{h_k} \left\{ \frac{g_{k-2} - g_{k-1}}{h_{k-1}} \right\}.$$
\begin{align*}
&= \frac{1}{h_\ell} \left\{ g_{t-1} + \frac{\ddot{g}(t_{\ell-1})}{2} (h_{\ell-1})^2 + O(h_{\ell-1}^3) + g_{t-1} \right\} \\
+ \frac{1}{h_\ell^2} \left\{ g_{t-1} + \frac{\ddot{g}(t_{\ell-1})}{2} (h_{\ell-1})^2 + O(h_{\ell-1}^3) + g_{t-1} \right\} \\
&= \frac{\ddot{g}(t_{\ell-1})}{2} + O(h_\ell) + \frac{\ddot{g}(t_{\ell-1})}{h_\ell} h_{\ell-1} \\
&= \frac{1}{2} \left[ 1 + \frac{h_{\ell-1}}{h_\ell} \right] \ddot{g}(t_{\ell-1}) + O(h_\ell) + \frac{O(h_{\ell-1}^2)}{h_\ell} \\
\end{align*}

(4.9)

The exact solution is \( x_{\ell+1} = \ddot{g}(t_{\ell}) \). The error is thus

\begin{align*}
e(\ell) &= \ddot{g}(t_{\ell}) - \frac{1}{2} \left[ 1 + \frac{h_{\ell-1}}{h_\ell} \right] \ddot{g}(t_{\ell-1}) + O(h_\ell) + \frac{O(h_{\ell-1}^2)}{h_\ell} \\
&= \ddot{g}(t_{\ell}) - \frac{1}{2} \left[ 1 + \frac{h_{\ell-1}}{h_\ell} \right] \left\{ \ddot{g}(t_{\ell}) - h_\ell \ddot{g}(t_{\ell}) + O(h_\ell^2) \right\} \\
&\quad + O(h_\ell) + \frac{O(h_{\ell-1}^2)}{h_\ell} \\
&= \frac{1}{2} \left[ 1 - \frac{h_{\ell-1}}{h_\ell} \right] \ddot{g}(t_{\ell}) + \frac{h_\ell + h_{\ell-1}}{2} \ddot{g}(t_{\ell}) + O(h_\ell) + \frac{O(h_{\ell-1}^2)}{h_\ell} \\
\end{align*}

(4.10)

Note that if the step size is constant, \( h_\ell = h \), then (10) reduces to

\( e(\ell) = 0(h) \) as given in Theorem 3.1. Suppose, however, that \( h_\ell \) varies with \( \ell \). Then \( e(\ell) \) may no longer be small. To illustrate this suppose that we pick \( h_\ell \) in an interval \([m\sigma, M\sigma]\). Then (10) becomes

\begin{align*}
e(\ell) &= \left[ 1 - \frac{h_{\ell-1}}{h_\ell} \right] g(t_{\ell}) + O(\sigma) \\
\end{align*}

(4.11)

Now fix \( t_{\ell} \) and let \( \sigma \to 0 \). Assuming \( \ddot{g}(t_{\ell}) \neq 0 \), \( e(\ell) \to 0 \) as \( \sigma \to 0 \) if and only if
Thus if we want $e(\ell)$ uniformly small on $[t_o + \varepsilon_o, t_1]$, $\varepsilon_o > 0$ then as $e \rightarrow 0$ the step size function $h$ must be "asymptotically constant," that is, have the behavior (12). In fact, (13) says that we get our best error estimate by taking $h_\ell = h_{\ell - 1}$ for all $\ell$ to begin with.

The preceding discussion shows that ordinary variable step methods are not to be used on linear systems if the index is greater than two. They can be implemented, with some care, if the index is less than three. We shall not do so since we are interested in the general index case. The interested reader is referred to [95].

In closing this section a comment on the $O(\cdot)$ notation is in order. To say $g(h) = O(h^r)$ means $g(h)h^{-r} \leq M_r$ for $h$ close to zero. Whether an $O(h^3)$ estimate is practically better than $O(h)$ depends on the sizes of $M_3$ and $M_1$. We do not intend to explicitly compute the constants $M_r$. By use of standard arguments the interested reader may derive the usual types of estimates. Suffice it to say that if the index is $k$, the estimates involve $A, B$ and the first $k + r$ derivatives of $f$.

5. NUMERICAL EXAMPLES

In this section we shall give several numerical examples that illustrate the preceding sections. The examples are of the form $Nx + x = f$, where $N$ is $5 \times 5$ and $f = \begin{bmatrix} t^3, t + t^4, t^2 + t^4, t + t^3, t^4 \end{bmatrix}^T$.

Since this system is already in canonical form, the errors that occur are intrinsic to the algorithms and not due to the specific example.

In our first example,
\[ N = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad x(0) = [1 \ 2 \ 3 \ 4 \ 5]^T. \]  

Using a backwards Euler's, the error in the first component of \( x, x_1 \), and the exact value of \( x_1 \) are given by

<table>
<thead>
<tr>
<th>TIME</th>
<th>ERROR WITH STEP = .05</th>
<th>ERROR WITH STEP = .01</th>
<th>ERROR WITH STEP = .005</th>
<th>VALUE OF EXACT SOLUTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>18.000</td>
<td>18.000</td>
<td>18.000</td>
<td>19.000</td>
</tr>
<tr>
<td>0.5</td>
<td>.495</td>
<td>.104</td>
<td>.052</td>
<td>21.625</td>
</tr>
<tr>
<td>1.0</td>
<td>.875</td>
<td>.179</td>
<td>.090</td>
<td>28.000</td>
</tr>
<tr>
<td>1.5</td>
<td>1.105</td>
<td>.224</td>
<td>.122</td>
<td>35.875</td>
</tr>
<tr>
<td>2.0</td>
<td>1.185</td>
<td>.239</td>
<td>.120</td>
<td>43.000</td>
</tr>
</tbody>
</table>

Table 4.5.1

Since a fourth order derivative is being taken, the low accuracy at \( h = .005 \) is not surprising.

For our next example we shall again use (1). This time, the first step is a backward Euler's. After that the second order method (3.7) is used. Again the value of \( x_1 \) is given.

<table>
<thead>
<tr>
<th>TIME</th>
<th>ERROR WITH STEP = .05</th>
<th>ERROR WITH STEP = .01</th>
<th>ERROR WITH STEP = .005</th>
<th>VALUE OF EXACT SOLUTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>18.0000</td>
<td>18.00000</td>
<td>18.00000</td>
<td>19.000</td>
</tr>
<tr>
<td>0.5</td>
<td>.0308</td>
<td>.001206</td>
<td>.0003005</td>
<td>21.625</td>
</tr>
<tr>
<td>1.0</td>
<td>.0208</td>
<td>.000806</td>
<td>.0002000</td>
<td>28.000</td>
</tr>
<tr>
<td>1.5</td>
<td>.0108</td>
<td>.000406</td>
<td>.0000053</td>
<td>35.875</td>
</tr>
<tr>
<td>2.0</td>
<td>.0001</td>
<td>.000005</td>
<td>.0001282</td>
<td>43.000</td>
</tr>
</tbody>
</table>

Table 4.5.2

108
Note the greater accuracy which is to be expected. The error that results in using a first order method for the first step is transient in nature for the system $\dot{x} + x = f$ and thus we still get the same increase in accuracy as if not only $x_0$ but $x_1$ where given.

For the initial condition $x(0) = [1 2 3 4 5]^T$ used in computing Table 1, the solution of $\dot{x} + x = f$ exhibits an impulse (distributional solution) at $t = 0$, and then gives the exact values listed. Since $x(0)$ is not close to the consistent initial condition, this impulse appears in the first $\text{Ind}(N)$ steps of the approximations. Thus, the discrete procedures will not only estimate the functional solutions but also any impulsive behavior. We give one example. Return to the situation of Table 1 where $\text{Ind}(N) = 5$. The solution $x_j$ involves $\delta^{(1)}$, for $0 \leq i \leq 5 - j$. The next table shows how the impulses are numerically estimated with a step size of .05.

<table>
<thead>
<tr>
<th>TIME</th>
<th>$x_1$ VALUE</th>
<th>$x_2$ VALUE</th>
<th>$x_3$ VALUE</th>
<th>$x_4$ VALUE</th>
<th>$x_5$ VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>1.</td>
<td>2.</td>
<td>3.</td>
<td>-4.</td>
<td>5.</td>
</tr>
<tr>
<td>0.05</td>
<td>-$769,560$</td>
<td>38,480.</td>
<td>-$1921.</td>
<td>100.05</td>
<td>$10^{-6}$</td>
</tr>
<tr>
<td>0.10</td>
<td>2,337,608.</td>
<td>-$78,400.$</td>
<td>1999.</td>
<td>.099</td>
<td>.0001</td>
</tr>
<tr>
<td>0.15</td>
<td>-$2,367,982.</td>
<td>39,999.</td>
<td>-.899</td>
<td>.145</td>
<td>.0005</td>
</tr>
<tr>
<td>0.20</td>
<td>800,019.</td>
<td>-$2.270.</td>
<td>-.776</td>
<td>.186</td>
<td>.0016</td>
</tr>
<tr>
<td>0.25</td>
<td>19.45</td>
<td>-$3.242.</td>
<td>-.601</td>
<td>.2195</td>
<td>.0039</td>
</tr>
</tbody>
</table>

Table 4.5.3

As noted earlier, if round off error is included, then the error term is of the form $O(h^r) + eh^{1-k}$. Thus we would expect that for fixed $r$, that as $h \to 0^+$, the error should decrease and then increase. To illustrate this consider...
\[
\begin{bmatrix}
0 & 0 & 0 \\
1 & 0 & 0 \\
0 & 1 & 0
\end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \sin t \\ 0 \\ \sin t \end{bmatrix}
\]

(5.2)

from [101]. The exact solution is \( \mathbf{x}^T = [-\sin t, -\cos t, 0]^T \). We shall use a backwards Euler's. The error estimate is then \( O(h) = \epsilon h^{-2} \). Let \( e(h) = Mh + \epsilon h^{-2} \). Then \( e(h) \) has a minimum at \( h = \left( \frac{2\epsilon}{M} \right)^{1/3} \). Since the entries of the solution and all their derivatives are bounded by 1, we would expect \( \epsilon^{1/3} \) to closely approximate the order of magnitude of the optimum \( h \).

For our example, \( \epsilon = 10^{-15} \), so that \( h \approx 10^{-5} \) should give the least error and for this \( h \), \( e(h) \) should be about \( M10^{-5} \approx 10^{-5} \).

The backwards Euler's gives

\[
\begin{align*}
x_{k+1} &= -\sin(t_{k+1}) \\
y_{k+1} &= -(\sin(t_{k+1}) - \sin(t_{k}))h^{-1} \\
z_{k+1} &= [2\sin(t_{k}) - \sin(t_{k+1}) - \sin(t_{k-1})]h^{-2} - \sin(t_{k+1}).
\end{align*}
\]

(5.3)

It is easy to show, using Taylor series, that \( z_{k} = O(h) \) which agrees to the accuracy expected with the solution \( z \equiv 0 \) ignoring round off error. However, using \( t_{k+1} = .4 \) the following values of \( z_{k+1} \) are computed from (3) for different \( h \).
\begin{table}
\centering
\begin{tabular}{|c|c|}
\hline
h & z(\cdot4) \\
\hline
0.1 & -0.094 \\
0.05 & -0.046 \\
0.001 & -0.00092 \\
0.0005 & -0.00046 \\
5 \times 10^{-5} & -0.000046 \\
10^{-5} & -0.0000094 \\
5 \times 10^{-6} & -0.0000048 \\
10^{-6} & 0.0000063 \\
5 \times 10^{-7} & 0.000048 \\
10^{-7} & 0.00055 \\
5 \times 10^{-8} & 0.0047 \\
10^{-8} & -0.25 \\
\hline
\end{tabular}
\caption{Table 4.5.4}
\end{table}

Note that the error is minimized around $5 \times 10^{-6}$ which is quite close to our rough estimate of $10^{-5}$. The error is also of the size to be expected.

Note also that if $r < k - 1$, as in our example, that $e(h)$ rises more rapidly after the minimum is reached than it initially decreases. For our example in Table 5, the error decreases linearly with $h$ at first but increases quite rapidly after $10^{-6}$. This suggests that in more complicated problems if $r < k - 1$, that the conservative approach (also the cheapest) would be to take slightly larger step sizes than a rough error analysis might suggest.

6. COMMENT ON $Ax = b$

The numerical procedures discussed in this chapter and Chapters 5 and 6 require the solution of the linear system;

\begin{equation}
\hat{E}_n x = \hat{b}.
\end{equation}
The $n \times n$ matrix $E_h$ is of the form,

$$E_h = \alpha A + \beta hB$$

where $A$ is singular, $\alpha, \beta$ are constants and $h$ is the step size.

Eigenvalues are often not the best way to analyze conditioning, but for (1), (2) they do provide an easy method of seeing the effect of $h$.

**Proposition 4.6.1** Suppose that $E_h$ is invertible for some $h$. Let $A = (\lambda A + B)^{-1}\hat{A}$, $k = \text{Ind}(\hat{A})$ where $\lambda$ is a scalar such that $\lambda A + B$ is invertible. Then

$$\| E_h^{-1} A^D A \| = O(1) \text{ as } h \to 0^+$$

$$\| E_h^{-1} (I - A^D A) \| = O(h^{-k}) \text{ as } h \to 0^+$$

**Proof** Observe that for any $n \times n$ matrix $C$

$$\left\| \frac{C}{(\lambda A + B)} \right\| \leq \left\| (\lambda A + B)^{-1} C \right\| \leq \left\| (\lambda A + B)^{-1} \right\| \| C \|.$$ 

Thus it suffices to consider $E_h = (\lambda A + B)^{-1} E_h = \alpha \hat{A} + \beta h(I - \lambda \hat{A})$. Then (3), (4) follow from the core-nilpotent decomposition of $A$ (or the Jordan form of $A$).

One immediate consequence of Proposition 1 is that if the index is bigger than one, then double precision is recommended, if not essential. Also for moderate sized systems, no more than a few hundred, unitary operations may be preferable to Gaussian elimination. These are incorporated into a number of packages such as LINPACK [38]. Another way to control ill-conditioning is to use higher order methods rather than reduce step size if greater accuracy is needed.
If the system arises from circuit equations as discussed in Chapter 3, then $E_h$ will usually be sparse (have mostly zero entries). For $n$ large, sparse matrix procedures can be used. Note that for the constant coefficient case studied in this chapter, that only $b$ changes in (1) at each step. Thus (1) need only be fully solved once. For example, if a LU factorization of $E_h$ is computed and saved, then further time steps only involve two sets of $n$ back substitutions.
5 Linear time varying systems

1. INTRODUCTION

The next chapter concerns autonomous non-linear systems of the form

$$\dot{x} + B(x) = f.$$  \hspace{1cm} (1.1)

This chapter will consider linear time varying systems of the forms

$$\dot{Ax}(t) + B(t)x(t) = f(t)$$ \hspace{1cm} (1.2)

and

$$A(t)\dot{x}(t) + B(t)x(t) = f(t).$$ \hspace{1cm} (1.3)

The systems (2) - (3) are, of course, of considerable interest in their own right. However, they are of fundamental interest when considering (1) for at least two reasons.

First, many of the numerical procedures used in solving (1) are either based on, or analyzed via, linear approximations. Secondly, (2) can be written as

$$\begin{bmatrix} A & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \dot{x} \\ \dot{z} \end{bmatrix} + \begin{bmatrix} B(x) \\ 0 \end{bmatrix} \begin{bmatrix} x \\ -1 \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}, \quad z(0) = 0,$$

which is in the form of (1). Thus linear time varying systems can be written as non-linear autonomous systems. Accordingly results about (1) will usually say something about (2). As we shall see, in this and the next chapter, many
of the additional difficulties present in moving from linear constant coefficients to (1) are already present in (2).

The system (2) has already been considered to some extent in [8, Chapter 6] where the reduction of (2) to constant coefficients and the decomposition of (2) (essentially by the approach of Luenberger [73]) are discussed. See also [42], [70], [102]. The concerns of this chapter are different. We shall be interested in properties of (2) that are important for understanding (1) and in discussing numerical procedures.

Throughout this chapter it will be assumed that the coefficients are real analytic in \( t \) unless stated otherwise. For many purposes, this assumption could be relaxed to sufficiently differentiable but at the expense of unduly complicating the discussion.

2. THE INDEX OF A LINEAR TIME VARYING SYSTEM

Our previous experience with the constant coefficient case suggests that the concept of index should be important in analyzing (1.3). This turns out to be true.

**Definition 5.2.1** Let \( \lambda \) be such that \( \lambda A + B \) is invertible at \( t = t_0 \). Then the system will be called **solvable** and the **index of the system** (1.3) at \( t_0 \) is \( \text{Ind}((\lambda A(t_0) + B(t_0))^{-1}A(t_0)) \), which will be denoted by \( \text{Ind}(A(t_0), B(t_0)) \) or just \( \text{Ind}(A,B) \).

Note that \( \text{Ind}(A,I) = \text{Ind}(A) \). Our first goal is to determine to what extent \( \text{Ind}(A,B) \) is an invariant of the system under coordinate changes.

A change of variables \( x = Py \), \( P \) invertible and constant will be called a **constant coordinate change**. If \( P \) depends on \( t \) we will call \( P \) a **coordinate change**. Let \( x = Py \). Then (1.3) becomes
\[ \dot{\lambda} P + (A\dot{P} + BP) = f \]  

(2.1)

P will be called an **allowable coordinate change** if at each value of \( t \) of interest, the system (2.1) is solvable, that is,

\[ \lambda AP + (A\dot{P} + BP) \text{ is invertible} \]  

(2.2)

for some scalar \( \lambda \) which may depend on \( t \). The coordinate change is **admissible** if

\[ A\dot{P} + BP \]  

(2.3)

is invertible for each value of \( t \) of interest.

Note that if \( P \) is an allowable coordinate change at \( t_0 \) and \( \lambda \) satisfies (2) at \( t_0 \), then \( e^{\lambda t}P \) is an admissible coordinate change in a neighborhood of \( t_0 \). Thus, without loss of generality, we shall limit ourselves to admissible coordinate changes. This will simplify our notation.

Suppose that \( P \) is an admissible coordinate change. Then (1) can be written as

\[ \hat{A}\hat{x} + \hat{x} = \hat{f} \]  

(2.4)

where \( \hat{f} = (A\dot{P} + BP)^{-1}f \) and \( \hat{A} = (A\dot{P} + BP)^{-1}AP = P^{-1}(APP^{-1} + B)^{-1}AP \). Let \( Q = PP^{-1} \). Note that \( Q(t) \) is arbitrary except for the requirement that \( \hat{A}Q + \hat{Q} \) is invertible since \( \hat{P} = QP \), \( P(0) = P_0 \) has a solution for any \( P_0 \). The index of (4) at \( t_0 \) is \( \text{Ind}(\hat{A}(t_0)) \).

Clearly the index of the system is not changed by constant coordinate changes or by admissible coordinate changes of the form \( e^{\lambda t}P_0 \), \( P_0 \) constant.
Example 5.2.1 Let $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$, $Q = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -1 & 0 \end{bmatrix}$, $P = QP$, $P(0) = I$, $B = I$. Then $\text{Ind}(A, B) = 3$, but $\text{Ind}((AQ + I)^{-1}A) = 2$. Thus admissible coordinate changes can sometimes raise or lower the index.

However, not every change is possible.

Theorem 5.2.1 Suppose that $A(t), B(t)$ are real analytic. Then the properties $\text{Ind}(A, B) \equiv 0$, $\text{Ind}(A, B) \equiv 1$, or $\text{Ind}(A, B) > 1$ at $t_o$ are invariant properties of (1.3) under real analytic admissible coordinate changes.

Proof That $\text{Ind}(A, B) \equiv 0$ (that is, $A(t)$ is invertible) is an invariant follows from the continuity of $A, B$. Thus it suffices to show $\text{Ind}(A, B) = 1$ at $t_o$ is an invariant.

Suppose then that $\text{Ind}(A, B) = 1$ at $t_o$ and (1.3) is written in the form (4). Also suppose that $A(t)$ is real analytic. Then the minors of $\hat{A}(t)$ are all real analytic. Thus $\text{rank}(\hat{A}(t))$ is identically constant except at isolated points where it may decrease. From [3] there is an analytic projection $R(t)$ defined on some interval $|t - t_o| \leq \epsilon$ such that $R(R(t)) = R(\hat{A}(t))$ for $0 < |t - t_o| < \epsilon$, $\text{rank}(\hat{A}(t))$ is constant for $0 < |t - t_o| < \epsilon$, and $R(t)$ has constant rank for $|t - t_o| < \epsilon$. Thus relative to the decomposition $R(R(t)) \oplus R(I - R(t))$ there is an analytic similarity $T(t)$ in a neighborhood of $t_o$ such that

$$T(t)\hat{A}(t)T^{-1}(t) = \begin{bmatrix} A_1(t) & A_2(t) \\ 0 & 0 \end{bmatrix}.$$ \hspace{1cm} (2.5)

Since $\hat{A}(t_o)$ has index one by assumption, we also have [21]

$$\text{Ind}(A_1(t_o)) = 1, \ R(A_2(t_o)) \subseteq R(A_1(t_o)).$$ \hspace{1cm} (2.6)
Let $T^{-1}QT = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$. Then

$$(AQ + I)^{-1}A = T \begin{bmatrix} A_{1}Q_{11} + A_{2}Q_{21} + I & A_{1}Q_{12} + A_{2}Q_{22} \\ 0 & I \end{bmatrix}^{-1} \begin{bmatrix} A_{1} & A_{2} \\ 0 & 0 \end{bmatrix} T^{-1}$$

$$= T \begin{bmatrix} (A_{1}Q_{11} + A_{2}Q_{21} + I)^{-1}A_{1} & (A_{1}Q_{11} + A_{2}Q_{21} + I)^{-1}A_{2} \\ 0 & 0 \end{bmatrix} T^{-1}. \tag{2.7}$$

To show (7) has index one also at $t_0$, we need to show that at $t_0$

$$\text{Ind}((A_{1}Q_{11} + A_{2}Q_{21} + I)^{-1}A_{1}) = 1 \tag{2.8}$$

and

$$R((A_{1}Q_{11} + A_{2}Q_{21} + I)^{-1}A_{1}) \supseteq R((A_{1}Q_{11} + A_{2}Q_{21} + I)^{-1}A_{2}). \tag{2.9}$$

(9) follows from $R(A_{2}(t_0)) \subset R(A_{1}(t_0))$. To see (8), note that since $A_{1}$ has index one, and $R(A_{2}) \subseteq R(A_{1})$ that there exists a constant matrix $W$ so that

$$A_{1}(t_0) = W^{-1}\begin{bmatrix} C_{1} & 0 \\ 0 & 0 \end{bmatrix} \text{W},$$

$$A_{2}(t_0) = W^{-1}\begin{bmatrix} C_{2} & C_{3} \\ 0 & 0 \end{bmatrix},$$

where $C_{1}$ is invertible. But then

$$(A_{1}Q_{11} + A_{2}Q_{21} + I)^{-1}A_{1} = W^{-1}\begin{bmatrix} X_{1} & X_{2} \end{bmatrix}^{-1} \begin{bmatrix} C_{1} & 0 \\ 0 & 0 \end{bmatrix} \text{W}$$

$$= W^{-1}\begin{bmatrix} X_{1}^{-1}C_{1} & 0 \\ 0 & 0 \end{bmatrix} \text{W}$$
which has index one.

There are two other invariants which are important. Clearly

**Proposition 5.2.1** Both rank(\(\hat{A}(t_0)\)) and \(N(\hat{A}(t_0))\) are invariant under admissible coordinate changes.

Example 1 shows the actual value of the index does not mean as much as it did for the constant coefficient case. After the change of coordinates given, the resulting system has index two but the solution involves second and not first derivatives of the functions \(f_1\). The exception to this is, according to Theorem 1, the index one case. As we shall show in the remainder of this chapter, the index one case has several other properties not shared by the other higher index cases.

3. THE INDEX ONE CASE

The first thing to notice is that there is a difference between (1.2) and (1.3). If (1.2) is put into the form (2.4), then \(N(\hat{A})\) is independent of \(t\) since \(N(\hat{A}(t)) = N(A(t))\). Conversely, suppose that \(N(\hat{A}(t))\) is independent of \(t\). Let \(M\) be a complement of \(N(\hat{A}(t))\). Then a constant coordinate change defined by the decomposition \(\Phi_n = M \oplus N(\hat{A}(t))\) gives

\[
\begin{bmatrix}
A_1(t) & 0 \\
A_2(t) & 0 \\
\end{bmatrix}
\begin{bmatrix}
\dot{x}_1 \\
\dot{x}_2 \\
\end{bmatrix}
+ 
\begin{bmatrix}
x_1 \\
x_2 \\
\end{bmatrix} =
\begin{bmatrix}
f_1 \\
f_2 \\
\end{bmatrix}.
\]  

(3.1a)

(3.1b)

The matrix \(\begin{bmatrix} A_1 & 0 \\ A_2 & 0 \end{bmatrix}\) has index one. Thus \(A_1\) has index one [21]. In fact \(A_1\) is invertible. To see this note that if \(A_1\phi = 0\) then

\[
\begin{bmatrix}
A_1 & 0 \\
A_2 & 0 \\
\end{bmatrix}
\begin{bmatrix}
\phi \\
0 \\
\end{bmatrix} =
\begin{bmatrix}
0 \\
0 \\
\end{bmatrix}.
\]

Thus if \(\begin{bmatrix} A_1 & 0 \\ A_2 & 0 \end{bmatrix}\) had index one, we must also have \(A_2\phi = 0\) which contradicts
\[
\begin{bmatrix}
\dot{\Phi} \\
0
\end{bmatrix} \notin N\left(\begin{bmatrix}
A_1 & 0 \\
A_2 & 0
\end{bmatrix}\right). \text{ Thus } A_1 \text{ is invertible. Multiply (1) by } \begin{bmatrix}
A_1^{-1} & 0 \\
-A_2A_1^{-1} & 1
\end{bmatrix}
to get
\]
\[
\begin{bmatrix}
1 & 0 \\
0 & 0
\end{bmatrix}\begin{bmatrix}
\dot{x}_1 \\
\dot{x}_2
\end{bmatrix} + \begin{bmatrix}
A_1^{-1}(t) & 0 \\
-A_2(t)A_1^{-1}(t) & 1
\end{bmatrix}\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix} = \begin{bmatrix}
A_1^{-1}(t)f_1 \\
-A_2(t)A_1^{-1}(t)f_1(t) + f_2(t)
\end{bmatrix} \quad (3.2)
\]

The preceding may be summarized in the following proposition.

**Proposition 5.3.1** If the index of the system (1.3) is one and \( N(A(t)) \) is independent of \( t \), then (1.3) may be written in the form (1.2) with a constant coordinate change.

The question of reducing (1.3) to the constant coefficient case is discussed in [8]. Two important special cases will now be considered. Note that for (1.2), \( \hat{A} \) also has constant rank but this is no longer equivalent to (1.2).

**Constant rank**

**Example 5.3.1** The system

\[
\begin{bmatrix}
1 & 0 & 0 \\
0 & 0 & t \\
0 & 0 & 0
\end{bmatrix} \begin{bmatrix}
x_1 \\
x_2 \\
x_3
\end{bmatrix} \dot{x} + x = 0 \quad (3.3)
\]

has index one at the point \( t_0 = 0 \) but does not have index one near \( t_0 \).

Note that in Example 1, rank \( \hat{A}(t) \) also changes at \( t_0 = 0 \). Since any system originally written as (1.2) has constant rank, the next few results can be useful.
Suppose that $A(t)$ in (1.3) has constant rank and the index of $\hat{A}(t_0)$ is one. Then after a constant similarity with respect to the decomposition $R(\hat{A}(t_0)) \oplus N(\hat{A}(t_0))$, we get

$$
\begin{bmatrix}
A_1 & A_2 \\
C_1A_1 & C_1A_2
\end{bmatrix}
\begin{bmatrix}
\dot{x}_1 \\
\dot{x}_2
\end{bmatrix}
+ 
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
= 
\begin{bmatrix}
f_1 \\
f_2
\end{bmatrix}
$$

(3.4a)

$$
\begin{bmatrix}
A_1 & A_2 \\
C_1A_1 & C_1A_2
\end{bmatrix}
\begin{bmatrix}
\dot{x}_1 \\
\dot{x}_2
\end{bmatrix}
+ 
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
= 
\begin{bmatrix}
f_1 \\
f_2
\end{bmatrix}
$$

(3.4b)

where $A_1(t_0)$ is invertible, $A_2(t_0) = 0$, $C_1(t_0) = 0$.

The decomposition (3.4) has several consequences. First, the behavior of Example 1 cannot occur if rank($\hat{A}(t)$) is constant.

**Proposition 5.3.2** If the index of the system (1.3) is one at a point $t_0$, $\hat{A}(t)$ is continuous and rank($A(t)$) is constant on a neighborhood of $t_0$, then the index is one in a neighborhood of $t_0$.

**Proof** Since $A_1(t_0)$ is invertible, $A(t)$ is continuous, and $C_1(t_0) = 0$, we have $A_1 + A_2C_1$ is invertible for $t$ near $t_0$. But then

$$
\begin{bmatrix}
I & 0 \\
-C_1 & I
\end{bmatrix}^{-1}
\begin{bmatrix}
A_1 & A_2 \\
C_1A_1 & C_1A_2
\end{bmatrix}
\begin{bmatrix}
I & 0 \\
C_1 & I
\end{bmatrix}
= 
\begin{bmatrix}
A_1 + A_2C_1 & A_2 \\
0 & 0
\end{bmatrix}
$$

(3.5)

The right hand side of (5) has index one if $A_1 + A_2C_1$ is invertible. Thus

$$
\begin{bmatrix}
A_1 & A_2 \\
C_1A_1 & C_1A_2
\end{bmatrix}
$$

has index one for $t$ near $t_0$ also. \[\Box\]

Now (4) is equivalent to

$$
A_1\dot{x}_1 + A_2\dot{x}_2 + x_1 = f_1, 
$$

(3.6a)

$$
\dot{x}_2 - C_1x_1 = f_2 - C_1f_1. 
$$

(3.6b)

Solving (6b) for $x_2$ and substituting in (6a) gives
\[
\begin{align*}
[A_1 + A_2 C_1] \dot{x}_1 + (A_2 \hat{C}_1 + I)x_1 &= f_1 - A_2 \dot{f}_2 + A_2 C_1 \dot{f}_1 + A_2 \hat{C}_1 f_1, \\
x_2 &= C_1 x_1 + \frac{f_2}{C_1} - \frac{f_1}{f_1}
\end{align*}
\] (3.7a)

As noted earlier, \(A_1 + A_2 C_1\) is invertible near \(t_0\) if the assumptions of Proposition 2 are met. Thus we have

**Theorem 5.3.1** Suppose that for (2.4) \(\hat{A}\) is continuous in a neighborhood of \(t_0\), \(\text{Ind}(\hat{A}(t_0)) = 1\), and \(\text{rank}(\hat{A}(t))\) is constant in a neighborhood of \(t_0\). Then the consistent initial conditions are characterized by

\[
(I - \hat{A}(t_0)^D \hat{A}(t_0))x(t_0) = (I - \hat{A}(t_0)^D \hat{A}(t_0))\frac{f}{f}(t_0),
\]

\(\hat{A}(t_0)^D \hat{A}(t_0)\) arbitrary. (3.8a)

Furthermore the change of variables \(x = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} y\) transforms (2.4) into

\[
\begin{align*}
(A_1 + A_2 C_1) \dot{y}_1 + A_2 \dot{y}_2 + y_1 &= f_1, \\
y_2 &= \frac{f_2}{C_1} - \frac{f_1}{f_1}.
\end{align*}
\] (3.9a)

Suppose that an \(r\)-th order backward difference scheme is used on (1.2). Then (9) shows that, ignoring round off error, \(y_2\) is computed exactly at time \(t_k\). The error in \(y_1\) will be \(O(h^{r+1})\) for each step and globally \(O(h^r)\) on any interval \([t_0, t_1]\) on which \(\text{Ind}(A(t)) \leq 1\) and \(\text{rank}(A(t))\) is constant.

Notice also that if the initial value is not consistent, then after one step, the procedure is on a solution since (9b) is solved exactly.

Of course the index can be identically one without the rank being constant.

Suppose that this occurs. That is, suppose that the \(\text{rank}(A(t)) = r\) for \(t \neq t_0\), \(\text{rank}(A(t_0)) < r\) and \(\text{Ind}(A(t)) \equiv 1\) for \(|t - t_0| < \epsilon\), some \(\epsilon > 0\). Suppose, in addition that \(A(t)\) is analytic for \(|t - t_0| < \epsilon\). Then there is
an analytic function $A^-(t)$ for $|t - t_0| < \varepsilon$ such that $A^-(t)$ is a $(1,2)$-inverse for $A(t)$ for $0 < |t - t_0| < \varepsilon$. Using a linearly independent set of columns of $AA^-$ and $I - AA^-$ to form the matrix $P(t)$ we have that

$$P^{-1}AP = \begin{bmatrix} A_1 & A_2 \\ 0 & 0 \end{bmatrix}$$

(3.10)

where $A_1$ is $r \times r$ and $A_1(t)$ is invertible for $0 < |t - t_0| < \varepsilon$. We may assume $A_2(0) = 0$. Let $x = P\dot{y}$. Then $A\dot{x} + x = f$ becomes

$$(P^{-1}AP)\ddot{y} + [(P^{-1}AP)P^{-1}P + I]y = P^{-1}f.$$  

(3.11)

That is,

$$\begin{bmatrix} A_1 & A_2 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \ddot{y}_1 \\ \dot{y}_2 \end{bmatrix} + \begin{bmatrix} A_1 & A_2 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}$$

or

$$A_1\ddot{y}_1 + A_2\ddot{y}_2 + (A_1Q_{11} + A_2Q_{21} + I)y_1 + (A_1Q_{11} + A_2Q_{22})y_2 = f_1$$

(3.12)

$$y_2 = f_2$$

(3.13)

Using (13), (12) becomes

$$A_1\ddot{y}_1 + (A_1Q_{11} + A_2Q_{21} + I)y_1 = f_1 - A_2\ddot{f}_2 - (A_1Q_{11} + A_2Q_{22})f_2$$

(3.14)

The matrix $A_1$ in (14) is invertible for $0 < |t - t_0| < \varepsilon$ and singular for $t = t_0$. Equations of this type will be considered in the next section.

Many problems, including a large class of circuits, when initially formulated are in the form (1.2). In this case not only is rank($\hat{A}(t)$) constant, but $N(\hat{A}(t))$ is constant as well. This does not mean that the index of the system is constant.
Example 5.3.2 Consider \[ \begin{bmatrix} t & 0 \\ 1 & 0 \end{bmatrix} \dot{x} + x = f. \] Then \( \text{rank}(\hat{A}(t)) = 1 \) and \( N(\hat{A}(t)) \) is a constant, but the index is 2 if \( t = 0 \) and 1 if \( t \neq 0 \).

One would expect that a constant nullspace should provide additional information. This will also be examined in the next section.

4. INDEX GREATER THAN ONE

For linear solvable systems, if the coefficients are constant, or the index is identically one, we have been able to establish necessary and sufficient conditions for uniqueness and consistency of solutions. The situation for a nonconstant higher index problem is quite different. The next example will be used to motivate much of this section.

Example 5.4.1 Consider the solvable system

\[
\begin{bmatrix}
0 & 1 \\
0 & 0
\end{bmatrix} \dot{\begin{bmatrix} x \\ y \end{bmatrix}} + \begin{bmatrix}
1 & t \\
t & 1
\end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}
\]

(4.1)

which is in the form (1.2). The system (1) is equivalent to the scalar equation

\[ \dot{x} + tx = g + (\dot{g} - f) t^{-1} \]

(4.2)

and

\[ y = g - tx. \]

(4.3)

In this example the index is one if \( t \neq 0 \) and two if \( t = 0 \). We are interested in the existence of solutions of (1) which are differentiable in a neighborhood of zero.
First of all, notice that if \( \dot{g}(0) - f(0) \neq 0 \), then (2) and hence (1) does not have a solution differentiable on \([0, t_1]\). Thus, unlike for the constant coefficient case, the system (1) need not have a solution for all smooth \( f \). On the other hand if \( \dot{g}(0) - f(0) = 0 \), then for any \( t_o \), (2) has a unique solution for any \( x_o = x(t_o) \), \( t_o \neq 0 \). \( y \) is then given by (3) so that \( y \) is at least as smooth as \( x \). Thus, in this example, even though the index changes at zero, the manifold of consistent initial conditions does not change dimension at \( t = 0 \).

**Example 5.4.2** Now consider the solvable system

\[
\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} + \begin{bmatrix} -2 & 0 \\ -t & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\]

(4.4)

or equivalently

\[
\frac{1}{2}\begin{bmatrix} 0 & -1 \\ 0 & -t \end{bmatrix}\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} + \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\]

(4.5)

Again, the system is index one except at \( t = 0 \) where the index is two.

The differentiable solutions to (4) are given by

\[
[x \ y]^T = [ct \ ct^2]^T
\]

(4.6)

At \( t = 0 \), consistent initial conditions no longer uniquely determine solutions since all the solutions pass through \( 0 \). Also the dimension of the manifold of consistent initial conditions drops from one to zero at \( t = 0 \).

Finally, consider
Example 5.4.3

\[
\begin{bmatrix}
0 & 1 \\ 0 & 0
\end{bmatrix}
\begin{bmatrix}
\dot{x} \\ \dot{y}
\end{bmatrix}
+ \begin{bmatrix}
-1 & 0 \\ t & 1
\end{bmatrix}
\begin{bmatrix}
x \\ y
\end{bmatrix}
= \begin{bmatrix}
0 \\ 0
\end{bmatrix}.
\] (4.7)

The only solution of (7) continuous on a neighborhood of \( t = 0 \) is the trivial solution \( x \equiv 0, y \equiv 0 \).

The remainder of this section will be concerned with examining more carefully the relationships between the various types of behavior that can occur in (1.2). While the form (1.2) is the direct generalization of (0.1.1) to the time varying case, there is an alternative format which often appears in the literature. To help put our discussion in perspective we shall use both formats. From the proof of Theorem 2.1,

**Proposition 5.4.1** The system (1.2) may be written as

\[
\dot{x} = G_{11}(t)x + G_{12}(t)y + f_1(t)
\] (4.8a)

\[
\dot{y} = G_{21}(t)x + G_{22}(t)y + f_2(t)
\] (4.8b)

by an admissible coordinate change if and only if \( \text{rank}(A(t)) \) is constant.

The system (8) is usually studied under the assumption that \( G_{22} \) is non-singular. The next proposition shows that this is equivalent to (8) being an index one system.

**Proposition 5.4.2** If \( G_{22}(t_o) \) is invertible, then there is a scalar \( \lambda \) such that at \( t_o \)

\[
\begin{bmatrix}
\lambda I + G_{11} & G_{12} \\ G_{21} & G_{22}
\end{bmatrix}
\] (4.9)

126
is invertible and the index of the system (8) is one. Conversely, if there is a \( \lambda \) such that (9) is invertible, and the index of the system (8) is one at \( t_0 \), then \( G_{22}(t_0) \) is invertible.

Proof Assume that \( G_{22}(t_0) \) is invertible. Then \( G_{22} \) is invertible for \( t \) near \( t_0 \). Let \( \Theta = \lambda I + G_{11} - G_{12}(G_{22}^{-1})G_{21} \). Let \( \lambda \) be such that \( \lambda I + G_{11} \) and \( \Theta \) are invertible. Such a \( \lambda \) always exists. Then (9) is invertible and

\[
\begin{bmatrix}
\lambda I + G_{11} & G_{12} \\
G_{21} & G_{22}
\end{bmatrix}^{-1} =
\begin{bmatrix}
\Theta^{-1} & \Theta^{-1}G_{11}G_{22}^{-1} \\
-G_{22}^{-1}G_{21}\Theta^{-1} & \phi
\end{bmatrix}
\]

(4.10)

where \( \phi = G_{22}^{-1} + G_{22}^{-1}(G_{21}\Theta^{-1}G_{12})G_{22}^{-1} \). Thus for (8)

\[
A = \begin{bmatrix}
\lambda I + G_{11} & G_{12} \\
G_{21} & G_{22}
\end{bmatrix}^{-1}
\begin{bmatrix}
I & 0 \\
0 & 0
\end{bmatrix} =
\begin{bmatrix}
\Theta^{-1} & 0 \\
X & 0
\end{bmatrix}
\]

which has index one since \( \Theta \) is nonsingular.

On the other hand suppose there is a \( \lambda \) such that (9) is invertible but \( G_{22}(t_0) \) is singular. Let \( u \) be such that \( G_{22}(t_0)u = 0 \). Then at \( t_0 \)

\[
\begin{bmatrix}
\lambda I + G_{11} & G_{12} \\
G_{21} & G_{22}
\end{bmatrix}
\begin{bmatrix}
0 \\
u
\end{bmatrix} =
\begin{bmatrix}
v \\
0
\end{bmatrix}
\]

where \( v \neq 0 \).

But then

\[
\begin{bmatrix}
\lambda I + G_{11} & G_{12} \\
G_{21} & G_{22}
\end{bmatrix}^{-1}
\begin{bmatrix}
I & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
v \\
0
\end{bmatrix} =
\begin{bmatrix}
0 \\
u
\end{bmatrix} \neq 0
\]

while

\[
\begin{bmatrix}
\lambda I + G_{11} & G_{12} \\
G_{21} & G_{22}
\end{bmatrix}^{-2}
\begin{bmatrix}
I & 0 \\
0 & 0
\end{bmatrix}^2
\begin{bmatrix}
v \\
0
\end{bmatrix} =
\begin{bmatrix}
0 \\
0
\end{bmatrix}
\]

127
so that the index is at least two. □

If the index is one, that is $G_{22}$ is invertible, then as noted in Section 3 it is easy to extend the usual results. Suppose then that $G_{22}$ is singular at $t_0$. Then (8) can be rewritten by taking the $y$ decomposition $N(G_{22}(t_0)) \perp N(G_{22}(t_0))$ and multiplying (8b) by another matrix determined by the decomposition $R(G_{22}(t_0)) \perp R(G_{22}(t_0))$ to get

\[
\begin{align*}
\dot{x} &= G_{11}x + G_{12}y_1 + G_{13}y_2 + f_1 \\
0 &= \tilde{G}_{21}x + \tilde{G}_{22}y_1 + \tilde{G}_{23}y_2 + \tilde{f}_2 \\
0 &= \tilde{G}_{31}x + \tilde{G}_{32}y_1 + \tilde{G}_{33}y_2 + \tilde{f}_3
\end{align*}
\] (4.11a, 4.11b, 4.11c)

where $\tilde{G}_{22}(t_0)$ is invertible and $\tilde{G}_{23}(t_0) = 0$, $\tilde{G}_{32}(t_0) = 0$, $G_{33}(t_0) = 0$.

The system (11) is still solvable if (8) is, since if $U, V$ are invertible,

\[
\begin{bmatrix}
\lambda I + C_{11} & C_{12}V \\
UC_{21} & UC_{22}V
\end{bmatrix} = 
\begin{bmatrix} I & 0 \\
0 & U \end{bmatrix}
\begin{bmatrix}
\lambda I + C_{11} & C_{12} \\
C_{21} & C_{22}
\end{bmatrix}
\begin{bmatrix}
I \\
0 & V
\end{bmatrix} = (4.12)
\]

shows that the left side of (12) is invertible for some $\lambda$ if and only if

\[
\begin{bmatrix}
\lambda I + C_{11} & C_{12} \\
C_{21} & C_{22}
\end{bmatrix}
\]
is invertible for the same $\lambda$. Solving (11b) for $y_1$ and substituting into (11a) gives the new system which is also still solvable,

\[
\begin{align*}
\dot{x} &= H_{11}x + H_{13}y_2 + \tilde{f}_1, \\
0 &= H_{31}x + H_{33}y_2 + \tilde{f}_3
\end{align*}
\] (4.13a, 4.13b)

The system (13) is again in the form (8) but with fewer dependent variables. There are four possibilities.
(i) \( H_{33}(t_0) \) is invertible

(ii) \( H_{33}(t_0) \) does not have constant rank in a neighborhood of \( t_0 \)

(iii) \( H_{33}(t_0) \equiv 0 \) on a neighborhood of \( t_0 \)

(iv) \( H_{33}(t_0) \neq 0 \), \( H_{33}(t_0) \) singular, and \( H_{33}(t_0) \) has constant rank on a neighborhood of \( t_0 \).

Consider first case (iii). Then (13) is

\[
\begin{align*}
\dot{x} &= H_{11}x + H_{12}y + f_1 \\
0 &= H_{21}x + f_2
\end{align*}
\] (4.14a)

The solvability of (13) at \( t_0 \) implies that \( H_{12}(t_0) \) is one to one (of full column rank), \( H_{21}(t_0) \) is onto (of full row rank). By continuity these properties hold on a neighborhood of \( t_0 \). Thus \( N(H_{21}) \) has constant dimension near \( t_0 \). Let \( T(t) \) be invertible near \( t_0 \) and such that \( H_{21}(t)T(t) \equiv [I \ 0] \).

Let \( x = Tz \). Then (14) is

\[
\begin{align*}
\dot{T}z + T\dot{z} &= H_{11}Tz + H_{12}y + f_1 \\
0 &= [I \ 0]z + f_2.
\end{align*}
\] (4.15)

Let \( z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \), and (15) becomes

\[
\begin{align*}
\dot{z}_1 &= \tilde{H}_{11}z_1 + \tilde{H}_{12}z_2 + \tilde{H}_{13}y + \tilde{f}_1 \\
\dot{z}_2 &= \tilde{H}_{21}z_1 + \tilde{H}_{22}z_2 + \tilde{H}_{23}y + \tilde{f}_2 \\
0 &= z_1 + \tilde{f}_3
\end{align*}
\] (4.16a)

Taking (16c) and substituting (16a), (16b) we have
\[ \dot{z}_2 = \tilde{H}_{22}z_2 + H_{23}\dot{y} + \tilde{f}_2 - \tilde{H}_{21}\tilde{f}_3 \]  
\[ 0 = \tilde{H}_{12}z_2 + H_{13}\dot{y} + \tilde{f}_1 - \tilde{H}_{11}\tilde{f}_3 + \tilde{f}_3 \]

The system (17) is again in the form (8). Is the system solvable? The system (14) is solvable at \( t_0 \) if and only if

\[ H_{21}(\lambda I + H_{11})^{-1}H_{12} \]  

is invertible for some \( \lambda \) at \( t = t_0 \).

The corresponding condition after the coordinate change is

\[ H_{21}(\lambda I + H_{11} + \dot{\tilde{T}}^{-1})^{-1}H_{21} \]

is invertible for some \( \lambda \). Let \( t = s/\alpha \) and define \( s_0 = \alpha t_0 \). Note that solvability of (8) is unaffected by such a change in time scale. The expression (19) at \( s_0 \) is

\[ H_{21}(\lambda I + H_{11} + \alpha \frac{dT}{ds} T^{-1})^{-1}H_{12} \]

Since (20) is invertible at \( \alpha = 0 \), it is invertible for all but a finite number of \( \alpha \). Thus solvability can always be preserved if changes in time scale are allowed. In many problems (19) will still be solvable without a time scale change.

If case (iv) holds, then the previous procedure used to derive (17) may be repeated.

Eventually then the problem reduces to either case (i) or case (ii). If we arrive at case (i) by only considering (iii), (iv) the results look much like the constant coefficient case. The dependent variable is decomposed into two components. One is given explicitly in terms of derivatives of forcing functions, the other is a classical ordinary differential equation.
In particular, solutions are defined on an interval about $t_0$, the number of linearly independent solutions is independent of $t$, and the solutions are uniquely determined by their values at $t_0$.

Suppose now, that at some stage, $H_{33}$ has a rank change at $t_0$. If $H_{33}$ is real analytic at $t_0$, then again using (3) and repeating the above process the system is reduced to (13) with $H_{22}$ invertible for $0 < |t - t_0| < \varepsilon$, some $\varepsilon > 0$, but $H_{22}(t_0)$ not invertible.

Multiplying by $\begin{bmatrix} \lambda I + H_{11} & H_{12} \\ H_21 & H_{22} \end{bmatrix}^{-1}$, (13) becomes

$$
\begin{bmatrix}
A_1 & 0 \\
A_2 & 0
\end{bmatrix} \begin{bmatrix}
\dot{x} \\
\dot{y}
\end{bmatrix} = \begin{bmatrix}
\lambda A_1 - I & 0 \\
\lambda A_2 - I
\end{bmatrix} \begin{bmatrix}
x \\
y
\end{bmatrix} + \begin{bmatrix}
\hat{f}_1 \\
\hat{f}_2
\end{bmatrix}.
$$

(4.21)

By Proposition 2, $A_1$ is invertible for $0 < |t - t_0| < \varepsilon$ but singular for $t = t_0$.

The subsystem (21) will be referred to frequently in what follows.

Consider the first equation in (21);

$$
A_1\dot{x} + (I - \lambda A_1)x = \hat{f}_1
$$

(4.22)

and say $A_1$ is $r \times r$. Then for $t_1 \neq t_0$, $t_1$ close to $t_0$ there are $r$ linearly independent solutions of (22). However, $x(t_0)$ is in the rank($A_1(t_0)$) dimensional linear manifold, $R(A_1(t_0)) + \hat{f}_1(t_0)$, so that initial conditions of the form $x(t_0) = \hat{x}_0$ cannot uniquely determine all solutions.

As noted with Example 1, the system (22) may, or may not have solutions continuous or differentiable at $t_0$. If (22) has at least one smooth solution, then the behavior of the rest of the solutions is determined by the associated homogeneous equation
\[ A_1 \dot{x} + (I - \lambda A_1)x = 0 \]  \hspace{1cm} (4.23)

where \( A_1 \) is \( r \times r \).

For our purposes, there are three cases of interest.

**Case I** There exist discontinuous (or nondifferentiable) solutions at \( t_0 \) of (23).

**Case II** There are \( r \) linearly independent solutions of (23) at \( t_0 \) which are continuous (or differentiable) but solutions are not uniquely determined by \( x(t_0), y(t_0) \).

**Case III** There are \( r \) linearly independent continuous (or differentiable) solutions to (23) which are uniquely determined by \( x(t_0), y(t_0) \).

**Example 5.4.1** (cont.) The equation (23) is equation (2). The associated homogeneous equation is

\[ \dot{x} + tx = 0. \]  \hspace{1cm} (4.24)

Thus solutions are uniquely determined by \( x(t_0) \) and there is a one-dimensional family of solutions for all values of \( t \). Note, however, that \( y(0) = g(0) \) so that \( x(0) \) is needed. Thus Example 1 is in Case III.

As shown by (4.6), Example 2 is in Case II. A simple example of Case I is \(-3t \dot{x} = x.\)

The next section considers the numerical solution of linear time varying singular systems. As to be expected errors are introduced. But for most of the problems considered the procedure steps from one solution curve to another. This is quite helpful for singular systems since the solutions form a proper submanifold of \( \mathbb{R}^n \times \mathbb{R} \). In Case I, it is to be expected that
the methods might fail near \( t_0 \). However, in Cases II and III, and sometimes I, the methods can often be designed to work.

If we are in Case II or Case III (and sometimes I) the system (23) has a regular singular point at \( t_0 \). That is, the solution is of the form 
\[ S(t)(t - t_0)^L \] where \( S \) is analytic at \( t_0 \) and \( L \) is a matrix. Turritan [108], [109] and more recently Lutz [76] [77] [78] [79] have developed a theory for determining whether \( t_0 \) is a regular singular point. Actually, Cases II and III are much stronger than regular singular since no logarithms are allowed in the solution. Hall [60], [61] has developed criterion for determining the number of analytic solutions to a system like (23).

In concluding this section, the results of Chapter 0 will be used to develop some useful formulas. Consider again the system (1.2). Let a zero subscript denote evaluation at \( t_0 \). Then (1.2) may be written

\[ A\dot{x} + B_0 x = (B_0 - B)x + f \] (4.25)

Suppose that \( x \) is an \( \text{Ind}(\hat{A}) \) times differentiable solution of (25). Let \( g = (B_0 - B)x + \mathcal{f} \). Then \( x \) is the solution of

\[ A\dot{x} + B_0 x = g \] (4.26)

From Theorem 0.1.3 (and taking \( t_0 = 0 \) for convenience)

\[ x = e^{-\hat{A}D_0^s}A\hat{x}_0 + \int_0^t e^{\hat{A}D_0^s(s-t)}\left(\hat{A}D^s_0(s-t)\right)dx ds + (I - A\hat{D}_0^s) \sum_{i=0}^{k-1} (-\hat{A}D_0^s)^i B_0 g(0), \] (4.27)

where \( k = \text{Ind}(\hat{A}) = \text{Ind}(A,B_0) \). The formulas (27), (28) while not really useful for explicitly solving (1.2), do shed light on questions such as uniqueness, existence, and numerical methods.
The consistent initial conditions satisfy

\[(I - \hat{A}^\Delta)x_o = (I - \hat{A}^\Delta)\sum_{i=0}^{k-1} \left[ -\hat{\alpha}_{o} \hat{D}_{i} \hat{D}_{o} \hat{D}_{o}^*(l) \right] \cdot (0). \tag{4.29}\]

If \( k = 1 \), this is just \( (I - \hat{A}^\Delta)x_o = (I - \hat{A}^\Delta)B_o \hat{f}_o \) which from Theorem 3.1, is also sufficient. For \( k = 2 \), (29) becomes

\[(I - \hat{A}^\Delta)x_o = (I - \hat{A}^\Delta)B_o \hat{f}_o \left( f_o - \hat{\alpha}_{o} \hat{D}_{o} \hat{f}_o - \hat{B}_{o} x_o \right). \tag{4.30}\]

Notice that in (30), the vectors \( x_o \) depend not only on \( B_o \) but \( \hat{B}_o = \hat{B}(t_o) \). Thus for systems of higher index, solutions of the time varying system can differ greatly from the constant approximation at \( t_o \) if \( \hat{B}_o \neq 0 \).

5. NUMERICAL SOLUTION OF LINEAR TIME VARYING SYSTEMS

If an implicit backward difference scheme, such as those in Chapter 4 is applied to (1.3), both of the coefficient matrices are evaluated at the same time. Thus even though the transformations of the preceding section may be too complicated to perform in practice, the approach of Section 4 or Section 6.4 of [8] can be useful in analyzing the numerical procedures.

If the system (1.2) has index one at \( t_o \) and an implicit scheme is used on (1.2) then for purposes of analysis, Proposition 3.1 and its proof show that it suffices to consider

\[
\dot{x} = G_{11}x + G_{12}y + f_1 \quad \tag{5.1a}
\]

\[
\dot{y} = G_{21}x + G_{22}y + f_2 \quad \tag{5.1b}
\]

with \( G_{22} \) invertible. Thus the numerical method is actually solving the classical problem

\[
\dot{x} = (G_{11} - G_{12}G_{22}^{-1}G_{21})x + f_1 - G_{22}^{-1}f_2 \quad \tag{5.2}
\]

134
with an implicit scheme and algebraically getting \( y \) as

\[
y = -G_{22}^{-1}(G_{21}x + f_2).
\]

(5.3)

The standard analysis applies to (2). In particular, the time step and order of the method can be varied.

The advantage of working with the singular system (1.2) or (1.3) directly rather than doing the analysis to get (1) was recognized at least ten years ago by Gear [53]. Gear’s original code has since been analyzed and modified by several authors. See Lininger [72], and Brayton, et. al. [7], for example. These codes actually also apply to non-linear systems which will be discussed in the next Chapter and use a Newton’s method or its equivalent to solve for \( y \) in (1b). They assume that \( G_{22} \) is invertible and thus, by Proposition 4.1, that the system has index one. It is not our intention to repeat this important work. Rather, the case when \( G_{22} \) is singular will be considered.

If at each step of the reduction of Section 4 (or Section 6.4 of [8]) either only a constant coordinate change or one just involving \( y \) is used and an \( r^{th} \) order method is used, then the numerical solution will approximate the actual solution with an error of \( O(h^r) \). If round off error is considered, there will be at most \( k \) 1 differentiations of combinations of the \( f_i \) (where \( k \) is the number of reductions) and an error term of the form \( \epsilon h^{1-k} \) must be considered where \( \epsilon \) depends on the machine round off error.

This is essentially the same result as for the time invariant linear systems of Chapter 4. As pointed out there, the time step \( h \) is assumed fixed if the index is greater than one.

Suppose, on the other hand, that at some stage, the appropriate \( G_{22} \) does not have constant rank. Thus the numerical method must be being used on a
subsystem of the form (4.21). The second equation of (4.21) algebraically
gives $y$ if $\dot{x}$, $x$ are known. The key equation is (4.22). Suppose, then to
begin with, that $X(t)$ is a fundamental matrix for (4.22) such that $X(t)$ is
sufficiently smooth on an interval about a point $t_0$, $X(t)$ is invertible for
$|t - t_0| > 0$, but $X(t_0)$ is singular.

For the time-invariant systems of Chapter 4 a backwards Euler’s worked
even for problems with index greater than one. The situation is now
different.

**Example 5.4.1 (cont.)** Again consider (4.1) but with $f = 1$, $g = e^t$ so
that $f(0) = \dot{g}(0)$ and there are solutions. Applying a backwards Euler’s gives

\begin{equation}
\begin{bmatrix}
0 & 1 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
x_{k+1} - x_k \\
y_{k+1} - y_k
\end{bmatrix}
+ \begin{bmatrix}
t_{k+1} & 1 \\
t_{k+1} & 1
\end{bmatrix}
\begin{bmatrix}
x_{k+1} \\
y_{k+1}
\end{bmatrix}
= \begin{bmatrix}
1 \\
1
\end{bmatrix}.
\end{equation}

Assume there is an $l$ so that $t_l = 0$. Then (4b) implies $y_l = e^{t_l} = 1$. Then
t_{l+1} = h and (4) gives $x_{l+1} = [h^2 e^h + e^h - h - 1]h^{-3}$, so that

\begin{equation}
x_{l+1} = \frac{3}{2} \left( \frac{1}{h} \right) + \frac{7}{6} + O(h).
\end{equation}

independent of $x_0$, $y_0$. The numerical solution provided by Euler’s is
exhibiting an impulse as $h \to 0^+$ even though all the solutions to (4.1) are
smooth and we are in Case III.

Notice that in Example 4.1, $y_l = 1$, and (5) holds regardless of $x_0$, $y_0$.
In some sense, Euler’s fails because at $t = 0$, the procedure cannot tell
which solution to follow. Note that the characterization of consistent
initial conditions (4.29) involves not only $x(t_0)$ but $x^{(i)}(t_0)$. This
suggests that higher order methods might be more effective since they use
values at several time steps. In the same vein Gantmacher [52, p. 160] has
noted that a solution to (4.22) that is regular at the regular singularity $t_0$ is sometimes uniquely determined by the values at $t_0$ of $x(t)$ and its first $s$ derivatives where $s$ is an integer determined by a matrix computed from $A_1(t)$. Now this result is only proved in [52] for the case when $A^{-1}(t)$ has a simple first order pole at $t_0$. However, (4.29) suggests that it is valid in some form for the general problem.

As an example utilizing a second order method, again consider (4.1) with $f = te^t + e^t + g$, $g = e^t + t^2$, $x(-1) = -1$, $y(-1) = e^{-1}$ so that the solution is $x = t$, $y = e^t$. Euler's method fails at $t_k = 0$. Use the difference scheme

$$
\dot{y}(t_k) \approx [\frac{3}{2} y_k - 2y_{k-1} + \frac{1}{2} y_{k-2}]h^{-1}
$$

(5.6)

discussed in Chapter 4. The error values for $h = .1$, taking the starting values as exact, are

<table>
<thead>
<tr>
<th>$h = .1$</th>
<th>TIME</th>
<th>x ERROR</th>
<th>y ERROR</th>
</tr>
</thead>
<tbody>
<tr>
<td>-.8</td>
<td>.00025</td>
<td>.00020</td>
<td></td>
</tr>
<tr>
<td>-.6</td>
<td>.00074</td>
<td>.00044</td>
<td></td>
</tr>
<tr>
<td>-.4</td>
<td>.00160</td>
<td>.00064</td>
<td></td>
</tr>
<tr>
<td>-.2</td>
<td>.00329</td>
<td>.00066</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>.00987</td>
<td>-.00000</td>
<td></td>
</tr>
<tr>
<td>.2</td>
<td>-.00359</td>
<td>.00072</td>
<td></td>
</tr>
<tr>
<td>.4</td>
<td>-.00592</td>
<td>.00237</td>
<td></td>
</tr>
<tr>
<td>.6</td>
<td>-.00725</td>
<td>.00435</td>
<td></td>
</tr>
<tr>
<td>.8</td>
<td>-.00793</td>
<td>.00635</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>-.00816</td>
<td>.00816</td>
<td></td>
</tr>
</tbody>
</table>

For comparison some of the values for $h = .01$ are given in the next table,
At \( h = .001 \), all error terms are zero to at least six places. Notice in both tables that the error appears to grow somewhat faster for \( t \) near zero but that the accuracy is still \( O(h^2) \). In fact, the analysis given later in this section, shows that this is the case for Example 4.1.

This and several other examples we've run suggests that the second order method almost always works provided those solutions that exist are all smooth enough at \( t_0 \). In fact, if one solution is smooth and the others blow up slowly enough at \( t_0 \), the method is still often able to estimate the one smooth solution.

We have not noticed much work in the literature on the numerical solution of (4.22). A complete analysis would take too much space and be too technical for this volume. Instead an important special case will be analyzed in some detail using the second order scheme (6).

Consider (4.22) under the assumption \( A_1(t_0) = 0 \). If this is not the case, the system may be further broken down. Suppose also that \( t_0 \) is a regular singularity of (4.22). That is, \( A_1 = (t - t_0)\tilde{A}_1(t) \) and \( \tilde{A}_1(t_0) \) is invertible. Finally suppose \( t_0 = 0 \) and the interval is \([\alpha, \beta] \), where \( \alpha < 0, \beta > 0 \). Thus we are supposing (4.22) may be written as
\[ \dot{x} + B(t)x = f \] (5.7)

and \( B(0) \) is invertible. The system (7) or its variants are used in Lutz [79], Hall [61] and Gantmacher [52]. Applying (6) to (7) gives

\[
t_{k+1} \left( \frac{3}{2} x_{k+1} - 2x_k + \frac{1}{2} x_{k-1} \right) + hB_{k+1}^x = h\dot{f}_{k+1}
\]

where \( x_{k+1} = x(t_{k+1}) \) for the desired solution \( x(t) \) and \( B_k = B(t_k) \), \( \dot{f}(t_k) = \dot{f}_k \), and \( h \) is the step size. Assume \( x(t) \) is at least twice differentiable. Assume also there is an \( \ell \) so that \( t_{k,\ell} = 0 \). Finally, let \( \{ x_k \} \) be the solution of

\[
t_{k+1} \left( \frac{3}{2} x_{k+1} - 2x_k + \frac{1}{2} x_{k-1} \right) + hB_{k+1}^x = h\dot{f}_{k+1}.
\]

The error on the \( k \)th step is \( e_k^x = x_k - x_k \). The error satisfies

\[
t_{k+1} \left( \frac{3}{2} e_{k+1}^x - 2e_k + \frac{1}{2} e_{k-1} \right) + hB_{k+1}e_{k+1} = t_{k+1} 0(h^3).
\] (5.8)

It will be convenient to also consider the equation (8) near the origin.

Note that \( t_{k,\ell} = \ell h \) where \( k = 0, \pm 1, \pm 2 \). Equation (8) is then

\[
k \left[ \frac{3}{2} e_{k+1}^x - 2e_{k+1} + \frac{1}{2} e_{k-2} \right] + B_{k+1}e_{k+1} = k0(h^3).
\] (5.9)

From (9) with \( k = 0 \), we get \( e_{0}^x = 0 \). For \( k \neq 0 \),

\[
e_{k+1}^x = \left[ \frac{3}{2} + \frac{1}{k} B_{k+1} \right]^{-1} 0(h^3) + 2e_{k+1} - \frac{1}{2} e_{k+2}.
\] (5.10)

Define \( \|B\|_\infty = \sup_{a \leq t \leq b} \|B(t)\| \) and take an integer \( K_0 \geq 0 \) so that

\[
e = \|B\|_{K_0}^{-1} < 1.
\] (5.11)

Note that \( K_0 \) is independent of \( h \). If \( \|e_{k-K_0}^x\| = O(h^r) \), \( \|e_{k-K_0-1}^x\| = O(h^r) \), then \( \|e_k^x\| = O(h^{\min(r,3)}) \) for \( |k - \ell| \leq K_0 \) since there are a fixed number of
steps. Thus to prove convergence it suffices to prove that \( ||e_i|| \leq M_1h^r \) for \( i \leq l - K_0 \) if \( e_0 = e_1 = 0 \) and \( ||e_i|| \leq M_1h^r \) for \( i \geq K_0 + l \) if \( ||e_{K_0+l}|| \).

\[
||e_{K_0+l+1}|| = O(h^r).
\]

Let \( X(t) \) be the fundamental solution of (7) such that \( X(a) = I \). Also suppose that the \( O(h^3) \) term in (10) is bounded by \( Mh^3 \). Then from (10),

\[
||e_2|| \leq \left( \frac{3}{2} - \epsilon \right)^{-1} Mh^3.
\]

Now treat \( x_2 \) as exact. Then \( x_1 \) is known within \( L_1h^3 \).

The local error at each step will be derived inductively as follows.

Suppose that one has \( x_k \) and \( x_k \) is assumed to be exact, that is \( e_k \) is assumed to be zero. Let \( c_k \) be such that \( X(t_k)c_k = x_k \). Then the value at \( t_{k-1} \) should be \( X(t_{k-1})c_k \) instead of \( x_{k-1} \). Assume that \( ||X(t_{k-1})c_k - x_{k-1}|| \leq L_kh^3 \). Then \( e_{k+1} \) gives the difference between \( x_{k+1} \) and the true value assuming \( x_k \) is correct which is \( X(t_{k+1})c_k \). Let \( x_{k+1} = X(t_{k+1})c_{k+1} \). Then from (10)

\[
||x_{k+1} - X(t_{k+1})c_k|| = ||X(t_{k+1})(c_{k+1} - c_k)||
\]

\[
= ||e_{k+1}|| \leq \left( \frac{3}{2} - \epsilon \right)^{-1} \left( \frac{1}{2} L_kh^3 + Mh^3 \right)
\]

for \( (k + 1) \leq l - K_0 \).

In order to determine how the \( L_k \) grow, suppose that there is a constant \( \theta \), independent of \( h \), so that

\[
||X(t_k)X^{-1}(t_{k+1})|| \leq \theta \text{ for } |k - l| \geq K_0 \quad (5.12)
\]

This assumption will be examined more carefully later. If \( x_{k+1} \) is assumed correct, the error in \( x_k \) is
\[ \|x(t_k)(c_{k+1} - c_k)\| = \|x(t_k)x(t_{k+1})^{-1}x(t_{k+1})(c_{k+1} - c_k)\| \]
\[ \leq \Theta \left( \frac{3}{2} - \varepsilon \right)^{-1} \left[ \frac{1}{2} L_k h^3 + M h^3 \right]. \]

That is,
\[ L_{k+1} = \Theta \left( \frac{3}{2} - \varepsilon \right)^{-1} \left[ \frac{1}{2} L_k + M \right]. \]

Let \( f(x) = \Theta \left( \frac{3}{2} - \varepsilon \right)^{-1} \left[ \frac{1}{2} x + M \right] \). This function is monotonically increasing since \( \varepsilon < 3/2 \). It will be a contraction and have a fixed point if
\[ \Theta \left( \frac{3}{2} - \varepsilon \right)^{-1} \frac{1}{2} < 1, \]
that is
\[ \Theta < 3 - 2\varepsilon. \] (5.13)

Since \( K_o \) can be chosen large (if \( h \) is small), (13) will be satisfied for small enough \( h \) if \( \Theta < 3 \). Suppose (13) is satisfied and let \( L_\infty = \sup_{0 \leq k \leq \ell - K_o} L_k \).

Then the error at each step for \( 0 \leq k \leq \ell - K_o \) is bounded by \( O(h^3) \). Now taking \( \varepsilon_{\ell - K_o} = 0, \| \varepsilon_{\ell - K_o - 1} \| = O(h^3) \), and using (10) for \( 2K_o \) iterations provided that the numbers \( \left\{ \frac{3}{2} k \right\}, k = 0, \pm 1, \ldots, \pm K_o \) are not in \( \sigma(B_{\ell}) \), the error \( \varepsilon_{\ell + K_o}, \varepsilon_{\ell + K_o + 1} \) will also be \( O(h^3) \). A similar argument as the preceding now gives a local error of \( O(h^3) \) for \( \ell + K_o \leq k \). Thus we have the following.

Proposition 5.5.1 If for the system (7), \( K_o \) is chosen so that (11) holds and (12), (13) hold for sufficiently small \( h \), where \( \Theta \) is independent of \( h \), and if \( \left\{ \frac{3}{2} k : |k| \leq K_o \right\} \cap \sigma(B_{\ell}) = \emptyset \), then the difference scheme (6) gives an \( O(h^2) \)-approximation on \([\alpha, \beta]\).
Note that while Proposition 1 says that the error is \( O(h^2) \) uniformly on \([a, b]\) it might still be a factor (independent of \( h \)) larger on \([a, b]\) than on \([a, a + K_o h]\). In fact the proof suggests that unless \( \| 2( \frac{3}{2} + B_2 )^{-1} \| < 1 \) this might well be the case.

The type of analysis used to prove Proposition 1 can be used to show that (6) can still be used sometimes to compute differentiable solutions of (7) even if the other solutions are not continuous.

To see this, again suppose that \( X(t) \) is a fundamental solution of \( \dot{x} + B(t)x = 0 \) and \( x_p \) is a particular solution of \( \dot{x} + B(t)x = f \) which is differentiable on \([a, b]\), \( 0 \in (a, b) \). Let \( \{x_k\} \) be the solution of (7) using (6). Then there are \( c_k \) so that if \( x_0 = x_p(t_0) \), \( x_1 = x_p(t_1) \), then

\[
x_k = X(t_k)c_k + x_p(t_k)
\] (5.14)

If (11), (12) still hold, then \( \| X(t_k)c_k \| = O(h^2) \) for \( 0 \leq k \leq \ell - K_o \). \( \{x_k\} \) will still approximate \( x_p \) if \( \| X(t_k) \| = o(t_k^{-2}) \) for \( |k - \ell| \leq K_o \). Rather than go through the details an illustrative example will be given.

**Example 5.5.1** Consider

\[
\begin{bmatrix}
0 & 1 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
\dot{x} \\
\dot{y}
\end{bmatrix}
+ \begin{bmatrix}
1 & t \\
3t & 1
\end{bmatrix}
\begin{bmatrix}
x \\
y
\end{bmatrix}
= \begin{bmatrix}
t + te^t + e^t \\
e^t + 3t^2
\end{bmatrix}
\] (5.15)

which has the particular solution

\[
x_p = t, \quad y_p = e^t.
\] (5.16)

The associated homogeneous system has the \( x \) equation

\[
3tx + (2 + 3t^2)x = 0
\] (5.17)
which is in the form of (7).

The solution of the associated homogeneous equation is

\[ x(t) = c \frac{e^{t^2/2}}{t^{2/3}}, \]

(5.18)

\[ y(t) = -3ct^{1/3} e^{t^2/2}. \]

(5.19)

Thus all solutions of (15) are discontinuous at the origin except for (16).

The condition (12) is

\[ \frac{x(t-h)}{x(t)} = \frac{e^{(t-h)^2/2}}{e^{t^2/2}} \left[ \frac{t}{t-h} \right]^{2/3} \leq \epsilon. \]

(5.20)

For small enough \( h \), \( |e^{(t-h)^2/2} e^{-t^2/2}| \leq 1 + \epsilon \) and \( 0 \leq \left[ \frac{t}{t-h} \right]^{2/3} \leq 1 \) for \( t < 0 \), and \( 0 < \delta < t \). Thus (13) holds with \( K_0 = 2 \). Also \( \|B\|_\infty = 1 + \delta \) for a small enough (but independent of \( h \)) interval around the origin. Thus from the proof of Proposition 1, one would expect \( \|X(t_k)c_k\| \leq Mh^{3} \) for \( 0 \leq k \leq t - K_0 \). But \( |x_h(t)| = 0 \left( \frac{1}{t^{2/3}} \right) \). Intuitively the error would be approximately \( M \frac{(B - \alpha)}{h} h^{3} h^{-2/3} = O(h^{4/3}) \).

In the table that follows, (15) was solved using (6) starting at \( t = -1 \). The starting values were taken as exact.

<table>
<thead>
<tr>
<th>VALUE OF h</th>
<th>ERROR IN X AT t = 0</th>
<th>( h^{4/3} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>. 1</td>
<td>-.0473705</td>
<td>.04642</td>
</tr>
<tr>
<td>. 01</td>
<td>-.0023512</td>
<td>.00215</td>
</tr>
<tr>
<td>. 001</td>
<td>-.0001102</td>
<td>.00010</td>
</tr>
</tbody>
</table>

The next table gives the error in the approximation to (15) using (6) on (-9,.9) starting at -1 with \( h = .01 \).
TABLE 5.5.4

<table>
<thead>
<tr>
<th>TIME</th>
<th>ERROR IN X</th>
<th>ERROR IN Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.9</td>
<td>-0.000062</td>
<td>-0.0000185</td>
</tr>
<tr>
<td>-0.7</td>
<td>-0.0000139</td>
<td>-0.0000334</td>
</tr>
<tr>
<td>-0.5</td>
<td>-0.0000212</td>
<td>-0.0000382</td>
</tr>
<tr>
<td>-0.3</td>
<td>-0.0000340</td>
<td>-0.0000408</td>
</tr>
<tr>
<td>-0.2</td>
<td>-0.0000452</td>
<td>-0.0000407</td>
</tr>
<tr>
<td>-0.1</td>
<td>-0.0000648</td>
<td>-0.0000389</td>
</tr>
<tr>
<td>0</td>
<td>-0.001128</td>
<td>-0.0000338</td>
</tr>
<tr>
<td>0.1</td>
<td>-0.0023512</td>
<td>-0.0000000</td>
</tr>
<tr>
<td>0.2</td>
<td>-0.000512</td>
<td>0.000154</td>
</tr>
<tr>
<td>0.3</td>
<td>-0.000248</td>
<td>0.000149</td>
</tr>
<tr>
<td>0.5</td>
<td>-0.000067</td>
<td>0.000081</td>
</tr>
<tr>
<td>0.7</td>
<td>0.000015</td>
<td>-0.000027</td>
</tr>
<tr>
<td>0.9</td>
<td>0.000065</td>
<td>-0.000155</td>
</tr>
</tbody>
</table>

As noted in Section 4.6, the preceding methods require solving a possibly ill conditioned system $E_h x = b$ at each step. Since the matrix $E_h$ now varies with time, this greatly increases the amount of work in comparison with the constant coefficient case. One traditional technique used to reduce the amount of computation is to only change $B_k$ at regular intervals and not at every time step. In essence $B(t)$ is approximated by a function that is piecewise constant. However, (4.28) suggests that this can lead to difficulties. Consider again Example 4.1 with $f = te^t + e^t + t$, $g = e^t + t^2$. Let $t_k = 0$. Then if $B_k$ is held constant, in effect the scheme (6) is being used on

$$
\begin{bmatrix}
0 & 1 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
x_r \\
y_r
\end{bmatrix} +
\begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
x_r \\
y_r
\end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}.
$$

(5.21)
The solution of (20) is \( x_r = te^t - 2t, y_r = e^t + t^2 \) while the solution of (4.1) is \( x = t, y = e^t \). They agree at \( t = 0 \). Since (6) provides \( O(h^2) \) for \( x_r \) and (20) provides \( O(h^2) \) for \( x \) in (4.1), the error in using (6) on (20) for more than two steps as an estimate for (4.1) is \( O(h) \) in the \( x \) variable.

There are other methods for solving differential equations numerically than the ones we have discussed. However, the methods we have given have an important characteristic. The constraint (1b) is solved almost exactly. As a consequence, the computed values, at least for the index one case, are very close to lying on solution curves. This was an important component of the proof of Proposition 1. A somewhat similar phenomena occurred in Section 1.6. All three Riccati equations in Section 1.6 produced functions that agreed on the manifold of interest but the \( K_R \) system seemed numerically less reliable due to error buildup in a component off the manifold.

In concluding this chapter, we shall consider a numerical method which does not solve (1b) almost exactly, works for nonsingular problems, but not for some singular systems. Integrate (1.1) from \( t_{k-m} \) to \( t_k \) to get

\[
A(x_k - x_{k-m}) = \int_{t_{k-m}}^{t_k} f(t) - B(t)x(t) \, dt \tag{5.22}
\]

Suppose the integral is evaluated using Simpson's rule (other integration rules could be used). Then (22) is

\[
A(x_k - x_{k-m}) = \frac{h}{3} \left[ g(t_{k-m}) + 4g(t_{k-m+1}) + \ldots + g(t_k) \right] \tag{5.23}
\]

where \( g(t) = f(t) - B(t)x(t) \). Given \( x_{k-m}, \ldots, x_{k-1} \), (23) is a linear equation in \( x_k \) which may be uniquely solved for \( x_k \) if \( [A + \frac{h}{3} B(t_k)] \) is invertible.

This method was tried on Example 4.1, again with the solutions \( x = t, y = e^t \), starting at \( t = -1 \), taking \( m = 6 \) and the first six \( x_k \) as exact. The

145
procedure produced large error \((10^6)\) in about 20-30 iterations using steps of \(h = .1, .01, .001, \) and \(.0001\). With the smaller time step, the system remained index one during the computer run. Note that (23) does not imply that \((x_k, y_k)\) lies on a solution even for the index one case. (22) also failed to work on some constant coefficient systems.

It is probably worth noting that if a backwards Euler's is used on a problem like Example 4.1 and \(t_k\) is never exactly equal to the place when the index is greater than one, then Euler's frequently produces an approximate solution. However, since the accuracy of Euler's is low and one usually does not know ahead of time where index changes will occur, this is not to be recommended.

The reader familiar with numerical procedures has probably noted a similarity between our observations and those usually made about "stiff" differential equations. This is not surprising since singularly perturbed systems often experience increasing stiffness as the perturbation parameter \(\varepsilon\) goes to zero and a singular system results when \(\varepsilon = 0\). In particular, implicit, backwards difference schemes are often used on stiff systems since they are "stiffly stable" [83]. Also some integration schemes, such as (22), are known to sometimes give slowly damped oscillatory errors [71]. There is an extensive literature on stiff differential equations and their numerical solution. For an expository treatment of stiffness see [101]. Different interpretations of stiffness may be found in [71] or [83]. A variety of numerical procedures for stiff differential equations and singular perturbation problems are given in [83].
6 Nonlinear singular systems

1. INTRODUCTION

In general, one can say little about analytic or approximate solutions of non-linear systems without making some assumptions about the nature of the nonlinearities. At worst, the assumptions should not be unreasonably restrictive. Ideally, they should be motivated by, or make possible, applications to other problems (pure or applied).

This chapter will consider systems of the form

\[ \dot{x} + B(x) = f \]  

(1.1)

where \( A \) is a singular \( n \times n \) matrix and \( B \) is a vector-valued function of \( x \) that is possibly nonlinear. As pointed out in Chapter 3, equations in the form (1) arise as models for nonlinear circuits. Sometimes the form (1) is used to preserve sparseness, because it involves variables of interest, or makes the analysis easier. Other times (1) cannot be reduced to the state variable form, \( \dot{z} = f(z, t) \). Not all circuits of interest have a state-variable representation [27]. The state-variable representation can fail to exist for many reasons. Among them are the presence of nonmonotonic characteristic curves for nonlinear elements of the circuit, negative resistors, or controlled sources. The models of most three terminal devices, such as transistors, invariably contain controlled sources. Examples of transistor networks not having normal form state equations may be found in [27, pp. 406-407]. A simple example of a nonlinear circuit with no state equation was given in Section 3.4. Newcomb [86] has shown that many circuits
which do not admit a state variable representation may be written in the form

\[
A \dot{x} + B(x,t) = f(t). \tag{1.2}
\]

As noted in the next section, (2) may be rewritten in the form (1) by adding an additional equation. Finally, (1) may occur as a reduced order model.

Section 2 will develop some of the basic concepts, notation, and facts to be used in studying (1). Section 3 will develop the index one case. Higher index systems are studied in Sections 4 and 6.

In many singular systems, the singularity appears in a linear subsystem. This is exploited in Section 5 where applications to both circuit and control theory are presented.

Sections 2 through 5 deal with continuous solutions of (1). Section 6 considers discontinuous or jump solutions.

2. GENERAL CONCEPTS

If \( f \) is a scalar valued function of the vector variable \( x \in \mathbb{R}^n \) (or \( \mathbb{C}^n \)), then the gradient of \( f \) is \( \nabla f = [f_{x_1}, f_{x_2}, \ldots, f_{x_n}] \) where \( f_{x_1} \) denotes the partial with respect to the variable \( x_1 \). Now let \( f \) be a vector valued function of the vector variable \( x \). Say that \( y = f(x) \) where \( x \in \mathbb{R}^n \), \( y \in \mathbb{R}^m \). Then the vector gradient of \( f = [f_1, \ldots, f_m]^T \) is

\[
\nabla f = \begin{bmatrix} f_{x_1} \\ \vdots \\ f_{x_m} \end{bmatrix} \in \mathbb{R}^{m \times n}.
\]
The vector gradient is also referred to as the Jacobian matrix of \( f \) or the derivative of \( f \).

It will be supposed throughout this chapter that \( B \) in (1.1) has all needed partials. If the system is initially in the form (1.2), it may be rewritten as (1.1) as by observing that

\[
\begin{bmatrix}
A & 0 \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
\dot{x} \\
\dot{z}
\end{bmatrix} +
\begin{bmatrix}
B(x, z) \\
0
\end{bmatrix} =
\begin{bmatrix}
f(t) \\
n(t)
\end{bmatrix}, \quad z(0) = 0,
\]

(2.1)

is equivalent to (1.2) and is in the form (1.1).

If \( B \) is at least \( r \) times continuously differentiable, and \( x \in \mathbb{R}^n \), expanding \( B(x) \) about \( x_o \) gives

\[
B(x) = B(x_o) + B_1(x - x_o) + B_2 + \ldots + B_{r-1} + o(|x - x_o|^r).
\]

(2.2)

Note that \( B(x_o) \) is a constant, \( B_1 = (\nabla B)(x_o) \), the vector gradient of \( B \) evaluated at \( x_o \), and \( B_{m} \) is \( m \)-linear in \( x - x_o \) for \( m \geq 2 \). Substituting (2) into (1.1) gives that for any solution \( x \),

\[
A \dot{x} + B_1 x = G(t)
\]

(2.3)

where

\[
G(t) = f(t) + B_1 x_o - (B(x_o) + B_2 + B_3 + \ldots).
\]

(2.4)

The notation of (2), (3), (4) will be used throughout Sections 2, 3, 4.

**Definition 6.2.1** The system (1.1) is solvable at \( x_o \) if there is a \( \lambda \) so that \( \lambda A + B_1 \) is invertible at \( x_o \). If (1.1) is solvable at \( x_o \), then the index of (1.1) at \( x_o \) is \( \text{Ind}(A, B) = \text{Ind}((\lambda A + B_1)^{-1}A) = \text{Ind}(A) \).
Definition 6.2.2 $x_0$ is a consistent initial condition for (1.1) at $t_o$ if there is a differentiable solution to (1.1), defined on some interval $[t_o, t_o + \gamma]$ such that $x(t_o) = x_0$.

Discontinuous solutions are also important. They will be discussed in Section 6.

The concept of solvability no longer has the intuitive meaning it had for the linear time invariant theory.

In [8] it is shown that solutions of $A\dot{x} + Bx = f$ are uniquely determined by consistent initial conditions if and only if $\lambda A + B$ is invertible for some scalar $\lambda$. This is no longer true for (1.1).

**Example 6.2.1** Consider the system

$$\dot{x} + x^2 - y^2 = 0,$$

$$x^2 = 0.$$  \hspace{2cm} (2.5a)

The initial condition $(0,0)$ is a consistent initial condition for (5) and there is only one solution through $(0,0)$. But at $(0,0)$ the system (5) has $B_1 = 0$ so that $\lambda A + B_1$ is singular for all $\lambda$.

**Example 6.2.2** The system

$$\dot{y} + x = 0,$$

$$2y - x^2 = 0,$$ \hspace{2cm} (2.6a)

has index one unless $x_o = 0$ where the index is two. The system is solvable everywhere. At $(0,0)$ the two solutions $(0,0)$ and $(-t, t^2/2)$ cross so that solvability does not imply uniqueness.
The crossing of solutions at an index change also occurred for the linear time varying systems. The next example has more uniquely nonlinear behavior.

**Example 6.2.3** For the system

\[
\begin{bmatrix}
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
x \\
y \\
z
\end{bmatrix}
+ \begin{bmatrix}
x \\
y + xy \\
z - \frac{y^2}{2}
\end{bmatrix} = \begin{bmatrix}
-1 \\
0 \\
0
\end{bmatrix},
\]

(2.7a)

(2.7b)

(2.7c)

\[
B_1 = \begin{bmatrix}
1 & 0 & 0 \\
y & 1 + x & 0 \\
0 & -y & 1
\end{bmatrix}
\]

\(B_1\) is invertible except for \(x = -1\), and (7) is solvable except at \(x = -1\). Assume \(x \neq -1\). Then

\[
\hat{A} = \begin{bmatrix}
1 & 0 & 0 \\
y & 1 + x & 0 \\
0 & -y & 1
\end{bmatrix}^{-1}
\begin{bmatrix}
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 0 & 0
\end{bmatrix}
= \frac{1}{1 + x}
\begin{bmatrix}
0 & 1 + x & 0 \\
0 & -y & 1 \\
0 & -y^2 & y
\end{bmatrix}.
\]

\(\text{Ind}(\hat{A}) = 3\) if \(y = 0\) and \(2\) if \(y \neq 0\). Substituting (7c) into (7b), there are two possibilities. One is \(y = 0\), \(x \equiv -1\), \(z \equiv 0\). The other is

\[
x(t) \text{ arbitrary},
\]

\[
y = - \int_0^t \frac{x(s) + y_0 s}{1 + x} ds,
\]

\[
z = \frac{y^2}{2}.
\]

Note that for this example no initial conditions uniquely determine solutions. Also there are infinitely differentiable solutions such that knowledge of \(x^{(i)}(t_0)\) for \(i = 0, 1, \ldots\) still does not determine \(x(t)\). This is in contrast to the observations about regular singular systems in Section 5.
In many important examples $\lambda A + B_1$ is nonsingular. The invertibility of $\lambda A + B_1$ is also essential to the implementation of the numerical procedures to be developed. Accordingly we shall assume that $\lambda A + B_1$ is invertible for some scalar $\lambda$ throughout this chapter.

As noted in the introduction, linear time varying systems can be written as nonlinear systems of the form (1.1). Thus all the difficulties present in Chapter 5 will also apply to (1.1). The index of a linear time varying system was also defined in Chapter 5.

**Proposition 6.2.1** For the linear time varying system

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

the two definitions of index are the same.

**Proof** Equation (8) may be written as

$$\begin{bmatrix} A & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \dot{x} \\ \dot{z} \end{bmatrix} + \begin{bmatrix} B(z) & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} = \begin{bmatrix} f \\ 1 \end{bmatrix}. \tag{2.8}$$

Thus $B_1 = \begin{bmatrix} B(z) \\ 0 \\ 0 \end{bmatrix}$ and the index of (9) at $(x_0, z_0)$, for $\lambda = 0$, is

$$\text{Ind} \begin{bmatrix} (\lambda A + B_1)^{-1}A & 0 \\ 0 & \lambda \end{bmatrix} = \text{Ind}[(\lambda A + B_1)^{-1}A]$$

which is the index of (8) at time $t_0$. $\square$

It is straightforward to show that the index of (1.1) is unchanged by constant coordinate changes of the form $x = Ty$ and that Theorem 5.2.1 on the invariance of $\text{Ind}(A, B)$ still holds.

Definitions 1 and 2 can be trivially extended to systems of the form
\[ A(x) \dot{x} + B(x) = f. \]  

(2.10)

While such systems will not be considered here, they naturally arise. For an example involving a neural processing model see [23]. Circuits with nonlinear resistors, inductors, and capacitors are also in the form (10) [26]. With these extended definitions, Theorem 5.2.1 now holds for coordinate changes \( x = T(t) \bar{y} \). That is, the index being 0,1, or greater than one, is an invariant property for (10) as discussed in Section 5.4. The proof is essentially that given in Section 5.4 and will be omitted.

The system (1.1) can be written as

\[ \dot{x} = F(x,y) + f_1(t), \quad (2.11a) \]

\[ \dot{y} = H(x,y) + f_2(t), \quad (2.11b) \]

by a constant coordinate change and conversely, (11) in the form (1.1).

Both formats (1.1) and (11) will be used for purposes of analysis. However, if the problem is originally in the form (1.1) and a numerical procedure is to be used, the transition to (11) could alter sparseness and the computation of \( E, H \) could be costly.

3. INDEX ONE SYSTEMS

The proof of Proposition 5.4.2 shows, with little change, that

Proposition 6.3.1 The system (2.11) has index one at \( x_0, y_0 \) if and only if \( (\nabla_y H)(x_0, y_0) \) is invertible.

If \( (\nabla_y H)(x_0, y_0) \) is invertible, and \( H(x_0, y_0) = 0 \), then by the implicit function theorem there exists a function \( y = F(x) \) defined near \( x_0 \) such that
\( y_o = F(x_o), \quad H(x, F(x)) = 0, \) and \( F \) is as smooth as \( H \) is. Thus (2.11), near \( x_o, y_o \) is equivalent to

\[
\dot{x} = F(x, F(x)), \tag{3.1a}
\]

\[
y = F(x). \tag{3.1b}
\]

This shows that:

**Proposition 6.3.2** If (2.11) has index one at \( (x_o, y_o) \), then \( (x_o, y_o) \) is a consistent initial condition if and only if \( H(x_o, y_o) = 0 \).

Note also that the proof of Proposition 5.4.2 still applies so that (2.11) is solvable at \( x_o, y_o \) if \( \nabla H \) is invertible there. An implicit scheme applied to (2.11) is, in essence, applied to (3.1a). The theory for (3.1a) is reasonably well developed. Numerical codes for the index one case date from [53]. See also [7], [72].

Since the index one case is fairly well understood, we shall concentrate on higher index problems.

4. **HIGHER ORDER SYSTEMS**

If the index is greater than one at \( x_o, y_o \), then by Proposition 3.1, \( \nabla H(x_o, y_o) \) is singular. However, this does not mean that the system cannot be written as (3.1). Rather it means that the assumptions of the implicit function theorem are not met.

**Example 6.4.1** Let \( E(x, y) \) be continuous and consider the system

\[
\dot{x} + E(x, y) = 0, \tag{4.1a}
\]

\[
x - y^3 = 0. \tag{4.1b}
\]
Since \( E(x, x^{1/3}) \) is a continuous function of \( x \), the consistent initial conditions for \( x_0 \neq 0 \) are given by (1b), \( x_0 = y_0^3 \). At \( x_0 = 0 \), \( x \), but not necessarily \( y \), is differentiable. Also (1) is solvable at \( y_0 = 0 \) if \( E_y(0,0) \neq 0 \).

If \( \nabla H(x_0, y_0) \) is singular, but there still exists a unique function \( y = F(x) \) such that \( y_0 = F(x_0) \), then \( y_0 \) cannot be found from \( x_0 \) by a standard Newton's method. Higher order methods can still sometimes be used. See Decker and Kelley [34], [35], [36] or [59].

The situation when no such unique \( F \) exists is important in equations which model hysteresis. This will be examined more carefully in Section 6.

Return now to (2.3). Since the left side of (2.3) is linear with constant coefficients, Theorem 0.1.3 may be applied to get

**Proposition 6.4.1** Assume that at time \( t_0 = 0 \) that \( \lambda A + B_1 \) is invertible for some \( \lambda \). Assume that \( x_0 \) is a consistent initial condition at time \( t_0 \) and \( x \) is the corresponding smooth solution. Then

\[
(I - \hat{A}^D\hat{A}) \left\{ \sum_{i=0}^{k-1} (-\hat{A}_B)_{i+1} \hat{D}_B(i)(0) - x_0 \right\} = 0
\]  

(4.2)

where \( \hat{A} = (\lambda A + B_1)^{-1}A \), \( \hat{B}_1 = (\lambda A + B_1)^{-1}B_1 \), \( k = \text{Ind}(\hat{A}) \), and \( \hat{G} = (\lambda A + B_1)^{-1}G \), \( G \) given by (2.4).

Note that condition (2) involves not only \( x_0 \) but also \( x^{(1)}(0) \) for an unknown solution \( x \). Similar formula were observed to apply in (5.4.29) for the linear time varying case. If the index is less than three, (2) depends only in \( A, B, x_0, f \), while if the index is three or greater, (2) leads to a correct characterization of the consistent initial conditions whereas certain other proposed methods do not.
Note that \( G(0) = \hat{f}(0) + B_1 x_0 - B(x_0) \), and \((I - \hat{A}^D \hat{B}) B_1 B_1 = I - \hat{A}^D \hat{B} \).
Thus (2) takes the form
\[
(I - \hat{A}^D \hat{B}) B_1 B_1^D (\hat{f}(0) - AB_1 \hat{f}(0) - B(x_0)) = 0, \quad \text{if} \; k \leq 2, \tag{4.3}
\]
or
\[
(I - \hat{A}^D \hat{B}) B_1 B_1^D \left\{ \sum_{i=0}^{k-1} (-\hat{A}B_1)^i \hat{f}(i)(0) - \hat{B}(x_0) - \sum_{i=2}^{k-1} (-\hat{A}B_1)^i \sum_{j=2}^{i} \hat{B}_j \right\} = 0 \tag{4.4}
\]
for \( k \geq 2 \), where \( \hat{B}_j^{(i)} \) is the \( i \)th derivative of \( \hat{B}_j \) with respect to \( t \) evaluated at \( 0 \). Note that \( \hat{B}_j^{(i)}(0) = 0 \) if \( j > k \).

To illustrate condition (2) a particular example will be considered in some detail. This example also illustrates certain numerical problems with semi-state equations which will be discussed later.

**Example 6.4.2** Consider the system
\[
\begin{bmatrix}
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
x \dot{} \\
y \dot{} \\
z \dot{}
\end{bmatrix}
+
\begin{bmatrix}
x - 1 \\
y \\
z + y^2/2
\end{bmatrix}
= \begin{bmatrix}
0 \\
0 \\
0
\end{bmatrix}. \tag{4.5}
\]

The one obvious restraint in (5) on initial conditions is \( z_0 + y_0^2/2 = 0 \). However there are others. At a possible initial condition \((x_0, y_0, z_0)\),
\[
B_1 = \begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & y_0 & 1
\end{bmatrix}
\]
Since \( B_1 \) is invertible, take \( \lambda = 0 \), and compute \( B_1^{-1}B_1 = I \),
\[
\hat{A} = B_1^{-1} A = \begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & -y_o & 1
\end{bmatrix}
\begin{bmatrix}
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 0
\end{bmatrix} = \begin{bmatrix}
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & -y_o
\end{bmatrix}.
\] (4.6)

There are two cases depending on whether \(y_o = 0\).

If \(y_o = 0\), \(\hat{A}\) is nilpotent of index 3, so that \(\hat{A}^3 = 0\). Condition (4) becomes

\[
\begin{bmatrix}
x_o - 1 \\
y_o \\
z_o + y_o^2 / 2
\end{bmatrix} - \begin{bmatrix}
0 & 0 & 1 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{bmatrix} \begin{bmatrix}
0 \\
\dot{y}(0) \\
[y(0)]^2
\end{bmatrix} = \begin{bmatrix}
0 \\
0 \\
0
\end{bmatrix}
\]

since the third component of \(\hat{B}_2\) is \(y^2/2\). Thus the restrictions are

\[
x_o - 1 + (\dot{y}(0))^2 = 0, \quad (4.7a)
\]

\[
z_o + y_o^2 / 2 = 0. \quad (4.7b)
\]

Since \(\dot{y}(0) = 1 - x_o\) from (5), (7a) becomes \((x_o - 1) + (1 - x_o)^2 = -x_o + x_o^2 = 0\). Thus \(x_o = 0\) or \(x_o = 1\). From (7b), \(z_o = 0\). Thus for (5), the only initial conditions satisfying (4) with \(y_o = 0\) are \((0,0,0)\) and \((1,0,0)\). Since \(x = 0\), \(y = t\), \(z = -t^2/2\), and \(x = 1\), \(y = 0\), \(z = 0\) are solutions of (5), we see that both \((0,0,0)\) and \((1,0,0)\) are consistent initial conditions.

Now suppose \(y_o \neq 0\). Then \(A\) in (6) has index 2 and [21]

\[
\begin{bmatrix}
0 & 0 & 1 \\
0 & 0 & -y_o \\
0 & 0 & y_o^2
\end{bmatrix}, \quad (4.6')
\]

\[
\hat{A}^D = \frac{1}{-y_o^3} \begin{bmatrix}
0 & 0 & y_o^{-2} \\
0 & 0 & y_o^{-2} \\
0 & 0 & y_o^{-1}
\end{bmatrix}, \quad \hat{A}^D \hat{A} = \begin{bmatrix}
0 & 0 & y_o^{-2} \\
0 & 0 & y_o^{-1} \\
0 & 0 & 1
\end{bmatrix}.
\]

157
From (3) the consistency requirement is \( (I - \hat{A}^D\hat{A})(B^{-1}_1B(x(0))) = 0 \). That is,

\[
\begin{bmatrix}
1 & 0 & -y_0^{-2} \\
0 & 1 & y_0^{-1} \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & -y_0 & 1
\end{bmatrix}
\begin{bmatrix}
x_0 - 1 \\
y_0 \\
z_0 + y_0^2/2
\end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}
\]

or

\[
x_0 - 1 + y_0^{-1}(y_0) - y_0^{-2}(z_0 + y_0^2/2) = 0, \tag{4.8a}
\]

\[
y_0^{-1}(z_0 + y_0^2/2) = 0. \tag{4.8b}
\]

Since \( y_0 \neq 0 \), (8) is equivalent to

\[
z_0 + y_0^2/2 = 0, \tag{4.9a}
\]

\[
x_0 = 0. \tag{4.9b}
\]

Notice that the solution mentioned earlier, \( x = 0, y = t, z = -t^2/2 \) passes through all points of the form (9) for some time \( t_0 \). Hence (9) describes the consistent initial conditions of (4) if \( y_0 \neq 0 \).

**Alternative conditions**

Another necessary condition for consistency of initial conditions has been proposed in [86]. In this section we shall show that that criterion is equivalent to (2) if either the system is linear or the index at \( x_0 \) is less than three. An example will be given to show that the condition is, in fact, not necessary for nonlinear systems of index at least three. This alternative condition is "derived" in a manner that makes it seem fairly natural. An analysis of how it can fail to hold is instructive and sheds light on the behavior of higher index systems.
The formula in [86] may be heuristically "derived" as follows. Suppose \( x \) is a solution of (1.1). Take \( t_0 = 0 \). Let \( \bar{x} = x(\delta), \bar{x}_0 = \bar{x}(0) \). Then from (1.1)

$\lim_{\delta \to 0} A \frac{(\bar{x} - \bar{x}_0)}{\delta} + \lim_{\delta \to 0} B(\bar{x}) = \lim_{\delta \to 0} f(\delta)$.

Then

$\lim_{\delta \to 0} \frac{1}{\delta} [A(\bar{x} - \bar{x}_0) + \delta \bar{B} - \delta f] = 0,$

$\lim_{\delta \to 0} \frac{1}{\delta} [A\bar{x} - A\bar{x}_0 + \delta [B(\bar{x}_0) + B_1(\bar{x} - \bar{x}_0) + \ldots] - \delta f] = 0,$

$\lim_{\delta \to 0} \frac{1}{\delta} [(A + \delta B_1)\bar{x} - (A + \delta B_1)\bar{x}_0 + \delta [B(\bar{x}_0) + B_2 + \ldots - f]] = 0,$

$\lim_{\delta \to 0} \frac{1}{\delta} [\bar{x} - \bar{x}_0 + (A + \delta B_1)^{-1} \delta [B(\bar{x}_0) + B_2 + \ldots - f]] = 0.$

The condition in [86] is that

$\lim_{\delta \to 0} \frac{1}{\delta} [\bar{x} - \bar{x}_0 + (A + \delta B_1)^{-1} \delta [B(\bar{x}_0) + B_2 + \ldots - f]] = 0,$  \hspace{1cm} (4.10)

that is,

$\lim_{\delta \to 0} (A + \delta B_1)^{-1} \delta [B(\bar{x}_0) + B_2 + \ldots - f] = 0.$  \hspace{1cm} (4.11)

Let \( \lambda \) be such that \( \lambda A + B_1 \) is invertible. Then (11) is, in the notation of Proposition 1,

$\lim_{\delta \to 0} (\hat{A} + \delta \hat{B}_1)^{-1} \delta [\hat{B}(\hat{x}_0) + \hat{B}_2 + \ldots - \hat{f}] = 0.$  \hspace{1cm} (4.12)

There exists an invertible matrix \( T \) such that

159
\[
\hat{\mathbf{T}} \hat{\mathbf{A}}^{-1} = \begin{bmatrix} C & 0 \\ 0 & N \end{bmatrix}, \text{ so that } \hat{\mathbf{T}} \hat{\mathbf{D}} \hat{\mathbf{T}}^{-1} = \begin{bmatrix} C^{-1} & 0 \\ 0 & 0 \end{bmatrix}
\] (4.13)

where \( C \) is invertible, if present, and \( N \) is nilpotent of index \( k \). Performing the similarity in (13) on (12) and using the fact that \( \hat{\lambda} \hat{\mathbf{A}} + \hat{\mathbf{B}}_1 = I \), (2) may be written as

\[
\lim_{\delta \to 0} \begin{bmatrix} (C + \delta E_1)^{-1} & 0 \\ 0 & (N + \delta E_2)^{-1} \end{bmatrix} \delta \begin{bmatrix} \hat{\mathbf{B}}_1^{[1]}(x_0) + \hat{\mathbf{B}}_2^{[1]} + \ldots + \hat{\mathbf{f}}^{[1]} \\ \hat{\mathbf{B}}_2^{[2]}(x_0) + \hat{\mathbf{B}}_2^{[2]} + \ldots + \hat{\mathbf{f}}^{[2]} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\]

where the \([ \ ]\) superscript denotes components after the similarity. Thus, for example, \( \hat{\mathbf{B}}_2^{[2]}(x_0) = \begin{bmatrix} \hat{\mathbf{B}}_2^{[1]}(x_0) \\ \hat{\mathbf{B}}_2^{[2]}(x_0) \end{bmatrix} \). But \( \delta (C + \delta E_1)^{-1} \to 0 \) as \( \delta \to 0 \).

Thus the only restriction imposed by (12) is that

\[
\lim_{\delta \to 0} (N + \delta(I - \lambda N))^{-1} \delta \begin{bmatrix} \hat{\mathbf{B}}_2^{[2]}(x_0) + \hat{\mathbf{B}}_2^{[2]} + \ldots + \hat{\mathbf{f}}^{[2]} \end{bmatrix} = 0
\]

or

\[
\lim_{\delta \to 0} (I + \delta^{-1}M)^{-1} \begin{bmatrix} \hat{\mathbf{B}}_2^{[2]}(x_0) + \hat{\mathbf{B}}_2^{[2]} + \ldots + \hat{\mathbf{f}}^{[2]} \end{bmatrix} = 0 \quad (4.14)
\]

where \( M = (I - \lambda N)^{-1}N \).

If \( \mathbf{B}(\mathbf{x}) \) is linear, then (10) is in essence a backwards Euler difference approximation and it was shown in Chapter 4 that the solution of this difference approximation to (1.1) converges to a solution of the differential equation (1.1).

If \( \mathbf{B} \) is nonlinear but the index is less than three, then (14) becomes

\[
\lim_{\delta \to 0} (I + \delta^{-1}M)(\hat{\mathbf{B}}_2^{[2]}(x_0) + \hat{\mathbf{B}}_2^{[2]} + \ldots + \hat{\mathbf{f}}^{[2]}) = 0
\]

(4.15)
But $\delta^{-1}\tilde{MB}^{-1} \rightarrow 0$ if $\tau \geq 2$. Hence (15) is

$$\lim_{\delta \to 0} \left[ \tilde{B}^{-1}(2)(x_o) - \tilde{f}^{-1}(2) - \delta^{-1}\tilde{MB}^{-1}(2)(x_o) + \delta^{-1}\tilde{MF}^{-1}(2) \right] = 0. \quad (4.16)$$

Thus $\tilde{MB}^{-1}(2)(x_o) = \tilde{MF}^{-1}(2)(0)$ and (16) implies that

$$\tilde{B}^{-1}(2)(x_o) - \tilde{f}^{-1}(2)(0) + \tilde{MF}^{-1}(2)(0) = 0$$

which is exactly (3).

Consider again Example 2 for $x_o = 0$, $y_o = 0$, $z_o = 0$ which was shown to be a consistent initial condition. Take $\lambda = 0$, so that $B_1 = \hat{B}_1 = I$ and $A = \hat{A}$. Thus (14) is

$$\lim_{\delta \to 0} \begin{bmatrix} 1 & \delta^{-1} & 0 \\ 0 & 1 & \delta^{-1} \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ y^2/2 \end{bmatrix} = 0$$

or

$$\lim_{\delta \to 0} \begin{bmatrix} 1 & -\delta^{-1} & \delta^{-2} \\ 0 & 1 & -\delta^{-1} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 0 \\ y^2/2 \end{bmatrix} = 0. \quad (4.17)$$

The first component of (17) gives

$$-1 + \frac{(y(0))^2}{2} = 0 \quad (4.18)$$

since $\lim_{\delta \to 0} y/\delta = \lim_{\delta \to 0} \frac{y - y_o}{\delta} = y(0)$. But $\dot{y}(0) = 1$ from (5) so that (18) is impossible. Thus (18) is not a necessary condition. The other consistent initial condition (1,0,0) does satisfy (11).
The expression appearing in (10) is of the type often used to motivate numerical procedures. Truncated series are then used as approximations. For example, (10) would suggest a possible iteration,

$$
\omega_{k+1} = \omega_k - (A + \delta B_1(\omega_k))^{-1}\delta[R(\omega_k) - f_{k+1}]
$$

(4.19)

for an explicit scheme with step size \( \delta \). This algorithm will be discussed later in this chapter.

However, the previous section shows that (19) need not always converge or even approximate solutions. For Example 2, (19) gives exact values if \( \omega_0 = (1,0,0) \). However, if \( \omega_0 = (0,0,0) \), then (19) gives that \( \omega_k = (1,0,0) \) for \( k \geq 1 \). Thus instead of approximating the solution to (5) with initial condition \((0,0,0)\) the scheme "jumps over" and approximates the solution with initial condition \((1,0,0)\). On the other hand, if \( \omega_0 \) lies on the solution \((0, t, -t^2/2)\) for some \( t \) significantly different from zero, then (19) approximates the solution as long as \( \omega_k \) is not close to \((0,0,0)\).

What happens near \((0,0,0)\) depends on the error that has accumulated in \( \omega_k \). If \( \omega_k \) is close to \((0,0,0)\), then (19) will usually drift over to the solution \((1,0,0)\). For example, with \( \omega_0 = (0,-1,-5) \), (this is \( t = -1 \)), with a step of .05, (19) approximates \((0, t, -t^2/2)\) until around \( t = -.5 \) \((k = 10)\). By \( t = .25 \) \((k = 25)\), the \( \omega_k \) are approximating \((1,0,0)\) to six places.

In Chapter 5 we saw that Euler's could fail to converge for a linear time varying system. This example shows that it can jump from one solution to another. As with the linear examples a second order backward difference scheme with fixed time step works well on Example 2 on time intervals including \( t = 0 \).
If (1.1) is to be solved by a numerical scheme involving backwards differences, then the next time step will be given by an equation of the form

\[ \alpha A x_{k+1} + \beta h B(x_{k+1}) = y \]  

(4.20)

where \( \alpha, \beta \) are constants depending on the difference scheme, \( h \) is the step size, and \( y \) depends on previous values of \( x \) and \( f \). Thus the equation to be solved is

\[ \alpha A u + \beta h B(u) = y. \]  

(4.21)

But

\[ \nabla_{u} (\alpha A u + \beta h B(u)) = \alpha A + \beta h B_{1}, \]  

(4.22)

which is invertible for most \( h \) by the assumption of solvability. If (22) is invertible, the implicit function theorem guarantees a solution of (21). The invertibility of (22) also means that Newton's method would be available to apply, at least in principle, to (20).

Taking \( x_{k} \) as the first estimate of \( x_{k+1} \) and applying Newton's method to (20) gives the estimate

\[ x_{k+1} = x_{k} - (\alpha A + \beta h B_{1}(x_{k}))^{-1}(\alpha A x_{k} + \beta h B(x_{k}) - y) \]  

(4.23)

For a backwards Euler, (20) becomes

\[ x_{k+1} = x_{k} - (A + h B_{1}(x_{k}))^{-1}(Ax_{k} + h B(x_{k}) - [Ax_{k} + hf_{k+1}]) \]

\[ = x_{k} - (A + h B_{1}(x_{k}))^{-1}(h B(x_{k}) - hf_{k+1}). \]  

(4.24)

Formula (24) is precisely (19).

The preceding suggests that a viable numerical strategy for semistate equations would be to carry though the time steps some indication of
whether one is near an index change. One such indicator would be the condition number of $G_{22}$ in (5.1b) for linear systems or the condition number of $\nabla H$ in (2.11) for nonlinear systems. As long as the index is one and the matrix is not ill conditioned, use a standard method such as those of Gear [53], Lininger, [72]. If the matrix $G_{22}$ or $\nabla H$ becomes "sufficiently" ill conditioned, where sufficiently would be numerically defined, then use a fixed time step second order or higher implicit scheme. Once the matrix is again sufficiently "well" conditioned a standard code can be returned to.

5. LINEAR SUBSYSTEMS

In many applications the singularity of

$$Ax + B(x) = f$$

(5.1)

may be viewed as being contained in a singular but linear subsystem. This section will discuss how this fact may be exploited.

In order for our procedure to work, the following assumptions will be made. Their significance will be discussed after the procedure is explained.

(i) $A$ is constant and square, say $n \times n$.

(ii) $B$ is twice differentiable.

There is a constant projection matrix $P$, with the properties

(iii.1) $PA = AP$

(iii.2) $PB(x) = B_o P x + B_o ((I - P)x)$, where $B_o$ is an $n \times n$ matrix, $B_o$ is a (possibly nonlinear) vector valued function of $(I - P)x$.

(iii.3) There is a scalar $\gamma$ such that rank $(\gamma P A P + P B_o P) = \text{rank}(P)$.

(iii.4) $(I - P)B(x) = B_o (P x) + B_o ((I - P)x)$ where $B_o$ is a linear function of $P x$ and $B_o$ may be nonlinear.
Now consider (1). The existence of $P$ satisfying assumptions (iii.1), (iii.2) says that there is a constant change of variables $x = Qz$ so that (1) becomes

$$
\begin{bmatrix}
A_1 & 0 \\
0 & A_2
\end{bmatrix}
\begin{bmatrix}
\dot{z}_1 \\
\dot{z}_2
\end{bmatrix}
+ \begin{bmatrix}
B_{11} & B_{12} \\
B_{21} & B_{22}
\end{bmatrix}
\begin{bmatrix}
z_1 \\
z_2
\end{bmatrix}
+ \begin{bmatrix}
B_1(z_2) \\
B_2(z_1, z_2)
\end{bmatrix}
= \begin{bmatrix}
f_1 \\
f_2
\end{bmatrix},
$$

(5.2)

with $B_1(0) = 0$, $B_2(0, z_2) = 0$, $B_2(z_1, 0) = 0$.

To see this, let $P$ be a projection so that (iii.1)-(iii.3) holds. This $B_0$ is not the same as that of Section 4. If rank $P = r$, let the first $r$ columns of $Q$ be a basis for the range of $P$, the last $n - r$ columns of $Q$ a basis for the range of $I - P$ (nullspace of $P$). Then

$$
Q^{-1}PQ = \begin{bmatrix}
I_r & 0 \\
0 & 0
\end{bmatrix},
Q^{-1}AQ = \begin{bmatrix}
A_1 & 0 \\
0 & A_2
\end{bmatrix}.
$$

Let $x = Qz$, and $z = \begin{bmatrix}
z_1 \\
z_2
\end{bmatrix}$ where $z_1$ is an $r$ vector. Note that $Px = PQz = Q \begin{bmatrix}
z_1 \\
0
\end{bmatrix}$ and

$$(I - P)x = (I - P)Qz = \begin{bmatrix}
0 \\
z_1
\end{bmatrix}.
$$

Now multiplying (1) on the left by $Q^{-1}$ and applying (iii.2), (iii.3) gives (2). Note that $\begin{bmatrix}
f_1 \\
f_2
\end{bmatrix} = Q^{-1}f$.

The first equation of (2) is

$$
A_1\dot{z}_1 + B_{11}z_1 = f_1 - B_1(z_2) - B_{12}z_2.
$$

(5.3)

By assumption (iii.3), $\gamma A_1 + B_{11}$ is invertible for some $\gamma$. Using Chapter 1, (3) is solved for $z_1$ in terms of $z_2$ as

$$
z_1(t) = \int_0^t e^{A_1(s-t)}\bar{D}_A(s)ds + e^{A_1(0)} + \sum_{i=1}^{k-1} [\le(AB)^i B^iA_1](0)
$$

(5.4)
where $\hat{A} = (\gamma A_1 + B_{11})^{-1} A_1$, $\hat{B} = (\gamma A_1 + B_{11})^{-1} B_{11}$,
\[ \hat{h} = (\gamma A_1 + B_{11})^{-1} (\check{f}_1 - B_1(z_2) - B_{12}z_2), \]
$k = \text{Ind}(\hat{A})$.

The second equation of (4) is
\[ A_2 \ddot{z}_2 + B_{21} \dot{z}_1 + B_{22} z_2 + B_2(z_1, z_2) = \check{f}_2, \] (5.5)
or, under assumption (iii.4),
\[ A_2 \ddot{z}_2 + B_{21} \dot{z}_1 + B_{22} \dot{z}_2 + B_2(z_2) = \check{f}_2. \] (5.6)

If differentiable solutions are sought, (6) can be replaced by
\[ A_2 \dddot{z}_2 + B_{21} \ddot{z}_1 + B_{22} \dot{z}_2 + B_2(z_2) \ddot{z}_2 = \check{f}_2. \] (5.7)

Substituting the solution for $z_1$ from (4) into (6) gives a nonlinear integro-differential equation solely in terms of $z_2$. In general, this equation can be quite hard to solve. However, as will be shown later, in many cases
\[ \text{Ind}(\hat{A}) = 1, \] (5.8)
and
\[ B_{21} \hat{A}^D \check{B} = 0. \] (5.9)

If (9) holds, (7) becomes
\[
B_{21} [\hat{A}^D \hat{h}(t) + (1 - \hat{A}^A) \sum_{i=0}^{k-1} [\hat{A}^D_i \hat{A}^D_{i+1}(t)] + A_2 z_2 + B_{22} \ddot{z}_2 + B_2(z_2) \dot{z}_2 = \check{f}_2. \] (5.10)

which is a nonlinear differential equation in $z_2$.

Under these additional assumptions (8), (9), if either $A_2 = 0$ and
\[ B_{22} + B_{22}(z_2(0)) + B_{21}(I - \hat{A}A^D)^{-1}B_1^{-1}(B_1(z_2(0)) - B_{12}) \]  

(5.11)

is invertible or \( A_2 \) is invertible, there will be at least local solutions at \( z_2(0) \). Here \( \frac{d(B_1(z_2))}{dt} = B_1(z_2)z_2'(t) \). That is, \( B_1(z_2) = \frac{v}{z_2}B_1(z_2) \).

As a simple example, consider the circuit (3.4.5) whose semi-state equations are given by (3.4.6). Let \( P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \). Then (3.4.6) may be written as a singular linear system,

\[ \begin{bmatrix} C & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -x_3 \\ u \end{bmatrix}, \]  

(5.12a)

which is in the form (3), and the nonlinear equation (algebraic in this example)

\[ x_2 - 3x_3 + x_3^3 = 0. \]  

(5.12b)

Now \( \gamma \begin{bmatrix} C & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} \) is invertible with \( \gamma = 1 \). Multiplying (12a) by

\[ \begin{bmatrix} C & 0 \\ 1 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} C^{-1} & 0 \\ -C^{-1} & 1 \end{bmatrix} \]

gives

\[ \begin{bmatrix} 1 & 0 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -C^{-1}x_3 \\ C^{-1}x_3 + u \end{bmatrix}. \]  

(5.13)

One may use (4) or observe directly that

\[ x_1 = -\int_0^t C^{-1}x_3 \, dt + x_1(0), \quad x_2 = C^{-1}x_3 + u = \int_0^t C^{-1}x_3 \, dt - x_1(0). \]  

(5.14)

Differentiating (12b) gives
\[ \dot{x}_2 - 3\dot{x}_3 + 3x_3^2 \dot{x}_3 = 0, \quad (5.15) \]

which is (7). Since \( \dot{x}_2 = C^{-1}\dot{x}_3 + \dot{u} + C^{-1}x_3 \) from (14), (15) is

\[ (3x_3^2 - 3 + C^{-1})\dot{x}_3 = -C^{-1}x_3 - \dot{u}, \quad (5.16) \]

which is the differential equation (10). Depending on \( u \) and \( C \), equation (16) has a unique solution for most \( x_3(0) \).

It is instructive to consider a slight modification of (3.4.5). Replace the equation \( x_2 - 3x_3 + x_3^3 = 0 \) by \( x_1 - 3x_3 + x_3^3 = 0 \). The differential equation (10) is now

\[ (3x_3^2 - 3)x_3 = C^{-1}x_3 \quad (5.17) \]

Provided \( \frac{x_3(0)^2}{2} - \ln(x_3(0)) \neq 1/2 \), that is \( x_3(0) \neq \pm 1 \), this equation has a unique solution on some interval \([0, t_0^*)\). Then the manifold of consistent initial conditions is

\[ x^T(0) = [3\alpha + \alpha^3, C^{-1}\alpha + u(0) - 3\alpha + \alpha^3, \alpha], \quad (5.18) \]

Since the behavior is similar for \( \alpha = 1 \) or \( \alpha = -1 \) we shall consider \( \alpha = 1 \). If \( x_3(0) = 1 \), and \( x \) is right differentiable, then (17) implies that \( x_3(0) = 0 \) which is a contradiction. Thus \( x_3(0) = 1 \) does not admit a solution for which \( x_3 \) is right differentiable at zero. In fact there are two continuous solutions to (17), differentiable for \((0, t_0^*)\) satisfying (18) with \( \alpha = 1 \). To see this note that the implicit solution of (17) for \( x_3(0) \neq 0 \), is

\[ \frac{x_3^2}{2} = \ln|x_3| = \frac{t}{3C} + \frac{x_3^2(0)}{2} - \ln|x_3(0)|. \quad (5.19) \]
If $x_3(0) = 1$, then for $t \geq 0$, there are two solutions $x_3^+, x_3^-$ of (17) with $x_3^+(0) = x_3^-(0) = 1$. In fact $\lim_{t \to \infty} x_3^+(t) = +\infty$, and $\lim_{t \to \infty} x_3^-(t) = 0$ and both $x_3^+, x_3^-$ are monotonic. Given $x_3^+, x_3^-$ define $x_1^+, x_1^-, x_2^+, x_2^-$ by (14).

If $x_3(0) > 0$ and $x_3(0) \neq 1$, then $\frac{x_3^2}{2} - \varepsilon x_3 |x_3| > 0$ and there is a unique $x_3$ (a translate of either $x_3^+$ or $x_3^-$ described above. Since these solutions are monotonic and move away from 1 we see solutions for this example exist for all $t \geq 0$ if $x(0) \neq 1$.

This example suggests that there is something intrinsically important about the nonlinear differential equation (7) in addition to the equation (5) which may be of an algebraic nature.

In order to carry out the procedure discussed in this section a variety of assumptions were made. Of course, for a particular example, this approach might work in modified form even if many of these assumptions did not hold.

The advantage of this approach is that it reduces the nonlinear semi-state equations to a nonlinear system of generally much smaller dimension. Often these reduced equations are not singular in that they can be written in the form $x^{(k)} = f(x, \ldots, x^{(k-1)})$. The ability to handle the singular nature of the equations via well understood linear theory is a definite advantage.

In some sense the largest rank of a $P$ which satisfies (iii) is a measure of how nonlinear the semi-state equation actually is.

Assumption (iii) is a technical assumption. There may not exist subsystems of this form even for linear equations. However, (iii) frequently is satisfied. Intuitively, (iii) says that it is possible to define a set of semi-state variables, so that the first $r = \text{rank } P$ equations are not
only linear in the first \( r \) variables, but can be solved for in terms of the remaining \( n - r \) variables.

Assumption (9) greatly simplifies the reduction process. To see that (9) admits a reasonably large class of systems, consider a circuit made up of linear capacitors, current controlled resistors and independent voltage sources. Let \( i_R, v_R, i_C, v_C, \varphi_C \) be the vectors of currents, voltages, and charges for the resistors and capacitors. Let \( E \) be the vector of independent voltages. The equations relating these variables are of the form

\[
\begin{align*}
A_R \dot{v}_R + A_C \dot{v}_C &= E = \begin{bmatrix} E \\ 0 \end{bmatrix} \\
B_R i_R + B_C i_C &= 0,
\end{align*}
\tag{5.20}
\]

\[
v_R = f(i_R),
\]

\[
v_C = C q_C, \text{ } C \text{ diagonal (or positive definite)}.
\]

Since \( \dot{i}_C = \dot{q}_C \) we get that \( i_C = C^{-1} \dot{v}_C \). Taking \( v_C, v_R, i_R \) as the semi-state variables, the semi-state equations are

\[
B_C C^{-1} \dot{v}_C + B_R i_R = 0,
\]

\[
A_C \dot{v}_C + A_R \dot{v}_R = E,
\]

\[
v_R = f(i_R),
\]

or equivalently

\[
\begin{bmatrix}
B C^{-1} & 0 & 0 \\
0 & 0 & B_R \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
\dot{v}_C \\
\dot{v}_R \\
\dot{i}_R
\end{bmatrix}
+ \begin{bmatrix}
A_C & A_R & 0 \\
0 & -I & 0 \\
0 & f(i_R) & 0
\end{bmatrix}
\begin{bmatrix}
v_C \\
v_R \\
i_R
\end{bmatrix}
= \begin{bmatrix} 0 \\ 0 \\ E \end{bmatrix}.
\]

A singular linear subsystem is
\[
\begin{bmatrix}
B_C^{-1} & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
\dot{v}_c \\
\dot{v}_R
\end{bmatrix}
+ 
\begin{bmatrix}
0 & 0 \\
A_C & A_R
\end{bmatrix}
\begin{bmatrix}
v_c \\
v_R
\end{bmatrix}
= 
\begin{bmatrix}
-B_{R_i} \\
0
\end{bmatrix}
\quad \text{(5.21)}
\]

If \( B_C \) and \( A_R \) are invertible, the approach of this section works. Condition (9) is

\[
[0, I] \left( \begin{bmatrix} B_C^{-1} & 0 \\ A_C & A_R \end{bmatrix}^{-1} \begin{bmatrix} B_C^{-1} & 0 \\ 0 & 0 \end{bmatrix} \right) D \begin{bmatrix} B_C^{-1} & 0 \\ A_C & A_R \end{bmatrix}^{-1} [0, 0] = [0, 0].
\quad \text{(5.22)}
\]

Letting \( X \) denote a possibly non-zero entry, (22) is

\[
[0, I] \left( \begin{bmatrix} B_C^{-1} & 0 \\ A_C & A_R \end{bmatrix}^{-1} \begin{bmatrix} B_C^{-1} & 0 \\ 0 & 0 \end{bmatrix} \right) D \begin{bmatrix} B_C^{-1} & 0 \\ A_C & A_R \end{bmatrix}^{-1} [0, 0] = [0, I] \begin{bmatrix} X & 0 \\ X & X \end{bmatrix} D \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = [0, I] \begin{bmatrix} X & 0 \\ X & X \end{bmatrix} = [0, I] \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = [0, 0]
\]
as desired.

Of course, if some of the resistors are linear, a larger subsystem then (22) could be used.

Control problems

Even if the underlying process, or circuit, possesses state equations, the solution of semi-state equation can become important in optimal control problems. As an example, consider the cheap control problem with process

\[
\dot{x} = f(x, u), \quad x(t_0) = x_0, \quad x(t_1) = x_1,
\quad \text{(5.23a)}
\]

and cost functional
\[ J[x, u] = \frac{1}{2} \int_{t_0}^{t_1} (x^T Q x + u^T R u) \, dt \]  

(5.23b)

where \( Q, R \) are constant positive semi-definite matrices. Both \( Q, R \) are allowed to be singular. The case when (23a) is linear has been studied in [8], or [9], [21]. This is a cheap control problem in the sense that certain controls, namely those for which \( R u = 0 \), do not contribute to the cost. (23b) may occur as the reduced order model in a singularly perturbed optimal control problem.

As noted in the previous section, often only some state variables enter in a nonlinear manner. Suppose then that (23) takes the form

\[
\begin{bmatrix}
\dot{x}_1 \\
\dot{x}_2
\end{bmatrix} =
\begin{bmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix} +
\begin{bmatrix}
B_1 \\
B_2
\end{bmatrix} u +
\begin{bmatrix}
f_1(x_2) \\
f_2(x_2)
\end{bmatrix},
\]

(5.24)

\[ J[x, u] = \frac{1}{2} \int_{t_0}^{t_1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix}
Q_{11} & Q_{12} \\
Q_{12} & Q_{22}
\end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + u^T R u \, dt. \]  

(5.25)

Let the vector \( \lambda \) be the costate variable. Then minimizing the Hamiltonian \( H = \frac{1}{2} x^T Q x + \frac{1}{2} u^T R u + \lambda^T R \tilde{f}(x, u) \) gives [1, p. 293],

\[
\begin{bmatrix}
\dot{\lambda}_1 \\
\dot{\lambda}_2
\end{bmatrix} =
\begin{bmatrix}
Q_{11} & Q_{12} \\
Q_{12} & Q_{22}
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix} +
\begin{bmatrix}
0 \\
0
\end{bmatrix} x_2 \begin{bmatrix} f_1(x_2) \\ f_2(x_2) \end{bmatrix}^T
\begin{bmatrix}
\lambda_1 \\
\lambda_2
\end{bmatrix} -
\begin{bmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{bmatrix}^T
\begin{bmatrix}
\lambda_1 \\
\lambda_2
\end{bmatrix},
\]

(5.26)

and

\[ R u + [B_1, B_2]^T \begin{bmatrix}
\lambda_1 \\
\lambda_2
\end{bmatrix} = 0. \]  

(5.27)

Combining (24), (26), and (27) gives

172
\[
\begin{bmatrix}
I & 0 & 0 & 0 \\
0 & I & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & I
\end{bmatrix}
\begin{bmatrix}
x_1 \\
\dot{\lambda}_1 \\
\dot{x}_2 \\
\dot{\lambda}_2
\end{bmatrix}
+ \begin{bmatrix}
-A_{11} & 0 & -B_1 \\
Q_{11} & A_{11}^T & 0 \\
0 & B_1^T & R \\
0 & 0 & B_2^T
\end{bmatrix}
\begin{bmatrix}
x_1 \\
\lambda_1 \\
x_2 \\
\lambda_2
\end{bmatrix}
= \begin{bmatrix}
-f_1(x_2) \\
0 \\
0 \\
-f_2(x_2)
\end{bmatrix}
\tag{5.28}
\]

where \( H_1 = A_{12}^T + \gamma \frac{f_1}{x_2} \). Now provided

\[
\begin{bmatrix}
\gamma I - A_{11} & 0 & -B_1 \\
Q_{11} & \gamma I + A_{11}^T & 0 \\
0 & B_1^T & R
\end{bmatrix}
\] is invertible for some scalar \( \gamma \) the first three equations of (28) can be solved for \( x_1, \lambda_1, u \) in terms of \( x_2, \lambda_2 \). The details may be found in [8, pp. 56-67] or [21, pp. 187-197]. Substituting this solution into the last two equations, again gives an equation just in \( x_2, \lambda_2 \). The advantage of this procedure is the following. If \( \mathbf{x} \) is \( m \)-dimensional, \( u \) is \( m \)-dimensional and \( r \) components of \( \mathbf{x} \) appear nonlinearily, (28) is a \( 2n + m \) dimensional singular system whereas, the reduced problem is, in general, not necessarily singular, and is of smaller order.

General useful conditions for the invertibility of (29) are difficult to come by. We note in passing that in the case of free controls, \( R = 0 \), a sufficient condition is the invertibility of \( B_1^T Q_{11} B_1 \). We conclude this section with a simple example. Take as the process,

\[
\begin{align*}
\dot{x}_1 &= x_1 + x_2 + u, \\
\dot{x}_2 &= x_1 - (x_2)^3 + 2u,
\end{align*}
\tag{5.30}
\]
where \( u \) is a scalar control. Let \( R = 0 \), and \( U = 1 \), so that controls are completely free. Then (28) is

\[
\begin{bmatrix}
1 & 0 & 0 & 0 & \dot{x}_1 \\
0 & 1 & 0 & 0 & \dot{x}_2 \\
0 & 0 & 0 & 1 & \ddot{\lambda}_1 \\
0 & 0 & 0 & 1 & \ddot{\lambda}_2
\end{bmatrix} + \begin{bmatrix}
-1 & 0 & -1 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 2 \\
0 & 0 & 0 & 1 & 3(x_2)^2
\end{bmatrix} \begin{bmatrix}
\dot{x}_1 \\
\dot{x}_2 \\
\ddot{\lambda}_1 \\
\ddot{\lambda}_2
\end{bmatrix} = \begin{bmatrix}
x_1 \\
x_2 \\
\lambda_1 \\
\lambda_2
\end{bmatrix}
\]

(5.31)

The singular subsystem is

\[
\begin{bmatrix}
1 & 0 & 0 & \dot{x}_1 \\
0 & 1 & 0 & \dot{x}_2 \\
0 & 0 & 1 & \ddot{\lambda}_1 \\
0 & 0 & 1 & \ddot{\lambda}_2
\end{bmatrix} + \begin{bmatrix}
-1 & 0 & -1 & 0 \\
1 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix} \begin{bmatrix}
\dot{x}_1 \\
\dot{x}_2 \\
\ddot{\lambda}_1 \\
\ddot{\lambda}_2
\end{bmatrix} = \begin{bmatrix}
0 \\
0 \\
-2\lambda_1 \\
-2\lambda_2
\end{bmatrix}
\]

(5.32)

Proceeding as in [8], [21], take \( \gamma = 0 \). Multiplying (32) by

\[
\begin{bmatrix}
-1 & 0 & -1 \\
1 & 1 & 0 \\
0 & 1 & 0
\end{bmatrix}^{-1} = \begin{bmatrix}
0 & 1 & -1 \\
0 & 0 & 1 \\
-1 & -1 & 1
\end{bmatrix}
\]

gives

\[
\begin{bmatrix}
0 & 1 & 0 & \dot{x}_1 \\
0 & 0 & 0 & \dot{\lambda}_1 \\
-1 & -1 & 0 & \ddot{u}
\end{bmatrix} + \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 1
\end{bmatrix} \begin{bmatrix}
\dot{x}_1 \\
\ddot{\lambda}_1 \\
\dddot{u}
\end{bmatrix} = \begin{bmatrix}
2\lambda_1 \\
-2\lambda_2 \\
-2\lambda_2
\end{bmatrix}
\]

(5.33)

The solution of (33) may be found from Theorem 0.3.1 as

\[
\lambda_1 = -2\lambda_2,
\]

\[
x_1 = 2\lambda_2 + 2\dot{\lambda}_2,
\]

\[
u_1 = -2\lambda_2 + 2\dot{\lambda}_2 + 2\ddot{\lambda}_2 - 2\dddot{\lambda}_2 = -2\lambda_2 + 2\dddot{\lambda}_2.
\]

(5.34)
Substituting into (31) now gives
\[ \dot{x}_2 - 2\lambda_2 - 2\dot{\lambda}_2 - 2\lambda_2 = (x_2)^3, \quad (5.35) \]
\[ \dot{\lambda}_2 + x_2 + 3(x_2)^2\lambda_2 = 0. \]

Of course, since this is a singular control problem, not all values of 
\( x_1(t_0), x_2(t_0), x_1(t_1), x_2(t_1) \) will admit functional solutions to (31). If
\( x_1(t_1), x_2(t_1) \) are specified and (35) has a solution \( x_2, \lambda_2 \), such that
\( x_1(t_0), x_2(t_1) \) are as required, then (33) will be consistent if
\[ x_1(t_0) = -2\lambda_2(t_0) + 2\dot{\lambda}_2(t_0), \quad x_1(t_1) = -2\lambda_2(t_1) + 2\dot{\lambda}_2(t_1). \]
\[ \lambda_1(t_0) = -2\lambda_2(t_0), \quad \lambda_1(t_1) = -2\lambda_2(t_1). \]

The system (35), or its equivalent form,
\[ \dot{x}_2 = (x_2)^3 + 4\lambda_2^2 + 2x_2 + 6(x_2)^2\lambda_2, \quad x_2(t_0), x_2(t_1) \]
\[ \dot{\lambda}_2 = -x_2 - (3x_2)^2\lambda_2 \]
is a classical nonlinear boundary value problem on which much has been 
written. A good summary of techniques may be found in [27].

Numerical Comment

The format (2) can also be exploited numerically. To simplify notation,
assume that the system is in the form
\[ E\dot{z}_1 + Fz_1 + G(z_2) = 0, \quad (5.36a) \]
\[ Hz_2 + J(z_1, z_2) = 0. \quad (5.36b) \]

It will be assumed that \( E \) is singular, \( \lambda E + F \) is invertible for some scalar
\( \lambda \), and \( H \) is invertible or zero. If \( H \) is zero, the additional assumption
will be made on \( J \) that \( \nabla \ J \) is a nonsingular matrix. The idea is that different procedures can be used on (36a) and (36b). To illustrate, let \( z_1(t_o + hk) = z_1(k) \). Given \( z_1(k), z_2(k) \), get \( z_1(k+1) \) from (36a) using a backwards difference scheme. Euler's will be used for illustration. Of course, Euler's should not be used if (36a) has index greater than one. Thus for \( z_1(k+1) \):

\[
E \frac{z_1(k+1) - z_1(k)}{h} + Fz_1(k+1) = -G(z_2(k))
\]

is solved for \( z_1(k+1) \). Then given \( z_1(k+1) \), solve (36b) for \( z_2(k+1) \) by whatever procedure is deemed best. In particular if \( H \) is nonsingular, then it would be possible to use either a higher or lower order difference scheme to find \( z_2(k+1) \).

Often it is the case in (36) that \( H = 0 \). An example is a circuit with current controlled nonlinear resistors and linear capacitors as mentioned earlier in this section. In this case (36b) becomes an algebraic equation. As noted, the procedure (4.19) is the same as solving

\[
J(z_1(k+1), z_2(k+1)) = 0
\]

for \( z_2(k+1) \), given \( z_1(k+1) \), in one step of a Newton's procedure given the initial guess, \( z_2(k) \). Since supposedly \( z_2(k+1) \) is usually not too far from \( z_2(k) \), the procedure often gives fairly good results.

However, given \( z_1(k+1) \) it is possible to solve (38) more accurately for \( z_2(k+1) \). This could become especially important in problems where there exist values of \( z_1(k+1) \) for which more than one \( z_2(k+1) \) may be a solution of (38). Such problems arise naturally and tend to appear in those systems that do not admit a state variable representation and exhibit hysteresis.
Solutions may be discontinuous at these places and $\nabla_{z_2} J(z_1, z_2)$ will be singular. In dealing with these problems, it is important not only that (38) be solved fairly accurately but that subroutines be set up which will determine the correct $z_2^{(k+1)}$ given $z_1^{(k+1)}$ in the event that $\nabla_{z_2} J(z_1, z_2)$ is singular. These problems will not be considered here.

As an example, consider the system

$$\begin{align*}
\dot{x} - x - y^2 &= 0, \quad x(0) = 4, \quad (5.39a) \\
\frac{1}{2} x - y &= 0, \quad y(0) = 2. \quad (5.39b)
\end{align*}$$

Of course (39) has the exact solution $x(t) = 4e^{2t}$. However, we wish to apply (4.19) to (39) and compare it to using (37) on (39a) and then solving (39b) by a different method, in this case, exactly. Using a step size of 0.01 gives:

<table>
<thead>
<tr>
<th>TIME</th>
<th>EXACT $x_1$ VALUE</th>
<th>ERROR IN (37)</th>
<th>ERROR IN (4.19)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>4.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.5</td>
<td>10.873</td>
<td>0.000</td>
<td>0.110</td>
</tr>
<tr>
<td>1.0</td>
<td>29.556</td>
<td>0.002</td>
<td>0.603</td>
</tr>
<tr>
<td>1.5</td>
<td>80.342</td>
<td>0.008</td>
<td>2.480</td>
</tr>
<tr>
<td>2.0</td>
<td>218.392</td>
<td>0.029</td>
<td>9.007</td>
</tr>
</tbody>
</table>

Table 6.5.1

Since they both involve first order differences, (4.19) and (37) often produce comparable results. However, Table 1 does suggest that for non-linear systems of the form (36) significant gains in accuracy can sometimes be achieved by exploiting the system structure. This could be preferable, in some instances, to using higher order versions of (4.3.7) or (37) or decreasing the step size.
6. HYSTERESIS AND IMPULSES

Suppose that one has the singular system

\[ \dot{x} = G(x, y) \quad (6.1a) \]
\[ 0 = H(x, y) \quad (6.1b) \]

and that \( x(t), y(t) \) is a solution on \([t_0, t_1]\), and

\[ \lim_{t \to t_1^+} x(t) = x^-, \quad \lim_{t \to t_1^+} y(t) = y^- \quad (6.2) \]

exist. If \( \nabla_y H(x^-, y^-) \) is nonsingular, then the solution may be uniquely continued to an interval \([t_0, t_2]\) where \( t_2 > t_1 \) using the results of Section 3. Suppose, however that \( \nabla_y H(x^-, y^-) \) is singular and there is not a unique continuous function \( F \) defined on a neighborhood of \((x^-, y^-)\) such that \( y^- = F(x^-) \) and \( H(x, F(x)) = 0 \). Then it may not be at all immediately obvious where the solution should "go next."

As a simple example consider the following circuit from [26, p. 487].

![Circuit Diagram](image)

where \( v_R = i_R^3 - 3i_R \). Note that \( v_C = v_R \). For reasons to become clear shortly \( v_C \) will be used as one of the semi-state variables. The \( v-i \) characteristic is given by the solid line in Figure 1.
In order to understand how the voltage acts, observe that $v_c = \frac{1}{C} q_c$ where $q_c$ is the charge on the capacitor. Then

$$\dot{v}_c = \frac{1}{C} \dot{q}_c = \frac{1}{C} i_c = -\frac{1}{C} i_R.$$ 

The semi-state equations are thus

$$\dot{v}_c = -\frac{1}{C} i_R,$$ (6.4a)

$$0 = v_c - i_R^3 + 3i_R.$$ (6.4b)

Note that (4) is solvable everywhere and index one except at the points $b, d$ in Figure 1 where the index is two. If $(v_c(0), i_R(0))$ lies on the curve (4b) anywhere except at the points $b, d$, then the solution to (4) starting at $(v_c(0), i_R(0))$ moves along the curve (4b) as indicated by the arrows in Figure 1. The point 0 is an equilibrium point.

Suppose now that $(v_c(0), i_R(0))$ lies to the left of $b$, say at $a$. Then $(v_c(t), i_R(t)) = (-1, 2) = b$. Since $\| (\dot{v}_c(t), \dot{i}_R(t)) \| > 1 > 0$, there is a time $t_1$ such that $(v_c(t), i_R(t))$ covers the half closed arc from $a$ and up to but not including $b$ during the time period $[0, t_1]$. What should happen at
t = t₁? There is no way the solution can be extended continuously beyond b and b is not an equilibrium point.

Chapter 3 discussed how the distributional solutions for the linear time invariant problem were only solutions for inconsistent initial conditions if properly interpreted. If, in fact, the system (4) is viewed as an idealized (reduced order) model, then the definition of solution should be linked with the type of reduction process.

To motivate what follows, consider again (3). Every device has at least a little inductance. Place then a "small" inductance L into the circuit (3) to get the circuit

\[ \begin{array}{c}
+ \\
\text{C} \\
- \\
+ \\
\end{array} \quad \begin{array}{c}
\text{L} \\
\text{L} \\
\text{R} \\
\text{R} \\
\end{array} \quad \begin{array}{c}
+ \\
\text{C} \\
- \\
+ \\
\end{array} \]

(6.5)

The system equations now have a state variable form

\[ \begin{align*}
\frac{dv}{dt} &= -\frac{1}{C} i_R, \\
L \frac{di_R}{dt} &= v_c - i_R^3 + 3i_R.
\end{align*} \]  

(6.6a)

(6.6b)

The resulting trajectory is;
The trajectory in [26] is pictured as periodic but it is not. As to be expected from (6b) the transition from near b to near c is quite rapid since L is small. The graphs in Figure 3 are for i in (6) against time for three values of L with C = 1 and i(0) = 1, v(0) = -2, that is point d. All are plotted to the same scale. The time interval is $0 \leq t \leq 9.88$. 
This suggests that for the original circuit (3) when at point b in Figure 1 the solution should jump to point c. From c it moves along the curve to d where it jumps to a and then goes up along the curve to b and repeats the process. The resulting current vs. time graph, on the same scale as Figure 3, is:

![Figure 6.6.4](image)

which is discontinuous. The curve in Figure 4 is the limit, as $\varepsilon \to 0^+$, of the curves in Figure 3.

How is this to be rigorously defined for larger systems? One approach is to assume that for (1), there is a potential $V(x,y)$ bounded below and proper and that the system (1) may be written as

\begin{align}
\dot{x} &= G(x,y), \\
0 &= \nabla V(x,y).
\end{align}

(6.7a)  
(6.7b)

The system (7) is then considered to be a reduced order model of
\[ \dot{x} = G(x, y), \quad (6.8a) \]

\[ \dot{y} = -V_y y(z, y). \quad (6.8b) \]

for \( \varepsilon > 0 \).

Intuitively the idea is that if \( x^-, y^- \) are such that \( V_y V(x^-, y^-) = 0 \) but \( G(x, y) \neq 0 \) and \( \varepsilon \geq 0 \), then if at time \( t \) the solution \( (x(t), y(t)) \) of (8) is close to, but unequal to (7b), then (8a) will displace it off the surface defined by (7b). If \( \varepsilon \) is small, then the \( \dot{y} \) term dominates \( \dot{x} \) and the trajectory moves roughly parallel to the surface \( \{(x^-, y) \ y \in \mathbb{R}^m\} \). This rapid motion should not significantly increase the potential. This approach has been rigorously defined and studied, for example, by Takens [107], who dealt directly with the system (1).

The Takens approach is as follows. Let \( M(x, y) = \{(x, z) \colon z \) arbitrary, \( \dim z = \dim y\} \). Let \( (x^-, y^-) \) be the point from whence we wish to continue our solution. Then the solution can continue from (jump to) \( (x^-, y^+) \) if \( (x^-, y^+) \) is a nondegenerate local minimum of \( V \) restricted to \( M(x^-, y^-) \) and there is a continuous curve in \( M(x^-, y^-) \) going from \( (x^-, y^-) \) to \( (x^-, y^+) \) along which \( V \) is decreasing.

Note that the assumption is not that \( V_y V = -H \) but rather that \( V_y V = 0 \) and \( H = 0 \) have the same solutions. The determination of \( V \) would come out of the physics of the problem. In [107], Takens shows that a potential exists for a large class of nonlinear RC circuits.

More recently Sastry, et. al., [100] have defined discontinuous solutions of (1) using a singular perturbations approach. Their definition applies to systems in the form (1) which cannot be written as (7). Space permits only a discussion of their approach here. The interested reader is encouraged to read [100] for additional details.
Let $M = \{ (x, y) \mid H(x, y) = 0 \}$ and assume that $M$ is an $n$-dimensional manifold in $\mathbb{R}^{n+m}$. A sufficient condition for this is that $\text{rank}[\nabla_x H, \nabla_y H] = n$ at all $(x, y) \in M$. Let $M_x = \{ x \mid (x, y) \in M \text{ for some } y \}$. For (4), $M$ is the curve in Figure 1 (one-dimensional manifold) and $M_x = \mathbb{R}$.

For simplicity of exposition assume that $\nabla_x H$ is nonsingular on $M$ except at a finite number of points where $\nabla_x H = 0$. If $\nabla_x H \neq 0$ but $\det(\nabla_x H) = 0$ at $(x_o, y_o)$, then after a constant coordinate change, some of the new $y$ variables may be solved for (implicit function theorem) and a new system in the form (1) results with $\nabla_x H = 0$ for this new system. Details are straightforward and can be found in [100].

Now consider the singularly perturbed system (augmented equation in [100])

\[ \dot{x} = G(x, y) \]  \hspace{1cm} (6.9a)

\[ \varepsilon \dot{y} = H(x, y). \]  \hspace{1cm} (6.9b)

If $(x_o, y_o) \in M$, then $y_o$ is an equilibrium of

\[ \varepsilon \dot{y} = H(x_o, y), \]  \hspace{1cm} (6.10)

the "fast dynamics." Assume that for each $x_o \in M_x$ that (10) has only a finite number of equilibrium points and that (10) is completely stable. That is, every solution of (10) converges to an equilibrium point. Finally let

\[ M_a = \{ (x, y) \in M \mid \text{Rloc}(\nabla_x H(x, y)) < 0 \}. \]

Note that if $(x_o, y_o) \in M_a$, then $y_o$ is an attractor for (10). In [100], $M_a$ is referred to as the physically measurable portion of $M$. The idea is that
if \((x(t), y(t))\) is near \(M_a\) the solution to (9) is driven by the fast dynamics quickly back towards the solution of (1). That is, the behavior is somewhat insensitive to small perturbations. For (4), (6), \(M_a\) is all of the curve in Figure 1 except the portion between points b and d.

It is now assumed that a definition is sought for a jump of a solution to (1) so \(x\) will be continuous and \(y\) is piecewise continuously differentiable so that at a jump time \(t_o\)

\[
\dot{x}(t_o^-) = G(x(t_o^-), y(t_o^-)), \\
\dot{x}(t_o^+) = G(x(t_o^+), y(t_o^+)).
\]

With the previous assumptions, we now have the definition

**Definition 6.6.1** [100] Suppose that if \(\lambda \in \sigma(H_\infty(x^-, y^-))\) and \(\text{Re}\lambda = 0\), then \(\lambda = 0\). Then the solution of the system (1) is said to jump from \((x^-,-y^-) \in M\) to \((x^+, y^+) \in M\) if there is an \(\alpha > 0\) such that given any \(\delta > 0\) there is an \(\varepsilon > 0, t_o > 0\) such that for \(0 < \varepsilon < \varepsilon_o\), if \((x^-_\varepsilon, y^-_\varepsilon)\) is a point so that

\[
|x^-_\varepsilon - x_o| + |y^-_\varepsilon - y_o| < \delta,
\]

then for \(\varepsilon t_o < t < \alpha\), one has

\[
|x(t, \varepsilon) - \dot{x}(t)| + |y(t, \varepsilon) - \dot{y}(t)| < \delta
\]

where \((x(t, \varepsilon), y(t, \varepsilon))\) is the solution to (9) starting at \((x^-_\varepsilon, y^-_\varepsilon)\) at \(t = 0\) and defined on \([0, \alpha)\), \((x, y)\) is a solution of (1) starting at \((x^+, y^+).\)

Intuitively, Definition 1 says that there is a neighborhood of \((x^-, y^-)\) so that any solution of the perturbed system (9); \((\varepsilon > 0)\) starting in that
neighborhood, uniformly approximates the solution of the unperturbed system (1) starting at \((x^-_0, y^+_0)\) within a time which is \(O(\epsilon)\).

The authors of [100] then go on to show that a large class of nonlinear circuits, such as (3) can be analyzed using this approach.

Note that if (1) is solvable and \(\frac{\partial}{\partial x} (x_0, y_0) = 0\), then (1) has index two at \((x_0, y_0)\). The perturbations used by Cobb as discussed in Chapter 2 are somewhat different than those of [100]. A full synthesis of the approach of this book and that of [100] incorporating higher index systems and independent sources remains to be developed. Such a theory would enable one to analyze semi-state models for nonlinear circuits involving several operational (differential) amplifiers.

Semi-state equations have also been used to study hysteresis by Newcomb [85], [86]. These papers are not only concerned with theoretically describing hysteresis but also experimentally displaying it. The following example is drawn from one developed in [85]. Consider,

\[
\begin{align*}
\dot{x}_1 + 2x_1 - x_2 &= u, \\
\dot{x}_2 - f(x_1 - a) &= 0,
\end{align*}
\]

(6.11a)

where

\[
f(x_1 - a) = \begin{cases} 
-1 & \text{if } x_1 < a - \delta \\
1 & \text{if } x_1 > a + \delta \\
\text{is monotonically increasing for } |x_1 - a| \leq \delta
\end{cases}
\]

(6.12)
Start with \( u = -2, x_2 = -1, x_1 = -3/2 \). If \( u \) is increased slightly and \( \epsilon \) is small, \( x_1 \) will quickly increase toward \( \frac{u - x_2}{2} = \frac{u + 1}{2} \) and \( x_2 \) stays at \(-1\). Continuing to increase \( u \) slowly, \( x_2 \) will stay at \(-1\) until \( \left| \frac{u + 1}{2} - \alpha \right| < \delta \) at which time \( x_2 \) begins to increase to \( 1 \). Continuing to increase \( u \) gives the solid curve in Figure 6. If the process is reversed, then \( x_2 \) stays at \( 1 \) until \( x_1 \) nears \( \alpha \). But as long as \( x_2 = 1 \), \( x_1 \) is tracking \( \frac{u - 1}{2} \) so that the resulting trajectory is the dashed one in Figure 6. Note that as \( u \) is varied, that the \( x_2 \) variable follows a different path. This can be useful, for example in preventing bouncing of automated typing systems due to noise [69]. A circuit which exhibits the behavior illustrated in Figure 6 is given in [85]. It is built using an operational amplifier.
The preceding analysis has much in common with the earlier discussion dealing with the first example. In some analyses $z$ is set equal to zero in (11a) and (11) is analyzed on a steady state basis.

This gives

\[ 2x_1 - x_2 = u, \]  
\[ (6.13a) \]

\[ x_2 = f(x_1 - a). \]  
\[ (6.13b) \]

Substituting (13b) into (13a) gives

\[ 2x_1 - f(x_1 - a) = u \]  
\[ (6.14) \]

which has the graph
The graph in Figure 7 closely resembles the quiescent characteristic curve of a tunnel diode [26]. Note that for \( u \) between \( \alpha \) and \( \beta \) there are three possible values of \( x_1 \) that satisfy (14) and it is not clear what \( x_1 \) should be. On the other hand the fuller, or augmented equations, as noted earlier, are easy to describe when the dynamics are taken into account. Notice also that the system (11) has index one everywhere.

The full model can effect what should be called a "solution." In [64], [105] an example is given from chemistry where the same reduced order model is perturbed to a deterministic model, as we have done with the circuit (1), and to the same differential equation but with a small amount of noise added. When the two full models are solved and the perturbation is allowed to go to zero, two different solutions result. The limit of the noise model does not exhibit hysteresis which agrees with physical observations for the chemical phase transition. This analysis is also applied to (4) in [64]. This suggests that in analyzing singular systems in problems where there is more "noise," that a different definition of discontinuous or distributional solution than that given in this chapter and Chapters 1, 2 may be needed.
7 Infinite dimensional linear systems

1. INTRODUCTION

This chapter will consider the problem of extending the results of Chapters 0, I, II, and [8] to linear differential systems of the form

$$\dot{Ax} + Bx = f$$

(1.1)

or discrete systems of the form

$$Ax_{n+1} + Bx_n = f_n, \quad n \geq 0$$

(1.2)

where $x(t)$ (or $x_n$) is a vector in an infinite dimensional space. There is no one general approach that seems best for every application. Rather, there are three interrelated ways of examining (1) or (2) each of which has its advantages. The first decision to be made is what type of linear spaces are $x$, $x_n$, $f$, $f_n$ supposed to be in. Secondly, are $A$, $B$ to be interpreted as matrices or as operators on spaces with some type of topology. If $A$, $B$ are to be operators, what types of continuity properties are to be used?

In Section 2 we consider the case when $A$, $B$ are infinite matrices, and $x$, $x_n$ are infinite vectors. No norms are used. Equations (1), (2) are assumed to hold component wise and derivatives are defined component wise.

Section 3 discussed spectral and resolvent approaches. This is the most traditional section.

Section 4 discusses a geometric approach that overlaps, but is independent of, Section 3.
2. SYSTEMS WITH INFINITE MATRICES

In some applications one does not wish to impose a norm on (1.1) or (1.2). Rather they are to be viewed as a collection of equations with an infinite number of dependent variables. This is often the situation, for example, when dealing with denumerable Markov chains. In these problems the spectral methods of the next section are inappropriate.

First, it is necessary to review the basic facts about infinite matrices. Unless stated otherwise, all matrices in this section are infinite matrices with finite entries from the field of complex numbers. If $T$ is a matrix, we shall write it as $T = [T_{ij}]$, $0 \leq i,j \leq \infty$. If $T_{ij} \geq 0$ for every $i,j$ (in particular, $T_{ij}$ is real), then $T \geq 0$. $|T|$ is defined by $|T|_{ij} = |T_{ij}|$. If $|T_{ij}| < \infty$ for all $i,j$, we write $|T| < \infty$. I denotes the identity matrix, $E$ a matrix with every entry equal to one, and $e$ a column matrix of ones. $C^*$ is the conjugate transpose of the matrix $C$. One could allow $i$ or $j$ to range over the integers in $(-\infty, \infty)$ instead of $[0, \infty)$ but $[0, \infty)$ is most appropriate for the applications we have in mind. If $S_n$ is a sequence of matrices, then

$$\lim_{n \to \infty} S_n = C$$

is taken to mean $\lim_{n \to \infty} (S_n)_{ij} = C_{ij}$, $0 \leq i \leq \infty$, $0 \leq j \leq \infty$. In particular, this applies if $S_n = \sum_{k=0}^{\infty} A_k$ for a sequence of matrices $\{A_k\}$.

Matrix multiplication and addition are defined in the usual way. A product or sum is well-defined if one does not encounter any divergent series (series may conditionally converge to $+\infty$ or $-\infty$) or sums of the form $\infty - \infty$. By convention, $0 \cdot \infty = 0$. Define

$$(FR) = \{ \text{all matrices such that each row has a finite number of nonzero entries} \},$$
(FC) = \{all matrices such that each column has a finite number of non-zero entries\},

(FRC) = (FR) \cap (FC).

Matrices in (FR) are called row finite. Matrices in (FC) are called column finite.

**Proposition 7.2.1** If A, B are matrices, then

(i) A \in (FR), B \in (FR) implies that AB \in (FR), and

(ii) A \in (FC), B \in (FC) implies that AB \in (FC).

Most of the algebraic manipulations done with finite matrices also work for infinite matrices with one crucial exception. Matrix multiplication is no longer associative. A standard example [67, p. 5] is \( e^* (Ae) = 0 \), \( (e^* A)e = 1 \); where A is defined to be 1 on the diagonal, -1 on the super-diagonal and zero elsewhere. Note that A \in (FRC). We shall give an example of a Markov chain with transition matrix P with \( I - P \) a scalar multiple of this A.

The uniqueness of most types of inverse depends upon associativity. Sufficient conditions for associativity are known. The following are taken from [67].

**Proposition 7.2.2** Let A, B, C be matrices which possibly have infinite entries. Then A(BC) = (AB)C if any one of the following conditions is met:

(i) A \geq 0, B \geq 0, and C \geq 0,

(ii) \( |A| |B| |C| < \infty \),

(iii) A \in (FR) and A(BC), (AB)C are well-defined,

(iv) C \in (FR) and A(BC), (AB)C are well-defined,

(v) B has finitely many nonzero entries and A(BC), (AB)C are well-defined.
Another useful corollary from [67, p. 7] is

**Corollary 7.2.1** If $A > 0$, $B > 0$, $C > 0$, $D > 0$, and either

(i) $|ABC| < \infty$, $|AB| < \infty$ and $|BD| < \infty$,

or

(ii) $|ACD| < \infty$, $|AC| < \infty$ and $|CD| < \infty$,

then $(A(B - C))D = A((B - C)D)$.

The definitions of commuting (1)- and (1,2)- inverses for infinite matrices are, except for an associativity requirement, the same as for finite dimensional matrices.

**Definition 7.2.1** A matrix $X$, $|X| < \infty$, is a C-(2) inverse of $A$, $|A| < \infty$, if

(i) $XA = AX$, and

(ii) $X(AX) = (XA)X = A$.

**Definition 7.2.2** A matrix $X$, $|X| < \infty$, is a C-(1,2) inverse of $A$, $|A| < \infty$, if $X$ is a C-(2) inverse of $A$ and

(iii) $A(XA) = (AX)A = A$.

The next definition provides a partial ordering on the set of C-(2) inverses for a given matrix.

**Definition 7.2.3** If $X,Y$ are C-(2) inverses for $A$ and

$$X(AY) = (XA)Y = X,$$  \hspace{1cm} (2.1)

and

$$Y(AX) = (YA)X = X,$$  \hspace{1cm} (2.2)
then \( X \subseteq Y \).

It is important to note that conditions (1) and (2) are independent. Intuitively, (2) implies that the "range" of \( X \) is in the "range" of \( Y \) while (1) says that the "nullspace" of \( X \) contains the nullspace of \( Y \).

**Example 7.2.1** Let

\[
A = I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}.
\]

Then \( X, Y \) are C-(2) inverses for \( A \) and (2) holds but not (1).

**Definition 7.2.4** If \( Y \) is a C-(2) inverse for \( A \) and \( X \subseteq Y \) for all C-(2) inverses of \( A \), \( Y \) is called maximal. If there exists a unique maximal (in the ordering of Definition 3) C-(2) inverse, it will be called the Drazin inverse of \( A \) and denoted \( A^D \). If \( A^D \) exists and is a C-(1,2) inverse, it is also called the group inverse of \( A \) and denoted \( A^\# \).

Note that \( X = 0 \) is a C-(2) inverse for any \( A \) and \( 0 \subseteq Y \) for any C-(2) inverse \( Y \) of \( A \).

**Proposition 7.2.3** If \( X \subseteq Y \subseteq Z \), where \( X, Y, Z \) are C-(2) inverses for \( A \), and \( XAYAX \) and \( ZAYAX \) are associative products, then \( X \subseteq Z \).

**Proof** Let \( X, Y, Z \) be C-(2) inverses for \( A \) such that \( X \subseteq Y \) and \( Y \subseteq Z \). Assume the products \( XAYAZ \) and \( ZAYAX \) are associative. Then

\[(XA)Z = X(AYZ) = (X(AY))(AZ) = X(A(YAZ)) = X(AY) = X.\]

Similarly \( Z(AX) = X \).

194
Proposition 7.2.4 If \(X, Y\) are \(C-(2)\) inverses for \(A\) such that \(X \subseteq Y\) and \(Y \subseteq X\), then \(X = Y\).

Proof Suppose \(X, Y\) are \(C-(2)\) inverses for \(A\) such that \(X \subseteq Y\) and \(Y \subseteq X\). Then \(X = XAY = Y\).

Corollary 7.2.2 If \(A\) has a maximal \(C-(2)\) inverse, then it is unique so that \(A^D\) exists.

Proposition 7.2.5 Suppose that \(X\) is a \(C-(1,2)\) inverse of \(A\) and \(Y\) is a \(C-(2)\) inverse of \(A\). If \(X(AY)\) and \(Y(AX)\) are associative products, and \(A \in (FR)\), then \(Y \subseteq X\).

Proof Suppose \(X, Y\) are as described. Then
\[
X(AY) = (XAX)(AY) = (X^2A)(AY) = (X^2A)(YA) = X^2[A(YA)] = X^2A = X.
\]
Similarly,
\[
(YA)X = (YA)(XAX) = (AY)(XAX) = (AY)(AX^2) = [(AY)AX^2 = AX^2 = X
\]
as desired.

Note that without some assumptions on \(A\), Proposition 5 is not valid since inverses are not always unique, but they are always \(C-(1,2)\) inverses.

Proposition 7.2.6 Suppose that \(X, A \in (FR)\) and \(X\) is a \(C-(1,2)\) inverse of \(A\). Then \(X\) is the unique \(C-(1,2)\) inverse of \(A\).
Since (FR) is associative, the proof of Proposition 6 is straightforward and will be omitted.

A similar result holds if $X, A \in (FC)$.

The next example will be referred to several times in what follows.

**Example 7.2.2** Let

$$
P = \begin{bmatrix}
\frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\
0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\
0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\
0 & 0 & 0 & 0 & 0
\end{bmatrix}, \quad A = I - P,
$$

$$
X = \begin{bmatrix}
2 & 2 & 2 & 2 & 2 \\
0 & 2 & 2 & 2 & 2 \\
0 & 0 & 2 & 2 & 2 \\
0 & 0 & 0 & 2 & 2
\end{bmatrix}
$$

Then $AX = XA = I$, $P \in (PCR)$, $X \in (FC)$.

Since $(XY)^* = Y^*X^*$ is valid for infinite matrices, $A^D$ exists if and only if $(A^*)^D$ exists and $(A^D)^* = (A^*)^D$.

Note that in Example 2, $A$ has an inverse and yet $A_0 = 0$. When working with infinite matrices, the concept of an inverse is not as directly related to that of nullspaces and ranges as in the case with finite matrices. For this reason, our development is more algebraic.

**Example 7.2.3** Let $S$ be defined by $S_{ij} = 1$ if $j = i + 1$, and $S_{ij} = 0$ if $j \neq i + 1$. The only C-(2) inverse for $S$ is 0 so that $S^D = 0$. Hence $(S^*)^D = 0$.  

196
If $S$ is thought of as acting on $\Delta$, the linear space of all sequences, then $S$ is the "unilateral" shift. Now $S, S^*$ have the property that viewed as operators on $\Delta$,

$$R(S^n) \supseteq R(S^{n+1}) \text{ for every } n, \text{ and } \bigcap_{n=0}^{\infty} R(S^n) = \{0\},$$

while

$$R(S^{*n}) = R(S^{*n+1}) = \Delta \text{ for every } n, \text{ and }$$

$$\bigcap_{n=0}^{\infty} R(S^{*n}) = \Delta.$$  

Notice that if $X$ is a $C-(2)$ or $C-(1,2)$ inverse for some $A$, then $X^*$ is a $C-(2)$ or $C-(1,2)$ inverse for $A^*$. On the other hand, if $X_1, X_2$ are $C-(2)$ inverses for $A$ and $X_1 \subseteq X_2$, then $X_1^* \subseteq X_2^*$.

**Proposition 7.2.7** If $A = A^*$ and $A^D$ exists, then $(A^D)^* = A^D$.

**Proof** If $A = A^*$, then $A^D, A^{D^*}$ are $C-(2)$ inverses of $A$ and $A^{D^*} \subseteq A^D$ by the definition of $A^D$. Thus $A^D \subseteq A^{D^*}$ and $A^{D^*} = A^D$. $\square$

While the preceding discussion is helpful in understanding the basic idea behind our definition of a Drazin inverse, it does not provide a means of showing that a given matrix has a Drazin inverse. We now develop one such criteria but first we need the following generalization of [67, Prop. I-63]. Its proof is the same as that of Proposition I-63 and will be omitted.

**Proposition 7.2.8** Let $(A_n)$ be a sequence of matrices. If there exists a matrix $C$ with finite entries such that $|A_n| < C$ for all $n$, then there is a convergent subsequence $(A_k)$ of $(A_n)$.
Given a matrix $A$, let $\tilde{A}_i = \begin{bmatrix} A_i & 0 \\ 0 & 0 \end{bmatrix}$, where $A = \begin{bmatrix} A_i & B_i \\ C_i & D_i \end{bmatrix}$, $A_i$ an $i \times i$ matrix (2.3). Then $\tilde{A}_i^D = \begin{bmatrix} A_i^D & 0 \\ 0 & 0 \end{bmatrix}$.

Theorem 7.2.1 Suppose that $A \in (FCR)$. Let $\tilde{A}_i$ be as in (3). If there exists a subsequence $\tilde{A}_k^D$ which converges, then $\lim_k \tilde{A}_k^D$ is a C-(2) inverse of $A$. If there exists a subsequence $A_k$ which has index $1$ and $\tilde{A}_k^\#$ converges, $\lim_k \tilde{A}_k^\#$ is a C-(1,2) inverse of $A$.

Proof Suppose that $Q = \lim_k \tilde{A}_k^D$ exists. Now $\tilde{A}_k AA_k = \tilde{A}_k^D$. Since $A \in (FCR)$, we have

$$Q = \lim_k \tilde{A}_k^D = \lim_k \tilde{A}_k^D AA_k^D = QAQ. \quad (2.4)$$

On the other hand,

$$\tilde{A}_k^D A - AA_k^D = \begin{bmatrix} A_k^D & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} A_k & B_k \\ C_k & D_k \end{bmatrix} \begin{bmatrix} A_k^D & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & A_k^D B_k \\ -C_k A_k^D & 0 \end{bmatrix}. \quad (2.5)$$

Thus $\lim_k (\tilde{A}_k^D A - AA_k^D) = 0$ since $(\tilde{A}_k^D A - AA_k^D)_{ij} = 0$ if $i, j \leq k$. Hence $QA - AQ = 0$, and $Q$ is a C-(2) inverse as desired. Now suppose that $Q = \lim_k \tilde{A}_k^\#$. Again (4) and (5) hold. Since $AA_k^\# A = \begin{bmatrix} A_k & 0 \\ 0 & 0 \end{bmatrix}$ and $\lim_k \begin{bmatrix} A_k & 0 \\ 0 & 0 \end{bmatrix} = A$, we have $AQA = A$ and $Q$ is a C-(1,2) inverse as desired. 

Note that it is possible to have the index of $A_i$ depend on $i$. 

198
Example 7.2.4 Let $A_{ij} = 1$ if $|i - j| = 1$. Then $A_i$ is invertible if and only if $i$ is even. Let $m = 2i$. Then

$$
\tilde{A}_2^D = \begin{bmatrix}
0 & 1 \\
1 & 0 \\
0 & 0 \\
\end{bmatrix},
\tilde{A}_m^D = \begin{bmatrix}
A_{m-2}^D & R & 0 \\
- & - & - \\
- & - & - \\
S & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 \\
\end{bmatrix}
$$

where $R, S$ are unspecified matrices. Thus $\tilde{A}_m^D$ converges to a matrix $Q$. In fact $Q$ is an inverse of $A$ and

$$
Q = \begin{bmatrix}
1 & 0 & 0 & 1 & 0 & -1 \\
0 & 1 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 & 0 \\
-1 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & -1 & 0 & -1 \\
\end{bmatrix}
$$

The next proposition, combined with Theorem 1 provides a means of not only proving the existence of a group inverse for a self-adjoint matrix but also calculating a given entry.

Proposition 7.2.9 Suppose that $A = A^*$ and let $\delta_i$ be the modulus of the smallest nonzero eigenvalue of $A_i$. If $\{\delta_i^{-1}\}$ is bounded by $m$, then

$$
-mE \leq \tilde{A}_i^D \leq mE \text{ for all } i.
$$

Proof Assume that $A = A^*$. Let $\rho$ denote the spectral radius. We may assume that $\rho(A_i) > 0$. But $\rho(\tilde{A}_i^D) = 1/\delta_i$ and
\[
\max_{k,\ell} |\tilde{A}_i^D|_{k,\ell} \leq \rho(\tilde{A}_i^D), \quad 0 \leq k, \ell \leq i. \quad \square
\]

Before discussing differential and difference equations, we will show that the definition of a Drazin inverse given in this section coincides with the usual one for finite dimensional matrices.

**Theorem 7.2.2** The definitions of the Drazin inverse given in this section and in Chapter 0 are equivalent for \(n \times n\) matrices.

**Proof** Suppose \(A \in \mathbb{C}^{n \times n}\). Then there exists an invertible matrix \(T\) such that

\[
T A T^{-1} = \begin{bmatrix} J & 0 \\ 0 & N \end{bmatrix},
\]

where \(J\) is invertible and \(N\) is nilpotent of index \(k\). Suppose that \(B\) commutes with \(A\). Let

\[
T B T^{-1} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},
\]

Then \(AB = BA\) implies that

\[
J B_{11} = B_{11} J, \quad (2.8)
\]

\[
J B_{12} = B_{12} N, \quad (2.9)
\]

\[
N B_{21} = B_{21} J, \quad (2.10)
\]

and

\[
N B_{22} = B_{22} N. \quad (2.11)
\]

200
From (9) we get \( J^k B_{12} = B_{12} N^k = 0 \) and hence \( B_{12} = 0 \). Similarly (10) gives that \( B_{21} J^k = N^k B_{21} = 0 \) so that \( B_{21} = 0 \). Using (11) and (12); \( N^{k-1} B_{22} = B_{22} N^{k-1} B_{22} = 0 \). Hence \( B_{22} N^{k-1} B_{22} = N^{k-2} B_{22} = 0 \). Continuing in this manner gives \( B_{22} = 0 \). Hence \( B \) in (7) is of the form

\[
TBT^{-1} = \begin{bmatrix} B_{11} & 0 \\ 0 & 0 \end{bmatrix}.
\tag{2.13}
\]

But \( TAA^{-1} T = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \), and \( BAA = A^D AB = B \) so that \( B \subseteq A^D \). Thus \( A^D \) is a maximal C-(2) inverse in the ordering of Definition 3. From Proposition 4 we have that it is the unique maximal one.

Note that the Drazin inverse of this section can be used to extend the results of [66] to denumerable Markov chains. See [10] for details.

Now consider the infinite system of differential equations

\[
\dot{Ax} + Bx = f.
\tag{2.14}
\]

For simplicity assume \( A, B \in \text{ (FCR) } \). Such equations occur, for example, in [106]. They can also occur as discretizations of partial differential equations or in the solution of partial differential equations using formal series.

In [8, p. 122] a system of the form (14) with \( A \) singular arose in computing the coefficients of asymptotic expansions in terms of \( \varepsilon \) for solutions of

\[
A(\varepsilon) \dot{x} + B(\varepsilon)x = f(\varepsilon).
\tag{2.15}
\]

The usual trick of letting \( x = e^{\lambda t} \gamma \) gives

\[
A \dot{\gamma} + (\lambda A + B)\gamma = e^{-\lambda t} f = g
\tag{2.16}
\]
The invertibility of $\lambda A + B$ is no longer sufficient to insure the existence of unique solutions.

**Example 7.2.5** Let $B = -I$ and $A$ be the $S$ of Example 3. Then $(A + B)$ is invertible and the inverse is unique. In fact, $[(A + B)^{-1}]_{ij} = -1$ if $i < j$ and the zero otherwise. Let $\mathbf{f} = \mathbf{0}$. Then (19) becomes $S\dot{x} - x = \mathbf{x}$. Define $x_n$ by $nSx_n = x_{n-1}$. Since $S$ is onto the vector space for all column matrices and is not one-to-one, the choice for $x_n$ is not unique. Let $x(t) = \sum_{n=0}^{\infty} t^n x_n$. By taking each $x_n \in \mathbb{C}^n$ such that $\|x_n\| < M$ for some fixed $M$, $x(t)$ will be a well-defined column vector whose entries are differential functions of $t$ on $(-1,1)$. The choice of $x_n$ is not unique. Note that in Example 5, $(A + B)^{-1}$ is not row finite.

In solving (15) in [8] with $\mathbf{f} = \mathbf{0}$ for the fast coefficients, the system (14) arises with

$$
A = \begin{bmatrix}
A_0 & 0 & 0 \\
A_1 & A_0 & 0 \\
A_2 & A_1 & A_0 \\
\vdots & \vdots & \vdots
\end{bmatrix}, \text{ and}
$$

$$
B = \begin{bmatrix}
0 & 0 & \cdots \\
\vdots & \vdots & \ddots \\
B_0 & 0 & \cdots \\
B_1 & B_0 & \cdots \\
\vdots & \vdots & \ddots
\end{bmatrix}
$$

$\ell$ rows of zeros

(2.17) (2.18)
where $A(\epsilon) = \sum_{i=0}^{\infty} \epsilon^i A_i$, $B(\epsilon) = \sum_{i=0}^{\infty} \epsilon^i B_i$ and $A(\epsilon)^{-1}B(\epsilon)$ has a pole of order $\ell$ at $\epsilon = 0$.

Since $A_0$ is singular, no $A_i$, $i \geq \text{rank } A_0$ of (17) is invertible. Note however (17) is row finite. As shown in [8], (17) often has a row finite left inverse. In the special case $A_k = 0$, $k \geq 2$, $B_j = 0$ if $j \geq 1$, a row finite left inverse is explicitly given in [8, Proposition 5.7.2].

Proposition 7.2.10 Suppose that $A, B$ are row finite, and $A$ has a row finite left-inverse $A^0$, and there exists a sequence $i_m \to \infty$ of integers such that $(A^0B)_{ij} = 0$ if $j > i_m$. Then if $x(0)$ is a consistent initial condition,

$$x(t) = e^{-A^0Bt}x(0).$$

(2.19)

Proof Suppose $x(0)$ is consistent and $x(t)$ is the solution of (14). Since $A^0, B, A^0B$ are row finite, Proposition 2, part (iii), implies

$$(A^0B)x = A^0(Bx)$$

and $x$ is a solution of

$$\dot{x} + A^0Bx = 0.$$  

(2.20)

Since $A^0B$ is row finite, $e^{-A^0Bt} = \sum_{i=0}^{\infty} (-A^0B)^i \frac{t^i}{i!}$ exists as a formal power series, (the powers $(-A^0B)^i$ are well defined). The existence of the sequence $\{i_m\}$ is equivalent to $A^0B$ being in the form

$$A^0B = \begin{bmatrix}
C_{11} & 0 & 0 \\
C_{21} & C_{22} & 0 \\
C_{31} & C_{32} & C_{33} \\
. & . & . \\
. & . & .
\end{bmatrix}$$

203
where each $C_{ii}$ is $n_i \times n_i$, $n_m = i_m - i_{m-1}$. Since the series for
\[ \exp \left( \begin{bmatrix} C_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ C_{rl} & \cdots & C_{rr} \end{bmatrix} t \right) \]
converges uniformly, the series for $e^{-A^0 B t}$ converges element wise. Thus (19) is well defined and satisfies (20).

In the previously mentioned example of (17), (18) from [8], $i_m$ was equal to $n$. In that example $A^0$ was not block lower triangular but $A^0 B$ was and that, as Proposition 10 shows, is what counts. Proposition 10 gives a form for the solutions. The next proposition characterizes the consistent initial conditions.

**Proposition 7.2.11** Under the assumptions of Proposition 10, $x(0)$ is a consistent initial condition if and only if
\[(I - AA^0 B) B^n x(0) = 0, \quad n = 0, 1, 2, \ldots \tag{2.21} \]

**Proof** Suppose $x = e^{-A^0 B t} q$. Then using the row finiteness and the existence of the $(i_m)$ we have
\[ A x + B x = [-AA^0 B + B] e^{-A^0 B t} q = [-AA^0 + I] B e^{-A^0 B t} q = 0. \]
This expression may be differentiated n-times and evaluated at $t = 0$ (again because of the row finiteness and block lower triangular nature of $A^0 B$).

As to be expected, the conclusions of Propositions 10 and 11 often hold without the block triangularity assumption. However, in many of these examples the coefficients possess continuity properties with respect to some norm.
Suppose now that \( A \) in (14) does not have a row finite left inverse. If \( \lambda A + B, A, \) and \( B \) are row finite and \( \lambda A + B \) has a row finite left inverse, then (16) is equivalent to the usual.

\[
\hat{A} \hat{x} + \hat{x} = \hat{f}
\]

(2.22)

where \( \hat{A} \) is row finite.

However, as noted in Example 5, it is possible in (22) to have \( \hat{A}^D = 0, \hat{f} = 0 \) and yet (22) to have non-trivial solutions and hence solutions not of the form \( e^{\hat{A}^D t} \hat{A}^D \hat{x} \). What is more, unlike the finite dimensional case when \( \hat{A} \) is nilpotent these solutions are not uniquely determined by \( \hat{f} \).

3. RESOLVENT TECHNIQUES

Suppose that \( x(t) \) is in some Banach space \( X \) (complete normed linear space) for each \( t \in [t_0, t_1] \) and \( A, B \) are operators from \( X \) into \( Y \) and \( f(t) \in Y \) for each \( t \in [t_0, t_1] \). There are several ways to interpret

\[
A \hat{x} + B \hat{x} = \hat{f}.
\]

(3.1)

One may or may not assume \( A, B \) are bounded (continuous). If they are not bounded, it is usually assumed that they are densely defined and closed operators, that is, their graphs are closed. Hence \( x_n \to x \), \( A x_n \to y \) implies \( A x = y \). Another variation [43] is to require \( x \) to be continuous and \( A \hat{x} \) to be differentiable and write (1) as

\[
\frac{d}{dt} (A \hat{x}) + B \hat{x} = \hat{f},
\]

(3.2)

**Example 7.3.1** Let \( B = I, f = 0 \) and \( A = \text{diag}(1, 1/2, 1/3, \ldots) \). Let \( X = Y = l^\infty \), the space of sequences with \( \| (x_n) \| = \sup_n |x_n| \) and let \( x = \{x_n(t)\} = \{e^{-nt}\} \). Then \( \hat{x} \) is not norm differentiable at \( t = 0 \) but \( \hat{A} \hat{x} = \{\frac{1}{n} e^{-nt}\} \) is.
Suppose for the remainder of this section that $A, B$ are closed linear operators with dense domains $D(A), D(B)$ such that $D(A) \cap D(B)$ is also dense in $X$. Inverses are assumed to be bounded with closed domain. We are not going to try and redevelop the theory of [44] [46] [99] in full detail and greatest generality but rather try and describe the general approach. Let $\rho(A, B) = \{\lambda \mid \lambda A + B \text{ is invertible}\}$. Let $\sigma(A, B) = \emptyset \setminus \rho(A, B)$. The sets $\rho, \sigma$ are called the resolvent set and spectrum of the pencil $\lambda A + B$. At this point there are two ways to proceed. In [99], [47] equation (1) is solved by Laplace transforms. In these approaches it is necessary that $\rho(A, B)$ contain a half-plane $\Re \lambda > \alpha$. Local Laplace transforms are also considered in [99]. The other method is to decompose the space, much as in Chapter 0, by the use of resolvent integrals. That method will now be discussed.

To begin with, suppose $T$ is bounded and $0$ is an isolated point of $\sigma(T)$. Then $(\lambda I - T)^{-1}$ is analytic in $\emptyset$ except for $\lambda \in \sigma(T)$.

Let $C_0$ be a contour around $0$ not including any of $\sigma(T) \setminus \{0\}$ in its interior. Then

$$P = \frac{1}{2\pi i} \int_{C_0} (\lambda I - T)^{-k} d\lambda$$

is a projection. The space $X$ decomposes into direct summands

$$X = R(I - P) \oplus R(P)$$

and with respect to these summands $T$ has the block matrix representation

$$T = \begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix}$$

(3.3)

where $\sigma(T_1) = \sigma(T) \setminus \{0\}$ and $\sigma(T_2) = \{0\}$. Operators $T_2$ for which $\sigma(T_2) = \emptyset$ are called quasi-nilpotent. The Volterra operator is a non-nilpotent example.
If $T_2$ were nilpotent, say $T_2^k = 0$, then Theorem 0.1.3 would give all the solutions of

$$T \dot{x} + x = f$$

with the one convention that

$$T^D = \frac{1}{2\pi i} \int_{C_1} \lambda^{-1} (\lambda I - T)^{-1} d\lambda,$$  \hspace{1cm} (3.5)

where $C_1$ is a contour around $\sigma(T) \setminus \{0\}$ not including zero. In terms of (3),

$$T^D = \begin{bmatrix} T_1^{-1} & 0 \\ 0 & 0 \end{bmatrix}. \hspace{1cm} (3.6)$$

We omit the proofs. If $T_2$ is nilpotent of index $k$,

$$\left(\lambda I - T_1\right)^{-1} = \begin{bmatrix} (\lambda I - T_1)^{-1} & 0 \\ 0 & (\lambda - T_2)^{-1} \end{bmatrix} = \begin{bmatrix} (\lambda I - T_1)^{-1} & 0 \\ 0 & \frac{1}{\lambda} \sum_{i=0}^{k-1} \left( \frac{T_2}{\lambda} \right) \end{bmatrix}$$

\hspace{1cm} (3.7)

and $(\lambda I - T)^{-1}$ would have a pole of order $k$ at $\lambda = 0$. The converse is also true.

**Proposition 7.3.1** Suppose $\lambda = 0$ is an isolated singularity of $(\lambda I - T)^{-1}$. Then $T_2$ in (3) is nilpotent of index $k$ if and only if $(\lambda I - T)^{-1}$ has a pole of order $k$ at $\lambda = 0$.

**Proof** This is a well known fact. We have already proved it one way. If $(\lambda I - T)^{-1}$ has a pole of order $k$ at $\lambda = 0$, then $\lambda^k (\lambda I - T)^{-1}$ is analytic at $\lambda = 0$. Thus $0 = \frac{1}{2\pi i} \int_{C_1} \lambda^k (\lambda I - T)^{-1} d\lambda = \begin{bmatrix} 0 & 0 \\ 0 & T_2^k \end{bmatrix}$ where $C_1$ is defined as before. $\square$
Now consider (1) again. Let $\mathbf{x} = e^{\lambda t} \mathbf{y}$. Note that $\mathbf{x}$ is differentiable if and only if $\mathbf{y}$ is. Then (1) again becomes

$$\mathbf{A} \dot{\mathbf{x}} + (\lambda \mathbf{A} + \mathbf{B}) \mathbf{x} = \mathbf{f}, \quad \dot{\mathbf{f}} = e^{-\lambda t} \mathbf{f}. \quad (3.8)$$

If $\lambda \in \rho(A,B)$ and if $(\lambda \mathbf{A} + \mathbf{B})^{-1} \mathbf{A}$ is bounded (but only densely defined), then the preceding analysis could be used on

$$(\lambda \mathbf{A} + \mathbf{B})^{-1} \mathbf{A} \dot{\mathbf{x}} + \mathbf{x} = (\lambda \mathbf{A} + \mathbf{B})^{-1} \mathbf{f} \quad (3.9)$$

by taking $\mathbf{T}$ a continuous, closed extension of $(\lambda \mathbf{A} + \mathbf{B})^{-1} \mathbf{A}$.

But how likely is $(\lambda \mathbf{A} + \mathbf{B})^{-1} \mathbf{A}$ to be bounded if $\mathbf{A}$ is not? Let

$$\hat{A}_\lambda = (\lambda \mathbf{A} + \mathbf{B})^{-1} \mathbf{A}.$$ 

Then on $\mathcal{D}(\mathbf{A})$, with $\mu \in \rho(A,B)$,

$$(\lambda \mathbf{A} + \mathbf{B})^{-1} \mathbf{A} = (\mu \mathbf{A} + \mathbf{B} + \gamma \mathbf{A})^{-1} \mathbf{A} = (I + \gamma \hat{A}_\mu)^{-1} \hat{A}_\mu \quad (3.10)$$

where $\gamma = \lambda - \mu$. But

$$\gamma (I + \gamma \hat{A}_\mu)^{-1} \hat{A}_\mu = I - (I + \gamma \hat{A}_\mu)^{-1} \quad (3.11)$$

on $\mathcal{D}(\hat{A}_\mu)$. Since the right side is bounded, so is the left. Thus we have,

**Proposition 7.3.2** If $\lambda \in \rho(A,B)$, then $\hat{A}_\lambda$ is a densely defined bounded operator.

Let $\overline{A}_\lambda$ be the continuous extension of $\hat{A}_\lambda$. From Proposition 1, in order to solve (9) by the formulas of Chapter 0, $\lambda$ should be an isolated point of $\sigma(\hat{A}_\mu)$. On a dense subset of $\mathcal{X}$,

$$(\gamma I - \hat{A}_\mu)^{-1} = \left( \frac{1}{\delta} I - \hat{A}_\mu \right)^{-1}$$

$$= \left( \frac{1}{\delta} I - (\mu \mathbf{A} + \mathbf{B})^{-1} \mathbf{A} \right)^{-1} = \left( \frac{1}{\delta} (\mu \mathbf{A} + \mathbf{B}) - \mathbf{A} \right)^{-1} (\mu \mathbf{A} + \mathbf{B})$$

$$= ((\frac{\mu}{\delta} - 1) \mathbf{A} + \frac{1}{\delta} \mathbf{B})^{-1} (\mu \mathbf{A} + \mathbf{B})$$

208
\[= \delta((\mu - \delta)A + B)^{-1}(\mu A + B - \delta A + \delta A)\]
\[= \delta I + \delta^2((\mu - \delta)A + B)^{-1}A.\]

That is,
\[(\gamma I - \hat{A}_\mu)^{-1} = \gamma^{-2}(\gamma I + ((\mu - \frac{1}{\gamma})A + B)^{-1}A),\] (3.12)

Think of \(\mu\) as fixed. But \(\gamma = 0\) will be an isolated pole of \((\gamma I - \hat{A}_\mu)^{-1}\) if and only if \(\gamma^m(\gamma I - \hat{A}_\mu)^{-1}\) is bounded for \(0 < |\gamma| < \varepsilon\) some \(\varepsilon > 0\), some integer \(m \geq 2\). From (12) this happens if and only if
\[\gamma^r[(\mu - \frac{1}{\gamma})A + B]^{-1}A\] (3.13)
is bounded for \(0 < |\gamma| < \varepsilon\), \(r = m - 2\). Letting \(\lambda = \mu - \frac{1}{\gamma}\) we have:

**Proposition 7.3.3** The closure of \(\hat{A}_\mu\) has an isolated eigenvalue at zero and \((\gamma I - \hat{A}_\mu)^{-1}\) has a pole at \(\gamma = 0\) if and only if there is an integer \(r\) such that
\[\|(\lambda A + B)^{-1}A\| \leq C|\mu - \lambda|^r\] (3.14)
for \(|\mu - \lambda| \geq K\), some constants \(C, K\).

As to be expected the least \(r\) useable in (14) is \(\text{Ind}(\hat{A}) - 2\) if \(\text{Ind}(\hat{A}) \geq 1\). Property (14) is the assumption usually made, see, for example [99], [47].

Note that if (1) is formally solved by Laplace transforms,
\[x(s) = (sA + B)^{-1}Ax(0) + (sA + B)^{-1}F(s).\] (3.15)
The growth rate of \((sA + B)^{-1}\) as \(|s| \to \infty\), \(\text{Re}(s) > 0\) is still important. If \(\text{F}(s) = 0\) then for (15) to define a function, one needs for \(\text{Re } s > 0\) that
\[\lim_{s \to \infty} (sA + B)^{-1}Ax(0) = 0.\]
For more information on the Laplace transform approach see [47] or [99].

The next example is simple but in some sense typical of what often happens.

**Example 7.3.2** Let $A = \text{diag}(1, 2, 3, \ldots)$ and $X = \ell^2$, $\|x_i\| = \left(\sum_i x_i^2\right)^{1/2}$. Take $\mathcal{D}(A)$ as all those $x \in \ell^2$ such that $Ax \in \ell^2$. Let $B = I$. Then $\sigma(A) = \{1, 2, \ldots\}$. For any $-1/\mu$ not a positive integer, $(\mu A + B)^{-1}A = \text{diag}\{(1 + \mu)^{-1}, 2(1 + 2\mu)^{-1}, 3(1 + 3\mu)^{-1}, \ldots\}$. Since $m(1 + m)^{-1} \leq 1$, let $\mu = 1$. Then if $|\lambda - 1| \geq 3$, we have

$$\| (\lambda - 1)(\lambda A + B)^{-1}A \| \leq 3.$$  

(3.16)

Note that (16) is (14) with $r = -1$. (The index is one.)

The reader interested in a more sophisticated example is referred to [47] who gives an example where $x = (x_1, \ldots, x_n)$ lies in an open domain in $\mathbb{R}^n$, $u(x, t)$ is the unknown scalar function, the differentiation in (2) is with respect to $t$, and $A$ is multiplication by a function $b(x)$, while

$$Bu = -\sum_{i,k=1}^n a_{ik}(x) \frac{\partial^2 u}{\partial x_i \partial x_k} + \sum_{i=1}^n a_i(x) \frac{\partial u}{\partial x_i} + a(x) u(x).$$

The Banach spaces $X, Y$, are weighted $L^p$ spaces. For these $A, B$, $\rho(A, B) \subset \{\lambda | \text{Re}\lambda > 0\}$ and $\|\lambda A(\lambda A + B)^{-1}\|$ is bounded just as in Example 2.

4. A GEOMETRIC APPROACH

The preceding two sections dealt with an algebraic and a resolvent approach. This section will consider an approach quite similar to that followed for the finite dimensional case but somewhat more geometrical in that it deals directly with ranges and null spaces. In particular we wish to allow zero to be an interior point of the spectrum and not just a pole of the resolvent.
This section will consider the differential equation

\[ \dot{Ax} + Bx = \dot{f}, \quad 0 \leq t \leq t_1 \]  

(4.1)

where \( A, B \) are continuous linear operators from a Banach space \( X \) into a Banach space \( V \). That is, \( A, B \in B(X, V) \).

Let \( D_{\leq t_1} (A, B) \) be the set of all \( x_0 \in X \) such that

\[ \dot{Ax} + Bx = 0, \]

(4.2)

with \( x(0) = x_0 \) has a solution on \([0, t_1] \) which is differentiable with respect to the norm of \( X \). The consistent initial conditions for (1) are a translate for those of (2), if (1) is consistent.

\( D_{\leq t_1} (A, B) \) has been characterized for both finite dimensional [8] and certain infinite dimensional [99] systems using Laplace transforms.

The Theorems of Chapter 0 and [8] provide a nice geometric characterization of \( D_{\leq t_1} (A, B) \) in the finite dimensional case if there exists a \( \lambda \) such that \( \lambda A + B \) has a right or left inverse. We shall establish infinite dimensional versions of these characterizations, then give an example showing the difficulty of further extension. Drazin and weak Drazin inverses for a different class of operators than those covered by the preceding sections will be defined.

To avoid technical difficulties we shall only work with bounded operators. For any operator \( T \) let \( R(T) \) be its range and \( N(T) \) be its null space. Let

\[ R(T^\infty) = \bigcap_{n=0}^\infty R(T^n). \]

A subspace is a closed linear manifold. The restriction of an operator \( T \) to an invariant subspace \( M \) is denoted \( T/M \).
Theorem 7.4.1 Suppose that \( A, B \in B(X,Y) \), and there exists a \( \lambda \) such that \( \lambda A + B \) is invertible. Let \( \hat{A}_\lambda = (\lambda A + B)^{-1}A \). Suppose that \( R(\hat{A}_\lambda^k) \) is closed for all \( k \geq 0 \). Then

\[
D_{t_1}(A,B) \subseteq R(\hat{A}_\lambda^\infty) \quad \text{for all } 0 < t_1 \leq \infty. \tag{4.3}
\]

If \( N(\hat{A}_\lambda) \cap R(\hat{A}_\lambda) \) is complemented in \( R(\hat{A}_\lambda^\infty) \), then

\[
D_{t_1}(A,B) = R(\hat{A}_\lambda^\infty) \quad \text{for all } 0 < t_1 \leq \infty. \tag{4.4}
\]

**Proof** Note that \( x \) is a solution of (2) if and only if

\[
\hat{A}_\lambda \dot{x} + (I - \lambda \hat{A}_\lambda)x = 0. \tag{4.5}
\]

From (5), \( x(t) \in R(\hat{A}_\lambda) \) for all \( t \). Hence \( \dot{x}(t) \in R(\hat{A}_\lambda^2) \) for all \( t \). Proceeding we get \( x(t) \in R(\hat{A}_\lambda^\infty) \) for all \( t \), and (3) follows. Now \( \hat{A}_\lambda \) maps \( R(\hat{A}_\lambda^\infty) \) onto itself. If \( N(\hat{A}_\lambda) \cap R(\hat{A}_\lambda^\infty) \) is complemented in \( R(\hat{A}_\lambda^\infty) \), then \( \hat{A}_\lambda \) restricted to \( R(\hat{A}_\lambda^\infty) \) has a bounded right inverse \( C \). Thus for any \( x_0 \in R(\hat{A}_\lambda^\infty) \),

\[
x(t) = e^{-C(I - \lambda \hat{A}_\lambda)t}x_0 \quad \text{is a solution of (5).} \]

Hence (4) follows. \( \square \)

Of course, in a Hilbert space, the complementation assumption is not needed. The assumptions on \( R(\hat{A}_\lambda^k) \) are needed in some form.

**Example 7.4.1** Take \( X = \ell^2 \), \( A = \text{diag}(1/n) \), \( B = -I \). Note that all assumptions of Theorem 1 are met except for \( R(\hat{A}_\lambda^k) \) being closed. Clearly \( R(\hat{A}_\lambda^\infty) = R(\hat{A}^\infty) \). (3) may be rewritten as \( \dot{x}_n - nx_n = 0 \), \( n \geq 1 \). Thus

\[
x_n(t) = e^{nt}x_n(0). \quad \text{Let } y_n = e^{-\sqrt{n}}. \quad \text{Then } \{y_n\} \in R(\hat{A}^\infty) \text{ since } \{n^k y_n\} \in \ell^2 \text{ for all } k \geq 0. \]

But \( \{x_n(t)\} = \{e^{nt}e^{-\sqrt{n}}\} \notin \ell^2 \) for any \( t > 0 \). Thus \( \{y_n\} \in R(\hat{A}^\infty) \), but \( \{y_n\} \notin D_{t_1}(A,B) \) for all \( t_1 > 0 \), so that (4) does not hold. Note that in this case, (3) does.
This example also shows that $D_{t_1} (A, B)$ can vary with $t_1$ since
$D_{t_1} (A, B) \neq D_{t_2} (A, B)$ if $t_2 > t_1$.

**Example 7.4.2** Again take $X = l^2$, $A$ as in Example 1, but $B = I$. Then
$R(A^\infty)$ is the same as before. But for any $\{y_n\} \in l^2$, \{e^{-nt}y_n\} satisfies
(3). Thus there exists $\{y_n\} \not\in R(A^\infty)$ such that $\{y_n\} \in D_{t_1} (A, B)$ for all
$t_1 > 0$.

Examples 1 and 2 show that as soon as one removes the assumption of
closed ranges, it becomes necessary to consider spectral properties of $A_{\lambda}$
and the pencil $\lambda A + B$ as done in [99]. If in Theorem 1, $\lambda A + B$ has a
bounded left inverse and $R(A_{\lambda}^k)$ is closed for all $k$, then (3) still holds.

In the finite-dimensional case, $R(A_{\lambda}^\infty)$ is independent of $\lambda$. The same
holds for the case covered by Theorem 1.

**Theorem 7.4.2** Suppose $A, B \in B(X,Y)$ and $\lambda, \mu$ are such that $\lambda A + B, \mu A + B$
are invertible. Then $R(A_{\lambda}^\infty) = R(A_{\mu}^\infty)$.

**Proof** $A_{\lambda} = (\lambda A + B)^{-1}A + (\lambda A_{\mu} + (I - \mu A_{\mu}))^{-1}(\mu A + B)^{-1}A$
$= (I + (\lambda - \mu)A_{\mu})^{-1}A_{\mu}$. Thus $R(A_{\lambda}^k) = R(A_{\mu}^k)$ since $A_{\mu}$
commutes with the invertible operator $(I + (\lambda - \mu)A_{\mu})$. \[ \]

With applications to (1) in mind, and motivated by the proof of Theorem
1, a Drazin-like inverse will now be defined.

**Definition 7.4.1** If $A \in B(X)$, $R(A^\infty)$, is closed, and $R(A^\infty)$ is complemented
in $X$, then a special type of weak Drazin inverse [22] is defined as follows.
Let $M$ be a complementary subspace to $R(A^\infty)$. Relative to $R(A^\infty) + M$. $A$ may
be written
\[ A = \begin{bmatrix} U & S \\ 0 & V \end{bmatrix} \]  

(4.6)

where \( U \) is an onto bounded operator and \( \mathcal{R}(V^\infty) = \{0\} \). If \( U \) has a bounded right inverse, then define

\[ A^d = \begin{bmatrix} U^0 & U^0Z \\ 0 & 0 \end{bmatrix} \]

where \( U^0 \) is a bounded right inverse of \( U \), \( Z \) is an arbitrary bounded operator.

Note that \( A^d \) can exist for operators for which \( 0 \) is not only not an isolated point of the spectrum of \( A \) but, in fact, an interior point of the spectrum of \( A \).

**Definition 7.4.2** If in (6), \( U \) is invertible and \( S = 0 \), then \( A^d \) is called the Drazin inverse of \( A \) and is denoted \( A^D \). \( A^D \) need not be unique.

**Proposition 7.4.1** Suppose \( A \in B(X) \), and \( \mathcal{R}(A^\infty) \) is closed and complemented by a subspace \( \mathcal{M} \). Let \( P \) be the projection onto \( \mathcal{M} \) along \( \mathcal{R}(A^\infty) \). If \( \sigma(A|\mathcal{R}(A^\infty)) \cap \sigma(PA|\mathcal{M}) = \emptyset \), then \( A^D \) exists.

**Proof** The assumptions of Proposition 1 imply \( \sigma(U) \cap \sigma(V) = \emptyset \) in (6).

Since \( 0 \in \sigma(V) \), \( U \) is invertible and \( A \) is similar to \( \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix} \) [98]. \( \square \)

Note that if the index of \( A \) is finite, then \( V \) in (6) is nilpotent of index \( k \) and Definition 2 gives the usual Drazin inverse. If \( T^d \) exists and \( V \) is not nilpotent, \( A \) will be said to be of infinite index.

**Proposition 7.4.2** Suppose \( A \in B(X) \), and \( A^D \) exists. Then
(i) $A^D$ is unique under the assumptions of Proposition 1.

(ii) $A^D AA = AA^D$

(iii) $R(A^D A) = R(A^\omega)$

(iv) $A^D AA^D = A^D$.

If $A^d$ exists, then

(v) $R(AA^d) = R(A^\omega)$,

(vi) $A^d AA^d = A^d$.

If $U$ in (6) is invertible, then

(vii) $R(A^d A) = R(A^\omega)$.

The solution of (2) is given by:

Corollary 7.4.1 If $A, B \in B(X,Y)$, there exists a $\lambda$ such that $\lambda A + B$ is invertible, $R(\hat{A}_{\lambda}^k)$ is closed for all $k$, $N(\hat{A}_{\lambda}) \cap R(\hat{A}_{\lambda}^\omega)$ is complemented in $R(\hat{A}_{\lambda}^\omega)$, and $R(\hat{A}_{\lambda}^\omega)$ is complemented in $X$, then

$$\hat{x}(t) = e^{-A_{\lambda}^d (I-\lambda A_{\lambda}) t} A_{\lambda} A_{\lambda}^d g$$

is a solution of (2) for all $g \in X$. Also $R(\hat{A}_{\lambda} A_{\lambda}^d) = D_{\ell_1} (A, B)$. Solutions are uniquely determined by consistent initial conditions if and only if $N(\hat{A}_{\lambda}) \cap R(\hat{A}_{\lambda}^\omega) = \{0\}$. In this case (7) gives all solutions of (2).

In solving (1), the appropriate formulas could be developed under the weaker assumption that $A_{\lambda}^d$ exists. For simplicity of exposition we shall only consider the case when $A_{\lambda}^D$ exists. For notational convenience the lambda subscript will be omitted from $A_{\lambda}$.

Proposition 7.4.3 Suppose that $A, B$ satisfy the assumption of Corollary 1 and $A^D$ exists. Then all solutions of (1) are of the form
\[ x = e^{-\hat{A}^D (I - \hat{\lambda} A) t} A^\lambda q + \hat{A}^D \int_0^t e^{\hat{A}^D (I - \hat{\lambda} A)(s-t)\hat{f}}(s)ds + (I - \hat{A}^D) x(t) \]  (4.8)

where \((I - \hat{A}^D) x(t)\) is the unique solution, if it exists, of

\[ \hat{A}[(I - \hat{A}^D) x] + (I - \hat{\lambda} A)(I - \hat{A}^D) x = (I - \hat{A}^D) \hat{f}. \]  (4.9)

From Theorem 0.1.3, the explicit solution of (9) is

\[ (I - \hat{A}^D) x = \sum_{i=0}^{k-1} [-\hat{A}^D (I - \hat{\lambda} A)]^i (I - \hat{\lambda} A) \hat{f} (i) \]  (4.10)

when \(X\) is finite dimensional. If \(\text{Ind}(\hat{A}(I - \hat{A}^D)) < \infty\) and \(X\) is a Hilbert space, then \(\hat{A}(I - \hat{A}^D)\) has the usual Jordan form since the subspaces \(R(\hat{A}^k)\) are assumed closed [58]. In general, one has \(\hat{A}(I - \hat{A}^D)\) has a strictly lower triangular block matrix representation and the entries of \((I - \hat{A}^D) x\) with respect to this decomposition may be computed recursively in a unique manner. The difficulty is in showing that \((I - \hat{A}^D) x\) when so calculated is in \(X\).

For purposes of illustration consider the case when (9) takes the form

\[
\begin{bmatrix}
0 & 0 & 0 & \vdots & \vdots & x_1 \\
\vdots & 0 & 0 & \vdots & \vdots & \vdots \\
0 & \vdots & 0 & \vdots & \vdots & \vdots \\
\end{bmatrix} + 
\begin{bmatrix}
I & 0 & 0 & \vdots & \vdots & \vdots \\
0 & I & 0 & \vdots & \vdots & \vdots \\
0 & 0 & I & \vdots & \vdots & \vdots \\
\end{bmatrix} \begin{bmatrix}
x_1 \\
x_2 \\
x_3 \\
\end{bmatrix} = 
\begin{bmatrix}
f_1 \\
f_2 \\
f_3 \\
\end{bmatrix}. 
\]  (4.11)

Then \(x_1 = f_1\) and \(V \dot{x}_1 + x_{i+1} = f_{i+1}\). This would occur, for example, if \(V\) were similar to a block weighted shift and \(I - \lambda V\) were invertible. Then the solution of (11) is

\[ x_i = \sum_{k=0}^{i-1} (-1)^k \prod_{j=i-1-k}^{i-1} V_j f_{i-k}^{(r)}(t). \]  (4.12)
Thus for (11) to be consistent, the $f_1$ must be infinitely differentiable.

The expression (12) may be written as

$$ x = \sum_{n=0}^{\infty} (-1)^n [V(I - \lambda V)^{-1}]^n (I - \lambda V)^{-1} (I - A^D A) f_1 $$

(4.13)

which is a direct analogue of (10). Note (11) need not be consistent for all $f$. For example, if $V_1 = 1, f_1 = 0$ for $i > 1, f_1(t) = e^t$, $x(t)$ would be in \( \ell^\infty \) but not in \( \ell^2 \).

At this point, it should be clear that one can use Proposition 3 to get an analogue of (10) where the $f_1$'s are restricted to those for which (1) is consistent except that the sum will be infinite as in (13).

We will now relate the $A^D$ from this section to the $A^D$ of Section 2.

**Proposition 7.4.4** Suppose that $A$ satisfies the assumptions of Definition 1,2 and Proposition 1. Let $A = \{Y \in B(X) | YA = AY, YAY = Y\}$ and order $A$ by $Y \leq Z$ if $YAZ = ZAY = Y$. Then $A^D$ is the unique maximal element of $A$.

**Proof** It suffices to show that $A^D \in A$ and is maximal. Suppose $Y \in A$. Then $R(A^\infty)$ is invariant under $Y$ so that subject to the appropriate decomposition of $X$, $Y = \begin{bmatrix} Y_1 & Y_2 \\ 0 & Y_3 \end{bmatrix}$, $A = \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix}$. Then $Y_3 V = V Y_3$, and $Y_3 V Y_3 = Y_3$.

Now $Y_3 V$ is a projection and $Y_3 V = (Y_3 V)^n = V^n Y_3^n$ so that $R(Y_3 V) \subseteq \sigma(R(V^n))$.

Hence $R(Y_3 V) = \{0\}$ and $Y_3 V = 0$. Thus $Y_3 = 0$. Since $A^D A = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} = A^D A$,

$Y \leq A^D$ if $Y_2 = 0$. That $Y_2 = 0$ follows from the assumption that $\sigma(A | R(A^\infty)) \cap \sigma(PA | M) = \emptyset$. If this condition does not hold, then it is sometimes possible to construct $Y \in A$ such that $Y_1 = U^{-1}, Y_2 \neq 0$ and neither $Y \leq A^D$ nor $A^D \leq Y$ holds. \[\square\]
We conclude this section by briefly considering unbounded operators. To follow the approach given above, a class of operators $A$ should be defined, an ordering $\subseteq$ introduced, and the maximal elements studied. The major difficulty is in deciding which $X$ to put in $A$. Let $D(X)$ be the domain of $X$. Is it best to have

$$AX =XA$$
or

$$AX =XA \text{ on } D(X)?$$

Similarly, should $X$ be required to be closed or not? To show how it can all make a difference, let us consider two definitions of $A$. Let $A_1$ be the set of all $X$ such that

$$D(X) \text{ is dense},$$

$$AX =XA \text{ on } D(X)$$

$$XAX =X,$$

and

$$\bigcup N(A^k) \subseteq D(X), \quad R(A^\infty) = \cap R(A^k) \subseteq D(X).$$

Let $A_2$ be the set of all $X$ such that $X$ satisfies (14), (15), (16), (17), and $X$ is closed. That is, the graph of $X$ is closed.

**Example 7.4.3** Take $B = l^2$ and $A = S^*$, where $S^*\{\alpha_0, \alpha_1, \alpha_2, \ldots\}$

$$= \{\alpha_1, \alpha_2, \alpha_3, \ldots\}.$$ Then

$$N(A^k) = \{\alpha_n \in l^2 : \alpha_n = 0 \text{ if } n > k\}.$$
Thus $\bigcup_k N(A^k)$ is dense in $l^2$. Using (15), (16), and (17) we can show that $\bigcup_k N(A^k) \subseteq N(X)$. But any closed operator has a closed nullspace. Thus $A_2 = \{0\}$. To see that $A_1 \neq \{0\}$, let $M$ be a finite-dimensional subspace of $l^2$, invariant under $S^*$, such that no vector in $M$ has only a finite number of nonzero components. Recall $S^*|M$ denotes the restriction of $S^*$ to $M$.

Then define $X$ as follows. Let $\mathcal{D}(X) = M + \bigcup_k N(A^k)$. Note that $M \cap \bigcup_k N(A^k) = \{0\}$. Set $X|M = (S^*|M)^{-1}$ and $X|\bigcup_k N(A^k) = 0$. Extend linearly to $\mathcal{D}(X)$. Then $X \neq 0$ and $X \in A_1$.

For the reader unfamiliar with $S^*$, the existence of the subspace $M$ may not be obvious. That it exists can be demonstrated as follows. Identify $l^2$ with $H^2$, the functions analytic in the open disc whose Taylor series have square summable coefficients. Let $b$ be a finite Blaschke product such that $b(0) \neq 0$. (b is an analytic function of modulus one on the circle with a finite number of zeros inside.) Let $M = H^2 \cap (bH^2)^\perp$. Then $M$ is invariant under $S^*$, and $S^*|M$ is invertible since $b(0) \neq 0$ [62], [65].

Now consider (17). How one defines $A^D$ will depend on how one wishes to use it. In the difference equation:

$$Ax_{n+1} = x_n, \quad \text{A singular, \quad } n \geq 0, \quad (4.18)$$

if $\{x_n\}$ is a solution, then

$$x_n = Ax_{n+1} = A^2x_{n+2} = \cdots = A^kx_{n+k}.$$

Note that $A$ maps $R(A^{\infty})$ onto itself. If $\{x_n\}$ is a solution of (18), then $x_n \in R(A^{\infty})$ for every $n$. If $A^D$ is to solve (18) we would need $R(A^{\infty}) \subseteq \mathcal{D}(X)$.

Whether or not $\bigcup_k N(A^k) \subseteq N(X)$ is necessary is a more difficult question. Its inclusion depends on whether or not the unicity of $A^D$ or the solving of (18) by using different $X \in A$ is the more desirable.
References


32. J. D. Cobb, Descriptor variable systems and optimal state regulation, preprint, 1981.


Laplace transform method for a class of degenerate evolution problems, Rend. Mat. 12(1979), 511-536.


Convergence in the boundary layer for singularly perturbed equations, Automatica (to appear).


On a class of optimal control problems with an almost cost free solution, preprint, 1980.


61. L. M. Hall, The number of analytic solutions of a singular differential system, preprint.


95. L. Petzold, Differential/algebraic equations are not ODE's, Sandia Report, SAND 81-8668, 1981.


Index

admissible (controls) 116 (13)
allowable 116
amplifier (operational) 80, 186

Blaschke product 219

capacitor (nonlinear) 76
characteristic (v-i) 75
circuits 78, 170
contour 7, 206
coordinate change 115
control (cheap) 32, 171
controllable (completely, R, C) 20
convergence (boundary) 61
infinite matrices 192
distributional 47
differences (first order) 92, 135, 146, 162
higher order 98, 134
variable step 105
direct sum 4
distributions (derivative, tempered) 8, 48, 49
solutions 46, 64
domains 206
eigenvalues 57, 65
error 97, 135
Euler's (forward or backward) 92, 135, 146, 162
expansions (asymptotic) 201
feedback 64
gradient 148
hysteresis 186
impulse 9, 50, 64, 109, 185
incidence matrix 77
index 5, 79, 115, 149
inductor (nonlinear) 76
initial conditions (consistent) 9, 94, 125, 134, 150, 155, 158, 204, 211
invariant properties 117
inverses
  C-(2), C-(1,2) 193
  Drazin (weak) 6, 194, (211)
group 7, 194
maximal 194
Jordan canonical form 5
jump solutions 185
Laplace transform 9, 51, 206
limiting solution 55
noise 189
null space 4
230
observable  21
parasitics  180
positive semi-definite  5, 199
projection  4
range  4, 67, 211
rank (constant)  120
reachable  15
regular singular  133
resistor (nonlinear)  75
resolvent  8, 205
Riccati  31
semi-stable  61
similar  5
singular perturbations  56, 184
solvable  9, 82, 115, 149
spectrum  5, 57, 141, 206
stabilizable  22
test functions  47
time optimal  27
totally singular  29
transistor  85
unicursal  87
List of symbols

\( \mathbf{x} \) vector

\( \mathbb{R}, \mathbb{C} \) real (complex) numbers, p. 4

\( \mathbb{R}^{m \times n}, \mathbb{C}^{m \times n} \) real (complex) \( m \times n \) matrices, p. 4

\( A^*, A^T \) (conjugate) transpose, p. 4

\( \langle x, y \rangle \) inner or scalar product, p. 4

\( \square \) end of proof, Halmos

\( \| \cdot \| \), \( M \) norm, subspace, p. 4

\( \text{dim} \) dimension, p. 4

\( \mathcal{R}(A), N(A) \) range, nullspace, p. 4

\( \bigoplus \) direct sum, p. 4

\( \mathcal{P}_{M,N} \) projection onto \( M \) along \( N \), p. 4

\( \mathcal{M}^\perp \) orthogonal complement, p. 4

\( \mathcal{P}_M \) orthogonal projection, p. 5

\( \det, \sigma(A) \) determinant, spectrum, p. 5

\( \rho(A) \) spectral radius, p. 5

\( \text{diag} \) diagonal matrix

\( \text{Ind} \) index, p. 5

\( \mathcal{A}^D \) Drazin inverse, p. 6

\( \mathcal{E}^A \) exponential, p. 7

\( \delta, \delta^{(1)} \) delta functions, p. 8

\( \mathcal{L} \) Laplace transform

\( \hat{A}, \hat{A}_\lambda \) p. 10

\( I, \Omega(I) \) p. 13

232
\( C^k(I), P^k(I), E(I) \) p. 14
\( \text{reach}\{x_0, I\} \) reach set p. 15
\( R_1, R_2 \) p. 16
\( AI \) p. 19
\( X_p \) particular solution, p. 25
\( X_h \) homogeneous solution, p. 25
\( E \) p. 27
\( T_0, T_1 \) p. 29
\( J(x, u), \lambda \) cost functional, costate, p. 32
\( N^\mu, M^\mu \) p. 34
\( K_R \) p. 32
\( K_E \) p. 37
\( K_S \) p. 41
\( T \) test functions, p. 47
\( S \) test functions of rapid descent, p. 47
\( \tilde{T}, \tilde{S} \) p. 47
\( D \) distributions, p. 48
\( SD \) tempered distributions, p. 48
\( \langle, \rangle \) p. 48
\( \underline{f}[m] \) p. 52
\( \text{Re} \) real part, p. 57
\( \text{Re} \) maximum real part, p. 57
\( \sigma(A, B) \) p. 64
\( \mathbb{E}^2, O(\cdotp) \) p. 91
\( D \) p. 99
\( \text{Ind}(A, B) \) Index of a system, p. 115
\( \nabla, \nabla_x \) gradient, vector gradient, p. 148
\( M(x, y) \)  p. 183
\( M, M_x \)  p. 184
\( (Fr) \)  p. 191
\( (Fc), (Frc) \)  p. 192
\( e \)  vector of ones, p. 192
\( |X| \)  p. 193
\( X \subseteq Y \)  p. 194
\( \Delta \)  sequence space
\( \ell^\infty \)  p. 205
\( X, Y \)  Banach spaces, p. 205
\( \mathcal{D}(A) \)  domain of \( A \), p. 206
\( \rho(A, B) \)  p. 206
\( A^d \)  weak Drazin, p. 213
\( H^2 \)  Hardy space, p. 219
\( B(X, Y), D_{r_1} (A, B) \)  p. 211
\( \mathcal{R}(T^\infty), T|\mathcal{M} \)  p. 211
Submission of proposals for consideration

Suggestions for publication, in the form of outlines and representative samples, are invited by the editorial board for assessment. Intending authors should contact either the main editor or another member of the editorial board, citing the relevant AMS subject classifications. Refereeing is by members of the board and other mathematical authorities in the topic concerned, located throughout the world.

Preparation of accepted manuscripts

On acceptance of a proposal, the publisher will supply full instructions for the preparation of manuscripts in a form suitable for direct photolithographic reproduction. Specially printed grid sheets are provided and a contribution is offered by the publisher towards the cost of typing.

Illustrations should be prepared by the authors, ready for direct reproduction without further improvement. The use of hand-drawn symbols should be avoided wherever possible, in order to maintain maximum clarity of the text.

The publisher will be pleased to give any guidance necessary during the preparation of a typescript, and will be happy to answer any queries.

Important note

In order to avoid later retyping, intending authors are strongly urged not to begin final preparation of a typescript before receiving the publisher's guidelines and special paper. In this way it is hoped to preserve the uniform appearance of the series.

Titles in this series (continued)

51 Subnormal operators
   J B Conway
52 Wave propagation in viscoelastic media
   F Marmardi
53 Nonlinear partial differential equations and
   their applications: Collège de France
   Seminar. Volume I
   H Brezis and J L Lions
54 Geometry of Coxeter groups
   H Hiller
55 Cusps of Gauss mappings
   T Banchoff, T Gaffney and C McCrory
56 An approach to algebraic K-theory
   A J Berrick
57 Convex analysis and optimization
   J-P Aubin and R B Vinter
58 Convex analysis with applications in
   differentiation of convex functions
   J R Giles
59 Weak and variational methods for moving
   boundary problems
   C M Elliott and J R Ockendon
60 Nonlinear partial differential equations and
   their applications. Collège de France
   Seminar. Volume II
   H Brezis and J Lions
61 Singular systems of differential equations II
   S L Campbell
ABOUT THIS VOLUME

Most of the material in this self-contained sequel to *Singular Systems of Differential Equations* (Research Note no. 40) has not previously appeared in book form and much of it is entirely new. It is concerned with singular systems of both linear and nonlinear differential equations and their applications in the theories of electrical circuits and optimal control.

Distinctive features of the work include: the discussion of both analytic and numerical techniques for handling nonlinear singular systems; further applications to nonlinear circuits; and a discussion of infinite dimensional systems and their applications to singular perturbations.

RESEARCH NOTES IN MATHEMATICS

The aim of this series is to disseminate important new material of a specialist nature in economic form. It ranges over the whole spectrum of mathematics and also reflects the changing momentum of dialogue between hitherto distinct areas of pure and applied parts of the discipline.

The editorial board has been chosen accordingly and will from time to time be recomposed to represent the full diversity of mathematics as covered by *Mathematical Reviews*.

This is a rapid means of publication for current material whose style of exposition is that of a developing subject. Work that is in most respects final and definitive, but not yet refined into a formal monograph, will also be considered for a place in the series. Normally homogeneous material is required, even if written by more than one author, thus multi-author works will be included provided that there is a strong linking theme or editorial pattern.

*Proposals and manuscripts*: See inside back cover.

Main Editors
A. Jeffrey, University of Newcastle-upon-Tyne
R. G. Douglas, State University of New York at Stony Brook

Editorial Board
F. F. Bonsall, University of Edinburgh
H. Brezis, Université de Paris
G. Fichera, Università di Roma
R. P. Gilbert, University of Delaware
K. Kirchgässner, Universität Stuttgart
R. E. Meyer, University of Wisconsin-Madison
J. Nitsche, Universität Freiburg
L. E. Payne, Cornell University
I. N. Stewart, University of Warwick
S. J. Taylor, University of Liverpool

ISBN 0 273 08516 6