ERROR ANALYSIS OF COLLOCATION METHODS

FOR THE

NUMERICAL SOLUTION

OF

ORDINARY DIFFERENTIAL EQUATIONS

D. M. CRUICKSHANK

Ph.D. Thesis

February 1974

UNIVERSITY OF NEWCASTLE UPON TYNE

ACKNOWLEDGEMENTS

I should like to thank my supervisor, Dr. Kenneth Wright, whose advice and enthusiasm throughout was very much appreciated.

I am also indebted to the typist, Mrs. Sandra Nicoll who bravely tackled this difficult task.

Throughout the period of research for this thesis the author was supported by the Science Research Council.

ABSTRACT

This thesis is concerned with an error analysis of numerical methods for two point boundary value problems and much of the investigation is concentrated on collocation methods from an 'a posteriori' point of view.

Most of the previous work on error bounds for boundary value problems has been of an 'a priori' nature, requiring knowledge of the inverse of the differential operator under consideration and furnishing convergence proofs and theoretical bounds on the error. There are however a few results of the converse nature and in this thesis means of determining error bounds in practice are developed, much of the analysis also applying to Fredholm integral equations of the second kind.

In more detail, having firstly considered certain preliminaries the setting for the theory and the principal results for later use are presented. It is demonstrated how the approximate solution by collocation of linear differential equations fits into this background and different 'a priori' approaches are examined by example and shown to be rather unsatisfactory.

The 'a posteriori' outlook is then considered and to achieve practical results the inverse of the approximating operator is related to the inverse of the collocation matrix. However the problem of obtaining a suitable bound on the norm of this inverse operator is encountered and after examination of the most obvious approach which proves unsatisfactory a convenient bound is developed.

Certain interesting computational properties of matrices involved in the process are discussed and a brief examination of condition numbers is given.

A different theoretical analysis using the concept of a 'collectively compact sequence of operators' is considered and it is demonstrated that the approximate solution by collocation of linear differential equations can be 'extended' to satisfy the conditions for this theory. Again the error bounds are reduced to a more practical level and subsequently a generalisation of the notion of this extension is suggested.

The implementation of the various practical error bounds which have been deduced is then considered in detail and formulae for their evaluation are presented. The numerical results of examples of this application are then given followed by a discussion of certain relevent points concerning the experiments.

In the final chapter certain possible extensions of the analysis herein are briefly examined and lastly a review of the work of this thesis with appropriate conclusions is given.

CONTENTS

INDEX OF TABLES

CHAPTER 1

INTRODUCTION AND PRELIMINARIES

1.1 Numerical Methods for Boundary Value Problems

In this section we survey the general background of numerical methods prior to the main part of the thesis which is concerned with error analysis.

We are primarily interested in certain aspects of the numerical solution of two point boundary value problems. A fairly general equation of this type may be regarded in the form

$$\frac{d^m x}{ds^m} + f(s,x,x^{(1)}, \ldots x^{(m-1)}) = 0 \qquad (1.1a)$$

over some interval [a,b] say with f a nonlinear function in the m+1 variables $s,x,x^{(1)}, \ldots x^{(m-1)}$ and will be subject to m boundary conditions, say

$$V_i(x,x^{(1)}, \ldots x^{(m-1)}) = 0 \quad (i = 1 \ldots m) \qquad (1.1b)$$

where the V_i are certain nonlinear functions in the m variables $x,x^{(1)}, \ldots x^{(m-1)}$ which are evaluated at either of the end points a or b.

However we deal mainly with linear equations which may be expressed as

$$Lx \equiv \frac{d^m x}{ds^m} + \sum_{j=0}^{m-1} p_j(s)x^{(j)}(s) = y(s) \qquad (1.2a)$$

subject to

$$U_i(x, x^{(1)}, \ldots x^{(m-1)}) = \gamma_i \qquad (i = 1 \ldots m) \qquad (1.2b)$$

where now the U_i are linear functions of the m variables again evaluated at either a or b and the γ_i are constants. We shall usually assume that $p_j(s)$ $(j = 0 \ldots m-1)$ and $y(s)$ are continuous and shall employ the abbreviation $U_i(x) = \gamma_i$ $(i = 1 \ldots m)$ for (1.2b).

Problems of either type are rarely solvable analytically and for this reason numerical methods of obtaining an approximate solution have been developed. There are a number of such approaches but they are all comprised of similar stages.

Consider for example the numerical solution of a linear problem of the form (1.2). Generally speaking any method for its approximate solution involves the following steps.

(a) A choice of a characterisation of the approximation in terms of certain unknown constants,

(b) A means of forming linear algebraic equations for the unknowns,

(c) A means of solving the algebraic equations, and sometimes the fourth stage

(d) Determination of the approximate solution from the constants.

For example the collocation and Rayleigh Ritz methods would involve all four processes with the numerical solution specified by some constants $a_1, a_2 \ldots a_n$ say and represented by a finite sum $\sum_{j=1}^{n} a_j \psi_j(s)$ for some independent set of functions $\{\psi_j\}_{j=1}^{n}$. The particular method then sets

up the equations which are subsequently solved by some means. The fourth step then determines the approximation by forming the finite sum at any desired point.

Finite difference approaches can also be viewed in this way with the numerical solution characterised by a set of its values at mesh points throughout the interval [a,b]. These point values are determined by applying a finite difference operator at the grid points to set up equations which may be solved for instance by a band-matrix algorithm. Since the unknowns are in fact values of the approximate solution no fourth stage is generally performed but one could visualise this if an interpolant of these point values were constructed.

Shooting methods may also be regarded in a similar manner to the finite difference case. We do not wish to consider this aspect in detail and it is in any case rather an unnatural way of looking at these methods.

In this thesis we concentrate on the collocation method and a detailed description of this is presented in section 1.3.

A consideration of finite difference and shooting methods is given by Keller (1968) or more recently by Roberts and Shipman (1972). An introduction to the Rayleigh Ritz and Galerkin methods may be found in Collatz (1960) with more detailed accounts of the Ritz method given by, for example, Gould (1957), Kantorovich and Krylov (1958), Mikhlin and Smolitskiy (1967) and Mikhlin (1970). There has been a considerable amount of recent work in this field, for example a series of papers by Ciarlet, Schultz and Varga with the latest in

1969, but it is not the aim of this thesis to discuss these developments.

When nonlinear problems of the form (1.1) are encountered we have a choice of procedure. Either nonlinear algebraic equations are set up and solved by an iterative technique or the problem itself is linearised and solved successively. Under certain circumstances these two approaches are equivalent.

An example of the second of these alternatives is Newton's method for operator equations. Application of this process to an equation of type (1.1) entails the successive approximate solution of linear differential equations

$$\frac{d^m x_{k+1}}{ds^m} + \sum_{j=0}^{m-1} \frac{\partial f}{\partial x^{(j)}} (s, x_k, x_k^{(1)}, \ldots x_k^{(m-1)}) x_{k+1}^{(j)}$$

$$= - f(s, x_k, x_k^{(1)} \ldots x_k^{(m-1)})$$

$$+ \sum_{j=0}^{m-1} \frac{\partial f}{\partial x^{(j)}} (s, x_k, x_k^{(1)}, \ldots x_k^{(m-1)}) x_k^{(j)} \qquad (1.3a)$$

subject to the linearised boundary conditions

$$\sum_{j=0}^{m-1} \frac{\partial V_i}{\partial x^{(j)}} (x_k, x_k^{(1)}, \ldots x_k^{(m-1)}) x_{k+1}^{(j)}$$

$$= - V_i (x_k, x_k^{(1)}, \ldots x_k^{(m-1)})$$

$$+ \sum_{j=0}^{m-1} \frac{\partial V_i}{\partial x^{(j)}} (x_k, x_k^{(1)}, \ldots x_k^{(m-1)}) x_k^{(j)}$$

$$(k \geq 0)(i=1 \ldots m) \qquad (1.3b)$$

That is, an initial guess x_o is chosen and then the problem (1.3) is solved by a numerical method for a first iterate $x_1(s)$ (or set of point values if a difference method is employed) and so on until some criterion, for example the proximity of successive iterates, is used for terminating the iteration.

The convergence of this process has been investigated by Kantorovich who gives sufficient conditions for local convergence - see for example Kantorovich and Akilov (1964, Chapter 18). Further discussions relating to the Newton (Kantorovich) method are given in Mikhlin and Smolitskiy (1967), Rall (1969) and Roberts and Shipman (1972). Alternatively if certain monotonicity properties are satisfied global convergence can be established, see Bellman and Kalaba (1965) and Collatz (1966).

Thus we see that for the approximate solution of any boundary value problem it is quite likely that linear differential equations would be encountered.

This completes a brief review of the most popular methods for the numerical solution of two point boundary value problems. In the next two sections we are more specific and consider a class known as projection methods and the collocation method in particular.

1.2 Projection Methods

As we have mentioned we are principally concerned with the numerical solution of differential equations, however much of the theory which we shall encounter utilises concepts of functional analysis and applies to

more general operator equations. Several methods for the approximate solution of such equations can be classified as projection methods and a brief description of these is given below. It is assumed that the reader is familiar with the basic concepts and notation of functional analysis.

Let X and Y be linear spaces with M a linear operator mapping $X \to Y$ and suppose we are given an equation

$$Mx = y \quad (y \in Y) \tag{1.4}$$

to solve for $x \in X$.

Let $\tilde{X}$ and $\tilde{Y}$ be subspaces of X and Y respectively of equal dimension. Let ϕ be a projection from $Y \to \tilde{Y}$ i.e. $\phi(Y) = \tilde{Y} = \phi(\tilde{Y})$.

With this background we shall define a <u>projection method</u> as a method which seeks an approximate solution $\tilde{x} \in \tilde{X}$ to (1.4) satisfying an equation

$$\phi(M\tilde{x} - y) = 0 \tag{1.5}$$

For any approximation $\hat{x}$ to the solution of (1.4) we should like the residual $M\hat{x} - y$ to be as close to zero as possible (since this is so for the true solution) and projection methods seek an $\tilde{x}$ such that the corresponding residual is mapped to zero under the influence of the projection operator.

There are other definitions of projection methods

but for the purposes of this thesis we shall adhere to the above specification.

1.3 Collocation

The collocation method is now described in detail and it is subsequently shown that it can usually be a projection method. Latterly the main literature on the subject is briefly reviewed.

Suppose we wish to solve numerically a problem of type (1.2). There are two essentially equivalent variations of the application of the collocation process and both are described.

In one approach the collocation method seeks an approximate solution $\tilde{x}$ in the form of a finite sum,

$$\tilde{x}(s) = \sum_{j=1}^{n+m} a_j \psi_j(s) \qquad (1.6)$$

where $\{a_j\}_{j=1}^{n+m}$ are real constants and the basis functions $\{\psi_j\}_{j=1}^{n+m}$ form a linearly independent set and are chosen by the user. An obvious choice for the $\{\psi_j\}$ is a set of polynomials, for example simple powers, Chebyshev polynomials or Legendre polynomials. Spline functions are another popular selection for the basis functions.

A set of n points $\{s_i\}_{i=1}^{n}$ known as the collocation points or nodes are chosen distributed throughout the interval [a,b]. When polynomial basis functions are employed the zeros of the n^{th} degree Chebyshev or Legendre polynomial are often taken as the nodes.

The method then sets up equations for the unknown constants by collocating on the selected points, that

is by requiring that the residual $L\tilde{x} - y$ vanish at the collocation points. This leads to n equations satisfied by the constants $\{a_j\}_{j=1}^{n+m}$, namely

$$\sum_{j=1}^{n+m} a_j L\,\psi_j\Big|_{s=s_i} = y(s_i) \qquad (i = 1 \ldots n) \qquad (1.7)$$

The remaining m equations needed to determine the unknowns are found by constraining the approximation (1.6) to satisfy the boundary conditions, i.e.

$$U_i(\tilde{x}) = \gamma_i \qquad (i = 1 \ldots m) \qquad (1.8)$$

Equations (1.7) and (1.8) together constitute n+m algebraic equations to be solved for the n+m constants. The best method of solution of these algebraic equations depends upon the form of the corresponding matrix, however Gaussian Elimination is very often the most suitable technique.

The process described so far has consisted of the appropriate steps (a), (b) and (c) of a general method discussed in Section 1.1.

Having determined the unknowns the approximate solution is then obtained at any point s by forming the sum (1.6).

The second approach which may yield the same answer as the former is to require that the approximation $\tilde{x}^*$ explicitly satisfies the boundary conditions (1.2b). That is, $\tilde{x}^*$ is sought in the form

$$\tilde{x}^*(s) = \xi(s) \sum_{j=1}^{n} a_j^* \psi_j(s) \qquad (1.9)$$

with the function $\xi(s)$ such that the equations $U_i(\tilde{x}^*) = \gamma_i$ $(i = 1 \ldots m)$ are automatically satisfied for all choices of constants $\{a_j^*\}_{j=1}^n$. If the conditions (1.2b) are complicated this may not be possible and it would be necessary to revert to the earlier approach. Further if these two representations are to furnish the same answer we must have that for any choices of $\{a_j\}$ and $\{a_j^*\}$ the two sets

$$\{\tilde{x}:\tilde{x} = \sum_{j=1}^{n+m} a_j\psi_j\} \cap \{\tilde{x}:U_i(\tilde{x}) = \gamma_i \ (i = 1 \ldots m)\}$$

and $\{\tilde{x}^*:\tilde{x}^* = \xi \sum_{j=1}^{n} a_j^*\psi_j\}$ are equivalent. With the representation (1.9) the same collocation points are used and the rest of the procedure is as before.

We shall now describe the usual manner in which the method is employed for our purposes. For example, suppose that (1.2a) is of even order $m = 2r$ over $[-1,1]$ and suppose that the end conditions (1.2b) are $x^{(i)}(-1) = x^{(i)}(+1) = 0$ $(i = 1 \ldots r)$.

We shall take the ψ_j as polynomials of degree $j-1$ $(j \geq 1)$ and the function $\xi(s)$ is taken as $(s^2 - 1)^r$ which satisfies the requirements. A popular representation of the form (1.9) is

$$\tilde{x}^* = (s^2 - 1)^r \sum_{j=0}^{n-1}{}' c_j T_j(s) \qquad (1.10)$$

where T_j is the Chebyshev polynomial of degree j and c_j is taken as a_{j+1}^* $(j = 0 \ldots n-1)$.

The symbol $\sum'$ means that the first term in the finite sum is to be halved. That is, the first term is now $\frac{c_o T_o}{2}$, this being a convenience and not a necessary condition.

It is now briefly demonstrated that this collocation process can be viewed as a projection method. This is considered in more detail in section 2.2.

Let Y be the space of continuous functions. $\tilde{Y}$ is to a large extent arbitrary and can be spanned by any n functions as long as the interpolation problem is soluble. With $\{s_i\}_{i=1}^n$ as the collocation points let ϕ be the projection $Y \to \tilde{Y}$ that maps each continuous function into its interpolant formed by interpolating at the nodes. That is, for a continuous function y, ϕy can be expressed as a combination of the n functions and is such that $(\phi y)(s_i) = y(s_i)$ $(i = 1 \ldots n)$. We do not specify the space X here but leave a more rigorous description until section 2.2. However we take $\tilde{X}$ as the set of functions of the form (1.10) and we see that both $\tilde{X}$ and $\tilde{Y}$ have dimension n. Then since the method requires that the residual vanish at the nodes, i.e. $(L\tilde{x}^* - y)\Big|_{s=s_i} = 0$ $(i = 1 \ldots n)$ this means that the polynomial of degree n-1 interpolating the residual at these n points must be identically zero, i.e. $\phi(L\tilde{x}^* - y) = 0$. Thus the approximation satisfies an equation of the form (1.5) showing that we have indeed a projection method. We have been fairly specific here but collocation is in fact a projection method under very general circumstances.

This concludes our description of the basic method.

The origins of the method are not clear but it seems that theoretical investigations relevent to collocation were first conducted in Russia by Kantorovich (1934,1948) although these have not been consulted. Other early results were obtained by Karpilovskaja (1953). In 1959 Kanorovich and Akilov (English transl. 1964) produced what is generally regarded as the major work on this and other topics, presenting convergence theorems for the approximate solution of a wide class of operator equations. Improved but more specific convergence results were achieved by Karpilovskaja (1963).

The use of a Chebyshev series in the approximation was considered by Lanczos (1938) and later other practical aspects and the application of the method to nonlinear problems were examined by Clenshaw and Norton (1963) and Wright (1964). A survey of the method of weighted residuals of which collocation is a particular case was given by Finlayson and Scriven (1966).

Theoretical results for nonlinear problems were later obtained by Vainikko (1965,1966,1969) with the paper in 1966 perhaps containing the most useful achievements. Other aspects of the method have been investigated by Shindler (e.g. 1969).

More recent studies of projection methods have been conducted by de Boor (1966), Phillips (1969,1972) and Coldrick (1972). Perhaps the most significant work within the last two years has been concerned with the use of splines in the approximation and the development of corresponding theoretical results. The main achievements are those of Lucas and Reddien (1972), Russell and

Shampine (1972) and the further advances of deBoor and Swartz (1973).

The numerical solution by collocation of linear partial differential equations has been investigated by Karpilovskaja (1970) who considers trignometric approximations and presents convergence results based on the theory of Kantorovich and Akilov (1964).

A theory of a different nature designed primarily for quadrature methods for integral equations has been developed by Anselone (1971) and in this thesis Anselone's work will emerge as a useful basis for further investigations.

1.4 Green's Functions

We now briefly introduce the idea of a Green's function. These functions will be used throughout to a great extent for both theoretical and practical purposes.

Consider for example the boundary value problem of (1.2a) subject to the homogeneous end conditions

$$U_i(x) = 0 \quad (i = 1 \ldots m) \tag{1.2c}$$

Then the Green's function $g(s,t)$, when it exists, is a function such that

$$x(s) = \int_a^b g(s,t)y(t)dt$$

This relationship has to hold for all continuous inhomogeneous terms $y(s)$.

The Green's function depends on the boundary conditions and knowledge of it enables us to invert the differential operator (1.2a) subject to the end conditions (1.2c).

By far the most common Green's function which we shall encounter is that for the differential operator $\frac{d^2}{ds^2}$ operating on x say, over [-1,1] subject to $x(-1) = x(+1) = 0$. The literature, for example Keller (1968, p.108) generally gives the Green's functions for interval [0,1] but when this is transformed to [-1,1] we have

$$g(s,t) = \begin{cases} \frac{1}{2}(s+1)(t-1) & s \leq t \\ \frac{1}{2}(s-1)(t+1) & s > t \end{cases}$$

For $s < t$ $\frac{\partial g}{\partial s}(s,t) = \frac{1}{2}(t-1)$ and for $s > t$ $\frac{\partial g}{\partial s}(s,t) = \frac{1}{2}(t+1)$.

We shall also have cause to use the quantities

$$\int_{-1}^{+1} |g(s,t)|dt \quad \text{and} \quad \int_{-1}^{+1} |\frac{\partial g}{\partial s}(s,t)|dt.$$

After elementary manipulation we obtain

$$\int_{-1}^{+1} |g(s,t)|dt = \frac{1}{2}(1 - s^2) \tag{1.11}$$

and

$$\int_{-1}^{+1} |\frac{\partial g}{\partial s}(s,t)|dt = \frac{1}{2}(1 + s^2) \tag{1.12}$$

1.5 Aim and Summary

Having introduced numerical methods for boundary value problems and considered certain preliminaries we now summarise the aim and content of this thesis.

As was mentioned in section 1.1 there are several methods for the numerical solution of differential equations. Having found an approximate solution by some means the following question arises. 'How good are our answers?'. This is the field of error analysis of which there are two basic types, 'a priori' which examines the error before the numerical problem is tackled and 'a posteriori' which is applied after the approximate solution has been computed and utilises this knowledge.

We are principally concerned with collocation methods and the literature cited in section 1.3 contains a considerable amount of work on error bounds which are usually expressed in more general functional analysis terms with the differential equation together with the boundary conditions treated as an operator equation. However most of these results are of an 'a priori' nature and are derived in terms of the inverse of the given operator. This approach leads to convergence and order of convergence proofs but is of little use if a computable bound on the error is required since knowledge of the inverse of the given differential operator is tantamount to knowing the true solution and is clearly not a very practical possibility.

There are some results of the converse 'a posteriori' nature but these seem to have remained as theoretical rather than practical bounds. It is the principal aim of this thesis to examine the 'a posteriori' theory and deduce, primarily for polynomial approximation, means of forming computable bounds which are subsequently applied

to sample two point boundary value problems. Much of the analysis given throughout is also pertinent to the numerical solution of Fredholm integral equations. Generally in these investigations the effect of rounding error is ignored, however at an appropriate stage relevent matrix condition numbers are given some consideration.

In Chapter 2 the functional analysis background for the theory is described and the main theorems are presented. In particular, the results of Kantorovich and Akilov (1964) are stated in a slightly simplified form for projection methods. These are followed by less involved but essentially similar theorems based on the work of Phillips (1969,1972) and Coldrick (1972). Finally the theory due to Anselone (1971) is summarised. The 'a posteriori' bounds given in these results are the object of our main investigation as a more practical approach is developed throughout the thesis.

The application of the theory for projection methods to the approximate solution by collocation of linear differential equations is considered in Chapter 3. Firstly it is demonstrated how to relate the numerical problem to the functional analysis setting and the study of 'a posteriori' approaches is motivated by examination of the 'a priori' results which are shown to be rather unsuitable. The main part of Chapter 3 is concerned with the 'a posteriori' bounds and various means of expressing these in terms of the inverse of the collocation matrix are examined. This investigation encounters awkward problems but eventually suitable results are achieved. During the course of this analysis interesting properties of certain matrices are

revealed and these are explored more fully in the final section.

In Chapter 4 the theory due to Anselone (1971) is studied and it is demonstrated how to 'extend' the collocation method to satisfy criteria necessary for the application of this theory. Again the problem of expressing the theoretical 'a posteriori' bounds in terms of computable quantities is successfully investigated. In the last section a generalisation of the earlier ideas is suggested.

Chapter 5 is concerned with the implementation on the machine of the computable bounds. The results derived in Chapters 3 and 4 based on the theorems of Chapter 2 are only applicable if a sufficiently large number of collocation points is employed. Actual values of this number presented later in the chapter for certain sample boundary value problems are sometimes found to be quite large and to avoid this difficulty more easily applicable estimates of the bounds are developed. In the last section the results of test applications of the different error bounding techniques are presented and compared with actual computed errors. This is followed by a discussion of certain pertinent points.

Chapter 6 examines certain areas where the analysis given might be usefully extended and ends by summarising appropriate conclusions to be drawn from this work.

This completes the summary of the thesis and for convenience we state here that all computations throughout

this work were performed on an IBM 360/67 computer using double length arithmetic.

CHAPTER 2

THEORY OF APPROXIMATION METHODS

2.1 Introduction

In this chapter we introduce the setting for certain operator equations and their approximate solution. In the former sections theorems based on the work of Kantorovich and Akilov (1964), Phillips (1969,1972) and Coldrick (1972) are given. These are both of an 'a priori' and an 'a posteriori' nature. In the latter sections theorems of a different type due to Anselone (1971) are presented.

These theorems are of a general nature with several possible areas of application. In later chapters we concentrate on the numerical solution by collocation methods of boundary value problems in ordinary differential equations, much of the analysis also being relevent for Fredholm integral equations. Other applications of the theory include Galerkin methods for both ordinary and partial differential equations and some of these topics are examined in Chapter 6.

We now introduce the background for the theory based on the work of Kantorovich and Akilov.

2.2 Setting for the Projection Method Theory

Let X and Y be normed linear spaces and let $\|\cdot\|$ and $\|\cdot\|_X$ denote the norms in the spaces Y and X respectively. Let [X,Y] denote the space of bounded linear operators mapping $X \to Y$ with the subordinate

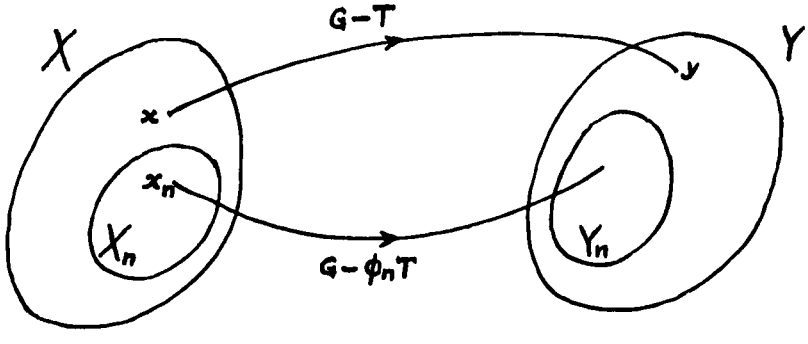
norm. Let X_n and Y_n be subspaces of X and Y respectively with ϕ_n a bounded linear projection $Y \to Y_n$. The subscript n will have significance later denoting the dimension of the subspaces but no restriction on dimensionality is made here.

Consider two equations

$$Gx - Tx = y \tag{2.1}$$

and $$Gx_n - \phi_n Tx_n = \phi_n y \tag{2.2}$$

where $x \epsilon X$, $x_n \epsilon X_n$ and $y \epsilon Y$. Here $G,T \epsilon [X,Y]$ and we further assume that G has a linear inverse and that G restricted to X_n establishes a bijection between X_n and Y_n. That is $G(X_n) = Y_n$ and $G^{-1}(Y_n) = X_n$. (2.1) is the given equation and we might seek an approximation to its solution x by solving (2.2) for $x_n \epsilon X_n$. (2.2) is the approximate equation and can be derived by seeking an $x_n \epsilon X_n$ such that $\phi_n\{(G-T)x_n - y\} = 0$ since $\phi_n Gx_n = Gx_n \epsilon Y_n$. An intuitive concept of the situation described is illustrated below.



$G - T: X \to Y$

$G - \phi_n T: X_n \to Y_n$

Note that $G - \phi_n T$ is regarded as being restricted to domain X_n.

Most of the theorems given later require the operators to satisfy the relationships above together with some extra conditions on the operators and spaces. However this is the basic setting and it is now shown that the numerical solution by collocation of a linear differential boundary value problem can be regarded in this way and we follow the description of Kantorovich and Akilov. For example, suppose we wish to solve the following $2m^{th}$ order linear equation over the interval $[-1,1]$:

$$L[x] \equiv \frac{d^{2m}x}{dt^{2m}} + p_{2m-1}(t)x^{(2m-1)}(t) + \ldots p_1(t)x^{(1)}(t)$$

$$+ p_0(t)x(t) = y(t) \qquad (2.3a)$$

subject to the boundary conditions

$$x^{(j)}(-1) = x^{(j)}(+1) = 0 \quad (j = 0 \ldots m-1) \qquad (2.3b)$$

In keeping with the description of section 1.3 we seek an approximate solution in the form

$$x_n(t) = (t^2 - 1)^m \sum_{k=0}^{n-1} a_k \psi_k(t) \qquad (2.4)$$

where $\{\psi_k\}_{k=0}^{n-1}$ are n independent polynomials of up to degree n-1. For example $\psi_k(t) = t^k$ or $\psi_k(t) = T_k(t)$ $(k = 0 \ldots n-1)$ could be selected. Let the chosen set of collocation points be $\{t_j\}_{j=1}^{n}$ and the method requires $L[x_n]|_{t=t_j} = y(t_j)$ $(j = 1 \ldots n)$. Let $C^{(q)}[-1,1]$ be

the space of functions which are q times continuously differentiable over [-1,1] with $C[-1,1] \equiv C^{(0)}[-1,1]$ and let B be the set of those continuous functions which satisfy the conditions (2.3b). Now define $X \equiv C^{(2m)}[-1,1] \cap B$ and let X_n be the space of functions of the form (2.4). Y is chosen as C[-1,1] and $Y_n \equiv P_{n-1}$ the set of polynomials of degree up to n-1. The projection ϕ_n is defined as the mapping projecting each continuous function into its unique interpolating polynomial at the collocation points. Define G and T by $Gx \equiv x^{(2m)}$ and $Tx \equiv -(p_{2m-1}x^{(2m-1)} + \ldots + p_1x^{(1)} + p_ox)$. Thus the differential equation (2.3a) plus the end conditions (2.3b) is equivalent to the operator equation $Gx - Tx = y$. Note that in principle G could be chosen differently but this would cause complications in the choice of subspaces and in knowledge of the inverse of G. This point is discussed again in Chapter 6. Kantorovich consideres a parameter λ in $(G - \lambda T)x = y$ but this is omitted explicitly for simplicity and can be considered as occurring in T. We choose the norm in the space Y as the infinity norm and the norm in X is chosen as $\|x\|_X = \|Gx\| = \|x^{(2m)}\|_\infty$ and we shall call this the X-norm. We require $p_i \epsilon C[-1,1]$ (i = 0, ... 2m-1) and this together with the above definition of $\|\cdot\|_X$ ensure that $G,T \epsilon [X,Y]$. This is shown later in more detail in section 3.2.

Clearly $G(X_n) = Y_n$. For $y \epsilon C[-1,1]$, $(G^{-1}y)(s) = \int_{-1}^{+1} g(s,t)y(t)dt$ where g(s,t) is the Green's function for the differential operator $\frac{d^{2m}}{dt^{2m}}$ subject to the conditions (2.3b) and is known explicitly. If

$\tilde{y} \epsilon Y_n$ then $G^{-1}\tilde{y} = \tilde{x}$ where $\tilde{x}$ is a polynomial of degree $2m + n - 1$ which must satisfy (2.3b) and so is of the form (2.4). Thus G is a bijection between X_n and Y_n.

As described in section 1.3 the application of the collocation method means that we seek an $x_n \epsilon X_n$ such that $(G - T)x_n|_{t=t_j} = y(t_j)$ $(j = 1 \ldots n)$. Thus $\phi_n\{(G - T)x_n - y\} = 0$ or $(G - \phi_n T)x_n = \phi_n y$ (since $Gx_n \epsilon Y_n$) and it has been shown that the approximate solution of a $2m^{th}$ order boundary value problem can be regarded in the functional analysis background given previously. As was mentioned earlier this is only one application of the theory and more general aspects are left until the final chapter.

2.3 Definitions of Compactness

Before proceeding to the statements of the theorems we introduce the concepts of compactness which will be used throughout this chapter. We follow the definitions given by Anselone (1971). Let S be a subset of a normed linear space X and let [X] be the space of bounded linear operators on X. Then S is compact iff every open cover of S has a finite subcover. S is said to be relatively compact iff the closure of S is compact. This situation differs slightly from that in Kantorovich and Akilov where, for sets, the term compact is equivalent to Anselone's relatively compact. The set S is sequentially compact iff each sequence in S has a convergent subsequence with the limit in X. The properties of relative and sequential compactness are equivalent.

Let U be the unit ball $\{z \in X : \|z\| \leq 1\}$ then $K \in [X]$ is compact iff the set KU is relatively compact (in X). This means in effect that a compact operator maps bounded sets onto relatively compact sets. This definition of a compact operator agrees with Kantorovich and Akilov's concept of a completely continuous operator.

2.4 The Theory of Kantorovich and Akilov

We now present in a slightly simplified form the theorems of Kantorovich and Akilov which apply to the solution of operator equations of the type (2.1) and (2.2) previously introduced. Firstly some further requirements must be satisfied. The norm in the space X is defined by $\|z\|_X = \|Gz\|$, $z \in X$. This is primarly for convenience in the theory but for the example of the approximate solution by collocation of differential boundary value problems is necessary to ensure bounded operators G and T (see sections 2.2 and 3.2). Subscripts on the norms $\|\cdot\|_X$ or $\|\cdot\|_Y$ will be used occasionally to clarify certain points. Also X_n and Y_n should be complete subspaces of X and Y respectively. This requirement holds trivially if X_n and Y_n are finite dimensional - see Brown and Page (1970, p.147).

The following three conditions are used:

I For every $z \in X$ there exists a $\tilde{y} \in Y_n$ such that
$\|Tz - \tilde{y}\| \leq \mu_1 \|z\|$ where μ_1 is independent of z.

II There exists an element $\tilde{y} \in Y_n$ such that
$\|y - \tilde{y}\| \leq \mu_2 \|y\|$ where μ_2 may depend on y.

III $G - \phi_n T$ satisfies the condition that the existence of a solution $\tilde{x}$ in X_n to $(G - \phi_n T)\tilde{x} = \tilde{y}$ for every $\tilde{y} \in Y_n$ implies its uniqueness.

Throughout the following four theorems $G - \phi_n T$ means $G - \phi_n T$ restricted to X_n and $(G - \phi_n T)^{-1}$ is an operation with domain Y_n.

We now state,

Theorem 1 (Kantorovich and Akilov)

If condition I holds, the linear operation $(G - T)^{-1}$ exists and $\delta = \mu_1 \| \phi_n (G - T) \| \| (G - T)^{-1} \| < 1$ then $(G - \phi_n T)\tilde{x} = \tilde{y}$ has a solution $\tilde{x}$ for all $\tilde{y} \epsilon Y_n$, with $\| \tilde{x} \| \leq \frac{D}{1-\delta} \| \tilde{y} \|$ where $D = (1 + \mu_1) \| (G - T)^{-1} \|$.

Further if condition III holds or in particular if $G^{-1}\phi_n T$ is a compact operator $\epsilon [X_n]$ then the linear operator $(G - \phi_n T)^{-1}$ exists and $\| (G - \phi_n T)^{-1} \| \leq \frac{D}{1-\delta}$ ■

Theorem 2 (Kantorovich and Akilov)

If conditions I, II and III are satisfied and equation (2.1) has the solution x then $\| x - x_n \| \leq \eta \| x \|$ where x_n is the solution of (2.2) and $\eta = (\mu_1 + \mu_2 \| (G - T) \|)(1 + \| (G - \phi_n T)^{-1} \phi_n (G - T) \|)$.

Alternatively if it is known that there exists an $\tilde{x} \epsilon X_n$ such that $\| x - \tilde{x} \| \leq \epsilon \| x \|$ then the above error bound holds without the use of conditions I and II, where now $\eta = \epsilon(1 + \| (G - \phi_n T)^{-1} \phi_n (G - T) \|)$.

If we have sequences of spaces X_n and Y_n $(n = 1,2...)$ with corresponding mappings then with the conditions of the theorem we have convergence in $\lim_{n \to \infty} \| x - x_n \| = 0$ provided $\lim_{n \to \infty} \mu_1 \| \phi_n \| = \lim_{n \to \infty} \mu_2 \| \phi_n \| = 0$ ■

Now a theorem of a slightly different nature is given.

Theorem 3 (Kantorovich and Akilov)

Given sequences of spaces X_n and Y_n and corresponding

approximate equations of the form (2.2) then if $(G - T)^{-1}$ exists, the space Y is complete, $\lim_{n\to\infty} \phi_n y = y$ $(y \epsilon Y)$ and $G^{-1}T$ is a compact operator $\epsilon[X]$ we have that the approximate equations are solvable for sufficiently large n and the approximate solutions converge to the exact solution■

The theorems presented so far are essentially of an 'a priori' nature. We now give a result which deduces information about the solubility of the given equation from the approximate equation.

Theorem 4 (Kantorovich and Akilov)

If the linear operation $(G - \phi_n T)^{-1}$ exists, condition I holds and $\delta = \mu_1(1 + \|(G - \phi_n T)^{-1}\phi_n(G - T)\|) < 1$ then $G - T$ has a linear left inverse with

$$\|(G - T)^{-1}\| < \frac{1 + \|(G - \phi_n T)^{-1}\phi_n\| + \|(G - \phi_n T)^{-1}\phi_n(G - T)\|}{1 - \delta}$$

Further if it is true that the uniqueness of the solution of equation (2.1) implies its solubility for every right hand side then the two-sided linear inverse $(G - T)^{-1}$ exists■

These then are the most relevent parts for our purposes of the theory of Kantorovich and Akilov. They actually consider a slightly more general situation with an operator $\tilde{T}\epsilon[X_n, Y_n]$ of which $\phi_n T$ restricted to X_n is a special case. However for the approximate solution of differential equations by collocation or Galerkin methods the theory reduces to this form.

The above results are proved by reducing the given and approximate equations to equations with operators mapping the space X into itself. This is done by applying G^{-1} to equations (2.1) and (2.2) to give $G^{-1}(G - T)x = G^{-1}y$

$$\Rightarrow \quad x - G^{-1}Tx = G^{-1}y \qquad (2.1')$$

and $x_n - G^{-1}\phi_n Tx_n = G^{-1}\phi_n y$ or

$$x_n - (G^{-1}\phi_n G)(G^{-1}T)x_n = (G^{-1}\phi_n G)G^{-1}y \qquad (2.2')$$

(2.1′) and (2.2′) are now in the form $(I - K)x = y_o$ and $(I - P_n K)x_n = P_n y_o$ with $y_o \epsilon X$, $K \epsilon [X]$, P_n a projection mapping $X \to X_n$ and I the identity operator on X. We shall consider later a similar process and shall not proceed further with this suffice it to say that once this form is achieved Kantorovich and Akilov then apply their theory for equations of the second kind to derive the theorems presented above.

It has been shown previously that the approximate solution by collocation methods of differential boundary value problems can be seen in the context of the theory. The extra conditions required for the application of the theorems are shown to the true in section 3.2 of the next chapter. Also an example of their 'a priori' application is considered proving the solubility of the approximate equation and finding the error bounds predicted by the theory.

2.5 Theory Developed from a More Recent Approach

We now however proceed to consider and modify recent work due to Coldrick (1972) of a related nature to that presented above. Similar investigations have been pursued by Phillips (1969,1972). The theory is designed for application to the numerical solution of integral equations but it is shown that this can be altered to prove results which can be later applied to the approximate solution of differential equations. This is achieved in a manner analagous to that which Kantorovich and Akilov use to reduce the equations (2.1) and (2.2) to the forms (2.1′) and (2.2′).

The approach seems less confusing than that of Kantorovich and leads to theorems of an 'a posteriori' character more suited to practical application than the theory above.

The setting for the theory initially described here is a normed linear space X (with norm denoted by $\|\cdot\|$) and $[X]$ is the space of bounded linear operators on X, with the subordinate norm. We now state a theorem which is standard when X is a Banach space but which is quoted from Coldrick (1972).

Theorem 5 (Coldrick (1972, p.14))

Let $K,L \in [X]$ and $(I - K)^{-1} \in [X]$. Suppose further that either K and L are compact or the linear space X is complete. Define $\delta = \|(I - K)^{-1}\| \|K - L\|$ and suppose $\delta < 1$, then $(I - L)^{-1} \in [X]$ and $\|(I - L)^{-1}\| \leq \frac{\|(I - K)^{-1}\|}{1 - \delta}$ ■

We are concerned with the approximate solution of an equation

$$(I - K)x = y \tag{2.5}$$

in X with $y \epsilon X$, I the identity operator on X and $K \epsilon [X]$ and we seek $x \epsilon X$. Let X_n be a subspace of X and P_n a linear projection mapping $X \to X_n$. We might hope to find an approximation x_n to x where $x_n \epsilon X_n$ by solving an approximate equation of the form

$$(I - P_n K)x_n = P_n y \quad \text{in } X_n. \tag{2.6}$$

With x_n satisfying (2.6) and seeking an x satisfying (2.5) we now give the following theorem.

Theorem 6 (This result is essentially given by Coldrick but with the roles of $I - K$ and $I - P_n K$ reversed).

Let X_n be a subspace of a normed linear space X and let P_n be a bounded linear projection mapping $X \to X_n$. Suppose that $K \epsilon [X]$ is compact and $(I - P_n K)^{-1} \epsilon [X]$. Then if $\delta_n = \|(I - P_n K)^{-1}\| \|(I - P_n)K\| < 1$ we have $(I - K)^{-1}$ exists $\epsilon [X]$ and

(a) $$\|(I - K)^{-1}\| \leq \frac{\|(I - P_n K)^{-1}\|}{1 - \delta_n},$$

(b) with x and x_n satisfying (2.5) and (2.6) respectively we have the error bound

$$\|x - x_n\| \leq \frac{\delta_n}{1 - \delta_n} \|x_n\| + \frac{\|(I - P_n K)^{-1}\|}{1 - \delta_n} \|(I - P_n)y\|.$$

This is a result of an 'a posteriori' nature. Notice that here $(I - P_n K)$ and $(I - P_n K)^{-1} \epsilon [X]$ and are not restricted to the subspace X_n.

Proof (a) Since K is compact and P_n is bounded, P_nK is compact - see Brown and Page (1970, p.245). Thus substituting P_nK for K and K for L in Theorem 5 we achieve the result (a).

(b) Thus there exists a unique x such that $(I - K)x = y$. Now $(I - K)(x - x_n) = y - (I - K)x_n$

$= y - P_ny + (K - P_nK)x_n$

$\Rightarrow x - x_n = (I - K)^{-1}(I - P_n)y + (I - K)^{-1}(I - P_n)Kx_n$

and (b) follows.

Corollary Let $\{X_n\}$ $(n = 1,2 \ldots)$ be a sequence of subspaces of the normed linear space X and let $\{P_n\}$ be a sequence of bounded, but not necessarily uniformly bounded, projections mapping $X \to X_n$ $(n = 1,2 \ldots)$. Suppose that for $n > n_o$, $(I - P_nK)^{-1}$ exists $\epsilon[X]$ and that for $n > n_1 > n_o$ $\delta_n = \|(I - P_nK)^{-1}\| \|(I - P_n)K\| < 1$, then $(I - K)^{-1}$ exists $\epsilon[X]$ and for $n > n_1$ (a) and (b) provide different bounds on $\|(I - K)^{-1}\|$ and error bounds for $\|x - x_n\|$ respectively■

So far for Theorems 5 and 6 we have only considered operators in one space X only. This situation is applied to integral equations of Fredholm type by Coldrick. Similar application is also considered by Kantorovich and Akilov.

We now consider two spaces X and Y with subspaces X_n and Y_n and exactly as described at the start of this chapter we wish to solve approximately a given equation of the form (2.1), namely $(G - T)x = y$ by means of an approximate equation of the form (2.2). The operators ϕ_n, G, T and their properties together with the rest of the setting is precisely as described earlier in section 2.2. It was shown that the numerical solution by collocation of a boundary value differential equation could be seen in this

light. To derive results analagous to those of Theorem 6 which can be applied to the approximate solution of differential equations we reduce the equations (2.1) and (2.2) to the form (2.1′) and (2.2′), wholly in X. This process is carried out by Kantorovich and Akilov to prove their results for the approximate solution of $(G - T)x = y$ and was mentioned briefly before. This is now described in more detail.

By operating on the left throughout equations (2.1) and (2.2) we derived the equations (2.1′) and (2.2′), namely $(I - G^{-1}T)x = G^{-1}y$ and $[I - (G^{-1}\phi_n G)(G^{-1}T)]x_n = (G^{-1}\phi_n G)G^{-1}y$ respectively. Since $G^{-1}: Y \to X$, $y_o = G^{-1}y \in X$ and $G^{-1}T$ maps $X \to X$. For $z \in X$, $Gz \in Y \Rightarrow \phi_n Gz \in Y_n \Rightarrow G^{-1}\phi_n Gz \in X_n$ and also $(G^{-1}\phi_n G)(G^{-1}\phi_n G) = G^{-1}\phi_n G$ thus proving that $G^{-1}\phi_n G$ is a projection from $X \to X_n$. Further if the norms in the spaces X and Y are related by $\|z\|_X = \|Gz\|_Y$ where $\|\cdot\|_Y$ represents the norm in the space Y then $\|G\| = \|G^{-1}\| = 1$ and if $T \in [X,Y]$ then $G^{-1}T\,[X]$. Thus writing $G^{-1}T$ as K, $G^{-1}\phi_n G$ as P_n and $G^{-1}y$ as y_o (2.1′) and (2.2′) are of the forms (2.5) and (2.6) respectively. Therefore transforming (2.1) and (2.2) in this way the situation is precisely as described before the statement of Theorem 6 which can now be used to give results for the approximate solution of (2.1). We now give

Theorem 7

Let X_n and Y_n be subspaces of X and Y respectively and let ϕ_n be a bounded linear projection mapping $Y \to Y_n$. Suppose that $T \in [X,Y]$ and $G^{-1}T$ is compact $\in [X]$. Suppose further that $(G - \phi_n T)^{-1} \in [Y,X]$ and $\delta_n = \|(G - \phi_n T)^{-1}\| \|(I - \phi_n)T\| < 1$. Then

$(G - T)^{-1}$ exists $\epsilon[Y,X]$ and (a) $\|(G - T)^{-1}\| \leq \dfrac{\|(G - \phi_n T^{-1}\|}{1 - \delta_n}$.

With x and x_n satisfying (2.1) and (2.2) respectively we have (b) $\|x - x_n\| \leq \dfrac{\delta_n}{1 - \delta_n}\|x_n\| + \dfrac{\|(G - \phi_n T)^{-1}\|}{1 - \delta_n}\|(I - \phi_n)y\|$ or more simply $\|x - x_n\| \leq \dfrac{\|(G - \phi_n T)^{-1}\|}{1 - \delta_n}\|(G - T)x_n - y\|$.

Notice that $G - \phi_n T$ and $(G - \phi_n T)^{-1}$ are regarded as operators between the whole spaces and not the subspaces.

Proof It was shown above how equations (2.1) and (2.2) could be transformed to the forms (2.5) and (2.6) and so with these relationships we have to show that the conditions of Theorem 6 are satisfied. We have $K \sim G^{-1}T$ and $P_n \sim G^{-1}\phi_n G$ and so $\|P_n\| = \|G^{-1}\phi_n G\| \leq \|\phi_n\|$. Thus if ϕ_n is bounded so also is P_n. Now $(I - P_n K)^{-1} \sim (I - G^{-1}\phi_n G G^{-1}T)^{-1} = (G - \phi_n T)^{-1}G$
$\Rightarrow (I - P_n K)^{-1}\epsilon[X]$. Also $\delta_n = \|(G - \phi_n T)^{-1}\|\|(I - \phi_n)T\| < 1$
$\Rightarrow \|(G - \phi_n T)^{-1}G\|\|G^{-1}(I - \phi_n GG^{-1})T\| < 1$
$\Rightarrow \|(I - P_n K)^{-1}\|\|(I - P_n)K\| < 1$ and this is the condition required for Theorem 6. Thus $(I - K)^{-1}$ exists $\epsilon[X]$
$\Rightarrow (I - G^{-1}T)^{-1}G^{-1} = (G - T)^{-1}$ exists $\epsilon[Y,X]$ and the results (a) and (b) follow on substitution. The latter result of (b) is derived from $(G - T)(x - x_n) = y - (G - T)x_n$ which implies $(x - x_n) = (G - T)^{-1}(y - (G - T)x_n)$ where $y - (G - T)x_n$ is the residual on substitution of x_n into the given equation■

In Theorem 7 we have occurring the quantity $\|(G - \phi_n T)^{-1}\|$ where $(G - \phi_n T)^{-1}\epsilon[Y,X]$ not $[Y_n,X_n]$. However we can employ the following argument to utilise $(G - \phi_n T)^{-1}$ restricted to Y_n, i.e. $(G - \phi_n T)^{-1}_{Y_n}$. $(G - \phi_n T)^{-1}(G - \phi_n T) = I$

$$\Rightarrow \quad (G - \phi_n T)^{-1} G = I + (G - \phi_n T)^{-1} \phi_n T$$

$$\Rightarrow \quad (G - \phi_n T)^{-1} = G^{-1} + (G - \phi_n T)^{-1} \phi_n T G^{-1}$$

Thus

$$\| (G - \phi_n T)^{-1} \| \leq 1 + \| (G - \phi_n T)_{Y_n}^{-1} \| \, \| \phi_n T \| \qquad (2.7)$$

Using (2.7) we can now state

Corollary If the conditions of Theorem 7 are satisfied and if B_n denotes $\| (G - \phi_n T)_{Y_n}^{-1} \|$ then provided $\delta_n = (1 + B_n \| \phi_n T \|) \| (I - \phi_n) T \| < 1$ we have the bound

$$\| (G - T)^{-1} \| \leq \frac{1 + B_n \| \phi_n T \|}{1 - (1 + B_n \| \phi_n T \|) \| (I - \phi_n) T \|} \quad \blacksquare$$

With this result the error bounds (b) of Theorem 7 may then be employed.

Phillips (1969,1972) considering in particular integral equations has presented similar results to those given above but does not use them in practice. We intend that these bounds be applied in an 'a posteriori' manner to the approximate solution by collocation of linear ordinary differential equations. The conditions required for these results are shown to hold in the section 3.5 of the next chapter. The bounds are then calculated by finding a bound on $B_n = \| (G - \phi_n T)_{Y_n}^{-1} \|$ in terms of the inverse matrix from the collocation equations.

Clearly 'a priori' bounds analagous to those of Theorem 7 could be given. Roughly, if $(G - T)^{-1}$ is known to exist then for sequences of subspaces with corresponding mappings, if $\lim_{n \to \infty} \| (I - \phi_n) T \| = 0$ and $\lim_{n \to \infty} \| (I - \phi_n) y \| = 0$ then with n sufficiently large $(G - \phi_n T)^{-1}$ exists $\epsilon [Y, X]$

and we have the 'a priori' error bound

$$\|x - x_n\| \leq \frac{\delta_n}{1 - \delta_n}\|x\| + \frac{\|(G - T)^{-1}\|}{1 - \delta_n}\|(I - \phi_n)y\| \qquad (2.8)$$

where $\delta_n = \|(G - T)^{-1}\|\,\|(I - \phi_n)T\|$ and $\lim_{n\to\infty} \|x - x_n\| = 0$.

2.6 Connections between the Conditions for 'a priori' Error Bounds

Theorem 2 due to Kantorovich and Akilov (1964) requires the conditions

I For every $z \in X$ there exists a $\tilde{y} \in Y_n$ such that

$\|Tz - \tilde{y}\| \leq \mu_1\|z\|$ and

II There exists an element $\tilde{y} \in Y_n$ such that

$\|y - \tilde{y}\| \leq \mu_2\|y\|$ and for convergence they demand $\lim_{n\to\infty} \mu_1\|\phi_n\| = \lim_{n\to\infty} \mu_2\|\phi_n\| = 0$.

The result (2.8) requires

$$\lim_{n\to\infty} \|(I - \phi_n)T\| = 0 \qquad (2.9a)$$

and $$\lim_{n\to\infty} \|(I - \phi_n)y\| = 0 \qquad (2.9b)$$

Suppose that the conditions (2.9) hold. Then (2.9a) $\Rightarrow \sup_{z\neq 0} \frac{\|(T - \phi_n T)z\|}{\|z\|} \to 0$ as $n \to \infty$ and for each n let

$\sup_{z\neq 0} \frac{\|(T - \phi_n T)z\|}{\|z\|} = \eta_n$. Thus for all $z \in X$

$\|(T - \phi_n T)z\| \leq \eta_n\|z\|$ and letting $\tilde{y} = \phi_n Tz$ we have $\|Tz - \tilde{y}\| \leq \eta_n\|z\|$, which is condition I required by Kantorovich and Akilov.

If (2.9b) holds then with $\zeta_n = \|y - \phi_n y\|$ we have $\lim_{n\to\infty} \zeta_n = 0$. Now for $y \neq 0$ it is true that $\|y - \phi_n y\| = \frac{\zeta_n}{\|y\|}\|y\|$. Thus writing $\xi_n = \frac{\zeta_n}{\|y\|}$ $\|y - \phi_n y\| \leq \xi_n \|y\|$ (for all y) and with $\tilde{y} = \phi_n y$ $\|y - \tilde{y}\| \leq \xi_n \|y\|$ which is condition II.

Conversely if the conditions of Kantorovich and Akilov hold then I implies that for all $z \epsilon X$ there exists a $\tilde{y} \epsilon Y_n$ such that $\|Tz - \tilde{y}\| \leq \mu_1 \|z\|$. Thus $\|Tz - \phi_n Tz\| = \|Tz - \tilde{y} + \tilde{y} - \phi_n Tz\| \leq \|Tz - \tilde{y}\| + \|\phi_n(\tilde{y} - Tz)\| \leq (1 + \|\phi_n\|)\mu_1 \|z\|$

$$\Rightarrow \frac{\|Tz - \phi_n Tz\|}{\|z\|} \leq \mu_1(1 + \|\phi_n\|) \qquad (z \neq 0)$$

$$\Rightarrow \sup_{z\neq 0} \frac{\|Tz - \phi_n Tz\|}{\|z\|} \leq \mu_1(1 + \|\phi_n\|)$$

$$\Rightarrow \|T - \phi_n T\| \leq \mu_1(1 + \|\phi_n\|) \to 0 \quad \text{as } n \to \infty$$

$$\text{if } \lim_{n\to\infty} \mu_1 \|\phi_n\| = 0.$$

If condition II holds then similarly $\|y - \phi_n y\| \leq (1 + \|\phi_n\|)\|y - \tilde{y}\| \leq \mu_2 \|y\| (1 + \|\phi_n\|)$ and if $\lim_{n\to\infty} \mu_2 \|\phi_n\| = 0 \Rightarrow \lim_{n\to\infty} \|y - \phi_n y\| = 0$.

This shows the relationship between the two sets of conditions required for convergence.

2.7 Background for Anselone's Theory

We now present a different theory for approximation methods due to Anselone (1971). As earlier in the chapter for the theory of Kantorovich and Akilov (1964) we first introduce the background for the results and then state the

theorems. These will be shown later in sections 4.2 and 4.3 to be suitable for application to the approximate solution by collocation of differential equations and will be used in practice in Chapter 5.

Thus following Anselone, let X be a real Banach space with [X] the Banach space of bounded linear operators on X with the subordinate norm and I as the identity operator on X.

Pointwise convergence of an operator sequence $\{S_n\}$ with $S_n \in [X]$ $(n \geq 1)$ to $S \in [X]$ is denoted by $S_n \to S$ and is defined by the requirement that $S_n z \to Sz$ as $n \to \infty$ for all $z \in X$. This is different from convergence in norm which means $\|S_n - S\| \to 0$. Anselone's theory uses the weaker pointwise convergence but requires that sequences of operators $\{S_n\}$ which will be used in some sense as approximations to a given operator satisfy additional compactness conditions. In section 2.3 the term compact applied to a single operator was defined. Anselone utilises an extension of this concept defined in the following manner. A set $V \subset [X]$ is <u>collectively compact</u> iff the set $VU = \{Sz : S \in V, z \in U\}$ is relatively compact, where U is the unit ball $\{z \in X : \|z\| \leq 1\}$. A sequence of operators in [X] is <u>collectively compact</u> iff the corresponding set is.

Before presenting the theorems we describe the types of equations to which they are applied. Let $y \in X$ and $K, K_n \in [X]$. We are concerned with the approximate solution of a given operator equation

$$(I - K)x = y \tag{2.10}$$

where the true solution x is given by $x = (I - K)^{-1}y$ when

the inverse exists. An approximation $x_n \epsilon X$ to x is sought satisfying an equation of the form

$$(I - K_n)x_n = y \qquad (2.11)$$

and $x_n = (I - K_n)^{-1}y$ when $I - K_n$ has an inverse. With this setting we are now in a position to state the theoretical results of Anselone which deal with sequences of approximations of the form (2.11) to the given equation (2.10).

2.8 Convergence Theorems and Error Bounds for Methods using a Sequence of Collectively Compact Operators to Approximate a Given Operator

Theorem 8 (Anselone (1971, p.10))

Let $K, K_n \epsilon [X]$ $(n = 1,2, \ldots)$ and assume that the three conditions, $K_n \to K$, K is compact and $\{K_n\}$ is collectively compact are satisfied. Suppose $(I - K)^{-1}$ exists and define $\Delta_n = \|(I - K)^{-1}\| \|(K_n - K)K_n\|$. Then $\Delta_n \to 0$ as $n \to \infty$ and for $\Delta_n < 1$, $(I - K_n)^{-1}$ exists $\epsilon [X]$ with

$$\|(I - K_n)^{-1}\| \leq \frac{1 + \|(I - K)^{-1}\| \|K_n\|}{1 - \Delta_n}.$$

Error bounds are given by

$$\text{(i)} \quad \|x - x_n\| \leq \frac{(1 + \|(I - K)^{-1}\| \|K_n\|)\|y - (I - K_n)x\|}{1 - \Delta_n} \quad \text{or}$$

$$\text{(ii)} \quad \|x - x_n\| \leq \frac{\|(I - K)^{-1}\| \|K_n y - Ky\| + \Delta_n \|x\|}{1 - \Delta_n} \quad \text{giving}$$

$\|x - x_n\| \to 0$ ■

This is a result of an 'a priori' nature since it depends upon knowledge of $(I - K)^{-1}$. Theorems 1, 2 and 3 require analagous assumptions. However we are primarily concerned in this thesis with, hopefully computable, ' a posteriori' error bounds and these are furnished by Theorems 9 and 10 to follow.

Theorem 9 (Anselone (1971, p.11))

Let $K, K_n \in [X]$ $(n = 1,2, \ldots)$ and assume that the same three conditions hold, namely $K_n \to K$, K is compact and $\{K_n\}$ is collectively compact. Whenever $(I - K_n)^{-1}$ exists define $\Delta^n = \|(I - K_n)^{-1}\| \|(K_n - K)K\|$. If for a particular value of n, such that $(I - K_n)^{-1}$ exists, we have $\Delta^n < 1$ then $(I - K)^{-1}$ exists with

$$\|(I - K)^{-1}\| \leq \frac{1 + \|(I - K_n)^{-1}\| \|K\|}{1 - \Delta^n}.$$

Error bounds are given by

$$\text{(i)} \quad \|x - x_n\| \leq \frac{(1 + \|(I - K_n)^{-1}\| \|K\|)\|y - (I - K)x_n\|}{1 - \Delta^n}$$

where $y - (I - K)x_n$ is the residual or

$$\text{(ii)} \quad \|x - x_n\| \leq \frac{\|(I - K_n)^{-1}\| \|K_n y - K y\| + \Delta^n \|x_n\|}{1 - \Delta^n} \quad \blacksquare$$

Nothing has so far been said concerning the uniform boundedness of the $(I - K_n)^{-1}$ or the possibility of convergence as $n \to \infty$. However having obtained by the above result that $(I - K)^{-1}$ exists we can then apply Theorem 8 to show that

$(I - K_n)^{-1}$ exists for all n sufficiently large and that its norms are uniformly bounded. Furthermore $\|x - x_n\| \to 0$ as $n \to \infty$ and the properties of a collectively compact sequence (Anselone (1970, p.8)) give $\Delta^n \to 0$. These deductions ensure that the estimates from Theorem 9 for $\|(I - K)^{-1}\|$ are uniformly bounded with respect to n as $n \to \infty$.

We shall later use the following generalisation which is a simple extension based on suggestions by Anselone.

Theorem 10

Let the operators K, K_n (n = 1,2, ...) satisfy the hypothesis of Theorem 9. Now however when $(I - K_n)^{-1}$ exists define $\Delta_d^n = \|(I - K_n)^{-1}\| \|(K_n - K)K^d\|$ (d integer ≥ 1) and if for a particular n $(I - K_n)^{-1}$ exists and $\Delta_d^n < 1$ then $(I - K)^{-1}$ exists with

$$\|(I - K)^{-1}\| \leq \frac{1 + \|K\| + \ldots + \|K^{d-1}\| + \|(I - K_n)^{-1}\| \|K^d\|}{1 - \Delta_d^n} .$$

The simplest error bound is $\|x - x_n\| \leq \|(I - K)^{-1}\| \|y - (I - K)x_n\|$ where $\|(I - K)^{-1}\|$ is bounded by the above expression■

As was mentioned earlier it will be shown in Chapter 4 (sections 4.2 and 4.3) that the approximate solution by collocation of linear differential equations can be modified so as to satisfy the criteria for Theorems 8, 9 and 10 and practical results will be given in Chapter 5.

CHAPTER 3

APPLICATION OF PROJECTION METHOD THEORY

3.1 Introduction

In this chapter we consider the application to the numerical solution of differential equations of the projection method theory given in sections 2.2-2.5 of the previous chapter. Firstly it is demonstrated that the solution by collocation of ordinary differential boundary value problems does indeed satisfy the conditions for the theory of Kantorovich and Akilov (1964). Next the 'a priori' approach is examined by example and it is shown that this is unsatisfactory not only because it requires knowledge of the inverse of the given operator but also due to the fact that error bounds are predicted which are far too conservative. An alternative approach is suggested which for fairly simple problems leads to improvements. The main part of the chapter is concerned with applying the 'a posteriori' results for projection method solution and the major problem is finding a realistic computable bound on the norm of the inverse of the approximate operator, i.e. a bound on $\| (G - \phi_n T)_{Y_n}^{-1} \|$ from the inverse collocation matrix. The 'a priori' theory predicts, subject to certain conditions, that these quantities be uniformly bounded as n increases but to devise practical bounds is seen to be an awkward problem. Interesting computational properties of matrices involved are examined and finally the use of row and column scaling to improve condition numbers is considered.

3.2 Application of Kantorovich and Akilov Theory to Boundary Value Problems

In section 2.2 it was shown in keeping with Kantorovich and Akilov how the approximate solution by collocation of an ordinary differential boundary value problem could be set in the functional analysis background for the theory.

Let us briefly remind ourselves of the situation described earlier. The example chosen was

$$\frac{d^{2m}x}{dt^{2m}} + P_{2m-1}(t)x^{(2m-1)}(t) + \ldots + P_1(t)x^{(1)}(t) + P_o(t)x(t) = y(t) \qquad (3.1a)$$

over say $[-1,1]$ subject to

$$x^{(j)}(-1) = x^{(j)}(+1) = 0 \quad (j = 0 \ldots m-1) \qquad (3.1b)$$

The $P_i(t)$ are assumed to be at least continuous $(i = 0 \ldots 2m-1)$. An approximation x_n of the form

$$x_n(t) = (t^2 - 1)^m \sum_{r=0}^{n-1} a_r \psi_r(t), \qquad (3.2)$$

where the $\psi_r(t)$ are polynomials of up to degree n-1, was sought by collocation at the n points $\{t_k\}_{k=1}^n$. The space X was chosen as the space of functions in $C^{(2m)}[-1,1]$ satisfying (3.1b) with X_n the subspace of functions of the form (3.2). Y was the space of continuous functions with Y_n as the space of polynomials of degree n-1. (3.1a) (3.1b) were shown to be equivalent to an operator equation of the form

$$(G - T)x = y \tag{3.3}$$

between the spaces X and Y, with $Gx = \frac{d^{2m}x}{dt^{2m}}$ and $Tx = -(P_{2m-1}x^{(2m-1)} + \ldots + P_o x)$. G is a bijection between X_n and Y_n and G^{-1} exists $\epsilon[Y,X]$. The approximate solution x_n satisfies the equation

$$(G - \phi_n T)x_n = \phi_n y \tag{3.4}$$

between X_n and Y_n where ϕ_n can be taken to be the projection mapping each continuous function to its interpolating polynomial of degree n-1 at the collocation points.

There is more than one choice of norm for the space Y e.g. l_∞, l_2 etc. but we shall use the infinity norm. In order that G,T be in [X,Y], and in particular be bounded we take the norms in the spaces X and Y to be related by $\|z\|_X = \|Gz\|_Y = \|z^{(2m)}\|_\infty$ $(z\epsilon X)$ and this point is considered shortly. We shall continue on occasions to use subscripts to emphasise with which norms we are dealing.

In order to apply their theory Kantorovich and Akilov show that the conditions we gave as I, II and III in section 2.4 hold. This is now described.

For $z\epsilon X$ we can say

$$z^{(j)}(s) = \int_{-1}^{+1} \frac{\partial^j g}{\partial s^j}(s,t)z^{(2m)}(t)dt \quad (j = 0 \ldots 2m-1)$$

where g(s,t) is the Green's function for the operator $\frac{d^{2m}}{dt^{2m}}$ subject to the homogeneous conditions (3.1b). Thus (Tz)(s) can be expressed as

$$-\{p_{2m-1}(s)\int_{-1}^{+1}\frac{\partial^{2m-1}g}{\partial s^{2m-1}}(s,t)z^{(2m)}(t)dt + \ldots$$

$$p_o(s)\int_{-1}^{+1} g(s,t)z^{(2m)}(t)dt\}$$

or

$$(Tz)(s) = \int_{-1}^{+1} k(s,t)z^{(2m)}(t)dt \qquad (3.5)$$

where

$$k(s,t) = -(p_{2m-1}(s)\frac{\partial^{2m-1}g}{\partial s^{2m-1}}(s,t) + \ldots$$

$$+ p_o(s)g(s,t)).$$

Since $k(s,t)$ has only a jump discontinuity at $s = t$ and $p_j(s)$ is continuous over $[-1,1]$ $(j = 0 \ldots 2m-1)$ we can be sure that $k(s,t)$ is bounded and integrable. Thus

$$|(Tz)(s)| \leq \int_{-1}^{+1} |k(s,t)|dt\|z^{(2m)}\|_\infty$$

and $\|Tz\| \leq k_o\|z\|_X$ giving T as a bounded operator with our choice of norms. (This verifies $T\epsilon[X,Y]$ as was mentioned in section 2.2).

Now

$$\frac{d}{dt}(Tz) = -\frac{d}{dt}(\sum_{i=0}^{2m-1} p_i z^{(i)})$$

$$= -\sum_{i=0}^{2m-1}(p_i' z^{(i)} + p_i z^{(i+1)})$$

provided

$$p_i(t) \epsilon C^{(1)}[-1,1] \quad (i = 0 \ldots 2m-1) \qquad (3.6)$$

Thus $\|(Tz)'\|_\infty \leq k_1\|z^{(2m)}\|_\infty$ for some constant k_1.

Therefore by Jackson's Theorem (Cheney (1966, p.147)) there exists a $\tilde{y} \epsilon Y_n$, i.e. a polynomial of degree n-1, such that $\|Tz - \tilde{y}\| \leq \frac{\pi}{2} \frac{k_1}{n} \|z^{(2m)}\|_\infty = \frac{\pi}{2} \frac{k_1}{n} \|z\|_X$ and condition I holds with

$$\mu_1 = \frac{\pi k_1}{2n} \tag{3.7}$$

Remark The assumption (3.6) is an important one and will be referred to later in this chapter in connection with a bound on the norm of the inverse of the approximate operator■

For condition II we can say that there exists a $\tilde{y} \epsilon Y_n$ (Cheney (1966, p.147)) such that

$$\|y - \tilde{y}\| \leq (\frac{\pi}{2})^k \frac{\|y^{(k)}\|}{n(n-1) \ldots (n-k+1)} \quad \text{if } y \epsilon C^{(k)}[-1,1] \quad (1 \leq k \leq n-1).$$

Thus $\|y - \tilde{y}\| \leq \mu_2 \|y\|$ where

$$\mu_2 = (\frac{\pi}{2})^k \frac{\|y^{(k)}\|}{n(n-1) \ldots (n-k+1)\|y\|} \tag{3.8}$$

and hence condition II holds.

If we can find a solution $\tilde{x} \epsilon X_n$ to $(G - \phi_n T)\tilde{x} = \tilde{y}$ for every $\tilde{y} \epsilon Y_n$ then this means there exists at least one set of coefficients a_o, a_1 ... a_{n-1} for every right hand vector in the linear collocation equations. But if the algebraic equations have a solution for every right hand vector it is well known that the solutions are unique. Thus there exists a unique $\tilde{x}$ such that $(G - \phi_n T)\tilde{x} = \tilde{y}$ for every $\tilde{y} \epsilon Y_n$, giving condition III. If Chebyshev zeros are used as collocation points we have $\|\phi_n\| \leq 8 + \frac{4}{\pi} \ln(n)$ (Natanson (1965, p.48)) whereas if Gauss points are employed

$\|\phi_n\| = O(n^{\frac{1}{2}})$.† So in either case provided the coefficients and right hand side in the differential equation have at least one continuous derivative we have $\lim_{n\to\infty} \mu_1\|\phi_n\| = 0$ and $\lim_{n\to\infty} \mu_2\|\phi_n\| = 0$.

We are now in a position to apply the theorems of Kantorovich and Akilov and in particular from Theorems 1 and 2 we have for sufficiently large n, that the inverse of the approximate operator exists. Further the approximate solutions converge to the exact solution with an error bound of at worst $O(\frac{\ln(n)}{n})$ for Chebyshev points or $O(n^{-\frac{1}{2}})$ for Gauss points. If $p_{2m-1} = p_{2m-2} = \ldots = p_{2m-k} = 0$ $(k \geq 1)$ and $y \epsilon C^{(j)}[-1,1]$ $(j \geq 2)$ then higher order convergence is guaranteed.

3.3 An 'a priori' Example

We now consider in some detail the 'a priori' application of the theory to a particular example to derive numerical bounds on the norms of the inverse operators and errors involved. These bounds hold for the number of collocation points being sufficiently large and these values of n are noted. The results predicted by this 'a priori' theory can be compared to those from an 'a posteriori' approach. (See TABLE 22).

The example examined is the problem

$$\frac{d^2x}{dt^2} - \lambda^2 x = y(t) \qquad (\lambda \text{ real}, > 0) \tag{3.9a}$$

with

$$x(-1) = x(+1) = 0 \tag{3.9b}$$

† Natanson (1965), p. 55.

Thus here $Gx = \frac{d^2x}{dt^2}$ and $Tx = \lambda^2 x$. The theoretical results are independent of the particular bases used for X_n and Y_n and depend primarily on the approximating properties of the subspaces. We shall be concerned with an approximation of the form .

$x_n(t) = (t^2 - 1) \sum_{r=0}^{n-1} a_r \psi_r(t)$ where the $\{\psi_r(t)\}_{r=0}^{n-1}$ are then any independent set of polynomials of up to degree n-1. The n collocation points used will be the zeros of the Chebyshev polynomial of degree n and Y_n will be the space of polynomials of degree n-1.

We shall use Theorem 1 to find the values of n required for applicability and also to bound the norm of the inverse of the approximate operator. Theorem 2 then gives the appropriate error bounds. All quantities occurring in Theorems 1 and 2 must therefore be bounded.

By Jackson's theorem (Cheney (1966, p.147)) there exists a $\tilde{y} \epsilon Y_n$ such that

$\|Tx - \tilde{y}\| = \|\lambda^2 x - \tilde{y}\| \leq (\frac{\pi}{2})^2 \frac{\lambda^2 \|x''\|}{n(n-1)}$. We can therefore choose $\mu_1 = (\frac{\pi}{2})^2 \frac{\lambda^2}{n(n-1)}$. If $y \epsilon C^{(k)}[-1,1]$ μ_2 is given by (3.8).

Examining the statements of the theorems it is seen that $\|\phi_n\|$, $\|(G - T)\|$, $\|(G - T)^{-1}\|$ and $\|(G - \phi_n T)^{-1}_{Y_n}\|$ have still to be bounded.

$$\|\phi_n\| \leq 8 + \frac{4}{\pi} \ln(n) \text{ by Natanson (1965, p.48)} \qquad (3.10)$$

$\|(G - T)\| = \sup_{\substack{x \epsilon X \\ \|x\|_X = 1}} \{\|(G - T)x\|_Y\}$. Now

$$(G-T)x(s) = x''(s) - \lambda^2 x(s) = x''(s) - \lambda^2 \int_{-1}^{+1} g(s,t)x''(t)dt$$

where $g(s,t)$ is the simple Green's function of section 1.4 for $\frac{d^2x}{dt^2}$ over $[-1,1]$ with the conditions (3.9b). Thus $\|(G-T)\| \leq (1+\lambda^2 \max_s \int_{-1}^{+1} |g(s,t)|dt)\|x''\|_\infty$. Therefore by (1.11)

$$\|(G-T)\| \leq 1 + \frac{\lambda^2}{2} \max_s (1 - s^2) = 1 + \frac{\lambda^2}{2} \tag{3.11}$$

We now show how to find $\|(G-T)^{-1}\|$.

$$\|(G-T)^{-1}\| = \sup_{\|y\|=1} \|(G-T)^{-1}y\|_X = \sup_{\|y\|=1} \|\frac{d^2}{ds^2} (G-T)^{-1}y(s)\|_\infty.$$

If $g_\lambda(s,t)$ is the Green's function for $x'' - \lambda^2 x$ over $[-1,1]$ subject to (3.9b) then

$$\|(G-T)^{-1}\| = \sup_{\|y\|_\infty=1} \|\frac{d^2}{ds^2} \int_{-1}^{+1} g_\lambda(s,t)y(t)dt\|_\infty.$$ Keller (1968, p.108) gives the Green's function for $x'' - \lambda^2 x$ over $[0,1]$ and on transformation to $[-1,1]$ we have

$$g_\lambda(s,t) = \frac{1}{\lambda \sinh 2\lambda} \begin{cases} \sinh\lambda(s+1)\sinh\lambda(t-1) & s \leq t. \\ \sinh\lambda(s-1)\sinh\lambda(t+1) & s \geq t. \end{cases}$$

To find $\frac{d^2}{ds^2} \int_{-1}^{+1} g_\lambda(s,t)y(t)dt$ we could split the range of integration and differentiate under the integral sign. However it is quicker to notice that from the differential equation

$$\frac{d^2}{ds^2} (G-T)^{-1}y(s) - \lambda^2 (G-T)^{-1}y(s) = y(s) \quad \text{and so}$$

$$\frac{d^2}{ds^2} (G-T)^{-1}y(s) = y(s) + \lambda^2 \int_{-1}^{+1} g_\lambda(s,t)y(t)dt$$

$$\Rightarrow \frac{d^2}{ds^2}(G-T)^{-1}y(s) = y(s) + \frac{\lambda\sinh\lambda(s-1)}{\sinh 2\lambda}\int_{-1}^{s}\sinh\lambda(t+1)y(t)dt$$

$$+ \frac{\lambda\sinh\lambda(s+1)}{\sinh 2\lambda}\int_{s}^{1}\sinh\lambda(t-1)y(t)dt$$

$$\Rightarrow \left|\frac{d^2}{ds^2}(G-T)^{-1}y(s)\right| \leq |y(s)| + \left\{\frac{\lambda\sinh\lambda(1-s)}{\sinh 2\lambda}\left[\frac{\cosh(t+1)}{\lambda}\right]_{-1}^{s}\right.$$

$$\left. + \frac{\lambda\sinh\lambda(s+1)}{\sinh 2\lambda}\left[-\frac{\cosh\lambda(1-t)}{\lambda}\right]_{s}^{1}\right\}\|y\|$$

using $|\sinh\lambda(s-1)| = \sinh\lambda(1-s)$ if $\lambda > 0$.

$$\Rightarrow \|(G-T)^{-1}\| \leq 1 + \max_{s}\left\{\frac{\sinh\lambda(1-s)}{\sinh 2\lambda}[\cosh\lambda(s+1) - 1]\right.$$

$$\left. + \frac{\sinh\lambda(s+1)}{\sinh 2\lambda}[-1 + \cosh\lambda(1-s)]\right\}.$$

After elementary manipulation we can achieve

$$\|(G-T)^{-1}\| \leq 2 - \frac{1}{\cosh\lambda} \tag{3.12}$$

Note that when $\lambda = 0$, $\|(G-T)^{-1}\|$ and $\|(G-T)\|$ are bounded by unity which is what we would expect since $\|G\| = \|G^{-1}\| = 1$. Also we see $\|(G-T)^{-1}\| \leq 2$ for all λ whereas $\|(G-T)\|$ is unbounded as $\lambda \to \infty$.

There now only remains $\|(G-\phi_n T)_{Y_n}^{-1}\|$ to be bounded and as in Theorems 1-4 we shall represent this by $\|(G-\phi_n T)^{-1}\|$. This is bounded from Theorem 1 by $\frac{D}{1-\delta}$ where $D = (1+\mu_1)\|(G-T)^{-1}\|$ provided $\delta = \mu_1\|\phi_n(G-T)\|\|(G-T)^{-1}\| < 1$. Thus we have to choose n large enough to give $\delta < 1$ and from (3.10), (3.11) and (3.12) we require

$$\left(\frac{\pi}{2}\right)^2 \cdot \frac{\lambda^2}{n(n-1)} \cdot \left(8 + \frac{4}{\pi}\ln(n)\right) \cdot \left(1 + \frac{\lambda^2}{2}\right) \cdot \left(2 - \frac{1}{\cosh\lambda}\right) < 1$$

Three values of λ where chosen and the table below shows the values of n needed to give $\delta < 1$.

Applicability of the Theory to an 'a priori' Example

λ^2	n required to give $\delta < 1$
0.5	5
1	8
2	14

TABLE 1

With n greater than the appropriate one of these values the error bound is now given by Theorem 2 as $\|x-x_n\| \leq \eta\|x\|$ with $\eta = (\mu_1+\mu_2\|(G-T)\|)(1+\|(G-\phi_n T)^{-1}\phi_n(G-T)\|)$ and $\|x\| \leq \|(G-T)^{-1}\|\|y\|$. Thus $\|x-x_n\|$ is less than

$$[(\tfrac{\pi}{2})^2 \cdot \frac{\lambda^2}{n(n-1)} + (\tfrac{\pi}{2})^{n-1} \frac{\|y^{(n-1)}\|\|G-T\|}{n!\,\|y\|}]\|(G-T)^{-1}\|\|y\|$$

$$(1+\|(G-\phi_n T)^{-1}\phi_n(G-T)\|)$$

if $y \in C^{(n-1)}[-1,1]$ as would often be the case. Note that the norm $\|x - x_n\|$ is the norm in the X space and so $\|x - x_n\|_X = \|x'' - x_n''\|_\infty$. To relate this to the error $\|x - x_n\|_\infty$ we use $(x - x_n)(s) = \int_{-1}^{+1} g(s,t)(x - x_n)''(t)dt$. where g(s,t) is the Green's function for $\frac{d^2}{dt^2}$ subject to (3.9b). Thus $|(x-x_n)(s)| < \int_{-1}^{+1} |g(s,t)|\|x-x_n\|_X \leq \frac{1}{2}(1-s^2)\|x-x_n\|_X$ by (1.11) $\Rightarrow \|x-x_n\|_\infty \leq \frac{1}{2}\|x-x_n\|_X$. Examining the error bound we can see this has the form

$$\|x-x_n\|_\infty \leq E_1(n)\|y\| + E_2(n)\|y^{(n-1)}\| \qquad (3.13)$$

where $E_1(n)$ is $O(\frac{\ln(n)}{n^2})$ and $E_2(n)$ is $O((\frac{\pi}{2})^{n-1} \frac{\ln(n)}{n!})$. Clearly the accuracy predicted by this 'a priori' approach is limited by the term $E_1(n)$ which depends on μ_1.

Values of E_1 and E_2 were calculated for the three values of λ^2 chosen above and the results are shown in TABLE 2 below.

Sample Results for an 'a priori' Error Bound

λ^2	n	$E_1(n)$	$E_2(n)$
0.5	8	0.36	1.2'-2
	10	0.19	2.8'-4
	12	0.12	4.8'-6
1	12	0.52	1.3'-5
	15	0.27	1.5'-8
	18	0.17	1.1'-11
2	18	1.1	4.6'-11
	20	0.71	2.4'-13
	25	0.35	2.8'-19

TABLE 2

The error bound (3.13) is very conservative. This can be seen by comparison of the above results with actual maximum errors computed by evaluation. Consider for instance the equation

$$\frac{d^2x}{dt^2} - \lambda^2 x = \cosh(1) \quad \text{with} \quad x(-1) = x(+1) = 0.$$

(When $\lambda = 1$ this has solution $x = \cosh(x) - \cosh(1)$).

For $\lambda^2 = 0.5$ we have $\begin{cases} \|x - x_8\| \leq 2.6'-10 \\ \|x - x_{10}\| \leq 2.2'-13 \\ \|x - x_{12}\| \leq 4.3'-16 \end{cases}$.

With the values of n in TABLE 2 for $\lambda^2 = 1, 2$ the actual errors are dominated by roundoff even using double length arithmetic and so are not given for comparison purposes. The 'a priori' bounds for this example are of the forms $\cosh(1).E_1(n)$ which are clearly far inferior to the true bounds. It will be seen later that certain bounds of the form (2.8) are restricted by the factor $\|(I-\phi_n)T\|$ which we saw in section 2.6 was very much connected with μ_1. It is for this reason and also the fact that we do not normally have an 'a priori' bound on the inverse of G-T that we are later concerned with developing more realistic computable 'a posteriori' bounds.

3.4 Alternative Approach

A different approach is now presented which could be used to give either 'a priori' or 'a posteriori' error bounds. We shall consider for simplicity second order differential equations although the analysis carries through in a similar manner for higher order problems.

Suppose we wish to solve approximately the equation $(G-T)x(t) \equiv x''(t)+p(t)x'(t)+q(t)x(t) = y(t)$ with $x(-1) = x(+1) = 0$, where p, q and $y \in C^{(\nu)}[-1,1]$, $(\nu \geq 0)$. This gives $x'' \in C^{(\nu)}[-1,1]$ by induction. If x_n is found by applying the collocation method as before then in keeping with the earlier notation we have by Theorem 2

that if there exists an $\tilde{x}\epsilon X_n$ such that $\|x-\tilde{x}\| \leq \varepsilon\|x\|$ then $\|x-x_n\| \leq \varepsilon(1+\|(G-\phi_n T)^{-1}\phi_n(G-T)\|\|x\|)$. This ε is used in the alternative conditions of Theorem 2 and like μ_1 and μ_2 is independent of the approximate method and depends on the approximating properties of the subspace Y_n.

However it is simpler to proceed directly as follows. If there exists an $\tilde{x}\epsilon X_n$ such that $\|x-\tilde{x}\| \leq \zeta$ then $\|x-x_n\| \leq \|x-\tilde{x}\| + \|x_n-\tilde{x}\| = \|x-\tilde{x}\| + \|(G-\phi_n T)^{-1}\phi_n(G-T)x - (G-\phi_n T)^{-1}\phi_n(G-T)\tilde{x}\| \leq (1 + \|(G-\phi_n T)^{-1}\phi_n(G-T)\|)\|x-\tilde{x}\|$. Thus

$$\|x-x_n\| \leq \zeta(1 + \|(G-\phi_n T)^{-1}\phi_n(G-T)\|) \qquad (3.14)$$

Now $\|x-\tilde{x}\|$ is a norm in the X space, i.e. $\|x-\tilde{x}\|_X = \|x"-\tilde{y}\|_\infty$ where $\tilde{y} = G\tilde{x}\epsilon Y_n$. So we are seeking a $\tilde{y}\epsilon Y_n = G^{-1}(X_n)$ to approximate x" and the corresponding $\tilde{x}$ is given by $G^{-1}\tilde{y}$. Now we are therefore approximating x" by a polynomial of degree n-1 so that Jackson's theorem can be applied. Since $x"\epsilon C^{(\nu)}[-1,1]$, by Jackson's theorem of Cheney (1966, p.147) there exists a polynomial $\tilde{y}$ of degree n-1 such that

$$\|x"-\tilde{y}\|_\infty \leq (\tfrac{\pi}{2})^\nu \frac{\|(x")^{(\nu)}\|_\infty}{n(n-1)\ \ldots\ (n-\nu+1)} \qquad (n \geq \nu+1)\ .$$

So if we assume henceforth that p, q and y are infinitely differentiable over [-1,1] then this result simplifies to $\|x"-\tilde{y}\|_\infty \leq (\tfrac{\pi}{2})^{n-1}\frac{\|(x")^{(n-1)}\|_\infty}{n!}$ (for all n). Hence with $\tilde{x} = G^{-1}\tilde{y}$ we have $\|x-\tilde{x}\|_X \leq \zeta$ where $\zeta = (\tfrac{\pi}{2})^{n-1}\frac{\|x^{(n+1)}\|_\infty}{n!}$. Thus we can apply the error bound (3.14) and this can then be modified to produce either 'a priori' or 'a posteriori' bounds.

Since we know $x''+px'+qx = y$ this enables us to express higher derivatives of the solution x in terms of lower ones. That is, $x^{(n+1)} = (x'')^{(n-1)} = (y-px'-qx)^{(n-1)}$ and so on until finally we reach $x^{(n+1)}(t) = A_n(t) + B_n(t)x(t)+C_n(t)x'(t)+D_n(t)x''(t)$. Now

$$x(s) = \int_{-1}^{+1} g(s,t)x''(t)dt \quad \text{and} \quad x'(s) = \int_{-1}^{+1} \frac{\partial g}{\partial s}(s,t)x''(t)dt$$

where $g(s,t)$ is the simple Green's function we have met before for $\frac{d^2z}{dt^2}$ with $z(-1) = z(+1) = 0$. Thus, in theory at least, we can find using (1.11) and (1.12) positive constants c_n and d_n such that $\|x^{(n+1)}\|_\infty \leq c_n+d_n\|x''\|_\infty$. We therefore have the error bound

$$\|x-x_n\|_X \leq (\tfrac{\pi}{2})^{n-1} \frac{(c_n+d_n\|x''\|_\infty)}{n!}[1+\|(G-\phi_nT)^{-1}\phi_n(G-T)\|]$$

or

$$\|x-x_n\|_X \leq e_n+f_n\|x\|_X \tag{3.15}$$

where $e_n = (\tfrac{\pi}{2})^{n-1}c_n(1+\|(G-\phi_nT)^{-1}\phi_n(G-T)\|)/n!$

and $f_n = (\tfrac{\pi}{2})^{n-1}d_n(1+\|(G-\phi_nT)^{-1}\phi_n(G-T)\|)/n!$.

From (3.15) we get the 'a priori' bound

$$\|x-x_n\|_X \leq e_n+f_n\|(G-T)^{-1}\|\|y\| \tag{3.16}$$

Using $\|x\| \leq \|x-x_n\| + \|x_n\|$ we deduce from (3.15)

$$\|x-x_n\| \leq \frac{e_n+f_n\|x_n\|}{(1-f_n)} \quad \text{provided} \quad f_n < 1,$$

and this is an 'a posteriori' bound not requiring a bound on $\|(G-T)^{-1}\|$ if $\|(G-\phi_nT)^{-1}\|$ is obtained independently.

As an illustration of the procedure the example used in section 3.3 is considered 'a priori'. Here $x''- \lambda^2 x = y \Rightarrow x'' = y + \lambda^2 x$. Thus $x^{(n+1)} = (x'')^{(n-1)} = (y + \lambda^2 x)^{(n-1)} = (y'' + \lambda^2 x'')^{(n-3)} = (y'' + \lambda^2 y + \lambda^4 x)^{(n-3)}$ and so on. Finally we reach, assuming for simplicity that n+1 is even,

$$x^{(n+1)} = \sum_{j=1}^{\frac{n-1}{2}} \lambda^{n-1-2j} y^{(2j)} + \lambda^{n-1} x \qquad (3.17)$$

Thus $\|x^{(n+1)}\| \leq \sum_{j=1}^{\frac{n-1}{2}} \lambda^{n-1-2j} \|y^{(2j)}\| + \lambda^{n-1} \|x''\|$

and $c_n = \sum_{j=1}^{\frac{n-1}{2}} \lambda^{n-1-2j} \|y^{(2j)}\|$ and $d_n = \lambda^{n-1}$.

If n+1 is odd n is even and (3.17) can be employed. After similar manipulation and utilising (1.12) we achieve

$$c_n = \sum_{j=1}^{\frac{n}{2}} \lambda^{n-2j} \|y^{(2j-1)}\| \text{ and } d_n = \lambda^n.$$

Now if we further take $y(t) = \cosh(1)$ as before then $c_n = 0$ and we have the error bound $\|x-x_n\|_\infty \leq \frac{1}{2}\|x-x_n\|_X$

$$\leq \frac{1}{2} \left(\frac{\pi}{2}\right)^{n-1} \frac{\lambda^{n'}}{n!} \left[1+\|(G-\phi_n T)^{-1}\phi_n (G-T)\|\right] \|(G-T)^{-1}\| \cosh(1)$$

where $n' = n-1$ if n is odd and n if n is even. Numerical values of this error bound are shown for various choices of λ^2 and n in TABLE 3 below. Theorem 1 is used to bound $\|(G-\phi_n T)^{-1}\|$ 'a priori' as in section 3.3 and these results are to be compared with those of the form $E_1(n)\cosh(1)$ derivable from TABLE 1.

Example of an Alternative 'a priori' Error Bound

$\lambda^2 = 0.5$		$\lambda^2 = 1$		$\lambda^2 = 2$	
n	$\|x-x_n\|_\infty$	n	$\|x-x_n\|_\infty$	n	$\|x-x_n\|_\infty$
8	3.8'-4	12	5.4'-6	18	7.6'-9
10	4.5'-6	15	6.3'-9	20	8.0'-11
12	3.9'-8	18	4.6'-12	25	3.8'-16

TABLE 3

Thus we see that great improvements can be made by this technique but still the results are fairly inaccurate compared with actual maximum errors (section 3.3). Of course often the differential equation will be too complicated to permit the successive differentiation required for this higher order result.

3.5 Application of 'a posteriori' Error Bounds

We have examined 'a priori' results and although they can be used for convergence proofs we have found them to be rather unsuitable for practical error bounds. We now for the major part of this chapter consider 'a posteriori' error bounds of the forms given by Theorem 7 and its corollary. However firstly we must show that the approximate solution of linear ordinary boundary value problems does indeed satisfy the required conditions for the theory. We have seen in section 2.2 and again briefly in section 3.2 how the collocation method applied to a $2m^{th}$ order differential equation fits into the functional analysis setting and assuming this knowledge it now only remains to show that the particular conditions of Theorem

7 are satisfied. In section 3.2 we verified the criteria necessary for the theory of Kantorovich and Akilov and this section is along similar lines.

The same equation as used previously is considered, namely (3.1a) subject to the boundary conditions (3.1b). Using the usual notation it was seen in section 3.2 that T was bounded $\epsilon[X,Y]$ and it has to be shown that $G^{-1}T$ is compact. For $G^{-1}T$ to be compact we need $G^{-1}T(U)$ to be relatively compact in X where U is the unit ball, $\{z\epsilon X : \|z\|_X \leq 1\}$ or equivalently $G^{-1}T(U)$ to be sequentially compact - see section 2.3. Let $\{z_n\}$ be a sequence in $G^{-1}T(U)$. Then $z_n \epsilon G^{-1}T(U) \Rightarrow Gz_n \epsilon T(U)$. So if we can show that any sequence in T(U) has a convergent subsequence then $\{Gz_n\}$ will have a convergent subsequence with limit v say. Then this gives $\{z_n\}$ containing a convergent subsequence with limit $G^{-1}v$ since $\|z_n - G^{-1}v\|_X = \|Gz_n - v\|_Y$. Thus it has to be shown that T(U) is relatively compact in $Y \equiv C[-1,1]$. Now $T(U) = \{u\epsilon Y : u = Tz \wedge \|z\|_X \leq 1\}$, so if $u\epsilon T(U) \Rightarrow |u| \leq \|Tz\| \leq \|T\|$, proving T(U) is uniformly bounded. Further if $t,t'\epsilon[-1,1]$ then if $z\epsilon U$ and $u = Tz$

$|u(t) - u(t')| = |(Tz)(t) - (Tz)(t')|$

$= \left| \int_{-1}^{+1} (k(t,\tau) - k(t',\tau))z''(\tau)d\tau \right|$ where k(s,t) is as defined in section 3.2. The range of integration can now be split by $\int_{-1}^{+1} = \int_{-1}^{t} + \int_{t}^{t'} + \int_{t'}^{+1}$ assuming without loss of generality that $t < t'$. In the intervals $[-1,t)$ and $(t',1]$ $k(s,\tau)$ is a continuous function of s, whereas for $\tau\epsilon[t,t']$ we can use the boundedness of $k(s,\tau)$ to get $\left| \int_{t}^{t'} (k(t,\tau) - k(t',\tau))\, z''(\tau)d\tau \right| \leq C|t'-t|\|z''\|$ for some constant C. Thus given any $\varepsilon > 0$ there exists a δ and

$|t'-t| < \delta \Rightarrow |u(t)-u(t')| < \varepsilon$ for all $z \in U$. This proves equicontinuity. Therefore by the Arzela Ascoli theorem (Kantorovich and Akilov (1964, p.22)) $T(U)$ is relatively compact in $Y = C[-1,1]$. Thus $T(U)$ is also sequentially compact and $G^{-1}T$ is a compact operator.

To apply the theory we need $\delta_n = \| (G-\phi_n T)^{-1} \| \| (I-\phi_n)T \| < 1$ and we would like $\| (I-\phi_n)T \|$ to get smaller as n increases. In section 2.6 we showed that $\| (I-\phi_n)T \| \leq \mu_1(1 + \|\phi_n\|)$ and for our polynomial approximation μ_1 was found via Jackson's theorem and is bounded by (3.7). If we are using Chebyshev points for collocation then $\|\phi_n\|$ is $O(\ln(n))$ and so as we choose n larger $\| (I-\phi_n)T \|$ is $O(\frac{\ln(n)}{n})$ which decreases.

We later consider the problem of bounding $\| (G-\phi_n T)^{-1} \|$ Basically if the collocation matrix is non singular then $(G-\phi_n T)_{Y_n}^{-1}$ exists and hence so does $(G-\phi_n T)_Y^{-1}$ and its norm is bounded by (2.7).

Remark The 'a priori' results of Theorem 1 which we discussed in section 3.2 would predict that for n sufficiently large $(G-\phi_n T)_{Y_n}^{-1}$ exists and its norms are uniformly bounded as n increases. Thus we would expect by taking enough collocation points to ensure $\delta_n = \| (G-\phi_n T)_Y^{-1} \| \| (I-\phi_n)T \| < 1$ for Theorem 7■

This theorem in (b) gives two possible error bounds. The former contains the term $\frac{\delta_n}{1-\delta_n} \| x_n \|$ and we have seen that $\| (I-\phi_n)T \|$ in δ_n is only $O(\frac{\ln(n)}{n})$ in general as n is chosen large (for Chebyshev points). This is clearly unsuitable being far too coarse if we are seeking a realistic computable error bound. Note that the 'a priori' result

(2.8) is similarly influenced by the factor $\|(I-\phi_n)T\|$, (c.f. section 3.3).

We are thus led to consider the latter bound from (b) of the form

$$\|x-x_n\| \leq \frac{\|(G-\phi_n T)^{-1}\|}{1-\delta_n}\|(G-T)x_n-y\| \text{ where } \frac{\|(G-\phi_n T)^{-1}\|}{1-\delta_n} \text{ is a}$$

bound on $\|(G-T)^{-1}\|$ and this one is used for numerical results given in Chapter 5.

Theorem 4 due to Kantorovich and Akilov gives another 'a posteriori' bound on $\|(G-T)^{-1}\|$ by

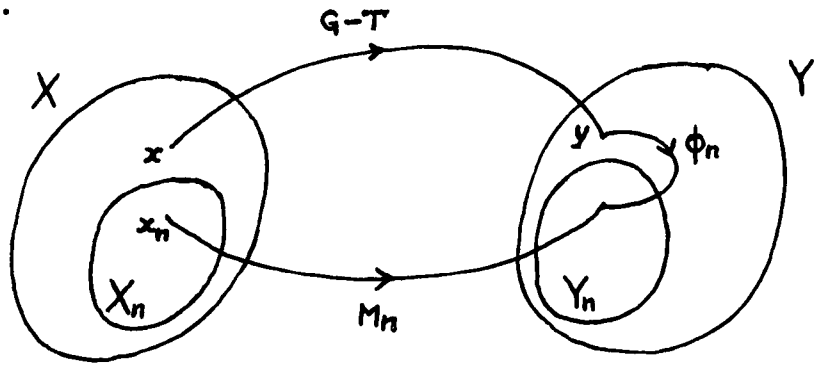
$$\|(G-T)^{-1}\| \leq \frac{1 + \|(G-\phi_n T)^{-1}\phi_n\| + \|(G-\phi_n T)^{-1}\phi_n(G-T)\|}{1-\delta} \text{ if}$$

$\delta = \mu_1 (1 + \|(G-\phi_n T)^{-1}\phi_n(G-T)\|) < 1$. However it is difficult to see how, with this, one can avoid using $\|(G-\phi_n T)^{-1}\phi_n\| \leq \|(G-\phi_n T)_{Y_n}^{-1}\|\|\phi_n\|$ and (with Chebyshev zeros) $\|\phi_n\| \leq 8 + \frac{4}{\pi}\ln(n)$ and is large if n is chosen large. Clearly δ will tend to zero very slowly and moreover we get a very poor bound on $\|(G-T)^{-1}\|$. It is for this reason that Theorem 7 is preferred in practice. This contains the term $\|\phi_n T\|$ but which is simply bounded by $\|\phi_n T\| \leq \|T\| + \|(I-\phi_n)T\|$.

3.6 Direct Approach to Bounding the Norm of the Inverse of the Approximate Operator

We showed in section 3.5 that the 'a posteriori' Theorem 7 and its corollary could be applied once $\|(G-\phi_n T)^{-1}\|$ is bounded. Equation (2.7) relates this to $\|(G-\phi_n T)_{Y_n}^{-1}\|$ and we now consider in detail the problem of finding a reasonable bound on $\|(G-\phi_n T)_{Y_n}^{-1}\|$ when polynomial approximations are sought.

Firstly a direct approach is examined. The abbreviations $M_n \equiv (G-\phi_n T)_{Y_n}^{-1}$ and $B_n = \|M_n^{-1}\|$ are introduced. Thus in the usual notation $M_n : X_n \to Y_n$ and as described in section 2.2 we have the situation illustrated below.



For simplicity the approximate solution by collocation of a second order differential equation with the solution being zero at the end points ± 1 is considered but higher order problems could be examined in a similar way.

Thus $X = \{z \epsilon C^{(2)}[-1,1] : z(-1) = 0 \wedge z(+1) = 0\}$ and we choose X_n as the space of functions of the form

$(t^2-1) \sum_{r=0}^{n-1} b_r \psi_r(t)$ where $\{\psi_r\}_{r=0}^{n-1}$ are a basis for P_{n-1} (the space of polynomials of up to degree n-1) and the $b_r (r = 0 \ldots n-1)$ are real numbers. $Y = C[-1,1]$ and $Y_n = P_{n-1}$. Let ϕ_n map each continuous $y \epsilon C[-1,1]$ into its interpolating polynomial at the collocation points $\{t_i\}_{i=1}^{n}$.

The aim of this section is to try to bound B_n by breaking up the operator M_n^{-1} into its different parts and then bounding these separately.

We have $B_n = \|M_n^{-1}\| = \sup_{\substack{\tilde{y} \epsilon Y_n \\ \|\tilde{y}\|=1}} \|M_n^{-1} \tilde{y}\|_X$

$$\text{or} \quad B_n = \sup_{\tilde{y}:\|\tilde{y}\|=1} \|GM_n^{-1}\tilde{y}\|_Y \tag{3.18}$$

When the collocation method is applied we have a set of linear equations

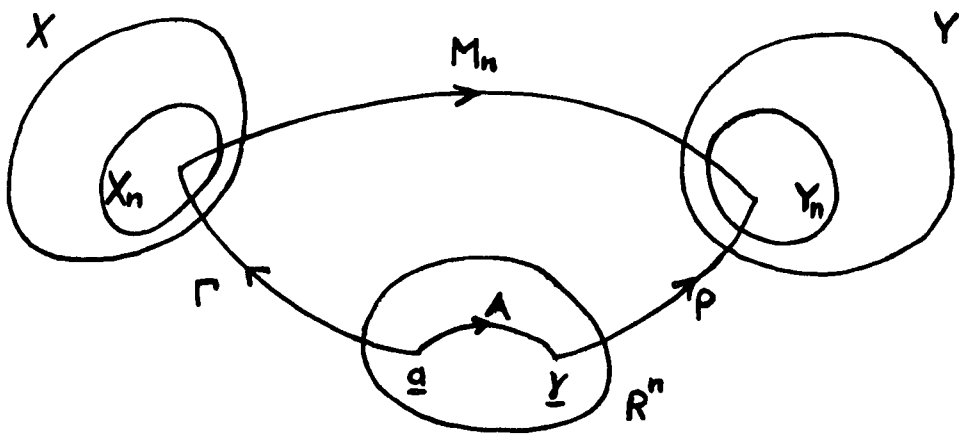
$$A\underline{a} = \underline{\gamma} \tag{3.19}$$

to solve for the coefficients of the approximation $x_n = (t^2-1)\sum_{r=0}^{n-1} a_r\psi_r$ to x. Here $\underline{a},\underline{\gamma} \in R^n$, the space of real n dimensional vectors. Define the mappings $\Gamma: R^n \to X_n$ and $\rho: R^n \to Y_n$ as follows:

$\Gamma(\underline{b}) = (t^2-1)\sum_{r=0}^{n-1} b_r\psi_r$ $(\underline{b} \in R^n)$ and $\rho(\underline{\beta})$ is the polynomial of degree n-1 such that $\rho\underline{\beta}(t_i) = \beta_i$ $(i = 1 \ldots n)$. That is, ρ constitutes polynomial interpolation and ρ^{-1} evaluation. Having solved the equations (3.19) for $\underline{a} = A^{-1}\underline{\gamma} = A^{-1}\rho^{-1}\phi_n y$ the approximate solution x_n is found by $x_n(t) = (\Gamma\underline{a})(t)$. Thus M_n^{-1} is related to the inverse collocation matrix by

$$M_n^{-1} \equiv \Gamma A^{-1}\rho^{-1} \tag{3.20}$$

and this is illustrated below.



We choose the norm in R^n by $\|\underline{b}\| = \max_i |b_i|$ for $\underline{b} = (b_o, b_1 \ldots b_{n-1}) \epsilon R^n$ and we have from (3.20)

$$B_n = \|\Gamma A^{-1} \rho^{-1}\| \leq \|\Gamma\| \|A^{-1}\| \|\rho^{-1}\| \qquad (3.21)$$

At first sight it is more obvious how to tackle the bound (3.21) than the form (3.18) although we shall see in section 3.7 that (3.18) can be utilised. We consider separately each factor of (3.21). Firstly,

$$\|\rho^{-1}\| = \sup_{\tilde{y}:\|\tilde{y}\|=1} \|\rho^{-1}\tilde{y}\| = \sup_{\|\tilde{y}\|=1} \max_{1\leq i\leq n} |\tilde{y}(t_i)| \leq 1.$$

As a slight digression from our present task we briefly mention some computational properties of the matrix A^{-1} when Chebyshev zeros are used as collocation points and the ψ_r $(r = 1, \ldots n-1)$ are Chebyshev polynomials and $\psi_o = \frac{T_o}{2}$.

It is found experimentally that for a given differential operator $\|A^{-1}\|$ remains virtually constant as n increases. (It is in fact the first row of A^{-1} which gives the maximum modulus row sum). Further if H is the n x n diagonal matrix $\text{diag}(h_1, h_2, \ldots h_n)$ with $h_1 = 1$ and $h_i = (i-1)^2$ $(i=2, \ldots n)$ then $\|HA^{-1}\|$ is roughly constant as n increases, i.e. as more collocation points are chosen. This is shown in TABLE 4 for the sample $Gx - Tx \equiv x'' + (1 + t^2)x$.

Constancy Property of the Norms of Certain Matrices

n	5	10	15	20	15
$\|A^{-1}\|$	1.807549	1.807561	→	→	→
$\|HA^{-1}\|$	1.807549	1.807561	→	→	→

TABLE 4

In the above table the symbol → means that the entry is the same as the one on its left.

If simple powers are used, instead of Chebyshev polynomials, in the basis for X_n it is found that the above matrix norms grow large as n increases and an example of this property is given in section 3.8. In the above table we notice $\|A^{-1}\| = \|HA^{-1}\|$. This holds for the particular operator chosen here and is not necessarily true in general as will be seen later.

These properties are relevant for following analysis and so are mentioned now but being away from the main theme of this and the next sections are left until section 3.8 to be considered in more detail.

We now return to the problem of bounding B_n and examine $\|\Gamma\|$ which occurs in (3.21).

$$\|\Gamma\| = \sup_{\substack{b\in R^n \\ \|\underline{b}\|=1}} \|\Gamma(\underline{b})\|_X = \sup_{\underline{b}:\max_i|b_i|=1} \{\max_t \left|\frac{d^2}{dt^2}(t^2-1)\sum_{r=0}^{n-1} b_r\psi_r(t)\right|\}$$

$$= \sup_{\underline{b}:\max_i|b_i|=1} \{\max_t |\tilde{x}''(t)|\} \text{ where } \tilde{x}(t) = (t^2-1)\sum_{r=0}^{n-1} b_r\psi_r(t).$$

We now consider two different choices for the $\psi_r(t)$. If $\psi_r(t)$ is taken as $t^r (r = 0 \ldots n-1)$ then we have

$$\tilde{x}''(t) = 2\sum_{r=0}^{n-1} b_r t^r + 4t\sum_{r=1}^{n-1} r b_r t^{r-1} + (t^2-1)\sum_{r=2}^{n-1} r(r-1) b_r t^{r-2} \qquad (3.22)$$

So $|\tilde{x}''(t)| \leq 2\sum_{r=0}^{n-1} 1 + 4\sum_{r=1}^{n-1} r + \sum_{r=2}^{n-1} r(r-1)$ and this expression is $O(n^3)$ and so would give a bound of order n^3 or $\|\Gamma\|$. This

is definitely unsuitable since in order to apply Theorem 7 we need $\delta_n = \|(G-\phi_n T)_Y^{-1}\| \|(I-\phi_n)T\| < 1$. $\|(G-\phi_n T)_Y^{-1}\|$ depends on B_n by (2.7) and clearly if $\|A^{-1}\|$ is constant or increasing with n and $\|\Gamma\|$ is $O(n^3)$ we are very likely unable to achieve $\delta_n < 1$. The remark of section 3.5 suggests that we should be able to construct bounds for B_n which do not increase with n and this is the basic problem which we tackle.

If Chebyshev polynomials are used with $\tilde{x}(t)$ of the form $(t^2-1)\left[\frac{b_o}{2}T_o + b_1T_1 + \ldots + b_{n-1}T_{n-1}\right]$ which we write as $\tilde{x}(t) = (t^2-1)\sum_{r=0}^{n-1}{}' b_r T_r(t)$ then

$$\tilde{x}''(t) = 2\sum_{r=0}^{n-1}{}' b_r T_r(t) + 4t\sum_{r=1}^{n-1} b_r T_r'(t) + (t^2-1)\sum_{r=2}^{n-1} b_r T_r''(t).$$

Now the Chebyshev polynomials satisfy the following differential equation - see for example Davis (1963, p.365): $(1-t^2)T_r'' - tT_r' + r^2T_r = 0$. Thus $(t^2-1)T_r'' = r^2T_r(t) - t\,T_r'(t)$ giving

$$\tilde{x}''(t) = 2\sum_{r=0}^{n-1}{}' b_r T_r(t) + \sum_{r=2}^{n-1} r^2 b_r T_r(t) + 4t\sum_{r=1}^{n-1} b_r T_r'(t) - \sum_{r=2}^{n-1} t b_r T_r'(t)$$

and rearranging we have

$$\tilde{x}''(t) = 2\sum_{r=0}^{n-1}{}' b_r T_r(t) + \sum_{r=2}^{n-1} r^2 b_r T_r(t) + 3t\sum_{r=1}^{n-1} b_r T_r'(t) + tb_1T_1'(t) \qquad (3.23)$$

Now by Markoff's theorem (Todd (1962, p.138)) $|T_r'(t)| \leq r^2$.

Thus $|\tilde{x}''(t)| \leq 2\sum_{r=0}^{n-1}{}' 1 + \sum_{r=2}^{n-1} r^2 + 3\sum_{r=1}^{n-1} r^2 + 1$ and this again is

an expression of $O(n^3)$ giving bounds on $\|\Gamma\|$ and B_n uunsuitable for practical purposes.

A variation of the above approach is now considered. Instead of saying $B_n = \|\Gamma A^{-1}\rho^{-1}\| \leq \|\Gamma\|\|A^{-1}\|\|\rho^{-1}\|$ as in (3.21) we investigate the possibility of using $B_n \leq \|\Gamma A^{-1}\|\|\rho^{-1}\|$. ΓA^{-1} is a mapping from R^n to X_n and is independent of the basis used in X_n but when bounding its norm the inequalities used still lead to different results.

$$\|\Gamma A^{-1}\| = \sup_{\substack{c\in R^n \\ \|\underline{c}\|=1}} \|\Gamma A^{-1}\underline{c}\|_X = \sup_{\|\underline{c}\|=1} \|\Gamma\underline{b}\|_X$$

where $\underline{b} = A^{-1}\underline{c}$. If we take t^r for $\psi_r(t)$ $(r=0 \ldots n-1)$ we can use (3.22) where $\tilde{x} = \Gamma\underline{b} = \Gamma A^{-1}\underline{c}$.

Now define $\underline{\beta} = (\beta_1, \beta_2, \ldots \beta_n)^t$ by $\beta_r = b_{r-1}$ $(r=1 \ldots n)$ then with $A^{-1} = (v_{ij})$ we have

$$\beta_r = \sum_{k=1}^{n} v_{rk}c_k \quad (r=1 \ldots n) \tag{3.24}$$

where $\underline{c} = (c_1, c_2, \ldots c_n)^t$. Thus

$$\tilde{x}''(t) = 2\sum_{r=1}^{n}\beta_r t^{r-1} + 4t\sum_{r=2}^{n}(r-1)\beta_r t^{r-2} + (t^2-1)\sum_{r=3}^{n}(r-1)(r-2)\beta_r t^{r-3}$$

and using (3.24) we have

$$\tilde{x}''(t) = 2\sum_{r=1}^{n}\sum_{k=1}^{n} v_{rk}c_k t^{r-1} + 4t\sum_{r=2}^{n}\sum_{k=1}^{n}(r-1)v_{rk}c_k t^{r-2}$$

$$+ (t^2-1)\sum_{r=3}^{n}\sum_{k=1}^{n}(r-1)(r-2)v_{rk}t^{r-3}$$

Therefore $\|\tilde{x}\|_X = \|\tilde{x}''\|_\infty$

$$\leq \sum_{k=1}^{n}\{2\sum_{r=1}^{n}|v_{rk}| + 4\sum_{r=2}^{n}(r-1)|v_{rk}| + \sum_{r=3}^{n}(r-1)(r-2)|v_{rk}|\}$$

if $\|\underline{c}\| \leq 1$ and

$$\|\Gamma A^{-1}\|_X \leq \sum_{k=1}^{n} \sum_{r=1}^{n} \{2|v_{rk}|+4(r-1)|v_{rk}|+(r-1)(r-2)|v_{rk}|\}$$

$$\Rightarrow \|\Gamma A^{-1}\|_X \leq \sum_{k=1}^{n} \sum_{r=1}^{n} (r^2+r)|v_{rk}| \qquad (3.25)$$

If Chebyshev polynomials are tried in the same way as before and using a similar definition of $\underline{\beta}$ then from (3.23) we have

$$\tilde{x}''(t) = 2\sum_{r=1}^{n} {}'\beta_r T_{r-1} + \sum_{r=3}^{n} (r-1)^2 \beta_r T_{r-1} + 3t \sum_{r=2}^{n} \beta_r T_{r-1}' + t\beta_2$$

We next employ (3.24) and take moduli throughout, utilising $|T_{r-1}'| \leq (r-1)^2$, to finally obtain

$$\|\Gamma A^{-1}\|_X \leq \sum_{k=1}^{n} \sum_{r=1}^{n} (4r^2-8r+6)|v_{rk}| \qquad (3.26)$$

Earlier in this section we mentioned certain computational properties of the collocation matrices when Chebyshev zeros are used as collocation points. Bearing these in mind we should expect that the bound (3.25) would increase wildly with n and this is shown by example in TABLE 5 below. The inequality (3.26) can be rewritten

$$\|\Gamma A^{-1}\| \leq 4\sum_{r=1}^{n} \left(\sum_{k=1}^{n} (r-1)^2|v_{rk}|\right) + 2\sum_{k=1}^{n} \sum_{r=1}^{n} |v_{rk}|$$

In view of the results for $\|HA^{-1}\|$ shown in TABLE 4 we anticipate that $\sum_{k=1}^{n} (r-1)^2 v_{rk}$ is roughly constant and therefore that $4\sum_{r=1}^{n} \left(\sum_{k=1}^{n} (r-1)^2|v_{rk}|\right)$ increases like $O(4n)$. This is also borne out by the computed results below.

Behaviour of the Direct Approach to Bounding the Norm of the Inverse of the Given Operator

$Gx - Tx \equiv x'' - x$					
n	5	10	15	20	25
Bound on $\|\Gamma A^{-1}\|$ (Powers)	22	1552	117373	9295153	7.4'8
Bound on $\|\Gamma A^{-1}\|$ (Chebyshev)	9.2	28.8	48.9	72.1	94.9

$Gx - Tx \equiv x'' + (1+t^2)x$				
n	10	15	20	25
Bound on $\|\Gamma A^{-1}\|$ (Chebyshev)	32.0	52.1	75.5	98.4

$Gx - Tx \equiv x'' + (8t^2+2t-1)x'+(4.5t^2+1.5t-1)x$				
n	10	15	20	25
Bound on $\|\Gamma A^{-1}\|$ (Chebyshev)	30.2	51.0	74.0	96.8

TABLE 5

Thus we see that although we can achieve better results by (3.26) the bound on B_n still increases with n and so is rather unsatisfactory.

Remark The remark of section 3.5 suggested we should be able by the 'a priori' consideration of Theorem 1 to bound the B_n uniformly as n is chosen larger. However we noted (remark in section 3.2) that we required the coefficients in the linear differential operator to have at least one continuous derivative in order to satisfy the conditions of the theorem and boundedness cannot be guaranteed if this does not hold. We have not required this property of the

coefficient functions in this section and so are unlikely to achieve a uniform bound on the B_n.

In the next section an approach is considered which does use the continuous differentiability and in which the functions $\tilde{x}''(t)$ are expressed in terms of the Lagrange interpolation basis polynomials corresponding to the collocation points.

3.7 Indirect Approach Using Second Derivative Values at the Collocation Points

For this section second order differential equations with their solutions being zero at the end points ± 1 are considered and the spaces and subspaces of section 3.6 are chosen.

Suppose the differential equation is of the form

$$Gx-Tx \equiv x''(t) + p(t)x'(t) + q(t)x(t) = y(t) \qquad (3.27)$$

with $x(-1) = x(+1) = 0$. Let $y \in C[-1,+1]$ but let $p,q \in C^{(1)}[-1,1]$ and this additional continuity will be used later. The analysis can be carried over to higher order problems. Here we have X_n as the space of functions of the form $(t^2-1)\sum_{r=0}^{n-1} b_r\psi_r(t)$ for some choice of n linearly independent polynomials $\psi_r(t)$ $(r = 0,1, \ldots n-1)$.

So far in trying to bound $B_n = \sup_{\substack{\tilde{y}\in Y_n \\ \|\tilde{y}\|=1}} \|GM_n^{-1}\tilde{y}\|_\infty$ by (3.18) we have expressed $M_n^{-1}\tilde{y}$ in the form $\Gamma A^{-1}\rho^{-1}\tilde{y}$ or $\Gamma(\underline{b})$ where $\underline{b} = A^{-1}\rho^{-1}\tilde{y}$ (employing the notation of the previous section). We have then used

$$(GM_n^{-1}\tilde{y})(t) = \frac{d^2}{dt^2}\{\Gamma(\underline{b})(t)\} = \frac{d^2}{dt^2}\{(t^2-1)\sum_{r=0}^{n-1} b_r\psi_r(t)\}.$$

That is, $M_n^{-1}\tilde{y}$ is formed in terms of the basis for X_n and then differentiated twice.

We now consider an approach which expresses $GM_n^{-1}\tilde{y} \in Y_n$ directly in terms of the Lagrange interpolating basis polynomials corresponding to the collocation points. That is, if $\tilde{x} = M_n^{-1}\tilde{y}$ for $\tilde{y} \in Y_n$ we write $\tilde{x}''(t) = \sum_{j=1}^{n} \tilde{x}''(t_j) l_j^n(t)$ and to determine $\tilde{x}''(t_j)$ $(j = 1 \ldots n)$ we proceed as below.

Define $\underline{\xi}(t) = (\xi_1(t), \xi_2(t), \ldots \xi_n(t))^t$ by $\xi_{r+1}(t) = \{(t^2-1)\psi_r(t)\}''(r = 0, \ldots n-1)$ and thus the second derivative of any function in X_n is of the form $\underline{\xi}^t(t)\underline{b}$. Let the choice of collocation points be $\{t_1, t_2, \ldots t_n\}$ and for any right hand side $y(t)$ we find by applying the collocation method an approximate solution $\tilde{x}(t) = (t^2-1)\sum_{r=0}^{n-1} a_r\psi_r(t)$ to an equation of the type (3.27) by solving the algebraic equations $A\underline{a} = \underline{\gamma}$ where $\underline{\gamma} = (\gamma_1, \gamma_2, \ldots \gamma_n)^t$ and $\gamma_i = y(t_i)$ $(i=1 \ldots n)$. The approach we now use expresses $\tilde{x}''$ in terms of $\{\tilde{x}''(t_j)\}_{j=1}^{n}$ instead of $\{a_j\}_{j=0}^{n-1}$ and we proceed as follows.

Consider firstly the equation $\tilde{x}'' = y(t)$ with $x(-1) = x(+1) = 0$. Let the matrix formed be A_o and the solution be $x_o(t)$ with coefficients $\underline{a}^{(o)}$. Therefore $\tilde{x}_o''(t) = \underline{\xi}^t.\underline{a}^{(o)} = \underline{\xi}^t(A_o^{-1}\underline{\gamma})$. Now in this case row j of A_o is exactly $\underline{\xi}^t(t_j)$ and so $\tilde{x}_o''(t_j) = \underline{\xi}^t(t_j)A_o^{-1}\underline{\gamma} = \underline{e}_j{}^t\underline{\gamma} = \gamma_j$ where $\underline{e}_j$ is the unit vector with unity in the j^{th} row and zeros elsewhere. This of course is what we would expect.

Now considering a differential equation of type (3.27) we have an approximate solution $\tilde{x}(t)$ with coefficient vector $\underline{a}$ corresponding to a right hand side $y(t)$ and $\tilde{x}''(t) = \underline{\xi}^t(t)\underline{a} = \underline{\xi}^t(t)A^{-1}\underline{\gamma} = \underline{\xi}^t(t)A_o^{-1}A_oA^{-1}\underline{\gamma}$. Thus

$$\tilde{x}''(t_j) = \underline{e}_j{}^t A_o A^{-1}\underline{\gamma} = \underline{e}_j{}^t W\underline{\gamma} \tag{3.28}$$

where $W = (W_{ij}) = A_oA^{-1}$. If $\underline{z} = (z_1, z_2, \ldots z_n)^t$ and $z_j = \tilde{x}''(t_j)$ $(j = 1 \ldots n)$ then $\underline{z} = W\underline{\gamma}$ determines the values of the second derivative of an approximate solution at the nodes. W is independent of $\{\psi_r(t)\}$ since the approximate solutions $\tilde{x}(t)$ are. This is discussed more fully later in this section.

Now from (3.18),

$$B_n = \sup_{\|\tilde{y}\|=1} \|GM_n^{-1}\tilde{y}\| = \sup_{\|\tilde{y}\|=1} \|\tilde{x}''(\tilde{y})\|$$

$$= \sup_{\|\tilde{y}\|=1} \Big\| \sum_{j=1}^{n} \tilde{x}''(t_j) l_j^n(t) \Big\|$$

$$\leq \lambda_n \sup_{\|\tilde{y}\|=1} \{\max_j |x''(t_j)|\} \text{ where } \lambda_n = \max_{-1\leq t\leq 1} \sum_{j=1}^{n} |l_j^n(t)|$$

$$\leq \lambda_n \sup_{\|\tilde{y}\|=1} \{\max_j |\text{row}_j(W)\underline{\gamma}|\} \text{ where}$$

$\underline{\gamma} = (\tilde{y}(t_1), \tilde{y}(t_2), \ldots \tilde{y}(t_n))^t$ now and $\text{row}_j(W)$ is the j^{th} row of W. Thus

$$B_n \leq \lambda_n \sup_{\|\tilde{y}\|=1} \{\max_j \Big| \sum_{i=1}^{n} W_{ji}\gamma_i \Big|\}$$

$$\leq \lambda_n \max_j \Big(\sum_{i=1}^{n} |W_{ji}| \Big) \leq \lambda_n \|W\|_\infty$$

It is found experimentally that $\|W\|_\infty$ is virtually constant with n and this is illustrated later in TABLE 6. However

$\lambda_n \leq 8 + \frac{4}{\pi} \ln(n)$ (Natanson (1965, p.48)) if for instance Chebyshev zeros were used as the nodes and consequently we would expect this bound on B_n to increase with n. We are led to consider an approach which utilises the fact that the coefficients in the differential equation have a continuous derivative and we now describe this.

If the collocation points are chosen as the zeros of a polynomial $\zeta_n(t)$ of degree n belonging to a set of orthogonal polynomials $\{\zeta_n(t)\}$ with weight function w(t) over [-1,1] then this implies (Natanson (1965, p.51)) that the set $\{l_j^n(t)\}_{j=1}^n$ of basis polynomials are also orthogonal with the same weight w(t). For instance $\{\zeta_n(t)\}$ could be Chebyshev polynomials with weight function $w(t) = (1-t^2)^{-\frac{1}{2}}$ or Legendre polynomials with weight w(t) = 1. As before $z_k = \tilde{x}''(t_k)$ (k = 1 ... n) and choosing the collocation points in the above way we have

$$\int_{-1}^{+1} w(t)\tilde{x}''(t) l_j^n(t)\,dt = \int_{-1}^{+1} w(t) \sum_{k=1}^{n} z_k l_k^n(t) l_j^n(t)\,dt = z_j U_j^n \text{ where}$$

$$U_j^n = \int_{-1}^{+1} w(t) (l_j^n(t))^2 dt. \quad \text{So} \int_{-1}^{+1} w(t) [\tilde{x}''(t)]^2 dt = \sum_{j=1}^{n} z_j^2 U_j^n.$$

Note that this result is precisely that of Gaussian quadrature since $[\tilde{x}''(t)]^2$ is a polynomial of degree 2n-2 and so quadrature with Gaussian nodes will be exact. The U_j^n are the weights at the nodes. This suggests a new norm $\| \ \|_{X2}$ say, which we introduce for convenience, defined by

$$\|z\|_{X2} = \{ \int_{-1}^{+1} w(t)[z''(t)]^2 dt \}^{\frac{1}{2}} \text{ for all } z \in X, \text{ the whole space.}$$

Note that this norm depends on the choice of collocation points whereas before $\| \ \|_X$ was independent of the nodes. This norm is well defined since $X \subset C^{(2)}[-1,1]$ and so z''(t)

is continuous and so integrable. Also the basic definitions of a norm are satisfied.

It is now shown how to find $\|M_n^{-1}\| = \|(G-\phi_n T)_{Y_n}^{-1}\|$ using the X2-norm in X and the infinity norm in Y.

$$\|M_n^{-1}\|_{X2} = \sup_{\substack{\tilde{y}\in Y_n \\ \|y\|=1}} \|M_n^{-1}\tilde{y}\|_{X2}$$

$$= \sup_{\tilde{y}:\|\tilde{y}\|=1} \|\tilde{x}(\tilde{y})\|_{X2} \quad (\text{where } \tilde{x}(\tilde{y}) = M_n^{-1}\tilde{y})$$

$$= \sup_{\tilde{y}:\|\tilde{y}\|=1} \{\int_{-1}^{+1} (\tilde{x}''(t))^2 w(t)\,dt\}^{\frac{1}{2}}$$

$$= \sup_{\tilde{y}:\|\tilde{y}\|=1} \{\sum_{j=1}^{n} z_j^2 U_j^n\}^{\frac{1}{2}}.$$

The U_j^n could be calculated individually but it is simpler to say

$$\|M_n^{-1}\|_{X2} \leq \sup_{\tilde{y}:\|\tilde{y}\|=1} \{(\max_k z_k^2) \sum_{j=1}^{n} U_j^n\}^{\frac{1}{2}}.$$

By Natanson (1965, p.52)

$$\sum_{j=1}^{n} U_j^n = \sum_{j=1}^{n} \int_{-1}^{+1} w(t)(l_j^n(t))^2 dt = \int_{-1}^{+1} w(t)\,dt = \Omega, \text{ say.}$$

So we have

$$\|M_n^{-1}\|_{X2} \leq \Omega^{\frac{1}{2}} \sup_{\tilde{y}:\|\tilde{y}\|=1} \{\max_k z_k^2\}^{\frac{1}{2}}$$

$$= \Omega^{\frac{1}{2}} \sup_{\tilde{y}:\|\tilde{y}\|=1} (\max_k |z_k|).$$

Now from (3.28) $z_k = \tilde{x}''(t_k) = \underline{e}_k^t A_o A^{-1} \underline{\gamma} = \underline{e}_j^t W \underline{\gamma}$ and since we are using the infinity norm $\|\tilde{y}\| = 1 \Rightarrow |\gamma_i| \leq 1$ $(i = 1 \ldots n)$.

Thus $\sup_{\tilde{y}:\|\tilde{y}\|=1} (\max_k |z_k|) \leq \max_k (\sum_{j=1}^{n} |W_{kj}|) = \|W\|_\infty$

and we arrive at the bound

$$\|M_n^{-1}\|_{X2} \leq \Omega^{\frac{1}{2}} \|W\|_\infty \qquad (3.29)$$

For Chebyshev nodes $\Omega = \int_{-1}^{+1} (1-t^2)^{-\frac{1}{2}} dt = \pi$ while for Legrendre points $\Omega = \int_{-1}^{+1} dt = 2$.

To illustrate the usefulness of this bound some examples for different operators of $\|W\| = \|A_o A^{-1}\|$ are shown in TABLE 6 for varying numbers, n, of collocation points. For these results Chebyshev zeros have been used and Chebyshev polynomials taken as the $\{\psi_r(t)\}$. It is seen experimentally that $\|A_o A^{-1}\|$ is virtually constant as n varies and this property is related to those discussed in section 3.6 and will be considered again in the next section.

Illustration of the Constancy of the Norm of the Matrix $A_o A^{-1}$

Differential Operator	n = 5	10	15	20	25	30
$x'' - x$	1.0234	1.1315	1.2014	1.2362	1.2594	1.2738
$x'' + (1+t^2)x$	1.9318	1.9306	1.9321	1.9318	1.9321	1.9320
$x'' + \frac{2}{(t+3)}x' - \frac{2}{(t+3)^2}x$	2.0570	2.1727	2.1956	2.2038	2.2075	2.2096
$x'' - \frac{t^2(t+1)}{4}x$	1.0148	1.0310	1.0388	1.0422	1.0441	1.0452

TABLE 6

Thus we have in (3.29) a bound on $\|M_n^{-1}\|_{X2}$ which does not increase significantly with n and this will be utilised later in this section and also in the next chapter.

Having given this analysis we turn again to the problem of bounding $B_n = \|M_n^{-1}\|$ in our original norm. It is found to be convenient to transform the equations $(G-T)x=y$ and $(G-\phi_n T)x_n = \phi_n y$ to integral equations as in section 3.2.

From (3.27) the given equation is

$$x''(s)+p(s)x'(s)+q(s)x(s) = y(s) \quad \text{with } x(-1) = x(+1) = 0,$$

$$\Rightarrow \quad x''(s)+p(s)\int_{-1}^{+1}\frac{\partial g}{\partial s}(s,t)x''(t)dt+q(s)\int_{-1}^{+1}g(s,t)x''(t)dt = y(s)$$

where $g(s,t)$ is the Green's function for $\frac{d^2z}{dt^2}$ subject to $z(-1) = z(+1) = 0$. Thus writing $u = x''$ we have

$$u(s) - \int_{-1}^{+1}k(s,t)u(t)dt = y(s) \tag{3.30}$$

where $k(s,t) = -p(s)\frac{\partial g}{\partial s}(s,t) - q(s)g(s,t)$.
Since $x\epsilon X \subset C^{(2)}[-1,1] \Rightarrow u\epsilon C[-1,1] = Y$ and u satisfies

$$(I-K)u = y \tag{3.31}$$

where K is a bounded linear operator on Y, i.e. $K\epsilon[Y]$. Similarly if $u_n = x_n''$ then $u_n \epsilon Y_n$ and satisfies

$$(I-\phi_n K)u_n = \phi_n y. \tag{3.32}$$

Now $B_n = \sup_{\|\tilde{y}\|=1}\|M_n^{-1}\tilde{y}\|_X = \sup_{\|\tilde{y}\|=1}\|\tilde{x}(\tilde{y})\|_X$

$$\Rightarrow B_n = \sup_{\|\tilde{y}\|=1}\|\tilde{u}(\tilde{y})\|_\infty \tag{3.33}$$

where $\tilde{u} = \tilde{x}''$, and from (3.32) $\tilde{u}$ satisfies

$$(I-\phi_n K)\tilde{u} = \tilde{y}. \tag{3.34}$$

Thus from (3.34) $\tilde{u} = \tilde{y} + \phi_n K\tilde{u}$, giving $\|\tilde{u}\| \leq \|\tilde{y}\| + \|\phi_n K\tilde{u}\|$. (For the remainder of this section unless otherwise specified infinity norms are used).

Therefore, using (3.33), we have

$B_n \leq \sup_{\|\tilde{y}\|=1} (\|\tilde{y}\| + \|\phi_n K\tilde{u}(\tilde{y})\|)$, giving

$$B_n \leq 1 + \sup_{\|\tilde{y}\|=1} \|\phi_n K\tilde{u}(\tilde{y})\|, \tag{3.35}$$

and we shall employ the inequality

$$\|\phi_n K\tilde{u}(\tilde{y})\| \leq \|K\tilde{u}(\tilde{y})\| + \|\phi_n K\tilde{u}(\tilde{y}) - K\tilde{u}(\tilde{y})\|. \tag{3.36}$$

Consider firstly $\|K\tilde{u}\|$. From (3.30) and (3.31)

$\|K\tilde{u}\| = \max_s \left| \int_{-1}^{+1} k(s,t)\tilde{u}(t)dt \right|$ and by Cauchy's inequality

$$\|K\tilde{u}\| \leq \max_s \{\int_{-1}^{+1} (k(s,t))^2 dt\}^{\frac{1}{2}} \{\int_{-1}^{+1} (\tilde{u}(t))^2 dt\}^{\frac{1}{2}}. \tag{3.37}$$

Now with $k_{max} = \max_{s,t} |k(s,t)|$,

$$\int_{-1}^{+1} (k(s,t))^2 dt \leq 2k_{max}^2 \tag{3.38}$$

and since $k(s,t) = -p(s)\frac{\partial g}{\partial s}(s,t) - q(s)g(s,t)$, k_{max} can be calculated. To utilise the second integral in (3.37) for bounding B_n we must find $\sup_{\|\tilde{y}\|=1} \{\int_{-1}^{+1} (\tilde{u}(t))^2 dt\}^{\frac{1}{2}}$. If we now choose the collocation points to be the roots of a polynomial belonging to an orthonormal set with weight function $w(t)$ then $\{\int_{-1}^{+1} (\tilde{u}(t))^2 dt\}^{\frac{1}{2}} \leq \{\int_{-1}^{+1} w(t)(\tilde{u}(t))^2 dt\}^{\frac{1}{2}}$ if $w(t) \geq 1$.

Now $\sup_{\|\tilde{y}\|=1} \{\int_{-1}^{+1} w(t)(\tilde{u}(t))^2 dt\}^{\frac{1}{2}} = \|M_n^{-1}\|_{X2}$ giving

$$\sup_{\|\tilde{y}\|=1} \{\int_{-1}^{+1} (\tilde{u}(t))^2 dt\}^{\frac{1}{2}} \leq \Omega^{\frac{1}{2}} \|A_o A^{-1}\| \tag{3.39}$$

(where $\Omega = \int_{-1}^{+1} w(t)dt$).

Thus by (3.38) and (3.39) we can bound in a reasonable manner the term $\sup_{\|\tilde{y}\|=1} \|K\tilde{u}(\tilde{y})\|$ which comes from (3.35) and (3.36). We have then

$$\sup_{\|\tilde{y}\|=1} \|K\tilde{u}(\tilde{y})\| \leq (2\Omega)^{\frac{1}{2}} k_{max} \|A_o A^{-1}\| . \tag{3.40}$$

We now have to consider the quantity

$$\sup_{\|\tilde{y}\|=1} \|\phi_n K\tilde{u}(\tilde{y}) - K\tilde{u}(\tilde{y})\|$$

$$= \sup_{\|\tilde{y}\|=1} \|\phi_n K\tilde{u}(\tilde{y}) - \tilde{v} + \tilde{v} - K\tilde{u}(\tilde{y})\| , \text{ for any } \tilde{v} \in Y_n = P_{n-1}$$

$$\leq \sup_{\|\tilde{y}\|=1} (1+\|\phi_n\|)(\|K\tilde{u}(\tilde{y}) - \tilde{v}\|) \text{ since } \phi_n \tilde{v} = \tilde{v}.$$

We now use the fact that $p, q \in C^{(1)}[-1,1]$. This point has been discussed in the remarks earlier in this chapter and the usefulness of this requirement is now seen.

By Jackson's theorem if $K\tilde{u}(\tilde{y}) \in C^{(1)}[-1,1]$ then for any $\tilde{y}$ there exists a $\tilde{v}(\tilde{y}) \in Y_n$ (i.e. a polynomial of degree n-1) such that

$$\|K\tilde{u}(\tilde{y}) - \tilde{v}(\tilde{y})\| \leq \frac{\pi}{2} \frac{\|(K\tilde{u}(\tilde{y}))'\|}{n} . \quad \text{Also for any } z \in X,$$

$$Kz'' = Tz \tag{3.41}$$

since $(Kz'')(s) = \int_{-1}^{+1} k(s,t)z''(t)dt$

$$= -\int_{-1}^{+1}\{p(s)\frac{\partial g}{\partial s}(s,t)+q(s)g(s,t)\}z''(t)dt$$

$$= -p(s)z'(s)-q(s)z(s) = (Tz)(s).$$

Thus $K\tilde{u}(\tilde{y}) = T\tilde{x}(\tilde{y}) = -p\tilde{x}'-q\tilde{x}$ and since we have assumed $p,q \in C^{(1)}[-1,1] \Rightarrow K\tilde{u} \in C^{(1)}[-1,1]$. Furthermore $(K\tilde{u})' = -p\tilde{x}''-p'\tilde{x}'-q\tilde{x}'-q'\tilde{x}$. By using the Green's function $g(s,t)$ as above we can therefore achieve

$$\|(K\tilde{u})'\| \leq k_1\|\tilde{x}''\|. \tag{3.42}$$

Thus using (3.42) we have

$$\sup_{\|\tilde{y}\|=1}\|\phi_n K\tilde{u}(\tilde{y})-K\tilde{u}(\tilde{y})\| \leq (1+\|\phi_n\|)\frac{\pi k_1}{2n}\|\tilde{x}''(\tilde{y})\|. \tag{3.43}$$

But $\sup_{\|\tilde{y}\|=1}\|\tilde{x}''(\tilde{y})\| = \|M_n^{-1}\|_X = B_n$ by (3.33) and so from (3.33), (3.35), (3.36), (3.40) and (3.43) we have

$$B_n \leq 1+(2\Omega)^{\frac{1}{2}}k_{max}\|A_oA^{-1}\|+(1+\|\phi_n\|)\frac{\pi k_1}{2n}B_n.$$

Therefore finally if $\epsilon_n = (1+\|\phi_n\|)\frac{\pi k_1}{2n} < 1$ we obtain

$$B_n \leq \frac{1+\sigma\|A_oA^{-1}\|}{1-\epsilon_n} \tag{3.44}$$

where $\sigma = (2\Omega)^{\frac{1}{2}}k_{max}$.

This is now a bound on B_n (for $\epsilon_n < 1$) which does not increase significantly with n provided $\|A_oA^{-1}\|$ is roughly constant.

It was seen from (3.28) that if $\underline{z} = (z_1, z_2, \ldots z_n)^t$ is such that $z_j = \tilde{x}''(t_j)$ $(j = 1 \ldots n)$ then $\underline{z} = A_o A^{-1} \underline{\gamma}$. Thus this means that AA_o^{-1} is the matrix to be inverted for $\underline{z}$ if we apply the collocation method to an equation of type (3.31) and seek an approximate solution in the form $\tilde{u}(t) = \sum_{j=1}^{n} z_j 1_j^n(t)$. $\tilde{u}$ will then satisfy the equation (3.34). This could be confirmed algebraically by forming the matrix $C = (c_{ij})$ with $c_{ij} = [(I-K)1_j^n]\ (t_i)$ $(i,j = 1 \ldots n)$ and verifying that AA_o^{-1} does indeed equal C. Thus the matrix $A_o A^{-1}$ is clearly independent of the basis in X_n.

Note that any differential equation $(G-T)x_n = y$ could actually be solved approximately by finding an approximation $\tilde{u}$ to $\tilde{x}''$ of the form $\tilde{u} = \sum_{j=1}^{n} \tilde{x}''(t_j) 1_j^n(t)$ by applying the collocation method to the corresponding integral equation of type $(I-K)u = y$. Then the approximation $\tilde{x}$ to x is obtained by integrating twice (subject to the end conditions) the polynomial $\tilde{u}(t) = \tilde{x}''(t)$. However although this is a convenient theoretical approach it is practically quite difficult since we have the problem of finding $(I-K)1_j^n$ $(j = 1 \ldots n)$ if this method is to be applied directly.

Summary and Conclusions

The main aim of this section (and indeed most of the latter part of this chapter) has been to show that if the inverse collocation matrix exists then M_n^{-1} exists and its norm can be bounded by (3.44). Thus we can use the 'a posteriori' theory and, in particular, apply Theorem 7 and its corollary.

We have seen in section 3.5 that the most suitable error bound from the theorem is $\|x-x_n\|_X \leq \|(G-T)^{-1}\| \|(G-T)x_n-y\|$ and using the corollary we have

$$\|(G-T)^{-1}\| \leq \frac{1+B_n\|\phi_n T\|}{1-\delta_n}$$

for $\delta_n = (1+B_n\|\phi_n T\|)(\|(I-\phi_n)T\|) < 1$.

In this expression B_n is bounded by (3.44) for n large enough to give $\epsilon_n < 1$. $\|\phi_n T\|$ is treated by $\|\phi_n T\| \leq \|T\|+\|(I-\phi_n)T\|$. Everything here is now calculable and for a sufficiently large number of collocation points we obtain computable error bounds by applying the above results. Numerical examples of the size of n required and of error bounds obtained by this technique are given in Chapter 5. This concludes the main theory of this chapter.

The bounds derived here and also in the next chapter ignore the effect of rounding error of which the condition number is a measure. In the next and final section some computational properties of matrices we have encountered which were mentioned briefly before are more fully analysed.

3.8 Computational Consideration of Matrices and Condition Numbers

In this section we consider some computational aspects of collocation methods by examining the structures and properties of matrices occurring in the application of the methods, the use of scaling and lastly the condition numbers.

In section 3.6 it was stated that, using Chebyshev zeros as collocation points, when powers were used in the basis for X_n the inverse collocation matrix has an unpredictable form and it's norm grows wildly as the chosen number of nodes increases. However when Chebyshev polynomials are used in the representation of the approximate solution it is found that the inverse matrix has a structure with the elements in any column generally decreasing in magnitude as the row number increases with the largest elements in the first row. Furthermore the infinity norm of the inverse matrix is more or less constant with different numbers of collocation points. This norm is in fact determined by the sum of the elements in the first row since these turn out to be positive.

An illustration of these properties is given in TABLES 7a-7d when the collocation method is applied using the zeros of the Chebyshev polynomial of degree 10 as the nodes to the sample operator $x''+(1+t^2)x$. The tables show the original and inverse matrices from using both powers and Chebyshev polynomials in the representation of the approximation. TABLE 7e shows for the same example the values to 3 significant figures of the norms of the two inverse matrices for varying numbers n of collocation points. This demonstrates how the norm of the inverse matrix from using powers increases with n.

1.95	5.88	9.66	13.3	16.8	20.2	23.4	26.5	29.5	32.3
1.63	5.02	7.23	8.54	9.15	9.23	8.93	8.35	7.58	6.70
1.25	3.71	3.62	2.56	1.31	0.221	-0.594	-1.13	-1.42	-1.54
1.04	2.29	0.276	-0.942	-1.24	-1.08	-0.792	-0.528	-0.330	-0.197
1.00	0.782	-1.73	-0.866	-0.276	-0.0727	-0.0172	-0.00377	-0.000789	-0.000159
1.00	-0.782	-1.73	0.866	-0.276	0.0727	-0.0172	0.00377	-0.000789	0.000159
1.04	-2.29	0.276	0.942	-1.24	1.08	-0.792	0.528	-0.330	0.197
1.25	-3.71	3.62	-2.56	1.31	-0.221	-0.594	1.13	-1.42	1.54
1.63	-5.02	7.23	-8.54	9.15	-9.23	8.93	-8.35	7.58	-6.70
1.95	-5.88	9.66	-13.3	16.8	-20.2	23.4	-26.5	29.5	-32.3

TABLE 7a

INVERSE MATRIX USING POWERS IN THE BASIS

0.000872	0.0157	0.0676	0.147	0.235	0.235	0.147	0.0676	0.0157	0.000872
0.000271	0.0102	0.0385	0.0924	0.118	-0.118	-0.0924	-0.0385	-0.0102	-0.000271
-0.00739	0.0331	-0.0161	0.171	-0.198	-0.198	0.171	-0.0161	0.0331	-0.00739
-0.00242	0.0180	0.00856	0.149	-0.574	0.574	-0.149	-0.00856	-0.0180	0.00242
0.0571	-0.175	0.385	-0.582	0.299	0.299	-0.582	0.385	-0.175	0.0571
0.0366	-0.121	0.347	-0.834	1.30	-1.30	0.834	-0.347	0.121	-0.0366
-0.127	0.402	-0.638	0.628	-0.265	-0.265	0.628	-0.638	0.402	-0.127
-0.0964	0.340	-0.682	1.05	-1.29	1.29	-1.05	0.682	-0.340	0.0964
0.0878	-0.231	0.288	-0.233	0.0889	0.0889	-0.233	0.288	-0.231	0.0878
0.0726	-0.212	0.332	-0.419	0.464	-0.464	0.419	-0.332	0.212	-0.0726

TABLE 7b

COLLOCATION MATRIX USING CHEBYSHEV POLYNOMIALS IN THE BASIS

0.976	5.88	17.4	35.6	59.1	86.0	114.	141.	164.	181.
0.815	5.02	12.8	19.1	17.0	1.99	-24.9	-56.5	-80.8	-84.7
0.625	3.71	6.0	-0.884	-17.2	-29.2	-18.0	20.7	65.2	77.3
0.521	2.29	-0.49	-10.6	-11.1	13.0	38.1	18.4	-45.4	-79.3
0.500	0.782	-4.46	-5.81	12.6	20.1	-19.4	-46.1	16.5	81.7
0.500	-0.782	-4.46	5.81	12.6	-20.1	-19.4	46.1	16.5	-81.7
0.521	-2.29	-0.49	10.6	-11.1	-13.0	38.1	-18.4	-45.4	79.3
0.625	-3.71	6.0	0.884	-17.3	29.2	-18.0	-20.7	65.2	-77.3
0.815	-5.02	12.	-19.1	17.0	-1.99	-24.9	56.5	-80.8	84.7
0.976	-5.88	17.4	-35.6	59.1	-86.0	114.	-141.	164.	-181.

TABLE 7c

INVERSE MATRIX USING CHEBYSHEV POLYNOMIALS IN THE BASIS

0.00565	0.0581	0.166	0.294	0.380	0.380	0.294	0.166	0.0581	0.00565
0.00432	0.0293	0.0522	0.0512	0.0212	-0.0212	-0.0512	-0.0522	-0.0293	-0.00432
0.00361	0.0164	0.0112	-0.0125	-0.0346	-0.0346	-0.0125	0.0112	0.0164	0.00361
0.00301	0.00848	-0.00428	-0.0161	-0.00952	0.00952	0.0161	0.00428	-0.00848	-0.00301
0.00248	0.00293	-0.00855	-0.00589	0.00722	0.00722	-0.00589	-0.00855	0.00293	0.00248
0.00195	-0.000241	-0.00621	0.00386	0.00504	-0.00504	-0.00386	0.00621	0.000241	-0.00195
0.00151	-0.00189	-0.00194	0.00506	-0.00272	-0.00272	0.00506	-0.00194	-0.00189	0.00151
0.00105	-0.00215	0.00102	0.00168	-0.00388	0.00388	-0.00168	-0.00102	0.00215	-0.00105
0.000686	-0.00181	0.00225	-0.00182	0.000695	0.000695	-0.00182	0.00225	-0.00181	0.000686
0.000284	-0.000828	0.0013	-0.00164	0.00181	-0.00181	0.00164	-0.00130	0.000828	-0.000284

TABLE 7d

Norms of Inverse Matrices

n	5	10	15	20	25
Powers	0.932	6.93	210	8.14'3	3.84'5
Chebyshev Polynomials	1.81	1.81	1.81	1.81	1.81

TABLE 7e

For brevity when powers are used to represent the approximate solution we shall call the inverse of the collocation matrix the "powers inverse matrix" and similarly when Chebyshev polynomials are used in the basis for X_n we shall call the corresponding matrix the "Chebyshev inverse matrix".

The above mentioned properties of the Chebyshev inverse matrix are not really surprising as the following discussion suggests.

Consider the collocation method applied to the problem Gx-Tx = f with the usual end conditions. We shall later choose f in an appropriate manner. With n collocation points let the linear equations to be solved for the coefficients of the approximation be $A\underline{a} = \underline{f}$ where $\underline{f} = (f_1, f_2 \ldots f_n)^t$ with $f_i = f(t_i)$ $(i = 1,2, \ldots n)$ and $\{t_i\}_{i=1}^n$ as the collocation nodes. Then $\underline{a} = A^{-1}\underline{f}$ and if $A^{-1} = (v_{ij})$

$$\Rightarrow \quad a_o = \sum_{j=1}^{n} v_{1j} f_j \tag{3.45}$$

If we take $f(t) \equiv 1$ then $a_o = \sum_{j=1}^{n} v_{1j}$, that is, the sum of the first row of A^{-1}. Now we would not expect that

a_o would vary greatly as the number of collocation points is increased and so we would anticipate that the sum of the elements in the first row of A^{-1} would be roughly constant. If these elements are positive then this would give that the sum of the moduli of the terms in the first row of A^{-1} was reasonably constant.

In particular we shall investigate the simple second order equation of the form $x'' = f$ subject to $x(-1) = x(+1) = 0$. With $f(t) \equiv 1$, $x(t) = \frac{1}{2}(t^2-1)$, so that if an approximation of the form $(t^2-1)\sum_{j=0}^{n-1}{}' a_j T_j(t)$ is sought and the collocation equations are $A_o\underline{a} = \underline{f}$ then clearly we must have $a_o = 1$, $a_j = 0$ $(j = 1 \ldots n-1)$. Thus

$$\sum_{j=1}^{n} \alpha_{1j} = 1 \tag{3.46}$$

where $A_o^{-1} = (\alpha_{ij})$ and we see that for this problem the sum of the elements in the first row of the inverse is constant.

It is now shown that the elements in the first row of A_o^{-1} all have the same sign so that $|\sum_{j=1}^{n} \alpha_{1j}| = \sum_{j=1}^{n} |\alpha_{1j}|$.

If $x(t)$ satisfies $x''(t) = f(t)$ with $f(t)$ a polynomial degree $\leq n-1$ and $x(t) = (t^2-1)z(t)$ then $z(t)$ must be a polynomial of up to degree $n-1$, $\sum_{j=0}^{n-1}{}' a_j T_j(t)$ say, so that

$$a_o = \frac{2}{\pi}\int_{-1}^{+1} \frac{z(t)T_o(t)}{\sqrt{1-t^2}}dt \text{ (using the orthogonality)}$$

$$\Rightarrow \quad a_o = -\frac{2}{\pi}\int_{-1}^{+1} (1-t^2)^{-\frac{3}{2}} x(t)dt. \tag{3.47}$$

Now from (3.47) with the substitution $t = \sin\tau$ and using

integration by parts and the end conditions on x(t) we can obtain

$$a_o = \frac{2}{\pi}\int_{-1}^{+1}(1-t^2)^{\frac{1}{2}}x''(t)dt = \frac{2}{\pi}\int_{-1}^{+1}(1-t^2)^{\frac{1}{2}}f(t)dt. \qquad (3.48)$$

For a particular value of n take f(t) as $l_k^n(t)$ which is a polynomial of degree n-1, then in this case $\underline{f}$ is such that $f_k = 1$, $f_j = 0$ $(j \neq k)$. Since the collocation method for this problem will give the true solution and this right hand side gives the special unit vector described above we must have that the coefficient vector $[a_o^{(k)}, a_1^{(k)}, \ldots a_{n-1}^{(k)}]^t$ say, in this case is equal to the k^{th} column of A_o^{-1}. In particular by (3.45) $a_o^{(k)} = \sum_{j=1}^{n} \alpha_{1j}f_j = \alpha_{1k}$. We thus have using (3.48) with $f(t) = l_k^n(t)$

$$a_o^{(k)} = \frac{2}{\pi}\int_{-1}^{+1}(1-t^2)^{\frac{1}{2}}l_k^n(t)dt = \frac{2}{\pi}\int_{-1}^{+1}(1-t^2)^{-\frac{1}{2}}(1-t^2)l_k^n(t)dt$$

$$= \frac{2}{\pi}\sum_{i=1}^{n}\frac{\pi}{n}(1-t_i^2)l_k^n(t_i) \text{ (since Gauss quadrature will}$$

be exact),

$$\Rightarrow a_o^{(k)} = \frac{2}{n}(1-t_k^2). \qquad (3.49)$$

Therefore as $|t_k| < 1$, $a_o^{(k)}$ is positive giving α_{1k} positive (k = 1, ... n) and the modulus of the sum of the first row of A_o^{-1} is equal to the sum of the moduli of the elements.

Equation (3.46) then gives $\sum_{j=1}^{n}|\alpha_{1j}| = 1$ and if we knew that for any column the elements of largest

magnitude were in the first row, this would give $\|A_o^{-1}\| = 1$ irrespective of the number of collocation points. Now

$$|a_j^{(k)}| = \frac{2}{\pi}\left| \int_{-1}^{+1} \frac{z(t)T_j(t)}{\sqrt{1-t^2}}dt \right|$$

$$= \frac{2}{\pi}\left| - \int_{-1}^{+1} (1-t^2)^{-\frac{3}{2}} T_j(t)x(t)dt \right|$$

$$\leq \frac{2}{\pi} \int_{-1}^{+1} (1-t^2)^{-\frac{3}{2}} |x(t)|dt$$

where $a_j^{(k)}$ is the a_j corresponding to the right hand side $l_k^n(t)$ and $a_j^{(k)} = \alpha_{j+1,k}$ $(j = 1 \ldots n-1)$ similarly as for $a_o^{(k)}$. So if $x(t)$ is of one sign then $|a_j^{(k)}| \leq |a_o^{(k)}|$. Now $x''(t) = l_k^n(t) \Rightarrow x(s) = \int_{-1}^{+1} g(s,t) l_k^n(t)dt$ where $g(s,t)$ is the Green's function of section 1.4 and applying Gauss-Chebyshev quadrature we have

$$x(s) \simeq x^*(s) = \sum_{i=1}^{n} \frac{\pi}{n} g(s,t_i) l_k^n(t_i)$$

$$= \begin{cases} \frac{\pi}{2n}(s+1)(t_k-1) & s \leq t_k \\ \frac{\pi}{2n}(s-1)(t_k+1) & s > t_k \end{cases}.$$

Thus in either case we can say that $x(s) \simeq x^*(s) < 0$, confirming that $|a_j^{(k)}|$ will usually be less than $|a_o^{(k)}|$.

This then suggests that the largest elements of any column of A_o^{-1} occur in the first row and together with (3.46) and (3.49) leads us to expect that $\|A_o^{-1}\|$ is constant (where in fact the constant is 1) with varying n.

We are generally concerned however with operators of the form G-T with T not the zero operator and for problems of this type often the Chebyshev collocation and inverse Chebyshev matrices are not of a substantially different structure to the simple case discussed. With this assumption the above analysis hints that again the norms of the inverse Chebyshev matrices, that is $\|A^{-1}\|$, might be reasonably constant with varied numbers of collocation points.

With powers in the basis for X_n we do not have the orthogonality result which has been utilised above and we are unable to come to similar possible conclusions.

Although we have been considering collocation with Chebyshev nodes, Legendre zeros lead in practice to similar results concerning the norm of the inverse Chebyshev matrix as is shown in TABLE 8 below for the sample operator $x''+(1+t^2)x$.

Norms of Inverse Chebyshev Matrices using Legendre Zeros

n	3	5	7	10	16
Norm of Inverse Chebyshev Matrix	1.761146	1.807759	1.807565	1.807561	1.807561

TABLE 8

These values can be compared to those given in TABLE 7e or in more detail to TABLE 4 when Chebyshev nodes are used. The similarity of the norms of the inverse matrices is probably due to the fact that for larger values of n the corresponding zeros of the Chebyshev and Legendre

polynomials of degree n are close.

Having discussed the form of the inverse collocation matrices we turn to a topic which utilises the above properties and consider the effect of column scaling. If we have a matrix A then column scaling of A is equivalent to postmultiplying A by a diagonal matrix D = diag $(d_1, d_2, \ldots d_n)$ say. That is, if B = AD then the elements in the j^{th} column of B are d_j times those in the corresponding column of A for j = 1, ... n.

In the notation we have used throughout consider the matrix A we obtain in the application of the collocation method to the approximate solution of a second order linear differential equation Gx-Tx = y with the usual end conditions by seeking an approximation x_n of the form $(t^2-1) \sum_{r=1}^{n} a_r \psi_r(t)$. The $\{\psi_r(t)\}$ are taken to be polynomials. Normally we have to solve the linear equations, $A\underline{a} = \underline{\gamma}$ say, for the coefficients $\underline{a}$. However we can solve a different set of equations $B\underline{b} = \underline{\gamma}$ where B = AD for D diagonal and represent x_n by $(t^2-1) \sum_{r=1}^{n} b_r \zeta_r(t)$ where $\{\zeta_r\}_{r=1}^{n}$ are some set of polynomials. Then since we must find the same approximate solution x_n this means $\zeta_r(t) = d_r \psi_r(t)$ (r = 1 ... n) since $b_r = \frac{1}{d_r} a_r$ and we see that column scaling of the collocation matrix A is equivalent to a certain transformation of the basis in X_n. That is, if $\{(t^2-1)\zeta_r(t)\}_{r=1}^{n}$ were chosen to represent the approximate solution then the collocation matrix would be B = AD.

For our purposes column scaling can be utilised in

principally two ways. The former of these concerns the mapping Γ introduced in section 3.6 and used in (3.21) to bound the norm of the inverse of the approximate operator. Our latter application of scaling is to improve the condition number of the linear collocation equations and this is dealt with later.

Consider now the mapping $\Gamma:R^n \to X_n$ such that $\Gamma(\underline{b}) = (t^2-1)\sum_{r=1}^{n} b_r\psi_r(t)$ $(\underline{b}\epsilon R^n)$ where $\{(t^2-1)\psi_r(t)\}_{r=1}^{n}$ is a basis for X_n. Equation (3.21) gave the bound $B_n \leq \|\Gamma\|\|A^{-1}\|\|\rho^{-1}\|$. Column scaling can now be used on the matrix A to determine a basis for X_n such that $\|\Gamma\|$ is greatly reduced in comparison with the original bases of section 3.6 and such that $\|A^{-1}\|$ remains bounded as more collocation points are chosen. This process is now described.

If we column scale with $D = \text{diag}\,(d_1,d_2 \ldots d_n)$ then we saw above that this means we change the basis in X_n from $\{(t^2-1)\psi_r(t)\}$ to $\{(t^2-1)\zeta_r(t)\} = \{(t^2-1)d_r\psi_r(t)\}$. Thus it is most likely that we shall have to choose $|d_r| \leq 1$ if we are to reduce $\|\Gamma\| = \|\Gamma_\zeta\|$ say, using the new basis since

$$\|\Gamma_\zeta\| = \sup_{\substack{b\epsilon R^n \\ \|\underline{b}\|=1}} \|(t^2-1)\sum_{r=1}^{n} b_r\zeta_r(t)\|_X.$$

Now when A is column scaled this means A^{-1} is row scaled since $(AD)^{-1} = D^{-1}A^{-1}$. We have discussed earlier the structure of the matrix A^{-1} when an approximate solution is sought in the form $(t^2-1)\sum_{r=0}^{n-1}{}'a_rT_r(t)$ and Chebyshev zeros are used as the nodes. Recall that the elements in any column

generally decreased in magnitude with increasing row number with the first row as the one with the largest elements. Further it was mentioned in section 3.6 (TABLE 4) that $\|HA^{-1}\|$ was found to be reasonably constant with varying n, where H was the matrix diag $(h_1, h_2, \ldots h_n)$ with $h_1 = 1$, $h_i = (i-1)^2$ $(i=2 \ldots n)$. Thus we can take $D = H^{-1}$ so that $d_1 = 1$, $d_i = \frac{1}{(i-1)^2}$ $(i=2 \ldots n)$ and with this choice we expect $\|B^{-1}\| = \|(AD)^{-1}\| = \|HA^{-1}\|$ to vary little as more collocation points are used. Two further examples, to support TABLE 4, of the variations of $\|A^{-1}\|$ and $\|B^{-1}\|$ are shown in TABLE 9 below.

With this choice of D we have determined a different basis $\{(t^2-1)\zeta_r(t)\}_{r=1}^{n}$ in X_n where $\zeta_1 = T_o$, $\zeta_r = \frac{T_{r-1}(t)}{(r-1)^2}$ $(r = 2 \ldots n)$. So now

$$\|\Gamma_\zeta\| = \sup_{\|\underline{b}\|=1} \|\frac{d^2}{dt^2}\{(t^2-1) \sum_{r=0}^{n-1}{}' b_r d_r T_{r-1}(t)\}\|_\infty$$

[where now $\underline{b} = (b_o, b_1 \ldots b_{n-1})^t$]. Substitution in (3.23) and rearranging gives

$$\|\Gamma_\zeta\| = \sup_{\|\underline{b}\|=1} \{b_o T_o(t) + 2 \sum_{r=1}^{n-1} \frac{b_r}{r^2} T_r(t) + \sum_{r=2}^{n-1} b_r T_r(t) + 3t \sum_{r=1}^{n-1} \frac{b_r}{r^2} T_r'(t) + t b_1 T_1'(t)\}.$$

On using $|T_r'(t)| \leq r^2$ this simplifies to

$$\|\Gamma_\zeta\| \leq 1 + 2 \sum_{r=1}^{n-1} \frac{1}{r^2} + 4(n-1) \leq 4n-1.$$

Thus employing the scaling we have

Variation of the Norms of the Original and Column Scaled Inverse Matrices

	n	5	10	15	20	25
Gx-Tx ≡x" -x	$\|A^{-1}\|$	.7325084	.7325075	.7325075	.7325075	.7325075
	$\|B^{-1}\|$	.7325084	.9346020	1.038456	1.099757	1.133578
Gx-Tx≡ x" +(8t^2+2t-1)x' +(4.5t^2+1.5t-1)x	$\|A^{-1}\|$	.7959828	.7714242	.7713679	.7713679	.7713679
	$\|B^{-1}\|$	.9267498	.9713072	1.064915	1.113972	1.144280

TABLE 9

$B_n \leq \|\Gamma_\zeta\| \|B\| = O(4n)$ giving an improvement upon the earlier results with the original basis. This, of course, is still unsuitable for application in the formulae for error bounds but illustrates the scope of column scaling.

The condition numbers of matrices occurring in the application and theory of the collocation method are now considered and it is shown how for certain matrices column scaling can be utilised to achieve improvements.

The condition number of a matrix A is defined by $\text{cond}(A) = \|A\| \|A^{-1}\|$ and the magnitude of this number is a guide to the effect of perturbations in the matrix upon the solution of algebraic equations which require inversion of A. (For a fuller explanation see e.g. Wilkinson (1965)). Smaller values of cond(A) suggest less possible perturbation in the solution. Gaussian elimination with row interchanges is in fact invariant under column scaling but the condition number is not and we are interested in finding a scaling which will reduce the condition number. This gives a more realistic indication of the effect of rounding errors.

Now we have seen how, using Chebyshev polynomials with Chebyshev zeros as collocation points, column scaling of the collocation matrix A by $D = \text{diag}\,(d_1, d_2 \ldots d_n)$, where $d_1 = 1$, $d_i = \frac{1}{(i-1)^2}$ $(i = 2, \ldots n)$ gave values of $\|B^{-1}\| = \|(AD)^{-1}\|$ which were reasonably constant with varying n. Moreover bearing in mind the form of the matrix A seen earlier in this section with, in any row, the larger elements occurring in the later columns we should therefore expect $\|B\|$ to be considerably less than $\|A\|$ giving a much improved condition number. This is

demonstrated in TABLE 10 below in which the following three operators are used as test examples:

Ex. 1 $x'' - x$

Ex. 2 $x'' + (1+t^2)x$

Ex. 3 $x'' + (8t^2+2t-1)x' + (4.5t^2+1.5t-1)x.$

The Use of Column Scaling to Improve Condition Numbers

	n	5	10	15	20	25
Ex. 1	cond(A)	68	591	2062	4973	9815
	cond(B)	12	31	52	73	93
Ex. 2	cond(A)	165	1457	5088	12272	24220
	cond(B)	29	60	90	119	149
Ex. 3	cond(A)	96	669	2241	5329	10450
	cond(B)	37	63	89	113	136

TABLE 10

The above table shows clearly the smaller values of the condition number when this column scaling is employed and suggests a more reasonable guide to the rounding errors.

Throughout this section we have for simplicity restricted our attention to the one particular choice of scaling above but slightly different selections e.g. $d_r = \frac{1}{r^2}$ $(r = 1 \ldots n)$ also lead to similar results.

Finally we consider the condition number of the matrix AA_o^{-1} of section 3.7 whose inverse A_oA^{-1} we saw was involved with the theory of that section concerned with bounding the inverse of the approximate operator. Also it was shown that AA_o^{-1} is the matrix to be inverted

if we are solving directly for the values at the collocation points of the second derivative of the approximate solution. Examples of the condition number of this matrix are given in TABLE 11 below for which the same three sample operators of TABLE 10 are again employed to illustrate the results.

Condition Numbers of the Matrix AA_o^{-1}

	n	5	10	15	20	25
Ex. 1	$\text{cond}(AA_o^{-1})$	1.54	1.68	1.80	1.85	1.89
Ex. 2	$\text{cond}(AA_o^{-1})$	2.07	2.37	2.53	2.61	2.67
Ex. 3	$\text{cond}(AA_o^{-1})$	63.6	84.7	90.9	93.1	94.2

TABLE 11

We observe that $\text{cond}(AA_o^{-1})$ does not grow substantially with n, unlike cond(A) and cond(B) above, presumably due to the fact that the second derivative approximations satisfy a type of integral equation.

This completes the work of this section on the consideration of the numerical properties of matrices occurring in the application of collocation methods. In particular we have seen how the structure of certain matrices can be utilised by the application of column scaling to reduce condition numbers.

CHAPTER 4

APPLICATION OF COLLECTIVELY COMPACT OPERATOR APPROXIMATION THEORY

4.1 Introduction

In this chapter we are primarily concerned with the application of the theory of Anselone (1971) to the approximate solution by collocation of linear ordinary differential equations but much of the theory will by default hold for Fredholm integral equations since the differential problem is regarded as an integral one.

The approximate solution by collocation of integral equations or of boundary value problems seen in this form, which was discussed in the earlier part of Chapter 2 and in Chapter 3 does not fit directly into the setting for Anselone's theory described in section 2.7. This is clearly seen from the fact that Anselone's approach requires from (2.11) that the approximating equation have a right hand side y whereas the theory of Kantorovich and Akilov has a projection of this term (c.f. (2.2), (2.2′)).

It is demonstrated in section 4.2 how to extend the collocation method to achieve equations of the appropriate form for the theory and in section 4.3 these 'extended' equations are shown to satisfy the required conditions of the theorems given in section 2.8. In section 4.4 convergence proofs for the usual polynomial collocation method are derived from this alternative theory and it is subsequently discussed how to relate the new concepts to the familiar previously considered quantities of

Chapter 3. In section 4.7 the applicability of Theorems 9, 10 and 7 is discussed and finally a generalisation of the earlier theory is given.

4.2 Adaptation of Collocation for Differential Equations to the Theoretical Background

We have seen (sections 3.2, 3.7) how a linear boundary value problem, e.g.

$$(G-T)x \equiv \frac{d^{2m}x}{dt^{2m}} + p_{2m-1}(t)x^{(2m-1)}(t) + \ldots$$

$$+ p_1(t)x^{(1)}(t) + p_o(t)x(t) = y(t) \qquad (4.1)$$

over $[-1,1]$ with $x^{(j)}(\pm 1) = 0$ $(j = 0,1 \ldots m-1)$ and $y(t)$, $p_i(t) \in C[-1,1]$ $(i = 0,1 \ldots 2m-1)$ can be transformed to an integral equation of the form

$$x^{(2m)}(s) - \int_{-1}^{+1} k(s,t)x^{(2m)}(t)dt = y(t) \text{ where}$$

$$k(s,t) = -\{p_{2m-1}(s)\frac{\partial g^{2m-1}}{\partial s^{2m-1}}(s,t) + \ldots + p_o(s)g(s,t)\}$$

and $g(s,t)$ is the Green's function for the operator $\frac{d^{2m}}{dt^{2m}}$ subject to the above homogeneous boundary conditions. If the solution $x(t)$ to the above differential problem exists it must have a continuous $(2m)^{th}$ derivative and $u \equiv x^{(2m)}$ satisfies the operator equation $(I-K)u = y$ in $C[-1,1]$, where

$$(Kv)(s) = \int_{-1}^{+1} k(s,t)v(t)dt \quad (v \in C[-1,1]).$$ Here $K \in [C]$ since $k(s,t)$ has only a jump discontinuity at $s = t$ in the closed interval $[-1,1]$ and

$\|Kv\|_\infty \leq \|v\|_\infty . \max_s \{\int_{-1}^{+1} |k(s,t)|dt\}$. Now the space of continuous functions with the infinity norm is a Banach space so that since we are attempting to fit our problem to the setting of section 2.7 we take the space X of that section as $C[-1,1]$ and we have $y \in X$ and $K \in [X]$. The given equation is $(I-K)u = y$, where we have u replacing the x of (2.10). If x_n is the usual approximation to x yielded by the collocation method applied to the differential equation (4.1) then we have seen in section 2.2 that x_n satisfies an equation of the form $Gx_n - \phi_n T x_n = \phi_n y$ (where ϕ_n constituted polynomial interpolation at the collocation points $\{t_j\}_{j=1}^n$, i.e. $\phi_n y = \sum_{j=1}^n l_j^n(t) y(t_j)$ for $y \in C[-1,1]$). Thus $u_n = x_n^{(2m)} = Gx_n$ satisfies

$$(I-\phi_n K)u_n = \phi_n y \qquad (4.2$$

since $T \equiv KG$.

To achieve the desired framework for Anselone's theory we need somehow to modify our collocation method to obtain approximating equations of the form $(I-K_n)z_n = y$ with $K_n \in [X]$ and $z_n \in X$ (replacing x_n in (2.11)) an approximation to u. This process is now described.

With u_n as the second derivative of the approximate solution found by straightforward application of our collocation method we make the following definitions. For each n = 1,2 ... define

$z_n \epsilon X$ by

$$z_n = y + Ku_n \tag{4.3}$$

$\Rightarrow \phi_n z_n = \phi_n y + \phi_n K u_n = u_n$ by (4.2). Let $K_n : X \to X$ be such that, with $v \epsilon X$

$$K_n v = K\phi_n v. \tag{4.4}$$

Then $(I-K_n)z_n = (I-K\phi_n)(y + Ku_n)$

$$= y - K\phi_n y + Ku_n - K\phi_n Ku_n$$

$$= y + K(u_n - \phi_n Ku_n - \phi_n y).$$

Thus by (4.2)

$$(I-K_n)z_n = y. \tag{4.5}$$

With these definitions we shall call z_n the 'extended' collocation approximation and (4.5) the 'extended' approximate equation.

This approach is similar to the Nyström extension of the quadrature method applied to Fredholm integral equations which is considered by Anselone and this is indicated as follows. If a quadrature is applied to an integral equation say

$u(s) - \int_{-1}^{+1} k(s,t)u(t)dt = y(s)$ for a general kernel $k(s,t)$ then let $\{v_j\}_{j=1}^{n}$ be obtained as approximate values to u at the nodes. The Nyström extension, $v(s)$ say, gives approximate values between the nodes $\{t_j\}_{j=1}^{n}$ by

$$v(s) = y(s) + \sum_{j=1}^{n} w_j k(s,t_j) v_j \qquad (4.6)$$

where the $\{w_j\}$ are the appropriate weights. Equation (4.6) is then analagous to (4.3) which may be rewritten as

$z_n(s) = y(s) + \int_{-1}^{+1} k(s,t) \sum_{j=1}^{n} l_j^n(t) u_n(t_j) dt$ (with $\{t_j\}$ as the collocation points). Rearranging gives

$z_n(s) = y(s) + \sum_{j=1}^{n} \left(\int_{-1}^{+1} k(s,t) l_j^n(t) dt \right) u_n(t_j)$, illustrating the similarity.

A further demonstration of the meaning of the extended approximation is to compare directly the equations satisfied by u_n and z_n. We have

$$(I-\phi_n K) u_n = \phi_n y,$$

$$(I-K\phi_n) z_n = y$$

and also $z_n - Ku_n = y$ by (4.3).

Thus $z_n(s) = y(s) + \int_{-1}^{+1} k(s,t) u_n(t) dt$

$$= y(s) + (Tx_n)(s)$$

$$= y(s) - \{p_{2m-1}(s) x_n^{(2m-1)}(s) + \ldots$$

$$+ p_o(s) x_n(s)\}$$

so that if $c_n(s)$ is such that $c_n^{(2m)}(s) = z_n(s)$ and c_n satisfies the end conditions of (4.1) then $c_n^{(2m)}(s) + p_{2m-1}(s) x_n^{(2m-1)}(s) + \ldots + p_o(s) x_n(s) = y(s)$. Notice that our extended approximation z_n is no longer a polynomial as was u_n and this has been necessary to satisfy an equation of the form (4.5) with right hand side y.

We shall not subsequently actually solve for z_n but the theory of its solution, in particular the inversion of $I-K_n$, can be used, as we shall see later in this chapter, in the theorems of Chapter 2 to bound the norm of $(I-K)^{-1}$. We can then relate the bound on $\|(I-K)^{-1}\|$ derived from $\|(I-K_n)^{-1}\|$ to $\|(G-T)^{-1}\|$ by the following argument. Recall the relationship between the operators T and K, viz. $T \equiv KG$. Thus $I-K = I-TG^{-1} = (G-T)G^{-1}$ and $(I-K)^{-1} = G(G-T)^{-1} \Rightarrow \|(I-K)^{-1}\|_\infty = \|G(G-T)^{-1}\|_\infty = \|(G-T)^{-1}\|_X$ where this last term is the usual norm of the inverse operator which we encountered in the former sections of Chapter 2 and in Chapter 3.

The error in the solution from the usual collocation method is $x-x_n = (G-T)^{-1}(y-(G-T)x_n)$

$$\Rightarrow \|x-x_n\|_X = \|(G-T)^{-1}\|_X \|y-(G-T)x_n\|, \qquad (4.7)$$

(c.f. Theorem 7). Thus we see that if by the theory of sections 2.7 and 2.8 we are able to bound $\|(I-K)^{-1}\|$ we can then bound $\|(G-T)^{-1}\|_X$ and hence obtain error bounds by (4.7).

Computational considerations and numerical results of applying this strategy are given in Chapter 5.

4.3 Satisfaction of the Criteria for the Application of the Theorems

We show here that the operators K, K_n $(n = 1,2 \ldots)$ defined in the previous section do indeed satisfy the conditions required for the theory of sections 2.7 and 2.8 provided we use orthogonal polynomial zeros as

collocation points, e.g. Chebyshev zeros, Legendre zeros etc.

We therefore wish to prove $K_n \to K$, K is compact and $\{K_n\}$ is collectively compact.

Lemma 1 The sequence $\{K_n\}$ is uniformly bounded.

Proof $\|K_n\| = \sup_{\substack{v\in X \\ \|v\|=1}} \|K_n v\| = \sup_{\|v\|=1} \max_s |(K_n v)(s)|$.

Now $|(K_n v)(s)| = |(K\phi_n v)(s)| = \left|\int_{-1}^{+1} k(s,t)(\phi_n v)(t)dt\right|$

$$\leq \left\{\int_{-1}^{+1}[k(s,t)]^2 dt\right\}^{\frac{1}{2}} \left\{\int_{-1}^{+1}[(\phi_n v)(t)]^2 dt\right\}^{\frac{1}{2}}$$ by Cauchy's

Inequality. $|k(s,t)|$ has been discussed previously and is bounded independently of n, of course. Now $(\phi_n v)(t)$ is a polynomial of degree n-1 and so Gaussian quadratures for the integration of $[(\phi_n v)(t)]^2$ will be exact. Thus if we choose Chebyshev zeros as the collocation points $\{t_j\}$ then

$$\int_{-1}^{+1}[(\phi_n v)(t)]^2 dt \leq \int_{-1}^{+1}(1-t^2)^{-\frac{1}{2}}[(\phi_n v)(t)]^2 dt = \sum_{i=1}^{n} w_i v^2(t_i)$$

where $\{w_i\}$ are the weights of the quadrature formula. Now we can say

$$\sup_{\|v\|=1}\left\{\sum_{i=1}^{n} w_i v^2(t_i)\right\}^{\frac{1}{2}} \leq \left\{\sum_{i=1}^{n} w_i\right\}^{\frac{1}{2}} = \pi^{\frac{1}{2}}$$

(see e.g. Natanson (1965, p.104)).

Thus $\|K_n\|$ can be bounded independently of n and $\{K_n\}$ is uniformly bounded■

Lemma 2 $K_n \to K$.

Proof To verify the above statement we have to show $K_n v \to Kv$ (for all $v \in X$). Now

$$\|K_n v - Kv\| = \max_s \left| \int_{-1}^{+1} k(s,t)[(\phi_n v)(t) - v(t)]\,dt \right|$$

$$\leq \max_s \left\{ \int_{-1}^{+1} [k(s,t)]^2\,dt \right\}^{\frac{1}{2}} \left\{ \int_{-1}^{+1} [(\phi_n v)(t) - v(t)]^2\,dt \right\}^{\frac{1}{2}}.$$

As before the first factor is independent of v and n. Now $v \in C[-1,1]$ and therefore we have

$$\lim_{n\to\infty} \int_{-1}^{+1} (1-t^2)^{-\frac{1}{2}} [(\phi_n v)(t) - v(t)]^2\,dt = 0$$

since this is the convergence result for the interpolation of continuous functions in the weighted L_2 norm, (see e.g. Natanson (1965, p.55)).

Lemma 3 K is compact and $\{K_n\}$ is collectively compact.

Proof For K to be compact we need KU to be relatively compact in X (where U is the unit ball $\{v \in X : \|v\| \leq 1\}$). To prove $\{K_n\}$ is collectively compact we require the set $\mathcal{K}U = \{K_n v : n \in N, v \in U\}$ to be relatively compact, (N being the set of positive integers).

These results are obtained by means of the Arzelà-Ascoli theorem, given for example by Kantorovich and Akilov (1964, p.22), by proving equicontinuity and uniform boundedness of the appropriate sets. Thus for $v \in U$, $\|Kv\| \leq \|K\|$ giving KU uniformly bounded. Now for $-1 \leq s_1 < s_2 \leq +1$,

$$|(Kv)(s_1)-(Kv)(s_2)| = \left|\int_{-1}^{+1}[k(s_1,t)-k(s_2,t)]v(t)dt\right|$$

$$\leq \int_{-1}^{+1}|k(s_1,t)-k(s_2,t)|dt.$$

In general $k(s,t)$ will have a discontinuity at $s=t$ and in view of this we split the above range of integration by $\int_{-1}^{+1} = \int_{-1}^{s_1} + \int_{s_1}^{s_2} + \int_{s_2}^{+1}$. For t in the intervals $[-1,s_1)$ and $(s_2,1]$ $k(s,t)$ is a continuous function of s for $s > s_1$ and $s < s_2$ respectively, and the corresponding integrals can be made arbitrarily small by choosing $|s_2-s_1|$ sufficiently small. Now $\int_{s_1}^{s_2}|k(s_1,t) - k(s_2,t)|dt \leq 2|s_2-s_1|\max_{s,t}|k(s,t)|$ and so this term can again be made arbitrarily small. Thus we have proved equicontinuity and by the Arzelà-Ascoli theorem we have that K is compact.

In Lemma 1 we showed $\{K_n\}$ was uniformly bounded and thus it only remains to satisfy the equicontinuity condition for $\mathcal{K}U$. As before with $-1 \leq s_1 < s_2 \leq 1$ say and $v \in U$

$$|(K_nv)(s_1)-(K_nv)(s_2)| = \left|\int_{-1}^{+1}[k(s_1,t)-k(s_2,t)](\phi_nv)(t)dt\right|$$

$$\leq \left\{\int_{-1}^{+1}[k(s_1,t)-k(s_2,t)]^2dt\right\}^{\frac{1}{2}}\left\{\int_{-1}^{+1}[(\phi_nv)(t)]^2dt\right\}^{\frac{1}{2}}.$$

We now deal with this expression by treating the former factor by splitting the range of integration as earlier in this lemma and the latter by the technique of Lemma 1. This proves equicontinuity and hence that $\mathcal{K}U$ is relatively compact in X, showing $\{K_n\}$ is collectively compact■

We have now satisfied the required conditions on the operators K, $\{K_n\}$ for Theorems 8, 9 and 10 so that if we assume that the appropriate inverse operators exist then these results can be applied to give convergence proofs and error bounds for $\|u-z_n\|$. As was mentioned in section 4.2 we do not actually solve for z_n but in the next section we see it can be used for convergence proofs for $\|x-x_n\|$ and in sections 5 and 6 we consider a more qualitative approach.

4.4 Convergence Proofs for the Usual Polynomial Collocation Method

We here give alternative convergence proofs to those of Kantorovich and Akilov type for the ordinary polynomial approximation x_n we have used in the earlier part of Chapter 2 and in Chapter 3. Recall that $u_n = Gx_n$ is also a polynomial.

Firstly we note that if we assume $(I-K)^{-1}$ exists then Theorem 8 gives, in the infinity norm, $\|u-z_n\| \to 0$. Now

$$u-u_n = u-\phi_n u+\phi_n u-u_n$$
$$= u-\phi_n u+\phi_n(u-z_n) \quad \text{(since } \phi_n z_n = u_n\text{)}.$$

So in any norm $\|u-u_n\| \leq \|u-\phi_n u\| + \|\phi_n(u-z_n)\|$. However if u is merely continuous then in the infinity norm $\|u-\phi_n u\| \not\to 0$ in general. This suggests the use of an L_2 norm which we take as the one with the Chebyshev weight function and we denote this norm by $\|\cdot\|_2$.

i.e. $\|v\|_2 = \{\int_{-1}^{+1} w(t)v^2(t)dt\}^{\frac{1}{2}}$ $(v \in C[-1,1])$

with $w(t) = (1-t^2)^{-\frac{1}{2}}$. Then in this norm we have $\|u-u_n\|_2 \leq \|u-\phi_n u\|_2 + \|\phi_n(u-z_n)\|_2$ and the term

$\|u-\phi_n u\|_2 \to 0$. (Natanson (1965, p.55)). Now

$$\|\phi_n(u-z_n)\|_2^2 = \int_{-1}^{+1} w(t)[\phi_n(u-z_n)(t)]^2 dt$$

$$= \sum_{j=1}^{n} w_j[u(t_j)-z_n(t_j)]^2 \text{ where } \{w_j\}_{j=1}^n \text{ and } \{t_j\}_{j=1}^n$$

are the weights and nodes respectively. Therefore

$$\|\phi(u-z_n)\|_2^2 \leq \pi \max_j [u(t_j)-z_n(t_j)]^2.$$

Now $\|u-z_n\|_\infty \to 0$ from above and therefore $\|u-u_n\|_2 \to 0$ as $n \to \infty$. Thus to emphasise this point we have $u_n \to u$ in the L_2 norm whereas $z_n \to u$ in the infinity norm.

From the theory of Kantorovich and Akilov applied in section 3.2 we had $\|x-x_n\|_X = \|u-u_n\|_\infty \to 0$ as $n \to \infty$ but this was only after we had required some extra continuity of derivatives from the coefficients and right hand side in the differential equation. This result of convergence in the weighted L_2 norm when no extra continuity conditions are assumed agrees with that of Vainikko (1965).

To obtain the convergence result $\|x-x_n\|_\infty \to 0$ as $n \to \infty$ we proceed as follows.

$$|x(s)-x_n(s)| = \int_{-1}^{+1} g(s,t)[u(t)-u_n(t)]dt \text{ where } g(s,t)$$

where $g(s,t)$ is the usual Green's function for G with the given homogeneous boundary conditions. We can now get $|x(s)-x_n(s)| \leq \{\int_{-1}^{+1}[g(s,t)]^2 dt\}^{\frac{1}{2}} \{\int_{-1}^{+1}[u(t)-u_n(t)]^2 dt\}^{\frac{1}{2}}$ by Cauchy's inequality. The former integral is bounded and the latter is less than

$$\left\{\int_{-1}^{+1} w(t)[u(t)-u_n(t)]^2\,dt\right\}^{\frac{1}{2}} = \|u-u_n\|_2 \to 0 \text{ as } n \to \infty.$$

This proves $\|x-x_n\|_\infty \to 0$ as $n \to \infty$.

4.5 The Relationship between the Inverses of the 'Extended' and the 'Usual' Approximate Operators

In section 4.2 we defined the extended approximation z_n and this was shown to satisfy $(I-K_n)z_n = y$ or $(I-K\phi_n)z_n = y$, whereas the usual polynomial approximation u_n to $u = Gx$ satisfies $(I-\phi_n K)u_n = \phi_n y$. In this section we establish the connection between the inverses of the operators $I-K_n$ and $I-\phi_n K$.

Assume that $(I-\phi_n K)$ restricted to the polynomial subspace of $C[-1,1]$ has an inverse denoted by $(I-\phi_n K)^{-1}$. Now take any $y \in X = C[-1,1]$ then

$$u_n = (I-\phi_n K)^{-1}\phi_n y \qquad (4.8)$$

satisfies $(I-\phi_n K)u_n = \phi_n y$. For this y and u_n define z_n by (4.3) then

$(I-K_n)z_n = y$ or $(I-K_n)(y+Ku_n) = y$

$\Rightarrow (I-K_n)(y+K(I-\phi_n K)^{-1}\phi_n y) = y$ by (4.8)

$\Rightarrow (I-K_n)(I+K(I-\phi_n K)^{-1}\phi_n)y = y$

and $I+K(I-\phi_n K)^{-1}\phi_n$ is a right side inverse of $I-K_n$.

Now we also wish to show that this operator is also a left side inverse and so is the unique inverse of $I-K_n$. Thus,

$$[I+K(I-\phi_n K)^{-1}\phi_n][I-K_n]y = [I+K(I-\phi_n K)^{-1}\phi_n][I-K\phi_n]y$$
$$= y+K(I-\phi_n K)^{-1}\phi_n y - K\phi_n y - K(I-\phi_n K)^{-1}\phi_n K\phi_n y$$
$$= y+K[(I-\phi_n K)^{-1}-I]\phi_n y - K(I-\phi_n K)^{-1}\phi_n K\phi_n y$$
$$= y+K(I-\phi_n K)^{-1}[I-(I-\phi_n K)]\phi_n y - K(I-\phi_n K)^{-1}\phi_n K\phi_n y$$
$$= y+K(I-\phi_n K)^{-1}\phi_n K\phi_n y - K(I-\phi_n K)^{-1}\phi_n K\phi_n y$$
$$= y$$

Therefore

$$(I-K_n)^{-1} = I+K(I-\phi_n K)^{-1}\phi_n \qquad (4.9)$$

This shows that whenever $(I-\phi_n K)^{-1}$ exists then $(I-K_n)^{-1}$ exists also and is expressed by (4.9). Now $(I-\phi_n K)^{-1} = G(G-\phi_n T)_{Y_n}^{-1}$ and in section 3.6 we gave the relationship (3.20) between $M_n^{-1} = (G-\phi_n T)_{Y_n}^{-1}$ and the inverse of the collocation matrix, A^{-1}. Thus we can employ the logical argument that if the collocation matrix is non singular, i.e. A^{-1} exists $\Rightarrow$ $(I-\phi_n K)^{-1}$ exists $\Rightarrow$ $(I-K_n)^{-1}$ exists. This approach will be used for the application of the 'a posteriori' theorems 9, 10 to bound $\|(I-K)^{-1}\|$ for use in (4.7). That is, the inverse matrix is known to exist and hence $(I-K_n)^{-1}$ exists also and so we have in conjunction with the results of section 4.3 the required conditions for the theory.

To apply theorems 9 and 10 practically we have to be able to bound $\|(I-K_n)^{-1}\|$. Equation (4.9) yields $\|(I-K_n)^{-1}\| \leq 1+\|K\|\|(I-\phi_n K)^{-1}\|\|\phi_n\|$, but $\|\phi_n\|$ is $O(\ln(n))$ for Chebyshev zeros for instance and this expression will increase as more collocation points are chosen making it difficult to achieve $\Delta^n < 1$ for Theorem 9

or $\Delta_d^n < 1$ for Theorem 10.

The next section shows how to find a more satisfactory bound.

4.6 A Bound on the Norm of the Inverse of the Extended Approximate Operator

In the previous section the relationship between the inverses of the extended and the usual approximate operators was seen. We here give a more practical means of bounding the norm of the inverse of our extended operator.

$\|(I-K_n)^{-1}\| = \sup_{\|y\|=1} \|(I-K_n)^{-1}y\| = \sup_{\|y\|=1} \|z_n(y)\|$ where

$z_n(y) = (I-K_n)y^{-1}$. Using the definition (4.3) we can say

$\|(I-K_n)^{-1}\| \leq \sup_{\|y\|=1} \{\|y\| + \|Ku_n(y)\|\}$ where $u_n = (I-\phi_n K)^{-1}\phi_n y$.

Thus

$$\|(I-K_n)^{-1}\| \leq 1 + \sup_{\|y\|=1} \left(\max_s \left\{\int_{-1}^{+1} [k(s,t)]^2 dt\right\}^{\frac{1}{2}} \cdot \right.$$

$$\left.\left\{\int_{-1}^{+1} [u_n(t)]^2 dt\right\}^{\frac{1}{2}}\right) \leq 1 + \sqrt{2}\, k_{max} \sup_{\|y\|=1} \left\{\int_{-1}^{+1} [u_n(t)]^2 dt\right\}^{\frac{1}{2}}$$

where $k_{max} = \max_{s,t} |k(s,t)|$.

Now $u_n(t)$ is a polynomial of degree n-1 and if our collocation points $\{t_i\}_{i=1}^n$ are the zeros of an n^{th} degree polynomial belonging to an orthogonal set with weight function $w(t) \geq 1$ then we can say

$$\int_{-1}^{+1} [u_n(t)]^2 dt \leq \int_{-1}^{+1} w(t)[u_n(t)]^2 dt$$

$$= \sum_{j=1}^{n} z_j^2 U_j^n \text{ using Gaussian quadrature.}$$

Here $z_j = u_n(t_j)$ and the U_j^n are the weights at the nodes ($j = 1 \ldots n$). This is a similar situation to that

discussed early in section 3.7 and used later to derive the result (3.39). Analagously to (3.28) (the derivation of which was shown for the example of second order problems), we would have (using the previous notation),

$$z_j = \underline{e}_j^t A_o A^{-1} \underline{\gamma} \qquad (j = 1 \ldots n)$$

with $\underline{\gamma} = [\gamma_1, \gamma_2, \ldots \gamma_n]^t$ and $\gamma_i = y(t_i)$ $(i = 1 \ldots n)$.

Now $\sup_{\|y\|=1} \{\sum_{j=1}^{n} z_j^2 U_j^n\}^{\frac{1}{2}} \leq \sup_{\|y\|=1} (\max_k |z_k|)(\sum_{j=1}^{n} U_j^n)^{\frac{1}{2}}$

and following the arguments of the early part of section 3.7,

$$\sup_{\|y\|=1} (\max_k |z_k|) \leq \|A_o A^{-1}\|$$

since $\|y\| = 1$ means $|\gamma_i| \leq 1$ $(i = 1 \ldots n)$. As before

$\sum_{j=1}^{n} U_j^n = \Omega$ where $\Omega = \int_{-1}^{+1} w(t)dt$.

Thus if for example we are using collocation with Chebyshev zeros we then have

$$\|(I-K_n)^{-1}\| \leq 1 + (2\Omega)^{\frac{1}{2}} k_{max} \|A_o A^{-1}\| \qquad (4.10)$$

(where $\Omega = \pi$ in this case).

Examples of $\|A_o A^{-1}\|$ were given in TABLE 6 in section 3.7 and it was illustrated that this quantity was virtually constant as more collocation points were chosen. Thus (4.10) provides a reasonable bound on $\|(I-K_n)^{-1}\|$ which can be used in Theorems 9 and 10 to obtain bounds on $\|(I-K)^{-1}\|$ for application to inequalities of the form (4.7). Chapter 5 contains some numerical examples of this process.

4.7 Comparison of Different Approaches

We have mentioned that Theorems 9 and 10 will be used in practice later and in this section a comparison of the applicability of these results and those of Theorem 7 is given.

Theorems 9 and 10 gave bounds on $\|(I-K)^{-1}\|$ in terms of $\|(I-K_n)^{-1}\|$ provided $\Delta^n = \|(I-K_n)^{-1}\|\,\|(K_n-K)K\| < 1$ and $\Delta_d^n = \|(I-K_n)^{-1}\|\,\|(K_n-K)K^d\| < 1$ respectively. The advantage of using the second result is now explained. (Note that $\Delta^n = \Delta_1^n$.)

Recall that $K_n \equiv K\phi_n$ by (4.4) so that

$$(K-K_n)K^d = K(I-\phi_n)K^d$$

$$\Rightarrow \|(K-K_n)K^d\| \leq \|K\|\,\|(I-\phi_n)K^d\|$$

$$= \|K\| \sup_{\|v\|=1} \{\|(I-\phi_n)K^d v\|\} \quad (v \in C[-1,1])$$

$$= \|K\| \sup_{\|v\|=1} \{\|(I-\phi_n)K^d v - \tilde{v} + \tilde{v}\|\}$$

for $\tilde{v}$ a polynomial of degree n-1. Thus

$$\|(K-K_n)K^d\| \leq \|K\|(1+\|\phi_n\|) \sup_{\|v\|=1} \|K^d v - \tilde{v}\|. \tag{4.11}$$

By Jackson's theorem (Cheney (1966, p.147)) there exists a polynomial $\tilde{v} \in P_{n-1}$ such that

$$\|K^d v - \tilde{v}\| \leq \left(\frac{\pi}{2}\right)^d \frac{\|(K^d v)^{(d)}\|}{n(n-1)\ldots(n-d+1)} \tag{4.12}$$

provided $K^d v \in C^{(d)}[-1,1]$.

We now prove that $K^d v$ does indeed have d continuous derivatives so that (4.12) can be applied in (4.11).

Lemma If $v \in C[-1,1]$ then $K^d v \in C^{(d)}[-1,1]$, provided that the coefficients in the differential equation are sufficiently differentiable.

Proof We use mathematical induction. Firstly $K^d v = K(K^{d-1}v) = Kw$ where $w = K^{d-1}v$. Now

$$(K^d v)(s) = \int_{-1}^{+1} k(s,t)w(t)dt$$

$$= p_{2m-1}(s)x_w^{(2m-1)}(s) + \ldots + p_1(s)x_w^{(1)}(s)$$

$$+ p_o(s)x_w(s)$$

if the differential equation involved in the definition of K was of type (4.1) and $x_w = G^{-1}w$. Thus $w \in C^{(j)}[-1,1]$ $\Rightarrow x_w \in C^{(2m+j)}[-1,1]$ and if we assume $p_j(s) \in C^{(d)}[-1,1]$ $(j=0,1\ldots 2m-1)$ then $w \in C^{(d-1)}[-1,1] \Rightarrow Kw \in C^{(d)}[-1,1]$. The case for d = 1 is certainly true and therefore by induction the lemma holds■

Thus provided that we have sufficient differentiability of the coefficients (4.12) can be utilised in (4.11) to give $\|(K-K_n)K^d\| \leq (\frac{\pi}{2})^d \frac{\|K\|(1+\|\phi_n\|)}{n(n-1)\ldots(n+d-1)} \sup_{\|v\|=1} \|(K^d v)^{(d)}\|$.

The problem now is to bound $(K^d v)^{(d)}$ in the manner $\|(K^d v)^{(d)}\| \leq k_d \|v\|$ for some constant k_d so that $\sup_{\|v\|=1} \|(K^d v)^{(d)}\| \leq k_d$. Note that we could have obtained similar results for $\|(K-K_n)K\|$ which would require evaluation of $\sup_{\|v\|=1} \|(Kv)^{(d)}\|$. However it is not possible in general for $d \geq 2$ to express $\|(Kv)^{(d)}\|$ in the form of a (constant) times $\|v\|$. This could be seen by considering an example of a second order equation and with d = 2 it would be clear that bounds on first derivatives of v were required. After sufficient algebraic manipulation however

we are able to achieve $\|(K^d v)^{(d)}\| \leq k_d \|v\|$. Roughly speaking this is possible because we perform d integrations of v and then d differentiations. In section 5.3 it is shown how to evaluate the constants k_d for d = 2.

Thus we see the advantage in using Theorem 10 because Δ_d^n is $O(\frac{\|\phi_n\|}{n(n-1)\ldots(n-d+1)})$ whereas Δ^n is $O(\frac{\|\phi_n\|}{n})$ and it is likely that the number of collocation points needed for applicability will be much less in the case of Theorem 10. Theorem 7 using the Kantorovich and Akilov approach requires $\delta_n = \|(G-\phi_n T)Y_n^{-1}\| \|(I-\phi_n)T\| < 1$ with $\|\cdot\|_X$ as the norm in the X space of the first part of Chapter 2 and of Chapter 3. Now

$$\|(I-\phi_n)T\| = \sup_{\|z\|_X=1} \|(I-\phi_n)Tz\|$$

$$\leq (1+\|\phi_n\|) \sup_{\|z\|_X=1} \|Tz-\tilde{v}\| \quad \text{for } \tilde{v} \epsilon P_{n-1}.$$

So that with $v = Gz$, $\|z\|_X = \|Gz\|_\infty = \|v\|$ we have $\|(I-\phi_n)T\| = \sup_{\|v\|=1} (1+\|\phi_n\|)\|Kv-\tilde{v}\|$ since $T \equiv KG$ and this is the same situation encountered as for the Anselone results of Theorem 9 and only yields δ_n as $O(\frac{\|\phi_n\|}{n})$.

To summarise then the work of this section, we expect that Theorem 10 will be applicable for much smaller numbers of collocation points than either Theorems 9 or 7 and so will be more suitable for practical bounds.

In the next Chapter in TABLES 13 and 14 the numerical results of some comparisons are given.

4.8 Generalisation of the Extension

To conclude this chapter we suggest a generalisation of the extended approximation discussed in section 4.2. The Nyström extension is implemented to improve upon the quadrature method for integral equations. However the extension we have introduced for collocation could be applied to any projection method to hopefully achieve results of a theoretical or practical nature.

We consider a general Banach space X instead of merely the space of continuous functions with the infinity norm. Let the given equation for $u \in X$ be

$$(I-K)u = y \qquad (4.13)$$

with I as the identity operator on X, $y \in X$ and $K \in [X]$. Let ϕ_n be any bounded linear projection of X into a subspace X_n of X, then we can regard $I-\phi_n K$ as a mapping from X_n to itself. When the operator $(I-\phi_n K)^{-1}$ exists $\in [X_n]$ we can make the following definitions:

$$u_n = (I-\phi_n K)^{-1}\phi_n y \qquad (\Rightarrow \; u_n \in X_n)$$

$$z_n = y + Ku_n \qquad (\Rightarrow \; z_n \in X)$$

and $K_n : X \to X$ is such that $K_n v = K\phi_n v$ for $v \in X$. Then the above three definitions imply $(I-K_n)z_n = y$ following the same argument as in section 4.2.

This extension could be applied to any projection method to define an 'extended projection' method and generalises the previous work. For example, if K is an integral operator of the form

$(Kv)(s) = \int_{-1}^{+1} k(s,t)v(t)dt$ (v integrable) for some kernel $k(s,t)$, we could consider the application of Galerkin's method. We choose a suitable function space X with the set P_{n-1} of polynomials of degree n-1 as the subspace X_n. In this case we could define ϕ_n by

$$(\phi_n v)(s) = \sum_{j=0}^{n-1} \left(\int_{-1}^{+1} L_j(t)v(t)dt \right) L_j(s) \quad (v \in X)$$

where $L_j(s)$ is the Legendre polynomial of degree j. If u_n is found from the Galerkin method in the usual way as an approximation to u satisfying (4.13) with K as above then with $z_n = y+Ku_n$ we have

$$z_n(s) - \int_{-1}^{+1} k(s,t) \left\{ \sum_{j=0}^{n-1} \int_{-1}^{+1} L_j(\tau) z_n(\tau) d\tau \right\} L_j(t)dt = y(s).$$

With suitable choices for the space X and its norm we might then hope to be able to apply the theory of sections 2.7 and 2.8 to deduce useful results concerning the Galerkin method.

This then illustrates one possible application of the generalisation.

CHAPTER 5

DETAILED CONSIDERATION OF ERROR BOUNDS AND NUMERICAL EXPERIMENTS

5.1 A Review of the Error Bounds and their Application

It was demonstrated in Chapter 3 and 4 that the theories of Chapter 2 could be applied to give error bounds for the approximate solution by collocation of linear ordinary differential boundary value problems. We are here concerned with the practical implementation of the 'a posteriori' bounds given in Theorems 7, 9 and 10, and this topic is considered in detail. We shall continue to employ the same notation as has been used throughout.

For the purposes of earlier chapters problems of the form (3.1) (or (4.1)) have been regarded as operator equations $Gx - Tx = y$ and it was seen (section 4.2) that the differential equation for x could be transformed to an integral equation for $u = Gx$ of the form $(I-K)u = y$ where $K \equiv TG^{-1}$.

With x_n as the polynomial approximation to x it was shown in section 3.5 that the most suitable means of bounding $\|x - x_n\|$ was to use the inequality

$$\|x - x_n\| \leq \|(G-T)^{-1}\| \, \|(G-T)x_n - y\| \qquad (5.1)$$

To utilise this in practice we have to find an 'a posteriori' bound on $\|(G-T)^{-1}\|$ and it was seen that Theorem 7 and its corollary provided a suitable result. Moreover $B_n = \|(G - \phi_n T)Y_n^{-1}\|$ occurring in the corollary was investigated in section 3.7 and was bounded by the inequality (3.44).

Other possible means of bounding $\|(G-T)^{-1}\|$ were discussed in section 4.2 where it was shown that $\|(G-T)^{-1}\|$ which is measured in the X-norm was equal to $\|(I-K)^{-1}\|$ in the infinity norm. Utilising the inequality (4.10) Theorems 9 and 10 can then be applied subject to certain conditions to produce 'a posteriori' bounds on $\|(I-K)^{-1}\|$ and hence on $\|(G-T)^{-1}\|$ for use in (5.1).

All the results furnished by Theorems 7, 9 and 10 require the number of collocation points n to be sufficiently large and some numerical examples of values of n required are given in section 5.5. Examples of computed error bounds along with estimates discussed in the next section are presented in section 5.6 and the results of further experiments are given in the Appendix.

Before the bounds are considered in detail in the following two sections care must be taken over the norms used for measuring the errors. Our bounds are derived from (5.1) in which we use the X-norm of Chapter 2. This has been necessary in order to be able to apply the theory of that chapter to bound $(G-T)^{-1}$ in some norm. An alternative would be to use $\|u-u_n\| \leq \|(I-K)^{-1}\| \|(I-K)u_n-y\|$ in the infinity norm where $u_n = Gx_n$. However this is of course equivalent to (5.1) since

$$\|u-u_n\|_\infty = \|G(x-x_n)\|_\infty = \|x-x_n\|_X.$$

Thus if we wish a bound on $\|x-x_n\|_\infty$ we are faced with the problem of bounding this from knowledge of $\|x-x_n\|_X$. This difficulty is treated by the following argument.

$$x(s) - x_n(s) = \int_{-1}^{+1} g(s,t)(x^{(2m)}(t) - x_n^{(2m)}(t))dt$$

where g(s,t) is the usual Green's function for the operator $\frac{d^{2m}}{dt^{2m}}$ subject to the homogeneous boundary conditions (3.1b). Thus

$$\|x - x_n\|_\infty \leq g^* \|x - x_n\|_X \tag{5.2}$$

where $g^* = \max_s \int_{-1}^{+1} |g(s,t)| dt$.

The inequality (5.2) will generally be a rather coarse bound on $\|x - x_n\|_\infty$ but seems unavoidable and its effect will be illustrated later in the results of section 5.6 where it will be seen that our 'a posteriori' error bounds are in better agreement with actual computed bounds when the X-norm is used compared with the infinity norm.

5.2 Detailed Formulation of the Error Bounds and their Estimates

The various means of deriving bounds on $\|(G-T)^{-1}\|$ were cited in the previous section and by combining the Theorems 7, 9 and 10 of Chapter 2 with the more practical results of Chapters 3 and 4 the detailed description of (5.3), (5.4) and (5.5) below can be given.

Chebyshev polynomials are used in the representation of the approximate solution and Chebyshev zeros are chosen as the collocation points. The bounds presented below are in fact independent of the basis used for the polynomial subspace but Chebyshev polynomials lead to more desirable condition number properties than simple powers and so are preferred.

With k(s,t) as the kernel when the differential problem is transformed to an integral one as in sections

3.2 and 4.2 and $K : C[-1,1] \to C[-1,1]$ such that for $v \in C[-1,1]$ $(Kv)(s) = \int_{-1}^{+1} k(s,t)v(t)dt$, we make the following definitions.

$k_{max} = \max_{-1 \le s,t \le 1} |k(s,t)|$

$k_o = \|K\|$. (Also $k_o = \|T\|$ since for $v \in C[-1,1]$ $Kv = Tz$ where $z = G^{-1}v$).

k_1 is such that for $v \in C[-1,1]$, $\|(Kv)'\| \le k_1\|v\|$.

k_2 is such that for $v \in C[-1,1]$, $\|(Kv)''\| \le k_2\|v\|$.

We must of course have sufficient differentiability of the coefficients in the differential equation for the latter two definitions.

In section 5.3 the problem of finding the constants k_{max}, k_o, k_1 and k_2 is considered but meantime three more quantities are defined using the above specifications as follows.

$\sigma = (2\pi)^{\frac{1}{2}}k_{max}$, $\epsilon_n = (1+\|\phi_n\|)\frac{\pi k_1}{2n}$ and

$\epsilon_n^{(2)} = (1+\|\phi_n\|)(\frac{\pi}{2})^2 \frac{k_2}{n(n-1)}$.

We have seen how there are three possible ways of bounding $\|(G-T)^{-1}\|$ for use in (5.1), namely by Theorems 7, 9 and 10. For Theorem 10 the choice of d = 2 is most suitable for practical purposes because otherwise with $d \ge 3$ the algebraic manipulation involved in bounding Δ_d^n can become lengthy.

Let the computable bounds on $\|(G-T)^{-1}\|$ furnished by Theorems 7, 9 and 10 for a particular value of n be denoted by $B_1(n)$, $B_2(n)$ and $B_3(n)$ respectively. Then we have

$$B_1(n) = \frac{1 + (k_o + \epsilon_n)B_n}{1 - \delta_n} \qquad (5.3)$$

for $\delta_n = [1+(k_o+\epsilon_n)B_n]\epsilon_n < 1$ where B_n is bounded by $\frac{1 + \sigma\|A_o A^{-1}\|}{1-\epsilon_n}$ provided $\epsilon_n < 1$,

$$B_2(n) = \frac{1 + k_o C_n}{1 - k_o C_n \epsilon_n} \tag{5.4}$$

for $\Delta^n = k_o C_n \epsilon_n < 1$ where

$$\|(I-K_n)^{-1}\| \leq C_n = 1 + \sigma\|A_o A^{-1}\|$$

and

$$B_3(n) = \frac{1 + k_o + k_o^2 C_n}{1 - k_o C_n \epsilon_n^{(2)}} \tag{5.5}$$

for $\Delta_2^n = k_o C_n \epsilon_n^{(2)} < 1$ as expressions suitable for numerical evaluation.

An example of the manner in which these bounds are applied in section 5.6 is now given. For instance, if n is large enough to give $\Delta_2^n < 1$ then

$$\|x - x_n\|_X \leq B_3(n)\|(G-T)x_n - y\|$$

provides an 'a posteriori' bound on $\|x - x_n\|_X$. Further we obtain

$$\|x - x_n\|_\infty \leq g^* B_3(n)\|(G-T)x_n - y\|$$

by (5.2). In such error bounds $\|(G-T)x_n - y\|$ is computed approximately by evaluation of the residual $(G-T)x_n - y$ at several points throughout the interval [1,-1] and taking the maximum of these values. This is not a rigorous bound but although it would be possible to

determine such a bound it is not thought worthwhile since this would involve a great deal of computing time and is not the main point of our analysis.

On examination it can be seen, as in section 4.7, that the bound from Theorem 10, namely $B_3(n)$, will generally be applicable for smaller numbers of collocation points than either of the others. That is, for any n Δ_2^n is likely to be less than Δ^n or δ_n and so the result (5.5) is able to be utilised for smaller values of n than either (5.3) or (5.4). Results comparing the values of n required are given in section 5.5 where it will be seen that they can be fairly large. This means that the number of collocation points used in practice to solve a problem might not be large enough to satisfy the conditions for the theory. In this case we would then have to increase n and invert a larger matrix (to compute $\|A_o A^{-1}\|$) in order to obtain a bound on $\|(G-T)^{-1}\|$. This bound could then be used for the original value of n, evaluating the residual appropriately. However having inverted the larger matrix we have essentially solved for a higher order approximation and could then obtain an error bound for this. This process would be rather unsatisfactory except perhaps for the situation where it was required to solve problems with the same differential operator but with a number of different right hand sides when it would be necessary only once to invert a large matrix, the residuals being recalculated each time.

To avoid such possible difficulties we now develop estimates of the bounds given in (5.3), (5.4) and (5.5) which do not require any stipulation about the size of n

and which are applicable for all numbers of collocation points.

It was seen in section 3.7 (for the second order examples chosen for TABLE 6) that $\|A_o A^{-1}\|$ was virtually constant with varying numbers of collocation points. This property can be utilised to derive estimates of the bounds given in (5.3), (5.4) and (5.5). For example, $B_2(n) = \frac{1+k_o C_n}{1-k_o C_n \epsilon_n} = \frac{1+k_o(1+\sigma\|A_o A^{-1}\|)}{1-k_o(1+\sigma\|A_o A^{-1}\|)\epsilon_n}$ and in this expression as n increases ϵ_n decreases whereas $\|A_o A^{-1}\|$ remains reasonably constant. Thus with n taken large the denominator will be close to unity and the numerator will be much the same as for smaller values $\hat{n}$ and we should expect that a good estimate of the bound $B_2(n)$ would be $\bar{B}_2(\hat{n}) = 1+k_o(1+\sigma\|A_o A^{-1}\|)\big|_{n=\hat{n}} = 1+k_o C_{\hat{n}}$. Here $\hat{n}$ is the smaller value of n actually being used in any calculation. That is, for the error bound the same value $\hat{n}$ of n would be employed for evaluation of both $\bar{B}_2(\hat{n})$ and the residual.

Similar estimates of bounds, $\bar{B}_1(\hat{n})$ and $\bar{B}_3(\hat{n})$ can be derived from $B_1(n)$ and $B_3(n)$ respectively. For $B_1(n)$ we have to implement the estimating scheme twice to achieve $B_n \simeq 1+\sigma\|A_o A^{-1}\| = C_n \simeq C_{\hat{n}}$ and then $B_1(n) \simeq \bar{B}_1(\hat{n}) = 1+k_o C_{\hat{n}}$ (since $\delta_n \simeq 0$ for n large). Thus for any number n of collocation points we should hope that

$$\begin{aligned} \bar{B}_1(n) &= 1+k_o C_n \\ \bar{B}_2(n) &= 1+k_o C_n \\ \text{and } \bar{B}_3(n) &= 1+k_o+k_o^2 C_n \end{aligned} \qquad (5.6)$$

would provide good estimates of the bounds on $\|(G-T)^{-1}\|$. (We notice that $B_1(n)$ and $B_2(n)$ both reduce to the same estimate).

The numerical experiments performed later indicate that the values $\bar{B}_1(n)$ and $\bar{B}_3(n)$ are indeed good estimates of their respective more rigorous bounds and are in fact likely to be closer to the actual norm of $(G-T)^{-1}$.

It is next shown how to calculate the remaining items needed for the various bounds.

5.3 Further Quantities Needed for the Numerical Evaluation of the Bounds

In this section it is demonstrated how to compute for a given differential equation the quantities k_{max}, k_o, k_1 and k_2 defined at the beginning of the previous section.

For simplicity we shall again consider second order equations of the form

$$(Gx-Tx)(s) = x''(s)+p(s)x'(s)+q(s)x(s) = y(s) \qquad (5.7)$$

with $x(-1) = x(+1) = 0$ and we shall require at times certain differentiability properties of the coefficients $p(s)$ and $q(s)$. For these problems $k(s,t) = -p(s)\frac{\partial g}{\partial s}(s,t) - q(s)g(s,t)$ with $g(s,t)$ the simple Green's function of section 1.4. The results of that section concerned with $g(s,t)$ will be used frequently and are restated here for convenience.

$$g(s,t) = \begin{cases} \frac{1}{2}(s+1)(t-1) & s \leq t \\ \frac{1}{2}(s-1)(t+1) & s > t \end{cases}$$

$$\int_{-1}^{+1} |g(s,t)|\,dt = \tfrac{1}{2}(1-s^2) \qquad \text{from (1.11)}$$

$$\int_{-1}^{+1} \left|\frac{\partial g}{\partial s}(s,t)\right| dt = \tfrac{1}{2}(1+s^2) \qquad \text{from (1.12)}$$

Further $\max_{-1\le s,t\le 1} |g(s,t)| = \tfrac{1}{2}$

and $\max_{-1\le s,t\le 1} \left|\frac{\partial g}{\partial s}(s,t)\right| = 1.$

To simplify the notation we define

$$p_j(s) = |p^{(j)}(s)| \quad \text{and} \quad q_j(s) = |q^{(j)}(s)|$$

provided $p,q \in C^{(j)}[-1,1]$ $(j = 0,1,2)$ (where $C^{(0)} \equiv C$).

We now show how to determine with a minimum of manipulations the constants k_{max}, k_o, k_1 and k_2 for a given differential operator.

$$k_{max} = \max_{s,t} |k(s,t)| = \max_{s,t} \left|p(s)\frac{\partial g}{\partial s}(s,t) + q(s)g(s,t)\right|$$

$$\Rightarrow k_{max} \le \max_s \{p_o(s) + \tfrac{1}{2}q_o(s)\} \qquad (5.8)$$

$$k_o = \|K\| = \sup_{\substack{v\in C[-1,1] \\ \|v\|=1}} \|Kv\|$$

$$= \sup_{\|v\|=1} \max_s \left|\int_{-1}^{+1} k(s,t)v(t)\,dt\right| \le \max_s \int_{-1}^{+1} |k(s,t)|\,dt$$

$$\Rightarrow k_o \le \max_s \{\tfrac{1}{2}p_o(s)(1+s^2) + \tfrac{1}{2}q_o(s)(1-s^2)\}. \qquad (5.9)$$

k_1 was such that for $v \in C[-1,1]$, $\|(Kv)'\| \le k_1\|v\|$.

Now $(Kv)' = (Tz)'$ where $z = G^{-1}v \Rightarrow z'' = v$. Thus

$$-(Kv)'(s) = p'(s)z'(s) + p(s)z''(s) + q'(s)z(s) + q(s)z'(s)$$

$$\Rightarrow -(Kv)'(s) = p(s)z'''(s) + \int_{-1}^{+1} \{[p'(s)+q(s)]\frac{\partial g}{\partial s}(s,t)$$

$$+ q'(s)g(s,t)\}z'''(t)dt \qquad (5.10)$$

and we can take

$$k_1 = \max_s [p_o(s) + \tfrac{1}{2}(1+s^2)|p'(s)+q(s)| + \tfrac{1}{2}(1-s^2)q_1(s)] \qquad (5.11)$$

We recall that k_2 was such that for $v \in C[-1,1]$ $\|(K^2v)''\| \leq k_2\|v\|$ and moreover $K^2v = K(Kv) = Kw''$ where $w(s) = (G^{-1}Kv)(s) = \int_{-1}^{+1} g(s,t)(Kv)(t)dt$. Thus

$$-(K^2v)'' = -(Tw)'' = p''w' + 2p'w'' + pw''' + q''w$$

$$+2q'w' + qw'' . \qquad (5.12)$$

Now $|w(s)| \leq \|Kv\| \int_{-1}^{+1} |g(s,t)|dt \leq \tfrac{1}{2}(1-s^2)\|Kv\|$ and

$$|w'(s)| \leq \|Kv\| \int_{-1}^{+1} |\frac{\partial g}{\partial s}(s,t)|dt \leq \tfrac{1}{2}(1+s^2)\|Kv\| .$$

Using $\|Kv\| \leq k_o\|v\|$, where k_o is given by (5.9), we can then bound every term in (5.12) except pw'''.
Now $w'''(s) = \frac{d}{ds}(Kv)(s) = \frac{d}{ds}(Tz)(s)$ and we can apply (5.10) and (5.11) to obtain

$$|w'''(s)| \leq k_1\|v\|$$

k_1 being given by (5.11).
Thus applying these bounds throughout (5.12) we have $|(K^2v)''(s)| \leq k_2\|v\|$ where

$$k_2 = \max_s \{k_o[\frac{(1-s^2)}{2}q_2(s)+\frac{(1+s^2)(p_2(s)+2q_1(s))}{2}$$

$$+ 2p_1(s)+q_o(s)] + k_1p_o(s)\}. \qquad (5.13)$$

Convenient means of determining the constants k_{max}, k_o, k_1 and k_2 are therefore provided by (5.8), (5.9), (5.11) and (5.13) respectively, and to illustrate the application of these we consider the sample operator $x''(s)+\alpha(1+s^2)x(s)$ with a parameter $\alpha > 0$, say. In this case

$$p(s) \equiv 0,$$

$$q(s) = \alpha(1+s^2) \Rightarrow q'(s)=2\alpha s \Rightarrow \quad q''(s) = 2\alpha,$$

and $q_o(s) = \alpha(1+s^2)$, $q_1(s) = 2\alpha|s|$, $q_2(s) = 2\alpha$.

Applying the formulae we have

$$k_{max} \leq \max_s \{\tfrac{\alpha}{2}(1+s^2)\} = \alpha,$$

$$k_o \leq \max_s \{\tfrac{\alpha}{2}(1+s^2)(1-s^2)\} = \max_s\{\tfrac{\alpha}{2}(1-s^4)\} = \tfrac{\alpha}{2},$$

$$k_1 = \max_s \{\tfrac{\alpha}{2}(1+s^2)(1+s^2)+\tfrac{1}{2}(1-s^2)2\alpha|s|\}$$

$$\leq \frac{\alpha}{2} \max_s \{1+2s^2+s^4+2-2s^2\} \leq 2\alpha .$$

$$\text{and } k_2 = \max_s \{k_o[\tfrac{1}{2}(1-s^2)2\alpha+\tfrac{1}{2}(1+s^2)4\alpha|s|+\alpha(1+s^2)]\}$$

$$= \max_s \{\frac{\alpha^2}{2}[1-s^2+2+2s^2+1+s^2]\}$$

$$= \frac{\alpha^2}{2} \max_s (2s^2+4) = 3\alpha^2.$$

In the calculations above we have been able at various points to make use of cancellation to obtain a lower bound than with straightforward minimisation of the individual terms. The next section contains a

specification of test problems to which the error bounding techniques discussed earlier are applied. For these or indeed any other problems when the formulae given here for finding the constants are used there will generally be possible a certain amount of cancellation. Further the bounds for the constants computed by a coarse implementation of these schemes may possibly be slightly refined if more sophisticated techniques are employed for determining the maximum value of functions over a given interval.

5.4 Specification of Test Problems

In this section test examples are described to which the strategies previously discussed for finding error bounds are later applied. The numerical results relating to this process are examined in sections 5.5 and 5.6 with further tables given in the Appendix.

We now present the six basic sample problems with a parameter α so that the equations are of the form $Gx-\alpha Tx = y$.

Problem 1

$$x''+\alpha(1+t^2)x = 1, \qquad x(-1) = x(+1) = 0$$

This example with $\alpha = 1$ is considered by Collatz (1960, p.143) and we take this particular problem and variations of it as ones to be discussed in detail in section 5.6 to illustrate the features of the different techniques for error bounds.

Problem 2

$$x''-\alpha x = \cosh(1), \qquad x(-1)=x(+1)=0$$

When $\alpha = 1$ this equation has the solution $\cosh(x)-\cosh(1)$ and is transformed from the example, $x''-4x = 4\cosh(1)$ over $[0,1]$ with $x(0) = x(1) = 0$, considered by Ciarlet, Schultz and Varga (1967, p.426).

Problem 3

$$x'' - \frac{2\alpha}{(t+5)^2}x = -\frac{1}{2(t+5)}, \qquad x(-1)=x(+1)=0$$

This problem with $\alpha = 1$ is a transformation to the interval $[-1,1]$ of the equation

$\frac{d^2x}{ds^2} - \frac{2}{s^2}x = -\frac{1}{s}$ subject to $x(2) = x(3) = 0$ with the exact solution $x(s) = \frac{1}{38}(19s - 5s^2 - \frac{36}{s})$ and is taken from Collatz (1960, p.178).

Problem 4

$$x''+\alpha\left[\frac{2x'}{(t+3)} - \frac{2x}{(t+3)^2}\right] = -\frac{1}{(t+3)}, \qquad x(-1)=x(+1)=0$$

The linear equation above is derived for $\alpha = 1$ after a certain amount of manipulation as a linearised version of the nonlinear problem

$\frac{d^2z}{ds^2} + \frac{1}{z}\left[1+\left(\frac{dz}{ds}\right)^2\right] = 0$ over $[0,1]$ with $z(0) = 1$, $z(1) = 2$

from Milne (1953, p.104). To achieve this we have used several adaptations. The nonlinear equation over $[0,1]$ is transformed to one over the interval $[-1,1]$ which is subsequently linearised in accordance with the process described in section 1.1. Into this equation, in the dependent variable $z(t)$ say, the function $z_o(t) = \frac{t+3}{2}$

is substituted as an initial approximation satisfying the boundary conditions and after the further substitution $x(t) = z(t) - \frac{1}{2}(t+3)$ we obtain after some manipulation that x satisfies Problem 4 with the above homogeneous boundary conditions.

Problem 5

$$x'' - \frac{3\alpha}{8} h^2(t)x = \frac{h^3(t)}{8} - \frac{3}{4}th^2(t), \qquad x(-1)=x(+1)=0,$$

where $h(t) = (t^2 + \frac{t}{2} + \frac{1}{2})$.

Again with $\alpha = 1$ this is a linearisation of a nonlinear problem which in this case is the following equation considered by Ciarlet, Schultz and Varga (1967, p.425).

$\frac{d^2z}{ds^2} = \frac{1}{2}(z+s+1)^3$ subject to $z(0) = z(1) = 0$. As for Problem 4 we change the variable from $s\epsilon[0,1]$ to $t\epsilon[-1,1]$ and linearise the nonlinear equation. The problem already has homogeneous boundary conditions and the substitution of $z_o(t) = t^2-1$ into the linearised equation yields our test example.

Problem 6

$$x'' + \frac{\alpha}{4}t^2(t+1)x = -\frac{1}{8}(t+1)t^2(t^2-2), \qquad x(-1)=x(+1)=0.$$

This problem is again the result of linearising a non-linear equation, namely

$\frac{d^2z}{ds^2} = \alpha sz^2$, $z(0) = z(1) = 1$, a form of which is considered by Collatz (1960, p.201). As before several transformations have been performed to derive the equation of Problem 6. In the linearised equation $z_o(t)$ is chosen

as t^2 and finally the substitution $x(t) = z(t) - 1$ gives $x(t)$ satisfying the desired equation.

This completes the description of the six basic examples chosen to demonstrate the results on application of the error bounding techniques.

For further illustrations we shall consider equations constructed using the same operators as Problems 1-6 but with, in turn, one of three additional fixed right hand sides. The extra right hand sides are as follows:

$$5 \sin(3t) \qquad \ldots\ldots \text{A},$$

$$\frac{1}{t^2 + 0.1} \qquad \ldots\ldots \text{B},$$

and
$$\begin{cases} t^3 + 2 + \sin(t) & -1 \leq t \leq 0 \\ (2-t)e^t & 0 < t \leq 1 \end{cases} \qquad \ldots\ldots \text{C}.$$

Each of these possesses a different property. The right hand side A is oscillatory in $[-1,1]$, B has a 'near singularity' at $t = 0$ and C has a discontinuous third derivative. We shall employ the notation that the problem formed by the operator J (J = 1,2 ... 6) and right hand side X (X = A,B or C) be denoted by Problem JX. (When X is absent this represents the original Problem J as before).

Before the numerical results are presented the values of the constants k_{max}, k_o, k_1 and k_2 for each of the six operators are given. It would be tedious

to include all the elementary manipulation involved and the numerical values are simply stated for brevity in TABLE 12. These have been calculated according to the strategies of section 5.3 as for Problem 1 which was used to demonstrate the process and as was mentioned there cancellation has been utilised where possible. Also these numbers are upper bounds and may possibly be refined by the application of more powerful techniques. However the estimates of bounds discussed in section 5.2 employ only the quantities k_{max} and k_o and in computing these simpler terms there is less scope for possible variation.

Values of the Constants k_{max}, k_o, k_1 and k_2 for the Test Problems

Operator	$\frac{k_{max}}{\alpha}$	$\frac{k_o}{\alpha}$	$\frac{k_1}{\alpha}$	$\frac{k_2}{\alpha^2}$
Problem 1	1	$\frac{1}{2}$	2	3
Problem 2	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{1}{2}$
Problem 3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{5}{32}$	0.018
Problem 4	$\frac{5}{4}$	$\frac{5}{4}$	$\frac{9}{4}$	6.26
Problem 5	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{27}{8}$	9.75
Problem 6	$\frac{1}{4}$	$\frac{4}{27}$	$\frac{9}{8}$	$\frac{16}{27}$

TABLE 12

5.5 Applicability of the Practical Bounds

We have seen that the practical bounds on the norm of the inverse of the operator G-T derivable from Theorems 7, 9 and 10 only hold for a sufficiently large

number of collocation points. Three bounds $B_1(n)$, $B_2(n)$ and $B_3(n)$ were formulated in section 5.2 and were given by (5.3), (5.4) and (5.5) respectively. In this section we are concerned with finding and comparing the actual values of n for which these formulae become applicable.

Firstly we introduce some notation. Each of the three bounds required some quantity say delta(n) less than unity in magnitude and for any of these let the value of n needed to give the corresponding delta $< \zeta$ be n_ζ. Thus for any of the 'a posteriori' bounds given by (5.3), (5.4) or (5.5) the appropriate value of n required for applicability is denoted by n_1, delta(n_1) being less than 1 in magnitude.

TABLE 13 below contains values of n_1 for the bounds $B_1(n)$ and $B_2(n)$ applied to the operators of Problems 1-6. To illustrate the dependence of the results upon the magnitude of the coefficients in the linear differential equations two values of the parameter α have been chosen. The values presented in TABLE 13 are not in fact exact but are not more than 5 greater than the precise n_1 and are intended more as a guide to illustrate the order of the sizes of n required.

Values of n_1 for $B_1(n)$ and $B_2(n)$ Applied to the Test Operators

	α \ Problem	1	2	3	4	5	6
$B_1(n)$	0.5	75	30	5	130	>100	30
	1	>100	90	5	>130	>100	65
$B_2(n)$	0.5	15	5	5	60	30	5
	1	>100	30	5	>100	>100	5

TABLE 13

It is seen from these examples that the values of n needed to apply the bound $B_2(n)$ are often significantly less than those for $B_1(n)$. From (5.3) and (5.4) $B_1(n)$ requires

$$\delta_n = [1+(k_o+\epsilon_n)B_n]\epsilon_n < 1 \qquad \text{(for } \epsilon_n < 1\text{)}$$

whereas for $B_2(n)$ we must have

$$\Delta^n = k_o C_n \epsilon_n < 1.$$

The better results for the bound from Theorem 9 are explained by the fact that for $\epsilon_n < 1$, $B_n \leq \frac{C_n}{1-\epsilon_n}$, and hence the bound on δ_n is

$$[1 + \frac{(k_o+\epsilon_n)C_n}{1-\epsilon_n}]\epsilon_n \geq k_o C_n \epsilon_n = \Delta^n.$$

It is clear from TABLE 13 that even for the better result large numbers of collocation points may still be involved. The applicability of the bound $B_3(n)$ from Theorem 10 is now considered and as was predicted in sections 4.7 and 5.2 this leads to improvements.

Now before these results are presented we consider the situation where $B_3(n)$ is to be used in practice in

the inequality (5.1). In this case it would not be satisfactory merely to choose the computed bound resulting from n equal to n_1, that is to take $\|(G-T)^{-1}\| \leq B_3(n_1)$ because clearly the corresponding value of Δ_2^n is close to unity and consequently $B_3(n_1)$ will be large. Instead we shall seek the numbers of collocation points required to give $\Delta_2^n < 0.2$, namely $n_{0.2}$, and with this value of n a much more reasonable bound would be expected.

Values of n_1 and $n_{0.2}$ along with $B_3(n_1)$ and $B_3(n_{0.2})$ are given in TABLE 14 below. To explain the format of the table a typical block under the heading of a problem operator with a particular value of the parameter α contains 4 entries which are the appropriate results for the 4 quantities mentioned in the previous sentence and are presented in the layout

n_1	$B_3(n_1)$
$n_{0.2}$	$B_3(n_{0.2})$

Applicability of the Bound $B_3(n)$

α \ Problem	1	2	3	4	5	6
0.5	5 4.43	2 3.46	2 1.03	11 44.6	8 23.5	2 1.31
	10 1.69	4 1.52	2 1.03	25 3.65	18 2.04	2 1.31
1.0	18 28.5	5 9.32	2 1.08	48 645	28 301	3 3.06
	39 3.70	10 2.62	2 1.08	>100 (19.3)	65 4.38	5 1.47
2.0	>100	17 487	2 1.34	>100	>100	8 9.29
		41 8.19	2 1.34			17 1.85

TABLE 14

(Note that for Problem 4 with $\alpha = 1$, $n_{0.2}$ is greater than 100 but however when $n = 100$ the value of $B_3(100)$ is 19.3).

On comparison of the results of TABLE 14 with those of TABLE 13 the clear improvement can be seen. The values $B_3(n_{0.2})$ are used where possible in the next section to provide error bounds. Nevertheless some of the numbers of collocation points required are still large and it is for this reason that the estimates discussed in section 5.2 were introduced.

Finally the values of n_1 from TABLE 14 for Problem 2 can be compared to the 'a priori' results of TABLE 1, for the same sample operator, which yielded numbers of roughly similar magnitude.

5.6 Error Bounds and Estimates of Bounds

In this section we present and discuss the numerical results when the error bounds and estimates we have derived are applied in practice. These are all based on the inequality (5.1) and utilise different means of bounding $\|(G-T)^{-1}\|$.

As was mentioned briefly in section 5.2 although it would be possible with a fair amount of work to find a strict bound on the residual it would be a deviation from the main aim of our analysis and the infinity norm of $(G-T)x_n - y$ is estimated accurately by evaluation of this residual at several points and by taking the maximum of these. 20 points equally spaced throughout the interval $[-1,1]$ are chosen for this purpose and for any value of n the resulting computed maximum is denoted by RES(n).

It was seen in the previous section that large numbers of collocation points were needed to apply the bounds $B_1(n)$ and $B_2(n)$ and for this reason they are not considered for practical purposes. The rigorous bound $B_3(n_{0.2})$ on $\|(G-T)^{-1}\|$ is utilised however where possible along with the estimates $\bar{B}_1(n)$ and $\bar{B}_3(n)$. The following notation is used.

$$E_3(n) = B_3(n_{0.2}) \times RES(n)$$
$$\bar{E}_1(n) = \bar{B}_1(n) \times RES(n)$$
$$\bar{E}_3(n) = \bar{B}_3(n) \times RES(n)$$

$E(n)$ is to represent the X-norm of the actual error, namely $\|x_n - x\|_X$ or $\|x_n'' - x''\|_\infty$ and is computed in the same way as the residual by evaluation at 20 equally spaced points in the interval $[-1,1]$, the exact solution x having been found by solving the problem with a large number of collocation points.

The above error bounds are all measured in the X-norm. However if we wish to predict results in the infinity norm we have to employ the rather coarse strategy, discussed in section 5.1, which produced the inequality (5.2). The quantity

$g^* = \max_s \int_{-1}^{+1} |g(s,t)|\,dt$ in that result is found from section 1.4 as $\max_s \frac{1}{2}(1-s^2) = \frac{1}{2}$. Thus the error bounds we obtain in the infinity norm are merely half those in the X-norm.

We shall employ the notation

$$F_3(n) = \frac{E_3(n)}{2},\quad \bar{F}_1(n) = \frac{\bar{E}_1(n)}{2} \text{ and } \bar{F}_3(n) = \frac{\bar{E}_3(n)}{2}.$$

Thus $F_3(n)$, $\bar{F}_1(n)$ and $\bar{F}_3(n)$ are computed 'a posteriori' results for bounding the error in the infinity norm. We shall represent the actual computed error in the functions x_n by $F(n)$.

In this section we give detailed results for Problem 1 and discuss certain general points, results for the other test examples being contained in the Appendix.

Finally before presentation of the tables two points concerning the notation should be clarified. From TABLE 14 it can be seen that for certain problems, with the parameter α equal to 2, n_1 can be greater than 100 (and consequently is not thought worthy of calculation). In this case we take the bound $B_3(n)$ to be inapplicable and hence we are unable to form the error bound $E_3(n)$. Should this situation arise the corresponding entry consists of the symbol ****.

The second point, perhaps an obvious one, is that the capital letter N now represents the number of collocation points and that the integer subscripts are now replaced by normal size numeric characters. Thus for example, $\bar{E}_1(n)$ is replaced by $\bar{E}1(N)$.

The sample tables illustrating the results on application of the different techniques for bounding the error are now presented. TABLES 15-17 demonstrate the results for Problem 1 with α = 0.5, 1.0 and 2.0 whereas TABLES 18-20 are concerned with Problems 1A, 1B and 1C respectively when α has been chosen as unity.

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 1 ALPHA= 0.5

N	6	8	10	12
$\bar{B}1(N)$	1.66	1.67	1.67	1.67
$\bar{B}3(N)$	1.42	1.42	1.42	1.42
RES(N)	5.06'-04	4.06'-05	3.33'-07	1.38'-08
E3(N)	8.55'-04	6.85'-05	5.63'-07	2.34'-08
$\bar{E}1(N)$	8.42'-04	6.75'-05	5.55'-07	2.31'-08
$\bar{E}3(N)$	7.16'-04	5.74'-05	4.72'-07	1.96'-08
E(N)	5.07'-04	4.09'-05	3.34'-07	1.39'-08
F3(N)	4.27'-04	3.43'-05	2.82'-07	1.17'-08
$\bar{F}1(N)$	4.21'-04	3.38'-05	2.77'-07	1.15'-08
$\bar{F}3(N)$	3.58'-04	2.87'-05	2.36'-07	9.81'-09
F(N)	1.42'-05	7.25'-07	3.02'-09	1.02'-10

TABLE 15

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 1 ALPHA= 1.0

N	6	8	10	12
$\bar{B}1(N)$	3.91	3.92	3.92	3.92
$\bar{B}3(N)$	2.96	2.96	2.96	2.96
RES(N)	2.87'-03	2.21'-04	3.79'-06	1.48'-07
E3(N)	1.06'-02	8.17'-04	1.40'-05	5.47'-07
$\bar{E}1(N)$	1.12'-02	8.65'-04	1.49'-05	5.80'-07
$\bar{E}3(N)$	8.49'-03	6.54'-04	1.12'-05	4.38'-07
F(N)	2.89'-03	2.25'-04	3.82'-06	1.49'-07
F3(N)	5.31'-03	4.09'-04	7.02'-06	2.74'-07
$\bar{F}1(N)$	5.62'-03	4.33'-04	7.43'-06	2.90'-07
$\bar{F}3(N)$	4.25'-03	3.27'-04	5.61'-06	2.19'-07
F(N)	9.13'-05	4.08'-06	3.55'-08	1.11'-09

TABLE 16

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 1 ALPHA= 2.0

N	6	8	10	12
$\bar{B}1(N)$	69.0	69.4	69.5	69.6
$\bar{B}3(N)$	70.0	70.4	70.5	70.6
RES(N)	7.56'-02	5.27'-03	2.01'-04	6.72'-06
E3(N)	****	****	****	****
$\bar{E}1(N)$	5.22'+00	3.66'-01	1.40'-02	4.67'-04
$\bar{E}3(N)$	5.29'+00	3.71'-01	1.42'-02	4.74'-04
E(N)	8.89'-02	5.54'-03	2.04'-04	6.84'-06
F3(N)	****	****	****	****
$\bar{F}1(N)$	2.61'+00	1.83'-01	6.98'-03	2.34'-04
$\bar{F}3(N)$	2.65'+00	1.86'-01	7.08'-03	2.37'-04
F(N)	7.25'-03	1.34'-04	2.75'-06	6.27'-08

TABLE 17

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 1A ALPHA= 1.0

N	6	8	10	12
$\bar{B}1(N)$	3.91	3.92	3.92	3.92
$\bar{B}3(N)$	2.96	2.96	2.96	2.96
RES(N)	2.58'-02	4.39'-04	7.28'-06	1.62'-07
E3(N)	9.54'-02	1.63'-03	2.69'-05	6.00'-07
$\bar{E}1(N)$	1.01'-01	1.72'-03	2.85'-05	6.35'-07
$\bar{E}3(N)$	7.62'-02	1.30'-03	2.16'-05	4.80'-07
E(N)	2.60'-02	4.40'-04	7.33'-06	1.63'-07
F3(N)	4.77'-02	8.13'-04	1.35'-05	3.00'-07
$\bar{F}1(N)$	5.05'-02	8.61'-04	1.43'-05	3.18'-07
$\bar{F}3(N)$	3.81'-02	6.50'-04	1.08'-05	2.40'-07
F(N)	6.67'-04	4.20'-06	4.01'-08	7.11'-10

TABLE 18

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 1B ALPHA= 1.0

N	15	20	25	30
$\bar{B}1(N)$	3.92	3.92	3.92	3.92
$\bar{B}3(N)$	2.96	2.96	2.96	2.96
RES(N)	9.29'-02	3.96'-02	4.06'-03	1.76'-03
E3(N)	3.44'-01	1.47'-01	1.50'-02	6.53'-03
$\bar{E}1(N)$	3.64'-01	1.55'-01	1.59'-02	6.92'-03
$\bar{E}3(N)$	2.75'-01	1.17'-01	1.20'-02	5.22'-03
E(N)	9.20'-02	3.98'-02	4.07'-03	1.76'-03
F3(N)	1.72'-01	7.33'-02	7.51'-03	3.26'-03
$\bar{F}1(N)$	1.82'-01	7.77'-02	7.96'-03	3.46'-03
$\bar{F}3(N)$	1.37'-01	5.87'-02	6.01'-03	2.61'-03
F(N)	1.97'-03	1.90'-04	8.53'-06	2.29'-06

TABLE 19

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 1C ALPHA= 1.0

N	5	8	12	15
$\bar{B}1(N)$	3.92	3.92	3.92	3.92
$\bar{B}3(N)$	2.96	2.96	2.96	2.96
RES(N)	3.29'-02	2.33'-03	7.66'-04	1.64'-04
E3(N)	1.22'-01	8.62'-03	2.83'-03	6.08'-04
$\bar{E}1(N)$	1.29'-01	9.13'-03	3.00'-03	6.44'-04
$\bar{E}3(N)$	9.75'-02	6.89'-03	2.27'-03	4.86'-04
E(N)	3.26'-02	2.51'-03	7.81'-04	1.74'-04
F3(N)	6.09'-02	4.31'-03	1.42'-03	3.04'-04
$\bar{F}1(N)$	6.45'-02	4.56'-03	1.50'-03	3.22'-04
$\bar{F}3(N)$	4.87'-02	3.45'-03	1.13'-03	2.43'-04
F(N)	2.80'-03	1.95'-04	3.45'-05	1.44'-05

TABLE 20

Several points concerning these results are now discussed.

As was suggested in section 5.1 it can be seen that error predictions in the X-norm are closer than those measured in the infinity norm. For example, the ratios $\bar{E}_3(n):E(n)$ given in TABLE 15 are less than 2:1 and from TABLE 16 are roughly 3:1. The corresponding values of $\bar{F}_3(n):F(n)$ however are greater and also increase with the number of collocation points.

The results for Problem 1 with $\alpha = 2$ are not so consistent as with the other choices of the parameter and this situation is a special case which we shall now consider. Examining TABLE 17 where $\alpha = 2$ we notice that there are quite large discrepancies between the predicted and the actual errors. The reason for this behaviour would appear to be that the problem $x'' + \lambda(1+t^2)x = 0$ with $x(-1) = x(+1) = 0$ has an eigenvalue λ with λ close to 2. The quantities $\bar{B}_1(n)$ and $\bar{B}_3(n)$ can be seen to be large, roughly 70 in magnitude and this is not surprising as these involve the norm of the matrix A_0A^{-1} which increases when α is near 2. The variation of $\|A_0A^{-1}\|$ with α is shown in TABLE 21 below where it can also be seen that the approximate constancy with n of these matrix norms still holds, two different numbers of collocation points having been chosen to illustrate this.

Variation of $\|A_0 A^{-1}\|$ in the Neighbourhood of an Eigenvalue

n \ α	1	2	2.1	2.15	2.19	2.2	2.5	3	4
10	1.931	13.47	31.09	88.62	187.2	105.4	7.59	3.92	3.16
20	1.932	13.49	31.15	88.82	187.6	105.6	7.88	4.36	3.32

TABLE 21

Thus these results explain why $\bar{B}_1(n)$ and $\bar{B}_3(n)$ are large and since both approximate operators (n = 10 and n = 20) have an eigenvalue λ with λ close to 2.19 this suggests that the original differential problem also has this eigenvalue. However this has not explained why there is such a large discrepancy between the estimated and the actual bounds, the latter being little affected by this eigenvalue.

Let us now consider the relative merits of the different results $E_3(n)$, $\bar{E}_1(n)$ and $\bar{E}_3(n)$. All of these employ the reliable estimate RES(n) of the norm of the residual. $E_3(n)$ utilises the rigorous bound $B_3(n_{0.2})$ on $\|(G-T)^{-1}\|$ which is slightly larger than the estimate $\bar{B}_3(n)$ and the errors from the approximate result can be seen to be closer to the actual computed errors. These tables demonstrate that the norm of the residual can be in fact close to the X-norm of the actual error, that is, the error in the second derivative and so RES(n) could be taken as an approximation to $\|x''-x_n''\|$. This process however is rather unsatisfactory and yields an unjustified estimate of the X-norm of the error as distinct from the more rigorous estimates of bounds on the corresponding error. Clearly for a

practical estimate we should choose the smaller of $\bar{B}_1(n)$ and $\bar{B}_3(n)$. For our examples $\bar{B}_3(n)$ usually yields slightly better results but the deciding factor is essentially the magnitude of the constant k_o defined in section 5.2. Recall that

$$\bar{B}_1(n) = 1+k_oC_n \quad \text{whereas } \bar{B}_3(n) = 1+k_o+k_o^2C_n$$

and certainly for $k_o > 1$ we should have $\bar{B}_1(n) < \bar{B}_3(n)$.

The errors in the infinity norm are of course related similarly to those discussed above.

A further interesting observation concerns the application of the schemes to problems with the right hand sides B or C. It can be seen from TABLE 19 and from TABLES 24, 28, 32, 36, 40 in the Appendix or from TABLE 20 with TABLES 25, 29, 33, 37, 41 in the Appendix that for large numbers of collocation points the residuals are very close irrespective of the differential operator. The actual errors in the second derivatives, E(n), are also fairly close for these values of n but the values of F(n) do not agree to the same extent. This behaviour appears due to the fact that when n is taken large the right hand sides B or C tend to dominate the collocation method which is essentially interpolating the particular right hand side independently of the operator T in the differential problem. That is, in applying collocation we are interpolating the right hand sides y by functions $(G-T)x_n$, with x_n polynomials and for large values of n with the right hand sides B or C the terms Gx_n which are polynomials seem to dominate the process so that the residual is approximately Gx_n-y. The right hand

terms in the basic examples and the function A are smoother and do not influence matters to the same degree. It should be pointed out however that the results of our bounding techniques do vary for all these problems.

Finally, a brief comparison of these 'a posteriori' bounds with 'a priori' values can be furnished by relating TABLE 22 in the Appendix to TABLES 2 and 3 of Chapter 3. The big improvement on using the 'a posteriori' approach is clearly seen.

This completes our discussion of the error bounds applied in practice and as has been mentioned previously the results of the additional numerical experiments for Problems 2-6 are given in the Appendix.

CHAPTER 6

EXTENSIONS AND CONCLUSIONS

6.1 Introduction

In this chapter several possible areas of extension for the application of the theory are discussed. In this thesis we have been primarily concerned with the approximate solution of collocation of two point linear boundary value problems and have considered the example (3.1) (or (4.1)) with G defined as the operator differentiating 2m times. Furthermore we have mainly been working with the infinity norm. The theory of Kantorovich and Akilov outlined in Chapter 2 can in fact be applied to a more general class of problems by suitable definitions of G and choices of norms in the spaces X and Y.

For example the numerical solution by collocation of a system of linear initial value problems adjusted to have homogeneous initial conditions can be placed in the setting of the theory. Let the r equations be

$$\frac{d\underline{x}}{ds} - \lambda M(s)\underline{x}(s) = \underline{y}(s) \qquad (6.1)$$

with $\underline{x}(o) = \underline{\theta}$. Here $\underline{x} = (z_1, z_2, \ldots z_r)^t$ say, M(s) is an r x r matrix with elements which are continuous functions of s and $\underline{y} = [y_1(s), y_2(s), \ldots y_r(s)]^t$ with $y_j(s)$ continuous (j = 1 ... r). $\underline{\theta}$ represents the zero vector and λ is a scalar.

Take X as the space of r x 1 vectors whose components are continuously differentiable and are zero at s = 0. Let Y be the space of vectors whose elements are continuous

then with $G : X \to Y$ as the operator differentiating componentwise we can write (6.1) as

$$G\underline{x} - \lambda T\underline{x} = \underline{Y} \qquad (6.2)$$

where $(T\underline{x})(s) = M(s)\underline{x}(s)$. Equation (6.2) is now in the form for the theory. An approximate solution can be sought in the form

$$\underline{x}_n = C\underline{\psi}$$

where $\underline{\psi} = [s\psi_1(s), s\psi_2(s), \ldots s\psi_n(s)]^t$ with $\{\psi_j\}_{j=1}^n$ n independent polynomials forming a basis for the polynomial subspace of degree n-1 (for some n) and where C is the n x n matrix of unknown coefficients. If $\underline{Y} = (y_1, y_2, \ldots y_r)^t \in Y$ let the norm in Y be such that

$$\|\underline{y}\|_Y = \max_{1 \le i \le r} \|y_i(s)\|_\infty$$

and for $\underline{x} = (z_1, z_2, \ldots z_r)^t \in X$ let

$$\|x\| = \max_{1 \le i \le r} \|z_i'(s)\|_\infty \quad (= \|G\underline{x}\|_Y).$$

Clearly we could choose suitable collocation points and define appropriate subspaces X_n and Y_n and projection $\phi_n : Y \to Y_n$ in a related manner to section 2.2. The appropriate results of Chapter 2 could then be applied from a practical as well as a theoretical point of view.

However collocation as a means of numerically solving initial value problems is unlikely to compare favourably

with the well developed and well known specialised methods and for this reason is not considered as a suitable topic for further investigation but nevertheless the above description illustrates the wide scope of the theory.

We have considered first order equations above but there are however extensions which can be applied in principle to second order boundary value problems of the type previously considered. These could possibly be furnished by choosing G as an operator different from $\frac{d^2}{ds^2}$. However for any choice of G we must be sure that the operator G^{-1} exists. For example we could in principle choose G such that for $x \in X$

$$(Gx)(s) = \frac{d^2x}{ds^2} - \mu x(s)$$

with μ a constant so that an equation

$$x''(s) + p(s)x'(s) + q(s)x(s) = y(s)$$

could be regarded in the form $Gx - Tx = y$ where $Tx = -px' - (q + \mu)x$. Having chosen the space X and the subspace X_n we would have to ensure that Y and Y_n were such that $G : X \to Y$ had a linear inverse, that is that μ was not an eigenvalue, and that G was a bijection between X_n and Y_n since these conditions are necessary for the application of the theory. Moreover in any application of the theory, for some choice of G, T and the norms, we would need to be able to approximate Tx for $x \in X$ and y by elements of Y_n. Thus these requirements could clearly cause problems. Since the norms in the

spaces X and Y will be co-ordinated by $\|x\|_X = \|Gx\|_Y$ the above approximation might be achieved by relating, for example, $\|Tx - \tilde{y}\|_Y$ for $\tilde{y} \epsilon Y_n$ to the norm $\|G^{-1}Tx - G^{-1}\tilde{y}\|$ in X and approximating in the X space.

An example of a more general definition of G is considered by Kantorovich and Akilov (1964, pp.590-595) where they discuss the equation

$$\frac{d}{dt}\left[p\frac{dx}{dt}\right] - \lambda\left\{\frac{d}{dt}[qx] + rx\right\} = y \qquad (6.3)$$

over [0,1] with $x(0) = x(1) = 0$ and define G as the operator $\frac{d}{dt}(p\,\frac{d}{dt})$. Such a choice of G might be useful in dealing with equations which contain a singularity but as was mentioned previously we must ensure that G has an inverse. We shall not give a detailed investigation of this example but shall present in the next section the main points of the argument.

In section 6.3 we consider work on the use of splines for two point boundary value problems and examine the possibility of employing 'a posteriori' error analysis.

Aspects of the application of the theory to nonlinear ordinary and linear partial differential equations are discussed briefly in sections 6.4 and 6.5 respectively and lastly a review of the work of this thesis with appropriate conclusions is given in the final section.

6.2 An Illustration of a More General Application

We here examine the important steps in the application of the theory given in sections 2.2 - 2.4 to an equation of the form (6.3) which is considered by Kantorovich and Akilov.

It is assumed that p, q are continuously differentiable with $p(t) > 0$ and that $r, y \in C[0,1]$. Galerkin's method is applied to determine an approximate solution in the form

$$x_n(t) = \sum_{j=1}^{n} a_j w_j(t) \tag{6.4}$$

where $w_j(t) \in C^{(1)}[0,1]$ and $w_j(0) = w_j(1) = 0 \quad (j = 1 \ldots n)$. With $w_0'(t) = \frac{1}{p(t)} \{ \int_0^1 \frac{ds}{p(s)} \}^{\frac{1}{2}}$ it is assumed that the system $\{w_k'\}$ $(k = 0,1, \ldots)$ is complete and orthonormal with respect to the weight $p(t)$.

$$\text{i.e. } \int_0^1 p(t) w_j'(t) w_k'(t) dt = \delta_{jk} \quad (j,k = 0,1, \ldots) \tag{6.5}$$

Now if X is the space of functions $z(t)$ say in $C^{(2)}[0,1]$ with $z(0) = z(1) = 0$ and Y is $C[0,1]$ then if $G \equiv \frac{d}{dt}(p \frac{d}{dt})$ equation (6.3) may be written as

$$Gx - \lambda Tx = y, \quad \text{with} \quad Tx = \frac{d}{dt}[qx] + rx.$$

The reason for requiring the condition (6.5) becomes clearer when an inner product $(\cdot, \cdot)$ is introduced on X such that for $z_1, z_2 \in X$,

$$(z_1, z_2) = \int_0^1 p(t) z_1'(t) z_2'(t) dt,$$

the norm being defined by $\|z\| = (z,z)^{\frac{1}{2}}$. Corresponding inner products and norms are introduced in Y relating them

to those in X by G^{-1}. That is,

$$(y_1,y_2)_Y = (G^{-1}y_1, G^{-1}y_2)_X$$

and $\|y\|_Y = \|G^{-1}y\|_X$.

(Note that for $y\epsilon Y$, $G^{-1}y$ is the element $z\epsilon X$ such that $z(o) = z(1) = 0$ and $\frac{d}{dt}(p\frac{dz}{dt}) = y$). The subspace X_n of X is chosen as the set of elements of the form (6.4) with Y_n as the set of functions of the form $\sum_{j=1}^{n} a_j Gw_j$.

To complete the specification of the spaces and mappings for the theory Kantorovich and Akilov take ϕ_n as the orthogonal projection of Y onto Y_n which means $\|\phi_n\| = 1$.

With the above definitions, for $y\epsilon Y$

$$\phi_n y = \sum_{k=1}^{n} (y,Gw_k)_Y \, Gw_k$$

$$= \sum_{k=1}^{n} (G^{-1}y,w_k)_X \, Gw_k$$

$$= \sum_{k=1}^{n} \{\int_o^1 p(t)(G^{-1}y)'w_k' \, dt\}Gw_k$$

and employing integration by parts it is shown that

$$\int_o^1 p(G^{-1}y)'w_k' \, dt = -\int_o^1 yw_k \, dt. \qquad (6.6)$$

Thus since Galerkin's method requires

$$\int_o^1 (Gx_n - \lambda Tx_n - y)w_k(t)dt = 0 \quad (k = 1 \ldots n)$$

this means from (6.6) that the method is equivalent to $\phi_n[Gx_n - \lambda Tx_n - y] = 0$, or

$$Gx_n - \lambda\phi_n Tx_n = \phi_n y. \tag{6.7}$$

Equation (6.7) is now in the form we have frequently encountered (apart from the constant λ which can be included in T) and we now have the framework for the theory and are in a position to examine the conditions required for its application.

To utilise the theorems it is required to find μ_1 and μ_2 for the conditions I and II of section 2.4. It is not thought necessary to present in detail the work of Kantorovich and Akilov on this topic and we simply give an outline of their analysis.

For condition I we need to find a μ_1 such that for all $z \epsilon X$ there exists a $\tilde{y} \epsilon Y_n$ and $\|Tz - \tilde{y}\| \leq \mu_1 \|z\|$. The strategy mentioned in the previous section is employed when $G^{-1}Tz$ is approximated by an element $\tilde{x}$ of X_n, since $\|Tz - \tilde{y}\|_Y = \|G^{-1}Tz - \tilde{x}\|_X$ where $\tilde{x} = G^{-1}\tilde{y}$. With $v = G^{-1}Tz$ it is shown that there is a kernel $K(s,t)$ such that

$$v'(s) = \int_0^1 K(s,t)z'(t)\,dt.$$

The approximation $\tilde{x}$ to v is found from

$$\tilde{x}'(s) = \int_0^1 K_n(s,t)z'(t)\,dt$$

where $K_n(s,t)$ is a partial sum of the Fourier expansion of

$K(s,t)$ in $\{w_j'(s)\}$ $(j = 1,2, \ldots)$. That is,

$$K_n(s,t) = \sum_{j=1}^{n} c_j(t) w_j'(s)$$

where $c_j(t) = \int_0^1 p(s)K(s,t)w_j'(s)ds$ and furthermore

$$\lim_{n\to\infty} \int_0^1 p(s)[K(s,t) - K_n(s,t)]^2 ds = 0 \qquad (0 \leq t \leq 1).$$

After further analysis it is demonstrated that a suitable μ_1 can indeed by obtained and that $\mu_1 \to 0$ as $n \to \infty$.

To find the μ_2 for condition II a similar approach is followed, approximating $G^{-1}y$ by an element of X_n.

Kantorovich and Akilov consider the example where trignometric functions are used as the $\{w_k\}$. In particular for

$$w_k(t) = \sin(k\pi t) \qquad (k = 1,2, \ldots)$$

it is found that μ_1 and μ_2 are both $O(n^{-\frac{1}{2}})$ so that with x as the true solution to (6.3) and x_n as the solution from the Galerkin method, Theorem 2 yields

$$\|x - x_n\|_X = O(n^{-\frac{1}{2}})$$

as a measure of convergence in this special norm. (Note that $\|\phi_n\| = 1$).

This then is a brief account of a possible generalisation investigated by Kantorovich and Akilov which is essentially an 'a priori' examination.

However we could attempt an 'a posteriori' error analysis of this problem and an approach similar to that of section 3.6 would probably be the most suitable means of bounding the norm of the inverse of the approximate operator. This would need careful definitions of the appropriate mappings and norm for the space R^n of vectors and there is clearly scope for further investigation into this topic.

6.3. The Use of Splines

There has been recently some very interesting work on two point linear and nonlinear boundary value problems concerned with the use of splines in the representation of the approximation.

In particular for an m^{th} order problem over $[a,b]$ say and given a partition $\pi_n : a = s_o < s_1 < \ldots < s_n = b$ of $[a,b]$, Russell and Shampine (1972) seek approximate spline solutions which for integer d are polynomials of degree $m + d$ in each subinterval of π_n and have m continuous derivatives throughout the whole interval. These splines are further required to satisfy the m given boundary conditions. To obtain the appropriate number of equations for determining the coefficients in the representation suitable collocation points are needed. These are furnished by subdividing each subinterval $[s_i, s_{i+1}]$, $(0 \leq i \leq n-1)$ by a further (d-1) similarly placed internal points so that there is a total of nd+1 points throughout $[a,b]$.

Russell and Shampine prove in particular for linear problems the uniform boundedness of the inverse of their approximate operator concerned with the m^{th} derivative of the approximate solution and also achieve convergence.

These results are analagous to those of our Theorems 1 and 2.

Further advances in the field of nonlinear problems have been achieved by deBoor and Swartz (1973) where they obtain improved rates of convergence over those given by Russell and Shampine (1972) by choosing the $\{s_j\}_{j=0}^{n}$ as a strict partition of $[a,b]$, and with Gauss points in each subinterval.

As with the case when polynomials are used as the approximating subspace the theory employing splines is mainly of an 'a priori' form with error bounds depending upon knowledge of the true solution. For linear problems at least it seems that the roles of the given and approximate operators in the theoretical results might be able to be interchanged to deduce 'a posteriori' bounds for the error of a similar nature to those of Theorem 7.

For such results splines would have, in theory, a definite advantage over polynomials since it is known (see Russell and Shampine (1972)) that the norms of the projections equivalent to our ϕ_n are uniformly bounded. For instance, if appropriately scaled Chebyshev zeros are used as the points over each subinterval then the norm of the projection is the Lebesgue constant $8 + \frac{4}{\pi} \ln(d + 1)$ and this is independent of n as the partition π_n is refined. This would mean that the applicability of the bounds given by results of similar form to Theorem 7 would probably be improved over the polynomial case since the corresponding δ_n would not involve a projection ϕ_n with $\|\phi_n\|$ as $O(\ln(n))$. However in practice this is not likely

to make a great difference because of the very gradual increase with n of the function ln(n).

Another advantage in using splines might be the computational properties of the band matrices if B-splines (which have compact support) are employed as the basis functions for the approximation.

In section 4.8 it was seen that for a projection method we could define an 'extended projection' method and it may be that this process could be applied to the usual polynomial spline solution to yield useful results.

Thus we see that there are areas where the work on the use of polynomials in this thesis might be able to be applied to splines and further research could be undertaken.

6.4 Nonlinear Problems

The approximate solution by polynomial collocation of nonlinear equations has been considered by Vainikko (1965, 1966, 1969). As was mentioned in the previous section spline approximations for such problems have been investigated by Russell and Shampine (1972) and deBoor and Swartz (1973). Results from the above work are essentially of an 'a priori' nature, assuming knowledge of the true solution and deriving order of convergence proofs.

We have not however examined the possibility of an 'a posteriori' error analysis and there are certain problems which would be encountered but it would seem that with further investigation advances might be achieved.

6.5 Elliptic Partial Differential Equations

Kantorovich and Akilov consider two linear elliptic problems and show with appropriate choices of spaces, mappings and norms that their theory can be applied to the Galerkin method. The two examples discussed are

$$\text{(i)} \quad \nabla^2 u + \lambda a(x,y)u \equiv \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \lambda a(x,y)u = v(x,y)$$

and

$$\text{(ii)} \quad \nabla^2 u + \lambda\{a(x,y)u + b(x,y)\frac{\partial u}{\partial x} + c(x,y)\frac{\partial u}{\partial y}\} = v(x,y)$$

with a, b and c continuously differentiable functions and in each case the equations hold over a domain D bounded by a circle Γ with the boundary condition that u vanishes on Γ.

We shall not discuss these in any depth and the reader is referred to Kantorovich and Akilov (1964, pp. 595-601) for a full description. We shall simply mention that for both problems G is taken to be the operator ∇^2, but for example (i) if u is twice continuously differentiable

$$\|u\| = \{\iint_D |\nabla^2 u|^2 dxdy\}^{\frac{1}{2}}$$

whereas for example (ii) the norm is such that

$$\|u\| = \{\iint_D [(\frac{\partial u}{\partial x})^2 + (\frac{\partial u}{\partial y})^2] dxdy\}^{\frac{1}{2}}.$$

To derive their results Kantorovich and Akilov seek an approximate solution which involves not only a polynomial in x and y but also another special function

required to satisfy certain conditions.

If, in some setting, an 'a posteriori' approach is to be examined it would appear that the main problem would be to find a suitable approximating subspace. That is, one for which there were results available concerning its approximating properties in the chosen norm. Also the Green's function for the differential operator G would have to be known explicitly. If these criteria were satisfied it would seem feasible that progress might be made.

Karpilovskaja (1970) has examined the collocation method with the possibility of approximating by trigonometric polynomials and utilises their properties to derive convergence results via the general theorems of Kantorovich and Akilov.

An 'a priori' application of the theory due to Anselone has been considered for the numerical solution of elliptic partial differential equations by Gilbert and Colton (1971).

6.6 Conclusions

The principal work of this thesis has been in developing algorithms for computing error bounds for the numerical solution by polynomial collocation of linear differential equations. This has been achieved by adapting the main theoretical results for 'a posteriori' bounds given in Chapter 2 to produce more readily practical formulae and has entailed relating the inverses of the approximating operators to the inverses of the matrices involved in determining the approximate solution.

The most suitable form of error bound was seen to be

$$\|x - x_n\| \leq \text{BND} \cdot (\text{norm of the residual})$$

where BND is a computed 'a posteriori' bound on the norm of the inverse of the given differential operator. Three rigorous expressions for BND were given by $B_1(n)$, $B_2(n)$ and $B_3(n)$ in equations (5.3, (5.4) and (5.5) respectively. Estimates of these bounds were shown to be $\bar{B}_1(n) = \bar{B}_2(n)$ and $\bar{B}_3(n)$ of (5.6).

All the different results for BND involve the norm of the matrix AoA^{-1} which we saw was independent of the basis used for the polynomial subspace and moreover varied little with the number of collocation points. The approximate solution x_n is of course invariant with the basis and thus so also are the error bounds furnished by our approaches if rounding errors are ignored. However rounding errors occur in practice and we have seen collocating on Chebyshev zeros that the inverse matrix with Chebyshev polynomials as a basis possessed an interesting structure with the property that its norm did not change greatly as larger values of n were taken. This leads to smaller condition numbers than when simple powers are employed, thus minimising the effect of roundoff. Also improvements in the condition number could be made by the use of column scaling and although this scaling does not affect the Gaussian elimination process, which may be employed in any application, it does lead to better bounds on the condition numbers and for these reasons Chebyshev polynomials are recommended as a suitable choice of basis functions. (It is quite likely

that Legendre polynomials would also be a convenient selection).

We now consider the question of which is the most suitable of the various expressions for BND to apply in practice. The rigorous formulae for this quantity only hold for a sufficiently large number of collocation points and, as a comparison, values of n required for these results were given in TABLES 13 and 14. It is concluded from these figures that the rigorous bounds $B_1(n)$ and $B_2(n)$ are not really a practical proposition whereas that $B_3(n)$ would be applicable in certain cases but not in others. To avoid this difficulty the estimates of these bounds were developed. These hold for any number of collocation points and the corresponding errors were seen from the tables to provide reliable results, being closer to the actual norm than the more rigorous bounds.

Clearly in any application the smaller of the two values $\bar{B}_1(n)$ and $\bar{B}_3(n)$ should be chosen. This is principally determined by the magnitude of the constant k_o for the operator under consideration. The size of k_o is in turn dependent upon the coefficients in the linear differential equation. Coefficients which are fairly small would give k_o small and hence $\bar{B}_3(n)$ as the lesser of the two. Conversely for larger functions in the operator we should expect $\bar{B}_1(n)$ to give the better result.

We discussed the point that the error bounds directly from the theory were measured in the X-norm which for second order problems was the infinity norm of the second derivative of the error. Very good predicted results were achieved in this norm but when we related these to values in the infinity

norm our bounds were not in such good agreement with the corresponding actual computed error due to coarse inequalities in the transformation.

We now mention two points concerning the implementation on the machine of our bounds.

We have said that the norm of the residual is calculated by evaluation at several points and by selection of the maximum in magnitude of these values. This is considerably less work than computing a strict bound but even this process does involve a certain amount of computing time and it would be convenient if reasonable bounds could be found which avoided this but it does not seem that this would be possible.

Secondly, we have seen that in obtaining values for our constants k_{max}, k_o, k_1 and k_2 from the formulae of section 5.3 cancellation within the algebraic expressions is often possible, yielding quite small results. However these formulae could be more automated, consequently requiring less work from the user but giving larger answers. The estimates of the bounds however involving only k_{max} and k_o, for which the expressions are simpler, may be more suited to automation since cancellation is less likely.

Finally we have examined briefly areas in which possible extensions or generalisations of our analysis might be applied. We suggest that the development of 'a posteriori' bounds when splines are chosen as the approximating functions would be the topic most likely to yield useful results, but there would seem to be several fields where further investigation could usefully be undertaken.

BIBLIOGRAPHY

1. P.M. Anselone (1971), 'Collectively Compact Operator Approximation Theory', Prentice-Hall.

2. R. Bellman and R.E. Kalaba (1965), 'Quazilinearisation and Nonlinear Boundary Value Problems', Elsevier.

3. C.R. deBoor (1966), 'The method of projections as applied to the numerical solution of two point boundary value problems using cubic splines', Doctoral thesis, University of Michigan, Ann Arbor.

4. C.R. deBoor and B. Swartz (1973), 'Collocation at Gaussian points', S.I.A.M. Journal of Numerical Analysis 10 p.582.

5. A.L. Brown and A. Page (1970), 'Elements of Functional Analysis', Van Nostrand.

6. E.W. Cheney (1966), 'Introduction to Approximation Theory', McGraw-Hill.

7. P.G. Ciarlet, M.H. Schultz and R.S. Varga (1967), 'Numerical methods of high order accuracy for the solution of boundary value problems. I. One dimensional problem', Numerische Mathematik 9, p.394.

8. P.G. Ciarlet, M.H. Schultz and R.S. Varga (1969), 'Numerical methods of high order accuracy for the solution of boundary value problems. V. Monotone operator theory', Numerische Mathematik 13, p.51.

9. C.W. Clenshaw and H.J. Norton (1963), 'The solution of nonlinear ordinary differential equations in Chebyshev series', Computer Journal 6, p.88.

10. D.B. Coldrick (1972), 'Methods for the numerical solution of integral equations of the second kind', Doctoral thesis, University of Toronto.

11. L. Collatz (1960), 'The Numerical Treatment of Differential Equations', Springer-Verlag.

12. L. Collatz (1966), 'Functional Analysis and Numerical Mathematics', Academic Press.

13. P.J. Davis (1963), 'Interpolation and Approximation', Blaisdell.

14. B.A. Finlayson and L.E. Scriven (1966), 'The method of weighted residuals - A review', Applied Mechanics Reviews 19, No. 9, pp.735-748.

15. R.P. Gilbert and D.L. Colton (1971), 'On the numerical treatment of partial differential equations by function theoretic methods', Proceedings of Synspade 1970, Numerical Solution of Partial Differential Equations - II, May, 1970, Maryland. (Ed. Hubbard), Academic Press.

16. S.H. Gould (1957), 'Variational Methods for Eigenvalue Problems', University of Toronto Press.

17. L.V. Kantorovich (1934), 'A method of approximate solution of partial differential equations', Doklady Akademii Nauk SSSR II, p.532.

18. L.V. Kantorovich (1948), 'Functional analysis and applied mathematics', Uspekhi Matem. Nauk. 3, No. 6, p.89.

19. L.V. Kantorovich and G.P. Akilov (1964), 'Functional Analysis in Normed Spaces', Pergamon.

20. L.V. Kantorovich and V.I. Krylov (1958), 'Approximate Methods of Higher Analysis', Noordhoff.

21. E.B. Karpilovskaja (1953), 'On the convergence of an interpolation method for ordinary differential equations', Uspekhi Matem. Nauk 8, No. 3, pp.111-118.

22. E.B. Karpilovskaja (1963), 'Convergence of the collocation method', Sov. Math. 4, p.1070.

23. E.B. Karpilovskaja (1970), 'A method of collocation for integro-differential equations with biharmonic principal part', U.S.S.R. Comp. Math. and Math. Phys. 10, No. 6, p.240.

24. H.B. Keller (1968), 'Numerical Methods for Two-Point Boundary Value Problems', Blaisdell.

25. C. Lanczos (1938), 'Trignometric interpolation of empirical and analytical functions', J. Math. Phys. 17, pp.123-129.

26. T.R. Lucas and G.W. Reddien Jr. (1972), 'Some collocation methods for nonlinear boundary value problems', S.I.A.M. Journal of Numerical Analysis 9, No. 2, pp.341-356.

27. S.G. Mikhlin (1970), 'The Numerical Performance of Variational Methods', Noordhoff.

28. S.G. Mikhlin and K.L. Smolitskiy (1967), 'Approximate Methods for the Solution of Differential and Integral Equations', Elsevier.

29. W.E. Milne (1953), 'The Numerical Solution of Differential Equations', Wiley.

30. I.P. Natanson (1965), 'Constructive Function Theory, Volume III', Ungar.

31. J.L. Phillips (1969), 'Collocation as a projection method for solving integral and other operator equations', Doctoral thesis, Purdue University, Lafayette, Ind.

32. J.L. Phillips (1972), title as above, S.I.A.M. Journal of Numerical Analysis 9, No. 1, p.14.

33. L.B. Rall (1969), 'Computational Solution of Non-linear Operator Equations', Wiley.

34. S.M. Roberts and J.S. Shipman (1972), 'Two Point Boundary Value Problems: Shooting Methods', Elsevier.

35. R.D. Russell and L.F. Shampine (1972), 'A collocation method for boundary value problems', Numerische Mathematik 19, pp.1-28.

36. A.A. Shindler (1969), 'Rate of convergence of the enriched collocation method for ordinary differential equations', Siberian Mathematical Journal 10, p.160.

37. J. Todd (1962), 'Survey of Numerical Analysis', McGraw-Hill.

38. G.M. Vainikko (1965), 'On the stability and convergence of the collocation method', Differentsial'nye Uravneniya 1, p.244.

39. G.M. Vainikko (1966), 'The convergence of the collocation method for nonlinear differential equations', U.S.S.R. Comp. Math. and Math. Phys. 6, No. 1, p.47.

40. G.M. Vainikko (1969), 'The compact approximation principle in the theory of approximation methods', U.S.S.R. Comp. Math. and Math. Phys. 9, No. 4, pp.1-32.

41. J.H. Wilkinson (1965), 'The Algebraic Eigenvalue Problem', Clarendon.

42. K. Wright (1964), 'Chebyshev collocation methods for ordinary differential equations', Computer Journal 6, p.358.

APPENDIX

Additional Numerical Examples of the Application of the Error Bounds and Estimates

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 2 ALPHA= 1.0

N	6	8	10	12
$\bar{B}1(N)$	2.16	2.18	2.21	2.23
$\bar{B}3(N)$	2.08	2.09	2.10	2.11
RES(N)	4.56'-05	2.01'-07	5.53'-10	1.04'-12
E3(N)	1.20'-04	5.26'-07	1.45'-09	2.74'-12
$\bar{E}1(N)$	9.85'-05	4.38'-07	1.22'-09	2.33'-12
$\bar{E}3(N)$	9.48'-05	4.20'-07	1.16'-09	2.21'-12
E(N)	4.45'-05	1.98'-07	5.48'-10	1.04'-12
F3(N)	5.98'-05	2.63'-07	7.25'-10	1.37'-12
$\bar{F}1(N)$	4.92'-05	2.19'-07	6.11'-10	1.16'-12
$\bar{F}3(N)$	4.74'-05	2.10'-07	5.82'-10	1.10'-12
F(N)	1.28'-06	3.38'-09	5.58'-12	7.22'-15

TABLE 22

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 2A ALPHA= 1.0

N	6	8	10	12
$\bar{B}1(N)$	2.16	2.18	2.21	2.23
$\bar{B}3(N)$	2.08	2.09	2.10	2.11
RES(N)	3.72'-02	1.22'-03	2.58'-05	3.37'-07
$\bar{E}3(N)$	9.76'-02	3.19'-03	6.75'-05	8.84'-07
$\bar{E}1(N)$	8.04'-02	2.65'-03	5.69'-05	7.52'-07
$\bar{E}3(N)$	7.74'-02	2.54'-03	5.42'-05	7.13'-07
E(N)	3.71'-02	1.22'-03	2.57'-05	3.36'-07
F3(N)	4.88'-02	1.59'-03	3.38'-05	4.42'-07
$\bar{F}1(N)$	4.02'-02	1.33'-03	2.85'-05	3.76'-07
$\bar{F}3(N)$	3.87'-02	1.27'-03	2.71'-05	3.57'-07
F(N)	8.61'-04	1.13'-05	1.39'-07	1.45'-09

TABLE 23

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 28 ALPHA= 1.0

N	15	20	25	30
$\bar{B}1(N)$	2.25	2.27	2.29	2.3[illegible]
$\bar{B}3(N)$	2.13	2.14	2.14	2.15
RES(N)	9.28'-02	3.96'-02	4.06'-03	1.76'-03
F3(N)	2.43'-01	1.04'-01	1.06'-02	4.62'-03
$\bar{E}1(N)$	2.09'-01	9.01'-02	9.29'-03	4.06'-03
$\bar{E}3(N)$	1.97'-01	8.47'-02	8.70'-03	3.79'-03
E(N)	9.28'-02	3.95'-02	4.05'-03	1.76'-03
F3(N)	1.22'-01	5.19'-02	5.32'-03	2.31'-03
$\bar{F}1(N)$	1.05'-01	4.51'-02	4.65'-03	2.[illegible]3'-03
$\bar{F}3(N)$	9.87'-02	4.24'-02	4.35'-03	1.90'-03
F(N)	1.07'-03	1.49'-04	6.93'-06	2.20'-06

TABLE 24

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 2C ALPHA= 1.0

N	5	8	12	15
$\bar{B}1(N)$	2.14	2.18	2.23	2.25
$\bar{B}3(N)$	2.07	2.09	2.11	2.13
RES(N)	6.22'-03	2.82'-03	7.65'-04	1.64'-04
E3(N)	1.63'-02	7.39'-03	2.00'-03	4.30'-04
$\bar{E}1(N)$	1.33'-02	6.15'-03	1.71'-03	3.70'-04
$\bar{E}3(N)$	1.29'-02	5.89'-03	1.62'-03	3.49'-04
L(N)	6.28'-03	2.69'-03	7.28'-04	1.64'-04
F3(N)	8.15'-03	3.69'-03	1.00'-03	2.15'-04
$\bar{F}1(N)$	6.66'-03	3.07'-03	8.53'-04	1.85'-04
$\bar{F}3(N)$	6.44'-03	2.95'-03	8.09'-04	1.75'-04
F(N)	6.83'-04	1.07'-04	1.76'-05	6.61'-06

TABLE 25

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 3 ALPHA= 1.0

N	6	8	10	12
$\bar{B}1(N)$	1.07	1.07	1.07	1.07
$\bar{B}3(N)$	1.07	1.07	1.07	1.07
RES(N)	2.20'-06	3.69'-08	5.61'-10	7.67'-12
E3(N)	2.38'-06	3.99'-08	6.06'-10	8.28'-12
$\bar{F}1(N)$	2.36'-06	3.96'-08	6.01'-10	8.23'-12
$\bar{E}3(N)$	2.35'-06	3.94'-08	5.98'-10	8.18'-12
F(N)	2.20'-06	3.69'-08	5.60'-10	7.67'-12
F3(N)	1.19'-06	2.00'-08	3.03'-10	4.14'-12
$\bar{F}1(N)$	1.18'-06	1.98'-08	3.01'-10	4.11'-12
$\bar{F}3(N)$	1.18'-06	1.97'-08	2.99'-10	4.09'-12
F(N)	5.78'-08	5.34'-10	4.83'-12	4.72'-14

TABLE 26

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 3A ALPHA= 1.0

N	6	8	10	12
$\bar{B}1(N)$	1.07	1.07	1.07	1.07
$\bar{B}3(N)$	1.07	1.07	1.07	1.07
RES(N)	4.16'-02	1.37'-03	2.93'-05	3.86'-07
E3(N)	4.49'-02	1.48'-03	3.17'-05	4.17'-07
$\bar{E}1(N)$	4.46'-02	1.47'-03	3.14'-05	4.15'-07
$\bar{E}3(N)$	4.44'-02	1.46'-03	3.13'-05	4.12'-07
E(N)	4.16'-02	1.37'-03	2.93'-05	3.86'-07
F3(N)	2.24'-02	7.41'-04	1.58'-05	2.09'-07
$\bar{F}1(N)$	2.23'-02	7.36'-04	1.57'-05	2.07'-07
$\bar{F}3(N)$	2.22'-02	7.32'-04	1.56'-05	2.06'-07
F(N)	1.02'-03	1.29'-05	1.59'-07	1.69'-09

TABLE 27

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 3B ALPHA= 1.0

N	15	20	25	30
$\bar{B}1(N)$	1.07	1.07	1.07	1.07
$\bar{B}3(N)$	1.07	1.07	1.07	1.07
RES(N)	9.29'-02	3.96'-02	4.06'-03	1.76'-03
E3(N)	1.00'-01	4.28'-02	4.38'-03	1.91'-03
$\bar{E}1(N)$	9.96'-02	4.25'-02	4.35'-03	1.89'-03
$\bar{E}3(N)$	9.91'-02	4.23'-02	4.33'-03	1.88'-03
E(N)	9.29'-02	3.96'-02	4.06'-03	1.76'-03
F3(N)	5.02'-02	2.14'-02	2.19'-03	9.53'-04
$\bar{F}1(N)$	4.98'-02	2.13'-02	2.18'-03	9.46'-04
$\bar{F}3(N)$	4.96'-02	2.11'-02	2.17'-03	9.42'-04
F(N)	1.29'-03	1.59'-04	7.30'-06	2.22'-06

TABLE 28

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 3C ALPHA= 1.0

N	5	8	12	15
$\bar{B}1(N)$	1.07	1.07	1.07	1.07
$\bar{B}3(N)$	1.07	1.07	1.07	1.07
RES(N)	8.01'-03	2.83'-03	7.66'-04	1.64'-04
E3(N)	8.65'-03	3.06'-03	8.27'-04	1.77'-04
$\bar{E}1(N)$	8.59'-03	3.04'-03	8.21'-04	1.76'-04
$\bar{E}3(N)$	8.55'-03	3.02'-03	8.17'-04	1.75'-04
E(N)	8.02'-03	2.80'-03	7.45'-04	1.61'-04
F3(N)	4.33'-03	1.53'-03	4.13'-04	8.87'-05
$\bar{F}1(N)$	4.29'-03	1.52'-03	4.11'-04	8.81'-05
$\bar{F}3(N)$	4.27'-03	1.51'-03	4.09'-04	8.77'-05
F(N)	9.71'-04	1.30'-04	2.16'-05	8.43'-06

TABLE 29

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 4 ALPHA= 1.0

N	6	8	10	12
$\bar{B}1(N)$	10.5	10.7	10.8	10.8
$\bar{B}3(N)$	14.1	14.3	14.4	14.5
RES(N)	1.16'-03	6.45'-05	3.38'-06	1.50'-07
E3(N)	2.24'-02	1.25'-03	6.52'-05	2.89'-06
$\bar{E}1(N)$	1.21'-02	6.88'-04	3.64'-05	1.62'-06
$\bar{E}3(N)$	1.63'-02	9.25'-04	4.88'-05	2.17'-06
E(N)	1.17'-03	6.31'-05	3.32'-06	1.48'-07
F3(N)	1.12'-02	6.23'-04	3.26'-05	1.45'-06
$\bar{F}1(N)$	6.07'-03	3.44'-04	1.82'-05	8.10'-07
$\bar{F}3(N)$	8.17'-03	4.62'-04	2.44'-05	1.09'-06
F(N)	2.64'-05	8.32'-07	2.62'-08	8.22'-10

TABLE 30

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 4A ALPHA= 1.0

N	6	8	10	12
$\bar{B}1(N)$	10.5	10.7	10.8	10.8
$\bar{B}3(N)$	14.1	14.3	14.4	14.5
RES(N)	4.95'-02	1.60'-03	3.59'-05	4.15'-07
E3(N)	9.55'-01	3.08'-02	6.92'-04	8.02'-06
$\bar{E}1(N)$	5.19'-01	1.70'-02	3.86'-04	4.49'-06
$\bar{E}3(N)$	6.98'-01	2.29'-02	5.18'-04	6.03'-06
E(N)	4.90'-02	1.63'-03	3.61'-05	4.17'-07
F3(N)	4.78'-01	1.54'-02	3.46'-04	4.01'-06
$\bar{F}1(N)$	2.60'-01	8.51'-03	1.93'-04	2.25'-06
$\bar{F}3(N)$	3.49'-01	1.14'-02	2.59'-04	3.01'-06
F(N)	1.34'-03	1.58'-05	2.11'-07	1.87'-09

TABLE 31

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 4B ALPHA= 1.0

N	15	20	25	30
$\bar{B}1(N)$	10.8	10.9	10.9	10.9
$\bar{B}3(N)$	14.6	14.6	14.6	14.6
RES(N)	9.50'-02	3.96'-02	4.12'-03	1.76'-03
E3(N)	1.83'+00	7.64'-01	7.94'-02	3.40'-02
$\bar{E}1(N)$	1.03'+00	4.31'-01	4.48'-02	1.92'-02
$\bar{E}3(N)$	1.38'+00	5.78'-01	6.02'-02	2.58'-02
E(N)	9.60'-02	3.95'-02	4.13'-03	1.76'-03
F3(N)	9.16'-01	3.82'-01	3.97'-02	1.70'-02
$\bar{F}1(N)$	5.15'-01	2.15'-01	2.24'-02	9.61'-03
$\bar{F}3(N)$	6.91'-01	2.89'-01	3.01'-02	1.29'-02
F(N)	1.24'-03	1.57'-04	7.23'-06	2.22'-06

TABLE 32

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 4C ALPHA= 1.0

N	5	8	12	15
$\bar{B}1(N)$	10.3	10.7	10.8	10.8
$\bar{B}3(N)$	13.9	14.3	14.5	14.6
RES(N)	1.13'-02	2.76'-03	7.64'-04	1.61'-04
E3(N)	2.18'-01	5.33'-02	1.48'-02	3.11'-03
$\bar{E}1(N)$	1.17'-01	2.95'-02	8.26'-03	1.75'-03
$\bar{E}3(N)$	1.57'-01	3.96'-02	1.11'-02	2.34'-03
E(N)	1.01'-02	2.68'-03	7.35'-04	1.65'-04
F3(N)	1.09'-01	2.67'-02	7.38'-03	1.55'-03
$\bar{F}1(N)$	5.83'-02	1.47'-02	4.13'-03	8.73'-04
$\bar{F}3(N)$	7.86'-02	1.98'-02	5.55'-03	1.17'-03
F(N)	1.06'-03	1.24'-04	2.08'-05	8.08'-06

TABLE 33

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 5 ALPHA= 1.0

N	6	8	10	12
$\bar{B}1(N)$	3.24	3.28	3.29	3.30
$\bar{B}3(N)$	3.43	3.46	3.47	3.48
RES(N)	2.91'-03	6.47'-04	3.06'-05	7.25'-07
E3(N)	1.27'-02	2.83'-03	1.34'-04	3.17'-06
$\bar{E}1(N)$	9.41'-03	2.12'-03	1.01'-04	2.40'-06
$\bar{E}3(N)$	9.97'-03	2.24'-03	1.06'-04	2.52'-06
E(N)	2.90'-03	6.46'-04	3.05'-05	7.23'-07
F3(N)	6.36'-03	1.42'-03	6.71'-05	1.59'-06
$\bar{F}1(N)$	4.71'-03	1.06'-03	5.04'-05	1.20'-06
$\bar{F}3(N)$	4.98'-03	1.12'-03	5.31'-05	1.26'-06
F(N)	6.42'-05	6.57'-06	2.30'-07	3.28'-09

TABLE 34

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 5A ALPHA= 1.0

N	6	8	10	12
$\bar{B}1(N)$	3.24	3.28	3.29	3.3
$\bar{B}3(N)$	3.43	3.46	3.47	3.48
RES(N)	4.55'-02	1.93'-03	1.90'-04	7.02'-06
E3(N)	1.99'-01	8.45'-03	8.32'-04	3.08'-05
$\bar{E}1(N)$	1.47'-01	6.32'-03	6.24'-04	2.32'-05
$\bar{E}3(N)$	1.56'-01	6.67'-03	6.58'-04	2.44'-05
E(N)	4.54'-02	1.93'-03	1.89'-04	7.01'-06
F3(N)	9.95'-02	4.23'-03	4.16'-04	1.54'-05
$\bar{F}1(N)$	7.36'-02	3.16'-03	3.12'-04	1.16'-05
$\bar{F}3(N)$	7.80'-02	3.34'-03	3.29'-04	1.22'-05
F(N)	1.08'-03	3.17'-05	1.66'-06	3.53'-08

TABLE 35

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 5B ALPHA= 1.0

N	15	20	25	30
$\bar{B}1(N)$	3.32	3.33	3.34	3.35
$\bar{B}3(N)$	3.49	3.50	3.50	3.51
RES(N)	9.29'-02	3.96'-02	4.06'-03	1.76'-03
E3(N)	4.07'-01	1.74'-01	1.78'-02	7.73'-03
$\bar{E}1(N)$	3.08'-01	1.32'-01	1.36'-02	5.91'-03
$\bar{E}3(N)$	3.24'-01	1.39'-01	1.42'-02	6.20'-03
E(N)	9.30'-02	3.96'-02	4.06'-03	1.76'-03
F3(N)	2.04'-01	8.68'-02	8.89'-03	3.87'-03
$\bar{F}1(N)$	1.54'-01	6.61'-02	6.78'-03	2.95'-03
$\bar{F}3(N)$	1.62'-01	6.94'-02	7.12'-03	3.10'-03
F(N)	1.27'-03	1.58'-04	7.28'-06	2.22'-06

TABLE 36

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 5C ALPHA= 1.0

N	5	8	12	15
$\bar{B}1(N)$	3.22	3.28	3.30	3.32
$\bar{B}3(N)$	3.41	3.46	3.48	3.49
RES(N)	5.07'-02	2.88'-03	7.64'-04	1.64'-04
E3(N)	2.22'-01	1.26'-02	3.35'-03	7.20'-04
$\bar{E}1(N)$	1.63'-01	9.45'-03	2.53'-03	5.45'-04
$\bar{E}3(N)$	1.73'-01	9.97'-03	2.66'-03	5.73'-04
E(N)	5.04'-02	2.85'-03	7.43'-04	1.61'-04
F3(N)	1.11'-01	6.32'-03	1.67'-03	3.60'-04
$\bar{F}1(N)$	8.16'-02	4.73'-03	1.26'-03	2.73'-04
$\bar{F}3(N)$	8.66'-02	4.99'-03	1.33'-03	2.87'-04
F(N)	2.55'-03	1.29'-04	2.12'-05	8.26'-06

TABLE 37

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 6 ALPHA= 1.0

N	6	8	10	12
$\bar{B}1(N)$	1.24	1.24	1.24	1.24
$\bar{B}3(N)$	1.18	1.18	1.18	1.18
RES(N)	3.13'-04	1.14'-05	1.32'-06	8.86'-09
E3(N)	4.60'-04	1.68'-05	1.95'-06	1.30'-08
$\bar{E}1(N)$	3.89'-04	1.42'-05	1.65'-06	1.10'-08
$\bar{E}3(N)$	3.71'-04	1.36'-05	1.57'-06	1.05'-08
E(N)	3.13'-04	1.14'-05	1.32'-06	8.86'-09
F3(N)	2.30'-04	8.41'-06	9.73'-07	6.51'-09
$\bar{F}1(N)$	1.95'-04	7.12'-06	8.23'-07	5.51'-09
$\bar{F}3(N)$	1.85'-04	6.78'-06	7.84'-07	5.25'-09
F(N)	9.02'-06	1.41'-07	1.30'-08	6.50'-11

TABLE 38

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 6A ALPHA= 1.0

N	6	8	10	12
$\bar{B}1(N)$	1.24	1.24	1.24	1.24
$\bar{B}3(N)$	1.18	1.18	1.18	1.18
RES(N)	4.99'-02	2.67'-03	1.19'-04	3.65'-06
E3(N)	7.33'-02	3.93'-03	1.75'-04	5.37'-06
$\bar{E}1(N)$	6.20'-02	3.32'-03	1.48'-04	4.55'-06
$\bar{E}3(N)$	5.91'-02	3.17'-03	1.41'-04	4.33'-06
E(N)	4.99'-02	2.67'-03	1.19'-04	3.65'-06
F3(N)	3.67'-02	1.96'-03	8.73'-05	2.69'-06
$\bar{F}1(N)$	3.10'-02	1.66'-03	7.39'-05	2.27'-06
$\bar{F}3(N)$	2.95'-02	1.58'-03	7.03'-05	2.16'-06
F(N)	1.27'-03	3.15'-05	1.03'-06	2.50'-08

TABLE 39

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 6B ALPHA= 1.0

N	15	20	25	30
$\bar{B}1(N)$	1.24	1.24	1.25	1.25
$\bar{B}3(N)$	1.18	1.18	1.18	1.18
RES(N)	9.29'-02	3.96'-02	4.06'-03	1.76'-03
E3(N)	1.37'-01	5.83'-02	5.97'-03	2.59'-03
$\bar{E}1(N)$	1.16'-01	4.94'-02	5.06'-03	2.20'-03
$\bar{E}3(N)$	1.10'-01	4.70'-02	4.81'-03	2.09'-03
E(N)	9.29'-02	3.96'-02	4.06'-03	1.76'-03
F3(N)	6.83'-02	2.91'-02	2.98'-03	1.30'-03
$\bar{F}1(N)$	5.78'-02	2.47'-02	2.53'-03	1.10'-03
$\bar{F}3(N)$	5.50'-02	2.35'-02	2.40'-03	1.05'-03
F(N)	1.30'-03	1.60'-04	7.34'-06	2.23'-06

TABLE 40

APPLICATION OF THE ERROR BOUNDS AND ESTIMATES

PROBLEM 6C ALPHA= 1.0

N	5	8	12	15
$\bar{B}1(N)$	1.24	1.24	1.24	1.24
$\bar{B}3(N)$	1.18	1.18	1.18	1.18
RES(N)	2.30'-02	2.82'-03	7.66'-04	1.64'-04
$\bar{E}3(N)$	3.38'-02	4.14'-03	1.13'-03	2.42'-04
$\bar{E}1(N)$	2.86'-02	3.50'-03	9.53'-04	2.05'-04
$\bar{E}3(N)$	2.72'-02	3.34'-03	9.07'-04	1.95'-04
E(N)	2.30'-02	2.80'-03	7.47'-04	1.61'-04
$\bar{F}3(N)$	1.69'-02	2.07'-03	5.63'-04	1.21'-04
$\bar{F}1(N)$	1.43'-02	1.75'-03	4.77'-04	1.02'-04
$\bar{F}3(N)$	1.36'-02	1.67'-03	4.54'-04	9.73'-05
F(N)	1.59'-03	1.32'-04	2.20'-05	8.59'-06

TABLE 41