DIFFERENCE METHODS FOR
ORDINARY DIFFERENTIAL EQUATIONS
WITH APPLICATIONS TO PARABOLIC EQUATIONS

by

Eusebius Jacobus Doedel
B.Sc., University of British Columbia, 1971
M.Sc., University of British Columbia, 1973

A Thesis Submitted in Partial Fulfilment of
the Requirements for the Degree of
Doctor of Philosophy

in the Institute
of
Applied Mathematics

We accept this thesis as conforming to the
required standard

THE UNIVERSITY OF BRITISH COLUMBIA
January, 1976
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Mathematics

The University of British Columbia
2075 Wesbrook Place
Vancouver, Canada
V6T 1W5

Date March 11 1976
The first chapter of the thesis is concerned with the construction of finite difference approximations to boundary value problems in linear nth order ordinary differential equations. This construction is based upon a local collocation procedure with polynomials, which is equivalent to a method of undetermined coefficients. It is shown that the coefficients of these finite difference approximations can be expressed as the determinants of matrices of relatively small dimension. A basic theorem states that these approximations are consistent, provided only that a certain normalization factor does not vanish. This is the case for compact difference equations and for difference equations with only one collocation point. The order of consistency may be improved by suitable choice of the collocation points. Several examples of known, as well as new difference approximations are given. Approximations to boundary conditions are also treated in detail. The stability theory of H. O. Kreiss is applied to investigate the stability of finite difference schemes based upon these approximations. A number of numerical examples are also given.

In the second chapter it is shown how the construction method of the first chapter can be extended to initial value problems for systems of linear first order ordinary differential equations. Specific examples are included and the well-known stability theory for these difference equations is summarized.

It is then shown how these difference methods may be applied to linear parabolic partial differential equations in one space variable after first discretizing in space by a suitable method from the first chapter.
The stability of such difference schemes for parabolic equations is investigated using an eigenvalue-eigenvector analysis. In particular, the effect of various approximations to the boundary conditions is considered. The relation of this analysis to the stability theory of J. M. Varah is indicated. Numerical examples are also included.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ABSTRACT</strong></td>
<td>i</td>
</tr>
<tr>
<td><strong>TABLE OF CONTENTS</strong></td>
<td>iv</td>
</tr>
<tr>
<td><strong>ACKNOWLEDGMENTS</strong></td>
<td>v</td>
</tr>
<tr>
<td>Chapter</td>
<td></td>
</tr>
<tr>
<td><strong>I. BOUNDARY VALUE PROBLEMS</strong></td>
<td></td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2. The Construction of Finite Difference Approximations</td>
<td>6</td>
</tr>
<tr>
<td>3. Effect of the Choice of Collocation Points</td>
<td>19</td>
</tr>
<tr>
<td>4. Examples of Finite Difference Approximations</td>
<td>21</td>
</tr>
<tr>
<td>5. Finite Difference Approximations to Boundary Conditions</td>
<td>29</td>
</tr>
<tr>
<td>6. Examples of Approximations to Boundary Conditions</td>
<td>34</td>
</tr>
<tr>
<td>7. The Stability of Finite Difference Schemes</td>
<td>37</td>
</tr>
<tr>
<td>8. Numerical Examples</td>
<td>46</td>
</tr>
<tr>
<td><strong>II. INITIAL VALUE PROBLEMS</strong></td>
<td></td>
</tr>
<tr>
<td>1. Introduction</td>
<td>54</td>
</tr>
<tr>
<td>2. Initial Value Problems in Ordinary Differential Equations</td>
<td>55</td>
</tr>
<tr>
<td>3. Examples of Finite Difference Approximations</td>
<td>58</td>
</tr>
<tr>
<td>4. The Stability of Finite Difference Approximations to Initial Value Problems</td>
<td>65</td>
</tr>
<tr>
<td>5. Initial Boundary Value Problems</td>
<td>71</td>
</tr>
<tr>
<td>6. The Stability of Finite Difference Approximations to Initial Boundary Value Problems</td>
<td>76</td>
</tr>
<tr>
<td>7. Numerical Examples</td>
<td>91</td>
</tr>
<tr>
<td><strong>BIBLIOGRAPHY</strong></td>
<td>117</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

The author wishes to thank Professor J. M. Varah for stimulating and guiding this research. He is also thankful to his wife Adrienne for her patience. Finally, he is indebted to Mrs. Vivian Davies for her excellent typing of the manuscript and to Dr. Fausto Milinazzo for proofreading the manuscript.
Chapter I

BOUNDARY VALUE PROBLEMS

1. INTRODUCTION

In this chapter the construction of finite difference approximations to the linear differential equation

\( Ly(x) = y^{(n)}(x) + \sum_{k=0}^{n-1} a_k(x)y^{(k)}(x) = f(x), \quad 0 \leq x \leq 1, \)

is investigated.

Define a mesh \( S_h = \{ x_j : 0 = x_0 < x_1 < \cdots < x_J = 1 \} \) with mesh-spacings \( h_j = x_j - x_{j-1}, \quad (1 \leq j \leq J) \). Let \( h = \max h_j \). It will always be assumed that \( \min h_j \geq \frac{1}{K} h \) for some constant \( K \). For any function \( w(x) \) on \( S_h \) let \( w_j = w(x_j) \).

Finite difference approximations to the differential equation (1.1) at the point \( x = x_j \) are assumed to have the form

\( L_h u_j = \sum_{i=-r}^{s} d_{ji} u_{j+i} = \tilde{f}_j, \quad k_0 \leq j \leq J - n + k_0, \quad k_0 \geq 0, \)

where \( \tilde{f}_j \) is some approximation to \( f_j \). Thus for each mesh \( S_h \) there are precisely \( J - n + 1 \) such equations. The width \( \omega_j \) of each finite difference approximation is defined as \( \omega_j = r_j + s_j + 1 \). If \( \omega_j = n + 1 \) then the approximation is said to be compact. The assumption is made that \( \omega_j \leq \omega_{\text{max}} \) for some integer \( \omega_{\text{max}} \) that does not depend on \( S_h \).

The local truncation error is given by \( \tau_j = L_h y_j - \tilde{f}_j \). If there is a
constant $C$ independent of $h$ such that $|\tau_j| \leq C h^n$ with $n_s > 0$, and $n_s$ as high as possible, then the difference approximation is said to be **consistent** with the differential equation and the order of consistency is equal to $n_s$.

To specify $y(x)$ completely it is necessary to impose boundary conditions, which are assumed to take the form

\[(1.1a) \quad B_k(0)y(0) = \sum_{i=0}^{n_k(0)} b_{k,i}(0)y(i)(0) = b_k(0), \quad 1 \leq k \leq n_0,\]

and

\[(1.1b) \quad B_k(1)y(1) = \sum_{i=0}^{n_k(1)} b_{k,i}(1)y(i)(1) = b_k(1), \quad 1 \leq k \leq n-n_0,\]

where

$n_k(0) < n$ and $n_k(1) < n$.

It will always be assumed that the differential equation (1.1) subject to the boundary conditions (1.1a,b) admits a unique solution that is as many times continuously differentiable as needed.

Similarly, to define a finite difference scheme that determines the approximations $u_j$ to $y_j$ completely one has approximations to the boundary conditions (1.1a,b) that look like

\[(1.2a) \quad B_{h,k}(0)u_0 = \sum_{i=0}^{s_k(0)} d_{k,i}(0)u_i = \tilde{b}_k(0), \quad 1 \leq k \leq n_0,\]

and

\[(1.2b) \quad B_{h,k}(1)u_1 = \sum_{i=0}^{s_k(1)} d_{k,i}(1)u_i = \tilde{b}_k(1), \quad 1 \leq k \leq n-n_0,\]
Here \( \tilde{b}_k(0) \) and \( \tilde{b}_k(1) \) are approximations to \( b_k(0) \) and \( b_k(1) \) respectively.

The definitions of truncation error and consistency are similar to the previous definitions and the equations are compact if \( s_k(0) = n_k(0) \) and \( r_k(1) = n_k(1) \). A finite difference scheme is said to be (accurate) of order \( s \) if the approximate solution \( u_j \), \( (0 \leq j \leq J) \), is uniquely defined by (1.2) and (1.2a,b) and \( |y_j - u_j| \leq K_1 h^s \) for some constant \( K_1 \) that does not depend on \( h \).

Numerous texts and papers are partly or entirely concerned with finite difference approximations to (1.1) and (1.1a,b). In particular, the books by Collatz (1960) and Keller (1968) need to be mentioned here. The general theory of difference approximations to systems of boundary value problems is contained in papers by Grigorieff (1970) and Kreiss (1972). An extensive survey of finite difference methods for boundary value problems can be found in a paper by Keller (1975).

Various finite difference approximations of the form (1.2) appear in the literature. To the second order differential equation

\[
y''(x) + a^1(x)y'(x) + a^0(x)y(x) = f(x),
\]

the simplest difference approximation is perhaps given by

\[
L_h u_j = D^2_h u_j + a^1_j D^1_h u_j + a^0 u_j = f_j.
\]
Here $D^2_{h}u_j = (u_{j+1} - 2u_j + u_{j-1}) / h^2$, $D^1_{h}u_j = (u_{j+1} - u_{j-1}) / 2h$ and the mesh is uniform, i.e., $h_j = h$. The order of consistency is equal to two, as is the order of accuracy of a difference scheme based upon (1.4), provided that the approximations to the boundary conditions are consistent.

Frequently it is desirable to use equations for which the order of consistency is more than two. One way of obtaining such equations is by allowing the width of the approximation to be greater than $n+1$. For example, a fourth order finite difference approximation to (1.3) is

$$L_h u_j = (I - \frac{h^2}{12} D^2_{h}) D^2_{h} u_j + a^1_j (I - \frac{h^2}{6} D^2_{h}) D^1_{h} u_j + a^0_j u_j = f_j.$$  

The width of this approximation is equal to five. If a difference approximation is not compact then an entire difference scheme cannot be based upon that equation alone. Near the boundaries other equations need to be used. In (1.5) for example, $L_h u_j$ is not defined if $j = 1$ or $j = J-1$ unless periodicity is assumed. Hence other approximations need to be used here. Examples of these are given by Shoosmith (1975). In this chapter such extra boundary conditions will be discussed in Sections 7 and 8.

It is not necessary to increase the width of the approximation to obtain higher order equations, for one can also require more "local information" to be available. In this case each difference equation involves the values of the coefficient functions and the inhomogeneous term at more than one point. For example, a fourth order approximation to

$$y''(x) + a^0(x)y(x) = f(x)$$
is given by

\[(1.7) \quad D_h^2 u_j + K_h(\frac{5}{6}) (a_j^0 u_j) = K_h(\frac{5}{6}) f_j,\]

with

\[K_h(c) w_j = \frac{1-c}{2} w_{j-1} + cw_j + \frac{1-c}{2} w_{j+1}.\]

The construction and analysis of such higher order compact approximations is also included in this chapter.

Usually the statement that a certain difference approximation has a given order of consistency can be justified by simply Taylor expanding. Systematic methods of deriving finite difference equations appear infrequently in the literature however, unless one considers finite element methods such as Galerkin, Ritz and collocation procedures with piecewise polynomials as finite difference methods. (For a discussion of these, see e.g., Varga (1966), Ciarlet Schultz and Varga (1967), Russell and Shampine (1972), deBoor and Swartz (1973), as well as the books of Strang and Fix (1973) and Schultz (1973)). Birkhoff and Gulati (1974) compare various compact equations, but no attempt is made to methodically construct such compact difference approximations. Swartz (1974) discusses a procedure for deriving difference approximations that resembles the procedure given in this thesis, but no complete analysis is given.

The main purpose of the present chapter is then to derive a large class of finite difference approximations to the problem (1.1) and to give their order of consistency. This is done in the next two sections. In Section 4 a number of examples is given, including known as well as new
difference equations. In Section 5 the construction of finite difference approximations to the boundary conditions is considered. Examples of these appear in Section 6. In Section 7 the theory of Kreiss (1972) is applied to investigate the stability of difference schemes based upon the difference approximations given in this chapter. Finally, in Section 8 the results of some numerical experiments are given.

2. THE CONSTRUCTION OF FINITE DIFFERENCE APPROXIMATIONS

To construct a finite difference approximation of the form (1.2) to the differential equation (1.1) one may proceed as follows. For given index \( j \) let \( z_1 < z_2 < \cdots < z_m \) be points in the interval \([x_{j-r}, x_{j+s}]\), satisfying \(|z_i - z_k| > \frac{1}{K} h\), \( i \neq k \). (To simplify the notation the subscript \( j \) denoting dependence on the particular difference approximation will be omitted here, so that, e.g., \( z_{j,i} \) becomes \( z_i \).

Let \( P_d \) denote the space of all polynomials with degree not exceeding \( d = r + s + m - 1 \). Assume that for some integer \( \ell \), \((-r \leq \ell \leq s)\), a polynomial \( p^\ell(x) \in P_d \) is determined uniquely by the \( r + s + m \) equations

\[
(2.1) \quad p^\ell(x_{j+i}) = u_{j+i}, \quad -r \leq i \leq s, \quad i \neq \ell
\]

\[
(2.2) \quad L p^\ell(z_i) = f(z_i), \quad 1 \leq i \leq m.
\]

(The equations (2.2) will be referred to as "collocation" equations.) Then the equation

\[
(2.3) \quad p^\ell(x_{j+\ell}) = u_{j+\ell},
\]
relates the quantities \( u_{j+i} \) \((-r \leq i \leq s)\), by an expression (difference equation) of the form

\[
(2.4) \quad L_n u_j = \sum_{i=-r}^{s} d_i u_{j+i} = \sum_{i=1}^{m} e_i f(z_i).
\]

In order to be able to exhibit explicit expressions for the coefficients \( d_i \) and \( e_i \) in (2.4), let the polynomials \( w^i(x) \), \((0 \leq i \leq d)\), form a basis of \( P_d \), so that one can write

\[
p^i(x) = \sum_{i=0}^{d} c_i w^i(x).
\]

Also, define a matrix \( D_L \) by

\[
(2.5) \quad D_L = \begin{bmatrix}
w^0(x_{j-r}) & w^1(x_{j-r}) & \cdots & w^d(x_{j-r}) & u_{j-r} \\
w^0(x_{j-r+1}) & w^1(x_{j-r+1}) & \cdots & w^d(x_{j-r+1}) & u_{j-r+1} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
w^0(x_{j+s}) & w^1(x_{j+s}) & \cdots & w^d(x_{j+s}) & u_{j+s} \\
Lw^0(z_1) & Lw^1(z_1) & \cdots & Lw^d(z_1) & f(z_1) \\
Lw^0(z_2) & Lw^1(z_2) & \cdots & Lw^d(z_2) & f(z_2) \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
Lw^0(z_m) & Lw^1(z_m) & \cdots & Lw^d(z_m) & f(z_m)
\end{bmatrix}
\]

If the relations (2.1), (2.2) and (2.3) are satisfied then \( \det(D_L) = 0 \).

Evaluating the determinant by the last column and introducing a normalizing factor \( E \) one finds that the coefficients \( d_i \) and \( e_i \) in (2.4) are given by

\[
(2.6) \quad d_i = \text{cof}_L [u_{j+i}] / E,
\]
and

(2.7) \[ e_i = -\text{cof}_L[f(z_i)] / E, \]

where \( \text{cof}_L[\cdot] \) is the cofactor of the given element in the matrix \( D_L \), with the operator \( L \) as in equation (1.1).

It is convenient to define the normalizing factor as

\[
E = - \sum_{i=1}^{m} \text{cof}_{L_0} [f(z_i)],
\]

where

\[
L_0 y(x) = y^{(n)}(x).
\]

If the basis functions \( w^i(x) \) are chosen such that

(2.8) \[ w^i(x_{j-r+k}) = \delta_{ik} \quad \text{for} \quad 0 \leq i \leq r + s, \]

and

(2.9) \[ w^i(x_{j+k}) = 0 \quad \text{for} \quad r + s + 1 \leq i \leq d, \quad -r \leq k \leq s, \]

then

\[
(2.10) \quad d_i = \frac{(-1)^m}{E} \begin{vmatrix}
L w^{r+i}(z_1) & L w^{r+s+1}(z_1) & \cdots & L w^{d}(z_1) \\
\vdots & \ddots & \ddots & \ddots \\
L w^{r+i}(z_m) & L w^{r+s+1}(z_m) & \cdots & L w^{d}(z_m)
\end{vmatrix}, \quad -r \leq i \leq s,
\]

and
With the above choice of basis functions, one can also express the normalizing factor as

\[
E = (-1)^m 
\begin{vmatrix}
1 & L_0 w^{r+s+1}(z_1) & \cdots & L_0 w^{r+s+m-1}(z_1) \\
\vdots & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \vdots \\
1 & L_0 w^{r+s+1}(z_m) & \cdots & L_0 w^{r+s+m-1}(z_m)
\end{vmatrix}
\]

If \( m = 1 \) then \( E = -1 \) and \( e_1 = 1 \). Further, the coefficients \( d_i \) and \( e_i \) are related by

\[
d_i = \sum_{k=1}^{m} L_k w^{r+i} e_k, \quad -r \leq i \leq s.
\]

**Remark.** The above formulas for \( d_i \) and \( e_i \) are essentially intended to show the form of these coefficients. In particular, note the type of products of coefficient functions that will appear. For actual application to numerical problems, one may first evaluate the determinants symbolically, collecting the various multiplicative and additive constants. Thereafter, the numerical evaluation of the \( d_i \) and \( e_i \) may be organized in such a way, that these coefficients are obtained in the smallest possible
number of arithmetic operations.

Basis functions $w^i(x)$ satisfying (2.8) and (2.9) are easily constructed. In fact, it is convenient to assume that the $w^i(x)$ are given by

$$w^i(x) = \frac{\sum_{k=0}^{r+s} (x-x_{j-r+k})^{r+s}}{\sum_{k=0}^{r+s} (x-x_{j-r+i}+x_{j-r+k})}, \quad 0 \leq i \leq r+s,$$

and

$$w^{r+s+i}(x) = \frac{\sum_{k=1}^{m-1} (x-t_k)^{m-1}}{\sum_{k=1}^{m-1} (t_i-t_k)}, \quad 1 \leq i \leq m-1.$$

Here the $t_i$, $(1 \leq i \leq m-1)$, are distinct points in the interval $[x_{j-r}x_{j+s}]$. It is not difficult to check that the points $t_i$ can be chosen in such a way that the polynomials $w^i(x)$, $(1 \leq i \leq d)$, form a linearly independent set.

The procedure leading up to the definition of the coefficients $d_i$ and $e_i$ is equivalent to a method of undetermined coefficients. To illustrate this consider the following example. Let $n=2$, $r=s=1$ and $m=2$; i.e., we are concerned with the construction of a difference approximation of the form
(2.15) \[ d_{-1}u_{j-1} + d_0u_j + d_1u_{j+1} = e_1f(z_1) + e_2f(z_2) , \]
to the differential equation
\[ Ly(x) = y''(x) + a^1(x)y'(x) + a^0(x)y(x) = f(x) . \]

This can be accomplished by requiring that (2.15) be satisfied for all polynomials \( p(x) \in P_3 \). (Replacing \( f(z_i) \) by \( Lp(z_i) \)). With \( w^i(x), \ (0 \leq i \leq 3) \), as in (2.13) and (2.14), this requirement leads to the following linear system of equations:

\[
\begin{bmatrix}
1 & 0 & 0 & -Lw^0(z_1) & -Lw^0(z_2) \\
0 & 1 & 0 & -Lw^1(z_1) & -Lw^1(z_2) \\
0 & 0 & 1 & -Lw^2(z_1) & -Lw^2(z_2) \\
0 & 0 & 0 & -Lw^3(z_1) & -Lw^3(z_2)
\end{bmatrix}
\begin{bmatrix}
d_{-1} \\
d_0 \\
d_1 \\
e_1 \\
e_2
\end{bmatrix}
= \begin{bmatrix}
0 \\
0 \\
0 \\
0 \\
0
\end{bmatrix} .
\]

Assume that the rank of the system is equal to four, so that one can rewrite the equations as

\[
\begin{bmatrix}
1 & 0 & 0 & -Lw^0(z_1) \\
0 & 1 & 0 & -Lw^1(z_1) \\
0 & 0 & 1 & -Lw^2(z_1) \\
0 & 0 & 0 & -Lw^3(z_1)
\end{bmatrix}
\begin{bmatrix}
d_{-1} \\
d_0 \\
d_1 \\
e_1
\end{bmatrix}
= \begin{bmatrix}
Lw^0(z_2)e_2 \\
Lw^1(z_2)e_2 \\
Lw^2(z_2)e_2 \\
Lw^3(z_2)e_2
\end{bmatrix} ,
\]

where the matrix on the left is nonsingular, i.e., \( Lw^3(z_1) \neq 0 \). Partition
this matrix into \[
\begin{bmatrix}
I & A \\
0 & B
\end{bmatrix}
\] as indicated. Then \[
\begin{bmatrix}
I & A \\
0 & B
\end{bmatrix}^{-1} = \begin{bmatrix}
I & -AB^{-1} \\
0 & B^{-1}
\end{bmatrix},
\]
and by elementary linear algebra
\[
B^{-1} = \frac{1}{\det(B)} C,
\]
where \(C\) is the matrix of cofactors of \(B^T\). In this simple example
\[
B^{-1} = \frac{-1}{Lw_3(z_1)}.
\]
Hence
\[
d_{-1} = [Lw^0(z_2) - Lw^0(z_1)Lw^3(z_2) / Lw^3(z_1)]e_2,
\]
\[
d_0 = [Lw^1(z_2) - Lw^1(z_1)Lw^3(z_2) / Lw^3(z_1)]e_2,
\]
\[
d_1 = [Lw^2(z_2) - Lw^2(z_1)Lw^3(z_2) / Lw^3(z_1)]e_2,
\]
and
\[
e_1 = [-1/Lw^3(z_1)]Lw^3(z_2)e_2.
\]
Let \(e_2 = \frac{-Lw^3(z_1)}{E}\). Then the coefficients are equal to
\[
d_{-1} = [Lw^0(z_1)Lw^3(z_2) - Lw^0(z_2)Lw^3(z_1)] / E,
\]
\[
d_0 = [Lw^1(z_1)Lw^3(z_2) - Lw^1(z_2)Lw^3(z_1)] / E,
\]
\[
d_1 = [Lw^2(z_1)Lw^3(z_2) - Lw^2(z_2)Lw^3(z_1)] / E,
\]
\[
e_1 = Lw^3(z_2) / E,
\]
\[
e_2 = -Lw^3(z_1) / E,
\]
which is in agreement with the equations (2.10) and (2.11).

More generally the above procedure is well-defined provided that the matrix
has rank equal to \( m-1 \). In particular if the normalizing factor \( E \) is non-zero then the matrix \( M_L \) has rank \( m-1 \) for all small enough \( h \).

With the definitions given this far, it is possible to state the basic theorem of this chapter.

**Theorem 2.1** Let the coefficients \( d_i \) and \( e_i \) and the normalizing factor \( E \) be given by (2.10), (2.11) and (2.12) respectively. Assume that \( E \neq 0 \) and that \( r + s > n \). If \( h \) is small enough then at least \( n + 1 \) of the coefficients \( d_i \) are nonzero and the order of consistency of the finite difference approximation (2.4) is at least equal to \( r + s + m - n \).

**Proof.** Define

\[
\tilde{d}_i = \text{cof}_{\mathcal{L}_0} [u_{j+i}] / E, \quad -r \leq i \leq s,
\]

\[
\tilde{e}_i = -\text{cof}_{\mathcal{L}_0} [f(z_i)] / E, \quad 1 \leq i \leq m,
\]

so that \( \sum_{i=1}^{m} \tilde{e}_i = 1 \).

First it will be shown that if \( E \neq 0 \) then at least \( n + 1 \) of the coefficients \( \tilde{d}_i \) are nonzero. For suppose on the contrary that \( \tilde{d}_{i_k} \neq 0 \), \( 1 \leq k \leq k_1 < n + 1 \) and \( \tilde{d}_i = 0 \) otherwise. Let ;
\[ p(x) \equiv \left[ \prod_{k=1}^{k_1} (x-x_{j+k}) \right] q(x), \text{ where } q(x) \in P_{n-k_1} \text{ is chosen such that } p(x) \text{ has degree } n \text{ and leading coefficient } \frac{1}{n!}. \] By assumption \( n \leq r + s \), so certainly \( n \leq d \). Hence by construction

\[ \sum_{i=-r}^{s} \bar{d}_i p_{j+i} - \sum_{i=-r}^{m} e_i L_0 p(z_i) = 0. \]

But from the definition of \( p(x) \) it follows that \( \sum_{i=-r}^{s} \bar{d}_i p_{j+i} = 0 \) and \( L_0 p(x) = 1 \). Therefore \( \sum_{i=-r}^{m} e_i = 0 \), which is a contradiction. So if \( E \neq 0 \) then \( n + 1 \) or more \( d_i \) are nonzero. Since \( d_i / \bar{d}_i = 1 + O(h) \), the same is true for the \( d_i \), provided \( h \) is small enough.

The truncation error \( \tau_j \) is defined as

\[ \tau_j = L_n y_j - \sum_{i=1}^{m} e_i f(z_i) = \sum_{i=-r}^{s} d_i y_{j+i} - \sum_{i=1}^{m} e_i f(z_i), \]

where \( y(x) \) is the solution to the differential equation (1.1) subject to the boundary conditions (1.1a, b). Taylor expand to get

\[ \tau_j = \sum_{i=-r}^{s} d_i \sum_{k=0}^{d} (x_{j+i} - x_j)^k \frac{y^{(k)}}{k!} - \sum_{i=1}^{m} e_i L \left[ \sum_{k=0}^{d} (z_i - x_j)^k \frac{y^{(k)}}{k!} \right] \]

\[ + \sum_{i=-r}^{s} d_i (x_{j+i} - x_j)^{d+1} \frac{y^{(d+1)}(s_i)}{(d+1)!} - \sum_{i=1}^{m} e_i L \left[ (z_i - x_j)^{d+1} \frac{y^{(d+1)}(t_i)}{(d+1)!} \right] \]

\[ = \sum_{k=0}^{d} \left[ \sum_{i=-r}^{s} d_i (x_{j+i} - x_j)^k - \sum_{i=1}^{m} e_i L (z_i - x_j)^k \right] \frac{y^{(k)}}{k!} + \]
for some \( s_i \) between \( x_j \) and \( x_{j+i} \) and \( t_i \) between \( x_j \) and \( z_i \). By construction the quantity between square brackets equals zero for each \( k \). From the definitions of the coefficients \( d_i \) and \( e_i \) and the basis functions \( w^1(x) \), together with the underlying assumption that \( \frac{h}{K} \leq h_j \leq h \), \( (1 \leq j \leq J) \), it follows that there are constants \( C_1, C_2 \) and \( C_3 \) that do not depend on \( h \), such that

\[
|d_i| \leq C_1 h^{-n}, \quad -r \leq i \leq s, \\
|e_i| \leq C_2, \quad 1 \leq i \leq m,
\]

and

\[
|L(z-x_j)^{d+1}| \leq C_3 h^{d+1-n}, \quad z_1 \leq z \leq z_m.
\]

With these estimates the second conclusion of the theorem follows immediately.

By means of two examples the necessity of the condition that \( E \) be nonzero will be illustrated.

**Example 2.2** Consider the case where \( n = 2, m = 2, r = 1, \) and \( s = 2 \).

Thus we are concerned with the construction of a finite difference approximation to the problem

\[
Ly \equiv y''(x) + a^1(x)y'(x) + a^0(x)y(x) = f(x)
\]

at the point \( x_j \), that involves the approximations \( u_{j-1}, u_j, u_{j+1}, u_{j+2} \) to \( y_{j-1}, y_j, y_{j+1}, y_{j+2} \). This is accomplished by the procedure of this section,
with two collocation points. For simplicity it is also assumed that the mesh is uniform. The basis functions \( w_i(x) \) are defined as in (2.13) and (2.14). In this case they can be written as

\[
\begin{align*}
  w^0(x) &= -(x-x_j)(x-x_{j+1})(x-x_{j+2}) / 6h^3, \\
  w^1(x) &= (x-x_{j-1})(x-x_j)(x-x_{j+1}) / 2h^3, \\
  w^2(x) &= -(x-x_{j-1})(x-x_{j+1})(x-x_{j+2}) / 2h^3, \\
  w^3(x) &= (x-x_{j-1})(x-x_{j})(x-x_{j+1}) / 6h^3, \\
  w^4(x) &= (x-x_{j-1})(x-x_{j})(x-x_{j+1})(x-x_{j+2}) / h^4.
\end{align*}
\]

Let the collocation points be defined by

\[
  z_1 = x_{j-1} + \xi_1 h \text{ and } z_2 = x_{j-1} + \xi_2 h.
\]

The normalizing factor \( E \) is in this example given by

\[
  E = L_0 w^4(z_2) - L_0 w^4(z_1) = (\xi_2 - \xi_1)(\xi_2 + \xi_1 - 3) / h^2.
\]

The collocation points \( z_1 \) are always assumed to be distinct, so \( \xi_2 \neq \xi_1 \). Hence \( E = 0 \) if and only if \( \xi_1 + \xi_2 = 3 \), i.e., when the points \( z_1 \) and \( z_2 \) are placed symmetrically about the midpoint \((x_j + x_{j+1})/2\), which would seem the most natural placement of these points. One can not alleviate the problem by redefining the normalizing factor \( E \). For example, with \( E = 1/h^2 \), \( \xi_1 = 1 \) and \( \xi_2 = 2 \) the formula

\[
  \bar{d}_i = h^2 \left| \begin{array}{cc}
  L_0 w^{1+i}(z_1) & L_0 w^4(z_1) \\
  L_0 w^{1+i}(z_2) & L_0 w^4(z_2)
  \end{array} \right|, \quad -1 \leq i \leq 2,
\]
gives $d_{-1} = -2/h^2$, $d_0 = 6/h^2$, $d_1 = -6/h^2$ and $d_2 = 2/h^2$. Further, $e_1 = h^2 L_0 w^4(z_2) = -2$ and $e_2 = -h^2 L_0 w^4(z_1) = 2$, so that for the problem $y'''(x) = f(x)$ the difference equation becomes

$$(-2 u_{j-1} + 6 u_j - 6 u_{j+1} + 2 u_{j+2}) / h^2 = 2f(x_{j+1}) - 2f(x_j).$$

Dividing by $2h$ one obtains

$$(-u_{j-1} + 3 u_j - 3 u_{j+1} + u_{j+2}) / h^3 = \left( f(x_{j+1}) - f(x_j) \right) / h,$$

which is consistent with $y'''(x) = f'(x)$, rather than $y''(x) = f(x)$.

**Example 2.3** As in the previous example let $n = 2$, $r = 1$ and $s = 2$, but take $m = 3$. So there are three collocation points. In addition to the basis functions $w^i(x)$, $(0 \leq i \leq 4)$, in example (2.2) take

$$w^5(x) = (x-x_{j-1})(x-x_j)(x-x_{j+1})(x-x_{j+1})(x-x_{j+2}) / h^5,$$

where $x_{j+1/2} = \frac{1}{2} (x_j + x_{j+1})$. Then the polynomials $w^i(x)$, $(0 \leq i \leq 5)$, form a basis of $P_5$. Now $E$ is given by

$$E = \begin{bmatrix}
L_0 w^4(z_2) & L_0 w^5(z_2) \\
L_0 w^4(z_3) & L_0 w^5(z_3)
\end{bmatrix} + \begin{bmatrix}
L_0 w^4(z_1) & L_0 w^5(z_1) \\
L_0 w^4(z_3) & L_0 w^5(z_3)
\end{bmatrix} - \begin{bmatrix}
L_0 w^4(z_2) & L_0 w^5(z_2) \\
L_0 w^4(z_2) & L_0 w^5(z_2)
\end{bmatrix}.$$

If the mesh is taken to be uniform and if the collocation points are distributed symmetrically, i.e., $z_1 = x_{j+1/2} - 3\xi h$, $z_2 = x_{j+1/2}$, $z_3 = x_{j+1/2} + 3\xi h$, then after some computation it is found that $E = c(4\xi^2 - 3)/h^5$, where $c$ is
some constant. Since the points \( z_i \) are distinct, \( \xi \) cannot be equal to zero. Hence if the points \( z_i \) are placed symmetrically, then \( E = 0 \) if and only if \( \xi = \frac{1}{2} \sqrt{3} \approx 0.866 \).

**Remark.** A necessary condition for the normalizing factor \( E \) to be non-zero is that \( r + s \geq n \). For example, the width of a difference approximation to a second order differential equation is at least three. This should not be surprising. To prove the assertion suppose \( r + s < n \). Then \( d^n w_i(x) / dx^n = 0 \), \( (0 \leq i \leq r + s) \), because the first \( r + s + 1 \) basis functions have degree \( r + s \). Hence the first column in the determinant defining \( \overline{d}_i \) consists of zeros for each \( i \). Thus \( \overline{d}_i = 0 \), \( (-r \leq i \leq s) \), and from the proof of theorem (2.1) it follows that \( E = 0 \).

Some sufficient conditions for \( E \) to be nonzero are given in the following theorem.

**Theorem 2.4** The normalizing factor \( E \) is nonzero if \( m = 1 \) and \( r + s \geq n \), or if \( r + s = n \). (i.e., when the difference equation is compact.)

**Proof.** If \( m = 1 \) then \( E = -1 \). If \( r + s = n \), \( m > 1 \) and \( E = 0 \), then it follows from (2.12) that there are constants \( c_{r+s} \), \( (r + s \leq i \leq d) \), not all zero, such that the polynomial \( q(x) \equiv \left( \sum_{i=r+s+1}^{d} c_i w_i(x) \right) \in P_d \) satisfies \( q^{(n)}(z_i) = c_{r+s}, (1 \leq i \leq m) \). Thus the polynomial \( \tilde{q}(x) \equiv q(x) - c_{r+s} x^n / n! \) satisfies \( \tilde{q}^{(n)}(z_i) = 0 \). But this is impossible since \( \tilde{q}^{(n)}(x) \in P_{m-1} \) because \( r + s = n \).

**Remark.** If the approximation is compact then for each \( n \) the coefficients \( \overline{d}_i \) are uniquely determined, independent of the number of collocation points. For example, if \( n = 2 \) and \( r + s = 2 \) then \( \overline{d}_{-1} = \frac{2}{h_j(h_j + h_j+1)} \).
The second derivative is always given by

\[
\begin{align*}
\frac{u_{j+1} - u_j}{h_{j+1}} - \frac{u_j - u_{j-1}}{h_j} = \frac{1}{2(h_j + h_{j+1})}.
\end{align*}
\]

### 3. EFFECT OF THE CHOICE OF COLLOCATION POINTS

It is well known, (see, e.g., Russell and Varah (1975)), that collocation procedures with certain piecewise polynomials for the numerical solution of boundary value problems have a higher order of accuracy if Gaussian points are used as collocation points. One will expect, that a similar statement applies to the finite difference equations discussed in the previous section. However, it may be surprising that the points \( z^*_i \) for which an extra high order of consistency is attained, do not generally coincide with the Gaussian points. Nevertheless, these special points can be characterized relatively easily as will now be shown.

Consider again the basis functions \( w^i(x) \) of \( P_{r+s+m-1} \) defined by equations (2.13) and (2.14). This linearly independent set may be extended to form a basis of \( P_{r+s+m} \) by adding a polynomial \( w^{r+s+m}(x) \epsilon P_{r+s+m} \) that vanishes at the mesh points. Thus this polynomial has the form
\[ w^{r+s+m}(x) \equiv \sum_{k=0}^{m-1} \frac{(x-t_k)}{(t_m-t_k)} + \sum_{k=0}^{r+s} \frac{(x-x_j-r+k)}{(t_m-x_j-r+k)} , \]

with \( t_m \neq t_k \) (\( 1 \leq k \leq m-1 \)), and \( t_m \neq x_j \) (\( -r \leq j \leq s \)). Expansion of \( y(x) \) in terms of the polynomials \( w^i(x) \) gives

\[ y(x) = \sum_{k=0}^{r+s+m} c_k w^k(x) + O(h^{r+s+m+1}) . \]

Substitution of this expression into the truncation error yields

\[ \tau_j = \sum_{i=-r}^{s} d_i y_{j+1} - \sum_{i=1}^{m} e_i f(z_i) \]

\[ \tau_j = \sum_{k=0}^{r+s+m} c_k \left[ \sum_{i=-r}^{s} d_i w^k(x_{j+1}) - \sum_{i=1}^{m} e_i L w^k(z_i) \right] + O(h^{r+s+m-n+1}) \]

The quantity between square brackets equals zero for \( 0 \leq k \leq r+s+m-1 \). Moreover, \( w^{r+s+m}(x_{j+1}) = 0 \), (\( -r \leq i \leq s \)). Hence

\[ \tau_j = -c_{r+s+m} \sum_{i=1}^{m} e_i L w^{r+s+m}(z_i) + O(h^{r+s+m-n+1}) . \]

Estimates as in the proof of theorem (2.1) again reveal that \( \tau_j = O(h^{r+s+m-n}) \). However, it is also clear from the above expression, that an additional order of consistency is obtained, if the collocation points \( z_i \) coincide with points where \( L_0 w^{r+s+m}(x) = \frac{d^n w^{r+s+m}(x)}{dx^n} = 0 \). Thus the following has been shown:
Theorem 3.1. Let the coefficients $d_i$ and $e_i$ and the normalizing factor $E$ be given by (2.10), (2.11) and (2.12) respectively. Assume that $E \neq 0$ and that $r + s \geq n$. Let $\tilde{w}^{r+s+m}(x) = \sum_{k=0}^{m-1} \prod_{j=1}^{r+s} \frac{x-x_j-r+k}{t_k}$ and assume that $L_0 \tilde{w}^{r+s+m}(z_i) = 0, (1 \leq i \leq m)$. Then the order of consistency of the finite difference approximation (2.4) is at least equal to $r + s + m - n + 1$.

If $m = 1$ and the difference approximation is compact, i.e., $r + s = n$, then $\tilde{w}^{r+s+m}(x) = \prod_{k=0}^{n} \frac{x-x_k-r+k}{t_k}$. In this case there is only one collocation point for which a higher order of consistency is obtained. If $m = 1$ and $r + s > n$ then there are $r + s + 1 - n$ possible choices of this point. (Assuming $E \neq 0$.) For the case $m > 1$, $r + s = n$ there is a $(m-1)$-parameter family of points $z_i$, $(1 \leq i \leq m)$, for which the improved order of consistency is obtained. The parameters in question are the points $t_k$ in the definition of $\tilde{w}^{r+s+m}(x)$ in theorem (3.1). Particular examples of these special collocation points will be included in the next section.

4. EXAMPLES OF FINITE DIFFERENCE APPROXIMATIONS

Example 4.1. Let $n = 1$, $r = 0$ and $s = 1$; i.e., we want to construct a two point finite difference approximation to the equation

\begin{equation}
Ly(x) = y'(x) + a^0(x)y(x) = f(x).
\end{equation}

Take $m = 1$ and $z_1 = x_{j+\frac{1}{2}} = \frac{1}{2}(x_j + x_{j+1})$. Then the basis functions $w_i(x)$ as defined by (2.13) and (2.14) are $w^0(x) = -(x-x_{j+1}) / h_j$, and $w^1(x) = (x-x_j) / h_j$. The difference approximation has the form

\[ d_0 u_j + d_1 u_{j+1} = f(x_{j+\frac{1}{2}}). \]
where the coefficients $d_0$, $d_1$ and $e_1$ can be obtained from (2.10) and (2.11), viz.

$$d_0 = Lw^0(z_1) = \frac{-1}{h_j} + \frac{1}{2} a_j^{0 \frac{1}{2}} ,$$

and

$$d_1 = Lw^1(z_1) = \frac{1}{h_j} + \frac{1}{2} a_j^{0 \frac{1}{2}} .$$

The difference approximation can then be written as

$$(4.2) \quad \frac{(u_{j+1} - u_j)}{h_j} + \frac{1}{2} a_j^{0 \frac{1}{2}} (u_j + u_{j+1}) = f_{j+\frac{1}{2}} ,$$

which is recognized as the centered Euler equation, or Box scheme of Keller (1969). According to theorem (2.1) the order of consistency is at least equal to one. That the order of consistency actually equals two follows from theorem (3.1), because $z_1$ is the critical point of $\tilde{w}^2(x) \equiv (x-x_j)(x-x_{j+1}).$

An equally simple difference equation, the trapezoidal method, is derived when taking $m = 2$, with $z_1 = x_j$ and $z_2 = x_{j+1}$. This equation has the form $d_0 u_j + d_1 u_{j+1} = e_1 f_j + e_2 f_{j+1}.$

With $w^2(x) = (x-x_j)(x-x_{j+1}) / h_j^2$ the coefficients are

$$d_0 = \frac{1}{E} \begin{vmatrix} Lw^0(z_1) & Lw^2(z_1) \\ Lw^0(z_2) & Lw^2(z_2) \end{vmatrix} , \quad d_1 = \frac{1}{E} \begin{vmatrix} Lw^1(z_1) & Lw^2(z_1) \\ Lw^1(z_2) & Lw^2(z_2) \end{vmatrix} ,$$
\[ e_1 = \frac{1}{E} \text{Lw}^2(z_2) \text{ and } e_2 = -\frac{1}{E} \text{Lw}^2(z_1), \text{ with } E = \text{L}_0 \text{w}^2(z_1) - \text{L}_0 \text{w}^2(z_2) \text{ and } \text{L}_0 y(x) = y'(x). \text{ The equation thus obtained is} \]

\[ \left( u_{j+1} - u_j \right) / h_j + \frac{1}{2} (a_{0j+1}^0 + a_{0j}^0) u_{j+1} + \frac{1}{2} (f_j + f_{j+1}). \]

By theorem (2.1) the order of consistency of this difference equation is equal to two.

Next, consider the general two point difference approximation with \( m = 2 \). That such a difference approximation is always consistent follows from theorem (2.4). Theorem (2.1) implies that the order of consistency is equal to two. From theorem (3.1) it follows that the order becomes three if the collocation points coincide with the critical points of \( \tilde{w}^3(x) = (x-x_1)(x-x_2)(x-x_{j+1}). \) Let \( t_j = x_j + \eta h_j, z_1 = x_j + \xi_1 h_j \) and \( z_2 = x_j + \xi_2 h_j. \) Then the special values of \( \xi_1 \) and \( \xi_2 \) that give a third order formula are the roots of \( p'(\xi) = 0, \) where \( p(\xi) = \xi(\xi-1)(\xi-\eta). \) These roots are

\[ \xi = \frac{n+1 \pm \sqrt{n^2 - 6n - 2}}{3}. \]

If \( \eta = \frac{1}{2} \) then the roots are \( \xi = \frac{1}{2} \pm \frac{1}{6} \sqrt{3} \). In that case the collocation points are placed symmetrically in the interval \([x_j, x_{j+1}]\). This results in another order of consistency being gained, so that the actual order of consistency is then equal to four. The coefficients of this approximation are

\[ d_0 = -\frac{1}{h_j} + \frac{1}{4} (a_0^0(z_1) + a_0^0(z_2)) - \frac{1}{12} h_j a_0^0(z_1) a_0^0(z_2), \]

\[ d_1 = \frac{1}{h_j} + \frac{1}{4} (a_0^1(z_1) + a_0^1(z_2)) + \frac{1}{12} h_j a_0^0(z_1) a_0^0(z_2). \]
\[ e_1 = \frac{1}{2} - h_j \frac{\sqrt{6}}{12} a_0(z_2) \text{ and } e_2 = \frac{1}{2} + h_j \frac{\sqrt{6}}{12} a_0(z_1). \]

It should be pointed out here, that the collocation points in this difference approximation are the Gaussian points with respect to the interval \([x_j, x_{j+1}]\). The procedure of Section 2 with \(m\) Gaussian points will yield a formula of order \(2m\). This is higher than predicted by theorem (3.1). However, this result does not generalize to higher order equations. The corresponding difference equations are equivalent to generalized \(m\)-stage implicit Runge-Kutta processes. (See Butcher (1964), Wright (1970).) It is also equivalent to a collocation procedure with \(C^0\) splines of degree \(m\) (Weiss (1974), Hulme (1971).)

Example 4.2. With \(r = 0\), \(s = 2\), \(m = 1\), \(z_1 = x_{j+1}\) and \(h_{j+1} = h_{j+2} = h\) the procedure of Section 2 gives a three point (two step) equation of the form

\[ d_0 u_j + d_1 u_{j+1} + d_2 u_{j+2} = f(x_{j+2}), \]

where \(d_i = Lw^i(z_1), \quad 0 \leq i \leq 2\).

The difference approximation to equation (4.1) is then found to be

\[ \frac{1}{2h} \left\{ u_j - 4u_{j+1} + 3u_{j+2} \right\} + a_0^0 u_{j+2} = f_{j+2}, \]

which is one of Gear's formulas, specially suited for stiff equations. (See Gear (1971), p. 217.) Higher order Gear's methods are obtained with \(r = 0\), \(s > 2\), \(m = 1\) and \(z_1 = x_{j+s}\). The order of consistency is equal to \(s\).

Example 4.3. Let \(n = 2\) and \(r = s = 1\), i.e., we want to construct a compact
finite difference approximation to the equation

(4.3) \[ Ly(x) = y''(x) + a^1(x)y'(x) + a^0(x)y(x) = f(x). \]

With \( m = 1 \) and \( z_1 = x_j \), the coefficients in the difference equation

\[ d_{-1} u_{j-1} + d_0 u_j + d_1 u_{j+1} = f_j, \]

are given by

\[ d_{-1} = Lw^0(x_j) = \frac{2-h_{j+1}}{h_j(h_j+h_{j+1})} a^1_j, \]
\[ d_0 = Lw^1(x_j) = \frac{-2(h_{j+1}-h_j)a^1_j}{h_jh_{j+1}} + a^0_j, \]
\[ d_1 = Lw^2(x_j) = \frac{2+h_ja^1_j}{h_{j+1}(h_j+h_{j+1})}. \]

From theorem (2.1) it follows that the order of consistency of this difference approximation is equal to one. The order of consistency becomes two if one lets \( z_1 \) coincide with the inflection point of \( \tilde{w}^3(x) = (x-x_{j-1})(x-x_j)(x-x_{j+1}) \). This point is given by \( z_1 = x_j + (h_{j+1}-h_j)/3 \). The coefficients \( d_i \) of the corresponding finite difference equation are given by

\[ d_{-1} = \frac{1}{h_j(h_j+h_{j+1})} \left[ 2 - \frac{(2h_j+h_{j+1})}{3} a^1(z_1) - \frac{(h_{j+1}-h_j)(2h_j+h_{j+1})}{9} a^0(z_1) \right], \]
\[ d_0 = \frac{1}{h_jh_{j+1}} \left[ -2 + \frac{(h_{j+1}-h_j)}{3} a^1(z) + \frac{2(h_j+h_{j+1})(h_j+2h_{j+1})}{9} a^0(z_1) \right], \]
$$d_{i+1} = \frac{1}{h_{j+1}(h_j+h_{j+1})} \left[ 2^+ \frac{(h_i+2h_{j+1})}{3} a^1(z_1) + \frac{(2h_i+h_{j+1})(h_{j+1}-h_i)}{9} a^0(z_1) \right].$$

Example 4.4. A third order three point formula for equation (4.3) is obtained with $m=3$, $z_1=x_{j-1}$, $z_2=x_j$ and $z_3=x_{j+1}$. The coefficients of the difference equation are somewhat complicated in this case. However, if the differential equation is given by

$$(4.4) \quad Ly(x) = y''(x) + a_0(x)y(x) = f(x),$$

then the resulting finite difference equation is identical to the "Mehrenstettenverfahren" mentioned in Section 1. (Equation (1.7).) There the mesh is assumed to be uniform. If this is not the case then the difference equation becomes

$$(4.5) \quad \frac{u_{j+1}-u_j}{h_{j+1}} - \frac{u_j-u_{j-1}}{h_j} + \frac{1}{2}(h_j+h_{j+1}) W_h(a_j u_j) = W_h(f_j),$$

where

$$W_h w_j = a_j w_{j-1} + (1-a_j-b_j) w_j + b_j w_{j+1},$$

with

$$a_j = \frac{1}{6} - \frac{h_{j+1}^2}{6h_j(h_j+h_{j+1})},$$

and
\[
\begin{align*}
\beta_j &= \frac{1}{6} - \frac{h_j^2}{6h_{j+1}^2(h_j + h_{j+1})}.
\end{align*}
\]

For a uniform mesh the order of consistency is equal to four.

**Example 4.5.** In this example, a number of choices of the collocation points \( z_i \) will be given, for which the order of consistency of the corresponding finite difference equations is greater than predicted by theorem (2.1). Again the differential equation under consideration is given by (4.3) and the finite difference approximations are assumed to be compact with \( r = s = 1 \). The mesh is taken uniform.

First let \( m = 2 \). This gives a difference approximation of order two. Let \( z_1 = x_j + \xi_1 h \), \( z_2 = x_j + \xi_2 h \) and \( t_1 = x_j + \eta h \). From theorem (3.1) it follows that the order becomes three if the values of \( \xi_1 \) and \( \xi_2 \) are taken to be the roots of \( p''(\xi) = 0 \), where \( p(\xi) = (\xi-\eta)(\xi-1)(\xi(\xi+1)) \). The collocation points will be symmetrically placed if one lets \( \eta = 0 \). This yields \( z_1 = x_j - \frac{h}{\sqrt{6}} \) and \( z_2 = x_j + \frac{h}{\sqrt{6}} \). Because of symmetry and the uniform mesh the order of this equation actually equals four.

Next, consider the case \( m = 4 \). This will in general lead to a fourth order formula. The order becomes five if the collocation points are the inflection points of \( w^6(x) = (x-t_1)(x-t_2)(x-t_3)(x-x_j-1)(x-x_j)(x-x_{j+1}) \). Now require the collocation points to have the form

\[
\begin{align*}
z_1 &= x_{j-\frac{1}{2}} - \xi h, & z_2 &= x_{j-\frac{1}{2}} + \xi h, \\
\qquad \quad \quad z_3 &= x_{j+\frac{1}{2}} - \xi h, & z_4 &= x_{j+\frac{1}{2}} + \xi h.
\end{align*}
\]
This choice has the advantage that the collocation points of consecutive difference equations partly coincide, thereby limiting the necessary total number of function evaluations. Upon some computation it is found that these points are defined by the value

\[ \xi = \sqrt{\frac{5}{12} - \frac{1}{2} \sqrt{\frac{23}{45}}} \approx 0.24 \quad \text{or} \quad \xi = \sqrt{\frac{5}{12} + \frac{1}{2} \sqrt{\frac{23}{45}}} \approx 0.88. \]

So with either value of \( \xi \) and with \( z_i \) as above one obtains a finite difference approximation for which the order of consistency is at least equal to five according to theorem (3.1). The actual order equals six because of symmetry and the uniform mesh.

**Example 4.6.** In this example some five point finite difference approximations to (4.3) will be discussed. So let \( r = s = 2 \). First take \( m = 1 \), with \( z_i = x_j \). The order of consistency of the corresponding equation is equal to three. This follows from theorem (2.1). If the mesh is uniform then this equation is identical to (2.5). By theorem (3.1) the order is then equal to four.

If \( m = 5 \) and if the collocation points \( z_i \), \( (1 \leq i \leq 5) \), coincide with the meshpoints, then the order of the resulting equation is equal to seven if the mesh is not uniform and equal to eight if the mesh is uniform. Of course the coefficients are very complicated in this case. However, if the differential equation does not involve the \( a^1(x)y'(x) \) term, i.e., when the equation is given by (4.4), then the finite difference approximation is the same as the five point "Mehrstellenverfahren" of Collatz (1960), p. 502, and can be expressed as

\[
(31u_{j-2} + 128u_{j-1} - 318u_j + 128u_{j+1} + 31u_{j+2})/(252h^2) + W_h(a_0u_j) = W_h f_j,
\]
where

\[ W_h w_j = \frac{(23w_{j-2} + 688w_{j-1} + 2358w_j + 688w_{j+1} + 23w_{j+2})}{3780}. \]

As in example (4.5) one may wish to determine the special points \( z_i \), for which the finite difference equation has an extra high order of consistency. If \( m = 2 \) and if the mesh is uniform, then a computation similar to those in the previous example shows that the order is equal to six if one lets \( z_1 = x_j - \xi h, \ z_2 = x_j + \xi h \), and if the value of \( \xi \) is either

\[ \sqrt{1 - \sqrt{\frac{11}{15}}} \approx 0.38 \quad \text{or} \quad \sqrt{1 + \sqrt{\frac{11}{15}}} \approx 1.36. \]

5. **FINITE DIFFERENCE APPROXIMATIONS TO BOUNDARY CONDITIONS**

To define a complete difference scheme, approximations to the boundary conditions are required. Let a boundary condition be given by

\[(5.1) \quad B y(0) \equiv y^{(k)}(0) + \sum_{i=0}^{k-1} b_i y^{(i)}(0) = b, \]

with \( k < n \). To construct a finite difference approximation to (5.1) one may proceed in a fashion that resembles the construction of difference equations in Section 2.

Let \( d = s + m \) and assume that a polynomial \( p(x) \in \mathbb{P}_d \) is uniquely determined by the equations

\[(5.2) \quad p(x_i) = u_i, \quad 0 \leq i \leq s, \]

\[(5.3) \quad L p(z_i) = f(z_i), \quad 1 \leq i \leq m. \]
Then the equation

\begin{equation}
B_p(0) = b
\end{equation}

generates a relation of the form

\begin{equation}
B_h u_0 \equiv \sum_{i=0}^{s} d_i u_i = e_0 b + \sum_{i=1}^{m} e_i f(z_i) .
\end{equation}

If the basis functions \( w^i(x) \), \( 0 \leq i \leq d \), of \( P_{d} \) are chosen as in Section 2, (Equations (2.13) and (2.14) with \( r = 0 \)), then one can show that the coefficients can be represented by the following determinants:

\begin{equation}
d_i = \frac{1}{E_0} \begin{vmatrix}
Bw^i(0) & Bw^{s+1}(0) & \cdots & Bw^d(0) \\
Lw^i(z_1) & Lw^{s+1}(z_1) & \cdots & Lw^d(z_1) \\
\vdots & \vdots & \ddots & \vdots \\
Lw^i(z_m) & Lw^{s+1}(z_m) & \cdots & Lw^d(z_m)
\end{vmatrix}, \quad 0 \leq i \leq s ,
\end{equation}

\begin{equation}
e_0 = \frac{1}{E_0} \begin{vmatrix}
Lw^{s+1}(z_1) & \cdots & Lw^d(z_1) \\
\vdots & \ddots & \vdots \\
Lw^{s+1}(z_m) & \cdots & Lw^d(z_m)
\end{vmatrix} ,
\end{equation}
It is convenient to define the normalizing factor $E_0$ by

$$E_0 \equiv (-1)^{m+1}$$

where $L_0 y(x) = y^{(n)}(x)$. If $m = 0$ define $E_0 = 1$. The proofs of the following theorems closely follow those given in Section 2.

**Theorem 5.1** Assume that $E_0 \neq 0$. If $m = 0$ let $s \geq k$; otherwise if $m > 0$, let $s + m \geq n$. Then at least $k + 1$ of the coefficients $d_i$ are non-zero for all small enough $h$, and the truncation error

$$\tau_0 \equiv B_h y(0) - e_0 b - \sum_{i=1}^{m} e_i f(z_i)$$

of the difference approximation (5.5) satisfies

$$|\tau_0| \leq Ch^{s+m-k+1},$$
i.e., the order of consistency of (5.5) is at least equal to \( s + m - k + 1 \).

**Proof.** Consider the equations

\[
\begin{align*}
\sum_{i=0}^{s} d_i w^l(x_i) - \sum_{i=1}^{m} e_i Lw^l(z_i) &= e_0 Bw^l(x_0), & 0 \leq l \leq d.
\end{align*}
\]

Here the polynomials \( w^l(x) \) are the same as those used in the definition of the coefficients. These equations can also be written as

\[
\begin{bmatrix}
-I & -A \\
0 & -C
\end{bmatrix}
\begin{bmatrix}
d_0 \\
\vdots \\
d_s \\
e_1 \\
\vdots \\
e_m
\end{bmatrix}
= e_0
\begin{bmatrix}
Bw^0(x_0) \\
\vdots \\
Bw^d(x_0)
\end{bmatrix},
\]

where \( I \) is the \((s+1) \times (s+1)\) identity matrix,

\[
A =
\begin{bmatrix}
Lw^0(z_1) & \ldots & Lw^0(z_m) \\
\vdots & \ddots & \vdots \\
Lw^s(z_1) & \ldots & Lw^s(z_m)
\end{bmatrix},
\]

and

\[
C =
\begin{bmatrix}
Lw^{s+1}(z_1) & \ldots & Lw^{s+1}(z_m) \\
\vdots & \ddots & \vdots \\
Lw^d(z_1) & \ldots & Lw^d(z_m)
\end{bmatrix}.
\]
Thus the equations (5.10) admit a unique solution only if \( \text{det}(C) \neq 0 \). In particular, the assumption that \( E_0 \neq 0 \) implies that \( \text{det}(C) \neq 0 \), provided that \( h \) is small enough.

Next let \( \overline{d}_i \), \((0 \leq i \leq s)\), be defined as the coefficients \( d_i \) in (5.6), but with \( L_0 \) and \( B_0 \) replacing \( L \) and \( B \). Here \( B_0 y(0) \equiv y^{(k)}(0) \). Assume that less than \( k+1 \) of the coefficients \( \overline{d}_i \) are nonzero. More precisely suppose that \( \overline{d}_{v} \neq 0 \), \((1 \leq v \leq k < k+1)\), and \( \overline{d}_i = 0 \) otherwise. Let

\[
q(x) = \prod_{\nu=1}^{k} (x-x_{\nu}) q_1(x), \quad \text{where } q_1(x) \in P_{k-k_1} \text{ has leading coefficient } 1/k!.
\]

Then \( q(x) \in P_k \), and by assumption \( k < d \). Therefore

\[
\sum_{i=0}^{s} \overline{d}_i q(x_i) - \sum_{i=1}^{m} \overline{e}_i L_0 q(z_i) = \overline{e}_0 B_0 q(x_0).
\]

The construction of \( q(x) \) makes the entire lefthand side vanish. Also, \( B_0 q(x_0) = 1 \). Hence \( \overline{e}_0 = 0 \). But this is a contradiction since \( \overline{e}_0 = (-1)^{m+1} \).

Thus if \( E_0 \neq 0 \) then at least \( k+1 \) of the coefficients \( \overline{d}_i \) are nonzero. If \( h \) is small the same can be said of the coefficients \( d_i \). This proves the first statement of the theorem. The proof of the second assertion closely follows that of part of theorem (2.1) and will be omitted.

**Theorem 5.2.** The normalizing factor is nonzero if \( m=0 \) or if \( s=k \). (In the latter case the difference equation is said to be compact.)

**Proof.** As already stated, one takes \( E_0 = 1 \) if \( m=0 \). If \( s=k \) and \( E_0 = 0 \), then it follows from (5.9) that there are constants \( c_i \), \((s+1 \leq i \leq d)\), such that the polynomial \( q(x) = \sum_{i=s+1}^{s+m} c_i w^i(x) \) satisfies \( L_0 q(z_i) = 0 \), \((1 \leq i \leq m)\). But \( L_0 q(x) \in P_{s+m-n} \) and \( s+m-n < m \) because \( s=k < n \). Hence \( L_0 q(x) \equiv q^{(n)}(x) = 0 \). So \( q(x) \equiv 0 \). This however is impossible, since not all \( c_i \) are equal.
to zero and since the polynomials \( w_i(x) \) are linearly independent.

As was the case for finite difference approximations to the differential equation, it is possible to identify collocation points \( z_i \), for which the approximation to the boundary condition attains a higher order of consistency, than predicted by theorem (5.1). A minor modification of the proof of theorem (3.1) shows the following.

**Theorem 5.3.** Let the coefficients \( d_i \) and \( e_i \) and the normalizing factor \( E_0 \) be given by equations (5.6), (5.7), (5.8) and (5.9). Assume that \( E_0 \neq 0 \) and that \( s + m \geq n \). Suppose \( \tilde{w}^{s+m+1}(x) \in P_{s+m+1} \) satisfies

\[
\tilde{w}^{s+m+1}(x_i) = 0, \quad 0 \leq i \leq s,
\]

and

\[
\frac{d^k \tilde{w}^{s+m+1}(0)}{dx^k} = 0.
\]

Also assume that the collocation points are chosen such that \( \tilde{L}_0 \tilde{w}^{s+m+1}(z_i) = 0, \quad (1 \leq i \leq m) \). Then the order of consistency of the finite difference approximation (5.5) is at least equal to \( s + m - k + 2 \).

6. **EXAMPLES OF APPROXIMATIONS TO BOUNDARY CONDITIONS**

**Example 6.1.** Let \( n = 2 \) and \( k = 1 \). Thus the differential equation is given by

\[
(6.1) \quad Ly(x) \equiv y''(x) + a^1(x)y'(x) + a^0(x)y(x) = f(x),
\]

and the boundary condition by
Let $m = 0$, so that the finite difference approximation to (6.2) does not involve the differential equation (6.1). With $s = 1$ the construction procedure of the previous section yields $e_0 = E_0 = 1$,

\[ d_0 = Bw^0(0) = B\left[-\frac{(x-x_1)}{h}\right] = \frac{-1}{h} + b_0, \]

and

\[ d_1 = Bw^1(0) = B\left[\frac{(x-x_0)}{h}\right] = \frac{1}{h}. \]

The approximation is therefore

\[ \frac{u_1-u_0}{h} + b_0 u_0 = b, \]

which has order of consistency equal to one.

If $s = 2$ and if the mesh is uniform, then one obtains an equally well-known 3-point equation, viz.

\[ \frac{u_2-4u_1+3u_0}{2h} + b_0 u_0 = b, \]

which is a second order formula.

Example 6.2. Now let $m = 1$ and $s = 1$, so that the approximation to (6.2) is compact and therefore consistent for any choice of $z_1$. Let $w^0(x) = -(x-x_1)/h$, $w^1(x) = (x-x_0)/h$ and $w^2(x) = (x-x_0)(x-x_1)/h^2$. Then the
normalizing factor $E_0$ is given by $E_0 = L_0 w^2(z_1) = \frac{2}{h^2}$. The coefficients of the difference approximation

\begin{equation}
(6.3) \quad d_0 u_0 + d_1 u_1 = e_0 b + e_1 f(z_1),
\end{equation}

are then equal to

\[
\begin{align*}
e_0 &= \frac{1}{E_0} Lw^2(z_1) = 1 + \frac{h}{2} a^1(z_1), \\
e_1 &= -\frac{1}{E_0} Bw^2(0) = \frac{h}{2}, \\
\end{align*}
\]

\[
\begin{vmatrix}
Bw^0(0) & Bw^2(0) \\
Lw^0(z_1) & Lw^2(z_1)
\end{vmatrix}
= \frac{-1}{h} a^1(z_1) + b_0 + \frac{h}{2} b_0 a^0(z_1),
\]

and

\[
\begin{vmatrix}
Bw^1(0) & Bw^2(0) \\
Lw^1(z_1) & Lw^2(z_1)
\end{vmatrix}
= \frac{1}{h} a^1(z_1) + \frac{h}{2} a^0(z_1).
\]

The order of consistency is equal to two. It is now of interest to determine for which choice of the collocation point $z_1$ the order increases to three. From theorem (5.3) it follows that this happens if $z_1$ is the inflection point of $\tilde{w}^3(x) = x^2(x-x_1)$. This point is $z_1 = \frac{x_1}{3} = \frac{h_1}{3}$.

**Example 6.3.** Let $s = 2$ and $m = 1$, so that the difference approximation to (6.2) has the form
Assume that the mesh is uniform and let \( z_1 = x_0 + \xi h \). Then \( E_0 = L_0 w^3(x_0 + \xi h) \) and \( E_0 = 0 \) if and only if \( \xi = 1 \). Hence one should not let \( z_1 = x_1 \). The order of consistency increases to four if \( z_1 \) is an inflection point of \( w^4(x) = x^2(x-x_1)(x-x_2) \). This happens if \( \xi = (9 - \sqrt{33})/12 \approx 0.27 \) or \( \xi = (9 + \sqrt{33})/12 \approx 1.23 \).

7. THE STABILITY OF FINITE DIFFERENCE SCHEMES.

In this section the stability theory of Kreiss (1972) will be used to investigate the stability of finite difference schemes based upon consistent finite difference approximations introduced in previous sections. For this purpose the mesh is assumed to be uniform, i.e., \( h_j = h = 1/J, \ 1 \leq j \leq J \).

A difference scheme consists of equations of the form

\[
(7.1) \quad L_h u_j = \sum_{i=-r}^{s} d_{j,i} u_{j+i} = \sum_{i=1}^{m} e_{j,i} f(z_{j,i}) \equiv \tilde{f}_j, \quad r \leq j \leq J-s,
\]

together with boundary conditions

\[
(7.2) \quad B_h, k(0)u_0 = \sum_{i=0}^{s_k(0)} d_{k,i}(0) u_j = e_{k,0}(0) b_k(0) + \sum_{i=1}^{m_k(0)} e_{k,i}(0) f(z_{k,i}(0)) \equiv \tilde{b}_k(0), \quad 1 \leq k \leq n_0.
\]

\[
(7.3) \quad B_h, k(1)u_J = \sum_{i=-r_k(1)}^{0} d_{k,i}(1) u_{j+i} = e_{k,0}(1) b_k(1) + \sum_{i=1}^{m_k(1)} e_{k,i}(1) f(z_{k,i}(1)) \equiv \tilde{b}_k(1), \quad 1 \leq k \leq n - n_0.
\]
In addition, if (7.1) is not compact, i.e., if \( r + s > n \) then \( r + s - n \) extra boundary conditions are required in order to match the number of equations and the number of unknowns. Although this is not necessary, it will be assumed that these extra equations are also consistent with the differential equation and given by

\[
(7.4) \quad L_h u_j = \sum_{i=0}^{s_j} d_{j,i} u_i + \sum_{i=1}^{m_j} e_{j,i} f(z_{j,i}) \equiv \tilde{f}_j, \quad k_0 \leq j \leq r-1, \quad k_0 \geq 0,
\]

\[
(7.5) \quad L_h u_j = \sum_{i=-r_j}^{0} d_{j,i} u_{j+i} + \sum_{i=1}^{m_j} e_{j,i} f(z_{j,i}) \equiv \tilde{f}_j, \quad J-s+1 \leq j \leq J-n+k_0.
\]

Let \( u_h = (u_0^T, \ldots, u_J^T) \). Then the equations (7.1) through (7.5) can be compactly written as

\[
(7.6) \quad L_h u_h = f_h.
\]

Here \( f_h \in \mathbb{R}^{J+1} \) is the appropriate right hand side vector and \( L_h \) is a \((J+1) \times (J+1)\) matrix. Consistency of the equations (7.1), (7.4) and (7.5) is easily seen to imply that \( |\tilde{f}_j - f(x_j)| \leq c h \). Also, let \( e_h \in \mathbb{R}^{J+1} \) be the error vector, i.e., \( e_h = (e_0^T, \ldots, e_J^T) \), with \( e_j = y_j - u_j \), and let \( \tau_h \) be the vector of truncation errors, \( \tau_h = (\tau_{1(0)}, \ldots, \tau_{n_0(0)}, \tau_{k_0}, \ldots, \tau_{J-n+k_0}, \tau_{1(1)}, \ldots, \tau_{n-n_0(1)})^T \), with

\[
\tau_k^{(0)} = B_h, k(0) y_0 - \tilde{b}_k(0), \quad 1 \leq k \leq n_0,
\]

\[
\tau_j = L_h y_j - \tilde{f}_j, \quad k_0 \leq j \leq J - n + k_0,
\]
and

\[ \tau_k(1) = B_{h,k}(1) y_J - \tilde{b}_{k}(1), \quad 1 \leq k \leq n - n_0. \]

For \( w_h \in \mathbb{R}^{J+1} \) let \( ||w_h|| = \max_{0 \leq j \leq J} |w_j| \). If \( A_h \) is a \((J+1) \times (J+1)\) matrix then \( ||A_h|| \) is the induced matrix norm, i.e., \( ||A_h|| = \max_{w_h \neq 0} \frac{||A_h w_h|||}{||w_h||} \).

The finite difference scheme (7.6) is said to be consistent if \( ||\tau_h|| \leq k_1 h \) as \( h \to 0 \) and stable if \( L_h^{-1} \) exists for all small enough \( h \) and \( ||L_h^{-1}|| \leq k_2 \). Here \( k_1 \) and \( k_2 \) are constants that do not depend on \( h \).

The stability property is essentially determined by the difference approximation to the highest derivative. To be more specific, the notion of characteristic polynomial of (7.1) is needed. Let \( E \) and \( d_i \) be as in Section 2, (Equations (2.12) and (2.16)), and define

\[ \bar{c}(t) = h^n \sum_{i=-r}^{s} d_i t^{r+i}. \]

It is not difficult to show that the equation \( \bar{c}(t) = 0 \) must have a root \( t=1 \) of multiplicity \( n \), because of consistency. Hence one can write \( \bar{c}(t) = (t-1)^n c(t) \), where \( c(t) \) is defined to be the characteristic polynomial associated with the difference equation (7.1). Assume that \( c(t) \) is explicitly given by

\[ c(t) = \sum_{i=0}^{N} a_i t^i, \quad \text{with} \quad N = r + s - n. \]

If the finite difference scheme (7.6) is not compact, then one also has \( r + s - n \) characteristic polynomials associated with the extra boundary conditions (7.4) and (7.5). It is assumed that these have the form
(7.8) \[ c_j(t) = \sum_{i=0}^{N_j} a_{j,i} t^i, \quad k_0 \leq j \leq r - 1, \]
and
(7.9) \[ c_j(t) = \sum_{i=0}^{N_j} a_{j,i} t^i, \quad J - s + 1 \leq j \leq J - n + k_0. \]

Finally, consider the homogeneous difference equation
(7.10) \[ \sum_{i=0}^{N} a_i v_{j+i} = 0, \quad 0 \leq j < \infty, \]
with boundary conditions
(7.11) \[ \sum_{i=0}^{N_j} a_{j,i} v_i = 0, \quad k_0 \leq j \leq r - 1, \quad \sup_{0 \leq j \leq \infty} |v_j| \leq \text{const.}, \quad 0 \leq j < \infty, \]
and
(7.12) \[ \sum_{i=0}^{N} a_{N-i} v_{J-j-i} = 0, \quad 0 \leq j < \infty, \]
subject to
(7.13) \[ \sum_{i=0}^{N_j} a_{j,N-j} v_{j-1} = 0, \quad J - s + 1 \leq j \leq J - n + k_0, \quad \sup_{0 \leq j < \infty} |v_{j-1}| \leq \text{const.} \]

Then one can state the following theorem due to Kreiss (1972).

Theorem 7.1. Let the homogeneous problem corresponding to (1.1) and (1.1a, b) only admit the trivial solution. Assume that the difference scheme (7.6) is consistent and that all roots \( t_i \) of the characteristic equation \( c(t) = 0 \)
satisfy \(|t_1| \neq 1\).

Further, if the difference scheme is not compact then also assume that the difference equations (7.10) with boundary conditions (7.11) and the difference equations (7.12) with boundary conditions (7.13) only have the trivial solution.

Then (7.6) has a unique solution for all small enough \(h\) and the difference scheme is stable.

We now investigate the stability properties of the approximation (7.1). If equation (7.1) is compact (and consistent) then consistency of the boundary conditions (7.2) and (7.3) is sufficient to guarantee stability. If (7.1) is not compact then the characteristic polynomial \(c(t)\) said to be symmetric if \(c(t) = t^{r+s-n}c(1/t)\) and strictly diagonally dominant if the degree of \(c(t)\) is even and if \(\sum_{i=0}^{r+s-n} \left|a_i\right| > \sum_{i \neq \ell} \left|a_i\right|\). (Here \(\ell \equiv (r+s-n)/2\).)

To motivate the last definition, consider the general form of a five point formula that is consistent with the second derivative. This approximation can be written as \(K_h^2 D_h^2 u_j\) where \(D_h^2 u_j = (u_{j+1} - 2u_j + u_{j-1})/h^2\) and \(K_h w_j = a_0 w_{j-1} + a_1 w_j + a_2 w_{j+1}\). The characteristic equation is \(c(t) = a_0 + a_1 t + a_2 t^2\) and \(c(t)\) is diagonally dominant if and only if the operator (matrix) \(K_h\) is diagonally dominant.

Normally one will construct the difference approximations (7.1) such that \(c(t)\) is symmetric. (But the characteristic equations associated with the extra boundary condition (7.4) and (7.5) need not be.)

Lemma 7.2. Assume that \(c(t)\) is symmetric. If the degree of \(c(t)\) is odd then \(c(-1) = 0\).
If the degree of \( c(t) \) is even and if \( c(t) \) is strictly diagonally dominant with positive coefficients, then there are no roots of \( c(t) = 0 \) on the unit circle.

Proof. The proof of the first statement is immediate. If the degree of \( c(t) \) equals \( 2N_1 \), then

\[
|c(e^{i\theta})| = |\sum_{k=0}^{2N_1} a_k e^{ik\theta}| = |(a_{N_1} + \sum_{k=0}^{N_1-1} a_k (e^{ik\theta} + e^{-ik\theta})) e^{iN_1\theta}|
\]

\[
= |a_{N_1} + 2 \sum_{k=0}^{N_1-1} a_k \cos k\theta| \geq a_{N_1} - 2 |\sum_{k=0}^{N_1-1} a_k \cos k\theta| \geq a_{N_1} - 2 |\sum_{k=0}^{N_1-1} a_k| > 0.
\]

Example 7.3. In the special case where \( c(t) = a_0 t^2 + (1-2a_0)t + a_0 \) the assumptions of the lemma hold if \( 0 < a_0 < \frac{1}{4} \). A simple computation shows that in fact there are no roots on the unit circle iff \( -\infty < a_0 < \frac{1}{4} \). This shows that the assumptions in the lemma are not strictly necessary, although perhaps desirable.

The effect of the choice of the collocation points \( z_i, (1 < i < m) \), on the roots of the characteristic equation \( c(t) = 0 \) will be illustrated by means of the next example.

Example 7.4. Let \( c(t) = a_0 t^2 + (1-2a_0)t + a_0 \) and let \( n = 2 \), i.e., we consider symmetric 5-point approximations to the second derivative. For \( c(t) \) to be symmetric it is necessary to place the collocation points symmetrically in the interval \( [x_{j-2}, x_{j+2}] \). Hence if \( m = 1 \) there is one collocation point \( z_1 = x_j \). The difference approximation to \( y'''(x) = f(x) \) is in this case given by \(( -u_{j-2} + 16u_{j-1} - 30u_j + 16u_{j+1} - u_{j+2} ) / 12h^2 = f(x_j) \). The left hand side can also be written as
so that the characteristic polynomial of this approximation is given by
\[ c(t) = -\frac{1}{12} t^2 + \frac{14}{12} t - \frac{1}{12}. \]
The characteristic equation has roots \( t_1 = 7 \pm 4 \sqrt{3} \), so that \( |t_1| \neq 1 \).

Next, if \( m = 2 \) let \( z_1 = x_j - qh \) and \( z_2 = x_j + qh \), \( q > 0 \). This generates a difference approximation to \( y''(x) = f(x) \) of the form
\[ \left( a_0 u_{j-2} + (1 - 4a_0) u_{j-1} + (-2 + 6a_0) u_j + (1 - 4a_0) u_{j+1} + a_0 u_{j+2} \right) / h^2 = \frac{1}{2} \left( f(z_1) + f(z_2) \right). \]

The corresponding characteristic polynomial is \( c(t) = a_0 t^2 + (1 - 2a_0) t + a_0 \).

It is easily shown that \( a_0 \) and \( q \) are related by \( a_0 = (6q^2 - 1) / 12 \). By example (7.3) there are no roots of \( c(t) = 0 \) on the unit circle if and only if
\[ -\infty < a_0 < \frac{1}{4}. \]
In terms of \( q \) this condition becomes \( 0 < q < \frac{1}{3} \sqrt{6} \approx 0.816 \).

If the collocation points are given by \( z_1 = x_j - q_0 h \) and \( z_2 = x_j + q_0 h \), where
\[ q_0 = \sqrt{1 - \frac{11}{15}} \approx 0.38 \quad \text{or} \quad q_0 = \sqrt{1 + \frac{11}{15}} \approx 1.36, \]
then the order of consistency of the corresponding finite difference approximation is equal to six. (See example (4.6).) Therefore the theory of Kreiss guarantees stability only for the first case, where \( q_0 = \sqrt{1 - \frac{11}{15}} \). Of course, this does not imply that a finite difference approximation based upon the second value of \( q_0 \) is necessarily unstable. That such an approximation may lead to a stable scheme is supported by the data obtained in the next section. (See example (8.1), experiment number 15.) The first value of \( q_0 \) appears to give a better error constant, however.

Finally, take \( m = 5 \) and let the collocation points \( z_i \), \( (1 \leq i \leq 5) \), coincide with the meshpoints \( x_{j+i} \), \( (-2 \leq i \leq 2) \). This generates the five point Mehrstellenverfahren of Collatz (1960), page 502, viz.
The associated characteristic equation is

\[
\frac{1}{252h^2} \left( 31u_{j-2} + 128u_{j-1} - 318u_j + 128u_{j+1} + 31u_{j+2} \right)
\]

\[= \frac{1}{3780} \left( 23f_{j-2} + 688f_{j-1} + 2358f_j + 688f_{j+1} + 23f_{j+2} \right).\]

The associated characteristic equation is

\[
\frac{1}{252} (31t^2 + 190t + 31) = 0 ,
\]

with roots \( t_1 \approx -0.168 \) and \( t_2 \approx -5.96 \). Hence these collocation points lead to a stable difference approximation.

Now consider the extra boundary conditions (7.4) and (7.5). Again, assume that the characteristic polynomial \( c(t) \) of (7.1) is symmetric and that the degree of \( c(t) \) is even and equal to \( 2N_1 \). If in addition \( c(t) \) is strictly diagonally dominant with positive coefficients then the characteristic equation \( c(t) = 0 \) has exactly \( N_1 \) roots inside the unit circle and \( N_1 \) roots outside the unit circle. A necessary condition for the difference equations (7.10), (7.11) and (7.12), (7.13) to admit the zero solution only is then that the number of extra boundary conditions at \( x = 0 \) is the same as the number at \( x = 1 \) and equal to \( N_1 = (r + s - n) / 2 \). (Hence \( k_0 = (r - s + n)/2 \) in (7.4) and (7.5)). It is also reasonable to assume now that the characteristic equations of the extra equations at \( x = 0 \) and \( x = 1 \) are related by

\[
(7.14) \quad c_{j+r-s-j}(t) = t \left( \frac{1}{j} \right) c_{j+1}(t) , \quad (r+s-n)/2 = k_0 \leq j \leq r-1 ,
\]

i.e., the conditions at \( x = 1 \) are the "reflection" of those at \( x = 0 \). For stability it is then sufficient to show that the difference equations (7.10)
subject to (7.11) have the zero solution only. For this purpose define polynomials $p_j(t)$ by

\[(7.15) \quad p_j(t) = \sum_{i=0}^{N} a_i t^{i+j}, \quad j \geq 0,\]

Here $N = r + s - n$ and the coefficients $a_i$ are the same as those of the characteristic polynomial $c(t)$ in (7.6). If (7.10) subject to (7.11) admits a nontrivial solution then it is easily seen that the polynomials $c_j(t)$, \((r+s-n)/2 = k_0 \leq j \leq r-1\), and $p_j(t)$, \(0 \leq j \leq \max N_j - N\), are linearly dependent. Hence we have shown the following

**Theorem (7.5).** Let the homogeneous problem corresponding to (1.1) and (1.1a,b) only have the trivial solution and let the difference scheme (7.6) be consistent. Assume that $c(t)$ is symmetric. Also suppose that the degree of $c(t)$ is even and that $c(t)$ is strictly diagonally dominant with positive coefficients. Let the characteristic polynomials of the extra boundary conditions be related as in Equation (7.14). If the polynomials $c_j(t)$ and $p_j(t)$ defined by (7.8) and (7.15) respectively are linearly independent then the difference scheme is stable. Hence there exists a constant $K$ independent of $h$ such that $||e_h|| \leq K ||\tau_h||$.

**Example 7.6.** Let $c(t) = a_0 t^2 + (1-2a_0) t + a_0$ with $-\infty < a_0 < \frac{1}{4}$. If the degree of the characteristic polynomial $c_1(t)$ of the extra boundary condition at $x = 0$ is also equal to two then the difference scheme is stable if $c_1(t) \neq c(t)$. If the degree of $c_1(t)$ is three then stability is guaranteed if there is no constant $b$ such that $c_1(t) = b(a_0 t^3 + (1-2a_0) t^2 + a_0 t) + (1-b) (a_0 t^2 + (1-2a_0) t + a_0)$. 
8. NUMERICAL EXAMPLES

The main purpose of the numerical examples given in this section is to check the correctness of statements in previous sections. They also give some indication as to what the relative accuracy of various discretizations is. All computations were carried out on an IBM 370/168, using double precision arithmetic. No attempt was made to optimize the efficiency of the computations, so that there will be no conclusions about the relative merit of various finite difference schemes.

Example 8.1. Let the differential equation be given by

\[ Ly(x) = y'' + y' - 2y = 2(1-6x)e^x \]

with boundary conditions

\[ y(0) = y(1) = 0. \]

The solution of this problem is

\[ y(x) = 2x(1-x)e^x. \]

The results of some numerical computations are given in Table (8.1). In this table r, s and m are as in the finite difference approximation (2.4). So the width of the approximation equals \( \omega = r + s + 1 \) and m denotes the number of collocation points. The letters A, B, C, D and E in the columns headed by "c" are a code indicating the location of the collocation points, viz.
A: The points $z_i$, $(1 \leq i \leq m)$, are optimal i.e., the order of consistency is as high as possible for the particular values of $r$, $s$ and $m$ considered.

B: The points $z_i$, $(1 \leq i \leq m)$, are optimal under the restrictions that each subinterval contain the same number of collocation points, and that these points are placed symmetrically with respect to the subinterval. (See the last case discussed in Example 5 of Section 4.)

C: The collocation points coincide with the meshpoints and $m = r + s + 1$.

D: $m = 1$ and $z_1 = x_j$.

E: The placement of the collocation points is not optimal, but they are placed symmetrically in the interval $[x_{j-r}, x_{j+s}]$.

The column headed by "o" gives the order of consistency of the finite difference approximation as predicted by theorems in this chapter. Columns 7 through 11 define the finite difference equations (2.4) for $r \leq j \leq J-s$. If the width of this approximation is equal to 5 then a special finite difference equation must be defined for $j = 1$. This is done in columns 2-6. (The special equation necessary for $j = J-1$ is assumed to be the "reflection" of the one for $j=1$.) The mesh is taken uniform in this example, so that $h_j = h = \frac{1}{J}$. For a number of values of $J$ the observed maximum error, i.e., $\max |u_j - y(x_j)|$, is given. The notation $0.438^{-1}$ means $0.438 \times 10^{-1}$ etc. In the final column, headed by "a", the observed asymptotic order of accuracy of the given finite difference scheme is listed. If it is not clear from the numerical results what this order is then the expected order is given between brackets.
Most of the results that appear in the table are self explanatory. The first seven experiments involve compact difference approximations. For experiments 5 and 6 the collocation points are given by $z_1 = x_{j+\frac{1}{2}} \pm \xi h$. There are two values of $\xi$ for which the optimal order of consistency is reached. (See the last case discussed in Example (4.5)). These values are $\xi = \sqrt{\frac{5}{12} - \frac{1}{2}\sqrt{\frac{23}{45}}}$, used in Experiment 5, and $\xi = \sqrt{\frac{5}{12} + \frac{1}{2}\sqrt{\frac{23}{45}}}$, used in 6.

Experiments 8-12 show the effect that various choices of the extra boundary conditions have on the overall accuracy. Note that even if the order of consistency of the extra finite difference equations is only equal to two, then the order of accuracy of the scheme remains four. This phenomenon is also explained in the paper of Kreiss (1972). (See also Bramble and Hubbard (1964) and Shoosmith (1975).) The actual accuracy however is seriously effected. (Amazingly this is not the case for experiment 9, but this must be considered exceptional.)

In 13 and 14 the collocation points of the main finite difference equations are $z_1 = x_j - \xi h$ and $z_2 = x_j + \xi h$. Again, as has been mentioned previously in Example (4.6), there are two values of $\xi$ for which the order of consistency becomes six. These are $\xi = \sqrt{1 - \sqrt{\frac{11}{15}}}$ used in 13 and $\xi = \sqrt{1 + \sqrt{\frac{11}{15}}}$ used in 14.

**Example 8.2.** Consider the equation

$$y'' + xy' - (1 + x)y = -(2 + x)e^x,$$

with boundary conditions

$$y'(0) = y(1) = 0.$$
<table>
<thead>
<tr>
<th>#</th>
<th>r s m c o</th>
<th>J = 4</th>
<th>J = 8</th>
<th>J = 16</th>
<th>J = 32</th>
<th>J = 64</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 1 2 D 2</td>
<td>.438^-1</td>
<td>.111^-1</td>
<td>.281^-2</td>
<td>.704^-3</td>
<td>.176^-3</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1 1 2 B 2</td>
<td>.111^-1</td>
<td>.279^-2</td>
<td>.700^-3</td>
<td>.175^-3</td>
<td>.439^-4</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1 1 2 A 4</td>
<td>.830^-4</td>
<td>.534^-5</td>
<td>.337^-6</td>
<td>.211^-7</td>
<td>.132^-8</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>1 1 3 C 4</td>
<td>.516^-3</td>
<td>.325^-4</td>
<td>.203^-5</td>
<td>.128^-6</td>
<td>.798^-8</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>1 1 4 B 6</td>
<td>.213^-6</td>
<td>.337^-8</td>
<td>.528^-10</td>
<td>.817^-12</td>
<td>**</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>1 1 5 E 6</td>
<td>.156^-6</td>
<td>.249^-8</td>
<td>.390^-10</td>
<td>.628^-12</td>
<td>**</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>1 1 1 D 2</td>
<td>2 2 1 D 4</td>
<td>.291^-1</td>
<td>.265^-2</td>
<td>.204^-3</td>
<td>.147^-4</td>
<td>.987^-6</td>
</tr>
<tr>
<td>9</td>
<td>1 3 1 D 3</td>
<td>2 2 1 D 4</td>
<td>.440^-2</td>
<td>.233^-3</td>
<td>.897^-5</td>
<td>.310^-6</td>
<td>.160^-7</td>
</tr>
<tr>
<td>10</td>
<td>1 4 1 D 4</td>
<td>2 2 1 D 4</td>
<td>.369^-4</td>
<td>.458^-5</td>
<td>.326^-6</td>
<td>.207^-7</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>1 1 2 A 4</td>
<td>2 2 1 D 4</td>
<td>.533^-3</td>
<td>.719^-4</td>
<td>.518^-5</td>
<td>.336^-6</td>
<td>.212^-7</td>
</tr>
<tr>
<td>12</td>
<td>1 1 3 C 4</td>
<td>2 2 1 D 4</td>
<td>.884^-3</td>
<td>.781^-4</td>
<td>.528^-5</td>
<td>.338^-6</td>
<td>.212^-7</td>
</tr>
<tr>
<td>13</td>
<td>1 1 5 E 6</td>
<td>2 2 2 A 6</td>
<td>.571^-7</td>
<td>.185^-8</td>
<td>.367^-10</td>
<td>.601^-12</td>
<td>**</td>
</tr>
</tbody>
</table>

** Contaminated by roundoff
The solution to this problem is \( y(x) = (1-x)e^x \).

In this example only compact approximations to the differential equations are considered. Numerical test calculations are performed with various difference approximations to the boundary condition \( y'(0) = 0 \). Results appear in Table (8.2). The notation used is the same as in the previous example. The approximation to the boundary condition is defined in columns 2-6, while the finite difference approximation to the differential equation is defined in the next five columns.

In Experiment 9 the collocation point for the boundary condition is \( z_1 = x_0 + \frac{h}{3} \). For this value of \( z_1 \) the order of consistency is three. (See Example (6.2).) In Experiments 13 and 14 this collocation point is \( z_1 = x_0 + \xi h \). In 13 the value of \( \xi \) is \((9 - \sqrt{33})/12\) and in 14 this value is \((9 + \sqrt{33})/12\). For these collocation points the order of consistency is equal to four rather than three. (See Example (6.3).) Note that the order of accuracy is not greater than the order of consistency of the discrete boundary condition. This differs from observations made about the extra boundary conditions in Example (8.1).

Example 8.3. The purpose of this example is to illustrate the usefulness of allowing nonuniform meshes. Let the differential equation be given by

\[ \varepsilon^2 y'' - x^2 y = 2\varepsilon^2 - x^4 - \varepsilon e^{-x^2/2\varepsilon} \] \[ \varepsilon = 0.001,\]

with boundary conditions

\[ y'(0) = 0 \quad \text{and} \quad y(1) = 1 + e^{-\frac{1}{2\varepsilon}} \approx 1.\]
**TABLE 8.2**

<table>
<thead>
<tr>
<th>#</th>
<th>r s m c o</th>
<th>r s m c o</th>
<th>$J = 4$</th>
<th>$J = 8$</th>
<th>$J = 16$</th>
<th>$J = 32$</th>
<th>$J = 64$</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 1 0 - 1</td>
<td>1 1 1 D 2</td>
<td>0.117</td>
<td>0.494</td>
<td>0.226</td>
<td>0.108</td>
<td>0.526</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0 2 0 - 2</td>
<td>1 1 1 D 2</td>
<td>0.240</td>
<td>0.492</td>
<td>0.110</td>
<td>0.261</td>
<td>0.633</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0 1 1 * 2</td>
<td>1 1 1 D 2</td>
<td>0.250</td>
<td>0.631</td>
<td>0.158</td>
<td>0.396</td>
<td>0.989</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>0 1 1 **2</td>
<td>1 1 1 D 2</td>
<td>0.240</td>
<td>0.492</td>
<td>0.110</td>
<td>0.261</td>
<td>0.633</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>0 1 0 - 1</td>
<td>1 1 2 A 4</td>
<td>0.106</td>
<td>0.466</td>
<td>0.219</td>
<td>0.106</td>
<td>0.521</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0 2 0 - 2</td>
<td>1 1 2 A 4</td>
<td>0.362</td>
<td>0.786</td>
<td>0.183</td>
<td>0.443</td>
<td>0.109</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>0 1 1 * 2</td>
<td>1 1 2 A 4</td>
<td>0.152</td>
<td>0.359</td>
<td>0.876</td>
<td>0.216</td>
<td>0.538</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>0 1 1 **2</td>
<td>1 1 2 A 4</td>
<td>0.338</td>
<td>0.764</td>
<td>0.181</td>
<td>0.440</td>
<td>0.108</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>0 1 1 A 3</td>
<td>1 1 2 A 4</td>
<td>0.531</td>
<td>0.602</td>
<td>0.711</td>
<td>0.862</td>
<td>0.106</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>0 1 2 C 3</td>
<td>1 1 2 A 4</td>
<td>0.146</td>
<td>0.171</td>
<td>0.207</td>
<td>0.254</td>
<td>0.316</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>0 4 0 - 4</td>
<td>1 1 2 A 4</td>
<td>0.344</td>
<td>0.164</td>
<td>0.898</td>
<td>0.526</td>
<td>0.317</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>0 1 3 E 4</td>
<td>1 1 2 A 4</td>
<td>0.482</td>
<td>0.324</td>
<td>0.208</td>
<td>0.132</td>
<td>0.828</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>0 2 1 A 4</td>
<td>1 1 2 A 4</td>
<td>0.165</td>
<td>0.108</td>
<td>0.730</td>
<td>0.472</td>
<td>0.300</td>
<td>4</td>
</tr>
<tr>
<td>14</td>
<td>0 2 1 A 4</td>
<td>1 1 2 A 4</td>
<td>0.563</td>
<td>0.249</td>
<td>0.129</td>
<td>0.734</td>
<td>0.437</td>
<td>4</td>
</tr>
</tbody>
</table>

* $z_1 = x_0$

** $z_1 = x_1$
As was the case in the previous examples, this problem has been constructed so it would have a known solution. Here the solution is

\[ y(x) = e^{-x^2/2c} + x^2, \]

which has a boundary layer at \( x = 0 \). One could equally well impose the boundary condition \( y(0) = 1 \), but from numerical point of view the condition \( y'(0) = 0 \) makes it considerably more difficult to obtain a reasonably accurate solution. Hence the given condition is the more interesting one here. Results are given in Table (8.3). In the nonuniform case the mesh is taken such that there are \( \frac{3}{4} J \) meshpoints, equally spaced, to the left of \( x = 0.1 \) and \( \frac{1}{4} J \) meshpoints, also uniformly spaced, to the right of \( x = 0.1 \). Hence for each experiment there is only one finite difference approximation that needs to take into account the nonuniformity of the mesh. The actual finite difference equations involved are the second case discussed in Example (4.3), (with \( a^+(x) \equiv 0 \)), and the equation given in Example (4.4).

| TABLE 8.3 |
| Uniform Mesh |
| \( \# \) | \( r s m c o \) | \( r s m c o \) | \( J=8 \) | \( J=16 \) | \( J=32 \) | \( J=64 \) |
| 1 | 0 2 0 - 2 | 1 1 1 D 2 | 1.000 \(^0\) | .954 \(^0\) | .486 \(^0\) | .923 \(^{-1}\) |
| 2 | 0 4 0 - 4 | 1 1 3 C 4 | .987 \(^0\) | .883 \(^0\) | .909 \(^{-1}\) | .782 \(^{-1}\) |

| Nonuniform Mesh |
| \( \# \) | \( r s m c o \) | \( r s m c o \) | \( J=8 \) | \( J=16 \) | \( J=32 \) | \( J=64 \) |
| 3 | 0 2 0 - 2 | 1 1 1 A 2 | .110 \(^0\) | .127 \(^{-1}\) | .508 \(^{-2}\) | .203 \(^{-2}\) |
| 4 | 0 4 0 - 4 | 1 1 3 C 3 \(^*\) | .934 \(^{-1}\) | .718 \(^{-2}\) | .223 \(^{-2}\) | .725 \(^{-3}\) |

\(^*\) 4 if the mesh is locally uniform
Example 8.4. Let the problem be the same as in Example (8.2). Approximate the differential equation by a compact finite difference equation with \( r = s = m = 1 \). Let \( z_1 = x_j \). The mesh is taken uniform. A one parameter set of approximations to the boundary condition \( y'(0) = 0 \) is given by

\[
\frac{u_0 - \alpha u_1 + (\alpha-1)u_2}{(\alpha - 2)h} = 0.
\]

This approximation is consistent if \( \alpha \neq 2 \) and the order of consistency is one. If \( \alpha = \frac{4}{3} \) then the order becomes two and the finite difference equation is in that case identical to the one obtained with the procedure of Section 5, letting \( r = 0, s = 2 \) and \( m = 0 \).

It will be shown in Section 6 of Chapter II that if \( \alpha > 2 \) then the finite difference operator \( L_h \) has an eigenvalue \( \lambda \to +\infty \) as \( h \to 0 \). This makes the approximation useless for time dependent problems. From the general theory of Kreiss (1972) it follows that the same approximation when used in the current boundary value problem, does not lead to an instability. This is supported by the following data, obtained with \( \alpha = 4 \).

**TABLE 8.4**

<table>
<thead>
<tr>
<th>J</th>
<th>J = 4</th>
<th>J = 8</th>
<th>J = 16</th>
<th>J = 32</th>
<th>J = 64</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>.784°</td>
<td>.248°</td>
<td>.100°</td>
<td>.453°</td>
<td>.216°</td>
<td>1</td>
</tr>
</tbody>
</table>
Chapter II

INITIAL VALUE PROBLEMS

1. **INTRODUCTION**

This chapter is concerned with the construction of finite difference approximations to initial value problems. In Section 2 the construction procedure of the previous chapter will be extended to systems of linear first order ordinary differential equations subject to initial conditions. Examples of this very general class of difference approximations appear in Section 3. In particular, a constant coefficient problem is discussed in detail, (Example (3.4)), in view of its application to the stability analysis of parabolic equations, (Section 6).

Well-known concepts in the stability analysis of finite difference approximations to initial value problems in ordinary differential equations are summarized in Section 4. Particular attention is paid to methods for "stiff" problems, because of their applicability to the numerical solution of parabolic equations. In Section 5 it is shown how the difference methods for boundary value problems and initial value problems may be combined to yield quite general difference schemes for the solution of linear parabolic equations in one space variable. The numerical stability of such schemes is investigated in Section 6. Particular attention is paid to the effect of finite difference approximations to the boundary conditions. Several examples are given. Finally, in Section 7 the results of some numerical experiments are given.
In this section the construction of finite difference approximations to the first order linear system

\[(2.1) \quad \forall w(t) = A(t)w'(t) - B(t)w(t) = f(t), \quad 0 \leq t \leq T,\]

will be discussed. Here \(w(t) = (w_1(t), w_2(t), \ldots, w_{J-1}(t))^T, f(t) = (f_1(t), f_2(t), \ldots, f_{J-1}(t))^T\) and \(A(t)\) and \(B(t)\) are continuous \((J-1) \times (J-1)\) matrices. It is assumed that the above system, together with the initial condition \(w(0) = w^0\), admits a unique solution \(w(t), (0 \leq t \leq T)\), that is as many times continuously differentiable as necessary. The construction procedure, that is an extension of the procedure discussed in the previous chapter, applies equally well to higher order systems, subject to more general boundary conditions. However, these will not be considered here.

Define a mesh \(S_h\) by

\[S_h = \{t_j : 0 = t_0 < t_1 < \cdots < t_N = T\}.\]

The mesh spacings are \((\Delta t)_n = t_n - t_{n-1}, (1 \leq n \leq N)\). Let \(\Delta t = \max(\Delta t)_n\), so that \((\Delta t)_n = \Delta t = T/N\) for uniform meshes. Assume that \(\min(\Delta t)_n \geq K^{-1} \Delta t\), for some constant \(K \geq 1\).

Finite difference approximations to (2.1) are assumed to have the form

\[(2.2) \quad L_n u^n = \sum_{\ell = -\rho_n}^{0} D_n,\ell u^{n+\ell} = \sum_{\ell = 1}^{\mu_n} E_n,\ell f(\zeta_n, \ell), \quad 1 \leq n \leq N,\]
with initial condition \( u^0 = w^0 \).

Here \( u^n = (u^n_1, u^n_2, \ldots, u^n_{J-1})^T \) is the approximation to \( w(t^n) \). For each \( n \) the points \( \xi_{n,k} \), \( 1 \leq k \leq \mu_n \), are distinct points, that will normally lie in the interval \([t_{n-\rho}, t_n]\). The coefficients \( D_{n,k} \) and \( E_{n,k} \) are \((J-1) \times (J-1)\) matrices. These coefficients may be determined by requiring that

\[
\sum_{k=-\rho}^{\mu} D_{n,k} p(t_{n+k}) - \sum_{k=1}^{\mu} E_{n,k} L_p(\xi_{n,k}) = 0 ,
\]

for all \( p(t) \in P_{\rho+\mu-1}^{J-1} \). Here \( P_{d}^{J-1} \) is the space of all \((J-1)\)-vector valued polynomials of degree at most \( d \). (For notational convenience the subscript \( n \) will frequently be omitted.) This requirement leads to the matrix equation

\[
\begin{bmatrix}
I & 0 & \cdots & 0 & -L_{1,1}^T & \cdots & -L_{0,\mu}^T \\
0 & I & 0 & -L_{1,1}^T & \cdots & -L_{1,\mu}^T \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & I & -L_{\rho,1}^T & \cdots & -L_{\rho,\mu}^T \\
0 & 0 & 0 & -L_{\rho+1,1}^T & \cdots & -L_{\rho+1,\mu}^T \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & 0 & -L_{\rho+\mu-1,1}^T & \cdots & -L_{\rho+\mu-1,\mu}^T \\
\end{bmatrix}
\begin{bmatrix}
D_{-\rho}^T \\
D_{-\rho+1}^T \\
\vdots \\
D_{0}^T \\
E_{1}^T \\
\vdots \\
E_{\mu}^T
\end{bmatrix}
= 0
\]

where \( I \) and \( 0 \) are the \((J-1) \times (J-1)\) identity and zero matrix respectively. \( L_{i,k}^T \) is the transpose of \( L_{i,k} \), where

\[
L_{i,k} = L_0 \omega^i(\xi_{n,k})A(\xi_{n,k}) - \omega^i(\xi_{n,k})B(\xi_{n,k}) ,
\]
with \( L_0 \omega^i(t) = \frac{d \omega^i(t)}{dt} \). The scalar polynomials \( \omega^i(t), \(0 \leq i \leq \rho + \mu - 1\), are linearly independent and chosen such that

\[
\omega^i(t_{n-\rho+k}) = \delta_{ik}, \quad 0 \leq i, k \leq \rho,
\]
and

\[
\omega^i(t_{n-\rho+k}) = 0, \quad \rho + 1 \leq i \leq \rho + \mu - 1, \quad 0 \leq k \leq \rho.
\]

For example, these polynomials may be defined in a similar fashion as the polynomials \( w^i(x) \) in Section (I.2). (Equations (1.2.13) and (1.2.14).)

The coefficient matrices \( D_k \) and \( E_k \) in (2.3) are uniquely defined up to a certain normalization, provided that the matrix

\[
C \equiv \begin{bmatrix}
L_{\rho+1,1} & \cdots & L_{\rho+1,\mu} \\
\vdots & \ddots & \vdots \\
L_{\rho+1,\mu} & \cdots & L_{\rho+1,\mu}
\end{bmatrix}
\]

has rank equal to \((\mu-1)(J-1)\). This is the case for all small enough \( \Delta t \), provided that the matrix

\[
C \equiv \begin{bmatrix}
L_0 \omega^{\rho+1}(\xi_1)A(\xi_1) & \cdots & L_0 \omega^{\rho+1}(\xi_1)A(\xi_1) \\
\vdots & \ddots & \vdots \\
L_0 \omega^{\rho+1}(\xi_\mu)A(\xi_\mu) & \cdots & L_0 \omega^{\rho+1}(\xi_\mu)A(\xi_\mu)
\end{bmatrix}
\]

has rank \((\mu-1)(J-1)\). This in turn is true if \( \mathcal{E} \neq 0 \) for all small enough \( \Delta t \).

Here \( \mathcal{E} \) is the analogue of the normalizing factor \( E \) in the previous chapter, (Equation (1.2.12)), and can be expressed as
Although it is not important for theoretical purposes in what manner the difference equation (2.2) is normalized, it may be assumed, that this is done in such a way that
\[ c \leq \| \sum_{\ell=0}^{\mu} E_{\ell} \| \leq C, \]
for all small \( \Delta t \). Here \( c \) and \( C \) are positive constants that do not depend on \( \Delta t \). The proofs of the following statements are then very much like those of Theorem (I.2.1), Theorem (I.2.4), and Theorem (I.3.1) and will be omitted.

Theorem 2.1. Assume that \( \rho \geq 1 \) and that \( E \neq 0 \). Then the order of consistency of the finite difference approximation (2.2) is at least equal to \( \rho + \mu - 1 \). In particular, if \( \mu = 1 \) or if \( \rho = 1 \) then \( E \neq 0 \). If the points \( \xi_{i} \) are the critical points of \( \tilde{\omega}^{\rho + \mu} (t) = \prod_{\ell=1}^{\mu-1} (t - \gamma_{\ell}) \prod_{\ell=1}^{0} (t - t_{n+\ell}) \) then the order is at least \( \rho + \mu \). (Thus there is a \((\mu-1)\)-parameter family of such points with parameters \( \gamma_{\ell} \), \( 1 \leq \ell \leq \mu-1 \).

3. EXAMPLES OF FINITE DIFFERENCE APPROXIMATIONS

In this section a number of specific examples of finite difference approximations to equation (2.1) will be discussed. Some of these have been mentioned before in Section (I.4), for the one dimensional case.

Example 3.1. With \( \rho = 1 \), \( \mu = 1 \) and \( \xi_{1} = t_{n+\frac{1}{2}} = \frac{1}{2} (t_{n} + t_{n+1}) \) one obtains the
Box scheme of Keller (1969). This difference approximation may be written as

\[ D_{-1}u^{n-1} + D_0u^n = f(t_{n-\frac{1}{2}}) , \]

where

\[ D_{-1} = \frac{-1}{\Delta t} A(t_{n-\frac{1}{2}}) - \frac{1}{2}B(t_{n-\frac{1}{2}}) , \]

and

\[ D_0 = \frac{1}{\Delta t} A(t_{n-\frac{1}{2}}) - \frac{1}{2}B(t_{n-\frac{1}{2}}) . \]

That the order of consistency is two follows from Theorem (2.1).

Next if \( p = 1 \) and \( \mu = 2 \), with \( \xi_1 = t_{n-1} \) and \( \xi_2 = t_n \) then the matrix equation (2.3) becomes

\[
\begin{bmatrix}
1 & 0 & -L_{0,1}^T & -L_{0,2}^T \\
0 & 1 & -L_{1,1}^T & -L_{1,2}^T \\
0 & 0 & -L_{2,1}^T & -L_{2,2}^T \\
\end{bmatrix}
\begin{bmatrix}
D_{-1}^T \\
D_0^T \\
E_1^T \\
E_2^T \\
\end{bmatrix} =
\begin{bmatrix}
0 \\
0 \\
0 \\
0 \\
\end{bmatrix},
\]

where

\[ L_{i,1} = L_0 \omega^i(t_{n-1})A(t_{n-1}) - \omega^i(t_{n-1})B(t_{n-1}), \quad 0 \leq i \leq 2 , \]

and

\[ L_{i,2} = L_0 \omega^i(t_n)A(t_n) - \omega^i(t_n)B(t_n), \quad 0 \leq i \leq 2 . \]

With \( \omega_0(t) = -(t-t_n) / \Delta t \), \( \omega^1(t) = (t-t_{n-1}) / \Delta t \) and \( \omega^2(t) = (t-t_{n-1})(t-t_n) / \Delta t^2 \), it is found that

\[ L_{0,1} = \frac{-1}{\Delta t} A(t_{n-1}) - B(t_{n-1}) , \quad L_{0,2} = \frac{-1}{\Delta t} A(t_n) , \]

\[ L_{1,1} = \frac{1}{\Delta t} A(t_{n-1}) , \quad L_{1,2} = \frac{1}{\Delta t} A(t_n) - B(t_n) . \]
\[
L_{2,1} = \frac{-1}{\Delta t} A(t_{n-1}) , \quad L_{2,2} = \frac{1}{\Delta t} A(t_n)
\]

From (3.1) it follows that \(E_1 A(t_{n-1}) = E_2 A(t_n)\). Take \(E_2 = I\). Then \(E_1\) is the solution of the matrix equation \(E_1 A(t_{n-1}) = A(t_n)\). Clearly, it is more difficult now to give explicit representations of the coefficients \(E_1\) than was the case for scalar equations in Chapter I. Of course, one can always obtain the coefficients numerically.

In the special case where \(A(t_{n-1})\) and \(A(t_n)\) commute, it is possible to state explicitly what the coefficients of the above finite difference equation are. In particular, if \(A\) is a constant matrix, then with \(E_2 = \frac{1}{2} I\) it is found that \(E_1 = \frac{1}{2} I\), \(D_{-1} = -\frac{1}{\Delta t} A - \frac{1}{2} B(t_{n-1})\) and \(D_0 = \frac{1}{\Delta t} A - \frac{1}{2} B(t_n)\).

Thus the difference equation may be written as

\[
A(\frac{u^n - u^{n-1}}{\Delta t}) - \frac{1}{2} \left( B(t_n)u^n + B(t_{n-1})u^{n-1} \right) = \frac{1}{3} \left( f(t_n) + f(t_{n-1}) \right)
\]

which is recognized as the usual trapezoidal method. The order of consistency of this equation is also equal to two.

**Example 3.2.** With \(p = 2\), \(\mu = 1\) and \(\xi_1 = t_{n-2} + \xi \Delta t\) the difference approximation to (2.1) becomes

\[
A(\xi_1) \left[ \frac{1}{2}(2\xi-1)u^{n-2}(\xi-1)u^{n-1} + \frac{1}{2}(2\xi-3)u^{n-2} \right] - B(\xi_1) \left[ \frac{1}{2}\xi(\xi-1)u^{n-}\xi(\xi-2)u^{n-1} + \right.
\]
\[
\left. + \frac{1}{2}(\xi-1)(\xi-2) \right] = f(\xi_1).
\]

(The mesh is uniform in this example.) The order of consistency of this approximation is equal to two. From Theorem (2.1) it follows that the order of consistency increases to three if one takes \(\xi = 1 + \frac{1}{3}\sqrt{3} \approx 1.58.\)
If $\xi = 2$ then the above formula reduces to

$$A(t_n) \frac{1}{\Delta t} \left[ \frac{3}{2} u^n - 2u^{n-1} + \frac{1}{2} u^{n-2} \right] - B(t_n)u^n = f(t_n),$$

which is one of Gear's backward differentiation formulas. (See Gear (1971), p. 217.)

**Example 3.3.** Consider the constant coefficient problem

$$Aw'(t) - Bw(t) = f(t), \quad 0 \leq t \leq T,$$

(3.2)

$$w(0) = w^0,$$

(3.2a)

where $A$ is nonsingular and where $A$ and $B$ commute. This occurs for example when $A = I$. Let $\mu = 2$ and $\rho = 1$. The construction procedure of Section 2 again leads to the matrix equation (3.1) given in Example (3.1), now for general choice of the points $\xi_1$ and $\xi_2$.

Let $E_1 = -L_0 \omega^2(\xi_2) A + \omega^2(\xi_2) B$. Then, using the fact that $A$ and $B$ commute, it is found that

$$E_2 = L_0 \omega^2(\xi_1) A - \omega^2(\xi_1) B,$$

$$D_{-1} = E_1 \left\{ L_0 \omega^0(\xi_1) A - \omega^0(\xi_1) B \right\} + E_2 \left\{ L_0 \omega^0(\xi_2) A - \omega^0(\xi_2) B \right\},$$

and

$$D_0 = E_1 \left\{ L_0 \omega^1(\xi_1) A - \omega^1(\xi_1) B \right\} + E_2 \left\{ L_0 \omega^1(\xi_2) A - \omega^1(\xi_2) B \right\},$$

where $L_0 \omega(t) = \omega'(t)$.

In particular if one uses the Gaussian points
\[ \zeta_1 = \frac{(t_{n-1} + t_n)}{2} - \Delta t \sqrt{3}/6 \quad \text{and} \quad \zeta_2 = \frac{(t_{n-1} + t_n)}{2} + \Delta t \sqrt{3}/6, \]

then the difference equation becomes

\[
(3.3) \quad \left\{ -\frac{1}{\Delta t} A^2 - \frac{1}{2} AB - \Delta t \frac{B^2}{12} \right\} u_n - 1 + \left\{ \frac{1}{\Delta t} A^2 - \frac{1}{2} AB + \Delta t \frac{B^2}{12} \right\} u_n =
\]

\[
= \left\{ \frac{1}{2} A - \frac{\Delta t \sqrt{3}}{12} B \right\} f(\zeta_1) + \left\{ \frac{1}{2} A + \frac{\Delta t \sqrt{3}}{12} B \right\} f(\zeta_2).
\]

The order of consistency of this equation is equal to four. If \( f(t) \equiv 0 \) then the finite difference approximation (3.3) can also be obtained from a particular Padé approximation to the solution \( w(t) = e^{A^{-1}B}w^0 \) of (3.2), (3.2a). These approximations, originally due to Padé (1892) have been extensively studied in connection with their stability properties. (See e.g., Varga (1961).)

**Example 3.4.** The preceding example suggests that the general \( \rho \)-step finite difference approximation to (3.2) with \( \mu \) collocation points has the form

\[
(3.4) \quad \sum_{\ell=-\rho}^{0} \left[ \sum_{\nu=0}^{\mu} \alpha^\ell_{\nu} A^{\mu-\nu} B^\nu \right] u_{n+\ell} = \sum_{\ell=1}^{\mu} \left[ \sum_{\nu=0}^{\mu-1} \beta^\ell_{\nu} A^{\mu-1-\nu} B^\nu \right] f(\zeta_\ell),
\]

where \( \alpha^\ell_{\nu} \) and \( \beta^\ell_{\nu} \) are constants that depend on the choice of the points \( \zeta_\ell \), \( (1 \leq \ell \leq \mu) \).

To verify this and to show how the coefficients \( \alpha^\ell_{\nu} \) and \( \beta^\ell_{\nu} \) may be found, consider the scalar equation

\[
(3.5) \quad Lw(t) \equiv w'(t) - \lambda w(t) = f(t),
\]
The finite difference approximation to (3.5) is given by

\[(3.6)\quad \sum_{\ell = -\rho}^{\mu} \delta_{\ell} u^{n+\ell} = \sum_{\ell = 1}^{\mu} \epsilon_{\ell} f(\xi_{\ell}) ,\]

where

\[(3.7)\quad \epsilon_{\ell} = \frac{(-1)^{\mu+\ell+1}}{\mathcal{E}} \begin{bmatrix} L\omega^p+1(\zeta_1) & \cdots & L\omega^p+1(\zeta_\mu) \\ \vdots & \ddots & \vdots \\ L\omega^p+1(\zeta_{\ell-1}) & \cdots & L\omega^p+1(\zeta_{\ell-1}) \\ L\omega^p+1(\zeta_{\ell+1}) & \cdots & L\omega^p+1(\zeta_{\ell+1}) \\ \vdots & \ddots & \vdots \\ L\omega^p+1(\zeta_{\mu}) & \cdots & L\omega^p+1(\zeta_{\mu}) \end{bmatrix} ,\]

and

\[(3.8)\quad \delta_{\ell} = \sum_{k=1}^{\mu} L\omega^p+\ell(\zeta_k) \epsilon_k , \quad -\rho \leq k \leq 0 .\]

The normalizing factor \(\mathcal{E}\) is given by (2.4). Considering (3.5), (3.7) and (3.8) it is observed that the coefficients have the form

\[\delta_{\ell} = \sum_{\nu=0}^{\mu-1} \alpha_{\nu}^\ell \lambda^\nu \quad \text{and} \quad \epsilon_{\ell} = \sum_{\nu=0}^{\mu-1} \beta_{\nu}^\ell \lambda^\nu ,\]

for certain constants \(\alpha_{\nu}^\ell\) and \(\beta_{\nu}^\ell\).

In order to show that these constants are the same as the constants in equation (3.4), note that by definition
\[
\sum_{\ell=1}^{\mu} \epsilon_\ell L_0^k(\xi_\ell) = 0, \quad \rho + 1 \leq k \leq \rho + \mu - 1,
\]
i.e.,
\[
q_k(\lambda) = \sum_{\ell=1}^{\mu} \left[ \sum_{\nu=0}^{\mu-1} \beta_\nu^\ell \lambda^\nu \right] \left[ L_0^k(\xi_\ell) - \omega^k(\xi_\ell) \lambda \right] = 0,
\]
for all \( \lambda \) and for \( \rho + 1 \leq k \leq \rho + \mu - 1 \).

So \( q_k(\lambda) = 0 \) and therefore \( q_k(C) = 0 \) for all matrices \( C \). In particular with \( C = A^{-1}B \) one obtains
\[
\sum_{\ell=1}^{\mu} \left[ \sum_{\nu=0}^{\mu-1} \beta_\nu^\ell (A^{-1}B)^\nu \right] \left[ L_0^kA^{-1}B - \omega^kA^{-1}B \right] = 0.
\]
Upon multiplication by \( A^\mu \) and using the assumption that \( A \) and \( B \) commute one obtains
\[
\sum_{\ell=1}^{\mu} \left[ \sum_{\nu=0}^{\mu-1} \beta_\nu^\ell A^{\mu-\nu}B^\nu \right] \left[ L_0^kA - \omega^kB \right] = 0, \quad \rho + 1 \leq k \leq \rho + \mu - 1.
\]
Thus if one defines
\[
E_\ell = \sum_{\nu=0}^{\mu-1} \beta_\nu^\ell A^{\mu-\nu}B^\nu,
\]
then the last \( \mu - 2 \) relations in the matrix equation (2.3) are satisfied. Further, the first \( \rho + 1 \) relations of (2.3) are automatically satisfied by letting
\[
D_\ell = \sum_{\nu=1}^{\mu} E_\nu \left[ L_0^{\rho + \ell}(\xi_\nu)A(\xi_\nu) - \omega^{\rho + \ell}(\xi_\nu)B(\xi_\nu) \right].
\]
Therefore it has been shown, that in case of the constant coefficient problem (3.2), with commuting A and B, one can explicitly give the coefficient matrices $D_f$ and $E_f$ of the difference approximation (2.2).

4. THE STABILITY OF FINITE DIFFERENCE APPROXIMATIONS TO INITIAL VALUE PROBLEMS

In this section some well-known concepts in the stability analysis of finite difference approximations to systems of first order ordinary differential equations will be summarized. To define stability of a finite difference approximation one only has to consider the form this difference equation takes when applied to the equation $w(t) = 0$. Here $w(t)$ is a scalar function. Thus for the type of approximations discussed in Section 2 the relevant difference equation is

\[(4.1) \quad \sum_{\ell=-\rho}^{0} \delta_{\ell} u^{n+\ell} = 0, \quad \rho \leq n \leq N,\]

where

\[
\delta_{\ell} = \frac{(-1)^{\ell}}{E} \begin{vmatrix} L_0 \omega^{+\ell}(\xi_1) & L_0 \omega^{+1}(\xi_1) & \cdots & L_0 \omega^{+\mu-1}(\xi_1) \\ \vdots & \vdots & \ddots & \vdots \\ L_0 \omega^{+\ell}(\xi_\mu) & L_0 \omega^{+1}(\xi_\mu) & \cdots & L_0 \omega^{+\mu-1}(\xi_\mu) \end{vmatrix}
\]

with $L_0 \omega(t) \equiv \omega'(t)$.

The mesh is taken to be uniform in this section, so that

\[(\Delta t)_n = \Delta t = 1/N, \quad (1 \leq n \leq N).\]

The approximation is said to be stable provided that for all given initial data $u^n$, $(0 \leq n \leq \rho-1)$, the solution to (4.1) satisfies
\[ |u^n| \leq \text{const} \sum_{\ell=0}^{\rho-1} |u^\ell|, \quad \rho \leq n \leq N. \]

This is the case if all roots \( \eta_i \) of the characteristic equation

\[ \sigma_0(\eta) = \Delta t \sum_{\ell=-\rho}^{0} \sigma_\ell \eta^{\rho + \ell} = 0 \]

satisfy

\[ |\eta_i| \leq 1, \quad 1 \leq i \leq \rho, \]

and

\[ |\eta_1| = 1 \text{ implies } \eta_1 \text{ is a simple root.} \]

These two conditions are usually referred to as the root condition. For practical purposes it is desirable that the approximation is strongly stable, i.e., the roots satisfy

\[ \eta_1 = 1, \]

\[ |\eta_i| < 1, \quad 2 \leq i \leq \rho. \]

**Example 4.1.** If \( \rho = 1 \) then the coefficients in (4.1) are given by \( \sigma_0 = -\frac{1}{\Delta t} \) and \( \sigma_1 = \frac{1}{\Delta t} \). The characteristic equation has one root \( \eta = 1 \) and thus this approximation is stable.

If \( \rho = 2 \) and \( \mu = 1 \) with \( \xi_1 = \frac{1}{n-2} + \xi \Delta t \), then the difference
approximation to \( w'(t) = 0 \) has the form

\[
\bar{\delta}_0 u^{n-2} + \bar{\delta}_1 u^{n-1} + \bar{\delta}_2 u^n = 0 ,
\]

where the coefficients \( \bar{\delta}_k \) are given by \( \bar{\delta}_0 = (2\xi-3)/2 \Delta t \), \( \bar{\delta}_1 = (4-4\xi)/2 \Delta t \) and \( \bar{\delta}_2 = (2\xi-1)/2 \Delta t \). The roots of the characteristic equation (4.2) are \( \eta_1 = 1 \) and \( \eta_2 = (2\xi-3)/(2\xi-1) \). Now \( |\eta_2| < 1 \) if and only if \( \xi > 1 \). Thus the approximation is strongly stable if and only if \( \xi > 1 \). In particular, the second order formula of Gear, for which \( \xi = 2 \), and the third order formula obtained if \( \xi = 1 + \frac{1}{3}\sqrt{3} \) are strongly stable. (See example (3.2).)

Now let \( \rho = 2 \) and \( \mu = 2 \) with \( \xi_1 = t_{n-2} + \xi_1 \Delta t \) and \( \xi_2 = t_{n-2} + \xi_2 \Delta t \). In this case the coefficients are

\[
\bar{\delta}_1 = \frac{\xi_2 - \xi_1}{\Delta t} \left[ 6(\xi_1 + \xi_2) - 6\xi_1 \xi_2 - 8 \right] ,
\]

\[
\bar{\delta}_2 = \frac{\xi_2 - \xi_1}{\Delta t} \left[ -3(\xi_1 + \xi_2) + 6\xi_1 \xi_2 + 2 \right] ,
\]

and

\[
\bar{\delta}_0 = -\bar{\delta}_1 - \bar{\delta}_2 .
\]

The characteristic equation \( \sigma_0(\eta) = 0 \) has roots \( \eta_1 = 1 \), and

\[
\eta_2 = -1 - \frac{\bar{\delta}_1}{\bar{\delta}_2} = \frac{-9(\xi_1 + \xi_2) + 6\xi_1 \xi_2 + 14}{-3(\xi_1 + \xi_2) + 6\xi_1 \xi_2 + 2} .
\]

For the approximation to be strongly stable it is necessary that \( |\eta_2| < 1 \). If \( \xi_2 = 2 \) then this is the case provided that \( \xi_1 < 0 \) or \( \xi_1 > \frac{2}{3} \). If \( \xi_2 = 1 \) then it is necessary that \( \xi_1 > \frac{5}{3} \) and if \( \xi_2 = 0 \) then one needs \( \frac{14}{9} < \xi_1 < 2 \).
Another well-known concept is that of stability region of a finite difference approximation. If in the system (2.1) the matrix $A^{-1}B$ has eigenvalues with large negative real part, then a large class of difference approximations requires $\Delta t$ to be very small for the approximation to be stable, even though this requirement may not be necessary for reasons of accuracy. Thus it is of interest to characterize those approximations that do not suffer from this shortcoming. For investigating the effect of the eigenvalues upon the stability it is customary to apply the finite difference approximations under consideration to the problem $\int \lambda w(t) = w'(t) - \lambda w(t) = 0$, where $w(t)$ is a scalar function. The analysis remains valid for the system $Aw'(t) = Bw(t)$, where $w$ is a vector and $A$ and $B$ are constant matrices, provided that $A^{-1}B$ has a complete set of eigenvectors. Finite difference approximations to the scalar equation $w'(t) = \lambda w(t)$ have the form

$$
\sum_{\ell=-\rho}^{0} \tilde{\delta}^n_{\ell} u^{n+\ell} = 0 , \quad \rho \leq n \leq N ,
$$

where the coefficients $\tilde{\delta}^n_{\ell}$ are given by

$$
(4.4) \quad \tilde{\delta}^n_{\ell} = \frac{(-1)^{\mu}}{\xi^n} \begin{vmatrix}
\int \lambda \omega^{\rho+\ell} (\xi_1) & \int \lambda \omega^{\rho+1} (\xi_1) & \cdots & \int \lambda \omega^{\rho+\mu-1} (\xi_1) \\
\vdots & \vdots & \ddots & \vdots \\
\int \lambda \omega^{\rho+\ell} (\xi_\mu) & \int \lambda \omega^{\rho+1} (\xi_\mu) & \cdots & \int \lambda \omega^{\rho+\mu-1} (\xi_\mu)
\end{vmatrix} .
$$

For stability it is again necessary that the roots $\eta_i$ of the characteristic equation $\tilde{q}(\eta) \equiv \Delta t \sum_{\ell=-\rho}^{0} \tilde{\delta}^n_{\ell} \eta^{\rho+\ell} = 0$ satisfy the root condition (4.3), (4.3a). From (4.4) it follows that $\tilde{q}(\eta)$ has the form
(4.5) \[ \tilde{q}(\eta) = \sigma_0(\eta) + \sum_{\nu=1}^{\mu} (\Delta t \lambda)^{\nu} \sigma_{\nu}(\eta), \]

where \( \sigma_0(\eta) \) is given by (4.2) and where \( \sigma_{\nu}(\eta) \in \mathbb{P} \).

The stability region \( S \) associated with a finite difference approximation of the form (2.2) is now defined as that region of the complex \( \Delta t \lambda \)-plane for which all roots \( \eta_\ell(\Delta t \lambda) \) of \( \tilde{q}(\eta) = 0 \) satisfy \( |\eta_\ell| < 1 \). Thus if a given difference approximation is applied to the equation \( w'(t) = \lambda w(t) \), and if \( \Delta t \lambda \in S \), then the approximate solution \( u^n \) will tend to zero as \( n \to \infty \) for any choice of initial data. For "stiff problems", i.e., for systems that have an eigenvalue \( \lambda \) with \( \text{Re}(\lambda) < 0 \), it is therefore desirable that the stability region \( S \) include a considerable portion of the negative half plane. In fact, for certain applications it is necessary that \( S \) contain the entire negative real axis. (See Section 6.) Note that an approximation satisfies the root condition if \( 0 \in \partial S \). Here \( \partial S \) is the boundary of \( S \). A necessary condition for the negative real axis to be contained in \( S \) is of course that the method is \textit{stable at } \infty, i.e., all roots of \( \sigma_\mu(\eta) = 0 \) lie on or inside the unit circle. (If the roots are strictly contained in the unit circle then the method will be called \textit{strictly stable at } \infty.) There are difference approximations however, that are strictly stable at \( \infty \), but for which the negative real axis is not contained in \( S \). (See Varah (1975).)

A related concept, introduced by Widlund (1967), is that of \textit{A(\alpha)} stability. A difference method is said to be \( \textit{A(\alpha)} - \text{stable} \) if \( S \) contains a region of the form

\[ S_{\alpha} \equiv \{ z : \pi - \alpha < \arg z < \pi + \alpha ; \ z \neq 0 \} \]
Thus if an approximation is $A(\alpha)$-stable, then the negative real axis is contained in $S_\alpha$ and hence in $S$.

The stability properties of expressions of the form (4.5), ("finite difference forms"), have been extensively investigated for the case $\mu = 1$ by, for example, Dahlquist (1959 and 1963), Gear (1971) and Varah (1975). The general case has been studied by Reimer (1968). The motivation for considering general finite difference forms is the fact that the stability analysis of many difference methods leads to studying such forms. Examples of these include methods based upon Padé rational approximation to the exponential, (see Varga (1961)), Runge Kutta methods and second derivative methods, (Enright (1974).) In this section it has been shown how these higher order finite difference forms arise in the stability analysis of the very general type of finite difference approximations considered in Sections 2 and 3. It is not the purpose of this section to extensively contribute to the investigations referred to above, but by means of some examples the effect that the choice of the points $\xi_k$ has on the stability region $S$ will be illustrated.

**Example 4.2.** Let $\rho = 1$ and $\mu = 1$ with $\xi_1 = t_{n-1} + \xi \Delta t$. So the difference approximation to $J \left( w(t) = w'(t) - \lambda w(t) = 0 \right)$ has the form

$$\tilde{\delta}_{-1} u^{n-1} + \tilde{\delta}_0 u^n = 0,$$

where the coefficients $\tilde{\delta}_{-1}$ and $\tilde{\delta}_0$ are found to be equal to

$$\tilde{\delta}_{-1} = \frac{-1}{\Delta t} - \lambda (1 - \xi) \quad \text{and} \quad \tilde{\delta}_0 = \frac{1}{\Delta t} - \lambda \xi .$$

Therefore the characteristic polynomial is given by
\( \tilde{q}(\eta) = \delta_0 \eta + \delta_{-1} = \sigma_0(\eta) + \Delta t \lambda \sigma_1(\eta) \), where \( \sigma_0(\eta) = \eta - 1 \) and \( \sigma_1(\eta) = -\xi \eta + \xi - 1 \).

The root of \( \sigma_1(\eta) = 0 \) is \( \eta_1 = (\xi - 1)/\xi \), so that \( |\eta_1| \leq 1 \) if and only if \( \xi \geq \frac{1}{2} \).

Thus this difference approximation is stable at \( \infty \) if \( \xi \geq \frac{1}{2} \). It is easy to check that the approximation is in fact A(\( \alpha \))-stable for such \( \xi \). The method is strictly stable at \( \infty \) only if \( \xi > \frac{1}{2} \). When applied to the general system (2.1) the choices \( \xi = 0, \xi = \frac{1}{2} \) and \( \xi = 1 \) correspond to the standard explicit one step Euler approximation, the Box scheme, (Keller (1969) ), and the totally implicit one step approximation.

**Example 4.3.** If one takes \( \rho = 2 \) and \( \mu = 1 \) with \( \xi_1 = t_{n-2} + \xi \Delta t \), then

\[ \tilde{q}(\eta) = \sigma_0(\eta) + \Delta t \lambda \sigma_1(\eta), \]

where \( \sigma_0(\eta) = (2 \xi - 1)\eta^2 + 4(1 - \xi)\eta + 2\xi - 3 \) and \( \sigma_1(\eta) = \xi (1 - \xi) \eta^2 + 2 \xi (\xi - 2)\eta - (\xi^2 - 3\xi + 2) \). In Example (4.1) it was found that this difference approximation is stable at zero if and only if \( \xi \geq 1 \), i.e., the roots of \( \sigma_0(\eta) = 0 \) lie on or inside the unit circle if and only if \( \xi \geq 1 \). The approximation is stable at \( \infty \) if the roots of \( \sigma_1(\eta) = 0 \) lie on or inside the unit circle. These roots are given by

\[ \eta = \frac{2\xi - \xi^2 \pm \sqrt{2\xi - \xi^2}}{\xi (1 - \xi)} \]

Some computation reveals that \( |\eta_1| \leq 1 \) provided \( \xi \geq 1 + \frac{1}{2} \sqrt{2} \approx 1.7 \). For example, the second order formula of Gear is strictly stable at \( \infty \), but the third order formula corresponding to \( \xi = 1 + \frac{1}{3} \sqrt{3} \), (see Example (3.2)), is not stable at \( \infty \).

5. **Initial Boundary Value Problems.**

This section is concerned with the construction of finite difference approximations to linear second order parabolic partial differential equations of the form
(5.1) \[ p(x, t) \frac{\partial y(x, t)}{\partial t} = L(t) y(x, t) + f(x, t), \quad 0 \leq x \leq 1, \quad 0 \leq t \leq T, \]

where

\[ L(t) y(x, t) = y_{xx}(x, t) + q(x, t) y_x(x, t) + r(x, t) y(x, t), \]

with initial condition

(5.1a) \[ y(x, 0) = g(x), \quad 0 \leq x \leq 1, \]

and boundary conditions

(5.1b) \[ a_0 y(0, t) + b_0 y_x(0, t) = g_0(t), \quad 0 \leq t \leq T, \]

and

(5.1c) \[ a_1 y(1, t) + b_1 y_x(1, t) = g_1(t), \quad 0 \leq t \leq T. \]

It is assumed that \( 0 < p < p(x, t) < p \leq \infty \), so that (1.4) is parabolic in the sense of Petrovskii (1938), and that the above problem has a unique solution, that is as many times continuously differentiable as necessary.

Define a mesh \( R_h(T) \) by

(5.2) \[ R_h(T) = \{(x_j, t_n) : 0 = x_0 < x_1 < \cdots < x_J = 1 ; 0 = t_0 < t_1 < \cdots < t_N = T \}. \]

The mesh spacings are \( h_j = x_j - x_{j-1}, \quad (1 \leq j \leq J) \), and \( (\Delta t)_n = t_n - t_{n-1}, \quad (1 \leq n \leq N) \).
Let \( h = \max_j h_j \) and \( \Delta t = \max_n (\Delta t)_n \). So for uniform meshes \( h_j = h = \frac{1}{J} \) and \( (\Delta t)_n = \Delta t = \frac{1}{N} \). For functions \( w(x,t) \) defined on \( \mathbb{R}^n \) let \( w^*_j = w(x_j, t_n) \),

\[
\begin{align*}
w^*_h &= (w^*_1, w^*_2, \ldots, w^*_n)^T, \\
\|w_h^*\| &= \max_j |w^*_j| \quad \text{and} \quad \|w_h^*\|_2 = \left\{ \sum_j (w^*_j)^2 \right\}^{1/2}.
\end{align*}
\]

Further, for matrices \( \|\cdot\| \) and \( \|\cdot\|_2 \) denote the matrix norms induced by the corresponding vector norms.

In principle it is possible to extend the construction method of Chapter I to partial differential equations. However, this approach will not be taken here, but instead discretization in space and time will be done separately. Discretizing in space first results in a system of linear first order ordinary differential equations. This system can then be solved by a difference method of the type discussed in Sections 2.3 and 4. More precisely this system is given by

\[
(5.3) \quad K^*_h(t) \left[ p(x_j, t)w'(x_j, t) \right] = L^*_h(t)w(x_j, t) + K^*_h(t)f(x_j, t), \quad 1 \leq j \leq J-1,
\]

where

\[
L^*_h(t)w(x_j, t) = \sum_{i=-r_j}^{s_j} d_{j,i}(t)w(x_{j+i}, t),
\]

and

\[
K^*_h(t)w(x_j, t) = \sum_{i=1}^{m_j} e_{j,i}(t)w(z_{j,i}, t),
\]

with initial data

\[
(5.3a) \quad w(x_j, 0) = g(x_j), \quad 0 \leq j \leq J.
\]
For each $t$ the coefficients $d_i(t)$ and $e_i(t)$ are defined in precisely the same manner as the coefficients $d_i$ and $e_i$ in Chapter I. Thus these coefficients are given by equations (1.2.10) and (I.2.11) with $L(t)$ replacing $L$. (To simplify the notation, the subscript $j$ in $d_{j,i}$, $e_{j,i}$ etc. will usually be omitted.)

The important assumption is made that the collocation points $z_i$ coincide with the mesh points $x_{j+i}$, i.e. if $z_i$ is a collocation point then $z_i = x_{j+i}$ for some integer $l$. For example, if $m = 1$ one may have

$$K_h(t) w(x_j, t) = e_1(t) w(x_j, t) = w(x_j, t), \quad 1 \leq j \leq J-1,$$

and if $m = 3$

$$K_h(t) w(x_j, t) = \sum_{i=-1}^{1} e_1(t) w(x_{j+i}, t), \quad 1 \leq j \leq J-1.$$

This restriction on the choice of the collocation points is necessary for (5.3) to constitute a proper system of ordinary differential equations.

The truncation error $\tau(t)$ associated with (5.3) is defined as

$$\tau(t) = \sum_{i=1}^{m} e_i(t) p(z_i, t) y_t(z_i, t) - \sum_{i=-r}^{s} d_i(t) y(x_{j+i}, t) - \sum_{i=1}^{m} e_i(t) f(z_i, t),$$

where $y(x, t)$ is the exact solution of (5.1) subject to (5.1a, b, c). Therefore, using the differential equation,

$$\tau(t) = \sum_{i=1}^{m} e_i(t) Ly(z_i, t) - \sum_{i=-r}^{s} d_i(t) y(x_{j+i}, t).$$

Since the coefficients $d_i(t)$ and $e_i(t)$ are defined as in Section (I.2) one may
apply Theorem (I. 2.1) and Theorem (I. 2.4) to arrive at the following.

**Theorem 5.2.** For each $t$, $(0 < t < T)$, let $d_i(t)$, $e_i(t)$ be defined by equations (I. 2.10) and (I. 2.11) respectively, with the normalizing factor $E$ as in (I. 2.12). Assume that for each $i$, $(1 \leq i \leq m)$, $z_i = x_{i+j\ell_i}$ for some integer $\ell_i$. Also assume that $r+s \geq n$ and that $E \neq 0$.

Then there exists a constant $C$ that does not depend on $h$, such that

$$|\tau(t)| \leq C h^{r+s+m-2}$$

In particular, if $r+s = n$ or if $m = 1$ then $E \neq 0$.

For each $t$, $(0 < t < T)$, there are also approximations to the boundary conditions (5.1b, c) of the form

$$(5.3b) \quad \sum_{i=0}^{s_0} d_{0,i} w(x_i, t) = g_0(t), \quad 0 \leq t \leq T,$$

and

$$(5.3c) \quad \sum_{i=-r_j}^{0} d_{j,i} w(x_{j+i}, t) = g_1(t), \quad 0 \leq t \leq T.$$ 

These approximations may be defined as in Section (I. 5), Equation (I.5.5), with $m = 0$. Thus these discrete boundary conditions are assumed to be independent of the differential equation. It is easy to show that in this case the coefficients $d_{0,0}$ and $d_{j,0}$ are nonzero. This allows elimination of the unknowns $w(x_0, t)$ and $w(x_j, t)$ in the system (5.3), so that one is left with a system of $J-1$ ordinary differential equations in the $J-1$ variables.
w(x_j, t), (1 \leq j \leq J-1). This reduced system will be compactly written as

\begin{equation}
K_h(t) w_h(t) = L_h(t) w_h(t) + \tilde{f}_h(t), \quad 0 \leq t \leq T,
\end{equation}

where \( w_h(t) \equiv (w(x_1, t), \cdots, w(x_{J-1}, t))^T \) and \( \tilde{f}_h(t) \) is the appropriate inhomogeneous term.

The initial data are

\begin{equation}
w_h(0) = g_h = (g(x_1), \cdots, g(x_{J-1}))^T.
\end{equation}

The discretization can now be completed by applying one of the methods for systems of ordinary differential equations discussed in Sections 2, 3 and 4. If the spatial discretization as well as the discretization in time are consistent, then it follows from the equivalence theorem of Lax, (see Richtmeyer and Morton (1967), p.45), that the solution of (5.4), (5.4a) converges to the solution of the continuous problem (5.1), (5.1a,b,c) if and only if the approximation is stable. The stability problem is discussed in the next section, whereas specific examples of finite difference approximations with numerical tests are given in Section 7.

6. THE STABILITY OF FINITE DIFFERENCE APPROXIMATIONS TO INITIAL BOUNDARY VALUE PROBLEMS

In this section the numerical stability of finite difference approximations to linear parabolic partial differential equations in one space variable will be investigated. For simplicity, we consider the diffusion equation
\[ y_t(x, t) = Ly(x, t) = y_{xx}(x, t), \quad 0 \leq x \leq 1, \quad 0 \leq t \leq T \]

with initial condition

\[ y(x, 0) = g(x), \quad 0 \leq x \leq 1, \]

and boundary conditions

\[ y(0, t) = 0, \quad 0 \leq t \leq T, \]

and

\[ y_x(1, t) = 0, \quad 0 \leq t \leq T. \]

Discretizing in space first as in Section 5 gives the equations

\[ \sum_{i=1}^{m_j} e_{j,i} w'(z_{j,i}, t) = \sum_{i=-r_j}^{s_j} d_{j,i} w(x_{j,i+1}, t), \quad 1 \leq j \leq J-1, \quad 0 \leq t \leq T, \]

where the coefficients \( d_{j,i} \) and \( e_{j,i} \) are given by equation (1.2.16) and (1.2.17) respectively, and where the collocation points \( z_{j,i} \) are required to coincide with the mesh points. Further, the mesh is assumed to be uniform, so that \( h_j = h = \frac{1}{J} \) and \( (\Delta t)_n = \Delta t = \frac{1}{N} \). The boundary condition (6.1c) is approximated by

\[ \sum_{i=-r_J}^{0} b_{j,i} w(x_{j+i}, t) = 0, \quad 0 \leq t \leq T. \]

As in the previous section, one may use the discrete boundary conditions
to eliminate \( w(x_0, t) \) and \( w(x_J, t) \) from (6.2). The reduced system of ordinary differential equations is then written as

\[
(6.3) \quad K_h w_h'(t) = L_h w_h(t),
\]

with initial data

\[
(6.3a) \quad w_h(0) = g_h.
\]

Before discretizing in time sufficient conditions for the solution of (6.3), (6.3a) to converge to the solution of (6.1), (6.1a, b, c) will be stated.

**Theorem 6.1.** The solution \( w_h(t) \) of (6.3), (6.3a) converges to the solution of (6.1), (6.1a, b, c), and this discretization is stable in the \( l_2 \)-norm with respect to perturbations in the initial data, provided that there exists a constant \( c_2 \) that does not depend on \( h \), such that for all small enough \( h \) the following conditions are satisfied.

1. The spatial discretization (6.3) is consistent with (6.1), (6.1b, c).
2. \( K_h \) is nonsingular.
3. The eigenvalue problem \( L_h v_h = \lambda_h K_h v_h \) admits a complete set of eigenvectors.
4. The eigenvalues \( \lambda_h, i \) \((1 \leq i \leq J-1)\), of the above eigenvalue problem satisfy \( \text{Re}(\lambda_h, i) \leq 0 \).
5. Let \( V_h \) be the matrix of eigenvectors. Then \( \|V_h\|_2 \|V_h^{-1}\|_2 \leq c_2 \).

**Proof.** Let \( e_h(t) = y_h(t) - w_h(t) \). Then \( K_h e_h'(t) = L_h e_h(t) + \tau_h(t) \). Therefore
\[ e_h(t) = e^{-tK_h^{-1}L_h}e_h(0) + \int_0^t e^{(t-r)K_h^{-1}L_h}\tau_h(r)\,dr \]
\[ = V_h e^{tA_h}V_h^{-1}e_h(0) + \int_0^t V_h e^{(t-r)A_h}V_h^{-1}\tau_h(r)\,dr, \]

where \( A_h \) is the diagonal matrix of eigenvalues.

If \( n_g \) is the order of consistency of (6.3), then the estimate
\[
\|e_h(t)\|_2 \leq c_2 \left\{ \|e_h(0)\|_2 + T \cdot \max_{0 \leq t \leq T} \|\tau_h(t)\|_2 \right\}
\]
follows immediately from the assumptions.

A fully discretized finite difference approximation to (6.1), (6.1a, b, c) is obtained by applying one of the class of methods discussed in Sections 2, 3 and 4 to the system (6.3). From the discussion in Example (3.4) it follows that if \( L_h \) and \( K_h \) commute then this approximation has the form
\[
(6.4) \quad \frac{1}{\Delta t} \sum_{\ell=0}^N \left[ \sum_{\nu=0}^\infty a^\ell_\nu (\Delta t)^\nu K_h^{\ell-\nu} L_h^\nu \right] u_n^{n+\ell} = 0, \quad 1 \leq n \leq N.
\]
The coefficients \( a^\ell_\nu \) are assumed to be independent of \( n \), so that the same difference approximation is used at each time-step. The initial data are \( u_0^h = g_h \). Further, if \( \rho > 1 \) then sufficiently accurate approximations to \( u_n^h \), \( (1 \leq n \leq \rho-1) \), must be given also. These may be obtained for example by initially using a one step method, \( (\rho=1) \), having the same order of consistency as (6.4).

The difference approximation (6.4) is said to be \textit{stable in the} \( l_2 \)-\textit{norm} if there exists a constant \( K \), that does not depend on \( h \) and \( \Delta t \), such that
for all small enough \( h \) and \( \Delta t \). Sufficient conditions for stability, and hence for convergence, are given in the following theorem.

**Theorem 6.2.** The solution \( u^n_h \) of (6.4), with appropriate initial conditions, converges to the solution of (6.1), (6.1a, b, c), and this difference scheme is stable in the \( l_2 \)-norm with respect to perturbation in the initial data, provided that there exist positive constants \( c_1 \) and \( c_2 \), that do not depend on \( h \), such that for all small enough \( h \) the following conditions are satisfied.

1. The difference scheme (6.4) is consistent.
2. \( K_h \) is nonsingular.
3a. \( K_h \) and \( L_h \) commute.
3. The eigenvalue problem \( L_h v_h = \lambda_h K_h v_h \) admits a complete set of eigenvectors.
4. The eigenvalues \( \lambda_{h,i} \) of the above eigenvalue problem lie in a region \( T_\alpha \) of the form

\[
T_\alpha = \{ z \in \mathbb{C} : \pi - \alpha < \arg z < \pi + \alpha ; \ z \neq 0 \}
\]

which has the property that if \( z \in T_\alpha \), then the roots \( \eta_{1i}(z) \) of the characteristic equation

\[
\tilde{q}(\eta) \equiv \sum_{\mu} \sum_{\ell=-\rho}^{\rho} a_{\nu} \nu^\ell \eta^{\rho+\ell} = 0
\]

satisfy
\[ |\eta_1(z)| < 1, \]
and
\[ |\eta_\ell(z)| \leq 1 - c_1, \quad 2 \leq \ell \leq \rho. \]

5. Let \( V_h \) be the matrix of eigenvectors. Then

\[ \|V_h\|_2 \|V_h^{-1}\|_2 \leq c_2. \]

**Proof.** By assumption 3 one can write \( u^n = V_h c^n_h \). Substitution into (6.4), also using 2 and 2a, gives

\[ \frac{1}{\Delta t} K_h^\mu \sum_{\ell=-\rho}^{\mu} \sum_{\nu=0}^\mu a^\ell \Delta t^\nu K_h^{-\nu} L_h^\nu V_h c^{n+\ell}_h \]

\[ = \frac{1}{\Delta t} K_h^\mu \sum_{\ell=-\rho}^{\mu} \sum_{\nu=0}^\mu a^\ell \Delta t^\nu V_h \Lambda_h^\nu c^{n+\ell}_h \]

\[ = \frac{1}{\Delta t} K_h^\mu V_h \sum_{\ell=-\rho}^{\mu} \sum_{\nu=0}^\mu a^\ell \Delta t^\nu \Lambda_h^\nu c^{n+\ell}_h = 0, \]

where \( \Lambda_h \) is the diagonal matrix of eigenvalues. Hence each component \( c^n_{h,j} \) of \( c^n_h \) satisfies the difference equation

\[ \sum_{\ell=-\rho}^{\mu} \sum_{\nu=0}^\mu a^\ell (\Delta t \lambda_{h,j})^\nu c^{n+\ell}_h, j = 0, \quad \rho \leq n \leq N = T/\Delta t. \]

Let \( c^n_{h,j} \) denote the vector \((c^n_{h,j}, c^{n-1}_{h,j}, \ldots, c^{n+\rho+1}_{h,j})^T\) and let

\[ \|c^n_{h,j}\|_2 = \left\{ \sum_{m=0}^{\rho-1} |c^{n-m}_{h,j}|^2 \right\}^{\frac{1}{2}}. \]

Define the \( \rho \times \rho \) matrix \( A(z) \) by
The eigenvalues $\eta_{m}(\Delta t \lambda_{h,j})$ of $A(\Delta t \lambda_{h,j})$ are precisely the roots of the characteristic equation

$$\tilde{q}(\eta) = \sum_{\ell=-\rho}^{\mu} \sum_{\nu=0}^{\mu} a_{\nu}^\ell (\Delta t \lambda_{h,j})^\nu \eta^{\rho+\ell} = 0$$

The definition of $T_{\alpha}$ implies that if $\lambda_{h,j} \in T_{\alpha}$ then $\Delta t \lambda_{h,j} \in T_{\alpha}$. From assumption 4 it now follows that

$$|\eta_{1}(\Delta t \lambda_{h,j})| \leq 1,$$

and

$$|\eta_{m}(\Delta t \lambda_{h,j})| \leq 1-c_{1}, \quad 2 \leq m \leq \rho,$$

for all $j$, ($1 \leq j \leq J-1$).

It is well known, (See Richtmeyer and Morton (1967), p.86), that
this implies that the family of matrices \( \{ A(z) : z \in T_\alpha \} \) satisfies

\[
\| (A(z))^n \|_2 \leq K_1, \quad \rho \leq n \leq N.
\]

Thus under these conditions one has

\[
|c_{h,j}^n| \leq \|c_{h,j}^n\|_2 = \| (A(\Delta t \lambda_h, j))^{n-\rho+1} \|
\]

and therefore

\[
\|c_h^n\|_2 = \left\{ \sum_{j=1}^{J-1} |c_{h,j}^n|^2 \right\}^{\frac{1}{2}} \leq K_1 \left\{ \sum_{j=1}^{J-1} \|c_{h,j}^{\rho-1}\|_2 \right\}^{\frac{1}{2}} = K_1 \left\{ \sum_{j=1}^{J-1} \sum_{m=0}^{\rho-1} |c_{h,j}^m|^2 \right\}^{\frac{1}{2}} = K_1 \left\{ \sum_{m=0}^{\rho-1} c_{h,m}^m \right\}^{\frac{1}{2}}
\]

so that

\[
\|u_h^n\|_2 = \|V_h \|c_h^n\|_2 \leq \|V_h\|_2 \|c_h^n\|_2 \leq \|V_h\|_2 \sum_{m=0}^{\rho-1} \|c_h^m\|_2 \frac{1}{2}
\]

\[
= \|V_h\|_2 K_1 \left\{ \sum_{m=0}^{\rho-1} \|V_h u_h^m\|_2 \right\}^{\frac{1}{2}} \leq \|V_h\|_2 \|V_h^{-1}\|_2 \sum_{m=0}^{\rho-1} \|u_h^m\|_2 \frac{1}{2}
\]

\[
\leq c_2 K_1 \left\{ \sum_{m=0}^{\rho-1} \|u_h^m\|_2 \right\}^{\frac{1}{2}}.
\]

This proves the theorem.

Before proceeding to analyze a number of examples, some remarks are in order. First, it should be noted, that conditions 3, 4 and 5 of Theorem (6.1) and Theorem (6.2) are in general difficult to verify analytically. However, a good check on whether a proposed discretization leads to a stable difference scheme, is obtained by numerical computation of the eigenvalues and the
condition number $\kappa(V_h) = \|V_h\|_2 \|V_h^{-1}\|_2$ for a few values of $h$.

The assumption that all eigenvalues lie in the left half plane is not strictly necessary, but reasonable in view of the differential equation under consideration. In problems where there are some, but not more than a fixed finite number of eigenvalues with positive real part, one need only add the requirement that the modulus of these eigenvalues can be uniformly bounded.

The essential difference between the stability analysis in this section and the stability analysis of difference methods for systems of ordinary differential equations in Section 4 is the fact that the current system, (equation (6.3)), does not have a fixed dimension. This results in the requirements that $\kappa(V_h)$ be bounded and that the stability region admit a region $T_a$ as in Theorem (6.2). A multistep method that admits such a region will be called strongly $\text{A}(a)$-stable. (See Section 4 for a definition of $\text{A}(a)$ stability.) Thus if a method is strongly $\text{A}(a)$-stable, then it is $\text{A}(a)$-stable, but the converse need not be true. All $\text{A}(a)$-stable one step methods are strongly $\text{A}(a)$-stable, because the characteristic equation $\tilde{q}(\eta) = 0$ has only one root in that case. Other examples of strongly $\text{A}(a)$-stable methods are the formulas of Gear, (see example (3.2) and example (I.4.2)), and the second derivative methods given in Enright (1974). These methods are used in the numerical examples of Section 7.

Condition 4 of Theorem (6.2) is closely related to the notion of parabolicity of a difference scheme, introduced by Widlund (1965 and 1966). In these papers the stability of very general multistep methods for linear parabolic systems without boundary conditions is considered, with Fourier transforms replacing the eigenvalue-eigenvector approach in this section. (See also Hakberg (1970) and Nordmark (1974).) The major disadvantage
of the current eigensystem analysis is perhaps the fact that the assumption of a complete set of eigenvectors may not be satisfied. An advantage is that the effect of various approximations to the boundary conditions can be investigated numerically. These effects have also been investigated by Varah for the scalar diffusion equation, (1970, 1971), and for parabolic systems, (1971a). For applications of the stability theory of Varah see Gottlieb and Gustafsson (1975). Earlier work on the stability of parabolic systems with boundary conditions was done by Osher (1968). The sufficient conditions for stability of one step methods, as given by Varah, are first that the scheme must be stable when applied to a pure initial value problem. (Thus the work of Widlund can be used to check this condition.) Further the amplification matrix $Q = [Q_0]^{-1}Q_{-1}$ of the difference scheme $Q_0 u^n_h = Q_{-1} u^{n-1}_h$ should have no eigenvalues $z$ with $|z| > 1$, and in addition, some condition on the spectrum of $Q$ near $z = 1$ must be satisfied. It is also shown by Varah how one can characterize eigenvalues $z$ with $|z| > 1$. Essentially the same characterization is obtained when analyzing the case where the eigenvalue problem $L_h v_h = \lambda_h K_h v_h$ of this section has eigenvalues $z$ with $\text{Re}(z) > 0$. This will be considered again in Example (6.3) and (6.4). It is not clear how the condition on the spectrum of $Q$ relates to the current eigensystem analysis.

The purpose of the next example is to indicate why it is desirable to restrict to discretizations in time which have a stability region that includes at least the entire negative real axis.

Example 6.3. It follows from the general theory of Kreiss (1972), that each eigenvalue of the eigenvalue problem

$$Ly = y'' = \lambda y, \quad y(0) = y'(1) = 0,$$
is approximated by an eigenvalue of the discrete eigenvalue problem

\[ L_h v_h = \lambda_h K_h v_h, \]

as \( h \to 0 \), provided that \( L_h u_h = K_h f_h \) defines a consistent and stable finite difference approximation to \( Ly = f \). (See Section (I.7).) Since the eigenvalues of (6.6) are given by \( \lambda_k = \left[ \frac{(2k-1)\pi}{2} \right]^2 \), \( k \geq 1 \), it follows that there are eigenvalues \( \lambda_h \) of (6.7) such that \( \lambda_h \to -\infty \) as \( h \to 0 \). Hence the restriction to discretizations in time with an unbounded region \( T_\alpha \) as in Theorem (6.2) is desirable. If \( T_\alpha \) were bounded then a restriction on the size of the timestep would have to be imposed. For example, consider the case where \( L_h u_j = \frac{1}{h^2} (u_{j-1} - 2u_j + u_{j+1}) \) and \( K_h u_j = u_j \), \( 1 \leq j \leq J-1 \). The discrete boundary conditions are \( u_0 = 0 \) and \( (u_{J} - u_{J-1})/h = 0 \). Then the eigenvalues of (6.7) are easily found to be given by

\[ \lambda_{h,k} = \frac{-4}{h^2} \sin^2 \left( \frac{(2k-1)\pi}{2(2J-1)} \right), \quad 1 \leq k \leq J-1. \]

Thus if the region \( T_\alpha \) includes only a part \((-\beta, 0)\) of the negative real axis then \( \Delta t \) must satisfy \( \Delta t \leq \frac{\beta}{4} h^2 \) in order to guarantee that \( \Delta t \lambda_{h,1} \notin T_\alpha \).

The stability of particular finite difference approximations will be investigated in the next two examples as well as in numerical examples in the next section. Since one can always apply a strongly \( A(\alpha) \)-stable method to the system (6.3), it is necessary only to study the effect of various spatial difference approximations on the eigenvalues of the eigenvalue problem (6.7). Specifically, it is important to check that this discrete eigenvalue problem does not allow eigenvalues \( \lambda_h \) that tend to \( \infty \) in the right half plane and that the condition number \( \kappa(V_h) \) is bounded.
Example 6.4. Although stability and consistency of the spatial difference approximation imply that the eigenvalues of (6.6) are approximated by those of the discrete eigenvalue problem (6.7), it need not be true that all eigenvalues of (6.7) converge to eigenvalues of (6.6).

Consider for example the case where \( L^h \) and \( K^h \) are as in the previous example, but where the boundary condition at \( x = 1 \) is approximated by

\[
\frac{u_{J-1} - (c-1) u_{J-2}}{(2-c) h} = 0
\]

This approximation is consistent with \( y'(1) = 0 \), provided that \( c \neq 2 \). The order of consistency is equal to one. If \( c = \frac{4}{3} \) then the order of consistency is equal to two. (See also Example 1.8.4') For this discretization the operator \( K^h \) in (6.3) is the identity and the operator \( L^h \) can be written as

\[
L^h = P^h \tilde{L}^h,
\]

where

\[
P^h = \begin{bmatrix}
1 & & \\
& 1 & \\
& & \ddots \\
& & & 1 \\
& & & & 1 - 2c
\end{bmatrix}
\quad \text{and} \quad
\tilde{L}^h = \begin{bmatrix}
-2 & 1 & & & \\
1 & -2 & 1 & & \\
& 1 & -2 & 1 & \\
& & \ddots & \ddots & \\
& & & 1 & -1
\end{bmatrix}
\]

Let \( v^h = P^h \tilde{v}^h \). Then the eigenvalue problem \( L^h v^h = \lambda^h K^h v^h \) becomes

\[
P^h \tilde{L}^h P^h v^h = \lambda^h \tilde{v}^h,
\]

which admits a complete set of orthonormal eigenvectors with real eigenvalues. The matrices \( v^h \) and \( \tilde{v}^h \) containing the eigenvectors \( v^h,i \) and \( \tilde{v}^h,i \) respectively by column are then related by \( V^h = P^h \tilde{V}^h \). Hence

\[
\kappa(V^h) \equiv \| V^h \|_2 \| V^{-1}^h \|_2 \leq \| \tilde{V}^h \|_2 \| \tilde{V}^{-1}^h \|_2 \| P^h \|_2 \| P^{-1}^h \|_2 \leq \max\{ \frac{1}{|c-2|}, \frac{1}{|c-2|} \},
\]
provided $c \neq 2$. This shows that the last condition of Theorem (6.1) and
Theorem (6.2) is satisfied for this discretization in space when $c \neq 2$. Next
consider the eigenvalues. Application of the Gersgorin theorem,
(Lancaster (1969), p. 226), shows that all eigenvalues are negative if and
only if $c < 2$. If $c = 2$ then there is an eigenvalue at zero and if $c > 0$
then there is a strictly positive eigenvalue. So assume that $c > 2$, $\lambda_h > 0$
and let $v_h$ be the eigenvector associated with $\lambda_h$. Thus $v_h$ satisfies
$$v_{j+1} - (2 + h^2 \lambda) v_j + v_{j-1} = 0, \quad 1 \leq j \leq J-1.$$ 

The solution to these difference equations is
$$v_j = c_1 z^j_h + c_2 z^{-j}_h, \quad 0 \leq j \leq J.$$

where $z_h$ and $1/z_h$ are the roots of
(6.8) \[ z^2 - (2 + h^2 \lambda)z + 1 = 0 \]

But $v_h$ must also satisfy the discrete boundary conditions. In particular,
the condition $v_0 = 0$ implies that $v_j = c_1 (z^j_h + z^{-j}_h), \quad (0 \leq j \leq J)$, and the
boundary condition at $x = 1$ leads to the equation
(6.9) \[ (z^J_h - z^{-J}_h - c(z^{J-1}_h - z^{-J+1}_h) + (c-1)(z^{J-2}_h - z^{-J+2}_h) = 0. \]

Consider now the equation obtained from (6.9) by omitting all negative
powers, viz.
$$z^2_h - cz_h + c-1 = 0. \]

This is the characteristic equation of the boundary condition at $x = 1$. Its
roots are $z_1 = 1$ and $z_2 = c-1$. Since by assumption $c-1 > 1$ it follows that
there is a root $z_h$ of (6.9) such that $z_h \to c-1$ as $h \to 0$. The
corresponding eigenvalue is obtained from (6.8) and asymptotically given by

$$\lambda_h = \frac{(z_h - 1)^2}{h^2 z_h} \approx \frac{(c-1)^2}{h^2 (c-1)}.$$ 

Thus $\lambda_h \to + \infty$ as $h \to 0$. This eigenvalue computation is essentially equivalent to the technique of Varah when applied to the current discretization (see Varah (1970), p. 33, and Varah (1971).) There the effect of the discrete boundary conditions is related directly to the eigenvalues of the amplification matrix, whereas here these effects are related to the eigenvalues of $K_h^{-1}L_h$. If the discretization is completed by applying a
multistep method to the system $K_h u_{h+1} = L_h u_h$, and if this multistep method is strongly A(α)-stable, but the positive real axis is not contained in its stability region, then the resulting scheme is stable if and only if $c < 2$. The simplest example of such a multistep method is the trapezoidal rule. (See Example (I.4.1).) For this method the stability region is the entire negative half-plane. That this scheme is unstable for $c > 2$ can be checked by applying it to the initial data $g_h = v_h$, where $v_h$ is the eigenvector associated with the positive eigenvalue $\lambda_h$. There are discretizations in time for which the complete difference scheme will be stable for small enough $h$, even if $c > 2$. In fact, all methods that are strictly stable at $\infty$ have this property.

**Example 6.5.** Again consider the eigenvalue problem (6.6) and its discrete version (6.7) with

$$L_h v_j = \frac{1}{h^2} (v_{j-1} - 2v_j + v_{j+1})$$
Thus this eigenvalue problem arises in the stability analysis when using the fourth order spatial difference approximation discussed in Example (1.4.4). Let the approximation to the boundary condition $y'(1) = 0$ have the form

\[ \sum_{i=-r_j}^{0} b_i u_{j+i} = 0 \]  

(6.10)

It is now difficult to check $\kappa(V_h)$, so only the eigenvalues will be considered. Proceeding as in the previous example, it is found that the eigenvalues are given by

\[ \lambda_h = \frac{12(z-1)^2}{h^2(z^2 + 10z + 1)} \]  

(6.11)

for each $z$ which is a root of

\[ \sum_{i=-r_j}^{0} b_i (z^{j+i} - z^{-j-i}) = 0. \]  

(6.12)

Note that if $z$ is a root of (6.12) then so is $\frac{1}{z}$. The characteristic polynomial of the discrete boundary condition (6.10) is given by

\[ c_j(z) = \sum_{i=-r_j}^{0} b_i z^{j+i} \]

Now equation (6.12) has a root $z$ satisfying $|z| > 1 + \epsilon$, $J$ large, if and only if the characteristic equation $c_j(z) = 0$ has a root $|z| > 1$. Further, if $z$ is a root of (6.12) with $|z| = 1$, then an easy computation shows that
the corresponding eigenvalue is real and negative. Thus it is sufficient to check that none of the roots of the characteristic equation of the discrete boundary condition exceeds one in absolute value.

As in the previous example the second order approximation

\[
\frac{1}{h} \left( \frac{3}{2} u_J - 2u_{J-1} + \frac{1}{2} u_{J-2} \right) = 0
\]

satisfies the eigenvalue condition because the roots of its characteristic equation are \( z_1 = 1 \) and \( z_2 = \frac{1}{3} \). Similarly, the characteristic equation of the third order approximation

\[
\frac{1}{6h} \left( 11 u_J - 18u_{J-1} + 9 u_{J-2} - 2u_{J-3} \right) = 0
\]

has roots \( z_1 = 1 \), \( z_2 = \frac{7 + \sqrt{39} i}{22} \) and \( z_3 = \frac{7 - \sqrt{39} i}{22} \) with \( |z_2| = |z_3| \approx 0.426 \). The characteristic equation of the fourth order approximation

\[
\frac{1}{12h} \left( 25 u_J - 48u_{J-1} + 36u_{J-2} - 16u_{J-3} + 3u_{J-4} \right) = 0
\]

has roots \( z_1 = 1 \), \( z_2 \approx 0.381 \), \( z_3 \approx 0.269 + 0.492 i \) and \( z_4 \approx 0.269 - 0.492 i \) with \( |z_3| = |z_4| \approx 0.561 \). Therefore, these difference equations also satisfy the eigenvalue condition.

7. **NUMERICAL EXAMPLES**

Example 7.1. The purpose of this example is to indicate the relative accuracy of various finite difference approximations. For this purpose consider the simple problem

\[
y_t = y_{xx} + (1 + \pi^2) y , \quad 0 \leq x \leq 1 , \quad 0 \leq t \leq 3 ,
\]
with initial condition

\[ y(x,0) = \sin \pi x, \quad 0 \leq x \leq 1, \]

and boundary conditions

\[ y(0,t) = y(1,t) = 0, \quad 0 \leq t \leq 3. \]

The solution is \( y(x,t) = e^{t\sin(\pi x)}. \) The maximum value this solution reaches is \( y(1/2,3) = e^3 \approx 20. \) Uniform meshes are used with \( h = \frac{1}{j} \) and \( \Delta t = \frac{1}{N}. \) Finite difference approximations have the form

\[
\frac{1}{\Delta t} \sum_{\ell=-\rho}^{0} \left[ \sum_{\nu=0}^{\mu} a_{\nu}^{\ell} (\Delta t)^{\nu} K_{h}^{u_{\nu}} L_{h}^{\nu} \right] u_{h}^{n+\ell} = 0,
\]

where

\[ L_{h}^{u_{j}} = \sum_{i=-s}^{s} d_{j} u_{i}^{n}, \]

and

\[ K_{h}^{u_{j}} = \sum_{i=1}^{m} e_{i} u^{n}(z_{j,i}). \]

The collocation points \( z_{j,i} \) coincide with the meshpoints. If \( m = 1 \) then \( z_{j,i} = x_{j} \) and if \( m = 3 \) then \( z_{j,i} = x_{j-2+i}, \) \((1 \leq i \leq 3). \) For the given problem the choice \( m = 3 \) corresponds to the "Mehrstellenverfahren" of Collatz. (See equation (I.1.7) and Example (I.4.4).)

Results appear in Table (7.1). The spatial difference approximation in experiment 5 is not compact. Thus near the boundaries other difference
| 1 | 1 1 1 2 | 1 C 2 | 1 1 1 2 | 2 G 2 | 1 1 3 4 | 1 P 4 | 1 1 3 4 | 4 G 4 | 2 2 1 4 | 4 G 4 | N |
|---|---|---|---|---|---|---|---|---|---|---|
| # | r s m o | p c o | J = 5 | J = 10 | J = 20 | J = 40 | N |
| 1 | .392±2 | .772±1 | .279±1 | .170±1 | 6 |
| 2 | .320±2 | .585±1 | .149±1 | .517 0 | 15 |
| 3 | .311±2 | .560±1 | .132±1 | .360 0 | 30 |
| 4 | .599±3 | .861±2 | .523±2 | .459±2 | 6 |
| 5 | .755±2 | .185±2 | .106±2 | .888±1 | 15 |
| 6 | .475±2 | .108±2 | .516±1 | .391±1 | 30 |
| 7 | .382±2 | .795±1 | .306±1 | .197±1 | 60 |
| 8 | .343±2 | .669±1 | .212±1 | .111±1 | 120 |
| 9 | .325±2 | .609±1 | .168±1 | .701 0 | 240 |
| 10 | .371 0 | .189±1 | .380±2 | .521±2 | 6 |
| 11 | .376 0 | .241±1 | .138±2 | .399±4 | 15 |
| 12 | .308 0 | .242±1 | .150±2 | .859±4 | 30 |
| 13 | .299 0 | .223±1 | .149±2 | .154±3 | 60 |
| 14 | .229±1 | .147±2 | .959±4 | .932±4 | 120 |
| 15 | .120±3 | .158 0 | .151 0 | .150 0 | 6 |
| 16 | .212±1 | .245±1 | .135±1 | .109 0 | 15 |
| 17 | .243±1 | .167±1 | .411±2 | .112±2 | 30 |
| 18 | .169±1 | .348±2 | .299±3 | .245±3 | 60 |
| 19 | .353±2 | .245±3 | .245±3 | .245±3 | 120 |
| 20 | .197±1 | .391±1 | .111±1 | .701 0 | 240 |
### TABLE 7.1 (Continued)

<table>
<thead>
<tr>
<th>#</th>
<th>r s m o</th>
<th>p c o</th>
<th>J = 5</th>
<th>J = 10</th>
<th>J = 20</th>
<th>J = 40</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1 1 3 4</td>
<td>2 S 4</td>
<td>.298&lt;sup&gt;0&lt;/sup&gt;</td>
<td>.527&lt;sup&gt;-2&lt;/sup&gt;</td>
<td>.136&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>.148&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.351&lt;sup&gt;0&lt;/sup&gt;</td>
<td>.222&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>.978&lt;sup&gt;-3&lt;/sup&gt;</td>
<td>.343&lt;sup&gt;-3&lt;/sup&gt;</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.234&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>.143&lt;sup&gt;-2&lt;/sup&gt;</td>
<td>.631&lt;sup&gt;-4&lt;/sup&gt;</td>
<td>.909&lt;sup&gt;-4&lt;/sup&gt;</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>1 1 3 4</td>
<td>1 P 6</td>
<td>.377&lt;sup&gt;0&lt;/sup&gt;</td>
<td>.243&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>.152&lt;sup&gt;-2&lt;/sup&gt;</td>
<td>.108&lt;sup&gt;-3&lt;/sup&gt;</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.377&lt;sup&gt;0&lt;/sup&gt;</td>
<td>.242&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>.151&lt;sup&gt;-2&lt;/sup&gt;</td>
<td>.947&lt;sup&gt;-4&lt;/sup&gt;</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.351&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>.342&lt;sup&gt;-2&lt;/sup&gt;</td>
<td>.330&lt;sup&gt;-2&lt;/sup&gt;</td>
<td>.944&lt;sup&gt;-4&lt;/sup&gt;</td>
<td>30</td>
</tr>
<tr>
<td>8</td>
<td>1 1 3 4</td>
<td>6 G 6</td>
<td>.351&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>.536&lt;sup&gt;-2&lt;/sup&gt;</td>
<td>.116&lt;sup&gt;-2&lt;/sup&gt;</td>
<td>.268&lt;sup&gt;-3&lt;/sup&gt;</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.238&lt;sup&gt;0&lt;/sup&gt;</td>
<td>.198&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>.124&lt;sup&gt;-2&lt;/sup&gt;</td>
<td>.824&lt;sup&gt;-4&lt;/sup&gt;</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.307&lt;sup&gt;0&lt;/sup&gt;</td>
<td>.156&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>.116&lt;sup&gt;-2&lt;/sup&gt;</td>
<td>.858&lt;sup&gt;-4&lt;/sup&gt;</td>
<td>30</td>
</tr>
<tr>
<td>9</td>
<td>1 1 3 4</td>
<td>4 S 6</td>
<td>.186&lt;sup&gt;0&lt;/sup&gt;</td>
<td>.116&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>.334&lt;sup&gt;-3&lt;/sup&gt;</td>
<td>.370&lt;sup&gt;-3&lt;/sup&gt;</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.300&lt;sup&gt;0&lt;/sup&gt;</td>
<td>.194&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>.120&lt;sup&gt;-2&lt;/sup&gt;</td>
<td>.719&lt;sup&gt;-4&lt;/sup&gt;</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.218&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>.136&lt;sup&gt;-2&lt;/sup&gt;</td>
<td>.848&lt;sup&gt;-4&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1 1 3 4</td>
<td>1 P 8</td>
<td>.377&lt;sup&gt;0&lt;/sup&gt;</td>
<td>.242&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>.151&lt;sup&gt;-2&lt;/sup&gt;</td>
<td>.663&lt;sup&gt;-4&lt;/sup&gt;</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.377&lt;sup&gt;0&lt;/sup&gt;</td>
<td>.242&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>.151&lt;sup&gt;-2&lt;/sup&gt;</td>
<td>.109&lt;sup&gt;-3&lt;/sup&gt;</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.101&lt;sup&gt;-3&lt;/sup&gt;</td>
<td>30</td>
</tr>
</tbody>
</table>
formulas must be used. The approximation at $x = x^*$ is defined by the values $r = 1$, $s = 4$ and $m = 1$, while for the approximation at $x = x^*_{J-1}$ these values are $r = 4$, $s = 1$ and $m = 1$. The order of consistency of these formulas is equal to four, as is the order of the main finite difference approximation. Stability of this particular discretization will be considered in Example (7.5). The discretization in time is defined by the number of steps "$p"$, the order of consistency "$o"$, and a code "c", indicating the type of method used.

"C" is the trapezoidal rule. (See Example (1.4.1) and Example (3.1).) This approximation is usually referred to as a Crank-Nicolson discretization in time.

"G" denotes a method of Gear. The coefficients $a^l_v$ of these methods can be found in Gear (1971), p. 217. (See also, Example (1.4.2) and Example (3.2).)

"P" stands for a Padé method, based upon a diagonal entry in the Padé table of rational approximations to the exponential function. The coefficients of this discretization in time may be obtained from Varga (1961). (See Example (3.3) for a fourth order Padé formula.)

Finally, "S" denotes a so-called second derivative method. These methods have the form (7.1) with $\mu = 2$ and can be found in Enright (1974). The coefficients $a^l_v$ are determined from consistency and $A(\alpha)$-stability requirements.

Example 7.2. In this example uniform as well as non-uniform meshes are employed to solve the equation

$$y_t = \frac{\epsilon}{2} x u_{xx} - x(1-x) u_x + xu, \quad 0 \leq x \leq 1, \quad t \geq 0.$$
with initial condition

\[ u(x,0) = \sin \pi x \quad 0 \leq x \leq 1 \]

and boundary conditions

\[ u(0,t) = u(1,t) = 0. \]

The discretization in space corresponds to the second case discussed in Example (1.4.3). If the mesh is uniform this formula reduces to the standard second order three-point formula. If the mesh is not uniform, then the collocation point \( z_{j,1} = x_j + \frac{(h_{j+1} - h_j)}{3} \) does not coincide with \( x_j \).

To circumvent this difficulty, the operator \( K_h \) is redefined in terms of the Lagrangian interpolation formula. More precisely, \( K_h \) is given by

\[ K_h u_j = w^0(z_{j,1})u_{j-1} + w^1(z_{j,1})u_j + w^2(z_{j,1})u_{j+1}, \]

where the Lagrangian basis functions \( w^i(x) \) are defined by

\[ w^0(x) = \frac{(x-x_j)(x-x_{j+1})}{h_j(h_j+h_{j+1})}, \quad w^1(x) = -\frac{(x-x_{j-1})(x-x_{j+1})}{h_j h_{j+1}} \quad \text{and} \]

\[ w^2(x) = \frac{(x-x_{j-1})(x-x_j)}{h_{j+1}(h_j+h_{j+1})}. \]

This yields the expression

\[ K_h u_j = \frac{h_{j+1} h_j}{9 h_j (h_j+h_{j+1})} u_{j-1} + \frac{h_{j+1} (2h_j+h_{j+1})}{9 h_j h_{j+1}} u_j + \frac{h_j (2h_{j+1}+h_j)}{9 h_{j+1} (h_j+h_{j+1})} u_{j+1}. \]

The resulting spatial difference approximation \( K_h u_h = L_h u_h \) remains second order. The discretization is completed by applying the trapezoidal rule in time. The complete difference scheme is then
Thus this scheme is a generalization of the well-known Crank-Nicolson scheme to non-uniform meshes. This generalization is by no means unique. Other second order formulas for non-uniform meshes that use more than one evaluation of each coefficient function may be derived also.

Figures (7.1a, b, c, d) show the approximate solution in the case of a uniform mesh, with \( h = \frac{1}{40} \) and \( \Delta t = \frac{1}{50} \). Plotted are the initial function along with the approximate solution at various times. The approximate solution is interpolated with a cubic spline.

Results of a computation on a non-uniform mesh are given in Figures (7.2a, b, c, d, and e). Now there are only 31 meshpoints distributed as indicated. The timestep is as before. Clearly, this computation is able to follow the shock profile for a longer time than the computation on the uniform mesh.

Example 7.3. In this example the stability of a number of finite difference schemes for solving the diffusion equation

\[
y_t(x,t) = Ly(x,t) \equiv y_{xx}(x,t),
\]

with \( y(x,0) = \hat{g}(x) \) and \( y(0,t) = y(1,t) = 0 \), is tested numerically. From Theorem (6.2) it follows that for this purpose one has to investigate the behavior of the eigenvalues \( \lambda_h \) of \( L_h v_h = \lambda_h K_h v_h \). Also, the condition number \( \kappa(V_h) \) must be bounded. In Table (7.2) a number of discretizations
Figure 7.1d
Figure 7.2a
Figure 7.2d

![Graph showing x-y coordinates with a label T = 4.4]
in space of the form (6.2) is given, for which the eigenvalues $\lambda_h$ were observed to be real with $\lambda_h < 0$. In fact, all eigenvalues converged to those of the continuous problem $Lw(x) = \lambda w(x)$, $w(0) = w(1) = 0$. These algebraic eigenvalue computations were performed with $h = \frac{1}{8}, \frac{1}{16}, \frac{1}{24}$. Also computed was $K(V_h)$ and as $h \to 0$ no significant increase in its value was observed. Thus the discretizations defined in Table (7.2) can be completed by applying any strongly $A(\alpha)$-stable multistep method.

Listed in this table are the main difference equation at $x_j$, ($3 \leq j \leq J-3$), and the equations at $x_1$ and $x_2$ which may differ from the main approximation. The equations at $x_{J-2}$ and $x_{J-1}$ are the "reflection" of those at $x_2$ and $x_1$ respectively. If $m = 1$ the collocation point is $z_1 = x_j$. In experiment 4 the collocation points for the approximation at $x_1$ are $z_i = x_{i-1}$, ($1 \leq i \leq 4$). The order of consistency is given under "o". The lower order equations near the boundaries do not affect the overall accuracy of the schemes.

Example 7.4. Consider again the algebraic eigenvalue problem of the previous example. Let the discretization in space be defined as follows.

<table>
<thead>
<tr>
<th></th>
<th>$j = 1$</th>
<th>$j = 2$</th>
<th>$3 \leq j \leq J-3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>r s m o</td>
<td>r s m o</td>
<td>r s m o</td>
</tr>
<tr>
<td>1</td>
<td>1 2 1 2</td>
<td>2 2 1 4</td>
<td>2 2 1 4</td>
</tr>
<tr>
<td>2</td>
<td>1 4 1 4</td>
<td>2 2 1 4</td>
<td>2 2 1 4</td>
</tr>
<tr>
<td>3</td>
<td>1 4 1 4</td>
<td>2 3 1 4</td>
<td>3 3 1 6</td>
</tr>
<tr>
<td>4</td>
<td>1 3 4 6</td>
<td>2 2 5 8</td>
<td>2 2 5 8</td>
</tr>
</tbody>
</table>

TABLE 7.2
For each approximation let there be only one collocation point \( z_1 = x_j \).
(So \( m = 1 \).) For the main difference approximation take \( r = s = 3 \). Thus this equation is a seven point formula of order six. The approximation at \( x_2 \) is defined by the values \( r = 2 \) and \( s = 5 \). At \( x_1 \) these values are \( r = 1 \) and \( s = 6 \). The difference equations at \( x_{J-2} \) and \( x_{J-1} \) are once more the "reflection" of those at \( x_2 \) and \( x_1 \). Thus the extra difference equations near the boundaries are also sixth order. Again the eigenvalues \( \lambda_h \) were computed for \( h = \frac{1}{8}, \frac{1}{16}, \frac{1}{24} \). For each \( h \) all eigenvalues were real and negative, with the exception of two conjugate pairs of complex eigenvalues listed below:

\[
\begin{align*}
  h = \frac{1}{8} : & \quad \lambda_1 = -14.9 \pm 7.18 \, i, \quad \lambda_2 = -20.5 \pm 6.50 \, i, \\
  h = \frac{1}{16} : & \quad \lambda_1 = -66.1 \pm 24.8 \, i, \quad \lambda_2 = -67.8 \pm 25.1 \, i, \\
  h = \frac{1}{24} : & \quad \lambda_1 = -150.5 \pm 55.8 \, i, \quad \lambda_2 = -150.8 \pm 56.0 \, i.
\end{align*}
\]

So apparently there is a double complex root as \( h \to 0 \), that is approximately given by

\[
\lambda_h \approx \frac{1}{h^2} \left( -0.26 \pm 0.096 \, i \right)
\]

Therefore, if the discretization is completed by applying a multistep method in time, then this method should be strongly \( A(\alpha) \)-stable with \( \tan \alpha > \frac{0.096}{0.26} \), or \( \alpha > 2.12^{\circ} \). There are strongly \( A(\alpha) \)-stable methods that do not satisfy this requirement. (See Varah (1975), p. 19).

**Example 7.5.** Consider the diffusion equation of Example (7.3), but now let the boundary conditions be given by \( y_x(0, t) = y(1, t) = 0 \). The approximation to the boundary condition at \( x = 0 \) is assumed to have the form
\[ \sum_{i=0}^{s_0} b_i w_i(t) = 0 \] with order of consistency equal to \( s_0 \). In Table (7.3) three spatial discretizations are given for which the discrete eigenvalue problem \( L_h v_h = \lambda_h K_h v_h \) was observed to have only real and negative eigenvalues. The computations were performed for \( h = \frac{1}{8}, \frac{1}{16}, \frac{1}{24} \). The condition number \( \kappa(V_h) \) was also computed and no increase of significance was seen for smaller \( h \). With the exception of the boundary condition at \( x = 0 \) the discretizations in Table (7.3) and Table (7.4) below are identical to those of Table (7.2).

**TABLE 7.3**

<table>
<thead>
<tr>
<th>#</th>
<th>( s_0 )</th>
<th>( j = 1 )</th>
<th>( j = 2 )</th>
<th>( 3 \leq j \leq J-3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>1 2 1 2</td>
<td>2 2 1 4</td>
<td>2 2 1 4</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1 4 1 4</td>
<td>2 2 1 4</td>
<td>2 2 1 4</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>1 4 1 4</td>
<td>2 3 1 4</td>
<td>3 3 1 6</td>
</tr>
</tbody>
</table>

Two discretizations for which the algebraic eigenvalue problem was observed to have complex eigenvalues are given in Table (7.4). In experiment 4 there are two distinct conjugate pairs of eigenvalues approximately given by \( \lambda_1 \approx \frac{1}{h^2}(-0.178 \pm 0.049i) \) and \( \lambda_2 \approx \frac{1}{h^2}(-0.261 \pm 0.097i) \). In experiment 5 there is only one conjugate pair given by \( \lambda_1 \approx \frac{1}{h^2}(-0.246 \pm 0.044i) \). The multistep differencing in time should therefore be strongly \( A(\alpha) \)-stable with \( \alpha \geq \alpha_{\min} \). The value \( \alpha_{\min} \) as computed from the complex eigenvalues above is also given in the table.
Example 7.6. The purpose of this example is to show the effect of using an improper approximation to a boundary condition. The equation under consideration is

\[ y_t(x, t) = \left( \frac{2}{3 \pi} \right)^2 y_{xx}(x, t), \quad 0 \leq x \leq 1, \quad t \geq 0, \]

with initial condition

\[ y(x, 0) = \cos\left(\frac{3\pi}{2} x\right), \quad 0 \leq x \leq 1, \]

and boundary conditions

\[ y_x(0, t) = y(1, t) = 0, \quad t \geq 0. \]

The solution is \( y(x, t) = e^{-t} \cos\left(\frac{3\pi}{2} x\right) \). Let the discretization in space be given by

\[ K_h w_j^t(x_j, t) = L_h w(x_j, t), \quad 1 \leq j \leq J-1, \]

where \( K_h w_j = \frac{1}{12} w_{j-1} + \frac{10}{12} w_j + \frac{1}{12} w_{j+1} \) and \( L_h w_j = \left(\frac{2}{3 \pi}\right)^2 \frac{1}{h^2} (w_{j-1} - 2w_j + w_{j+1}) \).

### Table 7.4

<table>
<thead>
<tr>
<th>#</th>
<th>s0</th>
<th>r</th>
<th>s</th>
<th>m</th>
<th>o</th>
<th>r</th>
<th>s</th>
<th>m</th>
<th>o</th>
<th>α_{min}</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>6</td>
<td>21°</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>11°</td>
</tr>
</tbody>
</table>
The order of consistency of this approximation is four. (See Example (I.4.4).)
A one parameter family of approximations to the boundary condition at
\( x = 0 \) is given by

\[
88w_0 - (5\alpha + 144)w_1 + (8\alpha + 72)w_2 - (3\alpha + 16)w_3 = 0
\]

This approximation is consistent if \( \alpha \neq 24 \). The order of consistency is
equal to two, except if \( \alpha = 0 \), when the order is three. The characteristic polynomial of the discrete boundary condition is given by

\[
c_0(z) = 88z^3 - (5\alpha + 144)z^2 + (8\alpha + 72)z - (3\alpha + 16).
\]

If \( \alpha > 24 \) then the characteristic equation \( c_0(z) = 0 \) has a real root \( z_1 > 1 \).
For such \( \alpha \) the eigenvalue problem \( K_h v_h = \lambda_h L_h v_h \) has a positive eigenvalue that is asymptotically given by

\[
\lambda_h \approx \frac{12(z_1 - 1)^2}{h^2(z_1^2 + 10z_1 + 1)}
\]

This may lead to an instability. (See Example (6.5).) For example, let
the discretization in time be the fourth order method of Gear. (See Gear
(1971), p.217.) This leads to the difference scheme

\[
\frac{1}{12\Delta t} K_h \left( 25u_h^n - 48u_h^{n-1} + 36u_h^{n-2} - 16u_h^{n-3} + 3u_h^{n-4} \right) = L_h u_h^n.
\]

Of the real axis only the interval \([0, 10^2/3]\) does not lie in the stability region
of this multistep method in time. (See Gear (1971), p.216.) Two
numerical computations were performed with \( \alpha = 30 \) and \( \Delta t = 0.025 \). The
first computation with \( h = 0.050 \) and the second with \( h = 0.025 \). The
values of $\Delta t \lambda_h$ for these data are equal to 2.03 and 8.12 respectively. Hence both schemes are unstable. The approximate solution at time $t = 0.5$, interpolated with a cubic spline, as well as the exact solution are shown in Figures (7.3a, b). The same computation was repeated with $\alpha = 20$. This is stable. Results at $t = 0.5$ are shown in Figures (7.4a, b).
BIBLIOGRAPHY


J. C. Butcher, Implicit Runge-Kutta processes, Math. of Comp. 18, 1964, pp. 50-64.


David Gottlieb and Bertil Gustafsson, Generalized Du Fort-Frankel methods for parabolic initial boundary value problems, ICASE Report no. 75-5, NASA's Langley Research Center, Hampton, VA.


S. J. Osher, Maximum norm stability for parabolic difference schemes in half-space, Hyperbolic equations and waves, Batelle Institute, Seattle, Washington, 1968, pp. 61-75.


J. M. Varah, Maximum norm stability of finite difference approximations to the mixed initial boundary value problem for the heat equation, Math. of Comp. 24, 1970, pp. 31-44.


