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

ABSTRACT

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

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

$$(1.1) \qquad Ly(x) \equiv y^{(n)}(x) + \sum_{k=0}^{n-1} a^k(x)y^{(k)}(x) = f(x), \quad 0 \le x \le 1 ,$$

is investigated.

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

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

$$(1.2) \qquad L_h u_j \equiv \sum_{i=-r_j}^{s_j} d_{j,i}\, u_{j+i} = \tilde{f}_j, \quad k_0 \le j \le J - n + k_0, \quad k_0 \ge 0 ,$$

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 ω_j of each finite difference approximation is defined as $\omega_j \equiv r_j + s_j + 1$. If $\omega_j = n + 1$ then the approximation is said to be compact. The assumption is made that $\omega_j \le \omega_{max}$ for some integer ω_{max} that does not depend on S_h.

The local truncation error is given by $\tau_j \equiv L_h y_j - \tilde{f}_j$. If there is a

constant C independent of h such that $|\tau_j| \leq Ch^{n_s}$ with $n_s > 0$, and n_s as high as possible, then the difference approximation is said to be <u>consistent</u> 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

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

and

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

where

$$n_k(0) < n \quad \text{and} \quad n_k(1) < n .$$

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 <u>finite difference scheme</u> that determines the approximations u_j to y_j completely one has approximations to the boundary conditions (1.1a, b) that look like

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

and

(1.2b) $$B_{h,k}(1)u_J \equiv \sum_{i=-r_k(1)}^{0} d_{k,i}(1)u_{J+i} = \tilde{b}_k(1), \quad 1 \le k \le n-n_0.$$

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) <u>of order s</u> if the approximate solution u_j, $(0 \le j \le J)$, is uniquely defined by (1.2) and (1.2a, b) and $|y_j - u_j| \le K_1 h^s$ 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

(1.3) $$y''(x) + a^1(x)y'(x) + a^0(x)y(x) = f(x) ,$$

the simplest difference approximation is perhaps given by

(1.4) $$L_h u_j \equiv D_h^2 u_j + a_j^1 D_h^1 u_j + a^0 u_j = f_j .$$

Here $D_h^2 u_j \equiv (u_{j+1} - 2u_j + u_{j-1}) / h^2$, $D_h^1 u_j \equiv (u_{j+1} - u_{j-1}) / 2h$ 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

$$(1.5) \quad L_h u_j \equiv (I - \frac{h^2}{12} D_h^2) D_h^2 u_j + a_j^1 (I - \frac{h^2}{6} D_h^2) D_h^1 u_j + a_j^0 u_j = f_j.$$

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

$$(1.6) \qquad y''(x) + a^0(x)y(x) = f(x)$$

is given by

$$D_h^2 u_j + K_h(\tfrac{5}{6})\ (a_j^0\, u_j) = K_h(\tfrac{5}{6})\, f_j, \tag{1.7}$$

with

$$K_h(c)w_j \equiv \frac{1-c}{2}\, w_{j-1} + cw_j + \frac{1-c}{2}\, w_{j+1}\ .$$

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 $|z_i - z_k| \geq \frac{1}{K'} h$, $i \neq k$. (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 $d \equiv r + s + m - 1$. Assume that for some integer ℓ, $(-r \leq \ell \leq s)$, a polynomial $p^\ell(x) \in P_d$ is determined uniquely by the $r + s + m$ equations

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

$$Lp^\ell(z_i) = f(z_i) , \quad 1 \leq i \leq m . \tag{2.2}$$

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

$$p^\ell(x_{j+\ell}) = u_{j+\ell} , \tag{2.3}$$

relates the quantities u_{j+i}, $(-r \leq i \leq s)$, by an expression (difference equation) of the form

$$(2.4) \qquad L_h u_j \equiv \sum_{i=-r}^{s} d_i u_{j+i} = \sum_{i=1}^{m} e_i f(z_i).$$

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

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

Also, define a matrix D_L by

$$(2.5) \quad D_L \equiv \begin{bmatrix} w^0(x_{j-r}) & w^1(x_{j-r}) & \cdot & \cdot & w^d(x_{j-r}) & u_{j-r} \\ w^0(x_{j-r+1}) & w^1(x_{j-r+1}) & & & w^d(x_{j-r+1}) & u_{j-r+1} \\ \cdot & \cdot & & & \cdot & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ w^0(x_{j+s}) & w^1(x_{j+s}) & \cdot & \cdot & w^d(x_{j+s}) & u_{j+s} \\ Lw^0(z_1) & Lw^1(z_1) & \cdot & \cdot & Lw^d(z_1) & f(z_1) \\ Lw^0(z_2) & Lw^1(z_2) & \cdot & \cdot & Lw^d(z_2) & f(z_2) \\ \cdot & \cdot & & & \cdot & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ Lw^0(z_m) & Lw^1(z_m) & \cdot & \cdot & Lw^d(z_m) & f(z_m) \end{bmatrix}$$

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

$$(2.6) \qquad d_i = \mathrm{cof}_L\,[u_{j+i}] \,/\, E\,,$$

and

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

where $\mathrm{cof}_L[\cdot]$ 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

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

where

$$L_0 y(x) \equiv y^{(n)}(x) \ .$$

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

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

and

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

then

$$d_i = \frac{(-1)^m}{E} \begin{vmatrix} Lw^{r+i}(z_1) & Lw^{r+s+1}(z_1) & \cdots & Lw^d(z_1) \\ \vdots & \vdots & & \vdots \\ Lw^{r+i}(z_m) & Lw^{r+s+1}(z_m) & \cdots & Lw^d(z_m) \end{vmatrix}, \ -r \leq i \leq s \ , \tag{2.10}$$

and

$$(2.11)\quad e_i = \frac{(-1)^{m+i+1}}{E} \begin{vmatrix} Lw^{r+s+1}(z_1) & \cdots & Lw^{d}(z_1) \\ \vdots & & \vdots \\ Lw^{r+s+1}(z_{i-1}) & \cdots & Lw^{d}(z_{i-1}) \\ Lw^{r+s+1}(z_{i+1}) & \cdots & Lw^{d}(z_{i+1}) \\ \vdots & & \vdots \\ Lw^{r+s+1}(z_m) & \cdots & Lw^{d}(z_m) \end{vmatrix}, \quad 1 \le i \le m .$$

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

$$(2.12)\qquad E = (-1)^m \begin{vmatrix} 1 & L_0 w^{r+s+1}(z_1) & \cdots & L_0 w^{r+s+m-1}(z_1) \\ \vdots & \vdots & & \vdots \\ 1 & L_0 w^{r+s+1}(z_m) & \cdots & L_0 w^{r+s+m-1}(z_m) \end{vmatrix} .$$

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

$$d_i = \sum_{k=1}^{m} Lw^{r+i}(z_k) e_k , \quad -r \le i \le s .$$

<u>Remark</u>. 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

$$(2.13) \qquad w^i(x) \equiv \frac{\prod_{\substack{k=0 \\ k\neq i}}^{r+s} (x-x_{j-r+k})}{\prod_{\substack{k=0 \\ k\neq i}}^{r+s} (x_{j-r+i}-x_{j-r+k})}, \qquad 0 \leq i \leq r+s,$$

and

$$(2.14) \qquad w^{r+s+i}(x) \equiv \left[\frac{\prod_{\substack{k=1 \\ k\neq i}}^{m-1} (x-t_k)}{\prod_{\substack{k=1 \\ k\neq i}}^{m-1} (t_i-t_k)}\right] \left[\frac{\prod_{k=0}^{r+s} (x-x_{j-r+k})}{\prod_{k=0}^{r+s} (t_i-x_{j-r+k})}\right], \qquad 1 \leq i \leq m-1.$$

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

$$d_{-1}u_{j-1} + d_0u_j + d_1u_{j+1} = e_1f(z_1) + e_2f(z_2) \ , \tag{2.15}$$

to the differential equation

$$Ly(x) \equiv y''(x) + a^1(x)y'(x) + a^0(x)y(x) = f(x) \ .$$

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:

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

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

$$\left[\begin{array}{cccc:c} 1 & 0 & 0 & & -Lw^0(z_1) \\ 0 & 1 & 0 & & -Lw^1(z_1) \\ 0 & 0 & 1 & & -Lw^2(z_1) \\ \hdashline 0 & 0 & 0 & & -Lw^3(z_1) \end{array}\right] \begin{bmatrix} d_{-1} \\ d_0 \\ d_1 \\ e_1 \end{bmatrix} = \begin{bmatrix} Lw^0(z_2)e_2 \\ Lw^1(z_2)e_2 \\ Lw^2(z_2)e_2 \\ Lw^3(z_2)e_2 \end{bmatrix} ,$$

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

$$M_L \equiv \begin{bmatrix} Lw^{r+s+1}(z_1) & \cdots & Lw^{r+s+m-1}(z_1) \\ \vdots & & \vdots \\ Lw^{r+s+1}(z_m) & \cdots & Lw^{r+s+m-1}(z_m) \end{bmatrix}$$

has rank equal to m-1. In particular if the normalizing factor E is nonzero 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.

<u>Theorem 2.1</u> 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 $r + s \geq n$. 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 $r + s + m - n$.

<u>Proof</u>. Define

$$\bar{d}_i \equiv \mathrm{cof}_{L_0}[u_{j+i}] / E, \quad -r \leq i \leq s, \tag{2.16}$$

$$\bar{e}_i \equiv -\mathrm{cof}_{L_0}[f(z_i)] / E, \quad 1 \leq i \leq m, \tag{2.17}$$

so that $\sum_{i=1}^{m} \bar{e}_i = 1$.

First it will be shown that if $E \neq 0$ then at least $n + 1$ of the coefficients $\bar{d}_i$ are nonzero. For suppose on the contrary that $\bar{d}_{\ell_k} \neq 0$, $1 \leq k \leq k_1 < n + 1$ and $\bar{d}_i = 0$ otherwise. Let

$p(x) \equiv \left[\prod_{k=1}^{k_1} (x-x_{j+\ell_k}) \right] q(x)$, where $q(x) \in P_{n-k_1}$ is chosen such that $p(x)$ has degree n and leading coefficient $\frac{1}{n!}$. By assumption $n \leq r + s$, so certainly $n \leq d$. Hence by construction

$$\sum_{i=-r}^{s} \bar{d}_i \, p_{j+i} - \sum_{i=1}^{m} \bar{e}_i L_0 p(z_i) = 0.$$

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 $\sum_{i=1}^{m} \bar{e}_i = 0$, which is a contradiction. So if $E \neq 0$ then $n + 1$ or more $\bar{d}_i$ are nonzero. Since $d_i / \bar{d}_i = 1 + 0(h)$, the same is true for the d_i, provided h is small enough.

The truncation error τ_j is defined as

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

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

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

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

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

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

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 $w^i(x)$, 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.

<u>Example 2.2</u> 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

$$w^0(x) = -(x-x_j)(x-x_{j+1})(x-x_{j+2}) / 6h^3 ,$$

$$w^1(x) = (x-x_{j-1})(x-x_{j+1})(x-x_{j+2}) / 2h^3 ,$$

$$w^2(x) = -(x-x_{j-1})(x-x_j)(x-x_{j+2}) / 2h^3 ,$$

$$w^3(x) = (x-x_{j-1})(x-x_j)(x-x_{j+1}) / 6h^3 ,$$

$$w^4(x) = (x-x_{j-1})(x-x_j)(x-x_{j+1})(x-x_{j+2}) / h^4 .$$

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_i 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

$$\bar{d}_i = h^2 \begin{vmatrix} L_0 w^{1+i}(z_1) & L_0 w^4(z_1) \\ L_0 w^{1+i}(z_2) & L_0 w^4(z_2) \end{vmatrix} , \qquad -1 \leq i \leq 2 ,$$

gives $\bar{d}_{-1} = -2/h^2$, $\bar{d}_0 = 6/h^2$, $\bar{d}_1 = -6/h^2$ and $\bar{d}_2 = 2/h^2$. Further, $\bar{e}_1 = h^2 L_0 w^4(z_2) = -2$ and $\bar{e}_2 = -h^2 L_0 w^4(z_1) = 2$, so that for the problem $y''(x) = f(x)$ the difference equation becomes

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

Dividing by 2h one obtains

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

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

<u>Example 2.3</u> 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 $w^i(x)$, $(0 \le i \le 4)$, in example (2.2) take

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

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

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

If the mesh is taken to be uniform and if the collocation points are distributed symmetrically, i.e., $z_1 = x_{j+\frac{1}{2}} - \xi h$, $z_2 = x_{j+\frac{1}{2}}$, $z_3 = x_{j+\frac{1}{2}} + \xi h$, then after some computation it is found that $E = c\xi^3(4\xi^2-3)/h^5$, where c is

some constant. Since the points z_i are distinct, ξ cannot be equal to zero. Hence if the points z_i are placed symmetrically, then $E = 0$ if and only if $\xi = \frac{1}{2}\sqrt{3} \cong 0.866$.

Remark. A necessary condition for the normalizing factor E to be nonzero 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 $d^n w^i(x) / dx^n \equiv 0$, $(0 \leq i \leq r + s)$, because the first $r + s + 1$ basis functions have degree $r + s$. Hence the first column in the determinant defining $\bar{d}_i$ consists of zeros for each i. Thus $\bar{d}_i = 0$, $(-r \leq i \leq s)$, 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_i, $(r + s \leq i \leq d)$, 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 $\bar{d}_i$ are uniquely determined, independent of the number of collocation points. For example, if $n = 2$ and $r + s = 2$ then $\bar{d}_{-1} = \frac{2}{h_j(h_j + h_{j+1})}$,

$\bar{d}_0 = \frac{-2}{h_j h_{j+1}}$ and $\bar{d}_1 = \frac{2}{h_{j+1}(h_j + h_{j+1})}$, so that the approximation to the second derivative is always given by

$$\frac{\frac{u_{j+1} - u_j}{h_{j+1}} - \frac{u_j - u_{j-1}}{h_j}}{\frac{1}{2}(h_j + h_{j+1})} .$$

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 $w^{r+s+m}(x) \in P_{r+s+m}$ that vanishes at the mesh points. Thus this polynomial has the form

$$w^{r+s+m}(x) \equiv \frac{\prod_{k=1}^{m-1} (x-t_k)}{\prod_{k=1}^{m-1} (t_m-t_k)} \quad \frac{\prod_{k=0}^{r+s} (x-x_{j-r+k})}{\prod_{k=0}^{r+s} (t_m-x_{j-r+k})} ,$$

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

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

The quantity between square brackets equals zero for $0 \leq k \leq r + s + m - 1$. Moreover, $w^{r+s+m}(x_{j+i}) = 0$, $(-r \leq i \leq s)$. 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 $L_0 w^{r+s+m}(x) \equiv \frac{d^n w^{r+s+m}(x)}{dx^n} = 0$. 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 $\tilde{w}^{r+s+m}(x) \equiv \prod_{k=1}^{m-1} (x-t_k) \prod_{k=0}^{r+s} (x-x_{j-r+k})$ 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 $\tilde{w}^{r+s+m}(x) = \prod_{k=0}^{n} (x-x_{j-r+k})$. 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

$$Ly(x) \equiv y'(x) + a^0(x)y(x) = f(x) \quad . \tag{4.1}$$

Take $m = 1$ and $z_1 = x_{j+\frac{1}{2}} \equiv \frac{1}{2}(x_j + x_{j+1})$. 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.

$$d_0 = Lw^0(z_1) = \frac{-1}{h_j} + \tfrac{1}{2}\, a^0_{j+\frac{1}{2}} ,$$

and

$$d_1 = Lw^1(z_1) = \frac{1}{h_j} + \tfrac{1}{2}\, a^0_{j+\frac{1}{2}} .$$

The difference approximation can then be written as

$$(u_{j+1} - u_j) / h_j + \tfrac{1}{2}\, a^0_{j+\frac{1}{2}} (u_j + u_{j+1}) = f_{j+\frac{1}{2}} , \tag{4.2}$$

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

$$d_0 = \frac{1}{E} \begin{vmatrix} Lw^0(z_1) & Lw^2(z_1) \\ Lw^0(z_2) & Lw^2(z_2) \end{vmatrix} , \qquad d_1 = \frac{1}{E} \begin{vmatrix} Lw^1(z_1) & Lw^2(z_1) \\ Lw^1(z_2) & Lw^2(z_2) \end{vmatrix} ,$$

$e_1 = \frac{1}{E} Lw^2(z_2)$ and $e_2 = \frac{-1}{E} Lw^2(z_1)$, with $E = L_0w^2(z_1) - L_0w^2(z_2)$ and $L_0y(x) = y'(x)$. The equation thus obtained is

$$(u_{j+1} - u_j) / h_j + \tfrac{1}{2}(a_j^0 u_j + a_{j+1}^0 u_{j+1}) = \tfrac{1}{2}(f_j + f_{j+1}).$$

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 $\tilde{w}^3(x) = (x-t_1)(x-x_j)(x-x_{j+1})$. Let $t_1 = x_j + \eta h_j$, $z_1 = x_j + \xi_1 h_j$ and $z_2 = x_j + \xi_2 h_j$. Then the special values of ξ_1 and ξ_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

$$\xi = \frac{\eta + 1 \pm \sqrt{\eta^2 - \eta + 1}}{3} .$$

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

$$d_0 = \frac{-1}{h_j} + \frac{1}{4}(a^0(z_1) + a^0(z_2)) - \frac{1}{12} h_j a^0(z_1) a^0(z_2),$$

$$d_1 = \frac{1}{h_j} + \frac{1}{4}(a^0(z_1) + a^0(z_2)) + \frac{1}{12} h_j a^0(z_1) a^0(z_2),$$

$$e_1 = \tfrac{1}{2} - h_j \frac{\sqrt{3}}{12} a^0(z_2) \text{ and } e_2 = \tfrac{1}{2} + h_j \frac{\sqrt{3}}{12} a^0(z_1).$$

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 2 m. 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 $r = 0$, $s = 2$, $m = 1$, $z_1 = x_{j+2}$ 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 $$d_i = Lw^i(z_1), \quad 0 \le i \le 2 .$$

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

$$\frac{1}{2h} \{u_j - 4u_{j+1} + 3u_{j+2}\} + a^0_{j+2} u_{j+2} = f_{j+2} ,$$

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

$$(4.3) \qquad Ly(x) \equiv y''(x) + a^1(x)y'(x) + a^0(x)y(x) = f(x).$$

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

$$d_{-1} = Lw^0(x_j) = \frac{2-h_{j+1}a_j^1}{h_j(h_j+h_{j+1})} ,$$

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

and

$$d_1 = Lw^2(x_j) = \frac{2+h_j a_j^1}{h_{j+1}(h_j+h_{j+1})} .$$

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

$$d_{-1} = \frac{1}{h_j(h_j+h_{j+1})}\left[2 - \frac{(2h_j+h_{j+1})}{3}a^1(z_1) - \frac{(h_{j+1}-h_j)(2h_{j+1}+h_j)}{9}a^0(z_1)\right]$$

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

$$d_1 = \frac{1}{h_{j+1}(h_j+h_{j+1})}\left[2+\frac{(h_j+2h_{j+1})}{3}a^1(z_1)+\frac{(2h_j+h_{j+1})(h_{j+1}-h_j)}{9}a^0(z_1)\right].$$

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

$$Ly(x) \equiv y''(x) + a^0(x)y(x) = f(x) , \tag{4.4}$$

then the resulting finite difference equation is identical to the "Mehrstellenverfahren" 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

$$\frac{\dfrac{u_{j+1}-u_j}{h_{j+1}} - \dfrac{u_j-u_{j-1}}{h_j}}{\frac{1}{2}(h_j+h_{j+1})} + W_h(a_j^0 u_j) = W_h(f_j) , \tag{4.5}$$

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

$$b_j = \frac{1}{6} - \frac{h_j^2}{6h_{j+1}(h_j + h_{j+1})} \quad .$$

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 ξ_1 and ξ_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

$$z_1 = x_{j-\frac{1}{2}} - \xi h \quad , \qquad z_2 = x_{j-\frac{1}{2}} + \xi h \ ,$$

$$z_3 = x_{j+\frac{1}{2}} - \xi h \quad , \qquad z_4 = x_{j+\frac{1}{2}} + \xi h \ .$$

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

$$\xi = \sqrt{\frac{5}{12} - \frac{1}{2}\sqrt{\frac{23}{45}}} \cong 0.24 \qquad \text{or} \qquad \xi = \sqrt{\frac{5}{12} + \frac{1}{2}\sqrt{\frac{23}{45}}} \cong 0.88 .$$

So with either value of ξ 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 $z_1 = x_j$. 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

$$(31u_{j-2} + 128u_{j-1} - 318u_j + 128u_{j+1} + 31u_{j+2})/(252h^2) + W_h(a_j^0 u_j) = W_h f_j,$$

where

$$W_h w_j \equiv (23w_{j-2} + 688w_{j-1} + 2358w_j + 688w_{j+1} + 23w_{j+2})/3780.$$

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 ξ 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

$$\text{(5.1)} \qquad By(0) \equiv y^{(k)}(0) + \sum_{i=0}^{k-1} b_i y^{(i)}(0) = b,$$

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 $d \doteq s + m$ and assume that a polynomial $p(x) \in P_d$ is uniquely determined by the equations

$$\text{(5.2)} \qquad p(x_i) = u_i \ , \qquad 0 \le i \le s \ ,$$

$$\text{(5.3)} \qquad Lp(z_i) = f(z_i) \ , \qquad 1 \le i \le m \ .$$

Then the equation

$$Bp(0) = b \tag{5.4}$$

generates a relation of the form

$$B_h u_0 \equiv \sum_{i=0}^{s} d_i u_i = e_0 b + \sum_{i=1}^{m} e_i f(z_i) \ . \tag{5.5}$$

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:

$$d_i = \frac{1}{E_0} \begin{vmatrix} Bw^i(0) & Bw^{s+1}(0) & \cdot\ \cdot & Bw^d(0) \\ Lw^i(z_1) & Lw^{s+1}(z_1) & \cdot\ \cdot & Lw^d(z_1) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ Lw^i(z_m) & Lw^{s+1}(z_m) & \cdot\ \cdot & Lw^d(z_m) \end{vmatrix} , \quad 0 \leq i \leq s \ , \tag{5.6}$$

$$e_0 = \frac{1}{E_0} \begin{vmatrix} Lw^{s+1}(z_1) & \cdot\ \cdot & Lw^d(z_1) \\ \cdot & & \cdot \\ \cdot & & \cdot \\ Lw^{s+1}(z_m) & \cdot\ \cdot & Lw^d(z_m) \end{vmatrix} , \tag{5.7}$$

$$(5.8) \qquad e_i = \frac{(-1)^i}{E_0} \begin{vmatrix} Bw^{s+1}(0) & \cdot\ \cdot & Bw^d(0) \\ Lw^{s+1}(z_1) & \cdot\ \cdot & Lw^d(z_1) \\ \cdot & & \cdot \\ \cdot & & \cdot \\ Lw^{s+1}(z_{i-1}) & \cdot\ \cdot & Lw^d(z_{i-1}) \\ Lw^{s+1}(z_{i+1}) & \cdot\ \cdot & Lw^d(z_{i+1}) \\ \cdot & & \cdot \\ \cdot & & \cdot \\ Lw^{s+1}(z_m) & \cdot\ \cdot & Lw^d(z_m) \end{vmatrix}, \quad 1 \le i \le m .$$

It is convenient to define the normalizing factor E_0 by

$$(5.9) \quad E_0 \equiv (-1)^{m+1} \begin{vmatrix} L_0 w^{s+1}(z_1) & \cdot\ \cdot & L_0 w^d(z_1) \\ \cdot & & \cdot \\ \cdot & & \cdot \\ L_0 w^{s+1}(z_m) & \cdot\ \cdot & L_0 w^d(z_m) \end{vmatrix} ,$$

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

<u>Theorem 5.1</u> Assume that $E_0 \neq 0$. If $m = 0$ let $s \ge k$; otherwise if $m > 0$, let $s + m \ge n$. 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

$$|\tau_0| \le Ch^{s+m-k+1} ,$$

i.e., the order of consistency of (5.5) is at least equal to $s+m-k+1$.

<u>Proof</u>. Consider the equations

$$(5.10) \qquad \sum_{i=0}^{s} d_i w^{\ell}(x_i) - \sum_{i=1}^{m} e_i L w^{\ell}(z_i) = e_0 B w^{\ell}(x_0), \quad 0 \le \ell \le d .$$

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

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

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

$$A \equiv \begin{bmatrix} Lw^0(z_1) & \cdots & Lw^0(z_m) \\ \cdot & & \cdot \\ \cdot & & \cdot \\ Lw^s(z_1) & \cdots & Lw^s(z_m) \end{bmatrix} ,$$

and

$$C \equiv \begin{bmatrix} Lw^{s+1}(z_1) & \cdots & Lw^{s+1}(z_m) \\ \cdot & & \cdot \\ \cdot & & \cdot \\ Lw^d(z_1) & \cdots & Lw^d(z_m) \end{bmatrix} .$$

Thus the equations (5.10) admit a unique solution only if $\det(C) \neq 0$. In particular, the assumption that $E_0 \neq 0$ implies that $\det(C) \neq 0$, provided that h is small enough.

Next let $\bar{d}_i$, $(0 \leq i \leq s)$, 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 $\bar{d}_i$ are nonzero. More precisely suppose that $\bar{d}_{\ell_\nu} \neq 0$, $(1 \leq \nu \leq k_1 < k+1)$, and $\bar{d}_i = 0$ otherwise. Let

$q(x) = \prod_{\nu=1}^{k_1} (x-x_{\ell_\nu}) q_1(x)$, where $q_1(x) \in P_{k-k_1}$ has leading coefficient $1/k!$. Then $q(x) \in P_k$, and by assumption $k \leq d$. Therefore

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

The construction of $q(x)$ makes the entire lefthand side vanish. Also, $B_0 q(x_0) = 1$. Hence $\bar{e}_0 = 0$. But this is a contradiction since $\bar{e}_0 = (-1)^{m+1}$. Thus if $E_0 \neq 0$ then at least $k+1$ of the coefficients $\bar{d}_i$ 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.

<u>Theorem 5.2</u>. The normalizing factor is nonzero if $m=0$ or if $s=k$. (In the latter case the difference equation is said to be compact.)

<u>Proof</u>. 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 $L_0 q(x) \equiv q^{(n)}(x) \equiv 0$. 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

$$\tilde{w}^{s+m+1}(x_i) = 0, \qquad 0 \leq i \leq s ,$$

and

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

Also assume that the collocation points are chosen such that $L_0 \tilde{w}^{s+m+1}(z_i) = 0$, $(1 \leq i \leq m)$. 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

$$Ly(x) \equiv y''(x) + a^1(x)y'(x) + a^0(x)y(x) = f(x), \tag{6.1}$$

and the boundary condition by

$$(6.2) \qquad By(0) \equiv y'(0) + b_0 y(0) = b.$$

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 $e_0 = E_0 = 1$,

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

and

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

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.

<u>Example 6.2.</u> 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 $w^0(x) \equiv -(x-x_1)/h$, $w^1(x) \equiv (x-x_0)/h$ and $w^2(x) \equiv (x-x_0)(x-x_1)/h^2$. Then the

normalizing factor E_0 is given by $E_0 \equiv L_0 w^2(z_1) = \frac{2}{h^2}$. The coefficients of the difference approximation

$$d_0 u_0 + d_1 u_1 = e_0 b + e_1 f(z_1) , \tag{6.3}$$

are then equal to

$$e_0 \equiv \frac{1}{E_0} \; Lw^2(z_1) = 1 + \frac{h}{2} a^1(z_1) ,$$

$$e_1 \equiv - \frac{1}{E_0} Bw^2(0) = \frac{h}{2} ,$$

$$d_0 \equiv \frac{1}{E_0} \begin{vmatrix} Bw^0(0) & Bw^2(0) \\ Lw^0(z_1) & Lw^2(z_1) \end{vmatrix} = \frac{-1}{h} - a^1(z_1) + b_0 + \frac{h}{2} b_0 a^0(z_1),$$

and

$$d_1 \equiv \frac{-1}{E_0} \begin{vmatrix} Bw^1(0) & Bw^2(0) \\ Lw^1(z_1) & Lw^2(z_1) \end{vmatrix} = \frac{1}{h} + a^1(z_1) + \frac{h}{2} a^0(z_1) .$$

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 $z_1 = \frac{x_1}{3} = \frac{h_1}{3}$.

<u>Example 6.3</u>. Let $s = 2$ and $m = 1$, so that the difference approximation to (6.2) has the form

$$(6.5) \qquad d_0u_0 + d_1u_1 + d_2u_2 = e_0b + e_1f(z_1) \ .$$

Assume that the mesh is uniform and let $z_1 = x_0 + \xi h$. Then $E_0 = L_0w^3(x_0 + \xi h)$ 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 $\xi = (9 - \sqrt{33})/12 \cong 0.27$ or $\xi = (9 + \sqrt{33})/12 \cong 1.23$.

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

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

together with boundary conditions

$$(7.2) \quad B_{h,k}(0)u_0 \equiv \sum_{i=0}^{s_k(0)} d_{k,i}(0)u_i = e_{k,0}(0)b_k(0) + \sum_{i=1}^{m_k(0)} e_{k,i}(0)f(z_{k,i}(0))$$

$$\equiv \tilde{b}_k(0), \quad 1 \leq k \leq n_0,$$

$$(7.3) \quad B_{h,k}(1)u_J \equiv \sum_{i=-r_k(1)}^{0} d_{k,i}(1)u_{J+i} = e_{k,0}(1)b_k(1) + \sum_{i=1}^{m_k(1)} e_{k,i}(1)f(z_{k,i}(1))$$

$$\equiv \tilde{b}_k(1), \quad 1 \leq k \leq n - n_0.$$

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

$$(7.4) \quad L_h u_j \equiv \sum_{i=0}^{s_j} d_{j,i} u_i = \sum_{i=1}^{m_j} e_{j,i} f(z_{j,i}) \equiv \tilde{f}_j, \quad k_0 \le j \le r-1, \quad k_0 \ge 0 ,$$

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

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

$$(7.6) \qquad L_h u_h = f_h.$$

Here $f_h \in R^{J+1}$ 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 $|\tilde{f}_j - f(x_j)| \le ch$. Also, let $e_h \in R^{J+1}$ be the error vector, i.e., $e_h = (e_0, \cdots, e_J)^T$ with $e_j = y_j - u_j$, and let τ_h be the vector of truncation errors, $\tau_h = (\tau_1(0), \cdots, \tau_{n_0}(0), \tau_{k_0}, \cdots, \tau_{J-n+k_0}, \tau_1(1), \cdots, \tau_{n-n_0}(1))^T$, with

$$\tau_k(0) \equiv B_{h,k}(0) y_0 - \tilde{b}_k(0), \qquad 1 \le k \le n_0,$$

$$\tau_j \equiv L_h y_j - \tilde{f}_j , \qquad k_0 \le j \le J - n + k_0 ,$$

and

$$\tau_k(1) \equiv B_{h,k}(1)y_J - \tilde{b}_k(1), \qquad 1 \le k \le n - n_0 .$$

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

The finite difference scheme (7.6) is said to be consistent if $||\tau_h|| \le k_1 h$ as $h \to 0$ and stable if L_h^{-1} exists for all small enough h and $||L_h^{-1}|| \le k_2$. 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 $\bar{d}_i$ be as in Section 2, (Equations (2.12) and (2.16)), and define

$$\bar{c}(t) \equiv h^n \sum_{i=-r}^{s} \bar{d}_i t^{r+i}.$$

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 <u>characteristic polynomial</u> associated with the difference equation (7.1). Assume that c(t) is explicitly given by

$$c(t) = \sum_{i=0}^{N} a_i t^i, \quad \text{with} \quad N = r + s - n. \tag{7.7}$$

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

$$(7.8) \qquad c_j(t) = \sum_{i=0}^{N_j} a_{j,i} t^i , \qquad k_0 \le j \le r - 1 ,$$

and

$$(7.9) \qquad c_j(t) = \sum_{i=0}^{N_j} a_{j,i} t^i , \qquad J - s + 1 \le j \le J - n + k_0 .$$

Finally, consider the homogeneous difference equation

$$(7.10) \qquad \sum_{i=0}^{N} a_i v_{j+i} = 0 , \qquad 0 \le j < \infty,$$

with boundary conditions

$$(7.11) \qquad \sum_{i=0}^{N_j} a_{j,i} v_i = 0, \; k_0 \le j \le r - 1, \qquad \sup_{0 \le j \le \infty} |v_j| \le \text{const.}, \; 0 \le j < \infty ,$$

and

$$(7.12) \qquad \sum_{i=0}^{N} a_{N-i} v_{J-j-i} = 0, \qquad 0 \le j < \infty ,$$

subject to

$$(7.13) \qquad \sum_{i=0}^{N_j} a_{j,N_j-i} v_{J-i} = 0, \; J - s + 1 \le j \le J - n + k_0, \; \sup_{0 \le j < \infty} |v_{J-i}| \le \text{const.}$$

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

<u>Theorem 7.1</u>. 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 $|t_i| \neq 1$.

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 <u>symmetric</u> if $c(t) = t^{r+s-n}\, c(\frac{1}{t})$ and <u>strictly diagonally dominant</u> if the degree of c(t) is even and if $|a_\ell| > \sum_{\substack{i=0 \\ i\neq \ell}}^{r+s-n} |a_i|$. (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 $K_h D_h^2 u_j$ 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.)

<u>Lemma 7.2.</u> 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

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

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

Example 7.3. In the special case where $c(t) = a_0t^2 + (1-2a_0)t + a_0$ the assumptions of the lemma hold if $0 \leq a_0 < \frac{1}{4}$. 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, $(1 \leq i \leq m)$, on the roots of the characteristic equation $c(t) = 0$ will be illustrated by means of the next example.

Example 7.4. Let $c(t) = a_0t^2 + (1-2a_0)t + a_0$ 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 $\left[x_{j-2}, x_{j+2}\right]$. 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 $\left(-u_{j-2} + 16u_{j-1} - 30u_j + 16u_{j+1} - u_{j+2}\right) / 12h^2 = f(x_j)$. The left hand side can also we written as

$$\frac{1}{12h^2}\left[-(u_{j-2}-2u_{j-1}+u_j)+14\,(u_{j-1}-2u_j+u_{j+1})-(u_j-2u_{j+1}+u_{j+2})\right],$$

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 $t_i = 7 \pm 4\sqrt{3}$, so that $|t_i| \neq 1$.

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

$$(a_0u_{j-2} + (1-4a_0)u_{j-1} + (-2+6a_0)u_j + (1-4a_0)u_{j+1} + a_0u_{j+2})/h^2 = \frac{1}{2}\Big(f(z_1)+f(z_2)\Big).$$

The corresponding characteristic polynomial is $c(t) = a_0t^2 + (1-2a_0)t + a_0$. 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 $z_1 = x_j - q_0h$ and $z_2 = x_j + q_0h$, where $q_0 = \sqrt{1-\sqrt{\frac{11}{15}}} \approx 0.38$ or $q_0 = \sqrt{1+\sqrt{\frac{11}{15}}} \approx 1.36$, 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 $q_0 = \sqrt{1-\sqrt{\frac{11}{15}}}$. 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.

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

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

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

$$c_{J+r-s-j}(t) = t^{N_j} c_j\left(\frac{1}{t}\right), \quad (r+s-n)/2 = k_0 \le j \le r-1 \ , \tag{7.14}$$

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

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

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 $c_j(t)$, $\left((r+s-n) / 2 = k_0 \leq j \leq r-1\right)$, and $p_j(t)$, $\left(0 \leq j \leq \max N_j - N\right)$, are linearly dependent. Hence we have shown the following

<u>Theorem (7.5)</u>. 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 $||e_h|| \leq K||\tau_h||$.

<u>Example 7.6</u>. 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 $c_1(t) \not\equiv c(t)$. If the degree of $c_1(t)$ is three then stability is guaranteed if there is no constant b such that $c_1(t) = b(a_0 t^3 + (1-2a_0)t^2 + a_0 t) + (1-b)\left(a_0 t^2 + (1-2a_0)t + a_0\right)$.

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

$$Ly(x) \equiv y'' + y' - 2y = 2(1-6x)e^{x}$$

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 $\omega \equiv r + s + 1$ 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 <u>optimal</u> 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 $\left[x_{j-r}, x_{j+s}\right]$.

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 $0.438\ 10^{-1}$ 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 $z_i = x_{j\pm\frac{1}{2}} \pm \xi h$. There are two values of ξ 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 ξ 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.

<u>Example 8.2</u>. Consider the equation

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

with boundary conditions

$$y'(0) = y(1) = 0 .$$

TABLE 8.1

#		r s m c o	J = 4	J = 8	J = 16	J = 32	J = 64	a
1		1 1 2 D 2	$.438^{-1}$	$.111^{-1}$	$.281^{-2}$	$.704^{-3}$	$.176^{-3}$	2
2		1 1 2 B 2	$.111^{-1}$	$.279^{-2}$	$.700^{-3}$	$.175^{-3}$	$.439^{-4}$	2
3		1 1 2 A 4	$.830^{-4}$	$.534^{-5}$	$.337^{-6}$	$.211^{-7}$	$.132^{-8}$	4
4		1 1 3 C 4	$.516^{-3}$	$.325^{-4}$	$.203^{-5}$	$.128^{-6}$	$.798^{-8}$	4
5		1 1 4 B 6	$.213^{-6}$	$.337^{-8}$	$.528^{-10}$	$.817^{-12}$	* *	6
6		1 1 4 B 6	$.704^{-6}$	$.106^{-7}$	$.165^{-9}$	$.257^{-11}$	* *	6
7		1 1 5 E 6	$.156^{-6}$	$.249^{-8}$	$.390^{-10}$	$.628^{-12}$	* *	6
8	1 1 1 D 2	2 2 1 D 4	$.291^{-1}$	$.265^{-2}$	$.204^{-3}$	$.147^{-4}$	$.987^{-6}$	4
9	1 3 1 D 3	2 2 1 D 4	$.440^{-2}$	$.233^{-3}$	$.897^{-5}$	$.310^{-6}$	$.160^{-7}$	(4)
10	1 4 1 D 4	2 2 1 D 4		$.369^{-4}$	$.458^{-5}$	$.326^{-6}$	$.207^{-7}$	4
11	1 1 2 A 4	2 2 1 D 4	$.533^{-3}$	$.719^{-4}$	$.518^{-5}$	$.336^{-6}$	$.212^{-7}$	4
12	1 1 3 C 4	2 2 1 D 4	$.884^{-3}$	$.781^{-4}$	$.528^{-5}$	$.338^{-6}$	$.212^{-7}$	4
13	1 1 5 E 6	2 2 2 A 6	$.571^{-7}$	$.185^{-8}$	$.367^{-10}$	$.601^{-12}$	* *	6
14	1 1 5 E 6	2 2 2 A 6	$.260^{-4}$	$.745^{-6}$	$.122^{-7}$	$.160^{-9}$	$.215^{-11}$	(6)

* * 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 ξ 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).

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

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

with boundary conditions

$$y'(0) = 0 \quad \text{and} \quad y(1) = 1 + e^{\frac{-1}{2\varepsilon}} \approx 1.$$

TABLE 8.2

#	r s m c o	r s m c o	J = 4	J = 8	J = 16	J = 32	J = 64	a
1	0 1 0 - 1	1 1 1 D 2	$.117^{0}$	$.494^{-1}$	$.226^{-1}$	$.108^{-1}$	$.526^{-2}$	1
2	0 2 0 - 2	1 1 1 D 2	$.240^{-1}$	$.492^{-2}$	$.110^{-2}$	$.261^{-3}$	$.633^{-4}$	2
3	0 1 1 * 2	1 1 1 D 2	$.250^{-1}$	$.631^{-2}$	$.158^{-2}$	$.396^{-3}$	$.989^{-4}$	2
4	0 1 1 ** 2	1 1 1 D 2	$.240^{-1}$	$.492^{-2}$	$.110^{-2}$	$.261^{-3}$	$.633^{-4}$	2
5	0 1 0 - 1	1 1 2 A 4	$.106^{0}$	$.466^{-1}$	$.219^{-1}$	$.106^{-1}$	$.521^{-2}$	1
6	0 2 0 - 2	1 1 2 A 4	$.362^{-1}$	$.786^{-2}$	$.183^{-2}$	$.443^{-3}$	$.109^{-3}$	2
7	0 1 1 * 2	1 1 2 A 4	$.152^{-1}$	$.359^{-2}$	$.876^{-3}$	$.216^{-3}$	$.538^{-4}$	2
8	0 1 1 ** 2	1 1 2 A 4	$.338^{-1}$	$.764^{-2}$	$.181^{-2}$	$.440^{-3}$	$.108^{-3}$	2
9	0 1 1 A 3	1 1 2 A 4	$.531^{-3}$	$.602^{-4}$	$.711^{-5}$	$.862^{-6}$	$.106^{-6}$	3
10	0 1 2 C 3	1 1 2 A 4	$.146^{-2}$	$.171^{-3}$	$.207^{-4}$	$.254^{-5}$	$.316^{-6}$	3
11	0 4 0 - 4	1 1 2 A 4	$.344^{-2}$	$.164^{-3}$	$.898^{-5}$	$.526^{-6}$	$.317^{-7}$	4
12	0 1 3 E 4	1 1 2 A 4	$.482^{-4}$	$.324^{-5}$	$.208^{-6}$	$.132^{-7}$	$.828^{-9}$	4
13	0 2 1 A 4	1 1 2 A 4	$.165^{-4}$	$.108^{-5}$	$.730^{-7}$	$.472^{-8}$	$.300^{-9}$	4
14	0 2 1 A 4	1 1 2 A 4	$.563^{-3}$	$.249^{-4}$	$.129^{-5}$	$.734^{-7}$	$.437^{-8}$	4

* $z_1 = x_0$

** $z_1 = x_1$

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

$$y(x) = e^{-x^2/2\varepsilon} + x^2 ,$$

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 $a^1(x) \equiv 0$), 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

J = 4	J = 8	J = 16	J = 32	J = 64	a
$.784^0$	$.248^0$	$.100^0$	$.453^{-1}$	$.216^{-1}$	1

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.

2. INITIAL VALUE PROBLEMS IN ORDINARY DIFFERENTIAL EQUATIONS

In this section the construction of finite difference approximations to the first order linear system

$$(2.1) \qquad \mathcal{L}\, w(t) \equiv A(t)w'(t) - B(t)w(t) = f(t), \quad 0 \le t \le T ,$$

will be discussed. Here $w(t) \equiv \left(w_1(t),\ w_2(t),\ \cdots, w_{J-1}(t)\right)^T$, $f(t) \equiv \left(f_1(t),\ f_2(t), \cdots,\ f_{J-1})\right)^T$ 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 $w(t)$, $(0 \le t \le T)$, 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 <u>mesh</u> S_h by

$$S_h \equiv \{t_j : 0 = t_0 < t_1 < \cdots < t_N = T\} .$$

The <u>meshspacings</u> are $(\Delta t)_n = t_n - t_{n-1}$, $(1 \le n \le N)$. Let $\Delta t \equiv \max_n (\Delta t)_n$, so that $(\Delta t)_n = \Delta t = T/N$ for uniform meshes. Assume that $\min(\Delta t)_n \ge \frac{1}{K}\Delta t$, for some constant $K \ge 1$.

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

$$(2.2) \qquad \mathcal{L}_h u^n \equiv \sum_{\ell=-\rho_n}^{0} D_{n,\ell} u^{n+\ell} = \sum_{\ell=1}^{\mu_n} E_{n,\ell} f(\zeta_{n,\ell}), \quad 1 \le n \le N ,$$

with initial condition $u^0 = w^0$.

Here $u^n \equiv (u_1^n, u_2^n, \cdots, u_{J-1}^n)^T$ is the approximation to $w(t_n)$. For each n the points $\zeta_{n,\ell}$, $(1 \le \ell \le \mu_n)$, are distinct points, that will normally lie in the interval $\left[t_{n-\rho_n}, t_n\right]$. The coefficients $D_{n,\ell}$ and $E_{n,\ell}$ are (J-1) $\times$ (J-1) matrices. These coefficients may be determined by requiring that

$$\sum_{\ell=-\rho}^{0} D_\ell p(t_{n+\ell}) - \sum_{\ell=1}^{\mu} E_\ell L p(\zeta_\ell) = 0 ,$$

for all $p(t) \in P^{J-1}_{\rho+\mu-1}$. Here P^{J-1}_d 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

$$(2.3)\quad \begin{bmatrix} I & 0 & \cdots & 0 & -L^T_{0,1} & \cdots & -L^T_{0,\mu} \\ 0 & I & & 0 & -L^T_{1,1} & \cdots & -L_{1,\mu} \\ \vdots & & & & & & \vdots \\ 0 & 0 & & I & -L^T_{\rho,1} & \cdots & -L^T_{\rho,\mu} \\ 0 & 0 & & 0 & -L^T_{\rho+1,1} & \cdots & -L^T_{\rho+1,\mu} \\ \vdots & & & & & & \vdots \\ 0 & 0 & & 0 & -L^T_{\rho+\mu-1,1} & \cdots & -L^T_{\rho+\mu-1,\mu} \end{bmatrix} \begin{bmatrix} D^T_{-\rho} \\ D^T_{-\rho+1} \\ \vdots \\ D^T_0 \\ E^T_1 \\ \vdots \\ E^T_\mu \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

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

$$L_{i,\ell} \equiv L_0\, \omega^i(\zeta_\ell)A(\zeta_\ell) - \omega^i(\zeta_\ell)B(\zeta_\ell) ,$$

with $L_0\omega^i(t) \equiv \frac{d\omega^i(t)}{dt}$. The scalar polynomials $\omega^i(t)$, $(0 \leq i \leq \rho+\mu-1)$, are linearly independent and chosen such that

$$\omega^i(t_{n-\rho+\ell}) = \delta_{i\ell}, \quad 0 \leq i, \ell \leq \rho \quad ,$$

and

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

For example, these polynomials may be defined in a similar fashion as the polynomials $w^i(x)$ in Section (I.2). (Equations (I.2.13) and (I.2.14).) The coefficient matrices D_ℓ and E_ℓ in (2.3) are uniquely defined up to a certain normalization, provided that the matrix

$$C \equiv \begin{bmatrix} L_{\rho+1,1} & \cdots & L_{\rho+\mu-1,1} \\ \vdots & & \vdots \\ L_{\rho+1,\mu} & \cdots & L_{\rho+\mu-1,\mu} \end{bmatrix}$$

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

$$\tilde{C} \equiv \begin{bmatrix} L_0\omega^{\rho+1}(\zeta_1)A(\zeta_1) & \cdots & L_0\omega^{\rho+\mu-1}(\zeta_1)A(\zeta_1) \\ \vdots & & \vdots \\ L_0\omega^{\rho+1}(\zeta_\mu)A(\zeta_\mu) & \cdots & L_0\omega^{\rho+\mu-1}(\zeta_\mu)A(\zeta_\mu) \end{bmatrix}$$

has rank $(\mu-1)(J-1)$. This in turn is true if $\mathcal{E} \neq 0$ for all small enough Δt. Here $\mathcal{E}$ is the analogue of the normalizing factor E in the previous chapter, (Equation (I.2.12)), and can be expressed as

$$(2.4) \qquad \mathcal{E} \equiv (-1)^{\mu} \begin{vmatrix} 1 & L_0\omega^{\rho+1}(\zeta_1) & \cdots & L_0\omega^{\rho+\mu-1}(\zeta_1) \\ & \vdots & & \vdots \\ 1 & L_0\omega^{\rho+1}(\zeta_\mu) & \cdots & L_0\omega^{\rho+\mu-1}(\zeta_\mu) \end{vmatrix} .$$

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 Δt. Here c and C are positive constants that do not depend on Δt. 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.

<u>Theorem 2.1</u>. Assume that $\rho \geq 1$ and that $\mathcal{E} \neq 0$. 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 $\mathcal{E} \neq 0$. If the points ζ_i are the critical points of $\tilde{\omega}^{\rho+\mu}(t) \equiv \prod_{\ell=1}^{\mu-1} (t-\gamma_\ell) \prod_{\ell=-\rho}^{0} (t-t_{n+\ell})$ then the order is at least $\rho+\mu$. (Thus there is a $(\mu-1)$-parameter family of such points with parameters γ_ℓ, $(1 \leq \ell \leq \mu-1)$.)

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.

<u>Example 3.1.</u> With $\rho = 1$, $\mu = 1$ and $\zeta_1 = t_{n-\frac{1}{2}} \equiv \frac{1}{2}(t_n + t_{n-1})$ one obtains the

Box scheme of Keller (1969). This difference approximation may be written as

$$D_{-1}u^{n-1} + D_0 u^n = f(t_{n-\frac{1}{2}}) ,$$

where $$D_{-1} = \frac{-1}{\Delta t} A(t_{n-\frac{1}{2}}) - \tfrac{1}{2}B(t_{n-\frac{1}{2}}) ,$$

and $$D_0 = \frac{1}{\Delta t} A(t_{n-\frac{1}{2}}) - \tfrac{1}{2}B(t_{n-\frac{1}{2}}) .$$

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

Next if $\rho = 1$ and $\mu = 2$, with $\zeta_1 = t_{n-1}$ and $\zeta_2 = t_n$ then the matrix equation (2.3) becomes

$$(3.1) \qquad \begin{bmatrix} I & 0 & -L_{0,1}^T & -L_{0,2}^T \\ 0 & I & -L_{1,1}^T & -L_{1,2}^T \\ 0 & 0 & -L_{2,1}^T & -L_{2,2}^T \end{bmatrix} \begin{bmatrix} D_{-1}^T \\ D_0^T \\ E_1^T \\ E_2^T \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} ,$$

where $$L_{i,1} \equiv L_0\omega^i(t_{n-1})A(t_{n-1}) - \omega^i(t_{n-1})B(t_{n-1}), \qquad 0 \le i \le 2 ,$$

and $$L_{i,2} = L_0\omega^i(t_n)A(t_n) - \omega^i(t_n)B(t_n), \qquad 0 \le i \le 2 .$$

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

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

$$L_{1,1} = \frac{1}{\Delta t} A(t_{n-1}) , \qquad L_{1,2} = \frac{1}{\Delta t} A(t_n) - B(t_n) ,$$

$$L_{2,1} = \frac{-1}{\Delta t} A(t_{n-1}) \quad , \qquad\qquad L_{2,2} = \frac{1}{\Delta t} A(t_n)$$

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_i 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

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

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

<u>Example 3.2</u>. With $\rho = 2$, $\mu = 1$ and $\zeta_1 = t_{n-2} + \xi\, \Delta t$ the difference approximation to (2.1) becomes

$$A(\zeta_1) \frac{1}{\Delta t}\left[\tfrac{1}{2}(2\xi-1)u^n - 2(\xi-1)u^{n-1} + \tfrac{1}{2}(2\xi-3)u^{n-2}\right] - B(\zeta_1)\left[\tfrac{1}{2}\xi(\xi-1)u^n - \xi(\xi-2)u^{n-1} + \right.$$
$$\left. + \tfrac{1}{2}(\xi-1)(\xi-2)\right] = f(\zeta_1) \; .$$

(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 $\xi = 1 \pm \frac{1}{3}\sqrt{3} \cong 1.58$.

If $\xi = 2$ then the above formula reduces to

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

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

<u>Example 3.3</u>. Consider the constant coefficient problem

$$Aw'(t) - Bw(t) = f(t)\ , \qquad 0 \le t \le T\ , \tag{3.2}$$

$$w(0) = w^0\ , \tag{3.2a}$$

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 ζ_1 and ζ_2.

Let $E_1 = -L_0\omega^2(\zeta_2)\,A + \omega^2(\zeta_2)\,B$. Then, using the fact that A and B commute, it is found that

$$E_2 = L_0\omega^2(\zeta_1)A - \omega^2(\zeta_1)B,$$

$$D_{-1} = E_1\{L_0\omega^0(\zeta_1)A - \omega^0(\zeta_1)B\} + E_2\{L_0\omega^0(\zeta_2)A - \omega^0(\zeta_2)B\}\ ,$$

and $$D_0 = E_1\{L_0\omega^1(\zeta_1)A - \omega^1(\zeta_1)B\} + E_2\{L_0\omega^1(\zeta_2)A - \omega^1(\zeta_2)B\},$$

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

In particular if one uses the Gaussian points

$$\zeta_1 = (t_{n-1} + t_n)/2 - \Delta t\sqrt{3}/6 \quad \text{and} \quad \zeta_2 = (t_{n-1}+t_n)/2 + \Delta t\sqrt{3}/6 ,$$

then the difference equation becomes

$$(3.3) \quad \left\{\frac{-1}{\Delta t} A^2 - \tfrac{1}{2}AB - \frac{\Delta t}{12} B^2\right\} u^{n-1} + \left\{\frac{1}{\Delta t} A^2 - \tfrac{1}{2}AB + \frac{\Delta t}{12} B^2\right\} u^n =$$

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

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).)

<u>Example 3.4</u>. The preceding example suggests that the general ρ-step finite difference approximation to (3.2) with μ collocation points has the form

$$(3.4) \quad \sum_{\ell=-\rho}^{0} \left[\sum_{\nu=0}^{\mu} \alpha_\nu^\ell A^{\mu-\nu} B^\nu \right] u^{n+\ell} = \sum_{\ell=1}^{\mu} \left[\sum_{\nu=0}^{\mu-1} \beta_\nu^\ell A^{\mu-1-\nu} B^\nu \right] f(\zeta_\ell) ,$$

where α_ν^ℓ and β_ν^ℓ are constants that depend on the choice of the points ζ_ℓ, $(1 \le \ell \le \mu)$.

To verify this and to show how the coefficients α_ν^ℓ and β_ν^ℓ may be found, consider the scalar equation

$$(3.5) \qquad Lw(t) \equiv w'(t) - \lambda\, w(t) = f(t) ,$$

The finite difference approximation to (3.5) is given by

$$(3.6) \qquad \sum_{\ell=-\rho}^{0} \delta_\ell u^{n+\ell} = \sum_{\ell=1}^{\mu} \epsilon_\ell f(\zeta_\ell) ,$$

where

$$(3.7) \quad \epsilon_\ell = \frac{(-1)^{\mu+\ell+1}}{\mathcal{E}} \begin{vmatrix} L\omega^{\rho+1}(\zeta_1) & \cdots & L\omega^{\rho+\mu-1}(\zeta_1) \\ \vdots & & \vdots \\ L\omega^{\rho+1}(\zeta_{\ell-1}) & \cdots & L\omega^{\rho+\mu-1}(\zeta_{\ell-1}) \\ L\omega^{\rho+1}(\zeta_{\ell+1}) & \cdots & L\omega^{\rho+\mu-1}(\zeta_{\ell+1}) \\ \vdots & & \vdots \\ L\omega^{\rho+1}(\zeta_\mu) & \cdots & L\omega^{\rho+\mu-1}(\zeta_\mu) \end{vmatrix}$$

and

$$(3.8) \qquad \delta_\ell = \sum_{k=1}^{\mu} L\omega^{\rho+\ell}(\zeta_k)\,\epsilon_k , \quad -\rho \le k \le 0 .$$

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

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

for certain constants α_ν^ℓ and β_ν^ℓ .

In order to show that these constants are the same as the constants in equation (3.4), note that by definition

$$\sum_{\ell=1}^{\mu} \epsilon_\ell L\omega^k(\zeta_\ell) = 0 , \qquad \rho+1 \leq k \leq \rho + \mu-1 ,$$

i.e.,

$$q_k(\lambda) \equiv \sum_{\ell=1}^{\mu} \left[\sum_{\nu=0}^{\mu-1} \beta_\nu^\ell \lambda^\nu \right] \left[L_0\omega^k(\zeta_\ell)-\omega^k(\zeta_\ell)\lambda \right] = 0 ,$$

for all λ and for $\rho + 1 \leq k \leq \rho + \mu - 1$.

So $q_k(\lambda) \equiv 0$ and therefore $q_k(C) \equiv 0$ for all matrices C. In particular with $C = A^{-1}B$ one obtains

$$\sum_{\ell=1}^{\mu} \left[\sum_{\nu=0}^{\mu-1} \beta_\nu^\ell (A^{-1}B)^\nu \right] \left[L_0\omega^k(\zeta_\ell)I-\omega^k(\zeta_\ell)A^{-1}B \right] = 0 .$$

Upon multiplication by A^μ and using the assumption that A and B commute one obtains

$$\sum_{\ell=1}^{\mu} \left[\sum_{\nu=0}^{\mu-1} \beta_\nu^\ell A^{\mu-1-\nu} B^\nu \right] \left[L_0\omega^k(\zeta_\ell)A-\omega^k(\zeta_\ell)B \right] = 0 , \quad \rho+1 \leq k \leq \rho + \mu-1 .$$

Thus if one defines

$$E_\ell \equiv \sum_{\nu=0}^{\mu-1} \beta_\nu^\ell A^{\mu-1-\nu} B^\nu ,$$

then the last μ-2 relations in the matrix equation (2.3) are satisfied. Further, the first ρ+1 relations of (2.3) are automatically satisfied by letting

$$D_\ell \equiv \sum_{\nu=1}^{\mu} E_\nu \left[L_0\omega^{\rho+\ell}(\zeta_\nu)A(\zeta_\nu)-\omega^{\rho+\ell}(\zeta_\nu)B(\zeta_\nu) \right] .$$

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_ℓ and E_ℓ 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

$$\sum_{\ell=-\rho}^{0} \bar{\delta}_\ell \, u^{n+\ell} = 0 \, , \qquad \rho \leq n \leq N \, , \tag{4.1}$$

where

$$\bar{\delta}_\ell = \frac{(-1)^\mu}{\mathcal{E}} \begin{vmatrix} L_0\omega^{\rho+\ell}(\zeta_1) & L_0\omega^{\rho+1}(\zeta_1) & \cdots & L_0\omega^{\rho+\mu-1}(\zeta_1) \\ \vdots & \vdots & & \vdots \\ L_0\omega^{\rho+\ell}(\zeta_\mu) & L_0\omega^{\rho+1}(\zeta_\mu) & \cdots & L_0\omega^{\rho+\mu-1}(\zeta_\mu) \end{vmatrix} .$$

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

The mesh is taken to be uniform in this section, so that $(\Delta t)_n = \Delta t = 1/N$, $(1 \leq n \leq N)$. The approximation is said to be <u>stable</u> 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 η_i of the <u>characteristic equation</u>

$$\sigma_0(\eta) \equiv \Delta t \sum_{\ell=-\rho}^{0} \bar{\delta}_\ell \, \eta^{\rho+\ell} = 0 \tag{4.2}$$

satisfy

$$|\eta_i| \leq 1 , \qquad 1 \leq i \leq \rho , \tag{4.3}$$

and

$$|\eta_i| = 1 \text{ implies } \eta_i \text{ is a simple root.} \tag{4.3a}$$

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

$$\eta_1 = 1 ,$$

$$|\eta_i| < 1 , \qquad 2 \leq i \leq \rho .$$

<u>Example 4.1</u>. If $\rho = 1$ then the coefficients in (4.1) are given by $\bar{\delta}_0 = \frac{-1}{\Delta t}$ and $\bar{\delta}_1 = \frac{1}{\Delta t}$. The characteristic equation has one root $\eta = 1$ and thus this approximation is stable.

If $\rho = 2$ and $\mu = 1$ with $\zeta_1 = t_{n-2} + \xi \Delta t$, 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 $\bar{\delta}_\ell$ 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 $\zeta_1 = t_{n-2} + \xi_1 \Delta t$ and $\zeta_2 = t_{n-2} + \xi_2 \Delta t$. 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 Δt 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 $\mathcal{L}_\lambda w(t) \equiv w'(t) - \lambda w(t) = 0$, 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

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

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

$$(4.4) \qquad \tilde{\delta}_\ell = \frac{(-1)^\mu}{\mathcal{E}} \begin{vmatrix} \mathcal{L}_\lambda \omega^{\rho+\ell}(\zeta_1) & \mathcal{L}_\lambda \omega^{\rho+1}(\zeta_1) & \cdots & \mathcal{L}_\lambda \omega^{\rho+\mu-1}(\zeta_1) \\ \vdots & \vdots & & \vdots \\ \mathcal{L}_\lambda \omega^{\rho+\ell}(\zeta_\mu) & \mathcal{L}_\lambda \omega^{\rho+1}(\zeta_\mu) & \cdots & \mathcal{L}_\lambda \omega^{\rho+\mu-1}(\zeta_\mu) \end{vmatrix} .$$

For stability it is again necessary that the roots η_i of the characteristic equation $\tilde{q}(\eta) \equiv \Delta t \sum_{\ell=-\rho}^{0} \tilde{\delta}_\ell \eta^{\rho+\ell} = 0$ satisfy the root condition (4.3), (4.3a). From (4.4) it follows that $\tilde{q}(\eta)$ has the form

$$(4.5) \qquad \tilde{q}(\eta) = \sigma_0(\eta) + \sum_{\nu=1}^{\mu} (\Delta t\,\lambda)^{\nu}\, \sigma_{\nu}(\eta) \ ,$$

where $\sigma_0(\eta)$ is given by (4.2) and where $\sigma_{\nu}(\eta) \in P_{\rho}$.

The <u>stability region</u> 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 λ with $\mathrm{Re}(\lambda) << 0$, 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 $0 \in \delta S$.. Here δS 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 <u>stable at ∞</u>, 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 <u>strictly stable at ∞</u>.) There are difference approximations however, that are strictly stable at ∞, 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 <u>$A(\alpha)$ stability</u>. A difference method is said to be $A(\alpha)$-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_α 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 ζ_ℓ has on the stability region S will be illustrated.

<u>Example 4.2</u>. Let $\rho = 1$ and $\mu = 1$ with $\zeta_1 = t_{n-1} + \xi\,\Delta t$. So the difference approximation to $\mathcal{L}_\lambda w(t) \equiv w'(t) - \lambda w(t) = 0$ 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

$\tilde{q}(\eta) = \tilde{\delta}_0\eta + \tilde{\delta}_{-1} = \sigma_0(\eta) + \Delta t\lambda\ \sigma_1(\eta)$, 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 ∞ if $\xi \geq \frac{1}{2}$. It is easy to check that the approximation is in fact $A(\alpha)$-stable for such ξ. The method is strictly stable at ∞ 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 $\zeta_1 = t_{n-2} + \xi\ \Delta t$, 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 ∞ 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 $|\eta_i| \leq 1$ provided $\xi \geq 1 + \frac{1}{2}\sqrt{2} \approx 1.7$. For example, the second order formula of Gear is strictly stable at ∞, but the third order formula corresponding to $\xi = 1 + \frac{1}{3}\sqrt{3}$, (see Example (3.2)), is not stable at ∞.

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) $p(x, t)\, y_t(x, t) = L(t)\, y(x, t) + f(x, t)\ , \qquad 0 \le x \le 1\ , \quad 0 \le t \le T\ ,$

where

$$L(t)\ y(x, t) \equiv y_{xx}(x, t) + q(x, t)\ y_x(x, t) + r(x, t)\ y(x, t),$$

with initial condition

(5.1a) $$y(x, 0) = g(x)\ , \qquad 0 \le x \le 1\ ,$$

and boundary conditions

(5.1b) $$a_0 y(0, t) + b_0 y_x(0, t) = g_0(t)\ , \qquad 0 \le t \le T\ ,$$

and

(5.1c) $$a_1 y(1, t) + b_1 y_x(1, t) = g_1(t), \qquad 0 \le t \le T\ .$$

It is assumed that $0 < p_* \le p(x, t) \le p^* < \infty$, so that (1.1) 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 <u>mesh</u> $R_h(T)$ by

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

The <u>mesh spacings</u> are $h_j \equiv x_j - x_{j-1}$, $(1 \le j \le J)$, and $(\Delta t)_n \equiv t_n - t_{n-1}$, $(1 \le n \le N)$.

Let $h \equiv \max_j h_j$ and $\Delta t \equiv \max_n (\Delta t)_n$. 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 $R_h(T)$ let $w_j^n \equiv w(x_j, t_n)$,

$$w_h^n \equiv (w_1^n, w_2^n, \cdots, w_{J-1}^n)^T, \quad \|w_h^n\| \equiv \max_j |w_j^n| \text{ and } \|w_h^n\|_2 \equiv \{\sum_{j=1}^{J-1} (w_j^n)^2\}^{\frac{1}{2}}.$$

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

$$(5.3) \quad K_h(t)\left[p(x_j, t)w'(x_j, t)\right] = L_h(t)w(x_j, t) + K_h(t)f(x_j, t), \qquad 1 \le j \le J-1 ,$$

where

$$L_h(t)\ w(x_j, t) \equiv \sum_{i=-r_j}^{s_j} d_{j,i}(t)w(x_{j+i}, t) ,$$

and

$$K_h(t)w(x_j, t) \equiv \sum_{i=1}^{m_j} e_{j,i}(t)w(z_{j,i}, t) ,$$

with initial data

$$(5.3a) \qquad w(x_j, 0) = g(x_j) , \qquad 0 \le j \le J.$$

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 (I.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 <u>the collocation points</u> z_i <u>coincide with the mesh points</u> x_{j+i}, i.e. if z_i is a collocation point then $z_i = x_{j+\ell}$ for some integer ℓ. For example, if m = 1 one may have

$$K_h(t)\ w(x_j, t) = e_1(t)\ w(x_j, t) = w(x_j, t), \qquad 1 \le j \le J-1 ,$$

and if m = 3

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

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

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

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.

<u>Theorem 5.2</u>. For each t, $(0 \leq t \leq T)$, 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, $(1 \leq i \leq m)$, $z_i = x_{j+\ell_i}$ for some integer ℓ_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, $(0 \leq t \leq T)$, there are also approximations to the boundary conditions (5.1b, c) of the form

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

and

$$\sum_{i=-r_J}^{0} d_{J,i}\, w(x_{J+i}, t) = g_1(t) \ , \qquad 0 \leq t \leq T \ . \tag{5.3c}$$

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

$$K_h(t)\, w_h'(t) = L_h(t)\, w_h(t) + \tilde{f}_h(t)\ , \qquad 0 \leq t \leq T\ , \tag{5.4}$$

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

$$w_h(0) = g_h = (g(x_1), \cdots, g(x_{J-1}))^T\ . \tag{5.4a}$$

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

(6.1) $$y_t(x, t) = Ly(x, t) \equiv y_{xx}(x, t) \quad , \quad 0 \leq x \leq 1, \quad 0 \leq t \leq T$$

with initial condition

(6.1a) $$y(x, 0) = g(x), \qquad 0 \leq x \leq 1 \ ,$$

and boundary conditions

(6.1b) $$y(0, t) = 0 \ , \qquad 0 \leq t \leq T \ ,$$

and

(6.1c) $$y_x(1, t) = 0 \ , \qquad 0 \leq t \leq T \ .$$

Discretizing in space first as in Section 5 gives the equations

(6.2) $$\sum_{i=1}^{m_j} \bar{e}_{j,i}\, w'(z_{j,i}, t) = \sum_{i=-r_j}^{s_j} \bar{d}_{j,i}\, w(x_{j+i}, t), \quad 1 \leq j \leq J-1, \quad 0 \leq t \leq T,$$

where the coefficients $\bar{d}_{j,i}$ and $\bar{e}_{j,i}$ are given by equation (I.2.16) and (I.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

$$\sum_{i=-r_J}^{0} b_i\, w(x_{J+i}, t) = 0 \ , \qquad 0 \leq t \leq T \ .$$

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

$$K_h \, w_h'(t) = L_h \, w_h(t) \, , \tag{6.3}$$

with initial data

$$w_h(0) = g_h \, . \tag{6.3a}$$

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.

<u>Theorem 6.1</u>. 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 $\mathrm{Re}(\lambda_{h,i}) \leq 0$.
5. Let V_h be the matrix of eigenvectors. Then $\|V_h\|_2 \; \|V_h^{-1}\|_2 \leq c_2$.

<u>Proof</u>. Let $e_h(t) \equiv y_h(t) - w_h(t)$. Then $K_h \, e_h'(t) = L_h e_h(t) + \tau_h(t)$. Therefore

$$e_h(t) = e^{tK_h^{-1}L_h} e_h(0) + \int_0^t e^{(t-r)K_h^{-1}L_h} \tau_h(r)\,dr$$

$$= V_h e^{t\Lambda_h} V_h^{-1} e_h(0) + \int_0^t V_h e^{(t-r)\Lambda_h} V_h^{-1} \tau_h(r)dr,$$

where Λ_h is the diagonal matrix of eigenvalues.

If n_s is the order of consistency of (6.3), then the estimate

$$\|e_h(t)\|_2 \le c_2 \left\{ \|e_h(0)\|_2 + T \max_{0 \le t \le T} \|\tau_h(t)\|_2 \right\}$$

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

$$\frac{1}{\Delta t} \sum_{\ell=-\rho}^{0} \left[\sum_{\nu=0}^{\mu} a_\nu^\ell (\Delta t)^\nu K_h^{\mu-\nu} L_h^\nu \right] u_h^{n+\ell} = 0 , \quad 1 \le n \le N . \tag{6.4}$$

The coefficients a_ν^ℓ are assumed to be independent of n, so that the same difference approximation is used at each time-step. The initial data are $u_h^0 = g_h$. Further, if $\rho > 1$ then sufficiently accurate approximations to u_h^n, $(1 \le n \le \rho - 1)$, 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 <u>stable in the ℓ_2-norm</u> if there exists a constant K, that does not depend on h and Δt, such that

$$\|u_h^n\|_2 \leq K\{\sum_{\ell=0}^{\rho-1} \|u_h^\ell\|_2^2\}^{1/2} , \quad \rho \leq n \leq N ,$$

for all small enough h and Δt. Sufficient conditions for stability, and hence for convergence, are given in the following theorem.

<u>Theorem 6.2</u>. The solution u_h^n 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.

2a. 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_α of the form

$$T_\alpha \equiv \{z \in C : \pi-\alpha < \arg z < \pi+\alpha ; \ z \neq 0\} ,$$

which has the property that if $z \in T_\alpha$, then the roots $\eta_i(z)$ of the characteristic equation

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

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

$$\|V_h\|_2 \; \|V_h^{-1}\|_2 \leq c_2 .$$

<u>Proof</u>. By assumption 3 one can write $u_h^n = V_h \, c_h^n$. Substitution into (6.4), also using 2 and 2a, gives

$$\frac{1}{\Delta t} K_h^\mu \sum_{\ell=-\rho}^{0} \sum_{\nu=0}^{\mu} a_\nu^\ell \, \Delta t^\nu \, K_h^{-\nu} \, L_h^\nu \, V_h \, c_h^{n+\ell}$$

$$= \frac{1}{\Delta t} K_h^\mu \sum_{\ell=-\rho}^{0} \sum_{\nu=0}^{\mu} a_\nu^\ell \, \Delta t^\nu \, V_h \, \Lambda_h^\nu \, c_h^{n+\ell}$$

$$= \frac{1}{\Delta t} K_h^\mu \, V_h \sum_{\ell=-\rho}^{0} \sum_{\nu=0}^{\mu} a_\nu^\ell \, \Delta t^\nu \, \Lambda_h^\nu \, c_h^{n+\ell} = 0 ,$$

where Λ_h is the diagonal matrix of eigenvalues. Hence each component $c_{h,j}^n$ of c_h^n satisfies the difference equation

$$\sum_{\ell=-\rho}^{0} \sum_{\nu=0}^{\mu} a_\nu^\ell \, (\Delta t \, \lambda_{h,i})^\nu \, c_{h,j}^{n+\ell} = 0, \quad \rho \leq n \leq N = T/\Delta t. \tag{6.5}$$

Let $\underline{c}_{h,j}^n$ denote the vector $(c_{h,j}^n , c_{h,j}^{n-1} , \cdots , c_{h,j}^{n-\rho+1})^T$ and let

$\|\underline{c}_{h,j}^n\|_2 \equiv \{ \sum_{m=0}^{\rho-1} | c_{h,j}^{n-m} |^2 \}^{\frac{1}{2}}$. Define the $\rho \times \rho$ matrix $A(z)$ by

$$A(z) \equiv \begin{bmatrix} \frac{-p_{\rho-1}(z)}{p_\rho(z)} & \frac{-p_{\rho-2}(z)}{p_\rho(z)} & \cdots & \frac{-p_0(z)}{p_\rho(z)} \\ 1 & 0 & & 0 \\ 0 & 1 & & 0 \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

where $p_m(z) \equiv \sum_{\nu=0}^{\mu} a_\nu^m z^\nu$, $0 \leq m \leq \rho$. Then (6.5) can be written as a one step difference equation

$$\underline{c}_{h,j}^n = A(\Delta t \lambda_{h,j})\, \underline{c}_{h,j}^{n-1} , \quad \rho \leq n \leq N , \quad 1 \leq j \leq J-1 .$$

The eigenvalues $\eta_m(\Delta t \lambda_{h,j})$ of $A(\Delta t \lambda_{h,j})$ are precisely the roots of the characteristic equation

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

The definition of T_α 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

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

Thus under these conditions one has

$$|c^n_{h,j}| \leq \|\underline{c}^n_{h,j}\|_2 = \| \left(A(\Delta t \lambda_{h,j})\right)^{n-\rho+1} \underline{c}^{\rho-1}_{h,j} \|_2 \leq K_1 \|\underline{c}^{\rho-1}_{h,j}\|_2$$

and therefore

$$\|c^n_h\|_2 = \{\sum_{j=1}^{J-1} |c^n_{h,j}|^2\} \leq K_1 \{\sum_{j=1}^{J-1} \|\underline{c}^{\rho-1}_{h,j}\|_2^2\}^{\frac{1}{2}} = K_1 \{\sum_{j=1}^{J-1} \sum_{m=0}^{\rho-1} |c^m_{h,j}|^2\}^{\frac{1}{2}} = K_1 \{\sum_{m=0}^{\rho-1} \|c^m_h\|_2^2\}^{\frac{1}{2}}$$

so that

$$\|u^n_h\|_2 = \|V_h c^n_h\|_2 \leq \|V_h\|_2 \|c^n_h\|_2 \leq \|V_h\|_2 \, K_1 \{\sum_{m=0}^{\rho-1} \|c^m_h\|_2^2\}^{\frac{1}{2}}$$

$$= \|V_h\|_2 K_1 \{\sum_{m=0}^{\rho-1} \|V_h^{-1} u^m_h\|_2^2\}^{\frac{1}{2}} \leq \|V_h\|_2 \|V_h^{-1}\|_2 K_1 \{\sum_{m=0}^{\rho-1} \|u^m_h\|_2^2\}^{\frac{1}{2}}$$

$$\leq c_2 K_1 \{\sum_{m=0}^{\rho-1} \|u^m_h\|_2^2\}^{\frac{1}{2}} \ .$$

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 $\kappa(V_h) \equiv \|V_h\|_2 \; \|V_h^{-1}\|_2$ 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_α as in Theorem (6.2). A multistep method that admits such a region will be called <u>strongly A(α)-stable</u>. (See Section 4 for a definition of A(α) stability.) Thus if a method is strongly A(α)-stable, then it is A(α)-stable, but the converse need not be true. All A(α)-stable one step methods are strongly A(α)-stable, because the characteristic equation $\tilde{q}(\eta) = 0$ has only one root in that case. Other examples of strongly A(α)-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 $Q \equiv [Q_0]^{-1}Q_{-1}$ of the difference scheme $Q_0 u_h^n = Q_{-1}u_h^{n-1}$ 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 $\mathrm{Re}(z) > 0$. 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.

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

$$Ly \equiv y'' = \lambda y \quad , \qquad y(0) = y'(1) = 0 \ , \tag{6.6}$$

is approximated by an eigenvalue of the discrete eigenvalue problem

$$L_h v_h = \lambda_h K_h v_h , \tag{6.7}$$

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 λ_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_α as in Theorem (6.2) is desirable. If T_α 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_α includes only a part $(-\beta, 0)$ of the negative real axis then Δt must satisfy $\Delta t \leq \frac{\beta}{4} h^2$ in order to guarantee that $\Delta t\, \lambda_{h,i} \in T_\alpha$.

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 λ_h that tend to ∞ 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

$$\frac{u_J - c\, u_{J-1} + (c-1)\, u_{J-2}}{(2-c)\, h} = 0$$

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 (I.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

$$P_h = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & 1 \\ & & & & & 2-c \end{bmatrix} \quad \text{and} \quad \tilde{L}_h = \begin{bmatrix} -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -1 \end{bmatrix}$$

Let $v_h = P_h^{\frac{1}{2}} \tilde{v}_h$. Then the eigenvalue problem $L_h v_h = \lambda_h K_h v_h$ becomes $P_h^{\frac{1}{2}} \tilde{L}_h P_h^{\frac{1}{2}} \tilde{v}_h = \lambda_h \tilde{v}_h$, 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 $V_h = P_h^{\frac{1}{2}} \tilde{V}_h$. Hence

$$\kappa(V_h) \equiv \|V_h\|_2 \, \|V_h^{-1}\|_2 \leq \|\tilde{V}_h\|_2 \, \|\tilde{V}_h^{-1}\|_2 \, \|P_h^{\frac{1}{2}}\|_2 \, \|P_h^{-\frac{1}{2}}\|_2 \leq \max\{|c-2|, \frac{1}{|c-2|}\},$$

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 λ_h. Thus v_h satisfies

$$v_{j+1} - (2 + h^2\lambda)v_j + v_{j-1} = 0, \qquad 1 \leq j \leq J-1.$$

The solution to these difference equations is

$$v_j = c_1 z_h^j + c_2 z_h^{-j}, \quad 0 \leq j \leq J.$$

where z_h and $1/z_h$ are the roots of

$$z^2 - (2 + h^2\lambda)z + 1 = 0 \tag{6.8}$$

But v_h must also satisfy the discrete boundary conditions. In particular, the condition $v_0 = 0$ implies that $v_j = c_1(z_h^j + z_h^{-j})$, $(0 \leq j \leq J)$, and the boundary condition at $x = 1$ leads to the equation

$$(z_h^J - z_h^{-J}) - c(z_h^{J-1} - z_h^{-J+1}) + (c-1)(z_h^{J-2} - z_h^{-J+2}) = 0. \tag{6.9}$$

Consider now the equation obtained from (6.9) by omitting all negative powers, viz.

$$z_h^2 - cz_h + c-1 = 0.$$

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

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

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 $K_h u_h' = L_h u_h$, and if this multistep method is strongly $A(\alpha)$-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 λ_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 ∞ have this property.

<u>Example 6.5</u>. Again consider the eigenvalue problem (6.6) and its discrete version (6.7) with

$$L_h v_j = \frac{1}{h^2} (v_{j-1} - 2v_j + v_{j+1})$$

and

$$K_h v_j = \frac{1}{12} v_{j-1} + \frac{10}{12} v_j + \frac{1}{12} v_{j+1} .$$

Thus this eigenvalue problem arises in the stability analysis when using the fourth order spatial difference approximation discussed in Example (I.4.4). Let the approximation to the boundary condition $y'(1) = 0$ have the form

$$\sum_{i=-r_J}^{0} b_i u_{J+i} = 0 \tag{6.10}$$

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

$$\lambda_h = \frac{12(z-1)^2}{h^2(z^2 + 10z + 1)} \tag{6.11}$$

for each z which is a root of

$$\sum_{i=-r_J}^{0} b_i(z^{J+i} - z^{-(J+i)}) = 0. \tag{6.12}$$

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

$$c_J(z) \equiv \sum_{i=-r_J}^{0} b_i z^{r_J+i}$$

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

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

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

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

has roots $z_1 = 1$, $z_2 \cong 0.381$, $z_3 \cong 0.269 + 0.492\,i$ and $z_4 \cong 0.269 - 0.492\,i$ with $|z_3| = |z_4| \cong 0.561$. 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

$$y_t = y_{xx} + (1 + \pi^2)y\ , \qquad 0 \le x \le 1\ , \qquad 0 \le t \le 3\ ,$$

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 $y(\frac{1}{2},3) = e^3 \cong 20$. Uniform meshes are used with $h = \frac{1}{J}$ and $\Delta t = \frac{1}{N}$. Finite difference approximations have the form

$$\frac{1}{\Delta t} \sum_{\ell=-\rho}^{0} \left[\sum_{\nu=0}^{\mu} a_\nu^\ell (\Delta t)^\nu \ K_h^{\mu-\nu} \ L_h^\nu \right] u_h^{n+\ell} = 0 \ , \tag{7.1}$$

where

$$L_h u_j^n = \sum_{i=-r}^{s} d_i u_{j+i}^n \ ,$$

and

$$K_h u_j^n = \sum_{i=1}^{m} e_i \ u^n(z_{j,i}).$$

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

TABLE 7.1

#	r s m o	ρ c o	J = 5	J = 10	J = 20	J = 40	N
1	1 1 1 2	1 C 2	$.392^{+2}$	$.772^{+1}$	$.279^{+1}$	$.170^{+1}$	6
			$.320^{+2}$	$.585^{+1}$	$.149^{+1}$	$.517^{0}$	15
			$.311^{+2}$	$.560^{+1}$	$.132^{+1}$	$.360^{0}$	30
					$.127^{+1}$	$.321^{0}$	60
						$.311^{0}$	120
2	1 1 1 2	2 G 2	$.599^{+3}$	$.861^{+2}$	$.523^{+2}$	$.459^{+2}$	6
			$.755^{+2}$	$.185^{+2}$	$.106^{+2}$	$.888^{+1}$	15
			$.475^{+2}$	$.108^{+2}$	$.516^{+1}$	$.391^{+1}$	30
			$.382^{+2}$	$.795^{+1}$	$.306^{+1}$	$.197^{+1}$	60
			$.343^{+2}$	$.669^{+1}$	$.212^{+1}$	$.111^{+1}$	120
			$.325^{+2}$	$.609^{+1}$	$.168^{+1}$	$.701^{0}$	240
3	1 1 3 4	1 P 4	$.371^{0}$	$.189^{-1}$	$.380^{-2}$	$.521^{-2}$	6
			$.376^{0}$	$.241^{-1}$	$.138^{-2}$	$.399^{-4}$	15
				$.242^{-1}$	$.150^{-2}$	$.859^{-4}$	30
4	1 1 3 4	4 G 4	$.308^{0}$	$.161^{0}$	$.151^{0}$	$.150^{0}$	6
			$.299^{0}$	$.293^{-1}$	$.118^{-1}$	$.108^{-1}$	15
				$.223^{-1}$	$.224^{-2}$	$.986^{-3}$	30
				$.229^{-1}$	$.149^{-2}$	$.154^{-3}$	60
					$.147^{-2}$	$.959^{-4}$	120
						$.932^{-4}$	240
5	2 2 1 4	4 G 4	$.120^{+3}$	$.158^{0}$	$.151^{0}$	$.150^{0}$	6
			$.212^{+1}$	$.245^{-1}$	$.135^{-1}$	$.109^{-1}$	15
			$.243^{+1}$	$.167^{-1}$	$.411^{-2}$	$.112^{-2}$	30
				$.169^{-1}$	$.348^{-2}$	$.299^{-3}$	60
					$.353^{-2}$	$.245^{-3}$	120
						$.245^{-3}$	240

TABLE 7.1 (Continued)

#	r s m o	ρ c o	J = 5	J = 10	J = 20	J = 40	N
6	1 1 3 4	2 S 4	$.298^{0}$	$.527^{-2}$	$.136^{-1}$	$.148^{-1}$	6
			$.351^{0}$	$.222^{-1}$	$.978^{-3}$	$.343^{-3}$	15
				$.234^{-1}$	$.143^{-2}$	$.631^{-4}$	30
					$.148^{-2}$	$.909^{-4}$	60
7	1 1 3 4	1 P 6	$.377^{0}$	$.243^{-1}$	$.152^{-2}$	$.108^{-3}$	6
			$.377^{0}$	$.242^{-1}$	$.151^{-2}$	$.947^{-4}$	15
						$.944^{-4}$	30
8	1 1 3 4	6 G 6	$.351^{-1}$	$.536^{-2}$	$.342^{-2}$	$.330^{-2}$	6
			$.238^{0}$	$.156^{-1}$	$.116^{-2}$	$.268^{-3}$	15
			$.307^{0}$	$.198^{-1}$	$.124^{-2}$	$.824^{-4}$	30
						$.858^{-4}$	60
9	1 1 3 4	4 S 6	$.186^{0}$	$.116^{-1}$	$.334^{-3}$	$.370^{-3}$	6
			$.300^{0}$	$.194^{-1}$	$.120^{-2}$	$.719^{-4}$	15
				$.218^{-1}$	$.136^{-2}$	$.848^{-4}$	30
10	1 1 3 4	1 P 8	$.377^{0}$	$.242^{-1}$	$.151^{-2}$	$.663^{-4}$	6
			$.377^{0}$	$.242^{-1}$	$.151^{-2}$	$.109^{-3}$	15
						$.101^{-3}$	30

formulas must be used. The approximation at $x = x_1$ 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 "ρ", the order of consistency "o", and a code "c", indicating the type of method used.

"C" is the trapezoidal rule. (See Example (I.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_ν^ℓ of these methods can be found in Gear (1971), p. 217. (See also, Example (I.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_ν^ℓ are determined from consistency and $A(\alpha)$-stability requirements.

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

$$y_t = \frac{\epsilon}{2} x u_{xx} - x(1-x) u_x + xu \ , \quad 0 \le x \le 1 \ , \ t \ge 0 \ ,$$

with initial condition

$$u(x,0) = \sin \pi x \ , \quad 0 \le x \le 1 \ ,$$

and boundary conditions

$$u(0,t) = u(1,t) = 0 \ .$$

The discretization in space corresponds to the second case discussed in Example (I.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 $z_{j,1} \equiv x_j + \frac{(h_{j+1}-h_j)}{3}$ 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

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

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

$$K_h \frac{(u_h^n - u_h^{n-1})}{\Delta t} = \frac{1}{2} L_h(u_h^n + u_h^{n-1}) ,$$

$$u_j^0 = \sin(\pi x_j).$$

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.

<u>Example 7.3.</u> 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 $y(x,0) = g(x)$ 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 λ_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 7.1a

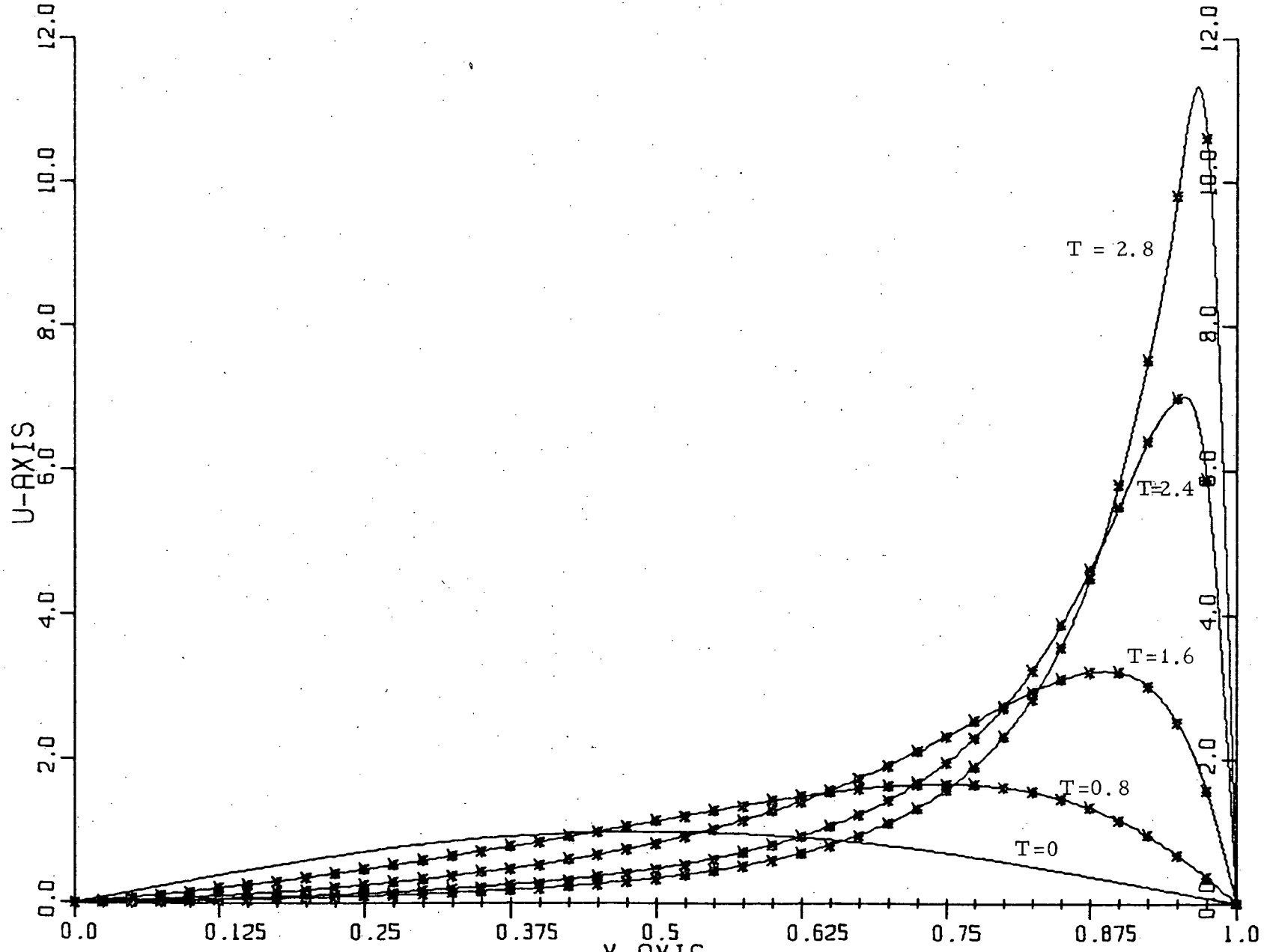

U-AXIS
X-AXIS
0.0
2.0
4.0
6.0
8.0
10.0
12.0
0.125
0.25
0.375
0.5
0.625
0.75
0.875
1.0
T = 2.8
T=2.4
T=1.6
T=0.8
T=0

Figure 7.1b

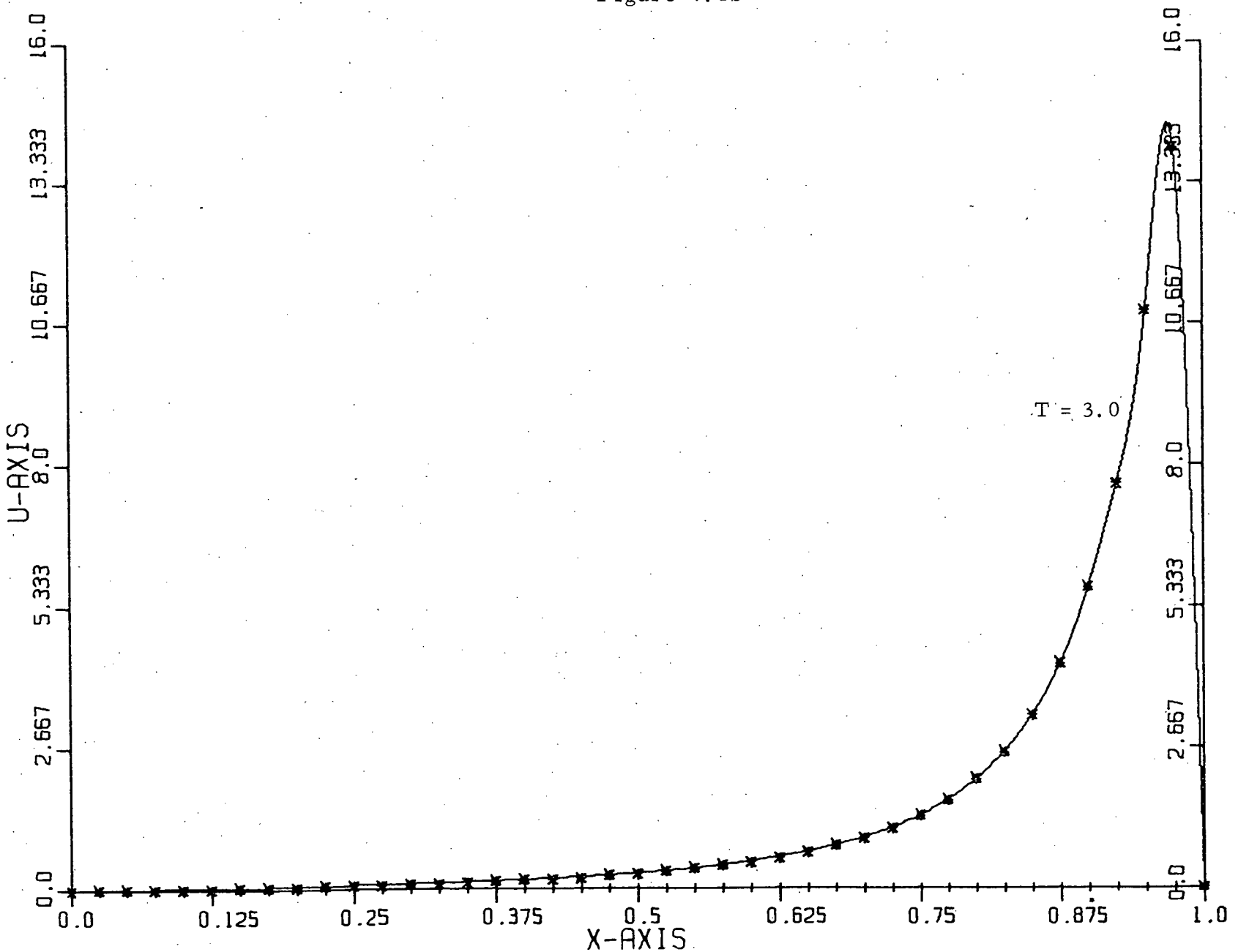

U-AXIS
X-AXIS
T = 3.0
0.0
2.667
5.333
8.0
10.667
13.333
16.0
0.125
0.25
0.375
0.5
0.625
0.75
0.875
1.0

Figure 7.1c

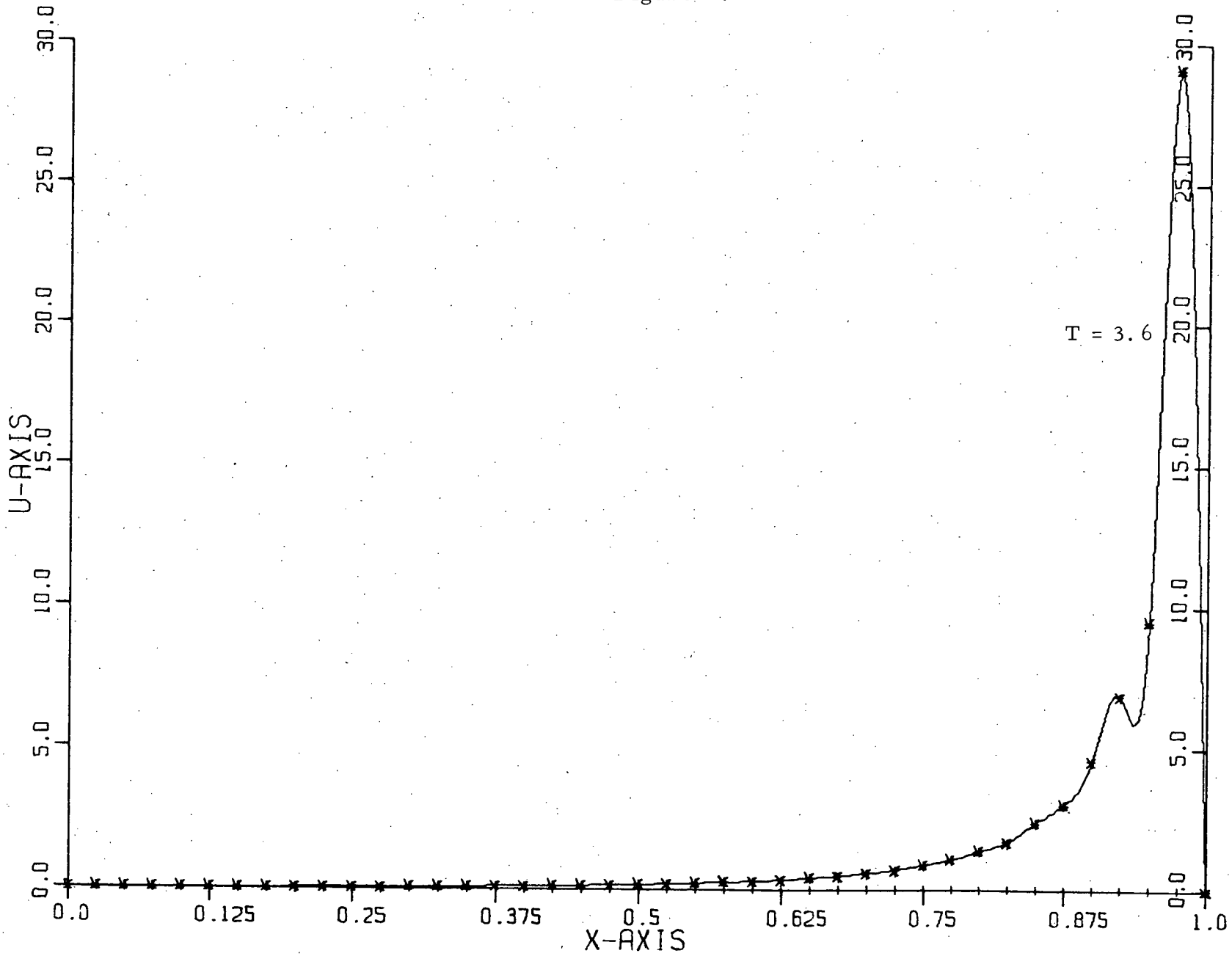

U-AXIS
X-AXIS
T = 3.6
0.0
5.0
10.0
15.0
20.0
25.0
30.0
0.125
0.25
0.375
0.5
0.625
0.75
0.875
1.0

Figure 7.1d

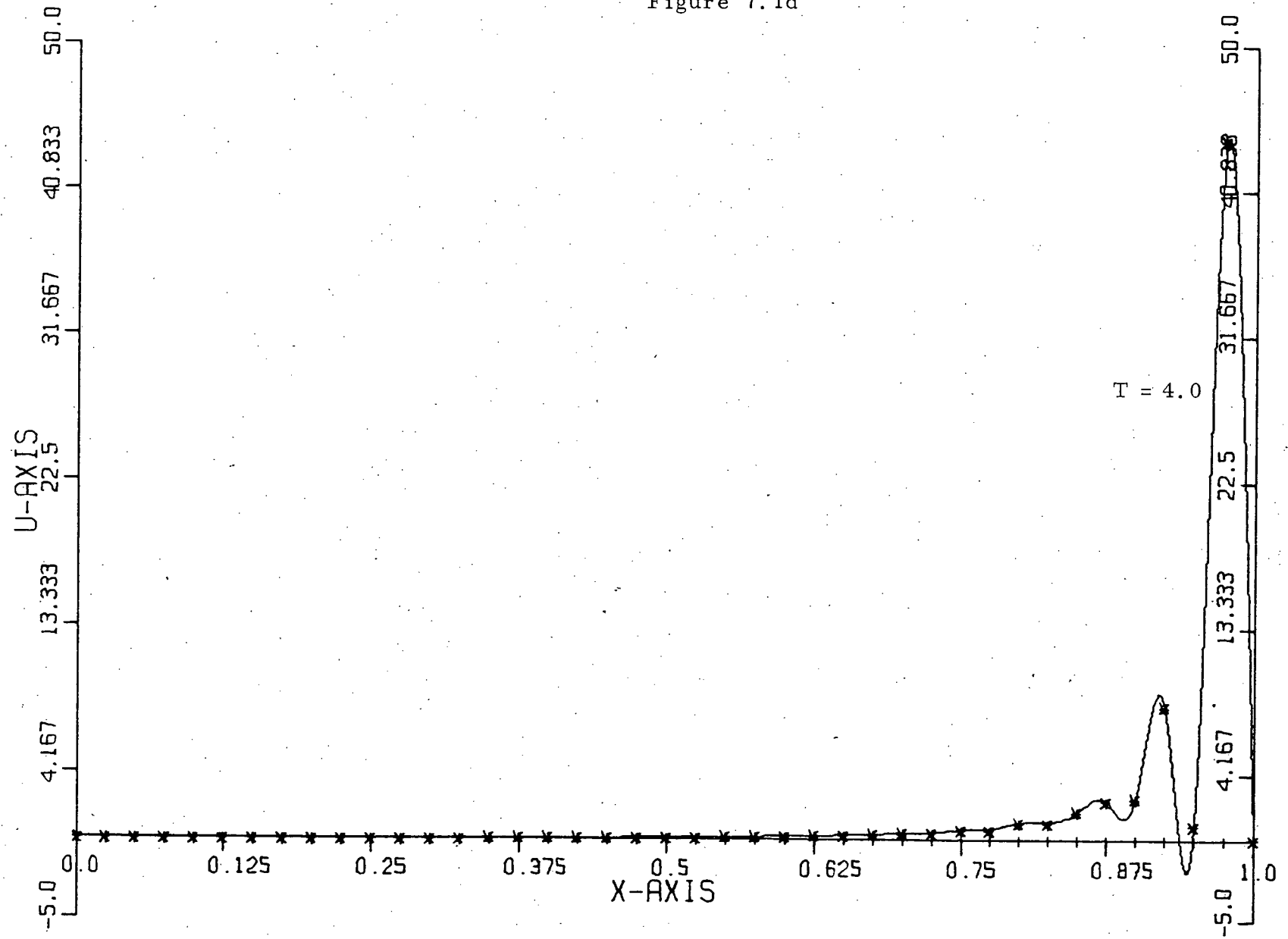

U-AXIS
50.0
40.833
31.667
22.5
13.333
4.167
-5.0
0.0
0.125
0.25
0.375
0.5
0.625
0.75
0.875
1.0
X-AXIS
T = 4.0

Figure 7.2a

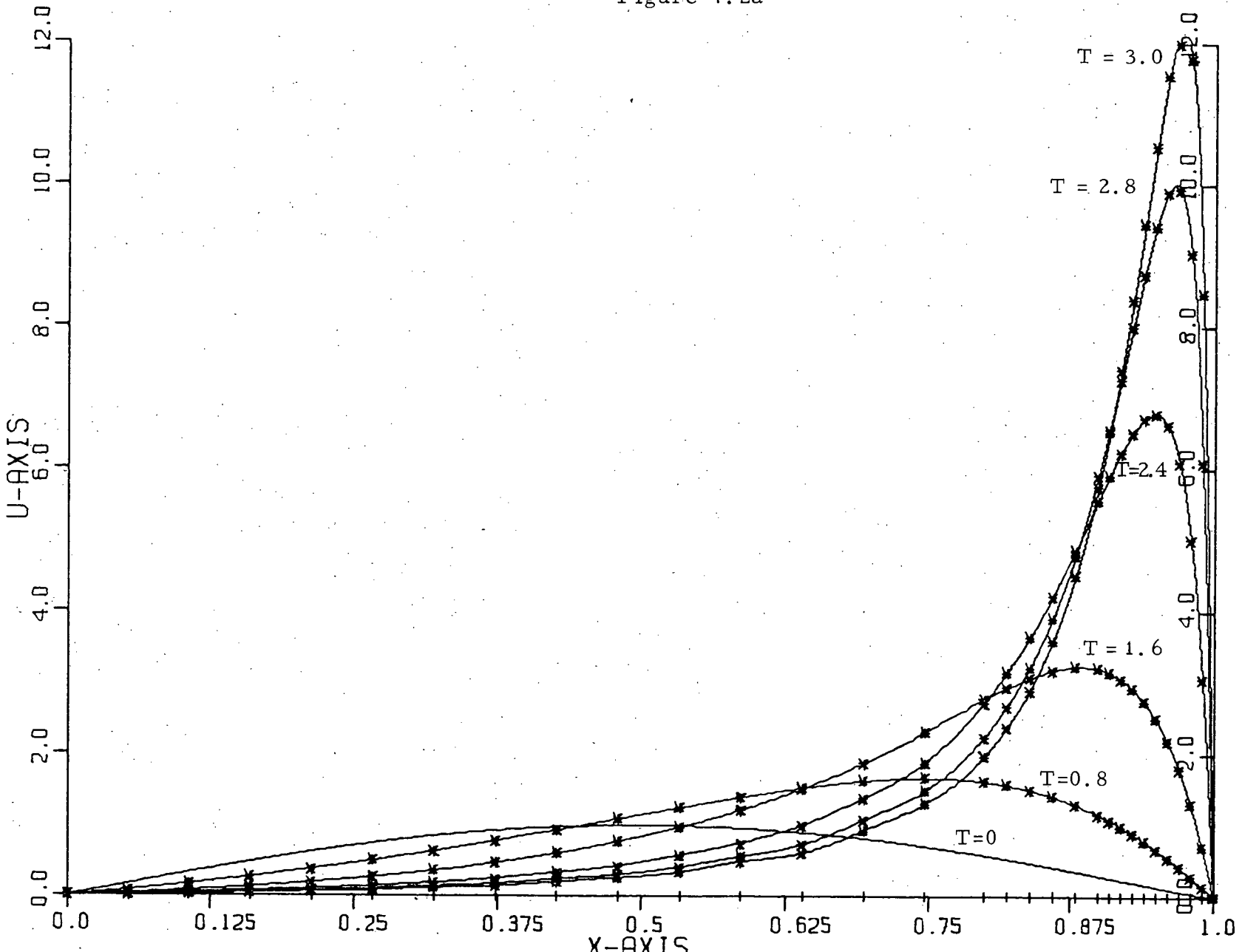

U-AXIS
X-AXIS
0.0
2.0
4.0
6.0
8.0
10.0
12.0
0.125
0.25
0.375
0.5
0.625
0.75
0.875
1.0
T = 3.0
T = 2.8
T=2.4
T = 1.6
T=0.8
T=0

Figure 7.2b

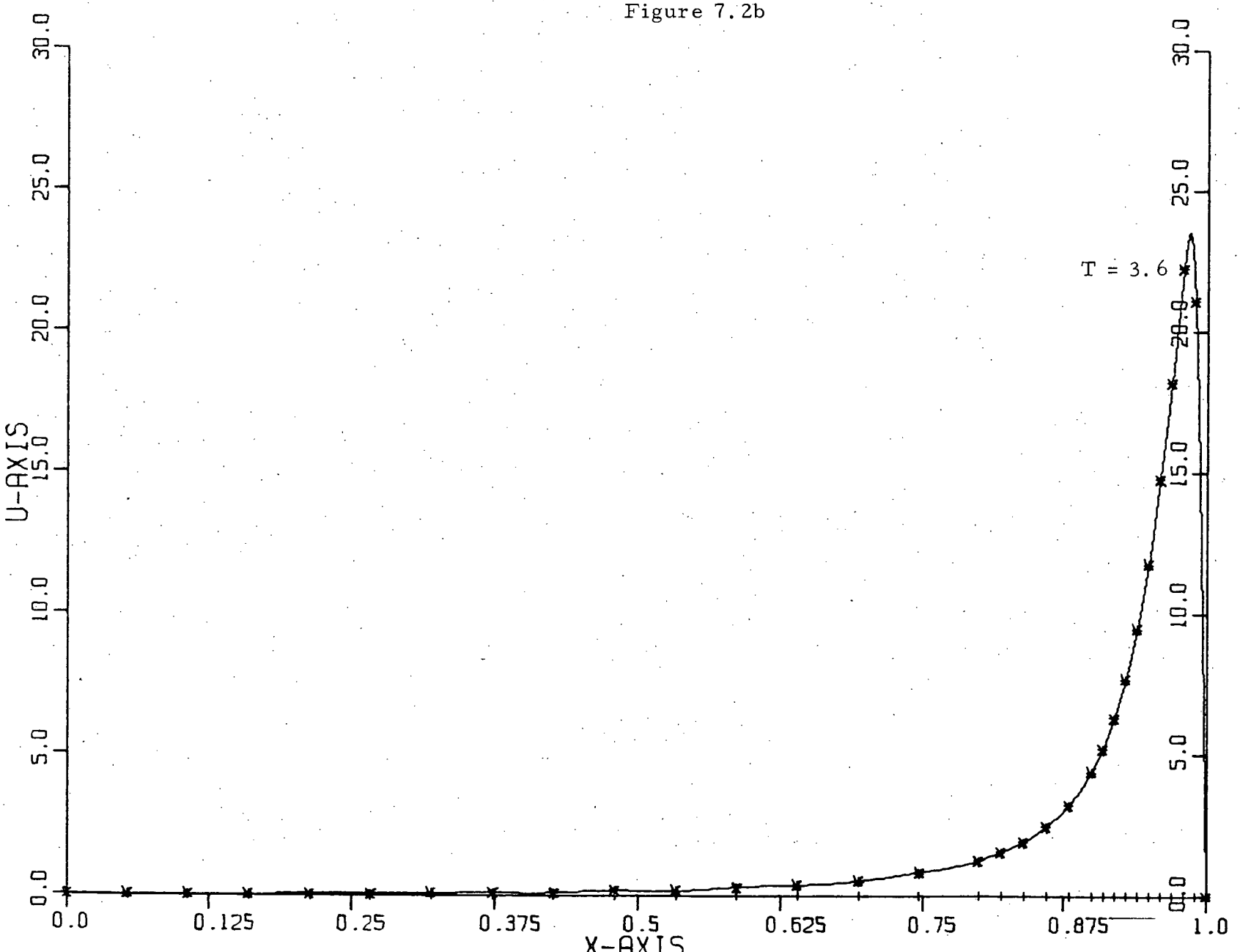

U-AXIS
X-AXIS
T = 3.6
0.0
5.0
10.0
15.0
20.0
25.0
30.0
0.125
0.25
0.375
0.5
0.625
0.75
0.875
1.0

Figure 7.2c

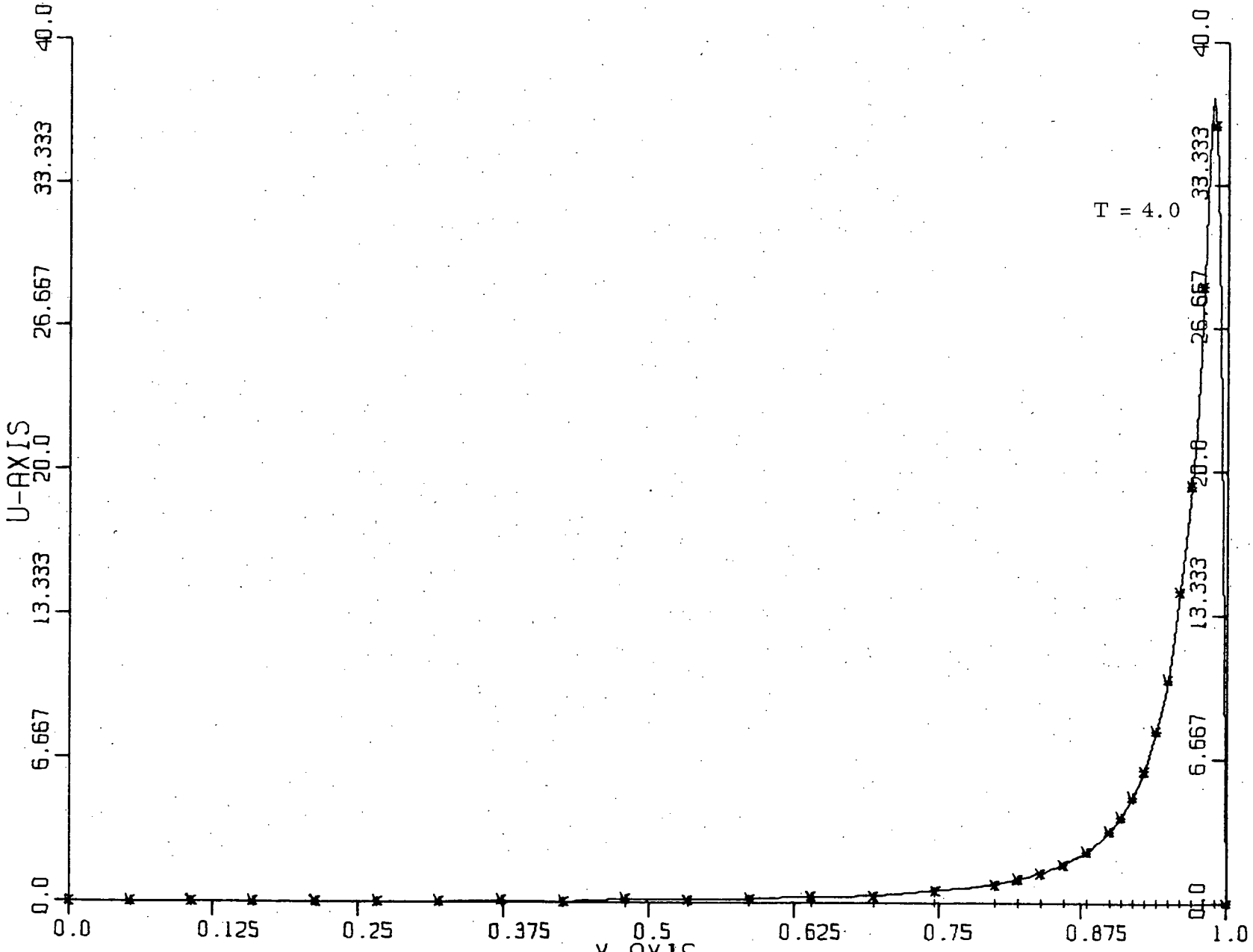

U-AXIS
0.0
6.667
13.333
20.0
26.667
33.333
40.0
T = 4.0
0.0
0.125
0.25
0.375
0.5
0.625
0.75
0.875
1.0

Figure 7.2d

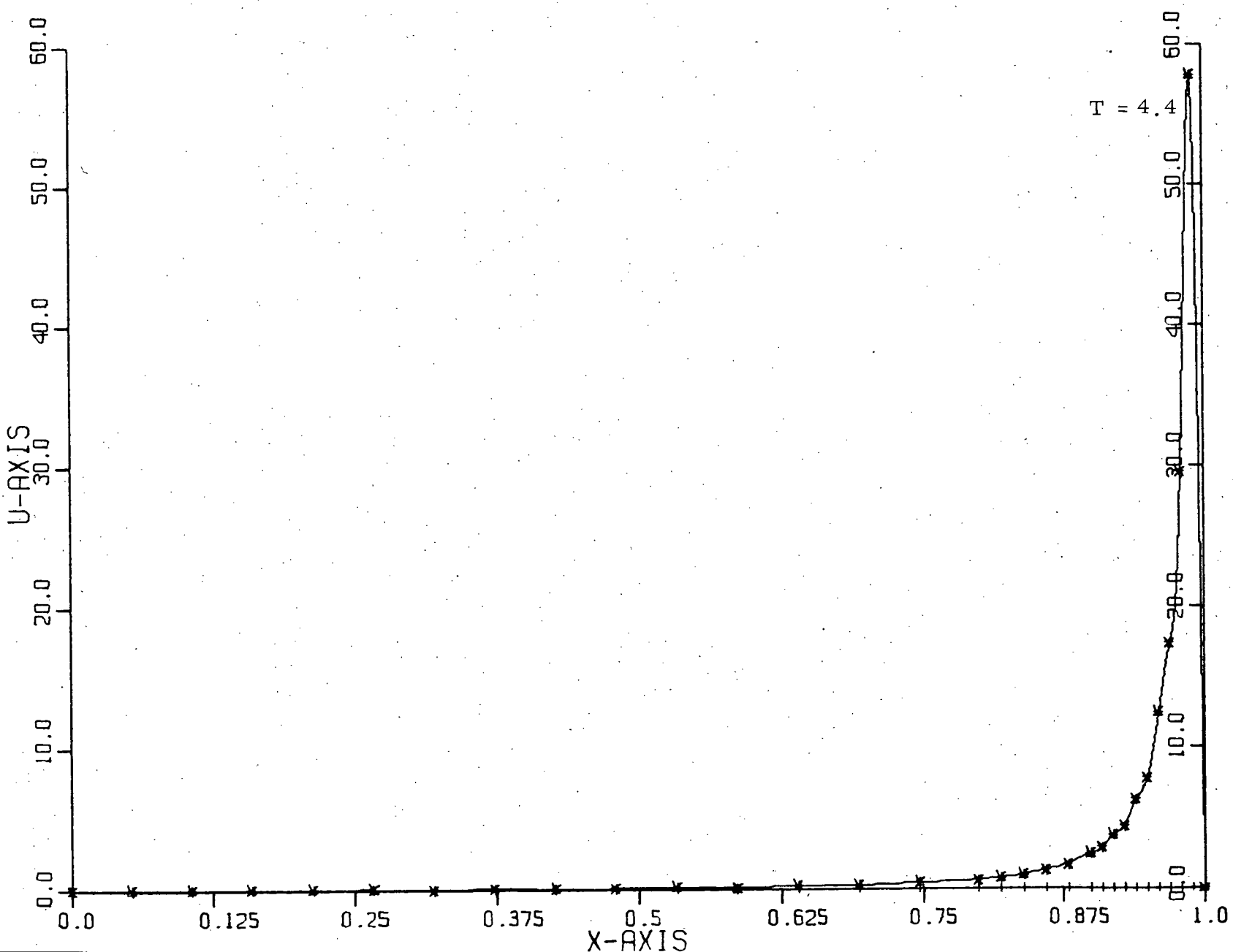

T = 4.4
U-AXIS
X-AXIS
0.0
10.0
20.0
30.0
40.0
50.0
60.0
0.125
0.25
0.375
0.5
0.625
0.75
0.875
1.0

Figure 7.2e

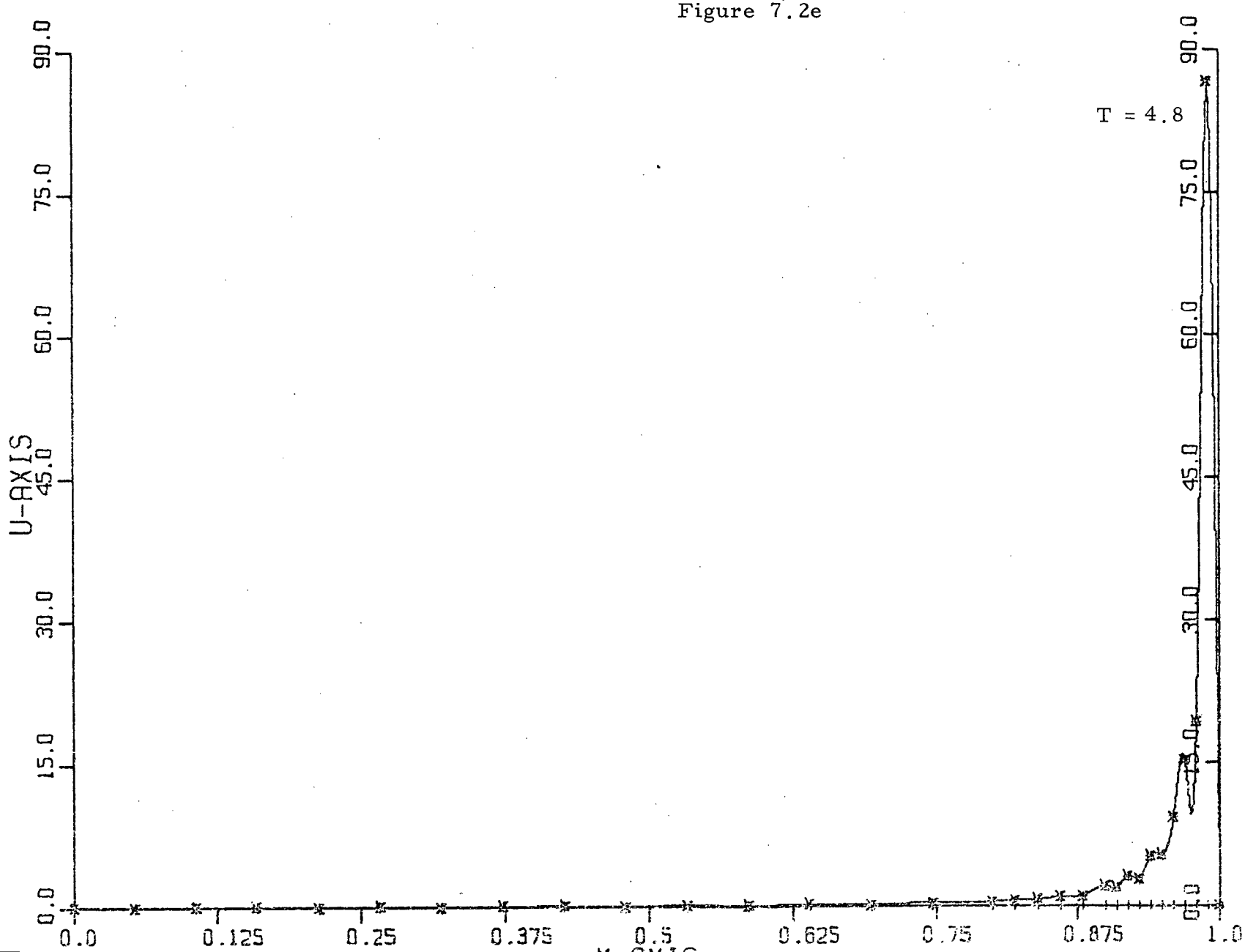

T = 4.8
U-AXIS
0.0
15.0
30.0
45.0
60.0
75.0
90.0
0.0
0.125
0.25
0.375
0.5
0.625
0.75
0.875
1.0

TABLE 7.2

	$j = 1$				$j = 2$				$3 \leq j \leq J-3$			
#	r	s	m	o	r	s	m	o	r	s	m	o
1	1	2	1	2	2	2	1	4	2	2	1	4
2	1	4	1	4	2	2	1	4	2	2	1	4
3	1	4	1	4	2	3	1	4	3	3	1	6
4	1	3	4	6	2	2	5	8	2	2	5	8

in space of the form (6.2) is given, for which the eigenvalues λ_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 $\kappa(V_h)$ 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.

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 λ_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:

$$h = \frac{1}{8} \quad : \quad \lambda_1 = -14.9 \pm 7.18\, i\,, \qquad \lambda_2 = -20.5 \pm 6.50\, i\,,$$

$$h = \frac{1}{16} \quad : \quad \lambda_1 = -66.1 \pm 24.8\, i\,, \qquad \lambda_2 = -67.8 \pm 25.1\, i\,,$$

$$h = \frac{1}{24} \quad : \quad \lambda_1 = -150.5 \pm 55.8\, i\,, \qquad \lambda_2 = -150.8 \pm 56.0\, i\,.$$

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

$$\lambda_h \cong \frac{1}{h^2}\,(-0.26 \pm 0.096\, i\,)$$

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 $\alpha > 2.12°$. There are strongly $A(\alpha)$-stable methods that do not satisfy this requirement. (See Varah (1975), p. 19).

<u>Example 7.5</u>. 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

		j = 1				j = 2				$3 \leq j \leq J-3$			
#	s_0	r	s	m	o	r	s	m	o	r	s	m	o
1	4	1	2	1	2	2	2	1	4	2	2	1	4
2	4	1	4	1	4	2	2	1	4	2	2	1	4
3	6	1	4	1	4	2	3	1	4	3	3	1	6

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 $\lambda_1 \cong \frac{1}{h^2}(-0.178 \pm 0.049\, i)$ and $\lambda_2 \cong \frac{1}{h^2}(-0.261 \pm 0.097\, i)$. In experiment 5 there is only one conjugate pair given by $\lambda_1 \cong \frac{1}{h^2}(-0.246 \pm 0.044\, i)$. The multistep differencing in time should therefore be strongly $A(\alpha)$-stable with $\alpha \geq \alpha_{min}$. The value α_{min} as computed from the complex eigenvalues above is also given in the table.

TABLE 7.4

		j = 1				J = 2				$3 \le j \le J-3$				
#	s_0	r	s	m	o	r	s	m	o	r	s	m	o	α_{min}
4	6	1	6	1	6	2	5	1	6	3	3	1	6	21°
5	8	1	3	4	6	2	2	5	8	2	2	5	8	11°

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

$$y_t(x,t) = \left(\frac{2}{3\pi}\right)^2 y_{xx}(x,t) \ , \quad 0 \le x \le 1 \ , \quad t \ge 0 \ ,$$

with initial condition

$$y(x,0) = \cos\left(\frac{3\pi}{2}x\right) \ , \quad 0 \le x \le 1 \ ,$$

and boundary conditions

$$y_x(0,t) = y(1,t) = 0 \ , \quad t \ge 0 \ .$$

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

$$K_h w'(x_j,t) = L_h w(x_j,t) \quad , \quad 1 \le j \le J-1 \ ,$$

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

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 α the eigenvalue problem $K_h v_h = \lambda_h L_h v_h$ has a positive eigenvalue that is asymptotically given by

$$\lambda_h \cong \frac{12(z_1 - 1)^2}{h^2(z_1^2 + 10z_1 + 1)}$$

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 $[0, 10\frac{2}{3}]$ 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).

Figure 7.3a

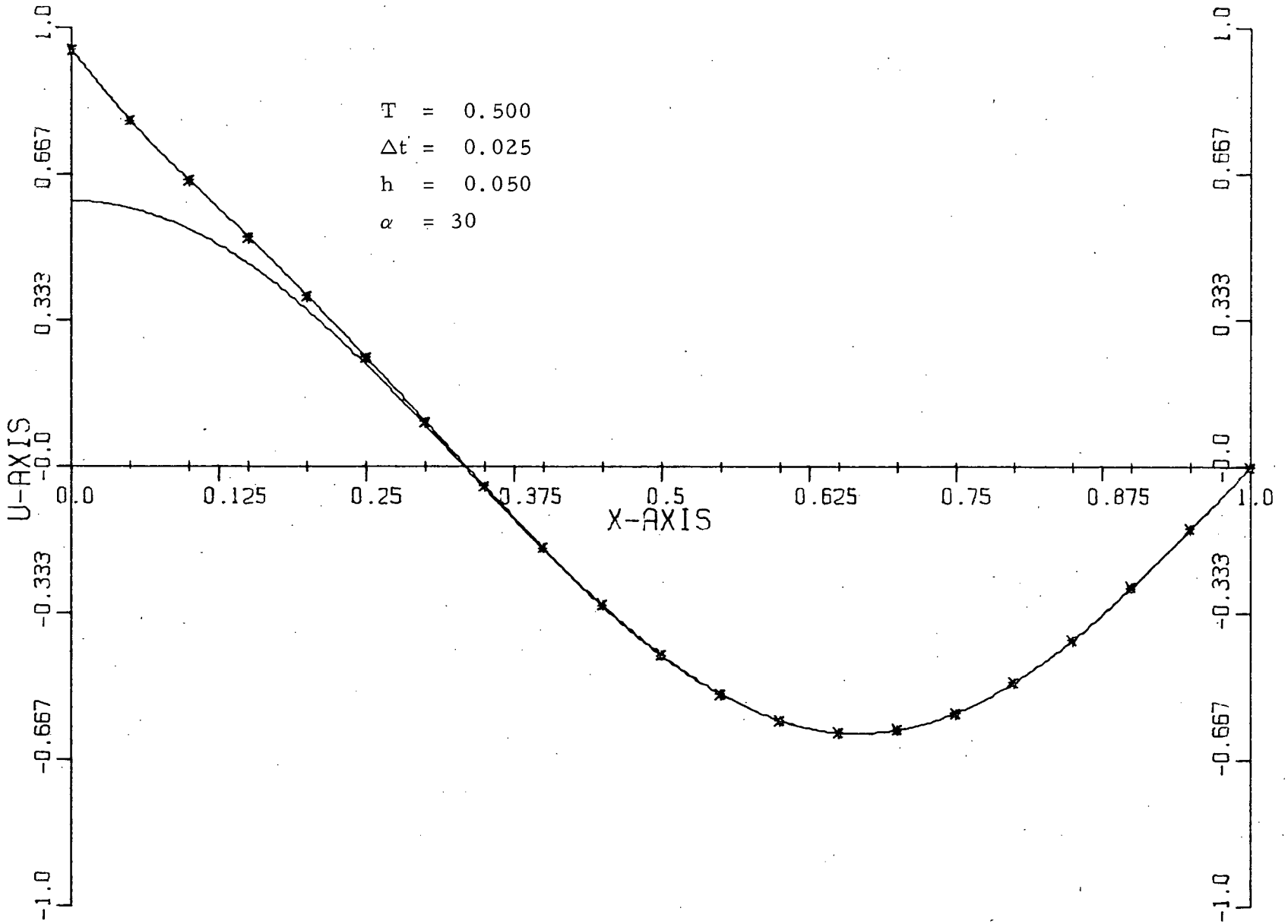

T = 0.500
Δt = 0.025
h = 0.050
α = 30
U-AXIS
X-AXIS
1.0
0.667
0.333
-0.0
-0.333
-0.667
-1.0
0.0
0.125
0.25
0.375
0.5
0.625
0.75
0.875
1.0

Figure 7.3b

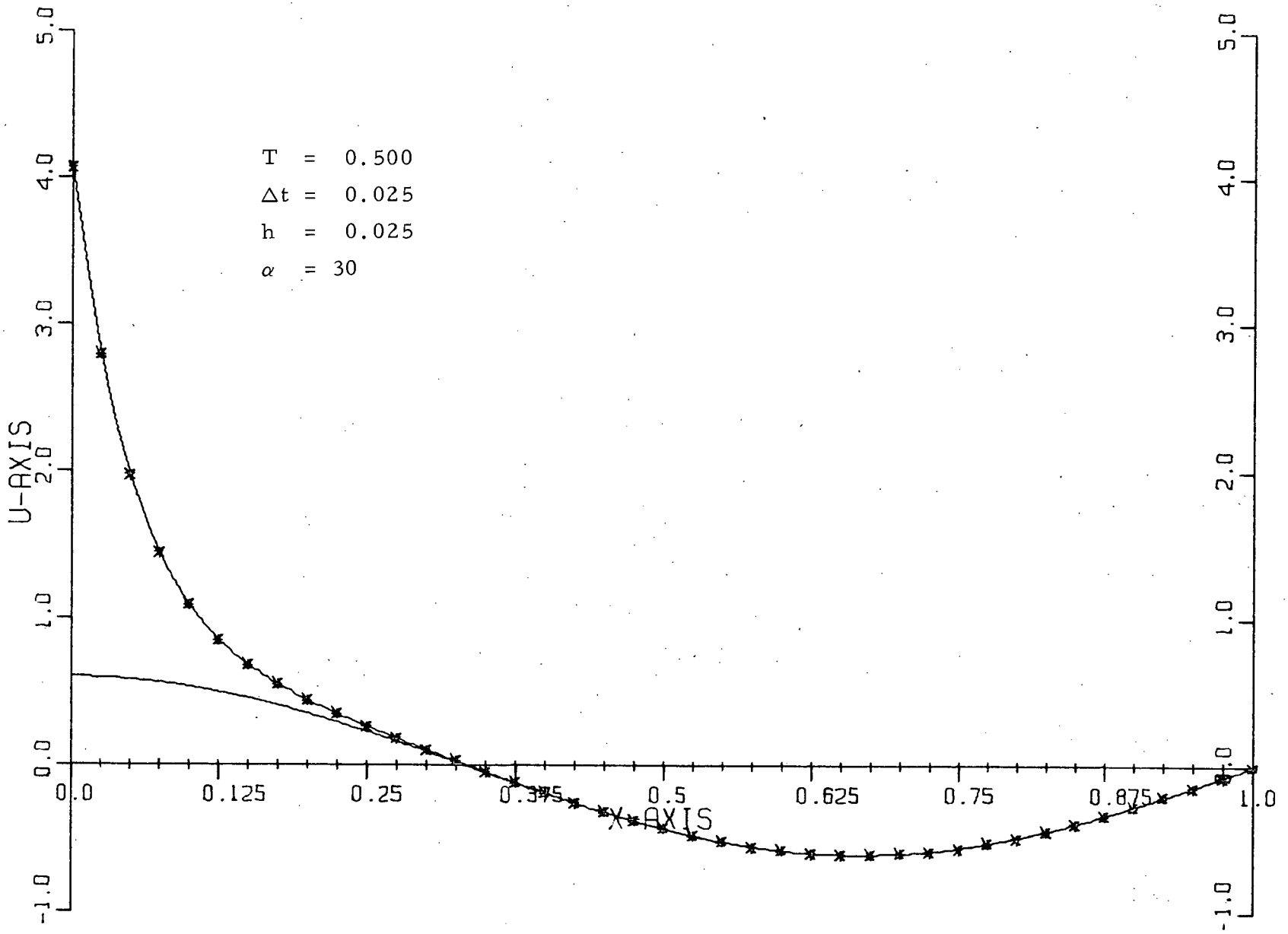

T = 0.500
Δt = 0.025
h = 0.025
α = 30
U-AXIS
X-AXIS
5.0
4.0
3.0
2.0
1.0
0.0
-1.0
0.0
0.125
0.25
0.375
0.5
0.625
0.75
0.875
1.0

Figure 7.4a

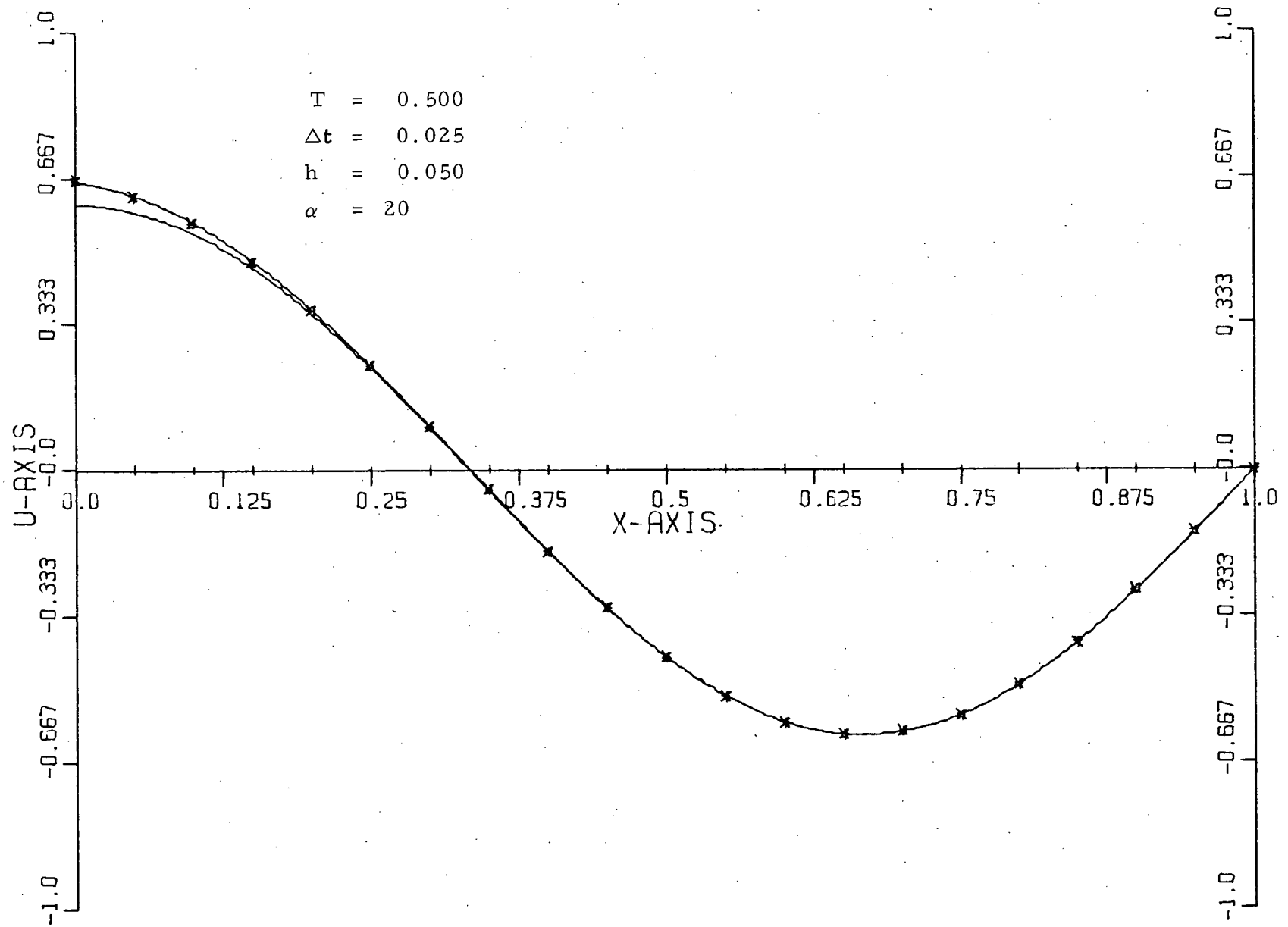
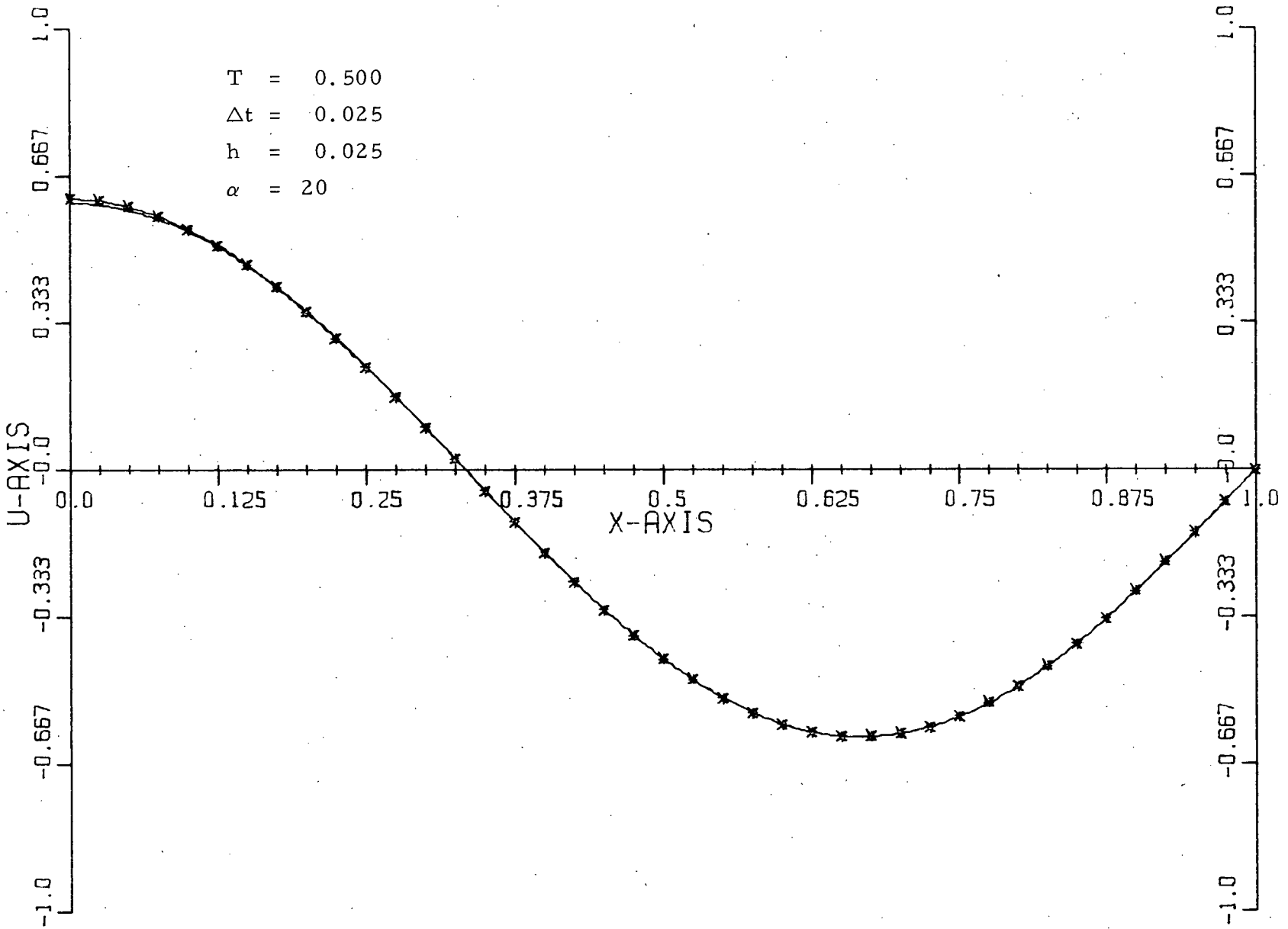

T = 0.500
Δt = 0.025
h = 0.050
α = 20
U-AXIS
X-AXIS
1.0
0.667
0.333
-0.0
-0.333
-0.667
-1.0
0.0
0.125
0.25
0.375
0.5
0.625
0.75
0.875
1.0

Figure 7.4b


T = 0.500
Δt = 0.025
h = 0.025
α = 20
U-AXIS
X-AXIS
1.0
0.667
0.333
-0.0
-0.333
-0.667
-1.0
0.0
0.125
0.25
0.375
0.5
0.625
0.75
0.875
1.0

BIBLIOGRAPHY

G. Birkhoff and S. Gulati, Optimal few-point discretizations of linear source problems, SIAM J. Numer. Anal. 11, 1974, pp. 700-728.

C. de Boor and B. Swartz, Collocation at Gaussian points, SIAM J. Numer. Anal. 10, 1973, pp. 582-606.

J. H. Bramble and B. E. Hubbard, On a finite difference analogue of an elliptic boundary value problem which is neither diagonally dominant nor of the nonnegative type, J. Math. and Phys. 43, 1964, pp. 117-132.

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

P. G. Ciarlet, M. H. Schultz and R. S. Varga, Numerical methods of high order accuracy for nonlinear boundary value problems, Numer. Math. 9, 1967, pp. 394-430.

L. Collatz, Numerische Behandlung von Differential-gleichungen, Springer-Verlag, Berlin, 1960.

G. Dahlquist, Stability and error bounds in the numerical integration of ordinary differential equations, Kungl. Tekn. Högsk. Stockholm, no. 130, 1959, 87 pp.

G. Dahlquist, A special stability problem for linear multistep methods, BIT, 3, 1963, pp. 27-43.

W. H. Enright, Second derivative multistep methods for stiff ordinary differential equations, SIAM J. Numer. Anal. 11, 1974, pp. 321-331.

C. W. Gear, Numerical initial value problems in ordinary differential equations, Prentice-Hall, 1971.

R. D. Grigorieff, Die Konvergenz des Rand-und Eigenwert problems linearer gewöhnlicher Differenzengleichungen, Numer. Math., 1970, pp. 15-48.

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.

B. Hakberg, Uniformly maximumnorm stable difference schemes, B.I.T. 10, 1970, pp. 266-276.

B. L. Hulme, Discrete Galerkin and related one step methods for ordinary differential equations, Research report, Sandia Laboratories, Albuquerque, N.M., 1971.

H. B. Keller, Numerical methods for two-point boundary value problems, Blaisdell, London, 1968.

H. B. Keller, Accurate difference methods for linear ordinary differential systems subject to linear constraints, SIAM J. Numer. Anal., 6, 1969, pp. 8-30.

H. B. Keller, Numerical solution of boundary value problems for ordinary differential equations, A. K. Aziz, ed., Academic Press, 1975, pp. 27-88.

H. O. Kreiss, Difference approximations for boundary and eigenvalue problems for ordinary differential equations, Math. of Comp. 26, 1972, pp. 605-624.

P. Lancaster, Theory of matrices, Academic Press, 1969.

S. Nordmark, Uniform stability of a class of parabolic difference operators, B.I.T. 14, 1974, pp. 314-325.

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.

H. Padé, Sur la représentation approchée d'une fonction par des fractions rationelles, thesis, Ann. de L'Éc. Nor., (3), 9, 1892.

I. G. Petrovskii, Über das Cauchysche Problem für ein System linearer partieller Differentialgleichungen im Gebiete der Nichtanalytischen Functionen, Bull. Univ. d'État Moscou, 1A, No. 7, 1938, pp. 1-74.

M. Reimer, Finite difference forms containing derivatives of higher order, SIAM J. Numer. Anal. 5, 1968, pp. 725-738.

R. D. Richtmeyer and K. W. Morton, Difference methods for initial value problems, Interscience Publishers, 1967.

R. D. Russell and L. F. Shampine, A collocation method for boundary value problems, Numer. Math. 19, 1972, pp. 1-28.

R. D. Russell and J. M. Varah, A comparison of global methods for linear two-point boundary value problems, Math. of Comp. 29, 1975, pp. 1-13.

M. H. Schultz, Spline Analysis, Prentice-Hall, 1973.

J. N. Shoosmith, A higher order finite difference method for the solution of two-point boundary value problems on a uniform mesh, Numerical solution of boundary value problems for ordinary differential equations, A. K. Aziz, ed., Academic Press, 1975, pp. 355-369.

G. Strang and G. J. Fix, An analysis of the finite element method, Prentice-Hall, 1973.

B. K. Swartz, The construction and comparison of finite difference analogs of some finite element schemes, report, Los Alamos Scientific Laboratory, Los Alamos, N. M., 1974.

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.

J. M. Varah, Stability of high order accurate difference methods for parabolic equations with boundary conditions, SIAM J. Numer. Anal. 8, 1971, pp. 598-615.

J. M. Varah, Stability of higher order accurate difference methods for parabolic equations with boundary conditions, SIAM J. Numer. Anal. 8, 1971a, pp. 569-574.

J. M. Varah, Stiffly stable linear multistep methods of extended order, Technical Report, Dept. of Comp. Sc., University of British Columbia, 1975.

R. S. Varga, On higher order stable implicit methods for solving parabolic partial differential equations, J. Math and Phys., 40, 1961, pp. 220-231.

R. S. Varga, Hermite interpolation type Ritz methods for two-point boundary value problems, Numerical solution of partial differential equations, J. H. Bramble, ed., Academic Press, 1966, pp. 365-373.

R. Weiss, The application of implicit Runge-Kutta and collocation methods to boundary value problems, Math. of Comp. 28, 1974, pp. 449-464.

O. B. Widlund, On the stability of parabolic difference schemes, Math. of Comp. 19, 1965, pp. 1-13.

O. B. Widlund, Stability of parabolic difference schemes in the maximum norm, Numer. Math. 8, 1966, pp. 186-202.

K. Wright, Some relationships between implicit Runge-Kutta, collocation and Lanczos τ methods, and their stability properties, B.I.T. 10, 1970, pp. 217-227.