NUMERICAL SOLUTION OF TWO-POINT BOUNDARY-VALUE PROBLEMS

Thesis By

Andrew Benjamin White, Jr.

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

California Institute of Technology
Pasadena, California 91109
1974
(Submitted March 13, 1974)

Acknowledgements

The author wishes to thank Professor Herbert B. Keller for his patient guidence and invaluable assistance in the preparation of this thesis. His encouragement and enthusiasm were essential to the completion of this work.

The opportunity to carry out this research was provided by California Institute of Technology Graduate Teaching Assistantships for which the author is grateful. He also wishes to express his appreciation to the entire Department of Applied Mathematics, both faculty and students, for providing an invigorating climate in which to pursue this work.

The author wishes to thank Mrs. Janet Smith and Mrs. Mary Beth Briggs for their expert typing of a detailed and tedious manuscript.

ABSTRACT

The approximation of two-point boundary-value problems by general finite difference schemes is treated. A necessary and sufficient condition for the stability of the linear discrete boundary-value problem is derived in terms of the associated discrete initial-value problem. Parallel shooting methods are shown to be equivalent to the discrete boundary-value problem. One-step difference schemes are considered in detail and a class of computationally efficient schemes of arbitrarily high order of accuracy is exhibited. Sufficient conditions are found to insure the convergence of discrete finite difference approximations to nonlinear boundary-value problems with isolated solutions. Newton's method is considered as a procedure for solving the resulting nonlinear algebraic equations. A new, efficient factorization scheme for block tridiagonal matrices is derived. The theory developed is applied to the numerical solution of plane Couette flow.

Table of Contents

Introduction

This thesis deals with the application of finite difference schemes to two-point boundary-value problems. The assumption is made throughout that these boundary-value problems have isolated solutions; that is, the homogeneous, linearized problem has only the trivial solution. The general theory developed places no restrictions on the form of the difference equations.

In Chapter 1, the application of an arbitrary, consistent difference scheme to a linear boundary-value problem is treated. In the main theorem of this chapter, Theorem 1.16, the stability of the discrete boundary-value problem is shown to be equivalent to the stability of the associated discrete initial-value problem. This associated initial-value problem employs the same difference equations to approximate the differential equation, but initial conditions replace the boundary conditions. From this result, it is clear that a simple shooting method is, in fact, a specific procedure for solving the discrete boundary-value problem.

Special emphasis is placed on one-step difference schemes in Chapter 2. High order accurate difference approximations are developed using both Taylor series and the integral form of the differential equation. In particular, one-step schemes of arbitrary order are derived which require the evaluation of a minimum number of new functions (e.g derivatives of $A(t)$, $f(t)$). The equivalence shown in Chapter 1 is used to examine the stability of triangular difference schemes.

In Chapter 3, the equivalence result of Theorem 1.16 is generalized to include all parallel shooting methods. Theorem 3.22 shows that these methods are each a particular procedure for solving the equations derived from approximating linear boundary-value problems. The Method of Complementary Functions is examined in detail as an example of methods for solving problems with separated boundary conditions. The Method of Adjoints is also considered and it is shown that this method is not in general equivalent to the discrete boundary-value problem.

Nonlinear boundary-value problems are dealt with in Chapter 4. The difference schemes examined in Chapter 2 are generalized to be applicable to nonlinear differential equations. Following Keller [6], existence and uniqueness of these discrete approximations is shown. We note that Newton's method converges quadratically.

Chapter 5 is concerned with the practical problem of solving the systems of algebraic equations arising from the approximation of boundary-value problems with separated boundary conditions. These equations are written in block tridiagonal form, $Mx = b$. The special zero structure of this system is exploited to show that, with an appropriate row switching strategy, such a matrix possesses a simple block LU decomposition if and only if M is nonsingular.

A numerical example is presented in Chapter 6. The equations considered model plane Couette flow. The Gap_4 scheme, as derived in Chapter 4, is used to discretize the nonlinear boundary-value problem and Newton's method is employed to solve the resulting set of nonlinear equations.

A consistent effort is made to use o to represent zero or a zero vector and 0 for zero matrices, except in tables or equation numbers. The numbering of theorems, equations, or tables is done consecutively throughout each chapter.

Chapter I

Linear Two-Point Boundary-Value Problems

1. Existence Theory

We consider the system of n first-order, linear ordinary differential equations:

$$\begin{aligned} &u'(t) - A(t)\, u(t) - f(t) = o \qquad t \in [o,1] \qquad &\text{a)} \\ &B_o\, u(o) + B_1\, u(1) - \beta = o \qquad &\text{b)} \end{aligned} \qquad (1.1)$$

where u, f, β are n-vectors and A, B_o, B_1 are n × n matrices. Before proceeding to the numerical approximation of (1.1a,b), we present an existence and uniqueness result convenient for our purposes.

Theorem 1.2. Let $A(t) \in C^m[o,1]$ for some $m \geq o$. Define the fundamental matrix X(t) as the solution of

$$X'(t) - A(t)\, X(t) = o \quad X(0) = I. \qquad (1.3)$$

Then for each $f(t) \in C^m[o,1]$ and $\beta \in E^n$, problem (1.1a,b) has a unique solution $y(t) \in C^{m+1}[o,1]$ iff $[B_o+B_1X(1)]$ is nonsingular.

Proof: The solution to the initial-value problem

$$u'(t) - A(t)\, u(t) - f(t) = o \quad u(o) = r$$

is in general

$$y(t) = X(t)r + X(t) \int_o^t X^{-1}(s)\ f(s)\ ds. \qquad (1.4)$$

Uniqueness for the initial-value problem insures that X(t) is nonsingular on [o,1]. The boundary-value problem (1.1a,b) has a solution if and only if we can define an n-vector r such that y(t) satisfies the boundary condition

$$B_o\ u(o) + B_1\ u(1) - \beta = o.$$

This requires

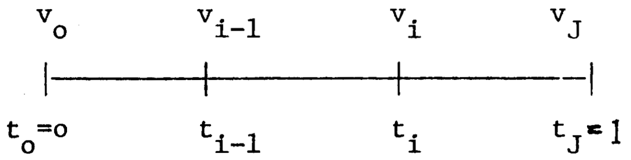
$$[B_o + B_1\ X(1)]r - \beta + B_1\ X(1) \int_o^1 X^{-1}(s)\ f(s)\ ds = o.$$

Thus, (1.1a,b) has a unique solution if and only if $[B_o + B_1\ X(1)]$ is nonsingular. That $y(t) \in C^{m+1}[o,1]$ is an observation from the form of differential equation (1.1a).

▨

2. Numerical Methods

Here we discuss some standard concepts of numerical analysis and develop some notation. In approximating the solution of (1.1a,b), we will employ a net of points $\{t_i\}_{i=o}^{i=J}$ on [o,1] and a net function $\{v_i\}_{i=o}^{i=J}$ defined on this net.



$t_i > t_{i-1}$

Each v_i is an n-vector and we define V such that

$$V = \begin{bmatrix} v_o \\ v_1 \\ \vdots \\ v_J \end{bmatrix}$$

where V is an n(J+1)-vector. We define the mesh widths $h_i = t_i - t_{i-1}$, i = 1,...,J, and $h_o = \max_{1 \le i \le J} h_i$. We further require that for each h_i there exists a $\lambda_i \in [\varepsilon,1]$, $\varepsilon > o$, such that

$$h_i = \lambda_i h_o.$$

This condition merely stipulates that

$$\varepsilon \le \min h_i/\max h_i \le 1$$

and thus the mesh becomes dense in [o,1] as $h_o \to o$.

The norm we will employ is $||V|| \equiv \max_{o \le i \le J} ||v_i||$ and for block matrices, the usual induced norm

$$||M|| = \max_{||V|| = 1} ||MV||.$$

In the n=1 case, this induced norm equals the maximum absolute row sum, however this result does not generalize to n > 1. We may easily produce the upper bound

$$||M|| \le \max_{o \le i \le J} \{\sum_{j=o}^{J} ||M_{ij}||\}$$

where M_{ij} is an n × n matrix element of the block matrix M.

An approximation to the differential equation (1.1a) may be written as

$$L_h\, v_i - r_i \equiv \sum_{j=o}^{J} M_{ij}\, v_j - r_i \qquad i = 1,\ldots,J. \tag{1.5a}$$

The boundary conditions (1.1b) are approximated by

$$B_o\, v_o + B_1\, v_J - \beta = o\ . \tag{1.5b}$$

We define the truncation error τ_i to be

$$\tau_i[y(t)] \equiv L_h\, y(t_i) - r_i \qquad i = 1,\ldots,J$$

where y(t) is any solution of the differential equation (1.1a). Similarly, we define a truncation error τ_o associated with the boundary conditions

$$\tau_o[\bar{y}(t)] \equiv B_o\bar{y}(o) + B_1\bar{y}(1) - \beta$$

where $\bar{y}(t)$ is any solution of (1.1b). Note that τ_o will always be zero. In considering the accuracy of approximations (1.5a), (1.5b), we are concerned with y(t) a solution of (1.1a,b) and we define

$$\tau[y(t)] \equiv \begin{bmatrix} \tau_o[y(t)] \\ \tau_1[y(t)] \\ \vdots \\ \tau_J[y(t)] \end{bmatrix} .$$

Combining the discrete approximations (1.5a) and (1.5b), the discrete boundary-value problem may be written as

$$\begin{bmatrix} B_o & 0 & \cdots & 0 & B_1 \\ M_{1o} & M_{11} & \cdots & M_{1J-1} & M_{1J} \\ \vdots & \vdots & & \vdots & \vdots \\ M_{Jo} & M_{J1} & \cdots & M_{JJ-1} & M_{JJ} \end{bmatrix} \begin{bmatrix} v_o \\ v_1 \\ \vdots \\ v_J \end{bmatrix} - \begin{bmatrix} \beta \\ r_1 \\ \vdots \\ r_J \end{bmatrix} = \begin{bmatrix} o \\ o \\ \vdots \\ o \end{bmatrix}. \qquad (1.7)$$

More briefly, we will write

$$\mathcal{B}_h V - r = o \qquad (1.8)$$

where $\mathcal{B}_h$ is an $n(J+1) \times n(J+1)$ matrix and r is an $n(J+1)$-vector with $r_o = \beta$. Employing Euler's method to approximate the differential equation (1.1a), this formulation of the discrete problem becomes

$$\begin{bmatrix} B_o & 0 & 0 & \cdots & B_1 \\ -\frac{1}{h_1} I - A(t_o) & \frac{1}{h_1} I & 0 & \cdots & 0 \\ 0 & -\frac{1}{h_2} I - A(t_1) & \frac{1}{h_2} I & \cdots & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & & -\frac{1}{h_J} I - A(t_{J-1}) & \frac{1}{h_J} I \end{bmatrix} \begin{bmatrix} v_o \\ v_1 \\ v_2 \\ \vdots \\ v_J \end{bmatrix} - \begin{bmatrix} \beta \\ f(t_o) \\ f(t_o) \\ \vdots \\ f(t_{J-1}) \end{bmatrix} = 0.$$

This example will be used throughout to illustrate various points of interest.

<u>Consistency</u>. The difference approximation (1.5a) is said to be consistent with the differential equation (1.1a) iff

$$||\tau[y(t)]|| \to o \text{ as } h_o \to o$$

where $y(t)$ is a solution of (1.1a,b). Euler's method provides an example. If $y(t) \in C^p[o,1]$, $p \geq 2$, then

$$\tau_i[y(t)] = h_i \sum_{K=2}^{P} h_i^{K-2} \frac{y^{(K)}(t_{i-1})}{K!} + O(h_i^p) .$$

Thus, Euler's method is consistent and we say that it is first-order accurate because the leading term in the truncation error expansion is linear in h_i.

Order of accuracy. A numerical scheme has an order of accuracy $p > o$ if p is the largest integer such that

$$||\tau[y(t)]|| \leq Kh_o^p$$

for all nets with $h_o \leq H$.

Stability. Let V be a solution to (1.8). The difference scheme (1.8) is said to be stable iff there exist constants $K_o \geq o$, and $H > o$, such that

$$||V|| \leq K_o \; ||r||$$

for all nets with $h_o \leq H$. For the linear case, this condition is equivalent to showing that a $K_o \geq o$ and $H > o$ exist such that

$$||B_h^{-1}|| \leq K_o$$

for all nets with $h_o \leq H$.

Convergence. The numerical scheme (1.8) is covergent iff

$$\max_{o \leq i \leq J} ||y(t_i) - v_i|| \to o \quad \text{as } h_o \to o \quad .$$

For convenience, we define the n(J+1)-vector Y to be

$$Y \equiv \begin{bmatrix} y(t_o) \\ y(t_1) \\ \vdots \\ y(t_J) \end{bmatrix} .$$

A standard result may now be stated.

<u>Lemma 1.10</u> Let the boundary-value problem (1.1a,b) have the exact solution y(t). Let the discrete boundary-value problem

$$\begin{aligned} &B_o v_o + B_1 v_J - \beta = o && && a) \\ &L_h v_i - r_i = o && i=1,\ldots,J && b) \end{aligned} \qquad (1.5)$$

be consistent with (1.1a,b) and stable. Then the method is convergent, that is,

$$||Y - V|| \to o \quad \text{as} \quad h_o \to o .$$

<u>Proof</u>: The definition of the truncation error $\tau[y(t)]$ combined with (1.5a,b) gives

$$\mathcal{B}_h(Y - V) = \tau .$$

By hypothesis, the difference scheme is stable, thus there exist constants $K_o \geq o$, $H > o$ such that

$$||Y - V|| \leq K_o||\tau||$$

for all nets with $h_o \leq H$. Also by hypothesis, the method is consistent, that is

$$||\tau|| \to o \quad \text{as} \quad h_o \to o .$$

Hence,

$$||Y - V|| = \max_{o \le i \le J} ||y(t_i) - v_i|| \to o \quad \text{as } h_o \to o$$

▨

For any particular scheme, we would like to be able to show that the net function approximates the exact solution at the net points: that the method is convergent. The truncation error and order of accuracy are generally derived via Taylor's Theorem, as was done in the example of Euler's method. In order to show that Euler's method is convergent for the discrete boundary-value problem, we need to prove that the matrix $\mathcal{B}_h$ in (1.9) has an inverse with uniformly bounded norm as $h_o \to o$. Then by Lemma 1.10, this numerical scheme would be convergent. However, even in this simple case, the nonsingularity of $\mathcal{B}_h$ is not obvious.

In Theorem 1.2 the initial-value problem is used to prove well-posedness of the boundary-value problem. As Keller [5] has shown for the case of the centered-Euler scheme, this approach is also useful in the numerical problem. If we approximate the solution of the initial-value problem

$$\begin{aligned} &u' - A(t)u - f(t) = o && \text{a)} \\ &u(o) = \beta && \text{b)} \end{aligned} \qquad (1.11)$$

by Euler's method, the following equations result

$$\begin{bmatrix} I & 0 & & \\ -\frac{1}{h_1} I - A(t_o) & \frac{1}{h_1} I & & \\ & \ddots & & \\ & \frac{1}{h_J} I - A(t_{J-1}) & \frac{1}{h_J} I \end{bmatrix} \begin{bmatrix} v_o \\ v_1 \\ \vdots \\ v_J \end{bmatrix} - \begin{bmatrix} \beta \\ f(t_o) \\ \vdots \\ f(t_{J-1}) \end{bmatrix} = 0.$$

Similarly, we define the discrete initial-value matrix I_h associated with any $\mathcal{B}_h$ (1.7) to be

$$I_h = \begin{bmatrix} I & 0 & \cdots & 0 \\ M_{1o} & M_{11} & \cdots & M_{1J} \\ \vdots & \vdots & \vdots & \vdots \\ M_{Jo} & M_{J1} & \cdots & M_{JJ} \end{bmatrix}.$$

To form I_h, we replace B_o, B_1 in the first block row of $\mathcal{B}_h$ with I,0 respectively.

3. Convergence for linear boundary-value problems

The main result of this chapter is that a numerical scheme applied to a boundary-value problem with a unique solution is stable and consistent if and only if the associated initial-value problem is stable and consistent. It will be useful to prove several lemmas first.

Lemma 1.12 (Factorization) Let I_h be the initial-value matrix associated with $\mathcal{B}_h$. Then

$$\mathcal{B}_h V - r \equiv I_h V + L\{NV - \xi\}$$

where L is a (J+1)n × n+1 matrix

$$L = \left[\begin{array}{c|c} I & o \\ 0 & r_1 \\ \vdots & \vdots \\ 0 & r_n \end{array}\right] ,$$

N is an (n+1) × (J+1)n matrix

$$N = \left[\begin{array}{ccccc} -I + B_o & 0 & \cdots & 0 & B_1 \\ \hline o & o & \cdots & o & o \end{array}\right] ,$$

and ξ is an (n+1)-vector

$$\xi = \left[\begin{array}{c} \beta \\ \hline 1 \end{array}\right] .$$

<u>Proof</u>: We take (1.7) to be the most general form of the boundary-value matrix $\mathcal{B}_h$.

$$\mathcal{B}_h = \begin{bmatrix} B_o & 0 & \cdots & B_1 \\ M_{1o} & M_{11} & \cdots & M_{1J} \\ \vdots & \vdots & & \vdots \\ M_{Jo} & M_{J1} & \cdots & M_{JJ} \end{bmatrix} .$$

The associated initial-value matrix is

$$\mathcal{I}_h = \begin{bmatrix} I & 0 & \cdots & 0 \\ M_{1o} & M_{11} & \cdots & M_{1J} \\ \vdots & \vdots & & \vdots \\ M_{Jo} & M_{J1} & \cdots & M_{JJ} \end{bmatrix} .$$

It may be directly verified that L,N,ξ are the proper factors, but the sequence of operations below may clarify the derivation.

$$\mathcal{B}_h V - r = \mathcal{I}_h V + \begin{bmatrix} I \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} [B_o - I \; 0 \; \ldots \; B_1]V \\ \\ -r \end{matrix}$$

$$= \mathcal{I}_h V + \left[\begin{array}{c|c} I & o \\ 0 & r_1 \\ \vdots & \vdots \\ 0 & r_J \end{array}\right] \left\{ \left[\begin{array}{cccc} B_o - I & 0 & \ldots & B_1 \\ \hline o & o & \ldots & o \end{array}\right] V - \left[\begin{array}{c} \beta \\ \hline 1 \end{array}\right] \right\}$$

The identity $r_o = \beta$ is used here.

▨

<u>Lemma 1.13 (Reducing)</u> Let L,N be p×q and q×p matrices respectively. Let x and b be a p-vector and a q-vector respectively. Further, let L have rank $q \leq p$. Then the matrix equation

$$(I + LN)x - Lb = o \qquad (1.14)$$

has a solution

$$x = Lw$$

if and only if

$$(I + NL)\, w - b = o \,,$$

<u>Proof</u>: The sufficiency is trivial on substitution of Lw into (1.14).

Since L is a p×q matrix with rank $q \leq p$, the columns of L are linearly independent. Therefore, there exists a $p \times (p-q)$ matrix $\bar{L}$ such that

$$\mathcal{R}(L) + \mathcal{R}(\bar{L}) = E^p.$$

We may write any vector $x \in E^p$ uniquely as

$$x = Lw + \bar{L}\bar{w} \quad . \tag{1.15}$$

Using this representation in (1.14),

$$L(w + NLw + N\bar{L}\bar{w} - b) + \bar{L}\bar{w} = o \quad .$$

The vector x is a solution of (1.14) if and only if

$$\bar{L}\bar{w} = o$$

and

$$w + NLw + N\bar{L}\bar{w} - b = o \quad .$$

The necessity follows.

▨

Note that this lemma reduces the solution of p simultaneous equations

$$(I + LN)\ x - Lb = o$$

to solution of q equations, $q \leq p$,

$$(I + NL)\ w - b = o.$$

Now we state and prove the main result of this chapter.

Theorem 1.16 Let the boundary-value problem (1.1a,b) have a unique solution $y(t) \in C^1[o,1]$. Then the following are equivalent:

a) The discrete boundary-value problem is stable, consistent (and convergent).

b) The associated initial-value problem is stable, consistent (and convergent).

Proof: (b ⟹ a) The Factorization Lemma (1.12) states that for the boundary-value problem (1.8)

$$\mathcal{B}_h V - r = \mathcal{I}_h V + L\{NV - \xi\} = o \; . \tag{1.17}$$

The initial-value problem is stable by hypothesis, therefore $\mathcal{I}_h$ is nonsingular for all $h_o \leq H$, say. Left multiplying (1.17) by $\mathcal{I}_h^{-1}$ we obtain

$$V + \mathcal{I}_h^{-1} L\{NV - \xi\} = o \; .$$

Define Z and z by

$$\mathcal{I}_h[Z \,\vdots\, z] = L \tag{1.18}$$

where Z is an $n(J+1)\times n$ matrix and z is an $n(J+1)$-vector. From (1.18), we note that Z satisfies

$$\mathcal{I}_h Z = \begin{bmatrix} I \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

and thus is an approximation to the fundamental matrix solution (1.3). By Lemma 1.10, the discrete initial-value problem is convergent

$$\max_{o\le i\le J} ||Z_i - X(t_i)|| \to o \quad \text{as } h_o \to o$$

and in particular

$$||Z_J - X(1)|| \to o \quad \text{as} \quad h_o \to o \, . \tag{1.19}$$

By virtue of (1.18), we may write (1.17) as

$$V + [Z \,\vdots\, z]\{NV - \xi\} = o \, . \tag{1.20}$$

The columns of the n(J+1)×n+1 matrix $[Z \,\vdots\, z]$ are linearly independent provided that $r_i \neq o$, $1 \le i \le J$. If the $(n+1)^{\underline{st}}$ column of L is the zero vector we remove it, replace $\xi = \begin{bmatrix} \beta \\ \hline 1 \end{bmatrix}$ by $\xi = [\beta]$ and the degeneracy is removed. We complete the proof assuming $r_j \neq o$, for some j.

By the Reducing Lemma, (1.20) has a solution if and only if $\begin{bmatrix} w \\ \hline \lambda \end{bmatrix}$ satisfies

$$\left\{ \left[\begin{array}{c|c} I_{n\times n} & o \\ \hline o & 1 \end{array} \right] + \left[\begin{array}{c|c} (B_o - I)Z_o + B_1 Z_J & B_1 z_J \\ \hline o & o \end{array} \right] \right\} \begin{bmatrix} w \\ \hline \lambda \end{bmatrix} - \begin{bmatrix} \beta \\ \hline 1 \end{bmatrix} = o \tag{1.21}$$

and the solution must be of the form

$$V = [Z \,\vdots\, z] \begin{bmatrix} w \\ \hline \lambda \end{bmatrix} \, .$$

We recall that $Z_o = I$ and take $\lambda = 1$, so that (1.21) is equivalent to

$$(B_o + B_1 Z_J)w - \beta + B_1 z_J = o \, . \tag{1.22}$$

Compare this with the condition derived in Theorem (1.2) for the solution of the boundary-value problem (1.1a,b). Existence and uniqueness of the solution, V, of the discrete boundary-value problem now hinge on the non-singularity of $(B_o + B_1 Z_J)$. We write

$$B_o + B_1 Z_J = [B_o + B_1 X(1)] - B_1(X(1) - Z_J) .$$

By hypothesis y(t) is unique, and Theorem 1.2 states that $[B_o + B_1 X(1)]$ must be non-singular. It has already been shown that

$$||X(1) - Z_J|| \to o \quad \text{as } h_o \to o$$

thus, by the Banach Lemma, $B_o + B_1 Z_J$ is non-singular for all nets with $h_o \leq H$, provided H is so small that for some $\rho \in (o,1)$

$$||(B_o + B_1 X(1))^{-1} B_1 \; (X(1) - Z_J)|| \leq 1 - \rho .$$

It is clear that the discrete boundary-value problem is stable since

$$||V|| \leq ||[Z \,\vdots\, z]|| \max \{||w||,1\}$$

and Z and z are solutions of discrete initial-value problems, which were assumed stable.

(a $\Rightarrow$ b) The proof of this part is essentially the same. Consider the discrete initial-value problem

$$\mathcal{I}_h V - r = o . \tag{1.23}$$

Following the Factorization Lemma (1.12), we may write

$$\mathcal{I}_h V - r = \mathcal{B}_h V - L\{NV + \xi\} .$$

By hypothesis, the discrete boundary-value problem is stable, thus

$\mathcal{B}_h$ is non-singular and we may define $[\hat{Z} \mid \hat{z}]$ by

$$\mathcal{B}_h[\hat{Z} \mid \hat{z}] = \left[\begin{array}{c|c} I & o \\ 0 & r_1 \\ \vdots & \vdots \\ 0 & r_J \end{array}\right] \equiv L \ . \qquad (1.24)$$

The Reducing Lemma (1.13) now implies that the solution V of (1.23) must be of the form

$$V = [\hat{Z} \mid \hat{z}] \left[\begin{array}{c} \hat{w} \\ \hline 1 \end{array}\right]$$

and exists if and only if

$$\hat{Z}_o \ \hat{w} - \beta + \hat{z}_o = o \, . \qquad (1.25)$$

Therefore, we wish to show that $\hat{Z}_o$ is non-singular for all nets with $h_o \leq H$, for H sufficiently small. Theorem 1.2 derives the solution of the boundary-value problem to be

$$y(t) = X(t)r + X(t) \int_o^t X^{-1}(s) \ f(s) \ ds \qquad (1.4)$$

where r must satisfy

$$[B_o + B_1 \ X(1)]r - \beta + B_1 \ X(1) \int_o^1 X^{-1}(s) \ f(s) \ ds = o \, .$$

Thus, the solution to the boundary-value problem

$$\begin{aligned} &\hat{X}'(t) - A(t) \ \hat{X}(t) = o && a) \\ &B_oX(o) + B_1X(1) - I = o && b) \end{aligned} \qquad (1.26)$$

is given by

$$\hat{X}(t) = X(t) \ R$$

where R is an n × n matrix and must satisfy

$$[B_o + B_1 X(1)]R - I = 0 .$$

By hypothesis y(t) is unique, thus the matrix R is non-singular, that is, $\hat{X}(o)$ is non-singular. We note that $\hat{Z}$ as defined in (1.24) is a convergent approximation to the equations (1.26), hence

$$||\hat{X}(o) - \hat{Z}_o|| \to o \text{ as } h_o \to o$$

and, as before, $\hat{Z}_o$ is non-singular for all nets with $h_o \leq H$, where H is sufficiently small. Now,

$$||V|| \leq ||[\hat{Z} \mid \hat{z}]|| \max \{||\hat{w}||, 1\}$$

and the discrete initial-value problem has a stable solution.

▨

We recall that the order of accuracy of a scheme is determined by the largest integer p such that

$$||\tau[y(t)]|| \leq K\, h_o^p \quad \text{for all } h_o \leq H.$$

<u>Corollary 1.27</u>. Let the boundary-value problem (1.1a,b) have a unique solution $y(t) \in C^{p+1}[o,1]$, for some integer $p \geq 1$. Let the discrete approximations (1.5a,b) be p-order accurate. Further, let the discrete initial-value problem associated with (1.5a,b) be stable. Then for all nets with $h_o \leq H$,

$$||Y - V|| \leq \bar{K}\, h_o^p$$

where V is the n(J+1)-vector solution of (1.5a,b) and $\bar{K}$ is a constant independent of h_o.

Proof. The proof follows that of Theorem 1.16.

▨

These results and the manipulations leading to them suggest several further topics. Chapter II will be concerned with the operational characteristics of various specific numerical schemes. Theorem 1.16 will be exploited to determine the stability properties of a particular class of schemes. We will examine asymptotic error expansions and discuss methods of increasing the accuracy of discrete approximations. In Chapter III, we will enlarge upon Theorem 1.16 and the equivalence of discrete initial-value and boundary-value problems.

Dr. H. O. Kreiss has recently published a paper [7] dealing with the stability of discrete approximations to arbitrary order systems of linear ordinary differential equations. Thus, it appears that this chapter is perhaps a special case of Kreiss' results. However, in Appendix A, we give an indication that Kreiss' work is complementary to our own and not inclusive. Further, Theorem 1.16 gives a necessary and sufficient condition for stability of the most general schemes, whereas Dr. Kreiss deals with k-step methods.

Chapter II

Difference Schemes

In this chapter, we will examine some specific difference schemes; of particular interest are one-step (two-point) methods. One-step methods for linear problems have the characteristic form

$$L_h v_i = M_{ii} v_i + M_{i,i-1} v_{i-1}.$$

These methods are of special interest for several reasons. They admit nonuniform nets and, as Keller [5] has shown, piecewise smooth solutions. In addition, the calculations involved in solving the matrix equation

$$\mathcal{B}_h V - r = o$$

for separated boundary conditions are particularly simple. In this case, the matrix $\mathcal{B}_h$ may be put into a block tridiagonal form which possesses a simple LU-decomposition.

One of two methods is usually employed to generate difference approximations to (1.1a) and evaluate their order of accuracy. The first approach is via Taylor series and necessitates evaluating the truncation error functions at some reference point $t_R(t_i)$ where

$$t_R(t_i) \to t_i \quad \text{as} \quad h_o \to o .$$

Quadrature formulae may also be used to generate useful approximations. If we write (1.1a) in the equivalent form

$$u(t) - u(\tau) - \int_\tau^t A(s)\ u(s) + f(s)\ ds = o \qquad (2.1)$$

the application of quadrature is immediately obvious.

1. Taylor Series

The centered-Euler or Box-scheme (Keller [5]) illustrates the use of Taylor series. Applied to (1.1a) the centered Euler scheme is

$$L_h v_i = \frac{1}{h_i}[I - \frac{h_i}{2} A(t_{i-1/2})]\, v_i + \frac{1}{h_i}[-I - \frac{h_i}{2} A(t_{i-1/2})]\, v_{i-1}$$

$$r_i = f(t_{i-1/2})$$

where $t_{i-1/2} = t_i - \frac{h_i}{2}$.

If we assume that $y(t) \in \mathcal{C}^{2m+1}[o,1]$ and let $t_R(t_i) = t_{i-1/2}$, the truncation error is shown to be

$$\tau_i[y(t)] = \sum_{k-1}^{m-1} (\frac{h_i}{2})^{2k} \frac{1}{(2k)!} \left\{ \frac{y^{(2k+1)}}{2k+1} - Ay^{(2k)} \right\} (t_R)$$

$$+ 0\left((\frac{h_i}{2})^{2m} \right) \tag{2.3a}$$

and the leading term, $0(h_i^2)$, is

$$(\frac{h_i}{2})^2 \frac{1}{2} \left\{ \frac{y^{(3)}}{3} - Ay^{(2)} \right\} (t_R) . \tag{2.3b}$$

Here the choice of $t_R = t_{i-1/2}$ is important due to the symmetry of (2.2). However, the first term of the truncation error expansion, (2.3b), is invariant and hence the order of accuracy of any scheme must be invariant to the reference point t_R. If we were to use $t_R = t_{i-1}$, the truncation error is shown to be

$$\tau_i[y(t)] = \sum_{k=2}^{2m-1} \left\{ \frac{h_i^k}{k!} \frac{y^{(2k+1)}}{k+1} - (\tfrac{1}{2})^{k+1} A^{(k)} y \right.$$

$$\left. - \sum_{p=o}^{k} (\tfrac{1}{2})^{p+1} \binom{k}{p} A^{(p)} y^{(k-p)} - (\tfrac{1}{2})^k f^{(k)} \right\} (t_R) \qquad (2.4a)$$

$$+ O(h_i^{2m})$$

and the leading term is

$$\frac{h_i^2}{2} \left\{ \frac{y^{(3)}}{3} - \frac{A^{(2)} y}{4} - \frac{1}{2} A y^{(2)} - \frac{1}{2} A^{(1)} y^{(1)} - \frac{1}{4} f^{(2)} \right\} (t_R). \quad (2.4b)$$

However, recalling that y(t) is an exact solution of (1.1a), this expression is shown to be equivalent to the following

$$\frac{h_i^2}{2} \left\{ \frac{1}{4} (y'-Ay-f)'' + \frac{1}{12} y^{(3)} - \frac{1}{4} Ay^{(2)} \right\} (t_R)$$

$$= \quad \frac{1}{2} \left(\frac{h_i}{2} \right)^2 \left\{ \frac{1}{3} y^{(3)} - Ay^{(2)} \right\} (t_R) .$$

Thus, the leading term of (2.3a) is exactly the same as (2.4a) with $t_R = t_{i-1/2}$ and $t_R = t_{i-1}$ respectively. Note that the next terms in each expansion are not the same.

The centered-Euler scheme may be generalized to an arbitrarily high order by employing the exact form of the truncation error. Incorporating the first term, $O(h_i^2)$, of the truncation error expansion

(2.3a) into the centered-Euler difference scheme produces a fourth-order scheme. Using the differential equation (1.1a), the h_i^2 term is

$$\frac{1}{2}\left\{\frac{1}{3}\,y^{(3)} - Ay^{(2)}\right\}\left(t_{i-1/2}\right)$$

$$= \frac{1}{6}\left\{\left[A^{(2)} + 2A^{(1)}\,A - 2\,AA^{(1)} - 2A^3\right] y\right.$$

$$\left. + \left[2A^{(1)}\,f^{(1)} + f^{(2)} - 2A^2 f - 2Af^{(1)}\right]\right\}\,(t_{i-1/2})$$

$$\equiv\ C(t_{i-1/2})\;y(t_{i-1/2}) + g(t_{i-1/2}).$$

Note the identity

$$y(t_{i-1/2}) = \frac{1}{2}\left[y(t_i) + y(t_{i-1})\right] + 0\left(\left(\frac{h_i}{2}\right)^2\right)\ .$$

A fourth-order accurate scheme may now be defined as

$$\frac{1}{h_i}\left[I - \frac{h_i}{2}\,A - \left(\frac{h_i}{2}\right)^3 C\right](t_{i-1/2})\;v_i + \frac{1}{h_i}\left[-I - \frac{h_i}{2}\,A\right.$$

$$\left. - \frac{h_i}{2}^3\,C\right](t_{i-1/2})\;v_{i-1} - f(t_{i-1/2}) - \left(\frac{h_i}{2}\right)^2 g(t_{i-1/2}) = 0. \qquad (2.5)$$

The truncation error is shown to be

$$\tau_i[y(t)] = \sum_{k=2}^{m-1}\left(\frac{h_i}{2}\right)^{2k}\left[\frac{y^{(2k+1)}}{(2k+1)!} - A\frac{y^{(2k)}}{(2k)!} - C\frac{y^{(2k-2)}}{(2k-2)!}\right]\;(t_{i-1/2})$$

$$+ \,0\left(\left(\frac{h_i}{2}\right)^{2m}\right)\ .$$

This augmented difference scheme has all the characteristics of the centered-Euler scheme and the advantage of being a fourth-order method. One disadvantage is the evaluation of C(t) and g(t); in addition to

$A(t)$ and $f(t)$, we must also evaluate $A^{(2)}(t)$, $A^{(1)}(t)$, $f^{(2)}(t)$, $f^{(1)}(t)$ and various products of these functions.

Since the functions $C(t)$ and $g(t)$ will continue to appear, we need to define these quantities in more generality. The usual procedure will be to evaluate $y^{(p)}(t)$ in terms of derivatives and combinations of $A(t)$, $f(t)$ and $y(t)$. For simplicity, we define

$$\frac{d^{\nu}y(t)}{dt^{\nu}} = C_{\nu}(t)y(t) + g_{\nu}(t) \qquad (2.6)$$

where $C_{\nu}(t)$ is an $n \times n$ matrix and $g_{\nu}(t)$ is an n-vector.

Thus,

$$y^{(o)}(t) = y(t) \Rightarrow C_o = I, \; g_o = o \,.$$

The terms of particular interest are contained in the following table.

ν	C_{ν}	g_{ν}
o	I	0
1	A	f
2	$A^{(1)}+A^2$	$Af+f^{(1)}$
3	$A^{(2)}+2A^{(1)}A+AA^{(1)}+A^2$	$(2A^{(1)}+A^2)f + Af^{(1)}+f^{(2)}$

Table 2.7

Employing this notation, the fourth-order generalized centered-Euler scheme (2.5) is

$$\frac{1}{h_i}\left\{C_o - \frac{h_i}{2} C_1 - \left(\frac{h_i}{2}\right)^3 \frac{1}{3!} (C_3-3C_2)\right\}(t_{i-1/2})v_i$$

$$-\frac{1}{h_i}\left\{C_o + \frac{h_i}{2} C_1 - \left(\frac{h_i}{2}\right)^3 \frac{1}{3!} (C_3-3C_2)\right\}(t_{i-1/2})v_{i-1} \qquad (2.8)$$

$$- \left\{g_1 + \left(\frac{h_i}{2}\right)^2 \frac{1}{3!} (g_3-3g_2)\right\}(t_{i-1/2}) = o .$$

In the next section, this method will be improved upon in the sense that the functions C_o, C_1, C_2, and C_3 will be used to define a sixth-order-method.

2. Quadrature Employing (2.1) we get

$$\frac{1}{h_i}\left[y(t_i)-y(t_{i-1})\right] - \frac{1}{h_i}\int_{t_{i-1}}^{t_i} [A(s)y(s) + f(s)]ds = o \qquad (2.9)$$

where y(t) is the exact solution to (1.1a,b). If we apply the trapezoidal rule to (2.9), the result is

$$\frac{1}{h_i}[y(t_i)-y(t_{i-1})] - \frac{1}{2}[A(t_i)y(t_i) + A(t_{i-1})y(t_{i-1})]$$

$$-\frac{1}{2}[f(t_i) + f(t_{i-1})] = O(h_i^2).$$

Thus a second order accurate difference scheme (Trapezoidal Rule) may be defined by

$$\frac{1}{h_i}\left[I - \frac{h_i}{2} A(t_i)\right]v_i + \frac{1}{h_i}\left[- I - \frac{h_i}{2} A(t_{i-1})\right] v_{i-1}$$

$$-\frac{1}{2}\left[f(t_i) + f(t_{i-1})\right] = o .$$

This may be equivalently written as

$$\frac{1}{h_i}\left\{C_o - \frac{h_i}{2}C_1\right\}(t_i)v_i + \frac{1}{h_i}\left\{-C_o - \frac{h_i}{2}C_1\right\}(t_{i-1})v_{i-1}$$

$$-\frac{1}{2}\left[g_1(t_i) + g_1(t_{i-1})\right] = o \tag{2.10}$$

using the functions in Table 2.7. This one-step method and others of arbitrarily high order may be derived in a more general setting by means of (2.9).

Lemma 2.11. Let $b(t) \in C^{m+1}[a,b]$. Also let polynomials $p_i(t)$ be defined by

$$\begin{aligned} p_o(t) &= 1 && \text{a)} \\ p_\nu(t) &= \int_{\xi_{\nu-1}}^{t} p_{\nu-1}(s)\,ds \qquad \nu=1,2,\ldots && \text{b)} \end{aligned} \tag{2.12}$$

where $\xi_i \in [a,b]$. Then

$$\text{a)}\quad \int_a^b b(s)ds = \left\{\sum_{k=o}^{m} (-1)^k\, p_{k+1}(t)\; b^{(k)}(t)\right|_a^b$$

$$+ (-1)^{m+1}\int_a^b p_{m+1}(s)\; b^{(m+1)}(s)\; ds\;,$$

$$\text{b)}\quad \max_{[a,b]}\; |p_\nu(t)| \le |b-a|^\nu.$$

Proof: Repeated integration by parts yields (a) immediately. Induction on ν yields (b). The induction hypothesis is started by

(2.12a) and, formally, is

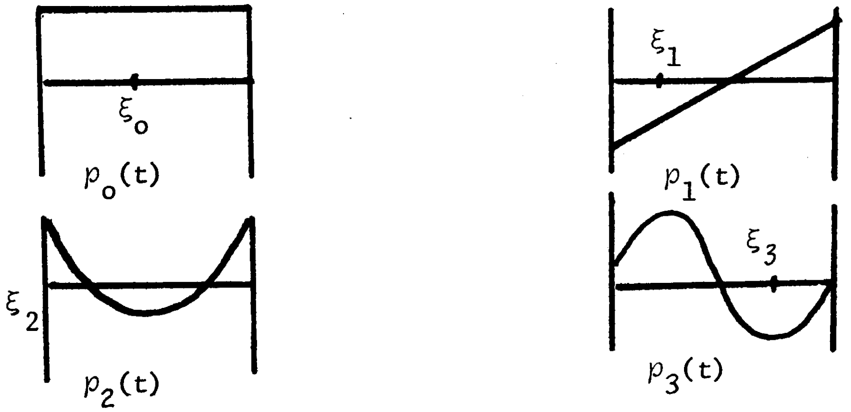
$$|p_{\nu-1}(t)| \le |b-a|^{\nu-1}.$$

Thus,

$$|p_{\nu}(t)| \le \int_{\xi_i}^{t} |p_{\nu-1}(s)|\,|ds| \le |b-a|^{\nu-1}|t-\xi_i| \le |b-a|^{\nu}$$

▨

Innumerable schemes present themselves by way of this Lemma. The Euler-Maclarin sum formula may be derived using polynomials which look like



where the scale for each polynomial is necessarily different. Thus, the Euler-Maclarin formula may be used to define a difference scheme of arbitrarily high order. A fifth-order method defined in this manner is

$$\begin{aligned}
&\frac{1}{h_i}\left\{C_o - \frac{h_i}{2} C_1 + \frac{h_i^2}{2} C_2 - \frac{h_i^4}{720} C_4\right\}(t_i)v_i \\
&\qquad - \frac{1}{h_i}\left\{C_o + \frac{h_i}{2} C_1 + \frac{h_i^2}{2} C_2 - \frac{h_i^4}{720} C_4\right\}(t_{i-1})v_{i-1} \\
&- \left\{\frac{1}{2}g_1 - \frac{h_i}{2} g_2 + \frac{h_i^3}{720} g_4\right\}(t_i) - \left\{\frac{1}{2} g_1 + \frac{h_i}{2} g_2 - \frac{h_i^3}{720} g_4\right\}(t_{i-1}) = 0.
\end{aligned} \qquad (2.13)$$

One undesirable aspect of (2.13) is evaluating $C_4(t)$ and $g_4(t)$. In order to employ (2.13) we must evaluate $A^{(3)}(t)$, $A^{(2)}(t)$, $A^{(1)}(t)$, $A(t)$, $f^{(3)}(t)$, $f^{(2)}(t)$, $f^{(1)}(t)$, $f(t)$ and assorted products and sums of these functions.

3. *Gap schemes* Thus far, we have considered several one-step schemes. The centered-Euler scheme was generalized in Section 1 to a fourth-order accurate scheme. This increase in accuracy is bought at the price of evaluating $A^{(2)}(t)$, $A^{(1)}(t)$, $f^{(2)}(t)$, and $f^{(1)}(t)$ in addition to $A(t)$ and $f(t)$. In Section 2, the Euler-Maclarin Sum Formula was used to define a fifth-order method, but this difference scheme required evaluation of $A^{(3)}(t)$, $f^{(3)}(t)$ and all lower order derivatives. Since each new function to be evaluated increases both programming complexity and computation time, we want to examine high-order, one-step methods with a minimum of derivatives required.

The centered-Euler scheme and the Trapezoidal Rule (2.10) possess another useful property. Each of these difference schemes has a truncation error expansion which contains only even powers of h_i.

For reasons that will become clear, see Section 6 on asymptotic error expansions, we want to preserve this property in considering high-order methods.

Recalling Lemma 2.11, we have the result

$$\int_a^b b(s)ds = \left\{ \sum_{k=o}^{m} (-1)^k \, p_{k+1}(t) \; b^{(k)}(t) \right|_a^b$$

$$+ (-1)^{m+1} \int_a^b p_{m+1}(s) \; b^{(m+1)}(s) \; ds.$$

If the sequence of polynomials $\{p_\nu(t)\}_{\nu=o}^{\nu=m}$ could be defined in such a way that

$$p_\nu(a) = p_\nu(b) = o \qquad\qquad \nu=n+1,\ldots,2n$$

then a difference method of order 2n can be defined which requires the evaluation of $A^{(n-1)}(t)$, $f^{(n-1)}(t)$, and all lower-order derivatives. For instance, a fourth-order method so defined requires evaluating $A^{(1)}(t)$ and $f^{(1)}(t)$ where as the generalized centered-Euler scheme needs $A^{(2)}(t)$, $f^{(2)}(t)$, etc.

<u>Lemma 2.14</u> Let $p_{2n}(t) = \frac{1}{n!}(t-a)^n(t-b)^n$. Define the sequence of polynomials $\{p_\nu(t)\}$, $\nu \geq o$, such that

$$p_\nu(t) = \frac{dp_{\nu+1}(t)}{dt} \qquad \nu=2n-1,\ldots,o \qquad a)$$

$$p_{\nu+1}(t) = \int_{\xi_\nu}^{t} p_\nu(s)ds \qquad \nu=2n,\ldots \qquad (2.15)$$

$$b)$$

$$\xi_{2r} = \frac{a+b}{2}, \; \xi_{2r+1} = a\,.$$

Then

a) $\{p_\nu(t)\}$ satisfy conclusion a) of Lemma 2.11.

b) $|p_\nu(t)| \leq |b-a|^\nu$ $\qquad \nu \geq 2n,$

c) $p_\nu(a) = p_\nu(b) = o$ $\qquad \nu=n+1,\ldots,2n,$

d) $p_{2r}(a) = p_{2r}(b) = o$ $\qquad r > n.$

Proof: We can replace (2.15a) with

$$p_\nu(t) = \int_{\xi_\nu}^{t} p_{\nu-1}(s)\,ds \qquad\qquad p_o(t)=1$$

and the sequence satisfies conclusion a) of Lemma 2.11.

b. $|p_{2n}(t)| \leq |b-a|^{2n}$ and the result is immediate by induction.

c. Immediate from (2.15a).

d. To prove this part of the Lemma, it is sufficient to show that $p_{2r}(t)$, $r \geq n$, is an even function about $t = \frac{b+a}{2}$. Without loss of generality, we assume that $\frac{b+a}{2} = o$, thus $\xi_{2r} = o$. By hypothesis, $p_{2n}(t)$ is even about $\frac{b+a}{2}$, hence we assume that $p_{2(\nu-1)}(t)$ is an even function.

$$p_{2\nu-1}(t) = \int_o^t p_{2\nu-2}(s)\,ds = -\int_o^{-t} p_{2(\nu-1)}(s)\,ds = -p_{2\nu-1}(-t).$$

Therefore, $p_{2\nu-1}(t)$ is an odd function.

$$p_{2\nu}(t) = \int_{-a}^{t} p_{2\nu-1}(s)\,ds = \int_{-a}^{-t} p_{2\nu-1}(s)\,ds + \int_{-t}^{t} p_{2\nu-1}(s)\,ds$$

$$= p_{2\nu}(-t).$$

Thus, by induction $p_{2r}(t)$, $r \geq n$, is an even function about $t = \frac{a+b}{2}$.

By hypothesis $p_{2r}(a) = o$, thus $p_{2r}(b) = o$.

▨

This lemma proves that the difference scheme

$$v_i - v_{i-1} + \sum_{k=1}^{n} (-1)^k \frac{d^{2n-k} p_{2n}(t_i)}{dt^{2n-k}} \left[C_k(t_i)v_i - (-1)^k C_k(t_{i-1})v_{i-1} \right]$$

$$+ \sum_{k=1}^{n} (-1)^k \frac{d^{2n-k} p_{2n}(t_i)}{dt^{2n-k}} \left[g_k(t_i) - (-1)^k g_k(t_{i-1}) \right] = 0 \qquad (2.16)$$

is of order 2n. We need to evaluate the derivatives $p_{2n}^{(2n-k)}(t_i)$.

$$p_{2n}(t) = (t-t_{i-1})^n (t-t_i)^n = P_+(t)P_-(t).$$

$$\frac{d^m p_{2n}(t)}{dt^m} = \frac{1}{2n!} \sum_{L=o}^{m} \frac{m!}{L!(m-L)!} P_+^{(m-L)}(t_i)\ P_-^{(L)}(t_i).$$

For $L > n$, $P_-^{(L)}(t) = o$, hence

$$\frac{d^m p_{2n(t_i)}}{dt^m} = \frac{1}{2n!} \sum_{L=o}^{n} \binom{m}{L} P_+^{(m-L)}(t_i)\ P_-^{(L)}(t_i), \qquad m \geq n.$$

Recalling the definition of $P_-(t)$, for $L < n, P_-^{(L)}(t_i) = o$.

Thus,

$$\frac{d^m p_{2n}(t_i)}{dt^m} = \frac{1}{2n!}\binom{m}{n} P_+^{(m-n)}(t_i)\ P_-^{(n)}(t_i)$$

$$= \frac{m!}{2n!(m-n)!}\ \frac{n!}{(2n-m)!}\ \left(h_i\right)^{2n-m} .$$

Define $\alpha_k^n \equiv (-1)^k \dfrac{(2n-k)!}{2n!(n-k)!}\ \dfrac{n!}{k!}$ and (2.16) becomes

$$v_i - v_{i-1} + \sum_{k=1}^{n} \alpha_k^n \left(\frac{h_i}{2}\right)^k \left[c_k(t_i)v_i - (-1)^k c_k(t_{i-1})v_{i-1}\right]$$

$$+ \sum_{k=1}^{n} \alpha_k^n \left(\frac{h_i}{2}\right)^k \left[g_k(t_i) - (-1)^k\ g_k(t_{i-1})\right] = 0 .$$

		k				
	α_k^n	1	2	3	4	5
	2	-1				
	4	-1	1/3			
2n	6	-1	2/5	-1/15		
	8	-1	3/7	-2/21	1/105	
	10	-1	4/9	-1/9	1/630	-1/945

Table 2.17

As examples, we write out the fourth-order and sixth-order schemes.

Fourth-order scheme:

$$\frac{1}{h_i}\left\{C_o - \frac{h_i}{2} C_1 + \frac{h_i^2}{12} C_2\right\}(t_i)v_i - \frac{1}{h_i}\left\{C_o + \frac{h_i}{2} C_1 + \frac{h_i^2}{12} C_2\right\}(t_{i-1})v_{i-1}$$

$$-\left\{\frac{1}{2}g_1 - \frac{h_i}{12}g_2\right\}(t_i) - \left\{\frac{1}{2}g_1 + \frac{h_i}{12}g_2\right\}(t_{i-1}) = 0. \tag{2.18}$$

Sixth-order scheme:

$$\frac{1}{h_i}\left\{C_o - \frac{h_i}{2} C_1 + \frac{h_i^2}{5} C_2 - \frac{h_i^3}{120} C_3\right\}(t_i)v_i$$

$$-\frac{1}{h_i}\left\{C_o + \frac{h_i}{2} C_1 + \frac{h_i^2}{5} C_2 + \frac{h_i^3}{120} C_3\right\}(t_{i-1})v_{i-1} \tag{2.19}$$

$$-\left\{\frac{1}{2}g_1 - \frac{h_i}{5}g_2 + \frac{h_i^2}{120}g_3\right\}(t_i) - \left\{\frac{1}{2}g_1 + \frac{h_i}{5}g_2 + \frac{h_i^2}{120}g_3\right\}(t_{i-1}) = 0 .$$

The matrices C_ν and the vectors g_ν are defined in Table 2.7.

Dr. Keller noted that these gap schemes might be derived using Hermite interpolation to estimate the function

$$b(t) = A(t)y(t) + f(t) .$$

For the fourth-order gap scheme (2.18), we write

$$b(t) = H(t) + (b(t) - H(t))$$

where H(t) is the lowest order polynomial that matches $b(t_i)$, $b^{(1)}(t_i)$, $b(t_{i-1})$, and $b^{(1)}(t_{i-1})$. In this case,

$$H(t) = b(t_{i-1})\,\frac{t_i-t}{h_i} + b(t_i)\,\frac{t-t_{i-1}}{h_i}$$

$$-\left[\frac{b(t_i) - b(t_{i-1})}{h_i} - b^{(1)}(t_{i-1})\right]\frac{(t_i-t)^2(t-t_{i-1})}{h_i^2}$$

$$+\left[\frac{b(t_i) - b(t_{i-1})}{h_i} - b^{(1)}(t_i)\right]\frac{(t_i-t)(t-t_{i-1})^2}{h_i^2} .$$

Recalling the integral equation (2.1), we have

$$y(t_i) - y(t_{i-1}) - \int_{t_{i-1}}^{t_i} H(s)\,ds = \int_{t_{i-1}}^{t_i} [b(s)-H(s)]ds .$$

$H(t)$ may be integrated exactly to get

$$y(t_i) - y(t_{i-1}) - \frac{h_i}{2}\,[b(t_i) + b(t_{i-1})] + \frac{h_i^2}{12}\,[b^{(1)}(t_i) - b^{(1)}(t_{i-1})]$$

$$= \int_{t_{i-1}}^{t_i} [b(s)-H(s)]ds .$$

Since $b(t) = y^{(1)}(t)$, this formulation will yield a difference scheme exactly the same as (2.18) with the truncation error

$$\tau_i[y(t)] = \frac{1}{h_i}\int_{t_{i-1}}^{t_i} [y^{(1)}(s)-H(s)]ds .$$

4. Other Schemes We are particularly interested in one-step difference schemes of the form (2.2), (2.5), and (2.18), thus only a brief mention will be made of other methods. The integral equation (2.1) suggests k-step methods of the form

$$\frac{1}{h_i}[v_i-v_{i-k}] - \sum_{p=o}^{k} M_{i,i-p}v_{i-p} - r_i = o \quad i=k,\dots,J. \tag{2.20}$$

Note that (2.20) represents (J-k+1)n equations in (J+1)n unknowns, and kn more equation must be found to determine V uniquely. These extra equations are given by the usual boundary conditions (1.5b) and by "starting" m-step difference schemes, $m < k$, usually of lower order than (2.20). In many cases the "starting" scheme is a one-step scheme. Combining all of these equations into the form

$$\mathcal{B}_h V - r = 0$$

the initial-value matrix, $\mathcal{I}_h$, is lower triangular. If the one-step "starting" scheme is of lower order than the k-step method, the mesh size, h(1), associated with the one-step method should be smaller than that of the k-step method, h(k). If the methods are of order p and q respectively ($p < q$), then Corollary 1.27 indicates that h(1) and h(k) should be related by

$$h(1) = h(k)^{q/p}.$$

5. Stability of Triangular Difference Schemes All of the methods mentioned here and most of those commonly used give rise to block triangular initial-value matrices $\mathcal{I}_h$. That is, they may be written

in the form

$$\mathcal{I}_h = \begin{bmatrix} I & 0 & \cdots & 0 \\ JM_{1o}+\bar{M}_{1o} & JM_{11}+\bar{M}_{11} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ JM_{Jo}+\bar{M}_{Jo} & M_{J1}+\bar{M}_{J1} & \cdots & JM_{JJ}+\bar{M}_{JJ} \end{bmatrix} . \qquad (2.21)$$

The factor of J appearing with all M_{ij}'s is a normalization of the inverse of the mesh size. We would like to examine the stability of families of matrices $\mathcal{I}_h$ of this form. The stability theory for discrete initial-value problems is well-developed provided that a k-step difference scheme is used and all M_{ij} can be written as $\mathit{M}_{ij}I$, M_{ij} a scalar. We are interested in a result which does not require these two restrictions on $\mathcal{I}_h$.

Lemma (2.22). Let $\mathcal{I}_h$ be a matrix of the form of (2.21). Define D_ν, N_ν, ν=o,1, by

$$JD_o+D_1 = \begin{bmatrix} J\left(\frac{1}{J}\ I\right) & & & \\ & JM_{11}+\bar{M}_{11} & & \\ & & \ddots & \\ & & & JM_{JJ}+\bar{M}_{JJ} \end{bmatrix} ,$$

$$-(JN_o+N_1) = \begin{bmatrix} 0 & & & \\ JM_{1o}+\bar{M}_{1o} & 0 & & \\ \vdots & & \ddots & \\ JM_{Jo}+M_{Jo} & JM_{J1}+\bar{M}_{J1} & \cdots & 0 \end{bmatrix} .$$

Let D_o be nonsingular and let $||D_o^{-1}|| = \frac{1}{d_o}$, $||D_1|| = d_1$, $||N_o|| = n_o$, and $||N_1|| = n_1$, independent of J. Further, let $\frac{n_o}{d_o} \leq 1$. Then

$$||\mathcal{I}_h^{-1}|| \leq e^{2\left(\frac{d_1}{d_o} + \frac{n_1}{n_o}\right)} \frac{2}{d_o}\left[1 + \left(\frac{d_1}{d_o} + \frac{n_1}{n_o}\right)^{-1}\right] \quad . \quad (2.23)$$

Proof: The matrix $\mathcal{I}_h$ is written as

$$\mathcal{I}_h = JD_o + D_1 - JN_o - N_1$$

or, equivalently

$$\mathcal{I}_h = \mathcal{J}(D_o + \mathcal{J}^{-1}D_1) - \mathcal{J}(N_o + \mathcal{J}^{-1}N_1)$$

where

$$\mathcal{J} = \begin{bmatrix} I & & & \\ & JI & & \\ & & \ddots & \\ & & & JI \end{bmatrix} \qquad \text{and}$$

$$\mathcal{I}_h^{-1} = [D_o + \mathcal{J}^{-1}D_1 - (N_o + \mathcal{J}^{-1}N_1)]^{-1} \mathcal{J}^{-1} .$$

By hypothesis and the Banach Lemma, $D_o + \mathcal{J}^{-1}D_1$ is nonsingular for $J > J_o = \frac{d_1}{d_o}$, and

$$\mathcal{I}_h^{-1} = [I - (D_o + \mathcal{J}^{-1}D_1)^{-1}(N_o + \mathcal{J}^{-1}N_1)]^{-1}(D_o + \mathcal{J}^{-1}D_1)^{-1}\mathcal{J}^{-1} .$$

Since $(D_o + \mathcal{J}^{-1}D_1)^{-1}$ is block diagonal and $(N_o + \mathcal{J}^{-1}N_1)$ is block nilpotent, the product $(D_o + \mathcal{J}^{-1}D_1)^{-1}$ $(N_o + \mathcal{J}^{-1}N_1)$ is block nilpotent.

Thus,

$$\mathcal{I}_h^{-1} = \sum_{k=o}^{J-1} \left[(D_o+\mathcal{J}^{-1}D_1)^{-1}(N_o+\mathcal{J}^{-1}N_1)\right]^k (D_o+\mathcal{J}^{-1}D_1)^{-1}\mathcal{J}^{-1}$$

$$= \sum_{k=o}^{J-1} \left[(D_o+\mathcal{J}^{-1}D_1)^{-1}(N_o+\mathcal{J}^{-1}N_1)\right]^k (D_o+\mathcal{J}^{-1}D_1)^{-1}(\mathcal{J}_1+\mathcal{J}_2)$$

where

$$\mathcal{J}_1 = \begin{bmatrix} I & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} , \qquad \mathcal{J}_2 = \begin{bmatrix} 0 & & & \\ & \frac{1}{J}\, I & & \\ & & \ddots & \\ & & & \frac{1}{J}\, I \end{bmatrix} .$$

$$||\mathcal{I}_h^{-1}|| \leq \frac{1}{J} \frac{1}{d_o - \frac{1}{J}d_1} \left|\left| \sum_{k=o}^{J-1} \left[(D_o+\mathcal{J}^{-1}D_1)^{-1} (N_o+\mathcal{J}N_1)\right]^k \right|\right|$$

$$+ \frac{1}{d_o - \frac{1}{J}d_1} \left|\left| \sum_{k=o}^{J-1} [(D_o+\mathcal{J}^{-1}D_1)^{-1}(N_o+\mathcal{J}^{-1}N_1)]^k \mathcal{J}_1 \right|\right|$$

$$\leq \frac{2}{d_o} \left[\frac{d_1}{d_o} + \frac{n_1}{n_o}\right]^{-1} \left[e^{2\left(\frac{d_1}{d_o} + \frac{n_1}{n_o}\right)} -1\right] + \frac{2}{d_o}\, e^{2\left(\frac{d_1}{d_o} + \frac{n_1}{n_o}\right)}$$

$$= \frac{2}{d_o} \left[1 + \left(\frac{d_1}{d_o} + \frac{n_1}{n_o}\right)\right]^{-1} e^{2\left(\frac{d_1}{d_o} + \frac{n_1}{n_o}\right)} .$$

▨

Combining Corollary 1.27 and Lemma (2.22), we state the following Theorem without proof.

Theorem 2.24. Let the linear boundary-value problem (1.1a,b) have a unique solution $y(t) \in C^1[0,1]$. Also, let the difference scheme be p^{th}-order accurate. Let $\mathcal{B}_h$ be of the form

$$\mathcal{B}_h = \begin{bmatrix} B_o & 0 & \cdots & B_1 \\ JM_{1o}+\overline{M}_{1o} & JM_{11}+\overline{M}_{11} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ JM_{Jo}+\overline{M}_{Jo} & JM_{J1}+\overline{M}_{J1} & \cdots & JM_{JJ}+\overline{M}_{JJ} \end{bmatrix}$$

where M_{ii} is nonsingular for every $i=0,\ldots,J$ and $||M_{\nu\nu}^{-1}|| \leq \frac{1}{d_o}$, $||\overline{M}_{\nu\nu}|| \leq d_1$, independent of J. Further, let

$$||(M_{\nu o}\ M_{\nu 1} \cdots M_{\nu,\nu-1}\ 0 \cdots 0)|| \leq n_o,$$

$$||(\overline{M}_{\nu o}\ \overline{M}_{\nu 1} \cdots \overline{M}_{\nu,\nu-1}\ 0 \cdots 0)|| \leq n_1,$$

and $\frac{n_o}{d_o} \leq 1$. Then for $h_o \leq H$, H sufficiently small,

$$||Y - V|| \leq \overline{K}\, h_o^p$$

▨

This Theorem gives a precise estimate on the convergence of finite difference solutions. However, we now want to look at a more detailed accounting of the error incurred by approximating (1.1a,b) with a difference scheme.

6. Asymptotic Error Expansions

In Section 3, the gap schemes were derived so that the truncation error expansion would have only even powers of h_i. The advantage of this arrangement will be evident in this discussion. The general difference scheme is defined as

$$L_h v_i - r_i = o \qquad i=1,\ldots,J$$

$$B_o v_o + B_1 v_J - \beta = o \ .$$

The truncation error is here defined to be

$$\tau_i[z(t)] \equiv L_h z(t_i) - r_i \qquad i=1,\ldots,J \qquad \text{a)}$$

$$\tau_o[z(t)] \equiv B_o z(t_o) + B_1 z(t_J) - \beta \qquad \text{b)} \qquad (2.25)$$

where $z(t) \in C^{p+1}[0,1]$.

For example, Euler's method has the truncation error, letting $t_R = t_{i-1}$,

$$\tau_i[z(t)] = Lz(t_R) - f(t_R) + \sum_{k=2}^{p} h_i^{k-1} \frac{z^{(k)}(t_{i-1})}{k!} + O(h_i^p) \ .$$

In general, the truncation error will be assumed to be of the form

$$\tau_i[z(t)] = Lz(t_R) - f(t_R)$$

$$+ \sum_{k=1}^{p} (h_i)^k F_k(A,f) z(t_R) + O(h_i^{p+1}) \qquad (2.26)$$

where each $\mathcal{F}_k(A,f)$ is a linear differential operator of at most order k+1 which depends, in general nonlinearly, upon $A^{(k)}(t_R)$, $f^{(k)}(t_R)$ and all lower order derivatives. For Euler's method,

$$\mathcal{F}_k = \frac{1}{(k+1)!}\,\frac{d^{(k+1)}}{dt^{k+1}}\,.$$

Theorem 2.27. Let the differential equation (1.1a,b) have a unique solution $y(t) \in C^{q+1}[0,1]$ where $A(t) \in C^q[0,1]$ and $f(t) \in C^q[0,1]$. Let

$$B_o v_o + B_1 v_J - \beta = 0$$

$$L_h v_i - r_i = 0 \qquad\qquad i=1,2,\ldots,J$$

be a stable, consistent difference approximation to (1.1a,b). Further, let the truncation error be given by

$$\tau_o[z(t)] = B_o z(t_o) + B_1 z(t_1) - \beta \qquad\qquad \text{a)}$$

$$\tau_i[z(t)] = Lz(t_R) - f(t_R) \qquad\qquad (2.28)$$

$$+ \sum_{k=1}^{p} (h_i)^k \mathcal{F}_k(A,f) z(t_R) + 0(h_i^{p+1}) \qquad\qquad \text{b)}$$

where $z(t) \in C^{p+1}[0,1]$ and $t_R = t_i + 0(h_i)$. Also let $\mathcal{F}_k(A(t), f(t))\, z(t) \in C^{\nu}[0,1]$ where $\nu = \min\{p-k,q-k\}$. Then for all nets with $h_o \leq H$, H sufficiently small,

$$v_i = \sum_{k=o}^{q} y_k(t_i)\, h_i^k + 0(h_o^{q+1}) \qquad\qquad (2.29)$$

where $y_o(t) = y(t)$ and $y_k(t)$, $k \geq 1$, is defined by

$$Ly_k(t) + \sum_{m=o}^{k-1} \mathcal{F}_{k-m}(A,f)y_m(t) = o \qquad \text{a)}$$

$$(2.30)$$

$$B_oy_k(o) + B_1y_k(1) = o \; . \qquad \text{b)}$$

Proof: By induction and the continuity of $\mathcal{F}_k(A,f)$, $A(t)$, and $f(t)$, it can be shown that

$$y_k(t) \in \mathcal{C}^{p-k+1}[0,1] \; \cdot$$

Immediately upon substituting

$$w(t) = y_o(t) + h_iy_1(t) + \ldots + (h_i)^q y_q(t)$$

into the truncation error expansion (2.28b), we get

$$\tau_i[w(t)] = \sum_{k=1}^{q} h_i^k \{Ly_k(t_R) + \sum_{m=o}^{k-1} \mathcal{F}_m(A,f)y_m(t_R)\} + 0(h_i^{q+1}).$$

Since $y_k(t)$ satisfy (2.30a), we have immediately

$$\tau_i[w(t)] = 0(h_i^{q+1})$$

and hence

$$L_h[v_i-w(t_i)] = 0(h_i^{q+1}) \; .$$

The boundary conditions and $\tau_o[w(t)]$ clearly yield

$$B_o[v_o-w(t_o)] + B_1[v_J-w(t_J)] = o \; .$$

By hypothesis, the difference scheme is stable, thus

$$\max_i \|v_i - z(t_i)\| \leq Mh_o^{q+1}, \qquad h_o \leq H.$$

Or equivalently,

$$v_i = y(t_i) + h_i y_1(t_i) + \ldots + h_i^q y_q(t_i) + O(h_o^{q+1}) \quad .$$

▨

The expression (2.29) for $v_i - y(t_i)$ is termed the asymptotic error expansion. In the following discussion we assume that $y(t) \in C^\infty[0,1]$, $A(t) \in C^\infty[0,1]$, and $f(t) \in C^\infty[0,1]$. Suppose that the truncation error $\tau_i[z(t)]$ has only even powers of h_i, that is, $\mathcal{F}_{2k+1} = o$, $k=0,1,\ldots.$ The function $y_1(t)$ is defined by

$$Ly_1(t) = o$$

$$B_o y_1(o) + B_1 y_1(1) = o.$$

Since the differential equation (1.1a,b) has a unique solution, $y_1(t) \equiv o$.

Assume that $y_{2k-1}(t) \equiv o$, $k=1,2,\ldots,\nu$, then $y_{2\nu+1}(t)$ is defined by

$$Ly_{2\nu+1}(t) + \sum_{k=o}^{2\nu} \mathcal{F}_{2\nu+1-k}(A,f)y_k(t) = o$$

$$B_o y_{2\nu+1}(o) + B_1 y_{2\nu+1}(1) = o. \tag{2.31}$$

For k even, $\mathcal{F}_{2\nu+1-k} \equiv o$, thus

$$Ly_{2\nu+1}(t) + \sum_{k=o}^{\nu-1} \mathcal{F}_{2(\nu-k)}(A,f)y_{2k+1}(t) = o .$$

But our induction assumption is that $y_{2k+1}(t) = o$, $k=0,2,\ldots,\nu-1$, thus $y_{2\nu+1}(t)$ is defined by

$$Ly_{2\nu+1}(t) = o$$

$$B_o y_{2\nu+1}(o) + B_1 y_{2\nu+1}(1) = o .$$

Clearly, then $y_{2k+1}(t) = o$, $k=0,1,2,\ldots$. This result indicates that if the truncation error contains only even powers of h_i, then the asymptotic error expansion will have the same property.

For example, the truncation error expansion for Gap_6 (2.19) is

$$\tau_i[z(t)] = Lz(t_R) - f(t_R)$$

$$+ \sum_{L=1}^{\infty} \left(\frac{h_i}{2}\right)^{2L} \left[\sum_{m=o}^{3} \frac{\alpha_m^3}{2L+1-m)!} \Big(C_m(t_R)z(t_R) + g_m(t_R) \Big)^{(2L+1-m)} \right] \quad (2.32)$$

where $t_R = t_{i-1/2}$. Putting this in the form of Theorem 2.27, we have

$$\mathcal{F}_{2k}(A,f)z(t_R) = \sum_{m=o}^{3} \frac{\alpha_m^3}{(2L+1-m)!} \Big(C_m(t_R)z(t_R)+g_m(t_R) \Big)^{(2L+1-m)} .$$

The previous discussion seems to indicate that the asymptotic error expansion of Gap_6 scheme is

$$v_i - y(t_i) = \left(\frac{h_i}{2}\right)^2 y_2(t_i) + O(h^4)$$

because the h_i^2 and h_i^4 terms are nonvanishing in (2.32). However, closer examination of $\mathcal{F}_2$ and $\mathcal{F}_4$ yields the following identities:

$$\mathcal{F}_2 z = \frac{1}{6}\left\{\frac{d^2}{dt^2}[Lz(t_R)-f(t_R)] - 2A\frac{d}{dt}[Lz(t_R)-f(t_R)] + \frac{2}{5}(A^2-4A^{(1)})[Lz(t_R)-f(t_R)]\right\}. \tag{2.33}$$

$$\begin{aligned}\mathcal{F}_4 z = \frac{1}{120}\Bigg\{&\frac{d^4}{dt^4}[Lz(t_R)-f(t_R)] - 4A\frac{d^3}{dt^3}[Lz(t_R)-f(t_R)]\\ &+ 4(A^2-2A^{(1)})\frac{d^2}{dt^2}[Lz(t_R)-f(t_R)]\\ &- 4(A^{(2)}-2AA^{(1)}-2A^{(1)}A)\frac{d}{dt}[Lz(t_R)-f(t_R)]\\ &+ 4(A^{(2)}A+AA^{(2)}+2A^{(1)}A^{(1)})[Lz(t_R)-f(t_R)]\Bigg\}.\end{aligned} \tag{2.34}$$

Thus, $y_2(t)$ is defined by

$$Ly_2(t) + \mathcal{F}_2 y(t) = o \ .$$

However, from (2.33) it can be seen that

$$\mathcal{F}_2 y(t) = o \ ,$$

thus, $y_2(t) = o$. In a similar fashion we can show that $y_4(t) = o$.

Theorem (2.27) and the identities (2.33) and (2.34) show that for the Gap_6 difference scheme

$$v_i - y(t_i) = \left(\frac{h_i}{2}\right)^6 y_6(t_i) + 0\left(\left[\frac{h_o}{2}\right]^8\right) \qquad \text{a)} \qquad (2.35)$$

where $y_6(t)$ is defined by

$$\frac{dy_6(t)}{dt} - A(t)y_6(t) + F_6(A,f)y(t) = o$$

$$\text{b)} \qquad (2.35)$$

$$B_o y_6(o) + B_1 y_6(1) = o\ .$$

7. <u>Increased Accuracy</u> We will consider two methods of increasing the accuracy of numerical solutions to the differential equation (1.1a,b). These methods are, by name, Richardson h → o extrapolation and Fox's method of Deferred Corrections (Fox [2], Pereyra [9]). Both methods rely on the entire truncation error series rather than the order of accuracy. Richardson extrapolation employs a solution of the discrete problem on two different nets to eliminate successive error terms; the method of Deferred Corrections uses a solution to the discrete problem to approximate the first truncation error term.

To illustrate the method of Deferred Corrections we look at the centered-Euler method (uniform mesh)

$$\frac{1}{h}[y(t_i)-y(t_{i-1})] - \frac{1}{2} A(t_{i-1/2})[y(t_i)+y(t_{i-1})] - f(t_{i-1/2})$$

$$(2.36)$$

$$= \frac{1}{2}\left(\frac{h}{2}\right)^2 \left[\frac{1}{3} y^{(3)}(t_{i-1/2}) - A(t_{i-1/2})y^{(2)}(t_{i-1/2})\right] + 0\left(\left(\frac{h}{2}\right)^4\right).$$

Define $\{v_i^o\}$ by

$$\mathcal{B}_h V^o - r = o \tag{2.37}$$

where the difference scheme used is centered-Euler. The equation (2.37) may be "corrected" by including the $O(h^2)$ error term in (2.36). For example,

$$\frac{1}{2}\left(\frac{h}{2}\right)^2\left[\frac{1}{3}y^{(3)}(t_{i-1/2}) - A(t_{i-1/2})y^{(2)}(t_{i-1/2})\right] =$$

$$\frac{1}{48}\left\{[A_{i+1}-3A_{i-1/2}]y_{i+1} - [A_i-3A_{i-1/2}]y_i - [A_{i-1}-3A_{i-1/2}]y_{i-1}\right.$$

$$\left. + [A_{i-2}-3A_{i-1/2}]y_{i-2}\right\} + \frac{1}{48}\left\{f_{i+1}-f_i-f_{i-1}+f_{i-2}\right\} + O(h^2)$$

where $f_i = f(t_i)$, $A_i = A(t_i)$, etc.. Employing this relation, a first correction to V_o can be defined by

$$\frac{1}{h}[v_i^1-v_{i-1}^1] - \frac{1}{2}A_{i-1/2}[v_i^1+v_{i-1}^1] - f_{i-1/2}$$

$$- \frac{1}{48}\left\{[A_{i+1}-3A_{i-1/2}]v_{i+1}^o-[A_i-3A_{i-1/2}]v_i^o-[A_{i-1}-3A_{i-1/2}]v_{i-1}^o\right.$$

$$\left. + [A_{i-2}-3A_{i-1/2}]v_{i-2}^o\right\} - \frac{1}{48}\left\{f_{i+1}-f_i-f_{i-1}+f_{i-2}\right\} = 0. \tag{2.38}$$

The sequence of operations (2.37), (2.38) yields a fourth-order accurate solution to (1.1a,b). We note that this procedure employs the same basic idea that was used to generate the fourth-order scheme (2.5). The method of Deferred Corrections has several disadvantages. At the boundary t=0, the quantity v_{-1}^{o} must be defined and thus we have to examine the solution outside the domain of interest. Also, the solution procedure for V^{o} is significantly different than (2.38) where the approximation to the first truncation error term is included. Evaluation of higher-order derivatives of y(t) is at best a delicate procedure and must be done with the utmost care.

Richardson extrapolation avoids these particular problems. Once two solutions are calculated, each on a different net, the power structure of the asymptotic error expansion is used to eliminate the leading error term at those points common to both nets. For example, the sixth-order gap scheme has an asymptotic error expansion of the form

$$v_i = y(t_i) + \left(\frac{h_i}{2}\right)^6 y_6(t_i) + 0(h^8) .$$

If we solve the discrete problem twice for $\{v_i(1)\}$ and $\{v_i(2)\}$, then

$$v_i(1) = y(t_i) + \left(\frac{h_i(1)}{2}\right)^6 y_6(t_i) + 0(h^8) ,$$

$$v_i(2) = y(t_i) + \left(\frac{h_i(2)}{2}\right)^6 y_6(t_i) + 0(h^8) \quad .$$

Let $t_i(1) = t_j(2)$, where $t_i(1)$ is the i^{th} point of the first net and $t_j(2)$ is the j^{th} point of the second net.

Define
$$w_i = \frac{h_i^6(1)v_j(2) - h_j^6(2)v_i(1)}{h_i^6(1) - h_j^6(2)}$$

then

$$w_i = y(t_i) + 0(h^8).$$

This elimination procedure may, of course, be carried on as long as is practical, that is, until the continuity of y(t) or round-off errors make further refinement unproductive.

Chapter III

Parallel Shooting

This chapter examines the relationship between implicit finite difference procedures and discrete initial-value techniques. The proof of Theorem 1.16 yields the following result: If both the boundary-value procedure

$$\mathcal{B}_h\, V^{BV} - r = o \tag{3.1}$$

and the initial-value procedure

$$\mathcal{I}_h[Z \,\vdots\, z] = \left[\begin{array}{c|c} I & o \\ 0 & r_1 \\ \vdots & \vdots \\ 0 & r_J \end{array}\right] \qquad \text{a)}$$

$$(B_o + B_1 Z_J)w = \beta - B_1 z_J \qquad \text{b)} \qquad (3.2)$$

$$V_i^{IV} = Z_i\, w + z_i \qquad \text{c)}$$

have unique solutions, then $V_i^{IV} = V_i^{BV}$, $i = o,\dots,J$. That is, the procedure (3.2a,b,c) is actually a method of solving the equations (3.1). The main result of this chapter states that any parallel shooting technique is also equivalent to (3.1) and, thus, to (3.2a,b,c).

Parallel shooting includes such procedures as simple shooting, Method of Complementary Functions, and the Godunov-Conte orthogonalization procedure (for example, see Section 2 on the Method of Complementary Functions). We include a separate treatment of the Method of Complementary Functions in order to illustrate the important case of separated boundary conditions. For completeness the Method of Adjoints will be discussed even though it has no simple relationship to (3.1).

1. Parallel Shooting

The boundary-value problem of interest is

$$\begin{aligned} &\frac{du}{dt} - A(t)u - f(t) = o \qquad t \in [o,1] \qquad &a)\\ &&(3.3)\\ &B_o\, u(o) + B_1\, u(1) - \beta = o. &b)\end{aligned}$$

The simple shooting method (Theorem 1.2) uses the solution of (n+1) initial-value problems to generate the solution of (3.3a,b). In parallel shooting, the interval [o,1] is divided into κ nonoverlapping subintervals, $[T_\nu, T_{\nu+1}]$, and separate initial-value problems are solved on each subinterval. The solution of (3.3a,b) is constructed by requiring continuity at T_ν, $\nu = 1,\dots,\kappa-1$, and satisfaction of the boundary conditions (3.3b).

Discrete parallel shooting methods mimic this procedure. We place a net $\{t_i^\nu\}_{i=o}^{J_\nu}$ on each subinterval $[T_{\nu-1}, T_\nu]$. On this net, a discrete initial-value problem is solved with initial conditions Z_o^ν, z_o^ν. These n+1 initial conditions are arbitrary except that we require the n × n matrix Z_o^ν be nonsingular. This initial-value solution can be represented as

$$\begin{bmatrix} I & 0 & \cdots & 0 \\ M^{\nu}_{1o} & M^{\nu}_{11} & \cdots & M^{\nu}_{1J_\nu} \\ \vdots & \vdots & & \\ M^{\nu}_{J_\nu o} & M^{\nu}_{J_\nu 1} & \cdots & M^{\nu}_{J_\nu J_\nu} \end{bmatrix} \left[\begin{array}{c|c} Z^{\nu}_{o} & z^{\nu}_{o} \\ Z^{\nu}_{1} & z^{\nu}_{1} \\ \vdots & \vdots \\ Z^{\nu}_{J_\nu} & z^{\nu}_{J_\nu} \end{array}\right] = \left[\begin{array}{c|c} Z^{\nu}_{o} & z^{\nu}_{o} \\ 0 & r^{\nu}_{1} \\ \vdots & \vdots \\ 0 & r^{\nu}_{J_\nu} \end{array}\right] \tag{3.4}$$

or, equivalently we write

$$I^{\nu}_{h}\,[Z^{\nu}\,|\,z^{\nu}] = [R^{\nu}\,|\,r^{\nu}]. \tag{3.5}$$

In parallel shooting, it is sometimes necessary to state an initial-value problem on some interval $[T_\nu, T_{\nu+1}]$ backward by formally giving n+1 "boundary" conditions at $T_{\nu+1}$ rather than "initial" conditions at T_ν. However, Theorem 1.16 states that for all nets with $h_o \le H$, H sufficiently small, the solution to this backward problem is equal to the solution of (3.4) with Z^{ν}_{o} and z^{ν}_{o} appropriately chosen. Thus, in order to examine the most general parallel shooting methods, it is sufficient to consider discrete initial-value problems of the form (3.4).

Defining Z^{ν}, z^{ν} by (3.4), the solution of any initial-value problem on the net $\{t^{\nu}_{i}\}^{J_\nu}_{i=o}$ is given by

$$v^{\nu}_{i} = Z^{\nu}_{i}\, w^{\nu} + z^{\nu}_{i} \qquad \nu = 1,\ldots,\kappa\ . \tag{3.6}$$

Thus, the continuity requirements for parallel shooting are

$$v^{\nu}_{J_\nu} = v^{\nu+1}_{o} \qquad \nu = 1,\ldots,\kappa-1\ . \tag{3.7}$$

In order to meet the boundary conditions, the following relationship must be satisfied

$$B_o \, v_o^1 + B_1 \, v_{J_\kappa}^\kappa - \beta = o. \tag{3.8}$$

Combining (3.5), (3.6), (3.7), and (3.8), the parallel shooting solution, V^{PS}, is given by

$$\mathcal{I}_h^\nu [Z^\nu \mid z^\nu] = [R^\nu \mid r^\nu] \qquad \nu = 1,\dots,\kappa \qquad \text{a)}$$

$$\begin{bmatrix} B_o Z_o^1 & 0 & \dots & 0 & B_1 Z_{J_\kappa}^\kappa \\ -Z_{J_1}^1 & Z_o^2 & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & -Z_{J_{\kappa-1}}^{\kappa-1} & Z_o^\kappa \end{bmatrix} \begin{bmatrix} w^1 \\ w^2 \\ \vdots \\ w^\kappa \end{bmatrix} = \begin{bmatrix} \beta - B_o z_o^1 - B_1 z_{J_\kappa}^\kappa \\ z_{J_1}^1 - z_o^2 \\ \vdots \\ z_{J_{\kappa-1}}^{\kappa-1} - z_o^\kappa \end{bmatrix} \qquad \text{b)} \qquad (3.9)$$

$$v_i^{PS} = Z_{i-S_{\nu-1}}^\nu \, w^\nu + z_{i-S_{\nu-1}}^\nu \qquad i = o,\dots,J$$

$$S_\nu = \sum_{i=1}^{\nu} J_i , \qquad J = S_\kappa . \qquad \text{c)}$$

<u>Stability</u>. For the purposes of this discussion, the discrete parallel initial value problem is said to be stable if

$$\|(\mathcal{I}_h^\nu)^{-1}\| \le \overline{K} \qquad \nu = 1,\dots,\kappa .$$

In the proofs to follow, define the $nJ_\nu \times n(J_\nu+1)$ matrix M^ν via the relationship

$$\mathcal{I}_h^\nu = \begin{bmatrix} I & 0 & \dots & 0 \\ & & M^\nu & \end{bmatrix} . \tag{3.10}$$

The initial-value matrix I_h associated with a parallel shooting procedure is

$$I_h = \begin{bmatrix} I & & & & \\ M^1_{10} \cdots & & & & \\ \vdots & \vdots & & & \\ \cdots & M_{J_1 J_1} & & & \\ & M^2_{10} & \cdots & & \\ & \vdots & \vdots & & \\ & \cdots & M^2_{J_2 J_2} & & \\ & & & \ddots & \end{bmatrix}. \qquad (3.11)$$

For convenience, we will write this matrix as ($\kappa = 3$)

$$I_h = \left[\begin{array}{c} I^1_h \\ \hline M_2 \\ \hline M^3 \end{array}\right] \qquad (3.12)$$

where the (1,o) block element of M^ν is the $(S_\nu+2,\ S_\nu+1)$ block element of I_h.

Theorem 3.13. Let (3.3a,b) have a unique solution $y(t) \in C^1[o,1]$. Let the difference scheme be consistent with (3.3a). Then the following are equivalent:

a) The discrete initial-value problem is stable .

b) The discrete parallel initial value problem is stable.

Proof: b) $\Rightarrow$ a) The matrix I_h can be factored in the following way

$$\mathcal{I}_h = \left[\begin{array}{c|c} \mathcal{I}_h^1 & \\ \hline & I \end{array}\right] \cdots \left[\begin{array}{c|c} I & \\ \hline & \mathcal{I}_h^\kappa \end{array}\right].$$

By hypothesis $||(\mathcal{I}_h^\nu)^{-1}|| \le \overline{K}$, $\nu = 1,\ldots,\kappa$, then

$$||(\mathcal{I}_h)^{-1}|| \le \max\,\{\overline{K}^\kappa, 1\}.$$

a) $\Longrightarrow$ b) Without loss of generality, we will consider the case in which the interval [o,1] is divided into 3 subintervals, $\kappa = 3$. The first step is to show that the matrix $\overline{\mathcal{I}}_h$, defined below, is nonsingular.

$$\overline{\mathcal{I}}_h = \left[\begin{array}{c} M_1 \\ \hline \mathcal{I}_h^2 \\ \hline M^3 \end{array}\right]$$

$$= \left[\begin{array}{cccccc} M^1_{1o} & \cdots & \cdot & & & \\ \vdots & & \vdots & & & \\ \cdot & \cdots & M^1_{J_1J_1} & & & \\ & & I & \cdots & 0 & \\ & & M^2_{1o} & \cdots & \cdot & \\ & & \vdots & & \vdots & \\ & & \cdot & \cdots & M^2_{J_2J_2} & \\ & & & & M^3_{1o} & \cdots \\ & & & & \vdots & \vdots \\ & & & & \cdot\ \cdots & M_{J_3J_3} \end{array}\right].$$

By using the Factorization Lemma (1.12), the Reducing Lemma (1.13), and the convergence of the initial-value problem, it can be shown that

$$||\,(\bar{I}_h)^{-1}|| \le K_1 \tag{3.14}$$

where K_1 is independent of h_o.

Let V^1, V^2, V^3 be nJ_1, $n(J_2+1)$, nJ_3-vectors respectively. Consider the $n(J+1)$ equations

$$\left[\begin{array}{ccc} M^1 & & \\ \hline & I_h^2 & \\ \hline & & M_3 \end{array}\right] \left[\begin{array}{c} V^1 \\ \hline V^2 \\ \hline V^3 \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline g^2 \\ \hline 0 \end{array}\right] \tag{3.15}$$

where g^2 is any $n(J_2+1)$-vector. Since $\bar{I}_h$ is nonsingular, for each g^2 there is a V^2 such that

$$I_h^2\, V^2 = g^2 \;.$$

Recalling (3.14), we find that

$$||V_2|| \le \left|\left| \left[\begin{array}{c} V_1 \\ \hline V_2 \\ \hline V_3 \end{array}\right] \right|\right| \le K_1 ||g^2|| \;. \tag{3.16}$$

Thus,

$$||(I_h^2)^{-1}|| \le K_1 .$$

This argument is valid for any $\nu = 1,\ldots,\kappa$, thus we have

$$||(I_h^\nu)^{-1}|| \le K_1 \qquad \nu = 1,\ldots,\kappa \;.$$

▨

Before proceeding with the theorem which will tie all three of these discrete methods together, it is convenient to present two lemmas.

Lemma 3.17. Let C_1, C_2 be nonsingular $m \times m$ matrices. Let C_1 and C_2 be related by

$$C_1 = C_2 + C_2 LN \tag{3.18}$$

where L, N are $m \times q$, $q \times m$ matrices respectively. Further, let L have rank $q \le m$. Then the $q \times q$ matrix $[I + NL]$ is nonsingular.

Proof. By hypothesis C_1 is nonsingular, therefore for each $b \in E^q$, there exists a unique m-vector, x, such that

$$C_1 x = C_2 Lb \quad .$$

Recalling (3.18), this equation can be rewritten as

$$\left[C_2 + C_2 LN\right] x = C_2 Lb$$

or equivalently as

$$(I + LN)x = Lb \ . \tag{3.19}$$

The Reducing Lemma (1.13) states that (3.19) only has solutions of the form

$$x = Lw$$

where

$$(I + NL)w = b.$$

Since a unique x, solution of (3.19), exists for every b, the matrix (I + NL) must be nonsingular.

▨

Lemma 3.20. Let the $n(J+\kappa) \times n(J+\kappa)$ expanded boundary-value matrix $\mathcal{B}_h^E$ be defined by

$$\mathcal{B}_h^E = \begin{bmatrix} B_o & 0 & & & \cdots & & & & B_1 \\ M^1_{1o} & \cdots & \vdots & & & & & & \\ \vdots & & \vdots & & & & & & \\ \cdots & & M^1_{J_1 J_1} & & & & & & \\ & & -I & I & & & & & \\ & & & M^2_{1o} & \cdots & \vdots & & & \\ & & & \vdots & & \vdots & & & \\ & & & \cdots & & M^2_{J_2 J_2} & & & \\ & & & & & -I & I & & \\ & & & & & & & \ddots & \end{bmatrix} . \qquad (3.21)$$

Let the boundary-value matrix $\mathcal{B}_h$ be nonsingular. Then $\mathcal{B}_h^E$ is nonsingular.

Proof. Suppose that $\mathcal{B}_h^E$ is singular, then there is a $n(J+\kappa)$-vector V^E such that

$$\mathcal{B}_h^E \, V^E = o \qquad ||V^E|| = 1 .$$

However, if we delete the κ-1 elements v_i^E, $i = S_\nu+\nu$, $\nu = 1,\ldots,\kappa-1$, from V^E and collapse V^E to form a $n(J+1)$-vector, V, then

$$\mathcal{B}_h V = o.$$

By hypothesis $\mathcal{B}_h$ is nonsingular, thus $||V|| = o$. Note that for each element v_i^E deleted from V^E to form V

$$v_i^E = v_{i-1}^E$$

and v_{i-1}^E appears as an element of V. Thus,

$$||V^E|| = o$$

and by contradiction, we have $\mathcal{B}_h^E$ nonsingular.

▨

Now, we state and prove the main theorem of this Chapter, which will include Theorem 1.16.

Theorem 3.22. Let (3.3a,b) have a unique solution $y(t) \in C^1[o,1]$. Let the difference scheme employed be consistent with (3.3a). Denote by BV, IV, and PS the following methods:

BV. Discrete boundary-value procedure, (3.1) ,

IV. Discrete initial-value procedure, (3.2a,b,c),

PS. Discrete parallel shooting procedure, (3.9a,b,c).

Let one of the discrete problems (3.1), (3.2a), and (3.9a) be stable.

Then, for all nets with $h_o \leq H$, H sufficiently small,

i) Each method uniquely defines an $n(J+1)$ vector as an approximation to $y(t)$ on that net, and

ii) This approximation is the same for each method.

Proof: Note that stability of (3.1) is equivalent to (3.2a) or (3.9a).

Re i). Theorem 1.16 proves that methods BV and IV uniquely define approximations to $y(t)$. To show that method PS uniquely defines an approximation it is sufficient to show that the $n\kappa \times n\kappa$ matrix in (3.9b) is nonsingular. Without loss of generality, we consider the case where $\kappa = 3$. Thus, we want to show that the $3n \times 3n$ matrix P, defined by

$$P \equiv \begin{bmatrix} B_o Z_o^1 & 0 & B_1 Z_{J_3}^3 \\ -Z_{J_1}^1 & Z_o^2 & 0 \\ 0 & -Z_{J_2}^2 & Z_o^3 \end{bmatrix},$$

is nonsingular. From conclusion i), $\mathcal{B}_h$ is nonsingular, hence by Lemma 3.20, $\mathcal{B}_h^E$ is nonsingular. Recall the $n(J+1)$-vector r in (3.1a), and define the $n(J+\kappa)$-vector r^E by

$$(r^E) = [\beta r_1 \ldots r_{S_1}\ o\ r_{S_1+1} \ldots r_{S_2}\ o\ r_{S_2+1} \ldots r_J]^T.$$

Consider the matrix equation

$$\mathcal{B}_h^E\, V^E = r^E. \tag{3.23}$$

Define the expanded parallel shooting matrix $\mathcal{I}_h^E$ by

$$\mathcal{I}_h^E = \left[\begin{array}{c:c:c} \mathcal{I}_h^1 & & \\ \hdashline & \mathcal{I}_h^2 & \\ \hdashline & & \mathcal{I}_h^3 \end{array}\right].$$

Recalling the parallel shooting method (3.9a,b,c), equation (3.23) can be written as

$$I_h^E\, V^E = I_h^E (I_h^E)^{-1} \bar{L}[N\, V^E + \xi] \tag{3.24}$$

where

$$\bar{L} = \left[\begin{array}{c|c|c|c} R^1 & 0 & 0 & r^1 \\ \hline 0 & R^2 & 0 & r^2 \\ \hline 0 & 0 & R^3 & r^3 \end{array}\right] \quad , \quad \xi = \left[\begin{array}{c} (Z_o^1)^{-1}\ (\beta - z_o^1) \\ \hline (Z_o^2)^{-1}\ \ z_o^2 \\ \hline (Z_o^3)^{-1}\ \ z_o^3 \\ \hline 1 \end{array}\right],$$

and

$$N = \left[\begin{array}{cccc|ccc|ccc} (Z_o^1)^{-1}(I - B_o) & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & -(Z_o^1)^{-1}B_1 \\ \hline 0 & & \cdots & (Z_o^2)^{-1} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \hline 0 & & \cdots & 0 & 0 & \cdots & (Z_o^3)^{-1} & 0 & \cdots & 0 \\ \hline o & & \cdots & o & o & \cdots & o & o & \cdots & o \end{array}\right] .$$

In general, $\bar{L}$ is a $n(J+\kappa)$x $n\kappa + 1$ matrix, ξ is an $(n\kappa + 1)$-vector, and N is a $n\kappa + 1 \times n(J+\kappa)$ matrix. Recalling Lemma 3.17, we note that

$$(I - NL)$$

is nonsingular, where

$$L = (I_h^E)^{-1}\bar{L} = \left[\begin{array}{c|c|c|c} Z^1 & 0 & 0 & z^1 \\ \hline 0 & Z^2 & 0 & z^2 \\ \hline 0 & 0 & Z^3 & z^3 \end{array}\right] . \tag{3.25}$$

When the proper substitutions are made, it is clear that (I - NL) is nonsingular if and only if P is nonsingular.

Re ii). Theorem 1.16 is sufficient to prove that

$V^{BV} = V^{IV}$. Recalling (3.23), the Reducing Lemma (1.13) states that

$$V^E = L \left[\begin{array}{c} w_1 \\ \hline w_2 \\ \hline w_3 \\ \hline 1 \end{array}\right]$$

where the n-vectors, w_i, must satisfy

$$P \left[\begin{array}{c} w_1 \\ \hline w_2 \\ \hline w_3 \end{array}\right] = \left[\begin{array}{c} \beta - B_o z_o^1 - B_1 z_{J_3}^3 \\ \hline z_{J_1}^1 - z_2^o \\ \hline z_{J_2}^2 - z_3^o \end{array}\right] .$$

Now, we have the relationship

$$v_i^{PS} = \begin{cases} v_i^E \ , & 0 \le i \le J_1 \\ v_{i+1}^E , & J_1 + 1 \le i \le S_2 \\ v_{i+2}^E , & S_2 + 1 \le i \le S_3 \end{cases}$$

Recalling (3.23) and the definition of $\mathcal{B}_h^E$, we note that

$$\mathcal{B}_h \begin{bmatrix} v_o^E \\ \vdots \\ v_{J_1}^E \\ v_{J_1+2}^E \\ \vdots \\ v_{J_1+J_2+1}^E \\ v_{J_1+J_2+3}^E \\ \vdots \\ v_{J_1+J_2+J_3+2}^E \end{bmatrix} = r$$

and thus it is clear that

$$V^{PS} = V^{BV} = V^{IV} .$$

▨

Unfortunately, Theorem (3.22) contains no information on the numerical stability of any parallel shooting technique. To say that a unique solution to the difference equations (3.9a,b,c) exists is not to say that it can be accurately approximated numerically. It is well-known that a simple shooting method may produce disastrous numerical solutions if the linearly independent solutions of (3.3a) grow at different rates. For example, consider the problem

$$\frac{du}{dt} - \begin{pmatrix} 60 & 0 \\ 0 & -60 \end{pmatrix} u = f(t)$$

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} u(o) + \begin{pmatrix} 1 & o \\ 1 & 1 \end{pmatrix} u(1) = \beta .$$

This problem has a unique solution for all $f(t) \in C[o,1]$ and $\beta \in E^2$, because the matrix

$$[B_o + B_1X(1)] = \begin{bmatrix} 1 + e^{60} & 1 \\ 1 + e^{60} & 1 + e^{-60} \end{bmatrix} \tag{3.26}$$

is nonsingular. However, if we try to invert this matrix numerically, on a machine with less than 85-bit accuracy, the computer will consider $(B_o + B_1X(1))$ singular, because $1 + e^{-60}$ will be represented as 1. Thus, no matter how close Z_J gets to $X(1)$, this matrix will be numerically singular due to round-off error. For this reason, methods which are equivalent in the sense of Theorem 3.22 may yield greatly different numerical solutions.

2. Method of Complementary Functions

The Method of Complementary Functions (Goodman and Lance [3]) is used to approximate the solution of (3.3a,b) when the boundary conditions are separated. That is, the boundary conditions (3.3b) can be written as

$$\begin{bmatrix} (B_o) \\ \hline [0] \end{bmatrix} u(o) + \begin{bmatrix} (0) \\ \hline [B_1] \end{bmatrix} u(1) = \begin{bmatrix} (\beta_o) \\ \hline [\beta_1] \end{bmatrix}$$

where B_o is a p × n matrix of rank p, B_1 is a q × n matrix of rank q, $p + q = n$, and β_o, β_1 are a p-vector, q-vector respectively. In this section, we use the convention that a single matrix in parentheses, (B_o), has p rows and a single matrix in brackets, $[B_1]$, has q rows. The original equations (3.3a,b) can now be rewritten as

$$\frac{du(t)}{dt} - A(t)\ u(t) - f(t) = o \qquad a)$$
$$(B_o)\ u(o) = (\beta_o) \qquad [B_1]\ u(1) = [\beta_1] \ , \qquad b) \qquad (3.27)$$

In what follows, we assume that the boundary-value problem (3.27a,b) has a unique solution y(t). If the boundary-value problem has a unique solution, then there exists a vector η_o such that

$$(B_o)\eta_o = (\beta_o)$$

where η_o is orthogonal to the null space of (B_o). Let the orthonormal set $\{\eta_i\}_{i=1}^q$ span the null space of (B_o), then

$$(\eta_i, \eta_o) = o \qquad 1 \le i \le q .$$

With these notions in hand, the solution of (3.27a,b) can be characterized in the following way. Solve the q homogeneous initial-value problems

$$\frac{dx^i(t)}{dt} - A(t)\, x^i(t) = o \qquad a)$$

$$x^i(o) = \eta_i \qquad b) \qquad (3.28)$$

and the inhomogeneous problem

$$\frac{dx^o(t)}{dt} - A(t)\, x^o(t) - f(t) = o \qquad a)$$

$$x^o(o) = \eta_o \ . \qquad b) \qquad (3.29)$$

The unique solution of (3.27a,b) is given by

$$y(t) = \sum_{i=1}^{q} \alpha_i\, x^i(t) + x^o(t) \qquad (3.30)$$

where the constants α_i, $i = 1,\ldots,q$, are determined from

$$[B_1] \sum_{i=1}^{q} \alpha_i\, x^i(1) + [B_1]\, x^o(1) = [\beta_1] \ . \qquad (3.31)$$

In the introduction to this chapter, it was stated that the Method of Complementary Functions is a special case of parallel shooting. This is the best point at which to make this relationship clear. Parallel shooting naturally contains the case of simple shooting. In order to characterize y(t) by means of simple shooting, solve the n homogeneous initial-value problems

$$\frac{d}{dt} X(t) = A(t) X(t) \qquad a)$$

$$X(o) = C \qquad b) \qquad (3.32)$$

and the inhomogeneous problem

$$\frac{dx(t)}{dt} = A(t) x(t) + f(t) \qquad a)$$

$$x(o) = c \qquad b) \qquad (3.33)$$

where C is a nonsingular n × n matrix and c is an n-vector. Now,

$$y(t) = X(t) \bar{\alpha} + x(t) \qquad (3.34)$$

where the n-vector $\bar{\alpha}$ is determined by

$$\left\{ \begin{bmatrix} (B_o) \\ \hline [0] \end{bmatrix} X(o) + \begin{bmatrix} (0) \\ \hline [B_1] \end{bmatrix} X(1) \right\} \bar{\alpha} =$$

$$\begin{bmatrix} (\beta_o) \\ \hline [\beta_1] \end{bmatrix} - \begin{bmatrix} (B_o) \\ \hline [0] \end{bmatrix} x(o) - \begin{bmatrix} (0) \\ \hline [B_1] \end{bmatrix} x(1) . \qquad (3.35)$$

The result of specializing the simple shooting procedure (3.32a,b), (3.33a,b), (3.34), (3.35) by taking

$$C = \begin{bmatrix} (B_o)^T & \eta_1 \cdots \eta_q \end{bmatrix} , \; c = \eta_o \qquad (3.36)$$

is the Method of Complementary Functions.

The discrete Method of Complementary Functions (hereafter reference to the Method of Complementary Functions will imply the discrete procedure) is derived by solving equations (3.28a,b), (3.29a,b), (3.30), and (3.31) numerically. Solve the q homogeneous discrete initial-value problems

$$\mathcal{I}_h\ z^i = \begin{bmatrix} \eta_i \\ o \\ \vdots \\ o \end{bmatrix} \qquad i = 1,\ldots,q \tag{3.32}$$

and the inhomogeneous discrete initial-value problem

$$\mathcal{I}_h\ z^o = \begin{bmatrix} \eta_o \\ r_1 \\ \vdots \\ r_J \end{bmatrix} . \tag{3.33}$$

For convenience, we write (3.32), (3.33) as

$$\mathcal{I}_h[Z \mid z^o] = \left[\begin{array}{c|c} Z_o & z_o^o \\ 0 & r_1 \\ \vdots & \vdots \\ 0 & r_J \end{array}\right] \tag{3.34}$$

where Z is a (J+1)n × q matrix with its i^{th} column being z^i. The discrete approximation to y(t) is given by

$$v_i^{CF} = Z_i\ w + z_i^o \tag{3.35}$$

where

$$[B_1]Z_J\ w = [\beta_1] - [B_1]z_J^o\ , \tag{3.36}$$

The initial-value equation (3.34) may be formally written as

$$\begin{bmatrix} \begin{matrix}(B_o)\\ [Z_o^T]\end{matrix} & 0 & \cdots & 0 \\ M_{1o} & M_{11} & \cdots & M_{1J} \\ \vdots & \vdots & \ddots & \vdots \\ M_{Jo} & M_{J1} & \cdots & M_{JJ} \end{bmatrix} \left[\begin{array}{c|c} Z_o & z_o^o \\ Z_1 & z_1^o \\ \vdots & \vdots \\ Z_J & z_J^o \end{array}\right] = \left[\begin{array}{c|c} (0) & (\beta_o) \\ [I] & [o] \\ 0 & o \\ \vdots & \vdots \\ 0, & o \end{array}\right] . \qquad (3.37)$$

The first n(q+1) equations of (3.37) state that the columns of Z_o are orthonormal and belong to the null-space of B_o, and also that $B_o\ z_o^o = \beta_o$ and z_o^o is orthogonal to the null-space of B_o. Now, we want to show that V_i^{CF} calculated via (3.37), (3.35), and (3.36) is a solution of the equations

$$\mathcal{B}_h\ V - r = o\ .$$

The discrete boundary-value problem is

$$\begin{bmatrix} \begin{matrix}(B_o)\\ [0]\end{matrix} & o & \cdots & o & \begin{matrix}(0)\\ [B_1]\end{matrix} \\ M_{1o} & M_{11} & \cdots & M_{1J-1} & M_{1J} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ M_{Jo} & M_{J1} & \cdots & M_{J,J-1} & M_{JJ} \end{bmatrix} \begin{bmatrix} v_o^{BV} \\ v_1^{BV} \\ \vdots \\ v_J^{BV} \end{bmatrix} = \begin{bmatrix} \begin{matrix}(\beta_o)\\ [\beta_1]\end{matrix} \\ r_1 \\ \vdots \\ r_J \end{bmatrix} . \qquad (3.38)$$

Recalling the Factorization Lemma (1.12), this matrix can be written as

$$\overline{I}_h\, V^{BV} = \begin{bmatrix} (0) & (\beta_o) \\ [I] & [o] \\ 0 & r_1 \\ \vdots & \vdots \\ 0 & r_J \end{bmatrix} \left\{ \left[\begin{array}{c} [Z_o^T][0]\ldots[-B_1] \\ \hline o \quad o \ldots \quad o \end{array}\right] V^{BV} + \left[\begin{array}{c} [\beta_1] \\ \hline 1 \end{array}\right] \right\}$$

where $\overline{I}_h$ is the n(J+1) × n(J+1) matrix in (3.37). Left-multiplying by $\overline{I}_h^{\,-1}$ and recalling (3.37), this equation is shown to be

$$V^{BV} = [Z \mid z^o] \left\{ \left[\begin{array}{c} [Z_o^T][0]\ldots[-B_1] \\ \hline o \quad o \qquad o \end{array}\right] V^{BV} + \left[\begin{array}{c} [\beta_1] \\ \hline 1 \end{array}\right] \right\}.$$

The Reducing Lemma (1.13) states that

$$V^{BV} = [Z \mid z^o] \left[\begin{array}{c} [w] \\ \hline \lambda \end{array}\right] \tag{3.39}$$

where w and λ are determined from

$$\left\{ I - \left[\begin{array}{c} [Z_o^T][0]\ldots[-B_1] \\ \hline o \quad o \ldots \quad o \end{array}\right] \left[Z \mid z^o \right] \right\} \left[\begin{array}{c} [w] \\ \hline \lambda \end{array}\right] = \left[\begin{array}{c} [\beta_1] \\ \hline 1 \end{array}\right].$$

When this is unraveled, we have

$$B_1\ Z_J\ w = \beta_1 - B_1 z_J^o,\ \lambda = 1\ . \tag{3.40}$$

Recall (3.35) and (3.36) and compare with (3.39) and (3.40) to show that

$$v_i^{BV} = v_i^{CF}\ , \qquad i = o,\ldots,J.$$

3. Method of Adjoints

The Method of Adjoints completes the study of solution procedures for discrete boundary-value problems. This method introduces the adjoint equation which can not be defined via linear transformations of the original problem. For this reason the equivalence exhibited between boundary-value techniques and parallel-shooting procedures is lacking here. A good explanation of the Method of Adjoints is found in Goodman and Lance [3].

The problem considered in [3] is a very special one:

$$\begin{aligned} &\frac{du}{dt} - A(t)u = f(t) && \text{a)} \\ &u_i(o) = \beta_i \qquad i = 1,\ldots,p && \text{b)} \qquad (3.41) \\ &u_i(1) = \beta_i \qquad i = p+1,\ldots,n. && \text{c)} \end{aligned}$$

The adjoint differential equation is

$$\frac{d\bar{u}}{dt} + A^*(t)\bar{u} = o \,. \qquad (3.42)$$

From (3.41), (3.42), the following relationship is derived

$$\frac{d}{dt}\,(\bar{u}^*u) - \bar{u}^*f = o \,. \qquad (3.43)$$

Integrating this expression, we obtain

$$\bar{u}^*(1)u(1) - \bar{u}^*(o)u(o) = \int_o^1 \bar{u}^*(s)f(s)\,ds \,, \qquad (3.44)$$

By solving the adjoint problem n-p times with initial conditions $\bar{u}^\nu(o) = \hat{e}_\nu$, $\nu = p+1,\ldots,n$, the equation (3.44) results in a set of n-p (or q) equations for the quantities $u_i(o)$, $i = p+1,\ldots,n$.

$$\sum_{L=p+1}^{n} \bar{u}_L^{\nu}(o)u_L(o) = \beta_\nu - \sum_{L=1}^{p} \bar{u}_L^{\nu}(o)\beta_L - \int_o^1 \bar{u}^*(s)f(s)ds$$

$$\nu = p+1,\dots,n .$$

Now, the initial conditions u(o) are explicitly known and the solution of a single initial-value problem determines y(t).

If we use the same discretization for (3.41a,b,c) and (3.42), (3.44), it is not in general true that the solution of the discrete Method of Adjoints, V^{MA}, is equal to V^{BV}. However, if we employ the centered-Euler scheme this equivalence does hold. The discrete analogue to (3.44) is naturally

$$\bar{v}_J^* v_J^{MA} - \bar{v}_o^* v_o^{MA} - \sum_{i=1}^{J} \frac{h_i}{2}(\bar{v}_i + \bar{v}_{i-1})^* f(t_{i-\frac{1}{2}}) = o. \quad (3.45)$$

From the difference equations

$$\frac{1}{h_i}(I + \frac{h_i}{2}A^*(t_{i-\frac{1}{2}}))\bar{v}_i - \frac{1}{h_i}(I - \frac{h_i}{2}A^*(t_{i-\frac{1}{2}})\bar{v}_{i-1} = o$$

and

$$\frac{1}{h_i}(I - \frac{h_i}{2}A(t_{i-\frac{1}{2}}))v_i^{BV} - \frac{1}{h_i}(I - \frac{h_i}{2}A(t_{i-\frac{1}{2}})v_i^{BV} = f(t_{i-\frac{1}{2}})$$

the following identity can be derived

$$\frac{1}{h_i}(\bar{v}_i^* v_i^{BV} - \bar{v}_{i-1}^* v_{i-1}^{BV}) = \frac{1}{2}(\bar{v}_i + \bar{v}_{i-1})^* f(t_{i-\frac{1}{2}}). \quad (3.46)$$

Upon summing from i = 1,...,n, this identity yields

$$\bar{v}_J^* v_J^{BV} - \bar{v}_o^* v_o^{BV} = \sum_{i=1}^{J} \frac{h_i}{2}(\bar{v}_i + \bar{v}_{i-1})^* f(t_{i-\frac{1}{2}}) . \quad (3.48)$$

Thus, V^{BV} also satisfies (3.46) and $v_i^{BV} = v_i^{MA}$, i = o,...,J.

Remember that this equivalence has been shown only if the centered-Euler scheme is used to discretize the Method of Adjoints. Identities similar to (3.47) can be derived for any number of other difference schemes, but the summation which yielded (3.48) does not in general show equivalence.

Chapter IV

Nonlinear Two-Point Boundary-Value Problems

This chapter deals with numerical approximations to the solution of nonlinear two-point boundary-value problems of the form

$$\frac{du}{dt} = f(u,t) \qquad t \in [o,1] \qquad \text{a)}$$

$$(4.1)$$

$$b(u(o),u(1)) = o \qquad \text{b)}$$

where u, f, b are n-vectors. After Keller [6], we are interested in isolated solutions of (4.1a,b). The solution of (4.1a,b) is isolated if the linearized boundary-value problem

$$\frac{dw}{dt} = f_1(y(t),t)w \qquad \text{a)}$$

$$(4.2)$$

$$b_1(y(o),y(1))w(o) + b_2(y(o),y(1))w(1) = o \qquad \text{b)}$$

has only the trivial solution. Note that we have used the notation

$$g_i(x_1,\ldots,x_n) = \frac{\partial g(x_1,\ldots,x_n)}{\partial x_i} .$$

When considering the discrete problem, we will also examine Newton's method as a means of calculating an approximation to y(t). The n(J+1)-vector Y is defined to be

$$Y = \begin{bmatrix} y(t_o) \\ \vdots \\ y(t_J) \end{bmatrix}$$

where y(t) is an isolated solution of (4.1a,b). We also define

$$\mathcal{S}_\rho[g(t)] = \{x \mid x \in E^n,\ ||x - g(t)|| \le \rho\},$$

$$S_\rho[g(t)] = \{(x_o,\dots,x_J) \mid x_i \in \mathcal{S}_\rho[g(t_i)],\ o \le i \le J\}.$$

1. Finite Difference Schemes

Each of the difference schemes discussed in Chapter 2 can be modified to make it applicable to nonlinear differential equations of the form (4.1a). Thus, the centered-Euler scheme (2.2) becomes

$$v_i - v_{i-1} = h_i\ f(\frac{1}{2}[v_i + v_{i-1}],\ t_{i-\frac{1}{2}}) \qquad (4.3)$$

for a general nonlinear equation. If equation (4.1a) is linear in u, that is,

$$f(u,t) = A(t)u + g(t),$$

then the difference equation (4.3) reduces to (2.2).

The nonlinear analogue of the integral equation (2.1) is clearly

$$u(t) - u(\tau) - \int_\tau^t f(u(s),s)ds = o\ . \qquad (4.4)$$

As in Chapter 2, quadrature formulae and integration by parts can be used to generate consistent numerical schemes. In this manner the general form of the trapezoidal rule is shown to be

$$v_i - v_{i-1} - \frac{h_i}{2} \quad [f(v_i,t_i) + f(v_{i-1},t_{i-1})] = o \cdot \qquad (4.5)$$

Again note that if f(u,t) is linear in u, then (4.5) reduces to (2.10), which is the trapezoidal rule applied to a linear problem.

The Gap schemes can be derived in the same manner as in Chapter 2. Recalling Lemma 2.11 and Lemma 2.14, it can be shown that

$$y(t_i) - y(t_{i-1}) + \sum_{K=1}^{n} \alpha_K^n \left(\frac{h_i}{2}\right)^K [y^{(K+1)}(t_i) - (-1)^K y^{(K+1)}(t_{i-1})]$$

$$= O(h_i^{2n})$$

where the constants $\{\alpha_K^n\}$ are given by Table 2.17. In order to derive Gap $_4$, we need to evaluate $\frac{d^2y(t)}{dt^2}$. Using the differential equation (4.1a),

$$\frac{d^2y}{dt^2} = f_1(y,t)\ f(y,t) + f_2(y,t) \cdot$$

Thus, we define the nonlinear Gap$_4$ scheme by

$$\frac{1}{h_i} [v_i - v_{i-1}] - \frac{1}{2} [f(v_i,t_i) + f(v_{i-1},t_{i-1})]$$

$$+ \frac{h_i}{12} [f_1(v_i,t_i)\ f(v_i,t_i) + f_2(v_i,t_i) - f_1(v_{i-1},t_{i-1})\ f(v_{i-1},t_{i-1})$$

$$- f_2(v_{i-1},t_{i-1})] = o \ . \qquad (4.6)$$

This is a fourth-order accurate scheme if $y(t) \in C^5[o,1]$.

Coupled with any of these difference schemes, the boundary conditions (4.1b) are approximated by $b(v_o,v_J) = o$. Employing the trapezoidal rule, an approximation to the solution of (4.1a,b) can be determined from the following set of $n(J+1)$ equations:

$$b(v_o,v_J) = o$$

$$N_h v_i = \frac{1}{h_i} [v_i - v_{i-1}] - \frac{1}{2} [f(v_i,t_i) + f(v_{i-1},t_{i-1})] = o \quad 1 \le i \le J . \tag{4.7}$$

More concisely we write

$$b(v_o,v_J) = o \qquad a)$$

$$N_h v_i = o \qquad b) \tag{4.8}$$

where $N_h v_i = o$ may represent any difference approximation of the differential equation (4.1a).

The truncation error associated with such a difference scheme is defined to be

$$\tau_i[y(t)] = N_h y(t_i) \qquad 1 \le i \le J$$

where $y(t)$ is a solution of (4.1a,b). The truncation error associated with the boundary condition is

$$\tau_o[y(t)] \equiv b(y(o), y(1))$$

and is always zero. Denote the $n(J+1)$-equations (4.8a,b) by

$$\eta(V) = o \tag{4.9}$$

where the $n(J+1)$-vector V is

$$V = \begin{bmatrix} v_o \\ \vdots \\ v_J \end{bmatrix} .$$

2. Existence of Solutions

We will now consider the question of existence of solutions to systems of nonlinear equations of the form (4.8a,b). As a prelude, let us examine the result of applying these nonlinear difference schemes to the linearized problem.

Lemma 4.10. Let the boundary-value problem (4.1a,b) have an isolated solution, y(t). Let

$$LV - r = o \tag{4.11}$$

be the n(J+1) equations which result from the application of the difference scheme (4.8a,b) to the linearized problem:

$$\frac{dw}{dt} = f_1(y,t)w + g(t) \qquad \text{a)}$$

$$b_1(y(o),y(1))w(o) + b_2(y(o),y(1))w(1) = \beta \qquad \text{b)} \qquad (4.12)$$

where g, β are n-vectors, and $g(t) \in C^1[o,1]$. Let the difference scheme be consistent and stable when applied to a linear two-point boundary-value problem with a unique solution, $z(t) \in C^2[o,1]$. Let $f(y,t) \in C^1[S_\rho[y(t)] \times [o,1]]$ and $b(x,y) \in C^1[S_\rho[y(o)] \times S_\rho[y(1)]]$. Then for all nets with $h_o \leq H$, H sufficiently small, L is nonsingular and

$$||\mathcal{L}^{-1}|| \le K_o$$

where K_o is independent of h_o.

<u>Proof</u>. By hypothesis, equation (4.12a,b) has a unique solution $z(t) \in C^2[o,1]$. Also, the difference scheme is stable when applied to (4.12a,b), thus there exist constants $K_o \ge o$, $H > o$ such that

$$||V|| \le K_o \; ||r|| \; .$$

for all nets with $h_o \le H$. This is equivalent to

$$||\mathcal{L}^{-1}|| \le K_o .$$

▨

<u>Theorem 4.13</u>. Let (4.1a,b) have an isolated solution $y(t) \in C^2[o,1]$. Let the hypothesis of Lemma 4.10 hold. Let the difference scheme (4.8a,b) be consistent with (4.1a,b), and let the Jacobian matrix $\eta_V(V)$ be continuous for $V \in S_{\rho_o}[y(t)]$. Further, let $\eta_V(V)$ be such that

$$||\mathcal{L} - \eta_V \; (Y)|| \to o \text{ as } h_o \to o \tag{4.14}$$

and

$$||\eta_V(V_1) - \eta_V(V_2)|| \le K_1 \; ||V_1 - V_2||$$

where (4.15)

$$V_1, V_2 \in S_{\rho_o}[y(t)].$$

Then for each ρ, $o < \rho \le \rho_o$, ρ_o sufficiently small and for all nets with $h_o \leq H(\rho)$, $H(\rho)$ sufficiently small, the $n(J+1)$ equations

$$\eta(V) = o$$

have a unique solution $W \in S_\rho[y(t)]$.

Proof: We define the vector function

$$\psi(V) = V - L^{-1}\,\eta(V).$$

If this has a solution, $\psi(W) = W$, then $\eta(W) = o$. The Contracting Mapping Theorem will be used to show that a unique solution exists. That is, we need to show that for some $\lambda \in (o,1)$

$$\text{i)}\quad ||Y - \psi(Y)|| \leq (1 - \lambda)\rho$$

$$\text{ii)}\quad ||\psi(V_1) - \psi(V_2)|| \leq \lambda||V_1 - V_2||, \text{ where}$$

$$V_1, V_2 \in S_\rho[y(t)].$$

Re i)

$$||Y - \psi(Y)|| = ||L^{-1}\eta(Y)|| \leq K_o||\tau[y(t)]||.$$

By the hypothesis the numerical scheme is consistent, hence

$$||\tau[y(t)]|| \to o \text{ as } h_o \to o .$$

Re ii)

$$||\psi(V_1) - \psi(V_2)|| = ||L^{-1}L(V_1 - V_2) - L^{-1}(\eta(V_1) - \eta(V_2))||$$

$$\leq ||L^{-1}|| \quad ||L(V_1 - V_2) - (\eta(V_1) - \eta(V_2))||.$$

Now, we employ the mean value theorem

$$\eta(V_1) - \eta(V_2) = \int_o^1 \eta_V(sV_1 + (1-s)V_2)ds\ (V_1 - V_2)$$

and immediately

$$\begin{aligned} ||\psi(V_1) - \psi(V_2)|| &\le K_o|| L - \int_o^1 \eta_V(sV_1 + (1-s)V_2)ds||\ ||V_1 - V_2|| \\ &\le K_o\{||L - \eta_V(Y)|| + ||\int_o^1 \eta_V(Y) - \eta_V(sV_1 + (1-s)V_2)\ ds||\} \times \\ &\quad ||V_1 - V_2||. \end{aligned}$$

By hypothesis $\eta_V(V)$ is continuous, thus

$$||\psi\ (V_1) - \psi\ (V_2)|| \le K_o\{K_1\rho + ||L - \eta_V(Y)||\}\ ||V_1 - V_2||.$$

By hypothesis, property (4.14),

$$||L - \eta_V(Y)|| \to o \text{ as } h_o \to o.$$

Thus, let ρ_o and $H(\rho_o)$ be sufficiently small such that

$$\lambda = K_o\{K_1\rho_o + ||L - \eta_V(Y)||\} \in (o,1)$$

and condition ii)is satisfied. Let $H(\rho)$ be sufficiently small such that

$$\frac{K_o\ ||\tau[y(t)]||}{\rho} \le (1-\lambda)\ . \tag{4.16}$$

Then $\psi(V)$ is a contracting map for $V \in S_\rho[y(t)]$ and there exists a unique solution $W \in S_\rho[y(t)]$.

▨

Corollary 4.17 (Convergence). Let the hypothesis of Theorem 4.13 hold. Then, if $\eta(W) = o$, $W \in S_\rho[y(t)]$,

$$||W - Y|| \to o \text{ as } h_o \to o.$$

Proof. It is sufficient to prove that for any $\varepsilon > o$, there exists an H such that

$$||W - Y|| \leq \varepsilon$$

for any net with $h_o \leq H$. However, this is the conclusion of Theorem 4.13 with ε replacing ρ.

▨

Corollary 4.18. Let the hypothesis of Theorem 4.13 hold. Let the difference scheme (4.8b) be p^th order accurate,

$$||\tau[y(t)]|| \leq K_2 \, h_o^p \, .$$

Then, for all nets with $h_o \leq H$, H sufficiently small,

$$||W - Y|| \leq K_3 \, h_o^p.$$

Proof. From the proof of Theorem 4.13, we know that $W \in S_\rho[y,t]$ if $\rho \leq \rho_o$, $h_o \leq H(\rho_o)$, and

$$\frac{K_o \; ||\tau[y(t)]||}{\rho} \leq 1-\lambda$$

where $\lambda \in (o,1)$ and fixed. By hypothesis

$$||\tau[y(t)]|| \leq K_2 \, h_o^p$$

and therefore

$$\frac{K_o \, ||\tau||}{\rho} \leq \frac{K_o K_2}{\rho} \, h_o^p \, .$$

Theorem 4.13 states that $W \epsilon S_\rho[y(t)]$ is a solution if

$$\frac{K_o K_2}{\rho} \, h_o^p \leq (1-\lambda) \, ,$$

thus it must be true for

$$\rho = \frac{K_o K_2}{1-\lambda} \, h_o^p \, .$$

Hence,

$$||W - Y|| \leq K_3 h_o^p \, .$$

▨

What difference schemes satisfy condition (4.14)? It would appear to be fortuitous indeed if any scheme had the property that

$$L = \eta_V(Y) \, ,$$

however this is the case for many common difference schemes. As examples, the centered-Euler (4.3) and the Gap_4 (4.6) schemes will be examined.

<u>Gap_4</u>. Application of the Gap_4 scheme to the linearized problem (4.12a,b) yields the following $n(J+1)$ equations

$$b_1(y(o),y(1))v_o + b_2(y(o),y(1))v_J = \beta \qquad \text{a)}$$

$$\frac{1}{h_i}\{I - \frac{h_i}{2} f_1(y_i,t_i) + \frac{h_i^2}{12}[f_{11}(y_i,t_i)f(y_i,t_i) + f_{12}(y_i,t_i) \qquad 4.19$$

$$+ f_1^2(y_i,t_i]\}v_i - \frac{1}{h_i}\{I + \frac{h_i}{2} f_1(y_{i-1},t_{i-1}) + \frac{h_i^2}{12}[f_{11}(y_{i-1}t_{i-1}) \times \qquad \text{b)}$$

$$f(y_{i-1},t_{i-1}) + f_{12}(y_{i-1},t_{i-1}) + f_1^2(y_{i-1},t_{i-1})]\}\, v_{i-1} = s_i \quad 1\le i\le J$$

where the right-hand side of (4.19b) is written simply as s_i because it is not of particular interest here. From (4.19a,b), we can completely characterize the elements (L_{ij}) of L.

Let the $(i,j)^{th}$ $n \times n$ block element of $\eta_V(Y)$ be N_{ij}. Then

$$N_{oo} = \left.\frac{\partial b(v_o,v_J)}{\partial v_o}\right|_{v_o=y(o),v_J=y(1)} = b_1(y(o),y(1))$$

and $N_{oo} = L_{oo}$. Similarly, $N_{oj} = L_{oj}$, $j = 1,\ldots,J$. From 4.6, it can be shown that

$$N_{i,i-1} = \left.\frac{\partial N_h v_i}{\partial v_{i-1}}\right|_{V=Y}$$

$$= -\frac{1}{h_i}\{I + \frac{h_i}{2} f_1(y_{i-1},t_{i-1}) + \frac{h_i^2}{12}[f_{11}(y_{i-1},t_{i-1})f(y_{i-1},t_{i-1})$$

$$+ f_1^2(y_{i-1},t_{i-1}) + f_{12}(y_{i-1},t_{i-1})]\} = L_{i,i-1}\ .$$

Similarly,

$$N_{ii} = L_{ii},\quad i = 1,\ldots,J.$$

Thus, the Gap_4 scheme satisfies condition (4.14) and in fact

$$L = \eta_V(Y) \text{ .}$$

Centered-Euler. The centered-Euler scheme satisfies equation (4.14), however L is not equal to $\eta_V(Y)$. Employing the centered-Euler scheme, (4.3), the Jacobian matrix $\eta_V(Y)$ has the following nonzero $n \times n$ block components:

$$\begin{aligned} &N_{oo} = b_1(y(o),y(1)) \qquad N_{oJ} = b_2(y(o),y(1)) && \text{a)} \\ &N_{i,i-1} = -\frac{1}{h_i}\{I + \frac{h_i}{2} f_1(\frac{1}{2}[y_i + y_{i-1}], t_i - \frac{1}{2}])\} && \text{b)} \qquad (4.20) \\ &N_{ii} = \frac{1}{h_i}\{I - \frac{h_i}{2} f_1(\frac{1}{2}[y_i + y_{i-1}], t_i - \frac{1}{2})\}. && \text{c)} \end{aligned}$$

The matrix L has these $n \times n$ block elements

$$\begin{aligned} &L_{oo} = b_1(y(o),y(1)) \qquad L_{oJ} = b_2(y(o),y(1)) && \text{a)} \\ &L_{i,i-1} = -\frac{1}{h_i}\{I + \frac{h_i}{2} f_1(y_{i-\frac{1}{2}}, t_{i-\frac{1}{2}})\} && \text{b)} \qquad (4.21) \\ &L_{ii} = \frac{1}{h_i}\{I - \frac{h_i}{2} f_1(y_{i-\frac{1}{2}}, t_{i-\frac{1}{2}})\}. && \text{c)} \end{aligned}$$

Thus, we have

$$\begin{aligned} ||L - \eta_V(Y)|| \le \max_i ||f_1(y_{i-\frac{1}{2}}, t_{i-\frac{1}{2}}) - f_1(y_{i-1}, t_{i-\frac{1}{2}})|| \\ + \max_i ||f_1(y_{i-\frac{1}{2}}, t_{i-\frac{1}{2}}) - f_1(y_i, t_{i-\frac{1}{2}})||. \end{aligned}$$

Clearly for $f(u,t) \in C^1[S_\rho[y(t)] \times [o,1]]$ and $y(t) \in C^1[o,1]$, property (4.14) holds. Note the quantity

$$||L - \eta_V(Y)||$$

need not approach zero like h_o^p in order for a p^{th} order accurate scheme to satisfy

$$||W - Y|| \leq K_3 h_o^p.$$

3. Solution of $\eta(V) = o$ by Newton's Method

In the previous section it was shown that under certain conditions a solution exists to the n(J+1) nonlinear equations

$$b(v_o, v_J) = o$$

$$N_h v_i = o, \qquad 1 \leq i \leq J \ .$$

We would like to examine a specific procedure for solving these equations, namely Newton's method. That is, we want to show that each iterate defined by

$$\eta_V(V^{\nu-1}) \ (V^\nu - V^{\nu-1}) + \eta(V^{\nu-1}) = o \qquad (4.22)$$

exists uniquely and that the sequence $\{V^\nu\}$ converges when V^o is judiciously chosen. The proof of the following theorem is identical to that given in Keller [6], with the appropriate generalizations.

Theorem 4.23 Let the hypothesis of Theorem 4.13 hold and let W be the n(J+1)-vector of that theorem where

$$\eta(W) = o.$$

Let V^o be such that $V^o \in S_\rho[y(t)]$ and $||V^o - W||$ is sufficiently small. Then for all nets with $h_o \leq H$, H sufficiently small, and for all ρ, $o < \rho \leq \rho_o$, ρ_o sufficiently small, the Newton iteration (4.22) uniquely defines a sequence $\{V^\nu\}$ and this sequence converges quadratically to W.

Proof: As noted before, see Keller [6].

▨

Chapters 1 and 3 examined the equivalence between shooting methods and implicit techniques for linear boundary-value problems. In the next section, we want to define a nonlinear shooting procedure and study its relationship to the method described here.

4. Equivalence of Shooting and Implicit Schemes for Nonlinear Problems

The nonlinear shooting scheme has the same underlying idea as in Chapter 1. The initial-value problem

$$\begin{aligned} &\frac{dx}{dt} = f(x(t),t) \qquad t \in [o,1] && a) \\ &x(o) = c && b) \end{aligned} \qquad (4.24)$$

is discretized, solved, and then the initial condition $x(o) = c$ is wiggled until the discrete boundary condition $b(c,u_J(c)) = o$ is satisfied.

For example, the following system of equations result if we use Euler's method to approximate (4.24a):

$$u_o = c \qquad \text{a)}$$

$$\frac{1}{h_i}[u_i - u_{i-1}] = f(u_{i-1},t_{i-1}), \quad 1 \le i \le J . \qquad \text{b)} \qquad (4.25)$$

Note that these equations can be solved explicitly, however an iteration scheme must be employed to satisfy the nonlinear boundary condition

$$b(u_o,u_J) = o . \qquad (4.26)$$

The application of Newton's Method to solve (4.26) is straight forward

$$[b_1(c^{\nu-1},u_J(c^{\nu-1})) + b_2(c^{\nu-1},u_J(c^{\nu-1})) \frac{\partial u_J(c^{\nu-1})}{\partial c}](c^{\nu}-c^{\nu-1})$$

$$+ b(c^{\nu-1},u_J(c^{\nu-1})) = o . \qquad (4.27)$$

However, the n × n matrix $\frac{\partial u_J}{\partial c}$ must be evaluated at each step. This matrix is determined by solving the variational equations

$$\frac{\partial u_o}{\partial c} = I \qquad \text{a)}$$

$$\frac{1}{h_i}[\frac{\partial u_i}{\partial c} - \frac{\partial u_{i-1}}{\partial c}] = f_1(u_{i-1},t_{i-1}) \frac{\partial u_{i-1}}{\partial c}, \quad 1 \le i \le J. \qquad \text{b)} \qquad (4.28)$$

A discrete shooting procedure results from combining (4.25a,b), (4.28a,b), and (4.27):

$$u_o^\nu = c^\nu \qquad \text{a)}$$

$$\frac{1}{h_i}[u_i^\nu - u_{i-1}^\nu] = f(u_{i-1}^\nu, t_{i-1}) \qquad \text{b)}$$

$$P_o^\nu = I \qquad \text{c)} \qquad (4.29)$$

$$\frac{1}{h_i}[P_i^\nu - P_{i-1}^\nu] = f_1(u_{i-1}^\nu, t_{i-1})\, P_{i-1}^\nu \qquad \text{d)}$$

$$\{b_1(c^\nu, u_J^\nu) + b_2(c^\nu, u_J^\nu)\, P_J^\nu\}\, [c^{\nu+1} - c^\nu] + b(c^\nu, u_J^\nu) = o \cdot \qquad \text{e)}$$

An alternative to this procedure is Newton's method as described in the previous section. That is, Newton's method is employed to solve the system of equations

$$b(v_o, v_J) = o \qquad \text{a)}$$

$$\frac{1}{h_i}[v_i - v_{i-1}] = f(v_{i-1}, t_{i-1}) \cdot \qquad \text{b)} \qquad (4.29)$$

This results in the following iteration scheme

$$b_1(v_o^\nu, v_J^\nu)[v_o^{\nu+1} - v_o^\nu] + b_2(v_o^\nu, v_J^\nu)[v_J^{\nu+1} - v_J^\nu] + b(v_o^\nu, v_J^\nu) = o \qquad \text{a)}$$

$$\frac{1}{h_i}[v_i^{\nu+1} - v_i^\nu] - \frac{1}{h_i}[v_{i-1}^\nu - v_{i-1}^\nu] = f_1(v_{i-1}^\nu, t_{i-1})[v_{i-1}^{\nu+1} - v_{i-1}^\nu] \qquad (4.31)$$

$$- \frac{1}{h_i}[v_i^\nu - v_{i-1}^\nu] + f(v_{i-1}^\nu, t_{i-1}), \quad 1 \le i \le J. \qquad \text{b)}$$

It has already been shown that $\{V^\nu\}$ exists and converges.

It is clear that in each case the method defines a solution of the same set of equations:

$$\eta(V) = o \;.$$

The implicit method (4.31a,b) applies Newton's method to each of the n(J+1) equations of (4.30a,b). The shooting method (4.29a,b,c,d,e) formally separates the first n equations from the remaining nJ,

$$\begin{aligned} &b(u_o,u_J) = o && \text{a)} \\ &\bar{\eta}(U) = o && \text{b)} \;. \end{aligned} \qquad (4.32)$$

In lieu of this set of equations, the shooting method substitutes the equations

$$\begin{aligned} &u_o = c && \text{a)} \\ &\bar{\eta}(U) = o && \text{b)} \end{aligned} \qquad (4.33)$$

which can be solved exactly when an explicit difference scheme is used to approximate (4.23a). Then the initial value c is varied until an initial value is found such that

$$b(c,u_J(c)) = o \;.$$

The principal difference is that the shooting procedure satisfies the last nJ equations exactly and the implicit Newton procedure does not. That is,

$$\bar{\eta}(U^\nu) = o \;, \qquad \bar{\eta}(V^\nu) = \bar{r}^\nu \;,$$

where in general $||\bar{r}^\nu|| \neq o$.

Theorem 4.34 Let the hypothesis of Theorem 4.16 hold. Let the shooting Newton iterates be $\{U^\nu\}$, $U^\nu \in S_\rho[y(t)]$, and let the implicit

Newton iterates be $\{V^\nu\}$. Let $\bar{\eta}(V^\nu) = \bar{r}^\nu$, $||\bar{r}^\nu|| > o$. Then $U^\nu \neq V^\nu$.

Proof: Define

$$r^\nu = \eta(U^\nu) - \eta(V^\nu)$$

As in Theorem 4.13, the mean value theorem is employed in the following way

$$r^\nu = \int_o^1 \eta_V(sU^\nu + (1-s)V^\nu)ds \ (U^\nu - V^\nu).$$

By hypothesis, η_V is continuous on $S_\rho[y(t)]$, hence

$$||r^\nu|| \leq Q \ ||U^\nu - V^\nu||$$

As in the proof of Theorem 4.13, it can be shown that the matrix $\int_o^1 \eta_V(sU^\nu + (1-s)\ V^\nu)ds$ is nonsingular, thus $Q > o$. Immediately from the above,

$$o < ||\bar{r}^\nu|| \leq ||r^\nu|| \leq Q \ ||U^\nu - V^\nu||$$

Thus, $||U^\nu - V^\nu|| > o \implies U^\nu \neq V^\nu$.

▨

This theorem states that the iteration sequences defined by implict Newton and shooting Newton are not the same, even though they are used to solve the same set of nonlinear equations.

Chapter V

Block Tridiagonal Matrices

This chapter develops a computational method for solving the systems of linear equations that have arisen in previous chapters. That is, we will develop a technique for solving the n(J+1) linear equations

$$\mathcal{B}_h V - r = o$$

or, when the boundary-value problem is nonlinear, the equations

$$\eta_V(V^\nu)(V^{\nu+1} - V^\nu) + \eta(V^\nu) = o$$

generated by Newton's method. More specifically we are interested in matrices which result from the approximation of two-point boundary-value problems with separated boundary conditions. That is, problems which may be written as

$$\begin{aligned} &u'(t) = f(u,t) \qquad t \in [o,1] && \text{a)} \\ &b_o(u(o)) = o && \text{b)} \qquad (5.1) \\ &b_1(u(1)) = o && \text{c)} \end{aligned}$$

where the boundary conditions (5.1b) and (5.1c) represent p and q conditions respectively, p + q = n.

To illustrate the form that these matrices will take, consider the approximation of

$$\begin{aligned} &u' = A(t)u + f(t) \qquad t \in [o,1] && \text{a)} \\ &B_o\, u(o) = \beta_o \qquad B_1\, u(1) = \beta_1 && \text{b)} \end{aligned} \qquad (5.2)$$

by Euler's method. The resulting finite difference equations (J=2) are

$$\begin{bmatrix} (B_o) & & \\ -\frac{1}{h_1}I - A(t_o) & \frac{1}{h_1}I & \\ & -\frac{1}{h_2}I - A(t_1) & \frac{1}{h_2}I \\ & & [B_1] \end{bmatrix} \begin{bmatrix} v_o \\ \\ v_1 \\ \\ v_2 \end{bmatrix} = \begin{bmatrix} (\beta_o) \\ f(t_o) \\ \\ f(t_1) \\ \\ [\beta_1] \end{bmatrix} \qquad (5.3)$$

where we have used the convention that a single matrix in parentheses, (B_o), has p rows and a single matrix in brackets, $[B_1]$, has q rows. The matrix of (5.3) can be rewritten in block tridiagonal form

$$\begin{bmatrix} A_o & B_1 & 0 \\ C_1 & A_1 & B_2 \\ 0 & C_2 & A_2 \end{bmatrix} \begin{bmatrix} v_o \\ v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} r_o \\ r_1 \\ r_2 \end{bmatrix}$$

where each A_i, B_i, C_i is an n × n matrix. Block matrices of this type have a very special zero structure: the first p rows of B_i and the last q rows of C_i are zero rows.

The solution procedure of interest is a block LU decomposition:

$$\begin{bmatrix} A_o & B_1 & \\ C_1 & A_1 & B_2 \\ & C_2 & A_2 \end{bmatrix} = \begin{bmatrix} I & & \\ L_1 & I & \\ & L_2 & I \end{bmatrix} \begin{bmatrix} D_o & B_1 & \\ & D_1 & B_2 \\ & & D_2 \end{bmatrix} .$$

Ideally, we would like to show that this decomposition exists if the matrix $\mathcal{B}_h$ is nonsingular. Unfortunately the result is not that clear. For example, if A(t) in (5.2a) is a 4 × 4 diagonal matrix and $(B_o) = (1\ o\ o\ o)$, then

$$A_o = \begin{bmatrix} x & o & o & o \\ x & o & o & o \\ o & x & o & o \\ o & o & x & o \end{bmatrix}$$

and A_o is clearly singular. No matter how small h_o becomes, this matrix will remain singular. However, if the 5^{th} row of $\mathcal{B}_h$ is interchanged with the 2^{nd} row, then

$$A_o = \begin{bmatrix} x & o & o & o \\ o & o & o & x \\ o & x & o & o \\ o & o & x & o \end{bmatrix} ,$$

A_o is nonsingular, and the LU factorization can proceed. Thus, a row switching strategy will be included in the decomposition procedure.

1. Block LU Decompostion

The block tridiagonal matrix M is considered in the following generality;

$$M = \begin{bmatrix} A_o & B_1 & & \\ C_1 & A_1 & \ddots & \\ & \ddots & \ddots & B_J \\ & & C_J & A_J \end{bmatrix} \qquad (5.4)$$

where each A_i, B_i, C_i is an n × n matrix. We also require that the first p rows of each B_i be zero and the last q rows of C_i be zero. Such a matrix M will be referred to as a block tridiagonal matrix with p/q zero structure.

The straightforward block decomposition is

$$M = \begin{bmatrix} I & & & \\ L_1 & I & & \\ & \ddots & \ddots & \\ & & L_J & I \end{bmatrix} \begin{bmatrix} D_o & B_1 & & \\ & D_1 & \ddots & \\ & & \ddots & B_J \\ & & & D_J \end{bmatrix} \qquad (5.5)$$

where the matrices L_i, D_i are defined by

$$D_o = A_o \qquad L_i = C_i D_{i-1}^{-1} , \qquad D_i = A_i - L_i B_i . \qquad (5.6)$$

Once the LU decomposition is completed, the standard procedure is adopted to solve MV = r:

$$LW = r , \qquad a)$$
$$UV = W . \qquad b) \qquad (5.7)$$

The n(J+1) vectors W and V are defined by the recursion relations:

$$
\begin{aligned}
w_o &= r_o \\
w_i &= r_i - L_i w_{i-1}
\end{aligned}
\tag{5.8}
$$

and

$$
\begin{aligned}
D_J v_J &= w_J , \\
D_i v_i &= w_i - B_{i+1} v_{i+1} .
\end{aligned}
\tag{5.9}
$$

If it should occur that some D_i is singular, then these recursion formulae will have to be altered to allow for the appropriate row interchanges.

Lemma 5.10. Let M^o, M denote n(m+1) × n(m+1) block tridiagonal matrices with p/q zero structure. Denote the block elements of M^o, M by A_i^o, B_i^o, C_i^o and A_i, B_i, C_i respectively. Let M^o be nonsingular. Then either A_o^o is nonsingular or by an interchange of rows (p + 1,...,p + n) of M^o a matrix M can be formed such that A_o is nonsingular.

Proof. $m = o$. Immediate because $M^o = A_o^o$.

$m > o$. Let (S_p) be the p × n matrix composed of the first p rows of A_o^o. Let S_n be the n × n matrix whose first q rows are the last q rows of A_o^o and whose last p rows are the first p rows of C_1^o. With these definitions in hand, the matrix M^o can be written as

$$M^o = \begin{bmatrix} (S_p) & (0) & \ldots & (0) \\ S_n & & & \\ 0 & & Q & \\ \vdots & & & \\ 0 & & & \end{bmatrix}$$

where Q is an (nm+q) × nm matrix. By hypothesis M^o is nonsingular, thus the rows of (S_p) are linearly independent and the (n+p) × n matrix S defined by

$$S = \begin{bmatrix} (S_p) \\ S_n \end{bmatrix}$$

has rank n. These two conditions imply that there are q rows of S_n which, together with the rows of (S_p), form a set of n linearly independent row vectors. Thus, by interchanging the appropriate rows of M^o we can form a matrix M such that A_o is nonsingular.

▨

Similar rows. Define rows of a block tridiagonal matrix with p/q zero structure to be similar if by interchanging these rows the matrix remains block tridiagonal with p/q zero structure.

Theorem 5.11. Let M^o be an n(J+1) × n(J+1) block tridiagonal matrix with p/q zero structure. Let M^o be nonsingular. Then by an interchange strategy among similar rows of M^o a matrix M can be formed such that M has a block LU decomposition of the form

$$M = \begin{bmatrix} I & & & \\ L_1 & I & & \\ & \ddots & \ddots & \\ & & L_J & I \end{bmatrix} \begin{bmatrix} D_o & B_1 & & \\ & D_1 & \ddots & \\ & & \ddots & B_J \\ & & & D_J \end{bmatrix}$$

where each D_i, $o \le i \le J$, is nonsingular.

Proof. Recalling Lemma 5.10, there is an interchange strategy among rows $(p + 1,\ldots,p + n)$ of M^o that forms an M^1 such that

$$M^1 = \begin{bmatrix} I & & & & \\ L_1^1 & I & & & \\ & & I & & \\ & & & \ddots & \\ & & & & I \end{bmatrix} \begin{bmatrix} D_o^1 & B_1^1 & & & \\ & D_1^1 & B_2^o & \ddots & \\ & C_2^o & A_2^o & & B_J^o \\ & & \ddots & \ddots & \\ & & & C_J^o & A_J^o \end{bmatrix}$$

where D_o^1 is nonsingular and M^1 is a block tridiagonal matrix with p/q zero structure. Assume that there exists a matrix M^k with the following properties:

i) $M^k = F^k G^k$

where

$$F^k = \begin{bmatrix} I & & & & & \\ L_1^1 & I & & & & \\ & \ddots & \ddots & & & \\ & & L_k^k & I & & \\ & & & 0 & I & \\ & & & & \ddots & \ddots \\ & & & & 0 & I \end{bmatrix}$$

and

$$G^k = \begin{bmatrix} D_o^1 & B_1^1 & & & & & \\ 0 & D_1^2 & \ddots & & & & \\ & \ddots & \ddots & B_k^k & & & \\ & & 0 & D_k^k & B_{k+1}^o & & \\ & & & C_{k+1}^o & A_{k+1}^o & \ddots & \\ & & & & \ddots & \ddots & B_J^o \\ & & & & & C_J^o & A_J^o \end{bmatrix}.$$

ii) Each D_i^{i+1}, $o \le i \le k - 1$, is nonsingular, and

iii) M^k is a block tridiagonal matrix with p/q zero structure formed by an interchange strategy of similar rows of M^o.

Since M^1 has properties i), ii), and iii) it is sufficient to show that from M^k we can proceed to M^{k+1}. By hypothesis M^o is nonsingular, thus M^k must be nonsingular. Define the $n(J+1-k) \times n(J+1-k)$ matrix H by

$$H = \begin{bmatrix} D_k^k & B_k^o & & \\ C_{k+1}^o & A_{k+1}^o & \ddots & \\ & \ddots & \ddots & B_J^o \\ & & C_J^o & A_J^o \end{bmatrix}$$

noting that H is block tridiagonal with p/q zero structure. The matrix M^k is nonsingular if and only if H is nonsingular. Recalling Lemma 5.10, there is an interchange strategy among rows $(p + 1,\ldots,p + n)$ of H which produces an $\bar{H}$ such that

$$\bar{H} = \begin{bmatrix} I & & & & \\ L^{k+1}_{k+1} & I & & & \\ & 0 & I & & \\ & & \ddots & \ddots & \\ & & & 0 & I \end{bmatrix} \begin{bmatrix} D^{k+1}_{k} & B^{k}_{k} & & & \\ 0 & D^{k+1}_{k+1} & B^{o}_{k+1} & & \\ & C^{o}_{k+2} & A^{o}_{k+2} & \ddots & \\ & & \ddots & \ddots & B^{o}_{J} \\ & & & C^{o}_{J} & A^{o}_{J} \end{bmatrix}$$

where D^{k+1}_{k} is nonsingular. Note that an interchange of any rows $(p+1,\ldots,p+n)$ of H is equivalent to an interchange of rows $(nk + p + 1,\ldots,nk + p + n)$ of M^{o}. By defining L^{k+1}_{k+1}, D^{k+1}_{k}, B^{k}_{k}, and D^{k+1}_{k+1} in this manner, we have found $M^{k+1} = F^{k+1}\ G^{k+1}$ which has properties i), ii), and iii).

▨

Corollary 5.12. Let M^{o} be an $n(J+1) \times n(J+1)$ block tridiagonal matrix with p/q zero structure. Let P be a permutation matrix of the form

$$P = \left[\begin{array}{c:c:c:c:c} I_p & & & & \\ \hdashline & P_1 & & & \\ \hdashline & & \ddots & & \\ \hdashline & & & P_J & \\ \hdashline & & & & I_q \end{array}\right]$$

where I_p is the $p \times p$ identity, I_q is the $q \times q$ identity, and each P_i is an $n \times n$ permutation matrix. Then the following are equivalent

a) M^o is nonsingular.

b) There exists a permutation matrix P such that

$$M = PM^o = \begin{bmatrix} I & & & \\ L_1 & I & & \\ & \ddots & \ddots & \\ & & L_J & I \end{bmatrix} \begin{bmatrix} D_o & B_1 & & \\ & D_1 & \ddots & \\ & & \ddots & B_J \\ & & & D_J \end{bmatrix}$$

where each D_i, $o \le i \le J$, is nonsingular.

Proof: a) ⇒ b). The proof is exactly that of Theorem 5.11.

b) ⇒ a). Immediate from hypothesis.

▨

Using this result, a block tridiagonal factorization can be designed to solve the equations

$$MV = r$$

where M is a block tridiagonal matrix with p/q zero structure. Recalling Corollary (5.12), the equations can be written as

$$(LU)V = P\ r$$

which allows solution via (5.6), (5.8), and (5.9). The only problem remaining is the practical one of determining each of the permutation matrices P_i comprising P. Each P_i can be found as the LU decomposition (5.6) proceeds.

Theorem 5.13. Let M be a block tridiagonal matrix with p/q zero structure. Let M be nonsingular. Let r, V, W be n(J+1) vectors defined by

$$r = \begin{bmatrix} r_o \\ \vdots \\ r_J \end{bmatrix}, \qquad V = \begin{bmatrix} v_o \\ \vdots \\ v_J \end{bmatrix}, \qquad W = \begin{bmatrix} w_o \\ \vdots \\ w_J \end{bmatrix} .$$

Then the following procedure can be used to find the solution of $MV = r$:

I. Forward substitution

A. Starting values: $D_o = A_o$, $w_o = r_o$.

B. Iteration procedure ($k = o,1,\ldots,J-1$)

1. If the current D_k is singular, interchange rows $(kn + p + 1,\ldots,kn + p + n)$ of M in order to form a nonsingular D_k and form new w_k, r_{k+1} by interchanging corresponding elements.

2. Compute: $L_{k+1} = C_{k+1} D_k^{-1}$

$$D_{k+1} = A_{k+1} - L_{k+1} B_{k+1}$$

$$w_{k+1} = r_{k+1} - L_{k+1} w_k$$

II. Back substitution ($k = J,J-1,\ldots,o$)

A. Compute v_k via $D_k v_k = w_k - B_{k+1} v_{k+1}$

Proof: The proof is the same as the proof of Theorem 5.11.

▨

Thus far only one method of decomposing the matrix M has been considered: the natural extension of the usual LU decomposition to block matrices. Consider an LU decomposition where U rather than L has a unit diagonal;

$$\begin{bmatrix} A_o & B_1 & \\ C_1 & A_1 & B_2 \\ & C_2 & A_2 \end{bmatrix} = \begin{bmatrix} D_o & & \\ C_1 & D_1 & \\ & C_2 & D_2 \end{bmatrix} \begin{bmatrix} I & U_1 & \\ & I & U_2 \\ & & I \end{bmatrix} .$$

The general recursion relations for D_i, U_i are

$$D_o = A_o \qquad U_i = D_{i-1}^{-1} B_i, \qquad D_i = A_i - C_iU_i. \tag{5.14}$$

Allowing for row interchanges, this decomposition can always be performed, however it has a practical disadvantage. In the LU decomposition (5.6), L_i has the same zero structure as C_i, namely the last q rows are zero. Thus, the matrices L, U can be stored in the same locations as the original matrix M. However, from (5.14) it is clear that U_i is in general a full n × n matrix, and this decomposition requires pnJ more storage locations than M.

2. A Split decomposition

Another possibility is the factorization

$$M = UL$$

where U has a unit diagonal. That is, in the case J = 2, we have

$$\begin{bmatrix} A_o & B_1 & \\ C_1 & A_1 & B_2 \\ & C_2 & A_2 \end{bmatrix} = \begin{bmatrix} I & U_1 & \\ & I & U_2 \\ & & I \end{bmatrix} \begin{bmatrix} D_o & & \\ C_1 & D_1 & \\ & C_2 & D_2 \end{bmatrix} .$$

<u>Corollary 5.15</u>. Let M^o be an $n(J+1) \times n(J+1)$ block tridiagonal matrix with p/q zero structure. Let P be a permutation matrix of the same form as Corollary 5.12. Then the following are equivalent:

a) M^o is nonsingular.

b) There exists a permutation matrix P such that

$$M = PM^o = \begin{bmatrix} I & U_1 & & \\ & I & \ddots & \\ & & \ddots & U_J \\ & & & I \end{bmatrix} \begin{bmatrix} D_o & & & \\ C_1 & D_1 & & \\ & \ddots & \ddots & \\ & & C_J & D_J \end{bmatrix}$$

where each D_i, $o \le i \le J$, is nonsingular.

<u>Proof</u>. b) ⟹ a) Immediate from hypothesis.

a) ⟹ b) Let Q be the $n(J+1) \times n(J+1)$ permutation matrix where $q_{i,n(J+1)-i+1} = 1$. Then $\bar{M}^o = QM^oQ$ is a nonsingular tridiagonal matrix with q/p zero structure. Recalling Corollary 5.11, there exists a permutation matrix $\bar{P}$ such that

$$\bar{M} = \bar{P}\,\bar{M}^o = \bar{L}\,\bar{U}$$

where $\bar{M}$ is a block tridiagonal matrix with q/p zero structure. Note that QQ = I, then we have

$$M = P\,M^o = UL$$

where $P = Q\bar{P}Q$, $U = Q\bar{L}$, and $L = \bar{U}Q$.

▨

Rather than pursuing this UL decomposition directly, we will examine a hybrid or split procedure. The matrix M might be factored in the following way:

$$\begin{bmatrix} A_o & B_o & \\ C_1 & A_1 & B_2 \\ & C_2 & A_2 \end{bmatrix} = \begin{bmatrix} I & & \\ L_1 & I & U_2 \\ & & I \end{bmatrix} \begin{bmatrix} D_o & B_1 & \\ & D_1 & \\ & C_2 & D_2 \end{bmatrix} . \tag{5.16}$$

This decomposition is a hybrid of the LU and the UL decompositions discussed previously.

<u>Theorem 5.19</u>. Let M^o be an $n(J+1) \times n(J+1)$ block tridiagonal matrix with p/q zero structure. Let P(s) be a permutation matrix of the same form as Corollary 5.12. Then the following are equivalent:

a) M^o is nonsingular.

b) For each integer s, $o \leq s \leq J$, there exists a permutation matrix P(s) such that $M = P(s)M^o$ has a split factorization of the form:

$$M = P(s)M^o = FG \tag{5.20}$$

where

$$F = \begin{bmatrix} I & & & & & \\ L_1 & I & & & & \\ & \ddots & \ddots & & & \\ & & L_s & I & U_{s+1} & \\ & & & & I & \ddots \\ & & & & & \ddots & U_J \\ & & & & & & I \end{bmatrix} ,$$

$$G = \begin{bmatrix} D_o & B_1 & & & & \\ & D_1 & \ddots & & & \\ & & \ddots & B_s & & \\ & & & D_s & & \\ & & & C_{s+1} & D_{s+1} & \\ & & & & \ddots & \ddots \\ & & & & C_J & D_J \end{bmatrix},$$

and each D_i, $o \le i \le J$, is nonsingular.

Proof: The proof is sketched here. The only element of F or G that is not guaranteed existence by Corollary (5.12) (L_i, $i = 1,\ldots,s$; D_i, $i = o,\ldots,s-1$) or Corollary (5.15) (U_i, $i = s+1,\ldots,J$; D_i, $i = s+1,\ldots,J$) is D_s. In order for the factorization (5.20) to exist, D_s must be defined to be

$$D_s = A_s - L_s B_s - U_{s+1} C_{s+1}.$$

Since D_s is given in terms of known quantities, the proof is complete.

▨

Note that this split procedure includes the LU factorization ($s = J$) and the UL factorization ($s = o$) as special cases. The general algorithm for solving $MV = r$ is given below where any quantity not previously defined is assumed to be zero.

Solution of $MV = r$ ($o \le s \le J$)

I. Decomposition procedure

A. Forward ($i = o,1,\ldots,s-1$)

1. Compute: $D_i = A_i - L_i B_i$

2. Interchange strategy: If D_i is singular interchange rows $(ni + p + 1,\ldots,ni + p + n)$ to form nonsingular D_i, and new B_{i+1}, C_{i+1}, A_{i+1}. Change r_i, r_{i+1} accordingly.

3. Compute: $w_i = r_i - L_i w_{i-1}$

$L_{i+1} = C_{i+1} D_i^{-1}$

B. Backward $(i = J,\ldots,s+1)$

1. Compute: $D_i = A_i - U_{i+1} C_{i+1}$

2. If D_i is singular, interchange rows $(ni - q - 1,\ldots,ni - q - n)$ to form nonsingular D_i, and new B_{i-1}, C_i, A_{i-1}. Change r_i, r_{i-1} accordingly.

3. Compute: $w_i = r_i - U_{i+1} w_{i+1}$

$U_i = B_i D_i^{-1}$

C. Split point

1. Compute: $D_s = A_s - L_s B_s - U_{s+1} C_{s+1}$

$w_s = r_s - L_s w_{s-1} - U_{s+1} w_{s+1}$

II. Solution routine

A. Split point

1. Solve for v_s: $D_s v_s = w_s$

B. Backward $(i = s-1,\ldots,o)$

1. Solve for v_i: $D_i v_i = w_i - B_{i+1} v_{i+1}$

C. Forward $(i = s+1,\ldots,J)$

1. Solve for v_i: $D_i v_i = w_i - C_i v_{i-1}$

The numerical example cited in Chapter 6 uses $s = o$ exclusively. In some cases however it has been found that the split point $s \cong (J+1)/2$ yields good numerical results when the $s = o$, $s = J$ decompositions are unsuitable.

Chapter VI

A Numerical Example: Plane Couette Flow

In this chapter, equations modeling plane Couette flow are used as a test case for some of the theory presented previously. This particular example was chosen because for certain special cases the exact solution is known and can be shown to be an isolated solution (see Chapter 4). The equations are given in F. K. Moore [8] as a similarity solution of the Navier-Stokes equations describing the motion of a compressible, viscous fluid contained between two parallel surfaces. These walls are in relative motion and each is kept at a constant temperature.

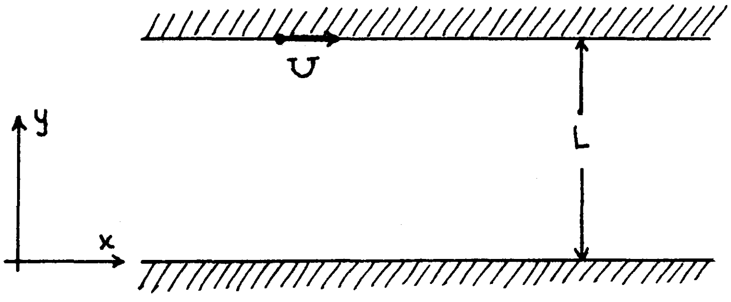


Fig. 6.1

The usual no-slip boundary conditions are imposed at the walls, and the variables of interest are T(t), the normalized temperature distribution, and u(t), the normalized x-component of velocity. In accordance with our previous notation, t is used as the independent variable. The equations will be considered in the following form:

$$\begin{aligned}
&\frac{d\bar{u}}{dt} = o && u(o) = o \\
&\frac{d\bar{T}}{dt} = \frac{1}{\mu(T)} K\bar{u}^2 && T(o) = \lambda \\
&\frac{dT}{dt} = \frac{1}{\mu(T)} \bar{T} && u(1) = 1 \\
&\frac{du}{dt} = \frac{1}{\mu(T)} \bar{u} && T(1) = 1
\end{aligned} \qquad (6.2)$$

where $\mu(T)$ is the viscosity-temperature relation.

1. Isolated Solution

In the case where $K = o$, the differential equations (6.2) reduce to

$$\begin{aligned}
&\frac{dT}{dt} = \frac{1}{\mu(T)}\bar{T}(o), && \bar{T} = \bar{T}(o), \\
&\frac{du}{dt} = \frac{1}{\mu(T)}\bar{u}(o), && \bar{u} = \bar{u}(o),
\end{aligned} \qquad (6.3)$$

with the same boundary conditions. These equations are easily integrated if we restrict the viscosity-temperature relation to be of the form

$$\mu(T) = T^{\alpha}.$$

An exact solution for (6.3) is shown to be

$$\begin{aligned}
\bar{u} &= \frac{1}{\alpha+1} \left[\frac{1 - \lambda^{\alpha+1}}{1 - \lambda}\right], \\
\bar{\bar{T}} &= \frac{1}{\alpha+1} \left[1 - \lambda^{\alpha+1}\right], \\
T &= [\lambda^{\alpha+1} (1-t) + t]^{\frac{1}{\alpha+1}}, \\
u &= \frac{1}{1-\lambda} \{[\lambda^{\alpha+1} (1-t) + t]^{\frac{1}{\alpha+1}} - \lambda\}.
\end{aligned} \tag{6.4}$$

The homogeneous, linearized problem becomes

$$\frac{dw}{dt} = A(t)w$$

$$B_o w(o) = o \qquad B_1 w(1) = o$$

where

$$A(t) = \begin{bmatrix} o & o & o & o \\ o & o & o & o \\ o & T^{-\alpha} & -\alpha T^{-\alpha+1} \bar{T} & o \\ T^{-\alpha} & o & -\alpha T^{-\alpha+1} \bar{u} & o \end{bmatrix}$$

and

$$B_o = B_1 = \begin{bmatrix} o & o & 1 & o \\ o & o & o & 1 \end{bmatrix}.$$

Substituting the exact solution from (6.4) into A(t), it can be shown that the linearized problem has only the trivial solution, and thus, the solution (6.4) is isolated.

2. Numerical Procedure

The Gap_4 will be employed to approximate the differential equations in (6.2). Substituting $\phi(T) = \mu^{-1}(T)$ and letting

$$y(t) \equiv \begin{bmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \\ y_4(t) \end{bmatrix} = \begin{bmatrix} \bar{u}(t) \\ \bar{T}(t) \\ T(t) \\ u(t) \end{bmatrix}$$

in (6.2), this two-point boundary-value problem can be written as:

$$\frac{dy}{dt} = f(y) \qquad \text{a)}$$

$$B_o y(o) = \begin{bmatrix} \lambda \\ 1 \end{bmatrix} \quad B_1 y(1) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \text{b)} \qquad (6.4)$$

where

$$f(y) = \begin{bmatrix} o \\ K\ \phi(y_3)y_1^2 \\ \phi(y_3)y_2 \\ \phi(y_3)y_1 \end{bmatrix}, \quad B_o = B_1 = \begin{bmatrix} o & o & 1 & o \\ o & o & o & 1 \end{bmatrix}.$$

In order to use the Gap_4 scheme, the vector $f_y f$ must be evaluated. Let $\alpha = (a\ b\ c\ d)^T$, then

$$f_y(\alpha)f(\alpha) \equiv F(\alpha) = \begin{bmatrix} o \\ K\ \phi(c)\ \phi'(c)ba^2 \\ a^2\ \phi(c)(K\phi(c) + \phi'(c)) \\ \phi(c)\ \phi'(c)\ ab \end{bmatrix} .$$

The Gap_4 difference approximation is given by

$$N_h v_i = v_i - v_{i-1} - \frac{h_i}{2}\,[f(v_i) + f(v_{i-1})] + \frac{h_i^2}{12}\,[F(v_i) - F(v_{i-1})] = o \qquad (6.5)$$

$$i = 1,\ldots,J.$$

Coupling these difference equations with the discrete boundary conditions,

$$B_o\ v_o = \begin{bmatrix} \lambda \\ 1 \end{bmatrix} \qquad B_1\ v_J = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad (6.6)$$

yields 4J+4 nonlinear equations which must be solved in order to evaluate V. Newton's method is employed to solve these equations.

Define the 4×4 matrices $N^1(\alpha)$, $N^2(\alpha)$ by

$$N^1(\alpha) = \frac{\partial f(\alpha)}{\partial\alpha}\ , \quad N^2(\alpha) = \frac{\partial F(\alpha)}{\partial\alpha}\ .$$

Thus, again letting $\alpha = (a\ b\ c\ d)^T$,

$$N^1(\alpha) = \begin{bmatrix} o & o & o & o \\ 2K\phi a & o & K\phi' a^2 & o \\ o & \phi & \phi' b & o \\ \phi & o & \phi' a & o \end{bmatrix}$$

and

$$N^2(\alpha) = \begin{bmatrix} o & o & o & o \\ 2ab\phi\phi' & K\phi\phi' a^2 & K(\phi\phi')' ba^2 & o \\ 2a\phi(K\phi+\phi') & o & a^2(K\phi^2+\phi\phi')' & o \\ \phi\phi' b & \phi\phi' a & o & o \end{bmatrix}$$

where ϕ, ϕ' are understood to stand for $\phi(c)$, $\phi'(c)$ respectively. Newton's method yields the following linear difference equations:

$$[I - \frac{h_i}{2} N^1(v_i^\nu) + \frac{h_i^2}{12} N^2(v_i^\nu)](v_i^{\nu+1} - v_i^\nu)$$

$$- [I + \frac{h_i}{2} N^1(v_{i-1}^\nu) + \frac{h_i^2}{12} N^2(v_{i-1}^\nu)](v_{i-1}^{\nu+1} - v_{i-1}^\nu)$$

$$= -N_h v_i^\nu.$$

These difference equations may be put into a form

$$\eta_V(V^\nu)(V^{\nu+1} - V^\nu) = -\eta(V^\nu),$$

where the matrix, $\eta_V(V^\nu)$, is block tridiagonal matrix with 2/2 zero structure. The split decomposition routine (s=o) was used to solve each of these equations for $(V^{\nu+1} - V^\nu)$. In each case considered, the initial guess, V^o, was taken to be

$$v_i^o = [\, o \quad o \quad \lambda + (1-\lambda)t_i \quad (1-\lambda)t_i]^T.$$

3. Numerical results

These computations were performed in double precision on an IBM 360/65. Three separate cases are reported here:

i) $K = o$, $\mu(T) = T$,

ii) $K = -1$, $\mu(T) = T$,

iii) $K = -1$, $\mu(T) = T^{3/2}$.

The left-hand boundary condition, $T(o) = \lambda$, was taken to be $T(o) = 1/2$.

In each case, ten uniformly spaced net points (J=9) were placed on the interval [o,1]. Newton's method was used to solve the equations (6.5), (6.6) and the iteration procedure was terminated when the residual, $||\eta(V^{\nu})||$, was acceptably small ($<10^{-8}$). The results of these calculations are summarized in Table 6.7.

(K,α^*)	ν	$\|\|\eta(V^{\nu})\|\|$	Time (sec.)
(0,1)	0 1 2 3	.11D 00 .7142D-02 .1169D-04 .5071D-10	2.51
(-1,1)	0 1 2 3 4	.1111D 00 .1072D 00 .5452D-02 .1363D-04 .1242D-09	3.67
(-1,3/2)	0 1 2 3 4	.1111D 00 .1004D 00 .1157D-01 .7149D-04 .6233D-08	4.64

*Represents the viscosity relation $\mu(T) = T^{\alpha}$

Table 6.7

Note that in each case the Newton iterates exhibit quadratic convergence.

An exact solution is known for case i), (o,1), so a more detailed inspection of the numerical solution is in order. The exact solution for K = o, $\mu(T) = T$ is

$$y(t) = \begin{bmatrix} 3/4 \\ 3/8 \\ (\frac{1}{4}\ [1-t] + t)^{1/2} \\ 2\{(\frac{1}{4}\ [1-t] + t^{1/2} - \frac{1}{2}\} \end{bmatrix} . \qquad (6.8)$$

The approximations to $y_1(t_i)$, v_{i1}, and $y_2(t_i)$, v_{i2}, are constants for all i:

$$v_{i1} = .750009065843 \text{ D } 00,$$

$$v_{i2} = .375004532921 \text{ D } 00.$$

Recalling the exact solution (6.8), the corresponding errors are .907 D-05 and .453 D-05 respectively. Let $|v_{i3} - y_3(t_i)| = e_{i3}$ and $|v_{i4} - y_4(t_i)| = e_{i4}$.

COMPARISON OF NUMERICAL SOLUTION
WITH EXACT SOLUTION

i	t_i	v_{i3}	$e_{i3}\times 10^5$	v_{i4}	$e_{i4}\times 10^5$
0	0	0.5	.000	0.0	.000
1	1/9	.5773 4657 9715	.370	.1546 9315 9431	.729
2	2/9	.6454 9323 1862	.395	.2909 8646 3724	.803
3	1/3	.7071 0324 9064	.351	.4142 0649 8128	.711
4	4/9	.7637 5972 0205	.295	.5275 1944 0410	.594
5	5/9	.8164 9433 7894	.229	.6329 8867 5788	.465
6	2/3	.8660 2378 3185	.168	.7320 4756 6369	.337
7	7/9	.9128 6988 9405	.031	.8257 3977 8811	.021
8	8/9	.9574 2660 7059	.050	.9148 5321 4118	.106
9	1	1.0	.000	1.0	.000

Table (6.9)

The maximum error occurs in the approximation of the constant function $y_1(t) = \bar{u}(t)$, thus

$$||V - Y|| = .907 \text{ D-05}.$$

For this example, the Gap$_4$ scheme yields an extremely accurate numerical solution on a comparatively sparse net. Also, Newton's method exhibits quadratic convergence once an "acceptable" initial guess has been found.

Appendix A

Ordinary Differential Equations of Higher Order

In a recent paper H. O. Kreiss [7] examined difference schemes used to approximate linear two-point boundary-value problems:

$$\frac{d^n y}{dt^n} + \sum_{k=o}^{n-1} A_k(t) \frac{d^k y}{dt^k} = F(t) \qquad \text{a)}$$

$$\sum_{k=o}^{L} \{B_{kL}(o)\, y^{(k)}(o) + B_{kL}\, (1)\, y^{(k)}(1)\} = g_L \qquad \text{b)} \qquad (A.1)$$

$$L = o,\dots,n-1$$

where A_k are n×n matrices and y,F,g are m-vectors. At first glance, a treatment of equations (A.1a,b) seems to be more general than the equations (1.1a,b) considered in Chapter 1. However, we give an indication here that the theory of Kreiss [7] and the theory developed in Chapter 1 are actually concerned with the same class of problems.

As an approximation of (A.1a), Kreiss considers difference schemes of the form

$$L_h\, v_i = S_o(h) D_+^n\, v_{i-r} + \sum_{k=o}^{n-1} \bar{A}_k(t_i,h) D_+^k\, v_{i-r} \qquad (A.2)$$

$$i = r,\dots,J-s$$

where D_+ is the forward difference operator $(D_+\, v_{i-1} = \frac{1}{h}(v_i - v_{i-1}))$.

The discrete boundary conditions are written

$$\sum_{k=o}^{L} \{\bar{B}_{kL}(o,h)D_+^k\ v_o + \bar{B}_{kL}(1,h)D_-^k\ v_J\} = \bar{g}_L(h) \qquad (A.3)$$

$$L = o,\ldots,n-1$$

where $D_-\ v_i = D_+\ v_{i-1}$.

We will consider only the case in which (A.2) is "as compact as possible", that is, r+s=n. With this restriction, (A.2) and (A.3) represent m(J+1) equations in m(J+1) unknowns, v_i, $i = 0,\ldots,J$. For this case, the difference equation (A.2) can be written as

$$L_h\ v_i = \bar{A}_n(t_i,h)D_+^h\ v_{i-r} + \sum_{k=o}^{n-1} \bar{A}_k(t_i,h)D_+^k\ v_{i-r} = F_i \qquad (A.4)$$

$$i = r,\ldots,J-s$$

where $\bar{A}_n(t_i,h)$ is an m×m matrix and replaces the difference operator $S_o(h)$.

The boundary-value problem (A.1a,b) can be written as a first-order system by defining

$$z^{k+1}(t) = y^{(k)}(t).$$

Thus, a first-order system of equations equivalent to (A.1a,b) is

$$\frac{dz^n(t)}{dt} + \sum_{k=o}^{n-1} A_k(t)\ z^{k+1}(t) = F(t) \qquad a)$$

$$\frac{dz^k(t)}{dt} = z^{k+1}(t) \qquad k = 1,\ldots,n-1 \qquad b) \qquad (A.5)$$

$$\sum_{k=o}^{L} \{B_{kL}(o)\ z^{k+1}(o) + B_{kL}(1)\ z^{k+1}(1)\} = g_L \qquad c)$$

$$L = o,\ldots,n-1.$$

It remains to show that, in some sense, (A.2), (A.3) can be considered as an approximation of (A.5a,b,c).

Analogous to the definitions (A.5b), define the m-vectors, w_i^k, k = 1,...,n, by

$$\begin{aligned} &w_i^1 = v_i && i = o,\ldots,J && a) \\ &D_+ w_i^k = w_i^{k+1} && i = o,\ldots,J-k. && b) \end{aligned} \qquad (A.6)$$

Substituting these quantities into (A.2) and (A.3), we get

$$\bar{A}_n D_+ w_{i-r}^n + \sum_{k=o}^{n-1} \bar{A}_k w_{i-r}^{k+1} = \bar{F}_i \qquad i = r,\ldots,J-s \qquad (A.7a)$$

and

$$\sum_{k=o}^{L} \{\bar{B}_{kL}(o,h)w_o^{k+1} + \bar{B}_{kL}(1,h)w_{J-k}^{k+1}\} = \bar{g}_L(h) \qquad L = o,\ldots,n-1. \qquad (A.7b)$$

Kreiss defines the difference scheme (A.4) to be consistent with (A.1a) if

$$|\bar{A}_n - I| + \sum_{k=o}^{n-1} |\bar{A}_k(t_i,h) - A_k(t_i)| = 0(h).$$

Thus, if the difference scheme (A.4) is consistent, then clearly (A.7a), (A.6b), and (A.7b) form a consistent approximation of (A.5a,b,c). Thus, we have N equations in N unknowns ($N = m(nJ - \frac{1}{2}[n^2 - n])$), and the only difference between (A.7a,b), (A.6b) and the approximations considered in Chapter 1 is that the variable w_i^k is only defined for i = o,...,J-k+1.

In order to mimic Chapter 1, we introduce $\frac{1}{2}(n^2+n)$ new equations. This can be formally done by expanding the range of i in (A.7a) from J-s to J+r-1 and the range of each i in (A.6b) from J-k to J-1. The only restriction we place on these new equations is that

$$\bar{A}_n D_+ w^n_{i-r} + \sum_{k=o}^{n-1} \bar{A}_k w^{k+1}_{i-r} = \bar{F}_i$$

$$i = J-s+1,\ldots,J+r-1$$

be consistent with the differential equation (A.5a).

$$\bar{A}_n D_+ w^n_{i-r} + \sum_{k=o}^{n-1} \bar{A}_k w^{k+1}_{i-r} = \bar{F}_i \qquad i = r,\ldots,J+r-1 \qquad \text{a)}$$

$$D_+ w^k_i = w^{k+1}_i \qquad k = 1,\ldots,n, \qquad i = o,\ldots,J-1 \qquad \text{b)}$$

$$\sum_{k=o}^{L} \{\bar{B}_{kL}(o,h)w^{k+1}_o + \bar{B}_{kL}(1,h)w^{k+1}_{J-k}\} = \bar{g}_L(h) \qquad \text{c)} \qquad (A.9)$$

$$L = o,\ldots,n-1$$

In this way equations (A.9a,b,c) define a consistent difference approximation of (A.5a,b,c) where $W^1 = V$, V being a solution of (A.2), (A.3). Thus, a stable solution, $\{W^1,W^2,\ldots,W^n\}$, of the difference approximations (A.9a,b,c) yields a stable solution, V, of (A.2) and (A.3). In addition, the theory presented in Chapter 1 provides a necessary condition for stability where no restrictions are placed on the form of the difference approximations.

References

[1] S. D. Conte, "The Numerical Solution of Linear Boundary-Value Problems", SIAM Review, Volume 8, No. 3, (1966), pp. 309-321.

[2] L. Fox, "Numerical Solution of Ordinary and Partial Differential Equations", Addison-Wesley, London, 1962.

[3] T. R. Goodman and G. N. Lance, "The Numerical Integration of Two-Point Boundary-Value Problems", Mathematics of Computations, Volume 10, (1956), pp. 82-86.

[4] E. Isaacson and H. B. Keller, "Analysis of Numerical Methods", John Wiley, New York, 1966.

[5] H. B. Keller, "Accurate Difference Methods for Linear Ordinary Differential Systems Subject to Linear Constraints", SIAM Journal on Numerical Analysis, Volume 6, No. 1, (1969), pp. 8-30.

[6] H. B. Keller, "Accurate Difference Methods for Nonlinear Two-Point Boundary Value Problems", to appear.

[7] H. O. Kreiss, "Difference Approximations for Boundary and Eigenvalue Problems for Ordinary Differential Equations", Mathematics of Computation, Volume 26, No. 119, (1972), pp. 605-624.

[8] F. K. Moore, editor, "Theory of Laminar Flows", Princeton University Press, Princeton, 1964, pp. 171-175.

[9] V. Pereyra, "Iterated Deferred Corrections for Nonlinear Boundary Value Problems", Numerische Mathematik, Volume 11, (1968), pp. 111-125.