Computer Sciences Department
The University of Wisconsin
1210 West Dayton Street
Madison, Wisconsin 53713

THE NUMERICAL SOLUTION OF BOUNDARY VALUE PROBLEMS FOR SECOND ORDER FUNCTIONAL DIFFERENTIAL EQUATIONS BY FINITE DIFFERENCES

by

Colin W. Cryer

Computer Sciences Technical Report #127

June 1971

CONTENTS

THE NUMERICAL SOLUTION OF BOUNDARY VALUE PROBLEMS FOR SECOND ORDER FUNCTIONAL DIFFERENTIAL EQUATIONS BY FINITE DIFFERENCES

by

Colin W. Cryer*

1. INTRODUCTION

In the present paper we consider numerical methods for computing the solution of the boundary value problem

$$\left.\begin{aligned} \ddot{x}(t) &= g(t,x(t)) + (\mathfrak{F}x)(t), \quad 0 < t < 1, \\ x(0) &= x(1) = 0, \end{aligned}\right\} \tag{1.1}$$

where, $\mathfrak{F} : \mathscr{C}[0,1] \to \mathscr{C}[0,1]$, and $g: R^2 \to R^1$.

It will be assumed that (1.1) has a unique twice continuously differentiable solution which will be denoted by x throughout the paper. It will also be assumed that g is continuously differentiable and that

$$\frac{\partial g(t,y)}{\partial y} \geq \beta > -\pi^2 , \tag{1.2}$$

for $t \in [0,1]$ and all $y \in R^1$. No explicit assumptions about $\mathfrak{F}$ will be made. However, it will be assumed that $\mathfrak{F}$ can be approximated (in a sense explained later) by a Lipschitz continuous mapping $\mathfrak{F}_h$.

Examples of boundary value problems which can easily be cast into the form (1.1) are:

---

*Sponsored by the Mathematics Research Center, United States Army, Madison, Wisconsin, under Contract No.: DA-31-124-ARO-D-462, and the Office of Naval Research under Contract No.: N00014-67-A-0128-0004. The computations were supported by the University of Wisconsin Grants Committee

1. The two-point boundary value problem

$$\left.\begin{aligned} y''(s) &= g(s,y(s)),\ a < s < b, \\ y(a) &= y(b) = 0; \end{aligned}\right\} \tag{1.3}$$

2. The boundary value problem

$$\left.\begin{aligned} y''(s) &= \frac{1}{16}\sin y(s) + s - (s+1)y(s-1),\ 0 < s < 2, \\ y(s) &= s - \frac{1}{2},\ \text{if}\ \ s \le 0, \\ y(2) &= -\frac{1}{2}, \end{aligned}\right\} \tag{1.4}$$

which is a special case of problems considered by Nevers and Schmitt [18];

3. The integro-differential equation

$$y''(s) = g(s) + \int_a^b y(u)\,f(s,\ du),\ \ a < s < b, \tag{1.5}$$

where the integral is a Stieltjes integral, and where $f{:}R^2 \to R^1$ is of bounded variation.

Concerning the theory of boundary value problems for functional differential equations see Cooke [2], El'sgol'ts [4], Fennell and Waltman [5,6], Grimm and Schmitt [10,11], Halanay [12], Halanay and Yorke [13], Hale [14], Kato [16], Norkin [19], Schmitt [21]. The numerical solution of boundary value problems for delay differential equations has been considered by Nevers and Schmitt [18], while the numerical solution of initial value problems for functional differential equations has been treated by Cryer and Tavernini [3], and Tavernini [22,23].

The numerical solution of the two-point boundary value problem (1.3) by finite differences has been extensively studied (see, for example, Ciarlet et al [1], Henrici [15], and Keller [17]), and this work has guided the present paper.

Acknowledgements

We have benefited from the stimulus provided by discussions with Professor Lucio Tavernini.

2. PRELIMINARIES

Throughout the paper matrices and vectors will be understood to be $n \times n$ matrices and n-vectors, respectively.

We set $h = \frac{1}{(n+1)}$. It will often be assumed that $h \le h_0$ where

$$h_0 = \min \left\{ \left[\frac{4}{|\beta|} \; \frac{\pi^2 - |\beta|}{\pi^2 + |\beta|} \right]^{\frac{1}{2}}, \; \left[\frac{1}{|\beta|} \right]^{\frac{1}{2}} \right\}, \; -\pi^2 < \beta < 0,$$
$$= \infty, \; \beta \ge 0. \tag{2.1}$$

If $\underline{Z} = (Z_i)$ is an n-vector, then $\|Z\| = \max_i |Z_i|$.

Throughout the remainder of this section, $\underline{A} = (a_{ij})$ will denote an $n \times n$ matrix. The following norms are used:

$$\|A\| = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|,$$

$$\|A\|_s = \sum_{j=1}^{n} \max_{1 \le i \le n} |a_{ij}|,$$

$$|A| = \max_{1 \le i, j \le n} |a_{ij}|.$$

If $a_{ij} \ge 0$ we write $\underline{A} \ge 0$ and say that $\underline{A}$ is non-negative. $\underline{A}$ is monotone if $\underline{A}^{-1}$ exists and $\underline{A}^{-1} \ge 0$, and $\underline{A}$ is an M-matrix if $\underline{A}$ is monotone, $a_{ii} > 0$, and $a_{ij} < 0$ for $i \ne j$. The following theorem (Ortega and Rheinboldt [20, p. 54]) will be useful:

Theorem 2.1

Assume that $\underline{A}$ is an M-matrix, that $\underline{\tilde{A}} \geq \underline{A}$, and that $\underline{\tilde{A}}$ satisfies the sign restrictions for M-matrices. Then $\underline{\tilde{A}}$ is an M-matrix and $\underline{A}^{-1} \geq \underline{\tilde{A}}^{-1} \geq 0$.

If $\underline{A}^{-1}$ exists we set $A^{-1} = (a^{*}_{ij})$. We say that $\underline{A}$ permits of an LU factorization if $\underline{A} = \underline{L}\,\underline{U}$ where $\underline{L}$ is a lower triangular matrix with unit diagonal and $\underline{U}$ is an upper triangular matrix. We denote by $\mathcal{L}(\underline{A})$ the strictly lower triangular matrix obtained by setting equal to zero the elements of $\underline{A}$ above and on the diagonal.

We denote by $\underline{\Lambda}$ the $n \times n$ matrix

$$\underline{\Lambda} = \begin{pmatrix} & & & & 1 \\ & 0 & & 1 & \\ & & \cdot & & \\ & \cdot & & 0 & \\ 1 & & & & \end{pmatrix},$$

and set $\underline{A}^{\Lambda} = \underline{\Lambda}\,\underline{A}\,\underline{\Lambda}$: $\underline{A}^{\Lambda}$ is obtained from $\underline{A}$ by inverting the order of the rows and columns of $\underline{A}$. The next lemma follows immediately from the definitions.

Lemma 2.2

(i) $\underline{\Lambda}^2 = \underline{I}$; (ii) $\|\underline{A}^{\Lambda}\| = \|\underline{A}\|$; (iii) If $\underline{A}$ is non-singular, $(\underline{A}^{\Lambda})^{-1} = (\underline{A}^{-1})^{\Lambda}$; (iv) If $\underline{A}$ is lower triangular (upper triangular) then $\underline{A}^{\Lambda}$ is upper triangular (lower triangular): (v) $(\underline{A}^{\Lambda})^{\Lambda} = \underline{A}$.

We will need the following elementary lemmas:

Lemma 2.3

Assume that $0 \leq \psi \leq \frac{\pi}{2}$ and $t \geq 1$. Then $\sin(t\psi) \leq t \sin \psi$.

Proof: It clearly suffices to consider the case when $t\psi \leq \frac{\pi}{2}$. Set

$$g(t) = \sin(t\psi) - t \sin \psi .$$

Then $g(1) = 0$ and

$$\begin{aligned} g'(t) &= \psi \cos(t\psi) - \sin \psi, \\ &\leq \psi \cos \psi - \sin \psi, \\ &= \cos \psi (\psi - \tan \psi), \\ &\leq 0. \end{aligned}$$

The lemma follows.

Lemma 2.4

Let $0 \leq a \leq 1$. Then $\operatorname{arc\,cos}(1 - a) \leq \left[\frac{2a}{1 - \frac{a}{2}}\right]^{\frac{1}{2}}$.

Proof:

$$\operatorname{arc\,cos}(1 - a) = \int_{1-a}^{1} \frac{dt}{[1 - t^2]^{\frac{1}{2}}} .$$

Setting $u = 1-t$,

$$\arc\cos(1-a) = \int_0^a \frac{du}{[2u - u^2]^{\frac{1}{2}}},$$

$$= \int_0^a \frac{du}{[2u]^{\frac{1}{2}}\,[1 - \frac{u}{2}]^{\frac{1}{2}}},$$

$$\leq \frac{1}{[1 - \frac{a}{2}]^{\frac{1}{2}}} \int_0^a \frac{du}{[2u]^{\frac{1}{2}}},$$

$$= \left[\frac{2a}{1 - \frac{a}{2}}\right]^{\frac{1}{2}}.$$

Lemma 2.5

Assume that $\psi \geq 0$ and $0 < a \leq 1$. Then

$$\frac{\sinh \psi}{\cosh \psi \ \sinh a\psi} \leq \frac{1}{a}.$$

Proof: Set

$$f(\psi) = \frac{\sinh \psi}{\cosh \psi \sinh a\psi}.$$

Then $f(0+) = \frac{1}{a}$, and

$$f'(\psi)\,[\cosh\psi \sinh a\psi]^2$$

$$= \cosh^2\psi \sinh a\psi - \sinh^2\psi \sinh a\psi - a \sinh\psi \cosh\psi \cosh a\psi,$$

$$= \sinh a\psi - a \sinh\psi \cosh\psi \cosh a\psi.$$

But, $\sinh a\psi \le a \sinh\psi$ so that $f'(\psi) \le 0$. The lemma follows.

Lemma 2.6

If $\underline{A}$ and $\underline{B}$ are $n \times n$ matrices then $\|\underline{A}\| \le n\,|\underline{A}|$; $\|\underline{A}\|_s \le n\,|\underline{A}|$; $\|\underline{A}\underline{B}\|_s \le \|\underline{A}\|\ \|\underline{B}\|_s$; $\|\underline{A}\| \le \|\underline{A}\|_s$; $|\underline{A}\underline{B}| \le \|\underline{A}\|\ |\underline{B}|$; and $|\underline{A}\underline{B}| \le |\underline{A}|\ \|\underline{B}^T\|$. If $\underline{H}$ is an $n \times n$ strictly lower triangular matrix then $\|\underline{H}\|_s \le (n-1)\ |\underline{H}|$.

Proof:

$$\|\underline{A}\underline{B}\|_s = \sum_{j=1}^{n} \max_{1\le i\le n} \left|\sum_{k=1}^{n} a_{ik} b_{kj}\right|,$$

$$\le \sum_{j=1}^{n} \max_{1\le i\le n} \|\underline{A}\| \max_{1\le k\le n} |b_{kj}|,$$

$$= \|\underline{A}\|\ \|\underline{B}\|_s,$$

as asserted. The other inequalities follow similarly.

3. LU FACTORIZATION OF JACOBI MATRICES

Let $\underline{J}$ be the tri-diagonal or Jacobi matrix,

$$\underline{J} = \begin{pmatrix} a_1 & -b_1 & 0 & 0 & . & . & . & 0 \\ -c_1 & a_2 & -b_2 & 0 & . & . & . & 0 \\ 0 & -c_2 & a_3 & -b_3 & . & . & . & 0 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & -b_{n-1} \\ 0 & 0 & \vdots & . & & & -c_{n-1} & a_n \end{pmatrix} . \tag{3.1}$$

For $1 \le k \le n$ we denote by D_k the determinant of the $k \times k$ submatrix of $\underline{J}$ formed by deleting the last $n - k$ rows and columns; we set $D_0 = 1$ and $D_{-1} = 0$.

Lemma 3.1

Assume that $D_k \neq 0,\ 1 \le k \le n$. Then $\underline{J}$ permits an LU factorization and $\underline{L} = (\ell_{ij})$ and $\underline{U} = (u_{ij})$ are tri-diagonal matrices with coefficients

$$\ell_{kk} = 1 \text{ and } u_{kk} = D_k/D_{k-1},\quad 1 \le k \le n;$$

$$\ell_{k+1,k} = -c_k\, D_{k-1}/D_k \text{ and } u_{k,k+1} = -b_k,\ 1 \le k \le n-1;$$

$$\ell^*_{ij} = (D_{j-1}/D_{i-1}) \prod_{k=j}^{i-1} c_k,\quad j \le i;$$

$$u^*_{ij} = (D_{i-1}/D_j) \prod_{k=i}^{j-1} b_k,\quad j \ge i;$$

where

$$\prod_{k=m+1}^{m} c_k = \prod_{k=m+1}^{m} b_k = 1$$

Proof: Follows immediately from Gantmacher [8, p. 35].

Lemma 3.2

Let $\underset{\sim}{J}$ and $\underset{\sim}{\tilde{J}}$ be Jacobi matrices with non-negative diagonal elements and non-positive off-diagonal elements. Assume that $\underset{\sim}{J} \geq \underset{\sim}{\tilde{J}}$ and that $\tilde{D}_k > 0$ for $1 \leq k \leq n$. Then: (i) $\underset{\sim}{J}$ and $\underset{\sim}{\tilde{J}}$ are M-matrices; (ii) $\underset{\sim}{J}$ and $\underset{\sim}{\tilde{J}}$ permit LU factorizations; (iii) $\underset{\sim}{U}$, $\underset{\sim}{\tilde{U}}$, $\underset{\sim}{L}$, and $\underset{\sim}{\tilde{L}}$ are Jacobi M-matrices and $\underset{\sim}{U} \geq \underset{\sim}{\tilde{U}}$, $\underset{\sim}{L} \geq \underset{\sim}{\tilde{L}}$; (iv) $\underset{\sim}{\tilde{J}}^{-1} \geq \underset{\sim}{J}^{-1} \geq 0$; $\underset{\sim}{\tilde{U}}^{-1} \geq \underset{\sim}{U}^{-1} \geq 0$; $\underset{\sim}{\tilde{L}}^{-1} \geq \underset{\sim}{L}^{-1} \geq 0$.

Proof: We assert that

$$D_{k+1}/D_k \geq \tilde{D}_{k+1}/\tilde{D}_k > 0, \tag{3.2}$$

for $0 \leq k \leq n-1$. To see this we first note that $D_1/D_0 = a_1/1 \geq \tilde{a}_1/1 = \tilde{D}_1/\tilde{D}_0 > 0$. so that (3.2) holds for $k = 0$. Assume that (3.2) holds for $0 \leq k \leq m-1$ where $m \geq 1$. Then, using the recurrence relations for D_k and $\tilde{D}_k$ (Gantmacher and Krein [9, p. 77]), and remembering that $a_k \geq \tilde{a}_k \geq 0$, $0 \leq b_k \leq \tilde{b}_k$, and $0 \leq c_k \leq \tilde{c}_k$,

$$\begin{aligned} D_{m+1}/D_m & \\ &= a_m - b_{m-1}\, c_{m-1}\, [D_m/D_{m-1}]^{-1}, \\ &\geq \tilde{a}_m - \tilde{b}_{m-1}\, \tilde{c}_{m-1}\, [D_m/D_{m-1}]^{-1}, \\ &\geq \tilde{a}_m - \tilde{b}_{m-1}\, \tilde{c}_{m-1}\, [\tilde{D}_m/\tilde{D}_{m-1}]^{-1}, \\ &= \tilde{D}_{m+1}/\tilde{D}_m, \end{aligned}$$

so that (3.2) holds for $k = m$. Using induction, the assertion follows.

Since $D_k > 0$ and $\tilde{D}_k > 0$ for $1 \leq k \leq n$, $\underline{J}$ and $\tilde{\underline{J}}$ are "sign-regular" (Gantmacher and Krein [9, p. 94], and hence $\underline{J}$ and $\tilde{\underline{J}}$ are M-matrices.

Since $D_k > 0$ and $\tilde{D}_k > 0$ for $1 \leq k \leq n$, it follows from Lemma 3.1 that $\underline{J}$ and $\tilde{\underline{J}}$ permit LU factorizations. Using (3.2) and the explicit representations for $\underline{U}$, $\underline{L}$, etc. in Lemma 3.1 it is easily seen that $\underline{U} \geq \tilde{\underline{U}}$, and $\underline{L} \geq \tilde{\underline{L}}$. Moreover, $\underline{U}^{-1}$, $\tilde{\underline{U}}^{-1}$, $\underline{L}^{-1}$, $\tilde{\underline{L}}^{-1} \geq 0$. Since the sign restrictions for M-matrices are satisfied, $\underline{U}$, $\tilde{\underline{U}}$, $\underline{L}$, and $\tilde{\underline{L}}$ are M-matrices. The remaining assertions of the lemma follow from Theorem 2.1.

4. PROPERTIES OF $\underline{J}_\alpha$.

In this section we consider the special $n \times n$ Jacobi matrices $\underline{J}_\alpha = (\gamma_{\alpha;ij})$,

$$\underline{J}_\alpha = \begin{pmatrix} \alpha & -1 & & & & & 0 \\ -1 & \alpha & -1 & & & & \\ & -1 & \cdot & \cdot & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \cdot & -1 & & \\ & 0 & & -1 & \alpha & -1 \\ & & & & -1 & \alpha \end{pmatrix} . \tag{4.1}$$

The corresponding determinants and LU factors are denoted by $D_{\alpha,k}$, $\underline{L}_\alpha = (\ell_{\alpha;ij})$, and $\underline{U}_\alpha = (u_{\alpha;ij})$, respectively.

We shall use the functions

$$\left.\begin{aligned} \theta = \theta(\alpha) &= \text{arc cosh}(\alpha/2), & \alpha > 2, \\ &= 0 & \alpha = 2, \\ &= \text{arc cos}\ (\alpha/2), & 0 < \alpha < 2, \end{aligned}\right\} \tag{4.2}$$

$$\varphi = \varphi(\alpha) = (n+1)\theta(\alpha) = \frac{1}{h}\ \theta(\alpha), \tag{4.3}$$

$$\left.\begin{aligned} d(\alpha, s) &= \sinh[(s+1)\theta(\alpha)]/\sinh[\theta(\alpha)], & \alpha > 2, \\ &= s + 1 & \alpha = 2, \\ &= \sin[(s+1)\theta(\alpha)]/\sin[\theta(\alpha)], & 0 < \alpha < 2. \end{aligned}\right\} \tag{4.4}$$

The following result is proved by Fischer and Usmani [7].

Lemma 4.1

Assume that $\alpha > 0$. Then: $D_{\alpha,k} = d(\alpha,k)$, for $-1 \le k \le n$;

$$\gamma^*_{\alpha;ij} = d(\alpha,j-1)d(\alpha,n-i)/d(\alpha,n), \qquad 1 \le j \le i \le n,$$
$$= d(\alpha,i-1)d(\alpha,n-j)/d(\alpha,n), \qquad 1 \le i \le j \le n.$$

Using Lemma 4.1 we obtain

Lemma 4.2

$$|J_\alpha^{-1}| \le \frac{\sinh(\varphi/2)}{2\sinh\theta\cosh(\varphi/2)} \le \frac{(n+1)}{4}, \qquad \alpha > 2,$$

$$\le (n+1)/4 \qquad , \qquad \alpha = 2,$$

$$\le \frac{\sin(\varphi/2)}{2\sin\theta\cos(\varphi/2)} \le \frac{(n+1)}{4\cos(\varphi/2)}, \quad 0 < \alpha < 2 \text{ and } 0 < \varphi \le \frac{\pi}{2},$$

$$\le \frac{n+1}{2\sqrt{2}\cos(\varphi/2)} \qquad , \quad 0 < \alpha < 2 \text{ and } \frac{\pi}{2} \le \varphi < \pi.$$

Proof: We begin by noting that if $\alpha \ge 2$ or $\alpha < 2$ and $\varphi \le \frac{\pi}{2}$ then $d(\alpha,s)$ is a monotone increasing function of s so that

$$|J_\alpha^{-1}| = \max_{1 \le i \le n} |\gamma^*_{\alpha;ii}| .$$

We now consider the four cases separately.

Case 1: $\alpha \geq 2$. Then

$$\gamma^*_{\alpha;ii} = \frac{\sinh(\varphi - i\theta)\sinh i\theta}{\sinh\theta \sinh\varphi},$$

$$= \frac{\cosh\varphi - \cosh(\varphi - 2i\theta)}{2\sinh\theta \sinh\varphi},$$

$$\leq \frac{\cosh\varphi - 1}{2\sinh\theta \sinh\varphi},$$

$$= \frac{2\sinh^2(\varphi/2)}{4\sinh\theta \sinh(\varphi/2)\cosh(\varphi/2)},$$

$$= \frac{\sinh(\varphi/2)}{2\sinh\theta \cosh(\varphi/2)}.$$

Appealing to Lemma 2.5 (or noting that $I_\alpha > I_2$ so that $I_\alpha^{-1} \leq I_2^{-1}$),

$$\gamma^*_{\alpha;ii} \leq \frac{1}{2}\,\frac{1}{\theta/(\varphi/2)},$$

$$= \frac{(n+1)}{4}.$$

Case 2: $\alpha = 2$. Then

$$\gamma^*_{\alpha;ii} = \frac{(n+1-i)\,i}{n+1},$$

$$\leq \frac{(n+1)}{4}.$$

Case 3: $\alpha < 2$ and $0 \le \varphi \le \frac{\pi}{2}$. Then

$$\gamma^*_{\alpha;ii} = \frac{\sin(\varphi - i\theta)\sin i\theta}{\sin\theta \quad \sin\varphi} ,$$

$$= \frac{\cos(\varphi - 2i\theta) - \cos\varphi}{2 \sin\theta \quad \sin\varphi} ,$$

$$\le \frac{1 - \cos\varphi}{2\sin\theta \sin\varphi} ,$$

$$= \frac{2\sin^2(\varphi/2)}{4\sin\theta\sin(\varphi/2)\cos(\varphi/2)} ,$$

$$= \frac{\sin(\varphi/2)}{2\sin\theta \quad \cos(\varphi/2)} .$$

Appealing to Lemma 2.3,

$$\gamma^*_{\alpha;ii} \le \frac{(\varphi/2\theta)\sin\theta}{2\sin\theta\cos(\varphi/2)} ,$$

$$= \frac{(n+1)}{4\cos(\varphi/2)} .$$

Case 4: $0 < \alpha < 2$ and $\pi/2 \le \varphi \le \pi$. Then

$$\gamma^*_{\alpha;ij} = \frac{\sin p\theta \quad \sin q\theta}{\sin\theta \quad \sin\varphi} ,$$

where p and q are positive integers depending on i and j satisfying $p + q \le n + 1$. Without loss of generality we may assume that $p \le (n+1)/2$. Using Lemma 2.3,

$$\begin{aligned}
\gamma^*_{\alpha;ij} &\le \frac{p \sin\theta \ \sin q\theta}{\sin\theta \ \sin\varphi}, \\
&\le \frac{p}{\sin\varphi}, \\
&= \frac{p}{2\sin(\varphi/2)\cos(\varphi/2)}, \\
&\le \frac{p}{\sqrt{2}\ \cos(\varphi/2)}, \\
&\le \frac{n+1}{2\sqrt{2}\ \cos(\varphi/2)} .
\end{aligned}$$

Theorem 4.3

Assume that $\alpha \ge 2 + h^2\beta$ and that $h \le h_0$. Then :(i) $\underset{\sim}{J}_\alpha$ is an M-matrix; (ii) $D_{\alpha,k} > 0$ for $1 \le k \le n$; (iii) $|\underset{\sim}{J}_\alpha^{-1}| \le K(\beta)/h$, where

$$\begin{aligned}
K(\beta) &= \frac{1}{4}, \quad \beta \ge 0, \\
&= \left\{ 2\sqrt{2} \ \cos\left(\frac{1}{2}\left[\frac{\pi^2 + |\beta|}{2}\right]^{\frac{1}{2}}\right)\right\}^{-1}, \quad 0 > \beta > -\pi^2 .
\end{aligned}$$

Proof: It follows from (2.1) that $\alpha \ge 1$. Hence, using (4.4) and Lemma 4.1 we see that $D_{\alpha,k} > 0$ for $1 \le k \le n$. Applying Lemma 3.2 with $\underset{\sim}{J} = \tilde{\underset{\sim}{J}} = \underset{\sim}{J}_\alpha$, it follows that $\underset{\sim}{J}_\alpha$ is an M-matrix.

If $\alpha \ge 2$ then, from Lemma 4.2, $|\underset{\sim}{J}_\alpha^{-1}| \le 1/(4h)$. On the other hand, if $\alpha < 2$ then $\beta < 0$ so that from (2.1) and Lemma 2.4,

$$\begin{aligned}
\varphi &= (n+1)\theta\,, \\
&= (n+1)\ \mathrm{arc\,cos}\ (\alpha/2)\,, \\
&\leq (n+1)\ \mathrm{arc\,cos}\ (1+\beta h^2/2)\,, \\
&\leq (n+1)\left[\frac{|\beta| h^2}{1-|\beta| h^2/4}\right]^{\frac{1}{2}}\,, \\
&= \left[\frac{|\beta|}{1-|\beta| h^2/4}\right]^{\frac{1}{2}}\,, \\
&\leq \left[\frac{|\beta|}{1-|\beta| h_0^2/4}\right]^{\frac{1}{2}}\,, \\
&\leq \left[\frac{\pi^2+|\beta|}{2}\right]^{\frac{1}{2}}\,,
\end{aligned}$$

so that $\varphi < \pi$. Hence, from Lemma 4.2,

$$|J_\alpha^{-1}| \leq \frac{(n+1)}{2\sqrt{2}\ \cos(\varphi/2)} \leq K(\beta)/h,$$

and the proof of the theorem is complete.

5. PERTURBATIONS OF MONOTONE MATRICES

Lemma 5.1

Assume that $\underline{H}$ is a strictly lower triangular matrix. Then $(\underline{I} - \underline{H})^{-1}$ and $(\underline{I} - \underline{H}^T)^{-1}$ exist and satisfy

$$\|(\underline{I} - \underline{H})^{-1}\| \le \exp[\|\underline{H}\|_s],$$
$$\|(\underline{I} - \underline{H}^T)^{-1}\| \le \exp[\|\underline{H}^T\|_s].$$

Proof: Let $\underline{H}^{(k)}$ denote the matrix obtained from $\underline{H}$ by setting equal to zero all the elements of $\underline{H}$ except those in the k-th column. Then

$$(\underline{I} + \underline{H}^{(n-1)}) \ldots (\underline{I} + \underline{H}^{(1)})(\underline{I} - \underline{H}) = \underline{I}\,.$$

Hence,

$$\begin{aligned}
&\|(\underline{I} - \underline{H})^{-1}\| \\
&= \|(\underline{I} + \underline{H}^{(n-1)}) \ldots (\underline{I} + \underline{H}^{(1)})\|, \\
&\le \prod_{k=1}^{n-1} \|\underline{I} + \underline{H}^{(k)}\|, \\
&\le \prod_{k=1}^{n-1} [1 + \|\underline{H}^{(k)}\|_s], \\
&\le \prod_{k=1}^{n-1} \exp[\|\underline{H}^{(k)}\|_s], \\
&= \exp[\sum_{k=1}^{n-1} \|\underline{H}^{(k)}\|_s], \\
&= \exp[\|\underline{H}\|_s],
\end{aligned}$$

as asserted, since $\|\underline{H}\|_s = \sum_{k=1}^{n-1} \|\underline{H}^{(k)}\|_s$.

Let $\underline{\tilde{H}}^{(k)}$ denote the matrix obtained from $\underline{H}^T$ by setting equal to zero all the elements of $\underline{H}^T$ except those in the k-th column. Then

$$(\underline{I} + \underline{\tilde{H}}^{(1)}) \ldots (\underline{I} + \underline{\tilde{H}}^{(n-1)}) \quad (\underline{I} - \underline{H}^T) = \underline{I} .$$

Repeating the previous arguments, the second inequality follows.

Theorem 5.2

Assume that $\underline{A}$, $\underline{P}$, $\underline{L}$, and $\underline{U}$ are matrices such that :(i) $\underline{P} = (p_{ij})$ is non-negative and lower triangular; (ii) $\underline{A} = \underline{U}\,\underline{L}$; (iii) $\underline{L} = (\ell_{ij})$ is a lower triangular M-matrix; (iv) $\underline{U}$ is an upper triangular tri-diagonal M-matrix with unit diagonal. Then $(\underline{A} + \underline{P})^{-1}$ exists and

$$\|(\underline{A} + \underline{P})^{-1}\| \le \|\underline{A}^{-1}\| \exp\left[\|\underline{P}\,\underline{A}^{-1}\|_s/(1+\kappa)\right]/(1+\kappa),$$

$$|(\underline{A} + \underline{P})^{-1}| \le |\underline{A}^{-1}| \exp\left[\|(\underline{P}\,\underline{A}^{-1})^T\|_s/(1+\kappa)\right],/(1+\kappa),$$

where

$$0 \le \kappa \le \min_i (\ell^*_{ii}\, p_{ii}) .$$

Proof: Let $E = (e_{ij}) = \underline{P}\,\underline{L}^{-1}$. Then $\underline{A} + \underline{P} = (\underline{U} + \underline{E})\,\underline{L}$. Since $\underline{P}$ and $\underline{L}^{-1}$ are non-negative lower triangular matrices, the same is true of E. Moreover, $e_{ii} = p_{ii}\, \ell^*_{ii} \ge \kappa$.

The off-diagonal elements elements of $\underline{U}$ will be denoted by $-u_i$. Thus, a typical $\underline{U} + \underline{E}$ is of the form

$$\begin{pmatrix} 1+e_{11} & -u_1 & 0 & 0 \\ e_{21} & 1+e_{22} & -u_2 & 0 \\ e_{31} & e_{32} & 1+e_{33} & -u_3 \\ e_{41} & e_{42} & e_{43} & 1+e_{44} \end{pmatrix}$$

Now consider the process whereby $\underline{U} + \underline{E}$ is transformed into a lower triangular matrix, $\underline{I} + \underline{G}$ say, by using column operations to successively "kill off" the elements $-u_1, -u_2, \ldots, -u_{n-1}$, of $\underline{U} + \underline{E}$. It is easily seen that

$$\underline{I} + \underline{G} = (\underline{U} + \underline{E})\,\underline{V},$$
$$\underline{V} = (\underline{I} + \underline{V}^{(1)}) \ldots (\underline{I} + \underline{V}^{(n-1)}),$$
$$0 \le \underline{V}^{(k)} \le \underline{Q}^{(k)}.$$

where $\underline{Q}^{(k)} = (q_{ij}^{(k)})$ is defined by

$$q_{ij}^{(k)} = u_k, \quad \text{if } i = k \text{ and } j = k+1,$$
$$= 0, \quad \text{otherwise.}$$

Hence,

$$\underline{V} \le (\underline{I} + \underline{Q}^{(1)}) \ldots (\underline{I} + \underline{Q}^{(n-1)}),$$
$$= \underline{U}^{-1}.$$

Therefore, remembering that $\underline{U}$ and $\underline{V}$ are upper triangular,

$$\underline{I} + \underline{G} = (\underline{U} + \underline{E})\,\underline{V},$$
$$= \underline{I} + \underline{D} + \mathcal{L}(\underline{E}\,\underline{V}),$$
$$= \underline{I} + \underline{D} + \mathcal{L}(\underline{P}\,\underline{L}^{-1}\,\underline{V}),$$
$$= (\underline{I} + \underline{D})\,(\underline{I} + \underline{H}), \text{ say,}$$

where

$$\underline{D} \geq \operatorname{diag}(\underline{U} + \underline{E}) \geq (1 + \kappa)\underline{I} \, ,$$

is a diagonal matrix, and

$$\begin{aligned} \underline{H} &= (\underline{I} + \underline{D})^{-1} \mathcal{L}(\underline{P}\,\underline{L}^{-1}\,\underline{V}), \\ &\leq \mathcal{L}(\underline{P}\,\underline{L}^{-1}\,\underline{U}^{-1})/(1 + \kappa), \\ &= \mathcal{L}(\underline{P}\,\underline{A}^{-1})/(1 + \kappa), \end{aligned}$$

is nonnegative and strictly lower triangular.

Now,

$$\begin{aligned} (\underline{A} + \underline{P}) &= (\underline{U} + \underline{E})\,\underline{L}, \\ &= (\underline{I} + \underline{G})\,\underline{V}^{-1}\,\underline{L}, \\ &= (\underline{I} + \underline{D})(\underline{I} + \underline{H})\,\underline{V}^{-1}\,\underline{L}. \end{aligned}$$

Hence, $(\underline{A} + \underline{P})^{-1}$ exists and

$$(\underline{A} + \underline{P})^{-1} = \underline{L}^{-1}\,\underline{V}\,(\underline{I} + \underline{H})^{-1}\,(\underline{I} + \underline{D})^{-1}.$$

Using Lemma 5.1,

$$\begin{aligned} \|(\underline{I} + \underline{H})^{-1}\| &\leq \exp[\|\underline{H}\|_s], \\ &\leq \exp[\|\underline{P}\,\underline{A}^{-1}\|_s/(1 + \kappa)], \end{aligned}$$

and

$$\begin{aligned} \|(\underline{I} + \underline{H}^T)^{-1}\| &\leq \exp[\|\underline{H}^T\|_s], \\ &\leq \exp[\|(\underline{P}\,\underline{A}^{-1})^T\|_s/(1 + \kappa)]. \end{aligned}$$

Hence, using Lemma 2.6,

$$\|(\underline{A} + \underline{P})^{-1}\|$$

$$\leq \|\underline{L}^{-1}\,\underline{V}\| \quad \|(\underline{I} + \underline{H})^{-1}\| \quad \|(\underline{I} + \underline{D})^{-1}\| \;,$$

$$\leq \|\underline{L}^{-1}\,\underline{U}^{-1}\| \quad \|(\underline{I} + \underline{H})^{-1}\| \quad /(1 + \kappa),$$

$$\leq \|\underline{A}^{-1}\| \;\exp\,[\,\|\underline{P}\;\underline{A}^{-1}\|_s/(1 + \kappa)]\,/\,(1 + \kappa),$$

and

$$|(\underline{A} + \underline{P})^{-1}|$$

$$\leq |\underline{L}^{-1}\,\underline{V}| \quad \|[(\underline{I} + \underline{H})^{-1}\,(\underline{I} + \underline{D})^{-1}]^T\|,$$

$$\leq |\underline{L}^{-1}\underline{U}^{-1}| \quad \|(\underline{I} + \underline{H}^T)^{-1}\| \quad \|(\underline{I} + \underline{D})^{-1}\|,$$

$$\leq |\underline{A}^{-1}| \;\exp\,[\,\|(\underline{P}\;\underline{A}^{-1})^T\|_s\,/\,(1 + \kappa)]\,/\,(1 + \kappa),$$

as asserted.

6. THE NUMERICAL METHOD

The interval $[0,1]$ is divided into $n+1$ subintervals each of length h the points of subdivision, or gridpoints, being denoted by $t_{h,i} = ih$, $0 \le i \le n+1$. The solution x of (1.1) is approximated at the n interior gridpoints.

The mappings $\varphi_h : \mathscr{C}[0,1] \to R^n$ and $G_h : R^n \to R^n$ are defined by

$$(\varphi_h\, y)_i = y(t_{h,i}), \quad 1 \le i \le n, \tag{6.1}$$

$$(G_h \underline{Y})_i = g(t_{h,i}, Y_i), \quad 1 \le i \le n, \tag{6.2}$$

while $\underline{J}_\alpha$ is the $n \times n$ matrix (4.1).

We assume that $F_h : R^n \to R^n$ is an approximation to $\mathfrak{F}$ such that

$$F_h \varphi_h\, x - \varphi_h \mathfrak{F} x = \underline{\eta}_h(x), \tag{6.3}$$

where $\|\underline{\eta}_h(x)\| \to 0$ as $h \to 0$. We observe that mappings F_h satisfying (6.3) are easily constructed for problems (1.3), (1.4), and (1.5).

Then the approximation to x at the n interior gridpoints is taken to be the solution $\underline{Z}_h \in R^n$ of the system of n nonlinear equations

$$\Phi_h\, \underline{Z}_h = \underline{J}_2 \underline{Z}_h + h^2 G_h \underline{Z}_h + h^2 F_h \underline{Z}_h = 0 \;, \tag{6.4}$$

provided that $\underline{Z}_h$ exists and is unique.

We set $\underline{X}_h = \varphi_h x$, and $\underline{E}_h = \underline{X}_h - \underline{Z}_h$. Since x is twice continuously differentiable,

$$h^2 \varphi_h \ddot{x} = -\underline{J}_2 \underline{X}_h + h^2 \underline{\tau}_h(x), \tag{6.5}$$

where $\|\underline{\tau}_h(x)\| \to 0$ as $h \to 0$. Using (1.1), (6.3), and (6.5), we see that

$$\Phi_h \underline{X}_h = h^2 \underline{\epsilon}_h(x), \tag{6.6}$$

where $\underline{\epsilon}_h(x) = \underline{\tau}_h(x) + \underline{\eta}_h(x)$ so that $\|\underline{\epsilon}_h(x)\| \to 0$ as $h \to 0$.

In general, $\mathfrak{F}x$ will have discontinuous derivatives at certain interior points of $[0,1]$. For example, for problem (1.4), $\mathfrak{F}x$ has a discontinuous derivative at $t = \frac{1}{2}$ (see (7.3)). To allow for this, we set

$$\underline{\epsilon}_h(x) = \underline{\epsilon}_h^{(1)}(x) + \underline{\epsilon}_h^{(2)}(x). \tag{6.7}$$

We assume that $\underline{\epsilon}^{(1)}(x)$ has at most m non-zero components, and also that

$$\|\underline{\epsilon}_h^{(1)}(x)\| \le K_1 h^{r-1}, \quad \|\underline{\epsilon}_h^{(2)}(x)\| \le K_2 h^r, \tag{6.8}$$

where m, K_1, K_2, and r are constants. The idea is that $\underline{\epsilon}^{(1)}(x)$ is the truncation error at the gridpoints where $\mathfrak{F}x$ is non-smooth, while $\underline{\epsilon}_h^{(2)}(x)$ is the truncation error at the remaining gridpoints.

<u>Theorem 6.1</u>

Assume that $h \le h_0$, that F_h has a continuous Frechet derivative F_h', and that for all $\underline{Y} \in R^n$ $F_h'(\underline{Y})$ is a non-negative lower triangular matrix satisfying $\|F_h'(\underline{Y})\| \le M$, where M is a constant independent of h and $\underline{Y}$.

Then, $\Phi_h'(\underline{Y})^{-1}$ exists for all $\underline{Y} \in R^n$ and

$$\|\Phi_h'(\underline{Y})^{-1}\| \le K(\beta) \exp [M \ K(\beta)]/h^2,$$

where $K(\beta)$ is defined in Theorem 4.3.

Also, there exists a unique solution $\underline{Z}_h$ of (6.4) and the error $\underline{E}_h$ satisfies

$$\|\underline{E}_h\| \le K(\beta) \exp [M K(\beta)] \| \underline{\epsilon}_h(x) \| .$$

If there is a constant M_T such that $\| (F_h'(\underline{Y}))^T \|_s \le M_T$ for all $\underline{Y}$, then

$$\|\underline{E}_h\| \le h^r K(\beta) \{mK_1 \exp[M_T K(\beta)] + K_2 \exp [M K(\beta)]\},$$

where K_1, K_2, m, and r, are as in (6.8).

Finally, if g and F_h are twice continuously differentiable, $\underline{Z}_h$ may be computed by Newton's method,

$$\underline{Z}_h^{(k+1)} = \underline{Z}_h^{(k)} - [\Phi_h' (\underline{Z}_h^{(k)})]^{-1} \quad \Phi_h (\underline{Z}_h^{(k)}), \tag{6.9}$$

provided that the initial approximation $\underline{Z}_h^{(0)}$ is sufficiently good.

<u>Proof</u>: For any $\underline{Y} \in R^n$,

$$\Phi_h'(\underline{Y}) = [\underline{J}_2 + h^2 G_h'(\underline{Y})] + h^2 F_h'(\underline{Y}),$$

$$= \underline{A} + \underline{P}, \text{ say.}$$

Set $\alpha = 2 + h^2\beta$. Since $G_h'(\underline{Y}) = \text{diag}\,(\frac{\partial g(t_{h,i}, Y_i)}{\partial y}) \ge \beta \underline{I}$, it follows that $\underline{A} \ge \underline{J}_\alpha$. Using Theorems 2.1 and 4.3 we see that $\underline{A}^{-1}$ exists and that $|\underline{A}^{-1}| \le |\underline{J}^{-1}_{2+h^2\beta}| \le K(\beta)/h$. From Lemma 2.6 it follows that $\|\underline{A}^{-1}\| \le K(\beta)/h^2$, $\|\underline{A}^{-1}\|_s \le K(\beta)/h^2$.

From Theorem 4.3 we know that $D_{\alpha,k} > 0$ for $1 \leq k \leq n$. Therefore, using Lemma 3.2 with $\underline{\tilde{J}} = \underline{J}_\alpha$ and $\underline{J} = \underline{A}$, we conclude that $\underline{A} = \underline{L}_1 \underline{U}_1$ where $\underline{L}_1$ and $\underline{U}_1$ are tri-diagonal M-matrices which are, respectively, lower and upper triangular. Invoking Lemma 2.2,

$$\underline{A} = \underline{A}^{\Lambda} = \underline{L}_1^{\Lambda} \, \underline{U}_1^{\Lambda} = (\underline{L}_1^{\Lambda} \, \underline{D})(\underline{D}^{-1} \underline{U}_1^{\Lambda}) = \underline{U} \, \underline{L}, \text{ say,}$$

where $\underline{D} = \text{diag}\,(\underline{U}_1^{\Lambda})$.

It is now easily seen that all the conditions of Theorem 5.2 are satisfied. Hence, using Lemma 2.6 and setting $\kappa = 0$, we see that $\Phi_h'(\underline{Y})^{-1}$ exists and satisfies

$$\begin{aligned} &\| \Phi_h'(\underline{Y})^{-1} \| \\ &\qquad \leq \|\underline{A}^{-1}\| \exp [\, \|\underline{P}\,\underline{A}^{-1}\|_s], \\ &\qquad \leq \|\underline{A}^{-1}\| \exp [\, \|\underline{P}\| \;\; \|\underline{A}^{-1}\|_s], \\ &\qquad \leq (K(\beta)\, h^{-2}) \exp [(h^2 M)(K(\beta) h^{-2})], \\ &\qquad = K(\beta) \exp [M\, K(\beta)] \,/h^2, \end{aligned}$$

as asserted. Moreover, remembering that $\underline{A}$ is symmetric,

$$\begin{aligned} &| \Phi_h'(\underline{Y})^{-1} | \\ &\qquad \leq |\underline{A}^{-1}| \;\; \exp [\, \|(\underline{P}\,\underline{A}^{-1})^T\|_s], \\ &\qquad \leq |\underline{A}^{-1}| \exp [\, \|(\underline{A}^{-1})^T\| \;\; \|\underline{P}^T\|_s], \\ &\qquad = K(\beta) \;\; \exp [K\,(\beta)\; M_T]/h. \end{aligned}$$

Since $\Phi_h'(\underline{Y})$ exists and is bounded for all $\underline{Y}$, it follows from the theorem of Hadamard (Ortega and Rheinboldt [20, p. 137]) that there exists a unique solution $\underline{Z}_h$ of (6.4).

From (6.4) and (6.6)

$$\Phi_h \underline{X}_h - \Phi_h \underline{Z}_h = h^2 \underline{\epsilon}_h(x),$$

so that (Ortega and Rheinboldt [20, p. 71])

$$(\tilde{\underline{A}} + \tilde{\underline{P}})\, \underline{E}_h = h^2 \underline{\epsilon}_h (x) \quad ,$$

where

$$\tilde{\underline{A}} = \underline{I}_2 + h^2 \int_0^1 G_h' (\underline{Z}_h + t\, (\underline{X}_h - \underline{Z}_h))dt,$$

$$\tilde{\underline{P}} = h^2 \int_0^1 F_h' \;(\underline{Z}_h + t(\underline{X}_h - \underline{Z}_h))\, dt.$$

Since $\tilde{\underline{A}}$ and $\tilde{\underline{P}}$ have the same properties as $\underline{A}$ and $\underline{P}$, respectively, the inverse $[\tilde{\underline{A}} + \tilde{\underline{P}}]^{-1}$ exists and satisfies the same inequalities as $(\underline{A} + \underline{P})^{-1}$.

Therefore,

$$\|\underline{E}_h\| = \|(\tilde{\underline{A}} + \tilde{\underline{P}})^{-1} h^2 \underline{\epsilon}_h(x)\|,$$

$$\leq K(\beta) \;\exp\,[M\; K(\beta)]\; \|\underline{\epsilon}_h(x)\| \;.$$

Also,

$$\begin{aligned}
\|\underline{E}_h\| &= \|(\tilde{\underline{A}} + \tilde{\underline{P}})^{-1} h^2 (\underline{\epsilon}_h^{(1)}(x) + \underline{\epsilon}_h^{(2)}(x))\|, \\
&\le mh^2 \,|\, (\tilde{\underline{A}} + \tilde{\underline{P}})^{-1} \,|\, \|\underline{\epsilon}_h^{(1)}(x)\| + h^2 \|(\tilde{\underline{A}} + \tilde{\underline{P}})^{-1}\| \|\underline{\epsilon}_h^{(2)}(x)\|, \\
&\le K(\beta)\, h^r \{mK_1 \exp [M_T K(\beta)] + K_2 \exp [M K(\beta)]\}.
\end{aligned}$$

Finally, the assertion that Newton's method can be used if g and F_h are twice continuously differentiable, is a trivial consequence of the fact that $\Phi_h'(\underline{Y})^{-1}$ exists and is bounded for all $\underline{Y}$, and Kantorovitch's theorem on the convergence of Newton's method (Henrici [15, p. 367]).

Theorem 6.2

Assume that $F_h'(\underline{Y})$ is independent of $\underline{Y}$, that

$$h^2 \|\underline{J}_2^{-1} \; F_h'(\underline{0})\| \le p_1 < 1,$$

that

$$h^2 \|\underline{J}_2^{-1} (G_h \underline{V} - G_h \underline{W})\| \le p_2 \|\underline{V} - \underline{W}\|,$$

for all $\underline{V}, \underline{W} \in R^n$, and that $p_3 = p_2/(1 - p_1) < 1$.

Then there exists a unique solution $\underline{Z}_h$ of (6.4) which can be found by successive approximation,

$$\underline{Z}_h^{(k)} = [\underline{I}_2 + h^2 F_h'(\underline{0})]^{-1} \; [-h^2 F_h \underline{0} - h^2 G_h \; Z_h^{(k-1)}], \qquad (6.10)$$

starting with any initial guess $\underline{Z}_h^{(1)}$.

The error $\underline{E}_h$ satisfies the inequalities

$$\|\underline{E}_h\| \le \|\underline{\epsilon}(x)\|/[4(1 - p_1)(1 - p_3)],$$

and

$$\|\underline{E}_h\| \le h^r [mK_1 + K_2]/[4(1-p_1)(1-p_3)] ,$$

where m, r, K_1, and K_2, are as in (6.8).

Proof: From the assumptions it follows $[\underline{I} + h^2 \underline{I}_2^{-1} F_h' (\underline{0})]^{-1}$ exists and has norm less than $1/(1 - p_1)$. Since $F_h'(Y)$ is independent of $\underline{Y}$, (6.4) may be rewritten in the equivalent form,

$$\underline{I}_2 \underline{Z}_h + h^2 F_h'(\underline{0}) \underline{Z}_h = -h^2 F_h \underline{0} - h^2 G_h \underline{Z}_h .$$

The first part of the theorem is now an immediate consequence of the contraction mapping theorem (Ortega and Rheinboldt [20, p. 383]).

The estimates for $\underline{E}_h$ follow from the observation that

$$\begin{aligned} \|\underline{E}_h\| & \\ &= \|\underline{X}_h - \underline{Z}_h\| , \\ &= \|[\underline{I}_2 + h^2 F_h'(\underline{0})]^{-1} h^2 \{G_h(\underline{Z}_h) - G_h(\underline{X}_h) + \underline{\epsilon}_h (x)\}\| , \\ &\le p_2 \|\underline{E}_h\|/(1 - p_1) + \|h^2 \underline{I}_2^{-1} \underline{\epsilon}_h(x)\|/(1 - p_1), \end{aligned}$$

and the fact that, from Lemma 4.2, $|\underline{I}_2^{-1}| \le 1/4h$.

Remarks

1. There are many possible variations of Theorems 6.1 and 6.2. Many of the results quoted by Ortega and Rheinboldt [20] could be applied to (6.4). It is possible to obtain sharper bounds for $\|L_\alpha^{-1}\|$, $\|L_\alpha^{-1}\|_s$, and $|L_\alpha^{-1}|$ (see Appendix A), and these can be used to sharpen the bounds in Theorem 6.1.

2. The following rather vague comments may be of some help in giving the reader a feel for Theorem 6.1. If $\mathfrak{F}$ is continuously differentiable and F_h is a "reasonable" approximation to $\mathfrak{F}$, then, noting (6.3), one sees that M will in general exist. The existence of M_T implies that the value of x(s) affects the values of x(t) for $t > s$ in a "moderate" fashion. For example, M_T will not exist if the equation (1.1) is a delay differential equation with delay $\Delta(t) = t - \frac{1}{2}$, since then $t - \Delta(t) = \frac{1}{2}$ so that the value of $x(\frac{1}{2})$ will greatly affect the values of x(t) for $t > \frac{1}{2}$.

7. A NUMERICAL EXAMPLE.

In this section we consider the numerical solution of (1.4), and compare our results with those of Nevers and Schmitt [18].

The existence of a unique solution y of (1.4) was established by Nevers and Schmitt [18].

Setting

$$x(t) = -y(2t) - \frac{1}{2}, \tag{7.1}$$

we find that x satisfies (1.1) with

$$g(t,x(t)) = -\frac{1}{4}\sin\left(x(t)+\frac{1}{2}\right) - (12t+2), \tag{7.2}$$

$$\left.\begin{aligned}(\mathcal{J}x)(t) &= -(8t+4)(1-2t), \quad t \le \frac{1}{2},\\ &= -(8t+4)x\left(t-\frac{1}{2}\right), \quad \frac{1}{2} \le t \le 1.\end{aligned}\right\} \tag{7.3}$$

Taking n to be odd we define F_h as follows:

$$\left.\begin{aligned}(F_h \underline{Y})_i &= -(8t_{h,i}+4)(1-2t_{h,i}), \quad i \le (n+1)/2,\\ &= -(8t_{h,i}+4)\, Y_{i-(n+1)/2}, \quad (n+1)/2 < i \le n,\end{aligned}\right\} \tag{7.4}$$

so that $\underline{\eta}_h(x) = 0$.

Since $x \in \mathscr{C}[0,1]$, $\mathcal{J}x \in \mathscr{C}[0,1]$ so that $x \in \mathscr{C}^{(2)}[0,1]$. Noting (7.3) it follows that x is four times continuously differentiable on $[0, \frac{1}{2}]$ and $[\frac{1}{2}, 1]$, but that $x^{(3)}(t)$ and $x^{(4)}(t)$ may have jump discontinuities at the point $t = \frac{1}{2}$. Let $\underline{\tau}_h(x) = \underline{\tau}_h^{(1)}(x) + \underline{\tau}_h^{(2)}(x)$, where

$$\left.\begin{aligned}\tau_{h,i}^{(1)}(x) &= \tau_{h,i}(x), \quad i = (n+1)/2,\\ &= 0, \quad \text{otherwise}.\end{aligned}\right\}$$

From (6.5) it follows that

$$\|\underline{\tau}^{(1)}(x)\| \le \frac{h}{3} \sup_{0\le t\le 1} |x^{(3)}(t)|,$$

$$\|\underline{\tau}^{(2)}(x)\| \le \frac{h^2}{12} \sup_{0\le t\le 1} |x^{(4)}(t)|,$$

so that (6.8) holds with $m = 1$, $r = 2$, $K_1 = \frac{1}{3} \sup_{0\le t\le 1} |x^{(3)}(t)|$,

and $K_2 = \frac{1}{12} \sup_{0\le t\le 1} |x^{(4)}(t)|$.

Since $F_h'(\underline{Y}) \le 0$, we cannot apply Theorem 6.1.

In order to apply Theorem 6.2 we need the following lemma.

<u>Lemma 7.1</u>

If F_h is as in (7.4) then, for all $\underline{Y} \in R^n$,

$$h^2 \|\underline{J}_2^{-1} F_h'(\underline{Y})\| \le 25/36.$$

<u>Proof:</u> Set $m = (n+3)/2$. From (7.4),

$$F_h'(\underline{Y}) = -4 \begin{pmatrix} \underline{O} & \underline{O} \\ \underline{D} & \underline{O} \end{pmatrix}$$

where $\underline{D}$ is the $(n+1-m) \times (n+1-m)$ diagonal matrix,

$$\underline{D} = \operatorname{diag}(d_j) = \operatorname{diag}(1 + 2jh), \quad m \le j \le n.$$

We denote by S_i the i-th row-sum of $[\underline{I}_2^{-1} F'_h(\underline{Y})]/(-4)$. Using Lemma 4.1,

$$S_i = \sum_{j=m}^{n} \gamma^*_{2;ij} \, d_j ,$$

$$= \sum_{\substack{j=m \\ j<i}}^{n} \frac{j(n+1-i)}{n+1} d_j + \sum_{\substack{j=m \\ j\geq i}}^{n} \frac{i(n+1-j)d_j}{n+1} .$$

Clearly, $S_i \leq S_m$ if $i \leq m$, so we may restrict ourselves to the case $i \geq m$. Then, if $ih = z$, and since $mh = \frac{1}{2} + h$, $nh = 1 - h$,

$$\sum_{j=1}^{n} j = n(n+1)/2,$$

$$\sum_{j=1}^{n} j^2 = n(n+1)(2n+1)/6,$$

$$S_i h^2 = (1 - z) \sum_{j=m}^{i-1} jh^2(1+2jh) + z \sum_{j=i}^{n} h(1-jh)(1+2jh),$$

$$= (1-z) S_i' + z S_i'' , \text{ say.}$$

Now,

$$S_i' = h^2 \sum_{j=m}^{i-1} j + 2h^3 [\sum_{j=1}^{i-1} j^2 - \sum_{j=1}^{m-1} j^2],$$

$$= h^2(i-m)(m+i-1)/2 + 2h^3 (i-1)(i)(2i-1)/6 - 2h^3(m-1)(m)(2m-1)/6,$$

$$= (z - \frac{1}{2} - h)(z + \frac{1}{2})/2 + (z-h)(z)(2z-h)/3 - (\frac{1}{2})(\frac{1}{2} + h)(1 + h)/3,$$

$$= \frac{z^2 - \frac{1}{4} - hz - \frac{h}{2}}{2} + \frac{2z^3 - 3hz^2 + h^2 z}{3} - \frac{2h^2 + 3h + 1}{12},$$

$$= \frac{1}{24} \{(16z^3 + 12z^2 - 5) + h(-24z^2 - 12z - 12) + h^2 (8z - 4)\}.$$

$$S_i'' = h \sum_{j=i}^{n} (1 + jh - 2h^2 j^2),$$

$$= h \sum_{j=i}^{n} (1 + jh) - 2h^3 [\sum_{j=1}^{n} j^2 - \sum_{j=1}^{i-1} j^2],$$

$$= h (n+1-i)(1+ih+1+nh)/2 - 2h^3 n (n+1)(2n+1)/6 + 2h^3 (i-1)(i)(2i-1)/6,$$

$$= (1-z)(3+z-h)/2 - (1-h)1(2-h)/3 + (z-h)z(2z-h)/3,$$

$$= (1-z)(3+z-h)/2 + (z-1)z(2z-h)/3 + \frac{1-h}{3} [z(2z-h) - (2-h)],$$

$$= (1-z)(3+z-h)/2 + (z-1)z (2z-h)/3 + \frac{1-h}{3} [(z-1)(2z-h) + (2z-h) - (2-h)],$$

$$= (1-z)\{\frac{3+z-h}{2} - \frac{2z^2-hz}{3} - \frac{(1-h)(2z+2-h)}{3}\},$$

$$= \frac{1-z}{6} \{9 + 3z - 3h - 4z^2 + 2hz - 4z - 4 + 2h + 4hz + 4h - 2h^2\},$$

$$= \frac{1-z}{6} \{(-4z^2 - z + 5) + h (6z + 3) - 2h^2\},$$

so that

$$\frac{z}{1-z} \; S_i'' = \frac{1}{24} \; \{(-16z^3 - 4z^2 + 20z) + h(24z^2 + 12z) - 8zh^2\}.$$

Hence,

$$24h^2 \; S_i \; / \; (1-z) = S_i' + z \; S_i'' / \; (1-z),$$

$$= A_0(z) + h \; A_1(z) + h^2 \; A_2(z),$$

where

$$A_0(z) = 8z^2 + 20z - 5,$$

$$A_1(z) = -12,$$

$$A_2(z) = -4.$$

Since $A_1(z), A_2(z) < 0$, it follows that

$$h^2 S_i \leq (1-z) \; A_0(z)/24,$$

$$= \emptyset(z), \text{ say.}$$

Now,

$$\emptyset(z) = (-8z^3 - 12z^2 + 25z - 5)/24,$$

$$\emptyset'(z) = -z^2 - z + \frac{25}{24} .$$

Since $\emptyset'(.5) > 0$, $\emptyset'(1) < 0$, and $\emptyset'(.64) < 0$, $\emptyset$ attains its maximum on $[\frac{1}{2}, 1]$ at a point $z_0 < .64$ satisfying $\emptyset'(z_0) = -z_0^2 - z_0 + \frac{25}{24} = 0$. Hence

$$\emptyset(z_0) = (1-z_0)(8z_0^{\ 2} + 20z_0 - 5)/24,$$

$$= (1-z_0)(-8z_0 + \frac{25}{3} + 20z_0 - 5)/24,$$

$$= (1-z_0)(18z_0 + 5)/36,$$

$$= (-18z_0^{\ 2} + 13z_0 + 5)/36,$$

$$= (18z_0 - \frac{75}{4} + 13z_0 + 5)/36,$$

$$= (31z_0 - \frac{55}{4})/36,$$

$$< ((31)(.64) - \frac{55}{4})/36$$

$$< \frac{25}{(4)(36)} \quad .$$

Combining the above results, the lemma follows.

From Theorem 4.3 it follows that $\|\underline{J}_2^{\ -1}\| \leq 1/(4h^2)$. However, it is well known (Henrici [15, p. 371]) that $\|\underline{J}_2^{\ -1}\| \leq 1/(8h^2)$. Using this stronger inequality we see that for all $\underline{V}$, $\underline{W} \in R^n$,

$$\|h^2 \underline{J}_2^{\ -1} (G_h \underline{V} - G_h \underline{W})\|$$

$$\leq \max \left| \frac{\partial g(t, y)}{\partial y} \right| . \ \|\underline{V} - \underline{W}\| \,/\, 8,$$

$$\leq \|\underline{V} - \underline{W}\|/32.$$

From this and Lemma 7.1 we see that Theorem 6.2 applies with $p_1 = 25/36$, $p_2 = 1/32$, and $p_3 = 9/88$. In particular, the iteration (6.10) can be used and $\|\underline{E}_h\| = O(h^2)$.

A program, NEVERS, was written to implement the algorithm (6.10) for problems of the form

$$y''(t) = g(t,y(t)) + c(t)\, y(t - \Delta(t)),\ 0 < t < 1,$$

$$y(t) = u(t),\quad t \le 0,$$

$$y(t) = v(t),\quad t \ge 1,$$

where $\Delta(t)$ may be positive or negative but must be a multiple of h. A listing of the program is given in Appendix B.

NEVERS was used to compute the solution $x(t)$ of (1.1) with g and $\mathfrak{F}$ defined by (7.2) and (7.3). The computations were performed using double-precision arithmetic (16 decimals) on the UNIVAC 1108 at the University of Wisconsin. The initial approximation, $\underline{Z}_h^{(0)}$, was taken to be zero. The iteration (6.10) was used until $\|\underline{Z}_h^{(k+1)} - \underline{Z}_h^{(k)}\| \le 1.10^{-10}$; this always occurred when k was less than or equal to 6. The approximation was computed for $n = 4, 8, 16, 32, 64$, and 128. Total computation time, including compilation, was 55 seconds.

Let $\underline{Y}_h = -\underline{Z}_h - .5$. Noting (7.1), we see that $\underline{Y}_h$ is an approximation to the solution $y(s)$ of (1.4). The components of $\underline{Y}_h$ corresponding to $s = .5$, 1.0, and 1.5, are given in Table 7.1. For comparison, Table 7.1 also contains the values computed by Nevers and Schmitt [18].

The following observations concerning the numerical results may be made. Firstly, the results of Table 7.1 confirm the assertion of Theorem 6.2 that $\|\underline{E}_h\| = O(h^2)$. Secondly, the rapid convergence of (6.10) is due to the fact that

$p_3 \doteq .1$. Thirdly, as can be seen from Table 7.1, the approximations $\underline{Y}_h$ decrease monotonely as $h \to 0$; this is connected with the fact that $G_h'(\underline{Y}) \leq 0$ and that $[\underline{I} + h^2 \underline{J}_2^{-1} F_h'(0)]^{-1} \geq 0$.

n+1	s = .5		s = 1.0		s = 1.5	
	Y	ΔY	Y	ΔY	Y	ΔY
4	-1.471368		-1.927188		-1.868366	
8	-1.524873	-53505	-2.042571	-115383	-1.938585	-70219
16	-1.538884	-14011	-2.072713	- 30142	-1.957158	-18573
32	-1.542430	- 3546	-2.080336	- 7623	-1.961869	- 4711
64	-1.543319	- 889	-2.082247	- 1911	-1.963052	- 1183
128	-1.543542	- 223	-2.082725	- 478	-1.963348	- 296
Nevers & Schmitt	-1.543053		-2.081821		-1.962343	

Table 7.1

Numerical solution of (1.4).

APPENDIX A

Further properties of J_α.

During the present investigation we obtained bounds for $\|J_\alpha^{-1}\|$, $\|L_\alpha^{-1}\|$, and $\|U_\alpha^{-1}\|$ which are sharper than those available in the literature. As it turned out, these bounds were not needed. We include them in the present appendix since they may be useful to other workers. We use the notation of section 4.

Fischer and Usmani [7] prove

Theorem A.1

$$\|J_\alpha^{-1}\| \le \frac{\sinh\varphi - 2\sinh[(\varphi-\theta)/2]}{(\alpha-2)\sinh\varphi}, \quad \alpha > 2,$$

$$\le (n+1)^2/8, \quad \alpha = 2,$$

$$\le \frac{n}{|\sin\theta \ \sin\varphi|}, \quad \alpha < 2.$$

The following theorem gives a different bound for $\|J_\alpha^{-1}\|$ when $\alpha < 2$.

Lemma A.2

Let $0 < \alpha < 2$ and $\varphi < \pi$. Then

$$\|J_\alpha^{-1}\| < \frac{(n+1)^2}{8\cos(\varphi/2)} .$$

Proof: Since $\varphi < \pi$, it follows from Lemma 4.1 that $J_\alpha^{-1} > 0$. Set

$$S_i = \sum_{j=1}^{n} |\gamma^*_{\alpha;ij}| = \sum_{j=1}^{n} \gamma^*_{\alpha;ij} .$$

Then (Fischer and Usmani [7, p. 132]),

$$S_i = \frac{D_{\alpha,n} - (D_{\alpha,n-i} + D_{\alpha,i-1})}{(\alpha - 2)\, D_{\alpha,n}} ,$$

$$= \frac{\sin\varphi - \sin(\varphi - i\theta) - \sin(i\theta)}{(\alpha - 2)\sin\varphi} ,$$

$$= \frac{2\sin(\varphi/2)\left[\cos(\varphi/2) - \cos[(\varphi/2) - i\theta]\right]}{2(\alpha - 2)\sin(\varphi/2)\cos(\varphi/2)} ,$$

$$= \frac{\sin[(\varphi - i\theta)/2]\ \sin[(i\theta)/2]}{2\sin^2(\theta/2)\cos(\varphi/2)} .$$

Using Lemma 2.3,

$$S_i \le \frac{(n+1-i)i}{2\cos(\varphi/2)} ,$$

$$\le \frac{(n+1)^2}{8\cos(\varphi/2)} ,$$

as asserted.

Lemma A.3

$$\|\underline{L}_\alpha^{-1}\| = \frac{1}{2} + \frac{1}{2}\ \frac{\tanh(n\theta/2)}{\tanh(\theta/2)} , \qquad \alpha > 2;$$

$$= (n+1)/2, \qquad \alpha = 2;$$

$$= \frac{1}{2} + \frac{1}{2}\ \frac{\tan(n\theta/2)}{\tan(\theta/2)} , \quad 0 < \alpha < 2 \text{ and } \varphi < \pi;$$

and

$$\|\underline{U}_\alpha^{-1}\| \le \frac{\tanh\varphi}{\theta}, \quad \alpha > 2;$$

$$\le (n+1)/e, \quad \alpha = 2;$$

$$\le \frac{1}{\sin\varphi} + \frac{n+1}{\varphi}\{1 + [\log\tan(\varphi/2)]_+\}, \quad 0 < \alpha < 2 \text{ and } \varphi < \pi;$$

where

$$[u]_+ = u, \text{ if } u \ge 0,$$

$$= 0, \text{ if } u < 0.$$

Proof: If $\underline{A} = (a_{ij})$ and

$$S_i = \sum_{j=1}^{n} |a_{ij}|$$

then

$$\|\underline{A}\| = \max_{1\le i\le n} |S_i| .$$

Using Lemmas 3.1 and 4.1 we investigate the different possible cases.

Case 1: $\underline{A} = \underline{L}_2^{-1}$. Then

$$S_i = \sum_{j=1}^{i} D_{2,j-1}/D_{2,i-1},$$

$$= \sum_{j=1}^{i} j/i,$$

$$= (i+1)/2,$$

so that

$$\|\underline{L}_2^{-1}\| = (n+1)/2.$$

<u>Case 2: $\underline{A} = \underline{L}_\alpha^{-1}$, $\alpha > 2$.</u> Then

$$S_i = \sum_{j=1}^{i} D_{\alpha,j-1}/D_{\alpha,i-1},$$

$$= \frac{1}{\sinh i\theta} \sum_{j=1}^{i} \sinh j\theta,$$

$$= \frac{1}{\sinh i\theta} \sum_{j=1}^{i} \frac{\cosh (j+\frac{1}{2})\theta - \cosh (j-\frac{1}{2})\theta}{2 \sinh(\theta/2)},$$

$$= \frac{\cosh (i+\frac{1}{2})\theta - \cosh \theta/2}{2 \sinh i\theta \sinh \theta/2},$$

$$= \frac{\cosh i\theta \cosh \theta/2 + \sinh i\theta \sinh \theta/2 - \cosh \theta/2}{2 \sinh i\theta \sinh \theta/2},$$

$$= \frac{1}{2} + \frac{(\cosh \theta/2)(\cosh i\theta - 1)}{2 \sinh i\theta \sinh \theta/2},$$

$$= \frac{1}{2} + \frac{(\cosh \theta/2)(2 \sinh^2 i\theta/2)}{2(\sinh \theta/2)(2 \sinh i\theta/2 \cosh i\theta/2)},$$

$$= \frac{1}{2} + \frac{1}{2} \frac{\tanh i\theta/2}{\tanh \theta/2},$$

so that

$$\|\underline{L}_\alpha^{-1}\| = \frac{1}{2} + \frac{1}{2} \frac{\tanh(n\theta/2)}{\tanh(\theta/2)}.$$

Case 3: $\underline{A} = \underline{L}_\alpha^{-1}$, $\alpha < 2$ and $\varphi < \pi$. Then

$$S_i = \sum_{j=1}^{i} D_{\alpha,j-1}/D_{\alpha,i-1},$$

$$= \frac{1}{\sin i\theta} \sum_{j=1}^{i} \sin j\theta,$$

$$= \frac{1}{\sin i\theta} \sum_{j=1}^{i} \frac{\cos (j-\frac{1}{2})\theta - \cos(j+\frac{1}{2})\theta}{2 \sin(\theta/2)} ,$$

$$= \frac{\cos(\theta/2) - \cos(i+\frac{1}{2})\theta}{2 \sin i\theta \sin \theta/2} ,$$

$$= \frac{\cos(\theta/2) - \cos i\theta \cos \theta/2 + \sin i\theta \sin \theta/2}{2 \sin i\theta \sin \theta/2} ,$$

$$= \frac{1}{2} + \frac{\cos \theta/2 (1 - \cos i\theta)}{2 \sin \theta/2 \sin i\theta} ,$$

$$= \frac{1}{2} + \frac{(\cos \theta/2)\, 2 \sin^2 (i\theta/2)}{2 \sin(\theta/2)\, (2 \sin i\theta/2 \cos i\theta/2)} ,$$

$$= \frac{1}{2} + \frac{1}{2} \frac{\tan(i\theta/2)}{\tan(\theta/2)} ,$$

so that

$$\|\underline{L}_\alpha^{-1}\| = \frac{1}{2} + \frac{1}{2} \frac{\tan(n\theta/2)}{\tan(\theta/2)} .$$

Case 4: $\underline{A} = \underline{U}_2^{-1}$. Then

$$S_i = \sum_{j=i}^{n} D_{2,i-1}/D_{2,j},$$

$$= \sum_{j=i}^{n} i/(j+1),$$

$$\le i \int_i^{n+1} \frac{1}{x}\, dx,$$

$$= i \log\left(\frac{n+1}{i}\right),$$

$$= f(i), \text{ say.}$$

Since

$$f'(x) = \log\left(\frac{n+1}{x}\right) - 1,$$

$f(x)$ attains its maximum in the interval $(0, n+1]$ at the point $x = (n+1)/e$.

Hence

$$\|\underline{U}_2^{-1}\| \le f\left(\frac{n+1}{e}\right) = (n+1)/e.$$

Case 5: $\underline{A} = \underline{U}_\alpha^{-1}$, $\alpha > 2$. Then

$$S_i = \sum_{j=i}^{n} D_{\alpha,i-1}/D_{\alpha,j},$$

$$= \sum_{j=i}^{n} \sinh i\theta/\sinh(j+1)\theta,$$

$$\leq \sinh(i\theta) \int_i^{n+1} \operatorname{csch}(t\theta)\, dt,$$

$$= \frac{\sinh(i\theta)}{\theta} \int_{i\theta}^{\varphi} \operatorname{csch} z \; dz,$$

$$= \frac{\sinh i\theta}{\theta} \log \left[\frac{\tanh(\varphi/2)}{\tanh(i\theta/2)}\right],$$

$$= \frac{2}{\theta} f(i\theta/2),$$

where

$$f(z) = \frac{1}{2} \sinh 2z \log \left[\frac{\tanh(\varphi/2)}{\tanh z}\right].$$

Now: (i) $f(\varphi/2) = 0$; (ii) $f(0+) = 0$;

(iii) $f(z) > 0$ if $0 < z < \varphi/2$. Hence, f attains its maximum in the interval $[0,\varphi/2]$ at an interior point $z = \zeta$, where $f'(\zeta) = 0$. Furthermore,

$$f'(z) = (\cosh 2z) \log \left(\frac{\tanh(\varphi/2)}{\tanh z}\right) - \frac{1}{2} \sinh 2z \frac{\operatorname{sech}^2 z}{\tanh z},$$

$$= (\cosh 2z) \log \left(\frac{\tanh(\varphi/2)}{\tanh z}\right) - 1.$$

Hence, if $f'(\zeta) = 0$,

$$\log\left(\frac{\tanh(\varphi/2)}{\tanh \zeta}\right) = \frac{1}{\cosh 2\zeta},$$

and

$$f(\zeta) = \frac{1}{2} \tanh 2\zeta .$$

Therefore,

$$\max_{0 \le z < (\varphi/2)} f(z) \le \frac{1}{2} \tanh \varphi .$$

Consequently,

$$\|\underline{U}_\alpha^{-1}\| \le [\tanh \varphi]/\theta .$$

The above bound for max f(z) is rather crude, but we have been unable to obtain a simple sharp bound.

Case 6: $\underline{A} = \underline{U}_\alpha^{-1}$, $0 < \alpha < 2$, $\varphi < \pi$. Then

$$S_i = \sum_{j=i}^{n} D_{\alpha,i-1}/D_{\alpha,j} ,$$

$$= \sum_{j=i}^{n} \sin i\theta / \sin (j+1)\theta ,$$

$$= \sin i\theta \left\{ \frac{1}{\sin(n+1)\theta} + \sum_{\substack{k=i+1 \\ k\theta > \frac{\pi}{2}}}^{n} \frac{1}{\sin k\theta} + \sum_{\substack{k=i+1 \\ k\theta \le \frac{\pi}{2}}}^{n} \frac{1}{\sin k\theta} \right\} ,$$

$$\le \sin i\theta \left\{ \frac{1}{\sin\varphi} + \left[\frac{1}{\theta} \int_{\frac{\pi}{2}}^{\varphi} \frac{dt}{\sin t} \right]_+ + \left[\frac{1}{\theta} \int_{i\theta}^{\frac{\pi}{2}} \frac{dt}{\sin t} \right]_+ \right\} ,$$

where

$$[u]_+ = \begin{cases} u, & \text{if } u \ge 0, \\ 0, & \text{if } u < 0, \end{cases}$$

and where we have used the fact that $\sin t$ is monotone increasing for $0 \le t \le \frac{\pi}{2}$ and monotone decreasing for $\frac{\pi}{2} \le t \le \pi$.

Since

$$\int \frac{dt}{\sin t} = \log \tan (t/2) ,$$

$$S_i \le \frac{1}{\sin\varphi} + (n+1) \left[\frac{\log \tan (\varphi/2)}{\varphi} \right]_+ + \frac{(n+1)}{\varphi} [f(i\theta)]_+ ,$$

where

$$f(z) = -\sin z\,[\log\, \tan\,(z/2)]\,,$$

$$= \frac{-\,2\,\tan\,(z/2)\,\log\tan\,(z/2)}{1+(\tan\,z/2)^2}\,,$$

$$= \frac{2\,\cot\,(z/2)\,\log\cot\,(z/2)}{1+[\cot\,(z/2)]^2}\,,$$

$$= 2\,g\,[\cot\,(z/2)]\,,$$

with

$$g(u) = \frac{u\,\log u}{1+u^2}\,.$$

Consider $g(u)$ on the interval $[1,\infty)$. We have that

$$g'(u) = \frac{u^2-1}{(u^2+1)^2}\,\left[\,1+\frac{2}{u^2-1} - \log u\right],$$

from which it is easily seen that g' has only one zero, $u = \sigma$ say, and that $e < \sigma < e^{3/2}$. Since $g(1) = g(\infty) = 0$, g attains its maximum at $u = \sigma$.

Hence

$$\max_{1\le i\le n}\,[f(i\theta)]_+ \le \max_{0\le z\le\frac{\pi}{2}} f(z)\,,$$

$$= 2\max_{1\le u<\infty} g(u)\,,$$

$$= 2\max_{e\le u\le e^{3/2}} g(u)\,,$$

$$\le 2(\log e^{3/2})\max_{e\le u\le\infty}\frac{u}{1+u^2}\,,$$

$$= \frac{3e}{1+e^2}\,,$$

$$< 1\,.$$

Combining the above results,

$$\|\underline{U}_\alpha^{-1}\| \le \frac{1}{\sin\varphi} + \frac{n+1}{\varphi}\{1 + [\log\tan(\varphi/2)]_+\} .$$

APPENDIX B

The program NEVERS.

A listing of NEVERS is given at the end of this appendix. Here we make a few comments to make the program more easily comprehensible to the reader.

NEVERS computes approximate solutions to problems of the form

$$\left.\begin{aligned} &\ddot{y}(t) = g(t,y(t)) + c(t)\ y(t-\Delta(t)),\ 0 < t < 1, \\ &y(t) = u(t),\ t \le 0, \\ &y(t) = v(t),\ t \ge 1, \end{aligned}\right\} \tag{A.1}$$

where $\Delta(t)$ may be positive or negative but must be a multiple of the stepsize h.

Setting

$$x(t) = y(t),\ 0 \le t \le 1, \tag{A.2}$$

(A.1) takes the form (1.1) with

$$(\mathfrak{F}x)(t) = \begin{cases} c(t)\ u(t-\Delta(t)), & \text{if } t-\Delta(t) \le 0, \\ c(t)\ x(t) & ,\ \text{if } 0 < t-\Delta(t) < 1, \\ c(t)\ v(t-\Delta(t)), & \text{if } 1 \le t-\Delta(t). \end{cases} \tag{A.3}$$

The approximate solution $\underline{Z}_h$ is computed using the iteration (6.10). The correspondence between the variables of (6.10) and the arrays used in NEVERS is as follows:

<u>Equation (6.10)</u>	<u>NEVERS</u>
$\underline{I}_2 + h^2\ F_h'(0)$	A
$-h^2\ F_h\ \underline{0}$	BC
$\underline{Z}_h^{(k+1)}$	X
$\underline{Z}_h^{(k)}$	XP

The initial guess $\underline{Z}_h^{(1)}$ is taken to be $\varphi_h x_0$ where $x_0 \in \mathscr{C}[0,1]$ is provided by the user.

The functions u(t), v(t), g(t,x), c(t), Δ(t), and x_0(t) must be provided by the user as procedures at the beginning of NEVERS. The iteration (6.10) is continued until $k = KMAX$ or $\|\underline{Z}_h^{(k)} - \underline{Z}_h^{(k-1)}\| \le \epsilon$, and the approximations $\underline{Z}_h$ are computed for $2 \le n \le 2$ **IMAX; in the listed program KMAX = 20, $\epsilon = 10^{-10}$, and IMAX = 7.

NEVERS makes use of the BUMP2 matrix package on the Univac 1108 at the University of Wisconsin. In addition to using the housekeeping subroutines MTADFM, MTMDEF, and MTMDFM, NEVERS uses the following BUMP2 subroutines:

MTCNST - Set up a constant matrix

MTMPRT - Print a matrix

MTINVD - Invert a double-precision matrix

MTMPY - Multiply two matrices

MTSUB - Subtract two matrices

MTNRMR - Compute the maximum row-sum norm of a matrix

MTMOVE - Move a matrix

The listed program uses double-precision arithmetic but is written so that it can be converted to single-precision by changing only three of the program's statements.

```
C       PROGRAM NEVERS
C       ***************************************************
C       PROGRAM TO SOLVE BOUNDARY VALUE PROBLEMS FOR EQUATIONS WITH
C       LINEAR PERTURBED ARGUMENT ON INTERVAL (0,1)
C       Y(T)=U(T) FOR T.LE.0    Y(T)=V(T) FOR T.GE.1
C       DDY(T)/DTT=G(T,Y(T)) + C(T)*Y(T-DELTA(T)) FOR 0.LT.T AND T.LT.1
C       SOLUTION IS COMPUTED BY ITERATION  INITIAL GUESS IS X0(T)
C       PROGRAM USES BUMP2 MATRIX PACKAGE ON UWCC 1108
C       SAMPLE PROBLEM IS PROBLEM OF NEVERS AND SCHMITT
C       ***************************************************
        IMPLICIT DOUBLE PRECISION ( A-H,O-Z)
        DIMENSION A(129,129),BC(129),X(129),B(129),XP(129)
        DIMENSION FORM(2),XDIF(129)
        INTEGER FORM,TYP
C
        U(T)=-2.*T
        V(T)=0.*T
        G(T,X)=-SIN(X+.5)/4. -(12.*T+2.)
        C(T)=-(8.*T+4.)
        DELTA(T)=.5+0.*T
        X0(T)=0.*T
C
        TYP = 1HD
        FORM(1)=6H(4D20.
        FORM(2)=6H8)
C       ***************************************************
C       TO RUN IN SINGLE PRECISION CHANGE TYP,FORM(1),IMPLICIT DOUBLE
C       PRECISION, AND MTINVD(STATEMENT NUMBER 600)
C       ***************************************************
        NA=129
        IMAX=7
C       ***************************************************
C       PROBLEM RUN WITH STEPSIZE H =1/4 .... 1/2**IMAX
C       NA = SIZE OF ARRAYS = 2**IMAX+1
C       ***************************************************
        KMAX=20
        EPS=1.E-10
C       ***************************************************
C       ITERATIONS CONTINUED UNTIL DIFFERENCE BETWEEN SUCCESSIVE
C       APPROXIMATIONS LESS THAN EPS
C       OR NUMBER OF ITERATIONS EQUAL TO KMAX
C       ***************************************************
        PRINT 50,NA,IMAX,KMAX
50      FORMAT('1 NA,IMAX,KMAX,EPS =  ',3I5)
        WRITE(6,FORM)EPS
C
        CALL MTADEM(1,A,NA,NA,TYP)
        CALL MTADEM(5,B,BC,XP,XDIF,X,NA,1,TYP)
C       ***************************************************
C       START OF OUTER LOOP
C       ***************************************************
        DO 9000 I=2,IMAX
        N=(2**I)-1
        ONE=1.
        H=ONE/(N+1)
        H2=H**2
```

```
      PRINT 100,I,N
100   FORMAT( '1 I,N,H= ' ,2I5)
      WRITE(6,FORM)H
      CALL MTMDEF(A,N,N,'GEN')
      CALL MTMDEM(3,BC,B,XP,N,1,'GEN')
C     ******************************************************
C     SET UP MATRIX A AND CONSTANT RIGHT HAND SIDE BC
C     ******************************************************
      ZERO=0.
      CALL MTCNST(A,ZERO,TYP)
      CALL MTCNST(BC,ZERO,TYP)
      DO 500 J=1,N
      T=J*H
      A(J,J)=2.
C
      JP1=J+1
      IF( (JP1).LE.N) A(J,JP1)=-1.
      IF( (JP1).GT.N) BC(J)=BC(J)+V(T+H)
C
      JM1=J-1
      IF( (JM1).GE.1) A(J,JM1)=-1.
      IF( (JM1).LT.1) BC(J)=BC(J)+U(T-H)
C
      JD= J- INT( (DELTA(T)+.5*H)/H)
      TD=T-DELTA(T)
      CD=C(T)*H2
      IF( (JD.GE.1) .AND. (JD.LE.N) ) A(J,JD)=A(J,JD)+CD
      IF(JD.LT.1)BC(J)=BC(J)-CD*U(TD)
      IF(JD.GT.N)BC(J)=BC(J)-CD*V(TD)
500   CONTINUE
C     ******************************************************
C     IF N.LE.4 PRINT MATRIX A FOR DEBUGGING PURPOSES
C     ******************************************************
      IF( N.LE.4 ) CALL MTMPRT(A,FORM,0,' MATRIX A .. ')
C
C     ******************************************************
C     INVERT A AND SET UP INITIAL GUESS XP
C     ******************************************************
600   CALL MTINVD( A,A ,$700)
      GO TO 800
700   STOP 700
800   CONTINUE
      DO 830 J=1,N
      T=J*H
830   XP(J)=X0(T)
C     ******************************************************
C     ITERATE
C     ******************************************************
      K=0
850   K=K+1
      DO 900 J=1,N
      T=J*H
900   B(J)=BC(J)-G(T,XP(J))*H2
      CALL MTMPY(A ,B,X)
C     ******************************************************
C     IF N.LE.4 AND K=1 PRINT VARIOUS MATRICES FOR DEBUGGING
```

```
C     *************************************************************
      IF( (N.GT.4) .OR. (K.GT.1) ) GO TO 925
      CALL MTMPRT(A ,FORM,0,' INVERSE OF MATRIX A ..')
      CALL MTMPRT(BC,FORM,0,' VECTOR BC..')
      CALL MTMPRT(XP,FORM,0,' INITIAL VECTOR XP..')
      CALL MTMPRT(B,FORM,0, ' VECTOR B..')
      CALL MTMPRT(X,FORM,0,' VECTOR X= ..')
925   CONTINUE
      CALL MTSUB(X,XP,XDIF)
      CALL MTNRMR(XDIF,RNXDIF)
      PRINT 950,K
950   FORMAT( '0 ITERATION NUMBER, ROW NORM OF XDIF =   ' , I5)
      WRITE(6,FORM)RNXDIF
      IF( RNXDIF.LE.EPS) GO TO 8900
      IF ( K.GE.KMAX ) GO TO 8800
      CALL MTMOVE(X,XP)
      GO TO 850
8800  PRINT 8801
8801  FORMAT( '0 ITERATION MAXIMUM REACHED ')
      GO TO 8950
8900  PRINT 8901
8901  FORMAT( '0ITERATIONS CONVERGED ')
      GO TO 8950
8950  CALL MTMPRT(X,FORM,0,' VECTOR X= ..')
9000  CONTINUE
      STOP
      END
```

REFERENCES

[1] Ciarlet, P. G., M. H. Schultz, and R. S. Varga: Numerical methods of high order accuracy for nonlinear boundary value problems. V. Monotone operator theory. Numer. Math. 13, 51-77 (1969).

[2] Cooke, K. L.: Some recent work on functional differential equations. Proceedings United States - Japan Seminar on Differential and Functional Equations. New York: Benjamin, 1968.

[3] Cryer, C. W., and L. Tavernini: The numerical solution of Volterra functional differential equations by Euler's method. SIAM J. Numer. Anal., to appear.

[4] El'sgol'ts, L. E.: Introduction to the Theory of Differential Equations with Deviating Arguments. San Francisco: Holden-Day, 1966.

[5] Fennell, R., and P. Waltman: Boundary value problems for functional differential equations. Bull. Amer. Math. Soc. 75 , 487-489 (1969).

[6] ______________: A boundary value problem for a system of nonlinear functional differential equations. J. Math. Anal. Appl. 26, 447-453 (1969).

[7] Fischer, C. F. and R. A. Usmani: Properties of some tridiagonal matrices and their application to boundary value problems. SIAM J. Numer. Analy. 6, 127-142 (1969).

[8] Gantmacher, F. R.: The Theory of Matrices, vol. I. New York: Chelsea, 1959.

[9] Gantmacher, F. R., and M. G Krein: Oszillationsmatrizen, Oszillationskerne und Kleine Schwingungen Mechanischer Systeme. Berlin: Akademie, 1960.

[10] Grimm, L. J., and K. Schmitt: Boundary value problems for delay-differential equations. Bull. Amer. Math. Soc. 74, 997-1000 (1968).

[11] ____________: Boundary value problems for differential equations with deviating arguments. Aequationes Math. 4, 176-190 (1970).

[12] Halanay, A.: On a boundary value problem for linear systems with time lag. J. Differential Equ. 2, 47-56 (1966).

[13] Halanay, A., and J. A. Yorke: Some new results and problems in the theory of differential-delay equations. SIAM Review 13, 55-80 (1971).

[14] Hale, J. K.: Functional Differential Equations. New York: Springer, 1971.

[15] Henrici, P.: Discrete Variable Methods in Ordinary Differential Equations. New York: Wiley, 1962.

[16] Kato, S.: Asymptotic behavior in functional differential equations. Tohoku Math. J. 18, 174-215 (1966).

[17] Keller, H. B.: Numerical Methods for Two-Point Boundary-Value Problems. Waltham, Mass.: Blaisdell, 1968.

[18] Nevers, K. de, and K. Schmitt: An application of the shooting method to boundary value problems for second order delay equations. J. Math. Anal. Appl., to appear.

[19] Norkin, S. B.: Differential Equations of Second Order with Deviating Arguments. Providence: Amer. Math. Soc., to appear.

[20] Ortega, J. M., and W. C. Rheinboldt: Iterative Solution of Nonlinear Equations in Several Variables. New York: Academic Press, 1970.

[21] Schmitt, K.: Comparison theorems for second order delay-differential equations. Rocky Mountain Math. J., to appear.

[22] Tavernini, L.: One-step methods for the numerical solution of Volterra functional differential equations. SIAM J. Numer. Anal., to appear.

[23] ____________: Linear multistep methods for the numerical solution of Volterra functional differential equations. Applicable Anal., to appear.