FINITE ELEMENT SOLUTIONS TO

BOUNDARY VALUE PROBLEMS

BY

P. MOORE

Thesis submitted for the degree of Doctor of Philosophy

Brunel University

1976

(Department of Mathematics.)

Abstract

This thesis consists of two distinct parts which deal with two-point boundary value problems and parabolic problems, respectively. In Section 1 we examine the numerical solution of a two-point boundary value problem by a collocation method based on the consistency relationship of regular splines. An existence and convergence result is established which generalises the $O(h^2)$ convergence result of the cubic spline collocation scheme for the problem in question. Contrary to most previously documented finite element schemes this method employs splines that may be non-linear in structure. Consequently, by a judicious choice of regular spline, the dominant terms of the true solution may be imitated more accurately than by the conventional polynomial based splines. The scheme is implemented by an algorithm that examines the suitability of various classes of regular splines and determines the subsequent deployment of them.

The second section investigates semi-discrete finite element schemes for approximating the linear parabolic equation. A standard finite element discretization is employed for the space variable whilst an A_0-stable, linear multistep, multiderivative discretization scheme, (L.M.S.D.) is used in time. We consider both the homogeneous and the nonhomogeneous linear parabolic equations and derive optimal convergence results for the above schemes. The convergence results achieved with a k-step L.M.S.D. scheme, incorporating the first m derivatives, generalise and extend the studies of several authors who concentrate on the particular cases of linear multistep formulae, m=1, and one-step schemes, k=1. A_o-stable L.M.S.D.'s are constructed and their implementation procedures examined. The suitability of selecting a L.M.S.D. method, with m, k > 1, in a semi-discrete Galerkin scheme is discussed, and its superiority over semi-discrete Galerkin schemes, that incorporate linear multistep or one-step formulae, is confirmed in several aspects.

Finally, a class of quasi-linear parabolic equations is solved by a semi-discrete Galerkin scheme that is third order accurate in time. This method is based on a particular third order L.M.S.D. scheme and requires the solution of linearly algebraic systems of equations at each time level. Thus, we improve on all the previously documented linearised schemes as they are only second order accurate in time. All the schemes described in Section 2 are unconditionally stable.

Acknowledgements

I wish to express my sincerest gratitude to Professor J.R. Whiteman for allowing me the opportunity to undertake the enclosed study, and for his encouragement and guidance during its composition. Furthermore, I wish to thank my friends and colleagues, especially Dr. N. Papamichael, for their assistance and good-humour throughout my period of study at Brunel University.

Contents

SECTION 1. REGULAR SPLINE SOLUTIONS TO A TWO-POINT BOUNDARY VALUE PROBLEM

INTRODUCTION

The past decade accounted for a vast literature of techniques and algorithms to solve numerically a variety of two-point boundary value problems. A rapid glance through any prominent journal of numerical analysis supports the opinion that the computer user is confronted with a wide, and even bewildering, choice of possibilities. Excluding the literature concerning 'shooting methods', and 'non-local approximants' (e.g. Chebyshev series) the user enters the extensive field of 'local approximations' encompassing all the documented finite difference and finite element schemes. For a thorough insight into methods for solving boundary value problems we recommend [4]. This book contains references which are far too extensive to include here.

The subclass of finite element schemes has recently received the concerted attention of numerical analysts. In particular the user will be aware of the existence of projection methods (including collocation methods), and schemes derived from a variational formulation of the problem. This variational formulation uses the property that the analytic solution to the boundary value problem strictly minimises a certain functional. Details may be found in [5], whilst computational aspects and rates of convergence are also considered in [13] and [20] , amongst others. The nomenclature 'projection' defines the underlying principle of projection methods. We project the problem into a finite dimensional subspace of an appropriate Hilbert space by some technique, and derive the approximant to the remodelled problem. In particular we may view the Galerkin procedure as a specific example of a projection method. The Galerkin method is employed by Douglas and Dupont [7] and Wheeler [24] to investigate a class of linear two-point boundary value problems. A superconvergence result at the knots is established in [7].

Projection methods of a collocation type for classes of nonlinear boundary value problems are studied in [9], [12], [14] and [16]. Collocation methods require the spline approximant to satisfy the differential equation at certain internal points.

The finite difference approach to nonlinear boundary value problems is illustrated in [10], [11] and [21]. Kreiss [11] develops a general but complete theory for the linear equation. In [10], Keller employs the centred Euler scheme to study a general nonlinear boundary value problem. He also notes that any scheme satisfying the theory of Kreiss may be extended to the nonlinear equation. Finally, we introduce the paper by Stepleman [21]. In particular for the two-point boundary value problem independent of the first derivative, and with disjoint boundary conditions, Stepleman notes that his method is the classical Numerov method.

Of fundamental significance is the structure of the approximant, or for finite differences, the structure of the difference formula. Independent of the approach employed, the numerical solution is dependent on a polynomial structure. The spline function spaces used in projection and variational formulations are piece-wise polynomials satisfying certain continuity constraints. Analogously, the finite difference schemes mentioned above are polynomial based. For example, the fourth order Numerov formula is derived by spanning the interval $[0, 2h]$ by a quartic polynomial, and collocating the values of the function $y(x)$ and its second derivative $y''(x)$ at the knots $x = 0, h, 2h$. Recently, some interest has surrounded the study of splines that are closer in structure to the function being approximated than the more conventional polynomial based splines. The classes of regular splines defined in Chapter 2 are but one example of an alternative structure. Such 'nonlinear' splines are developed in [17], [18] and [24].

In Chapter 1 we introduce the nonlinear boundary value problem and relevant results concerning a linear problem. These results are utilised, in Chapter 4, to establish an existence and convergence result for the collocation scheme of Chapter 3. The intrinsic characteristics of the class of regular splines described in Chapter 2 are exhibited by the cubic spline. In fact, the latter is a member of this class. Thus it is to be expected that the collocation scheme based on regular splines is a generalisation of the cubic spline collocation scheme. For schemes utilising the cubic spline see [2], [3], [15-17]. Finally, in Chapter 5 computational aspects of the regular spline collocation method are discussed and numerical examples evaluated and compared.

1. A Two-Point Boundary Value Problem

In this first section we consider the homogeneous boundary value problem

$$Ny(x) \equiv y''(x) - f(x,y(x)) = 0 \ , \ 0 < x < 1 \tag{1.1}$$

$$y(0) = y(1) = 0 \tag{1.2}$$

where f(x,y) is twice continuously differentiable, with respect to x and y, in a region D of the (x,y) plane intercepted by two lines $x = 0$ and $x = 1$. (For simplicity we denote $f_y \equiv \partial f/\partial y$).

Problems characterised by $Ny(x) = 0$ but defined over a general interval $a \leq x \leq b$ and incorporating non-homogeneous boundary data are equivalent to (1.1) - (1.2). Such problems may be reduced to our problem by the application of linear transformations in x and y.

To ensure that the solution, y(x), of (1.1) - (1.2) is unique in a subregion of D we follow Keller [10], and Urabe [22], by introducing the concept of an 'isolated solution'. A solution, y(x), of (1.1) - (1.2) is said to be isolated if and only if the linearised problem

$$L\,[y]\phi(x) \equiv \phi''(x) - f_y(x,y(x))\,\phi(x) \ , \ 0 < x < 1 \tag{1.3}$$

$$\phi(0) = \phi(1) = 0 \tag{1.4}$$

has only the trivial solution $\phi(x) \equiv 0$. Following the two aforementioned papers we note that an isolated solution is 'locally unique'; that is, no other solution to (1.1) - (1.2) exists in a sufficiently small neighbourhood of the isolated solution.

We summarise here some results concerning a linear two-point boundary value problem of the form

$$L\,u(x) \equiv u''(x) - A(x)u(x) = g(x) \ , \ 0 < x < 1 \tag{1.5}$$

$$u(0) = u(1) = 0 \tag{1.6}$$

where A(x) and g(x) are sufficiently smooth, say $A(x), g(x) \in C^1[0,1]$. It is well-known that the above problem has an unique solution if and only if $\lambda = 0$ is not an eigenvalue of the continuous eigenvalue problem $Lu = \lambda u$, $u(0) = u(1) = 0$. We now introduce a linear difference method for the problem (1.5) - (1.6). Define the set of equally spaced knots $\{x_j\}_{j=1}^{m+1}$ where $x_j = (j-1)h$ and $h = 1/m$. Let u_j , an approximant to $u(x_j)$, satisfy

$$u_1 = 0$$

$$u_{j-1}(-\frac{1}{h^2} + \frac{1}{6} A_{j-1}) + u_j(\frac{2}{h^2} + \frac{2}{3} A_j) + u_{j+1}(-\frac{1}{h^2} + \frac{1}{6} A_{j+1})$$

$$= \frac{1}{6} (g_{j-1} + 4g_j + g_{j+1}) \qquad j = 2, 3, \ldots, m$$

$$u_{m+1} = 0 \tag{1.7}$$

Note that for brevity of notation $A(x_j) \equiv A_j$ and $g(x_j) \equiv g_j$ etc. We use Theorems (3.1) and (3.3) of H. - O. Kreiss [11] to establish;

Lemma 1.1

Let $A(x), g(x) \in C^1[0,1]$ be such that (1.5) - (1.6) has an unique solution. Then there exist positive constants h_o and K_o such that for any $h \le h_o$ the linear difference equation (1.7) has an unique solution which is bounded by

$$\max_{2\le j\le m} |u_j| \le K_o \max_{1\le j\le m+1} |g_j| \tag{1.8}$$

The linear system (1.7) can be written in matrix-vector notation as

$$J_h \left[A(x)\right] \underline{u} = \underline{G} \tag{1.9$_1$}$$

where $J_h \left[A(x)\right]$ is the triple-diagonal matrix with elements $J(A)_{i,j}$:

$$\begin{aligned}
J(A)_{1,1} &= 1 \\
J(A)_{j,j-1} &= -\frac{1}{h^2} + \frac{1}{6} A_{j-1} && j = 2,3,\ldots,m \\
J(A)_{j,j} &= \frac{2}{h^2} + \frac{2}{3} A_j && j = 2,3,\ldots,m \\
J(A)_{j,j+1} &= \frac{-1}{h^2} + \frac{1}{6} A_{j+1} && j = 2,3,\ldots,m \\
J(A)_{m+1,\ m+1} &= 1
\end{aligned} \qquad (1.9_{11})$$

and the vectors $\underline{u}$ and $\underline{G}$ respectively denote

$$\underline{u} = (u_1, u_2, \ldots\ldots, u_{m+1})^T, \quad \underline{G} = (\overline{g}_1, \overline{g}_2, \ldots, \overline{g}_{m+1})^T$$

where $\overline{g}_j = \frac{1}{6}(g_{j-1} + 4g_j + g_{j+1})$, $j=2, 3, \ldots, m$; $\overline{g}_1 = \overline{g}_{m+1} = 0$

Lemma (1.1) implies that $J_h[A]$ is nonsingular for any $h \le h_o$. Define the matrix norm $\|\cdot\|$ by

$$\| B \| = \sup_{\underline{x} \in \mathbb{R}^{m+1}} \frac{\| B\underline{x} \|}{\|\underline{x}\|}$$

for any matrix B of dimensions (m+1) x (m+1), where for any vector $\underline{x} \in \mathbb{R}^{m+1}$

$$\| \underline{x} \| = \max_{1 \le j \le m+1} |x_j|$$

Thus by (1.8)

$$\| J_h [A]^{-1} \| \le K_o \qquad (1.10)$$

and we have established a bound over the family of matrices J_h^{-1}, $h \le h_o$.

2. Regular Splines

Regular splines, as employed in the following context, were introduced by Werner [23]. Let $0 = z_1 < z_2 \ldots < z_{n+1} = 1$, and consider functions $\hat{t}_\ell(x,c,d)$, defined for $x \in [z_\ell, z_{\ell+1}]$, depending on two parameters c,d where c and d are in certain prescribed intervals, for example $\mathbb{R}$ or $\mathbb{R}_+$. The functions, $\hat{t}_\ell$, are subject to conditions defined below.

The set of equally spaced knots $\{x_i\}_{i=1}^{m+1}$ is specified such that each z_ℓ, $\ell = 1, 2, \ldots, n+1$, coincides with a knot x_p for some $p \in \{i\}_{i=1}^{m+1}$. Hence each interval I_p, where $I_p \equiv [x_p, x_{p+1}]$, is contained in exactly one interval $[z_\ell, z_{\ell+1}]$. The notation $t_j(x,c,d)$ is used to denote the restriction of $\hat{t}_\ell(x,c,d)$ to I_j. In this way the functions $t_j(x,c,d)$ are well defined when the mesh of knots is refined.

For a given set of knots $\{x_i\}_{i=1}^{m+1}$ and classes of functions $\{\hat{t}_\ell\}_{\ell=1}^{n}$ which are twice continuously differentiable with respect to x, and continuous with respect to c and d, we define a spline by

$$\eta(x_1, x_2, \ldots, x_{m+1}; x) = \{u(x) \,|\, u(x) \in C^2[0,1],\ u|_{I_j} \equiv p_j + t_j(x, c_j, d_j),$$
$$p_j \in \Pi\ ;\ j = 1,2,\ldots,m\} \qquad (2.1)$$

where Π is the set of linear polynomials.

In the context of this paper we need the following assumptions on the classes of functions $t_j(x,c,d)$:

(A1) The classes $t_j(x,c,d)$ shall be <u>regular</u>.

i.e. any two functions of the same class either coincide or the difference of their second derivatives have at most one zero in I_j.

This assumption ensures that the functions $t_j(x,c_j,d_j)$ $j=1,2,\ldots,m$ can be parameterised in terms of $t_j''(x_j)$ and $t_j''(x_{j+1})$. To explain (A1) in greater detail it is necessary to quote a theorem by Werner [23] , namely

'If the family of splines is made up of regular functions then the interpolating spline is unique.'

Thus regularity is a sufficient condition for uniqueness of the spline which interpolates the data produced by the collocation method of Chapter 3, always assuming that the data is within the range of the spline.

If the classes of functions $t_j(x,c,d)$ are parameterised as above we adopt the notation

$$t_j \equiv t_j(x_j, x_{j+1}, M_j, M_{j+1} ; x)$$

where $\dfrac{d^2}{dx^2} t_j(x_j, x_{j+1}, M_j, M_{j+1} ; x_i) = M_i$ for $i = j$ and $j+1$

(A2) The functions t_j are <u>smooth</u>.

i.e. the functions t_j are four times continuously differentiable with respect to x, and these derivatives are twice continuously differentiable with respect to M_j and M_{j+1}.

(A3) The functions t_j are <u>4-bounded</u>.

i.e. the fourth derivative, t_j^{iv} , of t_j with respect to x shall be a twice continuously differentiable function of $\frac{1}{h} \cdot (M_{j+1} - M_j)$ and either M_j or M_{j+1}.

The assumption (A3) is motivated by the fact that only two parameters are needed to control the behaviour of the second and higher derivatives of the spline. It would be unwise to use bounds on M_j and M_{j+1} as when $h \to 0$ the two parameters become increasingly identified with each other. However M_j and $\frac{1}{h}(M_{j+1} - M_j)$ are well defined as $h \to 0$.

Examples of admissible classes of functions t_j are now given, cf. [19, pp.176].

Example 1 $\quad t_j = c_j(d_j - x + x_j)^k \qquad k \neq 0,1,2$

$$\text{eg.} \quad d_j = \frac{h}{1 - \left(\frac{M_{j+1}}{M_j}\right)^{\frac{1}{k-2}}} \qquad \text{and} \quad c_j = \frac{M_j d_j^{2-k}}{k(k-1)} \quad \text{whenever } M_j \neq 0$$

The above classes of functions yield splines of various structures for different values of k. For any $k < 0$ we have a rational spline. The standard cubic spline is derived by allocating to k the value three. The condition for (A1) - (A3) to hold is given by

$$\frac{M_j}{M_{j+1}} \in \mathbb{R}_+ \quad , \quad M_j \neq M_{j+1}$$

unless $k = 2n+1$ where n is any positive integer. For the latter cases (A1) - (A3) hold unconditionally whenever $M_j \neq M_{j+1}$.

Example 2 $\quad t_j = c_j e^{d_j(1+x-x_j)}$

where $\quad d_j = \frac{1}{h} \log\left(\frac{M_{j+1}}{M_j}\right) \quad$ and $\quad c_j = \frac{M_j}{d_j^2 e^{d_j}}$

Once again, the necessary and sufficient condition for (A1) - (A3) to hold is

$$\frac{M_j}{M_{j+1}} \in \mathbb{R}_+ , \quad M_j \neq M_{j+1}$$

Example 3 $\quad t_j = c_j \log(d_j - x + x_j)$

where $\quad d_j = \dfrac{h}{1 - \left(\dfrac{M_j}{M_{j+1}}\right)^{\frac{1}{2}}}$ and $c_j = -M_j d_j^2$

For this class of function the conditions (A1) - (A3) are satisfied whenever

$$0 < \frac{M_j}{M_{j+1}} < 1$$

Example 4 $\quad t_j = c_j \sin\left(\mu(x - x_j) + d_j\right) \qquad \mu \neq 0$

eg. $\quad d_j = \cot^{-1}\left\{ \dfrac{1}{\sin(\mu h)} \left[\dfrac{M_{j+1}}{M_j} - \cos(\mu h)\right]\right\}$

and $\mu^2 c_j = -M_j \operatorname{cosec} d_j$ whenever $M_j \neq 0$.

Functions of this class satisfy (A1) - (A3) unconditionally.

Following Schaback [19] and Werner [23] we define the difference operators

$$\Delta^1 (x_1, x_2) g(x) = \frac{g(x_2) - g(x_1)}{x_2 - x_1}$$

$$\Delta^2 (x_1, x_2, x_3) g(x) = \frac{1}{x_3 - x_1}\left[\Delta^1 (x_2, x_3) g(x) - \Delta^1 (x_1, x_2) g(x)\right]$$

where x_1, x_2, x_3 are piecewise disjoint and $g(x) \in C[0,1]$.

If the function $g(x)$ is differentiable we may allow x_1 and x_2 to coalesce and obtain

$$\Delta^2 (x_1, x_1, x_3) g(x) = \frac{1}{x_3 - x_1}\left[\frac{g(x_3) - g(x_1)}{x_3 - x_1} - g'(x_1)\right]$$

We adopt the following notation :

$$\Delta^2 (x_j, x_j, x_{j+1}) t_j \equiv p(x_j, x_{j+1}, M_j, M_{j+1}) \qquad j = 1, 2, \ldots, m.$$

The following lemma and corollary are by Werner [23] .

Lemma 2.1

Let $u(x) \in C^4[x_1,x_2]$, and $u''(x_i) = M_i$ for $i = 1$ and 2, then

$$\Delta^2(x_1|x_1,x_2)u(x) = \frac{1}{3}M_1 + \frac{1}{6}M_2 + R(x_1,x_2,u^{\overline{\underline{iv}}}(x))$$

where $R(x_1,x_2,u^{\overline{\underline{iv}}}(x)) = -\frac{h^2}{24} u^{\overline{\underline{iv}}}(\xi), \qquad x_1 < \xi < x_2.$

Corollary 2.2

If u(x) is four times continuously differentiable with respect to x, and these derivatives are twice continuously differentiable with respect to the parameters M_1 and M_2, then the remainder term above is also twice continuously differentiable with respect to these parameters.

Furthermore, if f(x,y) is a twice continuously differentiable function with respect to x and y, and $M_j \equiv f(x_j,y_j)$ for $j = 1$ and 2, then the remainder term is twice continuously differentiable with respect to y_j for $j = 1$ and 2.

3. A Regular Spline Collocation Method

In this chapter we derive a collocation method that yields a regular spline as an approximate solution to the problem (1.1) - (1.2).

From (2.1) any regular spline, u(x), satisfying (A1) can be expressed by

$$u(x)|_{I_j} = a_j + b_j x + t_j(x) \qquad j = 1,2,\ldots, m$$

where the parameters are still undetermined. Following Werner [23] the linear parameters a_j and b_j may be determined in terms of $u(x_j)$ and $u(x_{j+1})$ giving

$$u(x)|_{I_j} = u(x_j) - t_j(x_j) + (u(x_{j+1}) - u(x_j) + t_j(x_j) \qquad (3.1)$$

$$-t_j(x_{j+1}))\left(\frac{x-x_j}{h}\right) + t_j(x) \qquad j = 1,2,\ldots,m$$

The function u(x) and its second derivative are continuous for $x \in [0,1]$. Hence the conditions

$$u'(x_{j+1})|_{I_j} = u'(x_{j+1})|_{I_{j+1}} \qquad j = 1,2,\ldots,m-1 \qquad (3.2)$$

are necessary and sufficient for $u(x) \in C^2[0,1]$. The expression (3.2) applied to (3.1) yields a relationship analogous to the consistency relationship of cubic splines [1, pp 284], namely

$$p(x_j,x_{j-1}, M_j,M_{j-1}) + p(x_j,x_{j+1}, M_j,M_{j+1}) = 2\Delta^2(x_{j-1},x_j,x_{j+1})u(x)$$

$$j = 2,3,\ldots,m \qquad (3.3)$$

A collocation method is derived by fitting the equation (3.1) to the problem (1.1) - (1.2) at the set of knots $\{x_j\}_{j=1}^{m+1}$. Setting $u_j \equiv u(x_j)$ this can be written as

$$M_j = f(x_j,u_j) \qquad j = 1,2,\ldots,m+1 \qquad (3.4)$$

Equations (3.3) and (3.4), when combined, yield a non-linear system of equations, $\underline{F}(\underline{u}) = \underline{0}$, from which the knot values of the regular spline collocation solution are calculated. This system of m-1 equations in m-1 unknowns is given by

$$\begin{aligned}
u_1 &= 0 \\
N_h^j[\underline{u}] \equiv\ & p(x_j,x_{j-1},f(x_j,u_j),f(x_{j-1},u_{j-1})) \\
&+ p(x_j,x_{j+1},f(x_j,u_j),f(x_{j+1},u_{j+1})) \\
&- \frac{1}{h^2}(u_{j-1} - 2u_j + u_{j+1}) = 0 \qquad j = 2,\ldots,m \\
u_{m+1} &= 0 \qquad (3.5)
\end{aligned}$$

Let us assume that the system of equations (3.5) has an unique solution $\underline{u}^*$. The vector $\underline{u}^*$ yields values at the knots, which, combined with (3.1) and (3.4), construct a regular spline approximant, $y^*(x)$, to $y(x)$.

4. An Existence Theorem

This chapter is devoted to establishing the following theorem.

Theorem

Let the problem (1.1) - (1.2) be defined for a function f(x,y) that is twice continuously differentiable with respect to x and y in the region D of the (x,y) plane intercepted by two lines x = 0 and x = 1. Further assume that this problem has an isolated solution $y(x) \in C^4[0,1]$ in a region U where, for some $\tau > 0$.

$$U \equiv \{(x,y) \mid |y(x)-y| < \tau , \quad 0 \le x \le 1\} \subset D.$$

Define the closed sphere $S^{\rho}[y(x)]$, $\rho \le \tau$, by

$$S^{\rho}[y(x)] \equiv \{ \underline{V} \in \mathbb{R}^{m+1} \mid \underline{V}^{T} \equiv (0,V_2,\ldots,V_m,0), \ |V_j - y(x_j)| < \rho, \quad j = 2,3,\ldots,m\}$$

Then we select the classes of functions $\{t_j\}_{j=1}^{m}$ from the space of all functions, satisfying (A1) - (A3), that permit $\tilde{M}_i = f(x_i,(\underline{v})_i)$, i=j and j+1, as admissible values for M_j and M_{j+1} respectively, for all $\underline{v} \in S^{\rho}[y(x)]$. Finally, let ρ and h_o be sufficiently small so that, for some $\delta > 0$ and $h \le h_o$

$$|\tilde{M}_i - y''(x_i)| < \delta \qquad i = 1,2,\ldots,m+1$$

and

$$|\frac{1}{h} (\tilde{M}_{i+1} - \tilde{M}_i) - y'''(x_i)| < \delta \qquad i = 1,2,\ldots,m$$

Then for ρ and h_o sufficiently small

(i) the difference scheme $\underline{F}(\underline{u}) = \underline{0}$ (ie. (3.5)) has an unique solution $\underline{u}^* \in S^{\rho}[y(x)]$, for all $h \le h_o$.

(ii) $|y(x) - y^*(x)| = O(h^2)$ ($y^*(x)$ is the regular spline collocation solution).

Proof.

For any $\underline{u} \in S^{\rho}[y(x)]$ the regular spline u(x) that satisfies $u(x_i) = \underline{u}_i$, $u''(x_i) = f(x_i, \underline{u}_i) \equiv \tilde{M}_i$, i=j and j+1, is readily seen to be 4-bounded over I_j. Thus, by lemma 2.1 we may rewrite the system of equations $\underline{F}(\underline{u}) = \underline{0}$ as

$$u_1 = 0$$

$$N_h^j[\underline{u}] \equiv \frac{1}{6}\ f(x_{j-1},u_{j-1})+4f(x_j,u_j)+f(x_{j+1},u_{j+1}) \ - \frac{1}{h^2}(u_{j-1}-2u_j+u_{j+1})$$

$$+R(x_j,x_{j-1},u^{\overline{iv}}(x)) + R(x_j,x_{j+1},u^{\overline{iv}}(x)) = 0 \quad j=2,3\ldots,m$$

$$u_{m+1} = 0 \tag{4.1}$$

Since $J_h[A]$ of (1.9) is nonsingular for any $A(x) \in C^1[0,1]$, $h \le h_o$, let us select $A(x) \equiv f_y(x,y(x))$. Hence (4.1) is equivalent to

$$\underline{u} = \underline{u} - J_h[f_y]^{-1}\ \underline{F}(\underline{u}) \equiv \underline{\psi}[\underline{u}] \tag{4.2}$$

Let $\underline{v}$ and $\underline{w}$ be arbitrary vectors of $S^{\rho}[y(x)]$, then

$$\underline{\psi}[\underline{v}] - \underline{\psi}[\underline{w}] = \underline{v} - \underline{w} - J_h[f_y]^{-1}\left[\underline{F}(\underline{v}) - \underline{F}(\underline{w})\right] \tag{4.3}$$

To apply the mean value theorem on (4.3) we must first establish that $\underline{F} : S^{\rho}[y(x)] \to \mathbb{R}^{m+1}$ is a C^1 map; ie. (4.1) is continuously differentiable with respect to $\underline{u}$, $\underline{u} \in S^{\rho}[y(x)]$. The function f(x,y) is continuously differentiable with respect to y by definition. Consequently, the task reduces to proving that $R(x_j, x_i, u^{\overline{iv}}(x))$, i=j-1 and j+1, are continuously differentiable with respect to $\underline{u}$. The 4-bound on u(x), $\underline{u} \in S^{\rho}[y(x)]$ has already been established, and hence, from (A2) and corollary 2.2, R is continuously differentiable with respect to $\underline{u}$. Concluding that $\underline{F} : S^{\rho}[y(x)] \to \mathbb{R}^{m+1}$ is a C^1 map, and using the convexity of $S^{\rho}[y(x)]$, we deduce from the mean value theorem that, for any $\underline{v}$, $\underline{w} \in S^{\rho}[y(x)]$

$$\underline{F}\,(\underline{v}) - \underline{F}\,(\underline{w}) = \int_0^1 \frac{\partial}{\partial \underline{u}} \underline{F}(s\underline{v} + (1-s)\underline{w})ds \,\left[\underline{v} - \underline{w}\right]$$

$$\equiv \partial \underline{\tilde{F}} \left[\underline{v}, \underline{w}\right] \big/ \partial \underline{u} \,\left[\underline{v} - \underline{w}\right] \qquad (4.4)$$

Here $\partial \underline{F}(\underline{u}) / \partial \underline{u}$ is the Jacobian of the system (4.1). The non-zero elements of this matrix are

$$\left(\partial \underline{F}(\underline{u}) / \partial \underline{u} \right)_{1,1} = 1$$

$$\left(\partial \underline{F}(\underline{u}) / \partial \underline{u} \right)_{j,j-1} = -\frac{1}{h^2} + \frac{1}{6} f_y(x_{j-1}, u_{j-1}) + \frac{\partial}{\partial u_{j-1}} R(x_j, x_{j-1}, u^{\overline{iv}}(x))$$

$$\left(\partial \underline{F}(\underline{u}) / \partial \underline{u} \right)_{j,j} = \frac{2}{h^2} + \frac{2}{3} f_y(x_j, u_j) + \frac{\partial}{\partial u_j} R(x_j, x_{j-1}, u^{\overline{iv}}(x))$$

$$+ \frac{\partial}{\partial u_j} R(x_j, x_{j+1}, u^{\overline{iv}}(x)) \qquad (4.5)$$

$$\left(\partial \underline{F}(\underline{u}) / \partial \underline{u} \right)_{j,j+1} = -\frac{1}{h^2} + \frac{1}{6} f_y(x_{j+1}, u_{j+1}) + \frac{\partial}{\partial u_{j+1}} R(x_j, x_{j+1}, u^{\overline{iv}}(x))$$

$$\left(\partial \underline{F}(\underline{u}) / \partial \underline{u} \right)_{m+1,m+1} = 1 \qquad j = 2,3,\ldots,m$$

A simple manipulation of (4.3) and (4.4) yields

$$\underline{\psi}\left[\underline{v}\right] - \underline{\psi}\left[\underline{w}\right] = J_h\left[f_y\right]^{-1}\left[J_h\left[f_y\right] - \partial \underline{\tilde{F}}\left[\underline{v},\underline{w}\right] / \partial \underline{u}\right]\left[\underline{v} - \underline{w}\right]$$

whence, by (1.10)

$$\left\| \underline{\psi}\left[\underline{v}\right] - \underline{\psi}\left[\underline{w}\right] \right\| \le K_o \left\| J_h\left[f_y\right] - \partial \underline{\tilde{F}}\left[\underline{v},\underline{w}\right] / \partial \underline{u} \right\| \; \left\| \underline{v} - \underline{w} \right\| \qquad (4.6)$$

Note that;

(1) for some s^*, $0 < s^* < 1$

$$\left| f_y(x_j, y(x_j)) - \int_0^1 f_y(x_j, sv_j + (1-s)\, w_j)ds \right| =$$

$$\left| f_y(x_j, y(x_j)) - f_y\,(x_j, s^*v_j + (1-s^*)w_j) \right| \le$$

$$K\left| y(x_j) - s^*v_j - (1-s^*)w_j \right| \le K\,\rho \qquad (4.7)$$

where K is the Lipschitz constant for $f_y(x,.)$.

(ii) Let $u_s(x)$, $x \in I_j$, be the regular spline interpolating the values $u_s(x_i) = r_i(s)$, $u_s''(x_i) = f(x_i, r_i(s))$ where, for $i=j$ and $j+1$ $r_i(s) = sv_i + (1-s)\, w_i$. By the convexity of $S^\rho\left[y(x)\right]$, $u_s(x)$ is 4-bounded over I_j, whence by lemma 2.1 and corollary 2.2

$$\int_o^1 \frac{\partial}{\partial u_p} R(x_j, x_{j+1}, u_s^{\overline{iv}}(x))\, ds = O(h^2) \qquad p=j, j+1 \qquad (4.8)$$

Remembering that $J_h\left[f_y\right]$ is expressed by (1.9) with $A(x) \equiv f_y(x,y(x))$ we establish from (4.4) - (4.8) that

$$\| J_h\left[f_y\right] - \partial\tilde{\underline{F}}\left[\underline{v},\underline{w}\right] /\partial\underline{u} \| \le K\rho + K_1h^2 \qquad (4.9)$$

for some positive constant K_1, and h sufficiently small.

From (4.6) and (4.9) we conclude that for ρ, h sufficiently small

$$\| \underline{\psi}\left[\underline{v}\right] - \underline{\psi}\left[\underline{w}\right] \| \le \alpha \| \underline{v} - \underline{w} \| , \quad \alpha = K_oK\rho + K_oK_1h^2 < 1 \qquad (4\ .10)$$

The vector $\underline{Y} \equiv (0,y(x_2),\ldots,y(x_m),0)^T$ is the centre of the sphere $S^\rho\ [y(x)]$. Now, $\underline{F}\ (\underline{Y})$ may be estimated by (4.1)

i.e. $$\|\underline{F}\ (\underline{Y})\| = \max_{2\le j\le m} | N_h^j\left[\underline{Y}\right] |$$
$$\le K_2h^2 \qquad (4.11)$$

whenever h is sufficiently small. The bound (4.11) utilises (1.1), the continuity of y(x), the 4-bound over u(x), $x \in I_j$, for any vector $\underline{u} \in S^\rho\left[y(x)\right]$ and lemma 2.1. The expression (4.11) yields a bound on the local truncation error. Now, for any h sufficiently small, we have from (1.10), (4.2) and (4.11) that

$$\| \underline{Y} - \underline{\psi}\left[\underline{Y}\right]\| \le \| J_h\left[f_y\right]^{-1} \| \quad \|\underline{F}\ (\underline{Y})\|$$
$$\le K_o\ K_2\ h^2 \le (1-\alpha)\ \rho \qquad (4.12)$$

The expressions (4.10) and (4.12) verify that $\underline{\psi}\left[\underline{u}\right]$ takes $S^\rho\left[y(x)\right]$ into itself and is a contraction mapping whenever ρ,h_o

are sufficiently small. Thus $\underline{u} = \underline{\psi}\left[\underline{u}\right]$ has an unique solution $\underline{u}^* \in S^{\rho}\left[y(x)\right]$ from which we immediately deduce that $\underline{u}^*$ is the unique solution of $\underline{F}(\underline{u}) = \underline{0}$.

To determine an error estimate let the j^{th} component, ε_j, of the vector $\underline{\varepsilon}$ satisfy

$$\varepsilon_j = u_j^* - y(x_j) \quad , \quad j = 1,2,\ldots,m+1.$$

Now, as $\underline{u}^*$ is the unique solution of (3.5)

$$\underline{F}\ (\underline{u}^*) = \underline{F}\ (\underline{Y} + \underline{\varepsilon}) = \underline{0}$$

Consequently, the mean value theorem yields

$$\begin{aligned} -\underline{F}\ (\underline{Y}) &= \underline{F}\ (\underline{Y} + \varepsilon) - \underline{F}\ (\underline{Y}) \\ &= \partial\underline{\tilde{F}}\left[\underline{Y} + \underline{\varepsilon}, \underline{Y}\right]/\partial\underline{u}\ \underline{\varepsilon} \end{aligned}$$

and we may rewrite the above equation as

$$J_h\left[f_y\right]\underline{\varepsilon} = \left[J_h\left[f_y\right] - \partial\underline{\tilde{F}}\left[\underline{Y} + \underline{\varepsilon}, \underline{Y}\right]/\partial\underline{u}\right]\underline{\varepsilon} - \underline{F}\ (\underline{Y}) \tag{4.13}$$

By identical arguments employed when deriving (4.19), it is simple to show that

$$\left\| J_h\left[f_y\right] - \partial\underline{\tilde{F}}\left[\underline{Y} + \underline{\varepsilon}, \underline{Y}\right]/\partial\underline{u} \right\| \le K \left\|\underline{\varepsilon}\right\| + K_1h^2 \tag{4.14}$$

Since $J_h\left[f_y\right]$ has an uniformly bounded inverse, ie. (1.10), we use (4.11), (4.13) - (4.14) to deduce

$$\left\|\underline{\varepsilon}\right\| \le K_oK \left\|\underline{\varepsilon}\right\|^2 + K_oK_1h^2 \left\|\underline{\varepsilon}\right\| + K_oK_2h^2$$

whenever h is sufficiently small. Making h_o smaller if necessary we achieve

$$(1 - K_oK_1h^2)\left\|\underline{\varepsilon}\right\| \le K_oK\left\|\underline{\varepsilon}\right\|^2 + K_oK_2h^2 \ , \quad h \le h_o$$

The scalar equation $bx \le ax^2 + c$, with $a > 0$ and $4ac < b^2$, implies that either $x < x_-$ or $x > x_+$ where $x_{\underline{+}}$ are the two real roots of the equation $ax^2 - bx + c = 0$.

ie. $\underline{x_{\pm}} = (b \pm \sqrt{b^2 - 4ac})/2a$

Let h_o be smaller if necessary so that, say

$$4KK_o^2K_2h^2 < \frac{1}{2}(1 - K_oK_1h^2)^2 \quad , \quad h \le h_o$$

Similarly, as $\| \underline{\varepsilon} \| = \| \underline{Y} - \underline{u}^* \| \le \rho$, let ρ be sufficiently small to allow

$$\rho < (b + \sqrt{b^2 - 4KK_o^2K_2h^2})/2K_oK \text{ , where } b = 1 - K_oK_1h^2$$

then it follows that

$$\| \underline{\varepsilon} \| \le \left(\frac{1-K_oK_1h^2}{2K_oK} \right) \left(1 - \left[1 - \frac{4KK_o^2K_2h^2}{(1-K_oK_1h^2)^2} \right]^{\frac{1}{2}} \right)$$

Using the inequality, $1-x \le (1-x)^{\frac{1}{2}}$, $0 \le x \le 1$, we have proved that

$$\| \underline{\varepsilon} \| = \| \underline{Y} - u^* \| \le 2K_oK_2h^2 \Big/ (1-K_oK_1h^2) \quad \text{for any } h \le h_o \qquad (4.15)$$

The global error bound may now be established. Let $\varepsilon(x) \equiv y(x) - y^*(x)$, and note that by the continuity of $y(x)$ and the 4-bound on $y^*(x)$, $x \in I_j$, the second derivative $\varepsilon''(x)$ is uniformly bounded over I_j. Consequently,

$$\varepsilon(x) = \frac{\varepsilon_j(x_{j+1} - x) + \varepsilon_{j+1}(x - x_j)}{h} + \frac{(x-x_j)(x_{j+1}-x)\ \varepsilon''(\tilde{x})}{2}$$

$$= O(h^2)$$

for all $x \in I_j$, and some $\tilde{x}$, $x_j < \tilde{x} < x_{j+1}$. The proof is now complete.

Remark

The choice of $\rho = O(h)$ is compatible with the assumptions of the theorem. This order of accuracy is normally the minimum required by the starting value of any iterative method proposed to solve (3.5).

5. Computational Aspects and Examples

The selection of optimum classes of functions $\{\hat{t}_\ell\}_{\ell=1}^{n}$ is of fundamental importance. This process relies heavily on a preconceived notion of the analytic solution. Hopefully, this can be derived from the formulation of the problem. However, for a particular type of equation, we can be considerably influenced by the structure of the function f(x,y(x)) or by predetermining a characteristic of the analytic solution. A possibility is to assume a power series expansion for y(x)

$$\text{ie.} \quad y(x) = (x-a_1)^{\alpha}\,(a_2 + a_3(x-a_1) + \ldots\ldots\ldots) \qquad (5.1)$$

The exponent, α, is determined by substituting (5.1) into equation (1.1) and equating the least exponents of $(x-a_1)$ on either side of the equation. Consequently, a feasible solution may incorporate the function

$$\hat{t}_\ell(x,c,d) = c(x-d)^{\alpha} \quad , \quad \alpha \neq 0,1,2.$$

Flexibility of application is an important feature of regular splines and different classes may be deployed over consecutive intervals $[z_{\ell-1},z_\ell]$, $[z_\ell,z_{\ell+1}]$. Computationally, this is facilitated by expressing (3.5) in a simplified form. For an arbitrary regular spline, u(x), defined over I_j, we have by lemma 2.1 that

$$\Delta^2(x_j,x_j,x_{j+1})u(x) = \frac{M_j}{3} + \frac{M_{j+1}}{6} + A_{j,j+1}. \qquad (5.2)$$

With predetermined expressions for the A's, the terms (3.5) and (5.2) yield a computationally versatile system of equations, namely

$$u_1 = 0$$

$$\frac{1}{6} f\,(x_{j-1},u_{j-1}) + \frac{2}{3} f\,(x_j,u_j) + \frac{1}{6} f\,(x_{j+1},u_{j+1}) - \frac{1}{h^2}(u_{j-1} - 2u_j + u_{j+1})$$

$$+ A_{j,j+1} + A_{j,j-1} = 0 \qquad j = 2,3,\ldots,m$$

$$u_{m+1} = 0 \qquad (5.3)$$

The expressions $A_{j,j+1}$ for the examples of chapter 2 are

Example 1

$$A_{j,j+1} = \frac{M_{j+1}}{k(k-1)}\left(a^2 - 2a+1 - \frac{k(k-1)}{6}\right) - \frac{M_j}{k(k-1)}\left(a^2-ka + \frac{k(k-1)}{3}\right)$$

where $a = \dfrac{1}{1 - \left(\dfrac{M_{j+1}}{M_j}\right)\dfrac{1}{k-2}}$

When k = 3, the cubic spline, $A_{j,\ j+1} = 0$

Example 2

$$A_{j,j+1} = \sum_{n=4}^{\infty}\left(\frac{6-n^2+n}{6.n!}\right) M_j \left[\log\left(\frac{M_{j+1}}{M_j}\right)\right]^{n-2}$$

Example 3

$$A_{j,j+1} = - M_j\left(\frac{a^2}{2}\log\left(\frac{M_j}{M_{j+1}}\right) + a + \frac{1}{3}\right) - \frac{M_{j+1}}{6}$$

$$\text{where}\quad a = \frac{1}{1 - \left(\dfrac{M_j}{M_{j+1}}\right)^{\frac{1}{2}}}$$

Example 4

$$A_{j,j+1} = - M_j \sum_{n=1}^{\infty} (-1)^{n+1} (\mu h)^{2n}\left(\frac{1}{(2n+2)!} - \frac{1}{6.2n!}\right)$$

$$+ \frac{M_j \cos \mu h - M_{j+1}}{\sin \mu h} \sum_{n=1}^{\infty} (-1)^{n+1} (\mu h)^{2n+1}\left(\frac{1}{(2n+3)!} - \frac{1}{6.(2n+1)!}\right)$$

In examples 2 and 4, a truncation of the infinite series for A is used in numerical work.

A change in class of spline is frequently necessitated by the nature of the solution. The examples of chapter 2 illustrate that some classes of regular splines are not defined for all values of the second derivative. A common occurrence is that the sign of the second derivative must remain constant throughout the region of application. Such a spine, t(x), is invalid in a small neighbourhood of any point, η, where $t''(\eta) = 0$. The spline t(x) is obviously a 'bad fit' to the analytic solution in the neighbourhood of the point $x = \eta$. Consequently, we require a criterion to determine the deployment of the spline t(x).

To illustrate how a regular spline collocation scheme may be applied we investigate a hypothetical problem. Assume that y(x) has a singularity at $x = a_1$, $a_1 \notin [0,1]$, but is regular elsewhere, ie. y(x) is given by (5.1) with appropriate constants $\{a_i\}_{i=2}^{\infty}$. The exponent, α, is determined as previously stated and hence the splines to be incorporated in the solution include the rational spline,

$$t(x) = a_j + b_j\ (x-x_j) + \frac{c_j}{(d_j-x+x_j)^{\alpha}} \qquad x \in I_j$$

The effect of the singularity on y(x), $x \in [0,1]$, whether significant or not, lies chiefly in the region of a boundary point. The scheme proposed is to apply the rational spline over $[0,a']$, $[b',1]$ and the cubic spline over $[a',b']$, for suitable a',b'. In this way the spline solution can 'fit' the effect of the singularity and rid us of the necessity to use extremely small values of h if this effect is overwhelming (cf.Problem 1). The selection of a' and b' will be influenced by the function f(x,y(x)) and its values at x = 0 and x = 1.

A judicious choice for a',b' removes the obstacle associated with the sign of y"(x). The cubic spline is only one possibility for the interval $[a',b']$, and any class of splines that is defined unconditionally may be used instead (eg. Examples 1 and 4).

Solution of the appropriate system of equations (5.3) will yield information to formulate and solve a refined system. Using this information it is possible to realize the character of y(x) by evaluating certain structural parameters, q_j, derived by a direct comparison of the supposed structure of the analytic solution to the corresponding regular spline approximant. We qualify this process by referring to Examples 1-4 of chapter 2. Assume that for $x \in [\bar{a},\bar{b}] \subseteq [0,1]$, and constants e,f,g and p:

cf. Example 1

$$y(x) \simeq e + fx + g(p-x)^k \qquad k \neq 0,1,2$$

then $q_j \simeq d_j + x_j$ for any j such that $I_j \in [\bar{a}, \bar{b}]$.

Similarly, we have

Example 2

$$y(x) \simeq e + fx + g\,e^{px} \qquad q_j \simeq d_j$$

Example 3

$$y(x) \simeq e + fx + g\log(p-x) \qquad q_j \simeq d_j + x_j$$

Example 4

$$y(x) \simeq e + fx + g\sin(\mu x+p) \qquad q_j \simeq d_j - \mu x_j$$

Returning to our hypothetical example, let us assume that the parameters $\{q_j\}_{j=1}^{r}$ of the rational spline are closely grouped but for every other class of splines the associated parameters, $\{q_j\}_{n=r+1}^{m}$ vary substantially. We decide that, for $x \in [0,x_{r+1}]$, the rational spline is a good 'fit' whilst the cubic spline is probably best for $x \in [x_{r+1},1]$

Numerical criteria implementing the above ideas, are as follows: Let $\{q_j\}_{j=1}^{m}$ be the values of the parameter for an arbitrary class of regular splines

(C1)

$$\text{Let } r_j = \left| \frac{q_{j+1} - q_j}{\bar{q}_j} \right| \quad \text{where } \bar{q}_j = \max \; \{1, |q_j|\}$$

Then, if

$$|r_{p+1} - r_p| = \min_{1 \le j \le m-1} \{|r_{j+1} - r_j|\} \begin{cases} \le \text{ Ch, the spline is applicable} \\ > \text{ Ch, the spline is not applicable} \end{cases}$$

for some C $0 < C \le \frac{1}{2}$.

Normalise the values $\{q_j\}_{j=1}^{m}$ by

$$\hat{q}_j = \begin{cases} q_j & \text{if } q_p \le 1 \\ \dfrac{q_j}{q_p} & \text{if } q_p > 1 \end{cases}$$

(C2)

Apply the spline over the intervals $\left[x_r, x_{p+1}\right]$, and $\left[x_p, x_s\right]$ where the integers r and s satisfy

$$|\hat{q}_j - \hat{q}_p| < \cdot 2 \qquad j = r, r+1, \ldots, p$$

$$|\hat{q}_p - \hat{q}_j| < \cdot 2 \qquad j = p+1, p+2, \ldots, s-1$$

The effect on the solution of the parameter C in (C1) will be discussed later.

We may now define a remodelled system of equations (5.3) based on the criteria (C1) and (C2). The solution of the first system of equations is an excellent initial value to the remodelled problem, and comparatively little extra effort is required to solve this additional iterative problem.

Four problems are evaluated by the above criteria. For comparative purposes the problems are also evaluated by the cubic spline collocation method and by Numerov's method. As previously stated, the latter is a fourth-order finite difference scheme. For simplicity of notation we define by E the absolute error, and E_r the relative error. The parameter C of (C1) is taken to be $C = \frac{1}{2}$.

Problem 1
$$y''(x) = \frac{2(y(x)+x^2)^3}{1.01^2} - 2$$

$$y(0) = 101, \quad y(1) = 0$$

$$y(x) = \frac{1.01}{(x+.01)} - x^2$$

Table 1 The regular spline solution y*(x) uses the rational spline k = -1, and the cubic spline.

x	y*(x)	E	E_r	h
.05	16.8358	4.97×10^{-3}	2.95×10^{-4}	.1
.2	4.7728	3.32×10^{-3}	6.96×10^{-4}	.1
.5	1.7331	2.68×10^{-3}	1.55×10^{-3}	.1
.025	28.8576	1.08×10^{-3}	3.76×10^{-5}	.05
.2	4.7703	8.21×10^{-4}	1.72×10^{-4}	.05
.5	1.7309	5.41×10^{-4}	3.13×10^{-4}	.05

The values of the parameters $\{\hat{q}_j\}_{j=1}^{8}$, h = .1, are

$\hat{q}_1 = -0.00999$, $\hat{q}_2 = -0.00947$, $\hat{q}_3 = -0.00583$

$\hat{q}_4 = 0.00589 \quad , \quad \hat{q}_5 = 0.0330 \quad , \quad \hat{q}_6 = 0.0854$

$\hat{q}_7 = 0.175 \quad , \quad \hat{q}_8 = 0.318$

The cubic spline and Numerov's solutions are too inaccurate to give a useful comparison with the above values of h.

<u>**Problem 2**</u> $y''(x) = \frac{(y(x) + x^4)^3}{50} - 12x^2$, $y(-\frac{1}{2}) = 19.9375$, $y(2) = -\frac{38}{3}$

$$y(x) = \frac{10}{x+1} - x^4$$

The regular spline incorporated the rational spline, k = -1, the cubic spline, and the polynomial spline, k = 4.

		Regular Spline		Cubic Spline		Numerov's Method	
x	h	E	E_r	E	E_r	E	E_r
-.3	0.1	3.90×10^{-4}	2.73×10^{-5}	3.90×10^{-2}	2.73×10^{-3}	1.23×10^{-3}	8.63×10^{-5}
0.2	0.1	1.72×10^{-4}	2.07×10^{-5}	2.31×10^{-2}	2.78×10^{-3}	6.11×10^{-4}	7.33×10^{-5}
0.8	0.1	1.05×10^{-4}	2.04×10^{-5}	7.39×10^{-3}	1.44×10^{-3}	2.67×10^{-4}	5.19×10^{-5}
1.5	0.1	5.46×10^{-4}	5.14×10^{-4}	1.36×10^{-4}	1.28×10^{-4}	9.10×10^{-5}	8.57×10^{-5}
-.3	0.05	1.01×10^{-4}	7.04×10^{-6}	9.46×10^{-3}	6.63×10^{-4}	7.99×10^{-5}	5.60×10^{-6}
0.2	0.05	4.58×10^{-5}	5.50×10^{-6}	5.66×10^{-3}	6.79×10^{-4}	3.94×10^{-5}	4.73×10^{-6}
0.8	0.05	2.74×10^{-5}	5.33×10^{-6}	1.80×10^{-3}	3.49×10^{-4}	1.72×10^{-5}	3.34×10^{-6}
1.5	0.05	1.36×10^{-4}	1.28×10^{-4}	5.00×10^{-5}	4.70×10^{-5}	5.86×10^{-6}	5.51×10^{-6}

The parameters $\{\hat{q}_j\}_{j=1}^{5}$ of the rational spline with h = .1 , are

$\hat{q}_1 = -.99798$, $\hat{q}_2 = -1.00327$, $\hat{q}_3 = -1.0123$, $\hat{q}_4 = -1.0202$, $\hat{q}_5 = -1.0143$

<u>**Problem 3**</u> $y''(x) = \frac{1}{5}(y(x) + x^2(x-1)^2)^2 - (12x^2 + 2 - 12x)$, $y(0) = 120$, $y(1) = \frac{40}{3}$

$$y(x) = \frac{30}{(x+\frac{1}{2})^2} - x^2(x-1)^2$$

The regular spline solution incorporated the rational spline, $k = 2$, and the cubic spline.

		Regular Spline		Cubic Spline		Numerov's Method	
x	h	E	E_r	E	E_r	E	E_r
.2	.1	9.93 x 10^{-4}	1.62 x 10^{-5}	0.657	1.07 x 10^{-2}	2.90 x 10^{-2}	4.74 x 10^{-4}
.5	.1	7.96 x 10^{-6}	2.66 x 10^{-7}	0.360	1.20 x 10^{-2}	1.30 x 10^{-2}	4.34 x 10^{-4}
.8	.1	5.32 x 10^{-4}	3.00 x 10^{-5}	0.130	7.34 x 10^{-3}	4.27 x 10^{-3}	2.41 x 10^{-4}
.2	.05	2.50 x 10^{-4}	4.09 x 10^{-6}	0.157	2.56 x 10^{-3}	1.90 x 10^{-3}	3.10 x 10^{-5}
.5	.05	5.88 x 10^{-6}	1.96 x 10^{-7}	8.71 x 10^{-2}	2.91 x 10^{-3}	8.47 x 10^{-4}	2.83 x 10^{-5}
.8	.05	1.20 x 10^{-4}	6.75 x 10^{-6}	3.17 x 10^{-2}	1.79 x 10^{-3}	2.77 x 10^{-4}	1.57 x 10^{-5}

<u>Problem 4</u> $y''(x) = 16y - (\pi^2 + 16)\sin \pi x$, $y(1) = e^4$, $y(2) = e^8$

$y(x) = e^{4x} + \sin \pi x$

The exponential spline comprised the regular spline solution.

		Regular Spline		Cubic Spline		Numerov's Method	
x	h	E	E_r	E	E_r	E	E_r
1.2	.1	2.05×10^{-3}	1.69×10^{-5}	1.96	1.62×10^{-2}	1.54×10^{-2}	1.28×10^{-4}
1.5	.1	1.92×10^{-3}	4.76×10^{-6}	5.25	1.30×10^{-2}	4.12×10^{-2}	1.02×10^{-4}
1.8	.1	4.27×10^{-3}	3.19×10^{-6}	7.21	5.38×10^{-3}	5.64×10^{-2}	4.21×10^{-5}
1.2	.05	5.02×10^{-4}	4.15×10^{-6}	0.485	4.01×10^{-3}	9.68×10^{-4}	8.01×10^{-6}
1.5	.05	4.84×10^{-4}	1.20×10^{-6}	1.30	3.23×10^{-3}	2.59×10^{-3}	6.44×10^{-6}
1.8	.05	1.07×10^{-3}	7.99×10^{-7}	1.78	1.33×10^{-3}	3.54×10^{-3}	2.65×10^{-6}

Discussion

The previous chapters generalise the well-established theory of the cubic spline collocation scheme, for the problem (1.1) - (1₀2), to classes of regular spline collocation schemes₀ Consequently, we may now consider classes of schemes wherein, formally, only the cubic spline option existed.

Numerically, the versatility of the proposed scheme is of major importance. The classes of splines utilised depend on the ingenuity of the user₀ These may include the examples of chapter 2 or a class derived from an intuitive idea of the dominant terms of the true solution. Corresponding to the classes employed, structural parameters will be evaluated and these may yield desirable information, e.g. the location of a singularity₀ The numerical examples of chapter 5 illustrate the increased accuracy obtainable by a judicious application of regular splines compared with the cubic spline. Also, the results give a favourable comparison with Numerov's method, for the specified values of h. However, as h → 0, a fourth order method will converge faster than the second order collocation scheme and the comparison must favour the former₀ Yet, cf. problem 1, meaningful results may be obtained by the collocation scheme when the fourth order, polynomial based method is inapplicable₀

At this point we introduce the paper by Daniel and Swartz [6]₀ They derive a fourth-order, cubic spline scheme by collocating to a perturbed differential equation which is satisfied by the cubic spline interpolant of the true solution. The generalisation of their work to incorporate regular splines is a research possibility for the future.

We now consider the effect of varying the parameter C of (C1), chapter 5. Obviously as $C \to 0$ the number of splines satisfying (C1) will decrease and may equal zero. However, a small value of C will ensure applicability of a spline that imitates the dominant structure of the true solution in a subregion of $[0,1]$. In particular C=1/16 ensures applicability of the rational splines in problem 1 and 2, whilst C=1/100 is sufficient for the rational spline in problem 1. Note that the evaluation of problem 2 by Numerov's method is perfectly acceptable, and hence, to detect problems to which regular splines are especially recommended, we suggest a value of $C \simeq 1/30$. For comparison with the cubic spline collocation scheme the value $C = \frac{1}{2}$ is acceptable.

Let us conclude with the following comments. The regular spline collocation scheme is meaningful and interesting in itself, but note that the convergence is second order. Taking the parameter $C = \frac{1}{2}$ we achieve better results than those obtained by solely considering the cubic spline. However, if a polynomial based spline closely interpolates the true solution, without requiring excessively small values of h, it appears likely that a fourth order scheme is preferable. For problems not satisfying the above condition a suitable regular spline may ease the computations. Therefore, an interesting possibility is the production of computer packages for the problem (1.1) - (1.2) involving the regular spline collocation scheme and some fourth order method. The collocation method may be applied, with $C \simeq 1/30$, to remove the necessity of using excessively small values of h. Initially we employ the collocation scheme to investigate the suitability of appropriate classes of regular splines, and then switch to the fourth order scheme if none are revealed.

References

1. AHLBERG, J.H., NILSON, E.N., and WALSH, J.L. : The theory of splines and their applications; Academic Press. (1967)

2. ALBASINY, E.L., and HOSKINS, W.D. : Cubic spline solutions to two-point boundary value problems; Computer Journal, 12, pp. 151-153. (1969)

3. ALBASINY, E.L., and HOSKINS, W.D. : Increased accuracy cubic spline solutions to two-point boundary value problems; J. Inst. Maths. Applics., 9, pp. 47-55. (1972)

4. AZIZ, A.K. (Ed): Numerical solutions of boundary value problems for ordinary differential equations; Academic Press. (1975)

5. CIARLET, P.G., SCHULTZ, M.H., and VARGA, R.S. : Numerical methods of high order accuracy for non-linear boundary value problems. I. One dimensional problems; Numer. Math., 9, pp. 394-430. (1967)

6. DANIEL, J.W., and SWARTZ, B.K. : Extrapolated collocation for two-point boundary problems using cubic splines; J. Inst. Maths. Applics., 16, pp. 161-174. (1975)

7. DOUGLAS, J., Jr., and DUPONT, T. : Galerkin approximations for the two-point boundary problems using, continuous, piecewise polynomial spaces; Numer. Math., 22, pp. 99-109. (1974)

8. HENRICI, P. : Discrete variable methods in ordinary differential equations; Wiley & Sons, New York, London, Sydney. (1962)

9. KAMMERER, W.J., REDDIEN, G.W., and VARGA, R.S.: Quadratic interpolatory splines; Numer. Math., 22, pp. 241-259. (1974)

10。 KELLER, H.B. : Accurate difference methods for nonlinear two-point boundary value problems; SIAM J。 Numer. Anal., 11, pp. 305-320. (1974)

11。 KREISS, H。 - O。 : Difference approximations for boundary and eigenvalue problems for ordinary differential equations; Math。 Comp。, 26, pp. 605-624. (1972)

12. LUCAS, T.R。, and REDDIEN, G.W., Jr. : Some collocation methods for nonlinear boundary value problems; SIAM J. Numer. Anal., 9, pp. 341-356. (1972)

13。 PERRIN, F。M., PRICE, H.S, and VARGA, R.S. : On higher-order numerical methods for nonlinear two-point boundary value problems: Numer. Math., 13, pp. 180-198. (1969)

14. RUSSELL, R.D., and SHAMPINE, L.F. : A collocation method for boundary value problems; Numer. Math., 19, pp. 1-28。 (1972)

15. SAKAI, M. : Spline interpolation and two-point boundary value problems; Mem。 Fac. Sci., Kyushu Univ。, Ser. A, 24, pp。 17-34。 (1970)

16。 SAKAI。 M。 : Piecewise cubic interpolation and two-point boundary value problems; Publ. Res. Inst。 Math。 Sci., Kyoto Univ., 7, pp。 345-362. (1971)

17. SAKAI. M. : Ritz method for two point boundary value problems; Mem. Fac. Sci。, Kyushu Univ。, Ser. A, 27, pp. 83-97. (1973)

18. SCHABACK, R. : Spezielle rationale Splinefunktionen; J. Approx. Theory, 7, pp。 281-292. (1973)

19。 SCHABACK, R. : Interpolation mit nichtlinearen Klassen von Spline-Funktionen; J. Approx。 Theory, 8, pp. 173-188. (1973)

20. SCHULTZ, M.H., and Varga, R.S. : L-splines; Numer. Math., 10, pp. 345-369. (1967)

21. STEPLEMAN, R.S. : Tridiagonal fourth-order approximations to general two-point nonlinear boundary value problems with mixed boundary conditions; Math. Comp., 30, pp. 92-103. (1976)

22. URABE, M. : An existence theorem for multi-point boundary value problems; Funkcial. Fkvac., 9, pp. 43-60. (1966)

23. WERNER, H. : Interpolation and integration of initial value problems of ordinary differential equations by regular splines; SIAM J. Numer. Anal. 12, pp. 255-271. (1975)

24. WHEELER, M.F. : An optimal L_∞ error estimate for Galerkin approximations to solution of two point boundary value problems; SIAM J. Numer. Anal., 10, pp. 914-917. (1973)

SECTION 2. FINITE ELEMENT MULTISTEP MULTIDERIVATIVE SCHEMES FOR PARABOLIC EQUATIONS

Introduction

The feasibility of applying finite element methods to parabolic equations was apparent to engineers over a decade ago. Since those pioneering days the finite element application has been thoroughly investigated by mathematicians and, as a result, established by vigorous analysis. The intensity of activity in this field is evident from the extensive literature available.

The foundation of the finite element procedure is to express the problem by its variational formulation. The propriety of such an operation has been validated by several authors, e.g. [15]. The Galerkin procedure is to approximate this weak solution over a finite element space.

We consider finite element spaces that completely cover the region of definition. Thus, for a parabolic equation defined over an one-dimensional region we may select, for example, spaces of cubic splines or Hermite cubic splines, etc.; see [28] for details. A two dimensional region may be covered by triangular or quadrilateral elements depending on the boundary shape. For quadrilateral elements tensor products of one-dimensional splines are applicable. However, given a general curved boundary, the possibilities extend to curvilinear elements, see [20-21] and [24]. Each finite element space is associated with a parameter h. For one-dimensional regions h will be the maximum length of an interval, whereas in two dimensions it is the largest side of any triangular or quadrilateral element. The chief stipulation that the finite element spaces must obey is a result from approximation theory. We require the existence of an a priori bound of order h^{p+1}, p > 0, over the interpolatory error obtained by approximating a suitably continuous function by elements of that space. Superconvergence results at the knots for the Galerkin approximation have been established, e.g. [32] and [35].

The Galerkin procedure applied to a parabolic equation establishes, in the time variable, a stiff system of initial value problems in ordinary differential equations. The theoretical solution of this problem yields the so-called 'continuous time' Galerkin solution. In practise it is customary to discretize the initial value problem by some appropriate method. The stiffness of the system necessitates the use of discretization schemes that satisfy certain stability conditions. Dahlquist [5] introduced A-stability and investigated A-stable multistep methods, whereas Widlund [38] weakened the stability criterion to study the class of A(α) - stable multistep schemes. However, for our purposes the A_o-stability criterion of Cryer [4] is sufficient.

Various authors have considered the extension of A-stable multistep methods for initial value problems to those incorporating higher derivatives. Ehle [9], Makinson [19] and Norsett [26] investigate one-step methods incorporating higher derivatives whereas Enright [10] and Jeltsch [14] have considered multistep, multiderivative formulae. Alternatively Crouzeix [3] has studied A-stable Runge-Kutta methods for initial value problems.

The application of a discretization process to the system of initial value problems yields a 'discrete time' Galerkin solution. In particular for the linear parabolic equation, discrete time solutions have been evaluated, and error analysis established, for all the aforementioned processes. Zlámal [40-41, 43] applies A_o-stable, linear multistep methods to the system of differential equations, whereas Nassif[25], and Thomeé [31] utilise A-stable one-step Padé approximations, and Crouzieux [3], Zlámal [39, 42] amongst others apply A-stable Runge-Kutta schemes. The bulk of this section is devoted to analysing the discrete time Galerkin solution to linear parabolic equations by the application of A_o-stable linear multistep, multiderivative formulae to the Galerkin system of differential equations.

Chapters 1 - 2 introduce the linear homogeneous parabolic equation and examine the continuous time Galerkin solution. In chapter 3 we define A_o-stable linear multistep, multiderivative formulae and formulate the discrete time Galerkin solution. The theorems of chapter 4 are established by the vigorous analysis of chapter 5. These theorems state optimal convergence results in the L_2-norm under extremely general conditions. The nonhomogeneous linear parabolic equation is similarly analysed in chapter 6. Here we require more restrictive assumptions on the continuity of the analytic solution. However, note that we have relaxed the stipulation hitherto of the solution being q+1 times continuously differentiable with respect to time by an analogous assumption on the nonhomogeneous term, f(x,t). In chapter 7, we construct various A_o-stable, linear multistep, multiderivative schemes and investigate their implementation procedures. It is shown that optimal order schemes invariably necessitate the use of complex arithmetic but that by easing the requirement of optimal order the implementation procedure can be considerably simplified. Test problems are evaluated and analysed in chapter 8.

Finally, in chapter 9, we investigate the solution of a class of quasi-linear parabolic equations. The application of the Galerkin procedure to the quasi-linear equation has been studied by Thomeé and Wahlbin [33]. More general non-linear equations have been tackled by several authors and we refer, in particular, to the papers by Douglas and Dupont [7], and Wheeler [36]. The above authors obtain discrete time Galerkin solutions by employing one-step time discretisation schemes. *Two-step time discretization schemes* are utilised by Dupont, Fairweather and Johnson [8], and examined more generally by Zlámal [44]. Many of the discrete-time Galerkin schemes referred to above place ease of solution at a premium and, without a reduction in the order of convergence, avoid the necessity of solving a non-linear system of equations at each time level. However, even for the comparatively simple quasi-linear equation, only second order convergence in the time increment is achieved. In comparison, a fourth order finite difference scheme for a general non-linear parabolic equation is described by Watanabe and Flood [34], but this requires the solution of a non-linear system of equations at each time level.

Motivated by the previous chapters, we now utilise a third order linear multistep, multiderivative, method to achieve an unconditionally stable discrete time Galerkin solution to the quasi-linear equation. This solution will be established to be third order accurate in the time increment and is obtained by solving linearly algebraic systems of equations at each time level.

1. The Linear Homogeneous Parabolic Equation

We shall consider the initial boundary value problem

$$\frac{\partial u}{\partial t} = \sum_{i,j=1}^{N} \frac{\partial}{\partial x_i}\left(a_{ij}(x) \frac{\partial u}{\partial x_j}\right) - a(x)u \equiv Lu \;,\; (x,t) \in \Omega \times (0,\infty) \qquad (1.1a)$$

$$u(x,0) = g(x) \quad , \quad x \in \Omega \qquad (1.1b)$$

$$u(x,t) = 0 \quad , \quad (x,t) \in \Gamma \times [0,\infty) \qquad (1.1c)$$

where $x = (x_1,\dots,x_N)$ is a point of a bounded domain Ω, with boundary Γ, lying in the N-dimensional Euclidean space. Without loss of generality the boundary value is taken to be homogeneous Dirichlet. Non-homogeneous Dirichlet and Newmann boundary conditions apply with only minor adjustments.
For simplicity we allow

$$\{a_{ij}(x)\}_{ij=1}^{N} \;,\; a(x) \in C^{\infty}(\overline{\Omega}) \quad , \Gamma \in C^{\infty}$$

where $\overline{\Omega}$ is the closure of Ω. We also assume that

$$a(x) \geq 0 \qquad (1.2_1)$$

and the matrix $a_{ij}(x)$ is uniformly positive definite

$$\text{i.e.} \quad a_{ij}(x) = a_{ji}(x) \qquad 1 \leq i,\, j \leq N, \quad x \in \overline{\Omega}$$

$$\text{and} \quad \sum_{i,j=1}^{N} a_{ij}\xi_i\xi_j \geq \gamma \sum_{i=1}^{N} \xi_i^2 \quad \text{for some positive constant } \gamma \qquad (1.2_{11})$$

Before we can formulate the weak form of the problem (1.1) it is necessary to introduce Sobolev spaces. The Sobolev space $H^m(\Omega)$ is defined to be the space of real functions which, together with their first m generalised derivatives, are in $L_2(\Omega)$ the space of square integrable functions over Ω. The space $H^m(\Omega)$ is a Hilbert space, the inner product $(\cdot,\cdot)_m$ being given by

$$(u,v)_m = \sum_{|j|\leq m} \int_\Omega D^j u D^j v \, dx$$

where $j = (j_1,\ldots,j_N)$, $|j| = j_1 + \ldots + j_N$ and $D^j u = \dfrac{\partial^{|j|} u}{\partial x_1^{j_1} \ldots \partial x_N^{j_N}}$.

The associated norm, $\|\cdot\|_m$, is defined to be

$$\|v\|_m = (v,v)_m^{\frac{1}{2}}$$

The norm and inner product on $L_2(\Omega)$ are denoted respectively by $\|\cdot\|$ and $(\cdot,\cdot)$ where

$$\|v\| = \left(\int_\Omega v^2 dx \right)^{\frac{1}{2}} , \quad (u,v) = \int_\Omega uv \, dx$$

Further we denote by $H_o^1(\Omega)$ the space of all real functions v, where $v \in H^1(\Omega)$ and $v|_\Gamma = 0$ in the generalised sense. To formulate the weak problem associated with (1.1) we multiply the equation by an arbitrary function $v \in H_o^1(\Omega)$ and integrate over Ω. Using Green's theorem we get

$$\int_\Omega \frac{\partial u}{\partial t} v \, dx + \sum_{i,j=1}^{N} \int_\Omega a_{ij}(x) \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} dx + \int_\Omega a(x) uv dx = 0 \tag{1.3}$$

We adopt the notation

$$a(u,v) = \sum_{i,j=1}^{N} \int_\Omega a_{ij}(x) \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} dx + \int_\Omega a(x) uv \, dx$$

and consequently rewrite (1.3) as

$$\left(\frac{\partial u}{\partial t}, v\right) + a(u,v) = 0 \qquad \forall\, v \in H_0^1(\Omega),\ t > 0 \tag{1.4}$$

The weak solution of the problem (1.1) is the function $u(x,t) \in H_0^1(\Omega)$ which satisfies (1.4) for all $t > 0$ and the initial identity (1.1b).

We determine the asymptotic behaviour of $u(x,t)$ by employing the 'energy method'. Denoting $\alpha(t) \equiv \| u(\cdot,t) \|$ we have by applying (1.2) to the expression (1.4)with $v = u(x,t)$

$$\alpha(t)\frac{d}{dt}\alpha(t) + \gamma\left[\alpha(t)\right]^2 \le \left(\frac{\partial u}{\partial t}, u\right) + a(u,u) = 0$$

Cancelling throughout by $\alpha(t)$, multiplying by $e^{-\gamma t}$, and integrating from 0 to T we achieve

$$\| u(x,T)\| \le e^{-\gamma T} \| g(x) \| \tag{1.5}$$

2. The Galerkin Procedure

Let V^o be a finite dimensional subspace of $H^1_o(\Omega)$. The Galerkin method is to find an approximation, $U(x,t)$, to $u(x,t)$ of the form

$$U(x,t) = \sum_{i=1}^{d} C_i(t)\ V_i(x) \tag{2.1}$$

where $\{V_i(x)\}_{i=1}^d$ is a basis of V^o. The continuous-time Galerkin solution to (1.1) is the function (2.1) where the coefficients $\{C_i(t)\}_{i=1}^d$ are determined by the discrete analogue to (1.4), namely

$$\left(\frac{\partial U}{\partial t}, V\right) + a(U,V) = 0 \qquad \text{for any } V \in V^o,\ t > 0 \tag{2.2}$$

Substituting $\{V_i(x)\}_{i=1}^d$ in turn for V in (2.2), and assembling in matrix form, we see that

$$M \frac{d}{dt} \underline{C} + K\underline{C} = \underline{0} \tag{2.3}$$

where M and K are constant, positive-definite matrices. The elements of M and K are

$$M_{ij} = (V_i,V_j) \text{ and } K_{ij} = a(V_i,V_j) \quad , \quad 1 \le i,j \le d.$$

An appropriate initial condition is derived from a discretized form of the identity (1.1b). Let $\overline{g}(x) \in V^o$ be an approximation to g(x) and define $U(x,o) = \overline{g}(x)$. This yields an initial condition for $\underline{C}(o)$, say

$$\underline{C}(o) = \underline{a} \tag{2.4}$$

The equations (2.3) and (2.4) define the continuous time Galerkin solution.

Applying the energy method to (2.2) we have, by the previously described manipulations

$$\| U(x,T) \| \leq e^{-\gamma T} \| \bar{g}(x) \| \tag{2.5}$$

The expressions (1.5) and (2.5) will be influential in our choice of time discretization schemes to approximate U(x,t). Any method that preserves the asymptotic behaviour of the true solution is 'well-posed'. This concept of 'strong stability' or well-posedness is investigated by Crouzeix [3] and Nassif [25]. Following them we define a 'k-step' approximation method to be strongly stable if U^n, the approximant to $U(x,n\Delta_t)$, satisfies

$$\| U^n \| \leq C\, e^{-\alpha n \Delta_t} \sum_{j=o}^{k-1} \| U^j \| \tag{2.6}$$

where α is some positive constant. In the sequel we use C and c as generic constants, that may differ in successive lines.

We now impose a necessary property on the subspace V^o, namely $V^o \equiv V_h^p$, where V_h^p has the property that for any $\tilde{v} \in H^{p+1}(\Omega) \cap H_o^1(\Omega)$ there exists an element $v \in V_h^p$ such that whenever h is sufficiently small

$$\| \tilde{v} - v \| + h \| \tilde{v} - v \|_1 \leq Ch^{s+1} \| \tilde{v} \|_{s+1} , \quad s = 1,2,\ldots,p \tag{2.7}$$

Any function $\phi \in V_h^p$ can be expressed as $\phi = \underline{x}^T \underline{V}$ where $\underline{x}$ is a vector of constants and $\underline{V} = (V_1, V_2,\ldots,V_d)$. We assume that the space V_h^p exhibits the following properties

(P_1) $\quad \|V_i\|_1^2 \geq ch^{-2} \|V_i\|^2 \qquad i = 1,2,\ldots,d$

(P_{11}) $\quad a(\phi,\phi) \leq Ch^{-2} \| \phi \|^2 \qquad$ for any $\phi \in V_h^p$

The above properties are satisfied by the finite element subspaces used in practise.

Let $\tilde{\Lambda}$ be an eigenvalue of the matrix $S = M^{-1}K$. The matrix S is not symmetric but its eigenvalues are readily seen to be exactly those of the positive definite matrix $M^{-\frac{1}{2}}KM^{-\frac{1}{2}}$; ie. for some eigenvector $\underline{\tilde{x}}$

$$M^{-\frac{1}{2}}KM^{-\frac{1}{2}}\underline{\tilde{x}} = \tilde{\Lambda}\,\underline{\tilde{x}} \quad , \quad \tilde{\Lambda} \text{ real and positive}$$

Further, let $\tilde{\phi} \in V_h^P$ be defined as $\tilde{\phi} = (M^{-\frac{1}{2}}\underline{\tilde{x}})^T\underline{V}$,

whence

$$\tilde{\Lambda} = \frac{a(\tilde{\phi},\tilde{\phi})}{||\tilde{\phi}||^2}$$

Utilising (P_{11}) and (1.2) we establish that

$$\gamma \leq \tilde{\Lambda} \leq Ch^{-2}$$

However, by the minimax theorem for eigenvalues

$$\tilde{\Lambda}_{max} \equiv \max\{\tilde{\Lambda}\} = \max_{\phi \in V_h^P} \frac{a(\phi,\phi)}{||\phi||^2} \tag{2.8}$$

By (P_1) and (1.2) we now prove that $\tilde{\Lambda}_{max} \geq \gamma ch^{-2}$. Similarly we can see that $\tilde{\Lambda}_{min} \equiv \min\{\tilde{\Lambda}\}$ is bounded from above.

It is important to see that the eigenvalues of $S = M^{-1}K$ are positive and unbounded with respect to h. The largest eigenvalue of S, $\tilde{\Lambda}_{max}$, is of magnitude $\tilde{\Lambda}_{max} \sim Ch^{-2}$ whereas the smallest eigenvalue is bounded from above. Consequently, the system of differential equations (2.3) is a stiff system,

$$\text{ie.} \quad \Lambda_{max}/\Lambda_{min} \gg 0 \qquad \text{as } h \to 0$$

3. A_o - stable, linear multistep, multiderivative methods

Most classical methods for solving initial value problems of first order ordinary differential equations require, for reasons of stability, a condition of the form $|\Lambda_{max} \Delta_t| < C$; where Δ_t is the time increment and C a constant usually between one and ten. For the stiff system (2.3) this condition requires $h^{-2}\Delta_t$ to be small which imposes a severe limitation on the step length Δ_t. As we will be required to solve systems of linear equations at each time interval this restriction is prohibitive. We are thus lead to consider only methods where the region of absolute stability is unbounded. Since the eigenvalues Λ of the matrix S are real the classes of A_o-stable methods are sufficient. Zlámal [40-41, 43] employed the class of A_o-stable, linear multistep methods to solve the system (2.3). Other authors, including Nassif [25], Makinson [19], have studied various one-step methods for the solution of stiff systems. Following Obrechkoff (see [16,pp199]), Enright [10], Norsett [26] amongst others, we shall consider multistep formulae that incorporate the higher derivatives. We refer to such schemes as A_o-stable, linear multistep, multiderivative methods (L.M.S.D's). This follows the terminology of Genin [12] but we note that the title 'Obrechkoff methods' is also used, e.g. [16].

A L.M.S.D. method is of the type

$$\sum_{j=o}^{k} \alpha_j \, y_{n+j} = \sum_{j=o}^{k} \sum_{r=1}^{m} \beta_{rj} \, \Delta_t^r \, y_{n+j}^r \tag{3.1}$$

where $\alpha_k > 0$ and $y_n^r \equiv \frac{d^r}{dt^r} y \Big|_{t=n\Delta_t}$

Analogous to linear multistep methods (cf.[13 , pp. 221]) the method (3.1) is said to be of order q if, for Δ_t sufficiently small

$$L[y(t),\Delta_t] \equiv \sum_{j=o}^{k} \left\{\alpha_j y(t+j\Delta_t) - \sum_{r=1}^{m} \beta_{rj}\Delta_t^r y^r(t+j\Delta_t)\right\}$$

$$= C_{q+1}\,\Delta_t^{q+1} y^{q+1}(t) + O(\Delta_t^{q+2}) \tag{3.2}$$

for any sufficiently differentiable function y(t). Expanding $L[y(t),\Delta_t]$ by Taylor's theorem with integral form of the remainder we have (cf.[13, pp 247])

$$L[y(t),\Delta_t] = \Delta_t^{q+1} \int_o^k G(s)y^{q+1}(t+s\Delta_t)ds$$

$$\leq G\Delta_t^{q+1} \sup_{t\leq s\leq t+k\Delta_t} \left(|y^{q+1}(s)|\right) \tag{3.3}$$

where G(s) is the kernel function and $G = \int_o^k G(s)ds$.

The concept of A_o-stability was introduced by Cryer [4]. A multistep method is A_o-stable if, applied to the equation $\dot{y} = \lambda y$, $y(0) = 1$, for any real $\lambda > 0$, it gives approximate values y_n of $y(n\Delta_t)$ such that $y_n \to 0$ as $n \to \infty$. Considering (3.1), this is equivalent to the roots of $p(\xi,\tau)$ being of modulus less than one for $\tau > 0$, where

$$p(\xi,\tau) = \rho(\xi) + \sum_{r=1}^{m} \tau^r \sigma_r(\xi),$$

$$\rho(\xi) = \sum_{j=o}^{k} \alpha_j \xi^j \text{ and } \sigma_r(\xi) = \sum_{j=o}^{k} \beta_{rj}(-1)^{r-1}\xi^j \quad r=1,2,\ldots,m. \tag{3.4}$$

In addition we require that the L.M.S.D. methods satisfy the conditions of zero-stability and consistency, ([16 pp.30]).

Zero-stability dictates that the roots of $\rho(\xi)$ with modulus equal to one are simple. The consistency condition is maintained by

$$\sum_{j=o}^{k} \alpha_j = 0 \quad \text{and} \quad \sum_{j=o}^{k} j\alpha_j = \sum_{j=o}^{k} \beta_{1j}.$$

We shall always assume that the characteristic polynomials $\rho(\xi)$ and $\{\sigma_r(\xi)\}_{r=1}^{m}$ have no common factor. Similarly, the polynomials in τ, $\{u_j(\tau)\}_{j=o}^{k}$, where

$$\mu_j(\tau) = \alpha_j + \sum_{r=1}^{m} (-1)^{r-1} \beta_{rj} \tau^r$$

shall have no common factor. These assumptions are compatible with the L.M.S.D. scheme being irreducible to an equivalent scheme with a lower value for k or m.

The following two results, although required in the later analysis, are of interest in themselves.

<u>Lemma 1</u> Let the L.M.S.D. scheme (3.1) be A_o-stable, then there exists a positive constant μ, such that

$$\mu_k(\tau) > \mu \,, \quad \text{for all } \tau \geq 0$$

Proof: Since $\alpha_k > 0$ by definition the expression $\mu_k(\tau)$ is not identically zero. Let us assume that $\mu_k(\tau)$ has a root at $\tau = \bar{\tau}$. The function

$$f(\xi,\tau) \equiv \frac{p(\xi,\tau)}{\mu_k(\tau)} = \sum_{j=o}^{k} \frac{\mu_j(\tau)}{\mu_k(\tau)} \xi^j$$

is well defined except at the zeros of $\mu_k(\tau)$. As $\tau \to \bar{\tau}$ at least one of the coefficients of $f(\xi,\tau)$ must become unbounded since

$\tau = \bar{\tau}$ may not be a root of all $\{\mu_j(\tau)\}_{j=o}^{k-1}$. Consequently, as $\tau \to \bar{\tau}$, at least one of the roots of $f(\xi,\tau)$, and hence of $p(\xi,\tau)$, must become unbounded and have modulus greater than one. This contradicts the assumption of A_o-stability and we deduce that $\mu_k(\tau)$ must be bounded away from zero, $\tau > 0$. Since $\mu_k(0)=\alpha_k > 0$ the proof is complete.

Lemma 2. Let the L.M.S.D. scheme (3.1) be A_o-stable, then

$$\beta_{mk} \neq 0$$

Proof: Trivially, if $\{\beta_{mj}\}_{j=o}^{k-1}$ are all zero then $\beta_{mk} \neq 0$ otherwise the scheme will incorporate only the first (m-1) derivatives. Let us assume that at least one $\beta_{ms} \neq 0$, $0 \le s \le k-1$, and further that $\beta_{mk} = 0$. Using the function $f(\xi,\tau)$ of lemma 1 it is obvious that the coefficient of ξ^s must become unbounded as $\tau \to \infty$. Once again (cf.lemma 1) this comprises a contradiction in the initial assumption of A_o-stability and we deduce that $\beta_{mk} \neq 0$.

Corollary. Every A_o-stable L.M.S.D. scheme (3.1) must be implicit.

Finally, we investigate the approximate solution of (2.2) by the L.M.S.D. method (3.1). Let us again denote U^n to be an approximant to $U(x,n\Delta_t)$. Assuming that $\{U^j\}_{j=o}^{k-1}$ are given, the recurrence relationship for U^{n+k}, $n \ge 0$, is given by the system of difference equations

$$\left(\sum_{j=o}^{k} \alpha_j U^{n+j}, V\right) - \left(\sum_{j=o}^{k}\sum_{r=1}^{m} \beta_{rj}\, \Delta_t^r\, U_{(r)}^{n+j}, V\right) = 0 \qquad (3.5)$$

$$\left(U_{(r)}^{n+j}, V\right) + a\left(U_{(r-1)}^{n+j}, V\right) = 0 \qquad r = 1,2,\ldots,m \qquad (3.6)$$

The computational aspects of (3.5) and (3.6) will be investigated in chapter 6. The implementation procedures described there are equivalent to the solution of the linear system of equations

$$A\underline{U}^{n+k} \equiv \left[\alpha_k I + \sum_{r=1}^{m} (-1)^{r-1} \beta_{rk} \Delta_t^r (M^{-1}K)^r\right] \underline{U}^{n+k} = \underline{\tilde{U}}$$

for some predetermined vector $\underline{\tilde{U}}$. The condition number of the matrix A where

$$\mathrm{Cond}(A) = \frac{\max \{\Lambda_{[a]}\}}{\min \{\Lambda_{[a]}\}} ,$$

and $\{\Lambda_{[a]}\}$ is the set of eigenvalues of A, is readily seen by lemma 1 and the analysis of chapter 2 to satisfy

$$\mathrm{Cond}(A) = 0\left(h^{-2m} \Delta^m_{\ t}\right) .$$

Hence, by lemma 1, the matrix A is positive definite and, if we exclude the unrealistic case when $\Delta_t h^{-2} \to 0$, the condition number of A does not grow too fast for small m.

4. Theorems

The analysis of chapter 5 will prove the following theorems.

Theorem 1

Let the L.M.S.D. method (3.1) of order q be consistent, zero-stable and A_o-stable. Let the roots of the polynomial $\rho(\xi)$ with modulus equal to one be real, the modulus of the roots of the polynomial $\sigma_m(\xi)$ be less than one, and $\sigma_1(-1) \neq 0$ if $\rho(-1) = 0$. Further, let $g(x) \in L_2(\Omega)$. Then for any $t_o > 0$ there exists a positive constant $C(t_o)$ such that for $n\Delta_t \geq t_o$, and h,Δ_t sufficiently small

$$\| u(x,n\Delta_t) - U^n \| \leq C(t_o) \left\{ (\Delta_t^q + h^{p+1}) \| g \| + \sum_{j=o}^{k-1} \| u(x,j\Delta_t) - U^j \| \right\}$$

and

$$\| U^n \| \leq C\, e^{-\alpha n \Delta_t \lambda_1} \sum_{j=o}^{k-1} \| U^j \|$$

Corollary

If in addition we assume that U^o is the projection of $g(x)$ onto V_h^p by the L_2-inner product and $\{U^j\}_{j=1}^{k-1}$ are the values derived from a weakly A_o-stable Padé scheme of order q-1, then

$$\| u(x,n\Delta_t) - U^n \| \leq C(t_o) \left\{ \Delta_t^q + h^{p+1} \right\} \| g \|$$

and

$$\| U^n \| \leq C\, e^{-\alpha n \Delta_t \lambda_1} \| g \|$$

Theorem 2

Let us further restrict $w = 1$ to be the only root of $\rho(\xi)$ with modulus equal to one, then with the assumptions of theorem 1

$$\| u(x,n\Delta_t)-U^n\| \leq C(t_o,\beta)e^{-\beta n\Delta_t\lambda_1} \left\{(\Delta_t^q + h^{p+1})\| g\| + \sum_{j=o}^{k-1} \|u(x,j\Delta_t)-U^j\|\right\}$$

for some arbitrary positive constant $\beta, 0 < \beta < 1$.

Corollary

If the initial values are defined to be exactly those described in the corollary to theorem 1, then

$$\| u(x,n\Delta_t)-U^n\| \leq C(t_o,\beta)\ e^{-\beta n\Delta_t\lambda_1} \left\{\Delta_t^q + h^{p+1}\right\} \| g\|$$

5. Proof of Theorems

Let $\{\lambda_i\}_{i=1}^{\infty}$ and $\{\psi_i\}_{i=1}^{\infty}$ be respectively the eigenvalues (in increasing order) and the corresponding orthonormal eigenfunctions of the continuous eigenvalue problem

$$a(\psi,v) = \lambda(\psi,v) \qquad \forall\ v \in H_0^1(\Omega) \tag{5.1}$$

The eigenvalues are well-known to be positive and distinct. Further let $\{\Lambda_i\}_{i=1}^{d}$ and $\{\Psi_i\}_{i=1}^{d}$ be the eigenvalues (in increasing order) and the corresponding orthonormal eigenfunctions of the discrete eigenvalue problem

$$a(\Psi,V) = \Lambda(\Psi,V) \qquad \forall\ V \in V_h^p \tag{5.2}$$

Strang and Fix [29, Theorem 6.1, 6.2] have proved results for eigenvalues and eigenfunctions using subspaces, S_h, on a regular mesh. The only property of S_h utilised in the proof is the approximation property

$$\| u - Pu\|_s \leq Ch^{k-s}\| u \|_k \qquad s = 0, \text{ or } 1$$

where Pu is the Ritz approximation of u (ie. $a(u - Pu,V) = 0, \forall V \in S_h$)
A well-known consequence of (2.7) is that

$$\| u - Pu\| + h\| u-Pu\|_1 \leq Ch^{p+1}\| u \|_{p+1} \,.$$

Hence, for k=p+1, all conditions are satisfied and the theorems yield for h sufficiently small

$$0 \leq \Lambda_i - \lambda_i \leq Ch^{2p}\ \lambda_i^{p+1} \qquad , \quad i=1,2,\ldots,d \tag{5.3}$$

$$\| \Psi_i - \psi_i \| \leq Ch^{p+1}\ \lambda_i^{\frac{1}{2}(p+1)} \qquad , \quad i=1,2,\ldots,d \tag{5.4}$$

We adopt the following notations

$$v_i = (v,\psi_i) \;,\; \bar{v}_i = (v,\Psi_i) \;,\; v \in H^1_o(\Omega)$$

$$V_i = (V,\Psi_i) \;,\; \bar{V}_i = (V,\psi_i) \;,\; V \in V^p_h \tag{5.5}$$

We bound the error $u(x,n\Delta_t) - U^n$ by using the relationship

$$u(x,n\Delta_t) - U^n = e_1 + e_2$$

where $e_1 = u(x,n\Delta_t) - U(x,n\Delta_t)$, $e_2 = U(x,n\Delta_t) - U^n$ and proving bounds on e_1 and e_2.

The solution $u(x,t)$ of (1.1) can be expressed as

$$u(x,t) = \sum_{i=1}^{\infty} g_i \, e^{-\lambda_i t} \psi_i \tag{5.6}$$

where $\{g_i\}_{i=1}^{\infty}$ are the Fourier coefficients of $g(x)$.

Similarly, the solution $U(x,t)$ of the continuous Galerkin problem (2.2) can be expressed by

$$U(x,t) = \sum_{i=1}^{d} U^o_i \, e^{-\Lambda_i t} \Psi_i \tag{5.7}$$

where $\{U^o_i\}_{i=1}^{d}$ are the coefficients of $\bar{g}(x) \in V^p_h$ with respect to the basis $\{\Psi_i\}_{i=1}^{d}$.

a) Let $U^n = \sum_{i=1}^{d} U^n_i \Psi_i$. Using (5.7) we can write $e_2 = \sum_{i=1}^{d} \varepsilon^n_i \Psi_i$ and hence $\| e_2 \|^2 = \sum_{i=1}^{d} |\varepsilon^n_i|^2$, where

$$\varepsilon^n_i = U^o_i \, e^{-\Lambda_i n\Delta_t} - U^n_i \tag{5.8}$$

Also, let $U^n_{(r)} \equiv \sum_{i=1}^{d} U^n_{i,r} \Psi_i$ be the discrete approximation to $\frac{\partial^r}{\partial t^r} U(x,t)\Big|_{t=n\Delta_t}$. Substituting $U^n_{(r)}$ into (3.6) and (5.2) with $V = \Psi_i$ gives us the relationship

$$U^n_{i,r} + \Lambda_i U^n_{i,r-1} = 0 \quad , \quad r=1,\ldots,m \quad \text{where } U^n_{i,o} \equiv U^n_i$$

Consequently, we can construct the recurrence relationship

$$U^n_{i,r} = (-1)^r \Lambda^r_i U^n_i \qquad r=1,\ldots,m \tag{5.9}$$

Combining (5.9) and (3.5) with $V = \Psi_i$ yields

$$\sum_{j=o}^{k} (\alpha_j + \sum_{r=1}^{m} (-1)^{r-1} \beta_{rj} \Delta^r_t \Lambda^r_i) U^{n+j}_i = 0 \tag{5.10}$$

Define $\delta_j(\tau) = \mu_j(\tau)/\mu_k(\tau)$ where $\mu_j(\tau) = \alpha_j + \sum_{r=1}^{m} (-1)^{r-1}\beta_{rj} \tau^r$

and subsequently rewrite (5.10) as

$$\sum_{j=o}^{k} \delta_j(\Delta_t \Lambda_i) U^{n+j}_i = 0 \tag{5.11}$$

The expressions (5.8) and (5.11) combine to give

$$\sum_{j=o}^{k} \delta_j(\Delta_t \Lambda_i) \varepsilon^{n+j}_i = \sum_{j=o}^{k} \delta_j(\Delta_t \Lambda_i) U^o_i e^{-\Lambda_i (n+j)\Delta_t} \equiv d^n_i \tag{5.12}$$

We conclude this sub-section by bounding d^n_i. We see from (3.2) and (3.3) that

$$\sum_{j=o}^{k} \mu_j(\Delta_t \Lambda_i) U^o_i e^{-\Lambda_i (n+j)\Delta_t} \equiv L\left[U^o_i e^{-\Lambda_i t}, \Delta_t\right]_{t=n\Delta_t}$$

$$\leq G\Delta^{q+1}_t \Lambda^{q+1}_i |U^o_i| e^{-n\Delta_t \Lambda_i}$$

By lemma 1 a positive supremum of $\left[\mu_k(\tau)\right]^{-1}$, $\tau > 0$, must exist from which we conclude that

$$d_i^n \le C\ \Delta_t^{q+1}\Lambda_i^{q+1}|U_i^o|e^{-n\Delta_t\Lambda_i} \tag{5.13}$$

Alternatively, by lemmas 1 and 2, $\delta_j(\tau) j=0,1,\dots,k$, are bounded for any $\tau > 0$, thus

$$d_i^n \le C|U_i^o|e^{-n\Delta_t\Lambda_i} \tag{5.14}$$

b) This section uses a method employed by Henrici [13,pp242] and adapted by Zlâmal [40-41]. Define $\hat{p}(\xi,\tau)$ by

$$\hat{p}(\xi,\tau) = \delta_k(\tau) + \delta_{k-1}(\tau)\xi + \ . \ . \ . + \delta_o(\tau)\xi^k$$

Note that $\hat{p}(\xi,\tau) = \left[\mu_k(\tau)\right]^{-1}\xi^k\ p(\frac{1}{\xi},\tau)$ and hence the roots of $\hat{p}(\xi,\tau)$ are the reciprocals of the roots of $p(\xi,\tau)$. It is intuitively obvious that the roots of $p(\xi,\tau)$ approach the roots of $\rho(\xi)$ and $\sigma_m(\xi)$ as, respectively, $\tau \to 0$ and $\tau \to \infty$.

The essential roots of $\rho(\xi)$ (i.e. those of modulus one) are by assumption real, and by zero-stability single. The consistency condition dictates that $w = 1$ is always an essential root. Let us assume the most general situation when these essential roots are $w_1=1$, $w_2=-1$. Any other root $\{w_i\}_{i=3}^k$ of $\rho(\xi)$ has modulus less than one, say $|w_i| \le 1-\theta$, $0<\theta\le 1$. We employ a theorem from complex analysis, eg. [1, Theorem 11,pp. 131], to show that for each sufficiently shall $\varepsilon > 0$, there exists a $\tau_\varepsilon > 0$, such that the equation $p(\xi,\tau) = 0$, $\tau < \tau_\varepsilon$, has the same number of roots in the disc $|\xi-\xi_o|<\varepsilon$ as the equation $\rho(\xi) = 0$. Furthermore, if ξ_o is a root of $\rho(\xi)$ of multiplicity p then the p roots of $p(\xi,\tau)$ that approach it are distinct for τ sufficiently small. Hence no complications arise from a root of multiplicity greater than one.

We denote $w_{1,2}$ to be correspondingly w_1 or w_2. Selecting $\varepsilon < \frac{\theta}{2}$ we have that, for $\tau < \tau_{\theta/2}$, the equation $p(\xi,\tau)$ has only one root in the disc $|\xi - w_{1,2}| < \frac{\theta}{2}$. Let this root be $\xi_{w_{1,2}}(\tau)$. Rearranging the above we deduce that for any $0 < \varepsilon < \theta/_2$ there exists a τ_ε such that $|\xi_{w_{1,2}}(\tau) - w_{1,2}| < \varepsilon$ whenever $\tau < \tau_\varepsilon$. This is a definition for $\xi_{w_{1,2}}(\tau)$ to tend continuously to $w_{1,2}$ as τ tends to zero. Thus $\xi_{w_{1,2}}(\tau)$ can be expressed as an analytic function of τ,

$$\text{i.e. } \xi_{w_i}(\tau) = w_i + a_1^i \tau + a_2^i \tau^2 + \dots \qquad i = 1,2$$

Corresponding expressions hold for the other roots $\{\xi_{w_i}(\tau)\}_{i=3}^{k}$ of $p(\xi,\tau)$. Remembering that $|w_i| < 1-\theta$, $i=3,4,\dots,k$, we deduce that for τ sufficiently small, say $\tau < \tau_1$, $|\xi_{w_i}(\tau)| < 1 - \frac{\theta}{2}$, $i=3\dots,k$.

Expanding $p(\xi_{w_{1,2}}(\tau),\tau)$ about the point $w_{1,2}$ we see that

$$p(\xi_{w_i}(\tau),\tau) = \rho(w_i) + \tau a_1^i \rho'(w_i) + \tau\sigma_1(w_i) + O(\tau^2) = 0 \quad , i=1,2$$

and by comparing coefficients

$$a_1^i = -\frac{\sigma_1(w_i)}{\rho'(w_i)} \qquad i = 1,2 \tag{5.15}$$

We know that $\sigma_1(1) = \rho'(1)$ by the consistency condition, $\sigma_1(-1) \neq 0$ by assumption and $\rho'(w_{1,2}) \neq 0$ by zero-stability.

Thus,

$$\xi_{w_i}(\tau) = w_i + a_1^i \tau + O(\tau^2) \text{ where } a_1^i \text{ is real and non-zero}$$

and

$$|\xi_{w_i}(\tau)| = |1 + \frac{a_1^i\tau}{w_i} + O(\tau^2)| \qquad i = 1,2.$$

But as $|\xi_{w_i}(\tau)| < 1$ for $\tau > 0$ we must have $a_1^i/w_i < 0$. Consequently, for τ sufficiently small, say $\tau < \tau_2$

$$|\xi_{w_i}(\tau)| < 1 - \hat{\alpha}\tau \text{ , } i = 1,2, \text{ for some } \hat{\alpha} \text{ , } \hat{\alpha} \geq \frac{1}{2}\min\left\{|a_1^1| = 1, |a_2^1|\right\} .$$

Thus, we have shown that for $\tau < \hat{\tau}$, $\hat{\tau} = \min(\tau_1,\tau_2)$

$$|\xi_{w_i}(\tau)| < 1 - \alpha\tau \text{ , } \alpha > 0, \qquad i = 1,2,\ldots, k$$

and hence, for $\tau < \hat{\tau}$, all roots $\hat{\xi}(\tau)$ of $\hat{p}(\xi,\tau)$ satisfy

$$|\hat{\xi}(\tau)| > \frac{1}{1-\alpha\tau} .$$

Therefore, $\frac{1}{\hat{p}(\xi,\tau)}$ is holomorphic for $|\xi| \leq \frac{1}{1-\alpha\tau}$, $\tau < \hat{\tau}$, and the function can be expressed by a Taylor series expansion

i.e. $$\frac{1}{\hat{p}(\xi,\tau)} = \gamma_o(\tau) + \gamma_1(\tau)\xi + \gamma_2(\tau)\xi^2 + \ldots. \qquad \tau < \hat{\tau}$$

where, by Cauchy's estimate, eg. [1. pp. 122]

$$|\gamma_\ell(\tau)| \leq C (1 - \alpha\tau)^\ell \qquad \ell = 0,1,\ldots \text{ whenever } \tau < \hat{\tau}.$$

Similarly, let the roots of $\sigma_m(\xi)$ be $\{z_i\}_{i=1}^k$. These roots are by assumption less than one in modulus, say $|z_i| \leq 1 - \theta$, $0 < \theta \leq 1$. Applying the aforementioned theorem we prove that the equation $p(\xi,\tau) = 0$ has the same number of roots in the disc $|\xi - z_i| < \frac{\theta}{2}$ as the equation $\sigma_m(\xi) = 0$, whenever $\tau > \overline{C}$. Repeating the above argument we have that, for $\tau > \overline{C}$, the roots $\xi_i(\tau)$ of $p(\xi,\tau)$ satisfy

$$|\xi_i(\tau)| < 1 - \frac{\theta}{2}, \quad i = 1,2,\ldots,k.$$

This leaves a finite interval $[\hat{\tau},\overline{C}]$ where the roots $\xi_i(\tau)$ of $p(\xi,\tau)$ are known to be of modulus less than one. These roots are continuous functions of τ over a finite interval, hence

$$|\xi_i(\tau)| < 1 - \hat{\theta} \quad , \quad 0 < \hat{\theta} < 1, \text{ whenever } \hat{\tau} \le \tau \le \overline{C}.$$

and we conclude that there exists a constant $\overline{\alpha}$, $0 < \overline{\alpha} < 1$, such that

$$|\xi_i(\tau)| < 1 - \overline{\alpha} \qquad \text{whenever } \tau \ge \hat{\tau}.$$

By the previous argument we easily establish

$$|\gamma_\ell(\tau)| \le C(1-\overline{\alpha})^\ell \qquad \text{whenever } \tau \ge \hat{\tau} \quad \ell = 0,1,\ldots$$

Summarising, we have proved that,

$$|\gamma_\ell(\tau)| \le \begin{cases} C(1-\alpha\tau)^\ell \le C\, e^{-\alpha\ell\tau} & \tau < \hat{\tau} \\ C(1-\bar{\alpha})^\ell \le C\, e^{-\bar{\alpha}\ell}, \quad 0<\overline{\alpha}<1 & \tau \ge \hat{\tau} \end{cases}$$

Making $\hat{\tau}$ smaller if necessary we achieve $\overline{\alpha} = \alpha\hat{\tau}$. Denoting by i_* the smallest integer such that $\Delta_t\Lambda_i > \hat{\tau}$ we have

$$|\gamma_\ell(\Delta_t\Lambda_i)| \le \begin{cases} C\, e^{-\alpha\ell\Delta_t\Lambda_i} & i < i_* \\ C\, e^{-\alpha\ell\hat{\tau}} & i \ge i_* \end{cases} \tag{5.16}$$

c) We now assume that $w_1 = 1$ is the only essential root of $\rho(\xi)$. The value a_1 of (5.15) is now equal to -1 by the consistency relationship. Thus for $\Delta_t\Lambda_i$ sufficiently small

$$\xi_{w_1}(\Delta_t\Lambda_i) = 1 - \Delta_t\Lambda_i + O(\Delta_t^2\Lambda_i^2) = e^{-\Delta_t\Lambda_i} + g$$

where g is an analytic function of $\Delta_t\Lambda_i$ and $g = O(\Delta_t^2\Lambda_i^2)$ at $\Delta_t\Lambda_i = 0$

Expanding $p(\xi_{w_1},\Delta_t\Lambda_i)$ about the point $e^{-\Delta_t\Lambda_i}$ and equating to zero we have by (3.4) that

$$p(\xi_{w_1}(\Delta_t\Lambda_i),\Delta_t\Lambda_i) = \rho(e^{-\Delta_t\Lambda_i}) + \sum_{r=1}^{m} (\Delta_t\Lambda_i)^r\, \sigma_r(e^{-\Delta_t\Lambda_i}) + g\, \rho'(e^{-\Delta_t\Lambda_i})$$

$$+ O(g^2) + O(\Delta_t\Lambda_i g) = 0.$$

By substituting $y(t) = e^{-\Lambda_i t}$ into (3.2) and letting $t = 0$ we deduce

$$\rho(e^{-\Delta_t\Lambda_i}) + \sum_{r=1}^{m} (\Delta_t\Lambda_i)^r\sigma_r(e^{-\Delta_t\Lambda_i}) = C_{q+1}\Delta_t^{q+1}(-\Lambda_i)^{q+1} + O\left[(\Delta_t\Lambda_i)^{q+2}\right].$$

Consequently, by combining the above expressions

$$g\rho'(e^{-\Delta_t\Lambda_i}) = -C_{q+1}(-\Delta_t\Lambda_i)^{q+1} + O(\Delta_t\Lambda_i g) + O\left[(\Delta_t\Lambda_i)^{q+2}\right] + O(g^2)$$

and thus, using $\rho'(e^{-\Delta_t\Lambda_i}) = \rho'(1) + O(\Delta_t\Lambda_i)$

$$g = \frac{(-1)^q}{\rho'(1)} C_{q+1}(\Delta_t\Lambda_i)^{q+1} + O\left[(\Delta_t\Lambda_i)^{q+2}\right] = C(\Delta_t\Lambda_i)^{q+1} + O\left[(\Delta_t\Lambda_i)^{q+2}\right].$$

With the above expression of g we have established the bound,

$$\xi_{w_1}(\Delta_t\Lambda_i) \le e^{-\Delta_t\Lambda_i}\left[1 + C(\Delta_t\Lambda_i)^{q+1}\right] < 1$$

whenever $\Delta_t\Lambda_i$ is sufficiently small. Utilising a previous result, we realise that the other roots $\{\xi_{w_j}\}_{j=2}^{k}$ of $p(\xi,\Delta_t\Lambda_i)$ satisfy $|\xi_{w_j}| < 1 - \frac{\theta}{2}$, given $\Delta_t\Lambda_i$ sufficiently small. Therefore, we can select a value $\hat{\tau} > 0$ such that, for $0 < \Delta_t\Lambda_i < \hat{\tau}$

$$|\xi_{w_j}(\Delta_t\Lambda_i)| \le e^{-\Delta_t\Lambda_i}\left[1 + c(\Delta_t\Lambda_i)^{q+1}\right] < 1 \qquad j=1,2,\ldots,k.$$

Extending the argument as before we easily achieve

$$|\gamma_\ell(\Delta_t\Lambda_i)| \le C\ e^{-\ell\Delta_t\Lambda_i}\left[1 + c(\Delta_t\Lambda_i)^{q+1}\right]^\ell \quad , \quad \Delta_t\Lambda_i < \hat{\tau}$$

Hence, for $\Delta_t\Lambda_i < \hat{\tau}$ and β , $0 < \beta < 1$

$$e^{-\ell\Delta_t\Lambda_i}\left[1 + c(\Delta_t\Lambda_i)^{q+1}\right]^\ell$$

$$\le e^{-\frac{1}{2}(1+\beta)\ell\Delta_t\Lambda_i}\left\{ e^{-\frac{1}{2}(1-\beta)\Delta_t\Lambda_i}\left[1 + c(\Delta_t\Lambda_i)^{q+1}\right]\right\}^\ell$$

and since $1 + cx^{q+1} \le e^{\frac{1}{2}(1-\beta)x}$ whenever $x < \tau_\beta \le \hat{\tau}$

we have

$$|\gamma_\ell(\Delta_t\Lambda_i)| \le Ce^{-\frac{1}{2}(1+\beta)\ell\Delta_t\Lambda_i} \quad , \quad \Delta_t\Lambda_i < \tau_\beta$$

For $\Delta_t\Lambda_i \ge \tau_\beta$ we recall from a previous result that

$$|\gamma_\ell(\Delta_t\Lambda_i)| \le C\ e^{-\overline{\alpha}\ell} \quad , \quad 0 < \overline{\alpha} < 1$$

Making τ_β smaller if necessary we achieve $\overline{\alpha} = \frac{1}{2}(1+\beta)\tau_\beta$.

Denoting by $i_*(\beta)$ the smallest integer such that $\Delta_t\Lambda_i > \tau_\beta$ we see that

$$|\gamma_\ell(\Delta_t\Lambda_i)| \begin{cases} Ce^{-\frac{1}{2}(1+\beta)\ell\Delta_t\Lambda_i} & , \quad i < i_*(\beta) \\ Ce^{-\frac{1}{2}(1+\beta)\tau_\beta\ell} & , \quad i \ge i_*(\beta) \end{cases}$$

for some β , $0 < \beta < 1$

By comparing coefficients in the expansion of

$$\frac{1}{\hat{p}(\xi,\tau)} = \frac{1}{\delta_k(\tau) + \xi\delta_{k-1}(\tau) + \ldots + \xi^k\ \delta_o(\tau)} = \gamma_o(\tau) + \xi\gamma_1(\tau) + \ \ldots .$$

we establish

$$\delta_k(\tau)\gamma_\ell(\tau) + \delta_{k-1}(\tau)\gamma_{\ell-1}(\tau)+\ldots+\delta_o(\tau)\gamma_{\ell-k}(\tau) = \begin{cases} 1 & \ell = 0 \\ 0 & \ell > 0 \end{cases} \tag{5.18}$$

where $\gamma_\ell = 0$ for $\ell < 0$.

d) Henceforth, the following inequalities will be used extensively:

$$x\, e^{-\alpha x} \le (e\alpha)^{-1} < (2\alpha)^{-1} \tag{5.19}$$

$$x^p\, e^{-\alpha x} < \left(\frac{2\alpha}{p}\right)^{-p}$$

for any $x \ge 0$, $\alpha > 0$ and p a positive integer.

If we rewrite (5.12) with $n \equiv n-k-\ell$, multiply this by $\gamma_\ell(\Delta_t\Lambda_i)$, sum for $\ell = 0,1,\ldots, n-k$ and then apply (5.18) we prove

$$\begin{aligned} \varepsilon_i^n = &-\left[\delta_{k-1}(\Delta_t\Delta_i)\gamma_{n-k}(\Delta_t\Lambda_i) +\ldots+ \delta_o(\Delta_t\Lambda_i)\, \gamma_{n-2k+1}(\Delta_t\Lambda_i)\right] \varepsilon_i^{k-1} \\ &-\left[\delta_{k-2}(\Delta_t\Delta_i)\gamma_{n-k}(\Delta_t\Lambda_i) + \ldots+\delta_o(\Delta_t\Delta_i)\gamma_{n-2k+2}(\Delta_t\Lambda_i)\right] \varepsilon_i^{k-2} \\ &- \ldots - \delta_o(\Delta_t\Lambda_i)\gamma_{n-k}(\Delta_t\Lambda_i)\varepsilon_i^o + \sum_{\ell=o}^{n-k} d_i^{n-k-\ell}\gamma_\ell(\Delta_t\Lambda_i) \end{aligned} \tag{5.20}$$

Using (5.13), (5.16), (5.20) and the inequalities (5.19) a bound on $|\varepsilon_i^n|$ can be constructed as follows: for $i < i_*$

$$\begin{aligned} |\varepsilon_i^n| \le{}& Ce^{-\alpha(n-2k+1)\Delta_t\Lambda_i} \sum_{j=1}^{k-1} |\varepsilon_i^j| \\ &+ C\Delta_t^{q+1} \sum_{\ell=o}^{n-k} \Lambda_i^{q+1} |U_i^o|\, e^{-(n-k-\ell)\Delta_t\Lambda_i}\, e^{-\alpha\ell\Delta_t\Lambda_i} \end{aligned} \tag{5.21}$$

Note that for $n\Delta_t \ge t_o$ and $(2k-1)\Delta_t \le \frac{1}{2}t_o$

$$e^{-\alpha(n-2k+1)\Delta_t\Lambda_i} \le e^{-\frac{1}{2}\alpha t_o\Lambda_i} \le C(t_o)\Lambda_i^{-s} \tag{5.22}$$

where s will be determined later. For $\alpha - 1 \geq 0$

$$\Delta_t \, e^{-(n-k)\Delta_t\Lambda_i} \sum_{\ell=o}^{n-k} e^{-(\alpha-1)\ell\Delta_t\Lambda_i} \leq (n-k+1) \, \Delta_t \, e^{-(n-k)\Delta_t\Lambda_i}$$

$$\leq 2 \, (n-k)\Delta_t \, e^{-(n-k)\Delta_t\Lambda_i}$$

$$\leq \frac{C}{\Lambda_i} \, e^{-\frac{1}{2}(n-k)\Delta_t\Lambda_i} \leq \frac{C}{\Lambda_i} \, e^{-\frac{1}{4}t_o\Lambda_i}$$

$$\leq C(t_o)\Lambda_i^{-(q+1)}$$

For $\alpha - 1 < 0$

$$S \equiv \sum_{\ell=o}^{n-k} e^{-(\alpha-1)\ell\Delta_t\Lambda_i} \leq \frac{e^{(1-\alpha)(n-k+1)\Delta_t\Lambda_i}}{e^{(1-\alpha)\Delta_t\Lambda_i} - 1}$$

Hence,

$$S\Delta_t \, e^{-(n-k)\Delta_t\Lambda_i} \leq \frac{C\Delta_t \, e^{-\alpha(n-k)\Delta_t\Lambda_i}}{e^{(1-\alpha)\Delta_t\Lambda_i} - 1} \leq \frac{C \, e^{-\alpha(n-k)\Delta_t\Lambda_i}}{(1 - \alpha)\Lambda_i}$$

$$\leq \frac{Ce^{-\frac{1}{2}\alpha t_o\Lambda_i}}{\Lambda_i} \leq C(t_o)\Lambda_i^{-(q+1)}$$

Thus, we have shown that

$$\Delta_t \, e^{-(n-k)\Delta_t\Lambda_i} \sum_{\ell=o}^{n-k} e^{-(\alpha-1)\ell\Delta_t\Lambda_i} \leq C(t_o)\Lambda_i^{-(q+1)} \tag{5.23}$$

Collecting together (5.21)-(5.23), we conclude that whenever

$i < i_*$,

$$|\varepsilon_i^n| \leq C(t_o)\Lambda_i^{-s} \sum_{j=1}^{k-1} |\varepsilon_i^j| + C(t_o)\Delta_t^q|U_o^i| \tag{5.24}$$

For $i \geq i_*$, using (5.14) , (5.16) and (5.20)

$$|\varepsilon_i^n| \le C\, e^{-\alpha\hat{\tau}(n-2k+1)} \sum_{j=1}^{k-1} |\varepsilon_i^j| + C|U_i^o| \sum_{\ell=o}^{n-k} e^{-\Lambda_i(n-k-\ell)\Delta_t}\, e^{-\alpha\hat{\tau}\ell} \qquad (5.25)$$

But $$e^{-\alpha\hat{\tau}(n-2k+1)} < C\, e^{-\alpha\hat{\tau}n} \le Cn^{-q} \le C(t_o)\, \Delta_t^q \qquad (5.26)$$

as $n\, \Delta_t \ge t_o$. Also,

$$\bar{S} \equiv \sum_{\ell=o}^{n-k} e^{-\Lambda_i(n-k-\ell)\Delta_t}\, e^{-\alpha\hat{\tau}\ell} \le \sum_{\ell=o}^{n-k} e^{-\hat{\tau}(n-k-\ell+\alpha\ell)}$$

$$\le e^{-\hat{\tau}(n-k)} \sum_{\ell=o}^{n-k} e^{-\hat{\tau}(\alpha-1)\ell}$$

For $\alpha - 1 \ge 0$

$$\bar{S} \le (n-k+1)\, e^{-\hat{\tau}(n-k)} \le 2(n-k)\, e^{-\hat{\tau}(n-k)} \le C\, e^{-\frac{1}{2}\hat{\tau}(n-k)}$$

$$\le C\, e^{-\frac{1}{2}\hat{\tau}n} \le Cn^{-q} \le C(t_o)\, \Delta_t^q$$

Similarly, for $\alpha - 1 < 0$

$$\bar{S} \le \frac{e^{-\hat{\tau}(n-k)}\, e^{\hat{\tau}(1-\alpha)(n-k+1)}}{e^{\hat{\tau}(1-\alpha)} - 1} \le \frac{C\, e^{-\hat{\tau}\alpha(n-k+1)}}{\hat{\tau}(1-\alpha)} \le C\, e^{-\hat{\tau}\alpha n} \le C(t_o)\Delta_t^q.$$

Combining we have proved that

$$\sum_{\ell=o}^{n-k} e^{-\Lambda_i(n-k-\ell)\Delta_t}\, e^{-\alpha\hat{\tau}\ell} \le C(t_o)\Delta_t^q \qquad (5.27)$$

and the expressions (5.25) - (5.27) yield, for $i \ge i_*$

$$|\varepsilon_i^n| \le C(t_o)\Delta_t^q \left\{ \sum_{j=1}^{k-1} |\varepsilon_i^j| + |U_i^o| \right\} \qquad (5.28)$$

From the bounds (5.24) and (5.28) we achieve

$$\sum_{i=1}^{d} |\varepsilon_i^n|^2 \le C(t_o) \left\{ \Delta_t^{2q} \sum_{i=1}^{d} |U_i^o|^2 + \sum_{i<i_*} \Lambda_i^{-2s} \sum_{j=1}^{k-1} |\varepsilon_i^j|^2 + \Delta_t^{2q} \sum_{i\ge i_*} \sum_{j=1}^{k-1} |\varepsilon_i^j|^2 \right\}$$

Using $|\varepsilon_i^j| \le |U_i^o| + |U_i^j|$ we prove that

$$\| e_2 \|^2 = \sum_{i=1}^{d} |\varepsilon_i^n|^2 \le C(t_o) \left\{ \Delta_t^{2q} \sum_{j=o}^{k-1} \| U^j \|^2 + \sum_{i<i_*} \Lambda_i^{-2s} \sum_{j=1}^{k-1} |\varepsilon_i^j|^2 \right\} \quad (5.29)$$

Mihlin [22] has proved that $\Lambda_i \ge \lambda_i \ge c i^{\frac{2}{N}}$, c a positive constant. Thus for any $s > N$

$$\sum_{i=1}^{d} \Lambda_i^{-s} \le \sum_{i=1}^{\infty} \lambda_i^{-s} \le C.$$

We use this result frequently in the following analysis. Let $e_3 = \sum_{i<i_*} \sum_{j=1}^{k-1} \Lambda_i^{-2s} |\varepsilon_i^j|^2$. We can write ε_i^j as

$$\varepsilon_i^j = U_i^o \, e^{-j\Delta t \Lambda i} - U_i^j$$

$$= e^{-j\Delta t \Lambda i}(U_i^o - \overline{U}_i^o) + e^{-j\Delta t \Lambda i}(\overline{U}_i^o - u_i^o) + (e^{-j\Delta t \Lambda i} - e^{-j\Delta t \lambda i})\, u_i^o +$$

$$(u_i^j - \overline{U}_i^j) + (\overline{U}_i^j - U_i^j)$$

from which

$$|e_3| \le C \sum_{j=o}^{k-1} \sum_{i<i_*} \Lambda_i^{-2s} |u_i^j - \overline{U}_i^j|^2 + C \sum_{j=o}^{k-1} \sum_{i<i_*} \Lambda_i^{-2s} |\overline{U}_i^j - U_i^j|^2$$

$$+ C \sum_{j=1}^{k-1} \sum_{i<i_*} \Lambda_i^{-2s} \, |e^{-j\Delta t \Lambda i} - e^{-j\Delta t \lambda i}|^2 |u_i^o|^2 \quad (5.30)$$

The expression (5.30) can be investigated by using (5.3) and (5.4), whence,

$$e_4^j \equiv \sum_{i<i_*} \Lambda_i^{-2s} |u_i^j - \overline{U}_i^j|^2 \le \lambda_1^{-2s} \sum_{i=1}^{\infty} |u_i^j - \overline{U}_i^j|^2 \le c \| u^j - U^j \|^2$$

$e_5^j \equiv \sum_{i<i_*} \Lambda_i^{-2s} |\overline{U}_i^j - U_i^j|^2$. Now, $U_i^j - \overline{U}_i^j = \int_\Omega U^j(\Psi_i - \psi_i)\, dx$

i.e. $|\overline{U}_i^j - U_i^j|^2 \le C \| U^j \|^2 h^{2(p+1)} \lambda_i^{p+1}$, but as $\lambda_i \ge c i^{\frac{2}{N}}$ the series $\sum_{i=1}^{\infty} \lambda_i^{-(2s-p-1)}$ is convergent if we select 2s=p + 1 + N. Thus

$$e_5^j \le C \| U^j \|^2 h^{2(p+1)}$$

$$e_6^j \equiv \sum_{i<i_*} |e^{-j\Delta_t\Lambda_i} - e^{-j\Delta_t\lambda_i}|^2 \Lambda_i^{-2s} |u_i^o|^2$$

$$\le \sum_{i<i_*} \Lambda_i^{-2s} |j\Delta_t\Lambda_i - j\Delta_t\lambda_i|^2 |u_i^o|^2 \le j^2 \Delta_t^2 h^{4p} \sum_{i=1}^{\infty} \lambda_i^{-2(s-p+1)} |u_i^o|^2$$

and selecting $s = p+1 + \frac{N}{2}$ we have

$$e_6^j \le C\Delta_t^2 h^{4p} \| g \|^2 .$$

Substituting the above bounds in (5.30) we establish a bound on $|e_3|$, namely

$$|e_3| \le C \left\{ \sum_{j=o}^{k-1} \| u^j - U^j \|^2 + h^{2(p+1)} \sum_{j=o}^{k-1} \| U^j \|^2 + \Delta_t^2 h^{4p} \| g \|^2 \right\} \tag{5.31}$$

The desired result is obtained by substituting (5.31) into (5.29) and using the inequality

$$\| U^j \| \le \| U^j - u^j \| + \| u^j \| \le \| U^j - u^j \| + \| g \|$$

i.e. $$\| e_2 \| \le C(t_o) \left\{ \sum_{j=o}^{k-1} \| U^j - u^j \| + (h^{p+1} + \Delta_t^q) \| g \| \right\} \tag{5.32}$$

e) We now extend the analysis of section d to the situation when w = 1 is the only essential root of $\rho(\xi)$. Using (5.13), (5.17), (5.20) and the inequalities (5.19) a bound on ε_i^n is constructed as follows; for $i<i_*(\beta)$

$$|\varepsilon_i^n| \le C\, e^{-\frac{1}{2}(1+\beta)(n-2k+1)\Delta_t\Lambda_i} \sum_{j=1}^{k-1} |\varepsilon_i^j|$$

$$+ C\, \Delta_t^{q+1} \sum_{\ell=o}^{n-k} \Lambda_i^{q+1}|U_i^o| e^{-(n-k-\ell)\Delta_t\Lambda_i}\, e^{-\frac{1}{2}(1+\beta)\ell\Delta_t\Lambda_i} \tag{5.33}$$

Now for $n\, \Delta_t \ge t_o$ and $(2k-1)\, \Delta_t \le \frac{1}{2}t_o$

$$e^{-\frac{1}{2}(1+\beta)(n-2k+1)\Delta_t\Lambda_i} \le C\, e^{-\beta n\Delta_t\lambda_1} e^{-\frac{1}{4}(1-\beta)t_o\Lambda_i}$$

$$\le C(t_o,\beta) e^{-\beta n\Delta_t\lambda_1}\, \Lambda_i^{-s} \tag{5.34}$$

where as before, s will be determined later. Also, define

$$\tilde{S} \equiv \Delta_t\, e^{-(n-k)\Delta_t\Lambda_i} \sum_{\ell=o}^{n-k} e^{\frac{1}{2}(1-\beta)\ell\Delta_t\Lambda_i} \text{ and hence}$$

$$\tilde{S} \le \Delta_t\, e^{-(n-k)\Delta_t\Lambda_i}\, \frac{e^{\frac{1}{2}(1-\beta)(n-k+1)\Delta_t\Lambda_i}}{e^{\frac{1}{2}(1-\beta)\Delta_t\Lambda_i} - 1}$$

$$\le \frac{C\, e^{-\frac{1}{2}(1+\beta)(n-k)\Delta_t\Lambda_i}}{(1-\beta)\Lambda_i} \le \frac{C(\beta)}{\Lambda_i}\, e^{-\beta n\Delta_t\Lambda_i}\, e^{-\frac{1}{2}(1-\beta)(n-k)\Delta_t\Lambda_i}$$

$$\le \frac{C(\beta)}{\Lambda_i} e^{-\beta n\Delta_t\lambda_1}\, e^{-\frac{1}{4}(1-\beta)t_o\Lambda_i} \le C(t_o,\beta) e^{-\beta n\Delta_t\lambda_1}\, \Lambda_i^{-(q+1)} \tag{5.35}$$

From (5.33) - (5.35) we have whenever $i<i_*(\beta)$

$$|\varepsilon_i^n| \le C(t_o,\beta)\, e^{-\beta n\Delta_t\lambda_1} \left\{ \Lambda_i^{-s} \sum_{j=1}^{k-1} |\varepsilon_i^j| + \Delta_t^q\, |U_i^o| \right\} \tag{5.36}$$

Similarly, for $i \ge i_*(\beta)$, using (5.14), (5.18) and (5.20) we have

$$|\varepsilon_i^n| \le C\, e^{-\frac{1}{2}(1+\beta)\tau(n-2k+1)} \sum_{j=1}^{k-1} |\varepsilon_i^j| \tag{5.37}$$

$$+ C|U_i^o| \sum_{\ell=o}^{n-k} e^{-\Lambda_i(n-k-\ell)\Delta_t}\, e^{-\frac{1}{2}(1+\beta)\tau\ell}$$

where, for simplicity, we denote $\tau \equiv \tau_\beta$. But

$$e^{-\frac{1}{2}(1+\beta)\tau(n-2k+1)} \le C\, e^{-\beta n\tau}\, e^{-\frac{1}{2}(1-\beta)\tau n} \le C(\beta) e^{-\beta n\tau}\, n^{-q}$$

$$\le C(t_o,\beta)\, e^{-\beta n\tau}\, \Delta_t^q \tag{5.38}$$

Also, let $\hat{S} \equiv \sum_{\ell=o}^{n-k} e^{-\Lambda_i(n-k-\ell)\Delta_t}\; e^{-\frac{1}{2}(1+\beta)\tau\ell}$ and thus

$$\hat{S} \le e^{-\tau(n-k)} \sum_{\ell=o}^{n-k} e^{\frac{1}{2}\tau(1-\beta)\ell} \le e^{-\tau(n-k)}\; \frac{e^{\frac{1}{2}\tau(1-\beta)(n-k+1)}}{e^{\frac{1}{2}\tau(1-\beta)} - 1}$$

$$\le \frac{Ce^{-\frac{1}{2}(1+\beta)\tau n}}{\tau(1-\beta)} \qquad \le C(t_o,\beta)\, e^{-\beta n\tau}\, \Delta_t^q \tag{5.39}$$

The expressions (5.37) - (5.39) yield that, for $i \ge i_*(\beta)$

$$|\varepsilon_i^n| \le C(t_o,\beta)\Delta_t^q\, e^{-\beta n\Delta_t\lambda_1} \left\{ \sum_{j=1}^{k-1} |\varepsilon_i^j| + |U_i^o| \right\} \tag{5.40}$$

where we take Δ_t sufficiently small to allow $\Delta_t\lambda_1 < \tau$.

Following a course identical to section d we arrive at the result

$$\| e_2 \| \leq C(t_o,\beta) e^{-\beta n \Delta_t \lambda_1} \left\{ \sum_{j=o}^{k-1} \| U^j - u^j \| + \left(h^{p+1} + \Delta_t^q \right) \| g \| \right\} \qquad (5.41)$$

f) The error $e_1 = u(x,n\Delta_t) - U(x,n\Delta_t)$ will now be bounded. From (5.6) and (5.7) we have

$$e_1 = \sum_{i=1}^{\infty} g_i \, e^{-n\Delta_t \lambda_i} \psi_i - \sum_{i=1}^{d} U_i^o \, e^{-n\Delta_t \Lambda_i} \Psi_i$$

$$= \sum_{i>d} g_i \, e^{-n\Delta_t \lambda_i} \psi_i + \sum_{i=1}^{d} \left(e^{-n\Delta_t \lambda_i} - e^{-n\Delta_t \Lambda_i} \right) g_i \psi_i$$

$$+ \sum_{i=1}^{d} e^{-n\Delta_t \Lambda_i} (g_i - \bar{g}_i) \Psi_i + \sum_{i=1}^{d} e^{-n\Delta_t \Lambda_i} \bar{g}_i (\psi_i - \Psi_i)$$

$$+ \sum_{i=1}^{d} e^{-n\Delta_t \Lambda_i} (\bar{g}_i - U_i^o) \Psi_i \qquad (5.42)$$

Zlámal [41] uses a technique from Thomée [30] to show that $\lambda_{d+1} \geq ch^{-2}$. Hence, using (5.3) and (5.4) we have for some β, $0 \leq \beta < 1$.

$$e_7 \equiv \sum_{i>d} e^{-n\Delta_t \lambda_i} g_i \psi_i \leq e^{-\beta n \Delta_t \lambda_1} \sum_{i>d} e^{-(1-\beta) n \Delta_t \lambda_{d+1}} g_i \psi_i$$

$$\leq e^{-\beta n \Delta_t \lambda_1} e^{(1-\beta) t_o \lambda_{d+1}} \sum_{i=1}^{\infty} g_i \psi_i \leq C(t_o,\beta) \, e^{-\beta n \Delta_t \lambda_1} \lambda_{d+1}^{-\frac{1}{2}(p+1)} \sum_{i=1}^{\infty} g_i \psi_i$$

i.e. $\| e_7 \| \leq C(t_o,\beta) h^{p+1} e^{-\beta n \Delta_t \lambda_1} \| g \|$.

Let $e_8 \equiv \sum_{i=1}^{d} (e^{-n\Delta_t \lambda_i} - e^{-n\Delta_t \Lambda_i}) \, g_i \, \psi_i$. By the mean-value theorem

$$e_8 \le \sum_{i=1}^{d} n\Delta_t|\lambda_i - \Lambda_i|\; e^{-n\Delta_t\lambda_i}\; g_i\;\psi_i$$

$$\le Cn\,\Delta_t\, h^{2p}\; e^{-\beta n\Delta_t\lambda_i} \sum_{i=1}^{d} \lambda_i^{p+1}\; e^{-(1-\beta)n\Delta_t\lambda_i}\; g_i\psi_i$$

$$\le C(\beta)h^{2p}\; e^{-\beta n\Delta_t\lambda_1} \sum_{i=1}^{d} \lambda_i^{p}\; e^{-\frac{1}{2}(1-\beta)t_o\lambda_i}\; g_i\;\psi_i$$

$$\le C(t_o,\beta)h^{2p}\; e^{-\beta n\Delta_t\lambda_1} \sum_{i=1}^{\infty} g_i\;\psi_i$$

i.e. $\|e_8\| \le C(t_o,\beta)h^{2p}\; e^{-\beta n\Delta_t\lambda_1}\; \|g\|$

Let $e_9 \equiv \sum_{i=1}^{d} e^{-n\Delta_t\Lambda_i}(g_i - \bar{g}_i)\;\psi_i$. However $g_i - \bar{g}_i = \int_\Omega g(x)(\psi_i - \Psi_i)\,dx$

and thus by $n\Delta_t \ge t_o$, the Cauchy-Schwartz inequality and (5.19)

$$\|e_9\| \le C\,\|g\|\, h^{p+1}\; e^{-\beta n\Delta_t\lambda_1} \sum_{i=1}^{d} e^{-(1-\beta)n\Delta_t\lambda_i}\; \lambda_i^{\frac{1}{2}(p+1)}$$

$$\le C(t_o,\beta)\,\|g\|\, h^{p+1}\; e^{-\beta n\Delta_t\lambda_1} \sum_{i=1}^{\infty} \lambda_i^{-N}$$

$$\le C(t_o,\beta)\,\|g\|\, h^{p+1}\; e^{-\beta n\Delta_t\lambda_1}$$

Let $e_{10} = \sum_{i=1}^{d} e^{-n\Delta_t\Lambda_i}\;\bar{g}_i\;(\psi_i - \Psi_i)$. Thus

$$\|e_{10}\| \le Ch^{p+1}\; e^{-\beta n\Delta_t\lambda_1}\; \|g\| \sum_{i=1}^{d} e^{-(1-\beta)n\Delta_t\lambda_i}\; \lambda_i^{\frac{1}{2}(p+1)}$$

$$\le C(t_o,\beta)h^{p+1}\, e^{-\beta n\Delta_t\lambda_1}\|g\| \sum_{i=1}^{\infty} \lambda_i^{-N} \le C(t_o,\beta)h^{p+1}\, e^{-\beta n\Delta_t\lambda_1}\|g\|$$

Finally, $e_{11} \equiv \sum_{i=1}^{d} e^{-n\Delta_t\Lambda_i}\,(\bar{g}_i - U_i^o)\Psi_i$. If U^o is the orthogonal projection of $g(x)$ onto V_h^p with respect to the L_2-inner product then $e_{11} = 0$, otherwise

$$e_{11} = \sum_{i=1}^{d} e^{-n\Delta_t\Lambda_i} \left[(\bar{g}_i - g_i) + (g_i - U_i^o)\right] \Psi_i \quad \text{and hence}$$

$$\|e_{11}\| \le C(t_o,\beta)\, e^{-\beta n\Delta_t\lambda_1} \left\{\|g\|\, h^{p+1} + \|g - U^o\|\right\} \qquad \text{(cf.e9)}$$

Using (5.42) and the above bounds we conclude that

$$\|e_1\| \le \begin{cases} C(t_o,\beta)\, e^{-\beta n\Delta_t\lambda_1}\left(h^{p+1}\|g\| + \|g - U^o\|\right) \\ C(t_o,\beta)\, e^{-\beta n\Delta_t\lambda_1} h^{p+1}\|g\|. \text{ If } U^o \text{ is the } L_2\text{-inner product projection of } g(x) \text{ onto } V_h^p. \end{cases} \qquad (5.43)$$

for some arbitrary β, $0 \le \beta < 1$.

g) Returning to (5.11) we have

$$\sum_{j=o}^{k} \delta_j(\Delta_t\Lambda_i)U_i^{n+j} = 0.$$

Rewrite the above with $n = n - k - \ell$, multiply this by $\gamma_\ell(\Delta_t\Lambda_i)$, sum for $\ell = 0,1,\ldots,n-k$ and apply (5.18) to achieve the expression (5.20) with ϵ_i^j replaced by U_i^j, and $d_i \equiv 0$. Let us assume that Δ_t is sufficiently small so that $\Delta_t\lambda_1 < \hat{\tau}$. Using the remodelled expression of (5.20) and (5.16) we obtain

$$|U_i^n| \le C\, e^{-\alpha n\Delta_t\lambda_1} \sum_{j=o}^{k-1} |U_i^j|$$

from which it follows that

$$||U^n|| = \left(\sum_{i=1}^{d} |U_i^n|^2 \right)^{\frac{1}{2}} \leq C\, e^{-\alpha n\Delta_t\lambda_1} \sum_{j=o}^{k-1} ||U^j|| \tag{5.44}$$

which is the desired asymptotic result.

h) Initial Approximants U^j to $u(x,j\Delta_t)$, $j=0,1,\ldots, k-1$.

This section is concerned with the estimate $||e_2||$ under the assumption that U^o is the orthogonal projection of $g(x)$ onto V_h^P with respect to the L_2-inner product and $\{U^j\}_{j=1}^{k-1}$ are the approximate solutions of (2.2) at time $t = j\Delta_t$ obtained by a weakly A_o-stable Padé scheme of order $q - 1$.

Other viable methods for deriving these approximants include the weakly A_o-stable Runge-Kutta schemes. Such schemes have been thoroughly investigated by Crouzeix [3] and we refer the reader to his thesis for an account of these schemes.

A difference method derived from a Padé approximation of order $q - 1$ is a one-step method of the type

$$y_{n+1} - y_n = \sum_{s=o}^{1} \sum_{r=1}^{\tilde{m}} \tilde{\beta}_{rs}\Delta_t^r y_{n+s}^r \tag{5.45}$$

where

$$R(\tau) \equiv \frac{1 + \sum_{r=1}^{\tilde{m}} (-1)^r \tilde{\beta}_{ro}\, \tau^r}{1 + \sum_{r=1}^{\tilde{m}} (-1)^{r-1} \tilde{\beta}_{r1}\tau^r}$$ is an approximation to

$e^{-\tau}$, such that

$$|e^{-\tau} - R(\tau)| \leq C\,\tau^{q} \text{ as } \tau \to 0 \tag{5.46}$$

We note that any Padé scheme is a one-step, multiderivative method and satisfies (see (3.2)) the relation

$$y_{n+1} - y_{n} - \sum_{s=o}^{1} \sum_{r=1}^{\tilde{m}} \tilde{\beta}_{rs}\, \Delta_t^{r}\, y_{n+s}^{r} = \tilde{C}_q \Delta_t^{q}\, y^{q}(n\Delta_t) + O(\Delta_t^{q+1})$$

$$\leq \tilde{G}\, \Delta_t^{q} \sup_{o\leq s\leq 1} \left\{ |y^{q}(n+s)\Delta_t)| \right\} \tag{5.47}$$

A Padé scheme is said to be weakly A_o-stable (see [3]) if $|R(\tau)| \leq 1$, for any $\tau \geq 0$. The inequality (5.46) is stated to hold for small τ. However, as $|e^{-\tau} - R(\tau)| \leq 2\ \forall\tau \geq 0$,(5.46) is satisfied a fortiori for any $\tau \geq 0$. Applying the scheme (5.45) to the system of differential equations (2.2) we see immediately from an obvious adaptation of (5.10) that

$$\left(1 + \sum_{r=1}^{\tilde{m}} \Delta_t^{r} \tilde{\beta}_{r1} (-1)^{r-1}\, \Lambda_i^{r}\right) U_i^{j+1} - \left(1 + \sum_{r=1}^{\tilde{m}} \Delta_t^{r} \tilde{\beta}_{ro} (-1)^{r}\, \Lambda_i^{r}\right) U_i^{j} = 0$$

$$\text{or } U_i^{j+1} = R(\Delta_t \Lambda_i) U_i^{j} \quad , \quad j = 0,1,\ldots,k-2. \tag{5.48}$$

The recurrence equation (5.48) yields

$$U_i^{j+1} = \left[R(\Delta_t \Lambda_i)\right]^{j+1} \quad U_i^{o} \tag{5.49}$$

It is easily derived from (5.8) and (5.49) that

$$\varepsilon_i^{j+1} = U_i^{o} \left(e^{-\Lambda_i (j+1)\Delta_t} - \left[R(\Delta_t \Lambda_i)\right]^{j+1} \right)$$

and by using the definition of weak A_o-stability, and (5.46)

$$\left|\varepsilon_i^{j+1}\right| \leq \left|U_i^o\right| \left|e^{-\Lambda_i(j+1)\Delta_t} - \left[R\left(\Delta_t\Lambda_i\right)\right]^{j+1}\right|$$

$$\leq (j+1)\left|U_i^o\right| \left|e^{-\Lambda_i\Delta_t} - R(\Delta_t\Lambda_i)\right| \leq C\left|U_i^o\right|\Delta_t^q \Lambda_i^q$$

$$j = 0,1,\ldots,k-2 \qquad (5.50)$$

Consequently, returning to (5.29) we note

$$\sum_{j=1}^{k-1} \sum_{i<i_*} \Lambda_i^{-2s}\left|\varepsilon_i^j\right|^2 \leq C\,\Delta_t^{2q} \sum_{i<i_*} \Lambda_i^{-2(s-q)}\left|U_i^o\right|^2$$

$$\leq C\,\Delta_t^{2q}\,\|U^o\|^2 \text{ by selecting } s = q+\frac{N}{2} \qquad (5.51)$$

The initial approximant U^o to $g(x)$ is defined to be the projection of $g(x)$ onto V_h^p by the L_2-inner product, and is thus well known to satisfy,

$$\|U^o\| \leq \|g\|$$

Using the definition of weak A_o-stability, namely $|R(\tau)| \leq 1$, for $\tau \geq 0$ we have by (5.49)

$$\|U^j\|^2 = \sum_{i=1}^{d} |U_i^j|^2 \leq \sum_{i=1}^{d} |U_i^o|^2 = \|U^o\|^2 \leq \|g\|^2 \qquad j=1,2,\ldots,k-1 \qquad (5.52)$$

The expression (5.29) can now be reformulated by (5.51) and (5.52) to read

$$\|e_2\| \leq C(t_o)\,\Delta_t^q\,\|g\| \qquad (5.53)$$

We are able to deduce immediately the corresponding result when $w = 1$ is the only essential root of $\rho(\xi)$

$$\text{i.e.} \quad \| e_2 \| \leq C(t_o,\beta)\ e^{-\beta n\Delta_t\lambda_1}\ \Delta_t^q\ \| g \| \tag{5.54}$$

The theorems can now be established. Theorem 1 is determined from the relation $\| u(x,n\Delta_t) - U^n \| \leq \| e_1 \| + \| e_2 \|$ and the bounds (5.32), (5.43) with $\beta = 0$, and (5.44). Its corollary follows immediately by using (5.53) instead of (5.32). Theorem 2 and its corollary follow from the bounds (5.41), (5.43), and (5.54).

6. The Nonhomogeneous Linear Parabolic Equation

In this chapter we discuss the numerical solution of the nonhomogeneous equation

$$\frac{\partial u}{\partial t} = \sum_{i,j=1}^{N} \frac{\partial}{\partial x_i} \left(a_{i,j}(x) \frac{\partial u}{\partial x_j}\right) - a(x)u + f(x,t) \quad , \quad x \in \Omega, \ t > 0$$

$$\equiv \ Lu \ + \ f(x,t) \qquad (6.1a)$$

$$u(x,o) = g(x) \qquad x \in \Omega, \qquad (6.1b)$$

$$u(x,t) = 0 \qquad x \in \Gamma, \ t \geq 0 \qquad (6.1c)$$

We assume that t lies in a finite interval $[0,T]$. For the infinite interval $[0,\infty)$ a restriction on the growth of $\partial f(x,t)/_{\partial t}$ is required. For $v(x,t) \in H^{p+1}(\Omega)$, $t \geq 0$, and square integrable with respect to t, let us define the norm

$$\| v \|^2_{H^{p+1} \times L_2} \equiv \int_0^T \| v(x,t) \|^2_{H^{p+1}} dt.$$

Further, using the well-known inequality

$$\| v \|^2_{L_2} \leq C_\Omega \sum_{i=1}^{N} \int_\Omega \left(\frac{\partial v}{\partial x_i}\right)^2 dx \quad , \quad v \in H^1_o(\Omega)$$

we norm $H^1_o(\Omega)$ by

$$\| v \|^2_{H^1_o} = \sum_{i=1}^{N} \int_\Omega \left(\frac{\partial v}{\partial x_i}\right)^2 dx.$$

Analogously to chapter 1, the weak solution is immediately seen to satisfy

$$\left(\frac{\partial u}{\partial t}, v\right) + a(u,v) = (f(x,t),v) \qquad \forall\, v \in H_o^1(\Omega),\ t > 0 \qquad (6.2)$$

whilst the continuous time Galerkin solution U(x,t) is determined from

$$\left(\frac{\partial U}{\partial t}, V\right) + a(U,V) = (f(x,t),V) \qquad \forall\, V \in V_h^p,\ t > 0 \qquad (6.3)$$

Both (6.2) and (6.3) are associated with appropriate initial conditions.

Continuing as in Chapter 2, we easily achieve by the energy method that

$$\| u(x,t) \| \le e^{-\gamma t} \| g \| + \int_o^t e^{-\gamma(t-s)} \| f(x,s) \| ds$$

Assuming that f(x,t) is m-1 times continuously differentiable with respect to t we obtain from integration by parts

$$\| u(x,t) \| \le e^{-\gamma t} \| g \| + \sum_{r=1}^{m-1} \frac{(-1)^{r-1}}{\gamma^r} \left(\| f^{r-1}(x,t) \| - \| f^{r-1}(x,o) \| e^{-\gamma t} \right)$$

$$+ \frac{(-1)^{m-1}}{\gamma^{m-1}} \int_o^t e^{-\gamma(t-s)} \| f^{m-1}(x,s) \| ds \qquad (6.4)$$

where $f^p(x,t) \equiv \frac{\partial^p}{\partial t^p} f(x,t) \qquad p = 0,1,\ldots m-1.$

The continuous-time Galerkin solution satisfies an analogous bound (cf.(2.5)). Consequently, we define a time discretization scheme to be strongly stable if U^n satisfies:

$$\| U^n \| \le C\, e^{-\alpha n \Delta_t} \sum_{j=o}^{k-1} \| U^j \| + C \sum_{p=o}^{m-1} \frac{1}{\alpha^{p+1}} \sup_{o \le j \le n} \| f^p(x,j\Delta_t) \| \quad (6.5)$$

whenever Δ_t is sufficiently small, and m is an integer; $m \geq 1$.

Let us investigate the solution derived from an application of the L.M.S.D. method (3.1) to (6.3). The approximant U^n is easily seen to satisfy

$$\left(\sum_{j=o}^{k} \alpha_j U^{n+j} , V \right) - \left(\sum_{j=o}^{k} \sum_{r=1}^{m} \beta_{rj} \Delta_t^r U_{(r)}^{n+j} , V \right) = 0 \qquad (6.6_1)$$

where

$$\left(U^n_{(r)} , V \right) + a \left(U^n_{(r-1)} , V \right) = \left(f^{r-1} (x, n\Delta_t), V \right) \qquad (6.6_{11})$$

$$V\ V \in V_h^p , \quad r = 1,2,\ldots,m \ , \quad n = 0,1,\ldots$$

The results of this chapter are contained in the following theorem:

Theorem 3

Let the L.M.S.D. method (3.1) of order q be consistent, zero-stable and A_o-stable. Let the roots of the polynomial $\rho(\xi)$ with modulus equal to one be real, the modulus of the roots of the polynomial $\sigma_m(\xi)$ be less than one, and $\sigma_1(-1) \neq 0$ if $\rho(-1) = 0$. Further let $u(x,t)$, $\frac{\partial}{\partial t} u(x,t) \in H^{p+1}(\Omega)$, $t \in [0,T]$, and $f(x,t)$ be q+1 times continuously differentiable with respect to t and each such derivative $f^s(x,t) \in L_2(\Omega)$ $s = 0,1,\ldots, q+1$. Then, for any $t_o > 0$, there exists a positive constant $C(t_o)$ such that for $n\Delta_t \geq t_o$, $n\Delta_t \leq T$ and h, Δ_t sufficiently small

$$\| u(x,n\Delta_t) - U^n \| \leq C(t_o) \{ \Delta_t^q + h^{p+1} + \sum_{j=o}^{k-1} \| u(x,j\Delta_t) - U^j \| \}$$

and

$$\| U^n \| \leq C \, e^{-\alpha n\Delta_t\lambda_1} \sum_{j=o}^{k-1} \| U^j \| + C \sum_{p=o}^{m-1} \frac{1}{(\alpha\lambda_1)^{p+1}} \sup_{o\leq j\leq n} \| f^p(x,j\Delta_t) \|$$

The constant $C(t_o)$ depends on the parameter w defined by $\Delta_t = O(h^w)$. as $h \to 0$. We require that w be bounded away from zero. ie. $w \geq w_o > 0$.

Corollary

If in addition we assume that U^o is the projection of g(x) onto V_h^p by the L_2 - inner product and $\{U^j\}_{j=1}^{k-1}$ are the values derived from a weakly A_o - stable Padé scheme of order q-1, then

$$\| u(x,n\Delta_t) - U^n \| \leq C(t_o) \left\{ \Delta_t^q + h^{p+1} \right\}$$

and

$$\| U^n \| \leq C \, e^{-\alpha n\Delta_t\lambda_1} \left\{ \| g \| + \sum_{p=o}^{\tilde{m}-1} \frac{1}{(\alpha\lambda_1)^{p+1}} \sup_{o\leq s\leq k-1} \| f^p(x,s\Delta_t) \| \right\}$$

$$+ C \sum_{p=o}^{m-1} \frac{1}{(\alpha\lambda_1)^{p+1}} \sup_{o\leq s\leq n} \| f^p(x,s\Delta_t) \|$$

The latter closely resembles (6.4) in that the initial time values of $\| f^p(x,s\Delta_t) \|$, $p=0,1,\ldots,\tilde{m}-1$; $s=0,1,\ldots,k-1$ tend exponentially to zero as t increases.

Proof

This proof closely follows that of Theorem 1 but with complications arising from the extra term in the equation, ie. f(x,t).

The solution u(x,t) of (6.1) can be expressed as

$$u(x,t) = \sum_{i=1}^{\infty} g_i \, e^{-\lambda_i t} \psi_i + \sum_{i=1}^{\infty} (F_i(t), \psi_i)\psi_i \tag{6.7}$$

where $$F_i(t) = \int_o^t f(x,s) \, e^{-\lambda_i (t-s)} \, ds \tag{6.8}$$

Similarly the continuous-time Galerkin solution U(x,t) may be expressed as

$$U(x,t) = \sum_{i=1}^{d} U_i^o \, e^{-\Lambda_i t} \Psi_i + \sum_{i=1}^{d} (\overline{F}_i(t), \Psi_i)\Psi_i \tag{6.9}$$

where
$$\overline{F}_i(t) = \int_o^t f(x,s) \, e^{-\Lambda_i (t-s)} \, ds \tag{6.10}$$

a/

Using the same arguement as before we can write $e_2 = \sum_{i=1}^{d} \varepsilon_i^n \Psi_i$ where

$$\varepsilon_i^n = U_i^o \, e^{-\Lambda_i n \Delta_t} + (\overline{F}_i(n\Delta_t), \Psi_i) - U_i^n \tag{6.11}$$

Again let $U_{(r)}^n \equiv \sum_{i=1}^{d} U_{i,r}^n \Psi_i$ be the discrete approximation to $\frac{\partial^r}{\partial t^r} U(x,t)\Big|_{t=n\Delta_t}$. Substituting $U_{(r)}^n$ into (6.6) and letting $V = \Psi_i$ gives us the relationship

$$U_{i,r}^n + \Lambda_i U_{i,r-1}^n = \tilde{F}_n^{r-1} (\Psi_i) \qquad r = 1,\ldots,m$$

where $$\tilde{F}_n^p (\Psi_i) \equiv \int_\Omega f^p(x,n\Delta_t) \, \Psi_i \, dx$$

Consequently, we can construct a recurrence relationship, namely

$$U^n_{i,r} = (-1)^r \Lambda_i^r U_i^n + \sum_{p=o}^{r-1} (-\Lambda_i)^{r-1-p} \tilde{F}_n^p(\Psi_i) \tag{6.12}$$

Substituting (6.12) into (6.6_1) with $V = \Psi_i$ we achieve by a simple manipulation the relation, (cf. (5.11))

$$\sum_{j=o}^{k} \delta_j(\Delta_t\Lambda_i)\, U_i^{n+j} = \left[\mu_k(\Delta_t\Lambda_i)\right]^{-1} \left\{ \sum_{j=o}^{k} \sum_{r=1}^{m} \Delta_t^r \beta_{rj} \sum_{p=o}^{r-1} (-\Lambda_i)^{r-1-p} \tilde{F}_n^p(\Psi_i) \right\},$$

$$\equiv C_i^n \tag{6.13}$$

Combining (6.11) and (6.13) we have (cf. (5.12))

$$\sum_{j=o}^{k} \delta_j(\Delta_t\Lambda_i)\, \varepsilon_i^{n+j} = \sum_{j=o}^{k} \delta_j(\Delta_t\Lambda_i) \left\{ U_i^o\, e^{-\Lambda_i(n+j)\Delta_t} + (\bar{F}_i((n+j)\Delta_t), \Psi_i) \right\}$$

$$- c_i^n \equiv d_i^n \tag{6.14}$$

Note that by differentiation under the integral sign

$$\frac{\partial^r}{\partial t^r} \bar{F}_i(t) = (-\Lambda_i)^r \bar{F}_i(t) + \sum_{p=o}^{r-1} (-\Lambda_i)^{r-1-p} \frac{\partial^p}{\partial t^p} f(x,t) \tag{6.15}$$

and hence by (3.2) it is simple to deduce that

$$\mu_k(\Delta_t\Lambda_i) d_i^n \equiv L\left[U_i^o\, e^{-\Lambda_i t} + (\bar{F}_i(t), \Psi_i) \ , \Delta_t \right] \Big|_{t=n\,\Delta_t} \tag{6.16}$$

To bound (6.16) it is necessary to evaluate $\frac{\partial^{q+1}}{\partial t^{q+1}} \bar{F}_i(t)$ in an appropriate manner. Using integration by parts we show that

$$\bar{F}_i(t) = \sum_{p=o}^{q} \frac{(-1)^p}{\Lambda_i^{p+1}} \left\{ f^p(x,t) - f^p(x,o)\, e^{-\Lambda_i t} \right\}$$

$$+ \frac{(-1)^{q+1}}{\Lambda_i^{q+1}} \int_o^t f^{q+1}(x,s)\, e^{-\Lambda_i(t-s)}\, ds$$

and substituting this into (6.15) with r = q+1 we have

$$\frac{\partial^{q+1}}{\partial t^{q+1}} \overline{F}_i(t) = \sum_{p=o}^{q} (-1)^p \Lambda_i^p\, e^{-\Lambda_i t} f^{q-p}(x,o)$$

$$+ \int_o^t f^{q+1}(x,s)\, e^{-\Lambda_i(t-s)} ds \tag{6.17}$$

Combining (3.3), (6.16), (6.17) and lemma 1 we have

$$\mu_k(\Delta_t\Lambda_i) d_i^n \le G\, \Delta_t^{q+1} \left\{ \Lambda_i^{q+1} |U_i^o|\, e^{-\Lambda_i n\Delta_t} \right.$$

$$\left. + \sum_{p=o}^{q} \Lambda_i^p\, e^{-\Lambda_i n\Delta_t} \left|(f^{q-p}(x,o), \Psi_i)\right| \right\}$$

$$+ \Delta_t^{q+1} \int_o^k G(r) \left[\int_o^{(n+r)\Delta_t} (f^{q+1}(x,s), \Psi_i)\, e^{-\Lambda_i(n\Delta_t + r\Delta_t - s)} ds \right] dr$$

ie.
$$d_i^n \le C\, \Delta_t^{q+1} \left\{ \Lambda_i^{q+1} \left|U_i^o\right|\, e^{-\Lambda_i n\Delta_t} \right.$$

$$+ \sum_{p=o}^{q} \Lambda_i^p\, e^{-\Lambda_i n\Delta_t} \left| (f^{q-p}(x,o), \Psi_i)\right|$$

$$\left. + \int_0^T \left| (f^{q+1}(x,s), \Psi_i) \right| ds \right\} \tag{6.18}$$

Alternatively $\delta_j(\tau)$ is bounded for any $\tau > 0$, thus

$$d_i^n \le C\,|U_i^o|\,e^{-n\Delta_t\Lambda_i} + C\,\Delta_t^{q+1}\left\{\sum_{p=o}^{q} {}^{p}_{i}\,e^{-\Lambda_i n\Delta_t}\left|(f^{q-p}(x,o),\,\Psi_i)\right|\right.$$

$$\left. + \int_T^o \left|(f^{q+1}(x,s),\;\Psi_i\;)\right| ds\right\} \qquad (6.19)$$

Using (5.16), (5.20) with either (6.18) or (6.19) we prove by the method of chapter 5 that, for $i < i_*$

$$|\,\varepsilon_i^n\,| \le C(t_o)\,\Lambda_i^{-s}\sum_{j=1}^{k-1}|\,\varepsilon_i^j\,| + C(t_o)\,\Delta_t^q\left\{\,|\,U_o^i\,|\right.$$

$$\left. + \sum_{p=o}^{q}\left|\,(f^p(x,o),\,\Psi_i) + \int_o^T \left|\,(f^{q+1}(x,s),\,\Psi_i)\,\right| ds\right\} \qquad (6.20_I)$$

whilst for $i \ge i_*$

$$|\varepsilon_i^n| \le C(t_o)\,\Delta_t^q\left\{\sum_{j=1}^{k-1}\left|\varepsilon_i^j\right| + \left|U_i^o\right| + \int_o^T\left|\,(f^{q+1}\;(x,s),\,\Psi_i)\,\right|ds\right.$$

$$\left. + \sum_{p=o}^{q}\Lambda_i^p\,\Delta_t^{2s}\left|\,(f^{q-p}(x,o),\,\Psi_i)\right|\right\} \qquad (6.20_{II})$$

The only result not contained in chapter 5 that we used to establish (6.20) is a bound on $\Delta_t\sum_{\ell=o}^{n-k} e^{-\alpha\Delta_t\Lambda_i\ell}$ for any $i < i_*$,

ie.

$$\Delta_t\sum_{\ell=o}^{n-k} e^{-\alpha\Delta_t\Lambda_i\ell} \le \frac{\Delta_t}{1 - e^{-\alpha\Delta_t\Lambda_i}} \le \frac{C}{\alpha\lambda_1}$$

since $^x/(1-e^{-x})$ is an increasing function for $x > 0$ and $\Delta_t\Lambda_i < \hat{\tau}, i < i_*$.

From assumption (P_{11}) of chapter 2 we deduce immediately that $\Lambda_d \le Ch^{-2}$. We have assumed that $\Delta_t = O(h^w)$, $w \ge w_o > 0$. As $w = 1$ is compatible with large time increments this assumption is always satisfied in practise. Now,

$$\Lambda_i^p \cdot \Delta_t^{2s} \le \Lambda_d^p \, \Delta_t^{2s} \le Ch^{-2p} \, h^{2sw} \le C \quad , \quad s \ge {}^p/_w$$

and we obtain from (6.20) that

$$\| e_2 \|^2 = \sum_{i=1}^{d} |\varepsilon_i^n|^2 \le C(t_o) \, \Delta_t^{2q} \left\{ \| U^o \|^2 + \sum_{i \ge i_*} \sum_{j=1}^{k-1} |\varepsilon_i^j|^2 + \sum_{p=o}^{q} \| f^p(x,0) \|^2 \right.$$

$$\left. + \| f^{q+1}(x,t) \|^2_{L_2 \times L_2} \right\} + C(t_o) \sum_{i<i_*} \Lambda_i^{-2s} \sum_{j=1}^{k-1} |\varepsilon_i^j|^2 \qquad (6.21)$$

From (6.10) and (6.11) we see directly that

$$|\varepsilon_i^j| \le |U_i^o| + |U_i^j| + \frac{1}{\lambda_i} \sup_{o \le s \le j} |(f(x,s), \Psi_i)| \qquad (6.22)$$

and

$$\varepsilon_i^j = e^{-j\Delta_t\Lambda_i}(U_i^o - \overline{U}_i^o) + e^{-j\Delta_t\Lambda_i}(\overline{U}_i^o - u_i^o) + (e^{-j\Delta_t\Lambda_i} - e^{-j\Delta_t\lambda_i}) \, u_i^o$$

$$+ (u_i^j - \overline{U}_i^j) + (\overline{U}_i^j - U_i^j) + (\overline{F}_i(j\Delta_t) - F_i(j\Delta_t) \; , \; \Psi_i)$$

$$+ (F_i(j\Delta_t) \; , \; \Psi_i - \psi_i)$$

from which, cf. (5.30), if $e_3 = \sum_{i<i_*} \Lambda_i^{-2s} \sum_{j=1}^{k-1} |\varepsilon_i^j|^2$

$$|e_3| \le C \sum_{j=o}^{k-1} \sum_{i<i_*} \Lambda_i^{-2s} \, |u_i^j - \overline{U}_i^j|^2 + C \sum_{j=o}^{k-1} \sum_{i<i_*} \Lambda_i^{-2s} \, |\overline{U}_i^j - U_i^j|^2$$

$$+ C \sum_{j=1}^{k-1} \sum_{i<i_*} \Lambda_i^{-2s} \left| e^{-j\Delta_t\Lambda_i} - e^{-j\Delta_t\lambda_i} \right|^2 |u_i^o|^2$$

$$+ C \sum_{j=o}^{k-1} \sum_{i<i_*} \Lambda_i^{-2s} \left| (\overline{F}_i(j\Delta_t) - F_i(j\Delta_t), \Psi_i) \right|^2$$

$$+ C \sum_{j=o}^{k-1} \sum_{i<i_*} \Lambda_i^{-2s} \left| (F_i(j\Delta_t) , \Psi_i - \psi_i) \right|^2 \qquad (6.23)$$

Let $e_{12} = \sum_{i<i_*} \Lambda_i^{-2s} \left| (\overline{F}_i (j\Delta_t) - F_i (j\Delta_t), \Psi_i) \right|^2$

$$\leq \sum_{i<i_*} \Lambda_i^{-2s} \| \overline{F}_i (j\Delta_t) - F_i(j\Delta_t) \|^2$$

But $\overline{F}_i(t) - F_i(t) = \int_o^t f(x,s) \left(e^{-\Lambda_i(t-s)} - e^{-\lambda_i(t-s)} \right) ds$ and by the mean value theorem we have

$$\| \overline{F}_i (j\Delta_t) - F_i(j\Delta_t) \| \leq \sup_{o\leq s\leq j\Delta_t} \| f(x,s) \| \int_o^{j\Delta t} | j\Delta_t - s | \, ds \, | \Lambda_i - \lambda_i |$$

$$\leq C \sup_{o\leq s\leq j\Delta_t} \| f(x,s) \| \Delta_t^2 h^{2p} \lambda_i^{p+1}$$

Selecting $s = p+1 + \frac{N}{2}$ we have

$$e_{12} \leq C \Delta_t^4 h^{4p} \sup_{o\leq s\leq j\Delta_t} \| f(x,s) \|^2 \qquad (6.24)$$

Denote

$$e_{13} = \sum_{i<i_*} | (F_i(j\Delta_t), \Psi_i - \psi_i) |^2 \Lambda_i^{-2s}$$

But $\left| (F_i(j\Delta_t), \Psi_i - \psi_i) \right| \leq \| F_i(j\Delta_t) \| \; \| \Psi_i - \psi_i \|$

$$\leq \frac{C}{\lambda_i} h^{p+1} \lambda_i^{\frac{1}{2}(p+1)} \sup_{o \leq s \leq j\Delta_t} \| f(x,s) \|$$

Selecting $2s = p-1+N$ we have

$$e_{13} \leq Ch^{2(p+1)} \sup_{o \leq s \leq j\Delta_t} \| f(x,s) \|^2 \tag{6.25}$$

We construct a bound on $|e_3|$, (6.23), by using (5.31), (6.24) and (6.25). Hence, via (6.21), (6.22) and the inequality

$$\begin{aligned} \| U^j \| &\leq \| U^j - u^j \| + \| u^j \| \\ &\leq \| U^j - u^j \| + \| g \| + C \sup_{o \leq s \leq j\Delta_t} \| f(x,s) \| \end{aligned}$$

we prove

$$\begin{aligned} \| e_2 \| \leq C(t_o)\, \Delta_t^q & \left\{ \| g \| + \sup_{o \leq s \leq (k-1)\Delta_t} \| f(x,s) \| + \sum_{p=o}^{q} \| f^p(x,0) \| \right. \\ & \left. + \| f^{q+1}(x,t) \|_{L_2 x L_2} \right\} + C(t_o) h^{p+1} \left\{ \| g \| + \sup_{o \leq s \leq (k-1)\Delta_t} \| f(x,s) \| \right\} \\ & + C(t_o) \sum_{j=o}^{k-1} \| U^j - u^j \| \end{aligned} \tag{6.26}$$

b. In this section a bound on the L_2-norm of $e_1 \equiv u(x,n\Delta_t) - U(x,n\Delta_t)$ is derived by utilising a well-documented method based on the paper by Wheeler [36]. It appears to be essential to impose certain conditions on $u(x,t)$ to ensure validity of the error bound for the nonhomogeneous problem.

Let us assume certain 'smoothness' conditions on u(x,t), namely,

(A_1) $\quad u(x,t),\ \frac{\partial}{\partial t} u(x,t) \quad \epsilon \quad H^{p+1}(\Omega) \times L_2\ [0,T]$

(A_{11}) $\quad$ If $k \epsilon L_2(\Omega)$ and $\phi \epsilon H^2(\Omega)$ satisfy

$$a(\phi,V) = (K,V) \qquad \forall \quad V \epsilon V_h^p$$

then there exists a positive constant c such that

$$\| \phi \|_2 \leq C \| k \|$$

This result is standard given $\{a_{ij}(x)\}_{i,j=1}^N$ and Γ sufficiently smooth (Miranda [23]), and may even hold if Ω has certain corners (Wheeler [36]). Consequently, the initial assumptions $\{a_{ij}(x)\}_{ij=1}^N \epsilon C^\infty(\bar{\Omega})$, $\Gamma \epsilon C^\infty$ make (A_{11}) superfluous.

Following Wheeler [36] , Dupont, Fairweather and Johnson [8] , amongst others, let us define $W \epsilon V_h^p$, $\forall$ $t \geq 0$, by

$$a(u - W,V) = 0 \qquad \forall \ V \epsilon V_h^p \tag{6.27}$$

Obviously W exists, and is in fact the weighted H^1 projection of u(x,t) into V_h^p. The subsequent result follows directly from the proof of lemma (4.1) [8] , and the assumptions (A). Denote $\eta \equiv W - u$, then, whenever h is sufficiently small

$$\| \eta \|_r + \| \frac{\partial}{\partial t} \eta \|_r \leq Ch^{p+1-r} \{ \| u \|_{p+1} + \| \frac{\partial u}{\partial t} \|_{p+1} \} \quad r=0,1 \tag{6.28}$$

for any $t \epsilon [0,T]$. Let $\xi \equiv W - U \epsilon V_h^p$, then subtracting (6.3) from (6.2) we easily see that

$$(\frac{\partial}{\partial t} \xi , V) + a(\xi, V) = (\frac{\partial}{\partial t} \eta, V) \qquad \forall \ V \epsilon V_h^p, \quad t > 0$$

Selecting $V = \xi$, and applying (1.2) yields

$$\frac{d}{dt} \| \xi \|^2 + \gamma \| \xi \|^2_{H^1_o} \leq (\frac{\partial}{\partial t} \eta, \xi)$$

Now, using the inequality $|ab| \leq \frac{1}{2} a^2 + \frac{1}{2} b^2$ and integrating over $[0,t]$ we have

$$\| \xi(\cdot,t) \|^2 + \gamma \| \xi \|^2_{H^1_o x L_2} \leq \| \xi(\cdot,0) \|^2 + \frac{1}{2} \|\frac{\partial}{\partial t} \eta\|^2_{L_2 x L_2} + \frac{1}{2} \int_o^t \|\xi(\cdot,s)\|^2 ds \qquad (6.29)$$

Before we can proceed it is necessary to quote the following lemma; (see [2] for proof)

Gronwall's lemma

Let u(t), v(t) be non-negative for all $t \geq 0$ and further let C be a positive constant, if

$$u(t) \leq C + \int_o^t uvdt$$

then $$u(t) \leq C \exp \left(\int_o^t vdt \right)$$

Applying the above lemma to (6.29) and using (6.28) yields

$$\| \xi(\cdot,t) \| \leq C \left\{ \| \xi(\cdot,0) \| + h^{p+1} \|\frac{\partial u}{\partial t} \|_{H^{p+1} x L_2} \right\} \quad \forall \, t \in [0,T]$$

A simple application of the triangle inequality and (6.28) produces the desired result, namely

$$\| e_1 \| \leq C \{\| u(x,0) - U(x,o) \| + h^{p+1} \} \qquad (6.30)$$

c. Returning to (6.13) we have

$$\sum_{j=o}^{k} \delta_j(\Delta_t \Lambda_i) U_i^{n+j} = c_i^n$$

Rewrite the above equality with n = n-k-ℓ, multiply this by $\gamma_\ell(\Delta_t\Lambda_i)$, sum for ℓ = 0,1, ...,n-k, and apply (5.18) to achieve (5.20) with $\varepsilon_i^j \equiv U_i^j$ and $d_i^n \equiv c_i^n$. Let us assume that Δ_t is sufficiently small to ensure $\Delta_t\lambda_1 < \hat{\tau}$. Note that by (6.13)

$$\mu_k(\Delta_t\Lambda_i)c_i^n = \Delta_t \sum_{j=o}^{k} \sum_{r=1}^{m} \beta_{rj}\, \Delta_t^{r-1} (-\Lambda_i)^{r-1} \sum_{p=o}^{r-1} (-\Lambda_i)^{-p}\, \tilde{F}_n^p(\Psi_i)$$

and by lemmas 1 and 2 we deduce that

$$c_i^n \leq \Delta_t \sum_{p=o}^{m-1} \frac{1}{\lambda_1 p} \left| \tilde{F}_n^p (\Psi_i) \right| \tag{6.31}$$

Using (5.16), (6.31) and the variation of (5.20) we have

$$|U_i^n| \leq C\, e^{-\alpha n\Delta_t\lambda_1} \sum_{j=o}^{k-1} |U_i^j| + C\Delta_t \sum_{\ell=o}^{n-1} \sum_{p=o}^{m-1} \frac{1}{\lambda_1^{\,p}} \left| \tilde{F}^p_{n-k-\ell}(\Psi_i) \right| e^{-\alpha\ell\Delta_t\lambda_1}$$

As shown earlier, we note that for Δ_t sufficiently small

$$\Delta_t \sum_{\ell=0}^{n-k} e^{-\alpha\ell\Delta_t\lambda_1} \leq \frac{\Delta_t}{1-e^{-\alpha\Delta_t\lambda_1}} \leq \frac{C}{\alpha\lambda_1}$$

and hence

$$\| U^n \| = \left[\sum_{i=1}^{d} |U_i^n|^2 \right]^{\frac{1}{2}} \leq Ce^{-\alpha n\Delta_t\lambda_1} \sum_{j=o}^{k-1} \| U^j \| + \frac{C}{\alpha} \sum_{p=o}^{m-1} \frac{1}{\lambda_1^{\,p+1}} \sup_{o\leq s\leq n} \| f^P(x,s\Delta_t) \|$$

$$\leq Ce^{-\alpha n\Delta_t\lambda_1} \sum_{j=o}^{k-1} \| U^j \| + C(\alpha) \sum_{p=o}^{m-1} \frac{1}{(\alpha\lambda_1)^{p+1}} \sup_{o\leq s\leq n} \| f^P(x,s\Delta_t) \| \tag{6.32}$$

where $\alpha \geq \frac{1}{2} \min \{|a_1^1| = 1, |a_1^2|\}$

d. <u>Initial approximants</u>

Applying the derivative relations (6.12) to the weakly A_o-stable Padé scheme (5.45) yields a relation for the initial approximants U^{j+1}, $j=0,1,\dots, k-2$, namely

$$U_i^{j+1} = R(\Delta_t\Lambda_i)\; U_i^j + \frac{1}{\mu_1(\Delta_t\Lambda_i)} \sum_{s=o}^{1} \sum_{r=1}^{\tilde{m}} \Delta_t^r\, \tilde{\beta}_{rs} \sum_{p=o}^{r-1} (-1)^p \Lambda_i^p\, \tilde{F}_{j+s}^{(r-1-p)}(\Psi_i)$$

$$\equiv R(\Delta_t\Lambda_i)U_i^j + \frac{1}{\mu_1(\Delta_t\Lambda_i)} \quad \Phi_i^j \qquad (6.33_1)$$

where $\mu_1(\tau) = 1 + \sum_{r=1}^{\tilde{m}} (-1)^{r-1}\, \tilde{\beta}_{r1}\, \tau^r$

The recurrence relationship (6.33_1) gives us, (cf. (5.49))

$$U_i^{j+1} = \left[R(\Delta_t\Lambda_i)\right]^{j+1} \; U_i^o + \frac{1}{\mu_1(\Delta_t\Lambda_i)} \sum_{\ell=o}^{j} \left[R(\Delta_t\Lambda_i)\right]^{\ell} \; \Phi_i^{j-\ell} \qquad (6.33_{11})$$

and hence by (6.11)

$$\varepsilon_i^{j+1} = U_i^o \left\{e^{-(j+1)\Delta_t\Lambda_i} - \left[R(\Delta_t\Lambda_i)\right]^{\,j+1}\right\} \qquad (6.34)$$

$$+ \left(\overline{F}_i((j+1)\Delta_t) - \frac{1}{\mu_1(\Delta_t\Lambda_i)} \sum_{\ell=o}^{j} \left[R(\Delta_t\Lambda_i)\right]^{\ell} \Phi_i^{j-\ell},\; \Psi_i\right) \quad j=0,1,\dots,k-2$$

The expression (6.34) will be bounded by (5.50) and by investigating the error, at time $t = (j+1)\Delta_t$, when the Padé scheme is applied to the function $\overline{F}_i(t)$. Substituting (6.15) into (5.47) we have that

$$\overline{F}_i\,((j+1)\Delta_t) = R(\Delta_t\Lambda_i)\,\overline{F}_i\,(j\Delta_t) + \frac{1}{\mu_1(\Delta_t\Lambda_i)}\{\Phi_i^j + E_i^j\}$$

where E_i^j (cf. (3.3)) is the integral form of the remainder.

Consequently,

$$\overline{F}_i\,((j+1)\Delta_t) = \frac{1}{\mu_1(\Delta_t\Lambda_i)}\sum_{\ell=o}^{j}\left[R(\Delta_t\Lambda_i\,)\right]^{\ell}\left\{\Phi_i^{j-\ell}+E_i^{j-\ell}\right\} \qquad (6.35)$$

Comparing (6.34) and (6.35) we require a bound on

$$S_i^j \equiv \frac{1}{\mu_1(\Delta_t\Lambda_i)}\sum_{\ell=o}^{j}\left[R(\Delta_t\Lambda_i)\right]^{\ell}\left(E_i^{j-\ell}\,,\,\Psi_i\right) \qquad j = 0,1,\ldots,k-2$$

By lemma 1, the definition of weak A_o - stability, and (6.16) - (6.18) we have

$$S_i^j \le C\,\Delta_t^q\left\{\sum_{p=o}^{q}\Lambda_i^p\,e^{-j\Delta_t\Lambda_i}\left|(f^{q-p}\,(x,0)\,,\,\Psi_i)\right| + \int_0^T\left|\,(f^{q+1}\,(x,s),\,\Psi_i)\,\right|ds\right\}$$

and combining this with (5.50) we achieve

$$\left|\varepsilon_i^{j+1}\right| \le C\,\Delta_t^q\left\{\Lambda_i^q\,|U_i^o| + \sum_{p=o}^{q}\Lambda_i^p\,(f^{q-p}(x,0),\,\Psi_i) + \int_o^T|(f^{q+1}(x,s),\Psi_i)|ds\right\} \qquad (6.36)$$

Note that

(1) $\quad |U_i^j| \le |U_i^o| + C\sum_{p=o}^{\tilde{m}-1}\frac{1}{\lambda_1^{p+1}}\sup_{o\le s\le j}\left|(f^p\,(x,s\Delta_t),\,\Psi_i)\right| \qquad$ (cf. 6.33_{11}))

(11) $\quad \| U^o \| \le \| g \|$, if U^o is the projection of $g(x)$ into V_h^p by the L_2 - inner product

Consequently, substituting (6.36) into (6.21) with appropriate values of s and using (6.22), (6.32) with the above two inequalities we establish the following results

$$\| e_2 \| \leq C(t_o)\, \Delta_t^q \left\{ \| g \| + \sum_{p=o}^{q} \| f^p(x,0) \| + \sup_{\substack{o \leq s \leq k-1 \\ o \leq p \leq \tilde{m}-1}} \| f^p(x, s\Delta_t) \| \right.$$

$$\left. + \| f^{q+1}(x,t) \|_{L_2 \times L_2} \right\} \qquad (6.37)$$

$$\| U^n \| \leq C\, e^{-\alpha n \Delta_t \lambda_1} \left\{ \| g \| + \sum_{p=o}^{\tilde{m}-1} \frac{1}{(\alpha\lambda_1)^{p+1}} \sup_{o \leq s \leq k-1} \| f^p(x, s\Delta_t) \| \right\}$$

$$+ C \sum_{p=o}^{m-1} \frac{1}{(\alpha\lambda_1)^{p+1}} \sup_{o \leq s \leq n} \| f^p\, (x, s\Delta_t) \| \qquad (6.38)$$

The theorem and its corollary follow from (6.26), (6.30), (6.32), (6.37) and (6.38).

7. Examples and Implementation of L.M.S.D. Methods

a) Examples

The L.M.S.D. schemes with $k > 1$, $m > 1$, have significant advantages over the linear multistep schemes, $m = 1$, and the one-step schemes, $k = 1$, eg. the Padé approximants. The principle benefit is the availability of high order schemes for low values of m and k. Cryer [4] proves that an A_o-stable, linear multistep scheme, $m = 1$, $k > 2$ has order at most k. Similarly, it is well known that an A-stable Padé scheme has order $q = 2m$, but if stability at infinity is required then $q = 2m-1$. Note that A_o - stability is a weaker condition than A - stability. For comparative purposes we now quote several previously documented schemes.

To illustrate linear multistep schemes, $m = 1$, we quote from Zlámal [43]. A class of second order schemes, $k = 2$, is defined by

$$(-1 + \alpha)y_n + (1 - 2\alpha)y_{n+1} + \alpha y_{n+2} = \Delta_t\{(\tfrac{1}{2} - \alpha + \beta)y_n'$$

$$+ (\tfrac{1}{2} + \alpha - 2\beta)y_{n+1}' + \beta y_{n+2}'\} \qquad (7.1)$$

The necessary and sufficient conditions that the method be consistent, zero-stable, and A_o-stable are

$$\alpha \geq \frac{1}{2}, \qquad \beta > \frac{1}{2}\alpha$$

Defining $\beta = \frac{1}{2}\alpha + \varepsilon$ it can be shown that the error constant C_3 is given by $C_3 = -\left(\frac{1}{12} + \varepsilon\right)$. In the sequel we normalise the error constants, C_{q+1}, by defining $\sum_{j=o}^{k} \beta_{1j} = 1$, where β_{1j} is the coefficient of y_{n+j}'. Zlámal notes that for any $\varepsilon > 0$ the best stability at infinity is obtained by minimising, with respect to α, the roots of $\sigma_1(\xi)$.

ie. minimise $(2\varepsilon - \frac{1}{2} \pm \sqrt{(\alpha - \frac{1}{2})^2 - 4\varepsilon})\Big/(\alpha + 2\varepsilon)$

The class of third order schemes, k = 3, is defined by

$$\sum_{j=o}^{3} \alpha_j \, y_{n+j} = \Delta_t \sum_{j=o}^{3} \beta_j y'_{n+j} \qquad \text{where}$$

$$\begin{aligned} \alpha_3 &= \frac{1}{3} + \frac{1}{4}(\alpha + \beta) & \beta_3 &= \frac{1}{8}(1 + \alpha + \beta + w) \\ \alpha_2 &= -\frac{1}{4}(\alpha + 3\beta) & \beta_2 &= \frac{1}{8}(3 + \alpha - \beta - 3w) \\ \alpha_1 &= -\frac{1}{4}(\alpha - 3\beta) & \beta_1 &= \frac{1}{8}(3 - \alpha - \beta + 3w) \\ \alpha_o &= -\frac{1}{3} + \frac{1}{4}(\alpha - \beta) & \beta_o &= \frac{1}{8}(1 - \alpha + \beta - w) \end{aligned} \qquad (7.2_{11})$$

$$w = (1 - \gamma)\alpha\beta$$

Here, for the stability and consistency conditions, either

$$\gamma = 0, \qquad \beta \geq 0, \qquad \alpha = \frac{\sqrt{3}}{3}$$

or

$$0 < \gamma \leq 1, \qquad \beta > 0, \qquad \alpha > \sqrt{\beta\gamma + \frac{1}{3}} - \sqrt{\beta\gamma} \; ,$$

and the error constant is

$$C_4 = -\frac{1}{24}(\alpha + 3w)$$

We illustrate the class of consistent, zero -, and A_0 - stable, one-step, multiderivative methods, k = 1, m ≥ 1, by quoting the Padé (1,2), Padé (1,3) and Padé (2,3) schemes. The third order Padé (1,2) is given by

$$y_{n+1} - y_n = \Delta_t \left\{ \frac{2}{3} y'_{n+1} + \frac{1}{3} y'_n \right\} - \frac{\Delta_t^2}{6} y''_{n+1} \qquad (7.2_1)$$

and the error constant $C_4 = \frac{1}{72}$. The fourth order Padé (1,3) is defined by

$$y_{n+1} - y_n = \Delta_t \{ \frac{3}{4} y'_{n+1} + \frac{1}{4} y'_n \} - \frac{\Delta_t^2}{4} y''_{n+1} + \frac{\Delta_t^3}{24} y'''_{n+1} \qquad (7.2_{11})$$

where the error constant $C_5 = - \frac{1}{480}$. Finally we quote the fifth order Padé (2,3) namely

$$y_{n+1} - y_n = \Delta_t \{ \frac{3}{5} y'_{n+1} + \frac{2}{5} y'_n \} + \Delta_t^2 \{ - \frac{3}{20} y''_{n+1} + \frac{1}{20} y''_n \}$$
$$+ \frac{\Delta_t^3}{60} y'''_{n+1} \qquad (7.2_{111})$$

where the error constant $C_6 = - \frac{1}{7200}$

Other Padé approximations can be obtained by using the formulae for the rational approximations to e^{-x}, (eg. [27, 7.3 - 7.4]). We note that all Padé (α, β) approximants with $\beta > \alpha$ are stable at ∞. (i.e. modulus of roots of σ_m (ξ) are less than one).

To illustrate the multistep, multiderivative methods we select $k = m = 2$ and derive a family of fifth order, A_o - stable methods. Any fifth order method with $k = m = 2$ may be expressed as

$$(\alpha - 1)y_n + (1 - 2\alpha) y_{n+1} + \alpha y_{n+2} = \Delta_t \left\{ (\frac{7}{15} - \beta) y'_n + \frac{8}{15} y'_{n+1} \right.$$
$$\left. + \beta y'_n \right\} + \Delta_t^2 \left\{ \left(\frac{5}{72} + \frac{\alpha}{12} - \frac{\beta}{3} \right) y''_n + \left(- \frac{19}{180} + \frac{5\alpha}{6} - \frac{4\beta}{3} \right) y''_{n+1} \right.$$
$$\left. + \left(\frac{1}{360} + \frac{\alpha}{12} - \frac{\beta}{3} \right) y''_{n+2} \right\} \qquad (7.3)$$

We test for A_o-stability by employing the Routh-Hurwitz criterion, eg. [16 , pp 80] . For simplicity we employ a previous notation, namely:

$$\mu_2(\tau) = \alpha + \beta\tau + \left(\frac{\beta}{3} - \frac{\alpha}{12} - \frac{1}{360}\right)\tau^2$$

$$\mu_1(\tau) = (1 - 2\alpha) + \frac{8}{15}\tau + \left(\frac{4\beta}{3} - \frac{5\alpha}{6} + \frac{19}{180}\right)\tau^2$$

$$\mu_o(\tau) = (\alpha - 1) + \left(\frac{7}{15} - \beta\right)\tau + \left(\frac{\beta}{3} - \frac{\alpha}{12} - \frac{5}{72}\right)\tau^2$$

for any $\tau \geq 0$. By (3.4) we require the roots of the polynomial $p(\xi,\tau) = \sum_{j=o}^{2} \mu_j(\tau)\xi^j$ to be less than one in modulus for all $\tau > 0$.

By the Rough-Hurwitz criterion this requirement is satisfied if,

$\mu_2(\tau) > \mu_1(\tau) - \mu_o(\tau)$

$$\text{ie. } (4\alpha - 2) - \frac{\tau}{15} + \left(\frac{2\alpha}{3} - \frac{2\beta}{3} - \frac{8}{45}\right)\tau^2 > 0 \qquad (7.4_1)$$

$\mu_2(\tau) > \mu_o(\tau)$

$$\text{ie. } 1 + \left(2\beta - \frac{7}{15}\right)\tau + \tau^2/15 > 0 \qquad (7.4_{11})$$

$\mu_2(\tau) + \mu_1(\tau) + \mu_o(\tau) > 0$

$$\text{ie. } \tau + (2\beta - \alpha + \frac{1}{30})\tau^2 > 0 \qquad (7.4_{111})$$

for all $\tau > 0$. Note that, by lemmas 1 and 2, we also require

$$\mu_2(\tau) > 0 , \quad \left(\frac{\beta}{3} - \frac{\alpha}{12} - \frac{1}{360}\right) > 0 \qquad (7.4_{1v})$$

The inequalities (7.4) are satisfied if

$$\alpha > \frac{1}{2} , \qquad 2\beta - \alpha \geq -\frac{1}{30}$$

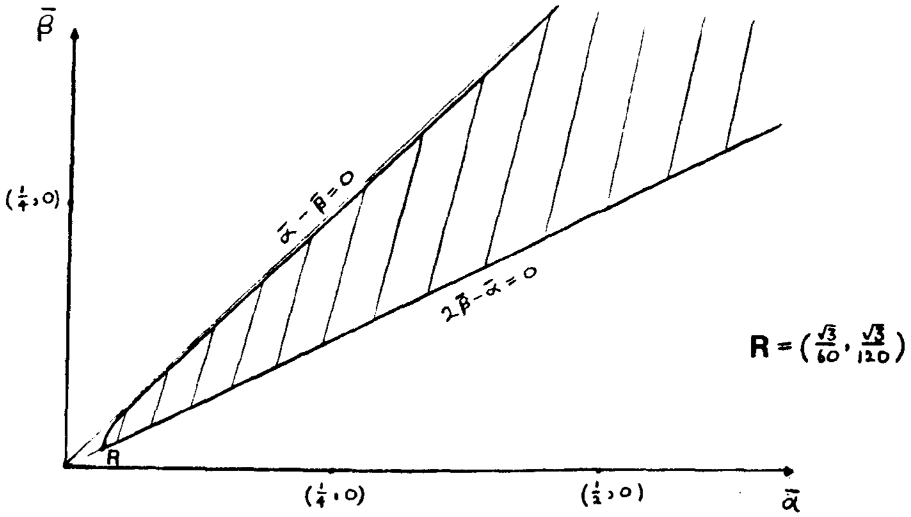
and $\left(\frac{1}{15}\right)^2 < 4(4\alpha - 2)(\frac{2\alpha}{3} - \frac{2\beta}{3} - \frac{8}{45})$

The region defined is best seen if we change the basis and let

$$\bar{\alpha} = \alpha - \frac{1}{2} , \qquad \bar{\beta} = \beta - \frac{7}{30}$$

from which we deduce that the inequalities (7.4) are equivalent to

$$\bar{\alpha} > 0 , \quad 2\bar{\beta} - \bar{\alpha} \geq 0 \text{ and } \left(\frac{1}{15}\right)^2 < \frac{32}{3} \bar{\alpha} (\bar{\alpha} - \bar{\beta})$$


$\bar{\beta}$
$(\frac{1}{4},0)$
$\bar{\alpha}-\bar{\beta}=0$
$2\bar{\beta}-\bar{\alpha}=0$
$R=(\frac{\sqrt{3}}{60},\frac{\sqrt{3}}{120})$
R
$(\frac{1}{4},0)$
$(\frac{1}{2},0)$
$\bar{\alpha}$

The shaded area of Diagram 1 contains the admissible values for $\bar{\alpha}$ and $\bar{\beta}$. We note here that the error constant of the L.M.S.D. scheme (7.5) is given by

$$C_6 = - {}^{11}/_{21600} - {}^{\alpha}/_{240} + {}^{\beta}/_{90}$$

The selection of particular values from the admissible range of the parameters α and β is now considered. Any scheme proposed to solve the stiff system of equations (6.3) should exhibit certain characteristics, of which, the principle is related to the nature of the analytic solution. The continuous time Galerkin solution, (6.9) - (6.10), can be divided into the steady-state solution, $\sum_{i=1}^{d} (\bar{F}_i(t),\Psi_i)\Psi_i$, and the transient solution $\sum_{i=1}^{d} U_i^o e^{-\Lambda_i t} \Psi_i$, so named as

$$\sum_{i=1}^{d} U_i^o e^{-\Lambda_i t} \Psi_i \rightarrow 0 \quad \text{as} \quad t \rightarrow \infty$$

Obviously, a desirable feature of any multistep method is that the component of the approximate solution corresponding to the transient solution should decay rapidly as t increases. Thus, for a stiff system, we favour schemes with an inherent ability to tackle effectively the components of the transient solution corresponding to large values of Λ_i.

Let us apply the scheme (3.1) to the scalar test equation $\dot{y} = -\lambda y$, $y(0) = 1$; $\lambda > 0$. By the definition of A_o - stability we know that the approximate solution $Y_n \rightarrow 0$ as $n \rightarrow \infty$. Importance is often attached to a scheme's 'stability at infinity'; that is, the behaviour of the approximate solution to the above scalar equation as $\lambda \rightarrow \infty$. For $\lambda >> 0$ the solution Y_n approaches the solution of the difference equation

$$\sum_{j=o}^{k} \beta_{mj} \overline{Y}_{n+j} = 0 \quad \text{as } \lambda \rightarrow \infty$$

Without loss of generality we shall assume that the roots $\{z_i\}_{i=1}^{k}$ of $\sigma_m(\xi)$, see (3.4), are real and distinct, then

$$Y_n \approx \sum_{i=1}^{k} a_i z_i^n \qquad \text{as} \quad \lambda \rightarrow \infty$$

where $\{a_i\}_{i=1}^{k}$ are constants determined by the initial values $\{Y_i\}_{i=o}^{k-1}$.

By assumption we know that $|z_i| < 1$, $i = 1,2,\ldots k$; hence $Y_n \rightarrow 0$ as $n \rightarrow \infty$ and the scheme is said to be stable at infinity. However, the rate of convergence of Y_n to zero may be improved by fixing the roots of $\sigma_m(\xi)$ to be equal, or close, to zero. Consequently, given a very stiff system of equations it is desirable to use a multistep scheme where the roots of $\sigma_m(\xi)$ are equal, or close, to zero.

Equally, we desire that the normalised error constant, C_{q+1}, is small

ie. C_{q+1} is defined by (3.2) with $\sum\limits_{j=o}^{k} \beta_{1j} = 1$

Consequently, we advance the following possibilities:

$$\alpha = 11/20 \;, \quad \beta = 79/300 \;, \quad C_6 = 1/8000 \;, \quad |\xi_1| \sim .86 \qquad (7.5a)$$

$$\alpha = 3/5 \;, \quad \beta = 7/24 \quad C_6 = 1/4320 \;, \quad |\xi_1| \sim .77 \qquad (7.5b)$$

$$\alpha = 2/3 \;, \quad \beta = 1/3 \quad C_6 = 1/2400 \;, \quad |\xi_1| \sim .57 \qquad (7.5c)$$

$$\alpha = 23/30 \;, \quad \beta = 2/5 \quad C_6 = 1/1350 \;, \quad \xi_1 = 0 \qquad (7.5d)$$

where ξ_1 is the largest root in modulus of $\sigma_2(\xi)$.

Higher order A_o - stable L.M.S.D. methods may be obtained by allowing either, or both, of m and k to be greater than two. Without reference to the general class of such schemes we note the following particular examples:

k = 2, m = 3

$$\frac{9}{10} y_{n+2} - \frac{4}{5} y_{n+1} - \frac{1}{10} y_n = \Delta_t\{\frac{23}{40} y'_{n+2} + \frac{2}{5} y'_{n+1} + \frac{1}{40} y'_n\} \qquad (7.6_1)$$

$$- \frac{3}{20} \Delta_t^2 y''_{n+2} + \frac{1}{60} \Delta_t^3 y'''_{n+2} \;, \quad q = 6 \;, \quad C_7 = -1/12600$$

$$\frac{15}{14} y_{n+2} - \frac{8}{7} y_{n+1} + \frac{1}{14} y_n = \Delta_t\{\frac{39}{70} y'_{n+2} + \frac{16}{35} y'_{n+1} - \frac{1}{70} y'_n \}$$

$$- \frac{4}{35} \Delta_t^2 \{y''_{n+2} - y''_{n+1}\} + \frac{1}{105} \Delta_t^3 y'''_{n+2} , \quad q = 7 , \quad C_8 = {}^{-1}/176400$$

$$(7.6_{11})$$

$k = 3, \quad m = 2$

$$\sum_{j=o}^{3} \alpha_j \, y_{n+j} = \Delta_t \sum_{j=o}^{3} \beta_j y'_{n+j} - c\Delta_t^2 \, y''_{n+3}$$

where

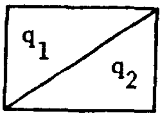
$$\begin{aligned} \alpha_3 &= {}^{11}/_{60} + {}^{39c}/_{4} & \beta_3 &= {}^{1}/_{20} + {}^{67c}/_{12} \\ \alpha_2 &= {}^{9}/_{20} - {}^{63c}/_{4} & \beta_2 &= {}^{9}/_{20} + {}^{9c}/_{4} \\ \alpha_1 &= - {}^{9}/_{20} + {}^{9c}/_{4} & \beta_1 &= {}^{9}/_{20} - {}^{27c}/_{4} \\ \alpha_o &= - {}^{11}/_{60} + {}^{15c}/_{4} & \beta_o &= {}^{1}/_{20} - {}^{13c}/_{12} \end{aligned} \qquad (7.6_{111})$$

This is a sixth order method with error constant $C_7 = \frac{1}{80} (c - \frac{1}{35})$ unless $c = 1/35$ which yields a seventh order scheme with error constant $C_8 = 1/19600$. A_o - stability is ensured by the condition

$$c > {}^{384}/_{17,275}$$

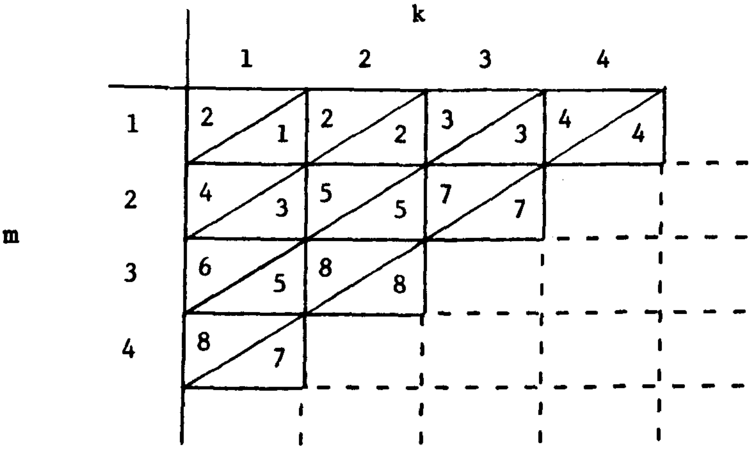
From the relevant theory, eg. Cryer [4] , or by direct evaluations we have established the following table concerning maximum orders of A_o - stable L.M.S.D. schemes.

The diagram expresses for $1 \leq m + k \leq 5$:



q_1 = maximum order of A_o - stable L.M.S.D. scheme for specific values of m and k.

q_2 is as q_1 but with added stipulation of stability at ∞.



m \ k	1	2	3	4
1	2 / 1	2 / 2	3 / 3	4 / 4
2	4 / 3	5 / 5	7 / 7	
3	6 / 5	8 / 8		
4	8 / 7			

b. Implementation

Any scheme proposed to solve the linear parabolic equation should be efficient in terms of computer storage and operations. For any finite element space V_h^p the matrices M and K are banded matrices, and thus an efficient method of solution should preserve and utilise this characteristic. In order to simplify the writing of the formulae let us consider the homogeneous problem (1.1). Defining

$U^n = \sum_{i=1}^{d} U_i^n V_i$ and $\underline{U}^n = (U_1^n, \dots, U_d^n)^T$ we have immediately from (3.5) and (3.6)

$$\sum_{j=o}^{k} \alpha_j M \underline{U}^{n+j} - \sum_{j=o}^{k} \sum_{r=1}^{m} \Delta_t^r \beta_{rj} M \underline{U}_{(r)}^{n+j} = \underline{0}$$

where

$$M \underline{U}_{(r)}^{n+j} = - K \underline{U}_{(r-1)}^{n+j} \qquad r = 1,2,\dots,m$$

By combining these two equations we achieve

$$\sum_{j=o}^{k} \alpha_j \underline{U}^{n+j} - \sum_{j=o}^{k} \sum_{r=1}^{m} \Delta_t^r \beta_{rj} (-1)^r (M^{-1}K)^r \underline{U}^{n+j} = \underline{0} \qquad (7.7)$$

The equation (7.7) is obviously impractical as it entails full matrices $M^{-1}K$, $(M^{-1}K)^2, \ldots, (M^{-1}K)^m$. However, by the use of complex arithmetic the sparseness of the matrices M and K is utilised. We illustrate this mode of implementation by reference to the family of equations (7.3). Equation (7.7) can be seen to be

$$\sum_{j=o}^{2} \tilde{\mu}_j (\Delta_t M^{-1}K) \underline{U}^{n+j} = \underline{0}$$

where $\tilde{\mu}_2(\tau) = \frac{\alpha}{\gamma} + \frac{\beta}{\gamma}\tau + \tau^2$, $\gamma = \frac{\beta}{3} - \frac{1}{360} - \frac{\alpha}{12}$

$$\tilde{\mu}_1(\tau) = \frac{(1-2\alpha)}{\gamma}\left[1 - \frac{8}{15(2\alpha-1)} + \frac{\left(-\frac{19}{360} + \frac{5\alpha}{6} - \frac{4\beta}{3}\right)\tau^2}{(2\alpha-1)}\right]$$

$$\tilde{\mu}_o(\tau) = \frac{\alpha-1}{\gamma}\left[1 - \frac{(7-15\beta)}{15(1-\alpha)}\tau + \frac{\left(\frac{5}{72} + \frac{\alpha}{12} - \frac{\beta}{3}\right)\tau^2}{1-\alpha}\right] \tag{7.8}$$

The roots of $\tilde{\mu}_2(\tau)$ are readily seen to be complex whenever α and β are admissible. Thus, let

$$\mu_2(\tau) = (z_2 - \tau)(\bar{z}_2 - \tau)$$

and further let $z_1^{(1)}$, $z_1^{(2)}$ and $z_o^{(1)}$, $z_o^{(2)}$ be respectively the roots of $\gamma\mu_1(\tau)/_{1-2\alpha}$ and $\gamma\mu_o(\tau)/_{\alpha - 1}$. Consequently, a simple manipulation shows that (7.8) is equivalent to

$$M\,\underline{U}^{n,1} = (M - z_o^{(1)}\Delta_t K)\,\underline{U}^{n}$$

$$M\,\underline{U}^{n,2} = (M - z_1^{(1)}\Delta_t K)\,\underline{U}^{n+1}$$

$$(z_2 M - \Delta_t K)\,\underline{U}^{n,3} = \left(\frac{2\alpha-1}{\gamma}\right)\left(M - z_o^{(2)}\Delta_t K\right)\underline{U}^{n,1}$$

$$+ \left(\frac{1-\alpha}{\gamma}\right) \left(M - z_1^{(2)} \Delta_t K \right) \underline{U}^{n,2}$$

$$\underline{U}^{n+2} = \frac{\text{Im } \underline{U}^{n,3}}{\text{Im } \bar{z}_2}$$

Although three intermediate steps are necessary at each time interval it is necessary to invert only two matrices. For the particular example (7.5d) only one intermediate step exists at each time interval, requiring the inversion of only one matrix. This fifth order scheme can be favourably, compared with the third order Padé (1,2). The latter may be implemented by

$$\frac{1}{6}(\alpha M + \Delta_t K) \underline{U}^{n,1} = (M - \frac{1}{3} \Delta_t K) \underline{U}^n$$

$$\underline{U}^{n+1} = \frac{\text{Im } \underline{U}^{n,1}}{\text{Im } \bar{\alpha}} \qquad \text{where} \quad \alpha = 2 + i\sqrt{2}$$

Both schemes require one intermediate step at each time interval and one matrix inversion, although the scheme (7.5d) necessitates a larger storage capacity and greater arithmetic operations per time interval. However, the latter disadvantage is easily compensated by the higher order and smaller error constant permitting larger time increments for comparable accuracy. Similarly, the schemes (7.6_1) and (7.6_{111}) have parallel modes of implementation to the Padé (1,3) and Padé (2,3) respectively. Consequently, arguing as before we can establish a preference for the schemes (7.6).

The use of complex arithmetic, and the extra storage necessary may be prohibitive. However, A-stable L.M.S.D. methods of arbitrary order have been investigated by several authors with the intention of simplifying the implementation. Of particular interest is the family of one-step Hermite formulae suggested by Makinson [19] and investigated fully by Norsett [26]. Norsett derived a family of A(O) - stable, one-step methods of order m+1 where the coefficient matrix, $G_m(M^{-1}K)$, of $\underline{U}^{n+1}$ is given by

$$G_m(M^{-1}K) \equiv (I + \frac{\Delta_t}{\gamma} M^{-1}K)^m, \text{ for a specified parameter } \gamma .$$

Continuing with the construction of L.M.S.D. methods with k = m = 2 we now establish a family of fourth order, A_o - stable methods where the coefficient matrix of $\underline{U}^{n+2}$ has the same characteristics as $G_2(M^{-1}K)$. The family of fourth order schemes with the above property is given by

$$\alpha y_{n+2} + (1-2\alpha)y_{n+1} + (\alpha-1)y_n = \Delta_t \{\beta\alpha y'_{n+2} + (\frac{1}{2} - \alpha + 4\beta\alpha - 3\beta^2\alpha)y'_{n+1} + (\alpha + \frac{1}{2} - 5\beta\alpha + 3\beta^2\alpha)y'_n\} + \Delta_t^2\{- \frac{\beta^2\alpha}{4} y''_{n+2} + (\frac{3\alpha}{2} - 4\beta\alpha + 2\beta^2\alpha - \frac{1}{12})y''_{n+1} + (\frac{\alpha}{2} + \frac{1}{12} - 2\beta\alpha + \frac{5}{4}\beta^2\alpha) y''_n \} \quad (7.9)$$

Applying the Routh-Hurwitz criterion we deduce that (7.9) is A_o-stable if for any $\alpha > \frac{1}{2}$

$$1 - \frac{\sqrt{12}}{6} < \beta < \min\left\{1 - \sqrt{\frac{1}{6\alpha}} , \frac{2}{3} - \frac{1}{6\alpha}\sqrt{4\alpha^2 - 2\alpha}\right\}$$

and $\alpha^2(3\beta^2 + 1 - 4\beta)^2 < (4\alpha - 2)(\beta^2\alpha - 2\beta\alpha + \alpha - \frac{1}{6})$

Alternatively, A_o - stability is ensured by $\alpha>\frac{1}{2}$ and

$$1 + \sqrt{\frac{1}{6\alpha}} < \beta < 1 + \frac{\sqrt{12}}{6}$$

The normalised error constant of the scheme (7.9) is expressed by

$$C_5 = \alpha\left(\frac{\beta}{6} - \frac{1}{24} - \frac{\beta^2}{8}\right) - \frac{1}{720}$$

As before we require that the choices of values for α and β yield a balance between the stability at infinity and the magnitude of the error constant. However, the A_o-stability requirement on β forces the modulus of the roots of $\sigma_2(\xi)$ to be extremely close to one for small values of C_5. The one important exception is when

$$\alpha = (5 + 16\sqrt{10})/90 \quad , \quad \beta = (12 - 2\sqrt{10})/13$$

ie. $$\left(\frac{5 + 16\sqrt{10}}{90}\right)y_{n+2} + \left(\frac{40 - 16\sqrt{10}}{45}\right)y_{n+1} + \left(\frac{16\sqrt{10} - 85}{90}\right)y_n =$$

$$\Delta_t\left\{\left(\frac{7\sqrt{10} - 10}{45}\right)y'_{n+2} + \left(\frac{40 - 4\sqrt{10}}{45}\right)y'_{n+1} + \left(\frac{5 - \sqrt{10}}{15}\right)y'_n\right\}$$

$$- \Delta_t^2\left(\frac{2\sqrt{10} - 5}{45}\right)y''_{n+2} \qquad (7.10)$$

and $C_5 = (4 - \sqrt{10})/270$

The scheme (7.10) has roots equal to zero at infinity. Its implementation is readily seen to be expressed by

$$(M + \frac{6 - \sqrt{10}}{13}\Delta_t K)\,\underline{U}^{n,1} = \left(\frac{112 - 88\sqrt{10}}{169}\right)\left(\frac{2 - \sqrt{10}}{3}M + \Delta_t K\right)\underline{U}^{n+1}$$

$$- \left(\frac{74 - 34\sqrt{10}}{169}\right)\left(\frac{53 + \sqrt{10}}{18}M - \Delta_t K\right)\underline{U}^{n}$$

$$(M + \frac{6 - \sqrt{10}}{13}\Delta_t K)\,\underline{U}^{n+2} = M\,\underline{U}^{n,1}$$

and requires the inversion of only one matrix. Similarly, the scheme (7.6_{111}) can be manipulated to exhibit the same characteristic, i.e. the polynomial $\mu_2(\tau)$ having a double root. This property is obtained by the value

$$c = 105\,(4\sqrt{2} - 3)/_{1127}$$

which yields a sixth order scheme.

We conclude this chapter with the following remarks:

(1) We conjecture that the maximum order of an A_o-stable L.M.S.D. scheme which is stable at infinity is

$$q = m(k + 1) - 1$$

Thus it is advisable to select m > 1 for the derivation of high order schemes.

(2) A clear advantage in increasing m rather than k results from the error constant decreasing more rapidly for m increasing than with k increasing, particularly if considered in conjunction with the rate of convergence at infinity. For example the third order schemes (7.2_{11}) have small error constants for $\gamma = 1$

$$\text{eg.} \quad \gamma = 1 , \quad \alpha = \frac{1}{4} , \quad \beta = \frac{1}{2} , \quad \text{whence } C_4 = {}^{-1}/96.$$

but these schemes are not stable at infinity. For the optimal rate of convergence at infinity, ie. when $\sigma_1(\xi) \equiv \xi^3$, the values are

$$\gamma = \frac{8}{9} , \quad \alpha = \beta = 3 , \quad \text{whence } C_4 = -\frac{1}{4}.$$

In comparison the third order Padé (1,2), with its equivalent, optimal rate of convergence at infinity, commands

$$C_4 = \frac{1}{72}$$

(3) With respect to the system of equations (6.3), maximum order, A_o - stable L.M.S.D. schemes, with m > 1, invariably require complex

arithmetic for their implementation. Ease of implementation, as characterised by (7.9), may only be obtained by relaxing the stipulation of maximum order. However, once this relaxation is operative we can derive high order A_0 - stable L.M.S.D.'s. that are simple to implement. We conjecture that schemes of order $q = mk$ can possess this property.

Given m fixed, let us compare the implementation procedures of two L.M.S.D.'s whose step numbers differ by one. Schemes incorporating $m > 1$ require the first $m - 1$ derivatives of $(f(x,t),V_i)$ at each time level. Thus for a k-step method these evaluations may be utilised $k + 1$ times which does not necessitate any further evaluations as k increases, although it requires a minor increase in storage capacity.

The number of intermediate stages at each time level employed by the complex arithmetic mode of implementation is readily seen to be equal to

$$\sum_{j=o}^{k} \left(\tilde{r}_j - 1 \right), \text{where } \tilde{r}_j = \sup \{ r \mid \beta_{rj} \neq 0 \; , \quad r = 1,\dots,m\}$$

for the homogeneous problem with, generally, an additional $m - 1$ intermediate stages for the non-homogeneous problem. Consequently, by increasing k by one but restricting the coefficients $\{\beta_{r,\, k+1}\}_{r=2}^{m}$ to be zero entails no additional intermediate stage, although the number of arithmetic operations and the required storage capacity will be increased by O(d). For example consider the schemes (7.2_1) (7.5d), and (7.6_{111}). All require one intermediate stage per time level but their orders are respectively 3, 5 and 7. The higher order, permitting larger time increments Δ_t, will easily compensate for the increase in arithmetic operations and storage. Finally, we note that within certain classes of schemes optimal stability at infinity is compatible with ease of implementation, i.e. (7.5).

With regard to the above remarks we advance the merits of the classes of L.M.S.D. schemes with $m = k - 1$, k, or $k + 1$ for $k \geq 2$. Such schemes may incorporate a balance between the above remarks.

8. Numerical Results

Complementing Zlámal [43] , and others, the following two test problems are studied in detail

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \quad , \qquad x \in (0,1) \ , \quad t > 0$$

$$u(0,t) = u(1,t) = 0 \quad , \qquad t \geq 0 \tag{8.1}$$

$$u(x,0) = g(x) \quad , \qquad x \in (0,1)$$

where g(x) is given respectively by

(1) $\qquad g(x) = 1 \quad , \qquad 0 < x < 1$

(2) $\qquad g(x) = \begin{cases} 2x \ , & 0 < x < \frac{1}{2} \\ 2(1-x), & \frac{1}{2} \leq x < 1 \end{cases}$

The analytic solution of (8.1) is given by $u(x,t) = \sum_{i=1}^{\infty} g_i \, e^{-\lambda_i t} \sin\sqrt{\lambda_i} x$ where $\lambda_i = \pi^2 i^2$ are the eigenvalues of $y'' = -\lambda y$, $y(o) = y(1) = 0$, and $\{g_i\}_{i=1}^{\infty}$ are the Fourier coefficients of the initial value g(x). The continuous time Galerkin solution has a similar form; ie. let $U(x,t) = \sum_{i=1}^{d} Y_i(t) \, \Psi_i(x)$, where as before $\Psi_i(x)$ are the eigenfunctions of the eigenvalue problem $a(\Psi,V) = \Lambda(\Psi,V)$, $\forall \ V \in V_h^p$. Consequently, it is easy to see that the equation (2.3) decomposes into the set of equations $Y_i' = - \Lambda_i \, Y_i$, $Y_i(0) = U_i^o$.

If the exact solution of (8.1) is smooth at t = 0 the Fourier coefficients g_i converge rapidly to zero as $i \to \infty$. Consequently, we expect the coefficients U_i^o to converge rapidly to zero as $i \to d$, whenever h is sufficiently small. Thus we anticipate that for g(x) very smooth a L.M.S.D. method with an error constant close to the minimal value should yield more accurate results than a scheme with a larger error constant but with improved stability at infinity. Conversely, for g(x) not smooth we anticipate a preference for a L.M.S.D. method with a near optimal rate of convergence at infinity.

The following results were derived by selecting $V_h^p \equiv \tilde{V}_h^3$, the space of cubic splines over [0,1], with a regular mesh of interval $h = 0.1$. The time increment $\Delta_t = 0.01$, and the initial value U(x,0) is taken to be the L_2 - projection of g(x) onto $\tilde{V}_h^3$. The Padé (1,2) or Padé (1,3) are used to determine the values $\{U^n\}_{n=1}^{k-1}$.

Problem 1

The Fourier coefficients $g_i = 2[1 - (-1)^i]/_{\pi i}$ converge very slowly to zero and consequently we expect to employ L.M.S.D. schemes with high rates of convergence at infinity. Numerical results have been obtained by the four multistep schemes (7.5). The Padé (1,3) has been used to evaluate the extra initial values; not to preserve the fifth order of convergence in Δ_t but because it can be shown to be more appropriate for small values of t. As the g_i's converge slowly to zero as $i \to \infty$ we are concerned for small t with components of the solution for which $\tau = \Delta_t \Lambda_i$ is large. Applying the Padé (1,2) and the Padé (1.3) to the test equation $Y' = \Lambda Y$, $Y(0) = 1$, where $\Lambda >> 0$, it is easy to see from (7.2) that:-

Padé (1,2) $$Y_{n+1} = \left(\frac{1 - \tau/3}{1 + 2\tau_{/3} + \tau^2_{/6}} \right) Y_n$$

$$\text{i.e.} \quad Y_{n+1} \simeq -\frac{2}{\tau} Y_n \quad \text{as } \tau \to \infty.$$

Padé (1,3) $$Y_{n+1} = \left(\frac{1 - {}^{\tau}/4}{1 + 3\tau_{/4} + \tau^2_{/4} + \tau^3_{/24}} \right) Y_n$$

$$\text{i.e.} \quad Y_{n+1} \simeq -\frac{6}{\tau^2} Y_n \qquad \text{as} \quad \tau \to \infty$$

whilst the L.M.S.D. method (7.5d) yields

$$\left(\frac{23}{30} + \frac{2}{5}\tau + \frac{1}{15}\tau^2 \right) Y_{n+2} + \left(-\frac{8}{15} + \frac{8}{15}\tau \right) Y_{n+1} + \left(-\frac{7}{30} + \frac{1}{15}\tau \right) Y_n = 0$$

$$\text{i.e.} \quad \left| Y_{n+2} \right| \simeq \frac{1}{\tau^2} \; Y_{n+1} \qquad \text{as} \quad \tau \to \infty$$

Since we wish to simulate the exponential decay $Y_{n+1} = e^{-\tau} Y_n$ the Padé (1,3) is preferred.

In the subsequent tables percentage errors are evaluated at the knots x = ih, h = 0.1, i = 1,2,..,5, for various time levels. Table 1 compares the Padé (1,α) , α = 2 and 3, for 0.01 ≤ t ≤ 0.1.

Table 1 Percentage errors. $\Delta_t = 0.01$

t	x=.1		x=.2		x=.3		x=.4		x=.5	
	α=2	α=3	α=2	α=3	α=2	α=3	α=2	α=3	α=2	α=3
0.01	1.83	.622	1.23	.477	.058	.082	.155	.035	.052	.012
0.02	.573	.092	.252	.007	.200	.043	.042	.001	.045	.026
0.03	.189	.023	.084	.004	.017	.004	.046	.004	.051	.000
0.05	.058	.005	.035	.002	.008	.000	.015	.002	.023	.003
0.1	.002	.000	.001	.000	.000	.000	.001	.000	.001	.000

Allocating the initial values at $t = 0$ and $t = \Delta_t$ to be respectively the L_2 - projection of $g(x)$ and the Padé (1.3) at $t = \Delta_t$, the problem has been computed by the four schemes (7.5). As expected the results improve as the respective roots of $\sigma_2(\xi)$ tend to zero. Table 2 gives the results at various time levels for the schemes (7.5b) and (7.5d).

As t increases, say $t \geq 0.08$, the exact solution and its first six derivatives become increasingly dominated by the lower values of i. Correspondingly, the maximum significant value of $\tau = \Delta_t \Lambda_i$ decreases. Thus with the discretization error becoming increasingly dominated by the lower values of i the system of equations tends to loose its stiffness and behaves as a non-stiff system. Let us use the L.M.S.D. schemes with a larger time increment, say $\Delta_t = 0.02$ (0.03), for $t \geq 0.08$ (0.09) with the initial six values from the Padé (1,3), $\Delta_t = 0.01$. It is to be expected that for $\Delta_t = 0.02$ the lack of high stability at infinity of the schemes (7.5a) - (7.5c) will be partly compensated by their smaller error constants compared with the scheme (7.5d). Numerically, this analysis is validated. For $\Delta_t = 0.02$, the schemes (7.5b) - (7.5d) yield extremely similar values whilst for $\Delta_t = 0.03$ (7.5d) is seen to be superior. Table 3 compares the schemes (7.5b) and (7.5d) for $\Delta_t = 0.02$. Table 4 illustrates the scheme (7.5d) with $\Delta_t = 0.03$.

Table 4. Percentage errors x 10^2 , $\Delta_t = 0.03$

t	x=0.1	x=0.2	x=0.3	x=0.4	x=0.5
0.09	11.18	6.714	1.439	2.659	4.184
0.12	4.826	2.924	0.669	1.039	1.668
0.15	3.023	0.782	0.352	0.756	1.164
0.27	0.352	0.187	0.021	0.117	0.166
0.39	0.052	0.012	0.003	0.021	0.027

Excellent results may similarly be obtained by the scheme (7.10). Tables 5 and 6 give the percentage errors when the scheme (7.10) is applied with respectively $\Delta_t = 0.02$, $\Delta_t = 0.03$ after the initial six (twelve) values were derived by the Padé (1,3), $\Delta_t = 0.01$, and the L_2 - projection of g(x).

Problem 2

The exact solution is not smooth at t = 0, but it is smoother than the previous example. The Fourier coefficients are given by $g_i = \frac{8}{\pi^2 i^2} \sin \frac{i\pi}{2}$, $i \geq 1$. The remarks concerning problem 1 are also seen to be appropriate to this example. Assuming that the initial values are derived as in the preceding example the numerical results show that

1). For $t \geq 0.02$, $\Delta_t = 0.01$, the results from the schemes (7.5b) - (7.5d) are barely distinguishable from each other.

2). For $t \geq 0.08$, $\Delta_t = 0.02$, the schemes, (7.5a) - (7.5d) are similar, with (7.5b) - (7.5d) almost identical.

3). For $t \geq 0.09$, $\Delta_t = 0.03$, the schemes (7.5b) - (7.5d) yield similar results, with (7.5c) and (7.5d) identical.

These numerical observations are anticipated as the solution is smoother than that of problem 1. Tables 7 and 8 illustrate the schemes (7.5b) and (7.5d) when, respectively, $\Delta_t = 0.02$, $\Delta_t = 0.03$, and the initial six values are derived by the Padé (1.3), $\Delta_t = 0.01$, and the L_2 - projection of g(x).

The numerical evidence and analysis for the homogeneous equation warrant the following conclusions. Firstly, the accuracy associated with the high order methods does not appear in the first few time steps. This coincides with the validity of the convergence results of theorems 1 and 2.

We have seen that extremely good results may be obtained by applying the Padé (1,2) or Padé (1,3) in the initial phrase and a high order L.M.S.D. scheme, with a larger time increment, in the further phase. The extent of the initial phase is determined by the smoothness of g(x), since g(x) controls the smoothness of u(x,t) for small t. For g(x), and hence U^o smooth, the discretization error is rapidly dominated by the first few values of i which enable us to progress rapidly to larger time increments; conversely for g(x) non-smooth.

The situation for the non-homogeneous equation is more complex. Corresponding to the homogeneous equation, the numerical results of a particular scheme will reflect the relationship between the smoothness of g(x) and the schemes stability at infinity. The discretization error at time t is now dependent on the quantities

$$\left\{ \Lambda_i^{p-q-1} \left(\frac{\partial^p}{\partial t^p} f(x,t), \Psi_i \right) \right\}_{p=0}^{q+1} , \qquad i = 1,2,\ldots,d ,$$

and thus, generally, it is ill-advised to employ a larger time increment in the further phase.

Table 2 Percentage errors $\Delta_t = 0.01$

t	x=0.1		x=0.2		x=0.3		x=0.4		x=0.5	
	(7.5c)	(7.5d)	(7.5c)	(7.5d)	(7.5c)	(7.5d)	(7.5c)	(7.5d)	(7.5c)	(7.5d)
0.02	3.506	0.585	1.103	0.059	0.501	0.010	0.321	0.031	0.279	0.021
0.03	1.374	0.189	0.315	0.129	0.129	0.096	0.068	0.013	0.046	0.018
0.04	0.912	0.059	0.209	0.112	0.082	0.058	0.052	0.014	0.038	0.005
0.10	0.012	0.002	0.005	0.003	0.004	0.001	0.001	0.001	0.001	0.002
0.20	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
0.30	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
0.40	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001

Table 3 Percentage Errors x 10^3 $\Delta_t = 0.02$

t	x=0.1		x=0.2		x=0.3		x=0.4		x=0.5	
	(7.5b)	(7.5d)	(7.5b)	(7.5d)	(7.5b)	(7.5d)	(7.5b)	(7.5d)	(7.5b)	(7.5d)
0.08	0.18	11.1	0.31	6.33	0.61	0.97	0.99	3.36	1.14	4.98
0.10	0.07	1.24	0.50	1.13	0.58	0.68	0.66	0.44	0.66	0.35
0.12	0.60	0.98	0.76	0.99	0.69	0.67	0.77	0.50	0.79	0.41
0.20	0.59	0.33	0.75	0.59	0.67	0.57	0.71	0.63	0.72	0.63
0.30	0.41	0.39	0.69	0.62	0.65	0.56	0.66	0.58	0.63	0.57
0.40	0.51	0.36	0.70	0.60	0.63	0.54	0.66	0.56	0.66	0.55

Table 5 Percentage Errors x 10^2 Δ_t=0.02

t	x=0.1		x=0.2		x=0.3		x=0.4		x=0.5	
	'6'	'12'	'6'	'12'	'6'	'12'	'6'	'12'	'6'	'12'
0.10	3.696	0.042	2.283	0.009	0.545	0.004	0.815	0.007	1.327	0.009
0.12	4.571	0.051	2.755	0.026	0.598	0.029	1.108	0.025	1.748	0.025
0.14	4.685	0.090	2.875	0.048	0.664	0.031	1.065	0.010	1.715	0.004
0.20	5.642	0.039	3.402	0.031	0.763	0.056	1.298	0.069	2.065	0.076
0.30	7.483	0.177	4.475	0.134	0.922	0.114	1.779	0.092	2.774	0.085
0.40	10.33	0.102	6.106	0.106	1.332	0.146	2.195	0.171	3.461	0.182

Table 6 Percentage Errors x 10 Δ_t=0.03

t	x=0.1		x=0.2		x=0.3		x=0.4		x=0.5	
	'6'	'12'	'6'	'12'	'6'	'12'	'6'	'12'	'6'	'12'
0.09	4.656	0.005	2.826	0.001	0.637	0.001	1.073	0.003	1.712	0.003
0.12	5.174	0.005	3.123	0.003	0.683	0.003	1.205	0.003	1.908	0.003
0.15	6.602	0.055	4.000	0.037	0.904	0.019	1.501	0.003	2.396	0.003
0.27	16.34	0.133	9.954	0.096	2.286	0.055	3.735	0.021	5.991	0.008
0.39	40.47	0.294	24.75	0.207	5.717	0.104	9.344	0.019	15.02	0.013

Table 7 Percentage Errors x 10^3 $\Delta_t = 0.02$

t	x=0.1		x=0.2		x=0.3		x=0.4		x=0.5	
	(7.5b)	(7.5d)	(7.5b)	(7.5d)	(7.5b)	(7.5d)	(7.5b)	(7.5d)	(7.5b)	(7.5d)
0.08	0.91	4.27	1.00	3.06	0.80	1.21	0.72	0.23	0.68	0.80
0.10	0.48	0.20	0.69	0.51	0.60	0.56	0.58	0.68	0.56	0.71
0.12	0.60	1.05	0.80	1.06	0.73	0.70	0.75	0.48	0.74	0.38
0.20	0.53	0.44	0.75	0.66	0.68	0.59	0.71	0.60	0.70	0.59
0.30	0.46	0.39	0.70	0.62	0.64	0.56	0.66	0.58	0.64	0.57
0.40	0.48	0.36	0.70	0.60	0.64	0.53	0.66	0.56	0.65	0.55

Table 8 Percentage Errors x 10^3 $\Delta_t=0.03$

t	x=0.1		x=0.2		x=0.3		x=0.4		x=0.5	
	(7.5b)	(7.5d)	(7.5b)	(7.5d)	(7.5b)	(7.5d)	(7.5b)	(7.5d)	(7.5b)	(7.5d)
0.09	1.84	36.6	1.16	23.34	0.05	6.19	1.75	7.97	2.48	13.47
0.12	2.65	14.8	2.50	8.96	1.30	1.84	0.16	4.20	0.85	6.54
0.15	1.77	10.1	1.09	6.70	0.05	2.02	1.64	1.81	2.32	3.32
0.27	2.30	1.31	1.47	1.10	1.39	0.47	0.36	0.03	1.16	0.16
0.39	2.90	0.06	1.88	0.24	0.28	0.11	1.66	0.08	2.51	0.04

9. The 'Quasi-Linear' Parabolic Equation

Using the notation of the previous chapters we shall investigate the numerical solution of the initial, boundary value problem

$$\rho(x)\frac{\partial u}{\partial t} - \sum_{i,j=1}^{N} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right) = f(x,t,u) + \operatorname{div} \underline{b}(x,t,u) \quad x \in \Omega,\ t > 0$$

$$u(x,0) = g(x) \qquad x \in \Omega \qquad (9.1)$$

$$u(x,t) = 0 \qquad x \in \Gamma\ ,\ t \geq 0$$

where $\underline{b}(x,t,u) \equiv (b_1(x,t,u),\dots,\ b_N(x,t,u))^T$

The weak solution is readily seen to satisfy

$$(\rho(x)\frac{\partial u}{\partial t},v) + a(u,v) = (f(x,t,u),v) - (\underline{b}(x,t,u)\ .\nabla v) \quad \forall v \in H_o^1(\Omega),\ t > 0$$

$$u(x,0) = g(x) \qquad (9.2)$$

where the bilinear functional $a(\cdot,\cdot)$ is now given by

$$a(u,v) = \sum_{i,j=1}^{N} \int_\Omega a_{ij}(x)\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i}\,dx$$

Defining the subspace V_h^p of $H_o^1\ (\Omega)$ as before, the continuous time Galerkin solution, $U(x,t)$, satisfies

$$(\rho(x)\frac{\partial U}{\partial t},\ V) + a(U,V) = (f(x,t,U),V) - (\underline{b}(x,t,U)\ .\ \nabla V) \quad \forall V \in V_h^p\ ,\ t > 0$$

$$U(x,0) = U^o(x) \qquad (9.3)$$

for a suitable approximation, $U^o(x)$, to $g(x)$.

Let $U(x,t) = \sum_{j=1}^{d} C_j(t)V_j$ and substitute $\{V_i\}_{i=1}^{d}$ in turn for V in (9.3). Then by assembling in matrix form we derive the initial value problem (cf. (2.3))

$$M \frac{d}{dt}\underline{C} + K\underline{C} = \underline{f}\,(\underline{C}) \quad , \quad \underline{C}(0) = \underline{a} \tag{9.4}$$

where the matrices M,K and the vector $\underline{f}\,(\underline{C})$ are given by

$$M_{ij} = (\rho(x)V_i \,,\, V_j) \quad , \quad K_{ij} = a(V_i \,,\, V_j) \quad , \; 1 \le i,j \le d$$

$$f_i(\underline{C}) = \left(f(x,t, \sum_{j=1}^{d} C_j(t)\,V_j) \,,\, V_i\right) - \left(\underline{b}(x,t, \sum_{j=1}^{d} C_j(t)\,V_j).\nabla V_i\right)$$

$$i=1,\ldots,d$$

The positive definiteness of M and K is ensured by the assumptions (B) defined overleaf. Approximating the solution of (9.4) by utilising a L.M.S.D. scheme with $m > 1$ is ill-advised, since it necessitates the differentiation of the non-linear term $\underline{f}(\underline{C})$. However, by an indirect application of a 3rd order, A_o-stable L.M.S.D. scheme it is possible to obtain an approximate solution to (9.4) by solving algebraically linear systems of equations at each time step. This particular 3rd order L.M.S.D. scheme is given by

$$\sum_{j=o}^{2} \alpha_j\, y_{n+j} = \Delta_t \sum_{j=o}^{2} \beta_j\, y'_{n+j} + c_2\Delta_t^2\, y''_{n+2}$$

where

$$\alpha_2 = \frac{1}{2} + \frac{\sqrt{3}}{3} \quad , \quad \beta_2 = \frac{1+\sqrt{3}}{6} + \frac{3\theta}{2} \quad , \quad c_2 = -\theta$$

$$\alpha_1 = -\frac{2\sqrt{3}}{3} \quad , \quad \beta_1 = \frac{2}{3} - 2\theta \tag{9.5}$$

$$\alpha_0 = \frac{\sqrt{3}}{3} - \frac{1}{2} \quad , \quad \beta_0 = \frac{1-\sqrt{3}}{6} + \frac{\theta}{2}$$

The above scheme is consistent and zero-stable. A_o-stability is ensured by the conditions

$$\theta > 0 \quad , \quad \frac{4\sqrt{3}}{3} + \tau(4\theta - \frac{1}{3}) + \theta\tau^2 > 0 \quad \text{for all } \tau > 0.$$

Both A_o-stability conditions hold a fortiori for $\theta > \frac{1}{12}$.

Consequently, the discrete time Galerkin solution, where U^n is an approximant to $U(x,n\Delta_t)$, is defined by

$$\sum_{j=o}^{2} \frac{\alpha_j}{\Delta_t} (\rho(x)U^{n+j}, V) + a\,(\sum_{j=o}^{2} \beta_j U^{n+j} + c_2\Delta_t Q^{n+2}, V) \tag{9.6$_1$}$$

$$= (f(x, \bar{t}^n, \bar{U}^n), V) - (\underline{b}(x,\bar{t}^n, \bar{U}^n).\nabla V) \qquad \forall V \in V_h^p \ , \quad n \geq 1$$

where $\bar{t}^n \equiv (n+1+ \frac{\sqrt{3}}{3})\Delta_t$,

$$\bar{U}^n \equiv \frac{1+\sqrt{3}}{6} U^{n-1} - \frac{1+2\sqrt{3}}{3} U^n + \frac{7+3\sqrt{3}}{6} U^{n+1}$$

and $Q^{n+2} \in V_h^p$ is defined by

$$(\rho(x)Q^{n+2}, V) + a(U^{n+2}, V) = \Big(f(x,(n+2)\Delta_t, \tilde{U}^{n+2}), V\Big)$$

$$- \Big(\underline{b}(x,(n+2)\Delta_t, \tilde{U}^{n+2}).\nabla V\Big) \qquad \forall V \in V_h^p \tag{9.6$_{11}$}$$

where $\tilde{U}^{n+2} = 3U^{n+1} - 3U^n + U^{n-1}$

For $\theta > \frac{1}{12}$ the scheme is unconditionally stable and, c.f. Theorem 4, third order accurate in Δ_t. Computational aspects related to the system (9.6) will be investigated at a later stage.

We shall impose the following assumptions (B)

(B_1) $u(x,t) \;,\; \frac{\partial}{\partial t} u(x,t) \;\epsilon\; H^{p+1}(\Omega) \;,\; t \in [0,T]$

(B_{11}) $u \in C^4(\bar{\Omega} \times [0,T])$ ie. four times continuously differentiable with respect to t, $(x,t) \in (\bar{\Omega} \times [0,T])$.

(B_{111}) the function $\rho(x)$ shall be bounded above and below by positive constants

ie. $0 < \eta \le \rho(x) \le C \;,\; x \in \Omega$

(B_{1v}) the matrix A, $A_{ij} = a_{ij}(x)$, is uniformly positive definite.

ie. $a_{ij}(x) = a_{ji}(x) \;,\; 1 \le i,j \le N$, and there exists a constant $C_o > 0$ such that

$$C_o^{-1} \sum_{i=1}^{N} \xi_i^2 \le \sum_{i,j=1}^{N} a_{ij}(x)\xi_1\xi_j \le C_o \sum_{i=1}^{N} \xi_i^2 \qquad \forall x \in \Omega$$

Further let the derivatives $\{\frac{\partial}{\partial x_i} a_{ij}(x)\}$ be bounded

(B_v) The functions f and b_i are uniformly Lipschitz continuous with respect to u

$$\text{ie. } \sum_{i=1}^{N} |b_i(x,t,u_1) - b_i(x,t,u_2)| + |f(x,t,u_1) - f(x,t,u_2)|$$

$$\le L\,|u_1 - u_2| \qquad (x,t) \in \Omega \times [0,T], \; -\infty < u_1,u_2 < \infty$$

To preserve continuity in a later proof we shall establish three results now. Let $\xi(x,t) \in H_o^1(\Omega)$ for each $t > 0$, and further denote $\xi^n \equiv \xi(x,n\Delta_t)$.

Lemma 3

Given $\theta > \frac{1}{12}$, then there exist positive constants c and C such that

$$\sum_{n=1}^{m} \left(\sum_{j=o}^{2} \rho(x)\, \alpha_j \xi^{n+j} \;,\; \sum_{j=o}^{2} \beta_j \xi^{n+j} \right) \geq c \,\|\xi^{m+2}\|^2 - C\left(\|\xi^1\|^2 + \|\xi^2\|^2 \right)$$

Proof

$$S^n \equiv \sum_{j=o}^{2} \alpha_j \xi^{n+j} \cdot \sum_{j=o}^{2} \beta_j \xi^{n+j}$$

$$= \left(\frac{\sqrt{3}}{6} - \frac{1}{8}\right)\left(2\theta - \frac{1}{6}\right)\left(\xi^{n+2} - 2\xi^{n+1} + \xi^n\right)^2 + \frac{1}{48}\left(\xi^{n+2} - \xi^n\right)^2 +$$

$$\left(\frac{\theta}{2} - \frac{1}{24}\right)\left((\xi^{n+2} - \xi^{n+1})^2 + (\xi^{n+1} - \xi^n)^2\right) - \theta\left(1+\frac{\sqrt{3}}{3}\right)\left(\xi^{n+2}\xi^{n+1} - \xi^{n+1}\xi^n\right)$$

$$+ \left(\frac{\theta}{2} + \frac{\theta\sqrt{3}}{6} + \frac{\sqrt{3}}{6} + \frac{1}{4}\right)\left(\xi^{n+2}\right)^2 - \frac{\sqrt{3}}{3}\left(\xi^{n+1}\right)^2 + \left(\frac{\sqrt{3}}{6} - \frac{\theta}{2} - \frac{\theta\sqrt{3}}{6} - \frac{1}{4}\right)\left(\xi^n\right)^2 \tag{9.7}$$

Thus there exist positive constants c and C such that

$$\sum_{n=1}^{m} S^n \geq c \left[\sum_{n=1}^{m} \left(\xi^{n+2} - 2\xi^{n+1} + \xi^n\right)^2 + \sum_{n=1}^{m} \left(\xi^{n+2} - \xi^n\right)^2 + \right.$$

$$\left. \sum_{n=1}^{m+1} \left(\xi^{n+1} - \xi^n\right)^2 \right] - \theta\left(1 + \frac{\sqrt{3}}{3}\right)\left(\xi^{m+2}\xi^{m+1} - \xi^2\xi^1\right)$$

$$+ \left(\frac{\theta}{2} + \theta\frac{\sqrt{3}}{6} + \frac{\sqrt{3}}{6} + \frac{1}{4}\right)\left(\xi^{m+2}\right)^2 + \left(\frac{\theta}{2} + \theta\frac{\sqrt{3}}{6} - \frac{\sqrt{3}}{6} + \frac{1}{4}\right)\left(\xi^{m+1}\right)^2 - C\left[\left(\xi^2\right)^2 + \left(\xi^1\right)^2\right]$$

Note that for $\theta > \frac{1}{12}$

$$D \equiv \left(\frac{\theta}{2} + \frac{\theta\sqrt{3}}{6} - \frac{\sqrt{3}}{6} + \frac{1}{4} \right) > 0$$

and thus using the inequality $|ab| \le \varepsilon a^2 + b^2/_{4\varepsilon}$, $\varepsilon > 0$, we find that

$$\left(\frac{\theta}{2} + \theta\frac{\sqrt{3}}{6} + \frac{\sqrt{3}}{6} + \frac{1}{4}\right)\left(\xi^{m+2}\right)^2 + \left(\frac{\theta}{2} + \theta\frac{\sqrt{3}}{6} - \frac{\sqrt{3}}{6} + \frac{1}{4}\right)\left(\xi^{m+1}\right)^2 - \theta\left(1 + \frac{\sqrt{3}}{3}\right)\xi^{m+2}\xi^{m+1}$$

$$\ge \left[\frac{\theta}{2} + \theta\frac{\sqrt{3}}{6} + \frac{\sqrt{3}}{6} + \frac{1}{4} - \theta^2\left(1 + \frac{\sqrt{3}}{3}\right)^2\Big/4D\right]\left(\xi^{m+2}\right)^2$$

$$= \frac{1}{4D}\left[\theta + \theta\frac{\sqrt{3}}{3} - \frac{1}{12}\right]\left(\xi^{m+2}\right)^2$$

Combining the above two inequalities we have shown that

$$\sum_{n=1}^{m} S^n \ge c\left[\left(\xi^{m+2}\right)^2 + \sum_{n=1}^{m}\left(\xi^{n+2} - 2\xi^{n+1} + \xi^n\right)^2 + \sum_{n=1}^{m}\left(\xi^{n+2} - \xi^n\right)^2 + \sum_{n=1}^{m+1}\left(\xi^{n+1} - \xi^n\right)^2\right] - C\left[\left(\xi^2\right)^2 + \left(\xi^1\right)^2\right] \qquad (9.8)$$

The desired result is obtained by multiplying (9.8) by $\rho(x) > 0$ and integrating over Ω.

Generally, lemma 3 can not be weakened to include any $\theta \le \frac{1}{12}$. For instance, if $\theta \le \frac{1}{12}$, $\xi^n \equiv (-1)^n \varepsilon(x)$, $n > 2$, where $\varepsilon(x) > 0$ for any $x \in \Omega$, and $\xi^1 = \xi^2 \equiv 0$ then $S^n \le 0$ which is contradictory to the lemma.

Lemma 4

Given $\theta > 0$, then there exist positive constants c, C and μ such that

$$- \sum_{n=1}^{m} \left[a\left(\sum_{j=o}^{2} \alpha_j \xi^{n+j} \;, c_2 \xi^{n+2} \right) + \mu \left\| \sum_{j=o}^{2} \alpha_j \xi^{n+j} \right\|_1^2 \right] \geq c \left\| \xi^{m+2} \right\|_1^2$$

$$- C \left\{ \| \xi^1 \|_1^2 + \| \xi^2 \|_1^2 \right\}$$

Proof

$$T^n = - a\left(\sum_{j=o}^{2} \alpha_j \xi^{n+j} \;, \; c_2 \xi^{n+2} \right)$$

$$= \theta\left(\frac{\sqrt{3}}{6} - \frac{1}{4} \right) a\left(\xi^{n+2} - 2\xi^{n+1} + \xi^n \;, \; \xi^{n+2} - 2\xi^{n+1} + \xi^n \right)$$

$$+ \theta\left(\frac{1}{4} - \frac{\sqrt{3}}{12} \right) a\left(\xi^{n+2} - \xi^{n+1} \;, \; \xi^{n+2} - \xi^{n+1} \right)$$

$$+ \theta\left(\frac{3}{4} - \frac{\sqrt{3}}{4} \right) a\left(\xi^{n+1} - \xi^n \;, \; \xi^{n+1} - \xi^n \right) - \theta\left(\frac{1}{2} + \frac{\sqrt{3}}{6} \right) a\left(\xi^{n+2} - \xi^n \;, \; \xi^{n+1} \right)$$

$$+ \theta\left(\frac{1}{2} + \frac{\sqrt{3}}{4} \right) a\left(\xi^{n+2} \;, \; \xi^{n+2} \right) - \theta \frac{\sqrt{3}}{3} \, a\left(\xi^{n+1} \;, \; \xi^{n+1} \right)$$

$$+ \theta\left(\frac{\sqrt{3}}{12} - \frac{1}{2} \right) a\left(\xi^n \;, \; \xi^n \right)$$

Summing T^n for $n = 1,\ldots,m$ we derive by (B_{1v})

$$\frac{1}{\theta} \sum_{n=1}^{m} T^n \geq C_o^{-1} \left(\frac{\sqrt{3}}{6} - \frac{1}{4} \right) \sum_{n=1}^{m} \| \xi^{n+2} - 2\xi^{n+1} + \xi^n \|_1^2$$

$$+ C_o^{-1} \left(\frac{3}{4} - \frac{\sqrt{3}}{4} \right) \sum_{n=1}^{m} \| \xi^{n+1} - \xi^n \|_1^2 + C_o^{-1} \left(\frac{1}{4} - \frac{\sqrt{3}}{12} \right) \sum_{n=1}^{m} \| \xi^{n+2} - \xi^{n+1} \|_1^2$$

$$+ \left(\frac{1}{2} + \frac{\sqrt{3}}{4} \right) a\left(\xi^{m+2} \;, \; \xi^{m+2} \right) + \left(\frac{1}{2} - \frac{\sqrt{3}}{12} \right) a\left(\xi^{m+1} \;, \; \xi^{m+1} \right)$$

$$- \left(\frac{1}{2} + \frac{\sqrt{3}}{6} \right) \left\{ a\left(\xi^{m+2} \;, \; \xi^{m+1} \right) - a\left(\xi^2 \;, \; \xi^1 \right) \right\}$$

$$- c\left\{ \| \xi^2\|_1^2 + \| \xi^1\|_1^2 \right\} \tag{9.9}$$

In the following we use the Cauchy-Schwartz inequality for positive definite forms, namely

$$\sum_{i,j=1}^{N} a_{ij}\xi^i\gamma^j \le \left(\sum_{i,j=1}^{N} a_{ij}\xi^i\xi^j \right)^{\frac{1}{2}} \left(\sum_{i,j=1}^{N} a_{ij}\gamma^i\gamma^j \right)^{\frac{1}{2}}$$

and thus by the Cauchy-Schwartz inequality for integrals

$$\int_\Omega \sum_{i,j=1}^{N} a_{ij}\xi^i\gamma^j dx \le \left(\int_\Omega \sum_{i,j=1}^{N} a_{ij}\xi^i\xi^j dx \right)^{\frac{1}{2}} \left(\int_\Omega \sum_{i,j=1}^{N} a_{ij}\gamma^i\gamma^j dx \right)^{\frac{1}{2}}$$

It now follows by using the inequality $|ab| \le \varepsilon a^2 + b^2/4\varepsilon$, $\varepsilon > 0$,

$$a\left(\xi^{m+2}, \xi^{m+1}\right) \le \varepsilon\, a\left(\xi^{m+1}, \xi^{m+1}\right) + \frac{1}{4\varepsilon}\, a\left(\xi^{m+2}, \xi^{m+2}\right) .$$

Selecting $\varepsilon = \left(\frac{1}{2} - \frac{\sqrt{3}}{12}\right) \Big/ \left(\frac{1}{2} + \frac{\sqrt{3}}{6}\right)$ we easily deduce that

$$\begin{aligned} &\left(\frac{1}{2} + \frac{\sqrt{3}}{4}\right) a\left(\xi^{m+2}, \xi^{m+2}\right) + \left(\frac{1}{2} - \frac{\sqrt{3}}{12}\right) a\left(\xi^{m+1}, \xi^{m+1}\right) - \left(\frac{1}{2} + \frac{\sqrt{3}}{6}\right) a\left(\xi^{m+2}, \xi^{m+1}\right) \\ &\qquad \ge \left\{ \frac{1}{2} + \frac{\sqrt{3}}{4} - \left(\frac{1}{2} + \frac{\sqrt{3}}{6}\right)^2 \Big/ \left(2 - \frac{\sqrt{3}}{3}\right) \right\} a\left(\xi^{m+2}, \xi^{m+2}\right) \\ &\qquad \ge c \| \xi^{m+2} \|_1^2 \end{aligned}$$

Applying the above inequality to (9.9) yields for constants c, C > 0

$$\sum_{n=1}^{m} T^n \ge c \left\{ \sum_{n=1}^{m} \| \xi^{n+2} - 2\xi^{n+1} + \xi^n \|_1^2 + \sum_{n=1}^{m+1} \| \xi^{n+1} - \xi^n \|_1^2 + \| \xi^{m+2} \|_1^2 \right\}$$

$$- c \left(\| \xi^1 \|_1^2 + \| \xi^2 \|_1^2 \right) \qquad (9.10)$$

Similarly,

$$\left(\sum_{j=o}^{2} \alpha_j \xi^{n+j} \right)^2 = \left(\xi^{n+2} - \xi^{n+1} \right)^2 + \frac{1}{12} \left(\xi^{n+2} - 2\xi^{n+1} + \xi^n \right)^2$$

$$+ \left(\frac{\sqrt{3}}{3} - \frac{1}{2} \right) \left\{ \left(\xi^{n+2} - \xi^{n+1} \right)^2 - \left(\xi^{n+1} - \xi^n \right)^2 \right\}$$

and hence:

$$\sum_{n=1}^{m} \| \sum_{j=o}^{2} \alpha_j \xi^{n+j} \|_1^2 \leq \left(\frac{1}{2} + \frac{\sqrt{3}}{3} \right) \sum_{n=1}^{m+1} \| \xi^{n+1} - \xi^n \|_1^2$$

$$+ \frac{1}{12} \sum_{n=1}^{m} \| \xi^{n+2} - 2\xi^{n+1} + \xi^n \|_1^2 \qquad (9.11)$$

The proof is concluded by subtracting an appropriate multiple of (9.11) from (9.10).

Lemma 5

Let u(x,t) be the solution of the weak problem (9.2) and suppose that the assumptions (B) are satisfied, then for Δ_t sufficiently small $u^n \equiv u(x, n\Delta_t)$, $0 \leq n \leq T/\Delta_t$, satisfies (cf. (9.6))

$$\sum_{j=o}^{2} \frac{\alpha_j}{\Delta_t} \left(\rho(x) u^{n+j}, v \right) + a\left(\sum_{j=o}^{2} \beta_j u^{n+j} + c_2 \Delta_t q^{n+2} , v \right) \qquad (9.12)$$

$$= \left(f(x, \bar{t}^n, \bar{u}^n), v \right) - \left(\underline{b}(x, \bar{t}^n, \bar{u}^n) . \nabla v \right) + r(v) \qquad \forall v \in H_o^1(\Omega), \quad n \geq 1$$

and $q^{n+2} \equiv \frac{\partial}{\partial t} u \Big|_{t = (n+2)\Delta_t}$ may be expressed by

$$\left(\rho(x)q^{n+2}\ ,\ v\right) + a\left(u^{n+2}\ ,\ v\right) = \left(f(x,(n+2)\Delta_t\ ,\ \tilde{u}^{n+2})\ ,\ v\right)$$

$$- \left\{\ \underline{b}(x,(n+2)\Delta_t\ ,\ \tilde{u}^{n+2})\ .\nabla\ v\ \right\} + \tilde{r}(v) \qquad \forall v \in H_o^1(\Omega) \tag{9.13}$$

where (1) $\| r(v) \| \leq c\Delta_t^3 \ \| v \|_1$

(2) $\| \tilde{r}(v) \| \leq C\Delta_t^3 \ \| v \|_1$

and the functions $\bar{u}^n$, $\tilde{u}^{n+2}$ are defined as

$$\bar{u}^n \equiv \frac{1+\sqrt{3}}{6}\ u^{n-1} - \frac{1+2\sqrt{3}}{3}\ u^n + \frac{7+3\sqrt{3}}{6}\ u^{n+1}$$

$$\tilde{u}^{n+2} \equiv 3u^{n+1} - 3u^n + u^{n-1}$$

Proof.

The coefficients of the 3rd order L.M.S.D. are constructed so that, for any sufficiently differentiable function y(t),

$$\sum_{j=o}^{2} \frac{\alpha_j}{\Delta_t}\, y_{n+j} = y'(\bar{t}^n) + E_1(y) \tag{9.14}$$

where

$$|E_1(y)| \leq C\ Y_4\Delta_t^3 \quad ; \qquad Y_\ell \equiv \sup_{o\leq t\leq T} \left|\frac{d^\ell}{dt^\ell}\, y(t)\right|$$

whenever Δ_t is sufficiently small. Similarly

$$\sum_{j=o}^{2} \beta_j y_{n+j} + \Delta_t c_2 y'_{n+2} = y(\bar{t}^n) + E_2(y) \tag{9.15}$$

where $|E_2(y)| \leq C\ Y_3\ \Delta_t^3$

Multiply the expression (9.14) by $\rho(x)v$, $v \in H^1_o(\Omega)$, and then integrate over Ω. If we select 'y' $\equiv u(x,t)$ we achieve by (9.2)

$$\sum_{j=o}^{2} \frac{\alpha_j}{\Delta_t} \left(\rho(x)u^{n+j} , v\right) + a\left(u(\bar{t}^n) , v\right) = \left(f(x,\bar{t}^n , \bar{u}^n), v\right)$$

$$- \left(\underline{b} (x, \bar{t}^n , u(\bar{t}^n)) . \nabla v \right) + \left(E_1(u) , \rho(x)v\right) \qquad (9.16)$$

Using (9.15) with 'y' $\equiv$ u, and (9.16) we deduce that

$$\sum_{j=o}^{2} \frac{\alpha_j}{\Delta_t} \left(\rho(x)u^{n+j} , v\right) + a\left(\sum_{j=o}^{2} \beta_j u^{n+j} + \Delta_t c_2 \frac{\partial}{\partial t} u^{n+2} , v\right) =$$

$$\left(f(x, \bar{t}^n , \bar{u}^n), v\right) - \left(\underline{b}(x, \bar{t}^n , \bar{u}^n) . \nabla v\right) + r(v)$$

where r(v) may be expressed as

$$r(v) = \left(E_1(u) , \rho(x)v\right) + a\left(E_2(u), v\right)$$

$$+ \left(f(x, \bar{t}^n , u(\bar{t}^n)) - f(x, \bar{t}^n , \bar{u}^n), v\right)$$

$$+ \left(\underline{b}(x , \bar{t}^n , \bar{u}^n) - \underline{b} (x , \bar{t}^n , u(\bar{t}^n)) . \nabla v\right)$$

Note that by assumptions (B) and the Cauchy-Schwartz inequality

(1) $\| u(\bar{t}^n) - \bar{u}^n \| \leq C\Delta_t^3$

(2) $\left(f(x , \bar{t}^n , u(\bar{t}^n)) - f(x , \bar{t}^n , \bar{u}^n) , v\right) \leq L\| u(\bar{t}^n) - \bar{u}^n\| \, \| v\|$

(3) $\left(\underline{b}(x , \bar{t}^n , \bar{u}^n) - \underline{b}(x , \bar{t}^n , u(\bar{t}^n)) . \nabla v\right)$

$$= \int_{\Omega} \sum_{i=1}^{N} \left[b_i(x\ ,\ \bar{t}^n\ ,\ \bar{u}^n) - b_i(x\ ,\ \bar{t}^n\ ,\ u(\bar{t}^n))\right] \frac{\partial v}{\partial x_i}\ dx$$

$$\leq \int_{\Omega} \left(\sum_{i=1}^{N} \left\{b_i(x\ ,\ \bar{t}^n\ ,\ \bar{u}^n) - b_i(x\ ,\ \bar{t}^n\ ,\ u(\bar{t}^n))\right\}^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^{N} \left(\frac{\partial v}{\partial x_i}\right)^2 \right)^{\frac{1}{2}}\ dx$$

$$\leq L \int_{\Omega} \left(\sum_{i=1}^{N} |\bar{u}^n - u(\bar{t}^n)|^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^{N} \left(\frac{\partial v}{\partial x_i}\right)^2 \right)^{\frac{1}{2}}\ dx$$

$$\leq L\ N^{\frac{1}{2}}\ \|\bar{u}^n - u(\bar{t}^n)\|\ \ \|v\|_1 \qquad (9.17)$$

Hence, employing the above inequalities, (9.14) and (9.15) it is simple to establish the bound

$$\|r(v)\| \leq c\Delta_t^3 \|v\|_1$$

A simple manipulation of (9.2) establishes

$$\left(\rho(x)\ \frac{\partial}{\partial t}\ u^{n+2},\ v\right) + a(u^{n+2}\ ,\ v) = \left(f(x,\ t_{n+2},\ \tilde{u}^{n+2}),\ v\right)$$

$$- \left(\underline{b}(x,t_{n+2},\ \tilde{u}^{n+2}).\ \nabla\ v\right) + \tilde{r}(v) \qquad \forall\ v \in H_o^1(\Omega)$$

where for simplicity of notation $t_{n+2} \equiv (n+2)\Delta_t$, and

$$\tilde{r}(v) = \left(f(x,t_{n+2},\ u^{n+2}) - f(x,t_{n+2},\ \tilde{u}^{n+2}),\ v\right)$$

$$-\left(\underline{b}(x,t_{n+2},\ u^{n+2}) - \underline{b}(x,t_{n+2},\ \tilde{u}^{n+2})\ .\ \nabla\ v\right)$$

Noting that $\|u^{n+2} - \tilde{u}^{n+2}\| \leq C\Delta_t^3$ we deduce by the assumptions (B) that

$$\|\tilde{r}(v)\| \leq C\Delta_t^3\ \|v\|_1$$

Our intention is to employ a method corresponding to chapter 6, section b. Hence, we need to impose the elliptic regularity condition (A_{11}). As before we define $W \in V_h^p$, $\forall\ t \geq 0$ by

$$a(u - W, V) = 0 \qquad \forall\ V \in V_h^p \tag{9.18}$$

Denote $\eta \equiv W - u$, whence (cf. (6.28)) for h sufficiently small

$$\|\eta\|_r + \|\frac{\partial}{\partial t}\eta\|_r \leq Ch^{p+1-r}\left\{\|u\|_{p+1} + \|\frac{\partial}{\partial t}u\|_{p+1}\right\} \quad r=0,1 \tag{9.19}$$

Analogously define $\bar{Q}^{n+2} \in V_h^p$ by

$$a(q^{n+2} - \bar{Q}^{n+2}, V) = 0 \qquad \forall\ V \in V_h^p \tag{9.20}$$

The assumption (B_1) testifies that $q^{n+2} \equiv \frac{\partial}{\partial t} u\Big|_{t=t_{n+2}} \in H^{p+1}(\Omega)$

and consequently by lemma (4.1) [8],

$$\| q^{n+2} - \bar{Q}^{n+2} \| \leq Ch^{p+1}\|\frac{\partial u}{\partial t}\|_{p+1} \tag{9.21}$$

The principle results of this chapter are contained in the following theorem and its corollary:

Theorem 4

Let u(x,t) be the solution of (9.1) and $\{U^n\}_{n=3}^{m+2}$ be defined by (9.6) with $\theta > \frac{1}{12}$. Further, let us suppose the assumptions (A_{11}), (B_1) - (B_v) are satisfied. Then for $3 \leq m + 2 \leq T/_{\Delta_t}$, and h, Δ_t sufficiently small

$$\|\xi^{m+2}\| \leq C\left\{h^{p+1} + \Delta_t^3 + \sum_{i=1}^{2}\left[\|\xi^i\| + \Delta_t^{\frac{1}{2}}\|\xi^i\|_1\right]\right\}$$

where $\xi \equiv W - U$, and $\xi^n \equiv \xi\Big|_{t=t_n}$

Proof.

For brevity of notation define $W^n \equiv W\Big|_{t=t_n}$

A straightforward manipulation of the expressions (9.12), (9.18) and (9.20) yields

$$\sum_{j=o}^{2} \frac{\alpha_j}{\Delta_t} (\rho(x)W^{n+j}, V) + a\left(\sum_{j=o}^{2} \beta_j W^{n+j} + c_2 \Delta_t \bar{Q}^{n+2} , V\right) = \sum_{j=o}^{2} \frac{\alpha_j}{\Delta_t} \left(\rho(x)\eta^{n+j}, V\right)$$

$$+ r(V) + (f(x, \bar{t}^n , \bar{u}^n), V) - (\underline{b}(x, \bar{t}^n , \bar{u}^n) . \nabla V)$$

$$\forall\ V \in V_h^p \qquad (9.22)$$

Subtracting (9.6_1) from (9.22) achieves, with $\tilde{Q}^{n+2} \equiv \bar{Q}^{n+2} - Q^{n+2}$,

$$\sum_{j=o}^{2} \frac{\alpha_j}{\Delta_t} (\rho(x)\xi^{n+j} , V) + a\left(\sum_{j=o}^{2} \beta_j \xi^{n+j} + c_2 \Delta_t \tilde{Q}^{n+2}, V\right) = \sum_{j=o}^{2} \frac{\alpha_j}{\Delta_t}\left(\rho(x)\eta^{n+j}, V\right)$$

$$+ r(V) + (f(x,\bar{t}^n, \bar{u}^n) - f(x,\bar{t}^n, \bar{U}^n), V) \qquad (9.23)$$

$$- \left(\underline{b}(x,\bar{t}^n, \bar{u}^n) - \underline{b}(x,\bar{t}^n, \bar{U}^n) . \nabla V \right) \qquad \forall\ V \in V_h^p$$

Note that, using the assumptions (B) and (9.19)

1) $(f(x,\bar{t}^n, \bar{u}^n) - f(x,\bar{t}^n, \bar{U}^n) , V) \le L \,||\bar{u}^n - \bar{U}^n||\ \ ||V||$

2) $\left(\underline{b}(x,\bar{t}^n, \bar{u}^n) - \underline{b}(x,\bar{t}^n, \bar{U}^n).\nabla V\right) \le L N^{\frac{1}{2}} ||\bar{u}^n - \bar{U}^n||\ \ ||V||_1$ (cf(9.17))

3) By the consistency relationship $\sum_{j=o}^{2} \alpha_j = 0$, hence

$$\sum_{j=o}^{2} \frac{\alpha_j}{\Delta_t} \eta^{n+j} = \frac{\alpha_2}{\Delta_t} (\eta^{n+2} - \eta^{n}) + \frac{\alpha_1}{\Delta_t} (\eta^{n+1} - \eta^{n})$$

i.e. $\left(\rho(x) \sum_{j=o}^{2} \frac{\alpha_j}{\Delta_t} \eta^{n+j}, V\right) \le C \sup_{o \le t \le T} \| \frac{\partial}{\partial t} \eta \| \; \| V \| \le C h^{p+1} \| V \|$

The above inequalities and lemma 5 produce a bound on the right hand side of (9.23), namely

$$\sum_{j=o}^{2} \frac{\alpha_j}{\Delta_t} (\rho(x) \xi^{n+j} , V) + a\left(\sum_{j=o}^{2} \beta_j \xi^{n+j} + c_2 \Delta_t \tilde{Q}^{n+2} , V\right)$$

$$\le C \left\{ h^{p+1} + \Delta_t^3 + \| \bar{u}^n - \bar{U}^n \| \right\} \| V \|_1$$

Select $V = \sum_{j=o}^{2} \beta_j \xi^{n+j} + c_2 \Delta_t \tilde{Q}^{n+2}$. Then using (B_{1v}) and the inequality $|ab| \le \varepsilon a^2 + b^2/4\varepsilon$, for suitable values of ε, we have shown that

$$\left(\frac{\rho(x)}{\Delta_t} \sum_{j=o}^{2} \alpha_j \xi^{n+j} , \sum_{j=o}^{2} \beta_j \xi^{n+j} + c_2 \Delta_t \tilde{Q}^{n+2}\right) + C_o^{-1} \| \sum_{j=o}^{2} \beta_j \xi^{n+j} + c_2 \Delta_t \tilde{Q}^{n+2} \|_1^2 \le \frac{C_o^{-1}}{2} \| \sum_{j=o}^{2} \beta_j \xi^{n+j} + c_2 \Delta_t \tilde{Q}^{n+2} \|_1^2 + C \left\{ h^{2(p+1)} + \Delta_t^6 + \| \bar{u}^n - \bar{U}^n \|^2 \right\} \qquad (9.24)$$

Subtracting (9.6_{11}) from (9.13), and using (9.18), establishes

$$\left(\rho(x) \{q^{n+2} - Q^{n+2}\} , v\right) + a \left(\xi^{n+2} , v\right) = \left(f(x, t_{n+2}, \tilde{u}^{n+2}) - f(x, t_{n+2}, \tilde{U}^{n+2}) , v\right) - \left(\underline{b}(x, t_{n+2}, \tilde{u}^{n+2}) - \underline{b}(x, t_{n+2}, \tilde{U}^{n+2}) . \nabla V\right)$$

$$+ \tilde{r}(V) \; \forall \; V \in V_h^p$$

Noting that $\tilde{Q}^{n+2} = (\bar{Q}^{n+2} - q^{n+2}) + (q^{n+2} - Q^{n+2})$ we achieve

$$\left(\rho(x)\tilde{Q}^{n+2} , V\right) + a \left(\xi^{n+2} , V\right) = \left(\rho(x)\{\bar{Q}^{n+2} - q^{n+2}\} , V\right) + \left(f(x, t_{n+2}, \tilde{u}^{n+2}) - f(x, t_{n+2}, \tilde{U}^{n+2}), V\right) - \left(\underline{b}(x, t_{n+2}, \tilde{u}^{n+2}) - \underline{b}(x, t_{n+2} , \tilde{U}^{n+2}).\nabla V\right) + \tilde{r}(V) \qquad (9.25)$$

Combining (9.24), and (9.25) with $V = \sum_{j=o}^{2} \alpha_j \xi^{n+j}$ and bounding the terms as before, achieves

$$\left(\frac{\rho(x)}{\Delta_t} \sum_{j=o}^{2} \alpha_j \xi^{n+j} , \sum_{j=o}^{2} \beta_j \xi^{n+j}\right) - c_2 \, a\left(\xi^{n+2} , \sum_{j=o}^{2} \alpha_j \xi^{n+j}\right) +$$

$$\frac{C_o^{-1}}{2} \left\| \sum_{j=o}^{2} \beta_j \xi^{n+j} + c_2 \Delta_t \tilde{Q}^{n+2} \right\|_1^2 \le \mu \left\| \sum_{j=o}^{2} \alpha_j \xi^{n+j} \right\|_1^2 + C \left\{ \| \bar{u}^n - \bar{U}^n \|^2 + \| \tilde{u}^{n+2} - \tilde{U}^{n+2} \|^2 + h^{2p+2} + \Delta_t^6 \right\}$$

Multiply the above expression by Δ_t , sum the result for $n=1,2,\dots,m$, use lemmas 3 and 4, $u - U = \xi - \eta$, and (9.19) to deduce

$$\| \xi^{m+2} \|^2 + \Delta_t \| \xi^{m+2} \|_1^2 + \Delta_t \frac{C_o^{-1}}{2} \sum_{n=1}^{m} \left\| \sum_{j=o}^{2} \beta_j \xi^{n+j} + c_2 \Delta_t \tilde{Q}^{n+2} \right\|_1^2$$

$$\le C \left\{ \Delta_t \sum_{n=o}^{m+1} \| \xi^n \|^2 + h^{2p+2} + \Delta_t^6 + \sum_{i=1}^{2} \left(\| \xi^i \|^2 + \Delta_t \| \xi^i \|_1^2 \right) \right\} \qquad (9.26)$$

The following lemma is now needed. (see [18] for proof).

Gronwall's lemma : a discrete analogue

Suppose that ϕ and χ are non-negative functions defined for $t = n\Delta_t$, $n = 0,1,\dots,m$, and that χ is non-decreasing. If

$$\phi^n \le \chi^n + C\Delta_t \sum_{r=o}^{n-1} \phi^r \quad , \qquad n = 1,2,\dots,m$$

where C is a positive constant, then

$$\phi^n \le \chi^n e^{Cn\Delta_t} \quad , \qquad n = 1,2\dots,m$$

A simple application of the above lemma to (9.26) yields

$$\| \xi^{m+2} \|^2 \le C \left\{ h^{2p+2} + \Delta_t^6 + \sum_{i=1}^{2} \left(\| \xi^i \|^2 + \Delta_t \| \xi^i \|_1^2 \right) \right\}$$

which immediately establishes the desired result.

Corollary

Assuming the criteria of Theorem 4, a bound on $u^{m+2} - U^{m+2}$ in the L_2 - norm is given by

$$\| u^{m+2} - U^{m+2} \| \le C \left\{ h^{p+1} + \Delta_t^3 + \Delta_t^{\frac{1}{2}} h^p + \sum_{i=1}^{2} \left[\| u^i - U^i \| + \Delta_t^{\frac{1}{2}} \| u^i - U^i \|_1 \right] \right\}$$

Proof.

This follows immediately from theorem 4, (9.19), and the triangle inequality on $u - U = \xi - \eta$.

The discrete scheme (9.6) is not self-starting and requires initial values for $\{U^i\}_{i=o}^{2}$. As before, we may select U^o to be the projection of g(x) onto V_h^p by the L_2 - inner product. The derivation of suitable values for U^1 and U^2 is not as trivial. A possibility is to derive these values by utilising an one-step method with a smaller time increment. Such one-step methods include variations of the θ - , and Crank-Nicholson methods, eg. [7] , [44]. Alternatively, we may apply Richardson's extrapolation technique to a one-step method. For example employing the backward difference method, ie. the θ-method with θ = 1, and Richardson's extrapolation twice we may employ the techniques of [11] and[36] to deduce that under certain continuity conditions.

$$\sum_{i=1}^{2} \left[\| u^i - U^i\| + \Delta_t^{\frac{1}{2}} \| u^i - U^i\|_1\right] \leq C\left\{ h^{p+1} + \Delta_t^{\frac{1}{2}} h^p + \Delta_t^3 + \| u^o - U^o\| \right\}$$

Finally we discuss the implementation of the scheme (9.6). Using the notation of chapter 7, and the definitions of M and K in (9.4), it is simple to see that the expressions (9.6_1) and (9.6_{11}) are equivalent to

$$\sum_{j=o}^{2} \alpha_j M\underline{U}^{n+j} + \Delta_t \sum_{j=o}^{2} \beta_j K\underline{U}^{n+j} + \Delta_t^2 c_2 K\underline{Q}^{n+2} = \Delta_t \underline{f}^n \tag{9.27}$$

$$M\underline{Q}^{n+2} + K\underline{U}^{n+2} = \underline{f}^{n,1}$$

where the vectors $\underline{f}^n$ and $\underline{f}^{n,1}$ are given by

$$\left(\underline{f}^n\right)_i = \left(f(x, \bar{t}^n, \bar{U}^n), v_i\right) - \left(\underline{b}(x, \bar{t}^n, \bar{U}^n).\nabla v_i\right) \qquad \text{and}$$

$$\left(\underline{f}^{n,1}\right)_i = \left(f(x, t_{n+2}, \tilde{U}^{n+2}), v_i\right) - \left(\underline{b}(x, t_{n+2}, \tilde{U}^{n+2}).\nabla v_i\right)$$

Eliminating $\underline{Q}^{n+2}$ from (9.27), and rearranging, we achieve

$$\left[\alpha_2\ I + \Delta_t\beta_2 M^{-1}K - \Delta_t^2\ c_2(M^{-1}K)^2\right]\underline{U}^{n+2} = -\left[\alpha_1\ I + \Delta_t\beta_1 M^{-1}K\right]\underline{U}^{n+1}$$

$$-\left[\alpha_o\ I + \Delta_t\beta_o M^{-1}K\right]\underline{U}^n + \Delta_t M^{-1}\underline{f}^n - \Delta_t^2 c_2 M^{-1}KM^{-1}\underline{f}^{n,1}$$

The mode of implementation depends on the character of the roots z_1 and z_2 of

$$-\frac{\alpha_2}{c_2} - \frac{\beta_2}{c_2}x + x^2$$

For $\frac{1}{12} < \theta < 1.2334587 \equiv \overline{\theta}$ the roots are complex whilst for $\theta \geq \overline{\theta}$ the roots are real, with a double root for $\theta = \overline{\theta}$.

Thus for $\frac{1}{12} < \theta < \overline{\theta}$ the implementation is equivalent to

$$M\underline{g}^n = \underline{f}^{n,1}$$

$$(z_1 M - \Delta_t K)\underline{U}^{n,1} = \frac{1}{c_2}\left[\alpha_1 M + \Delta_t B_1 K\right]\underline{U}^{n+1} + \frac{1}{c_2}\left[\alpha_o M + \Delta_t\ \beta_o K\right]\underline{U}^n$$

$$-\frac{1}{c_2}\Delta_t\underline{f}^n + \Delta_t^2\ K\underline{g}^n \equiv \underline{F}(\theta)$$

$$\underline{U}^{n+2} = \frac{\text{Im}\ \underline{U}^{n,1}}{\text{Im}\ \overline{z}_1} \tag{9.28}$$

In terms of the discretization error, the effect of varying θ is concentrated in the error E_2. For a sufficiently differentiable function $y(t)$

$$E_2(y) = \left(\frac{\sqrt{3}}{36} - \frac{\theta}{3}\right)\Delta_t^3\ y_{n+1}^{'''} + O(\Delta_t^4).$$

Hence, given $\theta = \frac{\sqrt{3}}{12}$, (9.5) is a fourth order scheme and $E_2(y)$ may be bounded in modulus by $C\Delta_t^4 Y_4$, whenever Δ_t is sufficiently small. Thus we are encouraged to expect optimal accuracy for $\theta = \frac{\sqrt{3}}{12}$ as $\Delta_t \to 0$.

The other distinctive value of θ is $\theta = \overline{\theta}$. This value facilitates the implementation procedure but has the disadvantage of producing a relatively large error constant, C_3, for E_2.

ie. $C_3 \equiv \left(\frac{\sqrt{3}}{36} - \frac{\overline{\theta}}{3}\right) \simeq -0.363$

Note that the implementation, with $\theta = \overline{\theta}$, is equivalent to

$$\begin{aligned} M\underline{g}^n &= \underline{f}^{n,1} \\ (z^*M - \Delta_t K)\,\underline{U}^{n,1} &= \underline{F}(\overline{\theta}) \qquad (9.29) \\ (z^*M - \Delta_t K)\,\underline{U}^{n+2} &= \underline{U}^{n,1} \end{aligned}$$

where $z^* \simeq 0.8734384$.

For $\theta \geq \overline{\theta}$ the error constant C_3 is prohibitively large. Thus restricting $\frac{1}{12} < \theta \leq \overline{\theta}$ we may approximate the solution of (9.1) by solving two algebraically linear systems of equations at each time step, c.f (9.28) - (9.29).

Discussion

As explained previously this section examines the application of a semi-discrete, Galerkin - L.M.S.D. scheme to the linear parabolic equation. The justification of such an application has been discussed and its merits established. High order, easily computable schemes are formulated that supercede, or rival, all the previously documented semi-discrete Galerkin schemes.

However, the suitability of applying a semi-discrete Galerkin - L.M.S.D. scheme to a general non-linear parabolic equation is doubtful. At each time level, a direct application generally requires the differentiation of non-linear systems of ordinary differential equations and the solution of a complicated non-linear system of algebraic equations. A linearisation process is needed to render the scheme more computationally attractive, but, error analysis suggests that this causes a reduction in the order of convergence. An important exception relates to the class of quasi-linear equations investigated in chapter 9. Here we have described a third order, unconditionally stable scheme that requires the solution of two systems of linear algebraic equations at each time level. Consequently, this improves on the order of accuracy of all the previously formulated linearised schemes.

REFERENCES

1. AHLFORS, L.V. : Complex analysis; McGraw-Hill, New York. (1966)
2. BELLMAN, R. : Stability Theory of Differential Equations; McGraw-Hill, New York. (1952).
3. CROUZEIX, M. : Sur l'approximation des equations differentielles operationnelles linéaries par des methods de RUNGE-KUTTA; Ph.d Thesis, Université Paris VI. (1975).
4. CRYER, C.W. : *A new class of highly stable methods : A_0-stable* methods; BIT, 13, pp. 153-159. (1973).
5. DAHLQUIST, G. : A special stability problem for linear multistep methods; BIT, 3, pp. 27-43 (1963).
6. DENDY, J.E., Jr. : An analysis of some Galerkin schemes for the solution of nonlinear time dependent problems; SIAM J. Numer. Anal., 12, pp. 541-565. (1975).
7. DOUGLAS, J., Jr., and DUPONT., T. : Galerkin methods for parabolic equations; SIAM J. Numer. Anal., 7, pp. 575-626. (1970).
8. DUPONT, T., FAIRWEATHER, G., and JOHNSON, J.P. : Three-level Galerkin methods for parabolic equations; SIAM J. Numer. Anal., 11, pp. 392-410. (1974).
9. EHLE, B.L. : High order A-stable methods for the numerical solution of systems of differential equations; BIT, 8, pp. 276-278. (1968).
10. ENRIGHT, W.H. : Second derivative multistep methods for stiff ordinary differential equations; SIAM J. Numer. Anal., 11, pp. 321-331. (1974).
11. FAIRWEATHER, G., and JOHNSON, J.P. : On the extrapolation of Galerkin methods for parabolic equations; Numer. Math., 23, pp. 269-287. (1975).

12. GENIN, Y. : An algebraic approach to A-stable linear multistep, multiderivative integration formulas; BIT, 14, pp.382-406.(1974).

13. HENRICI, P. : *Discrete variable methods in ordinary differential* equations; J. Wiley & Sons, New York - London - Sydney. (1962).

14. JELTSCH, R. : Note on A-stability of multistep, multiderivative methods; BIT, 16, pp. 74-78. (1976).

15. LADYŽENSKAHA, O.A., SOLONNIKOV, V.A., and URAL'CEVA, N.N. : Linear and quasilinear equations of parabolic type; Transl. Math. Monographs, vol. 23, Amer. Math. Soc., Providence, R.I. (1968).

16. LAMBERT, J.D. : Computational methods in ordinary differential equations; J. Wiley & Sons, London - New York - Sydney - Toronto. (1973).

17. LAMBERT, J.D. : Variable coefficient multistep methods for ordinary differential equations applied to parabolic partial differential equations; in 'Topics in Numerical Analysis II', Proceedings of the Royal Irish Academy Conference on Numerical Analysis, 1974. (Ed. MILLER, J.H.)

18. LEES, M. : A priori estimates for the solution of a difference approximation to parabolic differential equations; Duke Math. J., 27, pp. 297-312. (1960).

19. MAKINSON, G.J. : Stable high order implicit methods for the numerical solution of systems of differential equations; The Computer Journal, 11, pp. 305-310. (1968).

20. McLEOD, R.J.Y., and MITCHELL, A.R. : The construction of basis functions for curved elements in the finite element methods; J.I.M.A., 10, pp. 382-393. (1972).

21. McLEOD, R.J.Y., and Mitchell, A.R. : The use of parabolic arcs in matching curved boundaries in the finite element method; J.I.M.A., 16, pp. 239-246. (1975).

22. MIHLIN, S.G. : Mathematical Physics, An advanced course; Amsterdam, North - Holland. (1970).

23. MIRANDA, C. : Partial differential equations of elliptic type, (second rev. edition), Springer, Berlin-Heidelberg - New York. (1970).

24. MITCHELL, A.R., and McLeod, R. : Curved elements in the finite element method in'Proceedings of conference on numerical solution of differential equations', Lecture notes in mathematics, no. 363, Springer - Verlag, Berlin. (1974). (Ed. WATSON, G.A.).

25. NASSIF, N.R. : On the discretization of the time variable in parabolic partial differential equations; in 'The mathematics of finite elements and applications, II' Academic Press, (Ed. WHITEMAN, J.R.). To appear.

26. NORSETT, S.P. : One-step methods of Hermite type for numerical integration of stiff systems; BIT, 14, pp. 63-77. (1974).

27. RALSTON, A. : A first course in numerical analysis; McGraw-Hill. (1965).

28. SCHULTZ, M.H. : Spline analysis; Prentice-Hall, Inc., Englewood Cliffs, N.J. (1973).

29. STRAND, G., and FIX, G.J. : An analysis of the finite element method; Prentice Hall, Inc., Englewood Cliffs, N.J. (1973).

30. TITCHMARSH, E.C. : Eigenfunction expansions associated with second-order differential equations, Part 1; Oxford University Press. (1946).

31. THOMÉE, V. : Some convergence results for Galerkin methods for parabolic boundary value problems; in 'Mathematical aspects of finite elements in partial differential equations', Academic Press, Inc., (Ed. de Boor, C.) (1974).

32. THOMÉE, V. : Spline-Galerkin methods for initial-value problems with constant coefficients; in 'Proceedings of conference on numerical solution of differential equations', Lecture notes in mathematics, no. 363, Springer-Verlag, Berlin. (1974) (Ed. WATSON, G.A.).

33. THOMÉE, V., and WAHLBIN, L. : On Galerkin methods in semilinear parabolic problems; SIAM J. Numer. Anal., 12, pp. 378-389. (1975).

34. WATANABE, D.S., and Flood, J.R. : An implicit fourth order difference method for viscous flow; Math. Comp., 28, pp. 27-32. (1974).

35. WENDROFF, B. : Spline-Galerkin methods for initial-value problems with variable coefficients; in 'Proceedings of conference on numerical solution of differential equations', Lecture notes in mathematics, no. 363, Springer-Verlag, Berlin. (1974). (Ed. WATSON, G.A.).

36. WHEELER, M.F. : A priori L_2 error estimates for Galerkin approximations to parabolic partial differential equations; SIAM J. Numer. Anal., 10, pp. 723-759. (1973).

37. WHEELER, M.F. : L_∞ estimates of optimal orders for Galerkin methods for one-dimensional second order parabolic and hyperbolic equations; SIAM J. Numer. Anal., 10, pp. 908 - 913. (1973).

38. WIDLUND, O.B. : A note on unconditionally stable linear multistep methods; BIT, 7, pp.65-70. (1967).

39. ZLÁMAL, M. : Finite element methods for parabolic equations ; Math. Comp., 28, pp. 393-404. (1974).

40. ZLÁMAL, M. : Finite element multistep discretizations of parabolic boundary value problems; Math. Comp., 29, pp. 1-10. (1975).

41. ZLÁMAL, M. : Finite element multistep methods for parabolic equations; To appear.

42. ZLÁMAL, M. : Unconditionally stable finite element schemes for parabolic equations; in 'Topics of Numerical Analysis II', Proceedings of the Royal Irish Academy Conference on Numerical Analysis. 1974. (Ed. MILLER, J.H.).

43. ZLÁMAL, M. : Finite element methods in heat conduction problems; in 'The mathematics of finite elements and applications, II' Academic Press, (Ed. WHITEMAN, J.R.). To appear.

44. ZLÁMAL, M. : Finite element methods nonlinear parabolic equations; To appear.