CANADIAN THESES

THÈSES CANADIENNES

5/

NAME OF AUTHOR/*NOM DE L'AUTEUR* Yuh-shion Kuo

TITLE OF THESIS/*TITRE DE LA THÈSE* Equivalences Between Numerical Methods for Solving Differential Equations

UNIVERSITY/*UNIVERSITÉ* Simon Fraser University

DEGREE FOR WHICH THESIS WAS PRESENTED/ *GRADE POUR LEQUEL CETTE THÈSE FUT PRÉSENTÉE* Master of Science

YEAR THIS DEGREE CONFERRED/*ANNÉE D'OBTENTION DE CE DEGRÉ* 1983

NAME OF SUPERVISOR/*NOM DU DIRECTEUR DE THÈSE* R. D. Russell

DATED/*DATÉ* Aug 1 83 SIGNED/*SIGNÉ*

PERMANENT ADDRESS/*RÉSIDENCE FIXÉ* [redacted]

EQUIVALENCES BETWEEN NUMERICAL METHODS FOR SOLVING DIFFERENTIAL EQUATIONS

by

Yuh-Shiow Kuo

B.Sc. Fu-Jen Catholic University, 1978

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in the Department

of

Mathematics

SIMON FRASER UNIVERSITY

July 1983

APPROVAL

Name: Yuh-Shiow Kuo

Degree: MASTER OF SCIENCE

Title of thesis: EQUIVALENCES BETWEEN NUMERICAL METHODS FOR SOLVING DIFFERENTIAL EQUATIONS

Examining Committee:

Chairman: B.S. Thomson

R.D. Russell
Senior Supervisor

R.W. Lardner

E.M. Shoemaker

J. Paine
External Examiner
Visiting Assistant Professor
Department of Mathematics
Simon Fraser University

Date Approved: August 3, 1983

PARTIAL COPYRIGHT LICENSE

Title of Thesis/Dissertation:

Equivalences Between Numerical Methods for Solving Differential Equations.

Author:

(signature)

Yuh-Shiow Kuo.

(name)

Aug. 1 '83

(date)

ABSTRACT

A brief description of various types of boundary value problems and initial value problems is given. Existence and uniqueness theorems and stability properties of them are considered. One of our main purposes is to survey the well-known numerical methods such as initial value approaches, finite difference methods, and finite element methods for solving the problems . Then, various equivalences between the methods are shown. Lastly, the high order finite difference methods considered by Doedel and by Lynch-Rice are discussed from a different point of view. A relationship between the high order finite difference methods and collocation methods is presented.. Comparison of operation counts and numerical results for Doedel's methods, Lynch-Rice's methods, and collocation methods using B-splines and Gauss points is treated.

ACKNOWLEDGEMENTS

I am grateful to Dr. R. D. Russell for accepting me as his student and for suggesting the topics of this thesis. Because of his kind and patient guidance, I could finish my thesis and take care of my little one at the same time.

Thanks to Dr. John Paine for his kindly help in part of the thesis.

I also want to say thanks to the Mathematics Department S.F.U. for offering enough financial support and to Mrs. Kathy Hammes and Mrs. Sylvia Holmes for their kindness and help in arranging TAships without which my studies would not have been possible.

TABLE OF CONTENTS

LIST OF TABLES

1. Introduction

For solving boundary value problems (BVPs) for ordinary differential equations (ODEs), the most common numerical methods are: initial value approaches; finite difference methods; and finite element methods. One of the areas of considerable interest to numerical analysts is the relationships between these methods. Once some equivalence has been found, properties of a method (e.g. the property of being well-conditioned for a multiple shooting method) can thus be shown to apply to the equivalent ones.

The main purposes of this thesis are to show equivalences between some well-known numerical methods and to give a clear view of the high order finite difference methods of Doedel [9] and Lynch-Rice [16].

In Chapter 2, general forms for ODEs are given to help maintain basic understanding. Before introducing numerical methods, existence and uniqueness theorems for the solutions of general initial value problems (IVPs) and linear BVPs are mentioned. Since it is extremely difficult to establish existence and uniqueness theorems for general BVPs, there are only some for restricted BVPs in special cases. One of them that relates to second order BVPs is mentioned in Section 2.2. Basic stability properties of Ortega [18] for linear IVPs and BVPs are provided. It is shown in Lentini-Osborne-Russell [15] that the well-conditioning of a BVP is related to two bounding quantities: one involving the boundary conditions and the other involving the Green's function. Their result is described also. Having well-posed BVPs in hand, we consider numerical methods for solving them. In Chapter 3, to prepare the groundwork needed in the subsequent Chapters, the well-known methods are introduced, viz superposition, simple shooting, multiple shooting, and invariant imbedding for initial value approaches; trapezoidal rule, midpoint rule, and Runge-Kutta schemes for finite difference methods; collocation, Galerkin, least square, and Ritz methods for finite element methods. When presenting the methods, some equivalences between them are easily seen as they are introduced, e.g. equivalence between discrete Galerkin and collocation and equivalence between discrete least square and collocation.

Since finite difference methods involve unknowns which correspond directly to approximate solutions at mesh points, in Chapter 4 equivalences between finite difference methods and some other methods are emphasized. They include the trapezoidal rule and collocation; the midpoint rule and collocation; Runge-Kutta methods and collocation; the Box scheme and multiple shooting; multiple shooting and collocation; multiple shooting and invariant imbedding; and the simple Ritz and finite difference methods. While all of them are probably known, some are not in the literature. Although in theory, two methods are shown to be equivalent mathematically, i.e. they have the same solution set, in practice their properties (e.g. order of accuracy) can be different. This difference is particularly important from a computational point of view.

An investigation was made for high order finite difference methods considered by Doedel [9] and by Lynch-Rice [16]. The main difference between their methods is that Doedel's methods include noncompact approximations while Lynch-Rice do not. In Chapter 5, the construction of high order finite difference methods derived by Doedel is presented. Order of consistency and stability of the schemes are also discussed. For a general n-th order linear differential equation, choice of a set of auxiliary points which gives one higher order of accuracy is given. For the special n-th order differential equation $M=D^n$, formulae are provided to evaluate the uniquely determined right-hand-side coefficients of the approximations, and to find the location of the special auxiliary points which give the order of accuracy as high as possible. An obvious equivalence between these methods and collocation methods is given in Section 5.5. This equivalence appears to have not been previously observed and not shown in either Doedel [9] or in Lynch-Rice [16].

Work estimates for the finite difference methods for general n-th order differential equations are provided. Comparison of computational work of Doedel's schemes and of Lynch-Rice's schemes is offered for general second order differential equations. Lynch-Rice [16] also compared their methods with five other methods, but the comparison here is based on the Langrange polynomial interpolation basis functions rather than the set of basis functions they considered. From the operation counts, we conclude that Doedel's methods are more efficient than Lynch-Rice's schemes for same orders of accuracy if one takes as few auxiliary points as possible. In Section 5.8, numerical examples show that Doedel's schemes with one auxiliary point are competive with Lynch-Rice's schemes. Hence, methods using one auxiliary point are the most efficient of the high order finite difference methods. Comparing some numerical results of the finite difference methods with that of collocation methods using B-splines and Gauss points, one finds that the finite difference methods can require a large number of subintervals to achieve any significant accuracy.

The last Chapter of the thesis is Chapter 6, the conclusion.

2. IVPs and BVPs

We begin with a brief account of some of the basic prerequisites: general IVPs and BVPs, existence and uniqueness theorems, and problem stability.

2.1. Standard Problems for ODEs

In this section, some standard problems which correspond to the most common forms of BVPs are presented. Since ways to numerically solve BVPs can be closely connected with methods for solving IVPs, and since the theory of IVPs is closely related to that for BVPs, a treatment of IVPs is considered also.

2.1.1. IVPs

The general IVP can be written as a first order system

(2.1a) $\quad y'(t)=f(t,y(t)) \qquad t>a$

(2.1b) $\quad y(a)=\alpha$

where $y(t)=(y_1(t),y_2(t),\ldots,y_n(t))^T$ is the unknown function, $f(t,y)=(f_1(t,y),f_2(t,y),\ldots,f_n(t,y))^T$ is the nonlinear right hand side and α is a known n-vector of initial conditions which completely determes $y(t)$.

A high order ODE

$$y^{(m)}(t)=f(t,y,y',\ldots\ldots,y^{(m-1)})$$

can be converted to the first order form (2.1a), by letting

$$y_1(t)=y(t)$$

$$y_2(t)=y'(t)$$

$$\vdots$$

$$y_m(t)=y^{(m-1)}(t) .$$

The ODE has the equivalent form

$$y'_1=y_2$$

$$y'_2=y_3$$

$$\vdots$$

$$y'_{m-1}=y_m$$

$$y'_m=f(t,y_1,y_2,\ldots\ldots,y_m) .$$

Similarly, a system of high order ODEs can be reduced to a set of first order equations in this way.

In the simpler case where the IVP is linear, (2.1) is simplified to

(2.2a) $\quad \underset{\sim}{y}'(t)=A(t)\underset{\sim}{y}(t)+\underset{\sim}{f}(t) \qquad t>a$

(2.2b) $\quad \underset{\sim}{y}(a)=\underset{\sim}{\alpha}$

where $A(t)$ is an $n \times n$ matrix and $\underset{\sim}{f}(t)$ is an n-vector valued function of t.

The linear system (2.2a) is called **homogeneous** if $\underset{\sim}{f}(t)=0$, and **inhomogeneous** otherwise.

2.1.2. BVPs

Unlike IVPs, solutions to BVPs are not completely determined by initial information; the information is given at two or more points which normally correspond to the boundary of some physical region of interest. One basic form is the linear two point BVP

(2.3a) $\quad \underset{\sim}{y}'(t)=A(t)\underset{\sim}{y}(t)+\underset{\sim}{f}(t) \qquad a\le t\le b$

(2.3b) $\quad B_a\underset{\sim}{y}(a)+B_b\underset{\sim}{y}(b)=\underset{\sim}{\alpha}$

where $\underset{\sim}{y}(t)$ and $\underset{\sim}{f}(t)$ are n-vectors, $A(t)$ is an n x n matrix, a and b are finite or infinite and $\underset{\sim}{\alpha}$ is a constant n-vector, B_a and B_b are n x n matrices corresponding to n boundary conditions.

For (2.3) to have a unique solution, it is necessary but not sufficient that these boundary conditions be linearly independent, i.e. that the matrix $(B_a | B_b)$ have n linearly independent columns or simply $\mathrm{rank}(B_a, B_b)=n$.

BC of the general form (2.3b) are called **nonseparated** BC since each involves information about $\underset{\sim}{y}(t)$ at both end points. If $\mathrm{rank}(B_a)<n$ or $\mathrm{rank}(B_b)<n$, then the BC are called **partially separated**. The BC are called **separated** if they can be simplified

to

$$C_a y(a) = \alpha_1$$

$$C_b y(b) = \alpha_2$$

where C_a is a p x n matrix , C_b is an (n-p) x n matrix , and $p = \text{rank}(B_a)$.

A general linear multipoint BVP consists of the ODE (2.3a) and multipoint BC

(2.3c) $$\sum_{j=1}^{r} B_j y(\zeta_j) = \alpha$$

where $B_1, \ldots, B_r$ are n x n matrices, α is an n-vector and $a = \zeta_1 < \zeta_2 < \ldots < \zeta_r = b$.

A nonlinear two point BVP can normally be expressed in the form

(2.4a) $$y'(t) = f(t, y(t)) \qquad a \le t \le b$$

(2.4b) $$g(y(a), y(b)) = 0$$

where $g = (g_1, \ldots, g_n)^T$ and 0 is the zero n-vector.

A nonlinear mth order (scalar) BVP normally has the form

(2.5a) $$y^{(m)}(t) = f(t, y(t), y'(t), \ldots, y^{(m-1)}(t)) \qquad a \le t \le b$$

(2.5b) $$g(y(a), \ldots, y^{(m-1)}(a), y(b), \ldots, y^{(m-1)}(b)) = 0$$

where (2.5b) involves m-vectors g and 0 corresponding to the BCs.

In the linear case, (2.5a,b) simplifies to

$$\text{(2.6a)} \qquad y^{(m)}(t) = \sum_{j=1}^{m-1} a_j(t)\, y^{(j)}(t) + f(t) \qquad a \le t \le b$$

$$\text{(2.6b)} \qquad \sum_{l=0}^{m-1} [\, b_{jl}\, y^{(l)}(a) + c_{jl}\, y^{(l)}(b) \,] = \gamma_j \qquad 1 \le j \le m \quad .$$

As in the IVP case, (2.5a,b) and (2.6a,b) can be converted to the first order systems (2.4a,b) and (2.3a,b), respectively, where the unknown solution is $\underset{\sim}{y}(t) = (y, y', \ldots, y^{(m-1)})^T$.

The most general BVP we consider involves a system of ODEs of different orders with multipoint BC which is called a mixed order system and can be written as

$$\text{(2.7a)} \qquad y_i^{(m_i)}(t) = f_i(t, y_1, \ldots, y_1^{(m_1-1)}, y_2, \ldots, y_d^{(m_d-1)}) \qquad a \le t \le b$$

$$= f_i(t, \underset{\sim}{z}(\underset{\sim}{y}(t)))$$

$$\text{(2.7b)} \qquad g_i(\underset{\sim}{z}(\underset{\sim}{y}(t_i))) = 0 \qquad 1 \le i \le n$$

where $\underset{\sim}{y}(t) = (y_1(t), \ldots, y_d(t))^T$

$$\underset{\sim}{z}(\underset{\sim}{y}(t)) := (y_1(t), y_1'(t), \ldots, y_1^{(m_1-1)}(t), y_2(t), \ldots, y_2^{(m_2-1)}(t), \ldots, y_d^{(m_d-1)}(t))^T \qquad a = t_1 \le t_2 \le \ldots \le t_n = b \text{ and } n := \sum_{i=1}^{d} m_i .$$

2.2. Existence and Uniqueness Theory

Before introducing numerical methods for solving the above problems, existence and uniqueness theorems of the solutions of the problems are given in the following.

Thm 2.1: Let $\underset{\sim}{f}(t,\underset{\sim}{y})$ be continuous on $D=\{(t,\underset{\sim}{y}):\ a\le t\le b,\ \|y-\alpha\|\le R\}$, where $\|.\|$ is some vector norm, and satisfy a Lipschitz condition with respect to $\underset{\sim}{y}$ on D, that is, there is a constant K, such that for any $(t,\underset{\sim}{y})$ and $(t,\underset{\sim}{z})$ in D

$$\|\underset{\sim}{f}(t,\underset{\sim}{y})-\underset{\sim}{f}(t,\underset{\sim}{z})\| \le K\|\underset{\sim}{y}-\underset{\sim}{z}\|.$$

If $\|\underset{\sim}{f}(t,\underset{\sim}{y})\| \le M$ on D and $c=\min\{\,b-a, R/M\,\}$, then (2.1) has a unique solution for $a\le t\le a+c$ (see Keller [12]).

Unfortunately, since it is extremely difficult to provide a result like the above for general BVPs, there are only existence and uniqueness theorems restricted for BVPs in special cases. Many of them relate to an important class, the second order BVP

(2.8a) $\quad y''(t)=f(t,y,y') \qquad a\le t\le b$

(2.8b) $\quad y(a)=\alpha_1 \quad , \quad y(b)=\alpha_2 \qquad .$

For instance: (Bailey-Shampine-Waltman [5])

Thm.2.2 : Suppose that $f(t,y,y')$ is continuous on $D=[a,b]\times(-\infty,\infty)\times(-\infty,\infty)$ and satisfies there a Lipschitz condition, i.e. there exist constants L and K such that for every (t,y,y') and (t,z,z') in D

$$\|f(t,y,y')-f(t,z,z')\| \le K\|y-z\| + L\|y'-z'\|.$$

If $b-a<2\alpha(L,K)$ where

$$
\alpha(L,K)=\begin{cases}
\dfrac{2}{(4K-L^2)}\cos^{-1} L/2\,K & \text{if } 4K-L^2>0\\[2ex]
\dfrac{2}{(L^2-4K)}\cosh^{-1} L/2\,K & \text{if } 4K-L^2<0\\ & \quad L>0, K>0\\[2ex]
2/L & \text{if } 4K-L^2=0, \quad L>0\\[2ex]
+\infty & \text{otherwise}
\end{cases},
$$

then (2.8) has a unique solution. This result is the best possible.

While for general linear BVPs, when the problem is expressed in terms of an associated IVP, a general theorem is possible:

Thm. 2.3 Assume $A(t)$ and $\underset{\sim}{f}(t)$ are continuous in (2.3a). The BVP (2.3a,c) has a unique solution if and only if the matrix

$$
(2.9) \qquad Q := B_1 + \sum_{j=2}^{r} B_j Y(\eta_j)
$$

is nonsingular, in which case the solution is

$$
\underset{\sim}{y}(t)=Y(t)\,Q^{-1}\left(\underset{\sim}{\alpha}-\sum_{j=2}^{r} B_j Y(\eta_j)\int_a^{\eta_j} Y^{-1}(u)\,\underset{\sim}{f}(u)\,du\right)+\hat{\underset{\sim}{y}}(t).
$$

where $\hat{y}(t)=Y(t)\int_a^t Y^{-1}(u)\underset{\sim}{f}(u)\,du$, and $Y(t)$ is the fundamental solution of (2.3a) which satisfies $Y'(t)=A(t)Y(t)$ $a\le t\le b$ and $Y(a)=I$ (see Keller [12]).

Proof: Let $Y(t)$ be the fundamental solution that satisfies the above conditions. By direct substitution, it can be shown that

$$y(t)=Y(t)[\beta+\int_a^t Y^{-1}(u)f(u)du]=Y(t)\beta+\hat{y}(t)$$

is the unique solution of (2.3). To satisfy the BC (2.3c), must be chosen such that

$$Q\beta+\sum_{j=2}^{r} B_j Y(\xi_j)\int_a^{\xi_j} Y^{-1}(u)f(u)du=\alpha \quad .$$

Hence (2.3a,c) has a unique solution if and only if Q is nonsingular.

2.3. Problem Stability

In this section, we discuss the stability properties of IVPs and BVPs. In general, when computing a quantity y from data t by some numerical method M, a problem is called unstable or ill-conditioned if "small" changes in the data t produce "large" changes in the solution y even if the method M is executed with no rounding error. The method M is called numerically unstable if small rounding errors introduced when using M produces large errors in the solution y even when the data are exact. As a rule, one should not try to compute numerically unstable quantities, and one should not use numerically unstable methods (Franklin [10]).

Definition: Consider the initial value problem

(2.10a) $y'(t)=f(t,y)$ $t>a$

(2.10b) $y(a)=\alpha$

where $y(t)=(y_1(t),\ldots,y_n(t))^T$.

A solution $\underset{\sim}{y}(t)$ is called **stable** (with respect to change in the initial conditions $\underset{\sim}{y}(a)$) if given any $\varepsilon > 0$, there is a $\delta > 0$ so that any other solution $\underset{\sim}{y}(t)$ of (2.10a) for which

$$\| \underset{\sim}{y}(a) - \hat{\underset{\sim}{y}}(a) \| \leq \delta$$

satisties

(2.11) $\quad \| \underset{\sim}{y}(t) - \hat{\underset{\sim}{y}}(t) \| \leq \varepsilon \quad$ for all $t > a$.

The solution $\underset{\sim}{y}(t)$ is **asymptotically stable** if, in addition to (2.11)

(2.12) $\quad \| \underset{\sim}{y}(t) - \hat{\underset{\sim}{y}}(t) \| \to 0$ as $t \to \infty$

and $y(t)$ is **relatively stable** if, instead of (2.11)

(2.13) $\quad \| \underset{\sim}{y}(t) - \hat{\underset{\sim}{y}}(t) \| \leq \varepsilon \| \underset{\sim}{y}(t) \| \quad$ for all $t > a$

For the special case of the linear constant coefficient problem

(2.10c) $\quad \underset{\sim}{y}'(t) = A \underset{\sim}{y}(t) \quad\quad t > a$

stability can be fairly easily characterized (see Ortega [18]).

Thm 2.4 The unique solution of (2.10c,b) is stable if and only if all eigenvalues of A have nonpositive real part and any eigenvalues with zero real part belongs to a 1x1 Jordan block. Furthermore, $\underset{\sim}{y}(t)$ is asymptotically stable if and only if all eigenvalues of A have negative real part (Ortega [18] and Franklin [11]).

Thm.2.5 If s is the set of eigenvalues of A with maximal real part, then the solution $\underset{\sim}{y}(t)$ of (2.10c,b) is relatively stable if and only if $\underset{\sim}{\alpha}$ has a componant in the direction of a principal vector for some $\lambda_i \in s$, and this vector has the maximal degree associated with $\underset{\sim}{\alpha}$ (Ortega [18]).

When A is a function of t, the eigenvalues of the Jacobian of A(t) can change sign, so it is only possible to give a characterization of stability corresponding to Thm.2.4 which is local.

Now, consider the stability theory for BVPs. For simplicity, only the two point linear BVP (2.3a,b) is considered.

Assume (2.3 a,b) has a unique solution $\underset{\sim}{y}(t)$ and also, the matrices B_a and B_b in (2.3b) are scaled such that

(2.14) $$\|B_a\| = \|B_b\| = 1 \quad ,$$

then (2.3 a,b) is stable if for any $\varepsilon > 0$, there exists a $\delta > 0$ such that the following is satisfied:

if $\max\{\|\delta B_a\|, \|\delta B_b\|, \|\delta \underset{\sim}{\alpha}\|\} \le \delta$ there is a $b^* > a$ (independent of ε and δ) such that for any $b \ge b^*$ for which both (2.3 a,b) and the BVP (2.3 a),

(2.3b') $$\hat{B}_a \hat{\underset{\sim}{y}}(a) + \hat{B}_b \hat{\underset{\sim}{y}}(b) = \hat{\underset{\sim}{\alpha}}$$

where $\hat{B}_a = B_a + \delta B_a$, $\hat{B}_b = B_b + \delta B_b$ and $\hat{\underset{\sim}{\alpha}} = \underset{\sim}{\alpha} + \delta\underset{\sim}{\alpha}$ are well-posed, then the respective solution $\underset{\sim}{y}(t)$ and $\hat{\underset{\sim}{y}}(t)$ satisfy

(2.15) $$\|\underset{\sim}{y} - \hat{\underset{\sim}{y}}\| := \max_{a \le t \le b} \|\underset{\sim}{y}(t) - \hat{\underset{\sim}{y}}(t)\| < \varepsilon \quad .$$

If we express the solution $\underset{\sim}{y}(t)$ of (2.3 a,b) as

(2.16) $$\underset{\sim}{y}(t) = Y(t)\underset{\sim}{s} + \underset{\sim}{v}(t) \qquad a \le t \le b$$

where Y(t) is the n x n fundamental solution matrix which satisfies

(2.17) $Y' = A(t)Y \qquad a \le t \le b$

(2.17a) $Y(a) = I$

and $\underset{\sim}{v}(t)$ is a particular solution of (2.3 a).

Substituting (2.16) into the BC (2.3 b), we find that $\underset{\sim}{s}$ is required to satisfy

(2.18) $Q\underset{\sim}{s} = \hat{\underset{\sim}{\alpha}}$

where $Q = [B_a + B_b Y(b)]$

$$\hat{\underset{\sim}{\alpha}} = \underset{\sim}{\alpha} - B_a \underset{\sim}{v}(a) - B_b \underset{\sim}{v}(b) \quad .$$

Let $\tilde{y}(t)$ solve the differential equation (2.3 a) subject to the perturbed BC

(2.19) $\hat{B}_a \underset{\sim}{y}(a) + \hat{B}_b \underset{\sim}{y}(b) = \hat{\underset{\sim}{\alpha}}$

where $\delta B_a := \hat{B}_a - B_a$, $\delta B_b := \hat{B}_b - B_b$, and $\delta\underset{\sim}{\alpha} := \hat{\underset{\sim}{\alpha}} - \underset{\sim}{\alpha}$ are "small".

From Thm.2.3, since the nonsingularity of problem (2.3 a,b) is equivalent to the nonsingularity of the matrix Q defined in (2.18) which is the special case of (2.9) with r=2, $\xi_1 = a$, and $\xi_2 = b$, it is tempting to take the condition number of Q

$$\text{cond}(Q) := \|Q\| \, \|Q^{-1}\|$$

as an indication for the condition number of the BVP. This quantity however, turns out to be rather misleading at times. The reason is that Q contains the effects of the fundamental matrix Y(t) which is the solution of (2.17) subject to the initial values (2.17a), rather than the BC (2.3b). Thus, if the IVP behaves very differently than the BVP, Q may be ill-conditioned even when the problem (2.3a,b) is not.

In general, write $\hat{y}(t)$ as

$$\hat{y}(t) = Y(t)\hat{s} + v(t) \qquad a \le t \le b$$

and define $\delta s := \hat{s} - s$. We get

$$Q \cdot \delta s = \delta\alpha := \alpha - \hat{\alpha} + \delta B_a y(a) + \delta B_b y(b) .$$

Now, the relevant quantity is

$$\hat{y}(t) - y(t) = Y(t)\delta s, \qquad \text{not } \delta s \text{ alone,}$$

so an indication of the condition of the problem (2.3 a,b) is the number

$$(2.20) \qquad k_1 := \max_{a \le t \le b} \| Y(t) Q^{-1} \|$$

rather than cond(Q).

The solution $y^*(t)$ of (2.3a,b) can also be expressed as

$$y^*(t) = Y(t) Q^{-1} \alpha + \int_a^b G(t,s) f(s) \, ds$$

where Q is in (2.18), G(t,s) is the Green's function for (2.3a,b).

The perturbations in the BCs give perturbed solutions which are related to $\|y^*(a)\|$, $\|y^*(b)\|$, and hence to G(t,s). Therefore, conditioning of (2.3a,b) is also related to the boundedness of the Green's function G(t,s)

$$(2.21) \qquad k_2 = (b-a) \operatorname*{lub}_{\substack{a \le t \le b \\ a \le s \le b}} \| G(t,s) \|$$

(Lentini-Osborne-Russell [15]).

It is obvious that k_1 doesn't depend on the particular choice of fundamental matrix Y. If $\Phi(t)$ is any n x n fundamental matrix satisfies (2.17), then there is a nonsingular n x n matrix P such that $\Phi(t)=Y(t)P$. Hence,

$$k_1 = \max_{a\le t\le b} \|Y(t)Q^{-1}\| = \max_{a\le t\le b} \|Y(t)[B_a Y(a) + B_b Y(b)]^{-1}\|$$

$$= \max_t \|\Phi(t)P^{-1}[B_a \Phi(a)P^{-1} + B_b \Phi(b)P^{-1}]^{-1}\|$$

$$= \max_t \|\Phi(t)P^{-1}(P^{-1})^{-1}[B_a \Phi(a) + B_b \Phi(b)]^{-1}\|$$

$$= \max_t \|\Phi(t)[B_a \Phi(a) + B_b \Phi(b)]^{-1}\| .$$

Therefore, to estimate the condition number of (2.3 a,b), it is tempting to use the bound

$$\max_t \|\Phi(t)\| \; \|[B_a \Phi(a) + B_b \Phi(b)]^{-1}\| .$$

If we choose $\Phi(a)=I$, i.e. $\Phi(t)=Y(t)$, then this bound will frequently be misleading. Indeed, if $Y(t)$ increases exponentially as t increases, then this bound is approximately cond(Q). To obtain a more realistic estimate, $\Phi(t)$ must be properly scaled. In particular, let

$\Phi(t)=[\varphi_1(t),\ldots,\varphi_n(t)]$ be a fundamental solution matrix

$$\max_{a\le t\le b} \|\Phi(t)\| = 1 \qquad \text{and}$$

(2.22)

$$\max_t \|\varphi_i(t)\| = \max_t \|\varphi_j(t)\| \quad 1\le i,j\le n .$$

Then the condition number of (2.3 a,b) can be approximated by

$$(2.23) \qquad k_1 = k_1(b) := \left\| [B_a \Phi(a) + B_b \Phi(b)]^{-1} \right\| .$$

Hence the result follows:

Suppose the BVP (2.3 a,b) has a unique solution, then (2.3 a,b) is well-conditioned if $k_1(b)=o(1)$ in (2.23) as $b \to \infty$, and ill-conditioned if $1/k_1(b)=o(1)$ as $b \to \infty$.

Example: (Lentini-Osborne-Russell [15])

Consider the BVP

$$y'' + y = 0, \qquad y(0) = 0 = y(b) ,$$

then after some computation, $k_1(b) = (1+\cos b)/\sin b$,

which means that for b away from multiple of π, the solution is not sensitive to small changes in the BCs. However, when b gets close to a multiple of π, the problem is unstable.

Stability properties of BVPs can also be indicated by the condition of the problems.

After investigating the existence and uniqueness of the solution of given problems and the stability of the problems, we then consider numerical methods for solving these problems.

3. Numerical Methods for Solving BVPs

Most of the interesting equations which occur in practice require that their solutions be obtained by numerical means. In this chapter, some well-known numerical methods for solving BVPs are discussed. First, initial value techniques, then finite difference methods, and finally finite element methods are discussed.

3.1 Initial Value Approaches

Initial value techniques play an important role in the numerical solution of BVPs. The basic process is to solve BVPs by solving IVPs with some arbitrary initial conditions, then find the solutions by satisfying the given BCs. In this section, several initial approaches such as superposition, shooting, stabilized march, and invariant imbedding are considered.

3.1.1. Superposition

Consider the linear BVP

(3.1a) $\underset{\sim}{y}'(t) = A(t)\underset{\sim}{y}(t) + \underset{\sim}{f}(t) \qquad a \le t \le b$

(3.1b) $B_a \underset{\sim}{y}(a) + B_b \underset{\sim}{y}(b) = \underset{\sim}{\alpha}$.

Due to the linearity, the solution $\underset{\sim}{y}(t)$ of (3.1) can be written as

$$(3.2)\qquad y(t)=y(t;s)=Y(t)s+v(t)$$

where $Y(t)=Y(t;a)$ is the fundamental solution matrix satisfying

$$Y'(t)=A(t)Y(t)\qquad a\le t\le b$$

$$Y(a)=I$$

and $v(t)$ is a particular solution of (3.1).

Here, s is to be determined so that the BC (3.1 b) are satisfied, so

$$\alpha=B_a[Y(a)s+v(a)]+B_b[Y(b)s+v(b)]$$

$$=[B_a+B_bY(b)]s+B_av(a)+B_bv(b)$$

or

$$(3.3)\qquad Qs=\hat{\alpha}$$

where $Q=B_a+B_bY(b)$ and

$$\hat{\alpha}=\alpha-B_av(a)-B_bv(b)\quad .$$

If Q is nonsingular, then s can be obtained from (3.3) and so the solution $y(t)$ is constructed. The above procedure is called the method of superposition. Generally, initial value methods which involve solving the ODE over [a,b] as an IVP, as the above does, are called shooting methods (see section 3.1.2).

When Q is formed, it is often ill-conditioned. Sometimes, this can be corrected by scaling, but realize that Q is nonetheless still ill-conditioned. For instance, consider the problem

$$y''=10000y\qquad y(0)=1,\qquad y(1)=e^{100}\quad .$$

A short computation gives

$$Q=B_a+B_bY(b)=\begin{pmatrix}1&0\\0&0\end{pmatrix}+\begin{pmatrix}0&0\\1&0\end{pmatrix}\begin{pmatrix}\cosh 100 & (\sinh 100)/100\\ 100(\sinh 100) & \cosh 100\end{pmatrix}$$

$$=\begin{pmatrix}1 & 0\\ \cosh 100 & \frac{(\sinh 100)}{100}\end{pmatrix}\approx\begin{pmatrix}1 & 0\\ \frac{e^{100}}{2} & \frac{e^{100}}{200}\end{pmatrix}.$$

Q is ill-conditioned although by scaling (multiply row 2 by e^{-100}), we get a well-conditioned problem.

When the BCs are partially separated, a **reduced superposition** method as decribed below can be used.

Suppose the partially separated BCs can be written as

(3.4a) $\quad C_a\,\underset{\sim}{y}(a)=\underset{\sim}{\alpha}_1$

(3.4b) $\quad D_a\,\underset{\sim}{y}(a)+D_b\,\underset{\sim}{y}(b)=\alpha_2$

where C_a is a p x n matrix of rank p, D_a and D_b are q x n matrices with p+q=n, and $\underset{\sim}{\alpha}_1$ and $\underset{\sim}{\alpha}_2$ are p- and q- vectors.

Write the solution of (3.1 a) and (3.4) as

$$\underset{\sim}{y}(t)=\underset{\sim}{y}(t;\underset{\sim}{s})=\bar{Y}(t)\,\underset{\sim}{\bar{s}}+\underset{\sim}{v}(t)$$

where $\bar{Y}(t)$ is an n x q matrix of fundamental solutions sastisfying

(3.5a) $\quad \bar{Y}'(t)=A(t)\bar{Y}(t)$ and

(3.5b) $\quad C_a\bar{Y}(a)=[\,0\,]_{p\times q}$

and the particular solution $\underset{\sim}{v}(t)$ satisfies

(3.6) $\quad C_a\,\underset{\sim}{v}(a)=\alpha_1$.

The q-vector $\underset{\sim}{s}$ is determined so that the q boundary conditions

(3.4 b) are satisfied, i.e.

$$\bar{Q}\bar{s}=\bar{\alpha}$$

where $\bar{Q}=[D_a\bar{Y}(a)+D_b\bar{Y}(b)]$ is a q x q matrix and

$\bar{\alpha}=\alpha_2-D_a v(a)-D_b v(b)$ a q-vector.

We need n linearly independant initial conditions to determine $v(t)$ and each column of $\bar{Y}(t)$, while (3.5 b) or (3.6) only supply with p conditions. Hence we augment C_a by a q x n matrix G such that

$$\hat{B}_a=\begin{pmatrix} C_a \\ G \end{pmatrix} \quad \text{is nonsingular.}$$

Then, require the initial condition

$$\hat{B}_a v(a)=\begin{pmatrix} \alpha_1 \\ 0 \end{pmatrix} \qquad \hat{B}_a Y(a)=\begin{pmatrix} 0 \\ I_q \end{pmatrix}$$

where I_q is the q x q identity matrix.

If we partition

$$B_a^{-1}=[F_1 \mid F_2]$$

where F_1 has p columns and F_2 has q columns,

then we have $v(a)=F_1\alpha_1$ and $\bar{Y}(a)=F_2$.

After $\bar{s}$ is obtained, the solution $y(t)$ is determined.

Although the superposition method is conceptually simple and works in many instances, it frequently gives very ill-conditioned problems even when the BVP is well-conditioned.

It has two major drawbacks: (Scott-Watts [23])

1. due to the finite word length used by computers, the solutions may lose their numerical independence. The resulting matrix problem in (3.3) may be so poorly conditioned that s cannot be determined accurately,
2. related to the finite word length of the computer, a loss of significance can occur even if the linear combination vector s has been computed accurately. This will normally occur if the fundamental solution $Y(t)$ is large compared to the desired solution.

To overcome these difficulties, one can use multiple shooting (see section 3.1.3) or the stabilized march method (see section 3.1.4) to get a very well-conditioned matrix (Mattheij [17]) since they maintain stability by restricting the intergrations to smaller intervals, or by keeping the solutions nearly mutually orthogonal thus guaranteeing their independence over the entire interval. Also, normalizing the vectors at the initial point of each of the subintervals controls the growth of solutions (Scott-Watts [23]).

3.1.2. Simple Shooting

Simple shooting for solving

(3.7a) $\underset{\sim}{y}' = \underset{\sim}{f}(t, \underset{\sim}{y}) \qquad a \le t \le b$

(3.7b) $\underset{\sim}{g}(\underset{\sim}{y}(a), \underset{\sim}{y}(b)) = \underset{\sim}{0}$

is the following: First guess the unknown initial values at a, say $\underset{\sim}{s}$. If we denote $\underset{\sim}{y}(t;\underset{\sim}{s})$ as the solution of (3.7) subject to the initial conditions $\underset{\sim}{y}(a;\underset{\sim}{s})=\underset{\sim}{s}$, then the problem reduces to finding a solution s^* to a system of n nonlinear algebraic equations:

$$\underset{\sim}{F}(\underset{\sim}{s})=\underset{\sim}{0} \qquad \text{where } \underset{\sim}{F}(\underset{\sim}{s}):=\underset{\sim}{g}(\underset{\sim}{s},\underset{\sim}{y}(b;\underset{\sim}{s})).$$

One can solve the above system by Newton's method i.e. given s_0, then solve

(3.8) $$J(\underset{\sim}{s}_k)\Delta\underset{\sim}{s}_k=-\underset{\sim}{F}(\underset{\sim}{s}_k), \qquad \underset{\sim}{s}_{k+1}=\underset{\sim}{s}_k+\Delta\underset{\sim}{s}_k, \quad k=0,1,\ldots$$

where $J(s)=\partial\underset{\sim}{F}/\partial\underset{\sim}{s}$.

Define

$$B_a=\frac{\partial\underset{\sim}{g}(\underset{\sim}{s};\underset{\sim}{y}(b,\underset{\sim}{s}))}{\partial\underset{\sim}{s}}$$

$$B_b=\frac{\partial\underset{\sim}{g}(\underset{\sim}{s};\underset{\sim}{y}(b,\underset{\sim}{s}))}{\partial\underset{\sim}{y}(b,\underset{\sim}{s}))}.$$

By (3.7a),

$$\left(\frac{\partial\underset{\sim}{y}}{\partial\underset{\sim}{s}}\right)'=\left(\frac{\partial\underset{\sim}{y}'}{\partial\underset{\sim}{s}}\right)=\frac{\partial\underset{\sim}{f}}{\partial\underset{\sim}{s}}=\frac{\partial\underset{\sim}{f}}{\partial\underset{\sim}{y}}\frac{\partial\underset{\sim}{y}}{\partial\underset{\sim}{s}},$$

and by the definition of $y(a;s)$,

$$\frac{\partial\underset{\sim}{y}(a;\underset{\sim}{s})}{\partial\underset{\sim}{s}}=\frac{\partial\underset{\sim}{s}}{\partial\underset{\sim}{s}}=I.$$

Then

$$Y(t)=\frac{\partial\underset{\sim}{y}(t;\underset{\sim}{s})}{\partial\underset{\sim}{s}} \qquad \text{satisfies}$$

(3.9a) $$Y'(t)=\frac{\partial\underset{\sim}{f}(t,\underset{\sim}{y}(t;\underset{\sim}{s}))}{\partial\underset{\sim}{y}}Y(t) \qquad a\le t\le b$$

(3.9b) $$Y(a)=I,$$

where

$$J(s)=\frac{\partial\underset{\sim}{F}(\underset{\sim}{s})}{\partial\underset{\sim}{s}}=\frac{\partial\underset{\sim}{g}(\underset{\sim}{s},\underset{\sim}{y}(b;\underset{\sim}{s}))}{\partial\underset{\sim}{s}}$$

$$= \frac{\partial g(s, y(b;s))}{\partial s} \frac{\partial s}{\partial s} + \frac{\partial g(s, y(b;s))}{\partial y(b;s)} \frac{\partial y(b;s)}{\partial s}$$

$$= B_a I + B_b Y(b)$$

$$= B_a + B_b Y(b) .$$

Thus, J(s) for the nonlinear case is like Q in the linear case.

For this method and for superposition, one can use the multiple shooting method to prevent the fundamental solution Y from becoming numerically dependent or unbounded. This and alternative methods are described below.

3.1.3. Multiple Shooting

The multiple shooting method is a generalization of the shooting method which is designed to avoid the build-up of errors arising from the computation of fundamental solutions over large intervals (Keller [12]).

This is done by dividing [a,b] into J subintervals, solving IVPs involving fundamental solutions and a particular solution over each subinterval independently, and then taking the final solution as an appropriate combination of these solutions which satisfies the boundary conditions and is continuous across interior points connecting subintervals.

A. Linear case:

Multiple shooting for solving

(3.10a) $\quad y'(t)=A(t)y(t)+f(t) \qquad a\le t\le b$

(3.10b) $$B_a \underset{\sim}{y}(a) + B_b \underset{\sim}{y}(b) = \underset{\sim}{\alpha}$$

takes the following general form :

divide $[a,b]$ into J subintervals $[t_j, t_{j+1}]$ $(1 \le j \le J)$ where

(3.11) $$a = t_1 < t_2 < t_3 < \ldots < t_J < t_{J+1} = b \quad .$$

The fundamental solution $Y_j(t)$ and particular solution $\underset{\sim}{v}_j(t)$ on (t_j, t_{j+1}) are obtained by solving

(3.12a) $$Y'_j(t) = A(t)\, Y_j(t) \qquad t_j \le t \le t_{j+1}$$

(3.12b) $$Y_j(t_j) = F_j$$

for some given $n \times n$ matrix F_j, and

(3.13a) $$\underset{\sim}{v}'_j(t) = A(t)\, \underset{\sim}{v}_j(t) + \underset{\sim}{f}(t) \qquad t_j \le t \le t_{j+1}$$

(3.13b) $$\underset{\sim}{v}_j(t_j) = \underset{\sim}{v}^0_j$$

for some given vector $\underset{\sim}{v}^0_j$ $\quad 1 \le j \le J$.

Then constant vectors $\underset{\sim}{c}_1, \ldots, \underset{\sim}{c}_J$ are determined so that on $[t_j, t_{j+1}]$ the approximation solution $\underset{\sim}{y}(t)$ defined by

(3.14) $$\underset{\sim}{y}(t) := Y_j(t)\, \underset{\sim}{c}_j + \underset{\sim}{v}_j(t) \qquad 1 \le j \le J$$

is continuous and satisfies the BCs.

Requiring the BC,

$$B_a (Y_1(t_1)\underset{\sim}{c}_1 + \underset{\sim}{v}_1(t_1)) + B_b (Y_J(t_{J+1})\underset{\sim}{c}_J + \underset{\sim}{v}_J(t_{J+1})) = \underset{\sim}{\alpha}$$

to be satisfied,

(3.15) $$B_a F_1 \underset{\sim}{c}_1 + B_b Y_J(b)\underset{\sim}{c}_J = \underset{\sim}{\alpha} - B_a \underset{\sim}{v}_1 - B_b \underset{\sim}{v}_J(b) \quad ,$$

and requiring continuity at t_j $(2 \le j \le J)$, i.e.

$$Y_j(t_j)\underset{\sim}{c}_j + \underset{\sim}{v}_j(t_j) = \underset{\sim}{y}(t_j) = Y_{j-1}(t_j)\underset{\sim}{c}_{j-1} + \underset{\sim}{v}_{j-1}(t_j)$$

yields the conditions

(3.16) $$F_j \underset{\sim}{c}_j + \underset{\sim}{v}_j^0 = Y_{j-1}(t_j)\underset{\sim}{c}_{j-1} + \underset{\sim}{v}_{j-1}(t_j) \quad .$$

Writing (3.15) and (3.16) in matrix form, we get

(3.17) $$M\underset{\sim}{c} = \begin{bmatrix} B_1 F_1 & & & & B_2 Y_J(b) \\ -Y_1(t_2) & F_2 & & & \\ & -Y_2(t_3), F_3 & & & \\ & \cdot & \cdot & & \\ & & \cdot & \cdot & \\ & & & -Y_{J-1}(t_J) & F_J \end{bmatrix} \begin{bmatrix} \underset{\sim}{c}_1 \\ \underset{\sim}{c}_2 \\ \cdot \\ \cdot \\ \cdot \\ \underset{\sim}{c}_J \end{bmatrix}$$

$$= \begin{bmatrix} \underset{\sim}{\alpha} - B_1 \underset{\sim}{v}_1^0 - B_2 \underset{\sim}{v}_J(b) \\ \underset{\sim}{v}_1(t_2) - \underset{\sim}{v}_2^0 \\ \underset{\sim}{v}_2(t_3) - \underset{\sim}{v}_3^0 \\ \cdot \\ \cdot \\ \underset{\sim}{v}_{J-1}(t_J) - \underset{\sim}{v}_J^0 \end{bmatrix} \quad .$$

For the "standard" multiple shooting case, $F_j = I$ and $v_j^0 = 0$ $(1 \le j \le J)$ (and therefore $c_j = y(t_j)$).

The simplified form for this method is

$$
Mc = \begin{bmatrix} B_1 & & & & B_2 Y_J(b) \\ -Y_1(t_2) & I & & & \\ & -Y_2(t_3), I & & & \\ & & \cdot & \cdot & \\ & & & \cdot \quad \cdot & \\ & & & -Y_{J-1}(t_J) & I \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \cdot \\ \cdot \\ \cdot \\ c_J \end{bmatrix}
$$

$$
= \begin{bmatrix} d - B_2 v_J(b) \\ v_1(t_2) \\ \cdot \\ \cdot \\ v_{J-1}(t_J) \end{bmatrix} .
$$

B. Nonlinear case :

Consider the nonlinear BVP

$$y'(t) = f(t, y(t)) \qquad a \le t \le b$$

$$g(y(a), y(b)) = 0 \quad .$$

As in case A, $[a,b]$ is divided into J subintervals $[t_j,$

$t_{j+1}]$ $1 \le j \le J$, and the IVPs

(3.18a) $\quad y_j'(t) = f(t, y_j(t)) \qquad t_j \le t \le t_{j+1}$

(3.18b) $\quad y_j(t_j) = c_j \qquad 1 \le j \le J$

are solved independently.

The constants c_j , $1 \le j \le J$, are determined such that $y(t)$ which is defined by

$$y(t) := y_j(t) \quad \text{on } [t_j, t_{j+1}] \qquad 1 \le j \le J$$

satisfies continuity and the BCs, i.e.

$$y_j(t_{j+1}) - y_{j+1}(t_{j+1}) = y_j(t_{j+1}) - c_{j+1} = 0 \qquad 1 \le j \le J-1$$

and

$$g(y_1(a), y_J(b)) = g(c_1, y_J(b)) = 0 \quad .$$

Then the matrix form

$$\text{(3.19)} \qquad \phi(c) := \begin{bmatrix} g(c_1, y_J(b; c_J)) \\ \cdot \\ \cdot \\ y_1(t_2; c_1) - c_2 \\ \cdot \\ \cdot \\ y_{J-1}(t_J; c_{J-1}) - c_J \end{bmatrix} = 0$$

where $c = (c_1, \ldots, c_J)^T$ and $y_j(t; c_j)$ is a solution to (3.18).

We can solve (3.19) by Newton's method, i.e. given c^0 , let

$$J(\underset{\sim}{c}^{i})\Delta \underset{\sim}{c}^{i} = -\phi(\underset{\sim}{c}^{i}), \quad \underset{\sim}{c}^{i+1} = \underset{\sim}{c}^{i} + \Delta \underset{\sim}{c}^{i} \qquad i=0,1,\ldots$$

where the Jacobian matrix $J(\underset{\sim}{c}) = \partial\phi/\partial \underset{\sim}{c}$.

Define

$$g_a(\underset{\sim}{y}(\cdot,\underset{\sim}{c})) = \frac{\partial \underset{\sim}{g}(\underset{\sim}{y}_1(a;\underset{\sim}{c}_1);\underset{\sim}{y}_J(b;\underset{\sim}{c}_J))}{\partial \underset{\sim}{y}_1(a;\underset{\sim}{c}_1)}$$

$$g_b(\underset{\sim}{y}(\cdot,\underset{\sim}{c})) = \frac{\partial \underset{\sim}{g}(\underset{\sim}{y}_1(a;\underset{\sim}{c}_1);\underset{\sim}{y}_J(b;\underset{\sim}{c}_J))}{\partial \underset{\sim}{y}_J(b;\underset{\sim}{c}_J)} .$$

From (3.18 a), we have

$$\left(\frac{\partial \underset{\sim}{y}_j}{\partial \underset{\sim}{c}_j}\right)' = \frac{\partial \underset{\sim}{y}_j'}{\partial \underset{\sim}{c}_j} = \frac{\partial \underset{\sim}{f}}{\partial \underset{\sim}{y}_j}\frac{\partial \underset{\sim}{y}_j}{\partial \underset{\sim}{c}_j} ,$$

and (3.18 b) gives

$$\frac{\partial \underset{\sim}{y}_j(t_j;\underset{\sim}{c}_j)}{\partial \underset{\sim}{c}_j} = \frac{\partial \underset{\sim}{c}_j}{\partial \underset{\sim}{c}_j} = I .$$

Then $Y_j(t;\underset{\sim}{c}_j) = \dfrac{\partial \underset{\sim}{y}_j(t;\underset{\sim}{c}_j)}{\partial \underset{\sim}{c}_j}$ satisfies

$$Y_j'(t;\underset{\sim}{c}_j) = \frac{\partial \underset{\sim}{f}(t,\underset{\sim}{y}_j(t;\underset{\sim}{c}_j))}{\partial \underset{\sim}{y}_j} Y_j(t;\underset{\sim}{c}_j) \qquad t_j \le t \le t_{j+1}$$

$$Y_j(t_j;\underset{\sim}{c}_j) = I \qquad 1 \le j \le J ,$$

and

$$J(\underset{\sim}{c}) = \begin{bmatrix} g_a(\underset{\sim}{y}(\cdot;\underset{\sim}{c})) & & & & g_b(\underset{\sim}{y}(\cdot;\underset{\sim}{c}))Y_J(b;\underset{\sim}{c}_J) \\ Y_1(t_2;\underset{\sim}{c}_1) & -I & & & \\ & Y_2(t_3;\underset{\sim}{c}_2) & -I & & \\ & & \cdot & \cdot & \\ & & & \cdot & \cdot \\ & & & Y_{J-1}(t_J;\underset{\sim}{c}_J) & -I \end{bmatrix}.$$

Thus, to solve $\phi(\underset{\sim}{c})=0$ by Newton's method, each iteration involves solving a linear multiple shooting problem.

Indeed, to get a stable problem, several multiple shooting codes select the shooting points by attempting to satisfy $\|Y_j(t_{j+1})\| \le k$ for some constant k. In which case cond(M) = $\|M\|\,\|M^{-1}\| \le (k+1)_1(k_1+k_2 J)$ where k_1 and k_2 are given in (2.20) and (2.21).

The two sources of error using multiple shooting are

1. approximating the fundamental solution matrices and particular solutions, and
2. solving the linear system of algebraic equations.

3.1.4. Stabilized March

The stabilized march procedure, like multiple shooting, is designed to maintain numerical linear independence of fundamental and particular solutions. While with multiple shooting the fundamental solution $Y_j(t)$ and the particular solution $\underset{\sim}{v}_j(t)$ are computed independently on each subinterval, this is not done with the stabilized march. The stabilized march also intends to economize on the multiple shooting method by reducing the number of fundamental solution components which must be computed in the same way that reduced superposition economizes on superposition (Scott-Watts [21]).

The procedure starts with $Y_1(a)$ and $\underset{\sim}{v}_1(a)$ satisfying (3.5b) and (3.6) given. Suppose we are given a fundamental solution $Y_j(t)$ and a particular solution $\underset{\sim}{v}_j(t)$ satisfying (3.5) and (3.6) for $t \geq t_j$. Let $Y_j(t_j) = F_j$, where F_j is a n x q matrix of rank q. When the solutions are becoming linearly dependent, say for $t = t_{j+1}$, then a factorization

$$(3.20) \qquad Y_j(t_{j+1}) = F_{j+1} \begin{bmatrix} 0 \\ P_{j+1} \end{bmatrix} = \left[F^1_{j+1} \,\middle|\, F^2_{j+1} \right] \begin{bmatrix} 0 \\ P_{j+1} \end{bmatrix}$$

is performed, where F^1_{j+1}, F^2_{j+1}, P_{j+1} are n x p, n x q, and q x q matrices. Let

$$E_{j+1} = F^{-1}_{j+1} = \begin{bmatrix} E^1_{j+1} \\ E^2_{j+1} \end{bmatrix} .$$

The initial data for $\underset{\sim}{v}_{j+1}(t)$ is $\underset{\sim}{v}_{j+1}(t_{j+1}) = \underset{\sim}{w}_{j+1}$, where $\underset{\sim}{w}_{j+1}$ can be specified by $E_{j+1}(\underset{\sim}{v}_j(t_{j+1}) - \underset{\sim}{w}_{j+1}) = 0$ and p additional conditions.

The process of computing a fundamental solution $Y_i(t)$ and initial values F_i and v_i^0 at t_i is continued for $i=j+2,\ldots$ until eventually the point $t=b$ is reached. As for reduced superposition, the computed solution is

$$(3.22) \qquad y(t) := Y_j(t)c_j + v_j(t) \qquad t_j \le t \le t_{j+1} .$$

If $b=t_{J+1}$, then requiring $y(t)$ to satisfy continuity and the BCs gives

$$Y_j(t_{j+1})c_j + v_j(t_{j+1}) = Y_{j+1}(t_{j+1})c_{j+1} + v_{j+1}(t_{j+1}) \qquad 1 \le j \le J-1$$

or

$$(3.23) \qquad F_{j+1}\begin{bmatrix} 0 \\ P_{j+1} \end{bmatrix} c_j + v_j(t_{j+1}) = F_{j+1}c_{j+1} + w_{j+1} .$$

Multiplying by $\begin{bmatrix} 0 \\ E^2_{i+1} \end{bmatrix}$, then

$$\begin{bmatrix} 0 \\ P_{j+1} \end{bmatrix} c_j = \begin{bmatrix} 0 \\ I_q \end{bmatrix} c_{j+1} - \begin{bmatrix} 0 \\ E^2_{j+1} \end{bmatrix} (v_j(t_{j+1}) - w_{j+1}) .$$

The last q BCs are $D_a y(a) + D_b y(b) = D_a(F_1 c_1 + w_1) + D_b(Y_J c_J + v_J(b)) = \alpha_2$. In matrix form , we have

$$
Mc := \begin{bmatrix} D_a F_1 & & & & D_b Y_J(b) \\ -P_2 & I & & & \\ & -P_3 & I & & \\ & & \ddots & \ddots & \\ & & & -P_J & I \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_J \end{bmatrix} = \begin{bmatrix} \alpha_2 - D_a w_1 - D_b v_{J+1} \\ E_2^2 (v_2 - w_2) \\ E_3^2 (v_3 - w_3) \\ \vdots \\ E_J^2 (v_J - w_J) \end{bmatrix} .
$$

where $v_j = v_{j-1}(t_j)$ $(1 \leq j \leq J)$.

The stabilized march is not amenable to parallel processing like multiple shooting. However, it does compute a smaller solutions set and the multiple shooting points can be easily selected automatically. It also has the advantage that rapidly decaying fundamental solution information is only needed in the particular solution, viz one may be able to adapt an initial value solver such that it ignores these rapidly decreasing components in all integrations except that for this one solution vector (Lentini-Osborne-Russell [15]).

3.1.5. Invariant Imbedding

For simplicity, I only consider the BVP with linear separated BC

(3.24a) $\quad y'(t)=H(t)y(t)+h(t) \qquad a\le t\le b$

(3.24b) $\quad B_a y(a)=\alpha, \quad B_b y(b)=\beta$.

Suppose (3.24) can be reformulated as

(3.25a) $$\begin{pmatrix} u(t) \\ v(t) \end{pmatrix}' = \begin{bmatrix} A(t) & B(t) \\ C(t) & D(t) \end{bmatrix} \begin{pmatrix} u(t) \\ v(t) \end{pmatrix} = \begin{pmatrix} f(t) \\ g(t) \end{pmatrix}$$

with BC

(3.25b) $$(K_0, K_1)\begin{pmatrix} u(a) \\ v(a) \end{pmatrix}=\alpha, \qquad (K_2, K_3)\begin{pmatrix} u(b) \\ v(b) \end{pmatrix}=\beta$$

where $u(t)$, $f(t)$, α are p-vectors, $v(t)$, $g(t)$, β are q-vectors, n=p+q, $A(t)$, K_0 are p x p matrices, $B(t)$, K_1 are p x q matrices, $C(t)$, K_2 are q x p matrices, and $D(t)$, K_3 are q x q matrices.

The key idea in invariant imbedding is to replace a two point BVP by a set of initial value problems. One standard way is to first express the solution of (3.25 a) in the form

(3.26) $\quad u(t)=R(t)v(t)+x(t)$

where $R(t)$ is a p x q matrix and $x(t)$ is a p-vector. Substitute (3.26) into (3.25 a) to get

$$[R'+RCR+RD-AR-B]v+[x'+RCx+Rg-Ax-f]=0$$

$$v'=CRv+Cx+Dv+g.$$

If we require that the coefficients of $v(t)$ vanish in the first set above, and that (3.26) satisfies the BC (3.25 b) with the coefficient of $v(a)$ set to zero,

then the solutions are found by solving three IVPs

$$(3.27a)\quad \begin{cases} R' = [A(t) - RC(t)]R - RD(t) + B(t) \\ K_0 R(a) + K_1 = 0 \end{cases},$$

$$(3.27b)\quad \begin{cases} \underset{\sim}{x}' = [A(t) - RC(t)]\underset{\sim}{x} - R\underset{\sim}{g}(t) + \underset{\sim}{f}(t) \\ K_0 \underset{\sim}{x}(a) = \underset{\sim}{\alpha} \end{cases},$$

and

$$(3.27c)\quad \begin{cases} \underset{\sim}{v}' = [D(t) + C(t)R]\underset{\sim}{v} + C(t)\underset{\sim}{x} + \underset{\sim}{g}(t) \\ [K_2 R(b) + K_3]\underset{\sim}{v}(b) = \underset{\sim}{\beta} - K_2 \underset{\sim}{x}(b) \end{cases} \quad \text{for } a \le t \le b .$$

The first equation of (3.27 a) is called a matrix Riccati equation (See Reid [19]).

Suppose $Y(t)$ is a fundamental solution matrix of (3.24 a) ,that is

$$Y' = H(t)\,Y \qquad a \le t \le b \qquad \det(Y(a)) \ne 0$$

where $H(t) = \begin{pmatrix} A(t) & B(t) \\ C(t) & D(t) \end{pmatrix}$.

If $Y(t)$ can be partitioned into submatrices as $H(t)$ can, i.e.

$$Y(t) = \begin{pmatrix} Y_0(t) & Y_1(t) \\ Y_2(t) & Y_3(t) \end{pmatrix} \begin{matrix} \}p \\ \}q \end{matrix},$$
$$\qquad\qquad \underbrace{}_{p} \quad \underbrace{}_{q}$$

and $Y(t)$ satisfies

(3.28)

a) $Y_3(t)$ if nonsingular on $[a,b]$

b) $[K_0 Y_1(a) + K_1 Y_3(a)] \ne 0$.

then the solution $R(t)$ of (3.27 a) may be taken as $R(t) =$

$Y_1(t)Y_3^{-1}(t)$ (Keller-Lentini [13]).

After $R(t)$, $\underset{\sim}{x}(t)$, and $\underset{\sim}{v}(t)$ have been found, $\underset{\sim}{u}(t)$ is easily found as a linear combination of them, and the solution $\underset{\sim}{y}(t)$ of the BVP (3.24) is merely

$$\underset{\sim}{y}(t) = \begin{pmatrix} \underset{\sim}{u}(t) \\ \underset{\sim}{v}(t) \end{pmatrix} .$$

Even though invariant imbedding has the disadvantage of giving nonlinear initial value problems (for the matrix Riccati equation), it may well overcome the following two related difficulties of multiple shooting and stabilized march :

1. when the related IVPs are unstable, short subintervals are necessary, and
2. in the presence of rapidly growing/decreasing solutions, scaling can be a constant problem.

3.2. Finite Difference Methods

Choose a mesh $\pi : a=t_1<t_2<....<t_N<t_{N+1}=b$. The basic idea of the method involves finding approximate solution values at these mesh points t_j by the following :

1. form a set of algebraic equations for the approximate solution values by replacing derivatives with difference quotients in the differential equations and the boundary conditions ,
2. solve the resulting system of equations for the approximate solution. (Keller [12])

I start with a simple example.

3.2.1. A Simple Scheme for a Second Order Problem

Consider the scalar BVP

(3.29a) $Lu(t)=-u''+a_1(t)u'+a_0(t)u=b(t) \quad 0\le t\le 1$

(3.29b) $u(0)=\alpha \; , \; u(1)=\beta$

where $a_1(t)$, $a_0(t)$ and $b(t)$ are continuous functions on $[0,1]$.

Take a uniform mesh for this problem ,i.e. $t_i=(i-1)h$, $i=1,2,\dots,N+1$, $h=1/N$.

Assume the exact solution of (3.29) exists. A mesh function $\{u_j\}_{j=1}^{N+1}$ is sought such that $u_i=u(t_i)$, $i=1,2,\dots,N+1$.

The differential equation (3.29 a) is approximated by

(3.30a) $$Lu(t_i)=L_N u_i=-\frac{u_{i+1}-2u_i+u_{i-1}}{h^2}+a_1(t_i)\frac{u_{i+1}-u_{i-1}}{2h}+a_0(t_i)u_i=b(t_i) \qquad 2\le i\le N,$$

and the BCs (3.29 b) give

(3.30b) $u_1=\alpha, \quad u_{N+1}=\beta$.

If (3.30) is written in matrix form, then the following tridiagonal system is obtained :

$$
\begin{bmatrix}
2/h^2+a_0(t_2), & -1/h^2+a_1(t_2)/2h & & & \\
-1/h^2-a_1(t_3)/2h, & 2/h^2+a_0(t_3), & -1/h^2+a_1(t_3)/2h & & \\
\cdot & \cdot & \cdot & & \\
& \cdot & \cdot & \cdot & \\
& -1/h^2-a_1(t_{N-1})/2h, & 2/h^2+a_0(t_{N-1}), & -1/h^2+a_1(t_{N-1})/2h & \\
& & & -1/h^2-a_1(t_N)/2h, & 2/h^2+a_0(t_N)
\end{bmatrix}
\begin{bmatrix}
u_2 \\ u_3 \\ \cdot \\ \cdot \\ u_{N-1} \\ u_N
\end{bmatrix}
=
\begin{bmatrix}
b(t_2)+\alpha(1/h^2+a_1(t_2)/2h) \\ b(t_3) \\ \cdot \\ \cdot \\ b(t_{N-1}) \\ b(t_N)+\beta(1/h^2-a_1(t_N)/2h
\end{bmatrix}.
$$

If $a_0(t)$ is positive, the approximate mesh function $\{u_j\}_{j=1}^{N+1}$ exists, the matrix is positive definite, and the process of Gauss elimination without pivoting for tridiagonal matrices is extremely simple and efficient and stable (Ascher-Russell [3]).

3.2.2. One Step Schemes :Trapezoidal Rule and Midpoint Rule

For a given first order system of DEs, e.g. $\underset{\sim}{y}'=\underset{\sim}{f}(t,\underset{\sim}{y})$, if one integrates both sides of the equations, then the left hand side of the integration directly gives the solutions of the DE. Most finite difference methods involve converting to first order systems and then selecting a discretization.

Consider the linear first order system

(3.31a) $\quad L\underset{\sim}{y}(t)=\underset{\sim}{y}'(t)-A(t)\underset{\sim}{y}(t)=\underset{\sim}{b}(t) \qquad a\le t\le b$

(3.31b) $\quad B[\underset{\sim}{y}(a),\underset{\sim}{y}(b)]=B_a\underset{\sim}{y}(a)+B_b\underset{\sim}{y}(b)=\underset{\sim}{\beta}$

where $A(t)$, B_a, and B_b are $n \times n$ matrices.

The two simplest one step schemes, which use only information about the approximate solution at t_i to obtain the approximate solution at t_{i+1}, are the trapezoidal method and midpoint method.

On a mesh π, a numerical solution $\{\underset{\sim}{y}_i\}_{i=1}^{N+1}$ is sought where $\underset{\sim}{y}_i$ is to approximate componentwise the exact solution $\underset{\sim}{y}(t)$ at $t=t_i$ and is required to satisfy the BCs

$$B[\underset{\sim}{y}_1,\underset{\sim}{y}_{N+1}]=B_a\underset{\sim}{y}_1+B_b\underset{\sim}{y}_{N+1}=\underset{\sim}{\beta} \quad .$$

The trapezoidal method is defined as

$$L_\pi \underset{\sim}{y}_i=\frac{\underset{\sim}{y}_{i+1}-\underset{\sim}{y}_i}{h_i}-1/2[A(t_{i+1})\underset{\sim}{y}_{i+1}+A(t_i)\underset{\sim}{y}_i]$$
$$=1/2[\underset{\sim}{b}(t_{i+1})+\underset{\sim}{b}(t_i)] , \qquad h_i=t_{i+1}-t_i , \quad 1\le i\le N$$

and the midpoint method or the Box scheme is defined by

$$L_\pi \underset{\sim}{y}_i=\frac{\underset{\sim}{y}_{i+1}-\underset{\sim}{y}_i}{h_i}-1/2A(t_{i+1/2})(\underset{\sim}{y}_{i+1}+\underset{\sim}{y}_i))=\underset{\sim}{b}(t_{i+1/2})$$

where $h_i=t_{i+1}-t_i$ and $t_{i+1/2}:=t_i+h_i/2$, $\quad 1\le i\le n$.

For the nonlinear problem

(3.32a) $\quad N\underset{\sim}{y}(t)=\underset{\sim}{y}'(t)-\underset{\sim}{f}(t;\underset{\sim}{y})=\underset{\sim}{0} \qquad a\le t\le b$

(3.32b) $\quad \underset{\sim}{g}(\underset{\sim}{y}(a),\underset{\sim}{y}(b))=\underset{\sim}{0},$

the trapezoidal rule is given by

(3.33) $\quad \underset{\sim}{g}(\underset{\sim}{y}_1,\underset{\sim}{y}_{N+1})=\underset{\sim}{0}$

(3.34) $$N_\pi \underset{\sim}{y}_i = \frac{y_{i+1}-y_i}{h_i} - 1/2[\underset{\sim}{f}(t_{i+1},\underset{\sim}{y}_{i+1})+\underset{\sim}{f}(t_i,\underset{\sim}{y}_i)]=\underset{\sim}{0}$$

$$1\le i\le N$$

and the midpoint rule is given by (3.33) and

(3.35) $$N_\pi \underset{\sim}{y}_i = \frac{\underset{\sim}{y}_{i+1}-\underset{\sim}{y}_i}{h_i} - \underset{\sim}{f}(t_{i+1/2};1/2(\underset{\sim}{y}_i+\underset{\sim}{y}_{i+1}))=\underset{\sim}{0}$$

again $h_i = t_{i+1}-t_i$ and $t_{i+1/2} := t_i + h_i/2$, $1\le i\le N$.

For the nonlinear problem (3.32), it is not difficult to form the difference schemes (3.33), (3.34), and (3.35). The difficulty is in solving the resulting $n(N+1)$ nonlinear algebraic equations, where n is the order of the first order system (3.32a) and N is the number of subintervals. For instance, $n=5$ and $N=200$ gives more than 1000 equations. We consider Newton's method to solve the nonlinear problems. From (3.33) and (3.34),

$$N_\pi \underset{\sim}{y}_1 = \frac{\underset{\sim}{y}_2-\underset{\sim}{y}_1}{h_1} - 1/2[\underset{\sim}{f}(t_1,\underset{\sim}{y}_1)+\underset{\sim}{f}(t_2,\underset{\sim}{y}_2)]=\underset{\sim}{0}$$

$$N_\pi y_2 = \frac{y_3 - y_2}{h_2} - 1/2[f(t_2, y_2) + f(t_3, y_3)] = 0$$

(3.36)

$$\vdots$$

$$N_\pi y_N = \frac{y_{N+1} - y_N}{h_N} - 1/2[f(t_N, y_N) + f(t_{N+1}, y_{N+1})] = 0$$

$$g(y_N, y_{N+1}) = 0 \quad .$$

Letting the system of equations (3.36) be $F(y)=0$ where

$F = (N_\pi y_1, N_\pi y_2, \ldots\ldots)^T$, then

$$\frac{\partial F(t, y^*)}{\partial y} = \begin{bmatrix} -\frac{1}{h_1} I - \frac{1}{2}\frac{f(t_1, y_1)}{y}, & \frac{1}{h_1} I - \frac{1}{2}\frac{\partial f(t_2, y_2)}{\partial y} & & \\ & \ddots & \ddots & \\ & & -\frac{1}{h_N} I - \frac{1}{2}\frac{\partial f(t_N, y_N)}{\partial y}, & \frac{1}{h_N} I - \frac{1}{2}\frac{\partial f(t_{N+1}, y_{N+1})}{\partial y} \\ \frac{\partial g(y_1, y_{N+1})}{\partial y(a)} & & & \frac{\partial g(y_1, y_{N+1})}{\partial y(b)} \end{bmatrix}$$

$= J$.

Newton's method becomes solving for w from the following

$$(3.37)\qquad \frac{\partial F}{\partial Y}w = J\cdot\begin{bmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ w_{N+1} \end{bmatrix} = -F(y^m).$$

If written component-wise, (3.37) gives

$$(3.38)\qquad \frac{w_{i+1}-w_i}{h_i} - 1/2[A(t_{i+1})w_{i+1} + A(t_i)w_i] = -N_\pi y_i^m ,$$

and

$$(3.39)\qquad B_a w_1 + B_b w_{N+1} = -g(y_1^m, y_{N+1}^m)$$

where $\{y_i^m\}_{i=1}^{N+1}$ are known values from a former iteration,

$$(3.40)\qquad A(t_j) = \frac{\partial f}{\partial Y}(t_j, y_j^m),$$
$$B_a = \frac{\partial g}{\partial Y(a)}(y_1^m, y_{N+1}^m) \quad , \quad B_b = \frac{\partial g}{\partial Y(b)}(y_1^m, y_{N+1}^m),$$

and the next iterate is given by $y_i^{m+1} := y_i^m + w_i$, $i=1,\ldots,N+1$.

The system (3.38), (3.39), and (3.40) is a linear systems of equations for the correction vector $\{w_i\}_{i=1}^{N+1}$,which looks like a trapezoidal discretization of some linear problems. In each iteration we have performed two operations in succession, discretization and linearization.

Let $y^{(m)}(t)$ be an appropriately smooth function satisfying $y^{(m)}(t_i)=y_i^m$, $i=1,\ldots,N+1$. If we first linearize the differential problem

$$Ny = y' - f(t,y) = Dy - f(t,y),$$

Newton's method is :

given y^0, solve

$$\frac{\partial N y^m}{\partial y} w = -N y^m, \qquad (3.41)$$

and let $y^{m+1} = y^m + w$, $m=0,1,2,\ldots$.

Since $\frac{\partial N}{\partial y} = D\cdot - \frac{\partial f}{\partial y}$,

(3.41) is

$$Dw - \frac{\partial f(t,y^m)}{\partial y} w = w' - \frac{\partial f(t,y^m)}{\partial y} w$$

$$= -Ny^m = -y^{m\prime} + f(t,y^m).$$

Now, if discretization (trapezoidal rule) is applied to solve these equations, from linearizing the differential equation, we get (3.38) and (3.39). Thus, the two operations of linearization and discretization are commutative for the trapezoidal scheme. The iterative scheme where we first linearize and then discretize is called **quasilinearization**.

3.2.3. Runge-Kutta Schemes

The Runge-Kutta methods involve using only information about the approximate solution at t_i to obtain the approximate solution at t_{i+1}. The general form of a **k-stage Runge-Kutta** scheme for (3.32) on a mesh is

(3.42) $$\underset{\sim}{y}_{i+1} = \underset{\sim}{y}_i + h_i \sum_{j=1}^{k} \beta_j \underset{\sim}{f}_{ij}$$

(3.43) $$\underset{\sim}{f}_{ij} = \underset{\sim}{f}(t_{ij}; \underset{\sim}{y}_i + h_i \sum_{l=1}^{k} \alpha_{jl} \underset{\sim}{f}_{il})$$

(3.44) $$\underset{\sim}{g}(\underset{\sim}{y}_1, \underset{\sim}{y}_{N+1}) = \underset{\sim}{0}$$

where $\{t_{ij}\}$ are defined as $t_{ij} := t_i + h_i \rho_j$, $1 \le j \le k$, $1 \le i \le N$ with $0 \le \rho_1 < \rho_2 < \ldots < \rho_k \le 1$, the "canonical points".

The method is called **explicit** if $\alpha_{jl} = 0$ $j \ge l$ and **implicit** otherwise. Both the trapezoidal rule and midpoint rule are implicit Runge-Kutta schemes. The first is 2-stage and the second is 1-stage.

3.3. Finite Element Methods

Like the finite difference methods, finite element methods attempt to find approximate solutions of a BVP at a discrete set of points by satisfying the BC and ODE simultaneously throughout the interval. However, when solved by finite element methods, differential equations need not be converted to a first order system.

Consider the scalar problem

$$\text{(3.45a)} \qquad Ly(t) := y^{(m)}(t) - \sum_{j=0}^{m-1} a_j(t)\, y^{(j)}(t) = q(t) \qquad a \le t \le b$$

$$\text{(3.45b)} \qquad \sum_{l=0}^{m-1} [b_{jl}\, y^{(l)}(a) + c_{jl}\, y^{(l)}(b)] = r_j \qquad 1 \le j \le m.$$

For finite element methods, the approximate solution is a spline function $s(t) \in P_{k,\pi,l}$, where $P_{k,\pi,l}$ is a collection of spline functions which are of order k (degree less than or equal to k-1) in each subinterval of π: $a = t_1 < t_2 < \ldots < t_N < t_{N+1} = b$ and are lth order continuous at every mesh point. The advantages of selecting a spline space $P_{k,\pi,l}$ are that high order methods result and local basis representations produce banded matrix equations. For convenience, assume that $s(t) \in \overset{\circ}{P}_{k,\pi,l}$, the subspace of $P_{k,\pi,l}$ consisting of spline functions which satisfy the BC (3.45b). So the unknown solution parameters correspond to some representation for s(t). Letting the approximate solution be $s(t) = \sum_{j=1}^{M} \alpha_j \psi_j(t)$ where $\{\psi_j(t)\}$ is a basis of $\overset{\circ}{P}_{k,\pi,l}$, $M = \dim(\overset{\circ}{P}_{k,\pi,l})$, we determine α_j, j=1,...,M by requiring s(t) to satisfy the differential equations in one of several natural ways. The basic types of finite element methods are collocation, Galerkin, least squares, and Ritz methods which are described below.

3.3.1 Collocation Methods

For solving (3.45), the collocation solution $s(t)$ is required to satisfy the differential equation (3.45a) exactly at M points $\{z_i\}_{i=1}^{M}$ (called the collocation points) in $[a,b]$, i.e.

$$(3.46) \qquad Ls(z_i) = L\left(\sum_{j=1}^{M} \alpha_j \psi_j(z_i)\right) = q(z_i) \qquad 1 \le i \le M.$$

Thus collocation requires the residual $r(t) := Ls(t) - q(t)$ be set to zero at M points. In matrix form, (3.46) is

$$(3.47) \qquad C\underset{\sim}{\alpha} = \underset{\sim}{q}$$

where $C := (c_{ij})_{i,j=1}^{M} = (L\psi_j(z_i))_{i,j=1}^{M}$

$$\underset{\sim}{q} := (q(z_1), \ldots, q(z_M))^T$$

$$\underset{\sim}{\alpha} := (\alpha_1, \ldots, \alpha_M)^T \quad .$$

Suppose the collocation points z_k are in the jth subinterval and can be expressed as

$$z_k = \tau_{ji} = \tau_{j+1/2} + \frac{h_j}{2} \rho_i$$

where $h_j = t_{j+1} - t_j$, $\tau_{j+1/2} = t_j + h_j/2$.

Let $P_k(t)$ be the Gauss Legendre polynomial of degree k. If ρ_i are chosen to be the zeros of $P_l(t)$, $1 \le i \le l$, and to satisfy $-1 < \rho_1 < \rho_2 < \ldots < \rho_l < 1$, then z_k are called the **Gauss points**. If ρ_i is the (i-1)st zero of $[P_{l-1}(t) + P_l(t)]/(t-1)$,

$2 \le i \le l$, and $-1=\rho_1<\rho_2<\ldots<\rho_l<1$, then z_k are called the **Radau points**. In the case that ρ_i is the (i-1)st zero of $P'_{l-1}(t)$, $2 \le i \le l-1$, and $-1=\rho_1<\rho_2<\ldots<\rho_l=1$, then z_k are called the **Lobatto points**.

3.3.2. Galerkin Methods

For the Galerkin method, the approximate solution $\bar{s}(t)= \sum_{j=1}^{M} \bar{\alpha}_j \psi_j(t)$ is determined so that the differential equation (3.45a) is satisfied in the sense that

$$\int_a^b L\bar{s}(t)\psi_i(t)\,dt=\int_a^b q(t)\psi_i(t)\,dt \qquad 1 \le i \le M . \tag{3.48}$$

That is, one requires the residual $\bar{r}(t)=L\bar{s}(t)-q(t)$ to satisfy

$$\int_a^b \bar{r}(t)\eta(t)\,dt=0 \qquad \text{for all } \eta(t)\in \overset{\circ}{P}_{k,\pi,l} .$$

In matrix form, (3.48) is $G\underset{\sim}{\bar{\alpha}}=\underset{\sim}{\bar{q}}$,

where $G:=(g_{ij})_{i,j=1}^{M}=\left(\int_a^b L\psi_j(t)\psi_i(t)\,dt\right)_{i,j=1}^{M}$

$$\begin{aligned}\underset{\sim}{\bar{q}}&:=(\bar{q}_1,\ldots,\bar{q}_M)^T\\ &=\left(\int_a^b q(t)\psi_1(t)\,dt,\ldots,\int_a^b q(t)\psi_M(t)\,dt\right)^T\end{aligned} \tag{3.49}$$

$$\underset{\sim}{\bar{\alpha}}:=(\bar{\alpha}_1,\ldots,\bar{\alpha}_M)^T .$$

Unless the BVP is extremely simple, the elements of G and $\bar{q}$ must be approximated by a numerical quadrature. If the quadrature rule has the form

$$\int_a^b f(t)\,dt = \sum_{l=1}^{Q} w_l f(z_l) \quad , \tag{3.50}$$

the resulting discrete Galerkin method solution $\bar{s}^*(t) = \sum_{j=1}^{M} \bar{\alpha}_j^* \psi_j(t)$ is obtained from the system of equations

$$G^* \bar{\alpha}^* = \bar{q}^* \tag{3.51}$$

where

$$G^* = (g^*_{ij})_{i,j=1}^{M} = \left(\sum_{l=1}^{Q} w_l L\psi_j(z_l)\psi_i(z_l) \right)_{i,j=1}^{M}$$

$$\bar{q}^* = (\bar{q}_1^*, \ldots, \bar{q}_M^*)^T$$

$$= \left(\sum_{l=1}^{M} w_l q(z_l)\psi_1(z_l), \ldots, \sum_{l=1}^{M} w_l q(z_l)\psi_M(z_l) \right)^T$$

$$\bar{\alpha}^* = (\bar{\alpha}_1^*, \ldots, \bar{\alpha}_M^*)^T \quad .$$

The discrete Galerkin equations (3.51) can be written as

$$BDC\bar{\alpha}^* = BDq \tag{3.52}$$

where

$$B := (b_{ij})_{i=1,j=1}^{M,Q} = (\psi_i(z_j))_{i,j=1}^{M,Q}$$

$$D := \mathrm{diag}(w_i)_{i=1}^{Q}$$

and C, q are in (3.47).

If $Q=M$ and B is nonsingular, then the collocation and discrete Galerkin methods are equivalent. Compared with the collocation method, the Galerkin method has the disadvantage that the integral coefficients must be evaluated and the advantage that, for the same order of convergence, smoother spline functions can be used.

3.3.3. Least Squares Method

The Least squares method is to find $\hat{s}(t)=\sum_{j=1}^{M}\hat{\alpha}_j\psi_j(t)$ such that $E(\hat{s})=\min_{s\in P^{o}_{k,\pi,1}} E(s)$ where

$$E(s)=E(\alpha_1,\ldots,\alpha_M):=\int_a^b [Ls(t)-q(t)]^2\,dt \quad .$$

By setting $\frac{\partial E}{\partial \alpha_i}=0 \quad 1\le i\le M$,

we find this is equivalent to requiring

$$\int_a^b [L\hat{s}(t)-q(t)]L\psi_i(t)\,dt=0 \qquad 1\le i\le M$$

or

$$\int_a^b [L\hat{s}(t)]L\psi_i(t)\,dt=\int_a^b q(t)L\psi_i(t)\,dt \qquad 1\le i\le M \; . \tag{3.53}$$

In matrix form, (3.53) is

$$\hat{\underset{\sim}{H}}\hat{\alpha}=\hat{\underset{\sim}{q}} \tag{3.54}$$ where

$$\hat{H} = (h_{ij})_{i,j=1}^{M} = \left(\int_a^b L\psi_i(t) L\psi_j(t)\, dt\right)_{i,j=1}^{M}$$

(3.55) $$\hat{\underset{\sim}{q}} = (\hat{q}_1, \ldots, \hat{q}_M)^T$$

$$= \left(\int_a^b L\psi_1(t) q(t)\, dt, \ldots, \int_a^b L\psi_M(t) q(t)\, dt\right)^T$$

$$\hat{\underset{\sim}{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_M)^T \quad .$$

As with the Galerkin method, the discrete least squares method involves evaluating the integral coefficients in (3.55) by a numerical quadrature rule of the form (3.50).

The solution is $\hat{s}^*(t) = \sum_{j=1}^{M} \hat{\alpha}_j^* \psi_j(t)$ if

(3.56) $$C^T D C \hat{\alpha}^* = C^T D \underset{\sim}{q}$$

where D, C, and $\underset{\sim}{q}$ are as in (3.52), (3.47) and $\hat{\underset{\sim}{\alpha}}^* = (\alpha_1^*, \ldots, \alpha_M^*)^T$.

Clearly, if Q=M and the collocation matrix is nonsingular, the discrete least squares and collocation methods are equivalent.

3.3.4. Ritz Method

Consider the problem

(3.57a) $$Ly(t) = \sum_{i=0}^{m/2} (-1)^i D^i (\sigma_{2i}(t) D^i y(t)) = q(t)$$

$$a \le t \le b$$

(3.57b) $$D^{i-1} y(a) = D^{i-1} y(b) = 0 \qquad 1 \le i \le m/2$$

where $Dy(t) := y'(t)$ and the (smooth) coefficient functions satisfy $\sigma_{2i}(t) \ge 0$ $1 \le i \le m/2-1$ and $\sigma_m(t) \ge \eta > 0$ for $a \le t \le b$.

The operator L is called a **self-adjoint** operator, having the property that it satisfies

$$\int_a^b Lu(t)v(t)\,dt = \int_a^b u(t)Lv(t)\,dt$$

for any $u, v \in C_0^m[a,b]$, the space of functions in $C^m[a,b]$ which satisfy the BC (3.57b).

The Ritz method involves choosing an approximate subspace and letting the approximate solution be the function which minimizes the variational formulation I(u) for (3.57) over that subspace. Here

$$I(u) := \int_a^b \{ \sum_{i=0}^{m/2} \sigma_{2i}(t)(D^i u(t))^2 - 2q(t)u(t) \} dt.$$

The Ritz solution $\tilde{s}(t) = \sum_{j=1}^{M} \tilde{\alpha}_j \psi_j(t)$ satisfies

(3.58) $$I(\tilde{s}) = \min_{s \in P^0_{k,\pi,1}} I(s) \quad .$$

Setting $\frac{\partial I(s)}{\partial \alpha_j} = 0$ and $\frac{\partial^2 I}{\partial \alpha_j^2} > 0$ $1 \le j \le M$

the matrix form is

(3.59) $$\tilde{G}\tilde{\alpha} = \tilde{q} \qquad \text{where}$$

$$\tilde{G} = (\tilde{g}_{ij})_{i,j=1}^{M} = (e(\psi_j(t), \psi_i(t)))_{i,j=1}^{M}$$

(3.60) $$e(u,v) := \int_a^b \{ \sum_{i=0}^{m/2} \sigma_{2i}(t) D^i u(t) D^i v(t) \} dt$$

$$= \int_a^b Lu(t) v(t) dt$$

and $\tilde{q}$ is in (3.49).

It is clear that for solving (3.57), the Galerkin and Ritz methods are mathematically equivalent, at least if $k, l \geq m$. However, the discrete Ritz method for which the integral coefficients (3.60) are approximated by using a quadrature, is generally different from the discrete Galerkin method (3.51) and has the advantage of preserving the matrix symmetry in (3.59).

While certain methods have their particular advantage in special cases, we usually only consider the collocation method since it appears to be generally the most efficient and since software for this method has been developed. For a comparison, see Russell-Varah [20].

4. Equivalences Between These Methods

In this Chapter, various equivalences between the methods mentioned in Chapter 3 are presented.

4.1. Collocation and Finite Difference Methods

4.1.1. Collocation, Trapezoidal, and Midpoint Rules

Consider the BVP (3.7). The collocation methods using 2 Lobatto points and 1 Gauss point relate to the trapezoidal rule and the midpoint rule(or Box scheme), respectively.

In particular, when solving the BVP (3.7) by collocation with approximate solution s(t) in $P_{2,\pi,1}$, if the Gauss points are the collocation points then we have

(4.1) $\quad \underset{\sim}{s}'(t_{i+1/2}) = \underset{\sim}{f}(t_{i+1/2}, \underset{\sim}{s}(t_{i+1/2})) \qquad 1 \le i \le n$

(4.2) $\quad \underset{\sim}{g}(\underset{\sim}{s}(a), \underset{\sim}{s}(b)) = \underset{\sim}{0}$

where $t_{i+1/2} = (t_i + t_{i+1})/2$.

Since $\underset{\sim}{s}(t_i) = \underset{\sim}{y}_i$, $1 \le i \le N+1$, the Newton form of the interpolating polynomial $\underset{\sim}{s}(t)$ can be expressed as

$$\underset{\sim}{s}(t) = \underset{\sim}{y}_i + (\underset{\sim}{y}_{i+1} - \underset{\sim}{y}_i)(t - t_i)/h_i \quad ,$$

so

$$\underset{\sim}{s}'(t)=(\underset{\sim}{y}_{i+1}-\underset{\sim}{y}_i)/h_i \quad .$$

(i) Case I:

At the Gauss point $t_{i+1/2}$

$$\underset{\sim}{s}'(t_{i+1/2})=(\underset{\sim}{y}_{i+1}-\underset{\sim}{y}_i)/h_i$$

$$\underset{\sim}{s}(t_{i+1/2})=\underset{\sim}{y}_i+(\underset{\sim}{y}_{i+1}-\underset{\sim}{y}_i)(t_{i+1/2}-t_i)/h_i$$

$$=(\underset{\sim}{y}_{i+1}+\underset{\sim}{y}_i)/2 \quad .$$

Now (4.1) gives

$$N_\pi \underset{\sim}{y}_i = (\underset{\sim}{y}_{i+1}-\underset{\sim}{y}_i)/h_i-\underset{\sim}{f}(t_{i+1/2},1/2(\underset{\sim}{y}_i+\underset{\sim}{y}_{i+1}))=\underset{\sim}{0}$$

and (4.2) gives $\underset{\sim}{g}(\underset{\sim}{s}(\underset{\sim}{y}_1),\underset{\sim}{s}(\underset{\sim}{y}_{N+1}))=\underset{\sim}{0}$,

which are the midpoint rule (3.33) and (3.35) , one of the most widely used finite difference methods.

(ii) Case II:

When collocating at Lobatto points t_i, and t_{i+1},

$$\underset{\sim}{s}'(t_i)=\underset{\sim}{f}(t_i,\underset{\sim}{s}(t_i)), \quad \text{i.e.}$$

(4.3) $$(\underset{\sim}{y}_{i+1}-\underset{\sim}{y}_i)/h_i = \underset{\sim}{f}(t_i,\underset{\sim}{y}_i)$$

and

$$\underset{\sim}{s}'(t_{i+1})=\underset{\sim}{f}(t_{i+1},\underset{\sim}{s}(t_{i+1})), \quad \text{i.e.}$$

(4.4) $$(\underset{\sim}{y}_{i+1}-\underset{\sim}{y}_i)/h_i = \underset{\sim}{f}(t_{i+1},\underset{\sim}{y}_{i+1}) \quad .$$

From (4.3) and (4.4), we get

$$N_\pi \underset{\sim}{y}_i=(\underset{\sim}{y}_{i+1}-\underset{\sim}{y}_i)/h_i-1/2[\underset{\sim}{f}(t_i,\underset{\sim}{y}_i)+\underset{\sim}{f}(t_{i+1},\underset{\sim}{y}_{i+1})]=\underset{\sim}{0}$$

which is the trapezoidal rule in (3.34).

4.1.2. Collocation and Trapezoidal Rule

In the previous subsection, it has been shown that there is an equivalence between the trapezoidal rule and collocation method for $P_{2,\pi,1}$. Here, the same result is obtained for $P_{3,\pi,2}$.

Suppose $\quad \underset{\sim}{s}(t_i)=\underset{\sim}{y}_i$

and

$$\underset{\sim}{s}'(t_i)=\underset{\sim}{y}'_i \qquad 1\le i\le N+1 \quad ,$$

The Newton form of the interpolating polynomial s(t) gives

$$\underset{\sim}{s}(t)=\underset{\sim}{y}_i+\underset{\sim}{y}'_i(t-t_i)+(\underset{\sim}{y}_{i+1}-\underset{\sim}{y}_i-\underset{\sim}{y}'_i h_i)(t-t_i)^2/h_i^2 \quad .$$

Hence

$$\underset{\sim}{s}'(t)=\underset{\sim}{y}'_i+2(\underset{\sim}{y}_{i+1}-\underset{\sim}{y}_i-\underset{\sim}{y}'_i h_i)(t-t_i)/h_i^2,$$

so

$$\underset{\sim}{s}'(t_{i+1})=\underset{\sim}{y}'_i+2(\underset{\sim}{y}_{i+1}-\underset{\sim}{y}_i-\underset{\sim}{y}'_i h_i)/h_i=2(\underset{\sim}{y}_{i+1}-\underset{\sim}{y}_i)/h_i-\underset{\sim}{y}'_i \quad ,$$

and

$$\underset{\sim}{s}'(t_i)=\underset{\sim}{y}'_i \quad .$$

Since $\underset{\sim}{s}(t)$ satisfies the differential equation at collocation points, i.e.

$$\underset{\sim}{s}'(t_{i+1})=\underset{\sim}{f}(t_{i+1},\underset{\sim}{s}(t_{i+1})) \text{ and } \underset{\sim}{s}'(t_i)=\underset{\sim}{f}(t_i,\underset{\sim}{s}(t_i)),$$

from above

$$2(\underset{\sim}{y}_{i+1}-\underset{\sim}{y}_i)/h_i-\underset{\sim}{f}(t_i,\underset{\sim}{y}_i)=\underset{\sim}{f}(t_{i+1},\underset{\sim}{y}_{i+1}) \quad .$$

Hence, we have

$$N_\pi \underset{\sim}{y}_i=(\underset{\sim}{y}_{i+1}-\underset{\sim}{y}_i)/h_i-1/2[\underset{\sim}{f}(t_i,\underset{\sim}{y}_i)+\underset{\sim}{f}(t_{i+1},\underset{\sim}{y}_{i+1})]=\underset{\sim}{0}$$

which is the trapezoidal rule in (3.34).

4.2. Simple Ritz and Finite Difference

Consider the Ritz method for the simple second order BVP

(4.5a) $$-y''(t)+\sigma(t)y(t)=q(t) \qquad a\le t\le b$$

(4.5b) $$y(a)=0\ , \qquad y(b)=0$$

with $\tilde{s}(t)\in P_{2,\pi,1}$. For simplicity, consider a uniform mesh, i.e. $h_i=(b-a)/N$ for $1\le i\le N$. Let the piecewise linear B-spline basis functions which satisfy (4.5) be

$$B_j(t)=\begin{cases}(t-t_{j-1})/h & t\in[t_{j-1},t_j] \\ (t_{j+1}-t)/h & t\in[t_j,t_{j+1}] \quad 2\le j\le N \\ 0 & \text{otherwise}\end{cases}$$

(For the construction and evaluation of B-splines, see de Boor [6] [7], and Ascher-Russell [2]).

Then $\tilde{G}=(\tilde{g}_{ij})_{i,j=1,}^{N+1}$ $\tilde{q}_i$ in (3.59) can be shown to be

(4.6) $$\tilde{g}_{ij}=\begin{cases}\frac{2}{h}+\int_{t_{i-1}}^{t_i}\sigma(t)\left(\frac{t-t_{i-1}}{h}\right)^2dt+\int_{t_i}^{t_{i+1}}\sigma(t)\left(\frac{t_{i+1}-t}{h}\right)^2dt & j=i \\ \frac{-1}{h}+\int_{t_{i-1}}^{t_i}\sigma(t)\left(\frac{t_i-t}{h}\right)\left(\frac{t-t_{i-1}}{h}\right)dt & j=i\pm1 \\ 0 & \text{otherwise}\end{cases}$$

for $2\le i,j\le N$ and

$$(4.7)\quad \tilde{q}_i = \int_{t_{i-1}}^{t_i} q(t)\left(\frac{t-t_{i-1}}{h}\right)dt + \int_{t_i}^{t_{i+1}} q(t)\left(\frac{t_{i+1}-t}{h}\right)dt \qquad 2\le i\le N .$$

In this case (3.59) can be expressed as

$$\{-1/h + \int_{t_{i-1}}^{t_i} \sigma(t)\left(\frac{t_i - t}{h}\right)\left(\frac{t-t_{i-1}}{h}\right)dt\} y_{i-1} + \{2/h$$

$$+ \int_{t_{i-1}}^{t_i} \sigma(t)\left(\frac{t-t_{i-1}}{h}\right)^2 dt + \int_{t_i}^{t_{i+1}} \sigma(t)\left(\frac{t_{i+1}-t}{h}\right)^2 dt\} y_i$$

(4.8)

$$+\{-1/h + \int_{t_{i-1}}^{t_i} \sigma(t)\left(\frac{t_i - t}{h}\right)\left(\frac{t-t_{i-1}}{h}\right)dt\} y_{i+1}$$

$$= \int_{t_{i-1}}^{t_i} q(t)\left(\frac{t-t_{i-1}}{h}\right)dt + \int_{t_i}^{t_{i+1}} q(t)\left(\frac{t_{i+1}-t}{h}\right)dt$$

$$2\le i\le N .$$

If the trapezoidal rule is used to approximate the coefficients in (4.6) and (4.7), then (4.8) becomes

$$-y_{i-1}/h + \{2/h + h\sigma(t_i)\} y_i - y_{i+1}/h = hq(t_i) .$$

Dividing by h, we have

$$-\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + \sigma(t_i) y_i = q(t_i),$$

which is identical to a finite difference scheme for solving (4.5) (see Varah [22]).

4.3. Collocation and Implicit Runge-Kutta

When DEs are solved by most finite difference methods, they are converted to first order system. To relate collocation methods with Runge-Kutta methods, we consider the first order nonlinear DE

(4.9) $\quad \underset{\sim}{y}' = \underset{\sim}{f}(t, \underset{\sim}{y})$.

The collocation schemes for (4.9) are

$$\underset{\sim}{y}(t_i) = \underset{\sim}{y}_i \qquad , \quad \underset{\sim}{y}'(t_{ij}) = \underset{\sim}{f}(t_{ij}, \underset{\sim}{y}(t_{ij}))$$

where t_{ij} are collocation points which satisfy

$$t_{ij} = t_i + h_i \rho_j \ , \qquad i=1,\ldots,N \ , \ j=1,\ldots,k \ ,$$

and ρ_j are canonical points in [0,1] $0 \le \rho_1 < \rho_2 < \ldots < \rho_k \le 1$.

Let $\underset{\sim}{f}_{ij} = \underset{\sim}{f}(t_{ij}, \underset{\sim}{y}(t_{ij}))$, $j=1,\ldots,k$, and express $\underset{\sim}{y}$ in terms of interpolation to the values $\underset{\sim}{y}_i$, $\underset{\sim}{f}_{i1}, \ldots, \underset{\sim}{f}_{ik}$, i.e.

$$\underset{\sim}{y}(t) = \underset{\sim}{y}_i + h_i \sum_{j=1}^{k} \underset{\sim}{f}_{ij} \phi_j\left(\frac{t-t_i}{h_i}\right)$$

where $\phi_j(t)$ $j=1,\ldots,k$ are polynomials of degree at most k on [0,1] determined by interpolation conditions

(4.10) $\quad \phi_j(0) = 0 \qquad \phi_j'(\rho_l) = \delta_{jl} \qquad l=1,\ldots,k$

here δ_{ik} denotes the Kronecker delta function.

Let

$$(4.11) \qquad \beta_j = \phi_j(1) \qquad \alpha_{jl} = \phi_l(\rho_j) \quad .$$

Then we get the equivalent implicit Runge-Kutta method

$$(4.12) \qquad \underset{\sim}{y}_{i+1} = \underset{\sim}{y}_i + h_i \sum_{j=1}^{k} \beta_j \underset{\sim}{f}_{ij} \qquad 1 \le i \le N$$

$$\underset{\sim}{f}_{ij} = \underset{\sim}{f}(t_{ij}; \underset{\sim}{y}_i + h_i \sum_{l=1}^{k} \alpha_{jl} \underset{\sim}{f}_{il}) \qquad j=1,\ldots,k$$

where β_j, α_{jl} are given in (4.11) (Ascher-Weiss [4]).

Not every RK scheme is equivalent to a collocation scheme. But, among the most accurate RK schemes (using Gauss, Radau, and Lobatto points), the most important are in fact equivalent to collocation schemes.

When $k=1$, $\rho_1 = 1/2$, then by (4.10) and (4.11)

$$\beta_1 = \phi_1(1) = 1$$

$$\alpha_{11} = \phi_1(\rho_1) = 1/2 \quad ,$$

so (4.12) gives the midpoint rule, and the equivalence has been shown in Section 4.1.1.

When $k=2$, $\rho_1 = 0$, $\rho_2 = 1$, (4.10) and (4.11) give

$$\phi_1(t) = -t^2/2 + t$$

$$\phi_2(t) = t^2/2$$

and $\qquad \beta_1 = \beta_2 = 1/2 \quad .$

The equivalent trapezoidal rule was treated in Section 4.1.1 as well.

4.4. Collocation and Multiple Shooting

We now consider the collocation using monomial basis functions for solving the differential equation

(4.13a) $$Ly(t):=y^{(m)}(t)-\sum_{j=0}^{m-1}\sigma_j(t)\,y^{(j)}(t)=q(t) \qquad a\le t\le b$$

with the separated boundary conditions

(4.13b) $$\sum_{l=1}^{m} b_{jl}\,y^{(l-1)}(a)=0\ , \qquad \sum_{l=1}^{m} c_{jl}\,y^{(l-1)}(b)=0$$

$$1\le j\le m/2.$$

We consider collocation at Gaussian points with $s(t)\in P_{k,\pi,m}$ where $k\ge 2m$, and we assume that the order m of the DE is even.

In general, the spline solution is determined by two types of constraints, continuity conditions and discretization equations (collocation equations and BCs).

On $[t_i, t_{i+1}]$, the local monomial basis considered has the form

$$\left\{\frac{(t-t_i)^{j-1}}{(j-1)!}\right\}_{j=1}^{k}.$$

The collocation approximation s(t) can be written as

(4.14) $$s(t)=\sum_{j=1}^{m} z_{ij}\frac{(t-t_i)^{j-1}}{(j-1)!}+h_i^{m}\sum_{j=1}^{k-m} w_{ij}\frac{((t-t_i)/h_i)^{m+j-1}}{(m+j-1)!}$$

where $z_i = (z_{i1}, \ldots, z_{im})^T = (s(t_i), \ldots, s^{(m-1)}(t_i))^T$.

The scaling in the first sum and h_i^m in the second sum are only introduced for later notational convenience. Now, both continuity conditions and discretization conditions must be satisfied. For the continuity conditions

$s^{(r-1)}(t_{i+1}^+) = s^{(r-1)}(t_{i+1}^-)$ $1 \le r \le m, 1 \le i \le N$, (4.14) become

$$z_{i+1} = B_i z_i + D_i w_i \tag{4.15}$$

where $w_i = (w_{i1}, \ldots, w_{i,k-m})^T$,

$B_i = (B^i_{rj})$ is an $m \times m$ upper triangular matrix with entries

$$B^i_{rj} = (h_i^{j-r})/(j-r)! \qquad j \ge r, \tag{4.16}$$

and $D_i = (D^i_{rj})$ is an $m \times (k-m)$ matrix with entries

$$D^i_{rj} = h_i^{m+1-r}/(m+j-r)! = O(h_i). \tag{4.17}$$

The collocation conditions in (t_i, t_{i+1}) give

$$Ls(t_{ir}) = h_i^m \sum_{j=1}^{k-m} \frac{w_{ij}}{(m+j-1)!} L\left[\frac{\left(t_{ir} - t_i\right)^{m+j-1}}{h_i}\right] \tag{4.18}$$

$$- \sum_{l=1}^{m} \zeta_{l-1}(t_{ir}) \sum_{j=1}^{m} z_{ij} \frac{(t_{ir} - t_i)^{j-1}}{(j-1)!} = q(t_{ir})$$

where t_{ir} are the collocation points in (t_i, t_{i+1}) for

$1 \le r \le k-m$.

Writing (4.18) in matrix form, we have

$$H_i \underset{\sim}{z}_i + G_i \underset{\sim}{w}_i = \underset{\sim}{q}_i \tag{4.19}$$

where $\underset{\sim}{q}_i = (q(t_{i1}), \ldots, q(t_{i,k-m}))^T$,

$H_i = (H^i_{rj})$ is a $(k-m) \times m$ matrix with entries

$$H^i_{rj} = -\sum_{l=1}^{j} \sigma_{l-1}(t_{ir})[(\rho_r h_i)^{j-l}]/(j-l)! \quad , \tag{4.20}$$

and $G_i = (G^i_{rj})$ is a $(k-m) \times (k-m)$ matrix with entries

$$G^i_{rj} = \frac{(\rho_r)^{j-1}}{(j-1)!} - \sum_{l=1}^{m} \sigma_{l-1}(t_{ir}) h_i^{m+l-1} \frac{(\rho_r)^{m+j-1}}{(m+j-1)!} \ . \tag{4.21}$$

Thus the collocation matrix C corresponding to the unknowns

$(\underset{\sim}{z}_1, \underset{\sim}{w}_1, \underset{\sim}{z}_2, \underset{\sim}{w}_2, \ldots, \underset{\sim}{z}_N, \underset{\sim}{w}_N, \underset{\sim}{z}_{N+1})$ has the form

$$C = \begin{bmatrix} \begin{array}{|c|} \hline V_0 \\ \hline V_1 \\ \hline \end{array} & & & & & 0 \\ & \begin{array}{|c|} \hline V_2 \\ \hline \end{array} & & & & \\ & & \begin{array}{|c|} \hline V_3 \\ \hline \end{array} & & & \\ & & & \ddots & & \\ & 0 & & & & \begin{array}{|c|} \hline V_N \\ \hline V_{N+1} \\ \hline \end{array} \end{bmatrix} .$$

The blocks V_i $(1 \le i \le N)$ of the matrix consisting of collocation and continuity equations are $k \times (k+m)$ and have the structure

$$(4.22) \qquad V_i = \begin{bmatrix} H_i & G_i & 0 \\ -B_i & -D_i & I \end{bmatrix} \begin{matrix} \}k-m \\ \}m \end{matrix}$$

(column widths: m, $k-m$, m)

where I is the $m \times m$ identity matrix.

Consider the case where condensation of parameters is performed on V_i by removing columns $m+1$ to k (corresponding to the unknowns $\tilde{w}_i$) and rows 1 to $k-m$. From (4.19), eliminate $\tilde{w}_i$

$$\tilde{w}_i = G_i^{-1} \tilde{q}_i - G_i^{-1} H_i \tilde{z}_i \quad ,$$

and substitute this into the continuity equation (4.15) to obtain

$$(4.23) \qquad \tilde{z}_{i+1} = \Gamma_i \tilde{z}_i + \tilde{g}_i$$

where

$$(4.24) \qquad \Gamma_i = B_i - D_i G_i^{-1} H_i$$

$$\tilde{g}_i = D_i G_i^{-1} \tilde{q}_i \quad .$$

The coefficient matrix for $\{\tilde{z}_i\}_{i+1}^{N+1}$ corresponding to the BC (4.13) and (4.23) has the form

$$c' = \begin{bmatrix} V_0 & & & & \\ -\Gamma_1 & I & & & \\ & -\Gamma_2 & I & & \\ & & \ddots & & \\ & & & -\Gamma_N & I \\ & & & & V_{N+1} \end{bmatrix}.$$

Let $\{M_j(t;t_i)\}_{j=1}^{m}$ be the set of linearly independent solutions of $Ly=0$ subject to the initial conditions $M_j^{(l-1)}(t_i;t_i)=\delta_{jl}$, $1\le j,l\le m$, and let y be the solution to $Ly=0$ with $\underset{\sim}{z}(y(t_i))$ given, where $\underset{\sim}{z}(y(t_i)):=(y(t_i),y'(t_i),\ldots,y^{(m-1)}(t_i))^T$. Then we have

(4.25) $\quad \underset{\sim}{z}(y(t_{i+1}))=M(t_{i+1},t_i)\underset{\sim}{z}(y(t_i))$

where $M(t;t_i)$ is the fundamental matrix $M(t;t_i)=(\underset{\sim}{z}(M_1(t;t_i)),\ldots\underset{\sim}{z}(M_m(t;t_i))$. If $s(t)$ is the collocation approximation of $y(t)$, then

$$\left\| \underset{\sim}{z}(s(t_{i+1}))-\underset{\sim}{z}(y(t_{i+1}))\right\| =O(h_i^{2(k-m)+1}).$$

From the analysis of collocation, $\underset{\sim}{z}(s(t_{i+1}))=\Gamma_i\underset{\sim}{z}(s(t_i))$ and $\underset{\sim}{z}(y(t_{i+1})) = \Gamma_i\underset{\sim}{z}(y(t_i))+O(h^{2(k-m)+1})$. From (4.25), since

$\underset{\sim}{z}(y(t_i))$ was arbitrary, $\Gamma_i = Y(t_{i+1}, t_i) + O(h^{2(k-m)+1})$, i.e.

Γ_i is an approximation to the fundamental matrix

$Y(t;t_i)$ and c' is an $O(h^{2(k-m)+1})$ perturbation of a

multiple shooting matrix (Ascher-Russell [3]).

As stated before, after doing condensation, the collocation methods give multiple shooting like matrices which have the advantage of being well-conditioned.

4.5. Multiple Shooting and the Box Scheme

We now consider the relationship between multiple shooting and the Box scheme. Suppose the mesh is uniform, and h is small. The multiple shooting matrix (3.17) gives

$$-Y_i(t_{i+1})\underset{\sim}{c}_i + F_{i+1}\underset{\sim}{c}_{i+1} = \underset{\sim}{v}_i(t_{i+1}) - \underset{\sim}{v}^0_{i+1}$$

or

(4.26) $$-Y_i(t_{i+1})\underset{\sim}{c}_i + Y_{i+1}(t_{i+1})\underset{\sim}{c}_{i+1} = \underset{\sim}{v}_i(t_{i+1}) - \underset{\sim}{v}_{i+1}(t_{i+1})$$

$1 \le i \le J-1$.

If we set $Y_i(t_i) = I - (hA_{i-1/2})/2$, $1 \le i \le J-1$,

then using a Taylor expansion and the fact that $Y' = AY$, we have

$$Y_i(t_{i+1}) = Y_i(t_i) + hY_i'(t_i) + O(h^2)$$

$$= I - (hA_{i-1/2})/2 + hA_i Y_i(t_i) + O(h^2)$$

$$=I-(hA_{i-1/2})/2+hA_i(I-(hA_{i-1/2})/2)+O(h^2)$$

$$=I-(hA_{i-1/2})/2+hA_i+O(h^2)$$

$$=I-(h(A_{i+1/2}-hA_{i+1/2}))/2+hA_i+O(h^2)$$

$$=I-(hA_{i+1/2})/2+hA_i+O(h^2)$$

$$=I-(hA_{i+1/2})/2+h(A_{i+1/2}-(hA_{i+1/2})/2)+O(h^2)$$

$$=I+(hA_{i+1/2})/2+O(h^2) \quad .$$

Similarly, if $\tilde{v}_i(t_i)$ is a particular solution satsfying $\tilde{v}_i(t_i)=0$, then

$$\tilde{v}_i(t_{i+1})=\tilde{v}_i(t_i)+h\tilde{v}'_i(t_i)+O(h^2)=h(A_i(t_i)\tilde{v}_i(t_i)+\tilde{f}_i)+O(h^2)$$

$$=h\tilde{f}_i+O(h^2)$$

$$=h\tilde{f}_{i+1/2}+O(h^2).$$

Hence (4.26) gives

$$(4.27) \qquad -(I+(hA_{i+1/2})/2)\tilde{c}_i+(I-(hA_{i+1/2})/2)\tilde{c}_{i+1}=h\tilde{f}_{i+1/2}+O(h^2) \; .$$

Since $\tilde{y}_i$ is the multiple shooting approximation solution,

$$\tilde{y}_i := Y_i(t_i)\tilde{c}_i+\tilde{v}_i(t_i)=(I-(hA_{i-1/2})/2)\tilde{c}_i \; ,$$

and $\tilde{c}_i=\tilde{y}_i+O(h)$, (4.27) gives

$$-(I+(hA_{i+1/2})/2)\tilde{y}_i+(I-(hA_{i+1/2})/2)\tilde{y}_{i+1}=h\tilde{f}_{i+1/2}+O(h)$$

which can be written as

$$\frac{y_{i+1}-y_i}{h} = A_{i+1/2}\frac{(y_{i+1}+y_i)}{2} + f_{i+1/2} + O(1)$$

$$= A_{i+1/2}\, y_{i+1/2} + f_{i+1/2} + O(1) .$$

Thus, the Box scheme gives a matrix which is a discrete approximation to a multiple shooting matrix.

4.6. Invariant Imbedding and Multiple Shooting

It is shown in Keller-Lentini [14] that there is an equivalence between invariant imbedding and the box scheme in the sense that a specific algorithm for solving the difference equations is valid if and only if an appropriate imbedding is valid.But the equivalence was not obvious. Recently, Lentini-Osborne-Russell [15] presented an easier way of getting a close relationship between multiple shooting and invariant imbedding. The Keller-Lentini [14] result is a special case of their presentation. When solving a BVP with separated BC, the relationship between factorizations of the multiple shooting matrix and invariant imbedding formulations of the BVP are shown in [15]. The equivalence is described below.

Consider the problem (3.24). In (3.17), converting each block $[-Y_i(t_{i+1})\ F_{i+1}]$ to $[-F_{i+1}^{-1}\ Y_i(t_{i+1})\ I]$ gives M similar to the "standard" multiple shooting matrix. Therefore, it is sufficient to consider only the "standard" multiple shooting matrix. Suppose the fundamental solutions $Y_i(t)$ and particular solutions $v_i(t)$ at t_{i+1} can be partitioned as

$$
\underset{i}{Y}(t_{i+1}) = \begin{bmatrix} \overset{p}{Y_i^0} & \overset{q}{Y_i^1} \\ Y_i^2 & Y_i^3 \end{bmatrix} \begin{matrix} \}p \\ \}q \end{matrix}, \qquad v_i(t_{i+1}) = \begin{bmatrix} v_i^1 \\ v_i^2 \end{bmatrix} \begin{matrix} \}p \\ \}q \end{matrix} .
$$

For convenience, arrange the multiple shooting matrix so that it has the form

$$
(4.28) \quad
\left[\begin{array}{cc|cc|cc|cc}
K_0 & K_1 & 0 & 0 & & & & \\
-Y_1^2 & -Y_1^3 & 0 & I_q & & & & \\
\hline
-Y_1^0 & -Y_1^1 & I_p & 0 & & & & \\
0 & 0 & -Y_2^2 & -Y_2^3 & 0 & I_q & & \\
\hline
 & & -Y_2^0 & -Y_2^1 & I_p & 0 & & \\
 & & & & \ddots & & & \\
 & & & & -Y_N^2 & -Y_N^3 & 0 & I_q \\
\hline
 & & & & -Y_N^0 & -Y_N^1 & I_p & 0 \\
 & & & & 0 & 0 & K_2 & K_3
\end{array}\right]
\begin{bmatrix} u(t_1) \\ w(t_1) \\ u(t_2) \\ w(t_2) \\ u(t_3) \\ \vdots \\ w(t_N) \\ u(t_{N+1}) \\ w(t_{N+1}) \end{bmatrix}
=
\begin{bmatrix} \alpha \\ v_1^2 \\ v_1^1 \\ v_2^1 \\ v_2^2 \\ \vdots \\ v_{N+1}^2 \\ v_{N+1}^1 \\ \beta \end{bmatrix},
$$

or

$$
(4.29) \qquad \tilde{M} y = b .
$$

Now, consider the factorization of (4.28). If K_0 is nonsingular, multiply the first block by K_0^{-1} and perform the first step of elimination, then we have

$$\begin{bmatrix} I_p & K_0^{-1}K_1 & \vdots & & \\ 0 & -Y_1^3 + Y_1^2 K_0^{-1} K_1 & \vdots & 0 & I_q \\ \hline 0 & -Y_1^1 + Y_1^0 K_0^{-1} K_1 & \vdots & I_p & 0 \\ & & & & \ddots \end{bmatrix} \begin{bmatrix} \underset{\sim}{u}(t_1) \\ \underset{\sim}{w}(t_1) \\ \underset{\sim}{u}(t_2) \\ \vdots \end{bmatrix} = \begin{bmatrix} \underset{\sim}{x}_1 \\ \underset{\sim}{v}_1^2 + Y_1^2 \underset{\sim}{x}_1 \\ \underset{\sim}{v}_1^1 + Y_1^0 \underset{\sim}{x}_1 \\ \vdots \end{bmatrix}$$

where $\underset{\sim}{x}_1 = K_0^{-1} \underset{\sim}{\alpha}$.

Let $V_1(t) = \begin{bmatrix} v_1^1(t) \\ \\ v_1^3(t) \end{bmatrix} = Y_1(t) \begin{bmatrix} K_0^{-1} K_1 \\ \\ -I \end{bmatrix}$ be the column of complementary function solutions and $\underset{\sim}{z}_1 = \underset{\sim}{v}_1^2 + Y_1^2 \underset{\sim}{x}_1$.

The second step of elimination gives

$$\begin{bmatrix} I_p & -R(t_1) & \vdots & & \\ 0 & v_1^3(t_2) & \vdots & 0 & I_q \\ 0 & 0 & \vdots & I_p & -R(t_2) \\ & & & & \ddots \end{bmatrix} \begin{bmatrix} \underset{\sim}{u}(t_1) \\ \underset{\sim}{w}(t_1) \\ \underset{\sim}{u}(t_2) \\ \vdots \end{bmatrix} = \begin{bmatrix} \underset{\sim}{x}(t_1) \\ \underset{\sim}{z}_1 \\ \underset{\sim}{x}(t_2) \\ \vdots \end{bmatrix}$$

where from section 3.1.5, $R(t) = v_1^1(t)\, v_1^3(t)^{-1}$ is the Riccati matrix satisfying (3.27a) and $\underset{\sim}{x}(t)$ satisfies (3.26). The next step gives

$$\begin{bmatrix} \ddots & & & & \\ 0 & 0 & I_p & -R(t_2) & & \\ & 0 & -Y_2^3 - Y_2^2 R(t_2) & 0 & I_g \\ \hdashline & & -Y_2^1 - Y_2^0 R(t_2) & I_p & 0 \end{bmatrix} \begin{bmatrix} \vdots \\ u(t_2) \\ w(t_2) \\ u(t_3) \end{bmatrix} = \begin{bmatrix} \vdots \\ x(t_2) \\ z_2 \\ v_2^2 + Y_2^0 x(t_2) \end{bmatrix}$$

where

$$(4.30) \quad \begin{bmatrix} -Y_2^0 - Y_2^1 R(t_2) \\ \\ -Y_2^2 - Y_2^3 R(t_2) \end{bmatrix} = Y_2(t_3) \begin{bmatrix} -R(t_2) \\ \\ -I \end{bmatrix}$$

$$= Y_2(t_3) Y_1(t_2) \begin{bmatrix} K_0^{-1} K_1 \\ \\ -I \end{bmatrix} (Y_1^3 + Y_1^2 R(t_1))^{-1}$$

$$= Y_1(t_3) \begin{bmatrix} K_0^{-1} K_1 \\ \\ -I \end{bmatrix} (Y_1^3 + Y_1^2 R(t_1))^{-1}$$

$$= v_1(t_3)(Y_1^3 + Y_1^2 R(t_1))^{-1} = v_2(t_3) = \begin{bmatrix} v_2^1(t_3) \\ v_2^3(t_3) \end{bmatrix}.$$

Since $\begin{bmatrix} K_0 & K_1 \end{bmatrix} v_2(t_1) = 0$, $v_2(t)$ is a transformed set of complementary function solutions; therefore, $R(t_3) = v_2^1(t_3) v_2^3(t_3)^{-1}$. The process is continued until the stage that $v_i(t_{i+1})$ is singular or partial pivoting is performed on the block $\begin{bmatrix} v_i^3(t_{i+1}) \\ v_i^1(t_{i+1}) \end{bmatrix}$. Then reimbedding is needed. For this purpose, introduce a permutation M' of $\tilde{M}$ which preserves the zero structure in (4.28) and leaves the differential equation invariant (i.e. if $y'=Ay$ is modified by the change of variables $z=Py$, then $z'=\hat{A}z$ where $\hat{A}=PAP^{-1}$). Then P,Q in $M'=P\tilde{M}Q$ must have the form

$$P = \begin{bmatrix} I_p & & & & & \\ & P_1 & & & & \\ & & \cdot & & & \\ & & & \cdot & & \\ & & & & P_N & \\ & & & & & I_q \end{bmatrix} \qquad Q = \begin{bmatrix} I_n & & & & \\ & q_1 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & q_N^T \end{bmatrix},$$

where I_p, I_q, I_N are identity matrices of size p x p, q x q, and N x N, respectively, and $q_i=p_i$ are N x N permutation matrices which satisfy

$$(4.31) \qquad q_i = J^{-1} p_i J \qquad J := \begin{bmatrix} 0 & I_q \\ I_p & 0 \end{bmatrix}$$

(see Keller-Lentini[13], Lentini-Osborne-Russell[15]).

Using the same argument in (4.30), the theorem follows:

Thm.4.1 Suppose that the BVP (3.25) has a unique solution. Assume K_0 is nonsingular. Suppose further that the LU factorization for $P\tilde{M}Q$ exists where

$$(4.32) \qquad P=\begin{bmatrix} I_p & & & & & & & \\ & \ddots & & & & & & \\ & & I_N & & & & & \\ & & & p_1 & & & & \\ & & & & \ddots & & & \\ & & & & & p_1 & & \\ & & & & & & I_q \end{bmatrix}, \qquad Q=\begin{bmatrix} I_N & & & & & \\ & \ddots & & & & \\ & & I_N & & & \\ & & & q_1^T & & \\ & & & & \ddots & \\ & & & & & q_1^T \end{bmatrix}.$$

Here the $N \times N$ matrices p_1, q_1 occur $(n-i)$ times and satisfy (4.31). Then the factorization reduces (4.28) to the block upper triangular form

(4.33)

$$
\left[\begin{array}{cc|cc|ccc|cc|c}
I_p & -R_3(t_1) & 0 & 0 & & & & & & \\
0 & -v_1(t_2) & 0 & I_q & & & & & & \\
\hline
 & & I_p & -R_3(t_2) & & & & & & \\
 & & 0 & -v_2(t_3) & & & & & & \\
\hline
 & & & & \ddots & & & & & \\
 & & & & I_p & -R_3(t_i) & & & & \\
 & & & & 0 & -\hat{v}_i(t_{i+1}) & 0 & I_q & & \\
\hline
 & & & & & & I_p & -\tilde{R}_3(t_{i+1}) & & \\
 & & & & & & 0 & -\hat{v}_{i+1}(t_{i+2}) & & \\
\hline
 & & & & & & & & \ddots & \\
 & & & & & & & & I_p & -\tilde{R}(t_{N+1}) \\
 & & & & & & & & 0 & K_3 - K_2\tilde{R}(t_{N+1})
\end{array}\right]
\times
\begin{bmatrix}
\underset{\sim}{u}(t_1) \\
\underset{\sim}{w}(t_1) \\
\cdot \\
\underset{\sim}{u}(t_i) \\
\underset{\sim}{w}(t_i) \\
\widetilde{\underset{\sim}{u}}(t_{i+1}) \\
\widetilde{\underset{\sim}{w}}(t_{i+1}) \\
\cdot \\
\widetilde{\underset{\sim}{u}}(t_{N+1}) \\
\widetilde{\underset{\sim}{w}}(t_{N+1})
\end{bmatrix}
=
\begin{bmatrix}
\underset{\sim}{x}(t_1) \\
\underset{\sim}{z}_1 \\
\cdot \\
\underset{\sim}{x}(t_i) \\
\underset{\sim}{z}_i \\
\widetilde{\underset{\sim}{x}}(t_{i+1}) \\
\widetilde{\underset{\sim}{z}}_{i+1} \\
\cdot \\
\widetilde{\underset{\sim}{x}}(t_{N+1}) \\
\widetilde{\underset{\sim}{z}}_{N+1} - K_2\widetilde{\underset{\sim}{x}}(t_{N+1})
\end{bmatrix}
$$

Here $\begin{pmatrix} \tilde{u}(t) \\ \tilde{v}(t) \end{pmatrix} := q_1 \begin{pmatrix} u(t) \\ w(t) \end{pmatrix}$, $\hat{v}_j(t) = q_1 Y_j(t) \begin{bmatrix} R(t_j) \\ I \end{bmatrix}$,

$\tilde{v}_j(t) = q_1 Y_j(t) \begin{bmatrix} P(t_j) \\ I \end{bmatrix}$ $(1 \le j \le i)$, where $\tilde{R}(t)$ and $\tilde{x}(t)$

are the solutions to the invariant imbedding equations

corresponding to (3.27a, b) for the variables $\begin{bmatrix} \tilde{u}(t) \\ \tilde{v}(t) \end{bmatrix}$.

Proof: see Lentini-Osborne-Russell [15].

The factorization of $\tilde{M}$ to (4.33), when P=Q=I, can be interpreted as a forward elimination corresponding to finding solutions $R(t)$ and $x(t)$ to the IVP (3.27a), (3.27b). The back substitution on (4.33) to find w_j $(j=N,\ldots,0)$ then corresponds to finding the solution to (3.27c). For $q_i \neq I$, it corresponds to changing the invariant imbedding formulation at $t=t_{i+1}$. When P,Q have any number of adjacent blocks of permutaton matrices which are different, the matrix factorization corresponds to finding solutions for different imbedding formulations (Lentini-Osborne-Russell [15]). The equivalence between multiple shooting and invariant imbedding is therefore shown, and the Keller-Lentini [14] result concerning the equivalence between invariant imbedding and the Box scheme follows using basically the same argument as in this section.

5. Finite Differences for Solving High Order Differential Equations

In this chapter, the construction of finite difference methods which give high-accuracy approximations to the solution of a high order linear differential equation My=f subject to linear BCs is investigated.

Define a mesh π: $a=t_0<t_1<\ldots<t_J=b$ and for any function $w(t)$ on π, let $w_j=w(t_j)$. At mesh points, u_j is the estimate of y and u satisfies $M_h u=\widetilde{f}$, together with appropriate BCs, where $M_h u$ is a linear combination of values of u at stencil points (adjacent mesh points) and $\widetilde{f}$ is a linear combination of values of f at m auxiliary points close to the stencil points.

The construction is based on a local collocation procedure with polynomials, which is equivalent to the method of undetermined coefficients.

In section 1, the description of the discretization approximation is presented in the first part and some examples are given. In the second part the description of the finite difference approximations to boundary conditions follows. In section 2, the order of the truncation error is given when the location of the auxiliary points is independent of M. Stability of the scheme is discussed in section 3. And in section 4, methods with higher order of accuracy obtained by Doedel [9] and by Lynch-Rice [16] using special auxiliary points are presented and followed by examples. An obvious equivalence between the high order finite difference methods and collocation methods is shown in section 5. Section 6 contains a comparison of the computational effort of Doedel's schemes and Lynch-Rice's schemes. This was not done in either [9] or [16]. Comparison with collocation methods is performed. Numerical results for Doedel's methods, Lynch-Rice's methods, and collocation methods are provided for comparison.

5.1. Construction of the High Order Finite Difference Approximation

In this section, the construction of the high order finite difference methods is presented. We consider the interior subintervals first.

5.1.1. The Approximations for Interior Subintervals

Consider the nth order linear differential equation

$$My(t)=y^{(n)}(t)+\sum_{k=0}^{n-1} a_k(t)y^{(k)}(t)=f(t) \qquad a\le t\le b \tag{5.1}$$

and the mesh $a=t_0<t_1<\ldots<t_J=b$.

Let u_j be the approximate solution of (5.1) at $t=t_j$, and in the subinterval $[t_{j-r_j}, t_{j+s_j}]$ where r_j and s_j are positive constants. Let the difference operator M_h be

$$M_h u_j = \sum_{i=-r_j}^{s_j} d_{j,i} u_{j+i} .$$

The right hand side of the approximation equation $M_h u_j = \tilde{f}_j$ is

$$\tilde{f}_j = \sum_{i=1}^{m} e_{j,i} f(z_{j,i})$$

where $\{d_{j,i}\}$ and $\{e_{j,i}\}$ are known coefficients and

$z_{j,1} < z_{j,2} < \ldots < z_{j,m}$ are in $[t_{j-r_j}, t_{j+s_j}]$.

We will show how $\{d_{j,i}\}$ and $\{e_{j,i}\}$ are constructed below.

For simplicity, we frequently omit the subscript j, e.g. $z_{j,i}$, $d_{j,i}$, $e_{j,i}$, r_j and s_j become z_i, d_i, e_i, r, and s. Then the finite difference approximations to (5.1) at mesh points have the form

$$(5.2) \qquad M_h u_j = \sum_{i=-r}^{s} d_i u_{j+i} = \sum_{i=1}^{m} e_i f(z_i) = \tilde{f}_j \qquad r \le j \le J-s$$

where $h = \max_j h_j$, $h_j = t_j - t_{j-1}$.

Since there are J-s-r+1 equations in (5.2) and n BCs , and the number of unknowns is J+1, one requires that r+s≥n and also incorporates more constraints if necessary.

The coefficients d_i and e_i are determined so that the approximation is exact on P_L, the space of all polynomials of degree at most L. i.e. if $w^l(t)$ $(0 \le l \le L)$ form a basis for P_L then d_i, e_i, are made to satisfy the equations

$$\sum_{i=-r}^{s} d_i w^l(t_{j+i}) - \sum_{i=1}^{m} e_i M w^l(z_i) = 0 \tag{5.3}$$

$$l = 0, \ldots, L \quad .$$

The system (5.3) is homogeneous in d_i, e_i. Therefore, in addition to (5.3), we take some convenient normalization equation such as one of

(5.4)

a) $e_1 = 1$

b) $\sum_i |e_i| = 1$, or

c) $\sum_i e_i = 1$.

To uniquely determine the r+s+m+1 unknowns d_i and e_i from (5.3) and (5.4), L must be at least equal to r+s+m-1. BCs for u are obtained in a similar way, they are treated in the next section.

Let p(t) be the polynomial in $P_{r+s+m-1}$ which interpolates the solution y(t) and satisfies

$$p(t_{j+i}) = u_{j+i} \qquad -r \le i \le s \tag{5.5}$$

$$Mp(z_i) = f(z_i) \qquad 1 \le i \le m \quad . \tag{5.6}$$

Write p(t) as a linear combination of the basis functions

$$p(t) = \sum_{k=0}^{r+s+m-1} c_k w^k(t) \quad . \tag{5.7}$$

Then (5.5), (5.6), and (5.7) give

$$\sum_{k=0}^{r+s+m-1} c_k w^k(t_{j+i}) = u_{j+i} \qquad -r \le i \le s$$

$$\sum_{k=0}^{r+s+m-1} c_k M w^k (z_i) = f(z_i) \qquad 1 \le i \le m \ ,$$

or $\det(D_M) = 0$ where

$$D_M = \begin{bmatrix} w^0(t_{j-r}) & w^1(t_{j-r}) & \cdots & w^{r+s+m-1}(t_{j-r}) & u_{j-r} \\ w^0(t_{j-r+1}) & w^1(t_{j-r+1}) & & w^{r+s+m-1}(t_{j-r+1}) & u_{j-r+1} \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ w^0(t_{j+s}) & w^1(t_{j+s}) & & w^{r+s+m-1}(t_{j+s}) & u_{j+s} \\ Mw^0(z_1) & Mw^1(z_1) & \cdots & Mw^{r+s+m-1}(z_1) & f(z_1) \\ Mw^0(z_2) & Mw^1(z_2) & & Mw^{r+s+m-1}(z_2) & f(z_2) \\ \cdot & \cdot & & \cdot & \cdot \\ Mw^0(z_m) & Mw^1(z_m) & & Mw^{r+s+m-1}(z_m) & f(z_m) \end{bmatrix}$$

with the operator M as in (5.1).

One way to find d_i and e_i for a general set of z_i can be the following: Evaluate the determinant by expanding in terms of the last column and compare it with (5.2), introduce a normalizing factor E, then d_i and e_i are given by

(5.3a) $\qquad d_i = \mathrm{cof}_M [u_{j+i}]/E$

(5.3b) $\qquad e_i = -\mathrm{cof}_M [f(z_i)]/E$

where $\mathrm{cof}_M[\cdot]$ is the cofactor of the given element in D_M

and a convenient normalizing factor E can be chosen as

(5.9) $$E=-\sum_{i=1}^{m} \mathrm{cof}_{0_M}[f(z_i)]$$

with $M_1^0 y(t)=y^{(n)}(t)$.

If the $w^l(t)$ are chosen so that

(5.10)
$$w^{i}(t_{j-r+k})=\delta_{lk} \qquad 0\le l\le r+s$$
$$w^{l}(t_{j-r+k})=0 \qquad r+s+1\le l\le r+s+m-1,\ 0\le k\le r+s\ ,$$

then from (5.8) and (5.9), d_i and e_i can be calculated as

(5.11)
$$d_i=\frac{(-1)^m}{E}\begin{vmatrix} Mw^{r+i}(z_1) & Mw^{r+s+1}(z_1) & \cdots & Mw^{r+s+m-1}(z_1) \\ \vdots & \vdots & & \vdots \\ Mw^{r+i}(z_m) & Mw^{r+s+1}(z_m) & & Mw^{r+s+m-1}(z_m) \end{vmatrix}$$

$-r\le i\le s,$

$$
(5.12) \qquad e_i = \frac{(-1)^{m+i+1}}{E}
\begin{vmatrix}
Mw^{r+s+1}(z_1) & \cdots & Mw^{r+s+m-1}(z_1) \\
\vdots & & \vdots \\
Mw^{r+s+1}(z_{i-1}) & & Mw^{r+s+m-1}(z_{i-1}) \\
Mw^{r+s+1}(z_{i+1}) & & Mw^{r+s+m-1}(z_{i+1}) \\
\vdots & & \vdots \\
Mw^{r+s+1}(z_m) & & Mw^{r+s+m-1}(z_m)
\end{vmatrix}
$$

$$1 \le i \le m,$$

and

$$
(5.13) \qquad E = (-1)^m
\begin{vmatrix}
1 & M^0 w^{r+s+1}(z_1) & \cdots & M^0 w^{r+s+m-1}(z_1) \\
\vdots & \vdots & & \vdots \\
1 & M^0 w^{r+s+1}(z_m) & \cdots & M^0 w^{r+s+m-1}(z_m)
\end{vmatrix} .
$$

If $m=1$, then $E=-1$ and $e_1=1$.

To satisfy (5.10), $w^l(t)$, $0 \le l \le r+s+m-1$, can be chosen as

$$
(5.14a) \qquad w^l(t) = \prod_{k=0,\, k\neq l}^{r+s} \frac{(t-t_{j-r+k})}{(t_{j-r+l}-t_{j-r+k})}, \quad 0 \le l \le r+s ,
$$

$$
(5.14b) \qquad w^{r+s+l}(t) = (t-t_j)^{l-1} \prod_{k=0}^{r+s} (t-t_{j-r+k}) \,/\, h_j^{r+s+l}
$$

$$1 \le l \le m-1 .$$

If $r=0$ and $s=n$, then (5.2) is the higher-order difference approximation with identity expansions (HODIE) considered by Lynch-Rice [16].

In the following, we use d_i^0, e_i^0, M^0, f_i^0 for the coefficients and the operators when $M=D^n$.

Example 5.1

Consider $M=D^2+a^1(t)D+a^0(t)$ for $0\le t\le 1$.

Let $r=s=m=1$, suppose the mesh points are equally spaced and $z_1=t_j$. From (5.14a),

$$w^0(t)=\frac{(t-t_j)(t-t_{j+1})}{2h^2},$$

$$w^1(t)=\frac{(t-t_{j-1})(t-t_{j+1})}{-h^2},$$

$$w^2(t)=\frac{(t-t_{j-1})(t-t_j)}{2h^2},$$

so $d_{-1}=Mw^0(z_1)=1/h^2-a^1(t_j)/2h$,

$d_0=Mw^1(z_1)=-2/h^2+a^0(t_j)$,

$d_1=Mw^2(z_1)=1/h^2+a^1(t_j)/2h$,

and (5.2) becomes

$$\frac{u_{j+1}-2u_j+u_{j-1}}{h^2}+a^1_j\frac{u_{j+1}-u_{j-1}}{2h}+a^0_j u_j=f_j,$$

which is the usual divided difference approximation for

$D^2+a^1(t)D+a^0(t)$.

Example 5.2

Now, consider $M=D+a^0(t)$. Let $r=0$, $s=1$, $m=2$, $z_1=t_j$, and $z_2=t_{j+1}$. Choose the basis functions

$$w^0(t)=-(t-t_{j+1})/h_j \quad ,$$

$$w^1(t)=(t-t_j)/h_j \quad ,$$

$$w^2(t)=(t-t_j)(t-t_{j+1})/h_j^2 \quad ,$$

so

$$E=(-1)^2\begin{vmatrix} 1 & Mw^2(z_1) \\ 1 & Mw^2(z_2) \end{vmatrix} = Mw^2(z_2)-Mw^2(z_1)=2/h_j \quad ,$$

$$d_0=1/E\begin{vmatrix} Mw^0(z_1) & Mw^2(z_1) \\ Mw^0(z_2) & Mw^2(z_2) \end{vmatrix} = -1/h_j+a_j/2 \quad ,$$

$$d_1=1/E\begin{vmatrix} Mw^1(z_1) & Mw^2(z_1) \\ Mw^1(z_2) & Mw^2(z_2) \end{vmatrix} = 1/h_j+a^0_{j+1}/2 \quad ,$$

$$e_1=(Mw^2(z_2))/E=1/2 \quad ,$$

$$e_2 = (Mw^2(z_1))/2 = 1/2 ,$$

then (5.2) becomes

$$(u_{j+1} - u_j)/h_j + (a_j u_j + a_{j+1} u_{j+1})/2 = (f_j + f_{j+1})/2 ,$$

which is the trapozoidal rule.

If $m=1$ and $z_1 = t_{j+1/2}$ is taken, then

$$d_0 = Mw^0(z_1) = -1/h_j + a^0_{j+1/2}/2 ,$$

$$d_1 = Mw^1(z_1) = 1/h_j + a^0_{j+1/2}/2 .$$

The difference approximation is therefore

$$(u_{j+1} - u_j)/h_j + a_{j+1/2}(u_j + u_{j+1})/2 = f_{j+1/2}$$

which is the Box scheme.

5.1.2. The Approximation of Initial Conditions and Boundary Conditions

Approximation to initial or boundary conditions are required to complete the difference scheme. Here, only separated BCs are considered. Initial conditions can be considered as a special case of separated BCs.

Suppose two BCs for (5.1) are given by

(5.15a) $$B_a y(a) = y^{(k_1)}(a) + \sum_{i=0}^{k_1-1} b_i y^{(i)}(a) = \alpha \qquad k_1 < n ,$$

$$(5.15b) \qquad B_b\, y(b) = y^{(k_2)}(b) + \sum_{i=0}^{k_2-1} b_i\, y^{(i)}(b) = \beta \qquad k_2 < n .$$

The construction of a finite difference approximation to (5.15a) resembles that for (5.1) so we have

$$(5.16) \qquad B_{h,a}\, u_0 = \sum_{i=0}^{s} d_i u_i = e_0 \alpha + \sum_{i=1}^{m} e_i f(z_i) .$$

If the basis functions $w^l(t)$ $(0 \le l \le s+m)$ of P_{s+m} are chosen as in (5.10) for $r=0$, then the coefficients d_i and e_i can be calculated as

$$(5.17) \qquad d_i = 1/E_0 \begin{vmatrix} B_a w^i(a) & B_a w^{s+1}(a) & \cdots & B_a w^{s+m}(a) \\ Mw^i(z_1) & Mw^{s+1}(z_1) & \cdots & Mw^{s+m}(z_1) \\ \vdots & \vdots & & \vdots \\ Mw^i(z_m) & Mw^{s+1}(z_m) & \cdots & Mw^{s+m}(z_m) \end{vmatrix} ,$$

$$0 \le i \le s$$

$$(5.18) \qquad e_0 = 1/E_0 \begin{vmatrix} Mw^{s+1}(z_1) & \cdots & Mw^{s+m}(z_1) \\ \vdots & & \vdots \\ Mw^{s+1}(z_m) & \cdots & Mw^{s+m}(z_m) \end{vmatrix} ,$$

$$(5.19)\qquad e_i = \frac{(-1)^i}{E_0}\begin{vmatrix} B_a w^{s+1}(a) & \cdot\cdot & B_a w^{s+m}(a) \\ M w^{s+1}(z_1) & \cdot\cdot & M w^{s+m}(z_1) \\ \cdot & & \cdot \\ M w^{s+1}(z_{i-1}) & & M w^{s+m}(z_{i-1}) \\ M w^{s+1}(z_{i+1}) & & M w^{s+m}(z_{i+1}) \\ \cdot & & \cdot \\ M w^{s+1}(z_m) & \cdot\cdot & M w^{s+m}(z_m) \end{vmatrix},$$

$$1 \le i \le m$$

where E_0 can be chosen as

$$(5.20)\qquad E_0 = (-1)^{m+1}\begin{vmatrix} M^0 w^{s+1}(z_1) & \cdot\cdot & M^0 w^{s+m}(z_1) \\ \cdot & & \cdot \\ M^0 w^{s+1}(z_m) & \cdot\cdot & M^0 w^{s+m}(z_m) \end{vmatrix},$$

unless $m=0$, in which case, set $E_0=1$.

The BC (5.15b) is treated similarly. The finite difference approximation to (5.15b) is

$$(5.21)\qquad B_{h,b} u_J = \sum_{i=-r}^{0} d_i u_{J+i} = e_0 \beta + \sum_{i=1}^{m} e_i f(z_i)$$

and the coeficients d_i, e_i are

$$(5.22)\qquad d_i = 1/E_0 \begin{vmatrix} B_b w^{r+i}(b) & B_b w^{r+1}(b) & \cdots & B_b w^{r+m}(b) \\ Hw^{r+i}(z_1) & Hw^{r+1}(z_1) & & Hw^{r+m}(z_1) \\ \vdots & \vdots & & \vdots \\ Hw^{r+i}(z_m) & Hw^{r+1}(z_m) & & Hw^{r+m}(z_m) \end{vmatrix}$$

$$-r \le i \le 0$$

$$(5.23)\qquad e_0 = 1/E_0 \begin{vmatrix} Hw^{r+1}(z_1) & \cdots & Hw^{r+m}(z_1) \\ \vdots & & \vdots \\ Hw^{r+1}(z_m) & & Hw^{r+m}(z_m) \end{vmatrix}$$

$$(5.24)\qquad e_i = \frac{(-1)^i}{E_0} \begin{vmatrix} B_b w^{r+1}(b) & \cdots & B_b w^{r+m}(b) \\ Hw^{r+1}(z_1) & & Hw^{r+m}(z_1) \\ \vdots & & \vdots \\ Hw^{r+1}(z_{i-1}) & & Hw^{r+m}(z_{i-1}) \\ Hw^{r+1}(z_{i+1}) & & Hw^{r+m}(z_{i+1}) \\ \vdots & & \vdots \\ Hw^{r+1}(z_m) & \cdots & Hw^{r+m}(z_m) \end{vmatrix}$$

$$1 \le i \le m$$

and E_0 can be chosen as

$$(5.25) \qquad E_0 = (-1)^{m+1} \begin{vmatrix} M^0 w^{r+1}(z_1) & \cdots & M^0 w^{r+m}(z_1) \\ \vdots & & \vdots \\ M^0 w^{r+1}(z_m) & \cdots & M^0 w^{r+m}(z_m) \end{vmatrix}$$

__Example 5.3__

Suppose the BCs for Example 5.1 are given by

$$(5.26a) \qquad B_0 y(0) = y'(0) + b_0 y(0) = \alpha$$

$$(5.26b) \qquad B_1 y(1) = y'(1) + b_1 y(1) = \beta \quad .$$

Let $m=1$ and $s=1$ for (5.26a) and

$$w^0(t) = -(t-t_1)/h \quad ,$$

$$w^1(t) = (t-t_0)/h \quad ,$$

$$w^2(t) = (t-t_0)(t-t_1)/h^2 \quad ,$$

so that

$$E_0 = M^0 w^2(z_1) = 2/h^2 \quad ,$$

$$(5.27a) \qquad e_0 = Mw^2(z_1)/E_0 = 1 + a^1(z_1)h/2 \quad ,$$

$$(5.27b) \qquad e_1 = -B_0 w^2(0)/E_0 = h/2 \quad ,$$

(5.27c) $$d_0 = 1/E_0 \begin{vmatrix} B_0 w^0(0) & B_0 w^2(0) \\ Mw^0(z_1) & Mw^2(z_1) \end{vmatrix}$$

$$= -1/h - a^1(z_1) + b_0 + b_0 a^0(z_1) h/2 ,$$

and

(5.27d) $$d_1 = 1/E_0 \begin{vmatrix} B_0 w^1(0) & B_0 w^2(0) \\ Mw^1(z_1) & Mw^2(z_1) \end{vmatrix}$$

$$= 1/h + a^1(z_1) + a^0(z_1) h/2 .$$

The approximation to (5.26a) is

(5.28) $$d_0 u_0 + d_1 u_1 = e_0 \alpha + e_1 f(z_1).$$

If $r=1$ and $m=1$ at the right end point, the approximation to (5.26b) is similarly given by

(5.29) $$d'_{-1} u_{J-1} + d_0 u_J = e_0 \beta + e_1 f(z_1) .$$

Letting

$$w^0(t) = -(t - t_J)/h$$

$$w^1(t) = (t - t_{J-1})/h$$

$$w^2(t) = (t - t_{J-1})(t - t_J)/h^2 ,$$

by (5.22), (5.23), (5.24), and (5.25)

$$E_0 = M^{0} w^{2}(z_1) = 2/h^2$$

$$e_0 = Hw^2(z_1)/E_0 = 1 + a^1(z_1)h/2 \quad ,$$

$$e_1 = -B_1 w^2(1)/E_0 = h/2 \quad ,$$

$$d_{-1} = 1/E_0 \begin{vmatrix} B_1 w^0(1) & B_1 w^2(1) \\ Hw^0(z_1) & Hw^2(z_1) \end{vmatrix}$$

$$= -1/h - a^1(z_1) + b_1 + b_1 a^0(z_1)h/2 \quad ,$$

and

$$d_0 = 1/E_0 \begin{vmatrix} B_1 w^1(0) & B_1 w^2(0) \\ Hw^1(z_1) & Hw^2(z_1) \end{vmatrix}$$

$$= 1/h + a^1(z_1) + a^0(z_1)h/2 \quad .$$

5.2 The Order of Consistency

The local truncation error of (5.2) is defined as $\tau_j = H_h y_j - \tilde{f}_j$ where $y_j = y(t_j)$, $y(t)$ is the exact solution of (5.1) subject to appropriate BCs. If $\tau_j \to 0$, as $h_j \to 0$, then the finite difference approximation to (5.1) is said to be consistent. If there is a constant c and p is the largest integer such that

$$|\tau_j| \le ch^p \quad , \text{ as } h_j \to 0$$

then the difference approximation is said to be consistent of order p.

Let $y(t)$ be the exact solution of (5.1) subject to appropriate initial or boundary conditions and assume that $y(t)$ is unique. Taking a Taylor expansion of $y(t)$ at t_j, we get

$$\tau_j = \sum_{i=-r}^{s} d_i \sum_{k=0}^{L} (t_{j+i}-t_j)^k \frac{y_j^{(k)}}{k!} - \sum_{i=1}^{m} e_i H\Big[\sum_{k=0}^{L} (z_i-t_j)^k \frac{y_j^{(k)}}{k!}\Big]$$

$$+ \sum_{i=-r}^{s} d_i (t_{j+i}-t_j)^{L+1} \frac{y^{(L+1)}(b_i)}{(L+1)!} - \sum_{i=1}^{m} e_i H\Big[(z_i-t_j)^{L+1} \frac{y^{(L+1)}(g_i)}{(L+1)!}\Big]$$

$$= \sum_{k=0}^{L} \Big[\sum_{i=-r}^{s} d_i (t_{j+i}-t_j)^k - \sum_{i=1}^{m} e_i H(z_i-t_j)^k\Big] y_j^{(k)}/k!$$

$$+ \sum_{i=-r}^{s} d_i (t_{j+i}-t_j)^{L+1} \frac{y^{(L+1)}(b_i)}{(L+1)!} - \sum_{i=1}^{m} e_i \frac{y^{(L+1)}(g_i)}{(L+1)!} H(z_i-t_j)^{L+1}$$

for some $b_i \in [t_j, t_{j+i}]$, $g_i \in [t_j, z_i]$.

Since (5.2) is exact on P_L and $(t-t_j)^k \in P_L$ $0 \le k \le L$, the quantity in square brackets is zero for each k. If we make the assumption that $h/c \le h_j \le h$ $(1 \le j \le J)$ for some c (such a family of meshes is called <u>quasiuniform</u>), then from the calculation of d_i, e_i in (5.11), (5.12), and (5.13) with the basis functions $w^1(t)$ in (5.14a), (5.14b), it follows that there are constants c_1, c_2, and c_3 which are independent of h such that

$$|d_i| \le c_1 h^{-n} \qquad -r \le i \le s \quad ,$$

$$|e_i| \le c_2 \qquad 1 \le i \le m \quad ,$$

and

$$\left| H(z-t_j)^{L+1} \right| \le c_3 h^{L+1-n} \qquad z_1 \le z \le z_m \quad .$$

Recalling that L is at least r+s+m-1, we have the following theorem.

Thm.5.1 Let the coefficients d_i, e_i and the normalizing factor E be given by (5.11), (5.12), (5.13), respectively. Assume that E≠0, and r+s≥n. If the mesh is quasiuniform and h is small enough then at least n+1 of the d_i are nonzero and the order of consistency of the finite difference approximation (5.2) is greater than or equal to r+s+m-n (Doedel [8],[9] and Lynch-Rice [16]).

Example 5.4

In Example 5.1, r=1, s=1, m=1, and n=2; therefore, the order of the scheme is at least r+s+m-n= 1+1+1-2=1. In fact, it is of order 2 as we will see in Section 5.4.

Example 5.5

In Example 5.2, the order of the scheme is 0+1+2-1=2, which is the order of the trapezoidal rule.

Now, consider the approximation for BCs. The truncation error of (5.16) is

$$\tau_0 = B_{h,a}\, y(a) - e_0 \alpha - \sum_{i=1}^{m} e_i f(z_i)$$

where z_i are collocation points in $[t_0, t_s]$.

The truncation error of (5.21) is

$$\tau_J = B_{h,b}\, y(b) - e_0 \beta - \sum_{i=1}^{s} e_i f(z_i)$$

where z_i are the collocation points in $[t_{J-r}, t_J]$.

We have the following similar result to Thm.5.1.

Thm.5.2 Assume that $E_0 \neq 0$ in (5.20) ((5.25)). If $m=0$, let $s \geq k_1$ ($r \geq k_2$). If $m>0$, let $s+m \geq n$ ($r+m \geq n$). Then the order of consistency of the finite difference approximation to BCs (5.16) ((5.21)) is at least equal to $s+m-k_1+1$ ($r+m-k_2+1$) (Doedel [8][9]).

It follows that the order of consistency of (5.28),(5.29) in Example 5.3 is at least 1+1-1+1=2.

5.3. Stability of the Schemes

It is mentioned in Section 2.1.4 that one should not use numerically unstable methods. Once consistency of a new numerical method is established, for convergency it is necessary to establish stability. In order to guarantee convergence, we will now examine the stability of the schemes described, and in order to apply Kreiss' theory [14], we will assume that the mesh is uniform.

Consider the BVP (5.1) and the BCs

(5.30a) $$B_k(a)y(a) = \sum_{i=0}^{n_k(a)} b_{k,i}(a)\, y^{(i)}(a) = b_k(a) \qquad 1 \leq k \leq n_0$$

$$B_k(b)y(b) = \sum_{i=0}^{n_k(b)} b_{k,i}(b)y^{(i)}(b) = b_k(b) \qquad n_0+1 \le k \le n \tag{5.30b}$$

where $n_0 \ge 1$, $n_k(a) < n$, and $n_k(b) < n$.

A finite difference approximation for (5.1) and (5.30) has the form

$$L_h u_j' = \sum_{i=-r}^{s} d_{j,i} u_{j+i} = \sum_{i=1}^{m} e_{j,i} f(z_{j,i}) = \tilde{f}_j \qquad r \le j \le J-s \tag{5.31}$$

$r \le j \le J-s$

$$B_{h,k}(a)u_0 = \sum_{i=0}^{s_k(a)} d_{k,i}(a)u_i = e_{k,0}(a)b_k(a) \tag{5.32a}$$

$$+ \sum_{j=1}^{m_k(a)} e_{k,j}(a)f(z_{k,j}(a)) = b_k(a) \qquad 1 \le k \le n_0$$

$$B_{h,k}(b)u_J = \sum_{i=-r_k(b)}^{0} d_{k,i}(b)u_{J+i} = e_{k,0}(b)b_k(b) \tag{5.32b}$$

$$+ \sum_{j=1}^{m_k(b)} e_{k,j}(b)f(z_{k,j}(b)) = b_k(b) \qquad n_0+1 \le k \le n.$$

If the approximation is not compact, then $r+s-n$ extra difference equations are required to match the number of equations and the number of unknowns. Suppose these extra equations involve the differential equation and are given by

$$L_h u_j = \sum_{i=0}^{s_j} d_{j,i} u_i = \sum_{k=1}^{m_k} e_{j,k} f(z_{j,k}) = \tilde{f}_j \tag{5.33a}$$

$k_0 \le j \le r-1 \qquad k_0 \ge 0$

(5.33b) $$L_h u_j = \sum_{i=-r_j}^{0} d_{j,i}\, u_{J+i} = \sum_{k=1}^{m_k} e_{j,k}\, f(x_{j,k}) = \tilde{f}_j$$

$$J-s+1 \le j \le J-n+k_0 .$$

Letting $\underset{\sim}{u}_h = (u_0, \dots, u_J)^T$, then (5.31), (5.32), and (5.33) can be expressed as

(5.34) $$L_h \underset{\sim}{u}_h = \underset{\sim}{f}_h$$

where $\underset{\sim}{f}_h$ is the appropriate (J+1)-vector, and L_h is a (J+1)x(J+1) matrix.

Let $\tau_h = (\tau_1(a), \dots, \tau_{n_0}(a), \tau_{k_0}, \dots, \tau_{J-n+k_0}, \tau_{n_0-1}(b), \dots \tau_n(b))$

be the vector of the truncation errors with

$$\tau_k(a) = B_{h,k}(a)\, y_0 - b_k(a) \qquad 1 \le k \le n_0$$

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

and $$\tau_k(b) = B_{h,k}(b)\, y_J - b_k(b) \qquad n_0+1 \le k \le n .$$

The finite difference scheme (5.34) is said to be **stable** if for all sufficiently small h, L_h^{-1} exists and satisfies

$\|L_h^{-1}\| \le c$ for some constant c independent of h.

Here for any (J+1)-vector $\underset{\sim}{g} = (g_0, \dots, g_J)^T$, and for any (J+1)x(J+1) matrix A,

$$\|\underset{\sim}{g}\| = \max_{0 \le j \le J} |g_j|$$

and

$$\|A\| = \max_{g \neq 0} \frac{\|Ag\|}{\|g\|} .$$

Let R denote the translation operator, i.e. $R^j u_i = u_{i+j}$.

Define $D_+ = (R-I)/h$, and let $Int\ u_i$ be the polynomial of degree $\leq i$ which interpolates u at $u_0, u_1, \ldots u_i$.

For a compact scheme, we have

Thm.5.3 (Kreiss [14]) Assume the homogeneous problem corresponding to (5.1) and (5.30) only has the trival solution. If $r+s=n$, then there exists a constant k_1 such that for all solutions of (5.31) and (5.32) an a priori estimate

$$\|D_+^n u\| \leq K_1 (\|u\| + \|\tilde{f}\| + \sum_{k=0}^{n-1} |b_k|)$$

holds. (Here we define $\tilde{f}_i = 0$ for $i=0,1,\ldots r-1$ and $i=J-s+1,\ldots,J$). If the Equations (5.31) and (5.32) are consistent, then these equations have, for every $\tilde{f}$ and $\tilde{b}_k$ and all sufficiently small h, a unique solution u, and there is a constant k_2 such that

$$\|u\| \leq k_2 (\|\tilde{f}\| + \sum_{k=0}^{n-1} |\tilde{b}_k|) .$$

Furthermore, the interpolated function $Int\ u_n$ converges to the solution y of the differential equation. i.e.

$$\lim_{h \to 0} |Int\ u_n - y| = 0 .$$

We now consider the case when $r+s>n$. For later purpose, write (5.31) in the form

$$(5.31') \qquad L_h u_j = S(h) D_+^n u_{j-r} + \sum_{k=0}^{n-1} q_k D_+^k u_{j-r} = \tilde{f}_j$$

where $S(h)$ denotes a uniformly bounded difference operator of the form

$$S(h) = \sum_{k=0}^{r+s-n} S_k(h) E^k$$

and q_k are linear combinations of $d_{j,i}$.

For example, in example 5.1, $S(h)=1+a_j^1 h/2$, $q_1 = a_j^1 + a_j^0 h$, $q_0 = a_j^0$.

Let $x_j = D_+^n u_j$, $j=0,1,\ldots J-n$, then (5.31) can be written as

$$(5.35) \qquad S(h) x_{j-r} = g_j \qquad j=r,\ldots J-s \ .$$

In most applications (5.35) has constant coefficients, i.e. we can write it as

$$\sum_{i=-r}^{s-n} c'_i \hat{E}^i x_j = g_j \qquad j=r,\ldots,J-s$$

where c'_k are constants. In this case, define the characteristic polynomial $c(t)$ associated with (5.31) where

$$c(t) = \sum_{i=0}^{N} c_i t^i \quad \text{with } N=r+s-n, \text{ and } c_i = c'_{i-r} \ .$$

If (5.34) is not compact, then we need $r+s-n$ characteristic polynomials associated with the extra BCs (5.33). Let them have the form

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

$$k_0 \ge 0.$$

Also consider the homogeneous difference equations

(5.37a) $$\sum_{i=0}^{N} c_i v_{j+i-r} = 0 \qquad j=r,r+1,\ldots J,J+1,\ldots$$

with BCs

(5.37b) $$\sum_{i=0}^{N_j} c_{j,i} v_i = 0 \qquad k_0 \le j \le r-1$$

$$\sup_{r \le j < \infty} |v_j| \le \text{constant}$$

and

(5.38) $$\sum_{i=0}^{N} c_{N-i} v_{j+i-r} = 0 \qquad j=J-s,J-s-1,\ldots,0,-1,\ldots$$

with BCs

(5.38b) $$\sum_{i=0}^{j} c_{j,N_j-i} v_{J-i} = 0 \qquad J-s+1 \le j \le J-n+k_0$$

$$\sup_{s \le j < \infty} \left| v_{J-j} \right| \le \text{constant} \quad .$$

Then Kreiss [14] has also shown:

Thm.5.4 Suppose the homogeneous problem corresponding to (5.1) and (5.30) only has the trivial solution. Assume the difference scheme (5.34) is consistent and all roots t_i of the characteristic equation $c(t)=0$ satisfy $|t_i| \ne 1$. If the difference scheme is not compact, also suppose that the difference equations (5.37a,b) and (5.38a,b) have only the trivial solution. Then (5.34) has a unique solution for all sufficiently small h and the difference scheme is stable.

Example 5.6

Consider the problem

(5.39a) $\qquad y'(t)=f(t) \qquad 0\le t\le 1$

(5.39b) $\qquad y(0)=0 \quad .$

Let $r=m=s=1$ and $z_1 = t_{j+1/2}$,

then

$$d^0_{-1} = M^0 w^0(z_1) = \frac{2t_{j+1/2} - (t_j + t_{j+1})}{2h^2} = 0$$

$$d^0_0 = M^0 w^1(z_1) = \frac{2t_{j+1/2} - (t_{j-1} + t_{j+1})}{-h^2} = -1/h$$

$$d^0_1 = M^0 w^2(z_1) = \frac{2t_{j+1/2} - (t_{j-1} + t_j)}{2h^2} = 1/h \quad ,$$

hence $S(h)=R$ and $c(t)=t$. Therefore, $c(t)=0$ has no root on the unit circle. Since $c_0=0$ and $c_1=1$, (5.37a) and (5.38a) only have the trivial solution and by (5.33a) and w^0, w^1 in example 5.2, $d_{0,0}=-1/h$, $d_{0,1}=1/h$, $a_{0,0}=1$.

Hence (5.37b) only has the trivial solution, and so does the homogeneous problem corresponding to (5.39). Since noncompact approximations to (5.39) with $m=1$ are always consistent (see Doedel[9]), the scheme is stable.

Having some roots of $c(t)=0$ lie on the unit circle does not necessarily imply that the finite difference approximation is unstable. By numerical experience, it has been shown that such approximations may lead to stable schemes (Doedel [9]).

5.4. Improved Order with Particular Choice of z

In (5.2), if the z_i are chosen properly, higher order accuracy can be obtained. Doedel [9] only considered the choice of such z_i for which one higher order accuracy is obtained. The details of this analysis are presented below.

Let $w^0, w^1, \ldots, w^{r+s+m-1}$ be a basis of $P_{r+s+m-1}$ defined in (5.14a) and (5.14b). This linearly independent set can be extended to form a basis of P_{r+s+m} by adding a polynomial $w^{r+s+m}(t) \in P_{r+s+m}$ which vanishes at the mesh points. The extra polynomial can be of the form

$$w^{r+s+m}(t) = \prod_{k=1}^{m-1} (t-x_k) \prod_{k=0}^{r+s} (t-t_{j-r+k})$$

where x_k are in $[t_{j-r}, t_{j+s}]$, and satisfy $|x_k - t_j| \le ch$, $1 \le k \le m-1$. If we expand $y(t)$ in terms of $w^1(t)$, i.e.

$$y(t) = \sum_{k=0}^{r+s+m} c_k w^k(t) + O(h^{r+s+m+1}) ,$$

then the truncation error can be written in the form

$$\tau_j = \sum_{i=-r}^{s} d_i y_{j+i} - \sum_{i=1}^{m} e_i f(z_i) = \sum_{k=0}^{r+s+m} c_k \Big[\sum_{i=-r}^{s} d_i w^k(t_{j+i})$$

$$- \sum_{i=1}^{m} e_i M w^k(z_i)] + O(h^{r+s+m-n+1}).$$

The quantity between square brackets vanishes for $0 \le k \le r+s+m-1$, and since $w^{r+s+m}(t_{j+i}) = 0$ $(-r \le i \le s)$, τ_j becomes

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

Hence, it is clear that if the z_i are chosen so that $M^0 w^{r+s+m}(z_i) = 0$ $1 \le i \le m$ where $w^{r+s+m}(t)$ is in P_{r+s+m} and satisfies $w^{r+s+m}(t_i) = 0$ $0 \le i \le s$, then an extra order of consistency can be obtained.

Example 5.7

In Example 5.1, let

$$w^3(t) = (t-t_{j-1})(t-t_j)(t-t_{j+1})$$

and z_1 be the root of $M^0 w^3(t) = 6(t-t_j)$ i.e. $z_1 = t_j$, which is what we chose for z_1 in Example 5.1. Then, as stated in Example 5.4, the order of consistency of the scheme is 1+1=2.

However, if we treat the z_i in (5.2) as unknowns, one could expect that for the special operator M_h, higher orders up to $r+s+2m-n$ can be achieved. It is shown in Lynch-Rice [16] that such z_i's exist and they also offered the special choice of z_i for the case $M=M^0$. This will be discussed next.

For simplicity, we only consider compact schemes $(r+s=n)$ for $M=D^n=M^0$. For the general case of the variable coefficient operator M and sufficiently small h, it can be shown that there is a set of e's and a unique set of m auxiliary points z_i with $t_{j-r}<z_1<\ldots<z_m<t_{j+s}$ such that the high order scheme is exact on P_{2m+n-1} (see Lynch-Rice [16]). However, it is not clear how these z_i's can be found. Since their positions are problem dependent, it would not be practical to construct a high order scheme which is exact on P_{2m+n-1} for a general M. In the following, a generalized result of Lynch-Rice [16] is shown. The basic process is the same as theirs, but r and s are no longer restricted to 0 and n, respectively. We now find the special location of z_i's which would give the order of consistency as high as possible.

On the jth subinterval, since the $w^i(t)$ of (5.14a) are in P_n, their n-th derivatives are constants. When applying them to (5.3), we have

$$(5.40) \qquad d_i^0 - \frac{n!}{\prod_{k=-r,k\neq i}^{s}(t_{j+i}-t_{j+k})} \sum_{j=1}^{m} e_j^0 = 0.$$

If (5.4c) is used, (5.40) gives

$$d_i = \frac{n!}{\prod_{k=-r,k\neq i}^{s}(t_{j+i}-t_{j+k})}$$

which means the operator M_h^0 is n! times the usual divided difference approximation to $M^0=D^n$. Thus

(5.41) $$M_h^0 y(t_j) = \sum_{i=-r}^{s} d_i y(t_{j+i}) = n! \sum_{i=-r}^{s} y(t_{j+i}) / \left[\prod_{k=-r,k\neq i}^{s} (t_{j+i} - t_{j+k}) \right] = n! y[t_{j-r}, \ldots, t_{j+s}],$$

i.e., $M_h^0 y(t_j)$ is the nth derivative of the unique polynomial in P_n which interpolates the values $y(t_{j+i})$ at t_{j+i}, $i=-r,\ldots,s$.

By Taylor's Theorem, $y(t)$ can be expressed as

(5.42) $$y(t) = \sum_{i=0}^{n-1} \frac{D^i y(t_{j-r})(t-t_{j-r})^i}{i!} + \int_{t_{j-r}}^{t} \frac{(t-x)^{n-1} D^n y(t)}{(n-1)!} dx$$

(Goldberg[11]).

Substituting (5.42) into (5.41), since the nth divided difference of an element of P_{n-1} is zero, we get

(5.43) $$M_h^0 y(t_j) = n! \int_{t_{j-r}}^{t_{j+s}} B_n(\bar{t}_j; x) D^n y(x) dx$$

where $B_n(\bar{t}_j; x)$ is the nth divided difference

$g_n[t_{j-r}, \ldots, t_{j+s}; x]$ with respect to t of

$$g_n(t;x) = (t-x)_+^{n-1} = \begin{cases} (t-x)^{n-1}/(n-1)! & \text{if } t>x \\ 0 & \text{otherwise} \end{cases}.$$

Hence $B_n(\bar{t}_j; .)$ is the (n-1)st degree polynomial B-spline with joints at the stencil points. Therefore, the truncation error is

(5.44) $$\tau_j = M_h^0 y(t_j) - \tilde{f}_j$$

$$= n! \int_{t_{j-r}}^{t_{j+s}} B_n(\bar{t}_j;x) D^n y(x)\,dx - \sum_{i=1}^{m} e_i^0 D^n y(z_i)$$

$$= E_h^0 [D^n y]_j$$

where $E_h^0 D^n y$ is the quadrature error in using $\tilde{f}_j$ as an approximation to the intergral of $n! B_n(\bar{t}_j;x) D^n y$. Let

(5.45) $$v_0^i(t) = \prod_{k=0,k\neq i}^{m-1} \frac{(t-t_{j-r+k})}{(t_{j-r+i}-t_{j-r+k})} \qquad i=0,\ldots,m-1 \quad .$$

If e_i^0 are chosen such that

(5.46) $$e_i^0 = n! \int_{t_{j-r}}^{t_{j+s}} B_n(\bar{t}_j;x) v^{i-1}(x)\,dx \qquad i=1,\ldots,m \quad ,$$

then since $\sum_{i=1}^{m} v^{i-1}(t)=1$ and $\int_{t_{j-r}}^{t_{j+s}} B_n(\bar{t}_j;x)\,dx = 1/n!$,

we obtain $\sum_{i=1}^{m} e_i^0 = 1$. But for the e_i's in (5.46), and any y in P_{n+m-1} , $T_h y = E_h^0 [D^n y] = 0$. Since $B_n(\bar{t}_j;x)$ is positive on $[t_{j-r}, t_{j+s}]$, define the inner product

(5.47) $$(u,v) = \int_{t_{j-r}}^{t_{j+s}} B_n(\bar{t}_j;x) u(x) v(x)\,dx \quad .$$

Let $b_0, b_1, \ldots$ with $b_i \in P_i$ be the normalized orthogonal polynomials with respect to this inner product (call them the B-spline orthogonal polynomials). Each b_i has i distinct real zeros in (t_{j-r}, t_{j+s}) (call them B-spline Gauss points). If the m auxiliary points z_i are the B-spline Gauss points for b_m, since Gauss quadrature is exact on P_{2m-1} for Gauss points, the high order finite difference approximation is exact on P_{2m+n-1}.

A sequence of orthonormal polynomials can be generated by a 3-term recurrence relation as follows :

$$\tilde{b}_{-1}(t)=0 \quad ,$$

$$\tilde{b}_{0}(t)=1 \quad ,$$

(5.48)
$$\tilde{b}_{i+1}(t)=(t-B_i)\tilde{b}_i(t)-c_i\tilde{b}_{i-1}(t) \qquad i=0,1,2,\ldots \quad ,$$

where
$$B_i=\int_{t_{j-r}}^{t_{j+s}} x\tilde{b}_i^2(x)B_n(\bar{t}_j;x)\,dx/s_i \quad ,$$

$$c_i=\begin{cases} 0 & \text{if } i=0 \\ s_i/s_{i-1} & \text{if } i\neq 0 \end{cases} \quad ,$$

$$s_i=\int_{t_{j-r}}^{t_{j+s}} \tilde{b}_i^2(x)B_n(\bar{t}_j;x)\,dx \quad ,$$

and
$$b_i(t)=\tilde{b}_i(t)/s_i \qquad i=0,1,2,\ldots \quad .$$

Since the z_i are roots of $b_m(t)=0$, it is obvious that

$$\sum_{i=1}^{m} e_i^0 b_m^2(z_i)=0$$

and because (b_m, b_m) is positive, τ_j in (5.44) is not zero for polynomials in P_{2m+n}, that is, the approximation is not exact on P_{2m+n}. Hence the high order scheme has order at most $2m$. If only j of the m Gauss points are used, then the high order scheme is of order $m+j$.

Example 5.8

Consider the operator $M=D^2$. Let $m=2, r=s=1$. For convenience, consider a uniform mesh. Then

$$B_2(\bar{t}_j;x)=\begin{cases}(x-t_{j-1})/2h^2 & t_{j-1}<x<t_j\\(t_{j+1}-x)/2h^2 & t_j<x<t_{j+1}\\0 & \text{otherwise}\end{cases},$$

and (5.45) gives

$$v^0(t)=(t-t_j)/-h \quad ,$$

$$v^1(t)=(t-t_{j-1})/h \quad .$$

From (5.46), we get

$$e_1^0=2\int_{t_{j-1}}^{t_{j+1}} B_2(\bar{t}_j;x)v^0(x)\,dx=0 \quad ,$$

$$e_2^0 = 2\int_{t_{j-1}}^{t_{j+1}} B_2(t_j;x)\, v^1(x)\,dx = 1 ,$$

and by (5.48), we have

$$\tilde{b}_0 = 1$$

$$\tilde{b}_1 = t - t_j$$

$$\tilde{b}_2 = (t-B_1)(t-t_j) - c_1 = (t-t_j)(t-t_j) - h^2/6$$

$$= (t-(t_j + h/\sqrt{6}))(t-(t_j - h/\sqrt{6})) .$$

Hence the scheme

$$(u_{j-1} - 2u_j + u_{j+1})/h^2 = f(t_j + h/\sqrt{6}) \quad j=1,2,\ldots,J-1$$

for $y'' = f(t)$ (with appropriate BCs) is of order 4.

5.5. An Equivalence Between Finite Difference and Collocation Methods

In this section, an equivalence between the finite difference methods and collocation methods is presented.

Consider the case $m=1$. Let $w^1(t)$ in (5.10) be the set of basis functions for our collocation method. Then the collocation solution $s(t)$ is

$$s(t) = \sum_{l=0}^{r+s} c_l\, w^1_l(t) .$$

Since $w^1_l(t_{j-r+k}) = \delta_{lk}$ $0 \le l,k \le r+s,\ j=0,\ldots,J$,

$s(t_{j-r+1})=c_1$. If we write the collocation solution at t_{j-r+1} as u_{j-r+1}, then

$$s(t)=\sum_{l=0}^{r+s} w^l(t)\,u_{j-r+l} = \sum_{i=-r}^{s} w^{r+i}(t)\,u_{j+i} .$$

The collocation method for one collocation point and $w^1(t)$ in (5.10) requires u_{j+i} to satisfies

$$\text{(5.49)} \qquad \sum_{i=-r}^{s} Bw^{r+i}(t)\,u_{j+i} = f(z_1) .$$

On the other hand, from (5.11)

$$d_i = Bw^{r+i}(z_1) \qquad -r \le i \le s .$$

The left sum in (5.2) is

$$\text{(5.50)} \qquad \sum_{i=-r}^{s} d_i u_{j+i} = \sum_{i=-r}^{s} Bw^{r+i}(z_1)\,u_{j+i} ,$$

and right sum in (5.2) gives

$$\text{(5.51)} \qquad \sum_{i=1}^{m} e_i f(z_i) = f(z_1) .$$

By (5.50) and (5.51) we have

$$\sum_{i=-r}^{s} Bw^{r+i}(z_1)\,u_{j+i} = f(z_1)$$

which from (5.49) is collocation with one collocation point.

5.6. Work Estimates

In this section, the computational aspects of the higher order finite difference schemes are considered. Since the HODIE methods of Lynch-Rice [16] are similar to the above methods (call them Doedel's methods), comparison of the computational work for the same global accuracy is made. Comparison of efficiency with that of collocation methods using B-spline and Gaussian points is also considered.

To more efficiently evaluate d_i and e_i in (5.2), one can find the coefficients by solving the linear algebraic system (5.3) instead of calculating the determinents in (5.11) and (5.12). This is how the HODIE methods find the coefficients. If the basis functions are chosen such that they satisfy (5.10), then (5.3) gives

$$(5.52) \qquad d_i = \sum_{j=1}^{m} e_j Mw^{r+i}(z_j) \qquad -r\le i\le s$$

$$(5.53) \qquad \sum_{j=1}^{m} e_j Mw^{l}(z_j)=0 \qquad r+s+1\le l\le r+s+m-1$$

for the interior subintervals.

If the normalization equation is chosen such that $e_1=1$, then (5.53) becomes

$$(5.54) \qquad A\underset{\sim}{e}=-\underset{\sim}{b}$$

where $A=(a_{ij})_{i,j=1}^{m-1}=(Mw^{r+s+i-1}(z_j))_{i,j=2}^{m}$, $\underset{\sim}{e}=(e_2,\dots,e_m)^T$, and

$\underset{\sim}{b}=(Mw^{r+s+1}(z_1),\dots,Mw^{r+s+m-1}(z_1))^T$.

After the e_i are determined from (5.54), d_i can be found easily from (5.52), and the finite difference approximations at mesh points are the solutions u_j of (5.2).

Lynch-Rice [16] found that to obtain e_i's it is computationally more efficient to use a different set of basis functions ([16],p.363). As a result, (5.52) is no longer valid for the basis they considered, and so, one has to solve a system of r+s+1 algebraic equations to find the d_i. Their set of basis functions is therefore less efficient for evaluating d_i, and so, only basis functions satisfying (5.10) are considered here.

Consider the problem (5.1) subject to (5.30a,b) at a uniform partition $t_K=kh$, $k=0,\dots J$, and a general set of m auxiliary points. Let MT be a multiplication/division time and F be a function evaluation time. Suppose that the values $w^l(z_j)^{(i)}$,$0\le i\le n$, $0\le l\le r+s+m-1$, $1\le j\le m$, have been previously computed and stored (they do not depend on the subintervals), then the setup time for A and b in (5.54) is $((m-1)(m-1)n+n(m-1))MT+nmF$. The function evaluation is for a general set of z_i. If z_i are at mesh points for some i, function values at these points can be stored beforehand and do not need to be recomputed. If (5.54) is solved by Gaussian elimination without pivoting, the solution time is $[(m^2+m-3)(m-1)/3]MT$. In (5.52), for a fixed i, it takes nmMT time to evaluate $Mw^{r+i}(z_j)$ (since the $a_K(z_j)$ have already been calculated in (5.54)). Since $e_1=1$, it takes another $(m-1)MT$ time to find d_i. The total time required to evaluate d_i , $-r\le i\le s$ and e_i , $1\le i\le m$, is therefore

$[nm(m-1)+(m^2+m-3)(m-1)/3+(r+s+1)(nm+m-1)]MT+mnF.$

For simplicity, we only consider the case where there are the same number of auxiliary points in the consecutive subintervals involved in each row of (5.34). To get an order of $r+s+m-n$ for a complete scheme, in (5.32a) (or (5.32b)), one has to take i from 0 to at least $v=r+s-n+n_K(a)-1$ (or $-v'=-(r+s-n+n_K(b)-1$ to 0) for each k. v (or v') is nonpositive only when $r+s=n$ and $n_K(a)=0,1$ (or $n_K(b)=0,1$) (recall that $r+s\geq n$). For the case $n_K(a)=0$ (or $n_K(b)=0$), no approximation equation is needed since we have been given the exact solution at a (or b). If $n_K(a)=1$ (or $n_K(b)=1$), set $v=1$ (or $v'=1$) and pick $m-1$ auxiliary points in $[t_0, t_1]$ (or $[t_{J-1}, t_J]$). As in (5.52) and (5.53), e_i in (5.19) can be calculated more efficiently by solving the system of m equations

$$e_0 B_{k,a} w^k(a) + \sum_{j=1}^{m} e_j M w^k(z_j) = 0, \quad k=v+1,\dots,v+m, \tag{5.55}$$

and d_i in (5.17) can be evaluated by

$$d_i = e_0 B_{k,a} w^i(a) + \sum_{j=1}^{m} e_j M w^i(z_j) \qquad 0\leq i\leq v. \tag{5.56}$$

Then e_i in (5.24) can be calculated by solving

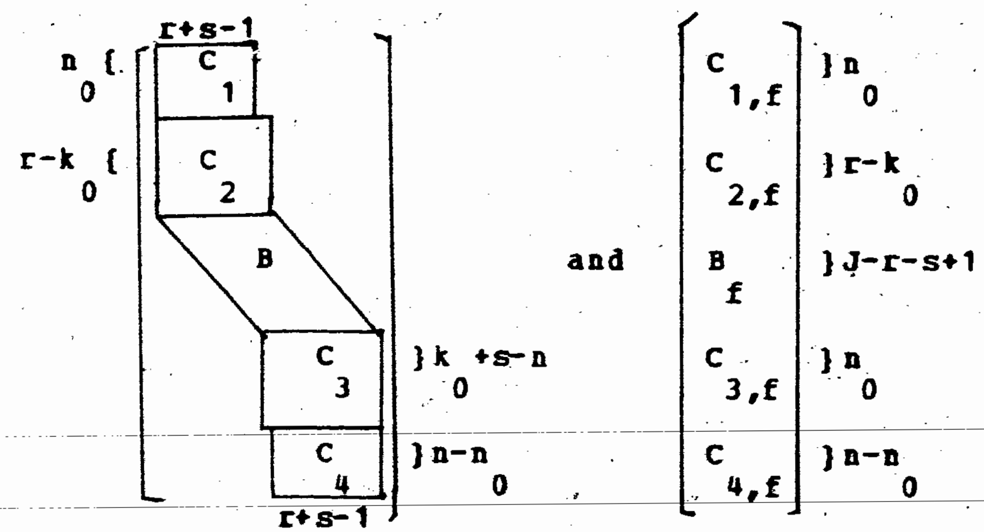
$$e_0 B_{k,b} w^k(b) + \sum_{j=1}^{m} e_j M w^k(z_j) = 0, \quad k=v'+1,\dots,v'+m, \tag{5.57}$$

and d_i in (5.22) can be evaluated from

$$d_i = e_0 B_{k,b} w^{r+i}(b) + \sum_{j=1}^{m} e_j M w^{r+i}(z_j) \qquad -v'\leq i\leq 0. \tag{5.58}$$

Using the same argument as before, when $n_k(a) = 1$ (or $n_k(b) = 1$), there are $2(m-1)+n(m-1)(m-1)+(m-1)(m^2+m-3)/3+2(m+n)$ multiplications to approximate the k-th boundary condition. If $v>0$ ($v'>0$), there are $(n_k(a)+1)m+nm^2+m(m^2+3m-1)/3+(v+1)(n_k(a)+n+m)$ (or $(n_k(b)+1)m+nm^2+m(m^2+3m-1)/3+(v'+1)(n_k(b)+n+m)$) multiplications to approximate k-th boundary condition. Denote the number of multiplications required for approximating a BC by MT_k. The number of function evaluations for approximating BCs also depends on $n_k(a)$ and $n_k(b)$, e.g. if all $n_k(a)$ are the same and not equal to 0 or 1 for $k=1,\ldots,n_0$, then only nmF is needed for the first n_0 BCs. Denote the number of function evaluations for approximating the first n_0 BCs by F_a and the last $n-n_0$ by F_b. In (5.33a) (or (5.33b)), i has to go from 0 to at least r+s (or from -(r+s) to 0) for each j. Since $n_k(a)$ (or $n_k(b)$) is less than n, $r+s-n+n_k(a)-1$ (or $r+s-n+n_k(b)-1$) is at most r+s-2. Hence, L_h and f_h in (5.34) have the form

(5.59)



where B is a trapezoid of width r+s+1 and height J-r-s+1. Depending upon $n_k(a)$ and $n_k(b)$, some elements of C_1 and C_4 may be zero. Note that one has to take different $r-k_0$ sets of auxiliary points for C_2 and different k_0+s-n sets of auxiliary points for C_3, otherwise, some rows of C_2 or C_3 will be identical and L_h will be singular. The setup time for L_h and $\underset{\sim}{f}_h$ is therefore

$C_1, C_{1,f}$	$F_a + \sum_{k=1}^{n_0} MT_k$
$C_2, B, C_{2,f}, B_f$ $C_3, C_{3,f}$	$(J-n+1)\{[q+(m-1)]MT+m(n+1)F\}$
$C_4, C_{4,f}$	$F_b + \sum_{k=n_0+1}^{n} MT_k$

where $q=nm(m-1)+(m^2+m-3)(m-1)/3+(r+s+1)(nm+m-1)$.

If Gaussian elimination without pivoting is used to solve (5.34), then the solution time is

$$(n_0+r-k_0)[\sum_{j=0}^{n_0-1}(r+s-j)+\sum_{j=2}^{k_0+2-r}(r+s-n_0+j)+(r+s+1-n_0)(J-2r-s$$

$$-1-n_0+k_0)+(r+s+1-n_0)(2n_0+2r+2s-2n-1)+\sum_{j=1}^{u}j(j+2)]MT$$

for getting an upper triangular matrix and

$$[\sum_{j=1}^{u}j+u(j-r-s)+\sum_{j=1}^{r-k_0}(u+j)+\sum_{j=1}^{n_0}(r+s-j)]MT$$ for back substitution, where $u=s+1+k_0-n_0$.

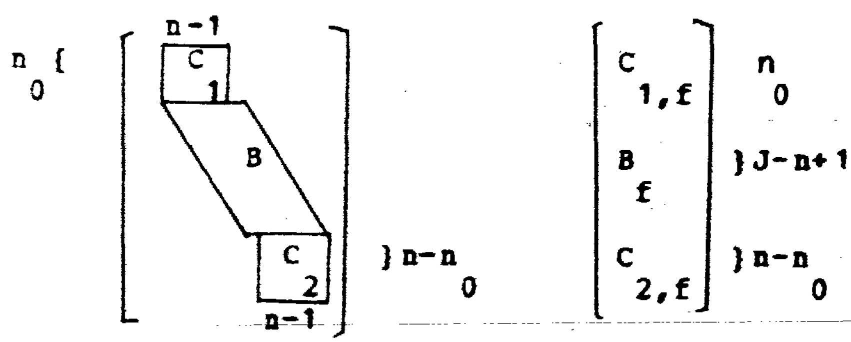
Note that the work estimates calculated are based on an $n_0 \times$

$(r+s-1)$ matrix C_1 and an $(n-n_0)\times(r+s-1)$ matrix C_3. Since some elements of C_1 and C_3 may be zero, the actual work required should be no more than the work calculated from the formula.

Now consider the HODIE methods. Recall that in the HODIE methods $r=0$ and $s=n$. Hence, to compare with Doedel's method which gives the same global accuracy of order $r+s+m-n$, one must select $p=r+s+m-n$ auxiliary points in each subinterval $[t_j, t_{j+n}]$ for a HODIE method.

Again, let $e_1=1$ and assume all values of $w^1(z_j)^{(i)}$ are already stored. For approximating BCs, consider the case when p points are collocated. Then in (5.32a) (or (5.32b)), i goes from 0 to $n_K(a)-1$ (or $n_K(b)-1$). As before, if $n_K(a)=0$, no approximation is needed. If $n_K(a)=1$, then one finds d_0, d_1 from $e_0, e_1, \ldots, e_{p-1}$. Let F_a' and F_b' be the function evaluation time for approximating BCs at a and b, respectively, and let MT_K' be the multiplication time for each BC. Since the HODIE methods are compact, no extra equations are required. L_h and $\underset{\sim}{f}_h$ of (5.34) then have the form

(5.60)



where B is a trapezoid of width n+1 and height J-n+1. Hence, setup time for (5.34) is

$C_1, C_{1,f}$	$F'_a + \sum_{k=1}^{n_0} MT'_k$
B, B_f	$(J-n+1)[p^2n+p(p^2+3p-1)/3)MT+pnF]$
$C_4, C_{4,f}$	$F'_b + \sum_{k=n_0+1}^{n} MT'_k$.

If Gaussian elimination without pivoting is used to solve (5.34), since some elements of C_1 and C_2 may be zero, the total time for the second part of the implementation is at

most $[n_0 \sum_{j=0}^{n_0-1}(n-j) + n_0(n+2-n_0)(J-n+1-n_0) + (n+2-n_0)(n+n_0-3)$

$+ \sum_{j=1}^{n-3}(n-j+2)(n-j)]MT$ to get an upper triangular matrix and

$[(n-n_0)(n-n_0+1)/2 + (n-n_0+1)(J-n+1) + (2n-n_0-1)n_0/2]MT$ to

obtain the solutions by back substitution.

The comparison below is done for second-order differential equations subject to Dirichlet BCs since they are the most important case and are simple. Consider the problem

$$My(t) = y''(t) + a_1(t)y'(t) + a_0(t)y(t) \qquad a \le t \le b$$

$$y(a) = 0 = y(b) \quad .$$

For this case, $n=2$, $n_0=1$, $n_1(a)=0=n_1(b)$, $(L_h)_{1,1} = (L_h)_{(J+1),(J+1)}$ $=1$, and $(f_h)_1 = (f_h)_{(J+1)} = 0$. The comparison is presented in Table 5.1 for three different orders of accuracy (4, 6, and 8).

The data for collocation using B-splines and Gauss points is derived from Russell-Varah [20].

Table 5.1

	order		
r+s-n=	4	6	8
0	(J-1) (82MT+12F)	(J-1) (186MT+18F)	(J-1) (354MT+24F)
1	(J-1) (60MT+9F)	(J-1) (144MT+15F)	(J-1) (284MT+21F)
2	(J-1) (42MT+6F)	(J-1) (110MT+12F)	(J-1) (226MT+18F)
3	(J-1) (27MT+3F)	(J-1) (83MT+9F)	(J-1) (179MT+15F)
4		(J-1) (60MT+6F)	(J-1) (140MT+12F)
collo-cation	J(32MT+4F)	J(82MT+6F)	J(164MT+8F)

The first row of Table 5.1 gives the computational work for the HODIE methods (Lynch-Rice [16]), and the first five rows give operation counts for five more general different Doedel's schemes. Lynch-Rice [16] picked $t_{j+1} \in [t_j, t_{j+2}]$ as one of the auxiliary points. Since the central mesh point of an odd-number-point difference operator is a zero of every odd-degree generalized B-spline orthogonal polynomial, one higher order of accuracy is obtained. In this case, one would expect that, for the same order of accuracy, the counts of the HODIE regular case in Lynch-Rice [16] is smaller than that of the HODIE methods we consider here. From Table 5.1, it is obvious that as r+s increases, the operation count for a given order decreases. It seems on the basis of operation counts that one should pick as few auxiliary points as possible and let $r+s$ be large. This view is supported in the numerical examples given in the next section.

If we compare the first row with the other rows of Table 5.1, Doedel's methods are more efficient than the HODIE methods for large $r+s-n$. By comparing the data in Table 5.1 with that of Table 9-2 in Lynch-Rice [16], we conclude that Doedel's methods are comparable even to the HODIE Gauss-type case when $r+s-n$ is large enough. Collocation with B-splines is also competitive in work for second-order differential equations.

5.7. Experimental Results

From Table 5.1, we have seen that if the number of auxiliary points is chosen as small as possible, Doedel's methods are much more efficient than the HODIE methods. Numerical experiments have been run to support both the theorems in the previous sections and the above conclusion. All computations were performed on the SFU IBM 3033 using double precision arithmetic. In each experiment, the mesh considered is equal spaced. For the HODIE methods, auxiliary points are chosen such that $z_{j,i}=t_j+(i-1)h/J$. For Doedel's methods, only one auxiliary point is used, it being chosen as the midpoint of $[t_j, t_{j+s}]$, $j=0,\ldots,J-s$. Note that the matrix L_h obtained when $r=r_0$, $s=s_0$, is identical with the matrix L_h^* obtained when $r=0$, $s=r_0+s_0$, and z_i' in $[t_j, t_{j+r_0+s_0}]$ are at the same location corresponding to z_i in $[t_{j-r_0}, t_{j+s_0}]$.

Since we consider the case $m=1$, if the order of consistency is required to be greater than one for Doedel's methods then $r+s$ must be greater than n, and hence $r+s-n$ extra finite difference equations are needed. In the experiments, they are divided into two sets. If $r+s-n$ is even, half of them are set to approximate the first $s+1$ solutions $u_0,\ldots,u_s$; if $r+s-m$ is odd, then $(r+s-n+1)/2$ equations are defined for $u_0,\ldots,u_s$. In both cases, the others are set for the last $s+1$ approximate solutions $u_{J-S},\ldots,u_J$. In $[t_0,t_S]$, the auxiliary point involved in the i-th equation of these extra equations is taken to be the (i+1)st mesh point. In $[t_{J-S},t_S]$, the auxiliary point is defined to be the reflection of the corresponding one in $[t_0,t_S]$.

In the following tables, numerical results are shown for a number of cases. Orders of consistency considered are 2, 4, 6, and 8. The results using Doedel's methods with $m=1$ are the rows marked s=. The rows marked m= are the results using the HODIE methods. For collocation methods, numbers of collocation points used are 2,3, and 4 for order 4,6,and 8, respectively. The results which use collocation methods with B-splines, Gaussian points, and uniform meshes (COLLO) are given in the last columns. The notation .220-1 stands for $.220\ 10^{-1}$. In the first column, (o) means the order of the schemes. J is the number of subintervals.

Example 5.9

Consider the DE

$$y''+y-2y=2(1-6t)e^{t}$$

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

The solution of this problem is $y(t)=2t(1-t)e^{t}$. The problem has been used by Doedel [8] [9]. The results are given in Table 5.2.

Table 5.2

(o)		J=8	J=16	J=32	J=64	J=128	COLLO
2	m=2	.220-1	.558-2	.140-2	.351-3	.877-4	
	s=3	.137-1	.348-2	.876-3	.219-3	.548-4	
4	m=4	.564-4	.354-5	.223-6	.139-7	.862-9	J=64
	s=5	.537-4	.623-5	.440-6	.283-7	.179-8	.735-9
6	m=6	.608-7	.958-9	.158-10	xx	xx	J=16
	s=7	.826-6	.123-7	.187-9	.307-11	.306-13	.800-11
8	m=8	.398-10	xx	xx	xx	xx	J=8
	s=9	--	.212-10	.626-13	xx	xx	.119-12

xx Contaminated by roundoff.

Example 5.10

Let the DE be given by

$$y''-4y=4\cosh(1)$$

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

The solution to this problem is $y=\cosh(2t-1)-\cosh(1)$. The problem has been used by Lynch-Rice [16] (note that they had a mistake in [16], $f(t)=4\cosh(1)$, not $f(t)=2\cosh(1)$). The results are given in the following table.

Table 5.3

(o)		J=8	J=16	J=32	J=64	J=128	COLLO
2	m=2	.795-2	.198-2	.496-3	.124-3	.310-4	
	s=3	.496-2	.124-2	.310-3	.775-4	.194-4	
4	m=4	.324-4	.203-5	.127-6	.794-8	.491-9	J=64
	s=5	.324-4	.359-5	.251-6	.162-7	.102-8	.168-9
6	m=6	.751-7	.118-8	.181-10	xx	xx	J=16
	s=7	.942-6	.148-7	.229-9	.364-11	.282-13	.360-11
8	m=8	.103-9	xx	xx	xx	xx	J=8
	s=9	--	.552-10	.144-12	xx	xx	.458-13

The last example shown is

Example 5.11

The DE $y''-400y=400\cos^2\pi t+2\pi^2\cos 2\pi t$

$$y(0)=0=y(1)$$

has solution $y=e^{-20}e^{20t}/(1+e^{-20})+e^{-20t}/(1+e^{-20})-\cos^2\pi t$.

The results are shown below.

Table 5.4

(o)		J=8	J=16	J=32	J=64	J=128	COLLO
2	m=2	.700+0	.149+0	.296-1	.665-2	.156-2	
	s=3	.724-1	.435-1	.153-1	.370-2	.928-3	
4	m=4	.137+0	.128-1	.723-3	.427-4	.253-5	J=16
	s=5	.270-1	.740-2	.512-3	.522-4	.431-5	.219-3
6	m=6	.265-1	.736-3	.105-4	.502-6	.230-8	J=16
	s=7	.208-1	.364-2	.137-3	.191-5	.269-7	.182-5
8	m=8	.380-2	.273-4	.974-7	.356-9	xx	J=16
	s=9	--	.173-2	.274-4	.117-6	.234-9	.891-8

From the results, the HODIE methods and Doedel's methods are quite competitive with each other. Comparing their results with that of COLLO, one requires larger values of J for the finite difference methods to achieve comparable accuracy. Taking the ratios of adjacent numbers in each row of the above tables, it is evident that both the HODIE methods and Doedel's methods give orders of consistency as predicted and there is no numerical instability. Since the implimentation of the HODIE methods involves solving an $(m-1)\times(m-1)$ matrix for each row of (5.34), the execution time of high order HODIE methods is much longer than of Doedel's methods for the same order of accuracy. From Table 5.1, we see that, when $m=1$, Doedel's methods are much cheaper than the HODIE methods.

6. Conclusion

General forms of BVPs and IVPs have been given, existence and uniqueness theorems for solutions of IVPs and BVPs have been provided. Stability properties of IVPs and two point BVPs have been discussed. Some well-known numerical methods for solving BVPs have been presented.

We have seen several equivalences between the most common numerical methods for solving differential equations. Some of them hold only in special circumstances. e.g. in Section 5.5, only one collocation point was considered, and there are many cases which have not been taken into account. Consequently, one remaining task would be to find more relationships between these methods (e.g., the HODIE methods and collocation when $m>1$)

We have discussed the high order finite difference methods thoroughly. As with finite element methods, when solved by the finite difference methods differential equations need not be converted to first order systems. Though one can get superconvergence using general B-spline and Gauss points, it is not practical for general n-th order differential equations. For the case $M^0=D^n$, one can find the B-spline Gauss points and the right-hand-side coefficients easily by using the formulae discussed before. While for general M, the location of the general B-spline Gauss points depends on M and the mesh points. Thus, it would be difficult to have a practical code for the finite difference methods which gives high order accuracy. However, the methods have been shown to be computationally efficient. Operations counts and numerical results have shown that, to find the approximations by Doedel's methods more efficiently, one should use schemes which only involve one auxiliary point.

REFERENCES

1. Ascher U. & Christiansen J. & Russell R.D., A Collocation Solver for Mixed Order Systems of Boundary Value Problems, Math. Comp.,No.33, 1979.

2. Ascher U. & Russell R.D.,Evaluation of B-splines for Solving Systems of BVPs, Technical Report 77-14,Nov.,1977.

3. Ascher U. & Russell R.D., Numerical Solutions of Boundary Value Problems, manuscript ,1981.

4. Ascher U. & Weiss R., Collocation for Singular Perturbation Problems I : First Order Systems with Constant Coefficients, To appear in SINUM.

5. Bailey P.B. & Shampine L.F. & Waltman P.E., Nonlinear Two Points Boundary Value Problems, Academic Press, New York, 1966.

6. DE Boor C.R., On Calculating with B-splines, J. of Approximation Theory v.6, pp50-62, 1972.

7. DE Boor C.R., Package for Calculating with B-splines, SIAM J. Numer. Anal. Vol 14, No. 3, June 1977.

8. Doedel E.J., The Construction of Finite Difference Approximations to Ordinary Differential Equations, SIAM J.Numer. Anal. Vol.15, No.3, June, 1978.

9. Doedel E.J., Difference Methods for Ordinary Differential Equations with Applocations to Parabolic Equations, Ph.D. Thesis UBC Jan., 1976.

10. Franklin J.N., Matrix Theory, Prentice-Hall Inc. Englewood Cliffs, N.J., 1968.

11. Goldberg R.R., Methods of Real Analysis, Xerox College Publishing, 1964.

12. Keller H.B., Numerical Methods for Two Points Boundary Value Problems, Blaisdell Pub. Co., Waltham, Mass., 1968.

13. Keller H.B. & Lentini M., Invariant Imbedding, the Box Scheme and an equivalence Between them, SINUM No. 19, 1982.

14. Kreiss H-O, Difference Approximations for Boundary and Eigenvalue Problems for Ordinary Differential Equations, Math. Comp. No. 26, v. 19, July, 1972.

15. Lentini M. & Osborne M. & Russell, R.D., The Close Relationships Between Methods for Solving ODE BVPs, manuscript, Jan. 1983.

16. Lynch R.E. & Rice J.R., A High-Order Difference Method for Differential Equations, Math. Comp. V.34, No.150, pp333-372, 1980.

17. Mattheij R.M.M., The Stability of LU-Decompositions of Block Tridiagonal Matrices, Manuscript, Mathematisch Institut, Katholicke Universiteit, Nijmegan, The Netherlands, 1982.

18. Ortega J.M., Numerical Analysis, 2nd ed., Academic Press, N.Y. & London, 1972.

19. Reid W.T., Riccati Differential Equations, Academic Press N.Y. & London, 1972.

20. Russell R.D. & Varah J.M., A Comparison of Global Methods for Linear Two-Point Boundary Value Problems, Math. Comp. v.29, No.132, 1975.

21. Scott M.R. & Watts H.A., Computational Solution of Linear Two-Point BVPs via Orthonormalization, SINUM v.14, No.1, 1977.

22. Varah J.M., A Comparison of Some Numerical Methods for Two Point Boundary Value Problems, Math. Comp. v.28, No.127, July, 1974.

23. Scott M.R. & Watts H.A., SUPORT- A Computer Code for Two-Point BVPs via Orthonormalization, SAND75-0198 Sandia

Laboratories, Albuquerque, New Mexico, 1975.