AN ANALYSIS OF DISCRETISATION METHODS FOR ORDINARY DIFFERENTIAL EQUATIONS .

Neil Pitcher ,

The Queen's College ,

Oxford .

D. Phil. thesis submitted in 1980 .

"Unto the pure all things are pure ."

The apostle Paul ... Titus 1:15 (A.V.)

BODLEIAN LIBRARY OXFORD
DEPOSITED
THESIS
22.4.81

ACKNOWLEDGEMENTS .

I should like to express my gratitude to my supervisor , Dr. Sean Mc.Kee , in the first place for suggesting this area of work , and for his helpful comments and suggestions throughout my time of research .

My thanks are also due to the Science Research Council for their financial support for my three years at The Queen's College , Oxford , which I have enjoyed very much .

Finally I should like to mention my parents , especially my late father , Mr. Rex Pitcher , to whom this thesis is dedicated , for their affection , help and encouragement over the many years of my education .

To my late father , Mr. Rex Pitcher .

An Analysis of Discretisation Methods for Ordinary Differential Equations , by Neil Pitcher of The Queen's College , Oxford .

A thesis for the degree of D.Phil. submitted in Trinity term 1980 .

ABSTRACT .

Numerical methods for solving initial value problems in ordinary differential equations are studied . A notation is introduced to represent cyclic methods in terms of two matrices , A_h and B_h , and this is developed to cover the very extensive class of m-block methods . Some stability results are obtained and convergence is analysed by means of a new consistency concept , namely optimal consistency . It is shown that optimal consistency allows one to give two-sided bounds on the global error , and examples are given to illustrate this . The form of the inverse of A_h is studied closely to give a criterion for the order of convergence to exceed that of consistency by one . Further convergence results are obtained , the first of which gives the orders of convergence for cases in which A_h and B_h have a special form , and the second of which gives rise to the possibility of the order of convergence exceeding that of consistency by two or more at some stages . In addition an alternative proof is given of the superconvergence result for collocation methods . In conclusion the work covered is set in the context of that done in recent years by various authors .

CONTENTS .

1. Introduction .

1.1. Preliminaries .

1.2. Linear Multistep Methods .

1.3. Cyclic Linear Multistep Methods .

2. Block Semi Explicit Methods .

2.1. Definitions .

2.2. Cyclic Linear Multistep Methods .

2.3. Predictor-Corrector Methods .

2.4. Explicit Runge-Kutta Methods .

2.5. Hybrid Methods .

2.6. Extrapolation .

3. Block Implicit Methods .

3.1. Definitions .

3.2. Advanced Linear Multistep Methods .

3.3. Implicit Runge-Kutta Methods .

3.4. One-Leg Methods .

3.5. Collocation Methods .

4. Block Methods .

4.1. Truncation Error .

4.2. The Numerical Stability of Block Semi Explicit Methods .

4.3. Zero Stability .

4.4. Convergence Theorems .

5. Two-Sided Error Bounds .

5.1 Optimal Consistency .

5.2 The Form of A_h^{-1} .

5.3 Improved Convergence .

6. Optimal Consistency of Orders $(p_1, p_2, \ldots, p_m)$.

6.1 Two-Sided Error Bounds .

6.2 A Special Class of Methods .

6.3 Two-Sided Error Bounds .

6.4 A Further Convergence Theorem .

6.5 Collocation .

7. Conclusions .

7.1 The Work of Spijker .

7.2 The Work of Albrecht .

7.3 Comparison with some Results of Stetter .

7.4 The Work of Skeel .

7.5 The Work of Cooper .

CHAPTER ONE .

INTRODUCTION.

1.1 Preliminaries.

Let T be some positive real number. We shall be concerned in this thesis with the numerical solution of the initial value ordinary differential equation $y'(t) = f(t,y(t))$ (1.1.1) with $y(0) = y_o$.

We take f to be in the class of Lipschitz continuous functions in which for every f there exists a Lipschitz constant L such that for any y_1, y_2 and $t \in [0,T]$,

$$|f(t,y_1) - f(t,y_2)| \leq L|y_1 - y_2| \quad (1.1.2) .$$

Let us use a basic steplength $h > 0$ along the interval $[0,T]$ such that $N = T/h$ is a positive integer . For any real number $i \in [0,N]$ we define $t_i = ih$, so that $t_o = 0$ and $t_N = T$.

A numerical method is a computational means of obtaining approximations to $y(t)$ at the points $\{t_i : i \in S\}$, where S is a finite subset of $[0,N]$. For any i we denote the approximation to $y(t_i)$ by y_i .

1.2 Linear Multistep Methods.

Any linear multistep method may be written in the form

$$\sum_{j=0}^{k} \alpha_{k-j} y_{i-j} = h \sum_{j=0}^{k} \beta_{k-j} f(t_{i-j}, y_{i-j}) \quad , \; i = k(1)N \quad (1.2.1) \quad ,$$

where $k, \alpha_o, \alpha_1, \ldots, \alpha_k, \beta_o, \beta_1, \ldots, \beta_k$ are real constants independent of h , and k is known as the stepnumber . A single application of this method computes y_i from the known values

$y_{i-1}, y_{i-2}, \cdots, y_{i-k}$.

Definition 1.2.1 .

A linear multistep method is CONVERGENT of ORDER $p \in \mathbb{N}$ if for $i = o(1)N$, $y_i - y(t_i) = O(h^p)$ as $h \to 0$.

Definition 1.2.2 .

A linear multistep method is ZERO STABLE if no root of the polynomial $\rho(\zeta) = \sum_{j=0}^{k} \alpha_j \zeta^j$ has modulus greater than one , and every root with modulus one is simple .

Definition 1.2.3 .

The LOCAL TRUNCATION ERROR at the $i^{th.}$ stage , $i \geq k$, is

$$\theta_i^* = \sum_{j=0}^{k} \alpha_{k-j} y(t_{i-j}) - h \sum_{j=0}^{k} \beta_{k-j} y'(t_{i-j}) .$$

Definition 1.2.4 .

A linear multistep method is CONSISTENT of ORDER $p \in \mathbb{N}$ if $\theta_i^* = O(h^{p+1})$ as $h \to 0$ for $i \geq k$.

If $\theta_i^* = C^{(1)} h^{p+1} y^{(p+1)}(t_i) + O(h^{p+2})$ for all $y(t) \in C^{(p+1)}[0,T]$, where $C^{(1)}$ is some constant independent of h , then $C^{(1)}$ is the ERROR CONSTANT .

Theorem 1.2.5 .

A linear multistep method is convergent of order p if and only if it is both zero stable and consistent of order p .

In the above linear multistep method (1.2.1) we may stipulate that $\alpha_k = 1$ and thus have 2k+1 free parameters $\alpha_0, \alpha_1, \ldots, \alpha_{k-1}, \beta_0, \beta_1, \cdots, \beta_k$ in constructing a method . Expanding θ_i^* as

a Taylor series about t_i and assuming $y(t)$ to have sufficient continuity we observe that it will be possible to choose these constants so as to give methods with order of consistency as high as $2k$. Yet the construction of linear multistep methods with order of convergence as high as this is precluded by considerations of zero stability , as the following theorem shows .

Theorem 1.2.6 . (6)

No zero stable linear multistep method of stepnumber k can have order of consistency exceeding $2[\frac{k}{2}] + 2$, where for any real x , $[x]$ denotes the integral part of x .

1.3 Cyclic Linear Multistep Methods .

Donelson and Hansen (9) proposed cyclic (linear multistep) methods as a means of surmounting the barrier presented by theorem 1.2.6 . Such a method consists of cycles of a finite number , say m , of linear multistep methods , and may be written in the form

$$\sum_{j=0}^{k} \alpha_{k-j}^{(\nu)} y_{i-j} = h \sum_{j=0}^{k} \beta_{k-j}^{(\nu)} f(t_{i-j}, y_{i-j}) \quad , \quad i = k(1)N \ ,$$

$$\nu \equiv i-k+1 \pmod{m} \ , \ 1 \leq \nu \leq m \ .$$

Suppose we are given k starting values $\tilde{y}_i$, approximating $y(t_i)$, for $i = 0(1)k-1$. Then we may represent the cyclic method in the form

$$\begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ \alpha_0^{(1)} & \cdots & \alpha_k^{(1)} & & \\ \ddots & & & \ddots & \\ \alpha_0^{(m)} & \cdots & \alpha_k^{(m)} & & \\ & \ddots & & & \end{pmatrix} \begin{pmatrix} y_0 \\ \vdots \\ y_{k-1} \\ y_k \\ \vdots \\ y_{k+m-1} \\ \vdots \end{pmatrix} =$$

$$h \begin{pmatrix} 0 & & & & \\ & \ddots & & & \\ & & 0 & & \\ \beta_0^{(1)} & \cdots & \beta_k^{(1)} & & \\ \ddots & & & \ddots & \\ \beta_0^{(m)} & \cdots & \beta_k^{(m)} & & \\ & \ddots & & & \end{pmatrix} \begin{pmatrix} f(t_0,y_0) \\ \vdots \\ f(t_{k-1},y_{k-1}) \\ f(t_k,y_k) \\ \vdots \\ f(t_{k+m-1},y_{k+m-1}) \\ \vdots \end{pmatrix}$$

$$+ (\tilde{y}_0, \ldots, \tilde{y}_{k-1}, 0, 0, \ldots, 0)^T \qquad (1.3.1) \ .$$

Let us define the vectors y , f , $g \in \mathbb{R}^{N+1}$ as follows , taking all vector and matrix suffices to begin at zero rather than one . The $i^{th.}$ component of y is y_i , $f_i = f(t_i,y_i)$ and $g_i =$

$$\begin{cases} \tilde{y}_i & \text{if } 0 \leq i \leq k-1 \ , \\ 0^i & \text{if } k \leq i \leq N \ . \end{cases}$$

We may write $A_h y = hB_h f + g$, where A_h and B_h are lower-triangular $(N+1)x(N+1)$ matrices with bandwidth $(k+1)$ independent of h. The arrays

$$\begin{matrix} \alpha_0^{(1)} & \cdots & \alpha_k^{(1)} \\ \ddots & & \\ \alpha_0^{(m)} & & \alpha_k^{(m)} \end{matrix}$$

and

$$\begin{matrix} \beta_0^{(1)} & \cdots & \beta_k^{(1)} \\ \ddots & & \\ \beta_0^{(m)} & & \beta_k^{(m)} \end{matrix}$$

(1.3.2) recur in A_h and B_h respectively .

Let us partition these arrays into m x m matrices $A_0, A_1, \ldots$

$\ldots, A_\tau, B_0, B_1, \ldots, B_\tau$, so that

$$(A_0, A_1, \ldots, A_\tau) = \begin{pmatrix} \alpha_0^{(1)} & \cdots & \alpha_k^{(1)} & & \\ & \ddots & & \ddots & \\ & & \alpha_0^{(m)} & \cdots & \alpha_k^{(m)} \end{pmatrix} \quad \text{and}$$

$$(B_0, B_1, \ldots, B_\tau) = \begin{pmatrix} \beta_0^{(1)} & \cdots & \beta_k^{(1)} & & \\ & \ddots & & \ddots & \\ & & \beta_0^{(m)} & \cdots & \beta_k^{(m)} \end{pmatrix} \quad \text{,where}$$

$\tau = 1 + [\frac{k-1}{m}]$ (24, p.218) .

So $A_h =$

$$\begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ A_0 & \cdots & A_\tau & & \\ & A_0 & \cdots & A_\tau & \\ & & \ddots & & \ddots \\ & & & A_0 & \cdots & A_\tau \end{pmatrix} \quad \text{and}$$

$B_h =$

$$\begin{pmatrix} 0 & & & & \\ & \ddots & & & \\ & & 0 & & \\ B_0 & \cdots & B_\tau & & \\ & B_0 & \cdots & B_\tau & \\ & & \ddots & & \ddots \\ & & & B_0 & \cdots & B_\tau \end{pmatrix} \quad .$$

Example 1.3.1 .

Consider the application of the following two linear multistep methods alternately .

$$y_i - y_{i-2} = \frac{1}{3}h(f_i + 4f_{i-1} + f_{i-2}) ,$$

$$y_{i+1} - y_i = \frac{1}{12}h(5f_{i+1} + 8f_i - f_{i-1}) .$$

$k = m = 2$, so $\tau = 1$.

$$A_0 = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}, \quad B_0 = \begin{pmatrix} 1/3 & 4/3 \\ 0 & -1/12 \end{pmatrix},$$

$$B_1 = \begin{pmatrix} 1/3 & 0 \\ 8/12 & 5/12 \end{pmatrix} .$$

Definition 1.3.2 .

A cyclic method is zero stable if $0 \leq \max_{i,j} \leq N \left| (A_h^{-1})_{ij} \right| = O(1)$ as $h \to 0$.

We note that any linear multistep method may be written as a cyclic method with $m = 1$, giving $\tau = k$, $A_\theta = \alpha_\theta$, $B_\theta = \beta_\theta$ for $\theta = O(1)\tau$. Definition 1.3.2 is a generalisation of definition 1.2.2 to take in the class of cyclic methods of which linear multistep methods form a subclass (18) . For cyclic methods the order of convergence is defined as in definition 1.2.1 . As for consistency , a cyclic method is consistent of order p if each of its constituent linear multistep methods is consistent of order at least p . We then have the following convergence theorem .

Theorem 1.3.3 .

Suppose we are given k starting values $\tilde{y}_0, \ldots, \tilde{y}_{k-1}$ such that $\tilde{y}_i - y(t_i) = O(h^p)$ as $h \to 0$ for $i = O(1)k-1$. Then any cyclic method using these starting values is convergent of order p if it is both zero stable and consistent of order p .

For the proof of this see theorem 2 of (9) .

Definition 1.3.4 .

For A_h of the form (1.3.1) we define the COMPANION MATRICES $G_1, G_2, \ldots, G_m$ and the matrix G by

$$G_\nu = \begin{bmatrix} 0 & 1 & & \\ & & \ddots & \\ & & & 1 \\ -\alpha_0^{(\nu)}/\alpha_k^{(\nu)} & \cdots & & -\alpha_{k-1}^{(\nu)}/\alpha_k^{(\nu)} \end{bmatrix} \quad , \text{ a } k \times k \text{ matrix} ,$$

$\nu = 1(1)m$, and

$$G = G_m G_{m-1} \cdots G_1 . \qquad (1.3.3)$$

Theorem 1.3.5 . (18)

$$0 \leq \max_{i,j} \leq N \left| \left(A_h^{-1}\right)_{ij} \right| = O(1) \text{ as } h \to 0$$ if no eigenvalue of G has modulus greater than one , and every eigenvalue with modulus one has linear elementary divisors .

It is possible to choose the m linear multistep methods to have order of consistency greater than the bounds imposed by theorem 1.2.6 , so that each individual linear multistep method is zero unstable and yet for there to be zero stability in the cyclic method , as the following example shows .

Example 1.3.6 . (9 , p. 149)

$$\begin{pmatrix} 1 & & & & & \\ 0 & 1 & & & & \\ 0 & 0 & 1 & & & \\ 0 & -57 & 24 & 33 & & \\ 0 & 1083 & -456 & -1347 & 720 & \\ 0 & 0 & 0 & -57 & 24 & 33 \\ & & & & \ddots & \ddots \end{pmatrix} \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ \vdots \end{pmatrix} =$$

$$h \begin{pmatrix} 0 & & & & & \\ 0 & 0 & & & & \\ 0 & 0 & 0 & & & \\ -1 & 24 & 57 & 10 & & \\ 0 & -350 & -1347 & 456 & 251 & \\ 0 & 0 & -1 & 24 & 57 & 10 \\ & & & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ \vdots \\ \vdots \end{pmatrix}$$

$+ (\tilde{y}_0, \tilde{y}_1, \tilde{y}_2, 0, 0, \ldots, 0)$.

By direct computation we obtain

$$G = \begin{pmatrix} 0 & 19/11 & -8/11 \\ 0 & 19/11 & -8/11 \\ 0 & 19/11 & -8/11 \end{pmatrix} .$$

$\det(G - \zeta I) = \zeta^2(\zeta - 1)$, so we have zero stability , yet for each constituent linear multistep method $k = 3$ and consistency is of order five .

CHAPTER TWO .

BLOCK SEMI EXPLICIT METHODS .

2.1 Definitions .

The matrix notation (1.3.1) can be extended to give a notation which will represent many different kinds of methods .

Definition 2.1.1 .

For any $\gamma \in \mathbb{R}^m$ and $r,n \in \mathbb{N}$, where r is independent of h , we define the RESTRICTION OPERATOR $\Delta_h : C[0,T] \rightarrow \mathbb{R}^{r+nm}$ as follows. For any $\xi \in C[0,T]$,

$$[\Delta_h \xi(t)]_i = \begin{cases} \xi(ih) \quad , \; i = 0(1)r-1 \; , \\ \\ \xi((r-1+s\gamma_m+\gamma_\nu)h) \; , \; \nu = 1(1)m \; , \; s = 0(1)n-1 \; , \\ \qquad\qquad\qquad\qquad\qquad\qquad i = r + sm + \nu - 1 \; . \end{cases}$$

For any r , I_r and 0_r will represent the r x r identity and zero matrices respectively .

Definition 2.1.2 .

A method is m-BLOCK SEMI EXPLICIT if $\gamma \in \mathbb{R}^m$ is given and we may represent the method in the form

$$A_h y^* = hB_h f^* + g \; ,$$

where A_h and B_h are lower triangular $(r+nm)x(r+nm)$ matrices with finite bandwidth independent of h and have the form

$$A_h = \left(\begin{array}{ccc|cccccc} & I_r & & & & & & & \\ \hline A_o^{(0)} & \cdots\cdots & A_{\tau-1}^{(0)} & A_\tau & & & & & \\ & A_o^{(1)} \cdots & A_{\tau-2}^{(1)} & A_{\tau-1} & A_\tau & & & & \\ & \ddots & \vdots & \vdots & & \ddots & & & \\ & & A_o^{(\tau-1)} & A_1 & \cdots & & A_\tau & & \\ & & & A_o & \cdots & & & A_\tau & \\ & & & & \ddots & & & & \ddots \\ & & & & & A_o & \cdots & & A_\tau \end{array}\right)$$

(2.1.1) and

$$B_h = \left(\begin{array}{ccc|cccccc} & 0_r & & & & & & & \\ \hline B_o^{(0)} & \cdots\cdots & B_{\tau-1}^{(0)} & B_\tau & & & & & \\ & \ddots & \vdots & & & & & & \\ & & B_o^{(\tau-1)} & B_1 & \cdots & & B_\tau & & \\ & & & B_o & \cdots & & & B_\tau & \\ & & & & \ddots & & & & \ddots \\ & & & & & B_o & \cdots & & B_\tau \end{array}\right)$$

(2.1.2) ,

where τ is a positive integer independent of h , $A_o, A_1, \ldots, A_\tau, B_o, B_1, \ldots, B_\tau$ are m x m matrices , $A_o^{(\theta)}, \ldots, A_{\tau-1-\theta}^{(\theta)}, B_o^{(\theta)}, \ldots, B_{\tau-1-\theta}^{(\theta)}$, $\theta = O(1)\tau-1$, are matrices with m rows(*), and y^* , f^* , $g \in \mathbb{R}^{r+nm}$. y_i^* is the computed approximation to $y([\Delta_h t]_i)$, $f_i^* = f([\Delta_h t]_i, y_i^*)$ and

$$g_i = \begin{cases} \tilde{y}_i , & i=O(1)r-1 , \\ 0 , & i \geq r . \end{cases}$$

Moreover $[\Delta_h t]_o = 0$ and $[\Delta_h t]_{r+nm-1} = (r-1+n\gamma_m)h = T$ (2.1.3) , so $N = r - 1 + n\gamma_m$.

Any block semi explicit method then may be fully represented

(*) The matrices $A_o^{(\theta)}, \ldots, A_{\tau-1-\theta}^{(\theta)}, B_o^{(\theta)}, \ldots, B_{\tau-1-\theta}^{(\theta)}$ are just $A_o, \ldots, A_{\tau-1-\theta}$, $B_o, \ldots B_{\tau-1-\theta}$ with some zero columns deleted , for $\theta = O(1)\tau-1$.

by the m x (1 + 2(τ + 1)m) array

$$(A_0, A_1, \ldots, A_\tau | B_0, B_1, \ldots, B_\tau | \gamma) \quad (2.1.4) .$$

For any block semi explicit method A_τ and B_τ are both lower triangular , and if B_τ is strictly lower triangular we shall call the method m-BLOCK EXPLICIT .

The class of m-block semi explicit methods is very large , and we now give a wide range of methods covered by this definition .

2.2 Cyclic Linear Multistep Methods .

The linear multistep method (1.2.1) is a 1-cyclic method and the array (2.1.4) takes the form $(\alpha_0, \alpha_1, \ldots, \alpha_k | \beta_0, \beta_1, \ldots, \beta_k | 1)$, and the method is called explicit if $\beta_k = 0$. For each i , $y_i^* = y_i$ and $N = r + n - 1$.

For example 1.3.1 we have the array

$$\left(\begin{array}{cccc|cccc|c} -1 & 0 & 1 & 0 & 1/3 & 4/3 & 1/3 & 0 & 1 \\ 0 & 0 & -1 & 1 & 0 & -1/12 & 8/12 & 5/12 & 2 \end{array}\right) ,$$

a 2-block semi explicit method . $y_i^* = y_i$ for each i and $N = r + 2n - 1$.

Example 2.2.1 .

Consider a cyclic method formed by alternating the explicit and implicit Euler linear multistep methods . This is a 2-block semi explicit method , and the array is

$$\left(\begin{array}{cccc|cccc|c} 0 & -1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 1 & 2 \end{array}\right) .$$

Example 2.2.2 .

Consider a cyclic method represented as a 3-block semi explicit method , for which $y_i^* = y_i$ for each i and $N = r + 3n - 1$.

$$\left(\begin{array}{cccccc|c|ccc} 0 & 3/2 & -5/2 & 1 & 0 & 0 & 0 & -17/24 & -1/3 & \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & -1/12 & \\ 0 & 0 & 0 & -6/5 & 1/5 & 1 & 0 & 0 & 0 & \end{array}\right.$$

$$\left.\begin{array}{ccc|c} 13/24 & 0 & 0 & 1 \\ 8/12 & 5/12 & 0 & 2 \\ 5/12 & 22/15 & 19/60 & 3 \end{array}\right) .$$

Example 2.2.3 .

Another such cyclic method is represented by

$$\left(\begin{array}{ccccccccc|} 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -3/2 & 1/2 & 1 \end{array}\right.$$

$$\left.\begin{array}{ccccccccc|c} 0 & 0 & 3 & -22/3 & 32/3 & -10/3 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & -17/12 & 15/4 & -13/4 & -1/12 & 1 & 0 & 2 \\ 0 & 0 & 0 & 0 & -21/16 & 155/48 & -61/48 & 41/48 & 1 & 3 \end{array}\right) .$$

Example 2.2.4 .

A cyclic method representable as a 2-block semi explicit method :

$$\left(\begin{array}{cccccc|cccccc|c} 0 & 0 & 0 & -1 & 1 & 0 & 0 & -1/6 & 1 & -5/6 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 13/6 & -1 & 5/6 & 2 \end{array}\right) .$$

Example 2.2.5 .

Another such cyclic method is

$$\left(\begin{array}{cccccc|cccccc|c} 0 & 0 & 0 & -1 & 1 & 0 & 1 & 1/6 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 1 & -1 & -1/3 & -1 & 1/3 & -1 & 5/6 & 2 \end{array}\right) .$$

We see then that for a cyclic method composed of m linear multistep methods $\gamma = (1,2, \ldots ,m)^T$. Each of the examples given will be referred to later .

2.3 Predictor-Corrector Methods .

Suppose we have an explicit linear multistep predictor $\sum_{j=0}^{k}\alpha^{*}_{k-j}y_{i-j} = h\sum_{j=0}^{k-1}\beta^{*}_{k-j}f_{i-j}$ followed by μ applications of the implicit corrector $\sum_{j=0}^{k}\alpha_{k-j}y_{i-j} = h\sum_{j=0}^{k}\beta_{k-j}f_{i-j}$ (2.3.1) .

Then in the array (2.1.4) we have $\tau = k$, $m = \mu + 1$,

$$A_\theta = \begin{pmatrix} 0 & \cdots & 0 & \alpha^{*}_\theta \\ 0 & \cdots & 0 & \alpha_\theta \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \alpha_\theta \end{pmatrix} , \quad \theta = 0(1)k-1 ,$$

$$A_k = \begin{pmatrix} \alpha^{*}_k & & & \\ & \alpha_k & & \\ & & \ddots & \\ & & & \alpha_k \end{pmatrix} , \quad \gamma = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} ,$$

$$B_k = \begin{pmatrix} 0 & & & \\ \beta_k & 0 & & \\ & \beta_k & 0 & \\ & & \ddots & \ddots \\ & & & \beta_k \quad 0 \end{pmatrix} ;$$

in $P(EC)^{\mu}E$ mode (see (15) , p.87)

$$B_\theta = \begin{pmatrix} 0 & \cdots & 0 & \beta^{*}_\theta \\ 0 & \cdots & 0 & \beta_\theta \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \beta_\theta \end{pmatrix} , \quad \theta = 0(1)k-1 ,$$

and in $P(EC)^{\mu}$ mode

$$B_\theta = \begin{pmatrix} 0 & \cdots & 0 & \beta^{*}_\theta & 0 \\ 0 & \cdots & 0 & \beta_\theta & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \beta_\theta & 0 \end{pmatrix} .$$

This then is a $(\mu+1)$-block explicit method , and

$$\Delta_h t = (0,h, \ldots ,(r-1)h,rh,rh, \ldots ,rh,(r+1)h, \ldots ,(r+1)h, \ldots$$
$$\underbrace{\qquad\qquad}_{(\mu+1)} \quad \underbrace{\qquad\qquad}_{(\mu+1)}$$
$$\ldots ,(r+n-1)h, \ldots ,(r+n-1)h)^T$$
$$\underbrace{\qquad\qquad}_{(\mu+1)}$$

and $N = r + n - 1$.

Example 2.3.1 .

Consider also a predictor-corrector method with a Milne modifier in PMECME mode (15 ,p.94) , in which we have

$$P : y_i^{[0]} - \hat{y}_{i-4}^{[1]} = \frac{4h}{3}[2f(t_{i-1},\hat{y}_{i-1}^{[1]}) - f(t_{i-2},\hat{y}_{i-2}^{[1]}) + 2f(t_{i-3},\hat{y}_{i-3}^{[1]})] ,$$

$$M : \hat{y}_i^{[0]} - y_i^{[0]} - \frac{112}{121}(y_{i-1}^{[1]} - y_{i-1}^{[0]}) = 0 ,$$

$$C : y_i^{[1]} - \frac{9}{8}\hat{y}_{i-1}^{[1]} + \frac{1}{8}\hat{y}_{i-3}^{[1]} = \frac{3h}{8}[f(t_i,\hat{y}_i^{[0]}) + 2f(t_{i-1},\hat{y}_{i-1}^{[1]}) - f(t_{i-2},\hat{y}_{i-2}^{[1]})] .$$

$$M : \hat{y}_i^{[1]} - y_i^{[1]} + \frac{9}{121}(y_i^{[1]} - y_i^{[0]}) = 0 .$$

This is a 4-block explicit method in which

$$A_o = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} , \quad A_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/8 \\ 0 & 0 & 0 & 0 \end{pmatrix} , \quad A_2 = 0_4 ,$$

$$A_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 112/121 & 0 & -112/121 & 0 \\ 0 & 0 & 0 & -9/8 \\ 0 & 0 & 0 & 0 \end{pmatrix} , \quad A_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -9/121 & 0 & -112/121 & 1 \end{pmatrix} ,$$

$$B_o = 0_4 , \quad B_1 = \begin{pmatrix} 0 & 0 & 0 & 8/3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} , \quad B_2 = \begin{pmatrix} 0 & 0 & 0 & -4/3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -3/8 \\ 0 & 0 & 0 & 0 \end{pmatrix} ,$$

$$B_3 = \begin{pmatrix} 0 & 0 & 0 & 8/3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3/4 \\ 0 & 0 & 0 & 0 \end{pmatrix} , \quad B_4 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 3/8 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} , \quad \gamma = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} .$$

2.4 Explicit Runge-Kutta Methods .

An m stage explicit Runge-Kutta method is usually written

$$\left.\begin{aligned} & y_i - y_{i-1} = h \sum_{j=1}^{m} c_j k_j , \\ & k_1 = f(t_{i-1}, y_{i-1}) , \\ & k_j = f\Big(t_{i-1} + h a_j , \; y_{i-1} + h \sum_{q=1}^{j-1} b_{jq} k_q\Big) , \quad j = 2(1)m , \\ & \text{where } a_j = \sum_{q=1}^{j-1} b_{jq} , \quad j = 2(1)m . \end{aligned}\right\} \qquad (2.4.1)$$

An alternative representation is

$$\begin{aligned} y_{i-1+a_2} &= y_{i-1} + h b_{21} f_{i-1} , \\ y_{i-1+a_3} &= y_{i-1} + h(b_{31} f_{i-1} + b_{32} f_{i-1+a_2}) , \\ &\cdots \\ y_{i-1+a_m} &= y_{i-1} + h(b_{m1} f_{i-1} + b_{m2} f_{i-1+a_2} + \cdots + b_{m,m-1} f_{i-1+a_{m-1}}) , \\ y_i &= y_{i-1} + h(c_1 f_{i-1} + c_2 f_{i-1+a_2} + \cdots + c_m f_{i-1+a_m}) , \end{aligned}$$

which can be viewed as an m-block explicit method with the array

$$\left(\begin{array}{c|c|c|c} \begin{matrix} 0 & \cdots & 0 & -1 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & -1 \end{matrix} & I_m & \begin{matrix} 0 & \cdots & 0 & b_{21} \\ 0 & \cdots & 0 & b_{31} \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & b_{m1} \\ 0 & \cdots & 0 & c_1 \end{matrix} & \begin{matrix} 0 & & & & \\ b_{32} & 0 & & & \\ \vdots & \ddots & \ddots & & \\ b_{m2} & \cdots & b_{m,m-1} & 0 & \\ c_2 & \cdots & c_{m-1} & c_m & 0 \end{matrix} & \begin{matrix} a_2 \\ a_3 \\ \vdots \\ a_m \\ 1 \end{matrix} \end{array}\right) . \qquad (2.4.2) .$$

$r = 1$, $\tau = 1$, $\gamma_\nu = \begin{cases} a_{\nu+1} , & = 1(1)m-1 , \\ 1 , & \nu = m , \end{cases}$ and $y^*_{sm+\nu}$ is an approximation to $y([\Delta_h t]_{sm+\nu}) = y((s+\gamma_\nu)h)$, $\nu = 1(1)m$, $s = 0(1)n-1$.

Example 2.4.1 .

Consider a semi explicit Runge-Kutta method , for which the formulation (2.4.1) has $k_j = f(t_{i-1} + ha_j, y_{i-1} + h\sum_{q=1}^{j} b_{jq}k_q)$, so that B_1 is lower triangular , but not strictly so . Lambert (15, p.154) quotes the scheme

$$y_i - y_{i-1} = \frac{1}{6}h(k_1 + 4k_2 + k_3) ,$$

$$k_1 = f_{i-1} ,$$

$$k_2 = f(t_{i-1} + h/2 , y_{i-1} + hk_1/4 + hk_2/4) ,$$

$$k_3 = f(t_{i-1} + h , y_{i-1} + hk_2) .$$

This is a 3-block semi explicit method with array

$$\left(\begin{array}{cccccc|cccccc|c} 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 1/4 & 1/4 & 0 & 0 & 1/2 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 1/6 & 2/3 & 1/6 & 0 & 1 \end{array}\right) .$$

2.5 Hybrid Methods .

A k-step hybrid formula takes the form

$$\sum_{j=0}^{k} \alpha_{k-j} y_{i-j} = h\sum_{j=0}^{k} \beta_{k-j} f_{i-j} + h\beta_\mu f_{i-k+\mu} \quad (2.5.1) \text{ , where } \mu \notin \{1,2, \ldots ,k\} .$$

Definition 2.5.1 .

The hybrid formula (2.5.1) has ORDER p and ERROR CONSTANT C if for all $y(t) \in C^{(p+1)}[0,T]$,

$$\sum_{j=0}^{k} \alpha_{k-j} y(t_{i-j}) - h\sum_{j=0}^{k} \beta_{k-j} y'(t_{i-j}) - h\beta_\mu y'(t_{i-k+\mu}) = Ch^{p+1} y^{(p+1)}(t_i) + O(h^{p+2}).$$

Example 2.5.2 .

In the case when $\beta_k = 0$ we need a predictor formula to obtain an estimate of $y_{i-k+\mu}$, and the hybrid method is formulated as a $P_\mu EP_H E$ algorithm . For example , on p.180 of (15) a $P_{2/3}EP_H E$ method is given :

$$P_{2/3} : y_{i-1/3} + \frac{25}{27}y_{i-1} - \frac{52}{27}y_{i-2} = \frac{h}{27}(50f_{i-1} + 20f_{i-2}) ,$$

$$P_H : y_i - y_{i-1} = \frac{h}{4}(3f_{i-1/3} + f_{i-1}) .$$

This is a 2-block explicit method with array

$$\left(\begin{array}{cccccc|cccccc|c} 0 & -52/27 & 0 & 25/27 & 1 & 0 & 0 & 20/27 & 0 & 50/27 & 0 & 0 & 2/3 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 1/4 & 3/4 & 0 & 1 \end{array}\right) .$$

r=2 and $\Delta_h t = (0,h,5h/3,2h,8h/3,3h, \ldots)^T$.

Example 2.5.3 .

When $\beta_k \neq 0$ we need in addition to a P_μ predictor a P_k predictor , and the hybrid method is formulated as a $P_\mu EP_k EC_H E$ algorithm . For example on p.175 of (15) a $P_{\frac{1}{2}}EP_k EC_H E$ scheme is given :

$$P_{\frac{1}{2}} : y_{i-\frac{1}{2}} - y^c_{i-2} = \frac{3h}{8}(3f^c_{i-1} + f^c_{i-2}) ,$$

$$P_k : y^p_i + 4y^c_{i-1} - 5y^c_{i-2} = h(4f^c_{i-1} + 2f^c_{i-2}) ,$$

$$C_H : y^c_i - y^c_{i-1} = \frac{h}{6}(f^p_i + f^c_{i-1} + 4f_{i-\frac{1}{2}}) ,$$

where for each j , $f^p_j = f(t_j,y^p_j)$ and $f^c_j = f(t_j,y^c_j)$. This is a 3-block explicit method with array

$$\left(\begin{array}{ccccccccc|ccccccccc|c} 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 3/8 & 0 & 0 & 9/8 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & -5 & 0 & 0 & 4 & 0 & 1 & 0 & 0 & 0 & 2 & 0 & 0 & 4 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1/6 & 2/3 & 1/6 & 0 & 1 \end{array}\right) .$$

r = 2 and $\Delta_h t = (0,h,3h/2,2h,2h,5h/2,3h,3h, \ldots)^T$.

Example 2.5.4 .

We may also use a $P_\mu EP_k EC_H$ mode . For the same $P_{\frac{1}{2}}$, P_k and C_H as in example 2.5.3 we have the same array , except that

$$B_0 = \begin{pmatrix} 0 & 3/8 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and } B_1 = \begin{pmatrix} 0 & 9/8 & 0 \\ 0 & 4 & 0 \\ 0 & 1/6 & 0 \end{pmatrix} .$$

Example 2.5.5 .

$P_\mu EP_H EC_H(E)$ methods may also be considered . For example Butcher (3 , p.130) quotes

$$P_{\frac{1}{2}} : y_{i-\frac{1}{2}} - y^c_{i-1} = \frac{h}{2} f^c_{i-1} ,$$

$$P_H : y^p_i - y^c_{i-1} = h(2f_{i-\frac{1}{2}} - f^c_{i-1}) ,$$

$$C_H : y^c_i - y^c_{i-1} = \frac{h}{6}(4f_{i-\frac{1}{2}} + f^p_i + f^c_{i-1}) ,$$

which is a 3-block explicit method with array

$$\left(\begin{array}{cccccc|cccccc|c} 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & -1 & 2 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 1/6 & 2/3 & 1/6 & 0 & 1 \end{array}\right) \quad \text{and } r = 1 .$$

We note that this is just a 3-stage explicit Runge-Kutta method .

Example 2.5.6 .

Kohfeld and Thompson (13) proposed another mode , $P_\lambda EP_\mu EP_H EC_H E$, in order to take advantage of the fact that hybrid methods have very small error constants . For instance Lambert (15 , pp.176-177) quotes the following $P_{0.7} EP_{0.5} EP_H EC_H E$ method , β being some constant .

$$P_{0.7} : y_{i-0.3} + 5.7937275y_{i-1} - 3.5721y_{i-2} - 3.2216275y_{i-3} = h(3.6869175f_{i-1} + 6.07257f_{i-2} + 0.9558675f_{i-3}) ,$$

$$P_{0.5} : y_{i-0.5} + (\frac{45}{64} - \frac{357}{25}\beta)y_{i-1} + (-\frac{25}{16} + \frac{3213}{250}\beta)y_{i-2} + (-\frac{9}{64} + \frac{357}{250}\beta)y_{i-3}$$

$$= h[(\frac{45}{32} - \frac{4063}{500}\beta)f_{i-1} + (\frac{15}{16} - \frac{4291}{500}\beta)f_{i-2} + \beta f_{i-0.3}] ,$$

$$P_H : y_i^p - y_{i-1} = h(\frac{221}{714}f_{i-1} - \frac{7}{714}f_{i-2} + \frac{500}{714}f_{i-0.3}) ,$$

$$C_H : y_i - y_{i-1} = \frac{h}{6}(f_i^p + f_{i-1} + 4f_{i-0.5}) .$$

This is a 4-block explicit method with

$$A_o = \begin{pmatrix} 0 & 0 & 0 & -3.2216275 \\ 0 & 0 & 0 & -\frac{9}{64} + \frac{357}{250}\beta \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} , \quad A_1 = \begin{pmatrix} 0 & 0 & 0 & -3.5721 \\ 0 & 0 & 0 & -\frac{25}{16} + \frac{3213}{250}\beta \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} ,$$

$$A_2 = \begin{pmatrix} 0 & 0 & 0 & 5.7937275 \\ 0 & 0 & 0 & \frac{45}{64} - \frac{357}{25}\beta \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 \end{pmatrix} , \quad A_3 = I_4 ,$$

$$B_o = \begin{pmatrix} 0 & 0 & 0 & 0.9558675 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} , \quad B_1 = \begin{pmatrix} 0 & 0 & 0 & 6.07257 \\ 0 & 0 & 0 & \frac{15}{16} - \frac{4291}{500}\beta \\ 0 & 0 & 0 & -\frac{7}{714} \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$B_2 = \begin{pmatrix} 0 & 0 & 0 & 3.6869175 \\ 0 & 0 & 0 & \frac{45}{32} - \frac{4063}{500}\beta \\ 0 & 0 & 0 & \frac{221}{714} \\ 0 & 0 & 0 & \frac{1}{6} \end{pmatrix} , \quad B_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \beta & 0 & 0 & 0 \\ \frac{500}{714} & 0 & 0 & 0 \\ 0 & \frac{2}{3} & \frac{1}{6} & 0 \end{pmatrix} , \quad \gamma = \begin{pmatrix} 0.7 \\ 0.5 \\ 1 \\ 1 \end{pmatrix} .$$

2.6 Extrapolation .

Example 2.6.1 .

In this example we apply the trapezoidal rule with steplength $\frac{1}{2}h$ twice , followed by the same formula with steplength h once , and then

extrapolate the two values obtained .

$$y_{i-\frac{1}{2}} - y_{i-1} = \frac{h}{4}(f_{i-\frac{1}{2}} + f_{i-1}) ,$$

$$y_i^{[0]} - y_{i-\frac{1}{2}} = \frac{h}{4}(f_i^{[0]} + f_{i-\frac{1}{2}}) ,$$

$$y_i^{[1]} - y_{i-1} = \frac{h}{2}(f_i^{[1]} + f_{i-1}) ,$$

$$y_i + \frac{1}{3}y_i^{[1]} - \frac{4}{3}y_i^{[0]} = 0 .$$

For each j and l , $f_j^{[l]} = f(t_j, y_j^{[l]})$. This is a 4-block semi explicit method with array

$$\left(\begin{array}{cccccccc|cccccccc|c}
0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 1/4 & 0 & 0 & 0 & \frac{1}{2} \\
0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 1/4 & 0 & 0 & 1 \\
0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 1/2 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & -4/3 & 1/3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}\right) .$$

Example 2.6.2 .

Consider a particular case of Gragg's method (15 , pp.191,192) .

We have $y_{i-1/2}^{[0]} - y_{i-1} = \frac{h}{2}f_{i-1}$,

$$y_i^{[0]} - y_{i-1} = hf_{i-1/2}^{[0]} ,$$

$$y_{i+1/2}^{[0]} - y_{i-1/2}^{[0]} = hf_i^{[0]} ,$$

$$\bar{y}_i - \frac{1}{4}y_{i+1/2}^{[0]} - \frac{1}{2}y_i^{[0]} - \frac{1}{4}y_{i-1/2}^{[0]} = 0 ,$$

$$y_{i-3/4}^{[1]} - y_{i-1} = \frac{h}{4}f_{i-1} ,$$

$$y_{i-1/2}^{[1]} - y_{i-1} = \frac{h}{2}f_{i-3/4}^{[1]} ,$$

$$y_{i-1/4}^{[1]} - y_{i-3/4}^{[1]} = \frac{h}{2}f_{i-1/2}^{[1]} ,$$

$$y_i^{[1]} - y_{i-1/2}^{[1]} = \frac{h}{2}f_{i-1/4}^{[1]} ,$$

$$y_{i+1/4}^{[1]} - y_{i-3/4}^{[1]} = \frac{h}{2}f_i^{[1]} ,$$

$$\bar{y}_i - \frac{1}{4}y^{[1]}_{i+1/4} - \frac{1}{2}y^{[1]}_i - \frac{1}{4}y^{[1]}_{i-1/4} = 0 \quad ,$$

$$y_i - \frac{4}{3}\bar{y}_i + \frac{1}{3}y_i = 0 \quad .$$

This is an 11-block explicit method with

$$A_o = \begin{pmatrix} 0 & \cdots & 0 & -1 \\ \vdots & & \vdots & -1 \\ \vdots & & \vdots & 0 \\ \vdots & & \vdots & 0 \\ \vdots & & \vdots & -1 \\ \vdots & & \vdots & -1 \\ \vdots & & \vdots & 0 \\ \vdots & & \vdots & 0 \\ \vdots & & \vdots & 0 \\ \vdots & & \vdots & 0 \\ 0 & \cdots & 0 & 0 \end{pmatrix}, \quad B_o = \begin{pmatrix} 0 & \cdots & 0 & 1/2 \\ \vdots & & \vdots & 0 \\ \vdots & & \vdots & 0 \\ \vdots & & \vdots & 0 \\ \vdots & & \vdots & 1/4 \\ \vdots & & \vdots & 0 \\ \vdots & & \vdots & 0 \\ \vdots & & \vdots & 0 \\ \vdots & & \vdots & 0 \\ \vdots & & \vdots & 0 \\ 0 & \cdots & 0 & 0 \end{pmatrix} ,$$

$$A_1 = \begin{pmatrix} 1 & & & & & & & & & & \\ 0 & 1 & & & & & & & & & \\ -1 & 0 & 1 & & & & & & & & \\ -1/4 & -1/2 & -1/4 & 1 & & & & & & & \\ & 0 & 0 & 0 & 1 & & & & & & \\ & & 0 & 0 & 0 & 1 & & & & & \\ & & & 0 & -1 & 0 & 1 & & & & \\ & & & & 0 & -1 & 0 & 1 & & & \\ & & & & & 0 & -1 & 0 & 1 & & \\ & & & & & & -1/4 & -1/2 & -1/4 & 1 & \\ 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & -4/3 & 1 \end{pmatrix} ,$$

$$B_1 = \begin{pmatrix}
0 & & & & & & & & & & \\
1 & 0 & & & & & & & & & \\
 & 1 & 0 & & & & & & & & \\
 & & 0 & 0 & & & & & & & \\
 & & & 0 & 0 & & & & & & \\
 & & & & \frac{1}{2} & 0 & & & & & \\
 & & & & & \frac{1}{2} & 0 & & & & \\
 & & & & & & \frac{1}{2} & 0 & & & \\
 & & & & & & & \frac{1}{2} & 0 & & \\
 & & & & & & & & 0 & 0 & \\
 & & & & & & & & & 0 & 0
\end{pmatrix},$$

$$\gamma = (1/2, 1, 3/2, 1, 1/4, 1/2, 3/4, 1, 5/4, 1, 1)^T .$$

CHAPTER THREE .

BLOCK IMPLICIT METHODS .

3.1 Definitions .

So far we have considered methods of the form $A_h y^* = hB_h f^* + g$ where A_h and B_h are both lower triangular . There are also classes of methods however which may be represented in this way , but with non-zero elements above the diagonal in these matrices .

Definition 3.1.1 .

A method is m-BLOCK IMPLICIT if $\gamma \in \mathbb{R}^m$ is given and it may be represented by the array $(A_0, A_1, \ldots, A_\tau \mid B_0, B_1, \ldots, B_\tau \mid \gamma)$ as (2.1.4) with at least one of A_τ and B_τ not lower triangular .

Again we shall have $A_h y^* = hB_h f^* + g$ and if A_τ is not lower triangular we may premultiply through by

$$\begin{pmatrix} I_r & & & & \\ & A_\tau^{-1} & & & \\ & & A_\tau^{-1} & & \\ & & & \ddots & \\ & & & & A_\tau^{-1} \end{pmatrix} \qquad (3.1.1)$$

to reduce A_h to lower triangular form , leaving B_h not lower triangular . Thus for any m-block implicit method we may consider only B_h to have non-zero elements above the diagonal , without loss of generality .

3.2 Advanced Linear Multistep Methods .

This class of methods has been studied by de Hoog and Williams (26) and more recently by McKee and Pitcher (17) .

Example 3.2.1 .

Consider Clippinger and Dimsdale's method (14) :

$$-\tfrac{1}{2}y_{i+1} + y_i - \tfrac{1}{2}y_{i-1} = \frac{h}{4}(f_{i-1} - f_{i+1}) ,$$

$$y_{i+1} - y_{i-1} = \frac{h}{3}(f_{i-1} + 4f_i + f_{i+1}) .$$

Having computed y_{i-1} we have implicit equations to solve for y_i and y_{i+1} together . This is a 2-block implicit method with array

$$\left(\begin{array}{cccc|cccc|c} 0 & -1/2 & 1 & -1/2 & 0 & 1/4 & 0 & -1/4 & 1 \\ 0 & -1 & 0 & 1 & 0 & 1/3 & 4/3 & 1/3 & 2 \end{array}\right) . \quad A_1^{-1} = \begin{pmatrix} 1 & 1/2 \\ 0 & 1 \end{pmatrix} ,$$

so premultiplying through by the matrix (3.1.1) we have also the array

$$\left(\begin{array}{cccc|cccc|c} 0 & -1 & 1 & 0 & 0 & 5/12 & 2/3 & -1/12 & 1 \\ 0 & -1 & 0 & 1 & 0 & 1/3 & 4/3 & 1/3 & 2 \end{array}\right) \qquad (3.2.1) .$$

3.3 Implicit Runge-Kutta Methods .

An implicit μ-stage Runge-Kutta method is defined by

$$\left.\begin{aligned} y_i - y_{i-1} &= h\sum_{j=1}^{\mu} c_j k_j , \\ k_j &= f\Big(t_{i-1} + ha_j , y_{i-1} + h\sum_{q=1}^{\mu} b_{jq}k_q\Big) , \quad j = 1(1)\mu , \end{aligned}\right\} \qquad (3.3.1)$$

where $a_j = \sum_{q=1}^{\mu} b_{jq}$, $j = 1(1)\mu$.

This is a $(\mu+1)$-block implicit method with array

$$\left(\begin{array}{ccc:c|c:ccc:c} 0 & \cdots & 0 & -1 & & & b_{11} & \cdots & b_{1\mu} & 0 & a_1 \\ \vdots & & \vdots & \vdots & & & \vdots & & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & -1 & I_{\mu+1} & 0_{\mu+1} & b_{\mu 1} & \cdots & b_{\mu\mu} & 0 & a_\mu \\ 0 & \cdots & 0 & -1 & & & c_1 & \cdots & c_\mu & 0 & 1 \end{array}\right) \qquad (3.3.2) .$$

Example 3.3.1 .

Lambert (15 , p.153) quotes the following 2-stage method :

$$y_i - y_{i-1} = \tfrac{1}{2}h(k_1 + k_2) ,$$

$$k_1 = f\Big(t_{i-1} + (\tfrac{1}{2} + \tfrac{\sqrt{3}}{6})h , y_{i-1} + \tfrac{h}{4}k_1 + (\tfrac{1}{4} + \tfrac{\sqrt{3}}{6})hk_2\Big) ,$$

$$k_2 = f(t_{i-1} + (\tfrac{1}{2} - \tfrac{\sqrt{3}}{6})h \ , \ y_{i-1} + (\tfrac{1}{4} - \tfrac{\sqrt{3}}{6})hk_1 + \tfrac{h}{4}k_2) \ ,$$

for which the array is

$$\left(\begin{array}{cccccc|ccccc|c} 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & (\frac{1}{4}+\frac{\sqrt{3}}{6}) & 0 & \frac{1}{2}+\frac{\sqrt{3}}{6} \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & (\frac{1}{4}-\frac{\sqrt{3}}{6}) & \frac{1}{4} & 0 & \frac{1}{2}-\frac{\sqrt{3}}{6} \\ 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 1 \end{array}\right) .$$

3.4 One Leg Methods .

If we consider the equation $y' = f(y)$, a one leg method is $\sum_{j=0}^{k} \alpha_{k-j} y_{i-j} = hf(\sum_{j=0}^{k} \alpha^*_{k-j} y_{i-j})$, which may be written as

$$\left.\begin{aligned} \hat{y}_{i-k} &= \sum_{j=0}^{k} \alpha^*_{k-j} y_{i-j} \ , \\ \sum_{j=0}^{k} \alpha_{k-j} y_{i-j} &= hf(\hat{y}_{i-k}) \end{aligned}\right\}$$, which is a 2-block implicit method with array

$$\left(\begin{array}{ccccccc|ccccc|c} -\alpha^*_0 & -\alpha^*_1 & \cdots & -\alpha^*_{k-1} & 1 & -\alpha^*_k & & 0 & \cdots & 0 & 0 & 0 & \gamma_1 \\ \alpha_0 & \alpha_1 & \cdots & \alpha_{k-1} & 0 & \alpha_k & & 0 & \cdots & 0 & 1 & 0 & 1 \end{array}\right) .$$

Example 3.4.1 .

Dahlquist (8) quotes the method $y_i - y_{i-1} = hf(\frac{1}{2}y_i + \frac{1}{2}y_{i-1})$. which has the array $\left(\begin{array}{cccc|cccc|c} 0 & -\frac{1}{2} & 1 & -\frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & -1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \end{array}\right)$.

3.5 Collocation Methods .

A collocation method consists of finding $u(t) \in \mathcal{P}_\mu$ such that $u(0) = y_0$ and $u'(a_\nu h) = f(a_\nu h, u(a_\nu h))$, $\nu = 1(1)\mu$, and then putting $y_1 = u(h)$ (3.5.1) . We then proceed in the same way to obtain y_2 , and so on . Nørsett (20) has shown that this is equivalent to the implicit Runge-Kutta method (3.3.2) with $b_{\nu q} = \int_0^{a_\nu} l_q(\eta)\,d\eta$ and

$c_\nu = \int_0^1 l_\nu(\eta)\,d\eta$, where $l_\nu(\eta) = \prod_{\substack{q=1 \\ q\neq\nu}}^{\mu} \frac{\eta - a_q}{a_\nu - a_q}$ are the Lagrange fundamental polynomials .

$a_1, a_2, \ldots, a_\mu$ are the nodes of the collocation and we have a $(\mu+1)$-block implicit method . The equivalence of the collocation and the Runge-Kutta method is in the sense that the Runge-Kutta method gives a discrete set of solutions which coincide exactly with the values taken by the continuous collocation polynomial at the points $0, a_1 h, a_2 h, \ldots, a_\mu h, h$.

Consider now a Galerkin method, which consists of finding $u(t) \in \mathcal{P}_\mu$ such that $u(0) = y_0$ and $\int_0^h [u'(\eta) - f(\eta, u(\eta))]\phi_\nu(\eta)\,d\eta = 0$ for $\nu = 0(1)\mu$ (3.5.2) , where the $\phi_\nu(\eta)$ are $\mu^{th.}$ degree polynomial basis functions and $\mu \geq 1$. We then put $y_1 = u(h)$ and then obtain y_2 by the same procedure , and so on . The integral in (3.5.2) is replaced by a quadrature formula , $\int_0^h v(\eta)\,d\eta \simeq h \sum_{\nu=1}^{\mu} w_\nu v(a_\nu h)$, where $0 \leq a_1 < a_2 < \ldots < a_\mu \leq 1$ and $w_\nu \neq 0$ for each ν .

Hulme (12) showed that this Galerkin method is equivalent to the collocation given above , and hence to an implicit Runge-Kutta method . Hence Galerkin methods may also be considered within the framework of m-block implicit methods .

In an interpolatory quadrature method (12) we find $u(t) \in \mathcal{P}_\mu$ such that $u(0) = y_0$ and $u(a_\nu h) = u(0) + h \sum_{j=1}^{\mu} b_{\nu q} f(a_q h, u(a_q h))$ for $\nu = 1(1)\mu$, where $b_{\nu q} = \int_0^{a_\nu} l_q(\eta)\,d\eta$, and then put $y_1 = u(h)$, and then proceed in the same way to find y_2 and so on .

Hulme (12) showed that this method too is equivalent to the collocation method given , so that interpolatory quadrature methods may also be considered in our formulation .

Consider now a perturbed collocation method (20) , for which we have a pre-assigned matrix

$$\tilde{P} = \begin{bmatrix} p_{01} & \cdots & p_{0,\mu} \\ \vdots & & \vdots \\ p_{\mu-1,1} & \cdots & p_{\mu-1,\mu} \end{bmatrix} ,$$

$N_j(t) = \frac{1}{j!}\sum_{i=0}^{\mu}(p_{ij} - \delta_{ij})t^i$, $p_{\mu j} = 0$ for $j = 1(1)\mu$, where δ_{ij} is the Krönecker δ , and for any $u(t) \in \mathcal{P}_\mu$,

$$(Pu)(t) = u(t) + \sum_{j=1}^{\mu} N_j(\frac{t}{h})h^j u^{(j)}(0) \in \mathcal{P}_\mu \quad .$$

A general perturbed collocation (PECO) method consists of finding $u(t) \in \mathcal{P}_\mu$ such that $u(0) = y_0$ and $u'(a_\nu h) = f(a_\nu h,(Pu)(a_\nu h))$ for $\nu = 1(1)\mu$, and then putting $y_1 = u(h)$. We proceed in the same way to obtain y_2 and so on .

Nørsett (20) showed that PECO is equivalent to the implicit Runge-Kutta method (3.3.1) with

$$b_{\nu q} = \int_0^a l_q(\eta)d\eta + \sum_{\alpha=1}^{\infty} N_\alpha(a_\nu) l_q^{(\alpha-1)}(0) \text{ and } c_\nu = \int_0^1 l_\nu(\eta)d\eta \ .$$

CHAPTER FOUR .

BLOCK METHODS .

Having seen the wide variety of methods which fall under the block semi explicit and block implicit umbrellas , we may define the class of block methods as the union of the block semi explicit and block implicit methods . Each block method then is characterised by an array $(A_0,A_1, \dots ,A_r \mid B_0,B_1, \dots ,B_r \mid \gamma)$.

4.1 Truncation Error .

From now on the symbols C and M , with or without subscripts or superscripts , will be assumed to be real , strictly positive constants , independent of h , unless otherwise stated .

Definition 4.1.1 .

The LOCAL TRUNCATION ERROR VECTOR $\Theta^* \varepsilon \mathbb{R}^{r+nm}$ is defined for an m-block method by $\Theta^* = A_h\Delta_h y(t) - hB_h\Delta_h y'(t) - g$.

Definition 4.1.2 .

The r starting values $\tilde{y}_0,\tilde{y}_1, \dots ,\tilde{y}_{r-1}$ are ACCURATE of order p ε N if there exist $C'_0,C'_1, \dots ,C'_{r-1}$ such that $\Theta^*_i = C'_i h^{p+O(h^{p+1})}$ for $i = 0(1)r-1$.

Definition 4.1.3 .

An m-block method is CONSISTENT of order p ε N if there exist $C_1,C_2, \dots ,C_m$, some of which may be zero , such that for every $y(t) \varepsilon C^{(p+1)}[0,T]$, $\Theta^*_{r+sm+\nu-1} = C_\nu[\Delta_h y^{(p+1)}(t)]_{r+sm} h^{p+1} + O(h^{p+2})$ as $h \to 0$, for $\nu = 1(1)m$ and $s = 0(1)n-1$.

For block methods zero stability is defined as for cyclic methods .

Definition 4.1.4 .

An m-block method is CONVERGENT of order p ε N if there exists C_U such that $\| e \|_\infty \leq C_U h^p$, where the GLOBAL ERROR VECTOR e is $y^* - \Delta_h y(t)$.

For each $\nu = 1(1)m$ we refer to the $\nu^{th.}$ STAGE of an m-block method as corresponding to the components $y^*_{r+sm+\nu-1}$, $s = 0(1)n-1$.

4.2 The Numerical Stability of Block Semi Explicit Methods .

For numerical stability we consider the differential equation $y' = \lambda y$ for some $\lambda < 0$. A method is absolutely stable for fixed $\bar{h} = h\lambda$ if y^*_{r+sm-1} is decreasing as $n \to \infty$. Now for an acceptable method it is desirable to have A-stability , which is the term given to denote the situation where we have absolute stability for all complex values of $\bar{h}$ with negative real part .

Any block semi explicit method may be written $A_h y^* = hB_h f^* + g$ $= h\lambda B_h y^* + g$, so $y^* = (A_h - \bar{h}B_h)^{-1} g$. Hence the solution is decreasing if there exists M^* such that $\|(A_h - \bar{h}B_h)^{-1}\|_\infty \leq M^*$.

Now A_h and B_h have recurring arrays like those for cyclic methods (1.3.2) , so we may define the companion matrices $G_1, G_2, \ldots, G_m$ and the matrix G for $A_h - \bar{h}B_h$ as in definition 1.3.4 . We now need the following theorem .

Theorem 4.2.1 .

If the elements of $(A_h - \bar{h}B_h)$ are independent of n , then there exists M^* such that $\|(A_h - \bar{h}B_h)^{-1}\|_\infty \leq M^*$ if and only if G has all its eigenvalues strictly inside the unit circle .

This result can be found in (18) . Hence we obtain the following theorem .

Theorem 4.2.2 .

An m-block semi explicit method is absolutely stable for some $\bar{h}$ if the eigenvalues of G all lie strictly within the unit circle .

Definition 4.2.3 .

For the linear multistep method (1.2.1) the FIRST and SECOND CHARACTERISTIC POLYNOMIALS are respectively

$$\rho(\zeta) = \sum_{j=0}^{k} \alpha_{k-j}\zeta^{k-j} \quad \text{and} \quad \sigma(\zeta) = \sum_{j=0}^{k} \beta_{k-j}\zeta^{k-j} \ .$$

Definition 4.2.4 .

For the linear multistep method (1.2.1) the STABILITY POLYNOMIAL is $\pi(\zeta,\bar{h}) = \rho(\zeta) - \bar{h}\sigma(\zeta)$.

Definition 4.2.5 .

For an m-block semi explicit method the stability polynomial is $\pi(\zeta,\bar{h}) = \det(G-\zeta I)$, where G is as in theorem 4.2.2 .

For a given $\bar{h}$ a method is absolutely stable if all the roots of the equation $\pi(\zeta,\bar{h}) = 0$ are less than one in modulus . In this section we shall consider the absolute stability of predictor-corrector methods , obtaining the stability polynomials given by Lambert (15 , pp.97,98) , and shall then examine the stability of certain extrapolation methods .

Proposition 4.2.6 .

For the predictor-corrector scheme (2.3.1) in $P(EC)^{\mu}E$ mode the stability polynomial is $\pi_{P(EC)^{\mu}E}(\zeta,\bar{h}) = \rho(\zeta) - \bar{h}\sigma(\zeta) + M_{\mu}(\bar{h})[\rho^{*}(\zeta) - \bar{h}\sigma^{*}(\zeta)]$, where $M_{\mu}(\bar{h}) = \dfrac{(\bar{h}\beta_k)^{\mu}(1 - \bar{h}\beta_k)}{1 - (\bar{h}\beta_k)^{\mu}}$ and $\rho^{*}(\zeta) - h\sigma^{*}(\zeta)$ and $\rho(\zeta) - h\sigma(\zeta)$ are the stability polynomials of the predictor and corrector respectively .

Proof .

$A_h - \bar{h}B_h$ is composed of the repeated array

$$\begin{array}{c|c|c|c|c}
\begin{matrix} 0 & \cdots & 0 & \alpha_0^*-\bar{h}\beta_0^* \\ & \ddots & \vdots & \alpha_0-\bar{h}\beta_0 \\ & & 0 & \vdots \\ & & & \alpha_0-\bar{h}\beta_0 \end{matrix} & \begin{matrix} 0 & \cdots & 0 & \alpha_1^*-\bar{h}\beta_1^* \\ 0 & \cdots & 0 & \alpha_1-\bar{h}\beta_1 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \alpha_1-\bar{h}\beta_1 \end{matrix} & \cdots & \begin{matrix} 0 & \cdots & 0 & \alpha_{k-1}^*-\bar{h}\beta_{k-1}^* \\ 0 & \cdots & 0 & \alpha_{k-1}-\bar{h}\beta_{k-1} \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \alpha_{k-1}-\bar{h}\beta_{k-1} \end{matrix} \\
\hline
\end{array}$$

$$\begin{bmatrix} 1 & & & & \\ -\bar{h}\beta_k & 1 & & & \\ & \ddots & \ddots & & \\ & 0 & \ddots & \ddots & \\ & & & -\bar{h}\beta_k & 1 \end{bmatrix}$$

where without loss of generality we take $\alpha_k^* = \alpha_k = 1$.

$$G_1 = \begin{bmatrix} 0 & 1 & & & & & & & & & \\ \vdots & & 1 & & & & & & & & \\ & & & 1 & & & & & & & \\ & & & & \ddots & & & & & & \\ & & & & & & & & & & 1 \\ \underbrace{0 \cdots 0}_{\mu} & -\alpha_0^*+\bar{h}\beta_0^* & \underbrace{0 \cdots 0}_{\mu} & -\alpha_1^*+\bar{h}\beta_1^* & \cdots & \underbrace{0 \cdots 0}_{\mu} & -\alpha_{k-1}^*+\bar{h}\beta_{k-1}^* \end{bmatrix},$$

$$G_\nu = \begin{bmatrix} 0 & 1 & & & & & & & & & \\ \vdots & & 1 & & & & & & & & \\ & & & 1 & & & & & & & \\ & & & & \ddots & & & & & & \\ & & & & & & & & & & 1 \\ \underbrace{0 \cdots 0}_{\mu-\nu+1} & -\alpha_0+\bar{h}\beta_0 & \underbrace{0 \cdots 0}_{\mu} & -\alpha_1+\bar{h}\beta_1 & \cdots & \underbrace{0 \cdots 0}_{\mu} & -\alpha_{k-1}+\bar{h}\beta_{k-1} & & & & \\ & & & & & & & \ddots & & & \\ & & & & & & & & & & 1 \\ & & & & & & & \underbrace{0 \cdots 0}_{\nu-2} & & & \bar{h}\beta_k \end{bmatrix},$$

$\nu = 2(1)\mu+1$. By direct computation we obtain $G = G_{\mu+1}G_\mu \cdots G_2G_1 =$

$$\begin{pmatrix} 0_{\mu+1} & I_{\mu+1} & & \\ 0_{\mu+1} & 0_{\mu+1} & I_{\mu+1} & \\ \vdots & & \ddots & \ddots \\ 0_{\mu+1} & \cdots & 0_{\mu+1} & I_{\mu+1} \\ \Gamma_0 & \Gamma_1 & \cdots & \Gamma_{k-1} \end{pmatrix} \quad \text{where } \Gamma_i = \begin{pmatrix} 0 & \cdots & 0 & \gamma_{i0} \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \gamma_{i\mu} \end{pmatrix} ,$$

a $(\mu+1)\times(\mu+1)$ matrix , and the γ's are defined by the relations

$$\gamma_{io} = -\alpha_i^* + \bar{h}\beta_i^* ,$$

$$\gamma_{i,j+1} = -\alpha_i + \bar{h}\beta_i + \bar{h}\beta_k\gamma_{ij} , \quad j = 0(1)\mu-1 .$$

Now using a result on pp.33,34 of (25) , we have

$$\pi_{P(EC)^\mu E}(\zeta,\bar{h}) = \det(G - \zeta I) \propto \det(\Gamma_0 + \Gamma_1\zeta + \cdots + \Gamma_{k-1}\zeta^{k-1} - I\zeta^k) \quad (*)$$

$$= \begin{vmatrix} -\zeta^k & & & \gamma_{00} + \zeta\gamma_{10} + \cdots + \zeta^{k-1}\gamma_{k-1,0} \\ & -\zeta^k & & \gamma_{01} + \zeta\gamma_{11} + \cdots + \zeta^{k-1}\gamma_{k-1,1} \\ & & \ddots & \vdots \\ & & -\zeta^k & \gamma_{0,\mu-1} + \zeta\gamma_{1,\mu-1} + \cdots + \zeta^{k-1}\gamma_{k-1,\mu-1} \\ & & & \gamma_{0\mu} + \zeta\gamma_{1\mu} + \cdots + \zeta^{k-1}\gamma_{k-1,\mu} - \zeta^k \end{vmatrix}$$

$$= (-\zeta^k)^\mu(\gamma_{0\mu} + \zeta\gamma_{1\mu} + \cdots + \zeta^{k-1}\gamma_{k-1,\mu} - \zeta^k)$$

$$\propto \zeta^k - \zeta^{k-1}\gamma_{k-1,\mu} - \cdots - \zeta\gamma_{1\mu} - \gamma_{0\mu} .$$

Now by direct computation

$$\gamma_{i\mu} = [1 + \bar{h}\beta_k + (\bar{h}\beta_k)^2 + \cdots + (\bar{h}\beta_k)^{\mu-1}](-\alpha_i + \bar{h}\beta_i) + (\bar{h}\beta_k)^\mu(-\alpha_i^* + \bar{h}\beta_i^*)$$

$$= \frac{1 - (\bar{h}\beta_k)^\mu}{1 - \bar{h}\beta_k}(-\alpha_i + \bar{h}\beta_i) + (\bar{h}\beta_k)^\mu(-\alpha_i^* + h\beta_i^*)$$

$$= \frac{1 - (\bar{h}\beta_k)^\mu}{1 - \bar{h}\beta_k}[-\alpha_i + \bar{h}\beta_i + M_\mu(\bar{h})(-\alpha_i^* + h\beta_i^*)] .$$

(*) By "$\det P(\zeta) \propto \det Q(\zeta)$" we mean that $\det P(\zeta) = K\zeta^q \det Q(\zeta)$ for some constant K and integer q . This implies that the solutions of the equation $\det P(\zeta) = 0$ are $\zeta = 0$ q times and the zeroes of $\det Q(\zeta)$.

Hence $\pi_{P(EC)^{\mu}E}(\zeta,h) \propto \zeta^k - \dfrac{1-(\bar{h}\beta_k)^{\mu}}{1-\bar{h}\beta_k}\sum_{j=0}^{k-1}[(-\alpha_j + \bar{h}\beta_j) + M_{\mu}(\bar{h})(-\alpha_j^* + h\beta_j^*)]\zeta^j$

$$= \zeta^k - \frac{1-(\bar{h}\beta_k)^{\mu}}{1-\bar{h}\beta_k}[(1-\bar{h}\beta_k)\zeta^k - \rho(\zeta) + \bar{h}\sigma(\zeta)$$

$$+ M_{\mu}(h)(\zeta^k - \rho^*(\zeta) + \bar{h}\sigma^*(\zeta))]$$

$$= \frac{1-(\bar{h}\beta_k)^{\mu}}{1-\bar{h}\beta_k}[\rho(\zeta) - \bar{h}\sigma(\zeta) + M_{\mu}(\bar{h})(\rho^*(\zeta) - \bar{h}\sigma^*(\zeta))] ,$$

and so the result follows.

Proposition 4.2.7 .

For the predictor-corrector scheme (2.3.1) in $P(EC)^{\mu}$ mode the stability polynomial is $\pi_{P(EC)^{\mu}}(\zeta,h) = \beta_k\zeta^k[\rho(\zeta) - \bar{h}\sigma(\zeta)]$

$$+ M_{\mu}(\bar{h})[\rho^*(\zeta)\sigma(\zeta) - \rho(\zeta)\sigma^*(\zeta)]$$

Proof .

Again taking $\alpha_k^* = \alpha_k = 1$ we now have in $A_h - \bar{h}B_h$ the repeated array

$$\left[\begin{array}{ccccc|ccccc|c|ccccc}
0 & \cdots & 0 & -\bar{h}\beta_0^* & \alpha_0^* & 0 & \cdots & 0 & -\bar{h}\beta_1^* & \alpha_1^* & & 0 & \cdots & 0 & -\bar{h}\beta_{k-1}^* & \alpha_{k-1}^* \\
 & \ddots & \vdots & -\bar{h}\beta_0 & \alpha_0 & 0 & \cdots & 0 & -\bar{h}\beta_1 & \alpha_1 & \cdots & 0 & \cdots & 0 & -\bar{h}\beta_{k-1} & \alpha_{k-1} \\
 & & 0 & \vdots & & \vdots & & \vdots & \vdots & \vdots & & \vdots & & \vdots & \vdots & \vdots \\
 & & & -\bar{h}\beta_0 & \alpha_0 & 0 & \cdots & 0 & -\bar{h}\beta_1 & \alpha_1 & & 0 & \cdots & 0 & -\bar{h}\beta_{k-1} & \alpha_{k-1}
\end{array}\right]$$

$$\left[\begin{array}{cccc}
1 & & & \\
-\bar{h}\beta_k & 1 & & \\
 & \ddots & \ddots & \\
 & & -\bar{h}\beta_k & 1
\end{array}\right] ,$$

and this gives

$$G_1 = \begin{bmatrix} 0 & 1 & & & & & & & & & & \\ \vdots & & 1 & & & & & & & & & \\ & & & 1 & & & & & & & & \\ & & & & \ddots & & & & & & & \\ & & & & & \ddots & & & & & & \\ & & & & & & \ddots & & & & & 1 \\ 0 \cdots 0 & \bar{h}\beta_0^* & -\alpha_0^* & 0 \cdots 0 & \bar{h}\beta_1^* & -\alpha_1^* & \cdots & 0 \cdots 0 & \bar{h}\beta_{k-1}^* & -\alpha_{k-1}^* \end{bmatrix}$$

(μ zeroes; $\mu-1$ zeroes; $\mu-1$ zeroes)

$$G_\nu = \begin{bmatrix} 0 & 1 & & & & & & & & & & \\ \vdots & & 1 & & & & & & & & & \\ & & & 1 & & & & & & & & \\ & & & & \ddots & & & & & & & \\ & & & & & \ddots & & & & & & \\ & & & & & & \ddots & & & & & \\ 0 \cdots 0 & \bar{h}\beta_0 & -\alpha_0 & 0 \cdots 0 & \bar{h}\beta_1 & -\alpha_1 & \cdots & 0 \cdots 0 & \bar{h}\beta_{k-1} & -\alpha_{k-1} & & \\ & & & & & & & & & & \ddots & 1 \\ & & & & & & & & & & 0 \cdots 0 & \bar{h}\beta_k \end{bmatrix},$$

($\mu-\nu+1$ zeroes; $\mu-1$ zeroes; $\mu-1$ zeroes; $\nu-2$ zeroes)

$\nu = 2(1)\mu+1$.

Hence by direct computation G is the $[1 + k(\mu+1)]x[1 + k(\mu+1)]$ matrix whose zero$^{th.}$ column is composed entirely of zeroes , whose zero$^{th.}$ row is composed of zeroes , except the $(\mu+1)^{th.}$ element , which is 1 , and whose remaining rows and columns are

$$\begin{bmatrix} 0_{\mu+1} & I_{\mu+1} & & & \\ 0_{\mu+1} & 0_{\mu+1} & I_{\mu+1} & & \\ \vdots & & \ddots & \ddots & \\ 0_{\mu+1} & 0_{\mu+1} & \cdots & 0_{\mu+1} & I_{\mu+1} \\ \Gamma_0^* & \Gamma_1^* & \cdots & \Gamma_{k-2}^* & \Gamma_{k-1}^* \end{bmatrix}$$

, where Γ_i^* is the $(\mu+1)x(\mu+1)$ matrix

$$\begin{bmatrix} 0 \cdots 0 & \delta_{i0} & \varepsilon_{i0} \\ 0 \cdots 0 & \delta_{i1} & \varepsilon_{i1} \\ \vdots & \vdots & \vdots \\ 0 \cdots 0 & \delta_{i\mu} & \varepsilon_{i\mu} \end{bmatrix}$$

and $\delta_{i0} = \bar{h}\beta_i^* \ , \ \delta_{i,j+1} = \bar{h}\beta_i + \bar{h}\beta_k \delta_{ij} \ ,$

$\varepsilon_{i0} = -\alpha_i^* \ , \ \varepsilon_{i,j+1} = -\alpha_i + \bar{h}\beta_k \varepsilon_{ij} \ , \ j = 0(1)\mu-1 \ .$

Therefore again using the result on pp.33,34 of (25) , we have

$$\pi_{P(EC)^\mu}(\zeta,\bar{h}) \propto \det(\Gamma_0^* + \Gamma_1^*\zeta + \ldots + \Gamma_{k-1}^*\zeta^{k-1} - I\zeta^k)$$

$$= \begin{vmatrix} -\zeta^k & & & x_0 & z_0 \\ & \ddots & & x_1 & z_1 \\ & & -\zeta^k & \vdots & \vdots \\ & & & x_{\mu-1}-\zeta^k & z_{\mu-1} \\ & & & x_\mu & z_\mu - \zeta^k \end{vmatrix} \ ,$$

where for $j = 0(1)\mu$, $x_j = \delta_{0j} + \zeta\delta_{1j} + \ldots + \zeta^{k-1}\delta_{k-1,j}$

and $z_j = \varepsilon_{0j} + \zeta\varepsilon_{1j} + \ldots + \zeta^{k-1}\varepsilon_{k-1,j}$.

Therefore $\pi_{P(EC)^\mu}(\zeta,\bar{h}) \propto (-\zeta^k)^{\mu-1}[(\zeta^k - x_{\mu-1})(\zeta^k - z_\mu) - z_{\mu-1}x_\mu]$ (4.2.1) .

Now by direct computation ,

$$\delta_{i,\mu-1} = \bar{h}\beta_i \frac{1 - (\bar{h}\beta_k)^{\mu-1}}{1 - \bar{h}\beta_k} + (\bar{h}\beta_k)^{\mu-1}(\bar{h}\beta_i^*) \ ,$$

$$\delta_{i\mu} = \bar{h}\beta_i \frac{1 - (\bar{h}\beta_k)^{\mu}}{1 - \bar{h}\beta_k} + (\bar{h}\beta_k)^{\mu}(\bar{h}\beta_i^*) \ ,$$

$$\varepsilon_{i,\mu-1} = -\alpha_i \frac{1 - (\bar{h}\beta_k)^{\mu-1}}{1 - \bar{h}\beta_k} - \alpha_i^*(\bar{h}\beta_k)^{\mu-1} \ ,$$

$$\varepsilon_{i\mu} = -\alpha_i \frac{1 - (\bar{h}\beta_k)^{\mu}}{1 - \bar{h}\beta_k} - \alpha_i^*(\bar{h}\beta_k)^{\mu} \ .$$

Hence $\zeta^k - x_{\mu-1} = \zeta^k \frac{1 - (\bar{h}\beta_k)^{\mu}}{1 - \bar{h}\beta_k} - \bar{h}\sigma(\zeta) \frac{1 - (\bar{h}\beta_k)^{\mu-1}}{1 - \bar{h}\beta_k} - (\bar{h}\beta_k)^{\mu-1}\bar{h}\sigma^*(\zeta) \ ,$

$$\zeta^k - z_\mu = \frac{1 - (\bar{h}\beta_k)^{\mu}}{1 - \bar{h}\beta_k}[\rho(\zeta) - \bar{h}\zeta^k\beta_k] + (\bar{h}\beta_k)^{\mu}\rho^*(\zeta) \ ,$$

$$z_{\mu-1} = \zeta^k \frac{1-(\bar{h}\beta_k)^{\mu}}{1-\bar{h}\beta_k} - \rho(\zeta)\frac{1-(\bar{h}\beta_k)^{\mu-1}}{1-\bar{h}\beta_k} - (\bar{h}\beta_k)^{\mu-1}\rho^*(\zeta) \quad \text{and}$$

$$x_{\mu} = \frac{1-(\bar{h}\beta_k)^{\mu}}{1-\bar{h}\beta_k}\bar{h}[\sigma(\zeta) - \beta_k\zeta^k] + (\bar{h}\beta_k)^{\mu}\bar{h}\sigma^*(\zeta) \quad .$$

By substituting these last four equations into (4.2.1) we obtain the result .

We now go on to examine the absolute stability of some extrapolation methods considered by Dahlquist (7) , beginning with example 2.6.1 , for which $A_h - \bar{h}B_h$ has the recurring array

$$\begin{array}{cccccc}
 & 0 & 0 & -1-\frac{\bar{h}}{4} & 1-\frac{\bar{h}}{4} & \\
 & 0 & 0 & -1-\frac{\bar{h}}{4} & 1-\frac{\bar{h}}{4} & \\
-1-\frac{\bar{h}}{2} & & 0 & & 0 & 1-\frac{\bar{h}}{2} \\
 & & 0 & -4/3 & 1/3 & 1
\end{array} \quad ,$$

giving $G_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \frac{4+\bar{h}}{4-\bar{h}} \end{bmatrix}$, $G_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \frac{4+\bar{h}}{4-\bar{h}} \end{bmatrix}$, $G_3 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{2+\bar{h}}{2-\bar{h}} & 0 & 0 \end{bmatrix}$

$G_4 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 4/3 & -1/3 \end{bmatrix}$, and $G = \begin{bmatrix} 0 & 0 & (\frac{4+\bar{h}}{4-\bar{h}})^2 \\ 0 & 0 & \frac{2+\bar{h}}{2-\bar{h}} \\ 0 & 0 & \frac{4}{3}(\frac{4+\bar{h}}{4-\bar{h}})^2 - \frac{1}{3}(\frac{2+\bar{h}}{2-\bar{h}}) \end{bmatrix}$.

Therefore the stability polynomial is $\zeta^3 - \zeta^2[\frac{4}{3}(\frac{4+\bar{h}}{4-\bar{h}})^2 - \frac{1}{3}(\frac{2+\bar{h}}{2-\bar{h}})]$, and so we have absolute stability for those $\bar{h}$ for which $\left|\frac{4}{3}(\frac{4+\bar{h}}{4-\bar{h}})^2 - \frac{1}{3}(\frac{2+\bar{h}}{2-\bar{h}})\right| < 1$.

Unfortunately $\lim_{x \to -\infty}[\frac{4}{3}(\frac{4+x}{4-x})^2 - \frac{1}{3}(\frac{2+x}{2-x})] = \frac{5}{3}$, so we do not have A-stability .

This then is the outcome when we use the extrapolated y_{i-1} to compute y_i . If instead we use one of the non-extrapolated $y_{i-1}^{[0]}$ or $y_{i-1}^{[1]}$, we do get A-stability , as we shall now see .

Example 4.2.7 .

Using $y_{i-1}^{[0]}$ the scheme is

$$y_{i-\frac{1}{2}} - y_{i-1}^{[0]} = \frac{h}{4}(f_{i-\frac{1}{2}} + f_{i-1}^{[0]}) ,$$
$$y_i^{[0]} - y_{i-\frac{1}{2}} = \frac{h}{4}(f_i^{[0]} + f_{i-\frac{1}{2}}) ,$$
$$y_i^{[1]} - y_{i-1}^{[0]} = \frac{h}{2}(f_i^{[1]} + f_{i-1}^{[0]}) ,$$
$$y_i + \tfrac{1}{3}y_i^{[1]} - \tfrac{4}{3}y_i^{[0]} = 0 .$$

This has the array

$$\left(\begin{array}{cccccccc|cccccccc|c}
0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1/4 & 0 & 0 & 1/4 & 0 & 0 & 0 & \frac{1}{2} \\
0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 1/4 & 0 & 0 & 1 \\
0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 1/2 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & -4/3 & 1/3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}\right) ,$$

and so in $A_h - \bar{h}B_h$ we have the recurring array

$$\begin{array}{cccccccc}
0 & 0 & -1 - \frac{\bar{h}}{4} & 0 & 0 & 1 - \frac{\bar{h}}{4} & & \\
 & 0 & 0 & 0 & 0 & -1 - \frac{\bar{h}}{4} & 1 - \frac{\bar{h}}{4} & \\
 & & -1 - \frac{\bar{h}}{2} & 0 & 0 & 0 & 0 & 1 - \frac{\bar{h}}{2} \\
 & & & 0 & 0 & 0 & -4/3 & 1/3 \quad 1
\end{array} ,$$

from which we obtain $G = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{4+\bar{h}}{4-\bar{h}} & 0 & 0 \\ 0 & 0 & (\frac{4+\bar{h}}{4-\bar{h}})^2 & 0 & 0 \\ 0 & 0 & \frac{2+\bar{h}}{2-\bar{h}} & 0 & 0 \\ 0 & 0 & \frac{4}{3}(\frac{4+\bar{h}}{4-\bar{h}})^2 - \frac{1}{3}(\frac{2+\bar{h}}{2-\bar{h}}) & 0 & 0 \end{pmatrix}$,

and the stability polynomial is $\zeta^5 - \zeta^4(\frac{4+\bar{h}}{4-\bar{h}})^2$. Since $\left|\frac{4+\bar{h}}{4-\bar{h}}\right| < 1$ for all complex $\bar{h}$ with negative real part , we have A-stability .

Example 4.2.8 .

Using $y_{i-1}^{[1]}$ and $f_{i-1}^{[1]}$ instead of $y_{i-1}^{[0]}$ and $f_{i-1}^{[0]}$ we have the array

$$\left(\begin{array}{cccccccc|cccccccc|c} 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1/4 & 0 & 1/4 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 1/4 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 1/2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & -4/3 & 1/3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right)$$

and so in $A_h - \bar{h}B_h$ we have the recurring array

$$\begin{array}{cccccccc} 0 & 0 & -1-\frac{\bar{h}}{4} & 0 & 1-\frac{\bar{h}}{4} & & & \\ & 0 & 0 & 0 & -1-\frac{\bar{h}}{4} & 1-\frac{\bar{h}}{4} & & \\ & & -1-\frac{\bar{h}}{2} & 0 & 0 & 0 & 1-\frac{\bar{h}}{2} & \\ & & & 0 & 0 & -4/3 & 1/3 & 1 \end{array}$$

from which we obtain $G = \begin{pmatrix} 0 & 0 & \frac{4+\bar{h}}{4-\bar{h}} & 0 \\ 0 & 0 & (\frac{4+\bar{h}}{4-\bar{h}})^2 & 0 \\ 0 & 0 & \frac{2+\bar{h}}{2-\bar{h}} & 0 \\ 0 & 0 & \frac{4}{3}(\frac{4+\bar{h}}{4-\bar{h}})^2 - \frac{1}{3}(\frac{2+\bar{h}}{2-\bar{h}}) & 0 \end{pmatrix}$,

so that the stability polynomial is $\zeta^4 - \zeta^3(\frac{2+\bar{h}}{2-\bar{h}})$. Since $\left|\frac{2+\bar{h}}{2-\bar{h}}\right| < 1$ for all complex $\bar{h}$ with negative real part , we have A-stability .

4.3 Zero Stability .

We recall that for block methods zero stability is defined as for cyclic methods .

Let us first consider Runge-Kutta methods . For both the explicit case (2.4.2) and the implicit case (3.3.2)

$$A_o = \begin{pmatrix} 0 & \cdots & 0 & -1 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & -1 \end{pmatrix} \quad \text{and } A_1 = I_m .$$

The companion matrices (1.3.3) are $G_\nu = \begin{pmatrix} 0 & 1 & & & \\ & \ddots & & & \\ & & \ddots & & \\ & & & \ddots & 1 \\ 0 & \cdots & 0 & 1 & 0 \cdots 0 \end{pmatrix}$,

with $m-\nu$ zeros before the 1 and $\nu-1$ zeros after it.

$\nu = 1(1)m$, and hence $G = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 1 \end{pmatrix}$, which has for eigenvalues $(m-1)$ zeroes and 1 , and so any Runge-Kutta method is zero stable .

We now consider the zero stability of the predictor-corrector schemes whose numerical stability we looked at in section 4.2 .

The matrix G for zero stability may be obtained from the corresponding matrix for absolute stability by putting $\bar{h} = 0$, since for absolute stability we demand that $\|(A_h - \bar{h}B_h)^{-1}\|_\infty \leq M^*$, while for zero stability we stipulate $\max_{i,j}|(A_h)^{-1}_{ij}| = 0(1)$. Accordingly if we take the stability polynomial of a given method and put $\bar{h} = 0$, we may test for zero stability by requiring that the resulting polynomial have none of its roots with modulus greater than one , and with those roots of modulus one having linear elementary divisors .

Doing this for $P(EC)^\mu E$ mode we obtain from proposition 4.2.6 the polynomial $\rho(\zeta)$, and so the predictor-corrector scheme is zero stable if and only if the corrector as a linear multistep method is . From proposition 4.2.7 we obtain for $P(EC)^\mu$ mode the polynomial $\beta_k \zeta^k \rho(\zeta)$, so this scheme is also zero stable if and only if the corrector is .

4.4 Convergence Theorems .

In this section we provide convergence proofs for block methods based on definition 4.1.3 of consistency .

Definition 4.4.1 .

For any vectors $u,v \in \mathbb{R}^{r+nm}$ we define the vector $f(u,v) \in \mathbb{R}^{r+nm}$ by $[f(u,v)]_i = f(u_i,v_i)$ for all i .

Thus $\Delta_h y'(t) = \Delta_h f(t,y(t)) = f(\Delta_h t,\Delta_h y(t))$ and $f^* = f(\Delta_h t,y^*)$.

Theorem 4.4.2 .

If an m-block semi explicit method is zero stable , consistent of order p and starting values are accurate of order p , then the method is convergent of order p .

Proof .

The method may be written $A_h y^* = hB_h f(\Delta_h t,y^*) + g$, and by definition 4.1.1 we have $A_h \Delta_h y(t) = hB_h f(\Delta_h t,\Delta_h y(t)) + g - \theta^*$

By subtraction we obtain $A_h e = \theta^* + hB_h[f(\Delta_h t,y^*) - f(\Delta_h t,\Delta_h y(t))]$,

$$\text{so } e = A_h^{-1}\theta^* + hA_h^{-1}B_h[f(\Delta_h t,y^*) - f(\Delta_h t,\Delta_h y(t))] .$$

Therefore for all i , by the triangle inequality and applying the Lipschitz condition (1.1.2) to f , $|e_i| \le |[A_h^{-1}\theta^*]_i| + hL\sum_{j=0}^{i} |(A_h^{-1}B_h)_{ij}||e_j|$.

The summation is up to i since A_h and B_h are both lower triangular and therefore so is $A_h^{-1}B_h$.

Now zero stability implies that there exists M_o such that $\max_{0\le i,j\le r+nm-1} |(A_h^{-1})_{ij}| = M_o$. Since A_h is lower triangular , so is A_h^{-1} , and therefore $|[A_h^{-1}\theta^*]_i| \le M_o\sum_{j=0}^{i}|\theta_j^*| = M_o[\sum_{j=0}^{r-1} C_j' h^p + \sum_{j=r}^{i}|\theta_j^*|] = O(h^p)$ for all i , where $\sum_{j=r}^{i} \equiv 0$ for $i < r$.

$$|(A_h^{-1}B_h)_{ij}| \le \sum_{q=0}^{r+nm-1} |(A_h^{-1})_{iq}||(B_h)_{qj}| \text{ by the triangle inequality ,}$$

$$\le M_o\sum_{q=0}^{r+nm-1}|(B_h)_{qj}| \le M_o\|B_h\|_1 \quad (4.4.1) ,$$

which is bounded since all elements of B_h are finite and B_h has bandwidth independent of h . Therefore there exist C' and M' such that

$$|e_i| \le C'h^p + hM' \sum_{j=0}^{i-1} |e_j| + hM'|e_i| \text{ , therefore}$$

$(1 - hM')|e_i| \le C'h^p + hM' \sum_{j=0}^{i-1} |e_j|$ .So for sufficiently small h there exist C and M such that

$$|e_i| \le Ch^p + hM \sum_{j=0}^{i-1} |e_j| \ .$$

We obtain the result $\|e\|_\infty \le C_U h^p$ for some C_U by making use of the following lemma .

Lemma 4.4.3 .

Let $x = (x_0, x_1, \dots, x_\omega)^T$ be a real vector with

$|x_i| \le \delta + hM \sum_{j=0}^{i-1} |x_j|$, $i = 1(1)\omega$, where M and δ are positive constants .

Then $\|x\|_\infty \le (hM|x_0| + \delta)\exp(M\omega h)$.

The proof of this lemma is elementary .

Theorem 4.4.4 .

If an m-block implicit method is zero stable , consistent of order p and starting values are accurate of order p , then it is convergent of order p .

Proof .

We may take A_h to be lower triangular without loss of generality . As in theorem 4.4.2 , $e = A_h^{-1}\theta^* + hA_h^{-1}B_h[f(\Delta_h t, y^*) - f(\Delta_h t, \Delta_h y(t))]$. Now B_h has the form (2.1.2) , where B_τ is an m x m matrix on the diagonal and is not lower triangular . Therefore due to the particular form of $A_h^{-1}B_h$, $|e_i| \le |[A_h^{-1}\theta^*]_i| + hL \sum_{j=0}^{i+m-\nu} |(A_h^{-1}B_h)_{ij}||e_j|$, where $\nu \equiv i-r+1 \pmod m$, $1 \le \nu \le m$ and $i \ge r$. As in theorem 4.4.2 , $[A_h^{-1}\theta^*]_i = O(h^p)$ for all i , so we

obtain , putting $i = r + sm + \nu - 1$,

$$|e_{r+sm+\nu-1}| \leq C'h^p + hL \sum_{j=0}^{r+(s+1)m-1} |(A_h^{-1}B_h)_{ij}|\,|e_j| \ , \ \nu = 1(1)\ m \ ,$$

$s = O(1)n-1$, for some C' . Again $|(A_h^{-1}B_h)_{ij}| \leq M_o\|B_h\|_1$, so there exists M' such that

$$|e_{r+sm+\nu-1}| \leq C'h^p + hM' \sum_{j=0}^{r+(s+1)m-1} |e_j| \ , \ \nu = 1(1)m \ , \ s = O(1)n-1 \ .$$

Using lemma 4.4.5 below we can reduce this inequality to

$$|e_{r+sm+\nu-1}| \leq Ch^p + hM \sum_{j=0}^{r+sm+\nu-1} |e_j| \ , \ \nu = 1(1)m \ , \ s = O(1)n-1 \text{ for}$$

sufficiently small h , for some C and M . That is ,

$|e_i| \leq Ch^p + hM \sum_{j=0}^{i} |e_j|$, which gives $\|e\|_\infty \leq C_U h^p$ for some C_U as in theorem 4.4.2 .

Lemma 4.4.5 .

Take any positive integers i and m and suppose that the real vector $(x_o, x_1, \ldots, x_{i+m})^T$ is such that for some real and positive M and δ

$|x_{i+\nu}| \leq \delta + hM \sum_{j=0}^{i+m} |x_j|$ for $\nu = 1(1)m$. Then for sufficiently small h

$$|x_{i+m-q}| \leq \frac{hM}{1 - qhM} \sum_{j=0}^{i+m-q} |x_j| + \frac{\delta}{1 - qhM} \quad \text{for } q = O(1)m-1 \ .$$

Proof .

We shall proceed by an induction argument on m . For m = 1 the result is immediate . Assume then that the lemma holds for m . Then for m+1 we have

$$|x_{i+\nu}| \leq \delta + hM \sum_{j=0}^{i+m+1} |x_j| \ , \ \nu = 1(1)m+1 \quad (4.4.2) \ .$$

Therefore when $\nu = m + 1$, $|x_{i+m+1}| \leq \delta + hM \sum_{j=0}^{i+m+1} |x_j|$, and so for sufficiently small h ,

$|x_{i+m+1}| \le \delta' + hM' \sum_{j=0}^{i+m} |x_j|$, where $M' = M/(1 - hM) \ge 0$ and $\delta' = \delta/(1 - hM) \ge 0$. Therefore for $\nu = 1(1)m$, eliminating $|x_{i+m+1}|$ we obtain

$$|x_{i+\nu}| \le \delta + hM \sum_{j=0}^{i+m} |x_j| + hM(\delta' + hM' \sum_{j=0}^{i+m} |x_j|)$$

$$\le \delta + hM\delta' + hM(1 - hM') \sum_{j=0}^{i+m} |x_j| = \frac{\delta}{1 - hM} + \frac{hM}{1 - hM} \sum_{j=0}^{i+m} |x_j|$$

$$\le \delta' + hM' \sum_{j=0}^{i+m} |x_j| .$$

Therefore by the assumed hypothesis , for $q = 0(1)m-1$ and for h sufficiently small ,

$$|x_{i+m-q}| \le \frac{hM'}{1 - qhM'} \sum_{j=0}^{i+m-q} |x_j| + \frac{\delta'}{1 - qhM'}$$

$$= \frac{\delta}{1 - (q+1)hM} + \frac{hM}{1 - (q+1)hM} \sum_{j=0}^{i+m-q} |x_j| .$$

That is ,

$$|x_{i+(m+1)-q}| \le \frac{\delta}{1 - qhM} + \frac{hM}{1 - qhM} \sum_{j=0}^{i+(m+1)-q} |x_j| , \quad q = 1(1)m .$$

The remaining case $q = 0$ is simply (4.4.2) .

CHAPTER FIVE .

TWO-SIDED ERROR BOUNDS .

5.1 Optimal Consistency .

Existing convergence proofs such as those of Stetter (24 , pp. 220,224,226) and the previous section usually give an upper bound on the infinity norm of the global error vector of the form

$$\| e \|_\infty \leq C_U h^p \qquad (5.1.1) .$$

This implies convergence of order at least p and holds provided the particular method is zero stable and consistent of order p and the r starting values are accurate of order p . Definition 4.1.3 of consistency however is not able to provide a two-sided error bound of the form

$$C_L h^p \leq \| e \|_\infty \leq C_U h^p \qquad (5.1.2) ,$$

as the following two examples will show .

Example 1.3.1 .

$$\theta^*_{2s+2} = -\frac{1}{90} h^5 y^v(t_{2s+2}) + O(h^6) ,$$

$$\theta^*_{2s+3} = -\frac{1}{24} h^4 y^{iv}(t_{2s+2}) + O(h^5) , \quad s = 0(1)n-1 .$$

Example 3.2.1 .

Taking the array (3.2.1) we have

$$\theta^*_{2s+1} = \frac{1}{24} h^4 y^{iv}(t_{2s+1}) + O(h^5) ,$$

$$\theta^*_{2s+2} = -\frac{1}{90} h^5 y^v(t_{2s+2}) + O(h^6) , \quad s = 0(1)n-1 .$$

Both examples then are consistent of order three , and yet computations show order four convergence , which we shall also show analytically .

A one-sided error bound of the form (5.1.1) then is not

sufficient to establish the exact order of convergence , whereas a two-sided error bound of the form (5.1.2) gives exact order of convergence p .

Albrecht (2) obtains a two-sided error bound by retaining a definition of consistency similar to definition 4.1.3 and stating conditions under which convergence is one order greater than consistency . Our strategy here will be different : we shall create a new concept , namely optimal consistency , which will permit a two-sided error bound , giving the correct order of convergence , q say , if and only if the method is optimally consistent of order q .

Definition 5.1.1 .

An m-block method is OPTIMALLY CONSISTENT of ORDERS $(p_1, p_2, \ldots, p_m)$ where $p_\nu \in \mathbb{N}$, $\nu = 1(1)m$, if for each $i \geq r$ there exists $C^{(i)} \neq 0$ such that $\left|[A_h^{-1}\theta^*]_i\right| = \left|[\Delta_h y(t) - hA_h^{-1}B_h\Delta_h y'(t) - A_h^{-1}g]_i\right| = C^{(i)}h^{p_\nu} + O(h^{p_\nu+1})$,

$$i = r+sm+\nu-1 \ , \ s = 0(1)n-1 \ , \ \nu = 1(1)m \ .$$

If $p = \min_{\nu=1(1)m} p_\nu$ then we say that the method is optimally consistent of order p .

Theorem 5.1.2 .

If an m-block method is optimally consistent of order p , zero stable and has r starting values accurate of order p then there exist C_L and C_U such that $C_L h^p \leq \|e\|_\infty \leq C_U h^p$.

Proof .

(i) The upper bound .

As in theorem 4.4.2 , $e = A_h^{-1}\theta^* + hA_h^{-1}B_h[f(\Delta_h t, y^*) - f(\Delta_h t, \Delta_h y(t))]$.

Therefore for each i , $|e_i| \leq \left|[A_h^{-1}\theta^*]_i\right| + hL\sum_{j=0}^{r+nm-1}\left|(A_h^{-1}B_h)_{ij}\right||e_j|$

$$= \left|[A_h^{-1}\theta^*]_i\right| + hL\sum_{j=0}^{i+m-\nu}\left|(A_h^{-1}B_h)_{ij}\right||e_j| \ ,$$

where $\nu \equiv i - r + 1 \pmod m$ and $1 \le \nu \le m$.

This applies to both block implicit and block semi explicit methods : for a block semi explicit method $(A_h^{-1}B_h)_{ij} = 0$ whenever $j > i$.

Therefore $|e_i| \le C^{(i)}h^{p_\nu} + hL\sum_{j=0}^{i+m-\nu}|(A_h^{-1}B_h)_{ij}||e_j|$ and using (4.4.1) and lemmas 4.4.3 and 4.4.5 again we obtain $\|e\|_\infty \le C_U h^p$.

(ii) The lower bound .

$A_h^{-1}\theta^* = e - hA_h^{-1}B_h[f(\Delta_h t, y^*) - f(\Delta_h t, \Delta_h y(t))]$ (5.1.3) , therefore

$$C^{(i)}h^{p_\nu} \le \|e\|_\infty + hL\sum_{j=0}^{r+nm-1}|(A_h^{-1}B_h)_{ij}||e_j|$$

$$\le \|e\|_\infty + hM'\sum_{j=0}^{r+nm-1}|e_j| \text{ for some } M' \text{ by (4.4.1) ,}$$

$\le \|e\|_\infty + hM'(r + nm)\|e\|_\infty$, for all i , where $\nu \equiv i - r + 1 \pmod m$ and $1 \le \nu \le m$. Now $(r + n\gamma_m - 1)h = T$ (2.1.3) , so there exist C" and M" such that $C"h^p \le (1 + M")\|e\|_\infty$, and hence we obtain the result , taking $C_L = C"/(1 + M")$.

Let us now see how this result applies to various examples .

Example 1.3.1 .

$$[\Delta_h y(t) - hA_h^{-1}B_h\Delta_h y'(t) - A_h^{-1}g\,]_i = y(t_i) - \tilde{y}_o - h[A_h^{-1}B_h\Delta_h y'(t)]_i$$

$$= \int_0^{t_i} y'(\eta)\,d\eta + O(h^4) - h[A_h^{-1}B_h\Delta_h y'(t)]_i ,$$

assuming a starting value $\tilde{y}_o$ accurate of order four .

$h[A_h^{-1}B_h\Delta_h y'(t)]_i$ is then a quadrature formula approximating $\int_0^{t_i} y'(\eta)\,d\eta$.

Let us now determine the form of $A_h^{-1}B_h$.

$$A_h = \begin{pmatrix} 1 & & & & & \\ 0 & 1 & & & & \\ -1 & 0 & 1 & & & \\ 0 & 0 & -1 & 1 & & \\ 0 & 0 & -1 & 0 & 1 & \\ 0 & 0 & 0 & 0 & -1 & 1 \\ & & \cdot & \cdot & \cdot & \end{pmatrix}$$, and by direct computation

$$A_h^{-1} = \begin{pmatrix} 1 & & & & & \\ 0 & 1 & & & & \\ 1 & 0 & 1 & & & \\ 1 & 0 & 1 & 1 & & \\ 1 & 0 & 1 & 0 & 1 & \\ 1 & 0 & 1 & 0 & 1 & 1 \\ & & \cdot & \cdot & \cdot & \end{pmatrix}$$, so $\max_{i,j} \left|(A_h^{-1})_{ij}\right| = 1$ and we have zero stability .

$$B_h = \begin{pmatrix} 0 & & & & & \\ 0 & 0 & & & & \\ 1/3 & 4/3 & 1/3 & & & \\ 0 & -1/12 & 8/12 & 5/12 & & \\ 0 & 0 & 1/3 & 4/3 & 1/3 & \\ 0 & 0 & 0 & -1/12 & 8/12 & 5/12 \\ & & & \cdot & \cdot & \cdot \end{pmatrix}$$, so

$$A_h^{-1}B_h = \begin{pmatrix} 0 & & & & & & & & \\ 0 & 0 & & & & & & & \\ \frac{1}{3} & \frac{4}{3} & \frac{1}{3} & & & & & & \\ \frac{1}{3} & \frac{4}{3} - \frac{1}{12} & \frac{1}{3} + \frac{8}{12} & \frac{5}{12} & & & & & \\ \frac{1}{3} & \frac{4}{3} & \frac{2}{3} & \frac{4}{3} & \frac{1}{3} & & & & \\ \frac{1}{3} & \frac{4}{3} & \frac{2}{3} & \frac{4}{3} - \frac{1}{12} & \frac{1}{3} + \frac{8}{12} & \frac{5}{12} & & & \\ \frac{1}{3} & \frac{4}{3} & \frac{2}{3} & \frac{4}{3} & \frac{2}{3} & \frac{4}{3} & \frac{1}{3} & & \\ \frac{1}{3} & \frac{4}{3} & \frac{2}{3} & \frac{4}{3} & \frac{2}{3} & \frac{4}{3} - \frac{1}{12} & \frac{1}{3} + \frac{8}{12} & \frac{5}{12} & \\ & & & & \cdot & \cdot & \cdot & & \end{pmatrix}$$

and we see that when i is even we have Simpson's rule applied globally , which has order four accuracy , and when i is odd we have again a global Simpson's rule with a local application of the quadrature rule

$$\int_{x-h}^{x} y'(\eta)d\eta \simeq h[\tfrac{5}{12}y'(x) + \tfrac{8}{12}y'(x-h) - \tfrac{1}{12}y'(x-2h)] ,$$

which has local accuracy of order four . Hence we have optimal consistency of order four , which implies convergence of order four , whereas consistency is only of order three .

Example 2.2.1 .

$r = 1$ and $\theta^*_{2s+1} = \frac{1}{2}h^2y''(t_{2s+1}) + O(h^3)$,

$\theta^*_{2s+2} = -\frac{1}{2}h^2y''(t_{2s+1}) + O(h^3)$, so we have order one consistency .

$$A_h = \begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ 0 & -1 & 1 & \\ 0 & 0 & -1 & 1 \\ & & \cdots & \end{pmatrix} ,$$ so by direct computation

$$A_h^{-1} = \begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ 1 & 1 & 1 & \\ 1 & 1 & 1 & 1 \\ & & \cdots & \end{pmatrix} ,$$ and $\max_{i,j}|(A_h^{-1})_{ij}| = 1$, giving zero stability .

Therefore $[\Delta_h y(t) - hA_h^{-1}B_h\Delta_h y'(t) - A_h^{-1}g]_i = y(t_i) - \tilde{y}_o - h[A_h^{-1}B_h\Delta_h y'(t)]_i$

$$= \int_0^{t_i} y'(\eta)d\eta + O(h^2) - h[A_h^{-1}B_h\Delta_h y'(t)]_i ,$$

assuming a starting value accurate of order two .

$$B_h = \begin{pmatrix} 0 & & & & \\ 1 & 0 & & & \\ 0 & 0 & 1 & & \\ 0 & 0 & 1 & 0 & \\ 0 & 0 & 0 & 0 & 1 \\ & & \cdots & & \end{pmatrix} , \text{ so } A_h^{-1}B_h = \begin{pmatrix} 0 & & & & & & \\ 1 & 0 & & & & & \\ 1 & 0 & 1 & & & & \\ 1 & 0 & 2 & 0 & & & \\ 1 & 0 & 2 & 0 & 1 & & \\ 1 & 0 & 2 & 0 & 2 & 0 & \\ 1 & 0 & 2 & 0 & 2 & 0 & 1 \\ & & & \cdots & & & \end{pmatrix} ,$$

and $h[A_h^{-1}B_h\Delta_h y'(t)]_i$ is the trapezoidal quadrature approximation to $\int_0^{t_i} y'(\eta)d\eta$, which has global accuracy of order two . Hence we have optimal consistency of order two , which implies convergence of order two .

Example 3.2.1 .

Taking the array (3.2.1) , $A_h = \begin{pmatrix} 1 & & & & \\ -1 & 1 & & & \\ -1 & 0 & 1 & & \\ 0 & 0 & -1 & 1 & \\ 0 & 0 & -1 & 0 & 1 \\ & & \cdots & & \end{pmatrix}$, so

$A_h^{-1} = \begin{pmatrix} 1 & & & & \\ 1 & 1 & & & \\ 1 & 0 & 1 & & \\ 1 & 0 & 1 & 1 & \\ 1 & 0 & 1 & 0 & 1 \\ & & \cdots & & \end{pmatrix}$ and again $\max_{i,j}\left|(A_h^{-1})_{ij}\right| = 1$, giving zero stability .

$$B_h = \begin{pmatrix} 0 & & & & \\ \frac{5}{12} & \frac{8}{12} & \frac{-1}{12} & & \\ \frac{1}{3} & \frac{4}{3} & \frac{1}{3} & & \\ 0 & 0 & \frac{5}{12} & \frac{2}{3} & \frac{-1}{12} \\ 0 & 0 & \frac{1}{3} & \frac{4}{3} & \frac{1}{3} \\ & & & \cdots & \end{pmatrix} , \text{ so } A_h^{-1}B_h = \begin{pmatrix} 0 & & & & \\ \frac{5}{12} & \frac{8}{12} & \frac{-1}{12} & & \\ \frac{1}{3} & \frac{4}{3} & \frac{1}{3} & & \\ \frac{1}{3} & \frac{4}{3} & \frac{1}{3}+\frac{5}{12} & \frac{2}{3} & \frac{-1}{12} \\ \frac{1}{3} & \frac{4}{3} & \frac{2}{3} & \frac{4}{3} & \frac{1}{3} \\ & & & \cdots & \end{pmatrix} .$$

$[\Delta_h y(t) - hA_h^{-1}B_h\Delta_h y'(t) - A_h^{-1}g]_i = y(t_i) - \tilde{y}_0 - h[A_h^{-1}B_h\Delta_h y'(t)]_i$, and again $h[A_h^{-1}B_h\Delta_h y'(t)]_i$ is a quadrature formula approximating $\int_0^{t_i} y'(\eta)d\eta$.

For i even we have a global application of Simpson's rule and for i odd we have the same quadrature formula together with a local application of $\int_{x-h}^{x} y'(\eta)d\eta \simeq h[\frac{5}{12}y'(x-h) + \frac{8}{12}y'(x) - \frac{1}{12}y'(x+h)]$, which has fourth order accuracy locally . Thus we have optimal consistency and hence convergence of order four , whereas we saw previously that consistency is only of order three .

5.2 The form of A_h^{-1} .

We may obtain further convergence results by examining in greater detail the form of A_h^{-1} . With A_h having the form (2.1.1) we define the nm x nm matrix D_h =

$$\begin{pmatrix} A_\gamma & & & & \\ A_{\gamma-1} & A_\gamma & & & \\ \vdots & & \ddots & & \\ A_1 & \cdots & & A_\gamma & \\ A_0 & \cdots & & A_\gamma & \\ & A_0 & \cdots & & A_\gamma \\ & & \ddots & & \ddots \\ & & & A_0 & \cdots & A_\gamma \end{pmatrix} .$$

Then $A_h = \begin{pmatrix} I_r & 0 \\ E_h & D_h \end{pmatrix}$ for some matrix E_h . By equating $A_h A_h^{-1}$ to I_{r+nm} we see that

$A_h^{-1} = \begin{pmatrix} I_r & 0 \\ -D_h^{-1}E_h & D_h^{-1} \end{pmatrix}$. If F_h is the nm x nm matrix

$$\begin{pmatrix} A_\gamma & & \\ & A_\gamma & 0 \\ & & \ddots \\ 0 & & A_\gamma \end{pmatrix} ,$$

then $D_h = F_h G_h^*$,

where $G_h^* =$

$$\begin{pmatrix} I_m & & & & & & \\ G^*_{\tau-1} & I_m & & & & & \\ \vdots & \ddots & \ddots & & & & \\ G^*_1 & \cdots & G^*_{\tau-1} & I_m & & & \\ G^*_o & & \cdots & G^*_{\tau-1} & I_m & & \\ & G^*_o & & \cdots & G^*_{\tau-1} & I_m & \\ & & \ddots & & \ddots & \ddots & \\ & & & G^*_o & \cdots & G^*_{\tau-1} & I_m \end{pmatrix}$$

and $G^*_\theta = A_\tau^{-1} A_\theta$ for $\theta = \alpha(1)\tau-1$.

$$D_h^{-1} = G_h^{*-1} F_h^{-1} \text{ and } F_h^{-1} = \begin{pmatrix} A_\tau^{-1} & & & \\ & A_\tau^{-1} & & \\ & & \ddots & \\ & & & A_\tau^{-1} \end{pmatrix},$$

so it remains for us to consider the inversion of G_h^* .

Lemma 5.2.1 .

Let W be an $(1+2)m \times (1+2)m$ matrix made up of $m \times m$ blocks

$X_o, X_1, \ldots, X_1 : W =$

$$\begin{pmatrix} I_m & & & & & \\ X_o & I_m & & & & \\ X_1 & X_o & I_m & & & \\ X_2 & X_1 & X_o & I_m & & \\ \vdots & & \ddots & \ddots & \ddots & \\ X_1 & & \cdots & & X_o & I_m \end{pmatrix} .$$

Let $W^{-1} =$

$$\begin{pmatrix} I_m & & & & & \\ H_o^{(0)} & I_m & & & & \\ H_o^{(1)} & H_1^{(1)} & I_m & & & \\ H_o^{(2)} & H_1^{(2)} & H_2^{(2)} & I_m & & \\ \vdots & & & \ddots & \ddots & \\ H_o^{(1)} & & \cdots & & H_1^{(1)} & I_m \end{pmatrix} .$$

Then $H_{\sigma}^{(s)} = H_{\sigma+1}^{(s+1)}$ for $\sigma = 0(1)s$, $s = 0(1)l-1$. (*)

Proof .

We shall use induction on s . For s = 0 we must prove that $H_o^{(0)} = H_1^{(1)}$. In performing the multiplication $WW^{-1} = I_{(l+2)m}$ we have $X_o + H_o^{(0)} = 0_m$ and $X_o + H_1^{(1)} = 0_m$, giving $H_o^{(0)} = H_1^{(1)} = -X_o$.

Let us assume now that $H_{\sigma}^{(k)} = H_{\sigma+1}^{(k+1)}$ for $\sigma = 0(1)k$, $k = 0(1)s-1$. Then in performing the multiplication $WW^{-1} = I_{(l+2)m}$ and considering rows $(s+1)m$ to $(s+2)m - 1$, we obtain

$$X_{s-k} + \sum_{q=0}^{s-k-1} X_q H_k^{(s-1-q)} + H_k^{(s)} = 0_m \ , \ k = 0(1)s \quad (5.2.1) \ .$$

Considering rows $(s+2)m$ to $(s+3)m - 1$ we obtain

$$X_{s+1-k} + \sum_{q=0}^{s-k} X_q H_k^{(s-q)} + H_k^{(s+1)} = 0_m \ , \ k = 0(1)s+1 \quad (5.2.2) \ .$$

Taking all the equations in (5.2.2) except the case k = 0 and putting $l = k-1$,

$$X_{s-l} + \sum_{q=0}^{s-l-1} X_q H_{l+1}^{(s-q)} + H_{l+1}^{(s+1)} = 0_m \text{ for } l = 0(1)s \quad (5.2.3) \ .$$

Then combining (5.2.1) and (5.2.3) and recalling that by assumption $H_{\sigma}^{(k)} = H_{\sigma+1}^{(k+1)}$ for $\sigma = 1(1)k$, $k = 0(1)s-1$, we obtain the result $H_{\sigma}^{(s)} = H_{\sigma+1}^{(s+1)}$ for $\sigma = 0(1)s$.

Corollary 5.2.2 .

G_h^{*-1} has the form

$$\begin{pmatrix} I_m & & & & & \\ P_o & I_m & & & & \\ P_1 & P_o & I_m & & & \\ P_2 & P_1 & P_o & I_m & & \\ \vdots & & & \ddots & \ddots & \\ P_{n-2} & & \cdots & & P_o & I_m \end{pmatrix}$$

(*) This lemma is a generalisation of the result for isoclinal matrices given in theorem 2.2 of Leroy J. Derr , Pac. J. of Maths., 37 , no. 1 , 41-43 (1971) .

This is an immediate consequence of lemma 5.2.1 .

Performing the multiplication $G^*_h G^{*-1}_h = I_{nm}$ we see that

$$
\begin{aligned}
&G^*_{\tau-1} + P_o && = 0_m\\
&G^*_{\tau-2} + G^*_{\tau-1}P_o + P_1 && = 0_m\\
&G^*_{\tau-3} + G^*_{\tau-2}P_o + G^*_{\tau-1}P_1 + P_2 && = 0_m\\
&\quad \cdots\\
&G^*_o + G^*_1 P_o + G^*_2 P_1 + \cdots + G^*_{\tau-1}P_{\tau-2} + P_{\tau-1} && = 0_m\\
&G^*_o P_o + G^*_1 P_1 + \cdots + G^*_{\tau-2}P_{\tau-2} + G^*_{\tau-1}P_{\tau-1} + P_\tau && = 0_m\\
&G^*_o P_1 + \cdots + G^*_{\tau-1}P_\tau + P_{\tau+1} && = 0_m
\end{aligned}
$$

and so on up to

$$G^*_o P_{n-\tau-2} + \cdots + G^*_{\tau-1}P_{n-3} + P_{n-2} = 0_m .$$

Therefore for each $q = 0(1)n-2$,

$$P_q = \sum_{l=1}^{u_q} \beta_l \prod_{\theta=0}^{\tau-1} G^{*\alpha^l_\theta}_\theta \qquad (5.2.4)$$

for some positive integers u_q, α^l_θ, $\theta = 0(1)\tau-1$, $l = 1(1)u_q$ and real scalars $\beta_1, \beta_2, \ldots, \beta_{u_q}$.

$$D_h^{-1} = G^{*-1}_h F_h^{-1} = \begin{pmatrix}
A_\tau^{-1} & & & \\
P_o A_\tau^{-1} & A_\tau^{-1} & & \\
P_1 A_\tau^{-1} & P_o A_\tau^{-1} & A_\tau^{-1} & \\
\vdots & & \ddots & \\
P_{n-2} A_\tau^{-1} & \cdots & P_o A_\tau^{-1} & A_\tau^{-1}
\end{pmatrix}$$

5.3 Improved Convergence .

Having determined the form of A_h^{-1} we are now in a position to formulate a criterion analogous to that of Albrecht (2, theorem 4.3) for the order of convergence to exceed that of consistency by one . This situation we refer to as improved convergence .

Theorem 5.3.1 .

Suppose we have an m-block method which is consistent of order p and zero stable , and starting values are accurate of order p+1 . Then we have exact order of convergence p+1 if optimal consistency is of order p+1 .

Proof .

If the method is optimally consistent of order p+1 we have exact order of convergence p+1 , due to theorem 5.1.2 .

If the method converges with exact order p+1 , assume that optimal consistency is of order p^* . Therefore by theorem 5.1.2 the exact order of convergence is p^* , so $p^* = p+1$.

Definition 5.3.2 .

The TRUNCATION ERROR VECTOR CONSTANTS $c^{(1)}, c^{(2)}, \ldots$ are defined for $y(t) \in C^{(p+q)}[0,T]$ by $c^{(q)} = (c_1^{(q)}, c_2^{(q)}, \ldots, c_m^{(q)})^T$, $q \geq 1$ and

$$\theta^*_{r+sm+\nu-1} = \sum_{q=1}^{\infty} c_\nu^{(q)} [\Delta_h y^{(p+q)}(t)]_{r+sm} h^{p+q} + O(h^{p+q+1}) \quad , \quad \nu = 1(1)m \ , \ s = 0(1)n-1$$

Theorem 5.3.3 .

Suppose we have an m-block method which is consistent of order p and zero stable and starting values are accurate of order p+1 . Then we have exact order of convergence p+1 if $A_\theta A_\tau^{-1} c^{(1)} = 0$, the zero vector , for $\theta = 0(1)\tau-1$.

Proof .

By virtue of theorem 5.3.1 we must prove optimal consistency of order p+1 . $A_h^{-1}\theta^* =$

$$\begin{pmatrix} I_r & 0 \\ \hline & \\ -D_h^{-1}E_h & D_h^{-1} \end{pmatrix} \begin{pmatrix} \theta^* \\ \vdots^0 \\ \vdots \\ \theta^*_{r-1} \\ \theta^* \\ \vdots^r \\ \vdots \\ \theta^*_{r+nm-1} \end{pmatrix} ,$$

so $[A_h^{-1}\theta^*]_{r+sm+\nu-1}$ is the $(sm+\nu-1)^{th.}$ component of

$$-D_h^{-1}E_h\begin{pmatrix}\theta^*_0\\ \vdots\\ \theta^*_{r-1}\end{pmatrix} + D_h^{-1}\begin{pmatrix}\theta^*_r\\ \vdots\\ \theta^*_{r+nm-1}\end{pmatrix} \quad .$$

Because of zero stability the elements of D_h^{-1} are bounded , therefore

$$-D_h^{-1}E_h\begin{pmatrix}\theta^*_0\\ \vdots\\ \theta^*_{r-1}\end{pmatrix} = -D_h^{-1}E_h\begin{pmatrix}C'_0\\ \vdots\\ C'_{r-1}\end{pmatrix} h^{p+1} \overset{+O(h^{p+2})}{\wedge} = \begin{pmatrix}C''_r\\ \vdots\\ C''_{r+nm-1}\end{pmatrix} h^{p+1} + O(h^{p+2})$$

for some constants $C''_0, \ldots, C''_{r+nm-1}$. Accordingly it remains to consider

$$D_h^{-1}\begin{pmatrix}\theta^*_r\\ \vdots\\ \theta^*_{r+nm-1}\end{pmatrix} \quad , \text{ whose elements } sm \text{ to } sm+m-1 \text{ are of the form}$$

$$A_\tau^{-1}c^{(1)}[\Delta_h y^{(p+1)}(t)]_{r+sm}h^{p+1} + \sum_{\sigma=0}^{s-1} P_\sigma A_\tau^{-1}c^{(1)}[\Delta_h y^{(p+1)}(t)]_{r+(s-\sigma-1)m}h^{p+1} + O(h^{p+2}) \ .$$

Since the method is zero stable , the elements of A_τ^{-1} are bounded , so $A_\tau^{-1}c^{(1)}[\Delta_h y^{(p+1)}(t)]_{r+sm}h^{p+1} = C'''h^{p+1}$ for some vector C''' . Therefore we have order p+1 convergence if and only if ..

$$\left\| \sum_{\sigma=0}^{n-2} P_\sigma A_\tau^{-1}c^{(1)}[\Delta_h y^{(p+1)}(t)]_{r+(n-\sigma-2)m} \right\|_\infty \leq C$$

for some C , for all $n \geq 2$ (5.3.1) .

Now for all σ , $P_\sigma A_\tau^{-1}c^{(1)} = \sum_{l=1}^{u_\sigma} \beta_l \prod_{\theta=0}^{\tau-1} G_\theta^{*\alpha_{\theta.}^{l}} A_\tau^{-1}c^{(1)}$ by (5.2.4)

$$= \sum_{l=1}^{u_\sigma} \beta_l \prod_{\theta=0}^{\tau-1} (A_\tau^{-1}A_\theta)^{\alpha_{\theta.}^{l}} A_\tau^{-1}c^{(1)}$$

$$= 0 \text{ if } A_\theta A_\tau^{-1}c^{(1)} = 0 \text{ for } \theta = 0(1)\tau-1 \ .$$

Hence (5.3.1) holds if $A_\theta A_\tau^{-1}c^{(1)} = 0$ for $\theta = 0(1)\tau-1$.

We now examine some applications of this result .

Example 1.3.1 .

$$A_o A_1^{-1} = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix} \text{ and } c^{(1)} = \begin{bmatrix} 0 \\ -1/24 \end{bmatrix} \text{ , so } A_o A_1^{-1} c^{(1)} = 0 \text{ ,}$$

implying fourth order convergence .

Example 2.2.1 .

$$A_o A_1^{-1} = \begin{bmatrix} -1 & -1 \\ 0 & 0 \end{bmatrix} \text{ and } c^{(1)} = \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{bmatrix} \text{ , so } A_o A_1^{-1} c^{(1)} = 0 \text{ , implying}$$

second order convergence .

Example 2.2.2 .

The companion matrices (1.3.3) are

$$G_1 = \begin{bmatrix} 0 & 1 \\ -3/2 & 5/2 \end{bmatrix} \text{ , } G_2 = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \text{ , } G_3 = \begin{bmatrix} 0 & 1 \\ 6/5 & -1/5 \end{bmatrix} \text{ ,}$$

giving $G = \begin{bmatrix} -3/2 & 5/2 \\ -3/2 & 5/2 \end{bmatrix}$, which has eigenvalues 0 and 1 , and so we have zero stability . Consistency is of order three and

$$c^{(1)} = \begin{bmatrix} -5/48 \\ -1/24 \\ 1/120 \end{bmatrix} \text{ . } A_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & -1/5 & 1 \end{bmatrix} \text{ , so } A_o A_1^{-1} = \begin{bmatrix} -1 & 2 & -5/2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \text{ ,}$$

and hence $A_o A_1^{-1} c^{(1)} = 0$, which implies fourth order convergence .

Example 2.2.3 .

The companion matrices (1.3.3) are

$$G_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \text{ , } G_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 2 \end{bmatrix} \text{ , } G_3 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 3/2 & -1/2 \end{bmatrix} \text{ ,}$$

$G = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & -1 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \end{bmatrix}$, which has eigenvalues $0, 0, -\frac{1}{2}, 1$ and so we have zero stability . Consistency is of order three

and $c^{(1)} = (2,-2,-2)^T$. $A_0A_2^{-1} = \begin{pmatrix} -\frac{1}{2} & \frac{1}{2} & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ and $A_1A_2^{-1} = \begin{pmatrix} 0 & 0 & 0 \\ \frac{1}{2} & -\frac{1}{2} & 1 \\ 0 & 0 & 0 \end{pmatrix}$,

so $A_0A_2^{-1}c^{(1)} = A_1A_2^{-1}c^{(1)} = 0$ and we have fourth order convergence .

Example 5.3.4 .

Consider the Runge-Kutta method quoted in (15, p.118) .

$$y_i - y_{i-1} = hk_2 ,$$
$$k_1 = f(t_{i-1}, y_{i-1}) ,$$
$$k_2 = f(t_{i-1} + \tfrac{1}{2}h, y_{i-1} + \tfrac{1}{2}hk_1) .$$

This is a two-block explicit method with array

$$\left(\begin{array}{cccc|cccc|c} 0 & -1 & 1 & 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & -1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \end{array}\right) .$$

Any Runge-Kutta method is zero stable , as we saw in section 4.3 . This example is consistent of order one and $c^{(1)} = \begin{pmatrix} 1/8 \\ 0 \end{pmatrix}$. $A_0A_1^{-1} = \begin{pmatrix} 0 & -1 \\ 0 & -1 \end{pmatrix}$,

so $A_0A_1^{-1}c^{(1)} = 0$, giving second order convergence .

Example 5.3.5 .

Another Runge-Kutta method (15, p.119) is

$$y_i - y_{i-1} = h(\tfrac{1}{2}k_1 + \tfrac{1}{2}k_2) ,$$
$$k_1 = f(t_{i-1}, y_{i-1}) ,$$
$$k_2 = f(t_{i-1} + h, y_{i-1} + hk_1) .$$

This is a 2-block explicit method with array

$$\left(\begin{array}{cccc|cccc|c} 0 & -1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & -1 & 0 & 1 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 1 \end{array}\right) .$$

Consistency is of order one and $c^{(1)} = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}$. Again $A_0A_1^{-1} = \begin{pmatrix} 0 & -1 \\ 0 & -1 \end{pmatrix}$,

so $A_0A_1^{-1}c^{(1)} = 0$, giving convergence of order two .

Example 5.3.6 .

For a predictor-corrector method in PECE mode ,

$A_\theta = \begin{bmatrix} 0 & \alpha^*_\theta \\ 0 & \alpha_\theta \end{bmatrix}$, $\theta = 0(1)k-1$ and $A_k = \begin{bmatrix} \alpha^*_k & 0 \\ 0 & \alpha_k \end{bmatrix}$.

Suppose we take a predictor consistent of order p-1 with error constant C* and a corrector of order p . Then the predictor-corrector scheme as a 2-block method is consistent of order p-1 and $c^{(1)} = \begin{bmatrix} C^* \\ 0 \end{bmatrix}$.

For each $\theta = 0(1)k-1$, $A_\theta A_k^{-1} = \begin{bmatrix} 0 & \alpha^*_\theta \\ 0 & \alpha_\theta \end{bmatrix} \begin{bmatrix} 1/\alpha^*_k & 0 \\ 0 & 1/\alpha_k \end{bmatrix} = \begin{bmatrix} 0 & \alpha^*_\theta/\alpha_k \\ 0 & \alpha_\theta/\alpha_k \end{bmatrix}$,

so $A_\theta A_k^{-1} c^{(1)} = 0$, and since any predictor-corrector method in $P(EC)^\mu(E)$ mode is zero stable if and only if the corrector is when considered as a linear multistep method , we have convergence of order p in both the predictor and the corrector , provided we take a zero stable corrector .

In PEC mode A_h is exactly as in PECE mode , and so we have the same result .

Example 3.2.1 .

Consistency is of order three and $c^{(1)} = \begin{bmatrix} 1/24 \\ 0 \end{bmatrix}$.

$A_o A_1^{-1} = \begin{bmatrix} 0 & -1 \\ 0 & -1 \end{bmatrix}$, so $A_o A_1^{-1} c^{(1)} = 0$, giving order four convergence .

Example 5.3.7 .

Consider the following generalised Nordsieck procedure (2, p.248) .

$$\bar{y}_{i-1} = \frac{1}{2}y_{i-1} + \frac{1}{2}\bar{y}_{i-2} + \frac{h}{24}(-f_i + 8f_{i-1} + 5f_{i-2}) ,$$

$$y_i = \frac{1}{2}y_{i-1} + \frac{1}{2}\bar{y}_{i-2} + \frac{h}{8}(3f_i + 8f_{i-1} + f_{i-2}) .$$

This is a 2-block implicit method with array

$$\left(\begin{array}{cccccc|cccccc|c} 0 & 0 & -1/2 & -1/2 & 1 & 0 & 0 & 5/24 & 0 & 1/3 & 0 & -1/24 & 0 \\ 0 & 0 & -1/2 & -1/2 & 0 & 1 & 0 & 1/8 & 0 & 1 & 0 & 3/8 & 1 \end{array}\right) .$$

Consistency is of order three and $c^{(1)} = \begin{pmatrix} 1/48 \\ -1/48 \end{pmatrix}$.

$$A_0A_2^{-1} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} , \text{ so } A_0A_2^{-1}c^{(1)} = 0 .$$

$$A_1A_2^{-1} = \begin{pmatrix} -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} \end{pmatrix} , \text{ so } A_1A_2^{-1}c^{(1)} = 0 .$$

$$G_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \text{ and } G_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} , \text{ so } G = \begin{pmatrix} 0 & 0 & 1 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} ,$$

which has eigenvalues 0,0,1 . Hence we have zero stability and fourth order convergence .

Example 5.3.8 .

Consider a $P_{2/3}EP_HE$ hybrid method (5, p.658) :

$$\begin{aligned} y_{i-1/3} - \tfrac{5}{9}y_{i-1} - \tfrac{4}{9}y_{i-2} &= \tfrac{10h}{9}f_{i-1} , \\ y_i - y_{i-1} &= \tfrac{h}{4}(3f_{i-1/3} + f_{i-1}) , \end{aligned}$$

which is a two block explicit method with array

$$\left(\begin{array}{cccccc|cccccc|c} 0 & -4/9 & 0 & -5/9 & 1 & 0 & 0 & 0 & 0 & 10/9 & 0 & 0 & 2/3 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 1/4 & 3/4 & 0 & 1 \end{array}\right) .$$

For zero stability the companion matrices are

$$G_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 4/9 & 0 & 5/9 \end{pmatrix} \text{ and } G_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} , \text{ so } G = \begin{pmatrix} 0 & 0 & 1 \\ 4/9 & 0 & 5/9 \\ 0 & 0 & 1 \end{pmatrix} ,$$

which has eigenvalues 0,0,1 , so we have zero stability .

Consistency is of order two and $c^{(1)} = \begin{pmatrix} 10/81 \\ 0 \end{pmatrix}$.

$$A_o A_2^{-1} = \begin{pmatrix} 0 & -4/9 \\ 0 & 0 \end{pmatrix} \quad , \text{ so } A_o A_2^{-1} c^{(1)} = 0$$

$$A_1 A_2^{-1} = \begin{pmatrix} 0 & -5/9 \\ 0 & -1 \end{pmatrix} \quad , \text{ so } A_1 A_2^{-1} c^{(1)} = 0 \text{ , and we have third order convergence .}$$

Example 5.3.9 .

Cooper also quotes the following example (5) .

$$y_{i-1/3} + 3y_{i-1} - 4y_{i-4/3} = 2f_{i-1} \text{ ,}$$

$$y_i + 3y_{i-1} - 4y_{i-4/3} = \frac{h}{12}(5f_{i-1/3} + 23f_{i-1}) \text{ ,}$$

which is a 2-block explicit method with array

$$\left(\begin{array}{cccc|cccc|c} -4 & 3 & 1 & 0 & 0 & 2 & 0 & 0 & 2/3 \\ -4 & 3 & 0 & 1 & 0 & 23/12 & 5/12 & 0 & 1 \end{array}\right) \text{ .}$$

For zero stability the companion matrices are

$$G_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 4 & -3 \end{pmatrix} \quad \text{and } G_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 4 & -3 & 0 \end{pmatrix} \text{ , so } G = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 4 & -3 \\ 0 & 4 & -3 \end{pmatrix} \text{ ,}$$

which has eigenvalues 0,0,1 , so we have zero stability .

Consistency is of order two and $c^{(1)} = \frac{2}{81}\begin{pmatrix} 3 \\ 4 \end{pmatrix}$. $A_o A_1^{-1} = \begin{pmatrix} -4 & 3 \\ -4 & 3 \end{pmatrix}$,

so $A_o A_1^{-1} c^{(1)} = 0$, giving order three convergence .

Example 5.3.10 .

Another method quoted in (5) is

$$y_{i-2/3} - y_{i-1} + \frac{2}{3}y^*_{i-1} - y_{i-5/3} + \frac{1}{3}y_{i-2} = \frac{2h}{3}f_{i-1} \text{ ,}$$

$$y_{i-1/2} - y_{i-1} = \frac{h}{8}(f_{i-1} + 3f_{i-2/3}) \text{ ,}$$

$$y^*_i - y_{i-1} = h(\frac{1}{2}f_{i-1} - \frac{3}{2}f_{i-2/3} + 2f_{i-1/2}) \text{ ,}$$

$$y_i - y_{i-1} = \frac{h}{6}(f_{i-1} + 4f_{i-1/2} + f^*_i) \text{ ,}$$

which is a 4-block explicit method with array

$$\left(\begin{array}{cccccccccccc|c}
0 & 0 & 0 & 1/3 & -1 & 0 & 2/3 & -1 & 1 & 0 & 0 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & \\
\hline
 & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2/3 & 0 & 0 & 0 & 0 & 1/3 \\
 & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/8 & 3/8 & 0 & 0 & 0 & 1/2 \\
 & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & -3/2 & 2 & 0 & 0 & 1 \\
 & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/6 & 0 & 2/3 & 1/6 & 0 & 1
\end{array}\right) .$$

$$G_1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ -1/3 & 1 & 0 & -2/3 & 1 \end{pmatrix} , \quad G_2 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} ,$$

$$G_3 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} , \quad G_4 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} , \text{ giving}$$

$$G = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ -1/3 & 1 & 0 & -2/3 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} , \text{ which is idempotent and has eigenvalues } 0,0,0,1,1 . \text{ Since G is idempotent , the eigenvalue 1 must have linear elementary divisors . Consistency is of order three and } c^{(1)} = \frac{1}{2^7 . 3^4} \begin{pmatrix} 64 \\ 3 \\ 96 \\ 0 \end{pmatrix} .$$

$$A_o A_2^{-1} = \begin{pmatrix} 0 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} , \text{ so } A_o A_2^{-1} c^{(1)} = 0 .$$

$$A_1A_2^{-1} = \begin{pmatrix} -1 & 0 & 2/3 & -1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \text{ so } A_1A_2^{-1}c^{(1)} = 0 \text{ , giving order four convergence .}$$

Example 5.3.11 .

Albrecht (1) quotes the method

$$\hat{y}_{i-1} = \frac{11}{30}y_{i-1} + \frac{19}{30}\hat{y}_{i-2} + \frac{19h}{720}(-f_i + 13f_{i-1} + 13f_{i-2} - f_{i-3}) ,$$

$$y_i = \frac{11}{30}y_{i-1} + \frac{19}{30}\hat{y}_{i-2} + \frac{h}{720}(251f_i + 817f_{i-1} + 97f_{i-2} + 11f_{i-3}) .$$

This is a 2-block implicit method with array

$$\left(\begin{array}{cccccccc|cccccc} 0 & 0 & 0 & 0 & -19/30 & -11/30 & 1 & 0 & 0 & -19/720 & 0 & 247/720 & 0 & 247/720 \\ 0 & 0 & 0 & 0 & -19/30 & -11/30 & 0 & 1 & 0 & 11/720 & 0 & 97/720 & 0 & 817/720 \\ & & & & & & & & 0 & -19/720 & 0 & 251/720 & 0 & \\ & & & & & & & & 0 & 251/720 & 1 & & & \end{array}\right) .$$

Consistency is of order four and $c^{(1)} = \frac{19}{180.120}\begin{pmatrix} 11 \\ -19 \end{pmatrix}$.

For zero stability the companion matrices are

$$G_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 19/30 & 11/30 \end{pmatrix} \text{ and } G_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 19/30 & 11/30 & 0 \end{pmatrix} ,$$

$$\text{so } G = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 19/30 & 11/30 \\ 0 & 19/30 & 11/30 \end{pmatrix} \text{ , which has eigenvalues 0,0,1 , so we have zero stability .}$$

$$A_0A_3^{-1} = A_1A_3^{-1} = 0_2 \text{ , so } A_0A_3^{-1}c^{(1)} = A_1A_3^{-1}c^{(1)} = 0 .$$

$$A_2A_3^{-1} = \begin{pmatrix} -19/30 & -11/30 \\ -19/30 & -11/30 \end{pmatrix} \text{ , so } A_2A_3^{-1}c^{(1)} = 0 \text{ and we have fifth order convergence .}$$

CHAPTER SIX .

OPTIMAL CONSISTENCY OF ORDERS $(p_1,p_2, \dots ,p_m)$.

6.1 Two-Sided Error Bounds .

In all the examples considered so far we have had optimal consistency of orders $p(1,1, \dots ,1)$ for some $p \in \mathbb{N}$. A more detailed analysis of convergence is needed however when we encounter methods whose optimal consistency is of orders $(p_1,p_2, \dots ,p_m)$ with not all the p_ν equal .

Definition 6.1.1 .

Let $a_1,a_2, \dots ,a_w$ be distinct positive integers : $1 \le a_1 < a_2 < \dots < a_w \le m$. Then we define the vector $e^{(a_1,a_2, \dots ,a_w)} \in \mathbb{R}^{r+nw}$ by $[e^{(a_1,a_2, \dots ,a_w)}]_i = e_i$, $i = 0(1)r-1$,

$$[e^{(a_1,a_2, \dots ,a_w)}]_{r+sw+q-1} = e_{r+sm+a_q-1} \quad , \quad q = 1(1)w \quad , \quad s = 0(1)n-1 \quad ,$$

and e is the global error vector as in definition 4.1.4 . Note that $e^{(1,2, \dots ,m)} = e$.

It is tempting to imagine that optimal consistency of orders $(p_1,p_2, \dots ,p_m)$ will imply $\|e^{(\nu)}\|_\infty \le C_U^{(\nu)} h^{p_\nu}$ for some $C_U^{(\nu)}$, $\nu = 1(1)m$, but this is not generally the case . The only thing we can say at this stage is that optimal consistency of orders $(p_1,p_2, \dots ,p_m)$ implies $C_L h^p \le \|e\|_\infty \le C_U h^p$, where $p = \min_{\nu=1(1)m} p_\nu$ (6.1.1) .

Our aim in this section is to obtain two-sided bounds of the form

$$C_L^{(a_1,a_2, \dots ,a_w)} h^{p^*} \le \|e^{(a_1,a_2, \dots ,a_w)}\|_\infty \le C_U^{(a_1,a_2, \dots ,a_w)} h^{p^*} \quad (6.1.2)$$

for some p^* , for some proper subsets $\{a_1,a_2, \dots ,a_w\}$ of $\{1,2, \dots ,m\}$, and thus obtain more information than that given by (6.1.1) . The analysis

we employ will retain the essential simplicity of the approach adopted in the preceding chapters , resorting only to optimal consistency and lemmas 4.4.3 and 4.4.5 .

Example 6.1.2 .

Let us consider a predictor-corrector scheme in $P(EC)^2E$ mode with predictor $y_i = y_{i-1} + hf_{i-1}$ and corrector $y_i = y_{i-1} + \frac{h}{12}(5f_i + 8f_{i-1} - f_{i-2})$.

This is a 3-block explicit method with array

$$\left(\begin{array}{ccccccccc|ccccccccc|c} 0&0&0&0&0&-1&1&0&0 & 0&0&0&0&0&0&0&0&0 & 1\\ 0&0&0&0&0&-1&0&1&0 & 0&0&-1/12&0&0&8/12&5/12&0&0 & 1\\ 0&0&0&0&0&-1&0&0&1 & 0&0&-1/12&0&0&8/12&0&5/12&0 & 1 \end{array}\right) .$$

By direct computation

$$A_h^{-1} = \left(\begin{array}{ccccccccccc} 1 & & & & & & & & & & \\ 0&1 & & & & & & & & & \\ 0&1&1 & & & & & & & & \\ 0&1&0&1 & & & & & & & \\ 0&1&0&0&1 & & & & & & \\ 0&1&0&0&1&1 & & & & & \\ 0&1&0&0&1&0&1 & & & & \\ 0&1&0&0&1&0&0&1 & & & \\ 0&1&0&0&1&0&0&1&1 & & \\ 0&1&0&0&1&0&0&1&1&1 & \\ & & & & & & & \cdots & & & \end{array}\right) .$$

Let us assume two starting values accurate of order three . Then $\theta_o^* = C_o' h^3$ and $\theta_1^* = C_1' h^3$.

$$\theta_{3s+2}^* = \frac{1}{2}h^2 y''(t_{s+2}) - \frac{1}{3}h^3 y'''(t_{s+2}) + \frac{1}{8}h^4 y^{iv}(t_{s+2}) + O(h^5) ,$$

$$\theta_{3s+1+\nu}^* = - \frac{1}{24}h^4 y^{iv}(t_{s+2}) + O(h^5) ,$$

$$\nu = 2,3 \ , \ s = 0(1)n-1 .$$

Hence by computing $A_h^{-1}\theta^*$ directly we obtain optimal consistency of orders $(2,3,3)$.

Example 6.1.3 .

Consider the 3-stage explicit Runge-Kutta method with array

$$\left(\begin{array}{cccccc|cccccc|c} 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 1/3 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 2/3 & 0 & 0 & 2/3 \\ 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 1/4 & 0 & 3/4 & 0 & 1 \end{array}\right) .$$

$$A_h^{-1} = \begin{pmatrix} 1 & & & & & & & \\ 1 & 1 & & & & & & \\ 1 & 0 & 1 & & & & & \\ 1 & 0 & 0 & 1 & & & & \\ 1 & 0 & 0 & 1 & 1 & & & \\ 1 & 0 & 0 & 1 & 0 & 1 & & \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ & & & & \cdots & & & \end{pmatrix} .$$

Let us assume a starting value accurate of order three . Then $\theta^*_o = C'_o h^3$,

$$\theta^*_{sm+1} = \frac{1}{18}h^2 y''(t_s) + \frac{1}{162}h^3 y'''(t_s) + \frac{1}{1944}h^4 y^{iv}(t_s) + O(h^5) ,$$

$$\theta^*_{sm+2} = \frac{1}{81}h^3 y'''(t_s) + \frac{1}{243}h^4 y^{iv}(t_s) + O(h^5) ,$$

$$\theta^*_{sm+3} = \frac{1}{216}h^4 y^{iv}(t_s) + O(h^5) ,$$

$$s = 0(1)n-1 .$$

Hence by computing $A_h^{-1}\theta^*$ directly we obtain optimal consistency of orders $(2,3,3)$.

6.2 A Special Class of Methods.

We shall be considering a class of methods characterised by A_h and B_h having certain forms , which will permit us to obtain bounds on the error of the form (6.1.2) .

Definition 6.2.1 .

A square matrix A is in the class $Q_{m,\alpha}$ for two positive integers m and α: $m \geq \alpha$ if

(i) A is lower triangular ;

(ii) if we denote the $(i,j)^{th.}$ element of A by A_{ij} , then there exist positive integers r,n,τ such that $A \in \mathbb{R}^{r+nm} \times \mathbb{R}^{r+nm}$ and for each

$\nu = 1(1)m-\alpha$, $s = 0(1)n-1$, $A_{i,r+sm+\nu-1} = 0$ for $i = r+sm+\nu(1)r+sm+m-1$, $i = r+\sigma m+q-1$, $q = 2(1)m$ and $\sigma = s+1(1)s+\tau$;

(iii) $A_{ii} \neq 0$ for all i ,

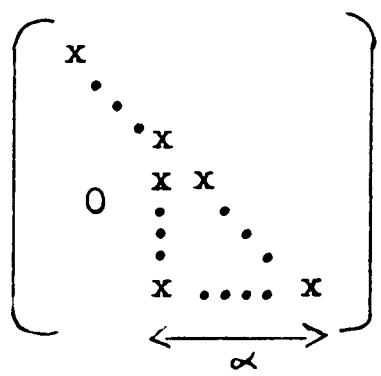
(iv) A is a band matrix : for each $\nu = 1(1)m$, $s = 0(1)n-1$, $A_{i,r+sm+\nu-1} = 0$ for $i \geq r+(s+\tau+1)m$ and for $j = 0(1)r-1$, $A_{ij} = 0$ for $i \geq r + \tau m$.

By way of illustration , when $m = 5$, $\alpha = 3$, $r = 2$, $n = 4$ and $\tau = 2$, $A \in Q_{5,3}$ has the form

x																			
x	x																		
x	x	x																	
x	x	0	x																
x	x	0	0	x															
x	x	0	0	x	x														
x	x	0	0	x	x	x													
x	x	x	x	x	x	x	x												
x	x	0	0	x	x	x	0	x											
x	x	0	0	x	x	x	0	0	x										
x	x	0	0	x	x	x	0	0	x	x									
x	x	0	0	x	x	x	0	0	x	x	x								
0	0	x	x	x	x	x	x	x	x	x	x	x							
0	0	0	0	x	x	x	0	0	x	x	x	0	x						
0	0	0	0	x	x	x	0	0	x	x	x	0	0	x					
0	0	0	0	x	x	x	0	0	x	x	x	0	0	x	x				
0	0	0	0	x	x	x	0	0	x	x	x	0	0	x	x	x			
0	0	0	0	0	0	0	x	x	x	x	x	x	x	x	x	x	x		
0	0	0	0	0	0	0	0	0	x	x	x	0	0	x	x	x	0	x	
0	0	0	0	0	0	0	0	0	x	x	x	0	0	x	x	x	0	0	x
0	0	0	0	0	0	0	0	0	x	x	x	0	0	x	x	x	0	0	x x
0	0	0	0	0	0	0	0	0	x	x	x	0	0	x	x	x	0	0	x x x

The zeroes represent elements which are necessarily zero , and all the elements on the diagonal are strictly non-zero .

We see that for an m-block method $A_h \in Q_{m,\alpha}$ if A_τ has the form



, all elements on the diagonal being strictly non-zero ,

and for $\theta = 0(1)\tau-1$ A_θ has the form $\begin{bmatrix} x & \dots & x & x & \dots & x \\ 0 & & & & & \\ & & \vdots & & \vdots & \\ & & x & \dots & x & \end{bmatrix}$ (α columns).

Definition 6.2.2 .

A square matrix B is in the class $Q^{(\beta)}_{m,\alpha}$ for three positive integers m,α,β : $\alpha,\beta \le m$ if

(i) it is lower triangular ;

(ii) there exist positive integers r,n such that $B \in \mathbb{R}^{r+nm} \times \mathbb{R}^{r+nm}$ and for each $\nu = 1(1)m-\alpha$, $s = 0(1)n-1$, $B_{i,r+sm+\nu-1} = 0$ for $i = r+sm+\nu-1+\beta(1)r+sm+m-1$, $i = r+\sigma m+q-1$, $q = 2(1)m$ and $\sigma = s+1(1)n-1$.

When $m = 5$, $\alpha = 3$, $\beta = 2$, $r = 2$ and $n = 4$, $B \in Q^{(2)}_{5,3}$ has the form

$$\left[\begin{array}{c|cc|ccccc|ccccccc|ccccccccc|ccccccccccc}
x & & & & & & & & & & & & & & & & & & & & & & & & & & & & & \\
x & x & & & & & & & & & & & & & & & & & & & & & & & & & & & & \\
\hline
x & x & x & & & & & & & & & & & & & & & & & & & & & & & & & & \\
x & x & x & x & & & & & & & & & & & & & & & & & & & & & & & & & \\
x & x & 0 & x & x & & & & & & & & & & & & & & & & & & & & & & & & \\
x & x & 0 & 0 & x & x & & & & & & & & & & & & & & & & & & & & & & & \\
x & x & 0 & 0 & x & x & x & & & & & & & & & & & & & & & & & & & & & & \\
\hline
x & x & x & x & x & x & x & x & & & & & & & & & & & & & & & & & & & & & \\
x & x & 0 & 0 & x & x & x & x & x & & & & & & & & & & & & & & & & & & & & \\
x & x & 0 & 0 & x & x & x & 0 & x & x & & & & & & & & & & & & & & & & & & & \\
x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & & & & & & & & & & & & & & & & & & \\
x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & & & & & & & & & & & & & & & & & \\
\hline
x & x & x & x & x & x & x & x & x & x & x & x & & & & & & & & & & & & & & & & & \\
x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & x & & & & & & & & & & & & & & & & \\
x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & x & & & & & & & & & & & & & & & \\
x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & 0 & x & & & & & & & & & & & & & & \\
x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & & & & & & & & & & & & & \\
\hline
x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & & & & & & & & & & & & \\
x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & x & & & & & & & & & & & \\
x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & x & x & & & & & & & & & \\
x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & & & & & & & & \\
x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & & & & & & & \\
\hline
x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & & & & & & \\
x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & x & x & x & x & x & x & x & & & & & \\
x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & x & x & x & 0 & x & x & & & & \\
x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & 0 & 0 & x & x & & & \\
x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & 0 & 0 & x & x & x & & \\
\hline
x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x & \\
x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & x & x & x & x & x & x & x & x & x & x & x & x \\
x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & x & x & x & 0 & x & x & x & 0 & x & x & \\
x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & 0 & 0 & x & x & 0 & 0 & x & x \\
x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & x & 0 & 0 & x & x & 0 & 0 & x & x & 0 & 0 & x & x
\end{array}\right] .$$

Lemma 6.2.3 .

Let U be any square $(l+1)x(l+1)$ matrix and suppose that there exist positive integers ρ,σ : $0 \le \rho,\sigma \le l$ such that for all integers i : $0 \le i \le l$, $u_{i\sigma} = u_{\rho\sigma}\delta_{\rho i}$, where $\delta_{\rho i}$ is

the Krönecker delta . Then $|\det U| = |u_{\rho\sigma}|\,|\det \tilde{U}|$, where $\tilde{U}$ is the matrix formed by deleting row ρ and column σ from U .

The proof follows trivially from considering how the determinant of a square matrix is defined .

Lemma 6.2.4 .

If $A \in Q_{m,\alpha}$ then A is nonsingular and $A^{-1} \in Q^{(1)}_{m,\alpha}$.

Proof .

$\det A = \prod_{i=0}^{r+nm-1} A_{ii}$ since A is lower triangular . Therefore $\det A \neq 0$ because of definition 6.2.1 (iii) , and so A is nonsingular .

Part (i) of definition 6.2.2 is fulfilled since the inverse of any nonsingular lower matrix is itself lower triangular .

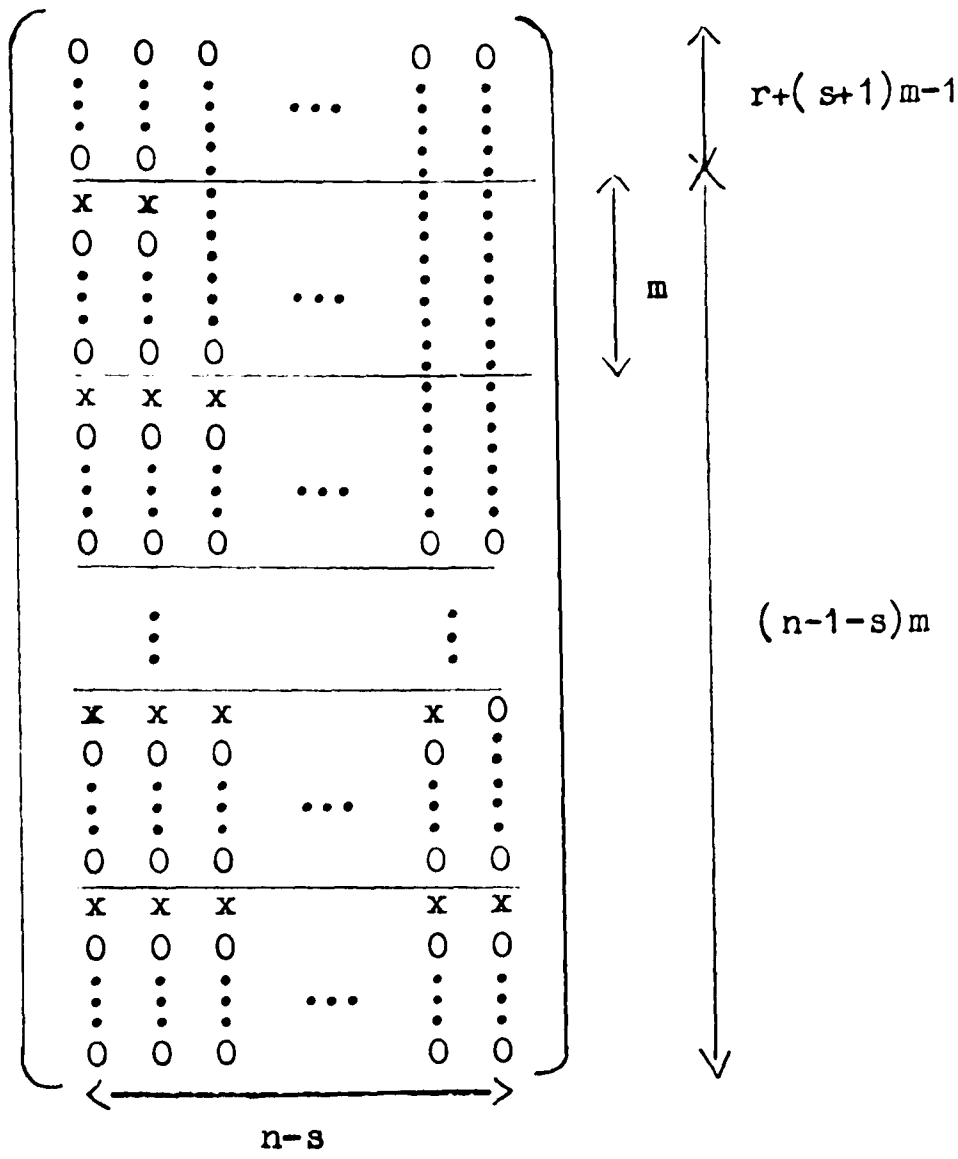
Now take any $\nu, s : 1 \leq \nu \leq m-\alpha$, $0 \leq s \leq n-1$. Let us denote the $(i,j)^{th.}$ element of A^{-1} by p_{ij} , and consider $p_{i_1, r+sm+\nu-1}$ for some $i_1 \geq r+sm+\nu$ such that $i_1 - r \not\equiv 0 \pmod m$.

Now $A^{-1} = \frac{1}{\det A} \operatorname{adj} A$, so $p_{i_1, r+sm+\nu-1} = \frac{1}{\det A} A^{*}_{r+sm+\nu-1, i_1}$, where A^{*}_{ji} is the $(j,i)^{th.}$ cofactor of A . Now $A^{*}_{r+sm+\nu-1, i_1}$ is the determinant of the matrix $\tilde{A}$ formed by deleting row $r+sm+\nu-1$ and column i_1 from A . Consider now the $(r+sm+\nu-1)^{th.}$ column of this resulting matrix $\tilde{A}$. $\tilde{A}_{i, r+sm+\nu-1} = 0$ whenever $i - r + 1 \not\equiv 0 \pmod m$ or $i < r + sm + \nu - 1$. That is , the only possible non-zero elements in the $(r + sm + \nu - 1)^{th.}$ column of $\tilde{A}$ are $\tilde{A}_{r+\sigma m-1, r+sm+\nu-1}$, where $\sigma = s+1(1)n-1$.

Consider now columns $r+\sigma m$, $\sigma = s+1(1)n-1$ of A . By definition , for each r the only possible non-zero elements in column $r+\sigma m$ are

$$A_{r+\sigma m, r+\sigma m} , A_{r+(\sigma+1)m, r+\sigma m} , \cdots , A_{r+(n-1)m, r+\sigma m} .$$

Let us now exchange columns zero and $r+sm+\nu-1$ of $\tilde{A}$, and then columns σ-s and $r+\sigma m-1$ of $\tilde{A}$, $\sigma = s+1(1)n-1$. In the resulting matrix $\tilde{A}^{(1)}$, whose determinant has the same magnitude as that of $\tilde{A}$, the first n-s columns are of the form



By lemma 6.2.3 $\det \tilde{A}^{(1)} \propto \det \tilde{A}^{(2)}$, where $\tilde{A}^{(2)}$ is the matrix formed by deleting column n-s-1 and row r+(n-1)m-1 of $\tilde{A}^{(1)}$. Since row r+(n-1)m-1 of $\tilde{A}^{(1)}$ is now deleted , column n-s-2 of $\tilde{A}^{(2)}$ has only one possibly non-zero element , in row r+(n-2)m-1 . Again lemma 6.2.3 gives $\det \tilde{A}^{(2)} \propto \det \tilde{A}^{(3)}$, where $\tilde{A}^{(3)}$ is formed by deleting column n-s-2 and row r+(n-2)m-1 of $\tilde{A}^{(2)}$. We continue in this manner until we obtain $\tilde{A}^{(n-s-1)}$, whose zero$^{th.}$ column is composed entirely of zeroes . Therefore $\det \tilde{A} = 0$, and so $A^{*}_{r+sm+\nu-1,i_1} = 0$ and we have the result .

Lemma 6.2.5 .

If $A \in Q_{m,\alpha_1}$ and $B \in Q^{(\beta)}_{m,\alpha_2}$, where α_1 , α_2 and β are positive integers less than m , then $A^{-1}B \in Q^{(\beta)}_{m,\alpha}$ where $\alpha = \min [m, \max (\alpha_2, \beta+\alpha_1-1)]$.

Proof .

By lemma 6.2.4 , $A^{-1} \in Q^{(1)}_{m,\alpha_1}$. Let us consider only rows and columns r to r+nm-1 of A^{-1} and B . We shall divide these elements of A^{-1} into m x m matrices :

$$\begin{pmatrix} \bar{A}_{oo} & & & \\ \bar{A}_{10} & \bar{A}_{11} & & \\ \vdots & & \ddots & \\ \bar{A}_{n-1,0} & \cdots & & \bar{A}_{n-1,n-1} \end{pmatrix} .$$

For s = O(1)n-1 , $\bar{A}_{ss}$ has the form

$$\begin{pmatrix} x & & & \\ & \ddots & & \\ & & x & \\ 0 & & x\ x & \\ & & \vdots \quad \ddots & \\ & & x \ \ldots\ldots \ x \end{pmatrix} ,$$

(α_1)

and $\bar{A}_{s\sigma}$ for $0 \le \sigma \le s-1$ has the form

$$\begin{pmatrix} x \ \ldots\ x & x & \ldots\ x \\ 0 \ \ldots\ 0 & x & \ldots\ x \\ \vdots \quad\ \vdots & \vdots & \quad\ \vdots \\ 0 \ \ldots\ 0 & x & \ldots\ x \end{pmatrix} .$$

(α_1)

In like manner we divide the corresponding elements of B into m x m matrices :

$$\begin{pmatrix} B_{oo} & & & \\ B_{10} & B_{11} & & \\ \vdots & & \ddots & \\ B_{n-1,0} & \cdots & & B_{n-1,n-1} \end{pmatrix} .$$

For s = O(1)n-1 , B_{ss} has the form

(β)

$$\begin{pmatrix} x & & & \\ \vdots & \ddots & & \\ x & & x & \\ & \ddots & \vdots & \ddots \\ 0 & & x & \\ & & \vdots & \\ & & x \ \ldots\ldots \ x \end{pmatrix}$$

(α')

with $\alpha' = \alpha_2$

and $B_{s\sigma}$ for $0 \leq \sigma \leq s-1$ has the form $\begin{pmatrix} x & \dots & x & x & \dots & x \\ 0 & \dots & 0 & x & \dots & x \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & x & \dots & x \end{pmatrix}$ (α_2 columns) .

Now by theorem 1.6.1 of (16) we obtain that rows $r+um$ to $r+(u+1)m-1$, columns $r+vm$ to $r+(v+1)m-1$, where $u \geq v$, of $A^{-1}B$ are composed of the $m \times m$ matrix $\sum_{w=v}^{u} \bar{A}_{uw} B_{wv}$.

Therefore the matrix on the diagonal of $A^{-1}B$ is $\bar{A}_{uu} B_{uu}$, which by direct computation we see to be of the same form as B_{ss} above , only with $\alpha' = \alpha$. For $v < u$, which corresponds to one of the $m \times m$ matrices below the diagonal of $A^{-1}B$, we observe by direct computation that $\bar{A}_{uu} B_{uv}$ has the form $\begin{pmatrix} x & \dots & x & x & \dots & x \\ 0 & \dots & 0 & x & \dots & x \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & x & \dots & x \end{pmatrix}$ (α_2 columns) , $\bar{A}_{uw} B_{wv}$ for $v < w < u$ has this form also and $\bar{A}_{uv} B_{vv}$ has the form

$$\begin{pmatrix} x & \dots & x & x & \dots & x \\ 0 & \dots & 0 & x & \dots & x \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & x & \dots & x \end{pmatrix}$$

($\min [m, \max (\alpha_2, \beta+\alpha_1-1)]$ columns) .

Hence we obtain the result .

Lemma 6.2.6 .

If $\{\xi_i\}$ is an increasing sequence of real numbers and the sequence $\{\eta_i\}$ is defined by $\eta_1 = \xi_1$, $\eta_i = \min (\xi_i, \eta_1 + 1)$, $i \geq 2$, then $\{\eta_i\}$ is also increasing .

Proof .

We shall show that $\eta_i \leq \eta_{i+1}$ for each i .

(i) $\underline{i = 1}$.

$\eta_2 = \min(\xi_2, \eta_1 + 1) = \min(\xi_2, \xi_1 + 1)$. Now $\xi_2 \geq \xi_1$ since $\{\xi_i\}$ is increasing , so $\eta_2 \geq \xi_1 = \eta_1$.

(ii) $\underline{i \geq 2}$.

$\eta_i = \min(\xi_i, \eta_1 + 1)$ and $\eta_{i+1} = \min(\xi_{i+1}, \eta_1 + 1)$.

There are four possibilities :

(a) $\eta_i = \xi_i$ and $\eta_{i+1} = \xi_{i+1}$. Then $\eta_{i+1} \geq \eta_i$ since $\{\xi_i\}$ is increasing .

(b) $\eta_i = \xi_i$ and $\eta_{i+1} = \eta_1 + 1$. Then $\xi_i = \min(\xi_i, \eta_1 + 1)$, and so $\xi_i \leq \eta_1 + 1 = \eta_{i+1}$, that is , $\eta_i \leq \eta_{i+1}$.

(c) $\eta_i = \eta_1 + 1$ and $\eta_{i+1} = \xi_{i+1}$. Then $\eta_1 + 1 = \min(\xi_i, \eta_1 + 1)$, and so $\eta_1 + 1 \leq \xi_i \leq \xi_{i+1}$ since $\{\xi_i\}$ is increasing . That is , $\eta_i \leq \eta_{i+1}$.

(d) $\eta_i = \eta_1 + 1$ and $\eta_{i+1} = \eta_1 + 1$. Then $\eta_i = \eta_{i+1}$.

In all four cases then , $\eta_i \leq \eta_{i+1}$.

$\underline{\text{Lemma 6.2.7}}$.

If $\{\xi_i\}$ is an increasing sequence of real numbers , ϱ is a positive integer greater than 2 and the sequence $\{\eta_i\}$ is defined by

$$\eta_1 = \xi_1 ,$$
$$\eta_i = \min(\xi_i , \eta_1 + 1) , \quad i = 2(1)\varrho ,$$
$$\eta_i = \min(\xi_i, \eta_{i-\varrho+1} + 1) , \quad i \geq \varrho+1 ,$$

then $\{\eta_i\}$ is also increasing .

$\underline{\text{Proof}}$.

Lemma 6.2.6 gives $\eta_1 \leq \eta_2 \leq \ldots \leq \eta_\varrho$ immediately .

Assume now that $\eta_1 \leq \eta_2 \leq \ldots \leq \eta_q$ (6.2.1) for some positive integer $q \geq \varrho$.

Then $\eta_q = \min(\xi_q, \eta_{q-\varrho+1} + 1)$ and $\eta_{q+1} = \min(\xi_{q+1}, \eta_{q-\varrho+2} + 1)$.

Again there are four possibilities :

(a) $\eta_q = \xi_q$ and $\eta_{q+1} = \xi_{q+1}$. Then $\eta_{q+1} \geq \eta_q$ since $\{\xi_i\}$ is increasing .

(b) $\eta_q = \xi_q$ and $\eta_{q+1} = \eta_{q-\varrho+2} + 1$. Then $\eta_q = \xi_q \leq \eta_{q-\varrho+1} + 1$ $\leq \eta_{q-\varrho+2} + 1$ by (6.2.1) , so $\eta_q \leq \eta_{q+1}$.

(c) $\eta_q = \eta_{q-\varrho+1}$ and $\eta_{q+1} = \xi_{q+1}$. Then $\eta_q = \eta_{q-\varrho+1} + 1 \leq \xi_q \leq \xi_{q+1}$ since $\{\xi_i\}$ is increasing , so $\eta_q \leq \eta_{q+1}$.

(d) $\eta_q = \eta_{q-\varrho+1}$ and $\eta_{q+1} = \eta_{q-\varrho+2} + 1$. $\eta_{q-\varrho+1} \leq \eta_{q-\varrho+2}$ by (6.2.1) , so $\eta_q \leq \eta_{q+1}$ and we have the result by induction .

Definition 6.2.8 .

Take any set of positive integers $\{p_1, p_2, \ldots, p_m\}$ and $2 \leq \beta \leq \alpha \leq m$. We define the set $\{\pi_1, \pi_2, \ldots, \pi_m\}$ by

$$\pi_1 = p_1 ,$$

$$\pi_\nu = \min (p_\nu, \pi_1 + 1) , \quad \nu = 2(1)\min(\beta, m-\alpha+1) ,$$

$$\pi_\nu = \min (p_\nu, \pi_{\nu-\beta+1} + 1) , \quad \nu = \beta+1(1)m-\alpha+1 \text{ if } \alpha \leq m-\beta ,$$

$$\pi_\nu = \pi_{m-\alpha+1} , \quad \nu = m-\alpha+2(1)m .$$

Lemma 6.2.9 .

In definition 6.2.8 if $p_1 \leq p_2 \leq \ldots \leq p_m$, then $\pi_1 \leq \pi_2 \leq \ldots \leq \pi_m$.

Proof .

This result follows immediately from putting $\varrho = \min (\beta, m-\alpha+1)$ in lemma 6.2.7 .

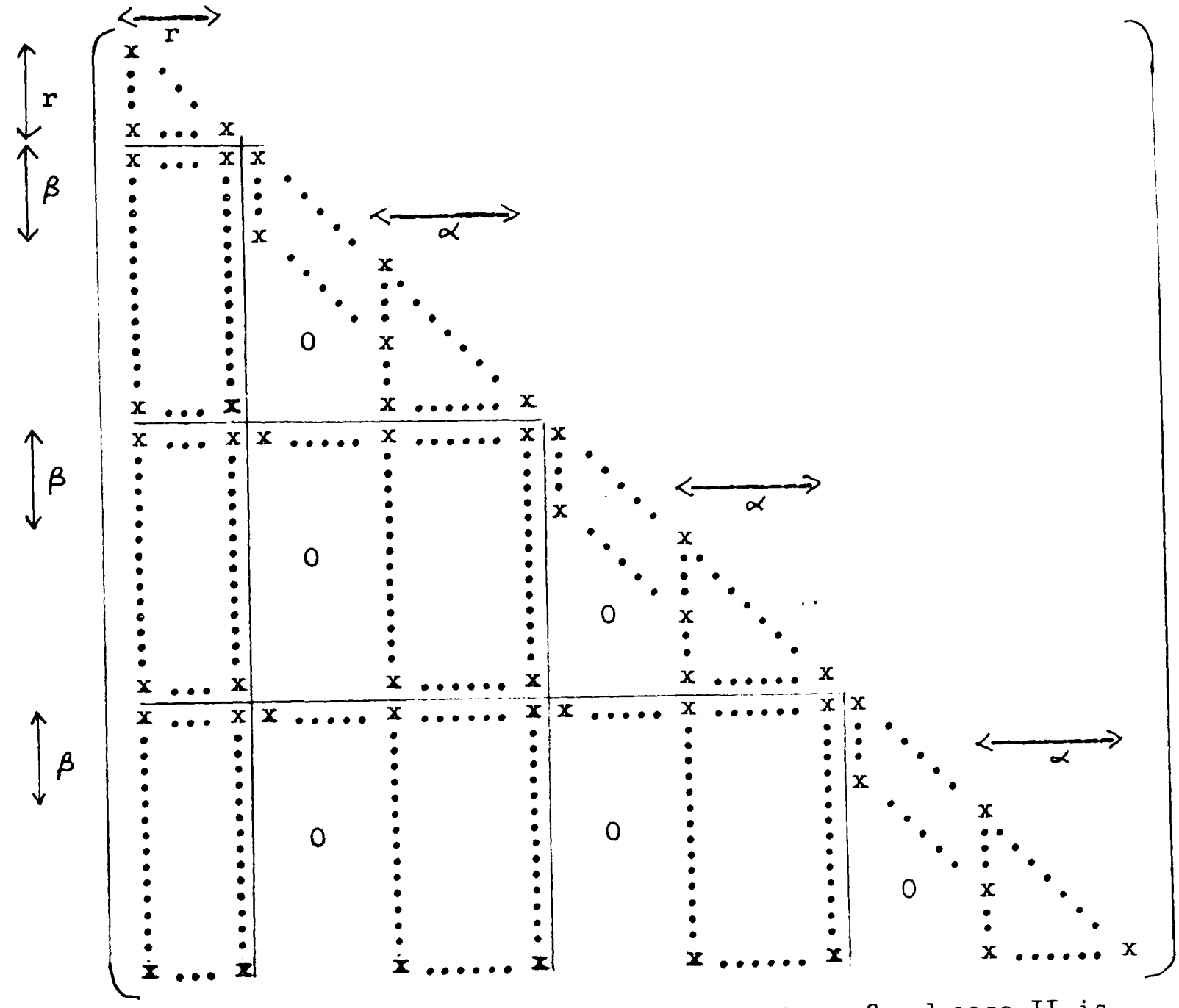
Theorem 6.2.10 .

Suppose we have an m-block semi explicit method which is zero stable and optimally consistent of orders $(p_1, p_2, \ldots, p_m)$ with $p_1 \leq p_2 \leq \ldots \leq p_m$, $m \geq 3$, with starting values accurate of order p_m . Suppose also that $A_h \in Q_{m,1}$ and $B_h \in Q^{(\beta)}_{m,\alpha}$ with $2 \leq \beta \leq \alpha < m$. Then for each $\nu = 1(1)m$

there exists $C_U^{(\nu)}$ such that $\|e^{(\nu)}\|_\infty \leq C_U^{(\nu)} h^{\pi_\nu}$, where $\pi_1, \pi_2, \ldots, \pi_m$ are defined as in definition 6.2.8 .

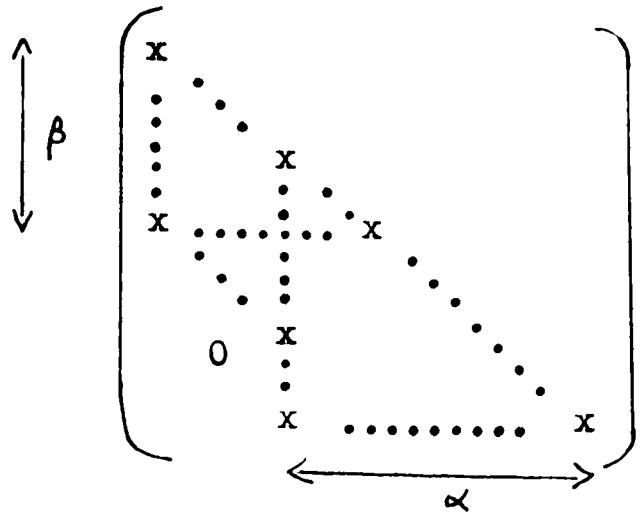
Proof .

By lemma 6.2.5 , $A_h^{-1}B_h \in Q_{m,\alpha_o}^{(\beta)}$, where $\alpha_o = \min[m, \max(\alpha,\beta)]$ $= \max(\alpha,\beta) = \alpha$ since $\beta \leq \alpha < m$. That is , $A_h^{-1}B_h$ has the form

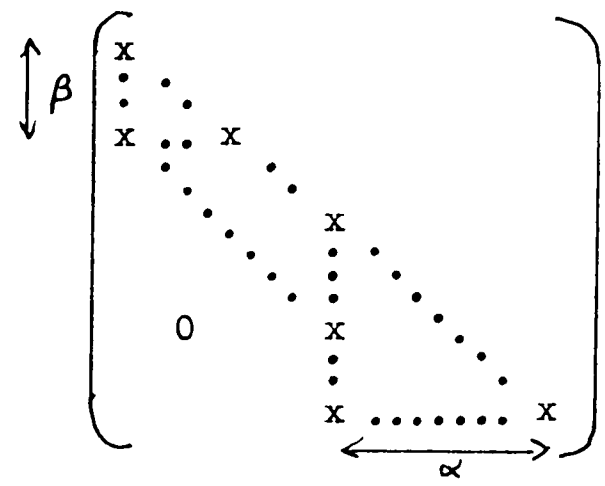


There are two cases to consider . Case I is when $\alpha > m-\beta$ and case II is when $\alpha \leq m-\beta$.

In case I the m x m matrices on the diagonal of $A_h^{-1}B_h$ have the form



(6.2.2) , and in case II they have the form



(6.2.3) .

As in theorem 4.4.2 , for each $\nu = 1(1)m$, $s = 0(1)n-1$, putting $i = r+sm+\nu-1$

$$|e_{r+sm+\nu-1}| \le |[A_h^{-1}\theta^*]_{r+sm+\nu-1}| + hL\sum_{j=0}^{r+sm+\nu-1}|(A_h^{-1}B_h)_{r+sm+\nu-1,j}||e_j|$$

$$= C^{(r+sm+\nu-1)}h^{p_\nu} + hL\sum_{j=0}^{r+sm+\nu-1}|(A_h^{-1}B_h)_{r+sm+\nu-1,j}||e_j| \ .$$

Now because of zero stability $|(A_h^{-1}B_h)_{ij}| \le M_o\|B_h\|_1$ for all i,j (4.4.1) ,

so $\sum_{j=0}^{r-1}|(A_h^{-1}B_h)_{r+sm+\nu-1,j}||e_j| \le M_o\|B_h\|_1\sum_{j=0}^{r-1}C_j'h^{p_m}$, and so for sufficiently small h , for each $s = 0(1)n-1$ there exist constants $C_1^{(r+sm+\nu-1)}$, M_1 such that

$$|e_{r+sm}| \le C_1^{(r+sm)}h^{p_1} + hM_1\sum_{j=r}^{r+sm-1}|e_j| \ ,$$

$$|e_{r+sm+\nu-1}| \le C_1^{(r+sm+\nu-1)}h^{p_\nu} + hM_1\sum_{\sigma=0}^{s-1}\sum_{q=m-\alpha+1}^{m}|e_{r+\sigma m+q-1}| + hM_1\sum_{\nu} \ ,$$

where

$$\sum_{\nu} = \begin{cases} \sum_{q=1}^{\nu-1} |e_{r+sm+q-1}| \ , \ \nu = 2(1)\beta \ , \\ \sum_{q=\nu-\beta+1}^{\nu-1} |e_{r+sm+q-1}| \ , \ \nu = \beta+1(1)m-\alpha+\beta-1 \ , \text{ if } \alpha \leq m-2 \ , \\ \sum_{q=m-\alpha+1}^{\nu-1} |e_{r+sm+q-1}| \ , \ \nu = m-\alpha+\beta(1)m \ . \end{cases} \qquad (6.2.4) \ ,$$

taking into account the form of $A_h^{-1}B_h$.

The result will follow once we show that for each $\nu = 1(1)m$ there exists $C_U^{(\nu)}$ such that $|e_{r+sm+\nu-1}| \leq C_U^{(\nu)} h^{\pi_\nu}$, $s = 0(1)n-1$ (6.2.5) .

To prove this we shall use induction on s . First however let us define the set $\{\pi_1^*, \pi_2^*, \ldots, \pi_m^*\}$ by $\pi_1^* = p_1$,

$$\pi_\nu^* = \min (p_\nu, \pi_1^* + 1) \ , \ \nu = 2(1)\beta \ ,$$

$$\pi_\nu^* = \min (p_\nu, \pi_{\nu-\beta+1}^* + 1) \ , \ \nu = \beta+1(1)m \ .$$

By lemma 6.2.7 , $\pi_1^* \leq \pi_2^* \leq \ldots \leq \pi_m^*$ (6.2.6) .

There are two possibilities : case I — $\alpha+\beta > m$;

case II — $\alpha+\beta \leq m$.

Case I .

In definition 6.2.8 we have $\pi_1 = p_1$,

$$\pi_\nu = \min (p_\nu, \pi_1 + 1) \ , \ \nu = 2(1)m-\alpha+1 \ ,$$

$$\pi_\nu = \pi_{m-\alpha+1} \ , \ \nu = m-\alpha+2(1)m \ .$$

Therefore $\pi_\nu = \pi_\nu^*$ for $\nu = 1(1)m-\alpha+1$ (6.2.7) and $\pi_\nu^* \geq \pi_\nu$ for $\nu = m-\alpha+1(1)m$ (6.2.8) .

(i) $s = 0$.

We must show that $|e_{r+\nu-1}| \leq C_U^{(0,\nu)} h^{\pi_\nu}$ (6.2.9) , $\nu = 1(1)m$, for some constants $C_U^{(0,\nu)}$. To do this , we use induction on ν .

(a) $\nu = 1$.

(6.2.4) gives $|e_r| \le C_1^{(r)} h^{p_1} = C_1^{(r)} h^{\pi_1}$.

(b) Assume that (6.2.9) holds for $\nu = 1(1)q : 1 \le q < m-\alpha+1$.

$\alpha + \beta > m$, so $\beta \ge m-\alpha+1$ and (6.2.4) gives

$$|e_{r+q}| \le C_1^{(r+q)} h^{p_{q+1}} + hM_1 \sum_{\lambda=1}^{q} |e_{r+\lambda-1}| \le C_1^{(r+q)} h^{p_{q+1}} + hM_1 \sum_{\lambda=1}^{q} C_U^{(0,\lambda)} h^{\pi_\lambda}$$

$$\le C_1^{(r+q)} h^{p_{q+1}} + M_{12}^{(0,q)} h^{\pi_1+1} \quad \text{for some } M_{12}^{(0,q)} \text{ , since by lemma 6.2.9}$$

$\pi_1 \le \pi_2 \le \ldots \le \pi_q$. So $|e_{r+q}| \le C_U^{(0,q+1)} h^{\pi_{q+1}}$ forsome $C_U^{(0,q+1)}$. This gives (6.2.5) for $s = 0$, $\nu = 1(1)m-\alpha+1$.

(c) Assume that (6.2.9) holds for $\nu = 1(1)q : m-\alpha+1 \le q \le \beta$.

(6.2.4) gives

$$|e_{r+q}| \le C_1^{(r+q)} h^{p_{q+1}} + hM_1 \sum_{\lambda=1}^{q} |e_{r+\lambda-1}|$$

$$\le C_1^{(r+q)} h^{p_{q+1}} + hM_1 \sum_{\lambda=1}^{q} C_U^{(0,\lambda)} h^{\pi_\lambda} \le C_U^{(0,q+1)} h^{\pi^*_{q+1}}$$

for some $C_U^{(0,q+1)}$. $\pi^*_{q+1} \ge \pi^*_{m-\alpha+1}$ by (6.2.6) . $\pi^*_{m-\alpha+1} = \pi_{m-\alpha+1}$ (6.2.7) .

$\pi_{m-\alpha+1} = \pi_{q+1}$ by definition , so we have $|e_{r+q}| \le C_U^{(0,q+1)} h^{\pi_{q+1}}$.

(d) Assume that (6.2.9) holds for $\nu = 1(1)q : \beta \le q < m-\alpha+\beta-1$.

(6.2.4) gives

$$|e_{r+q}| \le C_1^{(r+q)} h^{p_{q+1}} + hM_1 \sum_{\lambda=q-\beta+2}^{q} |e_{r+\lambda-1}|$$

$$\le C_1^{(r+q)} h^{p_{q+1}} + hM_1 \sum_{\lambda=q-\beta+2}^{q} C_U^{(0,\lambda)} h^{\pi_\lambda}$$

$$\le C_1^{(r+q)} h^{p_{q+1}} + M_{12}^{(0,q)} h^{\pi_{q-\beta+2}+1} \quad \text{for some } M_{12}^{(0,q)} \text{ ,}$$

since $\pi_{q-\beta+2} \le \pi_{q-\beta+3} \le \ldots \le \pi_q$ by lemma 6.2.9 .

Now for $\beta \le q < m-\alpha+\beta-1$, $2 \le q-\beta+2 < m-\alpha+1$, so $\pi_{q-\beta+2} = \pi^*_{q-\beta+2}$.

This then gives $|e_{r+q}| \le C_U^{(0,q+1)} h^{\pi^*_{q+1}}$ for some $C_U^{(0,q+1)}$. (6.2.8) implies

that $\pi^*_{q+1} \geq \pi_{q+1}$, so $|e_{r+q}| \leq C_U^{(0,q+1)} h^{\pi_{q+1}}$.

(e) Assume that (6.2.9) holds for $\nu = 1(1)q : m-\alpha+\beta-1 \leq q \leq m$.

(6.2.4) gives

$$|e_{r+q}| \leq C_1^{(r+q)} h^{p_{q+1}} + hM_1 \sum_{\lambda=m-\alpha+1}^{q} |e_{r+\lambda-1}|$$

$$\leq C_1^{(r+q)} h^{p_{q+1}} + hM_1 \sum_{\lambda=m-\alpha+1}^{q} C_U^{(0,\lambda)} h^{\pi_\lambda}$$

$$\leq C_1^{(r+q)} h^{p_{q+1}} + M_{12}^{(0,q)} h^{\pi_{m-\alpha+1} + 1} \text{ for some } M_{12}^{(0,q)}$$

since $\pi_{m-\alpha+1} = \pi_{m-\alpha+2} = \ldots = \pi_q$ by definition .

So by (6.2.7) ,

$$|e_{r+q}| \leq C_1^{(r+q)} h^{p_{q+1}} + M_{12}^{(0,q)} h^{\pi^*_{m-\alpha+1} + 1} .$$

By lemma 6.2.7 , $\pi^*_{m-\alpha+1} \leq \pi^*_{q-\beta+2}$, so

$$|e_{r+q}| \leq C_1^{(r+q)} h^{p_{q+1}} + M_{12}^{(0,q)} h^{\pi^*_{q-\beta+2} + 1} \leq C_U^{(0,q+1)} h^{\pi^*_{q+1}} \text{ for some } C_U^{(0,q+1)} .$$

Thus by (6.2.8) $|e_{r+q}| \leq C_U^{(0,q+1)} h^{\pi_{q+1}}$.

(ii) Assume that $|e_{r+\sigma m+\nu-1}| \leq C_U^{(\sigma,\nu)} h^{\pi_\nu}$, $\nu = 1(1)m$.,

$\sigma = 0(1)s$, for some constants $C_U^{(\sigma,\nu)}$.

We then need to show that $|e_{r+(s+1)m+\nu-1}| \leq C_U^{(s+1,\nu)} h^{\pi_\nu}$ (6.2.10) , $\nu = 1(1)m$,

for some constants $C_U^{(s+1,\nu)}$. Let us use induction on ν .

(a) $\nu = 1$.

(6.2.4) gives $|e_{r+(s+1)m}| \leq C_1^{(r+(s+1)m)} h^{p_1} + hM_1 \sum_{j=r}^{r+(s+1)m-1} |e_j|$

$$\leq C_1^{(r+(s+1)m)} h^{p_1} + M_{12}^{(s+1,q)} h^{\pi_1+1}$$

for some $M_{12}^{(s+1,q)}$. Hence $|e_{r+(s+1)m}| \leq C_U^{(s+1,0)} h^{\pi_1}$ for some $C_U^{(s+1,0)}$.

(b) Assume that (6.2.10) holds for $\nu = 1(1)q : 1 \leq q \leq m-\alpha+1$.

(6.2.4) gives

$$|e_{r+(s+1)m+q}| \leq C_1^{(r+(s+1)m+q)} h^{p_{q+1}} + hM_1 \sum_{\sigma=0}^{s} \sum_{\lambda=m-\alpha+1}^{m} |e_{r+\sigma m+\lambda-1}| + hM_1 \sum_{\lambda=1}^{q} |e_{r+(s+1)m+\lambda-1}|$$

$$\leq C_1^{(r+(s+1)m+q)} h^{p_{q+1}} + hM_1 \sum_{\sigma=0}^{s} \sum_{\lambda=m-\alpha+1}^{m} C_U^{(\sigma,\lambda)} h^{\pi_\lambda} + hM_1 \sum_{\lambda=1}^{q} C_U^{(s+1,\lambda)} h^{\pi_\lambda} .$$

So $|e_{r+(s+1)m+q}| \leq C_1^{(r+(s+1)m+q)} h^{p_{q+1}} + M_{12}^{(s+1,q)} h^{\pi_{m-\alpha+1}} + M_{12}^{(s+1,q)} h^{\pi_1+1}$

for some $M_{12}^{(s+1,q)}$, and hence

$$|e_{r+(s+1)m+q}| \leq M_{13}^{(s+1,q)} h^{\pi_{q+1}} + M_{12}^{(s+1,q)} h^{\pi_{m-\alpha+1}} \text{ for some } M_{13}^{(s+1,q)},$$

$$\leq C_U^{(s+1,q+1)} h^{\pi_{q+1}} \text{ for some } C_U^{(s+1,q+1)}, \text{ since } \pi_{q+1} \leq \pi_{m-\alpha+1} .$$

(c) Assume that (6.2.10) holds for $\nu = 1(1)q : m-\alpha+1 \leq q \leq \beta$.

(6.2.4) gives

$$|e_{r+(s+1)m+q}| \leq C_1^{(r+(s+1)m+q)} h^{p_{q+1}} + hM_1 \sum_{\sigma=0}^{s} \sum_{\lambda=m-\alpha+1}^{m} |e_{r+\sigma m+\lambda-1}| + hM_1 \sum_{\lambda=1}^{q} |e_{r+(s+1)m+\lambda-1}|$$

$$\leq C_1^{(r+(s+1)m+q)} h^{p_{q+1}} + M_{12}^{(s+1,q)} h^{\pi_{m-\alpha+1}} + M_{12}^{(s+1,q)} h^{\pi_1+1} \text{ as before ,}$$

$$\leq M_{13}^{(s+1,q)} h^{\pi^*_{q+1}} + M_{12}^{(s+1,q)} h^{\pi_{m-\alpha+1}} .$$

Now $\pi^*_{q+1} \geq \pi_{q+1}$ by (6.2.8) and $\pi_{q+1} \geq \pi_{m-\alpha+1}$ by definition , so

$$\left|e_{r+(s+1)m+q}\right| \le C_U^{(s+1,q)} h^{\pi_{q+1}} \text{ for some } C_U^{(s+1,q)} .$$

(d) *Assume that (6.2.10) holds for $\nu = 1(1)q : \beta \le q \le m-\alpha+\beta-1$.*

(6.2.4) gives

$$\left|e_{r+(s+1)m+q}\right| \le C_1^{(r+(s+1)m+q)} h^{p_{q+1}} + hM_1 \sum_{\sigma=0}^{s} \sum_{\lambda=m-\alpha+1}^{m} \left|e_{r+\sigma m+\lambda-1}\right| + hM_1 \sum_{\lambda=q-\beta+2}^{q} \left|e_{r+sm+\lambda-1}\right|$$

$$\le C_1^{(r+(s+1)m+q)} h^{p_{q+1}} + M_{12}^{(s+1,q)} h^{\pi_{m-\alpha+1}} + M_{12}^{(s+1,q)} h^{\pi_{q-\beta+2}+1}$$

$$\text{for some } M_{12}^{(s+1,q)},$$

$$\le M_{13}^{(s+1,q)} h^{\pi_{q+1}} + M_{12}^{(s+1,q)} h^{\pi_{m-\alpha+1}} \text{ for some } M_{13}^{(s+1,q)},$$

applying the same reasoning as in (i) (d) .

Then since $\pi_{q+1} \ge \pi_{m-\alpha+1}$ there exists $C_U^{(s+1,q+1)}$ such that

$$\left|e_{r+(s+1)m+q}\right| \le C_U^{(s+1,q+1)} h^{\pi_{q+1}} \text{ for some } C_U^{(s+1,q+1)}.$$

(e) *Assume that (6.2.10) holds for $\nu = 1(1)q : m-\alpha+\beta-1 \le q \le m$.*

(6.2.4) gives

$$\left|e_{r+(s+1)m+q}\right| \le C_1^{(r+(s+1)m+q)} h^{p_{q+1}} + hM_1 \sum_{\sigma=0}^{s} \sum_{\lambda=m-\alpha+1}^{m} \left|e_{r+\sigma m+\lambda-1}\right| + hM_1 \sum_{\lambda=m-\alpha+1}^{q} \left|e_{r+sm+\lambda-1}\right|$$

$$\le C_1^{(r+(s+1)m+q)} h^{p_{q+1}} + M_{12}^{(s+1,q)} h^{\pi_{m-\alpha+1}} + M_{12}^{(s+1,q)} h^{\pi_{m-\alpha+1}+1}$$

$$\text{for some } M_{12}^{(s+1,q)},$$

$$\le M_{13}^{(s+1,q)} h^{\pi_{q+1}} + M_{12}^{(s+1,q)} h^{\pi_{m-\alpha+1}} \text{ for some } M_{13}^{(s+1,q)},$$

applying the same reasoning as in (i) (e) . Then since $\pi_{q+1} \geq \pi_{m-\alpha+1}$, there exists $C_U^{(s+1,q+1)}$ such that $|e_{r+(s+1)m+q}| \leq C_U^{(s+1,q+1)} h^{\pi_{q+1}}$.

This completes the proof for case I .

Case II .

In definition 6.2.8 we have

$\pi_1 = p_1$,

$\pi_\nu = \min(p_\nu, \pi_1 + 1)$, $= 2(1)\beta$,

$\pi_\nu = \min(p_\nu, \pi_{\nu-\beta+1} + 1)$, $\nu = \beta+1(1)m-\alpha+1$,

$\pi_\nu = \pi_{m-\alpha+1}$, $\nu = m-\alpha+2(1)m$.

Again $\pi_\nu = \pi_\nu^*$ for $\nu = 1(1)m-\alpha+1$ (6.2.11) and $\pi_\nu^* \geq \pi_\nu$ for $\nu = m-\alpha+1(1)m$ (6.2.12) .

(i) $s = 0$.

(a) $\nu = 1$.

$|e_r| \leq C_U^{(0,1)} h^{\pi_1}$ as in case I .

(b) Assume that (6.2.9) holds for $\nu = 1(1)q : 1 \leq q < \beta$.

Then we obtain $|e_{r+q}| \leq C_U^{(0,q+1)} h^{\pi_{q+1}}$, as in case I .

(c) Assume that (6.2.9) holds for $\nu = 1(1)q : \beta \leq q < m-\alpha+\beta-1$.

We obtain $|e_{r+q}| \leq C_U^{(0,q+1)} h^{\pi_{q+1}}$ as in case I , (i) (d) .

(d) Assume that (6.2.9) holds for $\nu = 1(1)q : m-\alpha+\beta-1 \leq q < m$.

We obtain $|e_{r+q}| \leq C_U^{(0,q+1)} h^{\pi_{q+1}}$ as in case I , (i) (e) .

(ii) Assume that $|e_{r+\sigma m+\nu-1}| \leq C_U^{(\sigma,\nu)} h^{\pi_\nu}$, $\nu = 1(1)m$, $\sigma = 0(1)s$, for some constants $C_U^{(\sigma,\nu)}$.

Going through the steps (a) , (b) , (c) , (d) as above we obtain

the required result as in case I , and so we have the result for both cases .

This theorem gives a sequence of upper bounds on the various stages of the global error vector , and we can now apply this to various examples , and shall go on in the next section to give two-sided error bounds for some cases .

Example 6.1.2 .

$A_h \in Q_{3,1}$ and $B_h \in Q^{(2)}_{3,1}$ and we saw previously that optimal consistency is of orders $(2,3,3)$. We have also noted in section 4.3 that a predictor corrector scheme is zero stable if and only if the corrector is when considered as a single linear multistep method . This is the case in the method under consideration . Thus if we take starting values accurate of order three , the conditions of theorem 6.2.10 are fulfilled . $\beta = 2$ and hence we obtain $\pi_1 = 2$, $\pi_2 = \pi_3 = 3$, and

$$\|e^{(1)}\|_\infty \leq C_U^{(1)} h^2 ,$$

$$\|e^{(2)}\|_\infty \leq C_U^{(2)} h^3 ,$$

$$\|e^{(3)}\|_\infty \leq C_U^{(3)} h^3 \quad \text{for some } C_U^{(1)}, C_U^{(2)}, C_U^{(3)} .$$

Example 6.1.3 .

$A_h \in Q_{3,1}$ and $B_h \in Q^{(2)}_{3,1}$ and optimal consistency is of orders $(2,3,3)$. We have also seen that any Runge-Kutta method is zero stable . Thus if we take a starting value accurate of order three , the conditions of theorem 6.2.10 are fulfilled . Again $\beta = 2$, $\pi_1 = 2$, $\pi_2 = \pi_3 = 3$, and so

$$\| e^{(1)} \|_\infty \leq C_U^{(1)} h^2 ,$$

$$\| e^{(2)} \|_\infty \leq C_U^{(2)} h^3 ,$$

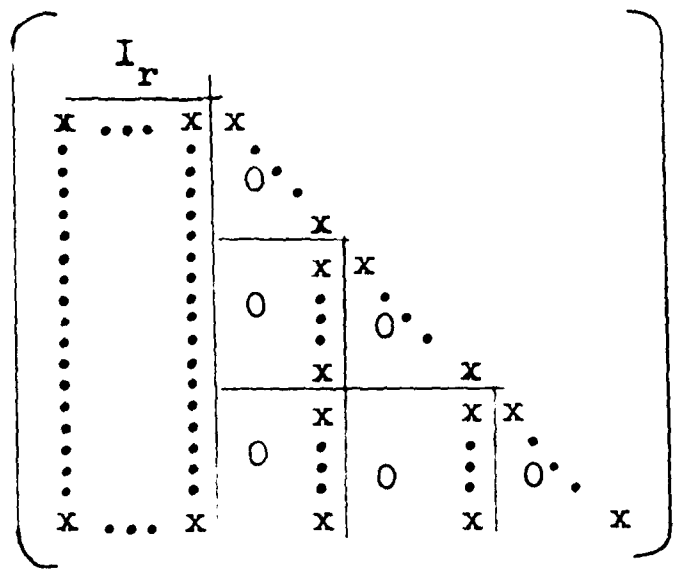
$$\| e^{(3)} \|_\infty \leq C_U^{(3)} h^3 .$$

Example 6.2.11 .

Suppose we have a predictor-corrector method in $P(EC)^\mu(E)$ mode using a predictor of order $p-\mu$ and a corrector of order p , with $\mu \geq 2$.

$A_h \in Q_{\mu+1,1}$ and in $P(EC)^\mu E$ mode $B_h \in Q^{(2)}_{\mu+1,1}$ and in $P(EC)^\mu$ mode $B_h \in Q^{(2)}_{\mu+1,2}$. $Q^{(2)}_{\mu+1,1} \subset Q^{(2)}_{\mu+1,2}$, so in both cases $B_h \in Q^{(2)}_{\mu+1,2}$.

By lemma 6.2.4 $A_h^{-1} \in Q^{(1)}_{\mu+1,1}$ and has the form



.

Taking starting values accurate of order p ,

$$\theta^* = (C'_0 h^p, \ \ldots \ , C'_{r-1} h^p, C^* h^{p-\mu+1}, Ch^{p+1}, \ \ldots \ , Ch^{p+1}, \ \ldots \ )^T$$

where C^* and C are the error constants of the predictor and the corrector respectively .

Therefore by computing $A_h^{-1}\theta^*$ directly we obtain optimal consistency of orders $(p-\mu+1, p, \ \ldots \ , p)$. Then provided we take a zero stable corrector , the conditions of theorem 6.2.10 are fulfilled with $\beta = \alpha = 2$, and so we obtain

$$\|e^{(1)}\|_\infty \le C_U^{(1)} h^{p-\mu+1} ,$$
$$\|e^{(2)}\|_\infty \le C_U^{(2)} h^{p-\mu+2} ,$$
$$\vdots$$
$$\|e^{(\mu-1)}\|_\infty \le C_U^{(\mu-1)} h^{p-1} ,$$
$$\|e^{(\mu)}\|_\infty \le C_U^{(\mu)} h^{p} ,$$
$$\|e^{(\mu+1)}\|_\infty \le C_U^{(\mu+1)} h^{p} .$$

We note that both $y^*_{r+s(\mu+1)+\mu}$ and $y^*_{r+s(\mu+1)+\mu-1}$ converge with order at least p . This implies that on the final application of the predictor-corrector scheme the $\mu^{th.}$ iteration of the corrector may be omitted and the same order of convergence obtained .

Example 6.2.12 .

If we take as predictor $y_i = y_{i-1} + hf_{i-1}$, and as corrector

$$y_i = y_{i-1} + \frac{h}{720}(251f_i + 646f_{i-1} - 264f_{i-2} + 106f_{i-3} - 19f_{i-4})$$

in $P(EC)^4(E)$ mode we have a special case of example 6.2.11 with $\mu = 4$ and $p = 5$. Hence

$$\|e^{(1)}\|_\infty \le C_U^{(1)} h^{2} ,$$
$$\|e^{(2)}\|_\infty \le C_U^{(2)} h^{3} ,$$
$$\|e^{(3)}\|_\infty \le C_U^{(3)} h^{4} ,$$
$$\|e^{(4)}\|_\infty \le C_U^{(4)} h^{5} ,$$
$$\|e^{(5)}\|_\infty \le C_U^{(5)} h^{5} .$$

Example 6.2.13 .

Consider a method composed of an explicit predictor of order $p-\mu$,

followed by μ different implicit correctors $C_1, C_2, \ldots, C_\mu$, of orders $p-\mu+1, p-\mu+2, \ldots, p$.

We then have a predictor-corrector scheme in $P\left(\prod_{\omega=1}^{\mu}(EC_\omega)\right)(E)$ mode. Again $A_h \in Q_{\mu+1,1}$ and $B_h \in Q^{(2)}_{\mu+1,2}$ and we have zero stability. Taking starting values accurate of order p we obtain

$$\theta^* = (C'_0 h^p, \ldots, C'_{r-1} h^p, C^* h^{p-\mu+1}, C_1 h^{p-\mu+2}, \ldots, C_{\mu-1} h^p, C_\mu h^{p+1})^T ,$$

where C^* is the error constant of the predictor and $C_1, \ldots, C_\mu$ are the error constants of the correctors. Again we obtain by direct computation $A_h^{-1}\theta^*$ optimal consistency of orders $(p-\mu+1, p-\mu+2, \ldots, p-1, p, p)$. The conditions of theorem 6.2.10 are again fulfilled with $\beta = \alpha = 2$, giving

$$\|e^{(1)}\|_\infty \leq C_U^{(1)} h^{p-\mu+1} ,$$

$$\|e^{(2)}\|_\infty \leq C_U^{(2)} h^{p-\mu+2} ,$$

$$\vdots$$

$$\|e^{(\mu-1)}\|_\infty \leq C_U^{(\mu-1)} h^{p-1} ,$$

$$\|e^{(\mu)}\|_\infty \leq C_U^{(\mu)} h^{p} ,$$

$$\|e^{(\mu+1)}\|_\infty \leq C_U^{(\mu+1)} h^{p} .$$

6.3 Two-Sided Error Bounds.

In some cases we may continue the analysis to obtain sequences of two-sided error bounds of the form (6.1.2).

Theorem 6.3.1.

Suppose we have an m-block semi explicit method for which the conditions of theorem 6.2.10 are fulfilled and $m \geq 3$.

Then $C_L h^{p_1} \leq \|e\|_\infty \leq C_U h^{p_1}$ for some C_L and C_U (6.3.1).

If in addition there exists $q \geq \beta+1$ such that $p_q = \pi_{q-\beta+1}$, then we have the two-sided error bound

$$C_L^{(q^*,q^*+1,\ \ldots\ ,m-1,m)} h^{\pi_{q^*}} \leq \left\| e^{(q^*,q^*+1,\ \ldots\ ,m-1,m)} \right\|_\infty \leq C_U^{(q^*,q^*+1,\ \ldots\ ,m-1,m)} h^{\pi_{q^*}} \qquad (6.3.2)$$

for some $C_L^{(q^*,q^*+1,\ \ldots\ ,m-1,m)}$ and $C_U^{(q^*,q^*+1,\ \ldots\ ,m-1,m)}$, where $q^* = \min(q-\beta+1, m-\alpha+1)$.

Proof .

(6.3.1) follows immediately from (6.1.1) . For each $q \geq \beta+1$,

$\left\| e^{(q^*,q^*+1,\ \ldots\ ,m-1,m)} \right\|_\infty = \max_{\nu=q^*(1)m} \left\| e^{(\nu)} \right\|_\infty \leq C_U^{(q^*)} h^{\pi_{q^*}}$ by theorem 6.2.10 , since $\pi_{q^*} \leq \pi_{q^*+1} \leq \ldots \leq \pi_m$ by lemma 6.2.9 . So we have the upper bound in (6.3.2) , taking $C_U^{(q^*,q^*+1,\ \ldots\ ,m-1,m)} = C_U^{(q^*)}$.

Now $A_h^{-1}\theta^* = e - hA_h^{-1}B_h[f(\Delta_h t, y^*) - f(\Delta_h t, \Delta_h y(t))]$ (5.1.3) , so for each $s = O(1)n-1$,

$$C^{(r+sm+q-1)} h^{p_q} \leq \left|e_{r+sm+q-1}\right| + h\left|[A_h^{-1}B_h(f(\Delta_h t, y^*) - f(\Delta_h t, \Delta_h y(t))]_{r+sm+q-1}\right|$$

$$\leq \left|e_{r+sm+q-1}\right| + hM' \sum_{j=0}^{r-1} \left|e_j\right| + hM' \sum_{\sigma=0}^{s-1} \sum_{\lambda=m-\alpha+1}^{m} \left|e_{r+\sigma m+\lambda-1}\right| + hM' \sum_{\lambda=q^*}^{\nu} \left|e_{r+sm+\lambda-1}\right| ,$$

taking into account the particular form of $A_h^{-1}B_h$. Now

$\left|e_{r+sm+q-1}\right| \leq \left\| e^{(q^*,q^*+1,\ \ldots\ ,m-1,m)} \right\|_\infty$,

$\left|e_i\right| \leq \left\| e^{(q^*,q^*+1,\ \ldots\ ,m-1,m)} \right\|_\infty$ for $i = O(1)r-1$,

$|e_{r+\sigma m+\lambda-1}| \leq \|e^{(q^*,q^*+1,\ \ldots\ ,m-1,m)}\|_\infty$ for $\sigma = O(1)n-1$, $\lambda = m-\alpha+1(1)m$,

$|e_{r+sm+\lambda-1}| \leq \|e^{(q^*,q^*+1,\ \ldots\ ,m-1,m)}\|_\infty$ for $s = O(1)n-1$, $\lambda = q^*(1)m$,

so we obtain $C^{(r+sm+q-1)}h^{p_q} \leq C''\|e^{(q^*,q^*+1,\ \ldots\ ,m-1,m)}\|_\infty$ for some C'' .

Now $q^* = \min(q-\beta+1, m-\alpha+1)$ and $\pi_\upsilon = \pi_{m-\alpha+1}$ for $\upsilon = m-\alpha+2(1)m$,

so $\pi_{q^*} = \pi_{q-\beta+1} = p_q$, and so the lower bound in (6.3.2) follows , if we take $C_L^{(q^*,q^*+1,\ \ldots\ ,m-1,m)} = \frac{1}{C''} \min_{s=O(1)n-1} \{C^{(r+sm+q-1)}\}$.

The first part of this theorem gives a two-sided bound on the maximum of all the global errors , and the second part will give two-sided bounds on the error in some of the more accurate stages in certain methods .

Example 6.1.2 .

(6.3.1) gives $C_L h^2 \leq \|e\|_\infty \leq C_U h^2$. $\beta = 2$, so $q-\beta+1 = q-1$, and for $q = 3$ we have $p_3 = \pi_2 = 3$, $q^* = \min(2,3) = 2$, $\pi_{q^*} = \pi_2 = 3$.

Hence (6.3.2) gives $C_L^{(2,3)}h^3 \leq \|e^{(2,3)}\|_\infty \leq C_U^{(2,3)}h^3$ (6.3.3) .

That is , we have exactly order three convergence in the second and third stages .

Example 6.1.3 .

Again we have $C_L h^2 \leq \|e\|_\infty \leq C_U h^2$ and $C_L^{(2,3)}h^3 \leq \|e^{(2,3)}\|_\infty \leq C_U^{(2,3)}h^3$ (6.3.4) by applying theorem 6.3.1 exactly as in the previous example .

Example 6.2.13 .

(6.3.1) gives $C_L h^{p-\mu+1} \leq \|e\|_\infty \leq C_U h^{p-\mu+1}$. $\beta = 2$, so for $q = \mu + 1$, $p_q = \pi_{q-1} = p$. $q^* = \min(\mu,\mu) = \mu$, so $\pi_{q^*} = \pi_\mu = p$. Hence (6.3.2) gives $C_L^{(\mu,\mu+1)}h^p \leq \|e^{(\mu,\mu+1)}\|_\infty \leq C_U^{(\mu,\mu+1)}h^p$ (6.3.5) , and so the $(\mu-1)^{\text{th}}$. and

μ^{th}. correctors converge with order exactly p .

Example 6.1.3 is Heun's third order formula (15 , p.119) , a Runge-Kutta method for which we have obtained the correct orders of convergence at each of the stages through optimal consistency , based on the calculation only of the local truncation errors of the following three linear multistep methods based on steplength h/3 :

$$y_{i-2/3} - y_{i-1} = \frac{h}{3} f_{i-1} ,$$

$$y_{i-1/3} - y_{i-1} = \frac{2}{3} h f_{i-2/3} ,$$

$$y_i - y_{i-1} = \frac{h}{4}(f_{i-1} + 3f_{i-1/3}) .$$

Our approach not only gives two-sided error bounds , but is also far simpler than the standard means of calculating the local truncation error $y(t_i) - y(t_{i-1}) - h\phi(t_{i-1}, y(t_{i-1}))$, where $\phi(t_{i-1}, y_{i-1}) = \frac{1}{4}(k_1 + 3k_3)$ and $k_1 = f(t_{i-1}, y_{i-1})$,

$$k_2 = f(t_{i-1} + h/3 , y_{i-1} + hk_1/3) ,$$

$$k_3 = f(t_{i-1} + 2h/3 , y_{i-1} + 2hk_2/3) .$$

Example 6.3.2 .

$$y_{i-\frac{1}{2}} - y_{i-2} = \frac{h}{8}(9f_{i-1} + 3f_{i-2}) ,$$

$$y_i^* - \frac{1}{5}(28y_{i-1} - 23y_{i-2}) = \frac{h}{75}(-255f_{i-1} - 160f_{i-2} + 136f_{i-\frac{1}{2}} + 9f_{i-3}) .$$

$$y_i - y_{i-1} = \frac{h}{720}(251f_i^* + 646f_{i-1} - 264f_{i-2} + 106f_{i-3} - 19f_{i-4}) .$$

This is a 3-block explicit method with array

$$\left(\begin{array}{cccccccccccc|ccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 23/5 & 0 & 0 & -28/5 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & -19/720
\end{array}\right.$$

$$\left.\begin{array}{cccccccccccc|c}
0 & 0 & 0 & 0 & 0 & 3/8 & 0 & 0 & 9/8 & 0 & 0 & 0 & \frac{1}{2} \\
0 & 0 & 3/25 & 0 & 0 & -32/15 & 0 & 0 & -17/5 & 136/75 & 0 & 0 & 1 \\
0 & 0 & 53/360 & 0 & 0 & -11/30 & 0 & 0 & 343/360 & 0 & 251/720 & 0 & 1
\end{array}\right) .$$

For zero stability we obtain $G = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -23/5 & 0 & 0 & 28/5 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$,

which has eigenvalues 0,0,0,0,1, giving zero stability . Assuming starting values accurate of order 5 ,

$$\theta^* = (C_0'h^5, C_1'h^5, C_2'h^5, C_3'h^5, \tfrac{3}{128}h^4y^{iv}(t_{7/2}) + O(h^5), \tfrac{-31}{1200}h^5y^{v}(t_{7/2}) + O(h^6),$$
$$\tfrac{-3}{160}h^6y^{vi}(t_{7/2}) + O(h^7), \ldots)^T .$$

$A_h \in Q_{3,1}$, so by lemma 6.2.4 $A_h^{-1} \in Q_{3,1}^{(1)}$ and by direct computation of $A_h^{-1}\theta^*$ we obtain optimal consistency of orders (4,5,5) .

$B_h \in Q_{3,1}^{(2)}$, so theorem 6.2.10 gives

$$C_L h^4 \leq \|e\|_\infty \leq C_U h^4$$

and since $\beta = 2$, $p_3 = \pi_2 = 5$ and $q^* = 2$,

$$C_L^{(2,3)}h^5 \leq \|e^{(2,3)}\|_\infty \leq C_U^{(2,3)}h^5 \quad (6.3.6) .$$

Example 6.2.11 .

(6.3.1) gives $C_L h^{p-\mu+1} \leq \|e\|_\infty \leq C_U h^{p-\mu+1}$. $\beta = 2$ and $p_{\mu+1} = \pi_\mu = p$, $q^* = \min(\mu,\mu) = \mu$. So $\pi_{q^*} = \pi_\mu = p$, and (6.3.2) gives

$$C_L^{(\mu,\mu+1)}h^p \leq \|e^{(\mu,\mu+1)}\|_\infty \leq C_U^{(\mu,\mu+1)}h^p \quad (6.3.7) .$$

Example 6.3.3 .

Consider the following $P_{2/3}EP_{1/2}EP_HEC_HE$ hybrid method :

$$y_{i-1/3} - y_{i-1} = \tfrac{h}{9}(8f_{i-1} - 2f_{i-2}) ,$$
$$y_{i-1/2} - y_{i-2} = \tfrac{h}{8}(9f_{i-1} + 3f_{i-2}) ,$$
$$y_i^* - y_{i-1} = \tfrac{h}{4}(3f_{i-1/3} + f_{i-1}) ,$$
$$y_i - y_{i-1} = \tfrac{h}{6}(f_i^* + f_{i-1} + 4f_{i-1/2}) .$$

This is a 4-block explicit method with array

$$\left(\begin{array}{cccccccccccc|cccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -2/9 & 0 & 0 & 0 & 8/9 \\
0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 3/8 & 0 & 0 & 0 & 9/8 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/6
\end{array}\right.$$

$$\left.\begin{array}{cccc|c}
0 & 0 & 0 & 0 & 2/3 \\
0 & 0 & 0 & 0 & 1/2 \\
3/4 & 0 & 0 & 0 & 1 \\
0 & 2/3 & 1/6 & 0 & 1
\end{array}\right) .$$

$G = \begin{pmatrix} 0&0&0&0&1&0 \\ 0&0&0&0&0&1 \\ 0&0&0&0&0&1 \\ 0&1&0&0&0&0 \\ 0&0&0&0&0&1 \\ 0&0&0&0&0&1 \end{pmatrix}$, which has eigenvalues 0,0,0,0,0,1, giving zero stability .

Assuming starting values accurate of order 4 , we have

$$\theta^* = (C_0' h^4, C_1' h^4, \tfrac{13}{81} h^3 y'''(t_{5/3}) + O(h^4), \tfrac{3}{128} h^4 y^{iv}(t_{5/3}) + O(h^5),$$

$$\tfrac{1}{216} h^4 y^{iv}(t_{5/3}) + O(h^5), \tfrac{-1}{2880} h^5 y^{v}(t_{5/3}) + O(h^6), \ \ldots\ )^T .$$

$A_h \in Q_{4,1}$, so $A_h^{-1} \in Q_{4,1}^{(1)}$ and direct computation of $A_h^{-1}\theta^*$ we obtain optimal consistency of orders (3,4,4,4) .

$B_h \in Q_{4,1}^{(3)} \subset Q_{4,3}^{(3)}$, and theorem 6.2.10 gives

$\|e^{(1)}\|_\infty \le C_U^{(1)} h^3$ and $\|e^{(\nu)}\|_\infty \le C_U^{(\nu)} h^4$, $\nu = 2,3,4$ (6.3.8) .

Also theorem 6.3.1 gives $C_L h^3 \le \|e\|_\infty \le C_U h^3$ and

$C_L^{(2,3,4)} h^4 \le \|e^{(2,3,4)}\|_\infty \le C_U^{(2,3,4)} h^4$ (6.3.9) .

Example 6.3.4 .

The seventh example given in the concluding section of (5) is a 4-block explicit method with array

$$\left(\begin{array}{cccccccc|cccccccc|c} 0 & -4/9 & 0 & -5/9 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 5/9 & 0 & 0 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1/8 & 3/8 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1/2 & -3/2 & 2 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1/6 & 0 & 2/3 & 1/6 & 0 & 1 \end{array}\right) .$$

$G = \begin{pmatrix} 0 & 4/9 & 0 & 5/9 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$, which has eigenvalues 0,0,0,1, giving zero-stability .

Assuming starting values accurate of order four , we have

$$\theta^* = (C_0' h^4, C_1' h^4, \tfrac{-1}{324} h^3 y'''(t_{5/6}) + O(h^4) , \tfrac{1}{3456} h^4 y^{iv}(t_{5/6}) + O(h^5) ,$$
$$\tfrac{1}{108} h^4 y^{iv}(t_{5/6}) + O(h^5) , \tfrac{-1}{2880} h^5 y^{v}(t_{5/6}) + O(h^6) , \ldots)^T .$$

$A_h \in Q_{4,1}$, so $A_h^{-1} \in Q^{(1)}_{4,1}$, and by direct computation of $A_h^{-1}\theta^*$ we obtain optimal consistency of orders (3,4,4,4) .

$B_h \in Q^{(3)}_{4,1} \subset Q^{(3)}_{4,3}$, and theorem 6.2.10 gives $\|e^{(1)}\|_\infty \leq C_U^{(1)} h^3$ and $\|e^{(\nu)}\|_\infty \leq C_U^{(\nu)} h^4$, $\nu = 2,3,4$ (6.3.10) .

Also theorem 6.3.1 gives $C_L h^3 \leq \|e\|_\infty \leq C_U h^3$ and $C_L^{(2,3,4)} h^4 \leq \|e^{(2,3,4)}\|_\infty \leq C_U^{(2,3,4)} h^4$ (6.3.11) .

6.4 A Further Convergence Theorem .

Definition 6.4.1 .

An m-block method is CONSISTENT of ORDER ZERO if there exist $C_1^{(1)}, C_2^{(1)}, \ldots, C_m^{(1)}$, some , but not all of which may be zero , such that for each $y \in C^{(1)}[0,T]$,

$$\theta^*_{r+sm+\nu-1} = C^{(1)}_{\nu}[\Delta_h y'(t)]_{r+sm} h + O(h^2) \ , \ \nu = 1(1)m \ , \ s = 0(1)n-1 \ .$$

We note that a method which is consistent of order zero is not consistent at all in the sense of definition 4.1.3 .

Definition 6.4.2 .

For any positive integer m we define the vectors $\delta_\nu \in \mathbb{R}^m$, $\nu = 1(1)m$, by $[\delta_\nu]_q = \delta_{\nu q}$, $q = 1(1)m$, where $\delta_{\nu q}$ is the Krönecker delta , and we are here beginning the vector suffices at one rather than zero .

Theorem 6.4.3 .

Suppose we have an m-block semi explicit method which is zero stable and consistent of order $p \geq 0$. Suppose also that

(i) $A_\theta A_\tau^{-1} c^{(j)} = 0$, $\theta = 0(1)\tau-1$, $j = 1(1)q$, where $q \geq 1$,

(ii) $\delta^T_{m_T} A_\tau^{-1} c^{(j)} = 0$, $j = 1(1)q-1$,

(iii) $\delta^T_\nu A_\tau^{-1} c^{(1)} \neq 0$, $\nu = 1(1)m-1$,

(iv) $\delta^T_\nu A_\theta A_\tau^{-1} B_\phi = \lambda^{(\theta,\phi)}_\nu \delta^T_m$ for some constants $\lambda^{(\theta,\phi)}_\nu$, independent of h ,

$$\nu = 1(1)m \ , \ \phi = 0(1)\tau \ , \ \theta = 0(1)\tau-1 \ ,$$

(v) δ^T_m is a left eigenvector of $A_\tau^{-1} B_\phi$, $\phi = 0(1)\tau$.

Then there exist C_L , C_U such that $C_L h^{p+1} \leq \|e\|_\infty \leq C_U h^{p+1}$ and $C_L h^{p+q'} \leq \|e^{(m)}\|_\infty \leq C_U h^{p+q'}$, where q' is an integer not less than q , assuming starting values accurate of order p+q' .

Proof .

(a) Let us first determine the orders of optimal consistency . As in the proof of theorem 5.3.3 we need only consider $D_h^{-1} \begin{pmatrix} \theta^*_r \\ \vdots \\ \theta^*_{r+nm-1} \end{pmatrix}$. For each s = 0(1)n-1 rows sm to sm+m-1 of this vector are

$$\sum_{\sigma=0}^{s-1} P_\sigma A_\tau^{-1} \sum_{j=1}^{\infty} c^{(j)} [\Delta_h y^{(p+j)}(t)]_{r+(s-\sigma-1)m} h^{p+j} + A_\tau^{-1} \sum_{j=1}^{\infty} c^{(j)} [\Delta_h y^{(p+j)}(t)]_{r+sm} h^{p+1} ,$$

assuming $y(t)$ has sufficient continuity .

Now by (i) $\sum_{\sigma=0}^{s-1} P_\sigma A_\tau^{-1} c^{(j)} = 0$ for $j = 1(1)q$, so

$$\sum_{\sigma=0}^{s-1} P_\sigma A_\tau^{-1} \sum_{j=1}^{\infty} c^{(j)} [\Delta_h y^{(p+j)}(t)]_{r+(s-\sigma-1)m} h^{p+j}$$

$$= \sum_{\sigma=0}^{s-1} P_\sigma A_\tau^{-1} \sum_{j=q+1}^{\infty} c^{(j)} [\Delta_h y^{(p+j)}(t)]_{r+(s-\sigma-1)m} h^{p+j} = O(h^{p+q}) ,$$

if we define $a = O(h^b)$ as $h \to 0$ for vector a and scalar b to mean that $\|a\|_\infty = O(h^b)$ as $h \to 0$.

(iii) implies that the first $m-1$ components of $A_\tau^{-1} c^{(1)}$ are non-zero . Elements sm to $sm+m-2$ of $D_h^{-1} \begin{pmatrix} \theta^*_r \\ \vdots \\ \theta^*_{r+nm-1} \end{pmatrix}$ are elements sm to $sm+m-2$ of $O(h^{p+q}) + A_\tau^{-1} c^{(1)} [\Delta_h y^{(p+1)}(t)]_{r+sm} h^{p+1} + O(h^{p+2})$, and so for $s = O(1)n-1$ and $\nu = 1(1)m-1$ there exist $C^{(r+sm+\nu-1)}$ such that $\left|[A_h^{-1}\theta^*]_{r+sm+\nu-1}\right| = C^{(r+sm+\nu-1)} h^{p+1}$.

(ii) implies that the $m^{th.}$ component of $A_\tau^{-1} c^{(j)}$ is zero for $j = 1(1)q-1$, hence the $m^{th.}$ component of $A_\tau^{-1} \sum_{j=1}^{\infty} c^{(j)} [\Delta_h y^{(p+j)}(t)]_{r+sm} h^{p+j}$ is $O(h^{p+q})$.

Therefore for each $s = O(1)n-1$, $[A_h^{-1}\theta^*]_{r+sm+m-1} = O(h^{p+q})$, and there exist $C^{(r+sm+m-1)}$ and some $q' \geq q$ such that

$$\left|[A_h^{-1}\theta^*]_{r+sm+m-1}\right| = C^{(r+sm+m-1)} h^{p+q'} .$$

So optimal consistency is of orders $(p+1, p+1, \ldots, p+1, p+q')$.

The result $C_L' h^{p+1} \leq \|e\|_\infty \leq C_U' h^{p+1}$ for some C_L' , C_U' now follows by (6.1.1) .

(b) B_h has the form (2.1.2) and so we may write

$$B_h = \left(\begin{array}{c|c} 0_r & 0 \\ \hline V_h & U_h \end{array}\right) \quad \text{where} \quad U_h = \begin{pmatrix} B_\tau & & & \\ \vdots & \ddots & & \\ B_0 & \ddots & \ddots & \\ & \ddots & \ddots & B_\tau \\ & & B_0 \cdots & B_\tau \end{pmatrix} .$$

$$\text{Hence } A_h^{-1}B_h = \left(\begin{array}{c|c} I_r & 0 \\ \hline -D_h^{-1}E_h & D_h^{-1} \end{array}\right)\left(\begin{array}{c|c} 0_r & 0 \\ \hline V_h & U_h \end{array}\right)$$

$$= \left(\begin{array}{c|c} 0_r & 0 \\ \hline D_h^{-1}V_h & D_h^{-1}U_h \end{array}\right) .$$

Let us now investigate $D_h^{-1}U_h$.

$$D_h^{-1}U_h = \begin{pmatrix} A_\tau^{-1} & & & \\ P_0A_\tau^{-1} & A_\tau^{-1} & & \\ \vdots & \ddots & \ddots & \\ P_{n-2}A_\tau^{-1} \cdots & & P_0A_\tau^{-1} & A_\tau^{-1} \end{pmatrix} \begin{pmatrix} B_\tau & & & \\ \vdots & \ddots & & \\ B_0 & \ddots & \ddots & \\ & \ddots & \ddots & \\ & & B_0 \cdots & B_\tau \end{pmatrix} .$$

Consider rows sm to $sm+m-1$, columns σm to $\sigma m+m-1$ of $D_h^{-1}U_h$ for $\sigma \le s$.

I. When $\sigma = s$ we have $A_\tau^{-1}B_\tau$.

II . When $s-\tau \le \sigma \le s-1$ we have $A_\tau^{-1}B_{\tau-s+\sigma} + \sum_{\varrho=0}^{s-\sigma-1} P_\varrho A_\tau^{-1} B_{\varrho-s+\sigma+\tau+1}$.

III . When $\sigma \le s-\tau-1$ we have $\sum_{\varrho=s-\sigma-\tau-1}^{s-\sigma-1} P_\varrho A_\tau^{-1} B_{\varrho-s+\sigma+\tau+1}$.

Now (iv) implies that for $\theta = O(1)\tau-1$, $\phi = O(1)\tau$,

$$A_\theta A_\tau^{-1} B_\phi = \begin{pmatrix} 0 \cdots 0 & \lambda_1^{(\theta,\phi)} \\ \vdots \quad \vdots & \vdots \\ 0 \cdots 0 & \lambda_m^{(\theta,\phi)} \end{pmatrix} .$$

(v) implies that for each $\beta = O(1)\tau$ there exists $\lambda^{(\beta)}$ such that $\delta_m^T A_\tau^{-1} B_\beta = \lambda^{(\beta)} \delta_m^T$, which in turn implies that the $m^{th.}$ row of $A_\tau^{-1} B_\beta$ is $(0, \ldots, 0, \lambda^{(\beta)})$.

Therefore for each s , row $sm+m-1$ of $D_h^{-1} U_h$ is of the form

$$(0 , \ldots , 0 , x , 0 , \ldots , 0 , x , \cdots\cdots , 0 , \ldots , 0 , x , 0 , \cdots\cdots , 0)$$

zero$^{th.}$ term; $(m-1)^{th.}$ term; $(2m-1)^{th.}$ term; $(sm+m-1)^{th.}$ term

where the crosses denote non-zero terms . Now

$$|e_i| \leq |[A_h^{-1}\theta^*]_i| + hL \sum_{j=0}^{i} |(A_h^{-1}B_h)_{ij}| |e_j| \quad \text{for all } i .$$

Therefore for each s

$$|e_{r+sm+m-1}| \leq C^{(r+sm+m-1)} h^{p+q'} + hM_1 \sum_{j=0}^{r-1} C'_j h^{p+q'} + hM_1 \sum_{\sigma=0}^{s} |e_{r+\sigma m+m-1}| .$$

Hence there exists C such that

$$|e^{(m)}_{r+s}| \leq Ch^{p+q'} + hM_1 \sum_{\sigma=0}^{s} |e^{(m)}_{r+\sigma}| , \quad s = O(1)n-1 ,$$

and lemma 4.4.3 gives $\|e^{(m)}\|_\infty \leq C''_U h^{p+q'}$ for some C''_U .

Now for all i ,

$$|[A_h^{-1}\theta^*]_i| \leq |e_i| + hL \sum_{j=0}^{i} |(A_h^{-1}B_h)_{ij}| |e_j| \quad \text{by (5.1.3)} .$$

Therefore for each s ,

$$C^{(r+sm+m-1)} h^{p+q'} \leq |e_{r+sm+m-1}| + hM_1 \sum_{\sigma=0}^{s} |e_{r+\sigma m+m-1}| + hM_1 \sum_{j=0}^{r-1} |e_j| .$$

So there exists C* such that

$$C^* h^{p+q'} \leq |e^{(m)}_{r+s}| + hM_1 \sum_{j=0}^{r-1} |e^{(m)}_j| + hM_1 \sum_{\sigma=0}^{s} |e^{(m)}_{r+\sigma}|$$

$\leq [1 + hM_1(r+n)]\|e^{(m)}\|_\infty \leq (1 + M')\|e^{(m)}\|_\infty$ for some M' .

Hence $\|e^{(m)}\|_\infty \geq C_L'' h^{p+q'}$ where $C_L'' = C'/(1 + M')$.

So the result follows , taking $C_L = \min(C_L', C_L'')$ and $C_U = \max(C_U', C_U'')$. Note that the order of convergence in the $m^{th.}$ stage is exactly the same as the order of optimal consistency , that is $p+q'$.

We may apply this result to various examples .

Example 2.2.4 .

$G = 1$, so we have zero stability . Consistency is of order one

and $c^{(1)} = \begin{bmatrix} 1/6 \\ -1/6 \end{bmatrix}$, $c^{(2)} = \begin{bmatrix} -2/3 \\ 2/3 \end{bmatrix}$, $c^{(3)} = \begin{bmatrix} 29/72 \\ -77/72 \end{bmatrix}$.

$$A_h = \begin{pmatrix} 1 & & & & & \\ 0 & 1 & & & & \\ 0 & 0 & 1 & & & \\ 0 & 0 & -1 & 1 & & \\ 0 & 0 & 0 & -1 & 1 & \\ 0 & 0 & 0 & 0 & -1 & 1 \\ & & & \cdots & & \end{pmatrix} \quad , \text{ so } A_h^{-1} = \begin{pmatrix} 1 & & & & & \\ 0 & 1 & & & & \\ 0 & 0 & 1 & & & \\ 0 & 0 & 1 & 1 & & \\ 0 & 0 & 1 & 1 & 1 & \\ 0 & 0 & 1 & 1 & 1 & 1 \\ & & & \cdots & & \end{pmatrix}$$

and by direct computation of $A_h^{-1}\theta^*$ we obtain optimal consistency of orders $(2,3)$. Let us take starting values accurate of order three .

$A_o A_2^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ and $A_1 A_2^{-1} = \begin{bmatrix} -1 & -1 \\ 0 & 0 \end{bmatrix}$.

(i) $A_o A_2^{-1} c^{(1)} = A_o A_2^{-1} c^{(2)} = A_1 A_2^{-1} c^{(1)} = A_1 A_2^{-1} c^{(2)} = 0$.

(ii) $A_2^{-1} c^{(1)} = (1/6, 0)^T$, so $\delta_2^T A_2^{-1} c^{(1)} = 0$.

(iii) $\delta_1^T A_2^{-1} c^{(1)} = 1/6$.

(iv) $A_o A_2^{-1} B_\phi = 0_2$ for $\phi = 0,1,2$, so $\lambda_\nu^{(0,\phi)} = 0$ for each ϕ, ν .

$A_1 A_2^{-1} B_o = \begin{bmatrix} 0 & 1/6 \\ 0 & 0 \end{bmatrix}$, $A_1 A_2^{-1} B_1 = \begin{bmatrix} 0 & -4/3 \\ 0 & 0 \end{bmatrix}$, $A_1 A_2^{-1} B_2 = \begin{bmatrix} 0 & -5/6 \\ 0 & 0 \end{bmatrix}$.

Hence $\lambda_1^{(1,0)} = 1/6$, $\lambda_1^{(1,1)} = -4/3$, $\lambda_1^{(1,2)} = -5/6$,

$\lambda_2^{(1,0)} = 0$, $\lambda_2^{(1,1)} = 0$, $\lambda_2^{(1,2)} = 0$.

(v) $A_2^{-1}B_o = \begin{pmatrix} 0 & -1/6 \\ 0 & -1/6 \end{pmatrix}$, so $\delta_2^T A_2^{-1} B_o = -\frac{1}{6}\delta_2^T$.

$A_2^{-1}B_1 = \begin{pmatrix} 1 & -5/6 \\ 0 & 4/3 \end{pmatrix}$, so $\delta_2^T A_2^{-1} B_1 = \frac{4}{3}\delta_2^T$.

$A_2^{-1}B_2 = \begin{pmatrix} 1 & 0 \\ 0 & 5/6 \end{pmatrix}$, so $\delta_2^T A_2^{-1} B_2 = \frac{5}{6}\delta_2^T$.

Hence by theorem 6.4.3 $C_L h^2 \leq \|e\|_\infty \leq C_U h^2$ and $C_L h^3 \leq \|e^{(2)}\|_\infty \leq C_U h^3$

(6.4.1) .

Example 2.2.5 .

G = 1 , so we have zero stability . Consistency is of order zero

and $c^{(1)} = \begin{pmatrix} -19/6 \\ 19/6 \end{pmatrix}$, $c^{(2)} = \begin{pmatrix} 7 \\ -7 \end{pmatrix}$, $c^{(3)} = \begin{pmatrix} -133/12 \\ 133/12 \end{pmatrix}$,

$c^{(4)} = \begin{pmatrix} 103/8 \\ -325/24 \end{pmatrix}$.

$A_h^{-1} = \begin{pmatrix} 1 & & & & & & \\ 0 & 1 & & & & & \\ 0 & 0 & 1 & & & & \\ 0 & 0 & 0 & 1 & & & \\ 0 & 0 & 0 & 1 & 1 & & \\ 0 & 0 & 0 & 1 & 1 & 1 & \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ & & & \cdots & & & \end{pmatrix}$ and by computing $A_h^{-1}\theta^*$ we obtain optimal consistency of orders (1,3) . Let us take starting values accurate of order three .

$A_o A_2^{-1} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ and $A_1 A_2^{-1} = \begin{pmatrix} -1 & -1 \\ 0 & 0 \end{pmatrix}$.

(i) $A_o A_2^{-1} c^{(1)} = A_o A_2^{-1} c^{(2)} = A_o A_2^{-1} c^{(3)} = A_1 A_2^{-1} c^{(1)} = A_1 A_2^{-1} c^{(2)} = A_1 A_2^{-1} c^{(3)} = 0$.

(ii) $A_2^{-1} c^{(1)} = (-19/6, 0)^T$, so $\delta_2^T A_2^{-1} c^{(1)} = 0$.

$A_2^{-1} c^{(2)} = (7,0)^T$, so $\delta_2^T A_2^{-1} c^{(2)} = 0$.

(iii) $\delta_1^T A_2^{-1} c^{(1)} = -19/6$.

(iv) $A_0A_2^{-1}B_\phi = 0_2$ for $\phi = 0,1,2$, so $\lambda_\nu^{(0,\phi)} = 0$ for each ϕ,ν .

$$A_1A_2^{-1}B_0 = \begin{bmatrix} 0 & 1/6 \\ 0 & 0 \end{bmatrix} , \quad A_1A_2^{-1}B_1 = \begin{bmatrix} 0 & -4/3 \\ 0 & 0 \end{bmatrix} ,$$

$$A_1A_2^{-1}B_2 = \begin{bmatrix} 0 & -5/6 \\ 0 & 0 \end{bmatrix} .$$

Hence $\lambda_1^{(1,0)} = 1/6$, $\lambda_2^{(1,0)} = 0$,

$\lambda_1^{(1,1)} = -4/3$, $\lambda_2^{(1,1)} = 0$,

$\lambda_1^{(1,2)} = -5/6$, $\lambda_2^{(1,2)} = 0$.

(v) $$A_2^{-1}B_0 = \begin{bmatrix} 1 & 1/6 \\ 0 & -1/6 \end{bmatrix} , \text{ so } \delta_2^T A_2^{-1}B_0 = -\frac{1}{6}\delta_2^T .$$

$$A_2^{-1}B_1 = \begin{bmatrix} 1 & 1 \\ 0 & 4/3 \end{bmatrix} , \text{ so } \delta_2^T A_2^{-1}B_1 = \frac{4}{3}\delta_2^T .$$

$$A_2^{-1}B_2 = \begin{bmatrix} 1 & 0 \\ 0 & 5/6 \end{bmatrix} , \text{ so } \delta_2^T A_2^{-1}B_2 = \frac{5}{6}\delta_2^T .$$

Hence by theorem 6.4.3 $C_L h \leq \|e\|_\infty \leq C_U h$ and $C_L h^3 \leq \|e^{(2)}\|_\infty \leq C_U h^3$ (6.4.2) .

Example 2.2.4 is a cyclic method which is consistent of order one , and yet convergent of order two in the first stage and order three in the second stage .

Example 2.2.5 is not consistent at all in the sense of definition 4.1.3 , and yet is convergent of order one in the first stage and order three in the second .

In both cases once the orders of optimal consistency are known we may use theorem 6.4.3 to give the correct orders of convergence .

Definition 4.1.3 of consistency is fitting for linear multistep methods , since it is able by theorem 1.2.5 to provide the correct order of convergence . When we consider general classes of m-block semi

explicit or implicit methods however , we see that it is an inadequate means of determining orders of convergence in general .

We saw in examples 1.3.1 , 2.2.1 , 3.2.1 , 5.3.4,5,6,7,8,9,10,11 that there are methods which have better convergence than consistency and theorems 4.4.2 and 4.4.4 would lead us to expect . In each of these cases however optimal consistency with theorems 5.1.2 , 5.3.1 and 5.3.3 give us the correct orders of convergence .

Moreover in example 2.2.4 we have convergence two orders better than consistency in one of the stages , and example 2.2.5 has order three convergence in one stage and is not even consistent according to definition 4.1.3 . Thus for m-block methods consistency is not a necessary condition for convergence , and indeed we shall proceed to show that it is possible to construct inconsistent methods with arbitrarily high orders of convergence in some of their stages .

We also see that optimal consistency is a superior concept to consistency for m-block methods in general , as it gives a two-sided error bound and a necessary and sufficient condition for convergence .

The importance of having a two-sided error bound of the form $C_L h^p \leq \|e\|_\infty \leq C_U h^p$ (6.4.3) rather than just the bound $\|e\|_\infty \leq C_U h^p$ (6.4.4) is that (6.4.3) gives us the exact order of convergence p . (6.4.4) on the other hand only tells us that we have order of convergence at least p , since there is always the possibility that also $\|e\|_\infty \leq C_U h^{p_1}$ for some $p_1 > p$, since $C_U h^p \leq C_U h^{p_1}$.

Proposition 6.4.4 .

For any positive integer z it is possible to construct a zero

stable m-block method which is not consistent , and yet for which $C_L h \leq \|e\|_\infty \leq C_U h$ and $C_L h^{z'} \leq \|e^{(m)}\|_\infty \leq C_U h^{z'}$ for some C_L , C_U and integer $z' \geq z$, assuming sufficient continuity in $y(t)$.

Proof .

Let us take $m = 2$ and consider a cyclic method composed of linear multistep methods with stepnumber $k = 2z - 3$.

We take

$$y_i - y_{i-1} = h(\beta_k^{(1)} f_i + \beta_{k-1}^{(1)} f_{i-1} + \ldots + \beta_o^{(1)} f_{i-k}) ,$$

$$y_{i+1} - y_i = h(\beta_k^{(2)} f_{i+1} + \beta_{k-1}^{(2)} f_i + \beta_{k-3}^{(2)} f_{i-2} + \beta_{k-5}^{(2)} f_{i-4} + \ldots + \beta_o^{(2)} f_{i-k+1}) .$$

$G = 1$, and hence we have zero stability .

$$\tau = 1 + [\tfrac{k-1}{2}] = 1 + [z-2] = z-1 .$$

$$A_\theta = 0_2 , \theta = 0(1)z-3 , A_{z-2} = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix} , A_{z-1} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} .$$

$$B_o = \begin{bmatrix} 0 & \beta_o^{(1)} \\ 0 & 0 \end{bmatrix} , B_\phi = \begin{bmatrix} \beta_{2\phi-1}^{(1)} & \beta_{2\phi}^{(1)} \\ \beta_{2\phi-2}^{(2)} & 0 \end{bmatrix} , \phi = 1(1)z-2 ,$$

$$B_{z-1} = \begin{bmatrix} \beta_k^{(1)} & 0 \\ \beta_{k-1}^{(2)} & \beta_k^{(2)} \end{bmatrix} .$$

By considering the local truncation error of the first linear multistep method we see that consistency is of order zero if

$$\sum_{j=0}^{k} \beta_{k-j}^{(1)} \neq 1 .$$

We then have $c^{(1)} = \begin{bmatrix} 1 - (\beta_o^{(1)} + \beta_1^{(1)} + \ldots + \beta_k^{(1)}) \\ 1 - (\beta_o^{(2)} + \beta_2^{(2)} + \beta_4^{(2)} + \ldots + \beta_{k-1}^{(2)} + \beta_k^{(2)}) \end{bmatrix}$

and

$$c^{(l)} = \frac{(-1)^l}{l!}\begin{pmatrix} l(\beta_{k-1}^{(1)} + 2^{l-1}\beta_{k-2}^{(1)} + 3^{l-1}\beta_{k-3}^{(1)} + \ldots + k^{l-1}\beta_o^{(1)}) - 1 \\ l((-1)^{l+1}\beta_k^{(2)} + 2^{l-1}\beta_{k-3}^{(2)} + 4^{l-1}\beta_{k-5}^{(2)} + 6^{l-1}\beta_{k-7}^{(2)} + \ldots \\ \ldots + (k-1)^{l-1}\beta_o^{(2)}) + (-1)^l \end{pmatrix}$$

for all $l \geq 2$.

We shall make use of theorem 6.4.3 to obtain the result .

(v) $$A_{z-1}^{-1}B_o = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 0 & \beta_o^{(1)} \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & \beta_o^{(1)} \\ 0 & \beta_o^{(1)} \end{pmatrix}$$, so

$\delta_2^T A_{z-1}^{-1}B_o = \beta_o^{(1)}\delta_2^T$.

For $\phi = 1(1)z-2$, $$A_{z-1}^{-1}B_\phi = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} \beta_{2\phi-1}^{(1)} & \beta_{2\phi}^{(1)} \\ \beta_{2\phi-2}^{(2)} & 0 \end{pmatrix}$$

$$= \begin{pmatrix} \beta_{2\phi-1}^{(1)} & \beta_{2\phi}^{(1)} \\ \beta_{2\phi-1}^{(1)} + \beta_{2\phi-2}^{(2)} & \beta_{2\phi}^{(1)} \end{pmatrix}$$, so

$\delta_2^T A_{z-1}^{-1}B_\phi = \beta_{2\phi}^{(1)}\delta_2^T$ if $\beta_{2\phi-1}^{(1)} + \beta_{2\phi-2}^{(2)} = 0$.

$$A_{z-1}^{-1}B_{z-1} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} \beta_k^{(1)} & 0 \\ \beta_{k-1}^{(2)} & \beta_k^{(2)} \end{pmatrix} = \begin{pmatrix} \beta_k^{(1)} & 0 \\ \beta_k^{(1)} + \beta_{k-1}^{(2)} & \beta_k^{(2)} \end{pmatrix}$$,

so $\delta_2^T A_{z-1}^{-1}B_{z-1} = \beta_k^{(2)}\delta_2^T$ if $\beta_k^{(1)} + \beta_{k-1}^{(2)} = 0$.

Then (v) holds if

$$\left.\begin{aligned} \beta_k^{(1)} + \beta_{k-1}^{(2)} &= 0 \\ \beta_{k-2}^{(1)} + \beta_{k-3}^{(2)} &= 0 \\ \beta_{k-4}^{(1)} + \beta_{k-5}^{(2)} &= 0 \\ &\vdots \\ \beta_1^{(1)} + \beta_o^{(2)} &= 0 \end{aligned}\right\} \qquad (6.4.5) \quad .$$

(iv) $A_\theta = 0_2$ for $\theta = 0(1)z-3$, so $\lambda_\nu^{(\theta,\phi)} = 0$ for all $\phi,\nu,\theta = 0(1)z-3$.

$$A_{z-2}A_{z-1}^{-1} = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ 0 & 0 \end{bmatrix} ,$$

$$\text{so } A_{z-2}A_{z-1}^{-1}B_0 = \begin{bmatrix} 0 & -\beta_0^{(1)} \\ 0 & 0 \end{bmatrix} \quad \text{and } \delta_1^T A_{z-2}A_{z-1}^{-1}B_0 = -\beta_0^{(1)}\delta_2^T ,$$

$$\delta_2^T A_{z-2}A_{z-1}^{-1}B_0 = 0 .$$

$$\text{For } \phi = 1(1)z-2 , \; A_{z-2}A_{z-1}^{-1}B_\phi = \begin{bmatrix} -\beta_{2\phi-1}^{(1)} & -\beta_{2\phi-2}^{(2)} & -\beta_{2\phi}^{(1)} \\ 0 & & 0 \end{bmatrix} ,$$

so $\delta_1^T A_{z-2}A_{z-1}^{-1}B_\phi = -\beta_{2\phi}^{(1)}\delta_2^T$ by (6.4.5) and $\delta_2^T A_{z-2}A_{z-1}^{-1}B_\phi = 0$.

$$A_{z-2}A_{z-1}^{-1}B_{z-1} = \begin{bmatrix} -\beta_k^{(1)} & -\beta_{k-1}^{(2)} & -\beta_k^{(2)} \\ 0 & & 0 \end{bmatrix} ,$$

so $\delta_1^T A_{z-2}A_{z-1}^{-1}B_{z-1} = -\beta_k^{(2)}\delta_2^T$ by (6.4.5) and $\delta_2^T A_{z-2}A_{z-1}^{-1}B_{z-1} = 0$.

Thus (iv) holds .

(i) Since $A_\theta = 0_2$ for $\theta = 0(1)z-3$, $A_\theta A_{z-1}^{-1}c^{(j)} = 0$ for all j , $\theta = 0(1)z-3$.

Now for any vector $v = (v_1,v_2)^T \in \mathbb{R}^2$, $A_{z-2}A_{z-1}^{-1}v = 0$ if and only if

$v_1 + v_2 = 0$, so $A_{z-2}A_{z-1}^{-1}c^{(1)} = 0$ if

$$(\beta_0^{(1)} + \beta_1^{(1)} + \ldots + \beta_k^{(1)}) + (\beta_0^{(2)} + \beta_2^{(2)} + \beta_4^{(2)} + \ldots + \beta_{k-1}^{(2)} + \beta_k^{(2)}) = 2 .$$

That is , by (6.4.5) , $\beta_0^{(1)} + \beta_2^{(1)} + \beta_4^{(1)} + \ldots + \beta_{k-1}^{(1)} + \beta_k^{(2)} = 2$ (6.4.6) .

For $l \geq 2$, $A_{z-2}A_{z-1}^{-1}c^{(l)} = 0$ if

$$\beta_{k-1}^{(1)} + 2^{l-1}\beta_{k-2}^{(1)} + 3^{l-1}\beta_{k-3}^{(1)} + \ldots + k^{l-1}\beta_0^{(1)} + (-1)^{l+1}\beta_k^{(2)} + 2^{l-1}\beta_{k-3}^{(2)} +$$

$$4^{l-1}\beta_{k-5}^{(2)} + 6^{l-1}\beta_{k-7}^{(2)} + \ldots + (k-1)^{l-1}\beta_0^{(2)} = \frac{1 + (-1)^{l+1}}{l} .$$

That is , due to (6.4.5) ,

$$\beta_{k-1}^{(1)} + 3^{l-1}\beta_{k-3}^{(1)} + 5^{l-1}\beta_{k-5}^{(1)} + \dots + k^{l-1}\beta_{o}^{(1)} + (-1)^{l+1}\beta_{k}^{(2)} = \frac{1 + (-1)^{l+1}}{l} \qquad (6.4.7) .$$

So (i) holds provided

$$\begin{pmatrix} 1 & 1 & 1 & 1 & \dots & 1 \\ -1 & 1 & 3 & 5 & \dots & k \\ 1 & 1 & 3^2 & 5^2 & \dots & k^2 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ (-1)^{z-1} & 1 & 3^{z-1} & 5^{z-1} & \dots & k^{z-1} \end{pmatrix} \begin{pmatrix} \beta_{k}^{(2)} \\ \beta_{k-1}^{(1)} \\ \beta_{k-3}^{(1)} \\ \vdots \\ \beta_{o}^{(1)} \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 2/3 \\ 0 \\ 2/5 \\ 0 \\ \vdots \\ \frac{1 + (-1)^{z+1}}{z} \end{pmatrix} . \qquad (6.4.8)$$

The matrix is square , since $k = 2z - 3$.

For any $v \in \mathbb{R}^2$, $\delta_2^T A_{z-1}^{-1} v = (0,1) \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = v_1 + v_2 = 0$

if and only if $v_1 + v_2 = 0$. Thus (ii) holds by virtue of (6.4.6) and (6.4.7) .

(iii) $\delta_1^T A_{z-1}^{-1} c^{(1)} = (1,0) \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 - (\beta_o^{(1)} + \beta_1^{(1)} + \dots + \beta_k^{(1)}) \\ 1 - (\beta_o^{(2)} + \beta_2^{(2)} + \beta_4^{(2)} + \dots + \beta_{k-1}^{(2)} + \beta_k^{(2)}) \end{pmatrix}$

$$= 1 - (\beta_o^{(1)} + \beta_1^{(1)} + \dots + \beta_k^{(1)}) \neq 0 ,$$

as was stipulated by consistency of order zero .

Therefore by theorem 6.4.3 we have the result provided we can

find a solution to (6.4.8) . Now the matrix in that equation is the Vandermonde matrix $V(-1,1,3,5, \ldots ,2z-3)$, which is nonsingular . Hence

$$\begin{bmatrix} \beta_o^{(2)} \\ \beta_{k-1}^{(1)} \\ \beta_{k-3}^{(1)} \\ \vdots \\ \beta_o^{(1)} \end{bmatrix} = [V(-1,1,3,5, \ldots ,k)]^{-1} \begin{bmatrix} 2 \\ 0 \\ 2/3 \\ \vdots \\ \dfrac{1 + (-1)^{z+1}}{z} \end{bmatrix} .$$

6.5 Collocation .

We saw in section 3.5 that any collocation method may be represented as an implicit Runge-Kutta method , and that Galerkin and interpolatory quadrature methods may be considered as collocation . Thus collocation , Galerkin and interpolatory quadrature methods all fall within the scope of our m-block methods , and we may analyse their convergence in a manner similar to the approach adopted so far .

For collocation methods the superconvergence result has been quite widely quoted (10,11,12,19,20) and we now give a proof of it using the notation of m-block implicit methods , considering the collocation in Runge-Kutta form .

Theorem 6.5.1 .

Suppose that for the collocation method (3.5.1) we have the quadrature formulas

$$\int_t^{t+h} \xi(\eta)d\eta = h \sum_{q=1}^{m} c_q \xi(t + a_q h) + C_Q h^{p+1} \quad (6.5.1) ,$$

$$\int_t^{t+a_\nu h} \xi(\eta)\,d\eta = h\sum_{q=1}^{m} b_{\nu q}\xi(t + a_q h) + C_Q^{(\nu)} h^{p_\nu} \quad (6.5.2)\ ,$$

for some $C_Q, C_Q^{(1)}, \ldots, C_Q^{(\mu)}$, where

$$b_{\nu q} = \int_0^{a_\nu} l_q(\eta)\,d\eta \text{ and } c_q = \int_0^1 l_q(\eta)\,d\eta\ .$$

Let $p_\nu^* = \min(p_\nu, p)$, $\nu = 1(1)\mu$, and assume a starting value accurate of order p . Then there exist $C_L, C_U, C_L^{(\nu)}, C_U^{(\nu)}$, $\nu = 1(1)\mu$, such that

$$C_L^{(\nu)} h^{p_\nu^*} \le \|e^{(\nu)}\|_\infty \le C_U^{(\nu)} h^{p_\nu^*}\ ,\ \nu = 1(1)\mu\ ,$$

$$C_L h^p \le \|e^{(\mu+1)}\|_\infty \le C_U h^p\ .$$

Proof .

Let us first obtain the orders of optimal consistency . We recall that the collocation method (3.5.1) is equivalent to the implicit Runge-Kutta method (3.3.2) , for which

$$A_h = \begin{pmatrix} 1 & & & & & & \\ -1 & 1 & & & & & \\ -1 & 0 & 1 & & & & \\ \vdots & \vdots & \vdots & \ddots & & & \\ -1 & 0 & 0 & \cdots & 0 & 1 & \\ 0 & 0 & 0 & \cdots & 0 & -1 & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \ddots \end{pmatrix} \text{ and}$$

$$B_h = \begin{pmatrix} 0 & & & & & & & & & \\ 0 & b_{11} & b_{12} & \cdots & b_{1\mu} & & & & & \\ 0 & b_{21} & b_{22} & \cdots & b_{2\mu} & & & & & \\ \vdots & \vdots & \vdots & & \vdots & & & & & \\ 0 & b_{\mu 1} & b_{\mu 2} & \cdots & b_{\mu\mu} & & & & & \\ 0 & c_1 & c_2 & \cdots & c_\mu & & & & & \\ 0 & 0 & 0 & \cdots & 0 & 0 & b_{11} & b_{12} & \cdots & b_{1\mu} \\ & & & & & & \cdots\cdots & & & \end{pmatrix}\ .$$

$$B_1 = \begin{pmatrix} b_{11} & \cdots & b_{1\mu} & 0 \\ \vdots & & \vdots & \vdots \\ b_{\mu 1} & \cdots & b_{\mu\mu} & 0 \\ c_1 & \cdots & c_\mu & 0 \end{pmatrix} \quad \text{and } B_o = 0_{\mu+1} \,. \quad \text{Let } \bar{B} = \begin{pmatrix} c_1 & c_2 & \cdots & c_\mu & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ c_1 & c_2 & \cdots & c_\mu & 0 \end{pmatrix} .$$

Then by direct computation we obtain

$$A_h^{-1}B_h = \begin{pmatrix} 0 & & & & & \\ \vdots & B_1 & & & & \\ \bar{B} & B_1 & & & & \\ \bar{B} & \bar{B} & B_1 & & & \\ \bar{B} & \bar{B} & \bar{B} & B_1 & & \\ \vdots & & & \ddots & \ddots & \\ 0\ \bar{B} & \cdots & \cdots & \bar{B} & B_1 \end{pmatrix} .$$

$\Delta_h y(t) = (y(0), y(a_1 h), y(a_2 h), \ldots, y(a_\mu h), y(h), y(h+a_1 h), \ldots)^T$, and since any Runge-Kutta method is zero stable , we have zero stability .

$m = \mu + 1$, and hence for $s = 0(1)n-1$ and $\nu = 1(1)\mu+1$,

$$[\Delta_h y(t) - hA_h^{-1}B_h\Delta_h y'(t) - A_h^{-1}g]_{sm+\nu}$$

$$= \begin{cases} y((s+a_\nu)h) - \tilde{y}_o - h\sum_{\sigma=0}^{s-1}\sum_{q=1}^{\mu} c_q y'((\sigma+a_q)h) - h\sum_{q=1}^{\mu} b_{\nu q} y'(s+a_q)h) \,, & \nu = 1(1)\mu \,, \\ y((s+1)h) - \tilde{y}_o - h\sum_{\sigma=0}^{s}\sum_{q=1}^{\mu} c_q y'(\sigma+a_q)h) \,, & \nu = \mu+1 \,, \end{cases}$$

$$= \begin{cases} \int_0^{(s+a_\nu)h} y'(\eta)\,d\eta + C_o' h^p - h\sum_{\sigma=0}^{s-1}\sum_{q=1}^{\mu} c_q y'((\sigma+a_q)h) - h\sum_{q=1}^{\mu} b_{\nu q} y'((s+a_q)h) \,, \\ \qquad\qquad \nu = 1(1)\mu \,, \\ \int_0^{(s+1)h} y'(\eta)\,d\eta + C_o' h^p - h\sum_{\sigma=0}^{s}\sum_{q=1}^{\mu} c_q y'((\sigma+a_q)h) \,, \quad \nu = \mu+1 \,. \end{cases}$$

At the $m^{th.}$ stage then we have a global application of the quadrature (6.5.1) , and at the previous stages we have a global application of (6.5.1) followed by a local application of (6.5.2) . Hence we have optimal consistency of orders $(p_1^*,p_2^*, \dots ,p_\mu^*,p)$. Note that by the Lagrange interpolation theorem $p_\nu \geq \mu + 1$, $\nu = 1(1)\mu$, and $p \geq \mu + 1$. Hence $p_\nu^* \geq \mu + 1$, $\nu = 1(1)\mu$.

Now $e = A_h^{-1}\theta^* + h A_h^{-1}B_h[f(\Delta_h t,y^*) - f(\Delta_h t,\Delta_h y(t))]$.

The Runge-Kutta method gives a discrete set of solutions represented here in the vector y^* which coincide exactly with the values taken by the continuous collocation polynomial $u(t)$ at the points $0,a_1h,a_2h, \dots ,a_\mu h,h,(1+a_1)h, \dots$, and so $y^* = \Delta_h u(t)$.

Therefore for each j ,

$$[f(\Delta_h t,y^*) - f(\Delta_h t,\Delta_h y(t))]_j = f([\Delta_h t]_j,[\Delta_h u(t)]_j) - f([\Delta_h t]_j,[\Delta_h y(t)]_j)$$

$$= f([\Delta_h t]_j,u([\Delta_h t]_j)) - f([\Delta_h t]_j,y([\Delta_h t]_j))$$

$$= u'([\Delta_h t]_j) - y'([\Delta_h t]_j)$$

$$= [\Delta_h(u'(t) - y'(t))]_j .$$

Thus $e = A_h^{-1}\theta^* + hA_h^{-1}B_h\Delta_h(u'(t) - y'(t))$. Due to the form of $A_h^{-1}B_h$ we have , for $s = 0(1)n-1$ and $\nu = 1(1)\mu$,

$$[A_h^{-1}B_h\Delta_h(u'(t) - y'(t))]_{sm+\nu} = \sum_{\sigma=0}^{s-1}\sum_{q=1}^{\mu} c_q(u'([\Delta_h t]_{\sigma m+q}) - y'([\Delta_h t]_{\sigma m+q}))$$

$$+ \sum_{q=1}^{\mu} b_{\nu q}(u'([\Delta_h t]_{sm+q}) - y'([\Delta_h t]_{sm+q}))$$

$$= \sum_{\sigma=0}^{s-1}\left[\int_{t_\sigma}^{t_\sigma+h}(u'(\eta) - y'(\eta))d\eta + C_Q h^{p+1}\right] + \int_{t_s}^{t_s+a_\nu h}(u'(\eta) - y'(\eta))d\eta$$

$$+ C_Q^{(\nu)}h^{p_\nu}$$

$$= \int_{t_o}^{t_s+a_\nu h} [u'(\eta) - y'(\eta)]d\eta + sC_Q h^{p+1} + C_Q^{(\nu)} h^{P_\nu}$$

$$= u(t_s+a_\nu h) - y(t_s+a_\nu h) - u(t_o) + y(t_o) + sC_Q h^{p+1} + C_Q^{(\nu)} h^{P_\nu}$$

$$= e_{sm+\nu} - e_o + sC_Q h^{p+1} + C_Q^{(\nu)} h^{P_\nu} .$$

Accordingly for $s = 0(1)n-1$ and $\nu = 1(1)\mu$,

$e_{sm+\nu} = [A_h^{-1}\theta^*]_{sm+\nu} + he_{sm+\nu} - hC_o' h^p + sC_Q h^{p+2} + C_Q^{(\nu)} h^{P_\nu+1}$, and so

$(1 - h)e_{sm+\nu} = [A_h^{-1}\theta^*]_{sm+\nu} - C_o' h^{p+1} + sC_Q h^{p+2} + C_Q^{(\nu)} h^{P_\nu+1}$, which gives

$$|(1 - h)e_{sm+\nu}| \le C^{(sm+\nu)} h^{p_\nu^*} + |C_o'| h^{p+1} + (n-1)|C_Q| h^{p+2} + |C_Q^{(\nu)}| h^{P_\nu+1} .$$

Therefore there exists $C_U^{(\nu)}$ such that for h sufficiently small ,

$|e_{sm+\nu}| \le C_U^{(\nu)} h^{p_\nu^*}$, and hence we obtain $\|e^{(\nu)}\|_\infty \le C_U^{(\nu)} h^{p_\nu^*}$.

Also , $|(1 - h)e_{sm+\nu}| \ge \left| C^{(sm+\nu)} h^{p_\nu^*} - |C_o' h^{p+1} - sC_Q h^{p+2} - C_Q^{(\nu)} h^{P_\nu+1}| \right|$

and therefore there exists $C_L^{(\nu)}$ such that for h sufficiently small ,

$|e_{sm+\nu}| \ge C_L^{(\nu)} h^{p_\nu^*}$, and so $\|e^{(\nu)}\|_\infty \ge C_L^{(\nu)} h^{p_\nu^*}$.

Thus we have for each $\nu = 1(1)\mu$,

$$C_L^{(\nu)} h^{p_\nu^*} \le \|e^{(\nu)}\|_\infty \le C_U^{(\nu)} h^{p_\nu^*} .$$

As for the case $\nu = \mu + 1$,

$$[A_h^{-1}B_h\Delta_h(u'(t) - y'(t))]_{sm+\mu+1} = \sum_{\sigma=0}^{s}\sum_{q=1}^{\mu} c_q[u'([\Delta_h t]_{\sigma m+q}) - y'([\Delta_h t]_{\sigma m+q})]$$

$$= \sum_{\sigma=0}^{s}\Big[\int_{t_\sigma}^{t_\sigma+h}(u'(\eta) - y'(\eta))d\eta + C_Q h^{p+1}\Big]$$

$$= \int_{t_o}^{t_{s+1}}[u'(\eta) - y'(\eta)]d\eta + (s+1)C_Q h^{p+1}$$

$$= u(t_{s+1}) - y(t_{s+1}) - u(t_o) + y(t_o) + (s+1)C_Q h^{p+1}$$

$$= e_{sm+\mu+1} - e_o + (s+1)C_Q h^{p+1}$$

$$= e_{sm+\mu+1} - C'_o h^p + (s+1)C_Q h^{p+1} .$$

Therefore $e_{sm+\mu+1} = [A_h^{-1}\theta^*]_{sm+\mu+1} + he_{sm+\mu+1} - C'_o h^{p+1} + (s+1)C_Q h^{p+2}$,

so $(1 - h)e_{sm+\mu+1} = [A_h^{-1}\theta^*]_{sm+\mu+1} - C'_o h^{p+1} + (s+1)C_Q h^{p+2}$

and $\left|(1 - h)e_{sm+\mu+1}\right| \le C^{(sm+\mu+1)}h^p + \left|C'_o\right| h^{p+1} + n\left|C_Q\right|h^{p+2}$.

Thus there exists C_U such that for h sufficiently small ,

$\left|e_{sm+\mu+1}\right| \le C_U h^p$, and hence $\left\|e^{(\mu+1)}\right\|_\infty \le C_U h^p$.

Also , $\left|(1 - h)e_{sm+\mu+1}\right| \ge \left|C^{(sm+\mu+1)}h^p - \left|C'_o h^{p+1} - (s+1)C_Q h^{p+2}\right|\right|$,

and so there exists C_L such that for h sufficiently small ,

$\left|e_{sm+\mu+1}\right| \ge C_L h^p$, giving $\left\|e^{(\mu+1)}\right\|_\infty \ge C_L h^p$.

So we have $C_L h^p \le \left\|e^{(\mu+1)}\right\|_\infty \le C_U h^p$.

CHAPTER SEVEN .

CONCLUSIONS .

7.1 The Work of Spijker .

We shall refer to (22) and (23) , two papers of M.N. Spijker . We consider the problem (1.1.1) with the Lipschitz condition (1.1.2) , to which we apply the m-block method

$$A_h y^* = hB_h f(\Delta_h t, y^*) + g \quad (7.1.1) .$$

Definition 7.1.1 .

For each $h > 0$ an m-block method may be represented by the DISCRETISATION $\Phi_h : \mathbb{R}^{r+nm-1} \rightarrow \mathbb{R}^{r+nm-1}$, where

$$\Phi_h y^* = A_h y^* - hB_h f(\Delta_h t, y^*) - g = 0 .$$

Suppose now that $\bar{y} \in \mathbb{R}^{r+nm-1}$ is a perturbed solution of the equation (7.1.1) with a perturbation $\bar{\theta} \in \mathbb{R}^{r+nm-1}$ such that $\Phi_h \bar{y} = \bar{\theta}$ Note that in the special case in which $\bar{y} = \Delta_h y(t)$, we have $\bar{\theta} = \bar{\theta}^*$, the local truncation error vector , as in definition 4.1.1 .

Definition 7.1.2 .

The discretisation Φ_h is STABLE with respect to the functional ψ if there exist positive constants K_U and h_1 such that for all vectors y^*, $\bar{y}$, $\bar{\theta} \in \mathbb{R}^{r+nm-1}$ satisfying $\Phi_h y^* = 0$ and $\Phi_h \bar{y} = \bar{\theta}$ with $h \leq h_1$, we have $\|\bar{y} - y^*\|_\infty \leq K_U \psi(\bar{\theta})$.

Definition 7.1.3 .

Let ϕ and ψ be real functionals on $\mathbb{R}^{r+nm-1}$ for each $h > 0$. Then $\phi \prec \psi$ if there exist K_U' and $h_1 > 0$ such that $\phi(\bar{\theta}) \leq K_U' \psi(\bar{\theta})$ for all $\bar{\theta} \in \mathbb{R}^{r+nm-1}$ and $h \leq h_1$.

Definition 7.1.4 .

A functional ψ on $\mathbb{R}^{r+nm-1}$ is a STABILITY FUNCTIONAL for Φ_h if Φ_h is stable with respect to ψ. A stability functional ϕ for Φ_h is a MINIMAL STABILITY FUNCTIONAL for Φ_h if all stability functionals ψ for Φ_h satisfy $\phi < \psi$.

We may show that assuming zero stability the stability functional $\psi(\bar{\theta}) = \|A_h^{-1}\bar{\theta}\|_\infty$ allows us to obtain a two-sided error bound of the form

$$K_L\psi(\bar{\theta}) \leq \|\bar{y} - y^*\|_\infty \leq K_U\psi(\bar{\theta}) \quad (7.1.2)$$

by making use of the same argument as in the proof of theorem 5.1.2 , only replacing $\Delta_h y(t)$ and θ^* by $\bar{y}$ and $\bar{\theta}$ respectively . The following theorem also shows that $\|A_h^{-1}\bar{\theta}\|_\infty$ is a minimal stability functional for Φ_h , provided Φ_h is zero stable .

Theorem 7.1.5 .

$\psi(\bar{\theta}) = \|A_h^{-1}\bar{\theta}\|_\infty$ is a minimal stability functional for Φ_h , assuming Φ_h to be zero stable .

Proof .

(i) We first show that ψ is a stability functional for Φ_h .

We have $\Phi_h y^* = 0$ (7.1.3) and $\Phi_h \bar{y} = \bar{\theta}$ (7.1.4) . Using the same analysis as in the proof of theorem 5.1.2 we have for each i

$$|\bar{y}_i - y_i^*| \leq |[A_h^{-1}\bar{\theta}]_i| + h|[A_h^{-1}B_h(f(\Delta_h t,\bar{y}) - f(\Delta_h t,y^*))]_i|$$

$$\leq \|A_h^{-1}\bar{\theta}\|_\infty + hM_o\|B_h\|_1 L \sum_{j=0}^{i+m-\nu} |\bar{y}_j - y_j^*| \text{ , where } i-r+1 \equiv \nu \pmod m ,$$

$1 \le \nu \le m$ and for an m-block semi explicit method the summation goes only as far as i . Let $M = M_o \|B_h\|_1 L$. Then by lemma 4.4.5 ,

$$|\bar{y}_i - y^*_i| \le \frac{\|A_h^{-1}\bar{\theta}\|_\infty}{1 - (m-1)hM} + \frac{hM}{1 - (m-1)hM} \sum_{j=0}^{i} |\bar{y}_j - y^*_j| ,$$

since $\frac{1}{1 - qhM} \le \frac{1}{1 - (m-1)hM}$ for $q = O(1)m-1$, if h is sufficiently small . Hence $|\bar{y}_i - y^*_i| \le \frac{\|A_h^{-1}\bar{\theta}\|_\infty}{1 - mhM} + \frac{hM}{1 - mhM} \sum_{j=0}^{i-1} |\bar{y}_j - y^*_j|$

and $\frac{1}{1 - mhM} > 0$ for $h < 1/mM$. Therefore by lemma 4.4.3

$$\|\bar{y} - y^*\|_\infty \le \frac{1}{1 - mhM}\left(hM|\bar{y}_o - y^*_o| + \|A_h^{-1}\bar{\theta}\|_\infty\right) \exp\left[\frac{M(r+nm-1)h}{1 - mhM}\right] .$$

Now by (2.1.3) there exists a positive Q independent of h such that $(r+nm-1)h = QT$, so since $|\bar{y}_o - y^*_o| = |(A_h^{-1}\bar{\theta})_o| \le \|A_h^{-1}\bar{\theta}\|_\infty$,

$\|\bar{y} - y^*\|_\infty \le K_U \psi(\bar{\theta})$ where $K_U = \frac{1 + hM}{1 - mhM} \exp\left[\frac{MQT}{1 - mhM}\right]$ (7.1.5) .

(ii) We now show that $\psi(\bar{\theta})$ is a minimal stability functional for $\bar{\Phi}_h$. By (7.1.3) and (7.1.4) , $\bar{\Phi}_h\bar{y} - \bar{\Phi}_h y^* = \bar{\theta}$, and so

$A_h(\bar{y} - y^*) - hB_h[f(\Delta_h t,\bar{y}) - f(\Delta_h t,y^*)] = \bar{\theta}$ (7.1.6) , and premultiplying through by A_h^{-1} we have

$$A_h^{-1}\bar{\theta} = \bar{y} - y^* - hA_h^{-1}B_h[f(\Delta_h t,\bar{y}) - f(\Delta_h t,y^*)] , \text{ so}$$

$$\|A_h^{-1}\bar{\theta}\|_\infty \le \|\bar{y} - y^*\|_\infty + h\|A_h^{-1}B_h[f(\Delta_h t,\bar{y}) - f(\Delta_h t,y^*)]\|_\infty .$$

Now since we have zero stability , $\max_{0\le i,j\le r+nm-1} |(A_h^{-1})_{ij}| = M_o$, and so for any $w \in \mathbb{R}^{r+nm-1}$, $h\|A_h^{-1}w\|_\infty \le hM_o(r+nm-1)\|w\|_\infty \le M_o QT\|w\|_\infty$.

Therefore $\left\| A_h^{-1}\bar{\theta} \right\|_\infty \leq \left\| \bar{y} - y^* \right\|_\infty + M_0 QT \left\| B_h [f(\Delta_h t, \bar{y}) - f(\Delta_h t, y^*)] \right\|_\infty$

$$\leq \left\| \bar{y} - y^* \right\|_\infty + M_0 QT \left\| B_h \right\|_\infty L \left\| \bar{y} - y^* \right\|_\infty$$

$$= (1 + M_0 QT \left\| B_h \right\|_\infty L) \left\| \bar{y} - y^* \right\|_\infty \quad (7.1.7) \ .$$

Suppose now that $\phi_1(\bar{\theta})$ is any stability functional for Φ_h . Then $\left\| \bar{y} - y^* \right\|_\infty \leq K_U' \phi_1(\bar{\theta})$ for some K_U' by definition 7.1.4 , and so

$\psi(\bar{\theta}) = \left\| A_h^{-1}\bar{\theta} \right\|_\infty \leq (1 + M_0 Q \left\| B_h \right\|_\infty L) K_U' \phi_1(\bar{\theta})$.

Thus $\psi < \phi_1$ and so ψ is a minimal stability functional for Φ_h .

Spijker's approach in (23) is to introduce the functional

$$\psi_s(\bar{\theta}) = \sum_{i=0}^{r-1} |\bar{\theta}_i| + \max_{r \leq i \leq r+nm-1} \left| h \sum_{j=r}^{i} \bar{\theta}_j \right|$$ and then put $A_h = P_h Q_h$ where P_h has a special form . He shows that ψ_s is a minimal stability functional for any Φ_h having the form $\Phi_h y^* \equiv P_h y^* - hB_h f(\Delta_h t, y^*) - g$, and that for an "optimally stable" Φ_h , ψ_s is a minimal stability functional subject to a certain condition on Q_h . It turns out however that the condition of "optimal stability" is too restrictive for a general discretisation Φ_h .

This is the motivation behind the work of Albrecht , which we turn to next .

7.2 The Work of Albrecht .

We refer in this section to (2) . Albrecht alludes to the work of Spijker and develops his ideas to find a minimal stability functional which is optimal , in a sense to be defined .

Definition 7.2.1 .

Let ψ be a stability functional for Φ_h with ψ independent of the right-hand side of (7.1.1) . Then ψ is OPTIMAL if the following conditions hold for all $\bar{\theta} \in \mathbb{R}^{r+nm-1}$ and all $h > 0$ sufficiently small :

(i) $\|\bar{y} - y^*\|_\infty = \psi(\bar{\theta})$ for all $\underline{\Phi}_h$ in which the right-hand side of (7.1.1) is independent of y^* .

(ii) For any $\varepsilon > 0$ there exists $\chi > 0$ such that

$$(1 - \varepsilon)\psi(\bar{\theta}) < \|\bar{y} - y^*\|_\infty < (1 + \varepsilon)\psi(\bar{\theta}) \text{ for all } f \text{ with } L < \chi .$$

Theorem 7.2.2 .

$\psi(\bar{\theta}) = \|A_h^{-1}\bar{\theta}\|_\infty$ is optimal .

Proof .

(i) For all $\underline{\Phi}_h$ in which the right-hand side of (7.1.1) is independent of y^* , (7.1.6) gives $A_h(\bar{y} - y^*) = \bar{\theta}$ and so $\|\bar{y} - y^*\|_\infty = \|A_h^{-1}\bar{\theta}\|_\infty = \psi(\bar{\theta})$.

(ii) (7.1.5) gives

$$\|\bar{y} - y^*\|_\infty \leq \frac{1 + hM}{1 - mhM}\exp\left[\frac{MQT}{1 - mhM}\right]\psi(\bar{\theta})$$

$$\leq (1 + hM)\exp(MQT)\psi(\bar{\theta}) \text{ if } 0 < h < 1/mM ,$$

and so the right-hand side of (ii) follows since $M = M_o\|B_h\|_1 L$.

(7.1.7) gives $\|\bar{y} - y^*\|_\infty \geq \frac{1}{1 + M_o Q\|B_h\|_\infty L}\psi(\bar{\theta})$

and the left-hand side of (ii) follows immediately .

The fact that $\|A_h^{-1}\bar{\theta}\|_\infty$ is an optimal stability functional is the motivation behind the term "optimal consistency" in section 5.1 . By using optimal consistency in our formulation of m-block methods we are able to analyse correctly the convergence of more methods than if we use Albrecht's $\hat{A}$-norm in his $\hat{A}$ formulation of discretisation algorithms .

Albrecht's theorem 4.3 in (2) is covered by our theorems 5.1.2 and 5.3.3 , which deal with optimal consistency of orders $p(1,1, \ldots ,1)$. In section 6 we have gone further than this , obtaining error bounds , some of which are two-sided , for methods having optimal consistency of orders $(p_1,p_2, \ldots ,p_m)$ with not all the p_ν being equal . Albrecht's theorem 4.3 could not obtain the bounds (6.3.3) to (6.3.11) inclusive , nor (6.4.1) , (6.4.2) , for instance .

Moreover our theorem 6.2.10 gives upper bounds on the various stages of the global error for an easily recognisable class of methods , which goes beyond the work of Albrecht .

To take an illustrative case , let us consider example 2.2.5 . This is an $\hat{A}$-method with

$$\begin{bmatrix} y^*_{2s+4} \\ y^*_{2s+5} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y^*_{2s+2} \\ y^*_{2s+3} \end{bmatrix} + h \begin{bmatrix} f^*_{2s+4} + f^*_{2s+3} + f^*_{2s+2} + \frac{1}{6}f^*_{2s+1} + f^*_{2s} \\ \frac{5}{6}f^*_{2s+5} + \frac{8}{6}f^*_{2s+3} - \frac{1}{6}f^*_{2s+1} \end{bmatrix}$$

and so the truncation error δ is given by

$$\delta = \begin{bmatrix} -19/6 \\ 0 \end{bmatrix} y'(t_{2s+5}) + O(h) \ .$$

The left eigenvector of $\hat{A}$ corresponding to the eigenvalue 1 is $(0,1)^T$, and so since $(0,1) \begin{bmatrix} -19/6 \\ 0 \end{bmatrix} = 0$ the condition of Albrecht's theorem 4.3 is satisfied , giving $C_L h \leq \|e\|_\infty \leq C_U h$.

Using our theorem 6.4.3 however we are able to obtain both this bound and $C_L h^3 \leq \left\|e^{(2)}\right\|_\infty \leq C_U h^3$.

7.3 Comparison with Some Results of Stetter .

In this section we refer to Stetter's book (24) . It appears that Stetter alludes just once to the possibility of two-sided error bounds , mentioning on p.84 that these can be obtained using Spijker's norm . As Albrecht has pointed out however , there are ways of examining convergence which are superior to this .

Generally , Stetter is concerned only with one-sided bounds on the error , defining convergence of order p at the $i^{th.}$ stage to hold when $e_i = O(h^p)$ as $h \rightarrow 0$.

Stetter's theorems 4.3.6 and 4.3.7 on pp. 224 and 226 deal with cyclic methods , giving a criterion which if fulfilled gives convergence one order better than consistency . He adopts the approach of Donelson and Hansen (9) . For an m-cyclic linear k-step method the m-cyclic linear mk-step auxiliary method is constructed and it is proved by Stetter that the order of convergence of the original method is the same as that of its auxiliary . This approach appears to be less concise and natural than ours , which constructs no auxiliary method and which for cyclic methods gives the exact order of convergence to be that of optimal consistency .

Consider for instance example 2.2.3 . This is a 3-cyclic 4-step method which we saw in section 5.3 to be consistent of order three and yet convergent of order four . The auxiliary method is a 3-cyclic 12-step method and to derive it and then determine its local truncation error to find its order of consistency is a cumbersome process .

In his section 5.2 Stetter studies predictor-corrector methods

and in theorem 5.2.3 on pp.288,289 obtains a result which gives the orders of convergence at the various stages in such schemes . The same is achieved also by our theorem 6.2.10 , as is seen in examples 6.2.11 and 6.2.13 . Our theorem 6.2.10 however is more general , as it applies to all methods which have $A_h \varepsilon Q_{m,1}$ and $B_h \varepsilon Q^{(\beta)}_{m,\alpha}$, for instance examples 6.1.2 , 6.1.3 , 6.3.2 , 6.3.3 and 6.3.4 . Moreover our theorem 6.3.1 gives in certain cases two-sided error bounds such as (6.3.3) to (6.3.7) inclusive , (6.3.9) and (6.3.11) .

7.4 The Work of Skeel .

We refer in this section to Skeel's paper (21) . He formulates his methods in the form

$$u_n = Su_{n-1} + h\,(t_{n-1},u_{n-1};h) \quad (7.4.1) \ , \ 1 \leq n \leq N \ , \ u_n \text{ and } u_{n-1} \varepsilon \mathbb{R}^k \ , \ S \varepsilon \mathbb{R}^k x \mathbb{R}^k \text{ for some integer } k \ .$$

Each of the methods which he quotes can be represented as an m-block method . Now an m-block method has a unique representation , whereas this is not the case for Skeel's formulation above . For instance , Skeel's example 2 on pp.667,668 is a predictor- corrector scheme in PEC mode , and he gives two alternative formulations of it .

Or we could look at his example 3 on p.668 , which in our notation has the array

$$\left(\begin{array}{cccccc|cccccc|c} 0 & 0 & -1 & 1 & 0 & 0 & -1/4 & 0 & 3/4 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & -1 & 0 & 1 & 0 & 1 & 0 & -2 & 2 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 1/6 & 2/3 & 1/6 & 0 & 1 \end{array}\right) \quad (7.4.2) \ .$$

Skeel's representation however lacks this compactness :

$$\begin{pmatrix} u_n \\ hu'_{n-1/2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} u_{n-1} \\ hu'_{n-3/2} \end{pmatrix} + h \begin{pmatrix} 1/6 & 2/3 & 1/6 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} f_{n-1} \\ f_{n-1/2} \\ f_n^p \end{pmatrix}$$

where $f_{n-1} = f(t_{n-1}, u_{n-1})$,

$$f_{n-1/2} = f(t_{n-1/2},\ u_{n-1} + \tfrac{3}{4}hf_{n-1} - \tfrac{1}{4}hu'_{n-3/2}) ,$$

$$f_n^p = f(t_n, u_{n-1} + 2hf_{n-1/2} - 2hf_{n-1} + hu'_{n-3/2}) \quad (7.4.3) .$$

Moreover this formulation is not unique , as we could equally well write

$$\begin{pmatrix} u_{n-3/2} \\ u_{n-1}^p \\ u_{n-1} \\ u_{n-1/2} \\ u_n^p \\ u_n \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_{n-5/2} \\ u_{n-2}^p \\ u_{n-2} \\ u_{n-3/2} \\ u_{n-1}^p \\ u_{n-1} \end{pmatrix}$$

$$+ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -1/4 & 0 & 3/4 & 0 & 0 & 0 \\ 1 & 0 & -2 & 2 & 0 & 0 \\ 0 & 0 & 1/6 & 2/3 & 1/6 & 0 \end{pmatrix} \begin{pmatrix} f_{n-3/2} \\ f_{n-1}^p \\ f_{n-1} \\ f_{n-1/2} \\ f_n^p \\ f_n \end{pmatrix} \quad (7.4.4) .$$

Our representation $A_h y^* = hB_h f^* + g$ is essentially linear and gives a natural and unique formulation of any method which can be represented in this way . (7.4.3) is a more complicated representation and the

question may be asked , by which criteria in Skeel's formulation do we decide which of the representations is the best .

Definition 7.4.1 .

If for each n the vector d_n denotes the local discretisation error then a method is QUASI-CONSISTENT of order p if

$$E = \lim_{n\to\infty} S^n \text{ and } \max_n \left|d_n\right| = O(h^p) \text{ and } \max_n \left|E(d_o + d_1 + \ldots + d_n)\right| = O(h^p) .$$

E is a k x k matrix and we define the $(N+1)k \times (N+1)k$ matrix [E] by

$$[E] = \begin{pmatrix} I_k & & & \\ -E & I_k & & 0 \\ & -E & I_k & \\ & 0 & \ddots & \ddots \\ & & & -E \quad I_k \end{pmatrix} \text{ , which gives}$$

$$[E]^{-1} = \begin{pmatrix} I_k & & & \\ E & I_k & & \\ E & E & I_k & \\ \vdots & \ddots & \ddots & \ddots \\ E & \ldots & E & I_k \end{pmatrix} \text{ , since by definition E is idempotent .}$$

Skeel takes various stability concepts , including a "moderately strict root condition" (MSRC) , and obtains his theorem 3.6 , which says that given a method satiafying the MSRC there exist positive constants c and C such that for sufficiently small h and any $R \in \mathbb{R}^{(N+1)k}$,

$c\left\|[E]^{-1}R\right\|_\infty \le \left\|\hat{U} - U\right\|_\infty \le C\left\|[E]^{-1}R\right\|_\infty$, where $\left\|\hat{U} - U\right\|_\infty = \max_{0\le n\le N} \left|u_n - \hat{u}_n\right|$,

u_n being the solution of (7.4.1) and $\hat{u}_n$ being a perturbed solution with perturbation r_n and $R = (r_o, r_1, \ldots, r_n)^T$.

Thus $\|[E]^{-1}R\|_\infty$ is a minimal stability functional , but as Albrecht (2) points out , it is not optimal .

Skeel uses the above result to show that a method satisfying the MSRC is convergent of order p if and only if it is quasi- consistent of order p , and he gives four examples of applications of this . First he considers his example 2 on pp.667,668 and shows that although the method is consistent of order three , it is quasi-consistent and hence convergent of order four . Considered as a 2-block method it has the array

$$\left[\begin{array}{cccccc|cccccc|c} 0 & 0 & 0 & -1 & 1 & 0 & -1/2 & 0 & 3/2 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 1 & -1/12 & 0 & 2/3 & 0 & 5/12 & 0 & 1 \end{array}\right] ,$$

and although we also get third order consistency , we see that this is just a special case of example 5.3.6 and so we get the "improved" convergence of order four .

Skeel's example 4 on p.668 is just our example 5.3.7 , and we have third order consistency but obtain fourth order convergence as he does . On p.680 he also quotes our example 1.3.1 , for which he obtains fourth order convergence as we do .

When it comes to Skeel's example 3 on p.668 , he obtains truncation error

$$d_n = h^4 \begin{Bmatrix} 0 \\ -\frac{5}{96}\frac{\partial f}{\partial y}(t_n, y_n)y_n'' \end{Bmatrix} + O(h^5)$$

if the method is given the formulation (7.4.3) , and he shows that quasi-consistency is of order four .

When we write this example as a 3-block method (7.4.2) we obtain our local truncation error by considering that of each of the hybrid methods

$$y_{i-1/2} - y_{i-1} = \frac{1}{4}h(3f_{i-1} - f_{i-3/2}) ,$$

$$y_i - y_{i-1} = h(2f_{i-1/2} - 2f_{i-1} + f_{i-3/2}) ,$$

$$y_i - y_{i-1} = \frac{1}{6}h(f_i + 4f_{i-1/2} + f_{i-1}) .$$

We are unable by this means to show fourth order convergence even using optimal consistency . This is because the m-block representation is entirely linear , whereas the method under consideration is non-linear , being essentially of Runge-Kutta form . By looking at the method in Skeel's way we are forced to perform complicated computations in order to derive the local truncation error , as is the case for example with an explicit Runge-Kutta method written in the form $y_i - y_{i-1} = h\phi(t_{i-1}, y_{i-1})$ where ϕ is a non-linear function . Thus the simplicity of looking at the local truncation error through just the individual hybrid methods is lost .

Moreover if we choose the formulation (7.4.4) of the method we obtain quasi-consistency only of order one , and so the effectiveness of Skeel's convergence theorem in giving the correct order of convergence for a particular method depends on the way chosen to represent it .

One of the advantages of our approach is the comparatively small amount of work needed to calculate the local truncation errors , as we saw with Heun's third order Runge-Kutta formula , example 6.1.3 . For this method we are able to obtain the bounds

$$\|e^{(1)}\|_\infty \leq C_U^{(1)} h^2 , \quad \|e^{(2)}\|_\infty \leq C_U^{(2)} h^3 , \quad \|e^{(3)}\|_\infty \leq C_U^{(3)} h^3 ,$$

$$C_L h^2 \leq \|e\|_\infty \leq C_U h^2 , \quad C_L^{(2,3)} h^3 \leq \|e^{(2,3)}\|_\infty \leq C_U^{(2,3)} h^3$$

by using theorems 6.2.10 and 6.3.1 , and the considering the local truncation errors of just the individual hybrid methods .

7.5 The Work of Cooper .

In this section we consider Cooper's approach as presented in (5) . He looks at linear methods , using the (A,B) formulation of Butcher in (4) . The class of methods of this kind is identical to our class of m-block methods .

To illustrate Cooper's concept of consistency we consider an (A,B) method in our m-block formulation with $A_h y^* = hB_h f^* + g$. We define the vector $z \in \mathbb{R}^{r+nm-1}$ by

$$z = (I_{r+nm-1} - A_h)\Delta_h y(t) + hB_h f(\Delta_h t, z) + g ,$$

where we may stipulate without loss of generality that A_h is lower triangular and has every element on the diagonal equal to one .

For each $s = 0(1)n-1$ the components $r + sm + \nu - 1$, $\nu = 1(1)m$ of $\Delta_h y(t) - z$ give the errors at each stage when all $y^*_{r+\sigma m+\nu-1}$ are assumed to have exact solution values for $\sigma = 0(1)s-1$, $\nu = 1(1)m$. It is this vector $\Delta_h y(t) - z$ from which Cooper obtains the order of consistency rather than our approach of examining the linear expression θ^* in definition 4.1.1 . Although $\Delta_h y(t) - z$ provides us with more information than θ^* , the calculation of the former vector presents more problems than that of the latter , and the algebraic conditions for consistency presented in Cooper's lemma 4 and theorem 5 are complicated .

In his final section Cooper looks at nine examples to which his theory is applied .

His first , second and sixth examples are our examples 5.3.8 , 5.3.9 and 5.3.10 respectively and he obtains the orders of convergence using his non-linear consistency concept as we do using optimal consistency .

Cooper's seventh example is our example 6.3.4 for which we obtained the bounds

$$\|e^{(1)}\|_\infty \le C_U^{(1)} h^3 \ , \ \|e^{(\nu)}\|_\infty \le C_U^{(\nu)} h^4 \ , \ \nu = 2,3,4 \ ,$$

$$C_L h^3 \le \|e\|_\infty \le C_U h^3 \ , \ C_L^{(2,3,4)} h^4 \le \|e^{(2,3,4)}\|_\infty \le C_U^{(2,3,4)} h^4 \ ,$$

giving the same orders of convergence as he does .

Note that in these examples we get two-sided error bounds , whereas all of Cooper's bounds are only one-sided .

Cooper's third example is a 3-block explicit method with array

$$\left(\begin{array}{ccccccccc|ccccccccc|c}
0 & 0 & -1/4 & 0 & 0 & -3/4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3/4 & 0 & 0 & 0 & 1/2 \\
0 & 0 & 1 & 0 & 0 & -2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -2 & 2 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1/6 & 2/3 & 1/6 & 0 & 1
\end{array}\right)$$

It is optimally consistent of orders (3,3,4) and so by (6.1.1)

$$C_L h^3 \le \|e\|_\infty \le C_U h^3 \ .$$

We are unable however to obtain fourth order convergence which Cooper's analysis shows in the third stage .

His fourth example is a 4-block explicit method with array

$$\left(\begin{array}{cccccccccccc|c}
0 & 0 & 0 & -11/27 & 0 & 0 & 0 & -16/27 & 1 & 0 & 0 & 0 & \\
0 & 0 & 0 & 20/27 & 0 & 0 & 0 & -47/27 & 0 & 1 & 0 & 0 & \\
0 & 0 & 0 & -23/10 & 0 & 0 & 0 & 13/10 & 0 & 0 & 1 & 0 & \\
0 & 0 & 0 & -1/49 & 0 & 0 & 0 & -48/49 & 0 & 0 & 0 & 1 &
\end{array}\right.$$

$$\left.\begin{array}{cccccccccccc|c}
0 & 0 & 0 & 4/27 & 0 & 0 & 0 & 16/27 & 0 & 0 & 0 & 0 & 1/3 \\
0 & 0 & 0 & -7/27 & 0 & 0 & 0 & -22/27 & 1 & 0 & 0 & 0 & 2/3 \\
0 & 0 & 0 & 61/80 & 0 & 0 & 0 & 71/20 & -189/80 & 27/20 & 0 & 0 & 1 \\
0 & 0 & 0 & 1/210 & 0 & 0 & 0 & 4/21 & 27/98 & 108/245 & 16/147 & 0 & 1
\end{array}\right) \ .$$

It is optimally consistent of orders (4,4,4,5) and so by (6.1.1)

$$C_L h^4 \le \|e\|_\infty \le C_U h^4 \ ,$$

but we are unable to show the sixth order convergence which Cooper

demonstrates in the fourth stage .

His eighth example is a 3-block explicit method with array

$$\left(\begin{array}{cccccc|cccccc|c} 0 & 0 & -1 & 1 & 0 & 0 & -1/4 & 0 & 3/4 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & -1 & 0 & 1 & 0 & 1 & 0 & -2 & 2 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 1/6 & 2/3 & 1/6 & 0 & 1 \end{array}\right)$$

It is optimally consistent of orders (3,3,4) and so as before

$$C_L h^3 \leq \|e\|_\infty \leq C_U h^3 ,$$

but we are unable to show the fourth order convergence which Cooper demonstrates in the third stage .

His ninth example is a 4-block implicit method with array

$$\left(\begin{array}{cccccccc|cccccccc|c} -5/81 & 0 & 0 & -76/81 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -14/27 & 0 & 1/9 & 0 & 0 & -\frac{1}{2} \\ -1/9 & 0 & 0 & -8/9 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 2/3 & 0 & 0 & 0 & 0 & \frac{1}{2} \\ 4/9 & 0 & 0 & -13/9 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -5/3 & 0 & 2 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1/6 & 0 & 2/3 & 1/6 & 0 & 1 \end{array}\right)$$

It is optimally consistent of orders (3,3,3,4) and so again we have

$$C_L h^3 \leq \|e\|_\infty \leq C_U h^3 ,$$

but we are unable to show fourth order convergence at the fourth stage as Cooper does . Cooper however points out that computations indicate fourth order convergence in the first stage , and like us he is unable to demonstrate this .

As we might have expected then Cooper's analysis gives correct orders of convergence for some methods which are beyond the scope of our approach as it now stands , because unlike Θ^* the vector $\Delta_h y(t) - z$ contains non-linear information . The advantage of our approach however is the comparative simplicity of the determination of the orders of optimal consistency , and in any case there are methods , such as the last one considered , for which neither Cooper nor we are able to obtain theoretically all the correct orders of convergence .

BIBLIOGRAPHY .

1. P. Albrecht : "On the Order of Composite Multistep Methods for Ordinary Differential Equations ."
   Numer. Math. 29 , 381-396 (1978) .
2. P. Albrecht : "Explicit , Optimal Stability Functionals and Their Application to Cyclic Discretisation Methods ."
   Comp. 19 , 233-249 (1978) .
3. J. C. Butcher : "A Modified Multistep Method for the Numerical Integration of Ordinary Differential Equations ."
   JACM 12 , 124-135 (1965) .
4. J. C. Butcher : "On the Convergence of Numerical Solutions to Ordinary Differential Equations ."
   Math. Comp. 20 , 1-10 (1966) .
5. G. J. Cooper : "The Order of Convergence of General Linear Methods for Ordinary Differential Equations ."
   SIAM J. Numer. Anal. 15 , No. 4 , 634-661 (1978) .
6. G. Dahlquist : "Convergence and Stability in the Numerical Integration of Ordinary Differential Equations ."
   Math. Scand. 4 , 33-53 (1956) .
7. G. Dahlquist : "A Special Stability Problem for Linear Multistep Methods ."
   BIT 3 , 27-43 (1963) .
8. G. Dahlquist : "Some Properties of Linear Multistep and One-Leg Methods for Ordinary Differential Equations ."
   Working Papers for the 1979 Signum Meeting on Numerical Ordinary Differential Equations , 1-1 to 1-4 .

9. J. Donelson , E. Hansen : "Cyclic Composite Multistep Predictor-Corrector Methods ."

SIAM J. Numer. Anal. 8 , 137-157 (1971) .

10. A. Guillon , F. L. Soulé : "La Resolution Numerique des Problèmes Differentielles aux Conditions Initials par des Méthodes de Collocation ."

RAIRO 3 , 17-44 (1969) .

11. B. L. Hulme : "One-Step Piecewise Polynomial Galerkin Methods for Initial Value Methods ."

Math. Comp. 26 , 415-426 (1972) .

12. B. L. Hulme : "Discrete Galerkin and Related One-Step Methods for Ordinary Differential Equations ."

Math. Comp. 26 , 881-891 (1972) .

13 J. J. Kohfeld , G. T. Thompson : "Multistep Methods with Modified Predictors and Correctors ."

JACM 14 , 155-166 (1967) .

14. K. S. Kunz : "Numerical Analysis ." First Edition .

Mc.Graw-Hill (1957) .

15. J. D. Lambert : "Computational Methods in Ordinary Differential Equations ."

John Wiley and Sons (1973) .

16. P. Lancaster : "Theory of Matrices ."

New York and c. (1969) .

17. S. Mc.Kee , N. Pitcher : "On the Convergence of Advanced Linear Multistep Methods ."

BIT 19 , 476-481 (1979) .

18. S. Mc.Kee : "The Linear Algebra of Discretisation Methods ."
Submitted for publication .

19. S. P. Norsett , G. Wanner : "The Real-Pole Sandwich for Rational Approximations and Oscillation Equations ."
BIT 19 , 79-94 (1979) .

20. S. P. Norsett : "Collocation and Perturbed Collocation Methods ."
Proceedings , Dundee conf. , Lec. Notes in Mathematics 773 (1979) .

21. R. Skeel : "Analysis of Fixed-Stepsize Methods ."
SIAM J. Numer. Anal. 13 , 271-300 (1976) .

22. M. N. Spijker : "On the Structure of Error Estimates for Finite-Difference Methods ."
Numer. Math. 18 , 73-100 (1971) .

23. M. N. Spijker : "On the Possibilty of Two-Sided Error Bounds in the Numerical Solution of Initial Value Problems ."
Numer. Math. 26 , 271-300 (1976) .

24. H. J. Stetter : "Analysis of Discretisation Methods for Ordinary Differential Equations ."
Berlin-Heidelberg-New York : Springer 1973 .

25. J. H. Wilkinson : "The Algebraic Eigenvalue Problem ."
Clarendon Press , Oxford (1965) .

26. J. Williams , F. de Hoog : "A Class of A-Stable Advanced Multistep Methods ."
Math. Comp. 28 , 163-177 (1974) .

APPENDIX

Here we verify computationally the orders of convergence claimed in the various methods we have quoted .

We consider the differential equation (1.1.1) with $f(t,y(t)) = -y(t)$ and $y_0 = 1$, which has the solution $y(t) = \exp(-t)$.

Using each particular method we integrate this using two steplengths h_1 and h_2 in turn and compare the computed solutions at corresponding points to determine the order of convergence . In the tables below the errors in the computed solutions are given at certain points for the pre-assigned h_1 and h_2 , and the orders of convergence at each of the stages are deduced .

Example 1.3.1 .

	$1^{st.}$ stage $t = 4$	$2^{nd.}$ stage $t = 4.1$
$h_1 = 1/10$	-2.6×10^{-7}	-1.6×10^{-7}
$h_2 = 1/30$	-3.1×10^{-9}	-2.0×10^{-9}
Orders of convergence	4	4

Example 2.2.1 .

	$1^{st.}$ stage $t = 1.9$	$2^{nd.}$ stage $t = 2$
$h_1 = 1/10$	-1.7×10^{-3}	-9.0×10^{-4}
$h_2 = 1/30$	-1.9×10^{-4}	-1.0×10^{-4}
Orders of convergence	2	2

Example 2.2.2 .

	$1^{st.}$ stage $t = 3.8$	$2^{nd.}$ stage $t = 3.9$	$3^{rd.}$ stage $t = 4$
$h_1 = 0.1$	-6.4×10^{-7}	-4.9×10^{-7}	-5.7×10^{-7}
$h_2 = 0.01$	5.4×10^{-11}	5.9×10^{-11}	4.4×10^{-11}
Orders of convergence	4	4	4

Example 2.2.3 .

	$1^{st.}$ stage $t = 2.9$	$2^{nd.}$ stage $t = 2.95$	$3^{rd.}$ stage $t = 3$
$h_1 = 0.05$	-2.6×10^{-6}	-3.0×10^{-6}	-1.5×10^{-6}
$h_2 = 0.005$	-2.0×10^{-10}	-1.8×10^{-10}	-1.2×10^{-10}
Orders of convergence	4	4	4

Example 2.2.4 .

	$1^{st.}$ stage $t = 0.95$	$2^{nd.}$ stage $t = 1$
$h_1 = 1/20$	1.6×10^{-4}	-1.4×10^{-5}
$h_2 = 1/60$	1.8×10^{-5}	-5.6×10^{-7}
Orders of convergence	2	3

Example 2.2.5 .

	$1^{st.}$ stage $t = 0.9$	$2^{nd.}$ stage $t = 0.95$
$h_1 = 1/20$	6.2×10^{-2}	-1.3×10^{-5}
$h_2 = 1/60$	2.1×10^{-2}	-5.4×10^{-7}
Orders of convergence	1	3

Example 3.2.1 .

	1st. stage t = 3.9	2nd. stage t = 4
$h_1 = 1/10$	9.1×10^{-8}	1.6×10^{-7}
$h_2 = 1/30$	1.1×10^{-9}	2.0×10^{-9}
Orders of convergence	4	4

Example 5.3.4 .

	1st. stage t = 1.95	2nd. stage t = 2
$h_1 = 1/10$	3.0×10^{-4}	4.9×10^{-4}
$h_2 = 1/30$	3.2×10^{-5}	5.1×10^{-5}
Orders of convergence	2	2

Example 5.3.5 .

	1st. stage t = 5	2nd. stage t = 5
$h_1 = 0.1$	2.3×10^{-5}	6.1×10^{-5}
$h_2 = 0.01$	2.3×10^{-7}	5.7×10^{-7}
Orders of convergence	2	2

Example 5.3.6 .

Let us take the third order predictor

$$y_i - y_{i-1} = \frac{h}{12}(23f_{i-1} - 16f_{i-2} + 5f_{i-3})$$ with the fourth order corrector

$$y_i - y_{i-1} = \frac{h}{12}(9f_i + 19f_{i-1} - 5f_{i-2} + f_{i-3}) .$$

The numerical results which we obtain are as follows .

	$1^{st.}$ stage $t = 3$	$2^{nd.}$ stage $t = 3$
$h_1 = 0.1$	-4.0×10^{-7}	1.9×10^{-6}
$h_2 = 0.01$	-1.7×10^{-11}	1.7×10^{-10}
Orders of convergence	4	4

Example 5.3.7 .

	$1^{st.}$ stage $t = 4.9$	$2^{nd.}$ stage $t = 5$
$h_1 = 1/10$	-1.2×10^{-7}	7.8×10^{-8}
$h_2 = 1/30$	-1.4×10^{-9}	-9.4×10^{-10}
Orders of convergence	4	4

Example 5.3.8 .

	$1^{st.}$ stage $t = 3.96$	$2^{nd.}$ stage $t = 4$
$h_1 = 0.1$	-5.5×10^{-6}	-8.0×20^{-6}
$h_2 = 0.01$	-5.0×10^{-9}	-7.2×10^{-9}
Orders of convergence	3	3

Example 5.3.9 .

	$1^{st.}$ stage $t = 4$	$2^{nd.}$ stage $t = 4.03$
$h_1 = 0.1$	-6.5×10^{-6}	-5.9×10^{-6}
$h_2 = 0.01$	-5.6×10^{-9}	-5.0×10^{-9}
Orders of convergence	3	3

Example 5.3.10 .

	1st. stage t = 3.06		2nd. stage t = 3.05	3rd. stage t = 3.1	4th. stage t = 3.1
$h_1 = 0.2$	3.2×10^{-7}	$h_1 = 1/10$	1.9×10^{-8}	-2.5×10^{-8}	2.0×10^{-8}
$h_2 = 0.02$	2.4×10^{-11}	$h_2 = 1/30$	2.1×10^{-10}	-3.1×10^{-10}	2.2×10^{-10}
Orders of convergence	4		4	4	4

Example 5.3.11 .

	1st. stage t = 3.9	2nd. stage t = 4
$h_1 = 1/10$	1.8×10^{-8}	1.1×10^{-8}
$h_2 = 1/30$	7.3×10^{-11}	4.7×10^{-11}
Orders of convergence	5	5

Example 6.1.2 .

	1st. stage t = 3	2nd. stage t = 3	3rd. stage t = 3
$h_1 = 0.1$	-2.7×10^{-4}	4.5×10^{-6}	-7.0×10^{-6}
$h_2 = 0.01$	-2.5×10^{-6}	3.7×10^{-9}	-6.8×10^{-9}
Orders of convergence	2	3	3

Example 6.1.3 .

	1st. stage t = 2.93	2nd. stage t = 2.96	3rd. stage t = 3
$h_1 = 0.1$	-3.8×10^{-5}	-4.1×10^{-6}	-6.7×10^{-6}
$h_2 = 0.01$	-3.0×10^{-7}	-3.8×10^{-9}	-6.3×10^{-9}
Orders of convergence	2	3	3

Example 6.2.12 .

	1st. stage t = 5	2nd. stage t = 5	3rd. stage t = 5	4th. stage t = 5	5th. stage t = 5
$h_1 = 0.5$	-1.2×10^{-3}	2.3×10^{-4}	-1.6×10^{-5}	2.7×10^{-5}	2.0×10^{-5}
$h_2 = 0.05$	-0.87×10^{-5}	1.5×10^{-7}	-2.5×10^{-9}	1.7×10^{-10}	1.3×10^{-10}
Orders of convergence	2	3	4	5	5

Example 6.2.13 .

Let us take the first order predictor

$y_i - y_{i-1} = hf_{i-1}$ with the second order corrector

$y_i - y_{i-1} = \frac{h}{2}(f_i + f_{i-1})$ followed by the third order

$y_i - y_{i-1} = \frac{h}{12}(5f_i + 8f_{i-1} - f_{i-2})$, the fourth order

$y_i - y_{i-1} = \frac{h}{24}(9f_i + 19f_{i-1} - 5f_{i-2} + f_{i-3})$ and the fifth order

$y_i - y_{i-1} = \frac{h}{720}(251f_i + 646f_{i-1} - 264f_{i-2} + 106f_{i-3} - 19f_{i-4})$.

The numerical results we obtain are as follows .

	1st. stage t = 5	2nd. stage t = 5	3rd. stage t = 5	4th. stage t = 5	5th. stage t = 5
$h_1 = 0.5$	-1.2×10^{-3}	2.4×10^{-4}	1.8×10^{-5}	2.6×10^{-5}	4.0×10^{-5}
$h_2 = 0.05$	-0.87×10^{-5}	1.5×10^{-7}	0.94×10^{-9}	2.3×10^{-10}	2.7×10^{-10}
Orders of convergence	2	3	4	5	5

Example 6.3.2 .

	1st. stage t = 3.95	2nd. stage t = 4	3rd. stage t = 4
$h_1 = 1/10$	-4.0×10^{-8}	1.3×10^{-8}	9.1×10^{-9}
$h_2 = 1/30$	-5.3×10^{-10}	5.6×10^{-11}	4.2×10^{-11}
Orders of convergence	4	5	5

Example 6.3.3 .

	1st. stage t = 2.46		2nd. stage t = 2.45	3rd. stage t = 2.5	4th. stage t = 2.5
$h_1 = 0.1$	1.6×10^{-5}	$h_1 = 1/10$	6.1×10^{-7}	-3.2×10^{-7}	8.7×10^{-7}
$h_2 = 0.01$	1.4×10^{-8}	$h_2 = 1/30$	7.1×10^{-9}	-3.7×10^{-9}	9.6×10^{-9}
Orders of convergence	3		4	4	4

Example 6.3.4 .

	1st. stage t = 2.483		2nd. stage t = 2.5	3rd. stage t = 2.55	4th. stage t = 2.55
$h_1 = 1/10$	1.3×10^{-6}	$h_1 = 1/10$	-2.7×10^{-8}	1.5×10^{-7}	2.5×10^{-8}
$h_2 = 1/70$	3.7×10^{-9}	$h_2 = 1/30$	-3.0×10^{-10}	1.7×10^{-9}	3.1×10^{-10}
Orders of convergence	3		4	4	4