ON PADÉ APPROXIMATIONS TO THE EXPONENTIAL FUNCTION AND A-STABLE METHODS FOR THE NUMERICAL SOLUTION OF INITIAL VALUE PROBLEMS

by

BYRON L. EHLE

A Thesis
Submitted in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

at the

UNIVERSITY OF WATERLOO

Waterloo, Ontario
Department of Applied Analysis
and
Computer Science

March, 1969.

ABSTRACT

The analytic solutions of many differential equations contain terms of the form $e^{q_i x}$ where each q_i is a comnlex number with negative real part. The system of differential equations $\vec{y}' = B\vec{y}$, where B is a real square matrix all of whose eigenvalues are in the left half plane provides an example of a class of problems where these are the only terms in the solution. If x is positive then these terms will approach zero. If the modulus of at least one of the q_i, say q_1, is large as compared with the others, then the differential equation is said to be stiff.

In attempting to solve such differential equations numerically, the exponential terms are approximated in some way. Let this approximation be denoted by $E(q_i x)$. If h is the step size of the method, then it is advisable that $|E(q_i h)| < 1$ for all i. If this condition is not satisfied the numerical approximation for $e^{q_i x}$ will grow instead of approaching zero and the total numerical solution may fail to give a satisfactory approximation to the analytic solution of the differential equation. Most numerical procedures restrict the step size h in order to meet this restriction on the modulus of $E(q_i h)$. Thus, for stiff equations, a small step size must be used to keep $|E(q_1 h)| < 1$ while the absence of $e^{q_1 x}$ would allow a much larger h to be used.

In order to overcome this restriction on the step size, numerical methods are needed which reduce to approximations of the exponential which are bounded by one in the entire left half plane. Such methods are commonly called A-stable methods. In this thesis it is shown that two new classes of methods, of arbitrarily high order, exist which reduce to the diagonal Padé approximations, $P_{n,n}(z)$, of the exponential function. These are A-stable since for all n, $P_{n,n}(z)$ is known to satisfy the condition

$$|P_{n,n}(z)| < 1 \qquad \text{for } Re(z) < 0.$$

Unfortunately the diagonal Padé approximations also satisfy the property that $|P_{n,n}(z)| \to 1$ as $|z| \to \infty$, so they do not produce good approximations for $|z|$ large, $Re(z) < 0$. In an attempt to solve this problem the subdiagonal Padé approximations to the exponential function are studied. In particular, it is shown that all the entries in the first and second subdiagonals of the Padé table satisfy the condition that they are bounded by one in the entire left half plane and also approach zero as the modulus of z becomes large. Examples of numerical methods which reduce to these approximations are given. These methods are also clearly A-stable.

Since many of the methods which are shown to be A-stable in this thesis are implicit, some consideration is given to showing how one of these methods could be implemented with sufficient efficiency to make it competitive with or superior to classical explicit methods which require reduced step size. Several numerical examples are included.

Finally it is shown that none of the entries in the third subdiagonal of the Padé table of the exponential function are bounded by one in the entire left half plane so methods which reduce to these approximations to the exponential would not be A-stable.

ACKNOWLEDGEMENTS

The author is indebted particularly to Professor J. D. Lawson for first suggesting this general topic and then for a number of helpful suggestions during its study. Sincere gratitude is also expressed to Professor J. Dieudonné since his private communication provided the key to the proof of the fundamental result given in Chapter 3.

Without the financial support provided by an assisted leave of absence from the University of Victoria during the early stages and a National Research Council grant during its final stages, this work would not have been possible.

INDEX

CHAPTER 1

THE PROBLEM

1.1 Introduction

The motivation for this thesis is the desire to obtain numeric, rather than analytic, solutions for the initial value problem

$$\frac{dy}{dx} = f(x,y), \quad y(a) = y_a \tag{1.1.1}$$

as x ranges from a to b.

We might attempt to obtain an approximation to the solution of (1.1.1) by replacing f(x,y) by its Taylor's series expansion about (a,y_a). Thus we obtain the equivalent initial value problem

$$\frac{dy}{dx} = f(a,y_a) + (x-a)\frac{\partial f}{\partial x}\bigg|_{(a,y_a)} + (y-y_a)\frac{\partial f}{\partial y}\bigg|_{(a,y_a)} + \cdots ,$$

$$y(a) = y_a. \tag{1.1.2}$$

Neglecting all but the first three terms on the right in (1.1.2), we obtain the initial value problem

$$\frac{dy}{dx} = A + Bx + Cy, \quad y(a) = y_a, \tag{1.1.3}$$

where

$$A = \left(f - a\,\frac{\partial f}{\partial x} - y_a\,\frac{\partial f}{\partial y}\right)\bigg|_{(a,y_a)} ,$$

$$B = \frac{\partial f}{\partial x}\bigg|_{(a,y_a)} ,$$

$$C = \frac{\partial f}{\partial y}\bigg|_{(a,y_a)}$$

whose solution for x close enough to a is an approximation to the solution of the original initial value problem (1.1.1).

Now the differential equation in (1.1.3) is a linear first order differential equation and assuming that $C \neq 0$ it can be solved exactly [53, p. 56] giving the solution

$$y = -\left[\frac{A}{C} + \frac{B}{C^2} + \frac{Bx}{C}\right] + C_1 e^{Cx} \qquad (1.1.4)$$

where C_1 is the constant of integration which is chosen to satisfy the initial condition $y(a) = y_a$.

If instead of solving (1.1.3) exactly, we were to solve it by some numerical method, we would want that method to produce a solution which was in close agreement with the actual solution given in (1.1.4). Restricting our attention for the moment to Runge-Kutta methods, we would require that the Runge-Kutta method give a satisfactory approximation to the linear term.

$$\frac{A}{C} + \frac{B}{C^2} + \frac{Bx}{C} .$$

Of course, this term will be given exactly by any Runge-Kutta process of order one or greater. Our Runge-Kutta process must also approximate $C_1 e^{Cx}$ with sufficient accuracy to satisfy our needs. Thus we are led to considering the special initial value problem

$$\frac{dy}{dx} = Cy, \quad y(0) = C_1 = y_0 \qquad (1.1.5)$$

whose solution is $y = C_1 e^{Cx}$. We now must study how well any particular numerical process solves this particular problem.

If we assume that $x \geq 0$, then our investigation naturally falls into two distinct parts, $C > 0$ and $C < 0$. Clearly any numerical approximation, $E_m(x)$, to the exponential function should satisfy at least certain basic requirements. For example, for $C > 0$ and x real we would like

(a) $E_m(Cx) - \exp(Cx) = 0((Cx)^{m+1}) \qquad m \geq 1$

(b) $E_m(Cx) \geq 1 \quad$ for $x \geq 0$

(c) $E_m(Cx) \to \infty \quad$ as $x \to \infty$,

while for $C < 0$ we would like property (a) to again be satisifed and we would also like

(d) $|E_m(Cx)| \leq 1 \quad$ for $x \geq 0$

(e) $E_m(Cx) \to 0 \quad$ as $x \to \infty$.

We adopt the notation that $x_n = nh$ and y_n is the corresponding approximation to $y(x_n)$ produced by some numerical procedure. Then, considering the classical 4th order Runge-Kutta process

$$\begin{aligned} K_1 &= f(x_n,y_n) \\ K_2 &= f(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}hK_1) \\ K_3 &= f(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}hK_2) \\ K_4 &= f(x_n + h, y_n + hK_3) \\ y_{n+1} &= y_n + \frac{h}{6}(K_1 + 2K_2 + 2K_3 + K_4) \end{aligned} \tag{1.1.6}$$

as an example, we obtain on solving (1.1.5) with a step size of h that $y(h) = C_1e^{Ch}$ is approximated by $y_1 = E_4(Ch)y_0$, where

$$E_4(Ch) = 1 + Ch + \frac{(Ch)^2}{2!} + \frac{(Ch)^3}{3!} + \frac{(Ch)^4}{4!} .$$

Now for $C > 0$ we see that properties (a), (b), and (c) are all satisfied. On the other hand, for $C < 0$ property (a) is satisfied but property (d) is satisfied only if $0 > Ch > -2.8$ (approx.) and property (e) is not satisfied at all.

If we were to make additional applications of the Runge-Kutta process we obtain the result that $y(nh) = C_1e^{nCh}$ is approximated by $y_n = C_1[E_4(Ch)]^n y_0$ and hence for $C < 0$, $Ch < -2.8$ (approx.) the sequence $\{y_n\}$ does not converge to zero but instead gets large without bound. We describe this unsatisfactory behavior by saying that the 4th order Runge-Kutta process is unstable for $Ch < -2.8$ (approx.).

As the above discussion would suggest, the classical 4th order Runge-Kutta process is not a very satisfactory procedure for solving (1.1.5) for values of C which are negative and large in absolute value since h must be chosen very small to satisfy property (d). Unfortunately, this same problem will occur with any explicit Runge-Kutta process because $E_m(Ch)$ will be a polynomial of degree m or greater and hence for $|Ch|$ large enough, $|E_m(Ch)| > 1$ and both property (d) and (e) will not be satisfied. Although we will not go into details here, it is noted that this same problem is encountered when using almost all of the classical one step and multistep methods for numerical solution of differential equations, two exceptions being

$$y_{n+1} = y_n + \frac{h}{2}\,[f(x_n,y_n) + f(x_{n+1},y_{n+1})] \qquad (1.1.7)$$

and

$$y_{n+1} = y_n + h\,f(x_{n+1},y_{n+1}) \qquad (1.1.8)$$

as noted by Dahlquist [11, 12].

Since it is reasonable to expect that problems will arise where C is negative and large in absolute value, a natural course is to attempt to devise methods which will give approximate solutions to (1.1.5) which do satisfy properties (d) and (e). Rather than carry this discussion farther, however, we will first generalize the problem and introduce some notation and definitions which will be used throughout the remainder of this thesis.

1.2 The Problem

In this thesis we will be interested in the obvious generalization of the problem given in (1.1.5), that is, in the numerical solution of the differential equation

$$\frac{d\vec{y}}{dx} = B\vec{y}, \qquad \vec{y}(0) = \vec{y}_0 \qquad (1.2.1)$$

where B is a real square matrix whose eigenvalues are distinct.

Since the solution of (1.2.1) is

$$\begin{aligned}\vec{y} &= \exp(Bx)\,\vec{y}_0\\ &= \sum_{i=1}^{n} \vec{a}_i e^{\lambda_i x} \qquad (1.2.2)\end{aligned}$$

where for each i, $\vec{a}_i$ is a column vector of coefficients and λ_i is the i^{th} eigenvalue of the n by n matrix B, we have an immediate generalization of the problem discussed in the introduction. That is, we might wish to find a numerical procedure which will approximate each of the exponential terms in the sum satisfactorily.

If the eigenvalues were all real and negative and we were, for example, to use the classical 4th order Runge-Kutta process in its vector form, we would need to choose h so that

$$h\lambda \max \le 2.8 \text{ (approx.)}$$

where

$$\lambda_{\max} = \max_{i=1,\ldots n} |\lambda_i| .$$

Thus, a very small step size might be required simply to keep errors in the computation of a term which was approaching zero quickly from destroying the accuracy of the entire solution.

Having generalized the problem given in (1.1.5) to the one given in (1.2.1) it is also natural to remove the restriction that C (and hence the λ_i) be real. In particular, we will be most interested in the case where all the eigenvalues of B have negative real parts and where the modulus of at least one of these eigenvalues is large as compared with the rest. Such a system of differential equations is said to be stiff [10].

Allowing complex eigenvalues requires a slight reformulation of the conditions which we would like an approximation to the exponential function to satisfy. We include these modifications in the following definitions.

Definition 1.1

An approximation $E_m(z)$ to the exponential function is A-acceptable if and only if for any complex number z, such that $Re(z) < 0$, we satisfy the following two properties:

Property (1)

$$E_m(z) - \exp(z) = 0(z^{m+1}), \quad m \geq 0$$

Property (2)

$$|E_m(z)| \leq 1 \qquad \text{for } Re(z) \leq 0.$$

Definition 1.2

An approximation $E_m(z)$ to the exponential function is L-acceptable (left-acceptable) if and only if it is A-acceptable and further it satisfies the following property:

Property (3)

$$|E_m(z)| \to 0 \qquad \text{as } Re(z) \to -\infty.$$

It is also natural to generalize the idea of stability which was introduced previously so that it also covers the entire left half plane. We do this by introducing Dahlquist's definition of A-stability [11].

Definition 1.3

A k-step method is called A-stable, if all of its solutions tend to zero, as $n \to \infty$, when the method is applied with fixed positive h to any differential equation of the form

$$\frac{dy}{dx} = qy \tag{1.2.3}$$

where q is a complex constant with negative real part.

Although Dahlquist introduced this definition in the context of a study of the stability of the linear multistep process

$$y_{n+k} = a_{k-1}y_{n+k-1} + \dots + a_0 y_n + h(b_k f_{n+k} + \dots + b_0 f_n)$$

it can clearly be applied to any numerical method for solving (1.2.3). In particular, any numerical method which reduces to an A-acceptable or an L-acceptable approximation to the exponential solution of (1.2.3) will be an A-stable method. On the other hand, a numeric procedure which is A-stable will produce an A-acceptable approximation but it need not reduce to an approximation which satisfies Property (3) and hence A-stability does not imply L-acceptability.

In this thesis it will be shown that there is a class of A-acceptable approximations to the exponential (Chapter 2) and that there is also a class of L-acceptable approximations to the exponential function (Chapter 3). Further, it will be shown that numerical procedures exist which reduce to A-acceptable and L-acceptable approximations to the exponential when solving (1.2.3) (Chapter 4). Finally, some indications of the practical application of these processes will be given (Chapter 5).

Before turning our attention to these items, some additional definitions and theorems which will be needed in later chapters of this thesis will be given.

1.3 Preliminary Definitions and Theorems

The following definitions and theorems are presented for reference. They are, for the most part, well known ideas and results which we shall require in later chapters of this thesis. Because of this, the proofs will be omitted and instead only a reference will be given where a proof of the theorem can be found.

Definition 1.4

The neighborhood of a point z_0 is the set of points z such that $|z - z_0| < \delta$, $\delta > 0$.

Definition 1.5

A set of points in the plane is called open if every point, z, of the set has a neighborhood lying entirely within the set.

Definition 1.6

A nonempty open set in the plane is called a domain if and only if any two points in the set can be joined by a polygon which lies in the set.

Definition 1.7

A function $f = f(z)$ is analytic at $z = z_0$ if and only if its derivative, $f'(z)$, exists not only at z_0 but at every point z in a neighborhood of z_0.

Definition 1.8

Let $f(z) = P(z)/Q(z)$ where $P(z_0) \neq 0$ if $Q(z_0) = 0$. Then the zeros of $Q(z)$ are called poles of $f(z)$.

Definition 1.9

A function f(z) which is analytic in a domain D, except for poles, is said to be meromorphic in D.

A major problem which will confront us in the next two chapters is finding the location of the zeros of a polynomial. The first six theorems which are given deal with this problem.

Theorem 1.1 (Lucus)

All the critical points of a non-constant polynomial f lie in the convex hull, H, of the set of zeros of f.

Proof: [47, p. 22]

Theorem 1.2 (Cauchy)

All the zeros of

$$f(z) = a_0 + a_1 z + \dots + a_n z^n, \quad a_n \neq 0,$$

lie in the circle

$$|z| = 1 + \max_k \left|\frac{a_k}{a_n}\right| , \ k = 0, 1, \dots, (n-1).$$

Proof: [47, p. 123]

Theorem 1.3

All the zeros of $f(z) = a_0 + a_1 z + \dots + a_n z^n$ lie on or outside the circle

$$|z| = \min_k \frac{|a_0|}{|a_0| + |a_k|} , \ k = 1, 2, \dots, n.$$

Proof: [47, p. 126]

Definition 1.10

A curve C is said to be a simple closed curve provided it is representable in the form

$$x = \emptyset(t),\ y = \Psi(t), \qquad t_1 \le t \le t_2 \qquad (1.3.1)$$

where $\emptyset(t)$ and $\Psi(t)$ are continuous and $z(t) = \emptyset(t) + i\Psi(t)$ satisfies the conditions

$$z(\alpha) \neq z(\beta),\quad \alpha \neq \beta \quad , t_1 < \alpha < \beta < t_2$$

and

$$z(t_1) = z(t_2), \quad t_1 \neq t_2.$$

Definition 1.11

If $\emptyset(t)$ and $\Psi(t)$ in (1.3.1) have continuous derivatives in $t_1 \le t \le t_2$ then the curve $z = z(t)$, $t_1 \le t \le t_2$ is said to be smooth.

Definition 1.12

If a curve C is composed of a finite number of arcs, each of which is smooth, we say that C is piecewise smooth.

Theorem 1.4

Let f(z) be meormorphic inside and on a simple closed curve C which does not pass through any of the zeros or poles of f(z). Then

$$\frac{1}{2\pi i} \int_C \frac{f'(z)}{f(z)}\, dz = N_C(f) - P_C(f)$$

where $N_C(f)$ and $P_C(f)$ are respectively the number of zeros and poles of f(z) inside C.

Proof: [1, p. 123], [15, p. 242]

Theorem 1.5

If $f(z) = \sum_{k=0}^{n} a_k z^k$, $\beta_1 \neq n$ and all the zeros of $f(z)$ lie in a circular region C, then every zero Z of the polynomial

$$f_1(z) = \beta_1 f(z) - z\, f'(z)$$

may be written in the form $Z = z$ or in the form

$$Z = \left[\frac{\beta_1}{\beta_1 - n}\right] z$$

where z is a point of C.

Proof: [47, p. 69]

Theorem 1.6

Let

$$P(z) = z^n + a_1 z^{n-1} + a_2 z^{n-2} + \ldots + a_n$$

be a polynomial with real coefficients and let

$$Q(z) = a_1 z^{n-1} + a_3 z^{n-3} + a_5 z^{n-5} + \ldots$$

and

$$R(z) = z^n + a_2 z^{n-2} + a_4 z^{n-4} + \ldots$$

so that $P(z) = R(z) + Q(z)$. Then all the zeros of $P(z)$ have negative real parts if and only if

$$\frac{Q(z)}{R(z)} = \cfrac{1}{c_1 z + \cfrac{1}{c_2 z + \cfrac{1}{c_3 z + \begin{matrix} \cdot \\ \quad\cdot \\ \qquad\cdot \end{matrix} + \cfrac{1}{c_n z}}}}$$

where the coefficients c_1, c_2, ..., c_n are all positive.

Proof: [60], [61, pp. 174-178], [24]

Theorem 1.7 (Maximum Modulus Theorem)

If f(z) is analytic inside and on a simple closed curve C and is not identically equal to a constant, then the maximum value of $|f(z)|$ occurs on C.

Proof: [1, p. 108]

We next define what we mean by the index of a point with respect to a curve γ.

Definition 1.13

If a piecewise differentiable closed curve γ does not pass through the point a, then the index of a with respect to γ, $n(a,\gamma)$, is defined to be

$$n(a,\gamma) = \frac{1}{2\pi i} \int_\gamma \frac{dz}{z - a} \quad .$$

Theorem 1.8

Let D be the domain bounded by a simple closed curve C and let γ be any piecewise smooth curve in D. Then if $a \notin D$, $n(a,\gamma) = 0$.

Proof: [1, p. 116]

The next theorem states when and how we may change the variable of integration in a complex integral.

Theorem 1.9

Let C_w be a piecewise smooth curve from w_1 to w_2 in the w plane and let f(w) be continuous on C_w. Let $w = g(z)$ be analytic in a domain D of the z plane and let C_z:

$$z = z(t), \qquad t_1 \le t \le t_2$$

be a piecewise smooth curve from $z_1 = z(t_1)$ to $z_2 = (t_2)$ in D. Let $g(z_1) = w_1$ and $g(z_2) = w_2$. As z traces C_z once in the given direction, let $w = g(z)$ trace C_w once in the given direction. Then

$$\int_{C_w\, w_1}^{w_2} f(w)\,dw = \int_{C_z\, z_1}^{z_2} f(g(z))\frac{dw}{dz}dz.$$

Proof: [32, p. 519], [15, p. 242]

Finally we give a result concerning the form of the Padé approximations to the exponential function which we will find very useful in subsequent chapters.

Theorem 1.10

Let $P_{j,k}(z)$ be the unique Padé approximation to the exponential function with numerator $N_{j,k}(z)$ of degree k and denominator $D_{j,k}(z)$ of degree j. Then

$$N_{j,k}(z) = \sum_{m=0}^{k} \frac{(j + k - m)!\ k!}{(j + k)!\ m!\ (k - m)!} z^m$$

$$D_{j,k}(z) = \sum_{m=0}^{j} \frac{(j + k - m)!\ j!}{(j + k)!\ m!\ (j - m)!} (-z)^m$$

Proof: [30], [50]

An immediate Corollary to Theorem 1.10 is the following.

Corollary 1.1

$$D_{j,k}(z) = N_{k,j}(-z).$$

This completes our list of preliminary theorems and definitions and we can now turn our attention to the problem of finding A-acceptable and L-acceptable approximations to the exponential function.

CHAPTER 2

PREVIOUSLY KNOWN RESULTS

2.1 Approximating the Exponential Function

Since the exponential function is encountered in a wide variety of situations, it is quite natural that a great deal of effort has been expended on devising ways of producing satisfactory numerical approximations for it under a wide variety of circumstances [9], [25], [45]. Most of this effort, however, has been directed to the evaluation of the exponential function for only real arguments. This is quite natural since exponentials with complex arguments which occur explicitly can be rewritten in terms of real arguments by employing the relation

$$e^{z} = e^{x+iy} = e^{x}(\cos[y] + i\ \sin[y]) \qquad (2.1.1)$$

where x and y are real [1]. While this relation would be useful when solving (1.2.3) by several of the techniques described in the last section of this chapter (see equations (2.5.13) and (2.5.16)), it would be of little use in solving the general problem (1.2.1).

A second major difficulty with most of the approximations which are considered is that they produce good approximations to the exponential only in some restricted interval. For example, in [9] the interval [-ln 2, ln 2] is used. For explicit computation on a binary machine this is quite satisfactory since by using the fact that

$$e^x = 2^n 2^r = 2^n(e^{\ln 2})^r = 2^n(e^{r \ln 2}) \qquad (2.1.2)$$

where

$$x \log_2 e = n + r, \ n \text{ integer}, \quad -1 < r < 1 \qquad (2.1.3)$$

accurate values of e^x can be easily produced if $e^{r \ln 2}$ can be computed, since the factor 2^n results in nothing more than a simple shift in exponent. As suggested by Lawson [36], this idea can be generalized to matrix problems through the use of an argument reduction scheme of the form

$$(\exp(2^{-m}A))^{2^m} = \exp(A). \qquad (2.1.4)$$

We note that this technique, although allowing us to effectively extend the stability region for the explicit computation of e^A, does not eliminate the need for A-acceptable and L-acceptable approximations for e^z.

Now the most general representation of the exponential function would be the Taylor's series

$$e^z = 1 + \sum_{n=1}^{\infty} \frac{z^n}{n!} \qquad (2.1.5)$$

which is often used as its definition [29, p. 138]. If (2.1.5) was used to compute e^z, computational limitations would require that we truncate the series expansion after a specified number of terms. But then, as noted previously, we do not obtain satisfactory values for e^z for values of z which satisfy the conditions; $|z|$ is large, $\text{Re}(z) < 0$. One method of partially overcoming this shortcoming is to employ a numerical procedure which, when solving (1.1.5), reduces to the Taylor's series expansion of the exponential to a specified number of terms plus one or more corrective terms which extend the stability region. Several papers by Lawson [35, 37] employ this feature. Thus, for example, Lawson's 5th order Runge-Kutta method when solving (1.1.5) gives

$$e^{Ch} = 1 + \sum_{n=1}^{5} \frac{(Ch)^n}{n!} + 0.5625 \frac{(Ch)^6}{6!} \qquad (2.1.6)$$

which agrees with the Taylor's series expansion to terms in h^5 but also contains one additional corrective term. As can be seen in [35] the extra term allows the dimensions of the stability region to be roughly doubled in size as compared with using only the Taylor's series to h^5.

Clearly methods which rely on such techniques can be of only limited effectiveness because the high order terms in $z = (Ch)$ will eventually dominate if $|z|$ is allowed to become large enough. Thus, such methods will never produce either an A-acceptable or an L-acceptable approximation to the exponential. The way to overcome this problem is to introduce rational approximations to the exponential.

Clearly some approximation criterion must be adopted when we introduce rational approximations to the exponential function. The order of the approximation is taken to be the exponent of z on the highest power term to which we attain agreement with the Taylor's series (2.1.4) when the indicated division is performed. Thus, the rational approximation

(1 + z/2)/(1 - z/2) is of order 2 since

$$\frac{1+\frac{z}{2}}{1-\frac{z}{2}} = 1 + z + \frac{z^2}{2} + \frac{z^3}{4} + \frac{\frac{z^4}{8}}{1-\frac{z}{2}}$$

while that of (1 + 2z)/(1 + z) is of order 1 because

$$\frac{1+2z}{1+z} = 1 + z + z^2 + \frac{z^3}{1+z} .$$

Naturally we would like to get the highest order of accuracy possible with a given expenditure of effort, so we are led quite naturally to the Padé approximations to the exponential.

2.2 Padé Approximations to the Exponential

The first few entries in the table of Padé approximations to the exponential function are given in Table 2.1 while the form of the general entry in the table was given previously in Theorem 1.10. As can be seen from the table, $P_{j,k}(z)$ has a numerator $N_{j,k}(z)$ of degree k and a denominator $D_{j,k}(z)$ of degree j.

The basic property of $P_{j,k}(z)$ for all $j \geq 0$, $k \geq 0$, is that each approximation is of order (j + k) [61, p. 394]. It can also be shown that this is the highest order obtainable with a rational approximation using polynomials of degree k and j respectively for the numerator and denominator [61, p. 378]. Furthermore, it can be shown that each of these approximations is unique except for a common constant factor which multiplies both the numerator and denominator [61, p. 378]. Since we are interested in finding A-acceptable and L-acceptable approximations for the exponential, it is clear that only some of the Padé

j \ k	0	1	2
0	1	$1 + z$	$1 + z + \frac{z^2}{2}$
1	$\frac{1}{1 - z}$	$\frac{1 + \frac{z}{2}}{1 - \frac{z}{2}}$	$\frac{1 + \frac{2z}{3} + \frac{z^2}{6}}{1 - \frac{z}{3}}$
2	$\frac{1}{1 - z + \frac{z^2}{2}}$	$\frac{1 + \frac{z}{3}}{1 - \frac{2z}{3} + \frac{z^2}{6}}$	$\frac{1 + \frac{z}{2} + \frac{z^2}{12}}{1 - \frac{z}{2} + \frac{z^2}{12}}$

Table 2.1

$P_{j,k}(z)$ - Padé Approximations to the Exponential

approximations will be of interest to us. In particular, only those approximations $P_{j,k}(z)$ for which $j \geq k$ need be considered in looking for A-acceptable approximations. This follows at once from the fact that if the numerator is of higher power than the denominator, then for some z, $|z|$ large, $Re(z) \leq 0$, the norm of the numerator will be larger than the norm of the denominator and Property (2), which is required for A-acceptablity, will not hold. By the same type of argument it also follows that if there are any Padé approximations to the exponential which are L-acceptable, they must come from the set $P_{j,k}(z)$, $j > k$.

2.3 A Set of A-acceptable Approximations

Studying the entries in Table 2.1, it is obvious at once that there is at least one A-acceptable approximation to the exponential, namely $P_{0,0}(z)$. A little further investigation and the computation of

$$|N_{j,j}(z)|^2 - |D_{j,j}(z)|^2 \qquad (2.2.1)$$

for $j = 1, 2, 3$ will show that $|P_{j,j}(z)| \leq 1$ for $Re(z) \leq 0$, $j = 1, 2, 3$ and thus $P_{1,1}(z)$, $P_{2,2}(z)$, and $P_{3,3}(z)$ are also A-acceptable approximations to the exponential function. Although this procedure could be continued with other values of j, a more general proof which establishes that $P_{j,j}(z)$, $j = 1, 2, 3, \ldots$ is A-acceptable can be given [3].

The basic idea of the proof is quite simple and is as follows. Since $P_{j,j}(z)$, for any j, is the quotient of two polynomials, it will be analytic in the entire left half plane if its denominator $D_{j,j}(z)$ has no zeros in the left half plane. Assuming that this is the case, by the Maximum Modulus Theorem (Theorem 1.7) we would then have that the maximum value of $|P_{j,j}(z)|$ occurs on the boundary of the semicircular region bounded by the imaginary axis from -R to R and the semicircle $|z| = R$, $Re(z) \leq 0$. From the form of $P_{j,j}(z)$ we have that

$$|P_{j,j}(z)| \to 1 \qquad \text{as } R \to \infty$$

and by Corollary 1.1

$$|P_{j,j}(iy)| = 1 \qquad \text{for all real } y$$

since $D_{j,j}(iy) = N_{j,j}(-iy)$. Thus $|P_{j,j}(z)| \leq 1$ for $Re(z) \leq 0$. As a consequence of the above we have

Lemma 2.1

For any j, $P_{j,j}(z)$ is A-acceptable provided it is analytic in the left half plane.

We must now determine which of the $P_{j,j}(z)$ are analytic in the entire left half plane. By Corollary 1.1 and the discussion preceding Lemma 2.1 this is equivalent to determining for which values of j we satisfy the condition that all the zeros of $N_{j,j}(z)$ are in the left half plane. We choose to work with $N_{j,j}(z)$ rather than $D_{j,j}(z)$ because there are no changes of sign to keep track of in $N_{j,j}(z)$. In particular, we shall prove

Theorem 2.1

For all $j \geq 0$, all the zeros of $N_{j,j}(z)$ are in the open left half plane.

Proof

This is a well known result from the theory of passive networks [3]. Decomposing $N_{j,j}(z)$ into two polynomials $e_j(z)$ and $f_j(z)$, which contain respectively only the even powered and odd powered terms of $N_{j,j}(z)$ we have that

$$e^z = \frac{N_{j,j}(z)}{N_{j,j}(-z)} + 0(|z|^{2j+1})$$

$$= \frac{e_j(z) + f_j(z)}{e_j(z) - f_j(z)} + 0(|z|^{2j+1})$$

$$= \frac{g_j(z) + 1}{g_j(z) - 1} + 0(|z|^{2j+1})$$

where $g_j(z) = e_j(z)/f_j(z)$.

Multiplying through by $(g_j(z) - 1)$ and collecting terms, we obtain

$$g_j(z) = \frac{e^z + 1}{e^z - 1} + 0(|z|^{2j})$$

$$= \coth(\tfrac{z}{2}) + 0(|z|^{2j}).$$

Now the function $\coth(z)$ has a continued fraction expansion given by

$$\mathrm{Coth}(z) = \frac{1}{z} + \cfrac{1}{\frac{3}{z} + \cfrac{1}{\frac{5}{z} + \cfrac{1}{\frac{7}{z} + \cdots}}}$$

[62, p. 303] and it can be shown [58] that $g_j(2z)$ is the j^{th} continued fraction approximation to $\coth(z)$. Using the fact that

$$\frac{f_j(z)}{e_j(z)} = \frac{1}{g_j(z)}$$

we have that

$$\frac{z^j f_j(\frac{1}{z})}{z^j e_j(\frac{1}{z})} = \frac{f_j(\frac{1}{z})}{e_j(\frac{1}{z})} = \cfrac{1}{z + \cfrac{1}{3z + \cfrac{1}{5z + \cdots + \cfrac{1}{(2j-1)z}}}}$$

and thus by Theorem 1.6

$$P(z) = z^j N_{j,j}(\frac{1}{z}) = z^j[f_j(\frac{1}{z}) + e_j(\frac{1}{z})]$$

has all of its zeros in the left half plane. But $z^n \neq 0$ in the left half plane, thus $N_{j,j}(\frac{1}{z})$ has all of its zeros in the left half plane. But the mapping $\frac{1}{z} \to z$ maps the left half plane into the left half plane so that it follows at once that $N_{j,j}(z)$ also has all of its zeros in the left half plane.

Combining the results of Theorem 2.1 with those of Lemma 2.1 and Corollary 1.1 we now have

Theorem 2.2

For all $j \geq 0$, $P_{j,j}(z)$ is A-acceptable.

Having established that all the diagonal Padé approximations to the exponential are A-acceptable, we quite naturally turn our attention to looking for L-acceptable approximations in the below diagonal entries in the Padé table.

2.4 Some Results on L-acceptable Approximations

Varga [59] has provided a preliminary lemma which might be useful in our search for L-acceptable approximations. It says

Lemma 2.2

$|P_{j,k}(z)| \leq 1$ for z real, $z \leq 0$ if and only if $j \geq k$.

Proof

Since $|P_{j,k}(z)| = 0(z^{k-j})$ for $z \to \infty$ $|P_{j,k}(z)|$ bounded implies that $j \geq k$. Conversely, if $j \geq k$, then $|D_{j,k}(z)| \geq |N_{j,k}(z)|$ if

$$\frac{(j + k - m)!\ j!}{(j + k)!\ m!\ (j - m)!} \geq \frac{(j + k - m)!\ k!}{(j + k)!\ m!\ (k - m)!},\quad 0 \leq m \leq k,$$

which is obviously true.

This lemma suggests that all below diagonal entries in the Padé table might be L-acceptable.

Looking again at the entries in Table 2.1 it is obvious at once that $P_{1,0}(z)$ is L-acceptable and hence also A-acceptable. Continuing down the first column of the table we have by direct computation of the roots of $N_{0,j}(z)$, $j = 1, 2, 3, \ldots, 23$, by Iverson [31] and Corollary 1.1 that $P_{2,0}(z)$, $P_{3,0}(z)$ and $P_{4,0}(z)$ might be L-acceptable since they are analytic in the entire left half plane and it might be possible to apply the Maximum Modulus Theorem to them just as we did to $P_{j,j}(z)$.

Unfortunately $P_{5,0}(z)$ cannot be either L-acceptable or A-acceptable since $N_{0,5}(z)$ has some of its zeros in the right half plane and thus $D_{5,0}(z)$ has zeros in the left half plane and therefore $|P_{5,0}(z)|$ is not bounded in the left half plane.

Since $N_{0,j}(z) = \frac{d}{dz} N_{0,j+1}(z)$ for all $j \geq 0$, it also follows from Theorem 1.1 that for all $j \geq 5$, $N_{0,j}(z)$ has zeros in the right half plane and hence for $j \geq 5$, $D_{j,0}(z)$ has zeros in the left half plane. Thus we have

Lemma 2.3

For all $j \geq 5$, $P_{j,0}(z)$ is neither L-acceptable nor A-acceptable.

It seems natural to attempt to extend Lemma 2.3 by considering $P_{j,k}(z)$ for values of k other than zero. Using the method of Routh [56, 57] as given in Wall [60] this author computed the number of zeros in the left half plane for all $D_{j,k}(z)$ of interest for $0 \leq j \leq 20$, $0 \leq k \leq 20$. The results of these computations are shown in Table 2.2

j \ k	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	0																				
1	0	0																			
2	0	0	0																		
3	0	0	0	0																	
4	0	0	0	0	0																
5	2	0	0	0	0	0															
6	2	0	0	0	0	0	0														
7	2	2	0	0	0	0	0	0													
8	2	2	2	0	0	0	0	0	0												
9	2	2	2	0	0	0	0	0	0	0											
10	4	2	2	2	0	0	0	0	0	0	0										
11	4	2	2	2	2	0	0	0	0	0	0	0									
12	4	4	2	2	2	2	0	0	0	0	0	0	0								
13	*	4	2	2	2	2	2	0	0	0	0	0	0	0							
14	*	*	4	2	2	2	2	0	0	0	0	0	0	0	0						
15	*	*	*	4	2	2	2	2	0	0	0	0	0	0	0	0					
16	*	*	*	*	2	2	2	2	2	0	0	0	0	0	0	0	0				
17	*	*	*	*	*	2	2	2	2	2	0	0	0	0	0	0	0	0			
18	*	*	*	*	*	*	2	2	2	2	2	0	0	0	0	0	0	0	0		
19	*	*	*	*	*	*	*	2	2	2	2	0	0	0	0	0	0	0	0	0	
20	*	*	*	*	*	*	*	*	2	2	2	2	0	0	0	0	0	0	0	0	0

Table 2.2

The number of zeros of $D_{j,k}(z)$ in the left half plane
(* indicates 2 or more zeros)

From the results in Table 2.2 it would seem that we might be safe in stating the following conjecture

Conjecture 2.1

All entries $P_{j,k}(z)$ on or above the fourth subdiagnoal in the Padé table of rational approximations to the exponential function are analytic in the entire left half plane.

From Table 2.2 this is seen to be true for $j \leq 20$, $k \leq 20$.

As a consequence we might suspect that the first four subdiagonal sets of Padé approximations might be L-acceptable. In the next chapter we will see that although this is not quite the case, at least the first two sets of subdiagonal approximations are L-acceptable. In establishing this result, we will also verify the conjecture for at least all first and second subdiagonal approximations.

2.5 Some A-stable Methods

The literature contains a number of examples of A-stable methods. Probably the best known A-stable methods are (1.1.7) which has an error of $0(h^3)$ and (1.1.8) which has an error of $0(h^2)$ and which reduce to $P_{1,1}(qh)$ and $P_{1,0}(qh)$ respectively when solving (1.2.3). If higher derivatives are allowed, then an obvious generalization of (1.1.7) would be the method

$$y_{n+1} = y_n + \frac{h}{2}(y'_n + y'_{n+1}) + \frac{h^2}{12}(y''_n - y''_{n+1}) \qquad (2.5.1)$$

which has an error of $0(h^5)$. Ralston [52, p. 212] has observed that this method is A-stable and this is clearly the case since it reduces to $P_{2,2}(qh)$ when solving (1.2.3).

Davidson [13] has recently proposed a method specifically designed for equations of the form

$$\frac{d\vec{y}}{dx} = A\vec{y} + \vec{u}(x) \tag{2.5.2}$$

which reduces to a form like that of (2.5.1), but with other coefficients, when solving (1.2.3). It is also A-stable. Since it requires third derivatives of y at x_n and x_{n+1} and also has an error of $O(h^5)$ it may not be as efficient as (2.5.1).

Both method (1.1.7) and (2.5.1) suffer from the unfortunate feature that when solving (1.2.3) the approximation to e^{qh} approaches one in absolute value as $h \to \infty$. (1.1.8) does not have this unfortunate property but does suffer from a larger error term. In an effort to overcome this difficulty, Makinson [46] considered the generalization of (2.5.1) to the form

$$\bar{y}_{n+1} = \bar{y}_n + \sum_{i=1}^{\nu} (\alpha_i \bar{y}_{n+1}^{(i)} + \beta_i \bar{y}_n^{(i)}) \frac{h^i}{i!} + T \tag{2.5.3}$$

where $\bar{y}_n = y(nh)$ and T is the truncation error in the process. With $\nu = 2$ he establishes that a one parameter family of the form

$$\begin{aligned}\bar{y}_{n+1} &= \bar{y}_n + h(\alpha_1 \bar{y}'_{n+1} + \{1 - \alpha_1\}\bar{y}'_n) \\ &+ \frac{h^2}{2}(\{\frac{1}{3} - \alpha_1\}\bar{y}''_{n+1} + \{\frac{2}{3} - \alpha_1\}\bar{y}''_n) + T\end{aligned} \tag{2.5.4}$$

can be found where

$$T = -\frac{(1 - 2\alpha_1)}{24} h^4 y^{(4)}(\theta) \tag{2.5.5}$$

where $x_n < \theta < x_{n+1}$ assuming that y is suitably differentiable. Setting $\alpha_1 = (\ 1 + \sqrt{3}/3)$ Makinson then establishes that the resulting process is A-stable and when solving (1.2.3) with large h the exponential is approximated by $(1 - \sqrt{3})$ which is a little better than either (1.1.7) or (2.5.1).

Repeating this process for $\nu = 3$, Makinson finds the following two parameter family with error of $O(h^5)$.

$$\begin{aligned}\bar{y}_{n+1} = \bar{y}_n &+ h(\alpha_1\bar{y}'_{n+1} + \{1 - \alpha_1\}\bar{y}'_n) \\ &+ \frac{h^2}{2}(\alpha_2\bar{y}''_{n+1} + \{1 - 2\alpha_1 - \alpha_2\}\bar{y}''_n) \qquad (2.5.6) \\ &+ \frac{h^3}{24}(\{1 - 4\alpha_1 - 6\alpha_2\}\bar{y}'''_{n+1} + \{3 - 8\alpha_1 - 6\alpha_2\}\bar{y}'''_n) \\ &+ T\end{aligned}$$

where

$$T = -(\frac{3}{2} - 5\alpha_1 - 5\alpha_2)\frac{h^5}{5!}y^{(5)}(\theta). \qquad (2.5.7)$$

Setting $\alpha_1 = 3.205$ (approx.) and $\alpha_2 = -6.851$ (approx.) he finds that the resulting process is also A-stable. In this case, for large values of h, e^{qh} is approximated by -.635.

Liniger and Willoughby [39] have also recently proposed several new A-stable methods based on using (2.5.3) with $\alpha_i \neq \beta_i$.

A somewhat different implicit technique for solving the differential equation $\frac{d\vec{y}}{dx} = \vec{f}(\vec{y})$ has been proposed by Rosenbrock [54], [55, p. 180]. Letting $A(\vec{y})$ denote the Jacobian matrix $(\frac{\partial\vec{f}}{\partial y})$ this method can be represented in the form

$$\vec{K}_i = h[\vec{f}(\vec{y}_n + \sum_{j=1}^{i-1} b_{ij}\vec{K}_j) + a_i A(\vec{y}_n + \sum_{j=1}^{i-1} c_{ij}\vec{K}_j)\vec{K}_i]$$

$$i = 1, 2, \ldots, \nu \qquad (2.5.8)$$

$$\vec{y}_{n+1} = \vec{y}_n + \sum_{i=1}^{\nu} \omega_i \vec{K}_i$$

where the a_i, b_{ij}, c_{ij}, and ω_i are constants whose values are determined by the accuracy desired in the process. In particular, for the two stage process

$$\vec{K}_1 = h[\vec{f}(\vec{y}_n) + a_1 A(\vec{y}_n)\vec{K}_1]$$

$$\vec{K}_2 = h[\vec{f}(\vec{y}_n + b_{21}\vec{K}_1) + a_2 A(\vec{y}_n + c_{21}\vec{K}_1)\vec{K}_2] \qquad (2.5.9)$$

$$\vec{y}_{n+1} = \vec{y}_n + \omega_1\vec{K}_1 + \omega_2\vec{K}_2$$

Rosenbrock establishes that the following conditions must be satisfied to attain agreement to the power of h indicated in the Taylor's series:

$$h: \quad \omega_1 + \omega_2 = 1$$

$$h^2: \quad \omega_1 a_1 + \omega_2(a_2 + b_{21}) = \frac{1}{2}$$

$$h^3: \begin{cases} \omega_1 a_1^2 + \omega_2(a_2 + (a_1 + a_2)b_{21}) = \frac{1}{6} \\ \omega_2(a_2 c_{21} + \frac{1}{2}b_{21}^2) = \frac{1}{6} \end{cases}$$

$$h^4: \begin{cases} \omega_1 a_1^3 + \omega_2(a_2^3 + (a_1^2 + a_1 a_2 + a_2^2)b_{21}) = \frac{1}{24} \\ \omega_2 a_2(a_2 c_{21} + \frac{1}{2}b_{21}^2) = \frac{1}{24} \\ \omega_2(a_1 a_2 c_{21} + a_2^2 c_{21} + a_2 b_{21} c_{21} + a_1 b_{21}^2) = \frac{1}{24} \\ \omega_2(\frac{1}{2}a_2 c_{21}^2 + \frac{1}{6}b_{21}^3) = \frac{1}{24} \, . \end{cases}$$

He then shows that with

$$\begin{aligned} a_1 &= 1 + \sqrt{6}/6 \\ a_2 &= 1 - \sqrt{6}/6 \\ b_{21} &= c_{21} = 0.173 \text{ (approx.)} \qquad\qquad (2.5.10) \\ \omega_1 &= -0.413 \text{ (approx.)} \\ \omega_2 &= 1.413 \text{ (approx.)} \end{aligned}$$

we obtain a process with error of $0(h^4)$ and the resulting process approximates $\exp(qh)$ with the rational expression

$$E_3(qh) = \frac{1 - qh - \frac{2}{3}(qh)^2}{1 - 2qh + \frac{5}{6}(qh)^2} \, .$$

It can be shown that $|E_3(qh)| \leq 1$ for all h if $Re(q) < 0$ and hence the process given by (2.5.8) and (2.5.10) is A-stable. As $h \to \infty$ we also see that e^{qh} is approximated by -0.8.

Calahan [8] has noted that letting

$$a_1 = a_2 = \sqrt{\frac{2 + \sqrt{3}}{6}} = 0.788 \text{ (approx.)}$$

$$b_{21} = -1.154 \text{ (approx.)}$$

$$c_{21} = 0$$

$$\omega_1 = 0.75$$

$$\omega_2 = 0.25$$

another method with error of $0(h^4)$ is obtained which is also A-stable. We note that the approximation to the exponential which this choice of constants produces is identical to that produced by Makinson's 3rd order method (2.5.3) with $\alpha_1 = 1 + \sqrt{3}/3$ which was discussed previously.

Calahan implies that this process would be computationally superior to other methods based on Rosenbrock's process because his choice of constants requires the computation of only one inverse matrix

$$(I + a_1 A(y_n))^{-1}$$

in order to compute $\vec{K}_1$ and $\vec{K}_2$ and produce a third order process while other methods, such as (2.5.10) require computing the above inverse as well as

$$(I + a_2 A(y_n + c_{21}\vec{K}_1))^{-1}$$

At least one other A-stable process [8] can be based on (2.5.8). It is given by letting

$$\begin{aligned} a_1 &= a_2 = 0.5 \\ b_{21} &= c_{21} = 0.0 \\ \omega_1 &= 1.0 \\ \omega_2 &= 0.0 \end{aligned} \qquad (2.5.12)$$

and can be seen to reduce to $P_{1,1}(qh)$ when solving (1.2.3). Although it also requires the computation of only one inverse matrix, it is only a second order process and thus is not as attractive as (2.5.11).

As another example, Lawson [36] has shown that explicit Runge-Kutta processes which can never be A-stable can be transformed into A-stable processes while retaining their basic Runge-Kutta qualities. Lawson's method for solving $\frac{d\vec{y}}{dx} = \vec{f}(x,\vec{y})$ is as follows.

$$\vec{K}_1^* = \vec{f}(x_n,\vec{y}_n) - A\,\vec{y}_n;$$

$$\vec{p}_i^* = \exp(c_i hA).\vec{y}_n + h\sum_{j=1}^{i-1} a_{ij}\exp[(c_i - c_j)hA]\vec{K}_j^*$$

$$\vec{K}_i^* = \vec{f}(x_n + c_i h,\vec{p}_i^*) - A.\vec{p}_i^*\,, \quad i = 1, 2, \ldots, \nu \qquad (2.5.13)$$

$$\vec{y}_{n+1} = \exp(hA).\vec{y}_n + h\sum_{i=1}^{\nu} b_i\exp[(1 - c_i)hA]\,\vec{K}_i^*$$

where the c_i, a_{ij}, b_i are chosen to give a ν stage Runge-Kutta process of order $\leq\nu$ when $A = 0$ and A is in general a real matrix whose entries will depend in some way on $\vec{f}(x,\vec{y})$.

For example, if

$$\vec{f}(x,\vec{y}) = B\vec{y} + \vec{u}(x,\vec{y})$$

where B is a constant matrix having eigenvalues with negative real parts and we assume that the spectral radius of B is much larger than a Lipschitz constant for $\vec{u}(x,\vec{y})$ then A would probably be set equal to B. Under this assumption, it is clear that solving (1.2.3) exactly using (2.5.13) we would obtain

$$y_{n+1} = [\exp(qh)]^{n+1}\ y_0\ . \qquad (2.5.14)$$

Now to carry out the steps in (2.5.13) numerically, an approximation E(z) to the exponential would be needed. Thus in place of (2.5.14) we would actually have that

$$y_{n+1} = [E(qh)]^{n+1}\ y_0\ . \qquad (2.5.15)$$

It follows at once, as noted by Lawson, that if we use one of the diagonal Padé approximations to the exponential, then we have an A-stable process.

Finally, if we restrict ourselves to equations of the form (2.5.2), then Pope [51], Kuo [33], Legras [38], and Calahan [7] have all proposed methods based on the fact that (2.5.2) has the exact solution

$$\vec{y}(x) = e^{Ax}\vec{y}(0) + \int_0^x e^{A(x-t)}\vec{u}(t)\ dt. \qquad (2.5.16)$$

The various methods proposed by the above authors result from differing interpretations on how the integral on the right in (2.5.16) should be evaluated. In each case, however, if an

A-acceptable approximation to the exponential is used, the resulting process will be A-stable.

CHAPTER 3

NEW RESULTS

3.1 Introduction

In Chapter 2 we exhibited a class of functions (the diagonal Padé approximations to e^z) which approximate e^z and, in particular, satisfy Properties (1) and (2). Unfortunately, these diagonal Padé approximations do not satisfy Property (3). In this chapter we shall prove that the first two subdiagonal Padé approximations to e^z, namely $P_{n+1,n}(z)$ and $P_{n+2,n}(z)$, $n \geq 0$, not only satisfy Properties (1) and (2) but also satisfy Property (3).

3.2 Properties (1) and (3)

Since $P_{n+1,n}(z)$ and $P_{n+2,n}(z)$ are rational functions of z both with polynomials of degree n as numerators and polynomials of degree n+1 and n+2, respectively, as denominators it follows at once that Property (3) will be satisfied. Since $P_{n+1,n}(z)$ and $P_{n+2,n}(z)$ are Padé approximations to e^z we also have at once that

$$P_{n+1,n}(z) - e^z = 0(z^{2n+1})$$

and

$$P_{n+2,n}(z) - e^z = 0(z^{2n+2})$$

as $z \to 0$. Hence Property (1) is satisfied.

We have only to establish Property (2) to complete the proof.

3.3 Establishing Property (2)

To establish that Property (2) holds for the first two subdiagonal Padé approximations to e^z we will proceed in a fashion similar to that of the previous chapter, in that we will show first that $P_{n+1,n}(z)$ and $P_{n+2,n}(z)$ are bounded by one on the imaginary axis and second that they are analytic in the left half plane. Then since they also satisfy Property (3) we can apply the maximum principle as was done in Chapter 2 to finally establish Property (2).

3.4 Some Preliminary Theorems

We begin by establishing the boundedness of $P_{n+1,n}(z)$ and $P_{n+2,n}(z)$ on the imaginary axis. Unfortunately neither $P_{n+1,n}(z)$ nor $P_{n+2,n}(z)$ exhibit the symmetry of $P_{n,n}(z)$ so our proof will require more effort that was the case for the diagonal Padé approximations. In particular, several preliminary theorems will be required.

Theorem 3.1

Let $N_{n,n}(z)$ be the numerator and $D_{n,n}(z)$ be the denominator of the nth diagonal Padé approximation $P_{n,n}(z)$ of e^z. Then

$$N_{n,n}(z) = N_{n-1,n-1}(z) + A\, z^2\, N_{n-2,n-2}(z)$$

and

$$D_{n,n}(z) = D_{n-1,n-1}(z) + A\, z^2\, D_{n-2,n-2}(z)$$

where

$$A = \frac{n!\,(2n-4)!}{(2n)!\,(n-2)!} = \frac{1}{4(2n-1)(2n-3)}$$

Proof

Since $D_{n,n}(z) = N_{n,n}(-z)$ for all $n \geq 0$ it suffices to prove the theorem only for $N_{n,n}(z)$.

By Theorem 1.10

$$N_{n,n} = \sum_{k=0}^{n} \frac{(2n-k)!\ n!}{(2n)!\ k!\ (n-k)!}\ z^k$$

$$= 1 + \frac{1}{2}z + \sum_{k=2}^{n-1} \frac{(2n-k)!\ n!}{(2n)!\ k!\ (n-k)!}\ z^k + \frac{n!}{(2n)!}\ z^n$$

Also from Theorem 1.10 we have

$$N_{n-1,n-1} + \frac{z^2\ n!\ (2n-4)!}{(2n)!\ (n-2)!}\ N_{n-2,n-2}$$

$$= \sum_{k=0}^{n-1} \frac{(2n-2-k)!\ (n-1)!}{(2n-2)!\ k!\ (n-1-k)!}\ z^k + \frac{z^2\ n!}{(2n)!} \sum_{k=0}^{n-2} \frac{(2n-4-k)!\ \cancel{(n-2)!}\ \cancel{(2n-4)!}}{\cancel{(2n-4)!}\ k!\ (n-2-k)!\ \cancel{(n-2)!}}\ z^k$$

$$= 1 + \frac{z}{2} + \sum_{k=2}^{n-1} \left\{ \frac{(2n-2-k)!\ (n-1)!}{(2n-2)!\ k!\ (n-1-k)!} + \frac{n!}{(2n)!}\ \frac{(2n-2-k)!}{(k-2)!\ (n-k)!} \right\} z^k + \frac{n!}{(2n)!}\ z^n$$

But

$$\left\{ \frac{(2n-2-k)!\ (n-1)!}{(2n-2)!\ k!\ (n-1-k)!} + \frac{n!\ (2n-2-k)!}{(2n)!\ (k-2)!\ (n-k)!} \right\}$$

$$= \frac{(2n-2-k)!\ n!}{(2n)!\ k!\ (n-k)!} \left\{ \frac{(2n-1)(2n)(n-k)}{n} + (k)(k-1) \right\}$$

$$= \frac{(2n-2-k)!\ n!}{(2n)!\ k!\ (n-k)!} \left\{ 4n^2 - 4nk - 2n + 2k + k^2 - k \right\}$$

$$= \frac{(2n-2-k)!\ n!}{(2n)!\ k!\ (n-k)!} \left\{ (2n-k-1)(2n-k) \right\}$$

$$= \frac{(2n-k)!\ n!}{(2n)!\ k!\ (n-k)!}\ .$$

But clearly this last equality completes the proof.

Lemma 3.1

Let $F_n(z) = N_{n,n}(z) \cdot D_{n,n}(z)$. Then $F_n(z)$ has no odd terms.

Proof

Since, by Corollary 1.1, $D_{nn}(z) = N_{nn}(-z)$, we have at once that $F_n(z) = F_n(-z)$. But since F_n is a polynomial it then has only even powers of z.

Lemma 3.2

$N_{n,n}(z)D_{n-1,n-1}(z) = D_{n,n}(-z)N_{n-1,n-1}(-z)$ for all $n \geq 1$.

Proof

The proof follows at once from Corollary 1.1.

Theorem 3.2

For all $n \geq 1$, the only term with an odd power of z in the product $D_{n,n}(z)N_{n-1,n-1}(z)$ is the term of highest power, namely,

$$\frac{(-1)^n \; n! \; (n-1)! \; z^{2n-1}}{(2n)! \; (2n-2)!} \; .$$

Proof

The proof is by induction. For $n = 1$ we have

$$D_{1,1}(z)N_{0,0}(z) = (1 - \tfrac{z}{2})(1)$$

and hence the theorem is true for $n = 1$. Now assume that the only odd term in $D_{n-1,n-1}(z)N_{n-2,n-2}(z)$ is

$$\frac{(-1)^{n-1} \; (n-1)! \; (n-2)! \; z^{2n-3}}{(2n-2)! \; (2n-4)!}$$

Then by Lemma 3.2 the only odd term in $N_{n-1,n-1}(z)D_{n-2,n-2}(z)$ is

$$\frac{-(-1)^{n-1}\ (n-1)!\ (n-2)!\ z^{2n-3}}{(2n-2)!\ (2n-4)!}\ .$$

From Theorem 3.1

$$D_{n,n}(z)N_{n-1,n-1}(z) = \left[D_{n-1,n-1} + \frac{z^2}{4(2n-1)(2n-3)}\ D_{n-2,n-2}\right] N_{n-1,n-1}$$

$$= D_{n-1,n-1}N_{n-1,n-1} + \frac{z^2}{4(2n-1)(2n-3)}\ D_{n-2,n-2}N_{n-1,n-1}.$$

Now by Lemma 3.1 the first term on the right has no odd terms hence by the induction hypothesis the product has only the odd term

$$\frac{n(n-1)\ z^2}{(2n)(2n-2)(2n-1)(2n-3)} \cdot \frac{(-1)^n\ (n-1)!\ (n-2)!\ z^{2n-3}}{(2n-2)!\ (2n-4)!} = \frac{(-1)^n\ n!\ (n-1)!}{(2n)!\ (2n-2)!}\ z^{2n-1}$$

which was to be shown.

Corollary 3.1

$$\frac{z}{(2n-1)}\ [D_{nn}(z)N_{n-1,n-1}(z) - N_{nn}(z)D_{n-1,n-1}(z)]$$

equals

$$(-1)^n \left[\frac{(n-1)!\ z^n}{(2n-1)!}\right]^2 .$$

Proof

The result follows at once from Theorem 3.2 and Lemma 3.2.

We next present four recurrence relations that hold among various numerators and denominators of the Padé approximation to e^z and which have been noted before [52, p. 310].

Lemma 3.3

The following four recurrence relations hold for the numerators and denominators of the Padé approximations of e^z for $j,k \geq 1$ and $A = \frac{j}{(j+k)(j+k-1)}$ and $B = \frac{-k}{(j+k)(j+k-1)}$

$$(1) \quad N_{j,k}(z) = N_{j,k-1}(z) + AzN_{j-1,k-1}(z)$$

$$(2) \quad D_{j,k}(z) = D_{j,k-1}(z) + AzD_{j-1,k-1}(z)$$

$$(3) \quad N_{j,k}(z) = N_{j-1,k}(z) + BzN_{j-1,k-1}(z)$$

$$(4) \quad D_{j,k}(z) = D_{j-1,k}(z) + BzD_{j-1,k-1}(z)$$

Proof

The proof follows at once from the equations defining D and N.

Lemma 3.4

If $z = iy$, y real, then $\overline{N_{nn}(iy)} = D_{nn}(iy)$ and hence $\overline{D_{nn}(iy)} = N_{nn}(iy)$ for all $n \geq 0$.

Proof

The proof follows at once from the equations for D and N.

Theorem 3.3

For all $n \geq 1$, if $z = iy$, y real, then

$$|D_{n,n-1}(z)|^2 - |N_{n,n-1}(z)|^2 = \left[\frac{(n-1)!}{(2n-1)!}\right]^2 y^{2n} \geq 0.$$

Proof

Using properties (1) and (2) of Lemma 3.3 we have that for any z

$$|D_{n,n-1}|^2 - |N_{n,n-1}|^2 = D_{n,n-1}\overline{D_{n,n-1}} - N_{n,n-1}\overline{N_{n,n-1}}$$

$$= \left(D_{nn} - \frac{z}{2(2n-1)} D_{n-1,n-1}\right)\left(\overline{D_{nn}} - \frac{\bar{z}}{2(2n-1)} \overline{D_{n-1,n-1}}\right)$$

$$- \left(N_{nn} - \frac{z}{2(2n-1)} N_{n-1,n-1}\right)\left(\overline{N_{nn}} - \frac{\bar{z}}{2(2n-1)} \overline{N_{n-1,n-1}}\right)$$

$$= \left\{|D_{nn}|^2 + \frac{|z|^2}{4(2n-1)^2}|D_{n-1,n-1}|^2 - \frac{z}{2(2n-1)} \overline{D_{nn}}D_{n-1,n-1}\right.$$

$$\left. - \frac{\bar{z}}{2(2n-1)} D_{nn}\overline{D_{n-1,n-1}}\right\}$$

$$- \left\{|N_{nn}|^2 + \frac{|z|^2}{4(2n-1)^2}|N_{n-1,n-1}|^2 - \frac{z}{2(2n-1)} \overline{N_{nn}}N_{n-1,n-1}\right.$$

$$\left. - \frac{\bar{z}}{2(2n-1)} N_{nn}\overline{N_{n-1,n-1}}\right\}$$

Now if z = iy by Corollary 1.1 we have that $|N_{nn}| = |D_{nn}|$ and $|N_{n-1,n-1}| = |D_{n-1,n-1}|$. Hence for z = iy we have

$$|D_{n,n-1}(iy)|^2 - |N_{n,n-1}(iy)|^2$$

$$= \frac{iy}{2(2n-1)}\left\{\overline{N_{nn}(iy)}\, N_{n-1,n-1}(iy) - N_{nn}(iy)\overline{N_{n-1,n-1}(iy)}\right.$$

$$\left. - \overline{D_{nn}(iy)}\, D_{n-1,n-1}(iy) + D_{nn}(iy)\overline{D_{n-1,n-1}(iy)}\right\} .$$

Applying Lemma 3.4 to remove all complements we have

$$|D_{n,n-1}(iy)|^2 - |N_{n,n-1}(iy)|^2$$

$$= \frac{iy}{(2n-1)} \left\{ D_{nn}(iy)N_{n-1,n-1}(iy) - N_{nn}(iy)D_{n-1,n-1}(iy) \right\} .$$

Now by Corollary 3.1

$$|D_{n,n-1}(iy)|^2 - |N_{n,n-1}(iy)|^2 = (-1)^n \left[\frac{(n-1)!\ (iy)^n}{(2n-1)!} \right]^2$$

$$= \left[\frac{(n-1)!}{(2n-1)!} \right]^2 y^{2n} \geq 0$$

and the theorem is proved.

3.5 A Property of $P_{n+1,n}(z)$

An immediate consequence of Theorem 3.3 is one of the results we wished to establish.

Corollary 3.2

The first subdiagonal Padé approximations $P_{n+1,n}(z)$ $(n \geq 0)$ of e^z are bounded by one on the imaginary axis.

Proof

$$|D_{n,n-1}|^2 - |N_{n,n-1}|^2 \geq 0 \text{ for } z = iy, \text{ y real}, n \geq 1,$$

implies

$$|P_{n+1,n}| = \frac{|N_{n+1,n}|}{|D_{n+1,n}|} \leq 1 \quad \text{for } z = iy, \text{ y real}, n \geq 0.$$

We can now turn our attention to showing that the second subdiagonal approximations $P_{n+2,n}(n \geq 0)$ are also bounded by 1 on the imaginary axis. To show this we will show that a theorem analogous to Theorem 3.3 can be established for the difference of the squares of the norms of $D_{n,n-2}$ and $N_{n,n-2}$ for $n > 2$. One preliminary theorem will be required.

Lemma 3.5

For all $n \geq 2$

$$(1) \quad N_{n,n-2}(z) = \frac{1}{2(n-1)} \left[(4n-2)N_{nn}(z) - (2n+z)N_{n-1,n-1}(z) \right]$$

and

$$(2) \quad D_{n,n-2}(z) = \frac{1}{2(n-1)} \left[(4n-2)D_{nn}(z) - (2n+z)D_{n-1,n-1}(z) \right] .$$

Proof

We begin by establishing (1). The proof of (2) will follow by simply replacing N wherever it occurs in the following proof by D. Using relation (1) of Lemma 3.3 with $j = n$ and $k = n-1$ and solving for $N_{n,n-2}$ we obtain

$$N_{n,n-2} = N_{n,n-1} - \frac{nz}{(2n-1)(2n-2)} N_{n-1,n-2}.$$

Now applying (1) of Lemma 3.3 twice more, first with $j = n$ and $k = n$ and then with $j = n-1$ and $k = n-1$ and solving for $N_{n,n-1}$ and $N_{n-1,n-2}$ we obtain

$$\begin{aligned} N_{n,n-2} &= \left[N_{nn} - \frac{nz}{(2n)(2n-1)} N_{n-1,n-1} \right] - \frac{nz}{(2n-1)(2n-2)} \left[N_{n-1,n-1} \right. \\ &\qquad \left. - \frac{(n-1)z}{(2n-2)(2n-3)} N_{n-2,n-2} \right] \\ &= N_{nn} - \frac{z}{(2n-2)} N_{n-1,n-1} + \frac{nz^2}{2(2n-1)(2n-2)(2n-3)} N_{n-2,n-2}. \end{aligned}$$

Now by Theorem 3.1

$$\frac{nz^2}{2(2n-1)(2n-2)(2n-3)} N_{n-2,n-2} = \frac{n}{(n-1)} \left[N_{nn} - N_{n-1,n-1} \right]$$

hence

$$N_{n,n-2} = N_{nn} - \frac{z}{2(n-1)} N_{n-1,n-1} + \frac{n}{(n-1)} \left[N_{nn} - N_{n-1,n-1} \right]$$

$$= \frac{1}{2(n-1)} \left[(4n-2)N_{nn} - (2n+z)N_{n-1,n-1} \right]$$

which establishes the lemma.

We now can prove the analog of Theorem 3.3 for the second subdiagonal Padé approximations.

Theorem 3.4

For all $n \geq 2$, if $z = iy$, y real, then

$$|D_{n,n-2}(z)|^2 - |N_{n,n-2}(z)|^2 = \left[\frac{(n-2)!}{(2n-2)!} \right]^2 y^{2n} \geq 0.$$

Proof

Using Lemma 3.5 we have that for any z

$$|D_{n,n-2}(z)|^2 - |N_{n,n-2}(z)|^2 = D_{n,n-2}\overline{D_{n,n-2}} - N_{n,n-2}\overline{N_{n,n-2}}$$

$$= \frac{1}{4(n-1)^2} \left\{ \left[(4n-2)D_{nn} - (2n+z)D_{n-1,n-1} \right] \left[(4n-2)\overline{D_{nn}} - \overline{(2n+z)}\,\overline{D_{n-1,n-1}} \right] \right.$$

$$\left. - \left[(4n-2)N_{nn} - (2n+z)N_{n-1,n-1} \right] \left[(4n-2)\overline{N_{nn}} - \overline{(2n+z)}\,\overline{N_{n-1,n-1}} \right] \right\}$$

$$= \frac{1}{4(n-1)^2} \Big\{ 4(2n-1)^2 |D_{nn}|^2 + |2n+z|^2 |D_{n-1,n-1}|^2$$

$$- 2(2n-1)\overline{(2n+z)} D_{nn} \overline{D_{n-1,n-1}} - 2(2n-1)(2n+z)\overline{D_{nn}} D_{n-1,n-1}$$

$$- 4(2n-1)^2 |N_{nn}|^2 - |2n+z|^2 |N_{n-1,n-1}|^2$$

$$+ 2(2n-1)\overline{(2n+z)} N_{nn} \overline{N_{n-1,n-1}} + 2(2n-1)(2n+z)\overline{N_{nn}} N_{n-1,n-1} \Big\}$$

Now, just as in Theorem 3.3 we restrict $z = iy$ and apply Corollary 1.1 and then Lemma 3.4 and obtain

$$|D_{n,n-2}(iy)|^2 - |N_{n,n-2}(iy)|^2$$

$$= \frac{2(2n-1)}{4(n-1)^2} \Big[-(2n-iy) D_{nn} N_{n-1,n-1} - (2n+iy) N_{nn} D_{n-1,n-1}$$

$$+(2n-iy) N_{nn} D_{n-1,n-1} + (2n+iy) D_{nn} N_{n-1,n-1} \Big]$$

$$= \frac{(2n-1)(iy)}{(n-1)^2} \Big[D_{nn} N_{n-1,n-1} - N_{nn} D_{n-1,n-1} \Big] \quad .$$

Finally by Corollary 3.1

$$|D_{n,n-2}(iy)|^2 - |N_{n,n-2}(iy)|^2$$

$$= \frac{(2n-1)^2}{(n-1)^2} (-1)^n \left[\frac{(n-1)!\ (iy)^n}{(2n-1)!} \right]^2$$

$$= \left[\frac{(n-2)!}{(2n-2)!} \right]^2 \quad y^{2n} > 0$$

and the theorem is proved.

3.6 A Property of $P_{n+2,n}(z)$

An immediate consequence of Theorem 3.4 is the second result we wished to establish.

Corollary 3.3

The second subdiagonal Padé approximations $P_{n+2,n}(z)$ $(n \geq 0)$ of e^z are bounded by one on the imaginary axis.

Proof

The proof is analogous to that of Corollary 3.2.

We might now ask, could we use the techniques employed above, namely reducing each subdiagonal numerator and denominator to a representation in terms of the diagonal numerators or denominators, to show that all subdiagonal approximations $P_{n+k,n}$ $(k \geq 1)$ are bounded by one on the imaginary axis? The answer is no. In fact all subdiagonal Padé approximations are not bounded by one on the imaginary axis as the following theorem shows.

Lemma 3.6

The Padé approximation $P_{3,0}(z)$ of e^z is not bounded by one for $z = iy$, y real.

Proof

Since $P_{3,0}(z) = \dfrac{1}{1 - z + \frac{z^2}{2} - \frac{z^3}{6}}$ we have that

$$|D_{3,0}(iy)|^2 - |N_{3,0}(iy)|^2 = \left[(1 - \frac{y^2}{2})^2 + (-y + \frac{y^3}{6})^2 \right] - \left[1 \right]$$

$$= \frac{y^4}{36} (y^2 - 3).$$

It is clear that for $-\sqrt{3} < y < \sqrt{3}$ $|D_{3,0}|^2 - |N_{3,0}|^2 \leq 0$ and hence

$$|P_{3,0}(iy)| = \frac{|N_{3,0}(iy)|}{|D_{3,0}(iy)|} \geq 1.$$

Thus the theorem is proved.

If we apply the techniques used to show that $P_{n+1,n}$ and $P_{n+2,n}$ were bounded on the imaginary axis to $P_{n+3,n}$ we will only succeed in generalizing the previous lemma as follows.

Theorem 3.5

For all $n \geq 3$,

$$|D_{n,n-3}(iy)|^2 - |N_{n,n-3}(iy)|^2 = (y^2 - n^2 + 2n) \left[\frac{(n-3)!\ y^{n-1}}{(2n-3)!}\right]^2$$

and hence $P_{n,n-3}(iy)$ is not bounded by one over the interval

$$-\sqrt{n^2 - 2n} < y < \sqrt{n^2 - 2n} \qquad \text{for } n \geq 3.$$

Proof

We will sketch only briefly the proof of this theorem since it is proved in the same basic way that Theorem 3.3 and Theorem 3.4 were proved. We begin by establishing that

$$N_{n,n-3}(z) = \frac{(n - z - 2)}{(n - 2)} N_{n-1,n-1}(z) + \frac{z(n + z)}{2(n-2)(2n-3)} N_{n-2,n-2}(z)$$

and

$$D_{n,n-3}(z) = \frac{(n - z - 2)}{(n - 2)} D_{n-1,n-1}(z) + \frac{z(n + z)}{2(n-2)(2n-3)} D_{n-2,n-2}(z)$$

by using Lemma 3.3, Lemma 3.5 and Theorem 3.1. With these relations for $N_{n,n-3}(z)$ and $D_{n,n-3}(z)$ we compute $|D_{n,n-3}|^2 - |N_{n,n-3}|^2$. Applying Corollary 1.1, Lemma 3.4, and finally Corollary 3.1 as we did in Theorem 3.3 we obtain the result given.

3.7 The Analyticity of $P_{n+1,n}(z)$ and $P_{n+2,n}(z)$

We now turn our attention to the second result which must be established if we are to prove that Property (2) holds. Specifically we must show that $P_{n+1,n}(z)$ and $P_{n+2,n}(z)$ for all $n \geq 0$ are analytic in the left half plane. This will require showing, just as it did for $P_{n,n}(z)$, that the denominators of $P_{n+1,n}(z)$ and $P_{n+2,n}(z)$ have no zeros in the left half plane or, what is equivalent by Corollary 1.1, showing that the numerators of $P_{n,n+1}(z)$ and $P_{n,n+2}(z)$ have all their zeros in the left half plane. We choose to work with $N_{n,n+1}(z)$ and $N_{n,n+2}(z)$ because all of their coefficients are positive while those of $D_{n+1,n}(z)$ and $D_{n+2,n}(z)$ alternate in sign.

The most obvious way to attempt to show that the zeros of $N_{n,n+1}(z)$ and $N_{n,n+2}(z)$ are in the left half plane is to apply the same technique which was used on $N_{n,n}(z)$ in Chapter 2 to show that its zeros were all in the left half plane. Unfortunately, a little investigation will show that the continued fractions produced by $N_{n,n+1}(z)$ and $N_{n,n+2}(z)$ using the technique of Chapter 2 are not easily related to one another, as were those of $N_{n,n}(z)$, and hence some other method of attack must be found. To this end we begin by establishing the following theorems.

Theorem 3.6

For all $j \geq 0$, $k \geq 0$,

$$N_{j,k}(z) = \frac{(j+k+1)}{(k+1)} N'_{j,k+1}(z) \quad .$$

Proof

From Theorem 1.10 we have that

$$N_{j,k+1}(z) = \sum_{m=0}^{k+1} \frac{(j+k+1-m)!\ (k+1)!}{(j+k+1)!\ m!\ (k+1-m)!} z^m$$

Thus

$$N'_{j,k+1}(z) = \sum_{m=0}^{k+1} \frac{(j+k+1-m)!\ (k+1)!\ m}{(j+k+1)!\ m!\ (k+1-m)!} z^{m-1}$$

$$= \sum_{m=1}^{k+1} \frac{(j+k+1-m)!\ (k+1)!}{(j+k+1)!\ (m-1)!\ (k+1-m)!} z^{m-1}$$

$$= \sum_{m=0}^{k} \frac{(j+k-m)!\ (k+1)!}{(j+k+1)!\ m!\ (k-m)!} z^m$$

$$= \frac{(k+1)}{(j+k+1)} \sum_{m=0}^{k} \frac{(j+k-m)!\ k!}{(j+k)!\ m!\ (k-m)!} z^m$$

$$= \frac{(k+1)}{(j+k+1)} N_{j,k}(z).$$

The theorem follows at once from this last result.

Lemma 3.7

The following two relations hold between various numerators of the Padé approximations of e^z for all $j \geq 0$, $k \geq 0$.

$$(1) \quad N_{j,k+1}(z) = N_{j+1,k}(z) + \frac{z}{j+k+1} \; N_{j,k}(z)$$

$$(2) \quad N_{j,k+1}(z) = \frac{j+k+2}{j+1} \, N_{j+1,k+1}(z) - \frac{k+1}{j+1} \, N_{j+1,k}(z)$$

Proof

For the first relation we have

$$N_{j+1,k}(z) + \frac{z}{j+k+1} N_{j,k}(z)$$

$$= \sum_{m=0}^{k} \frac{(j+k+1-m)! \; k! \; z^m}{(j+k+1)! \; m! \; (k-m)!} + \frac{z}{j+k+1} \sum_{m=0}^{k} \frac{(j+k-m)! \; k! \; z^m}{(j+k)! \; m! \; (k-m)!}$$

$$= 1 + \sum_{m=1}^{k} \left[\frac{(j+k+1-m)! \; k! \; z^m}{(j+k+1)! \; m! \; (k-m)!} \right] + \left[\sum_{m=0}^{k-1} \frac{(j+k-m)! \; k! \; z^{m+1}}{(j+k+1)! \; m! \; (k-m)!} + \frac{j! \; z^{k+1}}{(j+k+1)!} \right]$$

$$= 1 + \sum_{m=1}^{k} \frac{(j+k+1-m)! \; k! \; z^m}{(j+k+1)! \; m! \; (k-m)!} + \sum_{m=1}^{k} \frac{(j+k+1-m)! \; k! \; z^m}{(j+k+1)! \; (m-1)! \; (k+1-m)!} + \frac{j! \; z^{k+1}}{(j+k+1)!}$$

$$= 1 + \sum_{m=1}^{k} \frac{(j+k+1-m)! \; k!}{(j+k+1)! \; (m)! \; (k+1-m)!} \left\{ (k+1-m) + (m) \right\} z^m + \frac{j! \; z^{k+1}}{(j+k+1)!}$$

$$= 1 + \sum_{m=1}^{k} \frac{(j+k+1-m)! \; (k+1)!}{(j+k+1)! \; m! \; (k+1-m)!} \, z^m + \frac{j! \; z^{k+1}}{(j+k+1)!}$$

$$= \sum_{m=0}^{k+1} \frac{(j+k+1-m)! \; (k+1)!}{(j+k+1)! \; m! \; (k+1-m)!} \, z^m$$

$$= N_{j,k+1}(z) \; .$$

For the second relation we have

$$\frac{j+k+2}{j+1} N_{j+1,k+1}(z) - \frac{k+1}{j+1} N_{j+1,k}(z)$$

$$= \frac{j+k+2}{j+1} \sum_{m=0}^{k+1} \frac{(j+k+2-m)!\,(k+1)!\ z^m}{(j+k+2)!\ m!\ (k+1-m)!} - \frac{k+1}{j+1} \sum_{m=0}^{k} \frac{(j+k+1-m)!\,(k)!\ z^m}{(j+k+1)!\ m!\ (k-m)!}$$

$$= \frac{1}{j+1} \sum_{m=0}^{k} \left\{ \frac{(j+k+2-m)!\ (k+1)!}{(j+k+1)!\ m!\ (k+1-m)!} - \frac{(j+k+1-m)!\ (k+1)!}{(j+k+1)!\ m!\ (k-m)!} \right\} z^m$$

$$+ \frac{j!\ (k+1)!}{(j+k+1)!} z^{k+1}$$

$$= \frac{1}{j+1} \sum_{m=0}^{k} \frac{(j+k+1-m)!\ (k+1)!}{(j+k+1)!\ m!\ (k+1-m)!} \left\{ (j+k+2-m) - (k+1-m) \right\} z^m$$

$$+ \frac{j!\ (k+1)!}{(j+k+1)!} z^{k+1}$$

$$= \sum_{m=0}^{k+1} \frac{(j+k+1-m)!\ (k+1)!}{(j+k+1)!\ m!\ (k+1-m)!} z^m = N_{j,k+1}(z)$$

Applying Theorem 3.6 to relation (2) of Lemma 3.7 we have at once that

$$N_{j,k+1}(z) = \frac{j+k+2}{j+1} \left\{ N_{j+1,k+1}(z) - N'_{j+1,k+1}(z) \right\} .$$

Noting that this equation is also trivially satisifed if $k = -1$ we have the following corollary.

<u>Corollary 3.4</u>

For all $j \geq 0$, $k \geq 0$,

$$N_{j,k}(z) = \frac{j+k+1}{j+1}\left\{N_{j+1,k}(z) - N'_{j+1,k}(z)\right\} .$$

Substitution of the result of Corollary 3.4 into the first relation in Lemma 3.7 now gives a result which will be very useful, namely,

Theorem 3.7

For all $j \geq 0$, $k \geq 0$,

$$N_{j,k+1}(z) = \left[1 + \frac{z}{(j+1)}\right] N_{j+1,k}(z) - \frac{z}{(j+1)} N'_{j+1,k}(z).$$

In particular, we will use the fact that for $j = n$, $k = n+1$, $n \geq 0$, we have that

$$N_{n,n+2}(z) = \left[1 + \frac{z}{(n+1)}\right] N_{n+1,n+1}(z) - \frac{z}{(n+1)} N'_{n+1,n+1}(z).$$

If we now define $\theta_n(z)$ to be

$$\theta_n(z) = \frac{N_{n,n+2}(z)}{\frac{-z}{n+1} N_{n+1,n+1}(z)} = \left[\frac{-(n+1)}{z} -1\right] + \frac{N'_{n+1,n+1}(z)}{N_{n+1,n+1}(z)}$$

clearly the only zeros of $N_{n,n+2}(z)$ which are not zeros of $\theta_n(z)$ are zeros of $N_{n+1,n+1}(z)$.

Thus, since all the zeros of $\theta_n(z)$ are zeros of $N_{n,n+2}(z)$, if we can show that all the zeros of $\theta_n(z)$ are in the left half plane and also show that all the zeros of $N_{n+1,n+1}(z)$ are in the left half plane we will have shown that all the zeros of $N_{n,n+2}(z)$ are in the left half plane. The second condition has of course been established in Chapter 2, Theorem 2.1.

In order to study the zeros of $\theta_n(z)$ we consider the region S, bounded by the curve C, which is shown in Figure 3.1.

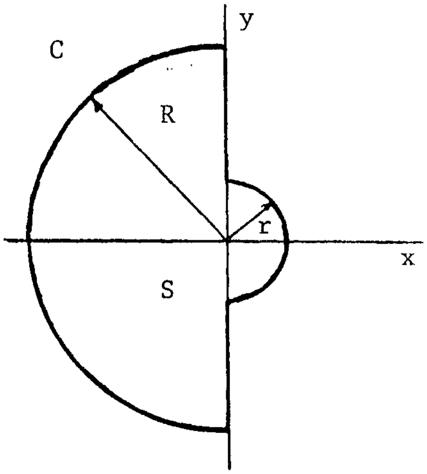

Figure 3.1

The boundary curve C is composed of the semicircle $|z| = R$, $\text{Re}(z) \leq 0$, R chosen so that all the zeros of $N_{n+1,n+1}(z)$ are inside $|z| = R$, the semicircle $|z| = r$, $r > 0$, $\text{Re}(z) \geq 0$, r chosen so that all the zeros of $N_{n,n+2}(z)$ are outside the circle $|z| = r$, and the imaginary axis from $-R \leq y \leq -r$ and $r \leq y \leq R$. That values of $R < \infty$ and $r > 0$ can be found which satisfy these conditions follows at once from Theorems 1.2 and 1.3 and the known form of the polynomials $N_{n+1,n+1}(z)$ and $N_{n,n+2}(z)$.

Applying Theorem 1.4 to $\theta_n(z)$ and assuming that all the zeros of $N_{n+1,n+1}(z)$ are in the left half plane we have that

$$\begin{aligned}\frac{1}{2\pi i} \int_C \frac{\theta_n'(z)}{\theta_n(z)}\, dz &= N_C(\theta_n) - P_C(\theta_n) \\ &= N_C(N_{n,n+2}(z)) - N_C(\tfrac{-z}{n+1} N_{n+1,n+1}(z)) \\ &= N_C\,(N_{n,n+2}(z)) - (n+2)\end{aligned}$$

Thus we have established the following.

Theorem 3.8

If all the zeros of $N_{n+1,n+1}(z)$ are in the left half plane then

$$N_c(N_{n,n+2}(z)) = (n+2) + \frac{1}{2\pi i} \int_C \frac{\theta_n'(z)}{\theta_n(z)} dz$$

where C is the curve in Figure 3.1.

Now, if we could show that $\int_C \frac{\theta_n'}{\theta_n} dz = 0$ we would have established that all the zeros of $N_{n,n+2}(z)$ are inside the region S provided all the zeros of $N_{n+1,n+1}(z)$ are in the left half plane. But since all the zeros of $N_{n,n+2}(z)$ are outside the circle of radius r, we would also be able to conclude that all the zeros of $N_{n,n+2}(z)$ are in the region S' given in Figure 3.2

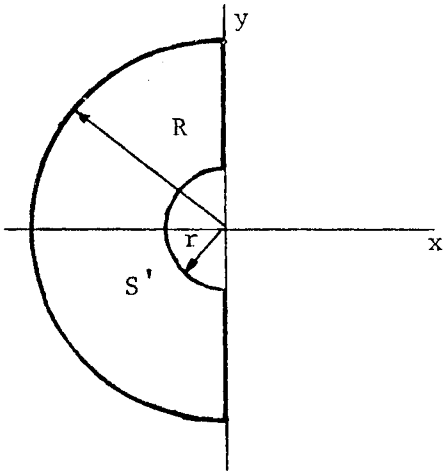

Figure 3.2

But since S' is entirely in the left half plane we would have shown that all the zeros of $N_{n,n+2}(z)$ are in the left half plane.

By Theorem 1.9, to evaluate $\int_C \frac{\theta_n'}{\theta_n} dz$ it is sufficient to determine the index of $\theta_n(C)$, that is, the index of the curve into which C is mapped by $\theta_n(z)$, with respect to the origin.* For convenience we shall call this new curve C*. We shall now show that as we travel along C* we have that its real part is always negative and hence by Theorem 1.8 its index with respect to the origin is zero.

Lemma 3.8

For $|z| = R$, R sufficiently large, $\mathrm{Re}(\theta_n(z)) < 0$.

Proof

For $|z|$ large enough, $\theta_n(z) = -1 + 0(\frac{1}{z})$, hence for sufficiently large R the result follows.

Lemma 3.9

For $|z| = r$, $r > 0$, r sufficiently small, $\mathrm{Re}(z) > 0$, $\mathrm{Re}(\theta_n(z)) < 0$.

Proof

Since as $|z| \to 0$, $\frac{N'_{n+1,n+1}(z)}{N_{n+1,n+1}(z)} \to \frac{1}{2}$ we have that for r sufficiently small $\mathrm{Re}(\frac{N'_{n+1,n+1}(z)}{N_{n+1,n+1}(z)}) \leq \frac{3}{4}$. Thus for r sufficiently small we have

$$\mathrm{Re}(\theta_n(z)) \leq \left\{\frac{-(n+1)\mathrm{Re}(z)}{r^2} -1\right\} + \frac{3}{4} < 0.$$

* (Dieudonné has used this idea [16] in establishing a result which is similar in spirit to the result we are attempting to establish [14], [47, p. 87].)

The problem now remains to establish that for $z = iy$, $r \leq y \leq R$ and $-R \leq y \leq -r$, that $\mathrm{Re}(\theta_n(z)) < 0$. Looking at $\theta_n(z)$ we see that for $z = iy$, $y \neq 0$, we would have the result if we could show that

$$\mathrm{Re}\left(\frac{N'_{n+1,n+1}(iy)}{N_{n+1,n+1}(iy)}\right) < 1.$$

To establish that this inequality does, in fact, hold, it will be sufficient to establish the following theorem.

Theorem 3.9

If $N^c_{n,n}(z) = N_{n,n}(\bar{z})$, that is $N^c_{n,n}(z)$ is the complement of $N_{n,n}(z)$, then for all $n \geq 0$ we have

$$2\mathrm{Re}(N'_{n,n}(z)N^c_{n,n}(z)) + \frac{(-1)^n \; z^{2n}(n!)^2}{[(2n)!]^2} = N_{n,n}(z)N^c_{n,n}(z)$$

when $z = iy$.

In order to establish Theorem 3.9 we will require two preliminary results. The first of these is,

Lemma 3.10

For all $n \geq 0$ and all z

$$N'_{n,n}(z)N^c_{n,n}(z) = \frac{1}{2}\left[N_{n,n}(z)N^c_{n,n}(z) - \frac{z}{2(2n-1)} N_{n-1,n-1}(z)N^c_{n,n}(z)\right].$$

Proof

We begin by observing that for $j = n$, $k = n$, relation (1) in Lemma 3.3 can be rewritten as

$$N_{n,n-1}(z) = N_{n,n}(z) - \frac{z}{2(2n-1)} N_{n-1,n-1}(z).$$

Now by Theorem 3.6

$$N'_{n,n}(z)N^c_{n,n}(z) = \frac{1}{2} N_{n,n-1}(z)N^c_{n,n}(z)$$

$$= \frac{1}{2}\left[N_{n,n}(z) - \frac{z}{2(2n-1)} N_{n-1,n-1}(z)\right] N^c_{n,n}(z)$$

and the lemma is proved.

The second result is,

Lemma 3.11

For $z = iy$, y real and all $n \geq 1$ the product $N_{n-1,n-1}(z)N^c_{n,n}(z)$ has only real terms of even powers of y except for its term of highest power which has the form

$$(-i) \frac{n!\ (n-1)!}{(2n)!\ (2n-2)!} y^{2n-1} .$$

Proof

Observing that for $z = iy$, $N^c_{n,n}(z) = D_{n,n}(z)$ we can apply Theorem 3.2 and the result follows at once. We are now ready to establish Theorem 3.9.

Proof (of Theorem 3.9)

By Lemma 3.10,

$$\mathrm{Re}(N_{n,n}(iy)N^c_{n,n}(iy)) = \frac{1}{2} \mathrm{Re}(N_{n,n}(iy)N^c_{n,n}(iy)) -$$

$$\frac{1}{2} \mathrm{Re}\left(\frac{iy}{2(2n-1)} N_{n-1,n-1}(iy)N^c_{n,n}(iy)\right) .$$

But

$$\mathrm{Re}(N_{n,n}(iy)N^c_{n,n}(iy)) = N_{n,n}(iy)N^c_{n,n}(iy)$$

and by Lemma 3.11

$$\operatorname{Re}\left(\frac{iy}{2(2n-1)} N_{n-1,n-1}(iy)N^c_{n,n}(iy)\right)$$

$$= \frac{n!\ (n-1)!\ y^{2n}}{2(2n-1)(2n)!\ (2n-2)!} = \frac{(n!)^2}{[(2n)!]^2}\, y^{2n}.$$

Thus,

$$\operatorname{Re}(N'_{n,n}(iy)N^c_{n,n}(iy)) = \frac{1}{2}\left\{N_{n,n}(iy)N^c_{n,n}(iy) - \frac{(n!)^2}{[(2n)!]^2}\, y^{2n}\right\}$$

which establishes the theorem.

Corollary 3.5

For all $n \geq 0$, y real, $y \neq 0$,

$$\operatorname{Re}\left(\frac{N'_{n+1,n+1}(iy)}{N_{n+1,n+1}(iy)}\right) < \frac{1}{2}\ .$$

Proof

From Theorem 3.9 we have

$$\frac{2\ \operatorname{Re}(N'_{n+1,n+1}(iy)N^c_{n+1,n+1}(iy)) + \dfrac{(n!)^2\ y^{2n}}{[(2n)!]^2}}{N_{n+1,n+1}(iy)N^c_{n+1,n+1}(iy)} = 1.$$

Thus

$$\frac{\operatorname{Re}(N'_{n+1,n+1}(iy)N^c_{n+1,n+1}(iy))}{N_{n+1,n+1}(iy)N^c_{n+1,n+1}(iy)} < \frac{1}{2}\ .$$

But

$$\frac{\operatorname{Re}(N'_{n,n}(z)N^c_{n,n}(z))}{N_{n,n}(z)N^c_{n,n}(z)} = \operatorname{Re}\left(\frac{N'_{n,n}(z)}{N_{n,n}(z)}\right) \quad \text{for all } n \geq 0$$

and hence the corollary is proved.

Corollary 3.5 immediately proves the following lemma.

Lemma 3.12

For $z = iy$, y real, $r \le |y| \le R$, $\text{Re}(\theta_n(z)) \le -\frac{1}{2}$.

Lemmas 3.8, 3.9 and 3.10 now establish the result that,

Lemma 3.13

The index of C* with respect to the origin is zero.

Finally, from Theorem 3.8, Lemma 3.13 and the discussion following Theorem 3.8 we have the following theorem.

Theorem 3.10

For any $n \ge 0$, if $N_{n,n}(z)$ has all of its zeros in the left half plane, then $N_{n,n+2}(z)$ has all of its zeros in the left half plane also.

Applying Theorem 2.1 to Theorem 3.10 results in the following corollary.

Corollary 3.6

For all $n \le 0$, $N_{n,n+2}(z)$ has all of its zeros in the left half plane.

As an immediate corollary to Corollary 3.6 we have

Corollary 3.7

For all $n \ge 0$, $N_{n,n+1}(z)$ has all of its zeros in the left half plane.

Proof

This result follows at once from Theorem 3.6 and Theorem 1.1.

We are now in a position to establish that the second condition which must be satisfied if we are to prove that Property (2) holds for the first two rows of subdiagonal Padé approximations is satisfied.

Theorem 3.11

For all $n \geq 0$, $P_{n+2,n}(z)$ and $P_{n+1,n}(z)$, the first and second subdiagonal Padé approximations to the exponential function, are analytic in the left half plane.

Proof

Since $D_{n+2,n}(z) = N_{n,n+2}(-z)$ and $D_{n+1,n}(z) = N_{n,n+1}(-z)$ by Corollary 3.6 and Corollary 3.7 we have that all the zeros of the denominators of $P_{n+2,n}(z)$ and $P_{n+1,n}(z)$ are in the right half plane. Thus, since $P_{n+2,n}(z)$ and $P_{n+1,n}(z)$ are quotients of polynomials and the only poles of the functions occur when the denominator is zero, we have that $P_{n+2,n}(z)$ and $P_{n+1,n}(z)$ are by definition analytic in the left half plane.

We note that Theorem 3.11 proves half of Conjecture 2.1.

3.8 The Fundamental Result

We can now establish the major results of this thesis in the following two theorems.

Theorem 3.12

For all $n \geq 0$, $P_{n+2,n}(z)$ and $P_{n+1,n}(z)$ satisfy Property (2).

Proof

The proof follows at once from Theorem 3.11, Corollary 3.2, Corollary 3.3, and Theorem 1.7.

It follows at once from Theorem 3.12 and from the discussion in Section 3.2 that we also have

Theorem 3.13

$P_{n+1,n}(z)$ and $P_{n+2,n}(z)$ are L-acceptable approximations to the exponential function for all $n \geq 0$.

Thus we have found that in addition to the diagonal Padé approximations to the exponential, the first two subdiagonal sets of Padé approximations to the exponential satisfy properties which make them of interest in methods involving stiff equations.

In the next chapter it will be shown that there are several different numerical methods which correspond to the various Padé approximations to the exponential. In particular, there are methods which correspond to the set of diagonal and first two sets of subdiagonal Padé approximations. Before turning our attention to applications of these results, however, we shall show that an alternative proof can be given that Property (2) holds which does not rely upon Theorem 2.1. To do this, two additional theorems will be required.

Lemma 3.14

For all $j \geq 0$, $k \geq 0$,

$$N_{j+1,k} = N_{j,k} - \frac{z}{j+k+1} N'_{j,k} .$$

Proof

The above result follows at once from Property (3) of Lemma 3.3 and Theorem 3.6.

Theorem 3.14

If for some $j,k \geq 0$, $N_{j,k}(z)$ has all of its zeros in the open left half plane, then for all $m \geq j$, $N_{m,k}(z)$ has all of its zeros in the open left half plane also.

Proof

By the assumption of the theorem we have that it is trivially true for $m = j$.

Now assume that for some m all of the zeros of $N_{m,k}(k)$ are in the open left half plane. Then all of the zeros of $N_{m,k}(z)$ are contained in a circular region C which lies entirely in the left half plane. Defining $\Omega_{m+1,k}(z)$ to be

$$\Omega_{m+1,k}(z) = (m+k+1)N_{m,k}(z) - zN'_{m,k}(z)$$

we have on applying Theorem 1.5, noting that $\beta_1 = m+j+1$ and $\eta = k$, that all the zeros of $\Omega_{m+1,k}(z)$ are contained in a circular region in the left half plane. But by Lemma 3.14

$$\Omega_{m+1,k}(z) = (m+k+1)N_{m+1,k}(z)$$

thus all the zeros of $N_{m+1,k}(z)$ are in the open left half plane and the theorem is proved by induction.

3.9 An Alternative Proof of the Fundamental Result

Using Theorem 3.14 it is now possible to give an alternative proof of Corollaries 3.6 and 3.7 without using Theorem 2.1 which specified the location of the zeros of the diagonal terms. In fact, this alternative proof gives a second proof of Theorem 2.1 which does not rely on continued fractions. The basis of the argument is suggested by the schematic diagram given in Figure 3.3.

$$\begin{tikzcd}
 & N_{0,2} \arrow[dd, dashed] & & & & \\
N_{1,1} \arrow[ur] & & N_{1,3} \arrow[dd, dashed] & & & \\
 & N_{2,2} \arrow[ur] & & N_{2,4} \arrow[dd, dashed] & & \\
 & & N_{3,3} \arrow[ur] & & N_{3,5} \arrow[d, dashed] & \\
 & & & N_{4,4} \arrow[ur] & {} \arrow[r, phantom] & {} \\
 & & & {} \arrow[uur] & & \ddots
\end{tikzcd}$$

↗ = Theorem 3.10

⇣ = Theorem 3.14

Figure 3.3

That is, by inspection $N_{1,1}(z)$ has all of its zeros in the left half plane. Thus by Theorem 3.10, $N_{0,2}(z)$ has all of its zeros in the left half plane. Then by applying Theorem 3.14 with $j = 0$, $k = 2$, and $m = 2$ we have that $N_{2,2}(z)$ has all of its zeros in the left half plane. An application of Theorem 3.10 now establishes that $N_{1,3}(z)$ has all of its zeros in the left half plane. The induction step is obvious and hence we have that both Theorem 2.1 and Corollary 3.6 are true for all n. Corollary 3.7 follows from Corollary 3.6 as before. The fundamental result, Theorem 3.13, also follows as before.

3.10 Conclusions

In conclusion we note that we have shown that all the entries on the diagonal and first two subdiagonals of the Padé table of the exponential function are A-acceptable approximations to the exponential and furthermore the first two sets of subdiagonal entries satisfy the stronger condition of being L-acceptable approximations. We have also shown that no entry in the third subdiagonal can satisfy either of these properties. Since no above diagonal entry in the Padé table can be A-acceptable, we can conclude our discussion of the Padé approximations to the exponential with the following conjecture.

Conjecture 3.1

The diagonal and first two subdiagonal sets of entries in the Padé table are the only A-acceptable Padé approximations to the exponential function.

CHAPTER 4

Some Examples of High Order A-stable Numerical Methods

4.1 Introduction

In this chapter we will show that there are at least two classes of one step methods of arbitrarily high order which satisfy Dahlquist's definition of A-stability (Definition 1.3). Examples of methods which reduce to A-acceptable and L-acceptable approximations will be given. Although most of these methods appear at the moment to be largely of theoretical interest, several examples of their practical application will be given in the chapter which follows.

4.2 Some Results of Butcher

In a series of papers [4], [5], [6], Butcher has studied a generalization of the well known ν stage Runge-Kutta process given by the equations

$$\vec{K}_i = \vec{f}(\vec{y}_n + h \sum_{j=1}^{i-1} \beta_{ij} \vec{K}_j) \qquad (i = 1, 2, \ldots, \nu) \qquad (4.2.1)$$

$$\vec{y}_{n+1} = \vec{y}_n + h \sum_{i=1}^{\nu} b_i \vec{K}_i \qquad (4.2.2)$$

by allowing the summation in (4.2.1) to be carried from 1 to ν for each $\vec{K}_i$ and assuming that not all the β_{ij} $(j \geq i)$ are zero. The resulting Runge-Kutta processes are called semi-explicit if $\beta_{ij} = 0$ $(j > i)$ and implicit if $\beta_{ij} \neq 0$ for at least one $j > i$.

Using an ν stage (implicit) Runge-Kutta process, Butcher was able to show that it is possible to obtain Runge-Kutta processes of order 2ν, $2\nu - 1$, and $2\nu - 2$. The values of β_{ij} and b_i $(i, j = 1, 2, \ldots, \nu)$ for all three cases for $\nu \leq 5$, 3, and 7 respectively are given in [5] and [6]. In particular, for $\nu = 2$, the unique 4th order process is given by

$$\begin{aligned} \beta_{11} &= \frac{1}{4} & \beta_{12} &= \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \beta_{21} &= \frac{1}{4} + \frac{\sqrt{3}}{6} & \beta_{22} &= \frac{1}{4} \\ b_1 &= \frac{1}{2} & b_2 &= \frac{1}{2} \end{aligned} \tag{4.2.3}$$

The two possible third order processes considered by Butcher for $\nu = 2$ are given by

$$\begin{aligned} \beta_{11} &= 0 & \beta_{12} &= 0 \\ \beta_{21} &= \frac{1}{3} & \beta_{22} &= \frac{1}{3} \\ b_1 &= \frac{1}{4} & b_2 &= \frac{3}{4} \end{aligned} \tag{4.2.4}$$

and

$$\begin{aligned} \beta_{11} &= \frac{1}{3} & \beta_{12} &= 0 \\ \beta_{21} &= 1 & \beta_{22} &= 0 \\ b_1 &= \frac{3}{4} & b_2 &= \frac{1}{4} \end{aligned} \tag{4.2.5}$$

while the second order process corresponding to $\nu = 2$ is given by

$$\begin{array}{lll} \beta_{11} = 0 & \beta_{12} = 0 & \\ \beta_{21} = 1 & \beta_{22} = 0 & (4.2.6) \\ b_1 = \frac{1}{2} & b_2 = \frac{1}{2} & \end{array}$$

and is seen to be just the improved Euler (or Heun) method [26, p. 67]. Clearly [26, p. 67] a wide variety of other second order 2 stage Runge-Kutta processes could have been included but were not since Butcher was basing his ν stage, order $2\nu - 2$, process on Lobatto quadrature formulas [52, p. 108].

4.3 Butcher's A-stable Methods

Now, it is not difficult to show that when Butcher's ν stage method is applied to (1.2.3) we obtain simply the ν^{th} diagonal Padé approximation $P_{\nu\nu}(z)$ of exp(z) [20]. This follows from the fact that Butcher's ν stage method applied to (1.2.3) reduces to the quotient of two polynomials of degree ν in qh and since this rational function must be an approximation of order 2ν of the exponential function it follows (Section 2.2) that it is the diagonal Padé approximation $P_{\nu\nu}(qh)$ of exp(qh).

For example, when Butcher's 2 stage method (4.2.3) is applied to (1.2.3) we have that

$$K_1 = q(y_n + \frac{h}{4} K_1 + (\frac{1}{4} - \frac{\sqrt{3}}{6}) hK_2),$$

$$K_2 = q(y_n + (\frac{1}{4} + \frac{\sqrt{3}}{6}) hK_1 + \frac{h}{4} K_2).$$

Solving for K_1 and K_2 using Cramer's rule [2, p. 306] gives

$$K_1 = qy_n(1 - \frac{qh\sqrt{3}}{6}) / \Delta \quad ,$$

$$K_2 = qy_n(1 + \frac{qh\sqrt{3}}{6}) / \Delta \quad ,$$

where

$$\Delta = (1 - \frac{1}{2}(qh) + \frac{(qh)^2}{12}).$$

Finally,

$$y_{n+1} = y_n + h(b_1K_1 + b_2K_2) \qquad (4.3.1)$$

$$= \left[\frac{1 + \frac{1}{2}(qh) + \frac{(qh)^2}{12}}{1 - \frac{1}{2}(qh) + \frac{(qh)^2}{12}}\right] y_n$$

$$= P_{22}(qh)\ y_n \ .$$

Clearly this same procedure can be applied to any of Butcher's ν stage processes of order 2ν and so we have in general that for any of Butcher's ν stage processes of order 2ν that

$$y_{n+1} = P_{\nu\nu}(qh)\ y_n.$$

when solving (1.2.3). Now, by Theorem 2.2, it follows that all of Butcher's ν stage processes of order 2ν are A-stable.

4.4 Methods of Butcher Which Are Not A-stable

It is natural to ask next if Butcher's ν stage methods of order $2\nu - 1$ and $2\nu - 2$ are also A-stable. If they reduce to entries in either the diagonal or first two subdiagonals of the Padé table when solving (1.2.3) then the answer would be yes.

Looking first at solving (1.2.3) with (4.2.4) we see that

$$K_1 = qy_n$$
$$K_2 = q(y_n + \frac{h}{3} K_1 + \frac{h}{3} K_2).$$

If we solve for K_1 and K_2 using Cramer's rule, which at the moment may seem like opening a peanut with a sledge hammer, we see that

$$K_1 = qy_n(1 - \frac{qh}{3}) \;/\; \Delta$$

$$K_2 = qy_n(1 + \frac{qh}{3}) \;/\; \Delta$$

where

$$\Delta = \begin{vmatrix} 1 & 0 \\ -\frac{qh}{3} & (1 - \frac{qh}{3}) \end{vmatrix} = (1 - \frac{qh}{3})$$

Substituting into (4.3.1) with the appropriate values of b_i we obtain

$$y_{n+1} = \left[\frac{1 + \frac{2}{3}(qh) + \frac{(qh)^2}{6}}{1 - \frac{qh}{3}} \right] y_n \qquad (4.4.1)$$

$$= \quad P_{1,2}(qh)\ y_n \ .$$

Unfortunately, $P_{1,2}(z)$ is an above diagonal Padé approximation to the exponential and hence by our previous discussion method (4.2.4) is not A-stable.

Now method (4.2.4) is typical of a class of ν stage solutions of order $2\nu - 1$ proposed by Butcher which are based on Radau quadrature [52, p. 108] with the left end fixed. In developing this whole class of methods, called type I methods by Butcher in [6], the assumption is made that $\beta_{1j} = 0$ for $j = 1, 2, \ldots, \nu$. Thus, for all type I methods the highest power of (qh) which can occur in the expansion of Δ in Cramer's rule is $(qh)^{\nu-1}$ since all terms involving (qh) are linear and the first row of Δ contains no terms in (qh). It follows that for all ν stage methods of type I we will have that

$$y_{n+1} = R_{\nu-1,\nu}(qh)\ y_n$$

where $R_{\nu-1,\nu}(qh)$ is a rational function of two polynomials in (qh) where the denominator is of degree at most $\nu - 1$ and the numerator is of degree at most ν. But since $R_{\nu-1,\nu}(z)$ must be a $2\nu - 1$ order approximation to the exponential we have that for all $\nu \leq 1$

$$R_{\nu-1,\nu}(z) = P_{\nu-1,\nu}(z)$$

and hence none of Butcher's type I methods are A-stable.

Now turning our attention to (4.2.5) we solve (1.2.3). This gives

$$K_1 = q(y_n + \frac{h}{3} K_1)$$

$$K_2 = q(y_n + h\ K_1) \ .$$

Again using Cramer's rule we obtain

$$K_1 = q\ y_n \ / \ \Delta$$

$$K_2 = q\ y_n(1 + \frac{2qh}{3}) \ / \ \Delta$$

where

$$\Delta = \begin{vmatrix} 1 - \frac{qh}{3} & 0 \\ -qh & 1 \end{vmatrix} = 1 - \frac{qh}{3} \ .$$

Substituting into (4.3.1) with the appropriate b_i we again obtain (4.4.1) and thus method (4.2.5) is not A-stable.

Now method (4.2.5) is typical of a class of ν stage methods of order $2\nu - 1$ proposed by Butcher which are based on Radau quadrature with the right end fixed. In developing this whole class of methods, called type II methods by Butcher in [6], the assumption was made that $\beta_{j,\nu} \equiv 0$ for $j = 1, 2, \ldots, \nu$. Thus, in a manner similar to that of type I methods we have that Δ can be of degree at most $\nu - 1$ in (qh) and thus all of Butcher's type II methods reduce to above diagonal approximations to the exponential and therefore are also not A-stable.

Finally we have method (4.2.6). By inspection we see that this reduces to

$$y_{n+1} = P_{0,2}(qh)\ y_n$$

when solving (1.2.3) and thus is not A-stable. As noted earlier, method (4.2.6) is based on Lobatto quadrature and it is typical of a class of ν stage methods or order $2\nu - 2$, called type III methods by Butcher in [6]. All type III methods satisfy the property that $\beta_{1,j} \equiv 0$ and $\beta_{j,\nu} \equiv 0$ for $j = 1, 2, \ldots, \nu$. It follows that when solving by Cramer's rule that Δ can be at most of degree $2\nu - 2$ in (qh) and hence a type III, ν stage, process reduces to

$$y_{n+1} = P_{\nu-2,\nu}(qh)\ y_n$$

and none of the type III processes of Butcher are A-stable.

We conclude this section by noting that although none of the methods based on Radau or Lobatto quadrature which were proposed by Butcher are A-stable they may still have considerably better stability properties than explicit processes of comparable or greater accuracy. To illustrate this, we plot in Figure 4.1 the stability regions for the improved Euler method, the classical 4^{th} order Runge-Kutta process (1.1.6) and Lawson's 5^{th} order process with extended region of stability [35] and the stability region for both of Butcher's 2 stage, 3rd order, processes (4.2.4) and (4.2.5). Except for values of q whose imaginary part is more than three times larger in absolute value than its real part, Butcher's method would give the best stability bounds.

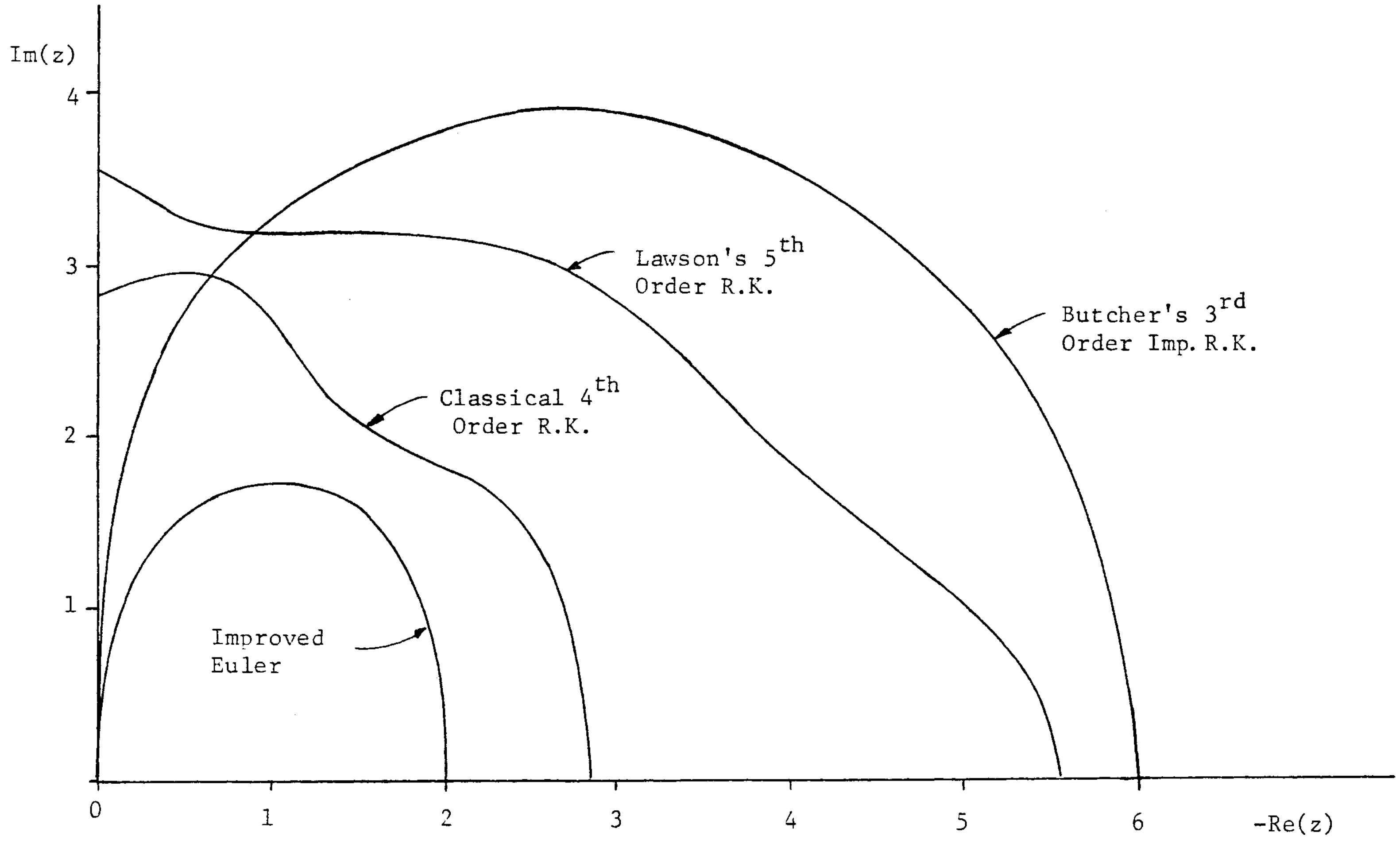

Figure 4.1

Upper Half of Region of Stability

4.5 A-stable Methods Based on Radau and Lobatto Quadrature

Since the methods based on Radau and Lobatto quadrature which were proposed by Butcher are not A-stable, it is natural to investigate whether methods based on these quadrature formulas can be found which are A-stable. The key to the search is to remove the assumptions that $\beta_{1,j} \equiv 0$ and/or $\beta_{j,\nu} \equiv 0$ for $j = 1, 2, \ldots, \nu$ which Butcher imposed to obtain explicit computations for K_1 and/or K_ν in his ν stage processes.

It is clear that when we solve (1.2.3) by a ν stage implicit Runge-Kutta process we obtain the relation

$$y_{n+1} = R_{\kappa,\eta}(qh)\; y_n \qquad (4.5.1)$$

where $R_{\kappa,\eta}(z)$ is the quotient of two polynomials with numerator of degree η and denominator of degree κ. Now the form of Δ when we solve for the K_i by Cramer's rule implies that $\kappa \leq \nu$. Thus if we demand that $\eta = \nu - 1$ when using a ν stage process of order $2\nu - 1$ or that $\eta = \nu - 2$ when using a ν stage process of order $2\nu - 2$, κ must equal ν in order to obtain the demanded accuracy and $R_{\kappa,\eta}(z)$ must equal either $P_{\nu,\nu-1}(z)$ or $P_{\nu,\nu-2}(z)$ depending on the order of the process. By the results of Chapter 3 we would then have A-stable processes.

The problem then is to see if by eliminating Butcher's assumptions about certain of the β's, can we find a ν stage process of order $2\nu - 1$ such that $\eta = \nu - 1$ and also a ν stage process of order $2\nu - 2$ for which $\eta = \nu - 2$ in (4.5.1). In order to determine if this is possible we will need to state the complete set of conditions which must be satisfied by the $\nu^2 - \nu$ parameters at our disposal in a ν stage implicit Runge-Kutta process.

In order to do this we first introduce several abbreviations to denote certain relationships which will be needed. These are patterned after the notation of Butcher [5]. The letters k and m are understood to be positive integers. For a ν stage Runge-Kutta process we then have letting

$$c_i = \sum_{j=1}^{\nu} \beta_{ij} :$$

A(ρ): The Runge-Kutta process is of order ρ,

$$B(\rho): \quad \sum_{i=1}^{\nu} b_i c_i^{k-1} = \frac{1}{k} \quad \text{for all } k \leq \rho ,$$

$$C(\rho): \quad \sum_{j=1}^{\nu} \beta_{ij} c_j^{k-1} = \frac{c_i^k}{k} \quad \text{for } i = 1, 2, \ldots, \nu \text{ and } k \leq \rho ,$$

$$D(\rho): \quad \sum_{i=1}^{\nu} b_i c_i^{k-1} \beta_{ij} = \frac{b_j(1-c_j^k)}{k} \quad \text{for } j = 1, 2, \ldots, \nu \text{ and } k \leq \rho ,$$

$$E(\rho,\xi): \quad \sum_{i=1}^{\nu} \sum_{j=1}^{\nu} b_i c_i^{k-1} \beta_{ij} c_j^{m-1} = \frac{1}{m(k+m)} \quad \text{for } k \leq \rho \text{ and } m \leq \xi .$$

In [5] Butcher then establishes the following theorems about any ν stage Runge-Kutta process.

Theorem 4.1

If A(ξ), then B(ξ).

Theorem 4.2

If $A(\xi + \eta)$, then $E(\xi,\eta)$.

Theorem 4.3

If $B(\xi + \eta)$ and $C(\eta)$, then $E(\xi,\eta)$.

Theorem 4.4

If $B(\xi + \eta)$ and $D(\xi)$, then $E(\xi,\eta)$.

Theorem 4.5

If $B(\nu + \eta)$ and $E(\nu,\eta)$, then $C(\eta)$.

Theorem 4.6

If $B(\xi + \nu)$ and $E(\xi,\nu)$, then $D(\xi)$.

Theorem 4.7

If $B(\rho)$, $C(\eta)$, and $D(\xi)$, where $\rho \le \xi + \eta + 1$, $\rho \le 2\eta + 2$, then $A(\rho)$.

We note that the following theorems are immediate consequences of Theorems 4.3 to 4.7.

Theorem 4.8

If $B(2\nu - 1)$ and $C(\nu)$, then $D(\nu - 1)$, $A(2\nu - 1)$.

Theorem 4.9

If $B(2\nu - 1)$ and $D(\nu)$, then $C(\nu - 1)$, $A(2\nu - 1)$.

Theorem 4.10

If $B(2\nu - 2)$ and $C(\nu)$, then $D(\nu - 2)$, $A(2\nu - 2)$.

Theorem 4.11

If $B(2\nu - 2)$ and $D(\nu)$, then $C(\nu - 2)$, $A(2\nu - 2)$.

Now it can be seen in [6] that Butcher's type I methods are designed to satisfy a theorem which is equivalent to Theorem 4.8

but not to satisfy Theorem 4.9. Similarly Butcher's type II methods were designed to satisfy a theorem equivalent to Theorem 4.9 but not Theorem 4.8. The obvious approach is then to switch theorems and attempt to find a type I process satisfying Theorem 4.9 and a type II process satisfying Theorem 4.8.

To show the feasibility of this approach we consider the case $\nu = 2$ for a type I process. We must then solve the set of equations specified by D(2) given that

$$c_1 = 0 \qquad c_2 = \frac{2}{3}$$
$$b_1 = \frac{1}{4} \qquad b_2 = \frac{3}{4} \tag{4.5.2}$$

imply B(3).

The equations D(2) are

$$\begin{aligned} b_1 c_1^0 \beta_{11} + b_2 c_2^0 \beta_{21} &= b_1(1 - c_1^1) \\ b_1 c_1^1 \beta_{11} + b_2 c_2^1 \beta_{21} &= \frac{b_1(1 - c_1^2)}{2} \end{aligned} \tag{4.5.3}$$

$$\begin{aligned} b_1 c_1^0 \beta_{12} + b_2 c_2^0 \beta_{22} &= b_2(1 - c_2^1) \\ b_1 c_1^1 \beta_{12} + b_2 c_2^1 \beta_{22} &= \frac{b_2(1 - c_2^2)}{2} \end{aligned} \tag{4.5.4}$$

and for the conditions given in (4.5.2) reduce to

$$\frac{1}{4}\beta_{11} + \frac{3}{4}\beta_{21} = \frac{1}{4}$$

$$\frac{1}{2}\beta_{21} = \frac{1}{8},$$

$$\frac{1}{4}\beta_{12} + \frac{3}{4}\beta_{22} = \frac{1}{4}$$

$$\frac{1}{2}\beta_{22} = \frac{5}{24}.$$

These equations are easily solved and we obtain

$$\begin{aligned} \beta_{11} &= \frac{1}{4} & \beta_{12} &= \frac{-1}{4} \\ \beta_{21} &= \frac{1}{4} & \beta_{22} &= \frac{5}{12} \end{aligned} \qquad (4.5.5)$$

Comparing these values of β_{ij} with those given by Butcher (equations (4.2.4)) for his type I process we see that we have a different method. Furthermore, if we solve (1.2.3) using the β_{ij} in (4.5.5) and the b_i in (4.5.2) we find that

$$\begin{aligned} y_{n+1} &= \left[\frac{1 + \frac{qh}{3}}{1 - \frac{2}{3}qh + \frac{(qh)^2}{6}}\right] y_n \\ &= P_{2,1}(qh)\ y_n \end{aligned}$$

and we have that (4.5.5) is the basis for a 2 stage A-stable implicit Runge-Kutta process of order 3.

Noting that for the Radau quadrature none of the b_i's are zero and all the c_i's are distinct, it follows that for any ν there is a unique set of β_{ij}'s determined by $D(\nu)$. This can be seen by observing that the determinant of the coefficients for each set of equations (grouped by values of j, as are those in (4.5.3) and (4.5.4)) can be written as a product of the b_i's times a Vandermonde determinant [52, p. 74] whose value is not zero. Thus Cramer's rule can be applied to solve for the β_{ij}. We shall denote ν stage implicit Runge-Kutta processes which are based on Radau quadrature with left end fixed and which satisfy $D(\nu)$ as type I_A processes.

Table 4.1 gives the values of c_i, β_{ij}, and b_i for $\nu = 2, 3$ for type I_A processes. Values for $\nu \geq 4$ are not given since exact values of c_i are not known. It is also noted that there is no one stage type I_A process since C(1) must be satisfied for all Runge-Kutta processes and this cannot be done with a one stage type I_A process.

By direct computation the case $\nu = 3$ can also be shown to reduce to a below diagonal Padé approximation to the exponential when solving (1.2.3.). Thus we have the following conjecture.

Conjecture 4.1

Each type I_A implicit Runge-Kutta process is A-stable.

Proof

Although this theorem has been proved for $\nu = 2, 3$ no proof of its general truth has as yet been found.

$$\nu = 2 \qquad \begin{array}{cc|c} \frac{1}{4} & \frac{-1}{4} & 0 \\ \frac{1}{4} & \frac{5}{12} & \frac{2}{3} \\ \hline \frac{1}{4} & \frac{3}{4} & \end{array}$$

$$\nu = 3 \qquad \begin{array}{ccc|c} \frac{1}{9} & \frac{-1-\sqrt{6}}{18} & \frac{-1+\sqrt{6}}{18} & 0 \\ \frac{1}{9} & \frac{88+7\sqrt{6}}{360} & \frac{88-43\sqrt{6}}{360} & \frac{6-\sqrt{6}}{10} \\ \frac{1}{9} & \frac{88+43\sqrt{6}}{360} & \frac{88-7\sqrt{6}}{360} & \frac{6+\sqrt{6}}{10} \\ \hline \frac{1}{9} & \frac{16+\sqrt{6}}{36} & \frac{16-\sqrt{6}}{36} & \end{array}$$

Table 4.1

Type I_A Processes

Repeating the same arguments with the Radau quadrature with right end fixed but requiring the C(ν) be satisfied instead of D(ν) we find the type II_A processes listed in Table 4.2. By direct computation each of these processes can also be shown to reduce to a below diagonal Padé approximation to the exponential and hence each is an A-stable process. The following conjecture is suggested.

$$\nu = 1 \qquad \begin{array}{c|c} 1 & 1 \\ \hline 1 & \end{array}$$

$$\nu = 2 \qquad \begin{array}{cc|c} \frac{5}{12} & \frac{-1}{12} & \frac{1}{3} \\[2ex] \frac{3}{4} & \frac{1}{4} & 1 \\[1ex] \hline \frac{3}{4} & \frac{1}{4} & \end{array}$$

$$\nu = 3 \qquad \begin{array}{ccc|c} \frac{88-7\sqrt{6}}{360} & \frac{296-169\sqrt{6}}{1800} & \frac{-2+3\sqrt{6}}{225} & \frac{4-\sqrt{6}}{10} \\[2ex] \frac{296+169\sqrt{6}}{1800} & \frac{88+7\sqrt{6}}{360} & \frac{-2-3\sqrt{6}}{225} & \frac{4+\sqrt{6}}{10} \\[2ex] \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} & 1 \\[1ex] \hline \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} & \end{array}$$

Table 4.2

Type II_A Processes

Conjecture 4.2

Each type II_A process is A-stable.

Proof

Although this theorem has been verified for $\nu = 1, 2, 3$ no proof of its general truth has as yet been found.

Turning our attention to Butcher's type III methods we observe that the β_{ij} which he finds satisfy neither Theorem 4.10 or 4.11 but do satisfy the condition that given $B(2\nu - 2)$, $C(\nu - 1)$ if and only if $D(\nu - 1)$ and either implies $A(2\nu - 2)$ [6]. Thus Theorem 4.10 and 4.11 give us the basis of finding other methods based on Lobatto quadrature.

We define a ν stage Runge-Kutta process which is based on Lobatto quadrature and which satisfies $C(\nu)$ as a type III_A process and one which satisfies $D(\nu)$ as a type III_B process. Table 4.3 shows type III_A processes for $\nu = 2, 3, 4$ while Table 4.4 shows type III_B processes for $\nu = 3, 4$. There is no type III_B process for $\nu = 2$.

An inspection of the β_{ij} for each type III_A and III_B process shows that there is either a row or a column of zeros and this will be true in general. It follows that a ν stage type III_A or type III_B process must reduce to an approximation of the exponential with denominator of power at most $\nu - 1$ and thus cannot produce a below diagonal Padé approximation to the exponential because of its required accuracy.

Direct computation shows that type III_A and type III_B processes do in fact reduct to diagonal Padé approximations $(P_{\nu-1,\nu-1}(qh))$ for $\nu \leq 4$. Thus we have the following conjecture.

$\nu = 2$

$$\begin{array}{cc|c}
0 & 0 & 0 \\
\frac{1}{2} & \frac{1}{2} & 1 \\
\hline
\frac{1}{2} & \frac{1}{2} &
\end{array}$$

$\nu = 3$

$$\begin{array}{ccc|c}
0 & 0 & 0 & 0 \\
\frac{5}{24} & \frac{1}{3} & \frac{-1}{24} & \frac{1}{2} \\
\frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 1 \\
\hline
\frac{1}{6} & \frac{2}{3} & \frac{1}{6} &
\end{array}$$

$\nu = 4$

$$\begin{array}{cccc|c}
0 & 0 & 0 & 0 & 0 \\
\frac{11+\sqrt{5}}{120} & \frac{25-\sqrt{5}}{120} & \frac{25-13\sqrt{5}}{120} & \frac{-1+\sqrt{5}}{120} & \frac{5-\sqrt{5}}{10} \\
\frac{11-\sqrt{5}}{120} & \frac{25+13\sqrt{5}}{120} & \frac{25+\sqrt{5}}{120} & \frac{-1-\sqrt{5}}{120} & \frac{5+\sqrt{5}}{10} \\
\frac{1}{12} & \frac{5}{12} & \frac{5}{12} & \frac{1}{12} & 1 \\
\hline
\frac{1}{12} & \frac{5}{12} & \frac{5}{12} & \frac{1}{12} &
\end{array}$$

Table 4.3

Type III_A Processes

$\nu = 3$	$\frac{1}{6}$	$\frac{-1}{6}$	0	0
	$\frac{1}{6}$	$\frac{1}{3}$	0	$\frac{1}{2}$
	$\frac{1}{6}$	$\frac{5}{6}$	0	1
	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$	

$\nu = 4$	$\frac{1}{12}$	$\frac{-1-\sqrt{5}}{24}$	$\frac{-1+\sqrt{5}}{24}$	0	0
	$\frac{1}{12}$	$\frac{25+\sqrt{5}}{120}$	$\frac{25-13\sqrt{5}}{120}$	0	$\frac{5-\sqrt{5}}{10}$
	$\frac{1}{12}$	$\frac{25+13\sqrt{5}}{120}$	$\frac{25-\sqrt{5}}{120}$	0	$\frac{5+\sqrt{5}}{10}$
	$\frac{1}{12}$	$\frac{11-\sqrt{5}}{24}$	$\frac{11+\sqrt{5}}{24}$	1	1
	$\frac{1}{12}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{1}{12}$	

Table 4.4

Type III_B Processes

Conjecture 4.3

Each type III_A and III_B process is A-stable.

Proof

Although this theorem has been verified for $\nu \leq 4$, no proof of its general truth has as yet been found.

It should be expected that since there are type III processes which reduce to above diagonal Padé approximations to the exponential and type III processes which reduce to diagonal Padé approximations that there should also be type III processes which reduce to below diagonal Padé approximations to the exponential. In Table 4.5 we give several examples of type III_C processes which satisfy the conditions that they are ν stage, order $2\nu - 2$, and reduce to $P_{\nu,\nu-2}(qh)$ when solving (1.2.3).

The key to finding type III_C processes is noting that Theorem 4.7 allows other conditions besides those given in Theorems 4.10 and 4.11 to produce a ν stage process of order $2\nu - 2$. The examples given in Table 4.5 satisfy both $D(\nu - 1)$ and $C(\nu - 1)$ and are therefore similar to Butcher's type III methods in many respects. No effort has been made to find a general set of conditions which will produce a type III_C process.

This completes our discussion of A-stable methods based on implicit Runge-Kutta methods.

$$\nu = 2 \qquad \begin{array}{cc|c} \frac{1}{2} & \frac{-1}{2} & 0 \\[1ex] \frac{1}{2} & \frac{1}{2} & 1 \\[1ex] \hline \frac{1}{2} & \frac{1}{2} & \end{array}$$

$$\nu = 3 \qquad \begin{array}{ccc|c} \frac{1}{6} & \frac{-1}{3} & \frac{1}{6} & 0 \\[1ex] \frac{1}{6} & \frac{5}{12} & \frac{-1}{12} & \frac{1}{2} \\[1ex] \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 1 \\[1ex] \hline \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & \end{array}$$

Table 4.5

Type III_C Processes

4.6 Additional A-stable Methods

If one is willing to allow the introduction of second and higher derivatives into the equations being used then a generalized class of linear one step methods first given by Hermite [27], usually credited to Obrechkoff [49] (see Hildebrand [28, p. 231], Henrici [26, p. 106], Milne [48, p. 79]) and which are also derived by Lanczos [34, p. 419] of the form

$$y_{n+1} = y_n + \sum_{i=1}^{\nu} \alpha_{i\nu} h^i \left(y_n^{(i)} - (-1)^i \, y_{n+1}^{(i)}\right) \qquad (4.6.1)$$

$$\nu = 1, 2, 3, \ldots$$

also exists which are of order 2ν when $\alpha_{i\nu}$ is the i^{th} coefficient in the numerator of the ν^{th} diagonal Padé approximation $P_{\nu\nu}(z)$ of e^z. Thus for (1.2.3), (4.6.1) reduces to $y_{n+1} = P_{\nu\nu}(qh)y_n$ and (4.6.1) is A-stable for all ν [17], [20]. We note that for $\nu = 1$ in (4.6.1) we have simply Dahlquist's result and for $\nu = 2$ we have method (2.5.1) which we already observed was A-stable. Milne [48, p. 78] has also considered (4.6.1) for the case $\nu = 3$ but no proof of stability appears to have been previously made for the cases $\nu \geq 3$. We note that Loscalzo [40, p. 79] says that he has proved that (4.6.1) is A-stable for $\nu \geq 3$ but the arguments leading to this statement are not complete. It is clear that he has considered only the case q real in (1.2.3) when we look at his remark on page 98 of [40].

Now (4.6.1) is a special case of the following more general result of Hermite [27], Obrechkoff [49], and Hummel and

Seebeck [30].

Theorem 4.12

Assume $y(t)$ has continuous derivatives for $a \leq t \leq x$ up to and including the derivative of order $(j+k+1)$. Let the binomial coefficients $p!/(q!\ (p-q)!)$ be denoted by ${}_pC_q$ with the understanding that ${}_pC_q = 0$ if q is greater than p. Then

$$y(x) = y(a) + \sum_{m=1} \frac{(k+j-m)!}{(k+j)!} \left[{}_kC_m y^{(m)}(a) - (-1)^m {}_jC_m y^{(m)}(x) \right] (x-a)^m + R \qquad (4.6.2)$$

where

$$R = (-1)^n \frac{k!\ j!\ (x-a)^{j+k+1}}{(j+k)!\ (j+k+1)!} f^{(j+k+1)}(\theta)$$

θ between a and x.

For the case $j = k$, $a = x_n$, $x = x_{n+1} = x_n + h$ (4.6.2) reduces to (4.6.1). For the case $k = j - 1$ we see that the coefficients are just those of the subdiagonal Padé approximation to the exponential, $P_{j,j-1}(x)$. For the case $k = j - 2$ it is also seen that the coefficients in (4.6.2) are just those of the Padé approximation to the exponential $P_{j,j-2}(x)$. Since we have already established that both of these classes of approximation are L-acceptable we have the following result.

Theorem 4.13

The quadrature formulas based on derivatives given by (4.6.2) for $k = j$, $j - 1$, and $j - 2$ and any $j \geq 0$, $j \geq 1$, $j \geq 2$ respectively, are all A-stable.

Loscalzo [40], [41], [42], [43], and [44] has recently done considerable work in the implementation of methods based on (4.6.1) which include the generation of spline functions $s_n(x)$ which give polynomial interpolations to the solution of the differential equation between the points (x_n, y_n) and (x_{n+1}, y_{n+1}) produced by (4.6.1). The polynomials are such that $s_{n-1}(x_n) = s_n(x_n)$ and $s_{n-1}^{(k)}(x_n) = s_n^{(k)}(x_n)$ for $k = 1, 2, \ldots, \nu$. Several examples of solving stiff equations are given in [40].

Very little work seems to have been done with the methods corresponding to $k = j - 1$ and $k = j - 2$. No previous discussion could be found in the literature about the stability of methods based on $k = j - 1$ except for the case $j = 1$ as noted previously in Chapter 1, equation (1.1.8). No literature at all could be found for methods where $k = j - 2$.

Recalling Makinson's results given in Chapter 2, we see that $k = j - 1$, $j = 2$ in (4.6.2) corresponds to $\alpha_1 = 2/3$ in (2.5.4). Comparing truncation errors, given by (2.5.5), we see that $\alpha_1 = 2/3$ gives a truncation error of

$$\frac{1}{72} h^4 y^{(4)}(\theta)$$

while $\alpha_1 = 1 + \ 3/3$ (Makinson's value) gives a truncation error of

$$\frac{3 + 2\sqrt{3}}{72} h^4 y^{(4)}(\theta).$$

This would suggest that the third order method based on (4.6.2) would give better results. In addition, as $h \to \infty$ we observe that the approximation to the exponential given by the method based on (4.6.2) approaches zero which Makinson's does not.

Corresponding to Makinson's 4^{th} order method (2.5.6) we would have $k = j - 2$, $j = 3$ in (4.6.2). It is easily seen that this implies $\alpha_1 = 3/4$ and $\alpha_2 = -1/2$. Equation (2.5.7) then gives that the truncation error would be

$$-0.25 \frac{h^5}{5!} y^{(5)}(\theta).$$

Using Makinson's values of α_1 and α_2 we obtain that the truncation error would be

$$19.73 \frac{h^5}{5!} y^{(5)}(\theta).$$

These values would suggest that (4.6.2) might be considerably better than the approximation proposed by Makinson. In addition the approximation based on (4.6.2) again goes to zero for large h while the approximation proposed by Makinson does not.

Turning our attention to Rosenbrock's method we observe that if we set

$$\begin{aligned} a_1 &= \frac{2 - \sqrt{2}\, i}{6} \\ a_2 &= \frac{2 + \sqrt{2}\, i}{6} \\ b_{21} &= \frac{1}{3} \\ c_{21} &= \frac{5}{18}(2 + \sqrt{2}\, i) \\ \omega_1 &= \omega_2 = \frac{1}{2} \end{aligned} \qquad (4.6.3)$$

we will produce a method with error of $O(h^4)$. Rosenbrock also gives an example using complex coefficients in [54] but it is only of order 2. Solving (1.2.3) with (2.5.9) and (4.6.3) we obtain

$$y_{n+1} = P_{2,1}(qh)\ y_n$$

and hence (2.5.9) with (4.6.3) produces an A-stable process where the approximation of the exponential approaches zero as h → ∞. This disproves a statement made by Rosenbrock [54] that it is not possible to have a third order process based on (2.5.9) with these properties.

Because of the complex coefficients, the method based on (4.6.3) is probably not as attractive as the method proposed by Calahan, (2.5.11), which was considered earlier.

Finally we note that both the method proposed by Lawson, (2.5.13), and those proposed by Pope, Kuo, Legras, and Calahan based on (2.5.16) can now be implemented in a wide variety of new ways using one of the L-acceptable approximations to the exponential given in Chapter 3 of this thesis.

CHAPTER 5

Numerical Examples Employing High Order A-stable Methods

5.1 Introduction

In this Chapter we consider the problem of implementing the implicit Runge-Kutta methods which were shown to be A-stable in Chapter 4. While not claiming that what follows constitutes a complete study of such processes, it is shown that through the use of proper starting values the method based on (4.2.3) can be as efficient as the classical 4th order Runge-Kutta process, (1.1.6), for small step sizes and retains this efficiencv for step sizes where the classical 4th order Runge-Kutta process would be unstable. Also inherent in the process is a prediction of the truncation error which has occurred at each step. Examples involving the solution of linear and nonlinear systems of equations are given.

5.2 Problems in Implementing Implicit R. K. Processes

Because of their implicit nature, several problems face the prospective user of an implicit Runge-Kutta process. First he must solve for the $\vec{K}_i$'s. If a constant coefficient system of equations is involved, such as (1.2.1), this can be done directly by solving the resulting system of linear equations for the $\vec{K}_i$'s. In most cases of interest, however, direct solution for the $\vec{K}_i$'s would probably not be possible. Finding the $\vec{K}_i$'s by iteration would seem to be the only practical method in most cases.

If we define δ as the greatest of

$$|\beta_{21}|, |\beta_{31}|+|\beta_{32}|, \ldots, |\beta_{\nu 1}|+|\beta_{\nu 2}| + \ldots + |\beta_{\nu,\nu-1}| ,$$

μ as the greatest of

$$|\beta_{11}|+|\beta_{12}|+ \ldots + |\beta_{1\nu}|, |\beta_{22}|+|\beta_{23}| + \ldots + |\beta_{2\nu}|, \ldots, |\beta_{\nu\nu}|$$

and σ as $\delta + \mu$ and for a vector $\vec{v} = (v_1, v_2, \ldots, v_n)$ we denote by $||v||$ the greatest of $|v_1|, |v_2|, \ldots, |v_n|$ we have the following result of Butcher's [5] concerning the convergence of the iterative technique

$$\vec{K}_i^{(N)} = \vec{f}(\vec{y}_n + h\left(\sum_{j=1}^{i-1} \beta_{ij} \vec{K}_j^{(N)} + \sum_{j=1}^{\nu} \beta_{ij} \vec{K}_j^{(N-1)} \right)) \qquad (5.2.1)$$

where $\vec{K}_i^{(N)}$ is the N^{th} iterate of $\vec{K}_i$.

Theorem 5.1

If $\vec{f}(\vec{y})$ satisfies a Lipschitz condition

$$||\vec{f}(\vec{y}) - \vec{f}(\vec{z})|| \leq L||\vec{y} - \vec{z}|| \qquad (5.2.2)$$

and $|h| < 1/(L\sigma)$, then the equations defining $\vec{K}_1, \vec{K}_2, \ldots, \vec{K}_\nu$ have a unique solution and $\vec{K}_1^{(N)}, \vec{K}_2^{(N)}, \ldots, \vec{K}_\nu^{(N)}$ defined by (5.2.1) tend to this solution as N tends to infinity.

Although the above theorem does provide sufficient conditions for convergence, it is not of much help to us. For example, in solving (1.2.1), (5.2.2) becomes

$$||By - Bz|| \leq ||B|| \cdot ||y - z|| .$$

Thus the Lipschitz constant would be just $||B||$. But the norm of B is at least as large as the spectral radius of B and by assumption the spectral radius of B is large. Thus Butcher's result would say that to guarantee convergence using (5.2.1) we would require that h be very small which is just what we are trying to avoid.

Recognizing that Butcher's theorem may imply severe limitations on the step size, we will attempt to develop an iterative method for finding the $\vec{K}_i$'s. This, of course, leads to another problem in that if very many iterations are required it would be less costly (in the sense of the number of function evaluations required) to simply solve the system of differential equations using an explicit Runge-Kutta process with a smaller step size. This problem can be largely overcome if a good initial guess for the $\vec{K}_i$'s can be obtained. As with predictor-corrector methods we could also reduce the number of function evaluations required if we were to stop the iteration when the iteration error was less than the truncation error of the method. Since we are working with Runge-Kutta methods, it would appear at first glance that the truncation error of the method is not easily obtainable. In the next three sections we will show how these difficulties can be handled.

To be specific in the discussion which follows, we will assume we are attempting to solve $\vec{y}' = \vec{f}(\vec{y})$ using the 4th order implicit Runge-Kutta process given by (4.2.3). The generalization to other implicit Runge-Kutta methods should be obvious.

5.3 Predicting the $\vec{K}_i^0$'s

An inspection of the β's suggests that good initial guesses of $\vec{K}_1$ and $\vec{K}_2$ would be given by computing the values of $\vec{y}$ corresponding to $x_n + \theta_1 h$ and $x_n + \theta_2 h$, where $\theta_1 = (\frac{1}{2} - \frac{\sqrt{3}}{6})$, $\theta_2 = (\frac{1}{2} + \frac{\sqrt{3}}{6})$ and $x_n = x_0 + nh$, and evaluating $\vec{f}$ at these points. If the resulting values of $\vec{K}_1^0$ and $\vec{K}_2^0$ are good enough, only one iteration will be required to establish their accuracy. A procedure which can be used to produce approximations to the required $\vec{y}$ values will now be given. The procedure depends upon whether we are taking our first step with a given step of h or our k^{th} step ($k \geq 2$).

(a) Step 1

Approximations to the values of $\vec{y}$ at $x_0 + \theta_1 h$ and $x_0 + \theta_2 h$ are obtained by employing a second order explicit Runge-Kutta process twice. Once to go from x_0 to $x_0 + \theta_1 h$ and again to go from $x_0 + \theta_1 h$ to $x_0 + \theta_2 h$. The resulting values of $\vec{y}$ are substituted into $\vec{f}$ to produce $\vec{K}_1^0$ and $\vec{K}_2^0$ for the first step.

(b) Step k ($k \geq 2$)

When an acceptable value of $\vec{y}_n$ ($n = k - 1$) has been found at $x = x_n$ we compute $\vec{f}(\vec{y}_n)$. We note that this is an "extra" derivative evaluation not required directly by the Runge-Kutta process but the information gained will be very useful in later steps.

The values of $\vec{y}$ at $x_n + \theta_1 h$ and $x_n + \theta_2 h$ are now estimated using the quadrature formula

$$\vec{y}(x_n + \theta h) = \vec{y}(x_{n-1}) + h \sum_{i=1}^{4} \omega_i \vec{y}\,' (x_{n-1} + \gamma_i h) \qquad (5.3.1)$$

where $\gamma_1 = 0$, $\gamma_2 = \theta_1$, $\gamma_3 = \theta_2$, and $\gamma_4 = 1$.

If the weights given in Table 5.1 are used, (5.3.1) will be a 4th order quadrature formula for $\theta = \theta_1$ and $\theta = \theta_2$.

$\omega_i \backslash \theta$	θ_1	θ_2
ω_1	-.04419540711644071	-1.872471259550258
ω_2	.5938203090566040	4.246669680947478
ω_3	.2533303190525340	-6.093820309056581
ω_4	.4083696444124900	5.508297022254174

Table 5.1

Observing that the four derivative values used in (5.3.1) correspond to $\vec{f}(\vec{y}_{n-1})$, $\vec{K}_1$, $\vec{K}_2$, and $\vec{f}(\vec{y}_n)$, all of which are known from the previous step, we should be able to produce reasonably good approximations to $\vec{y}$ at $x_n + \theta_1 h$ and $x_n + \theta_2 h$ using (5.3.1) with these values in place of $\vec{y}\,'(x_{n-1} + \gamma_i h)$.

5.4 The Iteration Procedure

Having determined starting values for the $\vec{K}_i$'s we compute an initial estimate for $\vec{y}$ at x_{n+1}, calling it $\vec{y}^{\,0}_{n+1}$, using (4.2.2). One iteration is done on the $\vec{K}_i$'s and a second value $\vec{y}^{\,1}_{n+1}$ is produced. If the maximum relative error

$$\max_i \left| \frac{y^1_{n+1,i} - y^0_{n+1,i}}{y^1_{n+1,i}} \right|$$

where $y_{n+1,i}$ is the ith component of $\vec{y}_{n+1}$, is less than a previously specified allowable relative iteration error, then we accept $\vec{y}_{n+1}^{1}$ as being satisfactory. If the condition is not satisfied we continue to iterate until two successive values $\vec{y}_{n+1}^{i}$ and $\vec{y}_{n+1}^{i+1}$ have a relative error less than the allowable error.

The procedure used to iterate for the $\vec{K}_i$'s will depend on the form of the system of equations being solved. Generally an iteration procedure something like that of (5.2.1) would need to be used. For the examples given in Section 5.6 of this thesis we have that the j^{th} function, f_j, of $\vec{f}$ is linear in the j^{th} variable, y_j. Hence, defining

$$K_{ij} = f_j(\vec{y}_n + h \sum_{\ell=1}^{\nu} \beta_{i\ell} K_\ell),$$

that is, K_{ij} is the j^{th} component in $\vec{K}_i$, we can solve for K_{ij} explicitly in terms of all the other K_{mn} and obtain

$$K_{ij} = \frac{f_j(\vec{y}_n + h \sum_{\ell=1}^{\nu} \beta_{i\ell} \vec{K}_\ell) - A_j \beta_{ij} hK_{ij}}{1 - A_j \beta_{ij} h}$$

where A_j is the coefficient of y_j in f_j. Letting $K_{ij}^{(N)}$ be the N^{th} iterate of K_{ij} and defining

$${}_i f_j(\vec{y}_n + h \sum_{\ell=1}^{\nu} \beta_{i\ell} K_\ell^{(N-1)})$$

to be

$$f_j(\vec{y}_n + h \sum_{\ell=1}^{\nu} \beta_{i\ell} \vec{\emptyset}_\ell^{ij}) - A_j \beta_{ij} h K_{ij}^{(N)}$$

where

$$\vec{\phi}_{\ell}^{ij} = \begin{cases} \vec{K}_{\ell}^{(N)} & \text{if } \ell < i \text{ and all } j, \\ (K_{i1}^{(N)}, K_{i2}^{(N)}, \ldots, K_{ij}^{(N)}, K_{i,j+1}^{(N-1)}, K_{i,j+2}^{(N-1)}, \ldots) & \\ \quad \text{when } i = \ell & \\ \vec{K}_{\ell}^{(N+1)} & \text{if } \ell > i \text{ and all } j \end{cases}$$

the generalized Gauss-Seidel iteration which is used to solve these problems is

$$K_{ij}^{(N)} = \frac{{}_{i}f_{j}(\vec{y}_n + h \sum_{\ell=1}^{\nu} \beta_{i\ell} \vec{K}_{\ell}^{(N-1)})}{1 - A_j \beta_{ij} h}, \quad j = 1, 2, 3, \ldots \quad (5.4.1)$$

with $i = 1, 2, \ldots, \nu$.

Clearly the above procedure of specifying the allowable relative iteration error will not be satisfactory unless we have some assurance that the actual relative truncation error of the process is of about the same size. If the truncation error could be estimated, we could from time to time change the step size so that the truncation error and the iteration error were of about the same magnitude.

5.5 Estimating the Truncation Error

As noted in [18], it is possible to combine multistep processes with Runge-Kutta processes. In [18] it is shown that, provided suitable conditions are satisfied, it is possible to obtain asymtotic error estimates for the truncation error of the Runge-Kutta process. In particular, in [18], it is shown that when a 4th order Runge-Kutta process is combined with

$$y_3 = \{y(0) + 18y(h) - 9y(2h)\}/10 + 3h\{3y'(h) + 6y'(2h) + y'(3h)\}/10 \tag{5.5.1}$$

the result is a 5th order process after three Runge-Kutta steps. The difference between the combined 5th order process and the Runge-Kutta result is an asymtotic estimate of the truncation error of the Runge-Kutta process.

The above idea is used to produce an estimate of the truncation error after three Runge-Kutta setps. If the estimated relative truncation error E_t satisfies the conditions that

$$E_i < |E_t| < 50\ E_i$$

where E_i is the allowed relative iteration error then three more steps are taken with the same step size and we repeat the checking process. If at some point it becomes necessary to change the step size, we must repeat step 1 in Section 5.3. If a change of step does not occur very often, then the added function evaluations required by this first step will not be significant. We note that in the evaluation of (5.5.1) we also use the information obtained from the "extra" derivative evaluation which is done at each step.

One small problem remains. What do we do over the first three steps when no estimate of the truncation error is available? It is recommended that an initial step size be specified which is somewhat smaller than might be expected to be used for the entire solution. This can be justified by the fact that the one or more relatively large eigenvalues involved in the system of equations being solved will probably cause one or more of the variables to change rapidly near the start of the interval of solution and accuracy will demand this smaller step size be used no matter what method of solution was being employed.

5.6 Numerical Examples

To illustrate that the above procedure can be used effectively, the following four problems are now considered.

Problem 1

$$y_1' = -y_1 + 95y_2, \qquad y_1(0) = 1,$$
$$y_2' = -y_1 - 97y_1, \qquad y_2(0) = 1.$$

Problem 2

$$y_1' = -y_1 + 95y_2, \qquad y_1(0) = 1,$$
$$y_2' = -y_1 - 97y_2, \qquad y_2(0) = -1/95.$$

Problem 3

$$y_1' = (-1 + y_2^2)y_1 + (1 + y_2)y_2, \qquad y_1(0) = -1,$$
$$y_2' = -y_1 + (-19 + 2y_1 + y_1^2)y_2, \qquad y_2(0) = 1.$$

Problem 4

$$y_1' = (-20 + 17y_2)y_1 + (76 - 36y_2 + 4y_2^2)y_2, \qquad y_1(0) = 3,$$
$$y_2' = (10 - y_1^2)y_1 + (-41 + 3y_1 + y_1^2)y_2, \qquad y_2(0) = 1.$$

Treating each of these problems as if they were of the form $\vec{y}' = B\vec{y}$, where B is a real matrix, we have that Problems 1 and 2 both have eigenvalues of -2 and -96 and solutions of

$$y_1 = (95e^{-2x} - 48e^{-96x})/47$$
$$y_2 = (48e^{-96x} - e^{-2x})/47$$

and

$$y_1 = e^{-2x}$$
$$y_2 = e^{-2x}/95$$

respectively.

Treating Problem 3 as if it were also a linear problem we see that it starts with eigenvalues of approximately -.1 and -19.9 and approaches the linear problem with eigenvalues of -1 and -19 as y_1 and $y_2 \to 0$. Treating problem 4 in the same way, we see that it begins with eigenvalues of -1 and -25 and as y_1 and $y_2 \to 0$ it behaves like a linear problem with eigenvalues of -1 and -60. Unfortunately this analysis for Problem 4 is not correct since it happens that for the given initial conditions $y_1 \to 1.65070477312$ (approx.) and $y_2 \to .360385998230$ (approx.) as $x \to \infty$ and Problem 4 actually approaches the linear problem with eigenvalues of 0 and -47. There is, of course, no way of knowing this until the solution is computed.

Since neither Problem 3 or 4 has a known analytic solution, "exact" solutions for these two problems were found by using the classical 4th order Runge-Kutta process with a step size of $1/2^{12}$.

In Figure 5.1 we plot the dominant (larger) relative error, which occurs at x = 10 in Problem 1, versus the number of function evaluations per unit step in x which produced that error. This is done for both the classical 4th order Runge-Kutta process and for the implicit method given by (4.2.3). Because of the eigenvalue -96, the largest step size permitted for the explicit Runge-Kutta process is $1/2^6$ which means a minimum of 2^8 function evaluations per unit step in x are required to obtain a solution. As can be seen from Figure 5.1, the implicit Runge-Kutta process is not quite as efficient as the explicit process for 2^8 and 2^9 function evaluations per unit step in x but produces solutions with about 2^5 function evaluations per unit step in x which the explicit process cannot do.

Figures 5.2, 5.3, and 5.4 give similar comparisons for the explicit and implicit Runge-Kutta solutions of Problems 2, 3 and 4. As can be seen, the implicit Runge-Kutta process compares favourably in those regions where the explicit Runge-Kutta process can be applied and extends the solution region considerably. In all four problems we note that the slope of the implicit Runge-Kutta process appears to be more like that of a 5th order Runge-Kutta process than that of the classical 4th order process.

5.7 Asymtotic Error Estimates

Since it is necessary to compute an approximation to the truncation error in order to implement the above process effectively, it seems natural to use this information in as many ways as possible since it is available. In particular, if we note that the approximation to the truncation error is also an approximation to the principal error function it is natural to compute an asymtotic error estimate by an Euler integration of the principal error function as suggested by Henrici [26, p. 136].

This requires a knowledge of the Jacobian matrix $J = (\frac{\partial f}{\partial y})$. Because of the stability problems involved, we are forced to use the first order implicit equivalent of Euler's formula, namely

$$(I - H(\frac{\partial f}{\partial y}))\ \vec{e}^{\,*}_{3n+3} = \vec{e}_{3n} \qquad (5.7.1)$$

to determine the propagation of the error e_{3n} at x_{3n} to the point x_{3n+3}. The error at x_{3n+3} is then given by

$$\vec{e}_{3n+3} = \vec{e}^{\,*}_{3n+3} + \text{(the estimated truncation error)} \qquad (5.7.2)$$

In (5.7.1), the Jacobian is evaluated at x_{3n+3} and $H = 3h$ where h is the Runge-Kutta step.

Figures 5.5, 5.6, 5.7, and 5.8 show the actual error and the asymtotic error based on (5.7.1) and (5.7.2) for the four problems considered in this paper. As can be seen, the asymtotic errors are generally in good agreement with the actual error. The irregularities in the error curves are caused by the automatic changing of the step size as the solution is being computed so that the truncation and iteration errors are of about the same magnitude.

5.8 Stability of the Combined Process

As noted above, the combination of the implicit Runge-Kutta process being considered with equation (5.5.1) produces a 5th order process. As an alternative to the approach presented above, the solution produced by the combined Runge-Kutta multistep process might be used as the solution. In that event, the stability

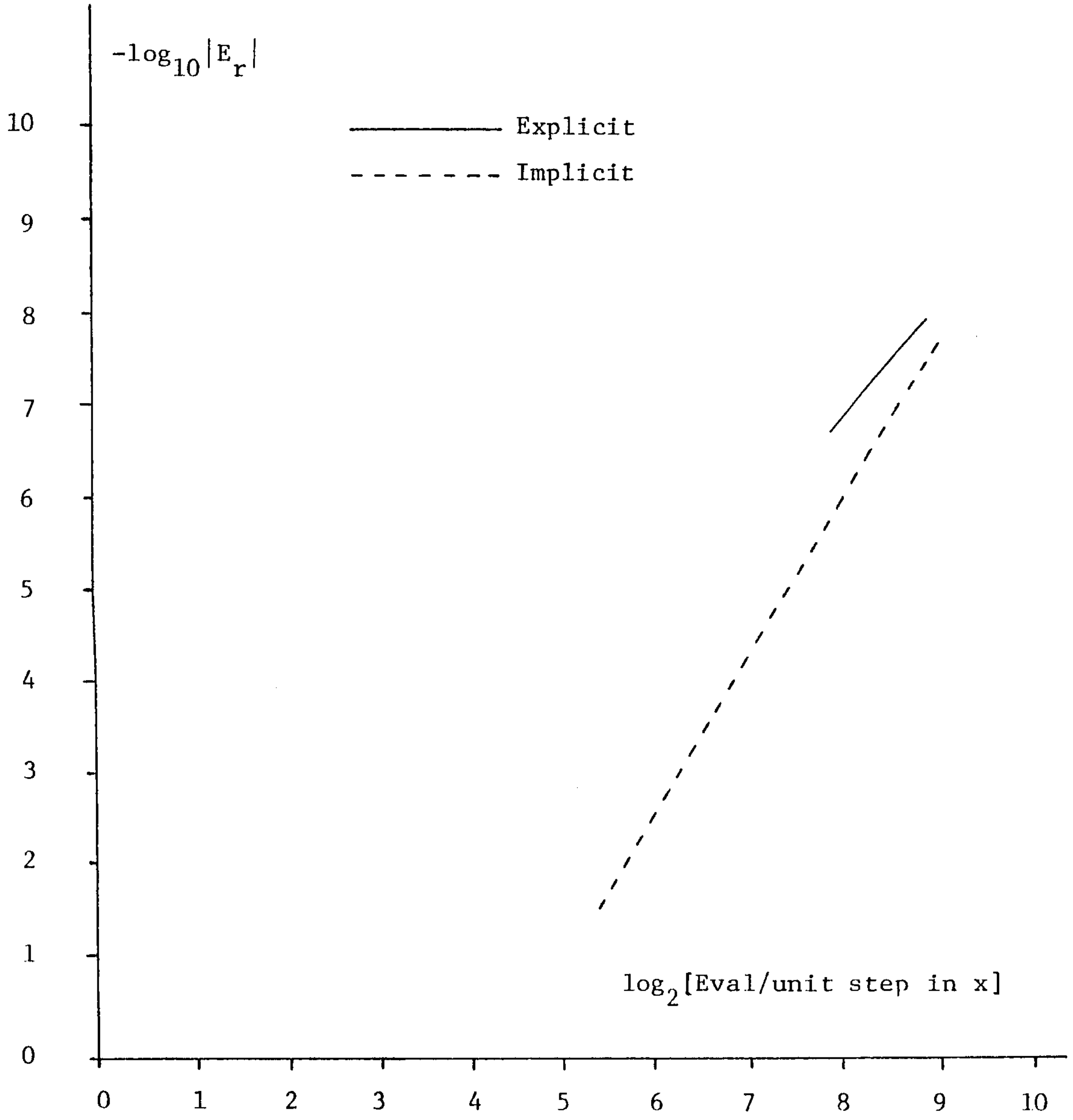

Figure 5.1

A graph of the efficiency of two methods of solving Problem 1 over the interval $0 \le x \le 10$. Relative errors plotted at $x = 10$.

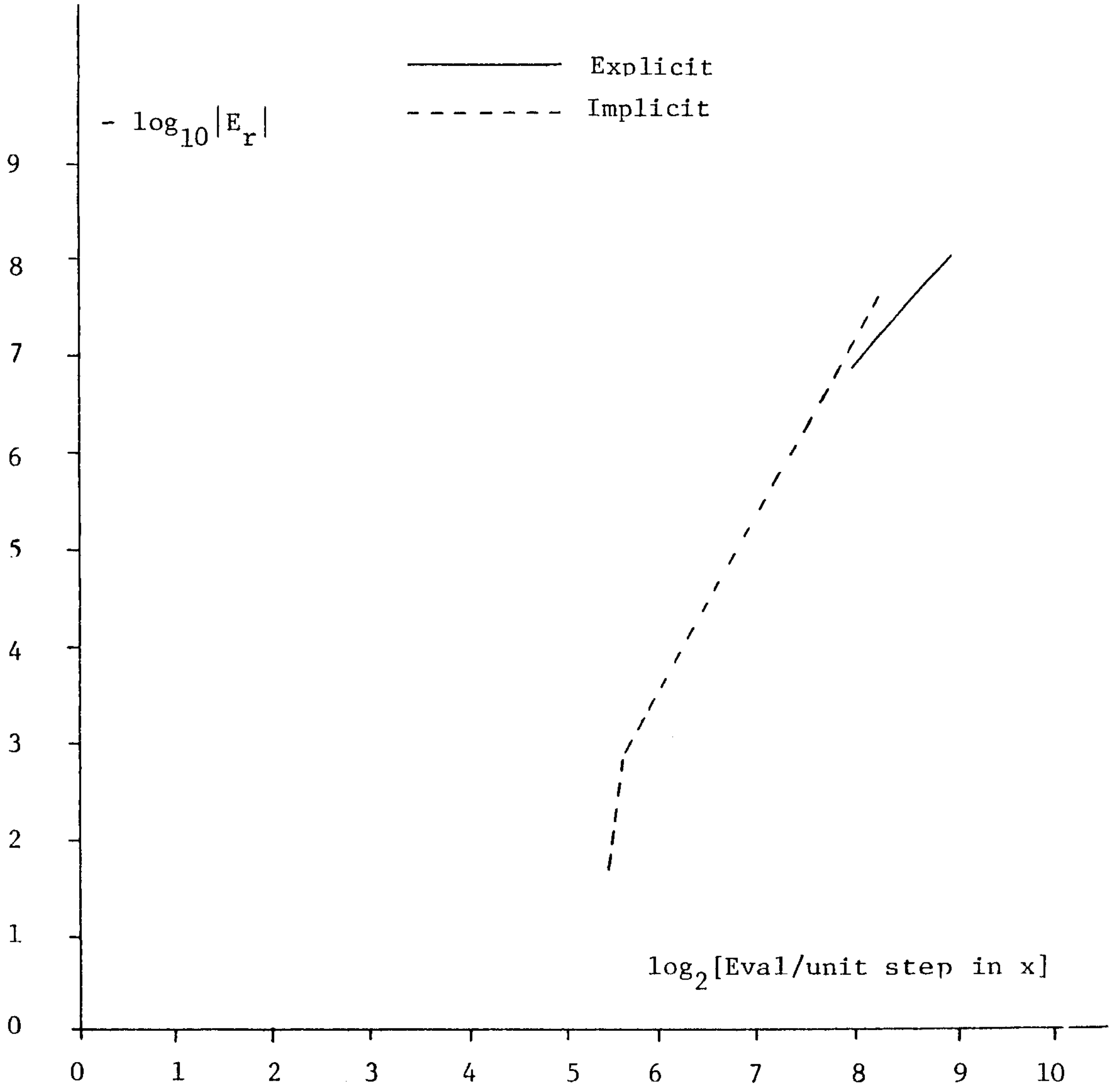

Figure 5.2

A graph of the efficiency of two methods of solving Problem 2 over the interval $0 \le x \le 10$. Relative errors plotted at $x = 10$.

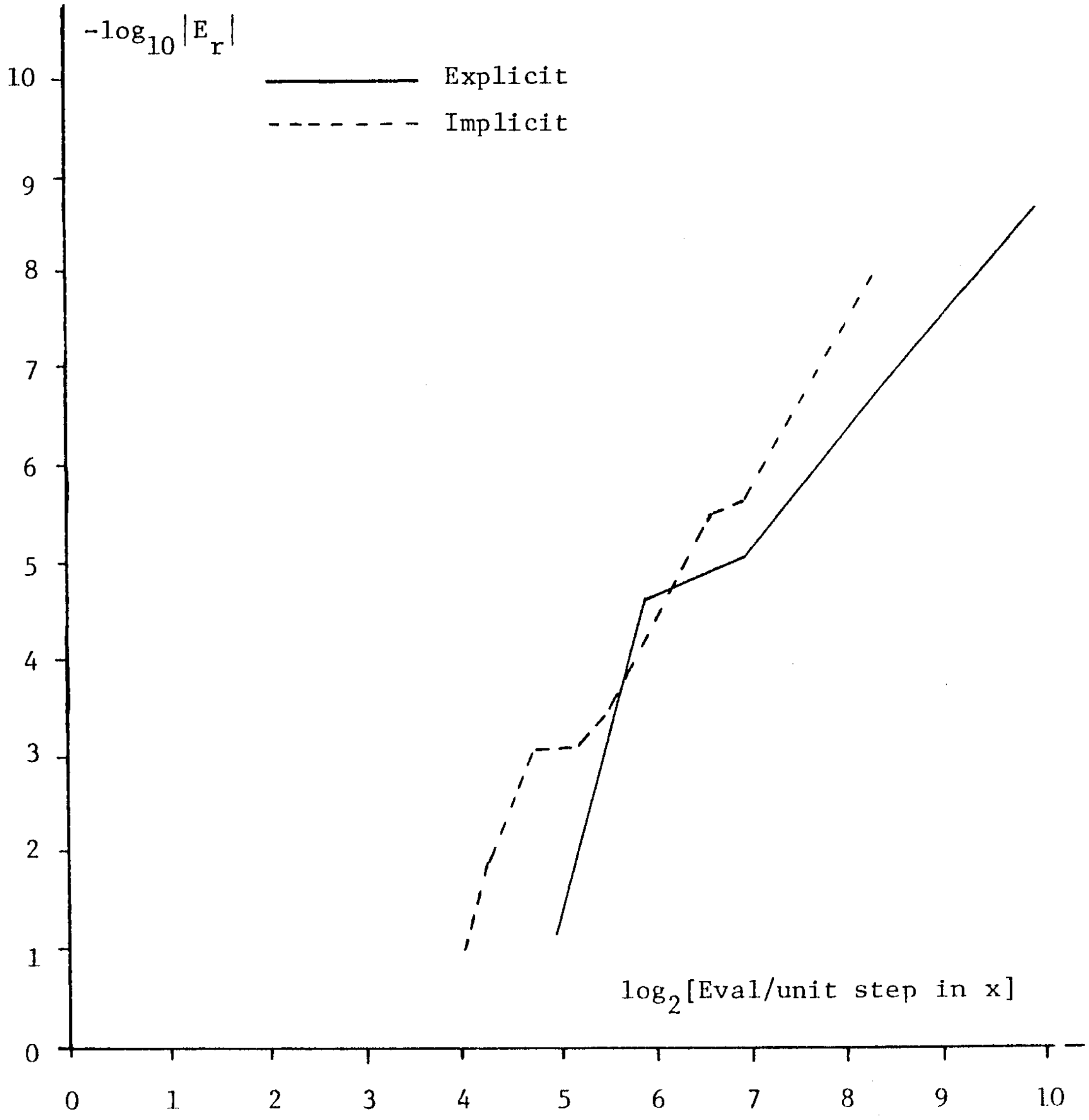

Figure 5.3

A graph of the efficiency of two methods of solving Problem 3 over the interval $0 \leq x \leq 10$. Relative errors plotted at $x = 10$.

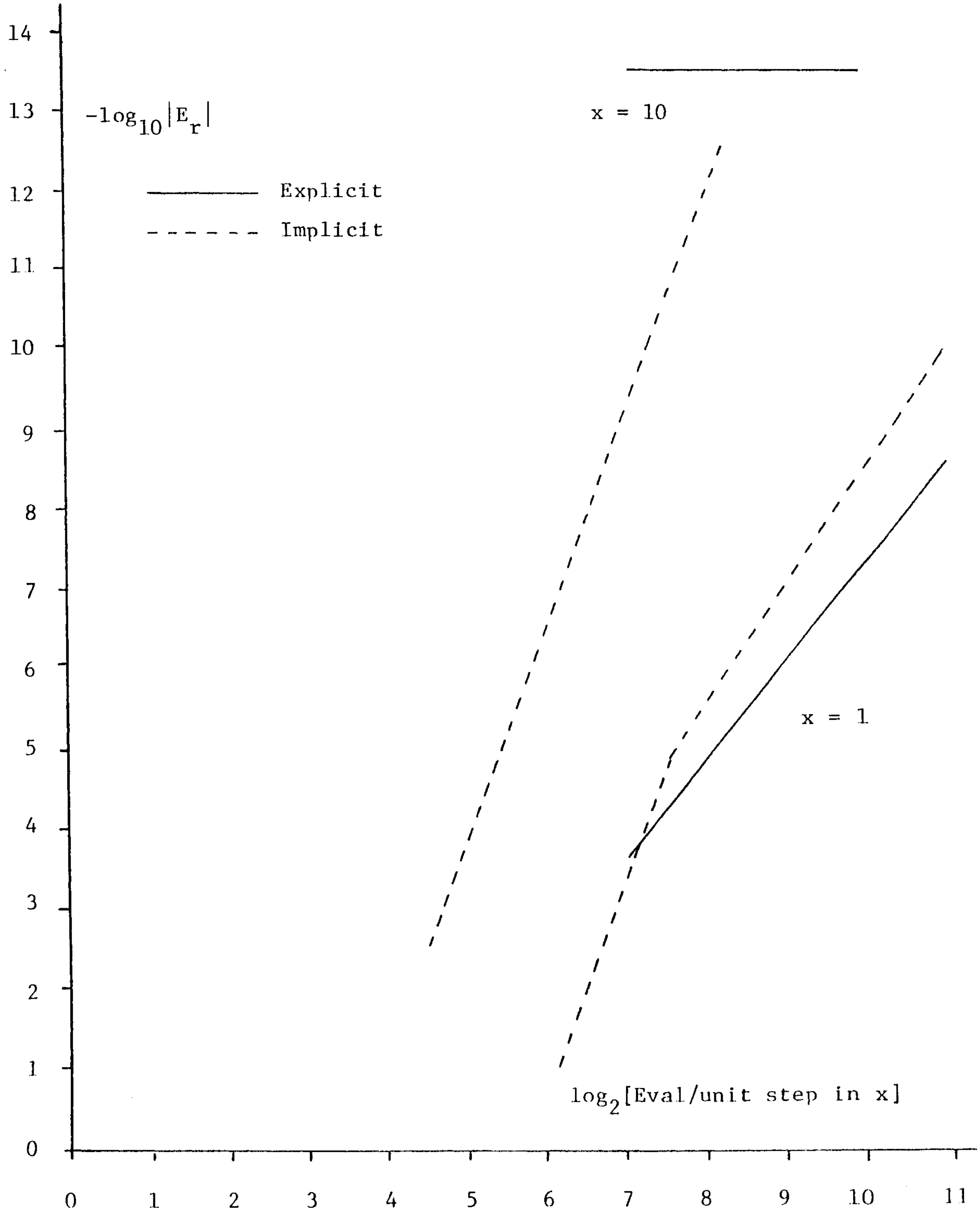

Figure 5.4

A graph of the efficiency of two methods of solving Problem 4 over the interval $0 \le x \le 10$. Relative errors plotted at x = 1 and x = 10.

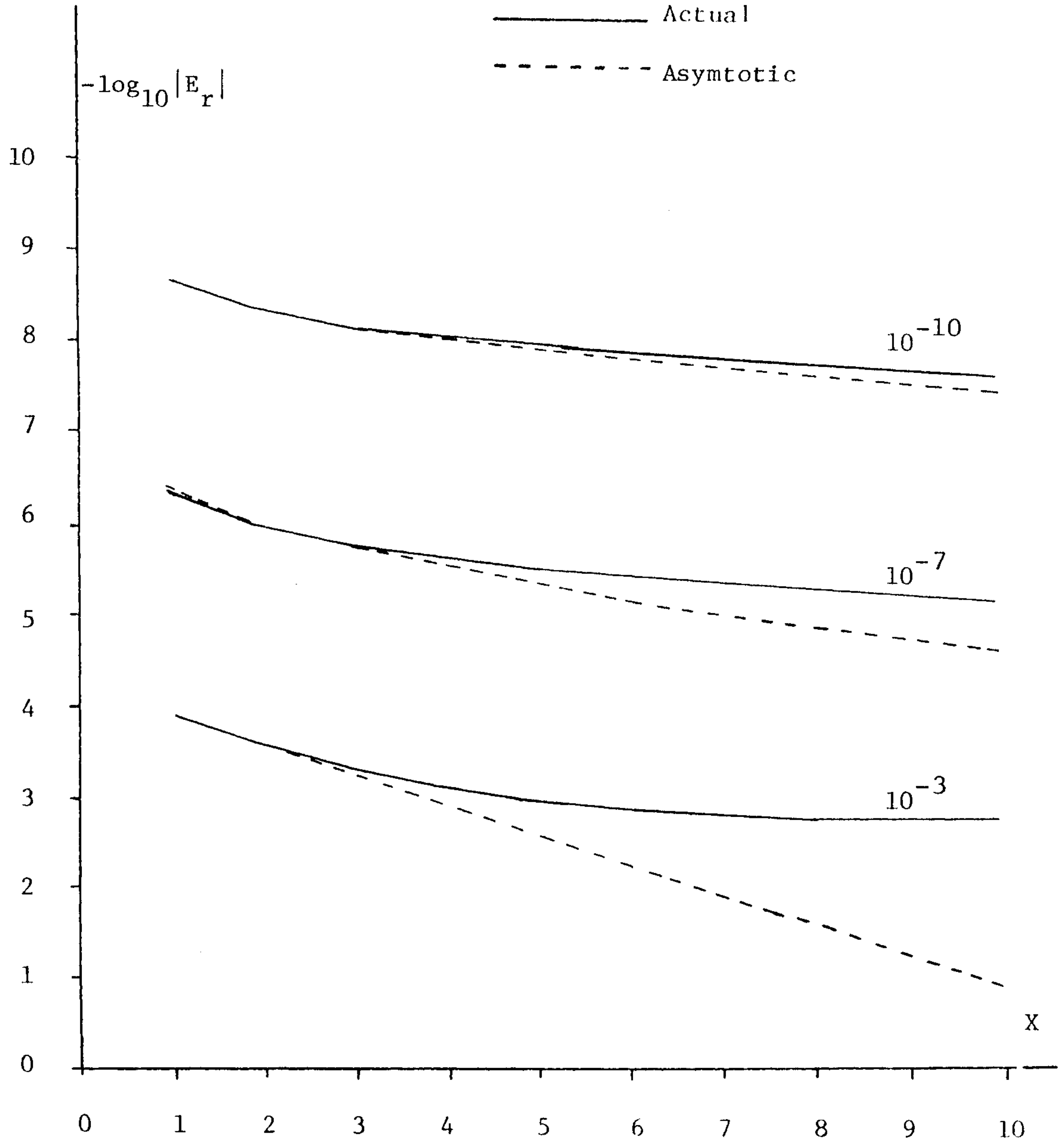

Figure 5.5

A comparison of the actual and predicted error for Problem 1 for various relative iteration errors for $1 \le x \le 10$.

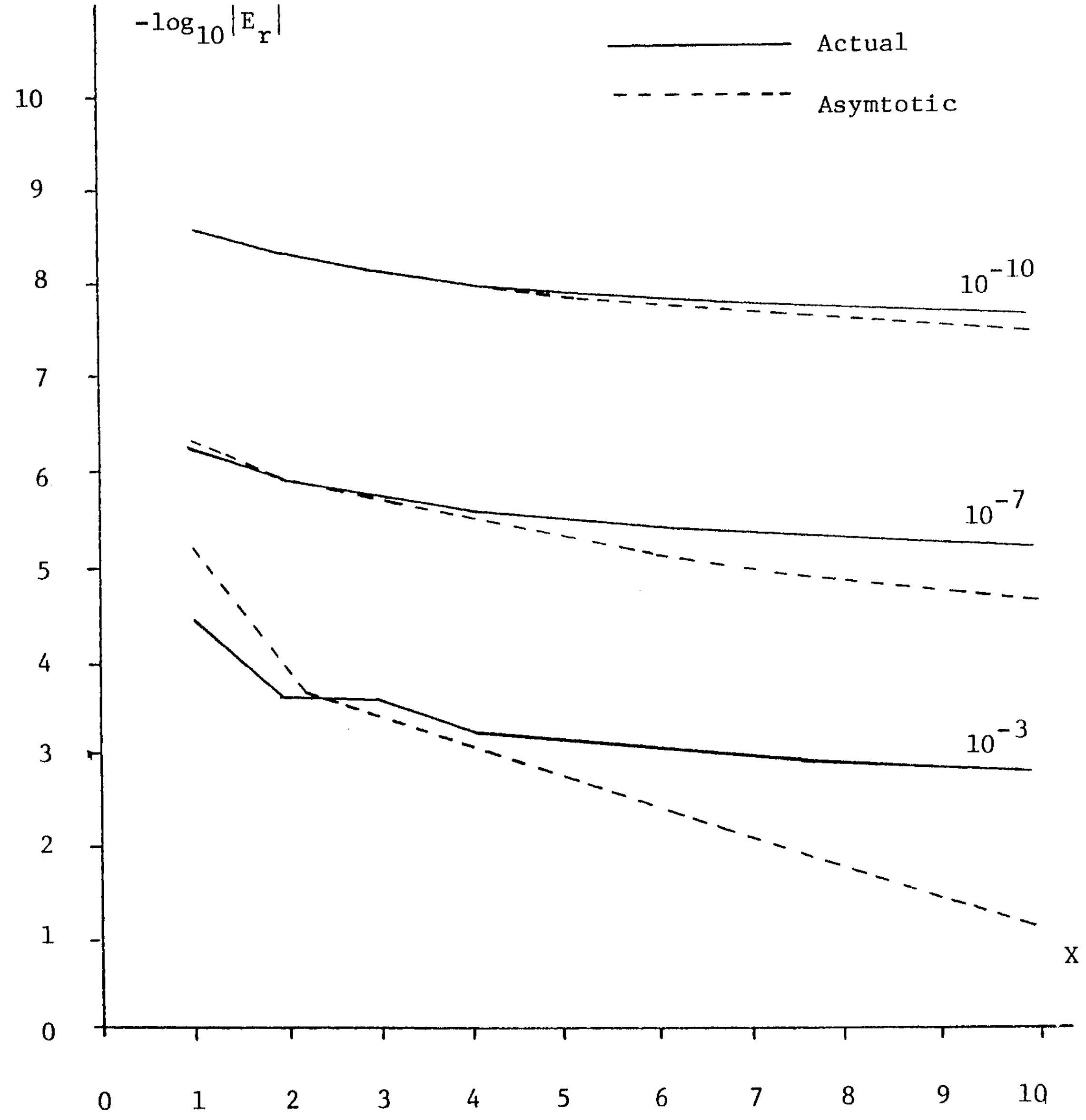

Figure 5.6

A comparison of the actual and predicted error for Problem 2 for various relative iteration errors for $1 \le x \le 10$.

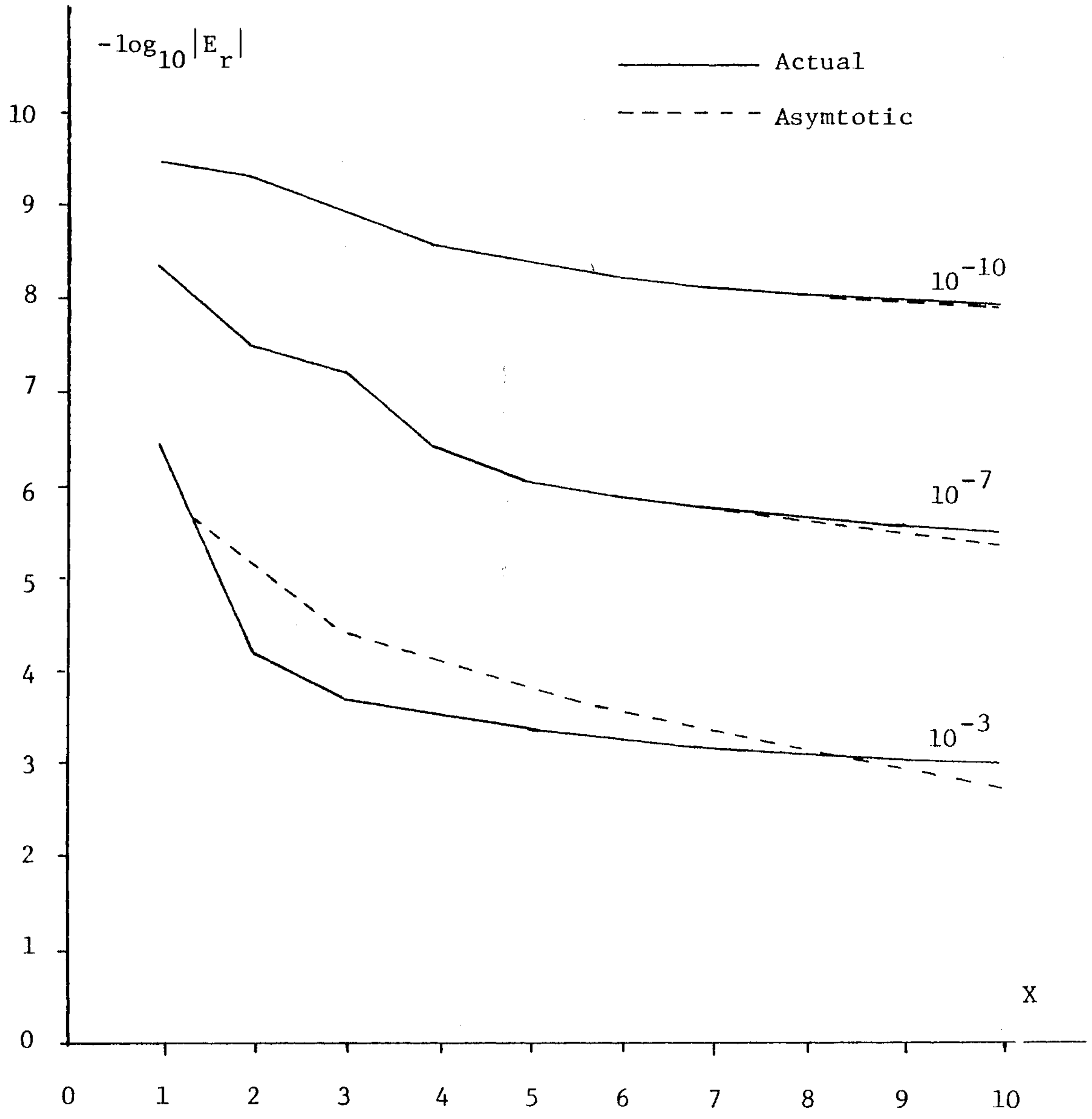

Figure 5.7

A comparison of the actual and predicted error for Problem 3 for various relative iteration errors for $1 \le x \le 10$.

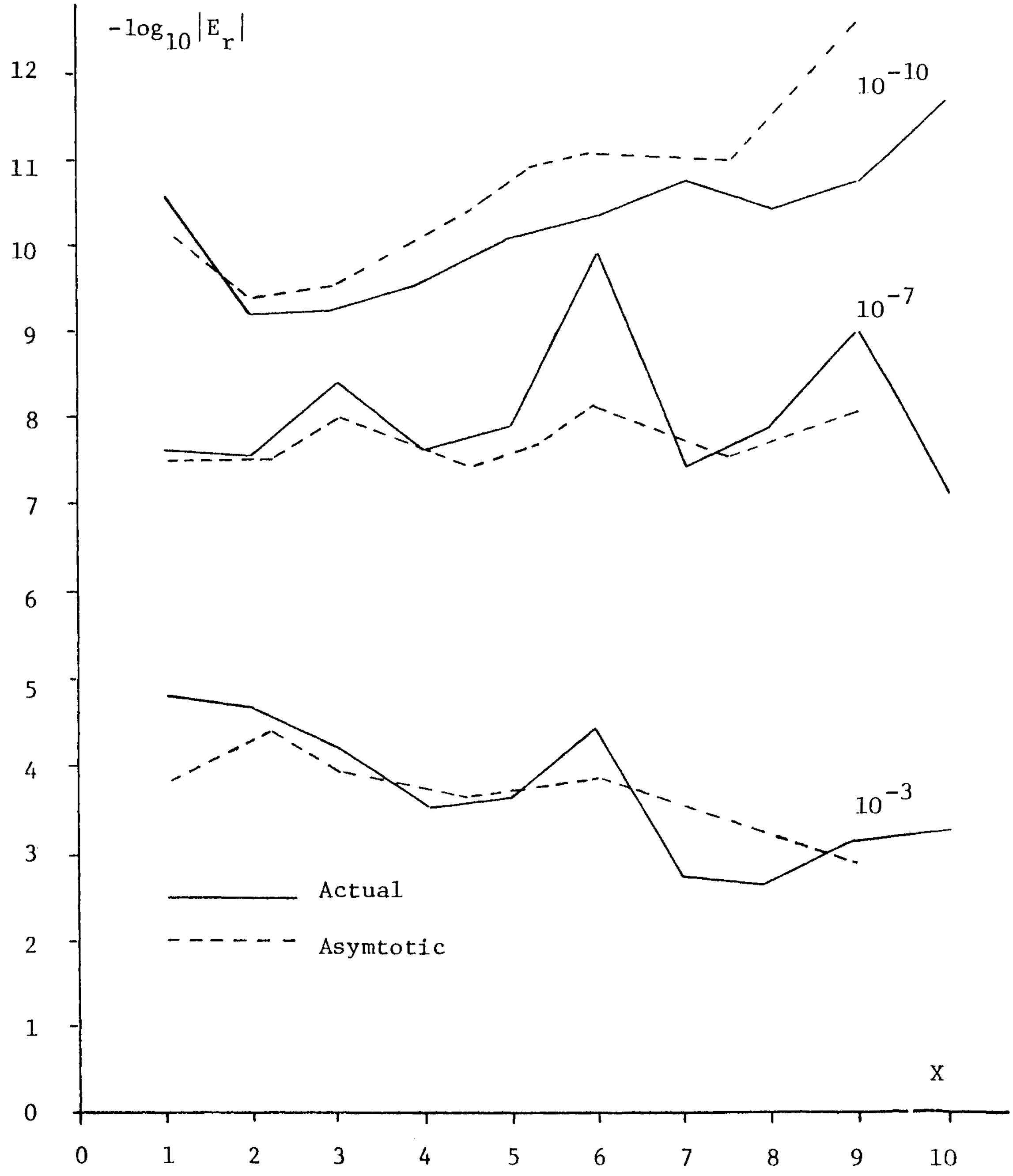

Figure 5.8

A comparison of the actual and predicted error for Problem 4 for various relative iteration errors for $1 \leq x \leq 10$.

of the combined process must be considered.

As noted by Ralston [52, p. 188], (5.5.1) would not be a stable corrector for a predictor-corrector process. In [18] it is shown that (5.5.1) is absolutely stable when used with the classical 4th order Runge-Kutta process provided the combined process satisfies the restriction $-6.6 < qH < 0$ when solving (1.2.3) where $H = 3h$.

In a similar way to that of [18] we can show that solving (1.2.3) using the exact solution given by the implicit Runge-Kutta process in combination with (5.5.1) that

$$y^*((n+1)H) = \frac{\sum_{i=0}^{7} c_i (qH)^i}{\sum_{j=0}^{6} d_j (qH)^j} y^*(nH) = E(qH) y^*(nH) \qquad (5.8.1)$$

where the c_i and d_j are appropriate constants and $y^*(nH)$ is the n^{th} value produced by the combined process.

It is easily verified that $|E(qH)| < 1$ provided $-19.5 < qH < 0$. The absolute stability constraint on the step size of the implicit Runge-Kutta process is thus $-6.5 < qh < 0$ if the combined process is used for the solution. We note that this constraint compares favorably with the explicit 5^{th} and 6^{th} order Runge-Kutta processes proposed by Lawson in [35] and [37].

For the procedure given in this thesis, where the combined process produces only a local error estimate, this stability constraint does not apply. Instead, we observe that the technique we have described in the previous sections is in fact a type of predictor-corrector process and its stability should be analyzed on that basis. For example, if we assume that only one iteration is done on the $\vec{K}_i$'s then the effect of the predictor should be included in the stability analysis of the total process, as noted by Fox [21, p. 55], since this can noticably reduce the stability region.

Unfortunately, in using the predictor (5.3.1), which allows us to make a 4th order prediction of y in only one step of h, we use function values at irrationally spaced values of x. This means that the resulting characteristic polynomial, $C(r) = 0$ [21], [52], [26] has terms involving r raised to powers which are irrational as well as rational and hence the usual theorems for the roots of $C(r) = 0$ do not apply. Thus, the usual stability analysis is not possible. We note that the results of the four problems considered suggests that stability is not a major problem in the process.

One way of avoiding the above difficulty and obtaining the usual stability analysis would be to use a predictor based on rationally spaced values of x. To do this would clearly lengthen the starting interval of the process and might not produce any better results. Additional study along these lines is possible but will not be considered here since as was indicated at the start of this chapter we were interested in showing that such processes could be implemented and not in being exhaustive in our study of them.

5.9 Improving the Estimates for the $\vec{K}_i$'s

As suggested by Ralston [52, p. 186], when using predictor-corrector methods it may be worthwhile to correct the predictor by using information about how badly it was in error on the previous prediction. This procedure was included in the program used to produce the results shown in Figures 5.1, 5.2, 5.3, and 5.4. It resulted in a slight reduction (averaging about 5%) in the number of function evaluations required per unit step in x as compared with a program which did not include this feature.

BIBLIOGRAPHY

1. Ahlfors, L.V., Complex Analysis, McGraw Hill, New York, 1953.

2. Birkhoff, G., and MacLane, S., A Survey of Modern Algebra, MacMillan, New York, 1953.

3. Birkhoff, G., and Varga, R.S., "Discretization Errors for Well-set Cauchy Problems, I", J. of Math and Physics, Vol. 44 (1965), pp. 1-23.

4. Butcher, J.C., "Coefficients for the Study of Runge-Kutta Integration Processes", J. of the Australian Mathematical Society, Vol. III (1963), pp. 185-201.

5. ____________, "Implicit Runge-Kutta Processes", Math Comp., Vol. 18 (1964), pp. 50-64.

6. ____________, "Integration Processes Based on Radau Quadrature Formulas", Math Comp., Vol. 18 (1964), pp. 233-244.

7. Calahan, D.A., "Numerical Solution of Linear Systems with Widely Separated Time Constants", Proc. IEEE, Nov. 1967, pp. 2016-2017.

8. ____________, "A Stable, Accurate Method of Numerical Integration for Nonlinear Systems", Proc. IEEE, April, 1968, p. 744.

9. Cody, W.J., and Ralston, A., "A Note on Computing Approximations to the Exponential Function", Comm. ACM, Vol. 10, No. 1 (1967), pp. 53-55.

10. Curtiss, C.F., and Hirschfelder, J.O., "Integration of Stiff Equations", Proc. Natl. Acad. Sci. U.S., Vol. 38 (1952), pp. 235-243.

11. Dahlquist, G., "A Special Stability Problem for Linear Multi-Step Methods", BIT, Vol. 3 (1963), pp. 27-43.

12. ____________, "Stability Questions for some Numerical Methods for Ordinary Differential Equations", Proceedings of Symposia in Applied Mathematics, American Math. Soc., Providence, Rhode Island, Vol. 15 (1963), pp. 147-158.

13. Davison, E.J., "A High-order Crank-Nicholson Technique for Solving Differential Equations", Computer Journal, Vol. 10 (1967), pp. 195-197.

14. Dieudonné, M.J., "Sur un Probléme de la Théorie des Polynomes", C. R. Acad. Sci. Paris, Vol. 199 (1934), pp. 999-1001.

15. ______________, Foundations of Modern Analysis, Academic Press, New York, 1960.

16. ______________, private communication.

17. Ehle, B.L., "High Order A-stable Methods for the Numerical Solution of Systems of D. E.'s", Dept. of AACS, University of Waterloo, Research Report, CSRR 2003, August, 1968.

18. __________, "Asymtotic Errors for Runge-Kutta Processes Using Multistep Formulas", Dept. of AACS, University of Waterloo, Research Report, CSRR 2006, October, 1968.

19. __________, "The Application of Implicit Runge-Kutta Processes to Stiff Systems", Dept. of AACS, University of Waterloo, Research Report, CSRR 2007, October, 1968.

20. __________, "High Order A-stable Methods for the Numerical Solution of Systems of D. E.'s", BIT, Vol, 8, No. 4 (1968), pp. 276-278.

21. Fox, L., Numerical Solution of Ordinary and Partial Differential Equations, Addison-Wesley Publishing Co., Reading, Massachusetts, 1962.

22. Gear, C.W., "Numerical Integration of Stiff Ordinary Differential Equations", Report #221, Department of Computer Science, University of Illinois, Urbana, 1967.

23. __________, "The Automatic Integration of Stiff Ordinary Differential Equations", Proceedings IFIP Congress 68, August, 1968, Edinburgh, Scotland, Book A, pp. 81-85.

24. Guillemin, E.A., The Mathematics of Circuit Analysis, John Wiley and Sons, New York, 1949, pp. 395-409.

25. Hart, J., et al, "Computer Approximation", SIAM, Ser. in App. Math, John Wiley and Sons, New York, 1968.

26. Henrici, P., Discrete Variable Methods in Ordinary Differential Equations, John Wiley and Sons, New York, 1962.

27. Hermite, C., "Sur la Formule d'interpolation de Lagrange", J. für Reine u. Angew. Math, Vol. 84 (1878), pp. 70-79; Oeuvres, Vol. 3, pp. 432-443.

28. Hildebrand, F.B., Introduction to Numerical Analysis, McGraw Hill, New York, 1956.

29. Hille, E., Analytic Function Theory, Ginn and Co., Boston, 1959, Vol. 1.

30. Hummel, P.M., and Seebeck, C.L., "A Generalization of Taylor's Theorem", Amer. Math. Monthly, Vol. 56 (1949), pp. 243-247.

31. Iverson, K.E., "The Zeros of the Partial Sums of e^z", Mathematical Tables and Other Aids to Computation Vol. 7, July (1953), pp. 165-168.

32. Kaplan, W., Advanced Calculus, Addison-Wesley Publishing Col, Reading, Massachusetts, 1952.

33. Kuo, F.F., and Kaiser, J.F., System Analysis by Digital Computer, John Wiley and Sons, New York, 1966, pp. 102-103.

34. Lanczos, C., Applied Analysis, Prentice Hall, Englewood Cliffs, N.J., 1956.

35. Lawson, J.D., "An Order Five Runge-Kutta Process with Extended Region of Stability", SIAM J. Numer. Anal., Vol. 3, No. 4 (1966), pp. 593-597.

36. ____________, "Generalized Runge-Kutta Processes For Stable Systems with Large Lipschitz Constants", SIAM J. Numer. Anal., Vol. 4, No. 3 (1967), pp. 372-380.

37. ____________, "An Order Six Runge-Kutta Process with Extended Region of Stability", SIAM J. Numer. Anal., Vol. 4, No. 4 (1967), pp. 620-625.

38. Legras, J., "Résolution Numérique des Grands Systèmes Différentiels Linéaires", Numerische Math., Vol. 8 (1966), pp. 14-28.

39. Liniger, W. and Willoughby, R., "Efficient Numerical Integration of Stiff Systems of Ordinary Differential Equations", IBM Labs, Yorktown Heights, N.Y., Report RC-1970.

40. Loscalzo, F.R., "On the Use of Spline Functions for the Numerical Solution of Ordinary Differential Equations", MRC Technical Summary Report No. 869, University of Wisconsin, May, 1968.

41. ______________, "Numerical Solution of Ordinary Differential Equations by Spline Functions (SPLINDIF)", MRC Technical Summary Report No. 842, January, 1968. University of Wisconsin.

42. Loscalzo, F.R., and Schoenberg, I.J., "On the use of Spline Functions for the Approximation of Solutions of Ordinary Differential Equations", MRC Technical Summary Report No. 723, University of Wisconsin, January, 1967.

43. Loscalzo, F.R., and Talbot, T.D., "Spline Function Approximations for Solutions of Ordinary Differential Equations", Bull. Amer. Math. Soc., Vol. 73 (1967), pp. 438-442.

44. ________________________, "Spline Function Approximations for Solutions of Ordinary Differential Equations", SIAM J. Numer. Anal., Vol. 4 (1967), pp. 433-445.

45. Lyusternik, L.A., Chervonenkis, O.A., and Yanpolskii, A.R., Handbook for Computing Elementary Functions, Pergamon Press, Oxford, 1965.

46. Makinson, G.J., "Stable High Order Implicit Methods for the Numerical Solution of Systems of Differential Equations", Computer Journal, Vol. 11, No. 3 (1968), pp. 305-310.

47. Marden, M., Geometry of Polynomials, American Mathematical Society, Providence, Rhode Island, 1966.

48. Milne, W.E., Numerical Solution of Differential Equations, John Wiley and Sons, New York, 1953.

49. Obrechkoff, N., "Sur les Quadratures Mecaniques", (Bulgarian, French summary), Spisanie Bulgar. Akad. Nauk, Vol. 65 (1942), pp. 191-289; MR, Vol. 10 (1949), p. 70.

50. Padé, H., "Sur la Représentation Approchée d'une Fonction par des Fractions Rationnelles", Thesis, Ann. de l'Ec. Nor., Vol. 9, No. 3 (1892).

51. Pope, D.A., "An Exponential Method of Numerical Integration of Ordinary Differential Equations", Comm. ACM, Vol. 6, No. 8 (1963), pp. 491-493.

52. Ralston, A., A First Course in Numerical Analysis, McGraw-Hill, New York, 1965.

53. Reddick, H.W., and Kibbey, D.E., Differential Equations, 3rd Edition, John Wiley and Sons, New York, 1956.

54. Rosenbrock, H.H., "Some General Implicit Processes for the Numerical Solution of Differential Equations", Computer Journal, Vol. 5 (1962-63), pp. 329-330.

55. Rosenbrock, H.H., and Storey, C., Computational Techniques for Chemical Engineers, Pergamon Press, London, 1966.

56. Routh, E.J., The Advanced Part of a Treatise on the Dynamics of a System of Rigid Bodies, MacMillan Co., London, 1892, pp. 194-202.

57. __________, The Advanced Part of a Treatise on the Dynamics of a System of Rigid Bodies, Dover, New York, 1955, pp. 221-227.

58. Storch, L., "Synthesis of Constant-Time Delay Ladder Networks Using Bessel Polynomials, Proc. I.R.E., Vol. 42 (1954) pp. 1666-1675.

59. Varga, R.S., "On Higher Order Stable Implicit Methods for Solving Parabolic Partial Differential Equations", J. Math Physics, Vol. 40 (1961), pp. 220-231.

60. Wall, H.S., "Polynomials Whose Zeros Have Negative Real Parts", American Math Monthly, Vol. 52 (1945), pp. 308-322.

61. __________, Analytic Theory of Continued Fractions, Van Nostrand, Princeton, New Jersey, 1948.

62. Watson, G.N., A Treatis on the Theory of Bessel Functions, Cambridge University Press, 1952.

63. Widlund, O.B., "A Note on Unconditionally Stable Linear Multistep Methods", BIT, Vol. 7 (1967), pp. 65-70.