A Class of Approximations to the Exponential Function for the Numerical Solution of Stiff Differential Equations

by
Terence Chun-Yat Lau

CS-74-13

Department of Applied Analysis and
Computer Science

University of Waterloo
Waterloo, Ontario, Canada

August, 1974

A CLASS OF APPROXIMATIONS TO THE EXPONENTIAL FUNCTION FOR THE NUMERICAL SOLUTION OF STIFF DIFFERENTIAL EQUATIONS

by

TERENCE CHUN-YAT LAU

A THESIS SUBMITTED
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

AT THE

UNIVERSITY OF WATERLOO

DEPARTMENT OF APPLIED ANALYSIS
AND COMPUTER SCIENCE,
FACULTY OF MATHEMATICS.

ONTARIO, CANADA.

JUNE, 1974.

I hereby declare that I am the sole author of this thesis.

I authorize the University of Waterloo to lend it to other institutions or individuals for the purpose of scholarly research.

Signature Lau Chun Yat.

Dedicated to

The Triune God,
Source of all Wisdom,
Lord of the Universe,
Hope of the World.

ACKNOWLEDGEMENTS

The author is especially grateful to Professor J.D. Lawson for the suggestion of the thesis topic and his many helpful comments. To Mrs. Teresa Miao goes my thanks for her splendid typing. And finally my greatest debt is to Rebecca for her cheerful acceptance of a part-time fiancé throughout the writing of this thesis.

Without the financial support provided by the National Research Council, Canada, this work would not have been possible.

ABSTRACT

Stiff differential equations form a class of problems which have very practical industrial significance. They are problems in which the time constants differ greatly in magnitude. The system of differential equations $\bar{y}' = B\bar{y}$, where B is a real square matrix, all of whose eigenvalues have negative real parts and such that the modulus of at least one of them is large as compared with the others, provides a typical example of this kind.

In attempting to solve such differential equations numerically, the exponential function is approximated in some way. Some of the popular approximations include the Padé approximations, the Chebyshev minimax approximations and the order-constrained minimax approximations. All these approximations are rational polynomials and many numerical methods employ the technique of factorizing the numerators and denominators of these rational functions into linear or quadratic factors. It is found that all (except perhaps one) roots of the denominator polynomial of almost every approximation in these families are complex. Hence, nearly all factors of the denominator will be quadratic.

This thesis is concerned with the developing of a class of rational approximations whose denominators can be factorised completely into linear factors. Let R_m^n be the set of all rational functions having such a property, and whose numerator and denominator polynomials are of degree less than or equal to m and n respectively. It is shown that the best approximation in the minimax sense to exp(-x) in R_0^n is that unique rational function whose denominator is a product of n identical linear factors, and that the error curve of approximation alternates exactly twice. Such an approximation can always be found in R_0^n, with a minimax error of the order $O(n^{-1})$ as n tends to infinity.

With the help of classical approximation theory, it is shown that the error curve of the minimax approximation in R_m^n, $m > 0$, will alternate at least $m' + 2$ times, where m' is the degree of its numerator. If the alternation is not more than $m' + 2$ times, then it is proved the denominator will have the same property as in the case of R_0^n. In all computational tests, error curves of approximations in R_m^n do exhibit an alternation of not more than $m' + 2$ times.

Motivated by the above results, an algorithm is then developed to compute rational functions of the form

$$\frac{a_0+a_1x+\ldots+a_mx^m}{(1+bx)^n} ,$$

whose error curves alternate $m + 2$ times. It is found that R_m^{m+2}-approximations are the most judicious choice when both degree of approximation and computation cost are taken into consideration together. For a fixed m, the error of the best approximation is inversely proportional to n^c, c being a constant that increases with m.

Order-constrained approximations of this form are also investigated. Most results are similar to the unconstrained case.

Finally, the merits of such approximations as applied to the numerical solution of stiff systems are discussed. On the real axis, they are obviously A-acceptable. It is shown that there is also much gain in operation cost, storage requirement and numerical stability.

TABLE OF CONTENTS

TABLE OF CONTENTS (cont'd)

CHAPTER 1

THE 'STIFF' PROBLEM

1.1 Introduction

In the numerical solution of differential equations, there is a class of problems classified as stiff systems; it appeared as early as 1952 [13] and has been associated with many important industrial problems [2,3,8,39,42,43]. The purpose of this thesis is to investigate a class of approximations to exp(x) that has special application in solving this kind of problem. The first chapter will describe what constitutes such a stiff system and the numerical difficulty that one can encounter when trying to solve it.

1.2 Stiff Ordinary Differential Equations

Consider the initial value problem

$$\frac{dy}{dx} = cy(x), \qquad (1.2.1)$$

$$y(a) = y_0,$$

where x ranges from a to b.

(1.2.1) has an exact analytic solution of the form

$$y = y_0 \exp(cx). \qquad (1.2.2)$$

If we try to solve (1.2.1) numerically by Euler's (point-slope) method, we have,

$$\begin{aligned} y_{n+1} &= y_n + hy'_n \\ &= (1 + hc)y_n \qquad (1.2.3) \\ &= (1 + hc)^{n+1}y_0 \end{aligned}$$

where $\quad h = x_{m+1} - x_m$, for all m,

is the step-size, which we shall assume to be uniform for $x \in [a,b]$.

We notice that, for (1.2.1), if $c < 0$ and $b = \infty$, then

$$y \to 0 \quad \text{as} \quad x \to \infty;$$

while the same behaviour can be found in (1.2.3) only if

$$|1 + hc| < 1.$$

In other words, it is necessary that

$$0 < h < \frac{2}{|c|} ,$$

for the numerical solution y_n to remain bounded as n tends to infinity.

This means, for the class of problems (1.2.1) with $c < 0$, we have to impose a restriction on the step-size of its numerical solution (by Euler's method) in order that the numerical method retains its stability. If $|c|$ is large, the restriction can be very severe. One may regard this as merely a statement about an appropriate time-scale for the solution of this simple example.

Unfortunately, such severe restrictions on step-size can exist in a problem which may look innocent enough and have a solution time-scale, which is not similarly constrained. For instance, the following illustration is from [18, p.1].

Consider the initial value problem

$$\frac{dy}{dx} = f(x,y),$$

$$y(a) = y_a, \qquad a < x < b. \tag{1.2.4}$$

If f(x,y) is expanded in its Taylor's series about the point (a,y_a), (1.2.4) can be transformed to

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

Ignoring all higher terms in (x-a) or $(y-y_a)$ on the right, we have

$$\frac{dy}{dx} = A + Bx + Cy,$$

$$y(a) = y_a, \tag{1.2.5}$$

where

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

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

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

If x is near a, we believe the solution of (1.2.5) will give us an approximate answer to (1.2.4). In fact, (1.2.5), being a linear first order differential equation, has an analytic solution of the form [47, p.56],

$$y(x) = -[\frac{A}{C} + \frac{B}{C^2} + \frac{Bx}{C}] + C_1\exp(cx), \qquad c = C, \qquad (1.2.6)$$

where C_1 is a constant of integration so that $y(a) = y_a$.

Looking at (1.2.6), we notice that the undesirable property of (1.2.1), when $C < 0$, may creep into the innocent-looking problem (1.2.4) through the last term in (1.2.6) if, in (1.2.4),

$$C = \left.\frac{\partial f}{\partial y}\right|_{(a,y_a)} < 0.$$

In fact, when we try to approximate the solution of (1.2.4) by applying some numerical method on (1.2.5), we expect the numerical answer obtained to be close to (1.2.6). This means the numerical method should be good enough to give a satisfactory approximation to both the linear term

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

as well as the exponential term

$$C_1\exp(cx).$$

Obviously, the first requirement will be met quite readily by many numerical methods, regardless of step-size.

However, to satisfy the second requirement, any numerical approximation $E(x)$ to the exponential function should satisfy certain properties. For example,

if $c > 0$, then

(a) $E(cx) \geq 1$, for $x \geq 0$,

(b) $E(cx) \to \infty$, as $x \to \infty$;

while if $c < 0$, then

(c) $E(cx) \leq 1$, for $x \geq 0$,

(d) $E(cx) \to 0$, as $x \to \infty$.

Conditions (a) and (b) are satisfied immediately by many numerical methods, e.g. the Runge-Kutta processes. However, in many practical situations, it is the case when $c < 0$ that is of interest.

Let us use the Runge-Kutta processes as an illustration to investigate further the latter case.

For the initial value problem (1.2.1), the exact solution at $x = a + ih$, $i = 1,2,3,\ldots$, is given by

$$y_{n+1} = y_n \exp(ch).$$

A Runge-Kutta process of order p using r substitutions and step-size h gives

$$y_{n+1} = E(a_{p+1},a_{p+2},\ldots,a_r;ch)y_n,$$

where

$$E(a_{p+1},\ldots,a_r;ch) = \sum_{i=0}^{p} \frac{(ch)^i}{i!} + \sum_{i=p+1}^{r} a_i \frac{(ch)^i}{i!}\,, \qquad (1.2.7)$$

and $a_{p+1},a_{p+2},\ldots,a_r$ are functions of the parameters of the Runge-Kutta process, and are not constrained by order requirement.

As we have just discussed, there is a restriction on the value of ch in (1.2.7) when it is applied to (1.2.1) if we want conditions (c) and (d) to be satisfied. Figure 1.2.1 shows such regions of restriction for the ordinary order 4 Runge-Kutta process (RK4), and Lawson's order 5 and 6 Runge-Kutta processes (RK5ES, RK6ES) [29,31] with extended region of stability.

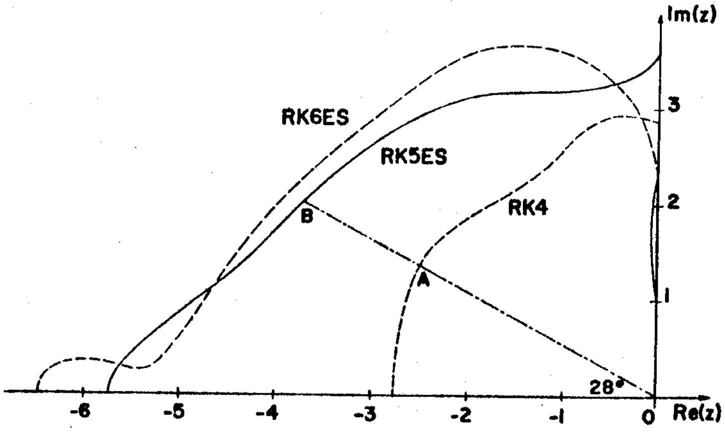

Fig.1.2.1 Stability regions for Runge-Kutta processes of orders 4, 5 and 6

For each process, ch has to be inside the corresponding region so as to satisfy condition (c). For any z = ch outside, the sequence

$$\begin{aligned} y_n &= E(ch)y_{n-1} \\ &= [E(ch)]^n y_0 \end{aligned}$$

does not converge to zero, but, instead, grows indefinitely as $n \to \infty$.

As the above discussion would suggest, when the classical Runge-Kutta processes are applied to (1.2.1) and (1.2.5) with

$$c < 0 \quad \text{and} \quad |c| >> 1,$$

the result could be extremely unsatisfactory unless the step-size is limited to a very small magnitude so that ch is within the region of stability. In fact, all explicit Runge-Kutta processes, when applied to

(1.2.1), will reduce to a polynomial in ch with positive degree in the form of (1.2.7). This means conditions (c) and (d) will not be satisfied for large ch. Actually, Dahlquist [14] pointed out that such will be the situation for all explicit linear multistep methods as well. And among the implicit multistep methods, we cannot expect anything better than order 2.

Nevertheless, (1.2.1) is a scalar differential equation and the value of h will usually be determined according to the allowable truncation error, rather than conditions (c) and (d). However, if (1.2.1) is generalised to a system of differential equations, we shall confront a worse situation.

1.3 Systems of Ordinary Differential Equations

Consider the following system of differential equations

$$\frac{d\bar{y}}{dx} = C\bar{y},$$
$$\bar{y}(a) = \bar{y}_a, \tag{1.3.1}$$

where C is a real square matrix whose eigenvalues are all distinct. (1.3.1) can be readily solved analytically. It can be shown [11, chap.3] that (1.3.1) has the fundamental matrix exp(xC) and its solution is of the form

$$\bar{y}(x) = \exp((x-a)C)\cdot\bar{y}_a,$$
$$|x| < \infty , \tag{1.3.2}$$

where, for any matrix A,

$$\exp(A) \equiv I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \dots ,$$

and the series is convergent for all A [11, p.65].

If C has the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, (1.3.2) can be written also as

$$\bar{y}(x) = \sum_{i=1}^{n} \exp(\lambda_i(x-a)) \cdot \bar{C}_i . \tag{1.3.3}$$

where $\bar{C}_i$ are column vectors of coefficients.

Now consider the numerical solutions of (1.3.1). We assume that all the eigenvalues of C have negative real parts. Extending the argument in the last section, the numerical scheme should be able to give a good approximation to each term in (1.3.3). If, for instance, K is the upper bound of the region of stability of the method used, we shall have to choose the step-size, h, such that

$$|\lambda_i h| < K, \quad i = 1,2,\ldots,n,$$

i.e. $$\lambda_{\max} h < K, \tag{1.3.4}$$

where $\lambda_{\max}$ is the eigenvalue with the maximum magnitude. Physical problems leading to this kind of equation appear very often, for example, in the flow of gas in chemical non-equilibrium, electronic circuit analysis, heat conduction, etc. [2,39,42]. Moreover, they have eigenvalues that are widely separated, i.e.

$$\max_i |\lambda_i| \gg \min_i |\lambda_i| .$$

Because of these large eigenvalues and the requirement (1.3.4), the step-size h must be severely restricted in order to achieve numerical stability , but the physical significance and contribution due to the large eigenvalues in the total solution fade out rapidly once the initial transient is over, i.e. when x is away from zero. This is obvious from (1.3.3) because, if $x >> 0$ and $\lambda_{max}, \lambda_{min} < 0$,

$$\exp(x\lambda_{max}) << \exp(x\lambda_{min}).$$

The dominant component of $\bar{y}(x)$ in (1.3.3), when x is large, will be the term associated with $\exp(x\lambda_{min})$.

Usually, we call those components associated with the large eigenvalues the highly damped components of the solution, and the whole system a stiff system [13]. And we see that the solving of a stiff system by an ordinary numerical method can be much handicapped by the presence of a component which is of only slight significance in the overall solution.

1.4 Partial Differential Equations

Consider the linear parabolic equation

$$\frac{1}{\beta}\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(\alpha(x)\frac{\partial u}{\partial x}\right) + \sum_{i=1}^{k} \gamma_i(t)\delta(x-x_i), \qquad (1.4.1)$$

$$0 < x < L, \quad t > 0,$$

subject to the homogeneous boundary conditions

$$u(0,t) = u(L,t) = 0, \quad t > 0, \qquad (1.4.2)$$

and the homogeneous initial condition

$$u(x,0) = u_0 = 0, \quad 0 \leq x \leq L, \qquad (1.4.3)$$

where $\alpha(x)$ is a positive, continuous or a piece-wise continuous function of x in $0 \le x \le L$, β a positive constant, $\delta(x-x_i)$ the Dirac delta function used to represent source and sink terms and $\gamma_i(t)$ a constant or positive, piece-wise constant. The class of physical problems which effectively reduces to (1.4.1)-(1.4.3) is very extensive, and includes problems associated with fluid flow in porous media [43], heat transfer [8], and mass transfer [3].

It is pointed out [9,12,52] that one way of solving problems of this kind is semi-discretization. That is, if the problem is discretized in space, but not in time, it can be transformed into a system of ordinary differential equations.

For instance, if we impose a partition on the x-interval [0,L] in (1.4.1)-(1.4.3), we can derive a system of spatial difference equations by replacing the differential equation at each mesh point with an appropriate finite difference equation [53]. If the standard three-point difference approximation [53, p.175] of the spatial derivatives in (1.4.1) is used, then the 'semi-discrete' approximation $\underline{u}(x_i,t) \equiv \underline{u}_i(t)$ satisfies the following ordinary matrix differential equation:

$$B(\frac{d\bar{u}}{dt}) = -A\bar{u} + \bar{g}(t), \quad t > 0, \tag{1.4.4}$$

where $\bar{u}(0) = \bar{u}_0 \equiv \bar{0}$,

and $\bar{u}(t) = (\underline{u}_1(t), \underline{u}_2(t), \ldots, \underline{u}_n(t))^T$.

The matrix B is a positive, real, diagonal, n × n matrix with diagonal elements $b_{ii} = 1/\beta$ and A is a real, symmetric, tri-diagonal, positive-definite, n × n matrix. The vector $\bar{g}(t)$ represents the source terms in (1.4.1).

For constant g(t), the solution $\bar{u}(t)$ of (1.4.4) can be verified [12] to be

$$\bar{u}(t) = A^{-1}\bar{g} + \exp(-tB^{-1}A)[\bar{u}_0 - A^{-1}\bar{g}], \quad t > 0. \qquad (1.4.5)$$

We notice immediately the term $\exp(-tB^{-1}A)$ in (1.4.5). Inevitably, we shall encounter the same problem as before if $(-tB^{-1}A)$ has the same undesirable behaviour as C in the last section, i.e. its eigenvalues have negative real parts and are widely separated. Unfortunately, this is often the case. In fact, (1.4.4) is in general stable, stiff and sparse, and the matrices have only real eigenvalues. This means some kind of special numerical method is required to deal with it.

*   *   *   *   *   *   *

In the discussion of this chapter, we confront a class of differential equations which is of great practical importance, but computationally extremely challenging. In fact, much research interest has been aroused trying to master the solution of stiff systems. In the next chapter, we shall summarise some important results in approximation theory which are required in many numerical methods for the solutions of stiff systems.

CHAPTER 2

SOME KNOWN RESULTS IN APPROXIMATION THEORY

As the discussion in Chapter 1 reveals, the stiffness of a system of differential equations is born out of the term exp(xA) in the exact solution of the system, in particular, when A has widely separated eigenvalues **in** the left complex-plane. Consequently, many of the current methods of solving stiff systems are designed with a motivation to tackle, either directly or indirectly, this troublesome term. In fact, it turns out that each of these methods, when applied to (1.2.1) or (1.3.1), is associated with a way of approximating the exponential function exp(x). Usually, the behaviour of the method depends on this basic approximation. We shall elaborate, in more detail, upon these methods and their exponential approximation behaviour in Chapter 3. We summarise some definitions and results in approximation theory that will be necessary for subsequent discussion.

2.1 Padé Approximations

Let us write any approximation to the exponential exp(z) as E(z) for any complex number z. The first three definitions are after Ehle [19].

Definition 2.1.1

E(z) is of order m iff $E(z)-\exp(z) = O(z^m)$ when $z \to 0$ for some $m \geq 1$.

Definition 2.1.2

E(z) is A-acceptable iff $|E(z)| < 1$ for $\mathrm{Re}(z) < 0$.

Definition 2.1.3

E(z) is L-acceptable iff it is A-acceptable and $|E(z)| \to 0$ as $\mathrm{Re}(-z) \to \infty$.

Let the (i,j)-th entry in the Padé table of approximations to exp(z) be denoted by

$$P_{ij}(z) = \frac{N_{ij}(z)}{D_{ij}(z)}, \quad i = 0,1,2,\dots, \quad j = 0,1,2,\dots,$$

where $D_{ij}(z)$ and $N_{ij}(z)$ are polynomials of degree i and j respectively.

$i \backslash j$	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 2.1.1 Padé Approximation to exp(z)

The next two theorems are due to Ehle [18].

Theorem 2.1.1

$P_{nn}(z)$ is A-acceptable of order (2n+1) for all non-negative integers n = 0,1,2,... .

Theorem 2.1.2

$P_{n+1,n}(z)$ and $P_{n+2,n}(z)$ are L-acceptable of order (2n+2) and (2n+3) respectively for all non-negative integers $n = 0,1,2,\ldots$.

The two theorems guarantee that the diagonal entries in the Padé table of approximation (Table 2.1.1) are A-acceptable, while the first and second sub-diagonal elements are L-acceptable.

2.2 Minimax and Order-constrained Minimax Approximations

We shall adopt the following notation and assumptions.

$r_{m,n}(x)$ - the rational function $p_m(x)/q_n(x)$, where $p_m(x)$ and $q_n(x)$ are polynomials of degree less than or equal to m and n respectively,

$$r_{m,n}(x) = \frac{p_m(x)}{q_n(x)} = \frac{a_0+a_1x+\ldots+a_mx^m}{b_0+b_1x+\ldots+b_nx^n} .$$

$\bar{a}(r_{m,n})$ - the vector of coefficients of $r_{m,n}(x)$,

$\bar{a}(r_{m,n}) = (a_0,a_1,\ldots,a_m,b_0,b_1,\ldots,b_n)$.

$R_{m,n}$ - the set of all rational functions $r_{m,n}(x)$,

$R_{m,n} = \{r_{m,n}(x)\}$.

$I_{a,b}$ - a closed subinterval of the real line,

$I_{a,b} = [a,b] = \{x:a \le x \le b\}$.

I_0 - the non-negative real axis,

$I_0 = [0,\infty) = \{x:0 \le x < \infty\}$.

Since our particular interest is in the approximation of the exponential function exp(-x) on I_0, we shall assume that any function or value is continuous and bounded in I_0. Besides, any value that depends on x has a limit as $x \to \infty$. That means we can add the point $x = \infty$ into I_0, and redefine

$$I_0 = [0,\infty] = \{x:0 \leq x \leq \infty\},$$

such that the value of a function at $x = \infty$ is defined to be its limit as $x \to \infty$. Hence, in the remaining discussion throughout this thesis, I_0 will be considered conceptually as a closed interval. If any one of these assumptions is not true or obvious, it will be discussed and dealt with explicitly.

Definition 2.2.1

The Chebyshev norm of f, defined, bounded and continuous in $I_{a,b}$, is

$$||f|| = \max_{x \in I_{a,b}} |f(x)|. \tag{2.2.1}$$

Definition 2.2.2

Suppose that f(x) is to be approximated by $r_{m,n}(x) \in S$ on $I_{a,b}$ where S is a subset of $R_{m,n}$. The minimal deviation $\rho_S(f)$ is defined by

$$\rho_S(f) = \inf_{r_{m,n} \in S} ||f(x)-r_{m,n}(x)||. \tag{2.2.2}$$

Definition 2.2.3

If there exists $r^*_{m,n}(x) \in S$, $S \subset R_{m,n}$, such that

$$\rho_S(f) = ||f(x)-r^*_{m,n}(x)||, \tag{2.2.3}$$

then $r^*_{m,n}(x)$ is called the best approximation to $f(x)$ with respect to (or in) S on $I_{a,b}$.

Definition 2.2.4

The error curve $f(x)-r_{m,n}(x)$ is said to alternate t times on $I_{a,b}$ if there are $t+1$ points

$$a \leq x_1 < x_2 < \dots < x_t < x_{t+1} \leq b,$$

such that

$$\begin{aligned} f(x_i)-r_{m,n}(x_i) &= -(f(x_{i+1})-r_{m,n}(x_{i+1})), \ i = 1,\dots,t \\ &= \pm||f(x)-r_{m,n}(x)|| \end{aligned} \tag{2.2.4}$$

on $I_{a,b}$.

Definition 2.2.5

The points x_i in definition 2.2.4 are called the extremal points.

Definition 2.2.6 [48, p.78]

A rational function $r_{m,n}(x)$ of the form

$$r_{m,n}(x) = \frac{a_0+a_1x+\dots+a_mx^m}{b_0+b_1x+\dots+b_nx^n}, \tag{2.2.5}$$

$$\sum_{i=0}^{n} b_ix^i \neq 0, \quad x \in I_{a,b},$$

is said to be of degree $m(\bar{a}) = m+n+1-d$ at $\bar{a}$ if $r_{m,n}(x)$ may be written as

$$r_{m,n}(x) = \frac{a_0'+a_1'x+\ldots+a_{m-p}'x^{m-p}}{b_0'+b_1'x+\ldots+b_{n-q}'x^{n-q}}, \qquad (2.2.6)$$

where $d = \min(p,q)$,

$a_{m-p}' \neq 0$,

$b_{n-q}' \neq 0$,

$\bar{a} = \bar{a}(r_{m,n}) = (a_0,\ldots,a_m,b_0,\ldots,b_n)$,

and the numerator and denominator of (2.2.6) have no common factor. If $r_{m,n} \equiv 0$, then $m(\bar{a}) = m+1$.

Theorem 2.2.1 (Rational Minimax Approximation)

Let $f(x)$ be continuous on $I_{a,b}$, and $w(x)$ a function, such that $w(x) \neq 0$, $x \in I_{a,b}$, then

(a) There exists $r_{m,n}^*(x) \in R_{m,n}$, such that

$\rho_{R_{m,n}}(f) = ||f(x)-w(x)\cdot r_{m,n}^*(x)||$,

i.e. $f(x)$ possesses a best rational Chebyshev approximation in $R_{m,n}$.

(b) $r_{m,n}^*$ is unique in the sense that two rational functions are identical when they coincide after being reduced to the lowest term as in (2.2.6).

(c) $r_{m,n}^*$ is the best rational Chebyshev approximation to $f(x)$ iff $f(x)-r_{m,n}^*(x)$ alternates at least $m(\bar{a}^*)$ times where $m(\bar{a}^*)$ is the degree of $r_{m,n}^*(x)$ and $\bar{a}^*$ the vector of coefficients of $r_{m,n}^*$.

Proof [1, p.53-57] or [48, p.74-80].

Definition 2.2.7

Let $f(x)$ be continuous on $I_{a,b}$, $k \geq 1$, and $f(x) \in C^{k-1}$ at $x = a$, then we define $R_{k,m,n}(f) \subset R_{m,n}$,

$$R_{k,m,n}(f) = \{r_{k,m,n}(x) \in R_{m,n}: \\ \frac{d^i}{dx^i} r_{k,m,n}(x)\Big|_{x=a} = \frac{d^i}{dx^i} f(x)\Big|_{x=a}, \ i = 0,1,\dots,k-1\}. \quad (2.2.7)$$

Theorem 2.2.2 (Order-constrained Rational Minimax Approximation)

Let $f(x)$ be continuous on $I_{a,b}$ and $f(x) \in C^{k-1}$ at $x = a$, and $w(x)$ a function such that $w(x) \neq 0$ in $I_{a,b}$, then

(a) There exists $r^*_{k,m,n}(x) \in R_{k,m,n}(f)$, such that

$$\rho_{R_{k,m,n}}(f) = ||f(x)-w(x) \cdot r^*_{k,m,n}(x)||,$$

i.e. $f(x)$ possesses a best order-constrained rational Chebyshev approximation in $R_{k,m,n}(f)$.

(b) $r^*_{k,m,n}(x)$ is completely characterised by the fact that $f(x)-r^*_{k,m,n}(x)$ alternates at least $m(\bar{a}^*)-k$ times where $m(\bar{a}^*)$ is the degree of $r^*_{k,m,n}(x)$ and $\bar{a}^*$ its vector of coefficients, $\bar{a}^* = \bar{a}(r^*_{k,m,n})$.

Proof [35].

CHAPTER 3

CURRENT STIFF SYSTEM SOLVERS

Presently, there is a large variety of numerical methods designed purposely for the solution of stiff systems. Table 3.0.1 gives a panorama of the situation, the discussion of which will be the essence of this chapter.

TABLE 3.0.1 Stiff System Solvers

- Linear Multistep Methods
  - No classical explicit A-stable methods [14]
  - Classical implicit methods of order ≤ 2 [14]
  - Stiffly stable methods [17,20]
  - Composite multistep methods [51]
  - Generalised multistep methods [27,33,36]
- One-Step Methods
  - Implicit methods
    - Runge-Kutta processes [10,18]
    - Methods using higher derivatives [15,18,22,23,38,40,44,46 p.212]
  - Explicit methods
    - Generalised Runge-Kutta processes [30]
    - Methods using the Jacobian matrices [7,21,49,50 p.180]
    - Others [6,26,37,45]

3.1 Linear Multistep Method

Definition 3.1.1

A linear multistep method is one of the form

$$y_{n+1} = \sum_{i=0}^{k} a_i y_{n-i} + h \sum_{i=-1}^{k} b_i y'_{n-i}, \quad |a_k| + |b_k| \neq 0, \tag{3.1.1}$$

applied to the solution of the initial value problem

$$y' = f(x,y), \quad y(0) = y_0. \tag{3.1.2}$$

Our discussion in this section will also be valid if (3.1.2) is a vector initial value problem, $\bar{y} = f(x,\bar{y})$, $\bar{y}(0) = \bar{y}_0$.

Definition 3.1.2 (Dahlquist [14])

An A-stable k-step method is a method of the form (3.1.1) which, when applied with fixed positive h to any differential equation of the form

$$y' = qy, \quad \text{Re}(q) < 0, \tag{3.1.3}$$

always gives solution that tends to zero as $n \to \infty$.

Definition 3.1.3

The region of stability of (3.1.1), when applied to any initial value problem of form (3.1.3), is that region S in the left half complex-plane such that if qh is in S, then all solutions of (3.1.1) tend to zero as $n \to \infty$.

Definitions (3.1.1) - (3.1.3) imply:

(A) The region of stability of an A-stable method is the whole left complex-plane;

(B) Using an A-stable linear multistep method to solve the system

$$\bar{y}' = A\bar{y},$$
$$\bar{y}(0) = \bar{y}_0,$$
$$\mathrm{Re}[\lambda_i(A)] < 0, \qquad (3.1.4)$$

for all eigenvalues $\lambda_i(A)$ of A, we do not have to care about the stability problems when choosing the step-size h. The sole consideration will be that of the truncation error.

However, in many cases, k-step methods are not a satisfactory tool for the solution of stiff systems because of the following results of Dahlquist [14] :

Theorem 3.1.1

An explicit k-step method ($b_{-1} = 0$ in (3.1.1)) cannot be A-stable.

Theorem 3.1.2

The order of an implicit k-step A-stable method cannot be greater than two.

Theorem 3.1.3

The trapezoidal rule has the smallest error coefficient of all A-stable linear multistep methods.

There are at least three different ways to overcome the order restriction imposed by Theorem 3.1.2, namely, "stiffly stable k-step methods", "composite multistep methods", and "generalised multistep methods".

Stiffly stable multistep methods

These were introduced by Gear [20,17]. They are methods of type (3.1.1). Figure 3.1.1 shows the typical appearance of their regions of stability. They are not strictly A-stable according to definition (3.1.2) since the region does not include the whole left complex-plane. However, only a small portion is left out. And at the expense of this small area, such methods can achieve order higher than two. If the b_i's in (3.1.1) all vanish, the maximum possible order is conjectured to be six [17], otherwise it is guessed to be eleven [24].

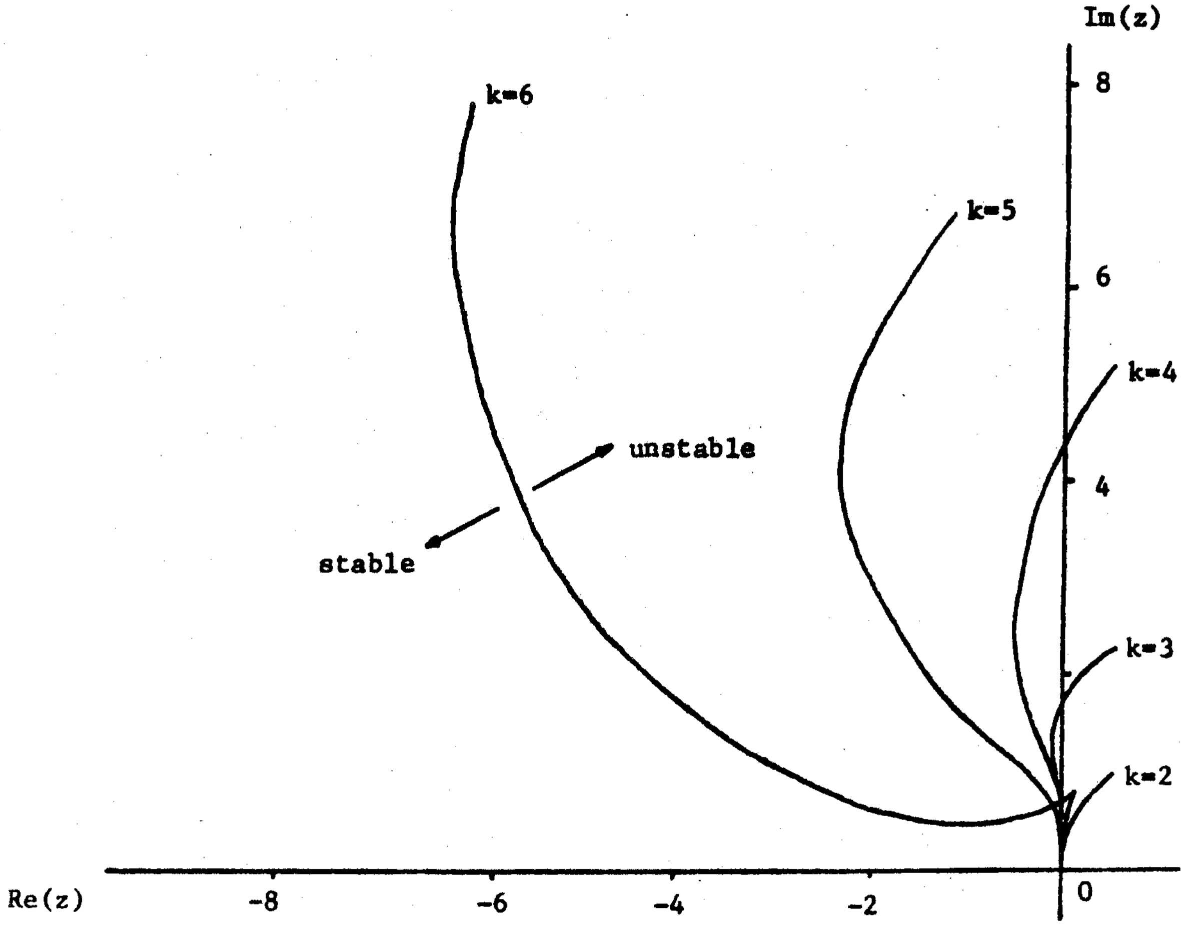

Fig.3.1.1 Stability regions for Gear's stiffly stable k-step methods

Composite multistep methods

These methods were proposed recently by Sloath and Bickart [51]. Instead of using only one equation of the form (3.1.1), these methods employ a set of p multistep formulae,

$$\sum_{j=-k}^{p-1} a_{ij}y_{mn+j} - h\sum_{j=-k}^{p-1} b_{ij}y'_{mn+j} = 0,\ i = 1,2,3,\dots,p. \qquad (3.1.5)$$

With the help of k past known values y_{mn+j}, $j = -k,\dots,-1$, a set of p new values y_{mn+j}, $j = 0,1,\dots,p-1$ are solved from (3.1.5). The first m values of this set, namely y_{mn+j}, $j = 0,\dots,m-1$, together with any other past values that may be necessary, are then used for the next iteration while the rest, y_{mn+j}, $j = m,\dots,p-1$ are discarded. It can be shown that if each formula used in (3.1.5) is of at least order r, then the whole method will be of order r too.

Generalised multistep methods

These formulae are studied by Lawson in [36]. They may be considered special cases of the methods discussed by Lambert and Sigurdson [27].

Let (3.1.2) be a vector initial value problem of dimension s, and A a square matrix of the same dimension. Applying a transformation

$$z(x) = \exp(-xA)y(x), \qquad (3.1.6)$$

to (3.1.2) will give us

$$\begin{aligned} z'(x) &= g(x,z) \\ &\equiv \exp(-xA)\{f[x,\exp(xA)\cdot z(x)]-A\cdot\exp(xA)\cdot z(x)\}, \\ z(0) &= y_0. \end{aligned} \qquad (3.1.7)$$

Integrating (3.1.7), we have

$$y(x+h) = \exp(hA)\cdot y(x) + h\int_0^1 \exp[(1-\tau)hA]\cdot u(x+\tau h)d\tau, \qquad (3.1.8)$$

where $u(x) = f(x,y(x))-A\cdot y(x).$ (3.1.9)

In [33], Lawson suggested the use of a quadrature formula to evaluate the integral in (3.1.8), such that, we have the approximation

$$y_{n+1} = E(hA)\cdot y_n + h \sum_{i=1}^{k} w_i(hA)u(x_n+\alpha_i h), \qquad (3.1.10)$$

where $w_i(hA)$ are the weight matrices to absorb the matrix weight function $\exp[(1-\tau)hA]$ in (3.1.8), α_i the distinct abscissae of the quadrature formula, and $E(hA)$ an approximation to $\exp(hA)$.

With the motivation to make (3.1.10) exact for vector polynomials $u(x)$, we require

$$\sum_{i=1}^{k} \bar{w}_i(hA)(\alpha_i)^j = \int_0^1 \exp[(1-\tau)hA]\tau^j d\tau \equiv \bar{M}_j(hA), \qquad (3.1.11)$$
$$j = 0,1,\ldots,k-1,$$

where $\bar{w}_i(hA)$ are the weight matrices in (3.1.10) when $E(hA) \equiv \exp(hA)$. A way to calculate $\bar{M}_j(hA)$ is

$$\bar{M}_0(hA) = (hA)^{-1}[\exp(hA)^{-1}],$$
$$\bar{M}_j(hA) = (hA)^{-1}[j\bar{M}_{j-1}(hA)-1], \quad j = 1,2,\ldots,k-1. \qquad (3.1.12)$$

If, instead of following (3.1.11), (3.1.12) exactly, we choose a basic rational approximation $E(hA)$ of order $(k+1)$ to $\exp(hA)$ and define

$$M_0(hA) = (hA)^{-1}[E(hA)-1],$$

$$M_j(hA) = (hA)^{-1}[jM_{j-1}(hA)-1],\ j = 2,\dots,k-1, \tag{3.1.13}$$

$$\sum_{i=1}^{k} w_i(hA)(\alpha_i)^j = M_j(hA),\quad j = 0,1,\dots,k-1, \tag{3.1.14}$$

putting back these to (3.1.10), we expect it will provide us with an approximate solution to (3.1.2). Moreover, we have the following three properties:

<u>Theorem 3.1.4</u>

If in (3.1.10), $E(hA)$ is an A-acceptable or L-acceptable approximation to $\exp(hA)$, and $w_i(hA)$ are calculated according to (3.1.13), (3.1.14), then the method is A-stable when applied to (3.1.4).

<u>Theorem 3.1.5</u>

If in (3.1.10), $E(hA)$ is an order $(k+1)$ approximation to $\exp(hA)$, and $w_i(hA)$ are calculated from (3.1.13), (3.1.14), then the method is exact for any particular integral of $y' = Ay + p(x)$, where $p(x)$ is an arbitrary vector polynomial of degree $k+1$ or less and A is a real non-singular square matrix.

<u>Theorem 3.1.6</u>

If (3.1.10) is A-stable and exact for the particular integral of $y' = By + p(x)$, where B is a stable square matrix of constants, then, regardless of step-size h,

$$\lim_{n\to\infty} [y_n - y(x_n)] = 0.$$

Proofs of the above three theorems can be found in [33].

If the abscissae $\{\alpha_i\}$ in (3.1.10) are chosen as follows:

$$\alpha_i = 2-i, \quad i = 1,2,\ldots,k, \qquad (3.1.15)$$

we have an implicit Adams-Moulton generalised multistep formula. If,

$$\alpha_i = 1-i, \quad i = 1,2,\ldots,k, \qquad (3.1.16)$$

we obtain an explicit Adams-Bashforth generalised multistep formula.

3.2 One-step Methods

Definition 3.2.1

A one-step method is a method of the form

$$y_{n+1} = \Phi(x_n,y_n,y_n',\ldots,y_n^{(i)};x_{n+1},y_{n+1},y_{n+1}',\ldots,y_{n+1}^{(j)}), \qquad (3.2.1)$$

for two non-negative integers i,j, applied for the solution of the scalar or vector initial value problem (3.1.2).

Definition 3.2.2

An explicit one-step method is a method of the form (3.2.1) in which Φ is independent of y_{n+1} and its derivatives. Otherwise, the method is implicit.

Definition 3.2.3

A one-step method is A-stable if, when applied to (3.1.3) with a fixed positive step-size h, it always gives solutions that approach zero as $n \to \infty$.

Another way to look at this definition is that, when the method is applied to (3.1.3), or (3.1.4), it can be represented in the form

$$y_{n+1} = E(qh)y_n, \quad (\text{or } \bar{y}_{n+1} = E(hA)\bar{y}_n) \qquad (3.2.2)$$

where E(qh) (or E(hA)) is an A-acceptable approximation to exp(qh) (or exp(hA)).

Definition 3.2.4

A one-step method is strongly A-stable when E(qh) in (3.2.2) is an L-acceptable approximation to exp(qh).

One-step A-stable methods can be classified as in Table 3.0.1 at the beginning of this chapter, and will be discussed in that same order here.

Implicit Runge-Kutta processes (IRK)

Runge-Kutta methods are of the form

$$y_{m+1} = y_m + h \sum_{i=1}^{v} w_i k_i , \qquad (3.2.3)$$

where

$$k_i = f(x_m + c_i h,\ y_m + \sum_{j=1}^{v} b_{ij} k_j),$$

$$c_i = \sum_{j=1}^{v} b_{ij}, \qquad (3.2.4)$$

w_i, b_{ij} being the coefficients of the process.

The process is [illegible] b_{ij} [illegible]

it is implicit [illegible]

[illegible]

[illegible]

and

$$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}). \tag{3.2.6}$$

These have errors of order $O(h^3)$ and $O(h^2)$ respectively. Their corresponding $E(qh)$ in (3.2.2) are in fact the $P_{1,1}(qh)$ and $P_{1,0}(qh)$ Padé approximation to exp(qh).

The highest attainable order of a v-stage IRK was found to be 2v by Butcher in [4]. He then gave four classes of IRK based on the Lobatto, Radau and Guassian quadrature formulae of arbitrary order. It can be verified that the corresponding E(qh) in (3.2.2) for the IRK based on the Guassian quadratures are in fact the diagonal entries in the Padé table of approximation to exp(qh). Noticing this fact, Ehle [18] started an investigation on the Padé table. His results are summarised in Theorem 2.1.1 and Theorem 2.1.2. These immediately establish the following two theorems.

Theorem 3.2.1

The IRK based on the Guassian quadrature and derived from the $P_{n,n}(qh)$ approximation to exp(qh) are A-stable.

Proof Ehle [18] or Wright [54].

Theorem 3.2.2

The IRK based on the Lobatto and Radau quadrature and derived from the $P_{n+1,n}(qh)$ and $P_{n+2,n}(qh)$ Padé approximation to exp(qh) are strongly A-stable.

Proof Chipman [10]. (These two classes of IRK are generated by Ehle in [18].)

These three classes of A-stable methods are of arbitrary order. Details of their implementation can be found in [10,18].

Recently, Lawson [32] pointed out two additional properties of these three classes of IRK exactly corresponding to Theorem 3.1.5 and Theorem 3.1.6 of the generalised multistep methods.

Implicit methods using higher derivatives

An obvious extension of (3.2.5) and (3.2.6) is the formula

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{2}[y_n'+y_{n+1}'] + \frac{h^2}{12}[y_n''-y_{n+1}''] \\ &= y_n + \frac{h}{2}[f(x_n,y_n) + f(x_{n+1},y_{n+1})] \\ &+ \frac{h^2}{12}[f'(x_n,y_n) - f'(x_{n+1},y_{n+1})]. \end{aligned} \tag{3.2.7}$$

This method is A-stable [46, p.212] and has an error of $O(h^5)$. However, when applied to (3.1.2), it requires the computation of the derivatives of $f(x,y)$.

Similar methods that involved such derivatives had been proposed by Davison [15], Makinson [40], Liniger and Willoughby [38], etc. They are all of the form

$$y_{n+1} = y_n + \sum_{i=1}^{v} \frac{h_i}{i!}(a_i y_{n+1}^{(i)} + b_i y_n^{(i)}). \tag{3.2.8}$$

They differ from each other by having truncation errors of different magnitude in h, and in that their corresponding E(qh) in (3.2.2) approaches exp(qh) differently as $h \to \infty$. Some of them are designed for differential equations of a particular form. Table 3.2.1 summarises their main properties.

Method	$O(h^n)$,n=	Derivatives used, v=	$\lim_{h\to\infty} E(qh)$
(3.2.7)	5	2	1
[15] Davison	5	3	1
[40] Makinson (a)	4	2	$1-\sqrt{3}$
" (b)	5	3	-0.635
[38] Liniger &(a)	2	1	determined by users
Willoughby(b)	3	2	
(c)	4	2	

TABLE 3.2.1 Properties of some A-stable implicit one-step methods involving higher derivatives

One interesting property of [38] is that the three classes of methods in it have parameters left to be chosen by the users to take care of those extreme eigenvalues in their stiff systems that are to be solved. The three classes are:

$$\text{(a)} \quad y_{n+1} = y_n + h[(1-a)y'_{n+1} + ay'_n] + O(h^2), \quad a \leq 1/2; \qquad (3.2.9)$$

$$\text{(b)} \quad y_{n+1} = y_n + \frac{h}{2}[(1+a)y'_{n+1}+(1-a)y'_n] + \frac{h^2}{4}[(b+a)y''_{n+1}-(b-a)y''_n], \quad a \geq 0,\ b \geq 0; \qquad (3.2.10)$$

$$\text{(c)} \quad y_{n+1} = y_n + \frac{h}{2}[(1+a)y'_{n+1} + (1-a)y'_n] + \frac{h^2}{12}[(1+3a)y''_{n+1} - (1-3a)y''_n], \quad a \geq 0. \qquad (3.2.11)$$

The restrictions on the parameters a and b are for the A-stability of the method.

If in (3.2.8),

$$b_i = (-1)^{i+1}a_i ,$$

we get a method first given by Hermite [22] (see also [23,44]), of the form

$$y_{n+1} = y_n + \sum_{i=1}^{v} a_{i,v}h^i(y_n^{(i)}-(-1)^i y_{n+1}^{(i)}). \qquad (3.2.12)$$

If $a_{i,v}$ is the i-th coefficient in the numerator of the v-th diagonal Padé approximation $P_{v,v}(z)$ to exp(z), then the corresponding E(qh) in (3.2.2) is also $P_{v,v}(qh)$. This means (3.2.12) is then a method of order (2v+1) and, by Theorem 2.1.1, A-stable for all v. In fact, this is a special case of the following more general result of Hermite [22], Obrechkoff [44], and Hummel and Seebeck [23].

Theorem 3.2.3

Let y(t) be continuously differentiable at least (j+k+1) times in [a,x]. Denote

$$C_q^p = \frac{p!}{q!(p-q)!} \quad , \quad \text{if } q \le p,$$

$$= 0 \qquad\qquad , \quad \text{if } q > p.$$

Then

$$y(x) = y(a) + \sum_{m=1}^{j} \frac{(k+j-m)!}{(k+j)!}[C_m^k\, y^{(m)}(a)$$

$$- (-1)^m C_m^j\, y^{(m)}(x)](x-a)^m + R, \qquad (3.2.13)$$

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

When $a = x_n$, $x = x_{n+1} = x_{n+h}$, (3.2.13) reduces to a method of the form (3.2.12). Furthermore, for the three cases, $j = k+i$, $i = 0,1,2$, the corresponding $E(qh)$ in (3.2.2) of the three methods are respectively the $P_{k,k}$, $P_{k+1,k}$, $P_{k+2,k}$, $k = 0,1,2,\ldots$ Padé approximations to exp(qh), Again, with the help of Theorems 2.1.1 and 2.1.2, we have (Ehle [18]),

Theorem 3.2.4

The quadrature formulae based on the derivatives given by (3.2.13) for $j = k,k+1,k+2$, $k = 0,1,2,\ldots$ respectively are all A-stable. In addition, the last two cases are strongly A-stable.

Explicit one-step methods

Various authors had produced explicit one-step methods that are A-stable, notably, Lawson's generalised Runge-Kutta process [30] and the method of Rosenbrook [49,50 p.180] and Calahan [7], with modifications by Haines [21].

The generalised Runge-Kutta methods produced by Lawson use an approach very similar to that of the generalised multistep methods previously described in section 3.1. By the same transformation (3.1.6), we have the analogue of (3.1.8):

$$y(x+h) = \exp(hA)y(x) + \int_0^h \exp[(h-\theta)A]\{f[x+\theta,y(x+\theta)]-A\cdot y(x+\theta)\}d\theta \qquad (3.2.14)$$

Instead of using a quadrature rule, the integral on the right is to be evaluated by a modified m-stage Runge-Kutta process as follows:

$$\begin{aligned} k_1^* &= f(x_n,y_n) - A\cdot y_n, \\ p_i^* &= \exp(c_i h\cdot A)\cdot y_n + h\sum_{j=1}^{i-1} a_{ij}\exp[(c_i-c_j)hA]k_j^*, \\ k_i^* &= f(x_n+c_i h,p_i^*) - A\cdot p_i^*, \quad i = 2,3,\ldots,m, \\ y_{n+1} &= \exp(hA)\cdot y_n + h\sum_{i=1}^{m} b_i\exp[(1-c_i)hA]\cdot k_i^*, \; n = 0,1,2,\ldots \end{aligned} \qquad (3.2.15)$$

The method would typically be applied to problems of the form,

$$\begin{aligned} y' &= By + u(t,y), \\ y(0) &= y_0, \end{aligned} \qquad (3.2.16)$$

where B is stable and stiff. If A = B and $u \equiv 0$, (3.2.15) gives

$$y_{n+1} = [\exp(h\cdot B)]^{n+1}y_0. \qquad (3.2.17)$$

This means that if an A-acceptable or L-acceptable approximation to exp(z) is used in (3.2.15), we have an A-stable process. Hence, by means of a transformation, a Runge-Kutta process can be made A-stable. It should be noted that the generalised Runge-Kutta processes (3.2.15) do not have the exactness properties analogous to Theorems 3.1.5 and 3.1.6. In [34], they are modified so as to induce this exactness.

The method of Rosenbrook with modifications by Haines uses a somewhat different technique. It requires not only values of f(x,y), but also of the Jacobian matrix $(\frac{\partial f}{\partial y})$ at each step in the process. It is A-stable, with error of $O(h^4)$ and $E(qh)$ approaches -0.8 as $h \to \infty$.

There are many other similar methods proposed by, among others, Pope [45], Kuo [26], Legras [37] and Calahan [6]. They are all designed for the particular problem of (3.2.16) and use the fact that its exact solution is

$$y(x) = \exp(Ax)\cdot y(0) + \int_0^x \exp[(x-t)\cdot A]u(t)dt. \qquad (3.2.18)$$

The difference is in their methods of evaluating the integral on the right. The resulting methods would be A-stable if A-acceptable approximation to the exponential is used.

*   *   *   *

As the discussions of section 3.1 and 3.2 indicate, there is a very intimate relationship between a numerical method of solving stiff equations and a numerical approximation to the exponential function. In particular, Gear had shown [16] the following property in the family of the Runge-Kutta processes with respect to rational exponential approximation :

Theorem 3.2.5

Let R(z) be any rational approximation, of the form

$$R(z) = \frac{P(z)}{Q(z)} ,$$

$$\text{degree of } P(z) \le \text{degree of } Q(z) \le n, \qquad (3.2.19)$$

to exp(z). If R(z) has distinct non-zero poles, there exists a n-stage Runge-Kutta process whose corresponding E(qh) in (3.2.2) is R(qh).

This means any A-acceptable rational approximation to exp(z) will give us an A-stable Runge-Kutta process. However, the resulting process of Theorem 3.2.5 is generally of a much lower order than the order of the rational approximation.

3.3 Methods for Partial Differential Equations

As mentioned in section 1.3, by means of semi-discretization, many linear parabolic partial differential equations (PDE) of type (1.4.1) commonly found in problems of heat conduction, fluid flow etc. can be transformed to a system of ordinary differential equations (ODE) having the form

$$\bar{y}'(t) = A\bar{y}(t) + \bar{f}(t),$$

$$\bar{y}(0) = \bar{y}_0, \qquad (3.3.1)$$

where A is a constant square matrix and f(t) is a forcing term from time-dependent boundary conditions or source and sink terms. In addition, A often has only real eigenvalues, and is stable, stiff and sparse.

A well-known numerical method for (3.3.1) is the Crank-Nicolson method

$$\frac{\bar{y}(t+h)-\bar{y}(t)}{h} = A\left[\frac{\bar{y}(t+h)+\bar{y}(t)}{2}\right] + \frac{\bar{f}(t+h)+\bar{f}(t)}{2}. \qquad (3.3.2)$$

Solving for $\bar{y}(t+h)$ in terms of $\bar{y}(t)$ and $\bar{f}(t)$, it reduces to

$$\begin{aligned}\bar{y}(t+h) &= (I-\tfrac{1}{2}hA)^{-1}(I+\tfrac{1}{2}hA)\bar{y}(t)\\ &\quad + h(I-\tfrac{1}{2}hA)^{-1}\left[\frac{\bar{f}(t+h)+\bar{f}(t)}{2}\right]. \qquad (3.3.3)\end{aligned}$$

We notice immediately this corresponds to the Padé approximation $P_{1,1}(z) = (1-x/2)/(1+x/2)$ to the exponential function $\exp(-x)$.

If $\bar{f}(t)$ of (3.3.1) is a polynomial of fixed degree in t, then the generalised multistep methods in section 3.1 [(3.1.10), (3.1.13)-(3.1.14)] with an A-acceptable exponential approximation, or the implicit Runge-Kutta processes based on the Padé approximation $P_{n,n}(z)$, $P_{n+1,n}(z)$, $P_{n+2,n}(z)$, $n = 0,1,\ldots,$ described in section 3.2 will give asymptotically exact solutions to (3.3.1). These methods exactly reproduce the particular integral and suppress the complementary function. Hence, the familiar Crank-Nicolson method (3.3.3) gives exact solution of the particular integral for linear f(t).

If $\bar{f}(t)$ is a constant and A has only real eigenvalues, the exact solution of (3.3.1) can be verified to be

$$\bar{y}(t+h) = \exp(hA)\cdot\bar{y}(t) + (I-\exp(hA))A^{-1}\bar{f}. \qquad (3.3.4)$$

Let us also define

$$r_{p,q}^{(m)}(z) = \sup_{-z\leq x\leq 0} |\exp(x)-P_{p,q}^{m}(\tfrac{x}{m})|, \tag{3.3.5}$$

where $P_{p,q}(z)$ is any Padé approximation to exp(z). Then we note the following result of Varga [52].

Theorem 3.3.1

If $\bar{w}_{p,q}(t)$ is the computed solution obtained from (3.3.4) using a Padé approximation $P_{p,q}(z)$ to exp(z), then

$$||\bar{y}(mh)-\bar{w}_{p,q}(mh)|| \leq r_{p,q}^{(m)}(mh\lambda(A))||\bar{y}_0-A^{-1}f||, \tag{3.3.6}$$

for $p = 0,1,2,\ldots,$

$q = 0,1,2,\ldots,$

$m = 1,2,\ldots,$

$t > 0,$

and $\lambda(A) = \max_i |\lambda_i(A)|,$

$\lambda_i(A)$ being the eigenvalues of A, and $||\cdot||$ denoting the Euclidean norm of a vector.

Theorem 3.3.1 is valid for all entries in the Padé approximation table. However, the method will be A-stable for matrices A with real eigenvalues only if $p \geq q$.

(3.3.6) implies that a computed answer **will** be accurate if $r_{p,q}^{(m)}$ $(mh\lambda(A))$ is sufficiently small. This, in turn, will mean the necessity of a small step-size h. But, by using a global minimax Chebyshev approximation in place of a Padé approximation to exp(z) in (3.3.4), we expect to be able to bound the error for any step-size h. This is made possible by the next theorem.

Theorem 3.3.2

The exact solution $\bar{y}(t)$ and the computed solution $\bar{w}_p(t)$ of (3.3.4), when the minimax Chebyshev approximation to exp(-x),

$$Q_p(x) = \frac{n_p(x)}{d_p(x)} ,$$

is used, have the following properties:

$$||\bar{y}(t)-\bar{w}_p(t)|| \le H_p||\bar{y}_0-A^{-1}\bar{f}||, \quad t > 0, \tag{3.3.7}$$

where $H_p = \rho_s(\exp(-x))$ (definition 2.2.2),

$S = \{r_{p,p}(x)$, rational functions whose numerator and denominator polynomials are of degree p or less$\}$.

Since H_p is usually of much smaller magnitude than $r_{p,q}^{(m)}$ for large z, using the minimax approximation would be "globally" superior to using the Padé approximation. This means, if $||\bar{y}_0-A^{-1}\bar{f}||$ is not too unreasonably large, we can still have a small error bound in (3.3.7) for any step-size h. Otherwise, Padé approximation with a small step-size would be necessary in order to have the error under control.

This leads to the study of order-constrained minimax approximation by Lawson [35]. It is a compromise between Padé approximation and minimax approximation. The result is in Theorem 2.2.2.

Although order-constrained approximation does not have as good a "global" error control as the true minimax approximation, it has the following two advantages:

(a) It provides a "best" uniform approximation to exp(-x) on $[0,\infty)$ with prescribed order at x = 0. If $||\bar{y}_0 - A^{-1}\bar{f}||$ of (3.3.7) is too large, a sufficient small step-size h will ensure the truncation error is bounded. Otherwise, we can use any step-size and are still guaranteed a global error bound.

(b) True minimax approximations are of order zero. Hence, they are not applicable to Theorems 3.1.5, 3.1.6 and their analogues in the case of one-step methods. Order-constrained approximations provide a remedy for this.

Implementation techniques for solving PDE of the type discussed in this section can be found in [9,12].

Lastly, it is worthwhile to mention that the 'Box Scheme' for parabolic mixed initial-boundary value problems in one space dimension described in [25] is also A-stable.

CHAPTER 4

A NEW APPROXIMATION TO THE EXPONENTIAL FUNCTION

In Chapter 3, it is shown that many of the methods for solving stiff systems are closely related to the numerical approximation of the exponential function. In this chapter, a new class of exponential approximations is introduced, which will have certain practical computational advantages when applied to the numerical methods of the previous chapter.

4.1 Definitions and Notations

We shall assume, in addition to those at the beginning of section 2.2, the following notation and definitions:

(4.1.1) m - an integer greater than -2, $m \in \{-1,0,1,2,\ldots\}$.

(4.1.2) n - a positive integer, $n \in \{1,2,3,\ldots\}$.

(4.1.3) b, b_i - some positive real numbers, $b, b_i > 0$.

(4.1.4) I_0 - the non-negative real axis, $I_0 = [0,\infty)$.

(4.1.5) $p_m(x)$ - a polynomial of degree m or less.

(4.1.6) $p_{-1}(x)$ - the constant function $\equiv 1$.

(4.1.7) $r_m^n(x)$ - a rational function of the form

$$r_m^n(x) = \frac{p_m(x)}{\prod\limits_{i=1}^{n} (1+b_i x)}, \quad x \in I_0.$$

(4.1.8) $\underline{r}_m^n(b,x)$- a rational function in x for a certain fixed b of the form

$$\underline{r}_m^n(b,x) = \frac{p_m(x)}{(1+bx)^n}, \; x \in I_0.$$

(4.1.9) R^n_m - the set of all rational functions $r^k_m(x)$

$$R^n_m = \{\frac{p_m(x)}{\prod_{i=1}^{k}(1+b_i x)} : x \in I_0,\ b_i > 0,\ n \geq k \geq \deg(p_m),\ \deg(p_m) \leq m\}.$$

(4.1.10) $\underline{R}^n_m$ - the set of all rational functions $\underline{r}^k_m(b,x)$

$$\underline{R}^n_m = \{\frac{p_m(x)}{(1+bx)^k} : x \in I_0,\ b > 0,\ n \geq k \geq \deg(p_m),\ \deg(p_m) \leq m\}.$$

(4.1.11) R^n_{-1} - $\{\frac{1}{\prod_{i=1}^{n}(1+b_i x)} : x \in I_0,\ b_i > 0,\ n > 0\}$.

(4.1.12) $\underline{R}^n_{-1}$ - $\{\frac{1}{(1+bx)^n} : x \in I_0,\ b > 0,\ n > 0\}$.

Notice the slight difference in the definitions (4.1.9) and (4.1.11). In (4.1.9), the degree of the denominator can be less than n, while in (4.1.11), we include only those that are of degree n. This same difference occurs between (4.1.10) and (4.1.12).

Our interest is in the approximation to exp(-x) on the non-negative real axis and hence all discussions will be confined to I_0, unless the contrary is explicitly stated. We shall start the investigation with approximations in the special form

$$r^n_{-1}(x) = \frac{1}{\prod_{i=1}^{n}(1+b_i x)} \quad \text{in } R^n_{-1}.$$

Since n = 1 implies the rational function has a linear denominator, the discussion reduces to a particular case of the usual Chebyshev rational approximation. We shall assume n > 1 in what follows.

4.2 Exponential Approximation in R^n_{-1}

In this section we shall try to establish the existence, uniqueness and characterisation of a best approximation in R^n_{-1} to exp(-x). In other words, we attempt to find a $r^*(x) \epsilon R^n_{-1}$, such that, (see Definition 2.2.2) ,

$$\rho_{R^n_{-1}}(f) = ||r^*(x) - f(x)||,$$
$$f(x) = \exp(-x),$$

to show it is unique and to find a characterisation by which it can be determined. Since the parameters $\{b_i\}$ of a function element $r^n_{-1}(x)$ in R^n_{-1} are nonlinearly related, the problem belongs to the class of nonlinear approximation. It is very unfortunate that $r^n_{-1}(x)$ is of such a particular form that many of the classical results of nonlinear approximations are not applicable to it. Hence, most of the results here have to be established independently.

The first two lemmas establish two basic important properties of the function g(x),

$$g(x) = \exp(x) - \prod_{i=1}^{n}(1+b_i x). \qquad (4.2.1)$$

Lemma 4.2.1

If $\sum_{i=1}^{n} b_i \leq 1$ then

(a) $g(0) = 0$,

(b) $g(x) > 0$, for $x > 0$.

Proof

(a) Obvious.

(b) Compare the Maclaurin series of exp(x) and the multinomial expansion of $\prod_{i=1}^{n} (1+b_i x)$,

$$\exp(x) = 1 + x + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!} + \ldots, \qquad (4.2.2)$$

$$\prod_{i=1}^{n} (1+b_i x) = 1+x \sum_{i} b_i+x^2 \sum_{\substack{i,j\\ i<j}} b_i b_j+x^3 \sum_{\substack{i,j,k\\ i<j<k}} b_i b_j b_k+\ldots + x^n b_1 b_2 \ldots b_n. \qquad (4.2.3)$$

Since $0 < \sum_{i} b_i \leq 1$,

we have, for any positive integer k, $k \leq n$,

$$0 < (\sum_{i=1}^{n} b_i)^k$$

$$= \sum_{n_1+n_2+\ldots+n_n=k} \frac{k!}{n_1!n_2!\ldots n_n!} b_1^{n_1} b_2^{n_2} \ldots b_n^{n_n}$$

$$= (\sum_{1\leq i_1<\ldots<i_k\leq n} k! b_{i_1} b_{i_2} \ldots b_{i_k} + \text{some other positive terms}) \leq 1.$$

This means

$$0 < \sum_{1\le i_1<\ldots<i_k\le n} k!b_{i_1}b_{i_2}\ldots b_{i_k} < 1,$$

or

$$0 < \sum_{1\le i_1<\ldots<i_k\le n} b_{i_1}b_{i_2}\ldots b_{i_k} < \frac{1}{k!}.$$

Putting back to (4.2.3) gives, for $x > 0$,

$$\prod_{i=1}^{n}(1+b_i x) < \sum_{i=0}^{n}\frac{x^i}{i!} < \exp(x). \qquad \text{Q.E.D.}$$

<u>Lemma 4.2.2</u>

If $\sum_{i=1}^{n} b_i > 1$ then

(a) $g(0) = 0$,

(b) there exists exactly one other point $\alpha > 0$, such that $g(\alpha) = 0$,

(c) $g(x) > 0$ iff $x > \alpha$.

<u>Proof</u>

(a) Obvious.

(b) and (c).

For small x near 0, $g(x) < 0$ because

$$\prod_{i=1}^{n}(1+b_i x) = 1 + x\cdot\sum b_i + O(x^2)$$
$$> 1 + x + O(x^2) = \exp(x).$$

For very large x, obviously, $g(x) > 0$. This means $g(x)$ has at least one positive zero at $x = \alpha$, $\alpha > 0$.

Differentiating g(x) n+1 times gives, for $x > 0$,

$$g^{(n+1)}(x) = e^x > 0,$$

which shows, by repeated application of Rolle's theorem, that the number of positive roots of g(x) is finite, at most n+1. Hence, it is possible to find its smallest positive root. Let it be α and $g(x) < 0$, $0 < x < \alpha$. We shall try to prove that $g(x) > 0$ if $x > \alpha$. For any $x > \alpha > 0$, let $x = c\alpha$, $c > 1$. Since

$$g(\alpha) = 0, \text{ or}$$

$$\exp(\alpha) = \prod_{i=1}^{n} (1+b_i\alpha),$$

we have, $\exp(x) = \exp(c\alpha)$

$$= [\exp(\alpha)]^c$$

$$= [\prod_{i=1}^{n} (1+b_i\alpha)]^c$$

$$= \prod_{i=1}^{n} (1+b_i\alpha)^c \quad .$$

Expanding $(1+b_i\alpha)^c$ in Taylor's series, we get

$$\exp(x) = \prod_{i=1}^{n} (1+cb_i\alpha + \frac{c(c-1)}{2!}(1+\xi_i)^{c-2}(b_i\alpha)^2)$$

$$> \prod_{i=1}^{n} (1+b_ic\alpha) = \prod_{i=1}^{n} (1+b_ix), \quad 0 < \xi_i < \alpha.b_i \quad .$$

Therefore, $g(x) > 0$ if $x > \alpha$ and $g(x) < 0$ if $0 < x < \alpha$. Q.E.D.

Since

$$\frac{1}{\prod_{i=1}^{n}(1+b_ix)} - \exp(-x) = \frac{\exp(x) - \prod_{i=1}^{n}(1+b_ix)}{\exp(x)\prod_{i=1}^{n}(1+b_ix)}$$

$$= \frac{g(x)}{d(x)} ,$$

where $d(x) \geq 1$ for $x \geq 0$, the curve of any $r^n_{-1}(x)$ will be of either one of the two patterns in Figure 4.1.

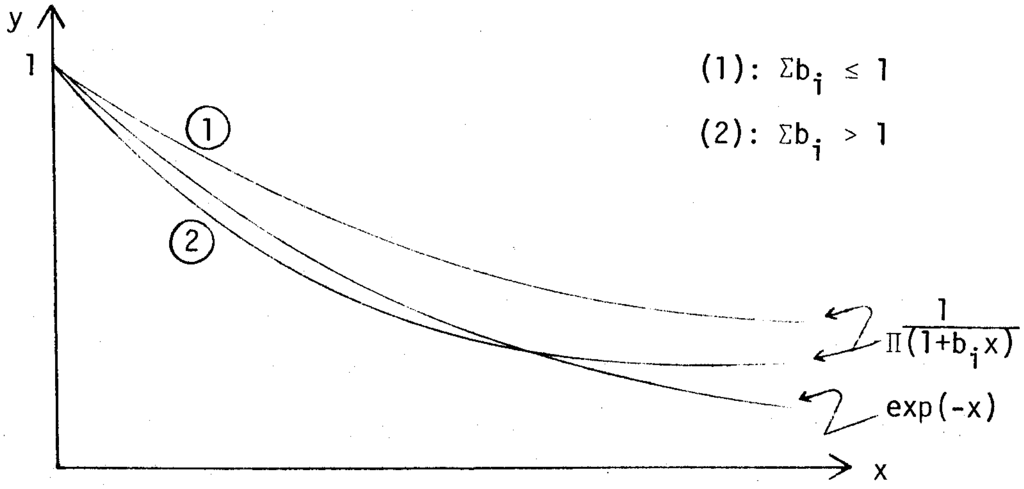

Fig.4.1 Graphs of exp(-x) and $\frac{1}{\Pi(1+b_ix)}$

The next lemma embodies a very important technique commonly used in approximation theory for establishing the necessary condition of a best approximation and will be used in subsequent discussions. Given an approximation r(x) to f(x), it provides a criterion to find, if possible, a $r^*(x)$ which will be a better approximation to f(x) than is r(x).

Lemma 4.2.3

If $r(x)$ approximates $f(x)$ on I_{ab} and $\{\xi_i\}$, $i = 1,\ldots,n$ is that sequence of points on I_{ab} at which the error curve attains its extrema alternately, i.e.

(a) $\xi_0 = a \leq \xi_1 < \xi_2 \ldots < \xi_n \leq b = \xi_{n+1}$,

(b) $\lambda_i = r(\xi_i) - f(\xi_i)$, $i = 1,2,\ldots,n$,

(c) $\lambda_{i-1} \cdot \lambda_i < 0$, $\quad i = 2,3,\ldots,n$,

(d) If $\lambda_i > (<) \; 0$, $\quad i = 1,2,\ldots,n$,

then $\lambda_i \geq (\leq) \; r(x) - f(x)$, $\xi_{i-1} \leq x \leq \xi_{i+1}$.

Let $r^*(x)$ be another function, such that

$$r^*(x) - r(x) = w(x) \prod_{i=1}^{n-1} (x-\beta_i), \tag{4.2.10}$$

where $\quad a = \beta_0 \leq \xi_1 < \beta_1 < \xi_2 < \beta_2 \ldots < \xi_{n-1} < \beta_{n-1} < \xi_n \leq \beta_n = b$,

and $w(x)$ is a continuous function satisfying

(a) $w(x) \neq 0$ for $x \in (a,b)$,

(b) that $r^*(x)-r(x)$ may vanish at an end point only if that end point is not an extremal point.

Then, if $w(x)$ can be made sufficiently small in I_{ab}, $r^*(x)$ is a better approximation to $f(x)$ than $r(x)$.

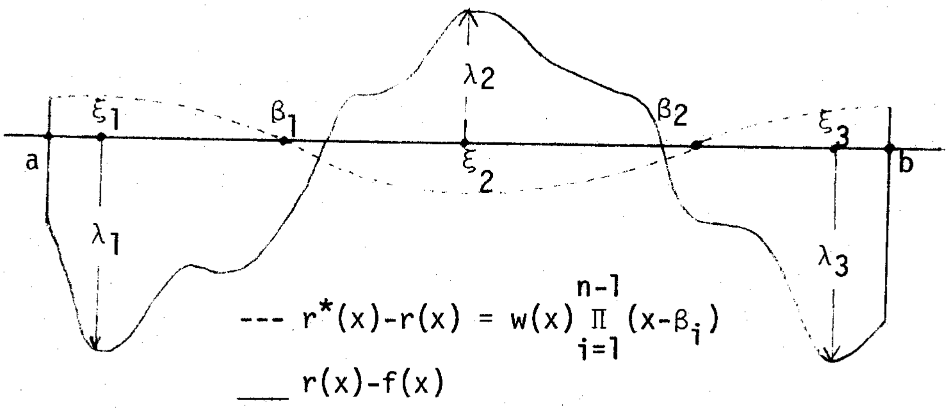

Fig. 4.2.1

Proof [1, p.55-56]

Pictorially, with reference to Fig.4.2.1, in each of the subintervals

$$[\beta_{i-1},\beta_i],\ i = 1,\dots,n,$$

in turn one of the following two inequalities is satisfied:

If $\lambda_i > 0$

$$\min(\lambda_{i-1},\lambda_{i+1}) + \alpha < r(x) - f(x) \leq \lambda_i.$$

If $\lambda_i < 0$

$$\max(\lambda_{i-1},\lambda_{i+1}) - \alpha > r(x) - f(x) \geq \lambda_i, \qquad (4.2.11)$$

where $\lambda_0 = \lambda_2$, $\lambda_{n+1} = \lambda_{n-1}$ and α is a certain small positive constant. Hence, if $w(x)$ is appropriately chosen, such that

$$|r^*(x)-r(x)| = |w(x) \prod_{i=1}^{n-1} (x-\beta_i)| \leq \alpha,$$

and the sign $r^*(x)-r(x)$ relative to $r(x)-f(x)$ is as illustrated in Fig.4.2.1, then $r^*(x)$ will be a better approximation than $r(x)$ as a consequence of the inequalities (4.2.11) and the equation

$$r^*(x) - f(x) = (r(x)-f(x)) + (r^*(x)-r(x)). \qquad \text{Q.E.D.}$$

We now continue to show a necessary condition for the best approximation to exp(-x) with respect to R^n_{-1}. This will then allow us to search for the best approximation within a much smaller subset of R^n_{-1}.

Lemma 4.2.4

If $r^n_{-1}(x) = \frac{1}{\prod_{i=1}^{n}(1+b_i x)}$ approximates exp(-x) best with reference to R^n_{-1}, then

(a) the error curve $r^n_{-1}(x)-\exp(-x)$ alternates exactly once,

(b) $b_1 = b_2 = \ldots = b_n = b$,

(c) $b > \frac{1}{n}$.

Proof

(a) If $r^n_{-1}(x)$ approximates exp(-x) best, then $\frac{1}{(1+b_1 x)}$ will also be the best weighted approximation to $\exp(-x)\cdot\prod_{i=2}^{n}(1+b_i x)$ with the weight function $\frac{1}{\prod_{i=2}^{n}(1+b_i x)}$. That is, the function $\frac{1}{1+b_1 x}$ should also minimize the weighted maximum norm

$$\left\| \frac{1}{\prod_{i=2}^{n}(1+b_i x)} \left((1+b_1 x)^{-1}-\exp(-x)\cdot\prod_{i=2}^{n}(1+b_i x)\right)\right\|$$

Classical Chebyshev rational approximation theory guarantees that the error curve alternates at least once. By Lemmas 4.2.1 and 4.2.2, we know the error curve cannot have more than one positive root. Therefore, it alternates exactly once.

(b) Suppose, that there exists a best approximation $r^*_{-1}(x)$, for which at least two of the b_i differ . Then, without loss of generality, let $b_1 < b_2$.

Let $\beta_3(x) = \prod_{i=3}^{n} (1+b_i x)$,

$$\beta_1(x) = (1+b_1x)(1+b_2x)\ \beta_3(x) = \frac{1}{r_{-1}^{n}(x)},$$

$$\beta_2(x) = (1+b_1'x)(1+b_2'x)\beta_3(x),$$

$$\beta_4(x) = \frac{\beta_1(x)\cdot\beta_2(x)}{\beta_3(x)},$$

where $b_1' = b_1 + \delta_1$,

$$b_2' = b_2 - \delta_2,$$

$$0 < \delta_1 < b_2 - b_1,$$

$$0 < \delta_2 = c\cdot\delta_1,\quad 1 < c < \frac{b_2}{b_1'},$$

and the point $d = \frac{c-1}{b_2-cb_1'}$ lies between the two extremal points of $r_{-1}^{n}(x)-\exp(-x)$.

The choice of c and d is possible because δ_1 is such that $\frac{b_2}{b_1'} > 1$ and, as c moves from 1 to $\frac{b_2}{b_1'}$, d moves from 0 to ∞. Now

$$\begin{aligned} d(x) &= \frac{1}{\beta_2(x)} - \frac{1}{\beta_1(x)} \\ &= \frac{1}{\beta_3(x)}\left[\frac{1}{(1+b_2'x)(1+b_1'x)} - \frac{1}{(1+b_2x)(1+b_1x)}\right] \\ &= \frac{1}{\beta_4(x)}[(1+b_2x)(1+b_1x)-(1+b_2x-\delta_2x)(1+b_1x+\delta_1x)] \\ &= \frac{x}{\beta_4(x)}[(\delta_2-\delta_1)+x(b_1\delta_2-b_2\delta_1+\delta_1\delta_2)]. \end{aligned}$$

If δ_1 is small, d(x) can be bounded sufficiently small. Also,

$$\delta_2-\delta_1 = \delta_1(c-1) > 0,$$

and $\quad b_1\delta_2 - b_2\delta_1 + \delta_1\delta_2$

$$= \delta_1[cb_1-b_2+c\delta_1]$$

$$= \delta_1[cb_1'-b_2] < 0$$

imply $d(x) > (<)\ 0$ for small (large) x and that $d(x)=0$ at $x=d$. Since, by Lemmas 4.2.1 and 4.2.2, the error curve $r^n_{-1}(x)-\exp(x)$ is negative for small x and positive for large x, $d(x)$ satisfies completely (4.2.10) of Lemma 4.2.3. Therefore, $\frac{1}{\beta_2(x)}$ is a better approximation to $\exp(-x)$ than $r^n_{-1}(x)$. Obviously, $\frac{1}{\beta_2(x)}$ is in R^n_{-1} too. We thus arrive at a contradiction.

(c) (b) implies $\sum_{i=1}^{n} b_i = nb$, and by Lemma 4.2.2,

$nb > 1$, i.e. $b > \frac{1}{n}$. Q.E.D.

The implication of Lemma 4.2.4 is that we can confine our search for the best approximation in R^n_{-1} to its subset $\underline{R}^n_{\underline{-1}}$,

$$\underline{R}^n_{\underline{-1}} = \{\underline{r}^n_{\underline{-1}}(b,x) = \frac{1}{(1+bx)^n}\}.$$

If a best approximation does exist in $\underline{R}^n_{\underline{-1}}$, it will be the best with respect to R^n_{-1} too.

Before we proceed on our search in $\underline{R}^n_{\underline{-1}}$, we need to define two functions $\lambda_1(b)$, $\lambda_2(b)$ as follows:

Consider, for any $b > 0$, the error function $e(b,x)$ of $\underline{r}^n_{\underline{-1}}(b,x)$,

$$e(b,x) = \underline{r}^n_{\underline{-1}}(b,x) - \exp(-x). \tag{4.2.13}$$

Since $\quad e(b,0) = 0,$

and $\quad \lim_{x\to\infty} e(b,x) = 0,$

$e(b,x)$ attains its minimum and maximum in $I_0 = [0,\infty)$, at say γ_1 and γ_2 respectively. Define:

$$\lambda_1(b) \equiv \min_{x\in I_0} e(b,x) = e(b,\gamma_1),$$

$$\lambda_2(b) \equiv \max_{x\in I_0} e(b,x) = e(b,\gamma_2). \tag{4.2.14}$$

If $nb \leq 1$, by Lemma 4.2.1,

$$0 = \gamma_1 < \gamma_2,$$

and $\quad \lambda_1(b) = 0 < \lambda_2(b).$ $\hfill$ (4.2.15)

If $nb > 1$, by Lemma 4.2.2,

$$0 < \gamma_1 < \gamma_2,$$

$$\lambda_1(b) < 0 < \lambda_2(b), \tag{4.2.16}$$

Since our interest is in the case when $e(b,x)$ alternates, we shall assume $nb > 1$.

Lemma 4.2.5

$\underline{r}^n_{\underline{-1}}(b^*,x)$ approximates $\exp(-x)$ best with respect to $\underline{R}^n_{\underline{-1}}$ iff $e(b^*,x)$ alternates exactly once.

Proof

(a) only if part

If $\underline{r}^n_{-1}(b^*,x)$ is the best in $\underline{R}^n_{-1}$, by Lemma 4.2.4, it is also the best in R^n_{-1} and hence $e(b^*,x)$ alternates exactly once.

(b) if part

If $e(b^*,x)$ alternates once, by Lemma 4.2.2, $nb > 1$ and (4.2.16) reduces to

$$0 < \gamma_1 < \gamma_2,$$

$$-e(b^*,\gamma_1) = -\lambda_1(b^*) = \lambda_2(b^*) = e(b^*,\gamma_2) > 0.$$

Consider any other $\underline{r}^n_{-1}(b,x)$, $b \neq b^*$.

If $b > b^*$, $\underline{r}^n_{-1}(b^*,x) - \underline{r}^n_{-1}(b,x) > 0, \quad x > 0,$

and
$$\begin{aligned}\lambda_1(b) &= \min e(b,x)\\ &= \min \{\underline{r}^n_{-1}(b,x)-\exp(-x)\}\\ &= \min \{[\underline{r}^n_{-1}(b^*,x)-\exp(-x)]-[\underline{r}^n_{-1}(b^*,x)-\underline{r}^n_{-1}(b,x)]\}\\ &\leq e(b^*,\gamma_1)-[\underline{r}^n_{-1}(b^*,\gamma_1)-\underline{r}^n_{-1}(b,\gamma_1)]\\ &< e(b^*,\gamma_1)\\ &= \lambda_1(b^*).\end{aligned}$$

Therefore,

$$\begin{aligned}||e(b,x)|| &= \max\{-\lambda_1(b),\lambda_2(b)\}\\ &> -\lambda_1(b^*)\\ &= ||e(b^*,x)||,\end{aligned}$$

and hence $\underline{r}^n_{-1}(b,x)$ is not as good an approximation to $\exp(-x)$ as $\underline{r}^n_{-1}(b^*,x)$. We get a similar result for any $b < b^*$. Therefore $\underline{r}^n_{-1}(b^*,x)$ is in fact the best approximation in $\underline{R}^n_{-1}$. Q.E.D.

Lemmas 4.2.6 and 4.2.7 establish the existence and uniqueness of the best approximation.

Lemma 4.2.6 (Existence lemma)

There exists a $b^* > 0$ such that $e(b^*,x)$ alternates exactly once.

Proof Consider the function

$$\lambda(b) = \lambda_1(b) + \lambda_2(b), \quad 0 < b < \infty.$$

Because, for any $b_2 \neq b_1$,

$$\begin{aligned} e(b_2,x) &= [\underline{r}^n_{-1}(b_1,x)-\exp(-x)]+[\underline{r}^n_{-1}(b_2,x)-\underline{r}^n_{-1}(b_1,x)] \\ &= e(b_1,x) + [\underline{r}^n_{-1}(b_2,x)-\underline{r}^n_{-1}(b_1,x)], \end{aligned} \tag{4.2.20}$$

As $b_2 \to b_1$, the second term of (4.2.20) tends to zero, so $e(b_2,x) \to e(b_1,x)$ and hence $e(b,x)$ is a function continuous with reference to b. Furthermore, all terms in (4.2.20) are bounded and it can easily be shown that $\lambda_1(b)$, $\lambda_2(b)$ are continuous functions with reference to b too. Now,

$$\begin{aligned} \lim_{b\to 0} \lambda(b) &= \lim_{b\to 0} \lambda_1(b) + \lim_{b\to 0} \lambda_2(b) \\ &= 0 + 1 = 1, \end{aligned}$$

and

$$\begin{aligned} \lim_{b\to\infty} \lambda(b) &= \lim_{b\to\infty} \lambda_1(b) + \lim_{b\to\infty} \lambda_2(b) \\ &= -1 + 0 = -1. \end{aligned}$$

That means there is a closed interval $[a,c]$, $c > a > 0$, such that $\lambda(b)$ is continuous in it and that

$$\lambda(a) > 0,$$

$$\lambda(c) < 0.$$

Clearly, we have a point $b^* > 0$ in the interval where

$$\lambda(b^*) = 0,$$

that is, $\lambda_2(b^*) = -\lambda_1(b^*)$.

At this point b^*, $e(b^*,x)$ alternates exactly once. Q.E.D.

Lemma 4.2.7 (Uniqueness lemma)

The best approximation to $\exp(-x)$ in $\underline{R}^n_{-1}$ is unique.

Proof If $\underline{r}^n_{-1}(b_1,x)$, $\underline{r}^n_{-1}(b_2,x)$ are two distinct best approximations to $\exp(-x)$, $b_1 \neq b_2$.

Let $\lambda_1(b_1) = e(b_1,\gamma_1) = \lambda_1(b_2)$,

$$\lambda_2(b_1) = e(b_1,\gamma_2) = \lambda_2(b_2).$$

Since $b_1 \neq b_2$,

$$\underline{r}^n_{-1}(b_1,\gamma_1)-\underline{r}^n_{-1}(b_2,\gamma_1)$$

$$= [\underline{r}^n_{-1}(b_1,\gamma_1)-\exp(-\gamma_1)]-[\underline{r}^n_{-1}(b_2,\gamma_1)-\exp(-\gamma_1)]$$

$$= \lambda_1(b_1)-[\underline{r}^n_{-1}(b_2,\gamma_1)-\exp(-\gamma_1)] < 0.$$

Similarly, $\underline{r}^n_{-1}(b_1,\gamma_2) - \underline{r}^n_{-1}(b_2,\gamma_2) > 0.$

Hence $\underline{r}^n_{-1}(b_1,x) - \underline{r}^n_{-1}(b_2,x)$

has a positive zero $x^* > 0$. But this is obviously a contradiction because its only zero is at $x = 0$. Q.E.D.

We summarise this section by the following theorem:

Theorem 4.2.1

Let R^n_{-1} be the set of all rational functions of the form

$$r^n_{-1}(x) = \frac{1}{\prod\limits_{i=1}^{n} (1+b_i x)}$$

where $b_i > 0$ and n is a positive integer, then

(a) there exists a best approximation to exp(-x) in R^n_{-1}.

(b) the best approximation is completely characterised by the facts that

(i) $b_1 = b_2 = \ldots = b_n$

(ii) the error curve $r^n_{-1}(x)$-exp(-x) alternates exactly once.

(c) the best approximation is unique.

Proof

(a) Lemmas 4.2.5 and 4.2.6 show that a best approximation exists in $\underline{R}^n_{-1}$. By Lemma 4.2.4, this is also the best approximation with reference to R^n_{-1}.

(b) Lemma 4.2.5.

(c) Lemma 4.2.7.

4.3 Error Estimation and Asymptotic Behaviour

In this section we investigate the quality of the approximation in R^n_{-1} to exp(-x) as a function of n. Because of Theorem 4.2.1, we shall denote, for any positive integer n, the best approximation to exp(-x) with reference to R^n_{-1} by

$$r^*_{-1}(b^*_n,x) = \frac{1}{(1+b^*_n x)^n} \ . \tag{4.3.1}$$

And the error curve

$$e(b^*_n,x) = r^*_{-1}(b^*_n,x)-\exp(-x) \tag{4.3.2}$$

has the two extremal points $\gamma_{n,1},\gamma_{n,2}$ such that

$$\begin{aligned} &0 < \gamma_{n,1} < \gamma_{n,2}, \\ &\lambda_1(b^*_n) = e(b^*_n,\gamma_{n,1}) < 0, \\ &\lambda_2(b^*_n) = e(b^*_n,\gamma_{n,2}) > 0, \end{aligned} \tag{4.3.3}$$

where $\lambda_1(\cdot)$ and $\lambda_2(\cdot)$ have the same meaning as (4.2.14). Because of the alternating property, we also have

$$-\lambda_1(b^*_n) = \lambda_2(b^*_n) > 0, \tag{4.3.4}$$

and
$$\rho_{R^n_{-1}}(\exp(-x)) = |\lambda_1(b^*_n)| = |\lambda_2(b^*_n)|. \tag{4.3.5}$$

Theorem 4.3.1

For any positive integer n, $b^*_n > b^*_{n+1} > 0$.

Proof

By Lemma 4.2.4(c), $b_n^* > 0$, $b_{n+1}^* > 0$. To establish a contradiction argument, suppose that $b_{n+1}^* \geq b_n^*$. Then, for $x > 0$,

$$r_{-1}^*(b_{n+1}^*,x) < r_{-1}^*(b_n^*,x).$$

In particular,

$$\begin{aligned}\lambda_1(b_{n+1}^*) &\leq e(b_{n+1}^*,\gamma_{n,1})\\ &= e(b_n^*,\gamma_{n,1}) + r_{-1}^*(b_{n+1}^*,\gamma_{n,1}) - r_{-1}^*(b_n^*,\gamma_{n,1})\\ &< \lambda_1(b_n^*).\end{aligned}$$

Similarly,

$$\begin{aligned}\lambda_2(b_{n+1}^*) &= e(b_{n+1}^*,\gamma_{n+1,2})\\ &= e(b_n^*,\gamma_{n+1,2}) + r_{-1}^*(b_{n+1}^*,\gamma_{n+1,2}) - r_{-1}^*(b_n^*,\gamma_{n+1,2})\\ &< e(b_n^*,\gamma_{n+1,2}) \leq \lambda_2(b_n^*).\end{aligned}$$

Since, as in (4.3.4),

$$-\lambda_1(b_n^*) = \lambda_2(b_n^*) > 0,$$

we have

$$-\lambda_1(b_{n+1}^*) > \lambda_2(b_{n+1}^*),$$

which is a contradiction to the alternating property of $e(b_{n+1}^*,x)$. Q.E.D.

The first implication of Theorem 4.3.1 is that the sequence $\{b_i^*\}$, $i = 1,2,\ldots$ is strictly monotonic decreasing, bounded below by zero. Therefore, $\lim\limits_{i\to\infty} b_i^*$ exists.

Secondly, it is useful when we try to determine the best approximation numerically. Once we obtain the best approximation in R^n_{-1}, and hence b^*_n, we know, for the best approximation in R^{n+1}_{-1}, b^*_{n+1} will be in the region $(1/n, b^*_n)$, by Lemma 4.2.4.

The next theorem tells us that the minimax error is also strictly decreasing with respect to n.

Theorem 4.3.2

For any positive integer n and f(x) = exp(-x),

$$\rho_{R^n_{-1}}(f) > \rho_{R^{n+1}_{-1}}(f) > 0.$$

Proof

Obviously, $\rho_{R^n_{-1}}(f) > 0$ and $\rho_{R^{n+1}_{-1}}(f) > 0$. Otherwise, $r^*_{-1}(b^*_n,x)$ or $r^*_{-1}(b^*_{n+1},x)$ is identical with exp(-x) which is impossible.

If we rewrite $r^*_{-1}(b^*_n,x)$ in the form

$$r^*_{-1}(b^*_n,x) = \frac{1}{(1+\underline{b}x)(1+b^*_n x)^n} \quad ,$$

where $\underline{b} = 0$ and can be considered as an extra parameter, then, applying the same technique as in Lemma 4.2.4(b) by perturbing slightly $\underline{b}$ and one b^*_n, we obtain a new approximation

$$\bar{r}^n_{-1}(x) = \frac{1}{(1+\underline{b}'x)(1+\underline{b}^*_n\,x)(1+b^*_n x)^{n-1}} \quad ,$$

where $0 = \underline{b} < \underline{b}' < \underline{b}^*_n < b^*_n$ and

$$||r^*_{-1}(b^*_n,x)-\exp(-x)|| > ||\bar{r}^n_{-1}(x)-\exp(-x)||.$$

Since $\bar{r}^n_{-1}(x) \in R^{n+1}_{-1}$ we have

$$\rho_{R^{n+1}_{-1}} \leq ||\bar{r}^n_{-1}(x)-\exp(-x)|| < \rho_{R^n_{-1}} . \qquad \text{Q.E.D.}$$

By the above theorem, we know the minimax approximation in R^n_{-1} is always better than those in R^k_{-1}, $k < n$. That means asymptotically as $n \to \infty$, we shall get the very best approximation. And the case of degeneracy does not happen, unlike the situation in classical rational approximation theory. Theorem 4.3.3 is an analogue of the de la Valleé Poussin Theorem in classical approximation theory. It gives an upper and a lower bound of the minimax error.

Theorem 4.3.3

If $r^*_{-1}(b^*_n,x)$ is the minimax approximation in R^n_{-1}, $\underline{r}^n_{-1}(b,x)$ an arbitrary member of $\underline{R}^n_{-1}$, and $f(x) = \exp(-x)$, then

$$\min\{|\lambda_1(b)|,|\lambda_2(b)|\} \leq \rho_{R^n_{-1}}(f) \leq \max\{|\lambda_1(b)|,|\lambda_2(b)|\}, \qquad (4.3.3)$$

where equality is attained iff $b = b^*_n$.

Proof By definition,

$$\rho_{R^n_{-1}}(f) = -\lambda_1(b^*_n) = \lambda_2(b^*_n) > 0.$$

Obviously, if $b = b^*_n$, we have equality in (4.3.3).

If $b > b^*_n$, for $x > 0$,

$$r^*_{-1}(b^*_n,x) - \underline{r}^n_{-1}(b,x) = \frac{1}{(1+b^*_n x)^n} - \frac{1}{(1+bx)^n} > 0.$$

At $x = \gamma_{n,1}$, the first extremal point of $r^{*}_{-1}(b^{*}_{n},x)-f(x)$,

$$\begin{aligned} \lambda_1(b) &\le \underline{r}^{n}_{-1}(b,\gamma_{n,1}) - \exp(-\gamma_{n,1}) \\ &= [r^{*}_{-1}(b^{*}_{n},\gamma_{n,1})-\exp(-\gamma_{n,1})]-[r^{*}_{-1}(b^{*}_{n},\gamma_{n,1})-\underline{r}^{n}_{-1}(b_n,\gamma_{n,1})] \\ &< \lambda_1(b^{*}_{n}) < 0, \end{aligned}$$

which means

$$|\lambda_1(b^{*}_{n})| < |\lambda_1(b)| \le \max\{|\lambda_1(b)|,|\lambda_2(b)|\}.$$

Similarly, at $x = \gamma_2$, the second extremal point of $\underline{r}^{n}_{-1}(b,x) - f(x)$,

$$\begin{aligned} \lambda_2(b) &= \underline{r}^{n}_{-1}(b,\gamma_2)-\exp(\gamma_2) \\ &= [r^{*}_{-1}(b^{*}_{n},\gamma_2)-\exp(\gamma_2)]-[r^{*}_{-1}(b^{*}_{n},\gamma_2)-\underline{r}^{n}_{-1}(b_n,\gamma_2)] \\ &< \lambda_2(b^{*}_{n}), \end{aligned}$$

and hence

$$|\lambda_2(b^{*}_{n})| > |\lambda_2(b)| \ge \min\{|\lambda_1(b)|,|\lambda_2(b)|\}.$$

If $b < b^{*}_{n}$, we can similarly prove

$$|\lambda_2(b^{*}_{n})| < |\lambda_2(b)| \le \max\{|\lambda_1(b)|,|\lambda_2(b)|\},$$

$$|\lambda_1(b^{*}_{n})| > |\lambda_1(b)| \ge \min\{|\lambda_1(b)|,|\lambda_2(b)|\},$$

where $\lambda_1(b)$ may be zero if $nb \le 1$. Q.E.D.

Before we proceed to find the asymptotic bound for the minimax error, we need the following three lemmas.

Lemma 4.3.1

For $0 \leq x \leq a < \infty$, as $n \to \infty$,

$$(1 + \frac{x}{n})^n = (1 - \frac{x^2}{2n})e^x + 0(\frac{1}{n^2}).$$

Proof Since x is bounded, as $n \to \infty$,

$$\begin{aligned}(1+ \frac{x}{n})^n &= 1+n\cdot\frac{x}{n} + \frac{n(n-1)}{n^2} \cdot \frac{x^2}{2!} +\ldots+ \frac{n!}{n^n} \cdot \frac{x^n}{n!} \\ &= [1+x+ \frac{x^2}{2!} +\ldots+ \frac{x^n}{n!}] \\ &\quad - \frac{1}{n}[\frac{x^2}{2!} + \frac{3.2}{2} \cdot\frac{x^3}{3!} +\ldots+ \frac{n(n-1)}{2}\cdot\frac{x^n}{n!}]+0(\frac{1}{n^2}) \\ &= [1+x+ \frac{x^2}{2!} +\ldots+ \frac{x^n}{n!}]- \frac{x^2}{2n}[1+x+\ldots+ \frac{x^{n-2}}{(n-2)!}]+0(\frac{1}{n^2}) \\ &= (1- \frac{x^2}{2n})e^x + 0(\frac{1}{n^2}).\end{aligned}$$

Q.E.D.

Lemma 4.3.2

The function

$$g(x) = (\frac{n-1}{n})(1+ \frac{x}{n-1})^{n+1} -\exp(x) \tag{4.3.5}$$

has exactly two positive real roots γ_i, $i = 1,2$, and

$$\lim_{n\to\infty} \gamma_i = 2\pm\surd 2.$$

Proof Consider the error function

$$e(x) = \frac{1}{(1+ \frac{x}{n-1})^n} - \exp(-x).$$

Since, $n\cdot\frac{1}{n-1} > 1$, by Lemma 4.2.2, there exists a $\xi > 0$, such that $e(\xi) = 0$. In addition, $e(0) \doteq \lim_{x\to\infty} e(x) = 0$. Hence, we have at least two distinct points $\gamma_2 > \gamma_1 > 0$ at which its derivative

$$e'(x) = \frac{-n}{(n-1)(1+\frac{x}{n-1})^{n+1}} + \exp(-x) = \frac{g(x)}{-(\frac{n-1}{n})(1+\frac{x}{n-1})^{n+1}\exp(x)}$$

vanishes, i.e. $g(x)$ has at least two positive roots. In Chapter 5, Lemmas 5.3.1, 5.3.2, it will be shown $g(x)$ has at most two positive roots.

Now, consider only the interval $[0,4]$. x is bounded in it. For large n, by Lemma 4.3.1,

$$\begin{aligned} g(x) &= (1-\tfrac{1}{n})(1+\tfrac{x}{n-1})^2(1+\tfrac{x}{n-1})^{n-1}-e^x \\ &= (1-\tfrac{1}{n})(1+\tfrac{x}{n-1})^2(1-\tfrac{x^2}{2(n-1)})e^x-e^x+0(\tfrac{1}{n^2}) \\ &= e^x[(1-\tfrac{1}{n})(1+\tfrac{2x}{n-1})(1-\tfrac{x^2}{2(n-1)})-1]+0(\tfrac{1}{n^2}) \\ &= e^x[\tfrac{-x^2}{2(n-1)}+\tfrac{2x}{n-1}-\tfrac{1}{n}]+0(\tfrac{1}{n^2}) \\ &= \tfrac{-e^x}{2(n-1)}[x^2-4x+2(1-\tfrac{1}{n})]+0(\tfrac{1}{n^2}). \end{aligned}$$

The roots of $[x^2-4x+2(1-\frac{1}{n})]$ are

$$\frac{4\pm\sqrt{16-8(1-\frac{1}{n})}}{2} \to 2\pm\sqrt{2}, \text{ as } n \to \infty.$$

Therefore, if we assume γ_i are continuous with n, then

$$\begin{aligned} \lim_{n\to\infty}\gamma_i &= 2\pm\sqrt{2}, \quad i = 1,2, \\ &= 3.4142,\ 0.5858. \end{aligned}$$

Q.E.D.

Lemma 4.3.3

Let exp(-x) be approximated with reference to R^n_{-1} by

$\underline{r}^n_{-1}(\frac{1}{n-1},x) \in R^n_{-1}$,

$$\underline{r}^n_{-1}(\frac{1}{n-1},x) = \frac{1}{(1+\frac{x}{n-1})^n} .$$

If $e(\frac{1}{n-1},x)$, $\gamma_1,\gamma_2,\lambda_1(\frac{1}{n-1})$, $\lambda_2(\frac{1}{n-1})$ are as defined by (4.2.13), (4.2.14), then, for i = 1,2,

(a) $$\lambda_i(\frac{1}{n-1}) = \frac{e^{-\gamma_i}(\gamma_i-1)}{n}$$

(b) $$\gamma_i \to 2\mp\sqrt{2}, \text{ as } n \to \infty,$$

(c) $$\lambda_i(\frac{1}{n-1}) \to \frac{-0.23}{n}, \frac{0.0794}{n}, \text{ as } n \to \infty.$$

Proof Consider, for i = 1,2, the error curve

$$e(\frac{1}{n-1},x) = \frac{1}{(1+\frac{x}{n-1})^n} - e^{-x}.$$

At $x = \gamma_i$,

$$e(\frac{1}{n-1},\gamma_i) = \lambda_i(\frac{1}{n-1}). \qquad (4.3.10)$$

Since $n(\frac{1}{n-1}) > 1$, by Lemma 4.2.2, γ_i do exist and $\gamma_2 > \gamma_1 > 0$. Furthermore, differentiating $e(\frac{1}{n-1}, x)$ onceat these two extremal points, which are inside $(0,\infty)$, we have

$$\frac{de}{dx}(\frac{1}{n-1},x)\Big|_{x=\gamma_i} = \frac{-n}{(n-1)(1+\frac{x}{n-1})^{n+1}} + e^{-x}\Big|_{x=\gamma_i} = 0 \ . \qquad (4.3.11)$$

Rewriting (4.3.11) as

$$\frac{1}{(1+\frac{\gamma_i}{n-1})^n} = (\frac{n-1}{n})(1+\frac{\gamma_i}{n-1})e^{-\gamma_i} \qquad (4.3.12)$$

gives, together with (4.3.10),

$$\begin{aligned}\lambda_i(\frac{1}{n-1}) &= e^{-\gamma_i}[(\frac{n-1}{n})(1+\frac{\gamma_i}{n-1})-1] \\ &= e^{-\gamma_i}[\frac{\gamma_i}{n-1} - \frac{1}{n} - \frac{\gamma_i}{n(n-1)}] \\ &= \frac{e^{-\gamma_i}(\gamma_i-1)}{n},\end{aligned}$$

proving (a).

$\lambda_2 > \lambda_1 > 0$, being the extremal points of the error curve, are roots of its derivative (4.3.11). Rewriting (4.3.11) in term of $g(x)$ in Lemma 4.3.2, we have,

$$\frac{d}{dx}e(\frac{1}{n-1},x) = \frac{g(x)}{-\exp(x)(\frac{n-1}{n})(1+\frac{x}{n-1})^{n+1}}.$$

Since the denominator is bounded for $0 < x < \infty$, γ_i, $i = 1,2$ are in fact the two roots of $g(x)$ in Lemma 4.3.2, proving (b).

(c) follows immediately from (a) and (b).

Q.E.D.

Similar to Lemma 4.3.3, we include also

Lemma 4.3.4

Let $\exp(-x)$ be approximated with reference to R^n_{-1} by $\underset{=}{r}^n_{-1}(\frac{1}{n},x) \in \underset{=}{R}^n_{-1}$,

$$\underset{=}{r}^n_{-1}(\frac{1}{n},x) = \frac{1}{(1+\frac{x}{n})^n},$$

then (a) $\gamma_1 = \lambda_1(\frac{1}{n}) = 0$,

(b) $\lambda_2(\frac{1}{n}) = \frac{\gamma_2 \exp(-\gamma_2)}{n}$.

Proof

(a) Because $n \cdot (\frac{1}{n}) = 1$, by Lemma 4.2.1, the error curve $\underline{r}^n_{-1}(\frac{1}{n},x)-\exp(-x)$ does not alternate and has only one maximum.

(b) Can be proved similarly as in Lemma 4.3.3(a).

Q.E.D.

We now establish the last theorem of this section.

Theorem 4.3.4

If $f(x) = \exp(-x)$ is to be approximated with reference to R^n_{-1}, then

(a) $\rho_{R^n_{-1}}(f) \le \frac{e^{-1}}{n} = \frac{0.3679}{n}$,

(b) $\min\limits_{i=1,2} \{|\frac{e^{-\gamma_i}(\gamma_i-1)}{n}|\} \le \rho_{R^n_{-1}}(f) \le \max\limits_{i=1,2} \{|\frac{e^{-\gamma_i}(\gamma_i-1)}{n}|\}$

where γ_i are the two positive real roots of (4.3.5),

(c) $\frac{0.0794}{n} < \rho_{R^n_{-1}}(f) < \frac{0.23}{n}$, as $n \to \infty$.

Proof

(a) Lemma 4.3.4(b) implies

$$\rho_{R^n_{-1}}(f) \le \max_{x\in I_0} \left|\frac{xe^{-x}}{n}\right| = \frac{e^{-1}}{n} \text{ at } x = 1$$

$$= \frac{0.3679}{n}$$

(b) and (c). Theorem 4.3.3 and Lemma 4.3.3.

Q.E.D.

Theorem 4.3.4 reveals that the minimax error converges rather slowly with reference to n in R^n_{-1}. In true minimax rational approximation, it is likely that the $r_{m,n}(x)$ minimax approximation has, as $(n+m) \to \infty$, an error [41, p.168] of

$$\frac{m!n!}{2^{m+n}(m+n)!(m+n+1)!}\,(1+0(1)).$$

This prompts the study of minimax approximation by rational functions of the form

$$\frac{p_m(x)}{(1+bx)^n}$$

in R^n_m, $m \geq 0$, in the next chapter.

CHAPTER 5

APPROXIMATING EXP(-X) IN R^n_m

In Chapter 4, the theory of approximating exp(-x) by rational functions $r^n_{-1}(x)$ of the form

$$r^n_{-1}(x) = \frac{1}{\prod_{i=1}^{n} (1+b_i x)} \tag{5.0.1}$$

is proposed and developed. Such approximations have errors of $O(\frac{1}{n})$, which is very large compared with true minimax approximation in the whole space of rational polynomials. For instance, it was shown in [12] that

$$\rho_{R_{n,n}}(\exp(-x)) \leq (2e^{\alpha})^{-n},\ \alpha = 0.1392,\ n = 0,1,2,\ldots\ .$$

Naturally, we would like to look into the set R^n_m. By allowing the numerator of (5.0.1) to be a polynomial of degree m or less, we expect to get a much better approximation. Such will be the theme of this chapter.

5.1 Existence of best approximation in R^n_m

The first question we attempt to answer is: Does there exist a best approximation to exp(-x) in R^n_m? In other words, is it possible to find a real vector

$$\bar{a}(r^{*n-\mu}_{m-\nu}) = (a_0,\ldots,a_{m-\nu},b_1,\ldots,b_{n-\mu}),\quad 0 \leq \nu \leq m, \quad 0 \leq \mu \leq n, \tag{5.1.1}$$

such that the rational polynomial

$$r^{*n-\mu}_{m-\nu}(x) = \frac{\sum_{j=0}^{m-\nu} a_j x^j}{\prod_{j=1}^{n-\mu} (1+b_i x)} \tag{5.1.2}$$

satisfies

$$\rho_{R^n_m}(\exp(-x)) = ||r^{*n-\mu}_{m-\nu}(x)-\exp(-x)|| \tag{5.1.3}$$

where $\rho_{R^n_m}$ and $||\cdot||$ are as defined in Definitions 2.2.1 and 2.2.2 for the interval $I_0 = [0,\infty)$? The answer is in the affirmative as will be shown in this section.

Theorem 5.1.1

Among the functions $r^k_m(x)$, $n \geq k \geq \deg(p_m)$, in R^n_m, there is at least one function for which

$$||r^k_m(x)-\exp(-x)|| \tag{5.1.4}$$

attains its minimum.

Proof

Step 1: Existence of a convergent subsequence in R^n_m.

Since (5.1.4) is bounded below by zero, it has a greatest lower bound

$$\begin{aligned}\rho_{R^n_m} &= \rho_{R^n_m}[\exp(-x)] \\ &= \inf_{r^n_m \in R^n_m} ||r^n_m(x)-\exp(-x)|| \geq 0.\end{aligned} \tag{5.1.5}$$

By definition, there exists an infinite sequence of functions $r_i(x)$ $(i = 1,2,\ldots)$ in R_m^n such that

$$\rho_{r_i} = ||r_i(x)-\exp(-x)|| \to \rho_{R_m^n}, \tag{5.1.6}$$

as $i \to \infty$.

Denoting $r_i(x)$ in the form

$$r_i(x) = \frac{\sum_{j=0}^{m} a_{ij}x^j}{\prod_{j=1}^{n} (1+b_{ij}x)}, \quad b_{ij} \geq 0, \tag{5.1.7}$$

we perform a normalisation on their coefficients as follows:

$$\begin{aligned} r_i(x) &= \frac{\sum_{j=0}^{m} a_{ij}x^j}{\prod_{j=1}^{n} (1+b_{ij}^2) \prod_{j=1}^{n} (\frac{1}{1+b_{ij}^2} + \frac{b_{ij}x}{1+b_{ij}^2})} \\ &= \frac{\sum_{j=0}^{m} u_{ij}x^j}{\prod_{j=1}^{n} (v_{ij}+w_{ij}x)} . \end{aligned}$$

Because $b_{ij} \geq 0$, $\{v_{ij}\}$, $\{w_{ij}\}$ are bounded as indicated:

$$\begin{aligned} &1 \geq v_{ij} > 0, \\ &1 > w_{ij} \geq 0, \quad j = 1,2,\ldots,n, \ i = 1,2,\ldots . \end{aligned} \tag{5.1.8}$$

The sequence $\{u_{ij}\}$ is also bounded as will be proved below.

Since the function

$$g(x) \equiv 0$$

is in R_m^n and

$$||g(x)-\exp(-x)|| = 1$$

$\rho_{R_m^n}$ of (5.1.5) is bounded by 1. Because

$$\lim_{i\to\infty} \rho_{r_i} = \rho_{R_m^n}$$

the sequence ρ_{r_i} has an upperbound G,

$$\rho_{r_i} < G. \tag{5.1.9}$$

Let $\{\xi_1,\xi_2,\ldots,\xi_{m+1}\}$ be a set of (m+1) distinct fixed points in I_0. We then have

$$\begin{aligned} |r_i(\xi)| &\le G + \max_{x\epsilon I_0} |\exp(-x)| \\ &= G + 1, \end{aligned} \tag{5.1.10}$$

where ξ is any of the points ξ_i. This at once implies

$$|\sum_{j=0}^{m} u_{ij}\xi^j| \le (G+1)|\prod_{j=1}^{n} (v_{ij}+w_{ij}\xi_j)| \tag{5.1.11}$$

for all i. As v_{ij}, w_{ij} are bounded (5.1.8), this means

$$|\sum_{j=0}^{m} u_{ij}\xi j| < K$$

for a certain constant K.

If the values of the polynomials

$$u_{i0} + u_{i1}x + \dots + u_{im}x^m$$

are bounded at (m+1) fixed distinct points, then all coefficients of these polynomials are also bounded.

We now make use of the fact that the sequence of vectors

$$\bar{a}_i = (u_{i0},\dots,u_{im},v_{i1},\dots,v_{in},w_{i1},\dots,w_{in}), \quad i = 1,2,\dots,$$

is bounded and select from it a certain convergent subsequence (which will likewise be denoted by $\{\bar{a}_i\}$), such that

$$\begin{aligned}\lim_{i\to\infty} \bar{a}_i &= \bar{a} \\ &= (u_0,\dots,u_m,v_i,\dots,v_n,w_1,\dots,w_n). \qquad (5.1.12)\end{aligned}$$

Define

$$r^*(x) = \frac{\sum_{j=0}^{m} u_j x^j}{\prod_{j=1}^{n} (v_j+w_jx)}. \qquad (5.1.13)$$

<u>Step 2</u> $r^*(x) \in R_m^n$.

Since $v_{ij} > 0$, $w_{ij} \geq 0$, we have

$$\begin{aligned}&v_j \geq 0, \\ &w_j \geq 0, \quad j = 1,\dots,n,\end{aligned}$$

and we shall consider separately the two cases:

Case A $\quad |v_k| + |w_k| = v_k + w_k = 0, \qquad (5.1.14)$

for a certain k, $0 \leq k \leq n$. This implies

$$\lim_{i\to\infty} v_{ik} = 0,$$
$$\lim_{i\to\infty} w_{ik} = 0, \qquad (5.1.15)$$

whence, because of (5.1.11),

$$\lim_{i\to\infty} u_{ij} = u_j = 0, \quad j = 1,2,\ldots,n.$$

$r^*(x)$ is thus the zero function

$$r^*(x) \equiv 0,$$

which obviously is in R^n_m.

Case B $\quad |v_j| + |w_j| = v_j + w_j > 0 \qquad (5.1.16)$

for all $j = 1,\ldots,n$.

In this case, $r^*(x)$ has no positive pole ; hence is defined for all $x > 0$. We shall prove it is also bounded.

Because of (5.1.12), pointwisely,

$$\lim_{i\to\infty} r_i(x) = r^*(x), \quad x > 0, \qquad (5.1.17)$$

and hence

$$\begin{aligned}|r^*(x)| &= |\exp(-x)-[\exp(-x)-r_i(x)]-[r_i(x)-r^*(x)]| \\ &\le |\exp(-x)|+|\exp(-x)-r_i(x)|+|r_i(x)-r^*(x)| \\ &\le \sup_{x>0}|\exp(-x)|+ \rho_{r_i}+\varepsilon_i, \end{aligned} \tag{5.1.18}$$

where $\varepsilon_i \to 0$ as $i \to \infty$.

Because of (5.1.9), and

$$\sup_{x>0}|\exp(-x)| = \exp(0) = 1,$$

(5.1.18) implies

$$|r^*(x)| \le M = 1 + G, \quad x > 0, \tag{5.1.19}$$

proving $r^*(x)$ is bounded in $(0,\infty)$.

Next, consider the linear factors in the denominator of $r^*(x)$. They are of the form

$$F_j = (v_j + w_jx), \ j = 1,2,\dots,n. \tag{5.1.20}$$

We can rewrite them in one of the three ways:

(i) If $v_j > 0$, $w_j > 0$,

$$F_j = v_j(1+ \frac{w_j}{v_j}x) = v_j(1+b_jx), \quad b_j > 0. \tag{5.1.21}$$

(ii) If $v_j = 0$,

$$F_j = w_jx.$$

(iii) If $w_j = 0$,

$$F_j = v_j. \tag{5.1.22}$$

Putting these back to $r^*(x)$ in (5.1.13) and after reducing the fraction to its lowest terms, $r^*(x)$ can assume any one of the three forms:

$$r^*(x) = \frac{\sum_{j=0}^{m-\mu} a_j x^j}{x^\eta \prod_{j=1}^{n-\nu} (1+b_j x)}, \qquad \eta > 0, \quad a_0 \neq 0, \quad a_{m-\mu} \neq 0,\ b_j > 0; \tag{5.1.23}$$

$$r^*(x) = \frac{\sum_{j=0}^{m-\mu} a_j x^j}{\prod_{j=1}^{n-\nu} (1+b_j x)}, \qquad a_{m-\mu} \neq 0,\ b_j > 0; \tag{5.1.24}$$

or

$$r^*(x) \equiv 0. \tag{5.1.25}$$

(5.1.23) is not possible because $r^*(x)$ can then be arbitrarily large as x approaches zero, contradicting the bounded property of $r^*(x)$ in (5.1.19). Similarly, in (5.1.24),

$$(m-\mu) \leq (n-\nu),$$

otherwise $r^*(x)$ grows unbounded as $x \to \infty$. Hence, $r^*(x)$ is of the form

$$r^*(x) = \frac{\sum_{j=0}^{m-\mu} a_j x^j}{\prod_{j=1}^{n-\nu} (1+b_j x)}, \quad (m-\mu) \leq (n-\nu),\ a_{m-\mu} \neq 0,\ b_j > 0, \tag{5.1.26}$$

or

$$r^*(x) \equiv 0 \tag{5.1.27}$$

both of which are in R_m^n.

Step 3 $r^*(x)$ is the best approximation.

Since $r^*(x)$ is in R_m^n, it is defined for all $x \geq 0$. Also, by definition,

$$\rho_{R_m^n} \leq ||r^*(x)-\exp(-x)||. \tag{5.1.28}$$

On the other hand,

$$\begin{aligned} ||r^*(x)-\exp(-x)|| &\leq ||r^*(x)-r_i(x)+r_i(x)-\exp(-x)|| \\ &\leq ||r^*(x)-r_i(x)||+||r_i(x)-\exp(-x)||. \end{aligned} \tag{5.1.29}$$

As i → ∞, (5.1.29) implies

$$||r^*(x)-\exp(-x)|| \leq \rho_{R_m^n}.$$

Hence,

$$||r^*(x)-\exp(-x)|| = \rho_{R_m^n},$$

and $r^*(x)$ attains the minimum of (5.1.4). Q.E.D.

5.2 Necessary Conditions for the Best Approximation

The significance of finding the necessary conditions is that it allows us to have a more specific search for the best approximation. Instead of investigating the whole set R_m^n, we only have to direct our attention to a particular subset of R_m^n which satisfies the necessary conditions because we are guaranteed that the best approximation will not lie outside it.

Lemma 5.2.1

Given

(i) $b_1, b_2, b_2 > b_1 \geq 0$;

(ii) a set of (m+2) distinct points $\{\beta_i\}$, $\beta_i > 0$, $i = 1,2,\ldots,m+2$;

(iii) a polynomial $p_m(x)$ of degree m;

$$p_m(\frac{-1}{b_1}) \neq 0, \text{ if } b_1 \neq 0,$$

$$p_m(\frac{-1}{b_2}) \neq 0;$$

(iv) $\Phi(x) = \prod_{j=1}^{m+2} (x-\beta_j)$;

then there exist

(i) b_1', b_2' and

(ii) a polynomial $q_m(x)$ of degree at most m, such that

$$(1+b_1x)(1+b_2x)q_m(x) - (1+b_1'x)(1+b_2'x)p_m(x) = w\Phi(x) \tag{5.2.1}$$

for certain sufficiently small w.

Proof

(A) Assume $b_1 > 0$.

Consider the function

$$g(x) = [1+(b_1+\delta_1)x][1+(b_2-\delta_2)x]p_m(x)+w\Phi(x). \tag{5.2.2}$$

We shall prove, for sufficiently small w, that there is a choice of δ_1, δ_2, such that g(x) has two zeroes at $x=-1/b_1$ and $-1/b_2$. Form the two equations

$$g(\frac{-1}{b_i}) = p_m(\frac{-1}{b_i})[1- \frac{b_1+\delta_1}{b_i}][1- \frac{b_2-\delta_2}{b_i}]+w\Phi(\frac{-1}{b_i}) = 0, \; i = 1,2. \qquad (5.2.3)$$

Since $p_m(\frac{-1}{b_i}) \neq 0$, $i = 1,2$, (5.2.3) reduces to

$$\delta_1(b_2-b_1-\delta_2) = \frac{-wb_1^2\Phi(\frac{-1}{b_1})}{p_m(\frac{-1}{b_1})} \; ,$$

$$\delta_2(b_2-b_1-\delta_1) = \frac{-wb_2^2\Phi(\frac{-1}{b_2})}{p_m(\frac{-1}{b_2})}, \qquad (5.2.4)$$

Or, in another form

$$\gamma\delta_i - \delta_1\delta_2 = \alpha_i, \quad i = 1,2, \qquad (5.2.5)$$

where

$$\gamma = b_2 - b_1 > 0,$$

$$\alpha_i = \frac{-wb_i^2\Phi(\frac{-1}{b_i})}{p_m(\frac{-1}{b_i})}. \qquad (5.2.6)$$

Because $\gamma > 0$, we can solve (5.2.5) for δ_1, δ_2:

$$\delta_2 = \delta_1 - \frac{\alpha_1-\alpha_2}{\gamma} \; , \qquad (5.2.7)$$

$$\gamma\delta_1 - \delta_1(\delta_1- \frac{\alpha_1-\alpha_2}{\gamma}) = \alpha_1, \qquad (5.2.8)$$

which reduces to

$$\gamma\delta_1^2 - (\gamma^2+\alpha_1-\alpha_2)\delta_1 + \gamma\alpha_1 = 0. \qquad (5.2.9)$$

Being a quadratic equation, (5.2.9) will have a real solution if its discriminant

$$\begin{aligned}&(\gamma^2+\alpha_1-\alpha_2)^2 - 4\gamma^2\alpha_1 \\ &= (\gamma^2-\alpha_1-\alpha_2)^2 - 4\alpha_1\alpha_2 \geq 0. \end{aligned} \qquad (5.2.10)$$

This can be satisfied easily if

$$\gamma^2 \geq 4 \max[|\alpha_1|,|\alpha_2|], \qquad (5.2.11)$$

that is, if

$$|w| \leq \varepsilon_1 = \frac{(b_2-b_1)^2}{4 \max\limits_{i=1,2}\left[\left|\frac{b_i^2\Phi(\frac{-1}{b_i})}{p_m(\frac{-1}{b_i})}\right|\right]} , \qquad (5.2.12)$$

where, by our given conditions, the denominator on the right never vanishes.

Hence, for any w whose magnitude satisfies (5.2.12), g(x) has two real zeroes at $-b_1^{-1}$ and $-b_2^{-1}$ for certain δ_1,δ_2. This means g(x) can be represented in the form

$$\begin{aligned} g(x) &= (1+b_1x)(1+b_2x)q_m(x) \\ &= (1+b_1'x)(1+b_2'x)p_m(x)+w\Phi(x) \end{aligned}$$

where δ_i, i = 1,2 satisfy (5.2.8), (5.2.9); $q_m(x)$, a polynomial of degree m or less; and

Proof

If $b_1 = 0$ and $b_1 = b_1'$, then the left-hand side of (5.2.1) will be a polynomial of degree (m+1) contradicting that $\Phi(x)$ on the right-hand side is of degree (m+2).

If in (5.2.1), $b_1 = b_1' \neq 0$ or $b_2 = b_2'$, then the expression on the left will have at least one real, negative zero at $-b_1^{-1}$ or $-b_2^{-1}$, contradicting the definition of $\Phi(x)$, whose zeroes are all real and positive.

Corollary 5.2.2

In (5.2.1), for any w satisfying (5.2.12), b_1' and b_2' are given by the equations

$$\delta_1 = \frac{(\gamma^2+\alpha_1-\alpha_2)\pm\sqrt{(\gamma^2-\alpha_1-\alpha_2)^2-4\alpha_1\alpha_2}}{2\gamma},$$

$$\delta_2 = \frac{(\gamma^2-\alpha_1+\alpha_2)\pm\sqrt{(\gamma^2-\alpha_1-\alpha_2)^2-4\alpha_1\alpha_2}}{2\gamma}, \qquad (5.2.15)$$

$$b_1' = b_1 + \delta_1,$$

$$b_2' = b_2 - \delta_2,$$

where γ, α_1, α_2 are as defined in (5.2.6), (5.2.7) or (5.2.13).

Proof This is obvious from the proof of Lemma 5.2.1, where b_1' and b_2' are obtained by solving the quadratic equations (5.2.5)-(5.2.9) for δ_1 and δ_2.

Q.E.D.

Corollary 5.2.3

(i) $b_1' \to b_1$ or b_2, as $w \to 0$;

(ii) $b_2' \to b_2$, as $w \to 0$, and $b_1' \to b_1$;

(iii) $b_2' \to b_1$, as $w \to 0$, and $b_1' \to b_2$.

Proof By definition, for $i = 1,2$, $\alpha_i \to 0$ as $w \to 0$. Therefore, from Corollary 5.2.2,

$$\delta_1 \to \frac{\gamma^2 \pm \gamma^2}{2\gamma} = \gamma \text{ or } 0, \text{ as } w \to 0,$$
$$= b_2 - b_1 \text{ or } 0. \tag{5.2.16}$$

Therefore,

$$b_1' = b_1 + \delta_1 \to b_2 \text{ or } b_1, \tag{5.2.17}$$

proving (i).

(ii) and (iii) follow immediately from (5.2.7). Q.E.D.

Corollary 5.2.4

$$b_2' - b_1' = \frac{\pm\sqrt{(\gamma^2-\alpha_1-\alpha_2)^2-4\alpha_1\alpha_2}}{\gamma}. \tag{5.2.18}$$

Proof This is obvious from Corollary 5.2.2 and the equation

$$b_2' - b_1' = (b_2-b_1) - (\delta_2+\delta_1)$$
$$= \gamma - \frac{\gamma^2 \pm \sqrt{(\gamma^2-\alpha_1-\alpha_2)^2-4\alpha_1\alpha_2}}{\gamma}$$
$$= \frac{\pm\sqrt{(\gamma^2-\alpha_1-\alpha_2)^2-4\alpha_1\alpha_2}}{\gamma}.$$

Q.E.D.

Corollary 5.2.5

The following are equivalent, if $w \neq 0$:

(i) $$|b_2'-b_1'| < |b_2-b_1|; \qquad (5.2.19)$$

(ii) $$(\alpha_1-\alpha_2)^2 < 2\gamma^2(\alpha_1+\alpha_2); \qquad (5.2.20)$$

(iii) $$|w| < \varepsilon_2 = \frac{2\gamma^2|\alpha_1'+\alpha_2'|}{(\alpha_1'-\alpha_2')^2}, \quad \text{if } \alpha_1' \neq \alpha_2', \qquad (5.2.21)$$

where $\gamma = (b_2-b_1)$,

$$\alpha_1' = \frac{-1}{a_m}, \quad \text{if} \quad b_1 = 0,$$

$$= \frac{b_1^2\Phi(\frac{-1}{b_1})}{p_m(\frac{-1}{b_1})}, \text{ otherwise,}$$

$$\alpha_2' = \frac{b_2^2\ (\frac{-1}{b_2})}{p_m(\frac{-1}{b_2})},$$

$\alpha_i = w\alpha_i'$, $i = 1,2$,

a_m = coefficient of x^m in $p_m(x)$, $a_m \neq 0$.

Proof From Corollary 5.2.4, we have

$$|b_2'-b_1'| < |b_2-b_1|, \qquad \text{(i)}$$

iff $$\frac{\sqrt{(\gamma^2-\alpha_1-\alpha_2)^2-4\alpha_1\alpha_2}}{\gamma} < \gamma, \quad \gamma > 0,$$

iff $$(\gamma^2-\alpha_1-\alpha_2)^2-4\alpha_1\alpha_2 < \gamma^4$$

iff $(\alpha_1-\alpha_2)^2 < 2\gamma^2(\alpha_1+\alpha_2)$, (ii)

iff $w^2(\alpha_1'-\alpha_2')^2 < 2w\gamma^2(\alpha_1'+\alpha_2')$

iff (iii) Q.E.D.

Let us denote the best approximation to exp(-x) in R_m^n by

$$r^{*n}_m(x) = \frac{a_0+a_1x+\ldots+a_{m-\mu}x^{m-\mu}}{\prod_{j=1}^{n-\nu}(1+b_jx)} = \frac{p_m(x)}{\prod_{j=1}^{n-\nu}(1+b_jx)} \qquad \begin{array}{l} m \geq 0, \\ n \geq 1, \\ n \geq m, \\ 0 \leq \mu \leq m, \\ 0 \leq \nu \leq n, \\ a_{m-\mu} \neq 0, \\ b_j > 0, \\ m-\mu \leq n-\nu, \end{array} \tag{5.2.22}$$

and let the error curve

$$r^{*n}_m(x)-\exp(-x) \tag{5.2.23}$$

alternate k times on $I_0 = [0,\infty)$. We assume also that $r^{*n}_m(x)$ is irreducible, i.e. $p_m(-b_j^{-1}) \neq 0$ $(j = 1,2,\ldots,n-\nu)$.

If we rewrite the error function

$$r^{*n}_m(x)-\exp(-x)$$

$$= \frac{1}{\prod_{j=2}^{n-\nu}(1+b_jx)} \left(\frac{p_m(x)}{(1+b_1x)} - \exp(-x)\prod_{j=2}^{n-\nu}(1+b_jx)\right), \tag{5.2.24}$$

we note that the rational polynomial

$$\frac{p_m(x)}{1+b_1x}$$

should also be the best weighted minimax approximation to

$$\exp(-x) \prod_{j=2}^{n-\nu} (1+b_jx)$$

with the weight function

$$\prod_{j=2}^{n-\nu} (1+b_jx)^{-1}.$$

Hence, by Theorem 2.2.1, (5.2.23) alternates at least m-μ+2 times, i.e.

$$k \geq m-\mu+2. \tag{5.2.25}$$

If $k = m-\mu+2$ then $\nu = 0$. Otherwise, if $\nu > 0$, by Lemma 5.2.1, we can find b_1', b_2' and a polynomial $q_m(x)$ of degree at most $(m-\mu)$ such that

$$(1+b_1x)q_m(x) - (1+b_1'x)(1+b_2'x)p_m(x) = w\Phi(x),$$

$$\Phi(x) = w \prod_{j=1}^{k} (x-\beta_j), \tag{5.2.26}$$

where β_j, $(j = 1,\ldots,k)$ are those points as depicted in Figure 4.2.1 of Lemma 4.2.3. (5.2.26) implies

$$\frac{q_m(x)}{(1+b_1'x)(1+b_2'x)\prod_{j=2}^{n-\nu}(1+b_jx)} - \frac{p_m(x)}{\prod_{j=1}^{n-\nu}(1+b_jx)}$$

$$= \frac{w}{(1+b_1'x)(1+b_2'x)\prod_{j=1}^{n-\nu}(1+b_jx)}\Phi(x)$$

$$= w(x)\Phi(x). \tag{5.2.27}$$

For sufficiently small w,by Corollary 5.2.2, δ_1, b_2' will satisfy

$$\delta_1 = b_1' > 0$$

$$b_2' = b_1 - \delta_2 > 0,$$

and hence

$$r_m^n(x) = \frac{q_m(x)}{(1+b_1'x)(1+b_2'x)\prod_{j=2}^{n-\nu}(1+b_jx)}$$

will be in R_m^n. Furthermore, since

$$(m-\mu) \leq (n-\nu)$$

(otherwise, $r^{*n}_m(x)$ will become unbounded as $x \to \infty$, contradicting that it is the best approximation) we have

$$\lim_{x\to\infty}|w(x)\Phi(x)|$$

$$= \lim_{x\to\infty}\left|\frac{w}{x^{(n-\nu)-(m-\mu)}b_1'b_2'\prod_{j=2}^{n-\nu} b_j}\right|$$

which is bounded because $(m-\mu) \leq (n-\nu)$. And because $w(x)$ is also bounded for any $0 \leq x < \infty$, hence,

$$|w(x)\Phi(x)| < |w|K , \qquad x \geq 0$$

for a certain constant K. If w is made small enough and its sign chosen appropriately, then, by Lemma 4.2.3, $r_m^n(x)$ will be a better approximation than $r^{*n}_m(x)$.

Similarly, we can prove

$$b_1 = b_2 = \dots = b_{n-\nu} \quad .$$

Otherwise, if there are say $b_2 > b_1 > 0$, then we can again construct a $r^n_m(x)$ which approximates $\exp(-x)$ better than $r^{*n}_m(x)$. Hence, we have proved

Theorem 5.2.1

Let the best approximation to $\exp(-x)$ in R^n_m be $r^{*n}_m(x)$ (5.2.22) and let its error curve (5.2.23) alternate k times on $[0,\infty)$, then

(i) $k \geq m-\mu+2$,

(ii) $\nu = 0$ and

$b_1 = b_2 = \dots = b_n$, if $k = m-\mu+2$.

In most cases of approximation, it is found $\mu = 0$ and $k = m+2$ (non-degenerate approximation). Because of Theorem 5.2.1, now it may be possible for us to search for the best approximation inside $\underline{R}^n_m$ instead of working on the whole set R^n_m. One implication of Lemma 5.2.1 and Corollaries 5.2.3 and 5.2.5 is that, given any rational function in R^n_m, they provide us a way to construct a better approximation using (5.2.15). Furthermore, since w is in general very small, (5.2.18) and (5.2.19) imply that the b_j's in the denominator of an approximation move towards one common point as the approximating function approaches the best approximation.

In the next section, it will be shown that the best approximation in R^n_0 is non-degenerate and the error curve alternates exactly twice.

In the numerical algorithm that finds the best approximation in R_m^n for general m,n, it will be assumed that the solution is nondegenerate and the corresponding error curve alternates exactly m+2 times. Such assumptions are found consistent with all computation tests attempted.

5.3 Sufficient Condition and Uniqueness in R_0^n

In this section, we attempt to complete the theory of approximation to exp(-x) by rational polynomials in R_0^n of the form

$$\frac{p_0(x)}{\prod_{j=1}^{n-\nu}(1+b_jx)}, \quad b_j > 0, \quad 0 \le \nu < n,$$

by establishing the sufficient condition for and uniqueness of the best approximation.

Recalling the function g(x) in (4.2.1),

$$g(x) = \exp(x)-\prod_{i=1}^{n}(1+b_ix), \qquad b_i > 0, \tag{5.3.1}$$

we have

Theorem 5.3.1

The following are equivalent for g(x) of (5.3.1):

(i) $\sum_{i=1}^{n} b_i \le 1$;

(ii) $g(x) > 0, \ x > 0$;

(iii) there exists $\varepsilon > 0$, such that $g(x) > 0, \ 0 < x < \varepsilon$.

Proof

(i) $\Rightarrow$ (ii) Lemma 4.2.1.

(ii) $\Rightarrow$ (i) If $g(x) > 0$, $x > 0$, then $g(x)$ has no positive zero. By Lemma 4.2.2, $\sum_{i=1}^{n} b_i \not> 1$, hence (i).

(ii) $\Rightarrow$ (iii) Trivial.

(iii) $\Rightarrow$ (i) For small x,

$$g(x) = (1-\sum_{i=1}^{n} b_i)x + O(x^2).$$

If, by contradiction, $\sum_{i=1}^{n} b_i > 1$, $g(x) < 0$ for x near zero, hence contradicting (iii). Q.E.D.

The next theorem is a counterpart of Theorem 5.3.1. The proof will be omitted since it would be very similar to that of Theorem 5.3.1.

Theorem 5.3.2

If $g(x)$ is as defined in (5.3.1), the following are equivalent

(i) $\sum_{i=1}^{n} b_i > 1$.

(ii) there is exactly one point $\xi > 0$ such that $g(\xi) = 0$.

(iii) there exists $\varepsilon > 0$ such that $g(x) < 0$, $0 < x < \varepsilon$.

Proof Lemmas 4.2.1, 4.2.2 and Theorem 5.3.1.

Lemma 5.3.1

Define

$$h(x) = a \exp(x) - \prod_{i=1}^{n} (1+b_i x),\ b_i > 0,\ x \geq 0,\ a > 0. \tag{5.3.2}$$

If $a < 1$, then $h(x)$ has exactly one positive root.

Proof Differentiating h(x) (n+1) times,

$$h^{(n+1)}(x) = a\exp(x) > 0, \quad x > 0.$$

Hence, h(x) has at most n+1 roots.

At x = 0,

$$h(0) = a-1 < 0.$$

As x → ∞, obviously,

$$a\exp(x) > \prod_{i=1}^{n}(1+b_ix),$$

and hence h(x) > 0. Therefore, h(x) has at least one positive zero. Let x = γ be its smallest positive zero, such that

$$h(x) < 0, \quad 0 < x < \gamma, \tag{5.3.3}$$

$$h(\gamma) = a\exp(\gamma)-\prod_{i=1}^{n}(1+b_i\gamma) = 0. \tag{5.3.4}$$

We shall show h(x) > 0, for x > γ.

Case A $\sum_{i=1}^{n} b_i \leq 1.$

By Lemma 4.2.1,

$$\exp(x) > \prod_{i=1}^{n}(1+b_ix) > 0, \quad x > 0.$$

Together with (5.3.4),

$$\begin{aligned} a\exp(\gamma)\exp(x) &> \prod_{i=1}^{n}(1+b_i\gamma)\prod_{i=1}^{n}(1+b_ix) \\ &= \prod_{i=1}^{n}[(1+b_i\gamma)(1+b_ix)] \\ &> \prod_{i=1}^{n}[1+b_i(x+\gamma)], \quad x > 0. \end{aligned} \tag{5.3.5}$$

Therefore,

$$a \exp(\gamma+x) > \prod_{i=1}^{n} [1+b_i(x+\gamma)], \quad x > 0,$$

hence, $h(x) > 0$, for $x > \gamma$.

<u>Case B</u> $\sum_{i=1}^{n} b_i > 1$.

Because of (5.3.3),

$$a \exp(x) < \prod_{i=1}^{n} (1+b_i x), \quad 0 \leq x < \gamma,$$

which gives, by the transformation $x = \gamma-y$,

$$\begin{aligned} a \exp(\gamma-y) &< \prod_{i=1}^{n} [1+b_i(\gamma-y)], \quad 0 < y < \gamma \\ &= \prod_{i=1}^{n} [(1+b_i\gamma)-b_i y] \\ &= \prod_{i=1}^{n} (1+b_i\gamma) \prod_{i=1}^{n} [1-\frac{b_i y}{1+b_i\gamma}]. \end{aligned}$$

Therefore, by (5.3.4),

$$0 < \exp(-y) < \prod_{i=1}^{n} [1-\frac{b_i y}{1+b_i\gamma}], \ 0 < y < \gamma$$

or
$$\begin{aligned} \exp(y) &> \frac{1}{\prod_{i=1}^{n} [1-\frac{b_i y}{1+b_i\gamma}]} \\ &= \prod_{i=1}^{n} [1+\frac{b_i y}{1+b_i\gamma} + (\frac{b_i y}{1+b_i\gamma})^2 + \ldots] \\ &> \prod_{i=1}^{n} [1+\frac{b_i y}{1+b_i\gamma}], \quad 0 < y < \gamma . \end{aligned}$$

By Theorem 5.3.1, this means

$$\exp(y) > \prod_{i=1}^{n} [1+ \frac{b_i y}{1+b_i\gamma}], \quad 0 < y < \infty$$

$$= \frac{\prod_{i=1}^{n} [1+b_i(y+\gamma)]}{\prod_{i=1}^{n} (1+b_i\gamma)}, \tag{5.3.6}$$

which gives

$$\exp(y) \prod_{i=1}^{n} (1+b_i\gamma)$$

$$= a \exp(y)\exp(\gamma)$$

$$= a \exp(y+\gamma) > \prod_{i=1}^{n} [1+b_i(y+\gamma)], \quad 0 < y < \infty ,$$

proving

$$a \exp(x) > \prod_{i=1}^{n} (1+b_i x), \qquad x > \gamma.$$

Hence, in both cases, γ is the only positive root. Q.E.D.

Lemma 5.3.2

If in (5.3.2), $a > 1$, then $h(x)$ has at most two positive zeroes.

Proof

Case A $\sum_{i=1}^{n} b_i \le 1$.

By Lemma 4.2.1,

$$\exp(x) > \prod_{i=1}^{n} (1+b_i x), \qquad x > 0.$$

Therefore,

$$a \exp(x) > \prod_{i=1}^{n} (1+b_i x) \; ,$$

hence, $h(x) > 0, \quad x > 0,$

and $h(x)$ has no positive roots.

Case B $\sum_{i=1}^{n} b_i > 1.$

By contradiction, let $h(x)$ have at least three positive roots, the first two of which being γ_i, $i = 1,2$, and

$$\gamma_2 > \gamma_1 > 0.$$

Since $a > 0$, $h(0) > 0$ and hence

$$h(x) > 0, \quad 0 < x < \gamma_1,$$

$$h(x) < 0, \quad \gamma_1 < x < \gamma_2.$$

Therefore,

$$a \exp(\gamma_2 - y) < \prod_{i=1}^{n} [1+b_i(\gamma_2 - y)], \quad 0 < y < (\gamma_2 - \gamma_1). \qquad (5.3.7)$$

By exactly the same argument as in proving case B of Lemma 5.3.1, it can be shown

$$h(x) > 0, \quad x > \gamma_2.$$

Hence $h(x)$ has at most two roots. Q.E.D.

Let the best approximation to $\exp(-x)$ in R_0^n be

$$r^{*n}_0(x) = \frac{a}{\prod_{i=1}^{n-\nu} (1+b_i x)} \; , \quad 0 \le \nu < n.$$

Obviously, $a \neq 0$ otherwise $r^{*n}_0(x) \equiv 0$ and we know this is not the best approximation. By Corollaries 5.3.1, 5.3.2 and Theorem 5.2.1, we know the error curve

$$\frac{a}{\prod\limits_{i=1}^{n-\nu}(1+b_i x)} - \exp(-x)$$

alternates exactly twice. Hence, $r^{*n}_0(x)$ will be found inside the subset $\underline{R}^n_0$ of R^n_0, and can be denoted in the form

$$r^{*n}_0(x) = \frac{a^*}{(1+b^*x)^n}\ ,\ b^* > 0.$$

To establish the sufficiency condition and uniqueness of $r^{*n}_0(x)$, we need the next lemma.

Lemma 5.3.3

The function

$$\frac{a_1}{(1+b_1x)^n} - \frac{a_2}{(1+b_2x)^n}\ ,\ b_1, b_2 > 0 \tag{5.3.8}$$

has at most one non-negative zero.

Proof Solving (5.3.8) for x, we have, assuming $a_2 \neq 0$,

$$\frac{1+b_1x}{1+b_2x} = (\frac{a_1}{a_2})^{1/n}\ . \tag{5.3.9}$$

If n is odd, (5.3.9) has at most one real root,

$$x = \frac{[(\frac{a_1}{a_2})^{1/n}-1]}{[b_1-b_2(\frac{a_1}{a_2})^{1/n}]}, \tag{5.3.10}$$

provided the denominator does not vanish.

If n is even, but $\frac{a_1}{a_2} < 0$, then (5.3.9) has no real root.

If n is even and $\frac{a_1}{a_2} > 0$. Let

$$(\frac{a_1}{a_2}) = (\pm\alpha)^n, \quad \alpha > 0.$$

(5.3.9) implies,

$$x = \frac{1-(\pm\alpha)}{\pm b_2\alpha-b_1}$$

$$= \frac{1-\alpha}{b_2\alpha-b_1} \text{ or } \frac{-(1+\alpha)}{b_2\alpha+b_1} . \tag{5.3.11}$$

Obviously, $\frac{-(1+\alpha)}{b_2\alpha+b_1} < 0$.

Hence, (5.3.9) has at most one non-negative root. Q.E.D.

Lemma 5.3.4 (Sufficiency Lemma)

If the error curve

$$\frac{a^*}{(1+b^*x)^n} - \exp(-x) = r^{*n}_m(x)-\exp(-x) \tag{5.3.12}$$

alternates twice, then $r^{*n}_0(x)$ is the best approximation to exp(-x) in $\underset{-}{R}^n_0-\underset{-}{R}^{n-1}_0$.

Proof If there is a better approximation

$$r_0^n(x) = \frac{a}{(1+bx)^n} \tag{5.3.13}$$

then, consider the function

$$\begin{aligned}\Delta(x) &= \frac{a^*}{(1+b^*x)^n} - \frac{a}{(1+bx)^n} \\ &= [\frac{a^*}{(1+b^*x)^n} - \exp(-x)] - [\frac{a}{(1+bx)^n} - \exp(-x)]\end{aligned} \tag{5.3.14}$$

At each extremal point of (5.3.12), the sign of (5.3.14) will be the sign of (5.3.12). Since (5.3.12) alternates twice, it means (5.3.14) will have at least two positive roots, contradicting Lemma 5.3.3. Q.E.D.

Lemma 5.3.5 (Uniqueness)

The best approximation in $R_{-0}^n - R_{-0}^{n-1}$ is unique.

Proof Let (5.3.12) and (5.3.13) both be best approximations. Consider again (5.3.14) the function $\Delta(x)$. If ξ_i $(i = 1,2,3)$ are the three extremal points of (5.3.12), we have

Case A $\Delta(\xi_i) \neq 0$, $i = 1,2,3$.

As in Lemma 5.3.4, this means $\Delta(x)$ has at least two positive zeroes,contradicting Lemma 5.3.3.

Case B $\Delta(\xi_i) = 0$ for more than one point.

This is impossible because $\Delta(x)$ cannot have more than one non-negative zero.

Case C $\Delta(\xi_i) = 0$ at one point.

If $\Delta(\xi_i) = 0$ at $i = 1$ or 3, there will be one more point θ, $\xi_1 < \theta < \xi_3$ such that $\Delta(\theta) = 0$.

Let $\Delta(\xi_i) = 0$ at $i = 2$. It can be easily shown that ξ_2 cannot be a double root. Hence, there will be a point $\xi_1 < \theta < \xi_3$ such that $\Delta(\theta) = 0$, $\theta \neq \xi_2$. In both cases, we get a contradiction to Lemma 5.3.3. Q.E.D.

Because of Theorem 5.2.1, Corollaries 5.3.1 and 5.3.2, the best approximation of exp(-x) in $\underset{\sim}{R}_0^n - \underset{\sim}{R}_0^{n-1}$ is in fact that in R_0^n. This allows us to conclude this section by

Theorem 5.3.3

Let R_0^n be the set of all rational polynomials of the form

$$r_0^k(x) = \frac{p_0(x)}{\prod_{j=1}^{k} (1+b_jx)}, \quad b_j > 0,\ k \leq n,$$

where $p_0(x)$ is a polynomial of degree 0, then

(a) there exists a best approximation to exp(-x) in R_0^n;

(b) the best approximation is completely characterized by

(i) $k = n$,

(ii) $b_1 = b_2 = \dots = b_n$,

(iii) the error curve $r_0^n(x)$-exp(-x) alternates exactly twice;

(c) the best approximation is unique.

5.4 "Sufficiency" in R_m^n

It turns out that the case is much more complicated when we try to extend the results of the last section for any $m > 0$.

The weighted approximation problem

$$\left\| \frac{1}{\prod_{i=2}^{n}(1+b_i x)} \left(\frac{p_m(x)}{1+b_1 x} - \prod_{i=2}^{n}(1+b_i x)\exp(-x) \right) \right\| \tag{5.4.1}$$

has been proved always non-degenerate for $m = 0$. That is, the error curve will alternate exactly $m+2$ times. However, for general $m > 0$, though in practice it is always found that the approximation is also non-degenerate, an analytic proof will be very difficult. Hence, we do not have, for $m > 0$, an analytic result similar to Theorem 5.3.3, although all computational tests do show that the best approximations are in $\underset{\sim}{R}_m^n$.

In fact, it can be shown that the sufficiency condition in Theorem 5.3.3 does not hold for $m > 0$.

By Theorem 2.2.1, we know, for any fixed $b > 0$, there is a polynomial $p_m^*(x)$ of degree at most m, such that

$$\left|\left|\frac{p_m^*(x)}{(1+bx)^n} - \exp(-x)\right|\right|$$

$$= \inf_{p_m(x)\in R_{m,0}} \left|\left|\frac{p_m(x)}{(1+bx)^n} - \exp(-x)\right|\right|. \tag{5.4.2}$$

Define

$$g_{m,n}(b,x) = \frac{p_m^*(x)}{(1+bx)^n} - \exp(-x), \quad x \geq 0,\ b > 0, \tag{5.4.3}$$

and

$$\phi_{m,n}(b) \quad = ||g_{m,n}(b,x)||. \tag{5.4.4}$$

Fig.5.4.1 is the graph of $\phi_{3,5}(b)$. Graphs of $\phi_{m,n}(b)$ for all other valid values of m and n are typically of the same shape. Theorem 2.2.1 guarantees that, for any $b > 0$, $g_{m,n}(b,x)$ alternates at least m+1 times for a nondegenerate $p_m^*(x)$. It is found that at those points where $\phi_{m,n}(b)$ attains its local minimum, $g_{m,n}(b,x)$ alternates one time more, i.e. m+2 times; and only m+1 times for other values of b. Hence, unlike Theorem 5.3.3, a value of b at which $g_{m,n}(b,x)$ alternates m+2 times cannot be a sufficient condition, that $\phi_{m,n}(b)$ is the global minimum. Instead, we would formulate

Conjecture 5.4.1

Let $\underset{=}{R}{}_m^n(b,\varepsilon) \subset \underset{-}{R}{}_m^n$ be defined as

$$\underset{=}{R}{}_m^n(b,\varepsilon) = \{\frac{p_m(x)}{(1+b'x)^n} : |b-b'| < \varepsilon\},\ b > \varepsilon > 0.$$

If the error curve

$$\frac{p_m^*(x)}{(1+b^*x)^n} - \exp(-x)$$

alternates at least m+2 times, then

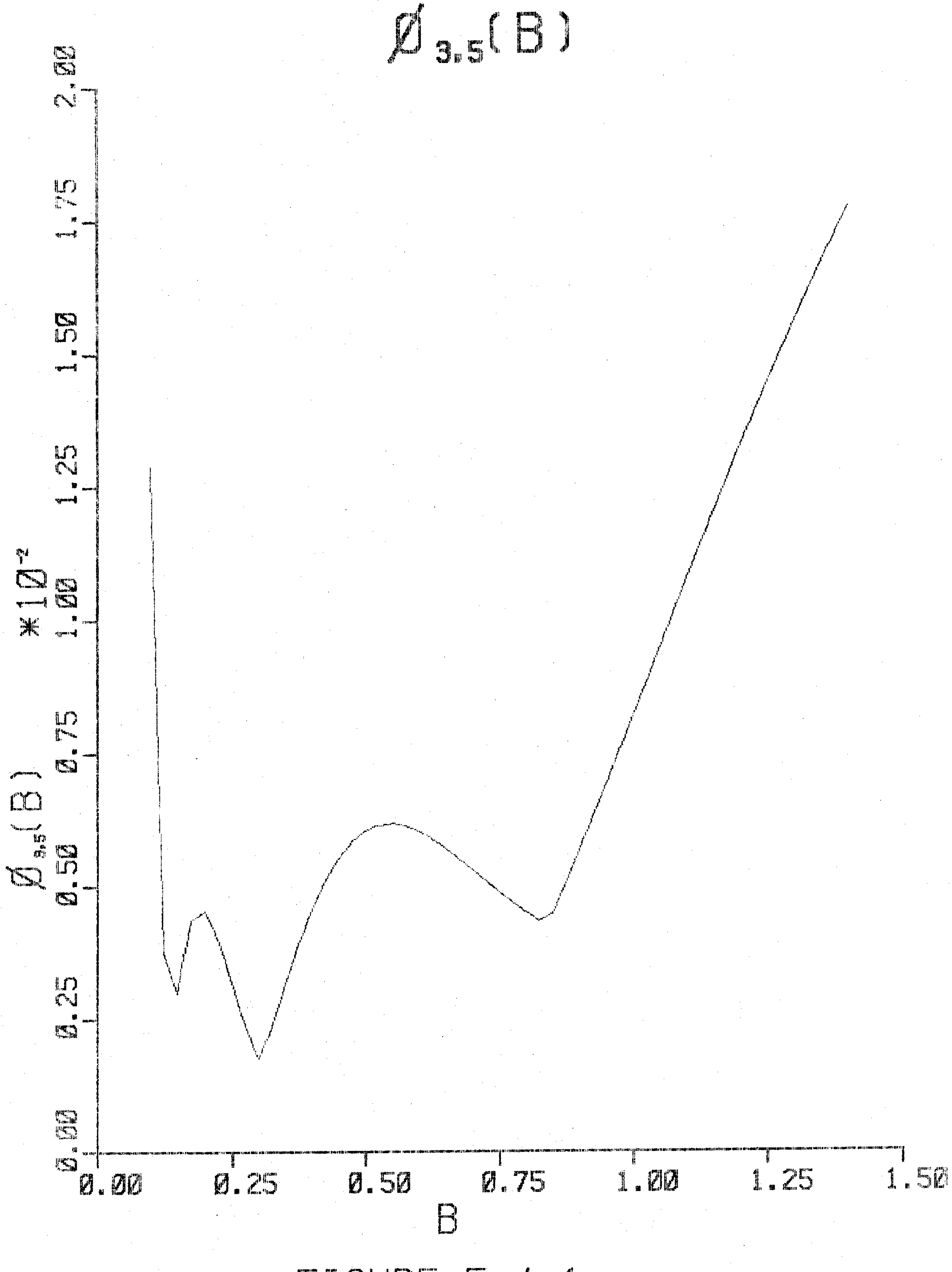
$\varnothing_{3.5}$(B)
$\varnothing_{3.5}$(B) *1Ø$^{-2}$
2.ØØ
1.75
1.5Ø
1.25
1.ØØ
Ø.75
Ø.5Ø
Ø.25
Ø.ØØ
Ø.ØØ
Ø.25
Ø.5Ø
Ø.75
1.ØØ
1.25
1.5Ø
B

FIGURE 5.4.1

$$\frac{p_m^*(x)}{(1+b^*x)^n}$$

is the best approximation to exp(-x) with reference to $\underline{R}_m^n(b^*,\varepsilon)$ for certain $\varepsilon > 0$.

Conjecture 5.4.1 means that an (m+2)-alternation can only be a sufficiency condition that

$$\frac{p_m^*(x)}{(1+b^*x)^n}$$

is a best approximation locally in a certain neighbourhood of b^*. To determine the global minimum for all $b > 0$, it would be necessary to compare the different local minima, just as in the case of finding the global minimum of a polynomial. If $\phi_{m,n}(b)$ attains its global minimum at two different points b_1 and b_2, i.e.

$$0 < \phi_{m,n}(b_1) = \phi_{m,n}(b_2) \leq \phi_{m,n}(b), \qquad b > 0$$

then the best approximation to exp(-x) with reference to $\underline{R}_m^n$ is not unique. Fortunately, all test cases performed so far, for $m = 0,1,\ldots,4$ $n = m,m+1,\ldots,m+7$ show that the global minimum, and hence the best approximation is unique in each case.

The graphs of some $\phi_{m,n}(x)$ have been plotted out numerically, and the following table gives the number of local minima found for each $\phi_{m,n}(x)$.

n \ m	0	1	2	3	4
0	-	-			
1	1	1	-		
2	1	2	2	-	
3	1	2	3	3	-
4	1	2	3	3	4
5	1	2	3	4	4
6		2	3	4	4
7		2	3	4	5
8			3	4	5
9				4	5
10					5

TABLE 5.4.1 Number of minima of $\phi_{m,n}(x)$

Table 5.4.1 seems to suggest that the number of local minima of $\phi_{m,m}$ is m and that of $\phi_{m,n}$ is at most m+1.

An interesting finding in plotting out $\phi_{m,n}(x)$ as Fig.5.4.1 is that, in all cases, the global minimum of $\phi_{m,n}(b)$ is either one of the two local minima that are on the two sides of $b = \frac{1}{n}$. This is consistent with the fact that $\lim_{n\to\infty}(1+ \frac{x}{n})^{-n} = e^{-x}$, and can be a useful guideline in the numerical computation of the global minimax approximation.

5.5 Order-constrained Approximations

In Section 3.3, order-constrained minimax approximation proposed by Lawson is mentioned, and its advantages when applied to the solution of stiff partial differential equations are discussed. It can be shown that many of the results obtained so far in this chapter can be extended readily to the case of order-constrained approximation.

Similar to Definition 2.2.7 is

Definition 5.5.1

Let f(x) be continuous on I_{ab}, and $f(x) \in C^{k-1}$ at $x = a$, then we define

$$R^n_{k,m}(f) \subset R^n_m$$

$$R^n_{k,m}(f) = \{ r^j_{k,m}(x) \in R^n_m :$$

$$\frac{d^i}{dx^i} r^j_{k,m}(x)\Big|_{x=a} = \frac{d^i}{dx^i} f(x)\Big|_{x=a},$$

$$i = 0,\dots,k-1\},\ 1 \le k \le m+1. \tag{5.5.1}$$

Theorem 5.5.1 (Existence theorem)

Among the functions $r^j_{k,m}(x)$, $n \ge j \ge \deg(p_m(x))$, in $R^n_{k,m}$, there is at least one function for which

$$||r^j_{k,m}(x)-\exp(-x)|| \tag{5.5.2}$$

attains its minimum.

Proof The proof is nearly identical with that of Theorem 5.1.1.

It can be shown there exists a convergent subsequence in $R^n_{k,m}$ whose limit function attains the minimum of (5.5.2). Using the same argument, the limit function will be in R^n_m. It will also be in $R^n_{k,m}$ because, for each member of the subsequence, its coefficients are such that they make the coefficients of the first k powers of x in the expression

$$\left(\sum_{i=0}^{\infty} \frac{x^i}{i!}\right) \prod_{i=1}^{j} (1+b_i x) - \sum_{i=0}^{m} a_i x^i$$

vanish. Obviously, the limit of this sequence of coefficients has such a property too. Hence, the limit function is in $R^n_{k,m}$. Q.E.D.

Theorem 5.5.2 (Necessary conditions)

Let the best approximation to exp(-x) in $R^n_{k,m}$ be as in (5.2.22),

$$r^{*n}_{k,m}(x) = \frac{\sum_{i=0}^{m-\mu} a_i x^i}{\sum_{i=1}^{n-\nu} (1+b_i x)}$$

and let its error curve (5.2.23) alternate j times on $[0,\infty)$, then

(i) $j \geq m+2-(\mu+k)$;

(ii) $\nu = 0$ and

$b_1 = b_2 = \ldots = b_n$ if $j = m+2-(\mu+k)$.

Proof The proof is again identical with that of Theorem 5.2.1 except that the function $\Phi(x)$ in Lemma 5.2.1 is replaced by

$$\Phi(x) = x^k \prod_{j=1}^{m+2-k} (x-\beta_j).$$ Q.E.D.

It should be noticed that rational functions in R^n_{-1} discussed in Chapter 4 are in fact order 1 approximations to exp(-x). Hence, for $R^n_{1,0}$, we have

Theorem 5.5.3

The best approximation in $R^n_{1,0}$ is unique and has the form

$$\frac{1}{(1+bx)^n},$$

completely characterised by the fact that the error curve

$$\frac{1}{(1+bx)^n} - \exp(-x)$$

alternates exactly once.

Proof Theorem 4.2.1. Q.E.D.

The above theorem implies that we can also apply the results of section 4.3 to provide an error estimate and analysis to the order-constrained minimax approximation in $R^n_{1,0}$. In particular, the minimax error will have bounds as in Theorem 4.3.4.

CHAPTER 6

NUMERICAL IMPLEMENTATION

6.1 Some Analytic Derivations

In this chapter, we shall describe an algorithm that implements some of the results of Chapters 4 and 5. The algorithm will try to compute a rational polynomial of the form

$$\frac{p_{k,m}(x)}{(1+bx)^n}, \qquad k \geq 0,\ n \geq m \geq 0,\ n > 0, \tag{6.1.1}$$

where $p_{k,m}(x)$ is a polynomial of degree m,

$$\frac{d^i}{dx^i}\left[\frac{p_{k,m}(x)}{(1+bx)^n}\right]_{x=0} = \frac{d^i}{dx^i}(\exp(-x))\Big|_{x=0} \tag{6.1.2}$$

$$i = 0,1,\ldots,k-1,\ 1 \leq k \leq m+1$$

and the error curve

$$\frac{p_{k,m}(x)}{(1+bx)^n} - \exp(-x) \tag{6.1.3}$$

alternates m+2-k times. If k = 0, it means we do not have the conditions in (6.1.2), and so the approximation will not have any order at the origin.

Let (6.1.1) be written in the form

$$(a_0+a_1x+\ldots+a_mx^m)/(1+bx)^n. \tag{6.1.4}$$

If $k \geq 1$, in order to satisfy (6.1.2), it is sufficient that a_i, $i = 0,\ldots,k-1$ agree with the coefficient of the term having the corresponding power in the series expansion of

$$(1+bx)^n \cdot \exp(-x),$$

i.e.

$$a_i = \sum_{j=0}^{i} \binom{n}{j} b^j \cdot \frac{(-1)^{i-j}}{(i-j)!} \quad , \quad i = 0,\ldots,k-1. \tag{6.1.5}$$

Because of the alternating property of (6.1.3), there is a set of m+3-k points $\{x_i\}$, at which (6.1.3) achieves its maximum value $|\lambda|$, i.e.

$$\frac{p_m(x_i)}{(1+bx_i)^n} - \exp(-x_i) = (-1)^{i+1}\lambda, \; i = 1,\ldots,m+3-k. \tag{6.1.6}$$

(6.1.6) can be rearranged to

$$\frac{p_m(x_i)}{(1+bx_i)^{n-1}} - (1+bx_i)\exp(-x_i) = (-1)^{i+1}\lambda(1+bx_i)$$

or

$$\sum_{j=0}^{m} a_j \frac{x_i^j}{(1+bx_i)^{n-1}} -b(x_i \exp(-x_i))-\lambda(-1)^{i+1}(1+bx_i) = \exp(-x_i)$$

$$i = 1,\ldots,m+3-k. \tag{6.1.7}$$

(6.1.7) is a system of equations linear in a_i, $i = 0,\ldots,m$ but not in b. Hence, we need two other auxiliary formulae for the computation of b.

At any extremal point x_i of (6.1.6), if x_i is in $(0,\infty)$, we know

$$\frac{d}{dx}\left[\frac{p_m(x)}{(1+bx)^n} - \exp(-x)\right]_{x=x_i} = 0. \tag{6.1.8}$$

From (6.1.6) and (6.1.8) we can solve for b:

$$\frac{1}{b} = \frac{n}{(\frac{p_m'(x_i)}{p_m(x_i)}) + \frac{1}{1+(-1)^{i+1}\lambda\exp(x_i)}} - x_i, \qquad (6.1.9)$$

where $p_m'(x_i) = \frac{dp_m(x)}{dx}\Big|_{x=x_i}$.

Also, if b', b"are very close to b, we have, at $x = x_i$,

$$\frac{p_m(x_i)}{(1+b'x_i)^n} - (\frac{1+b''x_i}{1+b'x_i})\exp(-x_i) \doteq (-1)^{i+1}\lambda. \qquad (6.1.10)$$

Assuming we have strict equality in (6.1.10) and using (6.1.6), we can again solve for b from

$$(\frac{1+bx_i}{1+b'x_i}) = \left[\frac{(\frac{1+b''x_i}{1+b'x_i})\exp(-x_i)+(-1)^{i+1}\lambda}{\exp(-x_i)+(-1)^{i+1}\lambda}\right]^{1/n} . \qquad (6.1.11)$$

When x_i is large, (6.1.11) can be approximated by

$$\begin{aligned} b &\doteq b'(\frac{b''}{b'})^{1/n} \\ &= b'[1+ \frac{b''-b'}{b'}]^{1/n} \\ &\doteq b'[1+ \frac{(b''-b')}{nb'}]. \qquad (6.1.12) \end{aligned}$$

We now propose

$$\frac{a_0^{(j+1)}+a_1^{(j+1)}x+\ldots+a_m^{(j+1)}x^m}{(1+b_h^{(0)}x)^n} - \frac{(1+b_h^{(j+1)}x)}{(1+b_h^{(0)}x)}\exp(-x). \tag{6.2.5}$$

When convergence is near, say

$$\left|\frac{b_h^{(j+1)}-b_h^{(0)}}{b_h^{(0)}}\right| < 50\varepsilon, \tag{6.2.6}$$

the coefficient term of exp(-x) in (6.2.5) can be replaced by 1.

Step 2 Using the latest values obtained in step 1

$$a_0^{(j+1)},\ldots,a_m^{(j+1)},b_h^{(0)},b_h^{(j+1)},\lambda^{(j+1)},$$

Compute a new b^* by one of the methods below:

2.1 Using the two vectors $(b_{h-1}^{(0)},b_{h-1}^{(j_{h-1}+1)})$ and

$(b_h^{(0)},b_h^{(j+1)})$, extrapolate linearly (b^*,b^*).

2.2 (Only at the start of the algorithm, or when 2.1 gives a $b^* \leq 0$).

With the help of (6.1.12), compute

$$b^* = b_h^{(0)}[1+\frac{b_h^{(j+1)}-b_h^{(0)}}{mb_h^{(0)}}] .$$

2.3 (Only if both 2.1 and 2.2 give a negative b^*.) Compute b^* according to (6.1.9) at an extremal points $x_i^{(j+1)}$, preferably the middle one.

If all three methods give a negative b^*, the algorithm fails.

Step 3 If the number of iterations exceeds a given limit, or

$|b_k^{(0)}-b_k^{(j+1)}| < \varepsilon$ (a stopping criterion)

terminate the algorithm; otherwise, set:

$h \leftarrow h+1;$ $x_i^{(0)} \leftarrow x_i^{(j+1)}$, $i = 1,2,\ldots,m+3-k;$

$j \leftarrow 0;$ $b_h^{(0)} \leftarrow b^*;$ Go back to Step 1.

In Appendix I, a listing of an ALGOL program implementing the above algorithm and tested on the Honeywell 6050 of the Mathematics Faculty, University of Waterloo can be found.

It should be noted that we can use, as an alternative to Algorithm 6.2, the more general algorithm studied by Barrodale, et.al. in [55] for computing best l_∞ approximations by functions nonlinear in one parameter. The idea of the algorithm is similar to that used in the plotting of Fig.5.4.1. Since an optimization technique like the Fibonacci method is used there to obtain the minima, their method guarantees convergence, although it will require more iterations than Algorithm 6.2.

6.3 Numerical Results: Unconstrained Case

When k = 0, the algorithm converges in nearly all cases to produce that rational function of form (6.1.1) whose error curve alternates m+2 times and whose value of b is that local minimum of $\phi_{m,n}(b)$ (5.4.4) nearest to the initial guess of b. Divergence occurs usually only when the initial guess b is too large or the resulting approximation is near-degenerate, i.e. some of the extremal points cluster very close together. The latter case can be remedied by a better initial guess of the critical points.

By using different starting values of b, we can compute all the local minima or at least the few smallest of $\phi_{m,n}(b)$, from which we can choose the best approximation in $\underset{\sim}{R}_m^n$. As mentioned in section 5.4, in all cases tested, the globally best approximation is at either b_1 or b_2, where

$$b_1 < \frac{1}{n} < b_2$$

and these are two local minima nearest to n^{-1}.

Let the globally best approximation in $\underset{\sim}{R}_m^n$ be

$$\underset{\sim}{r}_m^n = \frac{p_m(x)}{(1+bx)^n} \ . \qquad (6.3.1)$$

We know that if the best approximation in R_m^n does not degenerate, then it will be exactly (6.3.1). Searches have been done throughout some R_m^n and they all confirmed that the best approximations in those R_m^n do have the form (6.3.1). Let

$$e_{m,n} = ||\underset{\sim}{r}_m^n - \exp(-x)||. \qquad (6.3.2)$$

In Table 6.3.1, the values of $e_{0,n}$, $b_{0,n}$ and the theoretical error bounds computed according to Lemma 4.3.4 are listed. In all cases, $e_{0,n}$ are within the error bound. As $n \to \infty$, the column $(n \cdot e_{0n})$ tends to a constant, showing that the error is inversely proportional to n, as predicted in Theorem 4.3.4.

Table 6.3.2 tabulates the value of $e_{m,n}$ for various m,n. Fig.6.3.1 plots -log e_{mn} against log n for various fixed m. It is obvious immediately that the function -log e_{mn} is nearly linear in log n for each m. That is, we can describe $e_{m,n}$ and n by

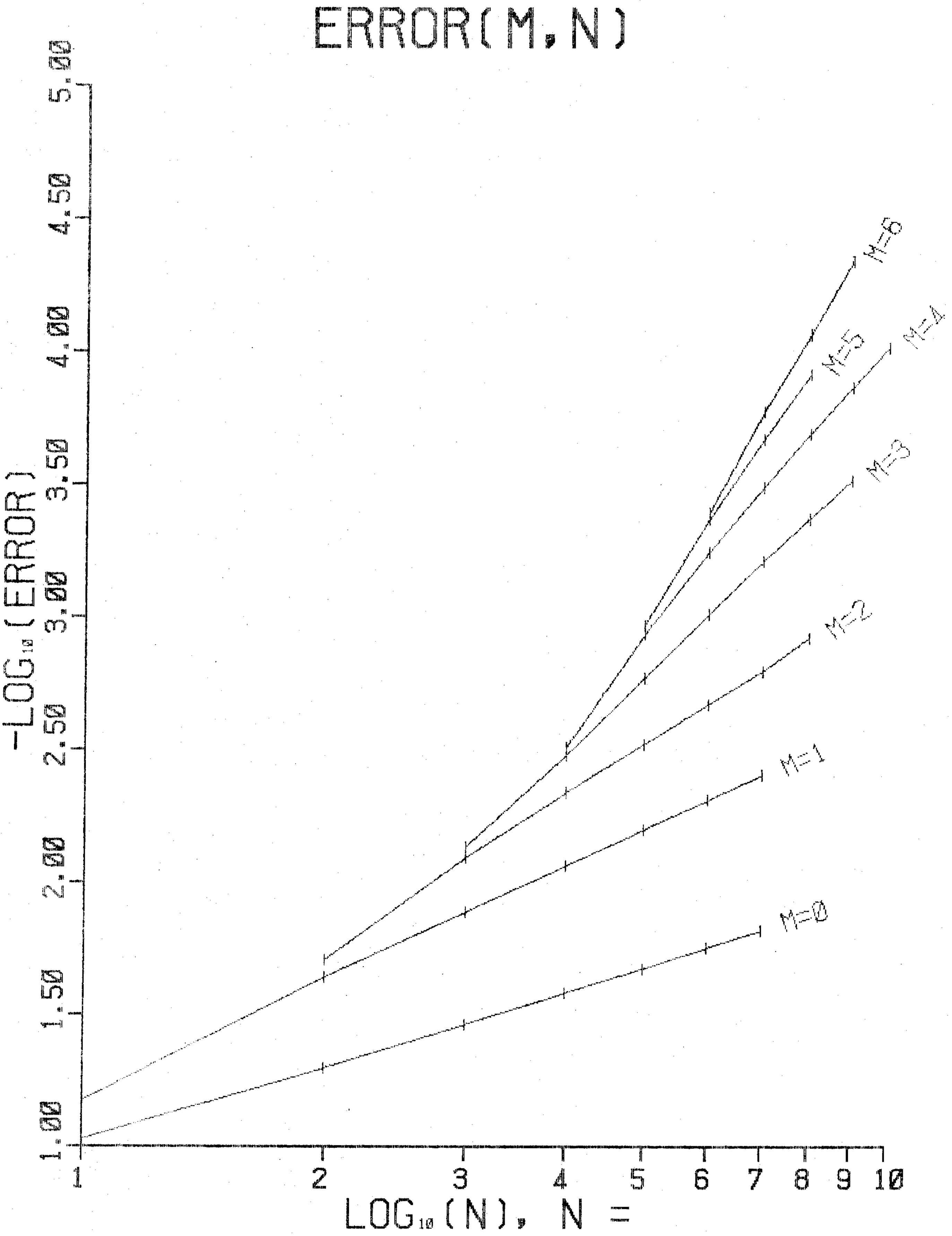
ERROR(M,N)
-LOG₁₀(ERROR)
1.00
1.50
2.00
2.50
3.00
3.50
4.00
4.50
5.00
M=6
M=5
M=4
M=3
M=2
M=1
M=Ø
1
2
3
4
5
6
7
8
9
10
LOG₁₀(N), N =

FIGURE 6.3.1

TABLE 6.3.1

n	e_{0n}	b_{0n}	error bound	$n.e_{0n}$
1	0.09357	2.2397	0.2036	0.094
2	0.05037	0.7485	0.1161	0.101
3	0.03442	0.4363	0.0812	0.102
4	0.02614	0.3060	0.0625	0.104
5	0.02107	0.2351	0.0508	0.105
6	0.01764	0.1907	0.0427	0.106
7	0.01517	0.1604	0.0369	0.106
8	0.01331	0.1383	0.0325	0.106
9	0.01186	0.1216	0.0290	0.106
10	0.01069	0.1084	0.0262	0.107
20	0.00538	0.0521	0.0133	0.108
80	0.00135	0.0126	0.0034	0.108
320	0.00034	0.0031	0.0008	0.108

$$-\log e_{mn} = c_m \log n + d_m, \tag{6.3.3}$$

which means

$$e_{mn} = \frac{k_m}{n^{c_m}}, \tag{6.3.4}$$

where k_m and c_m are constants for each m. As m increases, the gradient of (6.3.3) becomes steeper and hence c_m increases with m. So, for large m, the

e_{mn} decreases faster with n. Table 6.3.3 lists out some values of c_m and k_m.

Another useful observation is that as $m \to \infty$,

$$e_{m-1,m} \doteqdot e_{m,m} \; .$$

Furthermore, with m fixed, the decrease of $e_{m,n}$ with reference to n is fastest for the first few values of n. Hence, given due consideration to computational complexity and degree of approximation, the most recommended approximation to exp(-x) would be that in $\underset{\sim}{R}_m^{m+2}$, i.e. rational function of the form

$$\frac{p_m(x)}{(1+bx)^{m+2}} \; .$$

Some of the best approximations to exp(-x) in $\underset{\sim}{R}_m^n$ are listed in Appendix II.

6.4 Numerical Results: Constrained Case

The case for exponential approximation of the form

$$\frac{p_{k,m}(x)}{(1+bx)^n} \; , \quad k \geq 1,$$

with order k at the origin is very similar to the case with no order constraint.

As in section 5.4, for any m,n,k, given a fixed b, let

$$\frac{p^*_{k,m}(x)}{(1+bx)^n}, \quad m+1 \geq k \geq 1$$

n	$e_{0,n}$	$e_{1,n}$	$e_{2,n}$	$e_{3,n}$	$e_{4,n}$	$e_{5,n}$	$e_{6,n}$
0	0.5000						
1	0.0936	0.0668					
2	0.0504	0.0227	0.0195				
3	0.0344	0.0129	0.0080	0.0073			
4	0.0261	0.0086	0.0046	0.0033	0.00309		
5	0.0211	0.0063	0.0030	0.0017	0.00116	0.00107	
6	0.0176	0.0049	0.0021	0.0010	0.00057	0.00043	0.000401
7	0.0152	0.0039	0.0016	0.0006	0.00032	0.00021	0.000169
2^m	9.4'-2	2.3'-2	4.6'-3	4.2'-4	2.3'-5	5.1'-7	7.7'-9

TABLE 6.3.2 Minimax error $e_{m,n}$

m	c_m	k_m
0	1.0	1.1
1	1.4	0.06
2	1.9	0.07
3	2.9	0.17
4	3.5	0.27
5	4.5	1.42
6	5.1	3.40

TABLE 6.3.3 c_m, k_m in (6.3.4)

be an order-k approximation to exp(-x) which attains the minimum of

$$||\frac{p_{k,m}(x)}{(1+bx)^n} - \exp(-x)|| \tag{6.4.1}$$

for all $p_{k,m}(x)$ which makes

$$p_{k,m}(x)/(1+bx)^n$$

an order-k approximation to exp(-x). Define

$$g_{k,m,n}(b,x) = \frac{p^*_{k,m}(x)}{(1+bx)^n} - \exp(-x) \tag{6.4.2}$$

and

$$\phi_{k,m,n}(b) = ||g_{k,m,n}(b,x)||. \tag{6.4.3}$$

It should be noted that, with reference to (5.4.3) and (5.4.4),

$$g_{o,m,n}(b,x) \equiv g_{m,n}(b,x),$$

and $$\phi_{o,m,n}(b) \equiv \phi_{m,n}(b).$$

Fig.6.4.1 is the graph of $\phi_{k,3,5}(b)$ for k = 1,2,3,4. All of them look like $\phi_{3,5}(b)$ in Fig.5.4.1. Again, it is found that at each local minimum of $\phi_{k,m,n}(b)$, the error curve $g_{k,m,n}(b,x)$ alternates m+2-k times while at other points, only m+1-k times.

From Fig.6.4.1, it is obvious that there is a correspondence between the minima of each $\phi_{k,m,n}(b)$. All $\phi_{k,m,n}$, k = 0,1,2,...,m+1 seem to have the same number of local minima. However, for high order-

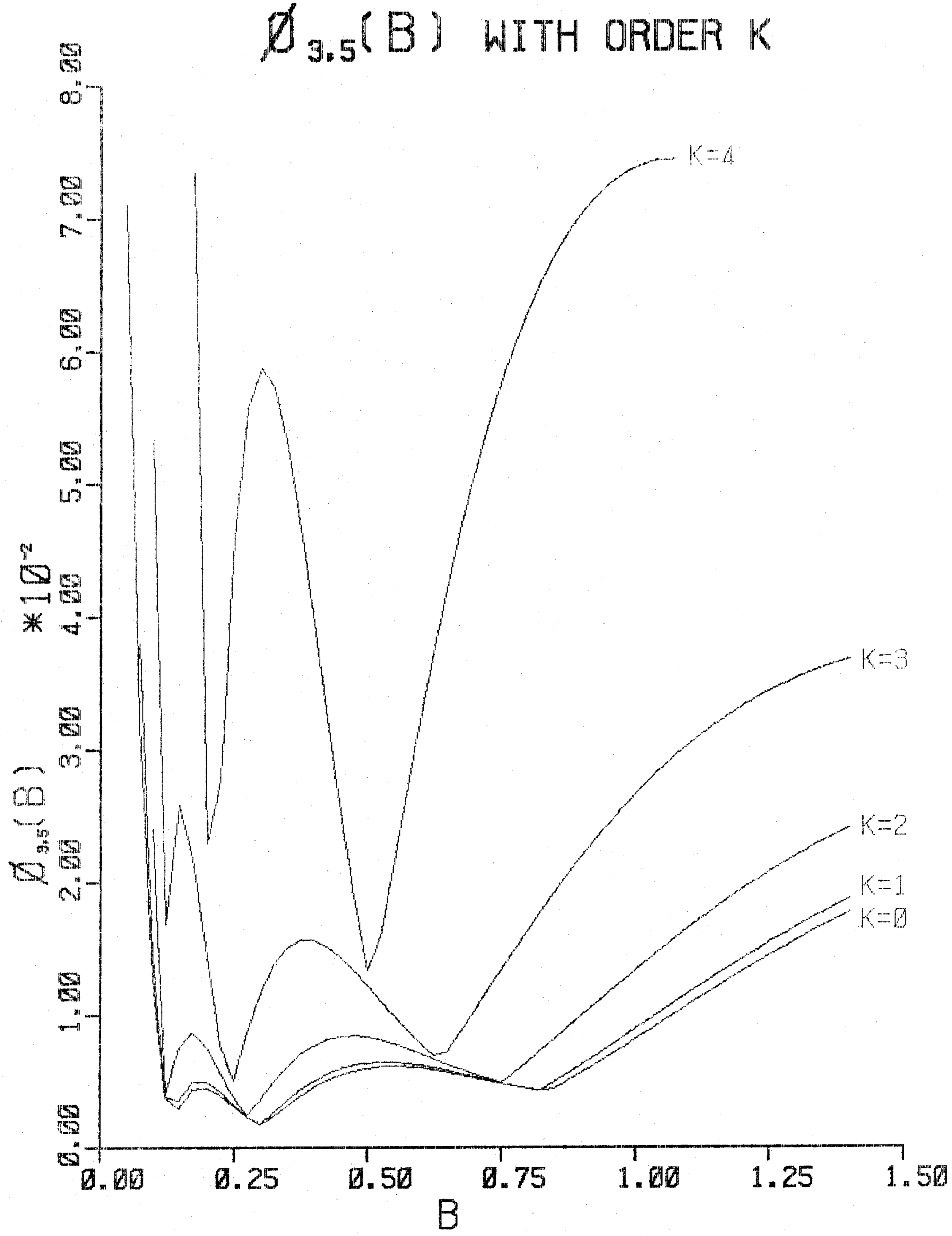
Ø3.5(B) WITH ORDER K
8.ØØ
7.ØØ
6.ØØ
5.ØØ
4.ØØ
3.ØØ
2.ØØ
1.ØØ
Ø.ØØ
Ø3.5(B) *1Ø-2
Ø.ØØ
Ø.25
Ø.5Ø
Ø.75
1.ØØ
1.25
1.5Ø
B
K=4
K=3
K=2
K=1
K=Ø

FIGURE 6.4.1

constrained cases, the globally best approximation may not be at that value of b nearest to n^{-1}. Furthermore, when k is large, the slope of $\phi_{k,m,n}(b)$ varies in a very drastic manner. Hence, when Algorithm 6.2 is applied to compute high order-constrained approximation, convergence is much more difficult unless a good starting initial guess of b is used.

Let us define, as in (6.3.2),

$$e_{k,m,n} = ||g_{k,m,n}(b)||, \quad b > 0 \tag{6.4.4}$$

In Table 6.4.1 are some values of $e_{k,m,n}$, providing us an idea how order-constrained approximations compare with the unconstrained ones (k = 0).

order k	m=3,n=5	m=4,n=6,	m=5,n=7
0	1.70'-3	5.69'-4	2.13'-4
1	1.80'-3	6.03'-4	2.26'-4
2	2.38'-3	8.00'-4	2.99'-4
3	4.26'-3	1.39'-3	4.69'-4
4	1.20'-2	2.85'-3	8.25'-4

TABLE 6.4.1 $e_{k,m,n}$

Some order-constrained approximations computed by Algorithm 6.2 are listed in Appendix II.

CHAPTER 7

APPLICATION IN STIFF SYSTEMS

In this chapter, we shall try to compare a $\underset{-}{R}_m^n$-approximation having the form

$$\frac{p_m(x)}{(1+bx)^n}$$

with other conventional exponential approximations whenthey are applied to the solution of stiff differential equations.

Consider the differential system

$$\bar{y}' = -A\bar{y}(x),$$

$$\bar{y}(0) = \bar{y}_0, \quad x \geq 0. \tag{7.0.1}$$

We assume that A has widely separated positive real eigenvalues.

We know the exact analytic solution of (7.0.1) is

$$\bar{y}(x) = \exp(-xA)\cdot\bar{y}_0. \tag{7.0.2}$$

If x is discretised by a step size h, such that,

$$x_0 = 0,$$

$$x_k = x_{k-1} + h, \quad k = 1,2,\ldots, \tag{7.0.3}$$

we have,

$$\bar{y}(x_k) = \exp(-hA)\cdot\bar{y}(x_{k-1}). \tag{7.0.4}$$

As discussed in Chapter 3, many numerical methods designed for the solution of stiff systems are based on the idea of replacing the matrix exponential in (7.0.4) by a rational approximation, say,

$$r_{m,n}(x) = [q_n(x)]^{-1}\cdot[p_m(x)], \tag{7.0.5}$$

where $p_m(x)$ and $q_n(x)$ are polynomials with real coefficients of degree m and n respectively. Hence, the exponential function in (7.0.4) will be approximated by

$$[q_n(hA)]^{-1}\cdot[p_m(hA)]. \tag{7.0.6}$$

In this manner, the (m,n)-th approximation $\bar{y}_k$ of $\bar{y}(x_k)$ is defined, in analogy with (7.0.4), for all $x \geq 0$ as

$$\begin{aligned} \bar{y}_k &= [q_n(hA)]^{-1}\cdot[p_m(hA)]\cdot\bar{y}_{k-1}, \quad k = 1,\ldots, \\ \bar{y}_0 &= \bar{y}(0). \end{aligned} \tag{7.0.7}$$

Cavendish, et al. [9] proposed to solve (7.0.7) as follows:

The fundamental theorem of algebra permits factorization of $q_n(x)$ and $p_m(x)$ into products of linear and quadratic polynomials with real coefficients, hence (7.0.7) can be represented in the alternate form

$$[\prod_{i=1}^{n_q} V_i(hA)]\cdot\bar{y}_k = [\prod_{i=1}^{m_p} W_i(hA)]\cdot\bar{y}_{k-1}, \tag{7.0.8}$$

where V_i and W_i are matrix polynomials of degree one or two. If we define the right-hand side of (7.0.8) to be $\bar{X}_0$, the solution $\bar{y}_k$ can be obtained by solving successively the n_q sets of linear equations for $\bar{X}_i$:

$$[V_i(hA)]\bar{X}_i = \bar{X}_{i-1}, \; i = 1,2,\ldots,n_q, \tag{7.0.9}$$

and the final $\bar{X}_{n_q} = \bar{y}_k$.

However, to obtain the answer in this manner can be numerically unstable because, if $\bar{y}_{n-1}$ is contaminated by an error $\bar{e}$ in the direction of the eigenvector that corresponds to the largest eigenvalue of A, the error involved in computing $\bar{X}_0$ would be many times that of $\bar{e}$, especially when we try to use a large step-size h. With an incorrect $\bar{X}_0$, we would not expect the final answer to be accurate after solving the n_q systems of linear equations.

An alternate method is to express the rational approximation as a partial fraction, such that (7.0.8) will look like

$$\bar{y}_k = \sum_{i=1}^{n_q} [U_i(hA)]^{-1}[T_i(hA)]\cdot\bar{y}_{k-1}, \qquad (7.0.10)$$

where $U_i(hA)$ are matrix polynomials ofdegree one or two and $T_i(hA)$ are matrix polynomials ofat least one degree less than $U_i(hA)$. Each term of the sum in (7.0.10) can be computed separately and these, together, give the final answer $\bar{y}_n$. Furthermore, (7.0.10) will be computationally more economical than (7.0.8).

We know when confined to the real axis, $\underset{-}{R}_m^n$-approximation is A-acceptable when m = n and L-acceptable when m < n. So, for a matrix with real eigenvalues, the numerical method will be A-stable. We now try to compare a $\underset{-}{R}_m^n$-approximation with an ordinary rational approximation with reference to (7.0.10).

Let the $\underset{-}{r}_m^n(x)$ and $r_{m,n}(x)$ exponential approximation have the following partial fraction representations respectively:

$$\underline{r}^n_m(x) \;=\; \sum_{i=1}^{n} \frac{a_i}{(1+bx)^i} \;, \qquad (7.0.11)$$

$$r_{m,n}(x) = \sum_{i=1}^{n_q} \frac{T_i(x)}{U_i(x)} \,. \qquad (7.0.12)$$

7.1 Operation Counts

In operation counts, we shall consider only multiplications and divisions for simplicity. We shall assume the order of the matrix A is d and, for the sake of simplicity, take only the term having the highest power of d in all counting.

In general, for an ordinary rational approximation (Chebyshev, Padé or order-constrained), all the $U_i(x)$ in (7.0.12) are quadratic, except one of them when n is odd. Hence, we shall get more or less the same result in the counting if we assume each term of (7.0.12) is of the form

$$\frac{T_i(x)}{U_i(x)} = \frac{t_{i,0}+t_{i,1}x}{1+u_{i,1}x+u_{i,2}x^2} \,. \qquad (7.1.1)$$

When applied to (7.0.10), it means we have to solve $n_q = n/2$ linear systems of the form

$$\bar{y}_{k_i} = (I+u_{i,1}hA+u_{i,2}h^2A^2)^{-1}\cdot(t_{i,0}+t_{i,1}hA)\cdot\bar{y}_{k-1},$$

$$i = 1,\ldots,n/2, \qquad (7.1.2)$$

which is equivalent to solving for $\bar{y}_{k_i}$ from

$$(I+bhA)\bar{y}_{k,1} = \bar{y}_{k-1},$$

$$(I+bhA)\bar{y}_{k,i} = \bar{y}_{k,i-1}, \ i = 2,\dots,n,$$

$$\bar{y}_k = \sum_{i=1}^{n} a_i \ \bar{y}_{k,i}, \qquad (7.1.5)$$

which is just n linear systems withdifferent right-hand side. It should be noted that the left-hand side of (7.1.5) is a tri-diagonal matrix compared with a five-diagonal matrix in (7.1.3).

In starting, it requires 2d operations to compute

$$(I + bhA) \qquad (7.1.6)$$

and 2d operations to perform a LDL^T decomposition on (7.1.6). Hence, the starting cost is 4d operations.

At each step k, for the n linear systems of (7.1.5), we need 3nd operations to solve for $\bar{y}_{k,i}$, i = 1,...,n and nd operations to compute the final $\bar{y}_k$, a total cost of 4nd operations.

Hence, both methods require about the same operation cost per step. However, using the $\underset{\sim}{R}_m^n$-approximation, the starting cost to compute A^2, the LDL^T decomposition etc. will be much reduced, by a factor of 2/3(n+1). Results for a band-matrix A will be similar.

If (7.0.1) is non-constant, i.e., A is a function of time, then LDL^T will have to be performed at every time-step. Hence, using the $\underset{\sim}{R}_m^n$-approximation will give further savings in operation cost.

7.2 Storage Requirement

When $\underset{\sim}{R}_m^n$-approximation is used, the matrix storage requirements are those for the matrix A, and the LDL^T decomposition of (1+bhA). However, for methods using ordinary approximations, we need the extra storage for A^2, $(t_{i,0}+t_{i,1}hA)$, and the LDL^T decomposition of $(1+u_{i,1}hA+u_{i,2}h^2A^2)$, $i = 1,2,\ldots,n/2$, (cf. (7.1.3)). This means $\underline{R}^n_m$-approximation allows a saving of about n matrix-core-requirements.

Such a saving can be practically very significant when A is large and dense. If it happens that it is not possible to store all the n/2 LDL^T decompositions,the decomposition will have to be performed at every step. In that case, $\underline{R}^n_m$-approximation will be far superior because it needs only one LDL^T decomposition while ordinary approximation will require n/2 every step. Storing the decomposition and A^2 is unnecessary for non-constant system when A changes at every time-step.

7.3 Numerical Stability

Let the spectral norm of a matrix M be defined as

$$||M|| = \max_i [\lambda_i(MM^T)]^{\frac{1}{2}}, \qquad (7.3.1)$$

where the notation $\lambda_i(MM^T)$ denotes an eigenvalue of MM^T. We shall also need to use the spectral radius, which is defined as

$$\rho(M) = \max_i |\lambda_i(M)|. \qquad (7.3.2)$$

Thus $$||M|| = [\rho(MM^T)]^{\frac{1}{2}}, \qquad (7.3.3)$$

and when M is symmetric,

$$||M|| = \rho(M). \qquad (7.3.4)$$

We shall assume that all matrices involved are symmetric for the discussion of the rest of this section.

We now proceed to perform an error analysis on the solving of the two different linear systems (7.1.3) and (7.1.5).

Let us denote the exact solution, the computed solution and their difference in (7.1.3) by $\bar{y}_{k_i}$, $\bar{y}'_{k_i}$ and $\bar{\varepsilon}_{k_i}$ respectively. This means

$$\bar{y}'_{k_i} = \bar{y}_{k_i} + \bar{\varepsilon}_{k_i} \quad . \tag{7.3.5}$$

From (7.0.10), we know

$$\bar{y}_{k+1} = \sum_{i=1}^{n/2} \bar{y}_{k_i} ,$$

and hence,

$$\bar{y}'_{k_i} = \bar{y}_k + \sum_{i=1}^{n/2} \bar{\varepsilon}_{k_i} . \tag{7.3.6}$$

From classical error analysis of matrix equations, the error in the solution of a linear system can be roughly bounded by

$$||\delta\bar{x}|| \leq \mathcal{O}(||A||\cdot||A^{-1}||)\cdot||\bar{x}||\cdot||\delta A|| . \tag{7.3.7}$$

If the matrix norm used is the spectral norm and A is symmetric, (7.3.7) can be written alternatively as

$$||\delta\bar{x}|| \leq \mathcal{O}(\rho(A)\cdot\rho(A^{-1}))\cdot||\bar{x}||\cdot||\delta A|| ,$$

or
$$||\delta\bar{x}|| \leq \mathcal{O}(\frac{|\lambda(A)|_{max}}{|\lambda(A)|_{min}}) \cdot||\bar{x}||\cdot||\delta A|| . \tag{7.3.8}$$

For each linear system in (7.1.3), if A has widely separated, positive real eigenvalues,

$$|\lambda(I+u_{i,1}hA+u_{i,2}h^2A^2)|_{\max} = \mathcal{O}(|\lambda(h^2A^2)|_{\max})$$

$$= \mathcal{O}(|\lambda(hA)|^2_{\max}),$$

and
$$|\lambda(I+u_{i,1}hA+u_{i,2}h^2A^2)|_{\min} = \mathcal{O}(|\lambda(I)|)$$

$$= \mathcal{O}(1). \qquad (7.3.9)$$

Using (7.3.5), (7.3.6) and (7.3.8), we have

$$||\bar{y}'_{k+1}-\bar{y}_{k+1}|| = \sum_{i=1}^{n/2} ||\bar{\varepsilon}_{k_i}||$$

$$\leq (\max_i||\delta(I+u_{i,1}hA+u_{i,2}h^2A^2)||)\cdot(\sum_{i=1}^{n/2}||\bar{y}_{k_i}||)\cdot\mathcal{O}(|\lambda(hA)|^2_{\max}). \quad (7.3.10)$$

Turning our attention to (7.1.5), we first notice that, since $b > 0$, $h > 0$ and all $\lambda_i(A)$ have positive real parts,

$$|\lambda(I+bhA)|_{\max} = \mathcal{O}(|\lambda(hA)|_{\max}), \qquad (7.3.11)$$

and

$$1 < |\lambda(I+bhA)|_{\min} = \mathcal{O}(1). \qquad (7.3.12)$$

This means, all $\bar{y}_{k,i}$, when computed according to (7.1.5), will be bounded. In addition,

$$\bar{y}'_{k,1} = \bar{y}_{k,1} + \bar{\varepsilon}_{k,1},$$

$$\bar{y}'_{k,i} = (I+bhA)^{-1}(\bar{y}_{k,i-1} + \bar{\varepsilon}_{k,i-1}) + \bar{\varepsilon}'_{k,i}$$

$$= \bar{y}_{k,i-1} + (I+bhA)^{-1}\,\bar{\varepsilon}_{k,i-1} + \bar{\varepsilon}'_{k,i}$$

$$= \bar{y}_{k,i-1} + \bar{\varepsilon}_{k,i}, \quad i = 2,\ldots,n, \qquad (7.3.13)$$

where $$\bar{\varepsilon}_{k,i} = (I+bhA)^{-1}\cdot\bar{\varepsilon}_{k,i-1} + \bar{\varepsilon}'_{k,i}$$

$$= \sum_{j=1}^{i} [(I+bhA)^{-1}]^{i-1}\cdot\bar{\varepsilon}'_{k,j}, \quad i = 2,\ldots,n,$$

$$\bar{\varepsilon}'_{k,1} = \bar{\varepsilon}_{k,1}, \tag{7.3.14}$$

and $\bar{\varepsilon}'_{k,i}$ is the computational error incurred when each linear system in (7.1.5) is solved. Because of (7.3.12), we know, for all i,

$$|\lambda_i(I+bhA)^{-1}| < 1,$$

and hence

$$||\bar{\varepsilon}_{k,i}|| \le O(1)\cdot\sum_{j=1}^{i} ||\bar{\varepsilon}'_{k,j}||. \tag{7.3.15}$$

From (7.3.8), (7.3.11) and (7.3.12), (7.3.15) gives

$$||\bar{\varepsilon}_{k,i}|| \le ||\delta(I+bhA)||\cdot(\sum_{j=1}^{i} ||\bar{y}_{k,i}||)\cdot O(|\lambda(hA)|_{max}). \tag{7.3.16}$$

Putting back to (7.3.13), and using (7.1.15), we get

$$||\bar{y}'_k - \bar{y}_k|| \le \sum_{i=1}^{n} |a_i| \; ||\bar{y}'_{k,i} - \bar{y}_{k,i}||$$

$$\le O(|\lambda(hA)|_{max})\cdot||\delta(I+bhA)||\cdot\sum_{i=1}^{n}(|a_i|\cdot\sum_{j=1}^{i}||\bar{y}_{k,i}||). \tag{7.3.17}$$

Comparing (7.3.10) and (7.3.17), we notice that the error in solving (7.1.3) differs from that in solving (7.1.5) by a factor of about

$$\frac{(n/2)\; O(|\lambda(hA)|^2_{max})}{(n^2/2)\; O(|\lambda(hA)|_{max})} = (\frac{1}{n})\; O(|\lambda(hA)|_{max}). \tag{7.3.18}$$

$E_{m,m}$ converges much faster than $e_{m,m}$; and hence is a better approximation.

However, because of the particular form of the $\underline{r}_m^n$ - approximations, the denominator of an $\underline{r}_m^{2^m}$ - approximation can in fact be computed in m multiplication steps. Therefore, if we measure the efficiency of the approximations in term of work done, a useful comparison is that between $E_{m,m}$ and $e_{m,2^m}$. Table 7.4.1 shows that the latter does converge much faster than $E_{m,m}$. This means it can be a more efficient scalar approximation to the exponential function than the Chebyshev approximation. Futhermore, it involves a smaller number of coefficients and therefore reduces the size of necessary computer storage.

Since the results of the last three sections can be applied directly to the comparison between $r_{m,m}$- and $\underline{r}_m^m$- approximations in the solving of stiff systems, we confine our attention here only to the comparison between $r_{m,m}$- and $\underline{r}_m^{2^m}$ - approximations. Using the latter, (7.1.5) becomes a set of 2^m linear systems as compared to only m/2 linear systems in (7.1.2)-(7.1.3). This implies a tremendous amount of work which can quite easily offset the saving gained during the LDL^T decomposition stage. Nevertheless, despite the big number of linear factors, they are identical and there is still a saving of storage area when constant system is being solved. Also, (7.3.18) becomes

$$\frac{m}{2^{2m}} O(|\lambda(hA)|_{max}) ,$$

which means the conditions of the two systems (7.1.3) and (7.1.5) may become about the same if m is large.

In conclusion, though $\underline{r}_m^{2^m}$-approximations are better approximations to the exponential function than $r_{m,m}$-approximations and in general $\underline{r}_m^n$-approximations have their advantages over $r_{m,n}$-approximations when applied to the solution of stiff systems, we have to be scrupulous when we try to increase the power of the denominator alone of the $\underline{r}_m^n$-approximation. Given a fixed m, it will be an interesting problem to find out, for what value of n we shall have the best choice of an $\underline{r}_m^n$-approximation for the solution of stiff systems of differential equations.

APPENDIX I

Implementation of Algorithm 6.2
in ALGOL

```
approx:
'begin'
'integer' l,m,order,nn,doj,itermax,scode,pcode;
'extended real' xlb,aa,bb,eps,btry,oldbtry;
'extended real''array'p[1:30],q[1:30];

'extended real' 'procedure' f(x);
'value' x; 'extended real' x;
'begin' f:=exp(-x); 'end' f;

'extended real' 'procedure' phi(x);
'value' x; 'extended real' x;
'begin' phi:=x/(1.0-x); 'end' phi;

'procedure' chebyshev
          (f,phi,l,m,order,eps,p,btry,xlb,alarm,itermax,scode,pcode);
'value' l,m,order, eps; 'integer' l,m,order,itermax,scode,pcode;
'extended real'  eps,xlb ; 'label' alarm;
'extended real' 'procedure' f,phi;'extended real''array' p ;
'begin' 'integer' i,itno,lk,larm,lp1,lp2,n,np1,itest,lk1,lpm2,mp1,mm1;
     'extended real' dn,dnp1,h1,sum,test,u,y,y2,y3,z,z1,z2,z3;
     'extended real' bbold,bold,evanew,evaold,expcon,btryold,eps20,
                     bdel,bbdel,evadel;
     'extended real''array' a[1:l+3,1:l+3],signum[1:l+4],b,err,h,
             xtest,x,xval[1:l+3];
     'comment'  this procedure determines the best fit (in the
                chebyshev sense) rational approximation r(x) of
                degree l (numerator) and m (denominator), with order
                lk at the origin, x=0,
                              p0+p1*x+........+pl*x**l
                of the form   ----------------------
                                    (1+b*x)**m
                to the function exp(-x) over the interval (0,+infinity).
                r(x) minimizes the maximum modulus of the
                error function del(x)=(r(x)-exp(-x)).
      'comment' the maximum  error of the best fit approxi-
                mation, xlb,is found with a relative error<eps.;
      'comment' l - degree of numerator polynomial, l => 0,
                m - degree of denominator polynomial, m > 0, m => l,
                order - order of approximation, order => o, s.t.
                     0, no order.
                p - computed coefficients of numerator polynomial.
                btry - input: initial guess of b,
                       output: computed b.
                xlb  - input: initial guess of max error,
                       output: max error of computed approximation.
                alarm - exit if algorithm diverges.
                itermax - max number of iterations allowed.
                scode - 1, if initial guess of extremal points to be
                           computed by routine, otherwise they have
```

```
                             to be supplied as inputs.
                 pcode - 1, if intermediate iteration printout
                          is not needed;


 'procedure' matinv;
 'begin' 'integer' i,j,k,m,nn,np1; 'extended real' temp,templ;
        np1:=lpm2;
       'for' nn:=1 'step' 1 'until' np1 'do'
       'begin' 'comment' partial pivoting;
            temp:=abs(a[nn,nn]); l:=nn;
            'for' k:=nn+1 'step' 1 'until' np1 'do'
            'begin' templ:=abs(a[k,nn]);
                 'if' templ'gr' temp 'then'
                 'begin' l:=k; temp:=templ;
                 'end';
            'end';
            'if' l'nq'nn 'then'
            'begin' 'for' m:=nn 'step' 1 'until' np1 'do'
                 'begin' temp:=a[nn,m]:a[nn,m]:=a[l,m];
                         a[l,m]:=temp;
                 'end';
            temp:=b[nn]:b[nn]:=b[l]; b[l]:=temp;
            'end';
            b[nn]:=b[nn]/a[nn,nn];
            'for' j:=np1 'step' -1 'until' nn 'do'
            a[nn,j]:=a[nn,j]/a[nn,nn];
            'for' i:=nn+1 'step' 1 'until' np1 'do'
            'begin'
            'for' j:=nn+1 'step' 1 'until' np1 'do'
                a[i,j]:=a[i,j]-a[i,nn]*a[nn,j];
                b[i]:=b[i]-a[i,nn]*b[nn]
            'end':
      'end';
      'comment' back substitution;
      'for' i:=np1 'step' -1 'until' 1 'do'
      'begin' b[i]:=b[i]/a[i,i];
           'for' k:=i-1 'step' -1 'until' 1 'do'
                b[k]:=b[k]-a[k,i]*b[i]
      'end';
'end' matinv;

'procedure' lineq;
'comment' sets up linear equations;
'begin' 'integer' i,j,ip1,ilk1; 'extended real' templ,t0,t1,s;

      'if' lk 'gq' 0 'then'
      'begin' 'for' i:=1 'step' 1 'until' lk1 'do'
                'for' j:=1 'step' 1 'until' lpm2 'do'
                a[i,j]:=0.0;
```

```
                 a[1,1]:=1.0; b[1]:=1.0;
                 t0:=1.0; t1:=1.0;
                 'if' lk 'gr' 0 'then'
                 'for' i:=1 'step' 1 'until' lk 'do'
                 'begin' t0:=-t0/i; t1:=t0: s:=t1; ip1:=i+1;
                 'for' j:=1 'step' 1 'until' i 'do'
                  'begin' t1:=-btry*t1*(mp1-j)*(ip1-j)/j:
                           s:=s+t1; 'end';
                 ilk1:=i+1;
                 a[ilk1,ilk1]:=1;
                 b[ilk1]:=s;
        'end'; 'end';

       'for' i:=1 'step' 1 'until' np1 'do'
       'begin' ilk1:=i+lk1; a[ilk1,1]:=1.0;
            templ:=(1.0+btry*xval[i])^mm1;
            'if' l'gr' 0 'then'
                 'for' j:=2 'step' 1 'until' lp1 'do'
                 a[ilk1,j]:=a[ilk1,j-1]*xval[i]:
                 'for' j:=1 'step' 1 'until' lp1 'do'
                  a[ilk1,j]:=a[ilk1,j]/templ;
                 b[ilk1]:=f(xval[i]);
             a[ilk1,lp2]:=-xval[i]*f(xval[i]);
             a[ilk1,lpm2]:=signum[i+1]*(1.0+btry*xval[i]);
       'end';
'end' lineq;

'extended real' 'procedure' del(x);
'value' x; 'extended real' x;
'comment'  computes approximation error at a'010x;
'begin' 'extended real' t1,t2,t3,t4; 'integer' i,k;
t2:=phi(x); t3:=b[lp1];
'if' itest 'gr' 0 'then' t1:=f(t2) 'else'
t1:=f(t2)*(1.0+b[lp2]*t2)/(1.0+btry*t2);
'for' i:=1 'step' 1 'until' l 'do'
      t3:=t3*t2+b[lp1-i]:
t4:=(1.0+btry*t2);
'if' ln(t4) 'ls' expcon 'then'
t2:=t3/(t4^m)  'else' t2:=0.0;
del:=(t2-t1);
'end' del;

'procedure' surmis1;
'comment'  reads in initial guesses of critical points ;
'begin'  'integer' k,i;
      'for' i:=1 'step' 1 'until' lpm2 'do' x[i]:=0.0:
      'for' i:=1 'step' 1 'until' np1 'do'
           input1(05,'('')',x[i]);
      output0(06,'('/'('initial guesses of critical points')'/')');
      'for' i:=1 'step' 1 'until' n+1 'do'
```

```
        output1(06,'('zz.ddddddd')',x[i]);
'end' surmis1;

'procedure' surmis;
'comment' computes initial critical points;
'begin' 'extended real' xi,b1,xm1,pi,z; 'integer' k,i;
    xi:=0.0;k:=n/2; b1:=(bb-aa)/2.0; xm1:=(bb+aa)/2.0:
    pi:=3.14159265;
    'for' i:=1 'step' 1 'until' lpm2 'do' x[i]:=0.0;
    'for' i:=1 'step' 1 'until' k 'do'
    'begin' xi:=xi+1.0;
        z:=cos(pi*(1.0-xi/n))*b1;
        x[i+1]:=z+xm1; x[n-i+1]:=xm1-z:
    'end';
    x[1]:=aa: x[n+1]:=bb;
    z:=lk1; z:=z/lpm2;
    'for' i:=1 'step' 1 'until' np1 'do'
     x[i]:=x[i]+(bb-x[i])*z;
    output0(06,'('/'('initial guesses of critical points')'/')');
    'for' i:=1 'step' 1 'until' n+1 'do'
        output1(06,'('zz.ddddddd')',x[i]);
'end' surmis;

'extended real' 'procedure' evalb;
'comment' evalb tries to evaluate b as a function of one
         of the extrema x[i], m, and the error at x[i]:
'begin' 'extended real' t1,t2,t3,t4,t5,tx; 'integer' tn,i,j:
   tn:=np1/2;
   tx:=x[tn]/(1.0-x[tn]);
   t1:=p[lp1]; t2:=0.0;
   'for' j:=1 'step' 1 'until' l 'do' 'begin'
    t1:=t1*tx+p[lp1-j];
    t2:=t2*tx+q[lp1-j]; 'end';
    t3:=exp(-tx): t4:=signum[tn]*b[np1];
    t2:=t2/t1; t4:=1.0+t4/t3;
    t5:=m/(t2+1.0/t4)-tx;
    evalb:=1.0/t5;
'end' evalb;

'procedure' search(x);
'comment' searches the extremal points of the error curve:
'extended real' 'array' x;
'begin'
stage2:  u:=sign(xlb); z1:=0.0;
    'comment'  search for new critical points:
    'if' n'lq' 1 'then'
    'begin'  h[1]:=0.015*(x[2]-x[1]);
             h[2]:=-h[1] 'end'
     '010'else' 'begin' 'for' i:=2 'step' 1 'until' n 'do'
        h[i]:=0.015*(x[i+1]-x[i-1]);
```

```
         h[1]:=h[2]*0.5; h[np1]:=-h[n]*0.5;
     'end';
     'for' i:=1 'step' 1 'until' np1 'do'
     'begin'  y2:=x[i]; h1:=h[i]; y3:=y2+h1;
         z2:=del(y2)*u; z3:=del(y3)*u;
         'if' z3 'lq' z2 'then'
         'begin' h1:=-h1; z:=z3; z3:=z2; z2:=z;
             y:=y3; y3:=y2;y2:=y 'end';
pace:  y:=y3+h1;
 'if' y<aa 'then' y:=aa 'else' 'if' y>bb 'then' y:=bb 'else'
     'begin' z:=del(y)*u;
         'if' z'gr'z3 'then'
         'begin' y2:=y3; y3:=y:z2:=z3;z3:=z;
                'go to' pace;
         'end' 'else' y:=-z3-z3+z+z2;
         'if' y'eq'0. 'then' y:=y3 'else'
            y:=(y2+y3)*0.5 + h1*(z2-z3)/y;
     'end';
     x[i]:=y; err[i]:=del(y); u:=-u;
     'if' i'gr'1 'then' 'begin' 'if' x[i]'lq'x[i-1] 'then'
         'begin' larm:=-1; 'go to' outout 'end':
         'end';
     z:=abs(err[i]);
     'if' z'gq' 10.0 'then'
         'begin' larm:=-1; 'go to' outout 'end':
     y:=abs(xlb); z:=abs(z-y)/y;
     'if' z1'ls' z 'then' z1:=z;
     'end':
'comment' search for one extra extremal point between the
        endpoints of the interval and the present
        critical points;
     'if' x[1]'gr' aa 'then'
     'begin' h1:=(x[1]-aa)*0.0625; u:=-sign(xlb):
         z3:=0.0; y:=aa;
         'for' i:=1 'step' 1 'until' 16 'do'
         'begin' z:=del(y)*u;
             'if' z'gr'z3 'then'
                 'begin' z3:=z; z2:=y 'end':
             y:=y+h1
         'end'; z:=abs(xlb);
         'if' z3'gr'z 'then'
         'begin'
             'for' i:=np1 'step' -1 'until' 2 'do'
             'begin' err[i]:=err[i-1]; x[i]:= x[i-1] 'end';
             x[1]:=z2; err[1]:=z3*u; 'go to' c2
         'end':
     'end';
     'if' x[np1]'ls' bb 'then'
     'begin'  h1:=(bb-x[np1])*0.0625: z3:=0.0; y:=bb;
         u:=-sign(err[np1]);
```

```
              'for' i:=1 'step' I 'until' 16 'do'
               'begin'
                    z:=del(y)*u;
                  'if' z'gr' z3 'then'
                  'begin' z3:=z; z2:=y 'end';
                  y:=y-h1
              'end'; z:=abs(xlb);
              'if' z3'gr' z 'then'
              'begin'
                  'for' i:=1 'step' 1 'until' n 'do'
                  'begin' err[i]:=err[i+1]; x[i]:=x[i+1] 'end';
                  x[np1]:=z2; err[np1]:=z3*u;
         c2:      xlb:=-xlb; z:=abs(z3-z)/z;
                  'if' z1'ls' z 'then' z1:=z;
              'end';
          'end';

     'end' search;


     output1(06,'('4/'('degree of numerator        =  ')',zzd')',l);
     output1(06,'('/'('degree of denominator      =  ')',zzd')',m):
     output1(06,'('/'('initial guess of max error=  ')',zd.8d')',xlb):
     output1(06,'('/'('initial guess of b         =  ')',zd.8d')',btry);

          aa:=0.0; bb:=0.995;
          'if' l 'eq' m 'then' bb:=0.999999;
          itno:=1; lk:=order-1; lp1:=l+1; lp2:=l+2; n:=lp1-lk; lk1:=lk+1;
          lpm2:=lp1+2; dn:=n; dnp1:=dn+1.0; np1:=n+1; mp1:=m+1; mm1:=m-1;
          bold:=-999.9; bbold:=-99999.0; evaold:=-99.0; evadel:=99.0;
          expcon:=35.0*ln(10.0)/m;
          larm:=0;
          eps20:=20.0*eps;
          'for' i:=1 'step' 1 'until' lp1 'do' p[i]:=0.0;
          'for' i:=1 'step' 1 'until' lp1 'do' q[i]:=0.0;
          'for' i:=1 'step' 1 'until' lpm2 'do' err[i]:=0.0;
              signum[1]:=1.0;
          'for' i:=1 'step' 1 'until' np1 'do'
              signum[i+1]:=-signum[i];
          'if' scode'eq'1 'then' surmis 'else' surmis1;

     stage1:
          'for' i:=1 'step' 1 'until' np1 'do' xval[i]:=phi(x[i]);
          lineq; matinv;
          'for' i:=1 'step' 1 'until' l 'do'
          'begin' p[i]:=b[i];
                  q[i]:=i*b[i+1]; 'end';
          bdel:=abs((btry-b[lp2])/btry);
          bbdel:=abs((bold-b[lp2])/bold);
          btryold:=btry;
```

```
        p[lp1]:=b[lp1];
        xlb:=b[lpm2];

        larm:=0;
        itest:=-1; search(x);


        'comment' check if computed b converges;
        'if' (10.0*bbdel 'gr' bdel 'and' bdel 'gr' eps)
             'or' (bbdel 'gr' 0.01)
             'then' 'go to' output;

        'comment' computes new b from old b;

         evanew:=(evaold*b[lp2]-btry*bbold)/
                ((evaold+b[lp2])-(btry+bbold));
         'if' (evanew 'lq' 0.0) 'or' (evaold 'lq' 0.0) 'then'
         evanew:=btry*(1.0+(b[lp2]-btry)/(m*btry));
         'if' evanew 'lq' 0.0 'then' evanew:=evalb;
         evaold:=btry; bbold:=b[lp2];
         btry:=evanew;
         evadel:=abs((evanew-evaold)/evanew);
          'if' itno 'gr' itermax 'or'
                btry 'ls' 0.0 'then' larm:=-1;

        'comment' search error curve using only new b;
        'if' bdel 'ls' eps20 'or' evadel 'ls' eps20 'then'
        'begin'
        'for' i:=1 'step' 1 'until' np1 'do' xtest[i]:=x[i];
        itest:=1; search(xtest);
        'if' z1 'ls' eps 'then' 'begin'
              larm:=1;
              'for' i:=1 'step' 1 'until' np1 'do'
              x[i]:=xtest[i];
              'go to' output; 'end'; 'end';

    output:
       xlb:=abs(err[1]); 'for' i:=2 'step' 1 'until' np1 'do'
       'begin' 'if' xlb'lq'abs(err[i]) 'then'
            xlb:=abs(err[i])  'end';

      'if' (larm 'nq' 0) 'or' (pcode 'nq' 1) 'then' 'begin'
       output2(06,"("//"("max. error & input b  =")",-d.8d"+d,
       3b-d.15d"+d")",xlb,btryold);
       output4(06,"("/"("deg. of num, denom, and order = ")",
                 3(zzzd),"(".   no. of iterations = ")",zzzd")",
                 l,m,order,itno);
       output0(06,"("/"("   crit. pts.   err. at crit. pts.")",
       "("   coeff. of num.   b & error ")"//")");
       'for' i:=1 'step' 1 'until' lp1 'do'
```

```
              output3(06,'('3z.8d,5b-d.8d'+d,5b-d.8d'+d/')',x[i],
              err[i],p[i]);
         output3(06,'('3z.8d,5b-d.8d'+d,22b-d.8d'+dd/')',x[lp2],err[lp2],
              b[lp2]);
         output3(06,'('3z.8d,5b-d.8d'+d,22b-d.8d'+dd/')',x[lpm2],
              err[lom2],b[lpm2]);
         'end';

   stage3:  'if' larm 'eq'0 'then'
              'begin' sum:=0.0;
                   'for' i:=1 'step' 1'until' np1 'do'
                        sum:=sum+signum[i]*err[i];
                   xlb:=sum/dnp1; itno:=itno+1;
                   bold:=b[lp2];
                   'go to' stage1;
              'end';

             itermax:=itno;

            'if' larm 'ls' 0 'then' 'go to' alarm;

   'end' chebyshev;

   'comment' for honeywell time-sharing system, add;
   sysparam(05,24,4); sysparam(06,24,4);
   sysparam(07,24,4); sysparam(07,6,85);

   aa:=0.0; bb:=1.0-0.000001;
   eps:=1.0'-7; scode:=1; pcode:=1:

   output0(06,'('////'('rational chebyshev approximation of exp(-x)')',
   '(' using: ( p(x) of degree l ) / (1+bx)**m ')'')');
   output0(06,'('///'('approximation of  exp(-x)')'')'):
   output2(06,'('/'('interval of approx. = (')',z.8d,'(',')',
      z.8d,'(')')'')',aa,bb);
   output1(06,'('/'('tolerance = ')',z.8d,///')',eps);

     start: input7(05,'('/3(zd),zd.2d,2(zd),z3d')',
                   l,m,order,btry,scode,pcode,itermax):
     'if' btry'nq'0.0 'then' 'begin'
     chebyshev(f,phi,l,m,order,eps,p,btry,xlb,help,itermax,scode,pcode);
     output5(07,'('/3(zd),b-d.2d'+d,bb-d.6d'+d')',
             l,m,order,xlb,btry);
     'for' doj:=1 'step' 1 'until' l+1 'do'
     output1(07,'('bb-d.6d'+d')',p[doj]);
     'goto' l1;
      help:   output0(06,'('//'('** help **')'//')');
      l1: 'goto' start;     'end';
   'end'
```

APPENDIX II

Minimax approximation of the form

$$\frac{(A_0+A_1x+\ldots+A_mX^m)}{(1+BX)^n}$$

to the exponential function

EXP(-X)

over the interval

$[0,\infty)$

with order k at X = 0.

M	N	K	ERROR	B	A0	A1	A2	A3
0	1	0	9.36E-2	2.239679E+0	1.093570E+0			
0	2	0	5.04E-2	7.485180E-1	1.050366E+0			
0	3	0	3.44E-2	4.363292E-1	1.034422E+0			
1	1	0	6.68E-2	1.727114E+0	1.066831E+0	-1.154257E-1		
1	2	0	2.27E-2	5.241638E-1	1.022710E+0	-1.853273E-1		
1	3	0	1.29E-2	3.123285E-1	1.012893E+0	-1.908452E-1		
2	2	0	1.95E-2	4.926232E-1	1.019500E+0	-2.174398E-1	4.732499E-3	
2	3	0	8.05E-3	2.712682E-1	1.008050E+0	-2.701013E-1	1.446912E-2	
2	4	0	4.59E-3	1.917889E-1	1.004586E+0	-2.814311E-1	1.774797E-2	
3	3	0	7.31E-3	2.639114E-1	1.007309E+0	-2.852386E-1	1.816549E-2	-1.343669E-4
3	4	0	3.31E-3	1.779711E-1	1.003308E+0	-3.245318E-1	2.994038E-2	-6.732805E-4
3	5	0	1.70E-3	3.037987E-1	9.982992E-1	5.539004E-1	-1.840110E-1	1.140349E-2

M	N	K	ERROR	B	A0	A1	A2	A3
					A4	A5	A6	A7
4	4	0	3.09E-3	1.752554E-1	1.003087E+0	-3.332364E-1	3.269673E-2	-8.820426E-4
					2.912989E-6			
4	5	0	1.16E-3	2.786610E-1	9.988394E-1	4.171811E-1	-1.788692E-1	1.668690E-2
					-3.757557E-4			
4	6	0	5.69E-4	2.035203E-1	9.994311E-1	2.324741E-1	-1.291854E-1	1.391546E-2
					-4.000130E-4			
5	5	0	1.07E-3	2.737463E-1	9.989295E-1	3.907711E-1	-1.770466E-1	1.768161E-2
					-4.787324E-4	1.646116E-6		
5	6	0	4.26E-4	1.929599E-1	9.995737E-1	1.663386E-1	-1.220215E-1	1.594582E-2
					-6.947541E-4	8.004538E-6		
5	7	0	2.13E-4	1.504733E-1	9.997866E-1	5.758252E-2	-8.931022E-2	1.345394E-2
					-7.038530E-4	1.099899E-5		
6	6	0	4.01E-4	1.908032E-1	9.995994E-1	1.528988E-1	-1.202848E-1	1.632410E-2
					-7.621695E-4	1.060042E-5	-1.934278E-8	
6	7	0	1.69E-4	1.450262E-1	9.998314E-1	1.859129E-2	-8.285686E-2	1.420051E-2
					-9.010176E-4	2.166595E-5	-1.453420E-7	
6	8	0	8.64E-5	1.181932E-1	9.999136E-1	-5.269584E-2	-5.933397E-2	1.192341E-2
					-8.768528E-4	2.608126E-5	-2.447076E-7	

M N K	ERROR	B	A0	A1	A2	A3
			A4	A5	A6	A7
3 5 1	1.80E-3	2.986777E-1	1.000000E+0	5.182229E-1	-1.769528E-1	1.108236E-2
3 5 2	2.38E-3	2.767995E-1	1.000000E+0	3.839973E-1	-1.511820E-1	9.985499E-3
3 5 3	4.26E-3	2.441207E-1	1.000000E+0	2.206034E-1	-1.246543E-1	9.295052E-3
4 6 1	6.03E-4	2.011216E-1	1.000000E+0	2.147669E-1	-1.251683E-1	1.364952E-2
			-3.950302E-4			
4 6 2	8.00E-4	1.907427E-1	1.000000E+0	1.444563E-1	-1.096371E-1	1.270371E-2
			-3.801531E-4			
4 6 3	1.39E-3	1.746767E-1	1.000000E+0	4.806008E-2	-9.038094E-2	1.193186E-2
			-3.836381E-4			
5 7 1	2.26E-4	1.491528E-1	1.000000E+0	4.707763E-2	-8.670149E-2	1.325067E-2
			-6.986138E-4	1.096754E-5		
5 7 2	2.99E-4	1.433903E-1	1.000000E+0	3.732231E-3	-7.612439E-2	1.248688E-2
			-6.835998E-4	1.097614E-5		
5 7 3	4.69E-4	2.440380E-1	1.000000E+0	7.082663E-1	4.237958E-2	-6.368383E-2
			7.612169E-3	-2.277150E-4		
5 7 4	8.25E-4	2.195571E-1	1.000000E+0	5.368996E-1	-2.458799E-2	-4.009482E-2
			5.734603E-3	-1.871455E-4		

BIBLIOGRAPHY

1. Achieser, N.I., Theory of Approximation, Frederick Ungar, N.Y., 1956.

2. Allen, R.H., and Pottle, C., "Stable Integration Methods for Electronic Circuit Analysis with Widely Separated Time Constants", Proc. Sixth Annual Allerton Conf. on Circuit and System Theory, U. of Illinois, Urbana, Ill., 1968, pp.311-320.

3. Bird, R.B., Stewart, W.E., and Lightfoot, E.N., Transport Phenomena, Wiley, N.Y., 1960.

4. Butcher, J.C., "Implicit Runge-Kutta Processes", Math. Comp., Vol.18 (1964), pp.50-64.

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

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

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

8. Carslaw, H.S. and Jaeger, J.C., Conduction of Heat in Solids, 2nd Edition, Oxford University Press, 1959.

9. Cavendish, J.C., Culham, W.E., and Varga, R.S., "A Comparison of Crank-Niclson & Chebyshev Rational Methods for Numerically Solving Linear Parabolic Equations", J. of Computational Physics, Vol.10, No.2, Oct. 1972, pp.354-367.

10. Chipman, F.H., "Numerical Solution of Initial Value Problems Using a Stable Runge-Kutta Processes", Research Report CSRR-2042, AA/CS U. of Waterloo, June 1971.

11. Coddington, E.A., and Levinson, N., Theory of O.D.E., McGraw-Hill, 1955.

12. Cody, W.J., Varga, R.S., and Meinardus, G., "Chebyshev Rational Approximations to e^{-x} in $[x,\infty)$ and Applications to Heat-Conduction Problems", J. of Approximation Theory, Vol.2, 1969, pp.50-65.

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

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

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

16. Dill, C., and Gear, C.W., "Rational Approximations by Implicit Runge-Kutta Schemes", BIT, Vol.10 (1970), pp.20-22.

17. _________________________, "A Graphical Search for Stiffly Stable Methods for Ordinary Differential Equations", J. ACM, Vol.18, (1971), pp.75-79.

18. Ehle, B.L., "On Padé Approximations to the Exponential Function and A-stable Methods for the Numerical Solution of Initial Value Problems", Ph.D. Thesis, U. of Waterloo, 1969.

19. __________,"Some Results on Exponential Approximation and Stiff Equation", In preparation.

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

21. Haines, C.F., "Implicit Integration Processes with Error Estimate for the Numerical Solution of Differential Equations", Computer Journal, Vol.12 (1969), pp.183-187.

22. 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.

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

24. Jain, M.K., and Srivastava, V.K., "High Order Stiffly Stable Methods for ODE", Report #394, U. of Illinois, Urbana, 1970.

25. Keller, H.B., "A New Difference Scheme for Parabolic Problems", Proceedings of the Second Symposium on the Numerical Solution of PDE, Hubbard, B. (editor), Academic Press, 1971.

26. Kuo, F.F., Kaiser, J.F., System Analysis by Digital Computer, John Wiley & Sons, N.Y. (1966), pp.102-103.

27. Lambert, J.D., and Sigurdson, S., "Multistep Formulas with Variable Matrix Coefficients", SIAM J. Num. Anal., Vol.9,No.4(1972), pp.715.

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

29. 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.

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

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

32. ____________, "On the Exactness of Implicit Runge-Kutta Processes for particular integrals", BIT, Vol.12, No.4, (1972), pp.586-589.

33. ____________, "Some Numerical Methods for Stiff ODE and PDE", Proc. 2nd Manitoba Conference in Num. Math. (1972), pp.27-34.

34. ____________ and Ehle, B.L., "Generalised Runge-Kutta Processes for Stiff Initial Value Problems", submitted for publication.

35. ____________, "Order-Constrained Chebyshev Rational Approximations", to appear, Math. Comp.

36. ____________, "Generalised Adams Methods for Stiff Systems of ODE" to appear, Utilitas Mathematics.

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

38. Liniger, W. and Willoughby, R., "Efficient Numerical Integration of Stiff Systems of Ordinary Differential Equations", SIAM Journal of Num. Anal., Vol.7, No.1 (1970), pp.47-66.

39. Lomax, H., and Bailey, H., "A Critical Analysis of Various Integration Methods for Computing the Flow of a Gas in Chemical Non-equilibrium", NASA Technical Note, NASA TND-4109.

40. 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.

41. Meinardus, G., Approximation of Functions: Theory and Numerical Methods, Springer Tracts in Natural Philosophy, Vol.13, Springer-Verlag, N.Y., 1967.

42. Moretti, G., "A New Technique for the Analysis of Non-equilibrium Flows", AIAA Journal, Vol.3 (1965), pp.223-229.

43. Muskat, M., The Flow of Homogeneous Fluids Through Porous Media, McGraw Hill, N.Y., 1937.

44. 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.

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

46. Ralston, A., A First Course in Numerical Analysis, McGraw-Hill, N.Y., 1965.

47. Reddick, H.W., and Kibbey, D.E., Differential Equations, 3rd edition, John Wiley & Sons, N.Y., 1956.

48. Rice, J.R., "The Approximation of Functions", Vol.1, Addison-Wesley, 1964.

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

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

51. Sloate, H.M., and Bickart, T.A., "A-stable Composite Multistep Methods", J.ACM, Vol.20, No.1, Jan.73.

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

53. __________, Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, N.J., 1962.

54. Wright, K., "Some Relationships between Implicit Runge-Kutta, Collocation, and Lanczos τ Methods, and their Stability Properties", BIT, Vol.10 (1970), pp.217-227.

55. Barrodale, I., Roberts, F.D.K., and Hunt, C.R., "Computing best L_p Approximations by Functions Nonlinear in one parameter", The Computer Journal, Vol.13, No.4 (1970), pp.382-386.