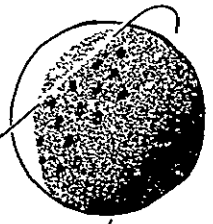
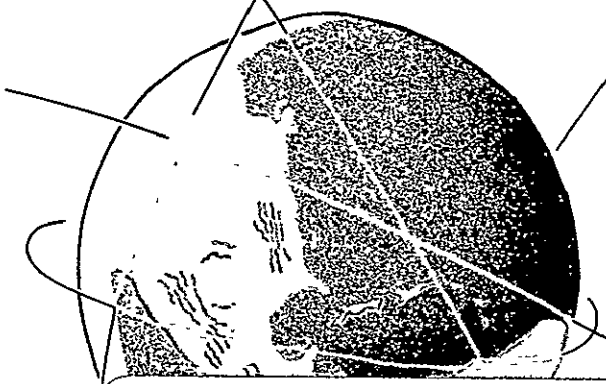


A NON-ITERATIVE METHOD OF OBTAINING TCHEBYCHEFF RATIONAL APPROXIMATIONS

BY
HAMILTON HAGAR, JR.

AMRL 1030 JULY 1970



APPLIED MECHANICS RESEARCH LABORATORY
THE UNIVERSITY OF TEXAS AT AUSTIN AUSTIN, TEXAS

A NON-ITERATIVE METHOD OF OBTAINING TCHEBYCHEFF RATIONAL APPROXIMATIONS

Hamilton Hagar, Jr.
The University of Texas at Austin
Austin, Texas

AMRL 1030
July 1970

Applied Mechanics Research Laboratory
The University of Texas at Austin
Austin, Texas

This report was prepared under

Contract No. NGL 44-012-008

for the

National Aeronautics and Space Administration

by the

Applied Mechanics Research Laboratory
The University of Texas at Austin
Austin, Texas

under the direction of

Byron D. Tapley
Chairman

ABSTRACT

A simple method of obtaining rational forms as approximations to functions which may be expressed as power series is developed. The method is near-optimal under the Tchebycheff norm. While most approaches are iterative in nature, the method presented here is free of this characteristic.

The technique is developed as a combination of Padé rational approximation and Lanczos' Tau method, which uses Tchebycheff polynomial properties for improving accuracy. In addition to some simple non-linear functions, some examples from the field of astrodynamics are used to illustrate the method.

ACKNOWLEDGMENTS

I extend my thanks to Dr. Ray Nachlinger for serving as chairman of my thesis committee and to committee members, Dr. Bart Childs, Dr. George Born, and Dr. Pat Hedgecoxe.

In particular I thank my wife, Jean, for her encouragement and for typing this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The concept of rational approximation is introduced and set in perspective with the more familiar polynomial approximation. The historical development is traced briefly and the scope of the investigation is set forth.

1.1 Rational Forms as Approximating Functions

A common method of approximating transcendental and other nonlinear functions is through the use of polynomial approximation. It is an attractive approach chiefly because of its simplicity, but generally requires polynomials of high degree for high accuracy. In recent years rational functions as approximating forms have come under investigation. Rational functions have been found to offer considerable flexibility and accuracy in the approximation of certain functions. The rational form

$$R_{mn}(x) = \frac{a_0 + a_1 x + \cdots + a_m x^m}{b_0 + b_1 x + \cdots + b_n x^n} \tag{1.1}$$

is roughly equivalent in its "curve-fitting" ability to a polynomial of degree m+n. In some instances, however, such rational forms are far superior to polynomials.

Heuristically, the consideration of rational functions as approximating forms is motivated by a comparison with polynomial approximation. Consider the approximation of a function, f, of a single independent variable, x, by a polynomial:

$$f(x) \simeq c_0 + c_1 x + \cdots + c_n x^n \tag{1.2}$$

The parameters to be determined, the c_k, enter the problem in a linear fashion. For a rational form, however, the parameters enter in a non-linear manner since

$$R_{mn}(x) = \frac{a_0 + a_1 x + \cdots + a_m x^m}{b_0 + b_1 x + \cdots + b_n x^n} \tag{1.3}$$

can be written as

$$R_{mn}(x) = r_0 + r_1 x + \cdots + r_m x^m \tag{1.4}$$

where

$$r_k = \frac{a_k}{b_0 + b_1 x + \cdots + b_n x^n} \tag{1.5}$$

Unlike the c_k, the r_k are not constant but vary with x. Hence one would expect greater flexibility in approximating with a rational form than with a polynomial. In fact, a polynomial is but a special case of a rational form, for if

$$b_0 + b_1 x + \cdots + b_n x^n = 1$$

then $r_k = a_k$ and one obtains the form of (1.4).

Algorithms for determining "best" rational approximations (optimum in some sense) have been found by various investigators (5), (6), (13). However, for optimality under the Tchebycheff norm these are, at best, involved and are usually iterative in nature. This investigation is concerned with the problem of approximating certain non-linear functions in a manner which allows greater facility in their handling. In particular, interest is in obtaining rational forms which approach optimality under the Tchebycheff norm in their approximation of such functions.

1.2 Background Study of Rational Approximation

One of the earliest successful attempts to obtain a method of analytically developing rational approximations was due to H. Padé in 1892. It is a simple but effective method based upon a series expansion of the function approximated. It suffers from the disadvantage of the Taylor expansion in that for a finite ordered approximation, the error increases as one progresses further from the origin. In spite of this, the Padé method forms the basis of the method developed in this investigation.

Apparently it was not until the late 1950's and early 1960's that rational approximation was extensively investigated. Shanks (17) in 1954 investigated several useful classes of non-linear transformations which yield rational forms from both converging and diverging sequences. Shanks' efforts also proved that Padé approximation is but a special case of his transformations.

Wynn (21) examined the rational approximation of functions defined by a power series and developed the so-called "epsilon algorithm," also a special case of Shanks' non-linear transformations.

Cheney (4) and Boehm (3) among others have developed considerable formal theory, having investigated existence, characterization, and convergence properties of rational Tchebycheff approximations (i.e., rational approximations which are optimal under the Tchebycheff norm). Both Cheney and Maehly (13) have developed a number of iterative algorithms for obtaining rational approximations which are optimal in the Tchebycheff sense. The work of these investigators is generally representative of the current level of development of optimal rational approximation.

1.3 Scope of the Investigation

For this investigation a restriction is introduced regarding the nature of the function to be approximated. Each function, f, must be expandable in a Taylor series,

$$f(x) = \sum_{k=0}^{\infty} \left[\frac{d^k f}{dx^k}\right]_{x=0} x^k = \sum_{k=0}^{\infty} c_k x^k \qquad (1.6)$$

The reason for this restriction rests with the Padé method which forms the basis for the development.

An important distinction should be made. The concern here is with obtaining rational forms as approximations to functions expressed as power series. The investigation is <u>not</u> concerned with approximating a function knowing only a set of its values.

Within the restrictions set forth above, the purpose of the investigation may be summarized. The purpose is to develop a relatively simple method for obtainjng rational forms as approximations to functions admitting a power series representation, and which are near-optimal in the Tchebycheff sense in their approximations of such functions. The motivating question is, "To what extent can such functions be accurately approximated by rational forms?"

CHAPTER 2

APPROXIMATION CHARACTERISTICS

This chapter sets forth some of the basic formal theory associated with the development of best rational approximations and is based on Cheney (4).

2.1 Existence of Best Rational Approximations

A family of rational functions, Γ_{mn}, is first defined.

$$\Gamma_{mn}[a,b] \equiv \left\{R_{mn} \equiv \frac{P}{Q}:\ \partial P \leqq m,\ \partial Q \leqq n,\ Q(x) > 0 \text{ on } [a,b]\right\} \tag{2.1}$$

Γ_{mn} is the class of rational functions defined on the closed interval $[a,b]$ where

$$\begin{aligned} P = P_m(x) &\equiv a_0 + a_1 x + \cdots + a_m x^m \\ Q = Q_n(x) &\equiv b_0 + b_1 x + \cdots + b_n x^n \end{aligned} \tag{2.2}$$

The polynomials P and Q are regular polynomials in the single independent variable, x. The degree (∂) of P and Q satisfy the inequalities

$$\partial P \leqq m, \quad \partial Q \leqq n \tag{2.3}$$

where m and n are integers specifying the order of the rational form, R_{mn}. The inequalities (2.3) admit the possibility that the polynomials comprising R_{mn} may have degrees less than m or n. The reason for this will be illustrated in the next chapter. P and Q are further restricted to have no factors in common other than constants. $R \equiv P/Q$ is then said to be in irreducible form. The members of Γ_{mn} must be bounded. Thus it

is both necessary and sufficient that Q have no roots on the interval [a,b]. This may be accomplished without loss of generality by requiring that $Q(x) > 0$.

The condition of optimality will be the minimization of the Tchebycheff norm (uniform norm, infinity norm),

$$\min \| \quad \| = \min \left\{ \max_{x \in [a,b]} | \quad | \right\} \tag{2.4}$$

Defining the error function (f is the function approximated),

$$\delta(x) \equiv f(x) - R_{mn}(x) \tag{2.5}$$

the scalar, λ, is obtained as

$$\lambda = \min \| \delta(x) \| = \min \left\{ \max_{x \in [a,b]} | \delta(x) | \right\} \tag{2.6}$$

Thus for R to be a best approximation the local minima and maxima of $\delta(x)$ for all $x \in [a,b]$ must have the same absolute value, λ (see Figure 2.1).

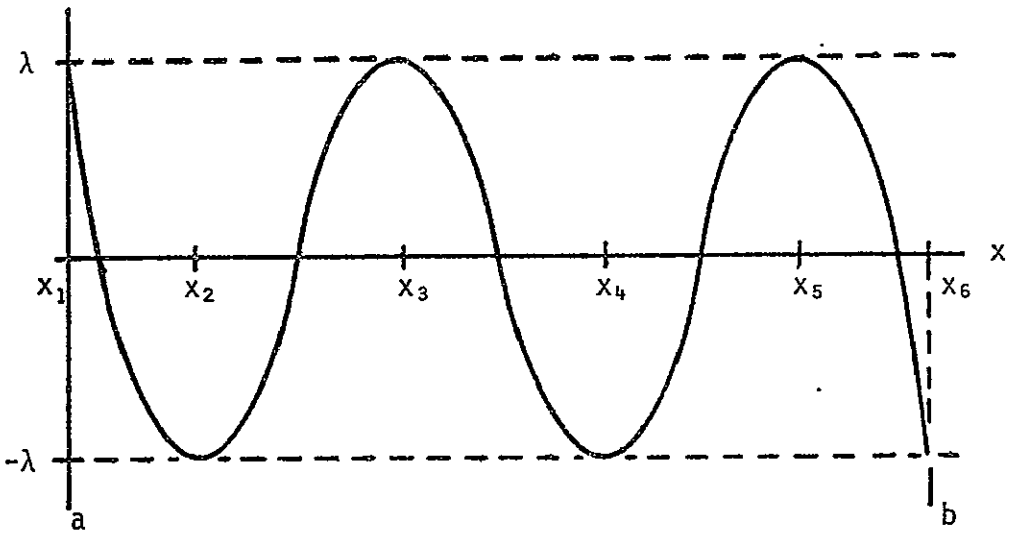


Figure 2.1 Condition of Optimality

Having defined the class Γ_{mn}, and having set forth the condition of optimality, an existence theorem is now given. This is due to Cheney (4) and is stated here without proof.

Existence Theorem: To each function, f, continuous on the closed interval [a,b], there corresponds at least one best rational approximation from the class Γ_{mn} [a,b].

2.2 Characterization of Best Rational Approximations

The condition of minimum Tchebycheff norm is now extended to more fully characterize best rational approximations. The concept of T-alternations is introduced by first stating that there exist k values for x, $x_1 < x_2 < \cdots < x_k$, such that $\delta(x_i)\ \delta(x_{i+1}) < 0$, $i = 1, 2, \cdots, k-1$. (In Figure 2.1 there are six values of x for which the condition holds.) Thus $\delta(x)$ changes sign k-1 times and is said to have k T-alternations. The condition of optimality may now be recast into the following criterion:

Criterion: In order for the irreducible rational function $R = P/Q \in \Gamma_{mn}$ to be a best approximation to the continuous function, f, it is necessary and sufficient that the error function, δ, have at least $2 + \max\left\{(m + \partial Q), (n + \partial P)\right\}$ T-alternations.

The reason for having to select the maximum of the bracketed terms is because of the inequalities (2.3). This is the criterion used in this investigation for judging best rational approximations.

In addition to proving the above, Cheney also proves the following:

Uniqueness Theorem: Best approximations in $\Gamma_{mn}[a,b]$ are always unique.

The representation of best rational approximations, however, is not unique since multiplication of the numerator and denominator polynomials by an arbitrary constant implies an infinite number of representations.

CHAPTER 3

PADÉ APPROXIMATION

The Padé method of obtaining rational approximations to functions defined by a power series is presented in this chapter. A matrix formulation is utilized and the method is illustrated with a simple example. An expression for the error is also derived.

3.1 Development of the Padé Method

Consider a function, f, which may be expanded in a Taylor series about the origin.

$$f(x) = \sum_{i=0}^{\infty} c_i \, x^i \tag{3.1}$$

A rational function of the form (1.1) is desired which will approximate f(x),

$$\sum_{i=0}^{\infty} c_i \, x^i = \frac{\sum_{k=0}^{m} a_k \, x^k}{\sum_{j=0}^{n} b_j \, x^j} + \delta(x) \tag{3.2}$$

where $\delta(x)$ is the error in the approximation. Multiplying (3.2) by the denominator on the right gives

$$\sum_{i=0}^{\infty} \sum_{j=0}^{n} b_j \, c_i \, x^{i+j} = \sum_{k=0}^{m} a_k \, x^k + \delta(x) \sum_{j=0}^{n} b_j \, x^j \tag{3.3}$$

which can be rewritten as

$$\sum_{k=0}^{\infty} \left(\sum_{j=0}^{n} b_j \, c_{k-j} - a_k \right) x^k = \delta(x) \sum_{j=0}^{n} b_j \, x^j \qquad (3.4)$$

with $a_{m+1} = a_{m+2} = \cdots = 0$, and $c_{k-j} = 0$ for $k < j$. Neglecting the error term, the coefficients of the first $n + m$ powers of x may be equated to zero to yield a system of $n + m + 1$ equations in $n + m + 2$ unknowns. This under-determinance may be relieved by normalizing one of the b's, say b_0, to 1. The result is then a system of $n + m + 1$ equations in $n + m + 1$ unknowns. For clarity, assume for the illustration that $m = n$. The resulting equations are

$$\begin{aligned}
c_0 &= a_0 \\
c_1 + c_0 \, b_1 &= a_1 \\
c_2 + c_1 \, b_1 + c_0 \, b_2 &= a_2 \\
&\vdots \\
c_n + \cdots + c_1 \, b_{n-1} + c_0 \, b_n &= a_n \\
&\vdots \\
c_{2n} + \cdots + c_{n+1} \, b_{n-1} + c_n \, b_n &= 0
\end{aligned} \qquad (3.5)$$

The system (3.5) may be rearranged and written in the following matrix notation:

$$\begin{bmatrix}
1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 & -c_0 & \cdots & 0 \\
\vdots & & \ddots & \vdots & \vdots & & \vdots \\
0 & \cdots & & 1 & -c_{n-1} & \cdots & -c_0 \\
0 & \cdots & & 0 & -c_n & \cdots & -c_1 \\
\vdots & & & \vdots & \vdots & & \vdots \\
0 & \cdots & & 0 & -c_{2n-1} & \cdots & -c_n
\end{bmatrix}
\begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \\ b_1 \\ \vdots \\ b_n \end{bmatrix}
=
\begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ \\ \\ \\ c_{2n} \end{bmatrix} \qquad (3.6)$$

In general the size of the square coefficient matrix is (m+n+1) x (m+n+1); the vector of unknown a's and b's, and the vector of c's are both (m+n+1) x 1. The matrix equation (3.6) may be partitioned as

$$\left[\begin{array}{c:c} I & H \\ \hdashline 0 & G \end{array}\right] \left[\begin{array}{c} a \\ \hdashline b \end{array}\right] = \left[\begin{array}{c} h \\ \hdashline g \end{array}\right] \tag{3.7}$$

where the sub-matrices are given below.

$[I]$ = (n+1) x (n+1) identity matrix,

$[0]$ = n x (n+1) zero matrix,

$$[H] = \begin{bmatrix} 0 & \cdots & 0 \\ -c_0 & \cdots & 0 \\ \vdots & & \vdots \\ -c_{n-1} & \cdots & -c_0 \end{bmatrix}, \quad (n+1) \text{ x } n$$

$$[G] = \begin{bmatrix} -c_n & \cdots & -c_1 \\ \vdots & & \vdots \\ -c_{2n-1} & \cdots & -c_n \end{bmatrix}, \quad n \text{ x } n$$

$$(a) = \begin{bmatrix} a_0 \\ \vdots \\ a_n \end{bmatrix}, \qquad (h) = \begin{bmatrix} c_0 \\ \vdots \\ c_n \end{bmatrix}, \quad (n+1) \text{ x } 1 \tag{3.8}$$

$$(b) = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}, \qquad (g) = \begin{bmatrix} c_{n+1} \\ \vdots \\ c_{2n} \end{bmatrix}, \quad n \text{ x } 1$$

Of course, in the general case the dimensions of these are

$[I]$, $(m+1) \times (m+1)$; $[0]$, $n \times (m+1)$

$[H]$, $(m+1) \times n$; $[G]$, $n \times n$

(a), (h), $(m+1) \times 1$; (b), (g), $n \times 1$

Because of the presence of the zero sub-matrix, the b coefficients may be solved independent of the a's as

$$(b) = [G]^{-1} (g) \tag{3.9}$$

Hence only an n x n system of linear equations need be solved simultaneously. The a coefficients may then be determined immediately from

$$(a) = (h) - [H] (b) \tag{3.10}$$

3.2 Example - $\mathrm{Tan}^{-1}(x)$

The Padé method may be effectively illustrated using the inverse tangent function,

$$f(x) = \tan^{-1}(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots \tag{3.11}$$

Let the order of the desired rational approximation be arbitrarily specified as $m = n = 2$. Note that the expansion for $\tan^{-1}(x)$ is not strictly of the form of (3.1). Placing it in this form may be accomplished in either of two ways. One is by factoring and making the change of variable, $y = x^2$,

$$f(x) = xg(y) = x\left(1 - \frac{y}{3} + \frac{y^2}{5} - \frac{y^3}{7} + \cdots\right) \tag{3.12}$$

and then approximating $g(y)$. The second approach is merely to include the missing even powers of x with zero coefficients. This latter approach may be designated as placing the series in fundamental form. It is particularly convenient to use this approach in Chapter 5, and provides additional insight in this example. Thus in fundamental form

$$f(x) = 0 + x + (0)x^2 - \frac{x^3}{3} + (0)x^4 - \frac{x^5}{5} + \cdots \tag{3.13}$$

From this the Padé equations are found to be

$$\begin{aligned} c_0 &= a_0 & c_0 &= 0 \\ c_1 + c_0 b_1 &= a_1 & c_1 &= 1 \\ c_2 + c_1 b_1 + c_0 b_2 &= a_2 \quad , & c_2 &= 0 \\ c_3 + c_2 b_1 + c_1 b_2 &= 0 & c_3 &= -1/3 \\ c_4 + c_3 b_1 + c_2 b_2 &= 0 & c_4 &= 0 \end{aligned} \tag{3.14}$$

(recall b_0 is normalized to 1). In matrix form,

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -c_0 & 0 \\ 0 & 0 & 1 & -c_1 & -c_0 \\ 0 & 0 & 0 & -c_2 & -c_1 \\ 0 & 0 & 0 & -c_3 & -c_2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix}$$

Corresponding to (3.8) the sub-matrices are

$$\begin{aligned} [H] &= \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ -1 & 0 \end{bmatrix} \quad , \quad (h) = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\ [G] &= \begin{bmatrix} 0 & -1 \\ 1/3 & 0 \end{bmatrix} \quad , \quad (g) = \begin{bmatrix} -1/3 \\ 0 \end{bmatrix} \end{aligned} \tag{3.16}$$

Utilizing (3.9) and (3.10) the solution vectors are

$$\begin{aligned} (b) &= [G]^{-1}(g) = \begin{bmatrix} 0 \\ 1/3 \end{bmatrix} \\ (a) &= (h) - [H](b) = \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} \end{aligned} \tag{3.17}$$

Hence the resulting rational approximation is

$$R_{12}(x) = \frac{x}{1 + 1/3\ x^2} \simeq \tan^{-1}x \tag{3.18}$$

Note that R_{12} from (3.18) belongs to Γ_{22} and that $\partial P = 1$ even though it was desired that $m = 2$. Thus here is an example for which strict inequality of the first of (2.3) holds.

Figure 3.1 illustrates the behavior of the error curves associated with R_{12} and with the Taylor expansion used to obtain R_{12}. Obviously for the given number of terms the accuracy has been improved. This is the advantage of the Padé method.

While its advantage is improved accuracy, the Padé method has an important similarity to the Taylor series. In fact, R_{mo}, obtained trivially from the Padé method is just the Taylor expansion to $m + 1$ terms. Thus Padé approximation is the rational analogue to the Taylor series.

An interesting example from Leibnitz is the evaluation of the expansion of $4\tan^{-1}(x)$ at $x = 1$. Thus,

$$4\tan^{-1}(x) = 4(1 - 1/3 + 1/5 - 1/7 + \cdots) = \pi \tag{3.19}$$

Consider now the rational approximation $R_{44}(x) \simeq \tan^{-1}(x)$. Solving the corresponding Padé equations formed from the expansion (3.13) gives

$$a_0 = 0 \qquad b_0 = 1$$

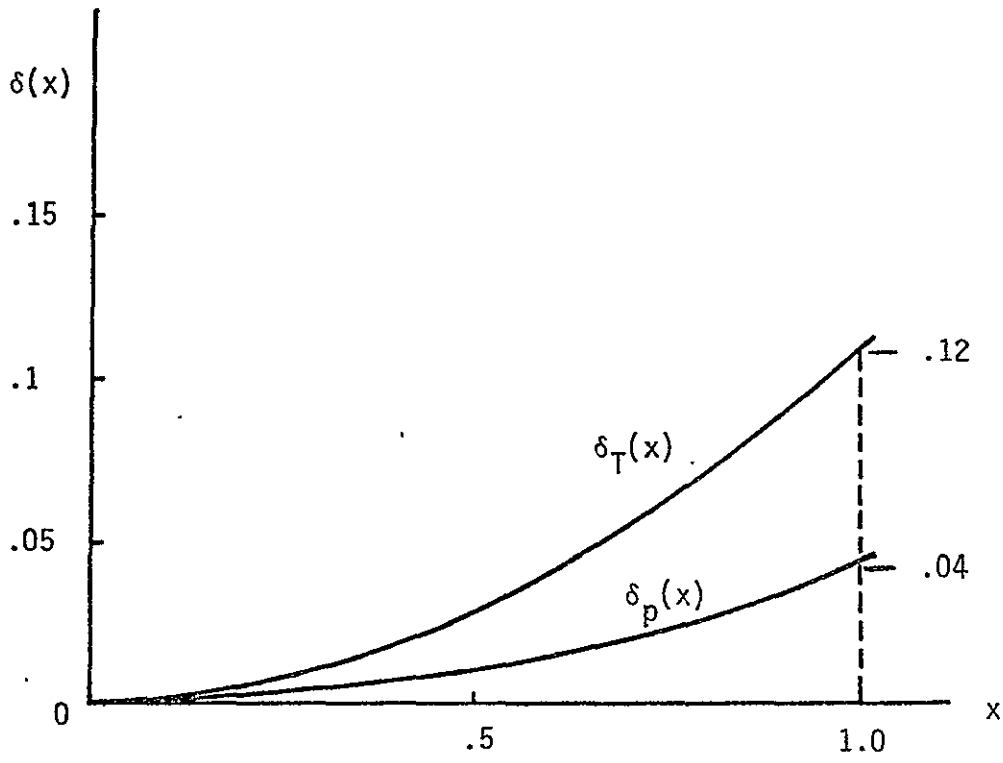
$$a_1 = 1 \qquad b_1 = 0$$

$$a_2 = 0 \qquad b_2 = .85714286$$

$$a_3 = .52380952 \qquad b_3 = 0$$

$$a_4 = 0 \qquad b_4 = 8.57142857 \times 10^{-2}$$

for which the resultant rational form is



$$\delta_T(x) = \tan^{-1}(x) - \left(x - \frac{x^3}{3}\right)$$

$$\delta_p(x) = \tan^{-1}(x) - \left(\frac{x}{1 + 1/3\ x^2}\right)$$

Figure 3.1 Error Curves for $\tan^{-1}(x)$ (Padé Method)

$$R_{34}(x) = \frac{a_1 x + a_3 x^3}{1 + b_2 x^2 + b_4 x^4} \simeq \tan^{-1}(x) \qquad (3.19)$$

Again, ∂P is less than the value sought, $m = 4$. Evaluating at $x = 1$,

$$4R_{34}(1) = 3.1372549$$

which yields an error of 4.3377×10^{-3}. To obtain this accuracy using the Taylor expansion requires 23 to 231 terms depending on summation schemes. Thus for many slowly convergent series, the Padé method provides a useful and accurate representation with little effort.

3.3 Error Expression

The basic equation from which the Padé coefficients are found is just the sum of the first (n+m+1) terms on the left of (3.4) set equal to zero:

$$\sum_{k=0}^{n+m} \sum_{j=0}^{n} b_j \, c_{k-j} \, x^k - \sum_{k=0}^{m} a_k \, x^k = 0 \qquad (3.20)$$

Substituting this expression back into (3.4) and solving for the error, δ, yields

$$\delta(x) = \frac{\sum_{k=n+m+1}^{\infty} \sum_{j=0}^{n} b_j \, c_{k-j} \, x^k}{\sum_{j=0}^{n} b_j \, x^j} \qquad (3.21)$$

For convergent series an approximation of the error may be obtained using the k = m+n+1 term in the numerator sum.

$$\delta(x) \simeq \frac{\left(c_{m+n+1} \, b_0 + c_{m+n} \, b_1 + \cdots + c_{m+1} \, b_n \right) x^{m+n+1}}{b_0 + b_1 x + \cdots + b_n \, x^n} \qquad (3.22)$$

In the previous example for $4\,R_{34}(1) \simeq 4\tan^{-1}(1) = \pi$,

$$\delta(1) \simeq \frac{4\,(c_9 + c_8b_1 + c_7b_2 + c_6b_3 + c_5b_4)\,x^9}{1 + b_1x + b_2x^2 + b_3x^3 + b_4x^4} \tag{3.23}$$

$$\delta(1) \simeq 4\left[\frac{.111 - (.143)(.857) + .2(8.57 \times 10^{-2})}{1 + .857 + 8.57 \times 10^{-2}}\right]$$

$$\simeq .01028$$

This is admittedly a crude approximation to the actual error. Taking a few more terms would certainly improve the value.

CHAPTER 4

THE TAU METHOD

This chapter reviews some properties of Tchebycheff polynomials, and then presents Lanczos' Tau method by a simple illustration employing a series solution to a simple linear differential equation.

4.1 Tchebycheff Polynomials

All the common orthogonal polynomials may be derived as solutions to the linear, second order differential equation due to Gauss,

$$x(1 - x)y'' + [\gamma - (\alpha + \beta + 1)x]y' - \alpha\beta y = 0 \qquad (4.1)$$

The solution to this equation is the hypergeometric function,

$$F(\alpha, \beta, \gamma; x) = 1 + \frac{\alpha\beta}{\gamma \cdot 1} x + \frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1)\cdot 1\cdot 2} x^2 + \frac{\alpha(\alpha+1)(\alpha+2)\beta(\beta+1)(\beta+2)}{\gamma(\gamma+1)(\gamma+2)\cdot 1\cdot 2\cdot 3} + \cdots \qquad (4.2)$$

where α, β and γ are constants. The series is convergent for $|x| < 1$ and is divergent for $|x| > 1$ except if the series terminates after a finite number of terms. When the constants are chosen as

$$\left.\begin{aligned} \gamma &= 1/2 \\ \beta &= n \\ \alpha &= -n \end{aligned}\right\} \quad n\text{: real, non-negative, integer}$$

and x is transformed to the new variable,

$$x = \frac{1 - \xi}{2}$$

the hypergeometric function yields the Tchebycheff polynomials.

$$F(-n, n, 1/2; \frac{1-\xi}{2}) = T_n(\xi) \tag{4.3}$$

The roots of these polynomials occur at

$$\xi_k = \cos\left[(2k-1)\pi/2n\right],$$

and their spacing is such that

$$\lambda = \min\left\{\max_{\xi \in [-1,1]} |T_n(\xi)|\right\} = \min ||T_n(\xi)|| = 1 \tag{4.4}$$

In other words, recalling section 2.2, each T_n possesses n+1 T-alternations. This is the property which makes the Tchebycheff polynomials useful in approximation, particularly in the following sections.

4.2 The Tau Method

Introduced in 1938 by C. Lanczos (11), the Tau method utilizes the uniform norm property of the Tchebycheff polynomials to improve the solution of various linear systems. The method is most easily presented using a series solution to a simple differential equation.

Consider the following first order, linear differential equation:

$$y' - y = 0 , \quad y(0) = 1 \tag{4.5}$$

Although the solution is simply $y = e^x$, a power series solution is assumed for the illustration as

$$y(x) = \sum_{k=0}^{\infty} c_k x^k \tag{4.6}$$

Substituting into (4.5) and equating powers of x yields a system of coefficient equations which leads to the recurrence relation,

$$c_n = \frac{c_{n-1}}{n} = \frac{c_0}{n!}$$

Then, in the presence of the boundary condition, $y(0) = 1$, (4.7) provides the solution to the differential equation (4.5),

$$y(x) = 1 + x + \frac{x^2}{2} + \cdots + \frac{x^n}{n!} \tag{4.8}$$

the limit of which, as $n \to \infty$, is obviously e^x. Since a series solution is usually exact only in the limit, practically speaking it cannot satisfy the differential equation exactly. Substituting (4.8) into (4.5) produces an error of $(x^n/n!)$. Thus there will always be an error which may be made arbitrarily small but never zero.

Now Lanczos reasoned that the error incurred by truncating the series solution could be countered by introducing another variable, τ_n, into the differential equation. Thus the original equation, which was solved approximately, is perturbed to result in one which may be solved exactly. At the same time, judiciously choosing the way in which the new variable would be introduced could improve the accuracy of the series solution.

Suppose the new variable, τ_n, is multiplied by a Tchebycheff polynomial of order n and the result added to the right side of the differential equation. In effect this is an approximation of the error, and the coefficients which result actually appear to improve the accuracy of the series solution. Returning to the example, this operation modifies (4.5) to yield

$$y' - y = \tau_n T_n \tag{4.9}$$

Now an additional requirement is that the range of x must be known so that the Tchebycheff polynomials may be appropriately scaled. Hence for $x \in [a,b]$,

$$T_n = T_n\left(\frac{2x - a - b}{b - a}\right) \tag{4.10}$$

For the example let $x \in [0,1]$. Then (4.10) yields the so-called "shifted" Tchebycheff polynomials, T_n^*. Letting $n = 2$,

$$y' - y = \tau_2 T_2^* \tag{4.11}$$

and the assumed solution becomes a quadratic:

$$y(x) = c_0 + c_1 x + c_2 x^2 \tag{4.12}$$

Substituting this into (4.11) and using T_2^* in its polynomial form ($T_2^* = 8x^2 - 8x + 2$)[1] the desired system of coefficient equations is obtained as

$$\begin{aligned} c_1 - c_0 &= \tau_2 \\ 2c_2 - c_1 &= -8\tau_2 \\ -c_2 &= 8\tau_2 \end{aligned} \tag{4.13}$$

The solution of this system yields

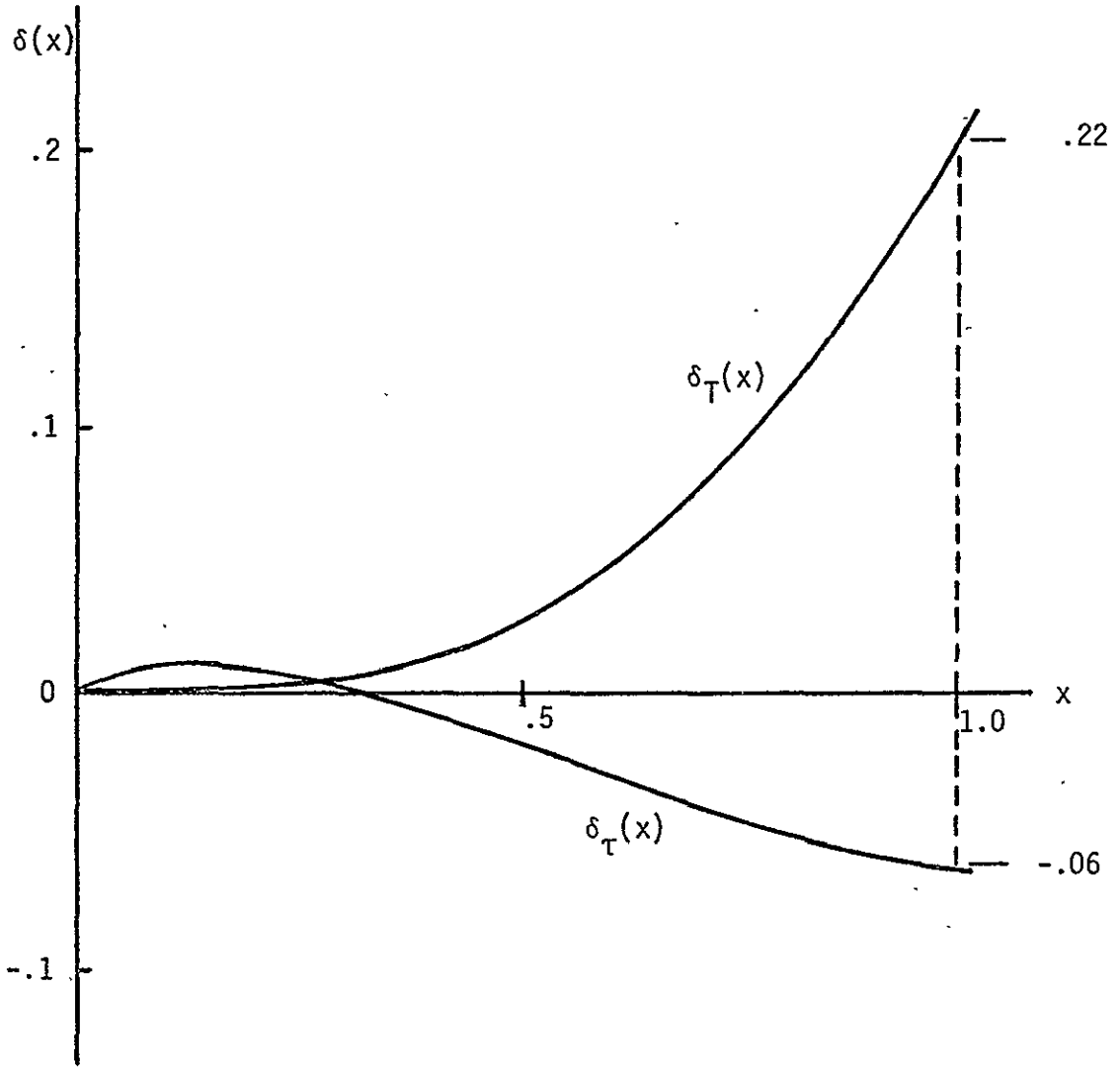
$$c_0 = 1, \quad c_1 = c_2 = 8/9, \quad \tau_2 = -1/9$$

for which

$$y(x) = 1 + 8/9x + 8/9x^2, \quad x \in [0,1] \tag{4.14}$$

Comparison of the Tau method with the corresponding series solution is shown in Figure 4.1, giving the corresponding error curves. Note that gaining in accuracy toward the right side of the interval requires sacrificing some of the accuracy toward the left. In spite of this, the net result is a gain in accuracy, under the Tchebycheff norm, over the interval.

The foregoing illustration of the Tau method utilized a simple first order, linear differential equation for which one τ-variable was



$$\delta_T(x) = e^x - (1 + x + 1/2\ x^2)$$

$$\delta_\tau(x) = e^x - (1 + 8/9\ x + 8/9\ x^2)$$

Figure 4.1 Error Curves for e^x (Tau Method)

used. In the case of higher order equations additional τ's may be required. For example, a second order equation containing derivatives of only second order would require two τ-variables. Also, Lanczos pointed out that the method is applicable to any system of linear equations, and use is made of this fact in the next chapter.

CHAPTER 5

TAU AND PADÉ METHODS COMBINED

As a system of linear algebraic equations, the Padé coefficient equations are well-suited for application of the Tau method. This chapter develops the procedure, again using matrix notation, and derives the associated error expression. The nature of the method is examined as applied to various transcendental functions. Approximate solutions and representations are obtained for the classic transcendental equation of Kepler.

5.1 Development of the Tau-Padé Equations

Following the development of the Padé method, a function, f, expandable in a Taylor series about the origin is to be approximated by a rational function of the form (1.1).

$$f(x) = \sum_{i=0}^{\infty} c_i\, x^i = \frac{\sum_{k=0}^{m} a_k\, x^k}{\sum_{j=0}^{n} b_j\, x^j} + \delta(x) \qquad (5.1)$$

where, as before, δ is the error. The function is assumed to be expressable in the fundamental form introduced in Chapter 3. As before, multiplying by the denominator term and rewriting yields equation (3.4). Neglecting the error term and taking the first N terms allows (3.4) to be written as

$$\sum_{k=0}^{N} \sum_{j=0}^{n} b_j \, c_{k-j} \, x^k - \sum_{k=0}^{m} a_k \, x^k = 0 \ , \quad c_{k-j} = 0, \ k < j \tag{5.2}$$

N is an integer, greater than n+m, whose value will be determined later. Now, since N > n+m, equating coefficients of like powers of x results in an overdetermined system. According to the Tau method, the procedure would be to introduce a τ-variable as the coefficient of a Tchebycheff polynomial. Here, however, greater flexibility is available because any number of the c_i's are already available. Thus not just one but any number, ℓ, of τ-variables may be introduced. If ℓ is determined as $\ell = N-m-n$, then the ℓ τ-variables may be added to (5.2) resulting in

$$\sum_{k=0}^{N} \sum_{j=0}^{n} b_j \, c_{k-j} \, x^k = \sum_{k=0}^{m} a_k \, x^k + \sum_{k=n+m+1}^{N} \tau_k \, T_k \, (x) \tag{5.3}$$

The value of N is thus seen to be just $n+m+\ell$. The Tchebycheff polynomials must, of course, be scaled to the appropriate interval of x. Writing the Tchebycheff polynomials in their powers of x gives

$$T_k \, (x) = \sum_{r=0}^{k} s_{kr} \, x^r \tag{5.4}$$

where the coefficients s_{kr} are found according to the relation

$$s_{k,2r} = (-1)^r \, \frac{k}{k-r} \, \frac{(k-r)!}{(k-2r)! \ r!} \, 2^{(k-2r-1)} \ ,$$

$$s_{k,2r+1} = 0, \quad r = 0, \ 1, \ \cdots \ k/2 \tag{5.5}$$

In this notation (5.3) becomes

$$\sum_{k=0}^{N} \sum_{j=0}^{n} b_j \, c_{k-j} \, x^k = \sum_{k=0}^{m} a_k \, x^k + \sum_{k=n+m+1}^{N} \sum_{r=0}^{k} \tau_k \, s_{kr} \, x^r \tag{5.6}$$

The double sum on the right may be reversed if the upper limit on r is set to N. Then (5.6) may be rewritten to yield the basic equation for the Tau-Padé combination,

$$\sum_{k=0}^{N} \sum_{j=0}^{n} b_j \, c_{k-j} \, x^k = \sum_{k=0}^{m} a_k \, x^k + \sum_{r=0}^{N} \sum_{k=n+m+1}^{N} \tau_k \, s_{kr} \, x^r \qquad (5.7)$$

with the understanding that $c_{k-j} = 0$ for $k < j$ and $s_{kr} = 0$ for $k < r$. If a step function of the indeces is defined as

$$u(k,j) = \begin{cases} 0, & k < j \\ 1, & k \geq j \end{cases} \qquad (5.8)$$

then the restrictions on c_{k-j} and s_{kr} may be expressed qualitatively in the basic equation:

$$\sum_{k=0}^{N} \sum_{j=0}^{n} b_j \, c_{k-j} \, u(k,j) \, x^k = \sum_{k=0}^{m} a_k \, x^k + \sum_{r=0}^{N} \sum_{k=n+m+1}^{N} \tau_k \, s_{kr} \, u(k,r) \, x^r \qquad (5.9)$$

The use of the step function, u, is most advantageous from a programming standpoint.

Following the line of development of the Padé method, the coefficients of like powers of x may be equated to yield the necessary system of linear coefficient equations. These may also be written in matrix notation. As before, let m=n for clarity, and normalize the b_0 coefficient to one. The resulting system is then given in Figure 5.1. Note that the coefficient matrix is essentially the same as for the Padé method except that now it is augmented by the inclusion of the scaled Tchebycheff coefficients. Correspondingly, the vector of unknowns is

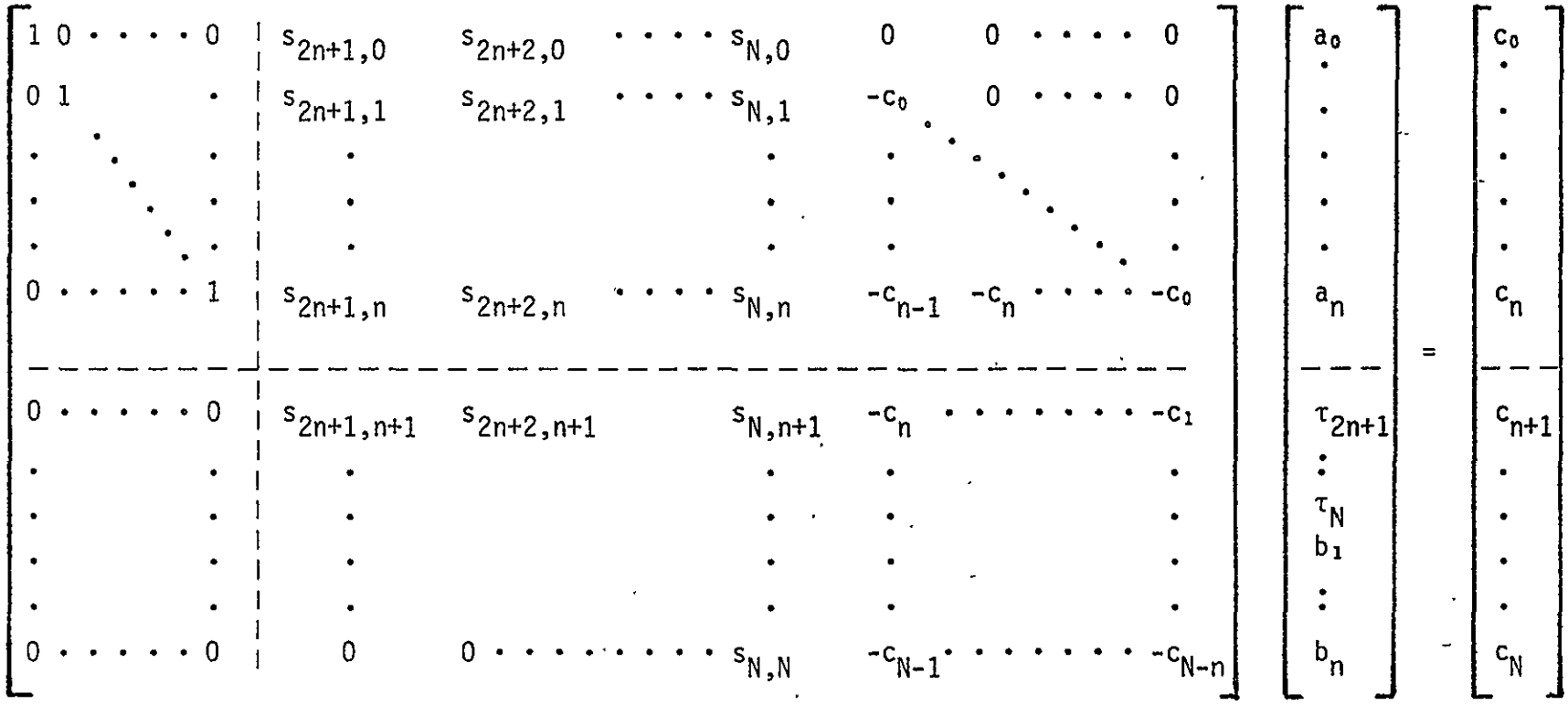


Figure 5.1 Matrix Form of Tau-Padé Coefficient Equations

augmented by the addition of the τ's. Partitioning the system yields, as before,

$$\left[\begin{array}{c|c} I & H \\ \hline 0 & G \end{array}\right] \left[\begin{array}{c} a \\ \hline b \end{array}\right] = \left[\begin{array}{c} h \\ \hline g \end{array}\right] \tag{5.10}$$

with the appropriate sub-matrices defined as follows:

$[I]$ = identity matrix, $(n+1) \times (n+1)$

$[0]$ = zero matrix, $(N-n) \times (n+1)$

$$[H] = \begin{bmatrix} s_{2n+1,0} & \cdots & s_{N,0} & 0 & \cdots & 0 \\ s_{2n+1,1} & \cdots & s_{N,1} & -c_0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ s_{2n+1,n} & \cdots & s_{N,n} & -c_{n-1} & \cdots & -c_0 \end{bmatrix}, \quad (n+1) \times (N-n) \tag{5.11}$$

$$[G] = \begin{bmatrix} s_{2n+1,n+1} & \cdots & s_{N,n+1} & -c_n & \cdots & -c_1 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots\cdots & s_{N,N} & -c_{N-1} & \cdots & -c_{N-n} \end{bmatrix}, \quad (N-n) \times (N-n)$$

$$(h) = \begin{bmatrix} c_0 \\ \vdots \\ c_n \end{bmatrix}, \quad (n+1) \times 1; \qquad (g) = \begin{bmatrix} c_{n+1} \\ \vdots \\ c_N \end{bmatrix}, \quad (N-n) \times 1$$

$$(a) = \begin{bmatrix} a_0 \\ \vdots \\ a_n \end{bmatrix}, \quad (n+1) \times 1; \qquad (b) = \begin{bmatrix} \tau_{2n+1} \\ \vdots \\ \tau_N \\ b_1 \\ \vdots \\ b_n \end{bmatrix}$$

In the general case the dimensions of the sub-matrices are

$$\begin{aligned} &[I],\ (m+1)\times(m+1);\quad [0],\ (N-m-1)\times(m+1) \\ &[H],\ (m+1)\times(N-m-1);\quad [G],\ (N-m-1)\times(N-m-1) \\ &(a),\ (h),\ (m+1)\times 1;\quad (b),\ (g),\ (N-m-1)\times 1 \end{aligned} \tag{5.12}$$

The corresponding expressions for the solution of the augmented system are identical with those for the Padé coefficients,

$$(b) = [G]^{-1}(a),\ (a) = (h) - [H](b) \tag{5.13}$$

Since the application of the Tau method adds ℓ τ-variables to the Padé equations, it is convenient to alter the notation slightly to reflect the number of τ-variables added. This is done by merely adding a third subscript to the symbol R_{mn}. Hence a rational approximation of order m,n found from the Tau-Padé equations using ℓ τ-variables is indicated by $R_{mn\ell}$.

5.2 Examples

The following examples serve to illustrate the Tau-Padé combination and to bring out some important aspects of the method.

$f(x) = e^x$. This first example already has its Taylor expansion in the fundamental form introduced in Chapter 3.

$$f(x) = e^x = 1 + x + 1/2!\ x^2 + 1/3!\ x^3 + \cdots \tag{5.14}$$

For the desired rational form let $m = n = 2$. Let the number of τ-variables, ℓ, be six. Thus, $N = \ell + m + n = 10$. The choice of ℓ is rather arbitrary, being suggested primarily by experience. For rational forms of order (2,2) rarely will more than six τ's be required, and four will often suffice. For this example let $x \in [0,1]$; then the scaled Tchebycheff coefficients

correspond to the coefficients of the shifted Tchebycheff polynomials,

$$s_{kr} = (-1)^r \frac{k}{k-r} \frac{(k-r)!}{(k-2r)!\, r!} 2^{(k-2r-1)}, \qquad (5.15)$$

$$r = 0, 1, 2, \cdots, k$$

Under these specifications the components of the partitioned augmented coefficient matrix are

$$[H] = \begin{bmatrix} s_{50} & \cdots & 0 & 0 \\ \vdots & & -c_0 & 0 \\ s_{52} & \cdots & -c_1 & -c_0 \end{bmatrix}, \quad 3 \times 8$$

$$[G] = \begin{bmatrix} s_{53} & \cdots & s_{10,3} & -c_2 & -c_1 \\ \vdots & & \vdots & & \vdots \\ s_{55} & & & & \\ 0 & \ddots & \vdots & & \vdots \\ \vdots & & & & \\ 0 & \cdots & s_{10,10} & -c_9 & -c_8 \end{bmatrix}, \quad 8 \times 8 \qquad (5.16)$$

$$(h)^T = (c_0 \; c_1 \; c_2)$$

$$(g)^T = (c_3 \; c_4 \; c_5 \; c_6 \; c_7 \; c_8 \; c_9 \; c_{10})$$

Solving the system according to (5.13) yields the set of values listed below. For comparison the corresponding Padé coefficients are also given.

Tau-Padé	Padé
$a_0 = 1.0000031$	$a_0 = 1.0$
$a_1 = .54164234$	$a_1 = .5$
$a_2 = .10792084$	$a_2 = .0833333$

$\tau_5 = 3.49986928 \times 10^{-6}$

$\tau_6 = 4.36506101 \times 10^{-7}$

$\tau_7 = 3.09266682 \times 10^{-8}$

$\tau_8 = 1.55708614 \times 10^{-9}$

$\tau_9 = 5.66904584 \times 10^{-11}$

$\tau_{10} = 1.19460073 \times 10^{-12}$

$b_0 = 1.0$	$b_0 = 1.0$
$b_1 = -.45821125$	$b_1 = -.5$
$b_2 = 6.50542644 \times 10^{-2}$	$b_2 = .0833333$

$$e^x \simeq R_{226} = \frac{a_0 + a_1 x + a_2 x^2}{b_0 + b_1 x + b_2 x^2}, \quad x \quad 0,1 \qquad (5.17)$$

Figure 5.2 plots the error functions associated with the Taylor series (δ_T), the Padé method (δ_p), and the Tau-Padé combination (δ_τ). For the Taylor and Padé approximations five terms of the expansion (5.14) were used. Although eleven terms were used to determine the coefficients in (5.17), the results require no more computational effort for evaluation of the rational form than for the standard Padé form. The maximum of the Tau-Padé error, δ_τ, is three orders of magnitude less than for δ_p and δ_T. Under the criterion for an optimal approximation (Chapter 2), the Tau-Padé approximation is nearly optimum. The error curve has $2 + \max(m + \partial Q, n + \partial P) = 6$ alternations which approach T-alternations:

$$\min ||\delta_\tau|| = 3.09 \times 10^{-6}, \quad \max ||\delta_\tau|| = 6.68 \times 10^{-6}$$

$\underline{f(x) = \ell n(1+x)}$. As a second illustration, consider the series expansion for $\ell n(1+x)$ whose fundamental form is

$$f(x) = \ell n(1+x) = 0 + x - 1/2\, x^2 + 1/3\, x^3 - 1/4\, x^4 + \cdots \qquad (5.18)$$

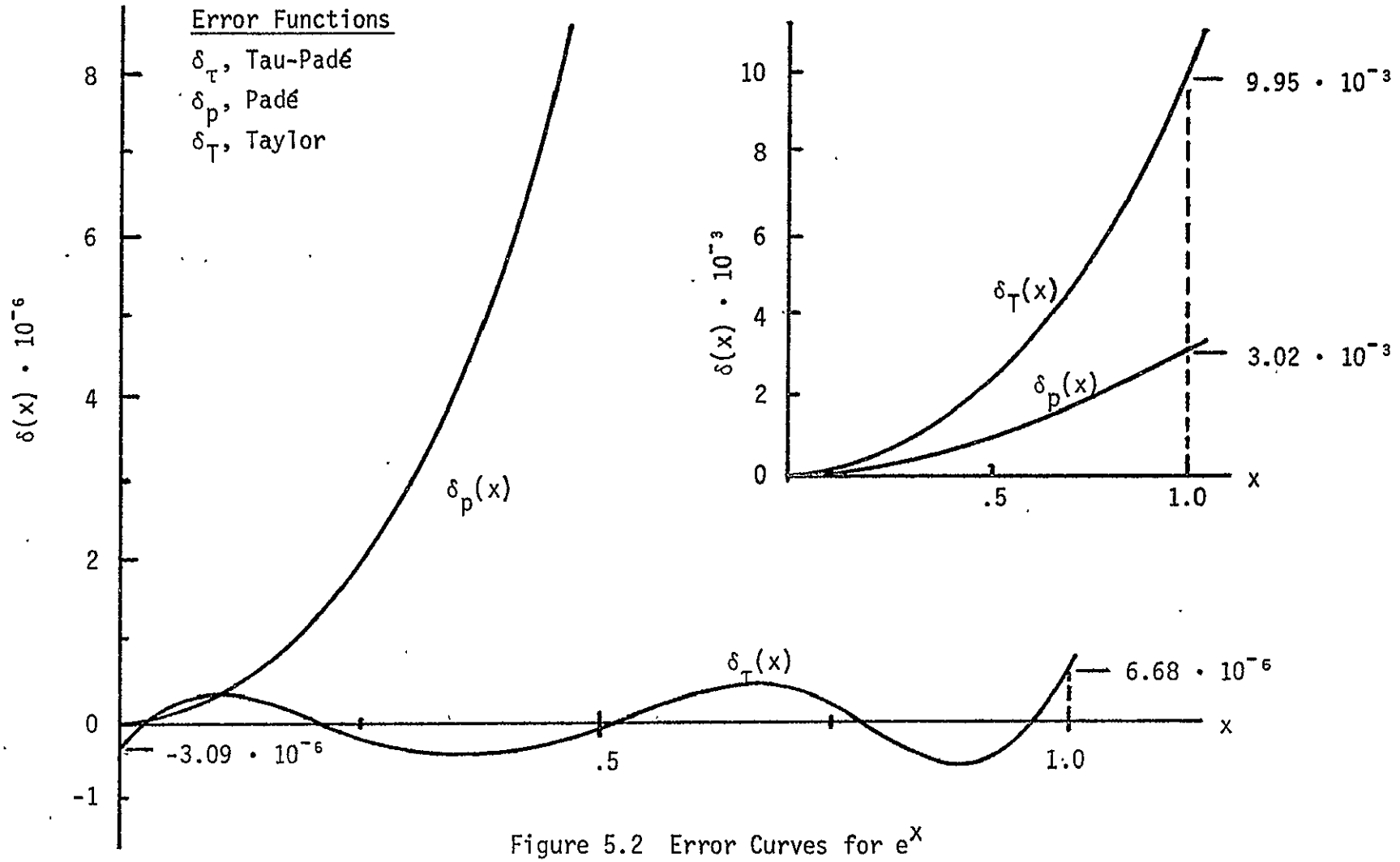


Figure 5.2 Error Curves for e^x

As before an approximation of order 2,2 for $x \in [0,1]$ will be obtained using six τ-variables. Solution of the system of augmented coefficient equations yields the following values which are compared with the standard Padé method:

Tau-Padé	Padé
$a_0 = -2.05651975 \times 10^{-5}$	$a_0 = 0.0$
$a_1 = 1.0009656$	$a_1 = 1.0$
$a_2 = .62730344$	$a_2 = .5$
$\tau_5 = -2.47728605 \times 10^{-5}$	
$\tau_6 = -5.59521770 \times 10^{-6}$	
$\tau_7 = -1.62740556 \times 10^{-6}$	
$\tau_8 = -2.70964554 \times 10^{-7}$	
$\tau_9 = -3.27816790 \times 10^{-8}$	
$\tau_{10} = -1.63798206 \times 10^{-9}$	
$b_0 = 1.0$	$b_0 = 1.0$
$b_1 = 1.1344666$	$b_1 = 1.0$
$b_2 = .21541081$	$b_2 = .16666667$

Figure 5.3 shows the various error functions, δ_τ, δ_p, δ_T, for the approximations to $\ell n(1+x)$. In this example the Tau-Padé error behaves similar to an optimal error curve up to $x \simeq 0.8$. Thereafter, however, it grows rapidly, much as the Padé error. Also there is not as extensive an advantage with the Tau-Padé combination in this case as in the previous example, δ_p and δ_τ being closer. The Taylor error appears to grow extremely rapidly. This behavior is, of course, inherent in the nature of the defining series. Consideration of the convergence of the series offers a possible explanation. Because the convergence is quite slow,

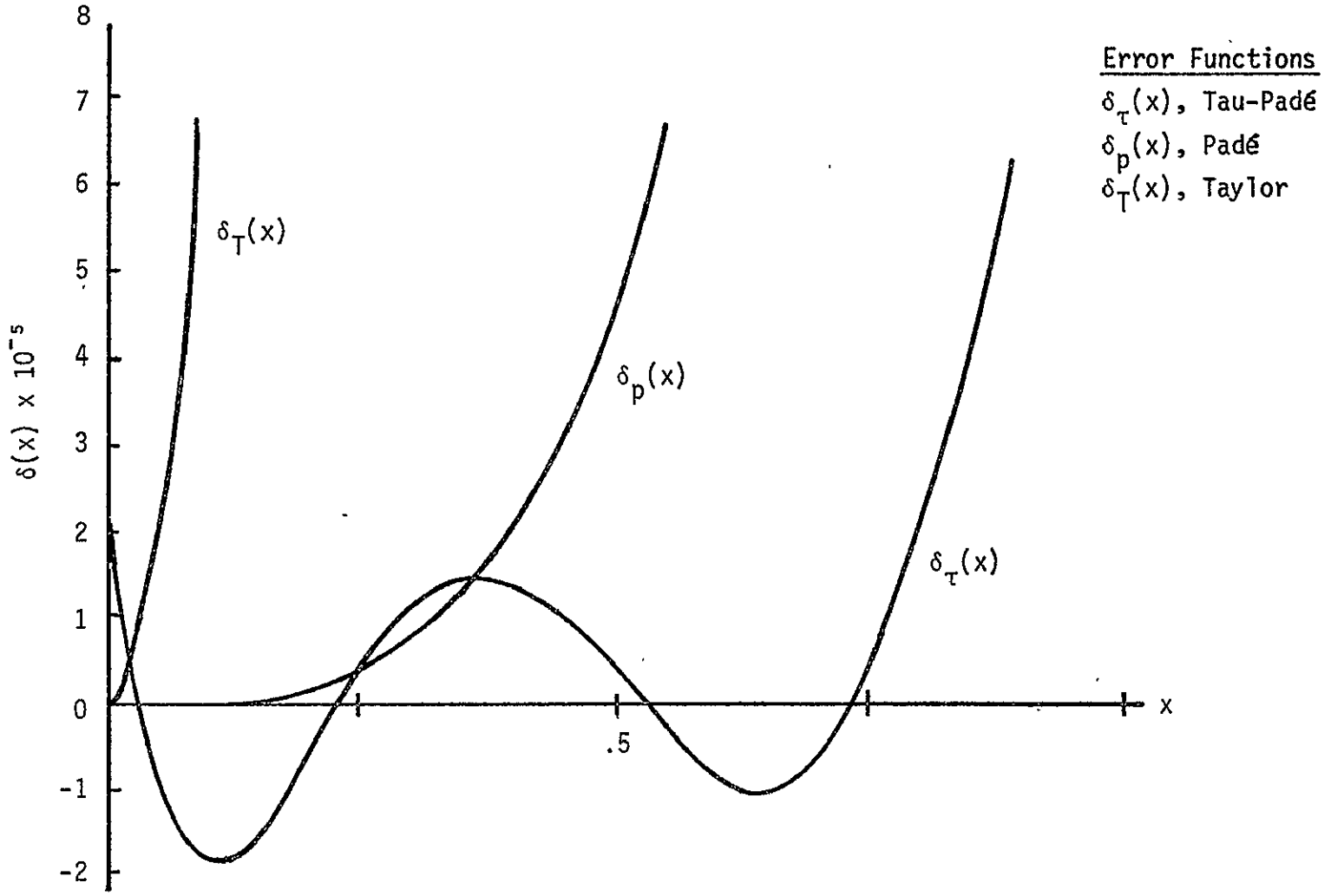


Figure 5.3 Error Curves for ℓn(1+x)

six τ-variables may be not be enough to obtain the desired near-optimal behavior. Also, near the upper limit of the interval the convergence becomes extremely slow, divergence occurring at $x = 1.0$.

$f(x) = 1/(1 + 2x + x^2)$. An example is now offered which is actually an irreducible rational function. Now one would expect that both the Padé and Tau-Padé methods should produce the function exactly. If the indicated division is carried out, the following series results:

$$f(x) = 1 - 2x + 3x^2 - 4x^3 + \cdots \tag{5.14}$$

For m=0, n=2, the Padé method gives back the rational form of f(x). In fact, in all cases where f(x) is itself an irreducible rational function of the form (1.1), and m and n have the same values for the approximation as for the function, $\hat{m}$ and $\hat{n}$, the Padé method simply returns the original function. Similarly, using the Tau-Padé combination with any number of τ-variables gives the exact form of f(x), the τ-variables being zero. In other words, if the function to be approximated is itself a member of Γ_{mn} then the Tau-Padé method yields the function exactly regardless of the number of τ-variables which may be introduced.

A natural question is to inquire about the performance of the method when m and n do not happen to coincide with the actual order, $\hat{m}$, $\hat{n}$, of the rational function, f. First if either or both m and n are less, the resulting approximation merely will be a lower order approximation. Now if $m > \hat{m}$, $n = \hat{n}$ or if $m = \hat{m}$, $n > \hat{n}$ the Tau-Padé method will yield the correct a's and b's with

$$a_{\hat{m}+1} = \cdots = a_m = 0 \text{ for } m > \hat{m},$$

$$b_{\hat{n}+1} = \cdots = b_n = 0 \text{ for } n > \hat{n}.$$

up through x^N be considered even though it does not appear explicitly in the series. The most consistent way to insure this is through reduction to fundamental form.

Proceeding with the example, let $\ell=4$ and again $x \in [0,1]$. Then the Tau-Padé equations yield the following parameter values:

$$a_0 = 8.06609950 \times 10^{-8} \qquad b_0 = 1.0$$

$$a_1 = .9998780 \qquad b_1 = .34753262$$

$$a_2 = .34783041 \qquad b_2 = 1.0408238$$

$$a_3 = .70477263 \qquad b_3 = .29418535$$

$$a_4 = .19018504 \qquad b_4 = .17348426$$

$$\tau_9 = 1.06674826 \times 10^{-7}$$

$$\tau_{10} = 2.89900133 \times 10^{-8}$$

$$\tau_{11} = 3.10652367 \times 10^{-9}$$

$$\tau_{12} = 1.30341767 \times 10^{-10}$$

The error functions for both the standard Padé and Tau-Padé methods are shown in Figure 5.4.

An evaluation of π may again be obtained,

$$4\, R_{444}(1) = 3.141114136$$

with an error of 4.775×10^{-4}. To obtain the same accuracy by merely summing the expansion requires 2090 terms.

5.3 Error Expression

The error expression for the Tau-Padé combination is found the same way as for the standard Padé method. Substituting the basic equation (5.7) into the general expression (3.4) and solving for $\delta(x)$ yields

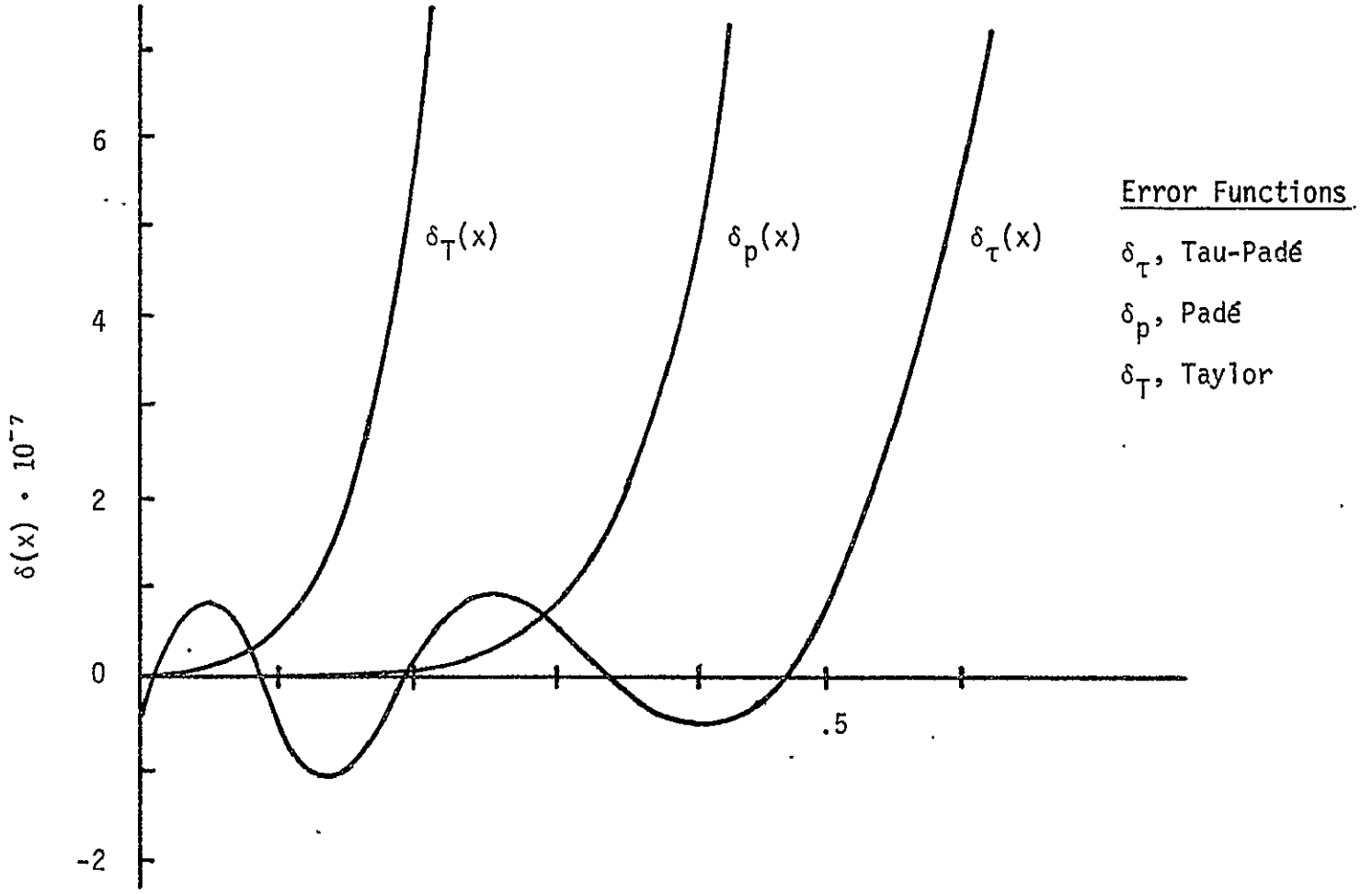


Figure 5.4 Error Curves for $\tan^{-1}(x)$

$$\delta(x) = \frac{\sum_{k=N+1}^{\infty} \sum_{j=0}^{n} b_j \, c_{k-j} \, x^k + \sum_{r=0}^{N} \sum_{k=n+m+1}^{N} \tau_k \, s_{kr} \, x^r}{\sum_{j=0}^{n} b_j \, x^j}, \quad (s_{kr} = 0, \; k < r) \tag{5.21}$$

An interesting result is the limit of (5.16) as $N \to \infty$:

$$\delta_\infty(x) = \lim_{N \to \infty} \delta(x) = \frac{\sum_{k=n+m+1}^{\infty} \tau_k \, T_k(x)}{\sum_{j=0}^{n} b_j \, x_j} \tag{5.22}$$

For convergent series a rough approximation of the order of magnitude of the error may be obtained by this expression. In general the maximum value of each T_k occurs at the interval limits and is ± 1.0. When x is normalized such that $|x| \in [0,1]$, (5.22) reduces to

$$\delta_\infty(1) = \frac{\sum_{k=n+m+1}^{\infty} \tau_k}{\sum_{j=0}^{n} b_j} \tag{5.23}$$

Considering the example for e^x, (5.23) yields

$$\delta_\infty(1) = \frac{3.969 \times 10^{-6}}{.607} = 6.539 \times 10^{-6}$$

which is slightly less than the actual maximum error, 6.68×10^{-6}. Caution must be used when using this method, however, since it loses validity when the approximations are not close to the optimal.

5.4 Applications to Kepler's Equation

A fundamental equation of astrodynamics is the transcendental relation between time and angular position of an orbiting body.

For elliptic motion: $M = E - e \sin E$ (5.24a)

For hyperbolic motion: $M = e \sinh H - H$ (5.24b)

M is the mean anomaly (a function of the mean motion and the time), e is the eccentricity, and E and H are the eccentric and hyperbolic anomalies, respectively. As transcendental equations, (5.24a,b) generally are solved for E and H using some iterative technique such as the method of successive approximations or the Newton-Raphson method. Such algorithms require reasonable starting values which are obtained by some method of approximation. In this section attention is given to the problem of obtaining solutions for E or H using the Tau-Padé equations. In the final paragraphs, Battin's modification (2) of Herrick's universal variable formulation of Kepler's equation will be introduced, and a different application of the Tau-Padé combination will be made.

Approximations to Sin and Sinh. If rational approximations for the sine and hyperbolic sine functions are obtained in the form

$$R_{21}(x) = \frac{a_0 + a_1 x + a_2 x^2}{1 + b_1 x} \tag{5.25}$$

then (5.24a) and (5.24b) can be written in the common form,

$$M = x - e\left(\frac{a_0 + a_1 x + a_2 x^2}{1 + b_1 x}\right) \tag{5.26}$$

where x is E or H as required. Further, it is necessary to obtain the approximations only over the positive values of x, since both functions

are odd, and the sign on x is positive or negative according to the sign on M. The identification of elliptic or hyperbolic motion may be made depending on whether $e < 1$ or $e > 1$, respectively (the case $e = 1$ represents parabolic motion; the time-position relation for this motion requires no iterative technique for solution). Thus (5.26) may be re-written as

$$\gamma M = x - \sigma e \left(\frac{a_0 + a_1 x + a_2 x^2}{1 + b_1 x} \right) \tag{5.27}$$

where

$$\sigma = M/|M| \ , \quad \gamma = (1-e)/|1-e|$$

Multiplying (5.27) by $1 + b_1 x$ and combining coefficients of like powers of x gives the quadratic,

$$Ax^2 + Bx + C = 0 \tag{5.28}$$

where

$$A = b_1 - \sigma e a_2$$

$$B = 1 - \sigma e a_1 - \gamma M b$$

$$C = -\sigma e a_0 - \gamma M$$

which may then be solved for x.

The values of the a and b coefficients will now be found for the trigonometric approximations. First, scaling of the Tchebycheff coefficients is done for $x \in [0,\pi]$ since sine is periodic and odd. For sinh the interval $x \in [0,\pi/2]$ is chosen with the value $\pi/2$ being selected arbitrarily since the hyperbolic sine is not periodic. The number of τ-variables selected is eight. The results are summarized below.

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

$a_0 = -2.738634 \times 10^{-2}$ $\qquad$ $b_0 = 1.0$

$a_1 = 1.271591$ $\qquad$ $b_1 = -1.155438 \times 10^{-4}$

$a_2 = -0.404789$

$$\sinh x = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \cdots$$

$a_0 = 2.565447 \times 10^{-3}$ $\qquad$ $b_0 = 1.0$

$a_1 = 0.948262$ $\qquad$ $b_1 = -0.356916$

$a_2 = -0.193801$

The plots of the error curves for these approximations are shown in Figures 5.5a,b. Note that the errors are relatively large. One reason is because the primary contributor to the approximations is the first term in the Taylor expansions. Another important effect is that generally, the larger the interval, the harder it is to approximate the function. In view of this, one might expect that the resulting solutions of (5.28) would not yield very accurate values of x. This is indeed the case as seen in Figures 5.6a,b.* The question as to which root of the quadratic to use has not been formally answered. However, experience has shown that using

$$x = (-B + u\sqrt{B^2 - 4AC}\,)\,/2A,$$

$$u = \begin{cases} +1, & \text{ellipse} \\ -1, & \text{hyperbola} \end{cases}$$

---

*Undoubtedly a better approximation would result if one found $R_{32\ell}$ for sin and sinh, and recast (5.28) as a cubic. However, cubic equations are not particularly convenient to solve.

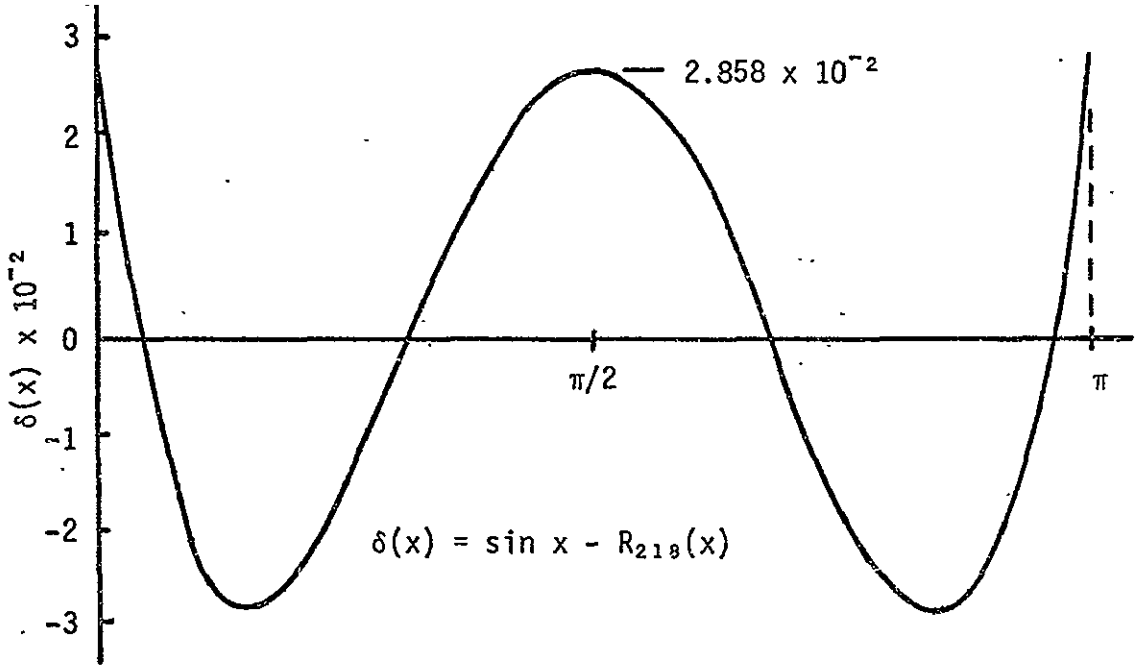


Figure 5.5a Error for Sin x

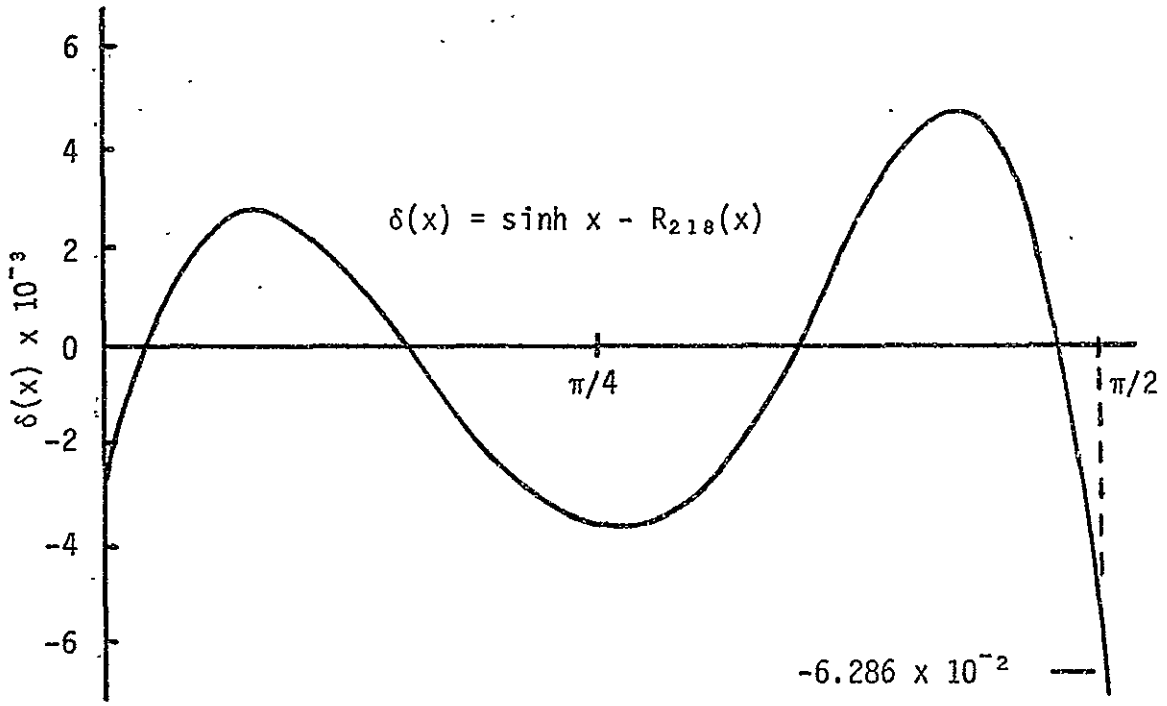


Figure 5.5b Error for Sinh x

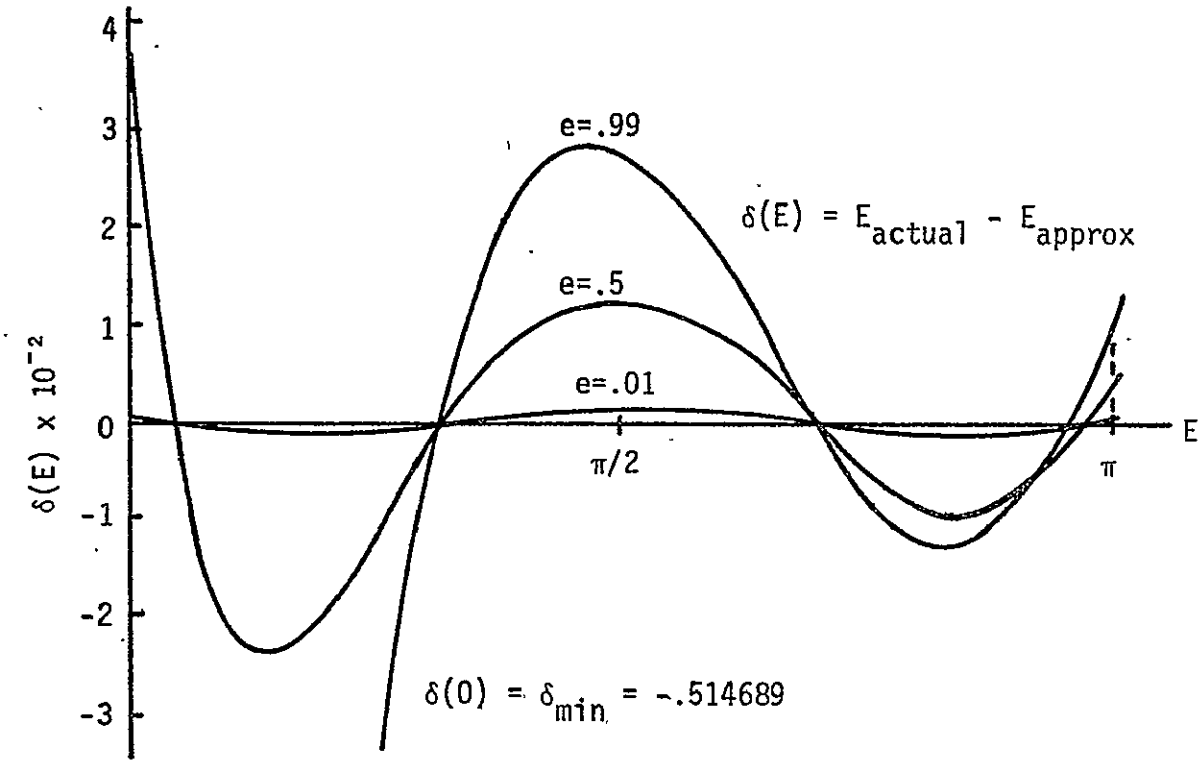


Figure 5.6a Error Curves for E

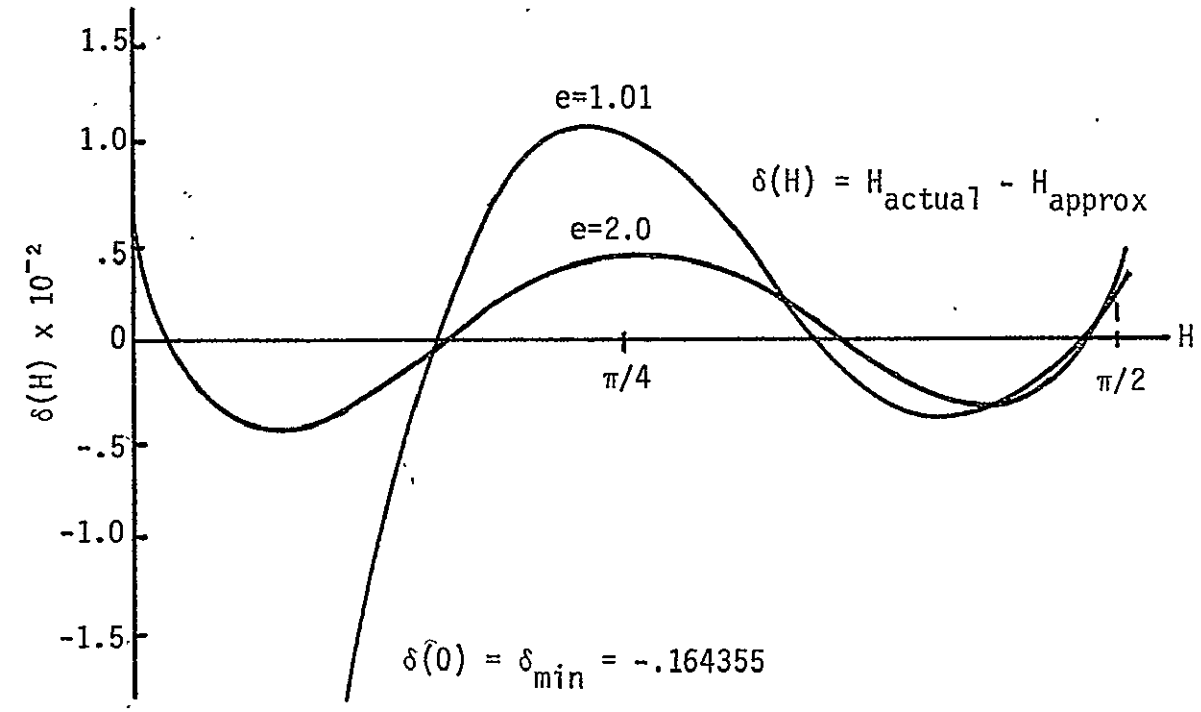


Figure 5.6b Error Curves for H

consistently yields compatible results.

In Figures 5.6a,b there is another effect present which is due to the nature of Kepler's equation itself. As $e \to 1$ and $x \to 0$, a region of non-uniform behavior is approached, emphasized by the fact that the error curves go off scale. Moulton (14) has investigated this for the elliptic case, and Russell (16) has shown, using perturbation methods, that the non-uniform region is approached when $1 - e \cos E \to 0$. One empirical way of countering this is to set $x = M$ when

$$1 - e \cos M < \varepsilon_1, \text{ elliptic motion,}$$

$$e \cosh M - 1 < \varepsilon_2, \text{ hyperbolic motion,}$$

where ε_1 and ε_2 are some judiciously chosen values on the order of .01.

Variable Transformation. A second way to improve the accuracy for elliptic motion is one motivated by a suggestion by Gottlieb (8). Define a new variable,

$$w \equiv E - M = e \sin E \tag{5.30}$$

The limiting values are obviously $-1 \leq w \leq 1$. Kepler's equation may be rewritten in terms of w as

$$w = e \sin (w + M) \tag{5.31}$$

or

$$w = e(\sin w \cos M + \cos w \sin M) \tag{5.32}$$

Now since the interval of interest is now smaller, one could expect better accuracy. However, to retain the quadratic form (5.28) polynomials must be used to approximate sin w and cos w which partially offsets the advantage of a smaller interval. Approximating sine and cosine by R_{208},

$$\begin{aligned} \sin w &\simeq s_0 + s_1 w + s_2 w^2 \\ \cos w &\simeq c_0 + c_1 w + c_2 w^2 \end{aligned} \tag{5.33}$$

and defining

$$S_M \equiv \sin M, \quad C_M \equiv \cos M \tag{5.33}$$

(5.32) becomes

$$w = e\, C_M(s_0 + s_1 w + s_2 w^2) + e\, S_M(c_0 + c_1 w + c_2 w^2)$$

or

$$(e\, C_M\, s_2 + e\, S_M\, c_2)w^2 + (e\, C_M\, s_1 + e\, S_M\, c_1 - 1)w + (e\, C_M\, s_0 + e\, S_M\, c_0) = 0 \tag{5.35}$$

which may then be solved for w, and E subsequently found. Recalling that sine is an odd function, and cosine, even, the shifted Tchebycheff polynomial coefficients may be used in the Tau-Padé equations. Then the sine approximations must be multiplied by the σ defined in (5.27) so that (5.35) becomes

$$A\, w^2 + B\, w + C = 0, \tag{5.36}$$

with

$$A = e(C_M\, \sigma\, s_2 + S_M\, c_2)$$
$$B = e\, C_M\, \sigma\, s_1 + e\, S_M\, c_1 - 1$$
$$C = e\, (C_M\, \sigma\, s_0 + S_M\, c_0)$$

Here, the negative sign is used with the radical in the expression for the root, i.e., u = -1 in (5.29). Finally, solution of the corresponding Tau-Padé equations for coefficients of the R_{208} approximations yields

$$s_0 = -4.63945315 \times 10^{-3}, \qquad c_0 = 1.0021690$$
$$s_1 = 1.0851999, \qquad c_1 = -3.48778236 \times 10^{-2}$$
$$s_2 = -.23475756, \qquad c_2 = -.42972092$$

The associated error curves are shown in Figure 5.7, exhibiting the near-optimal characteristics (maxima of approximately equal magnitude, and

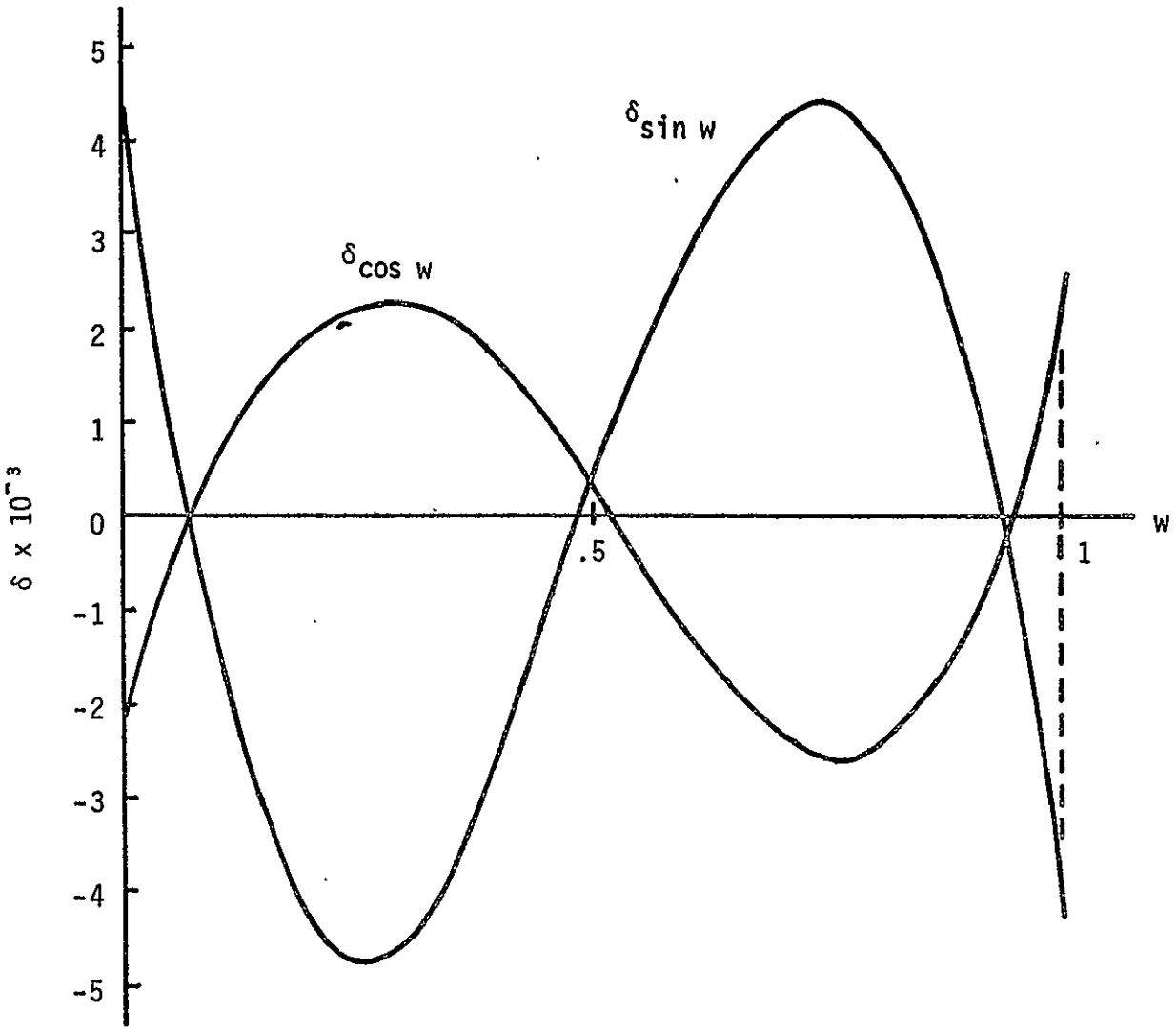


Figure 5.7 Error for Quadratic Approximation to Sin and Cos

four T-alternations). The reason that the sine error is larger than the cosine error is due to the number of terms employed in the Taylor series, only one term being taken for the sine as opposed to two for the cosine. Now even though the sine and cosine functions are approximated by R_{208} instead of a higher order $R_{mn\ell}$, the advantage of a smaller interval of approximation predominates. This is seen in Figure 5.8, showing the improved error behavior for E with e = .5 and e = .99. Immediately obvious, however, is the irregular behavior of these error curves. This is due to the fact that w is formed as a combination of the two quadratic approximations for sine and cosine, as shown by (5.32). This in turn maps into the solution of (5.36) in a non-linear fashion and hence the optimality exhibited in Figure 5.7 is not present in Figure 5.8.

Transcendental Functions for the Universal Variable Formulation. To this point, concern has been with obtaining approximate solutions of Kepler's equation for the eccentric or hyperbolic anomaly. A different application of the Tau-Padé combination is now made to a universal variable formulation of Kepler's equation. Such a form, originally introduced by S. Herrick, allows writing just one time -- angular position relationship for both elliptic and hyperbolic motion. A modified formulation is given by Battin (2) as

$$\sqrt{\mu}\, t = \frac{\bar{r}_0 \cdot \bar{v}_0}{\sqrt{\mu}} x^2\, C(\alpha_0 x^2) + (1 - r_0 \alpha_0) x^3\, S(\alpha_0 x^2) + r_0 x \tag{5.37}$$

where

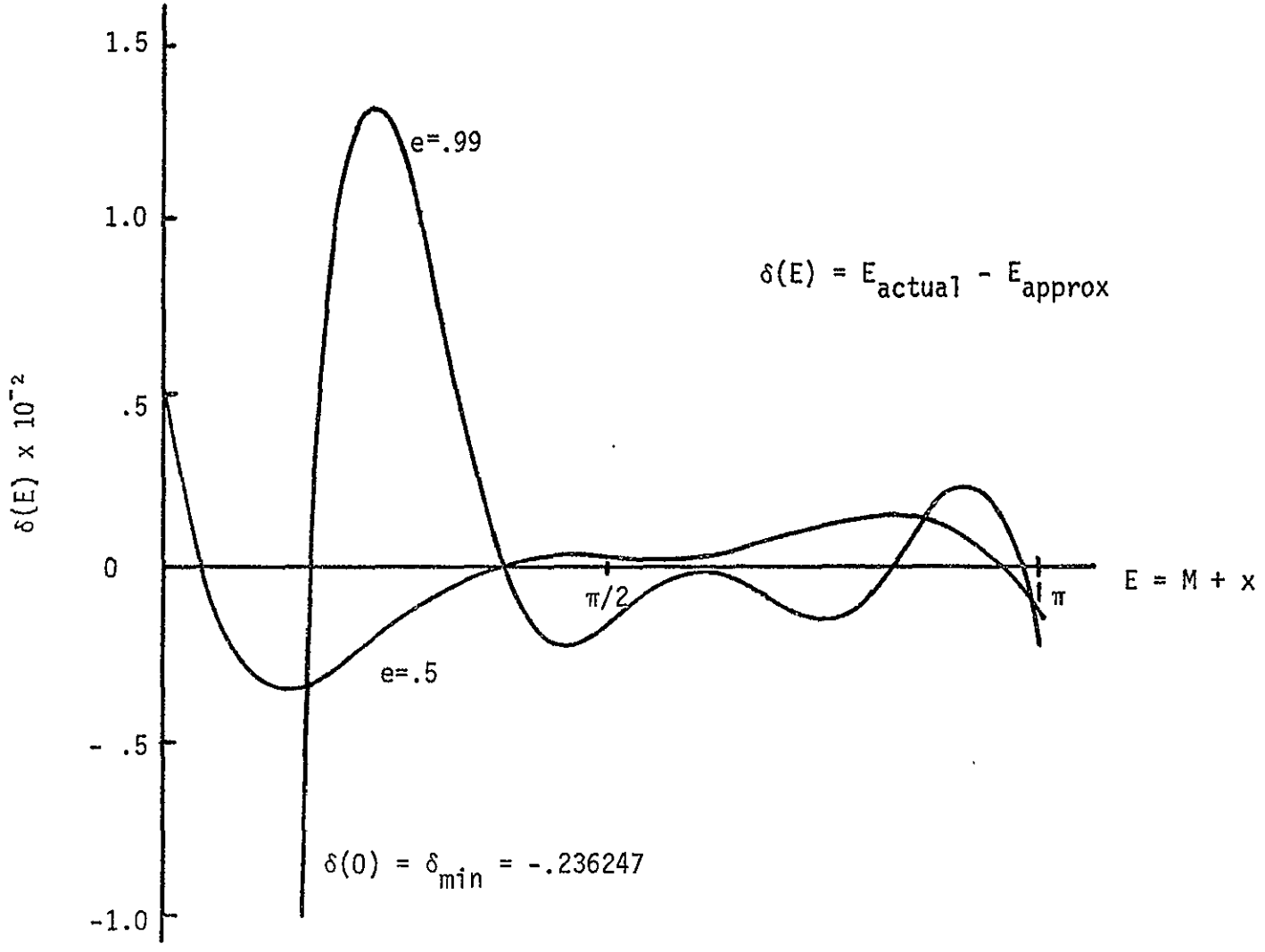


Figure 5.8 Eccentric Anomaly Error

$\bar{r}_0$, $\bar{v}_0$ = initial position and velocity vectors,

t = time from initial to final position,

μ = gravitational parameter of central body,

$\alpha_0 = 2/r_0 - v_0^2/\mu$; $r_0 = |\bar{r}_0|$, $v_0 = |\bar{v}_0|$

The parameter, x, is termed the universal variable, and defined as

$$x \equiv \begin{cases} \dfrac{E - E_0}{\sqrt{\alpha_0}}, & \text{elliptic motion} \\ \dfrac{H - H_0}{\sqrt{-\alpha_0}}, & \text{hyperbolic motion} \end{cases} \tag{5.38}$$

with E_0, E, H_0, H being the initial and final eccentric and hyperbolic anomalies respectively. The C and S are transcendental functions defined by Battin (2) as

$$S(u) = \frac{1}{3!} - \frac{u}{5!} + \frac{u^2}{7!} - \frac{u^3}{9!} + \cdots \tag{5.39}$$

$$C(u) = \frac{1}{2!} - \frac{u}{4!} + \frac{u^2}{6!} - \frac{u^3}{8!} + \cdots \tag{5.40}$$

While these series converge rapidly, over large intervals the development of their Tau-Padé approximants to high order is useful and informative. First, let $u = \alpha_0 x^2 \in \left[-(2\pi)^2, (2\pi)^2\right]$. Here, no symmetry of the S(u) and C(u) exists. Hence no advantage of symmetry is present, and the full range must be used. This presents a problem because the interval is large, and scaling of the Tchebycheff coefficients requires multiplying each coefficient by some power of the interval length. This quickly leads to numerical difficulties for high order approximations since, for example, the twelfth order Tchebycheff polynomial has a scaled coefficient of x^{10}

on the order of 10^{16}, while the lowest order coefficient is 1. This problem (which can be encountered often for large intervals) is readily circumvented by the alternative of scaling u to $[-1,1]$. Hence,

$$u^n \equiv (40w)^n \tag{5.41}$$

so that

$$S(w) = \frac{1}{3!} - \frac{40w}{5!} + \frac{(40w)^2}{7!} - \cdots \tag{5.42}$$

$$C(w) = \frac{1}{2!} - \frac{40w}{4!} + \frac{(40w)^2}{8!} - \cdots \tag{5.43}$$

for $w \in [-1,1]$. With these changes, solution of the Tau-Padé equations for the coefficients of R_{448} as approximations for C and S gives

C:

$a_0 = .50000005$

$a_1 = -1.3329425$

$a_2 = 1.2170535$

$a_3 = -.44140671$

$a_4 = 5.73370383 \times 10^{-2}$

$b_0 = 1.0$

$b_1 = .66744544$

$b_2 = .21448521$

$b_3 = 4.03536694 \times 10^{-2}$

$b_4 = 3.82170297 \times 10^{-3}$

$\tau_9 = -1.67491428 \times 10^{-7}$

$\tau_{10} = 4.70365685 \times 10^{-8}$

$\tau_{11} = -6.53462358 \times 10^{-9}$

$\tau_{12} = 5.91614229 \times 10^{-10}$

$\tau_{13} = -3.89852249 \times 10^{-11}$

$\tau_{14} = 1.98790997 \times 10^{-12}$

$\tau_{15} = -8.01643151 \times 10^{-14}$

$\tau_{16} = 2.72493162 \times 10^{-15}$

S:

$a_0 = .16666667$ $\quad$ $\tau_9 = -5.63359592 \times 10^{-9}$

$a_1 = -.23582030$ $\quad$ $\tau_{10} = 1.42027615 \times 10^{-9}$

$a_2 = .14947671$ $\quad$ $\tau_{11} = -1.78816313 \times 10^{-10}$

$a_3 = -4.03986018 \times 10^{-2}$ $\quad$ $\tau_{12} = 1.47893824 \times 10^{-11}$

$a_4 = 4.16704687 \times 10^{-3}$ $\quad$ $\tau_{13} = -8.96491447 \times 10^{-13}$

$b_0 = 1.0$ $\quad$ $\tau_{14} = 4.22974700 \times 10^{-14}$

$b_1 = .58507792$ $\quad$ $\tau_{15} = -1.58967531 \times 10^{-15}$

$b_2 = .16225461$ $\quad$ $\tau_{16} = 5.04612583 \times 10^{-17}$

$b_3 = 2.58883507 \times 10^{-2}$

$b_4 = 2.04894379 \times 10^{-3}$

The associated error curves for $w \in [-1,1]$ are shown in Figures 5.9a,b. Because of the rapid convergence of the series, the Padé and Taylor results are practically the same. Hence the two errors are shown as one.

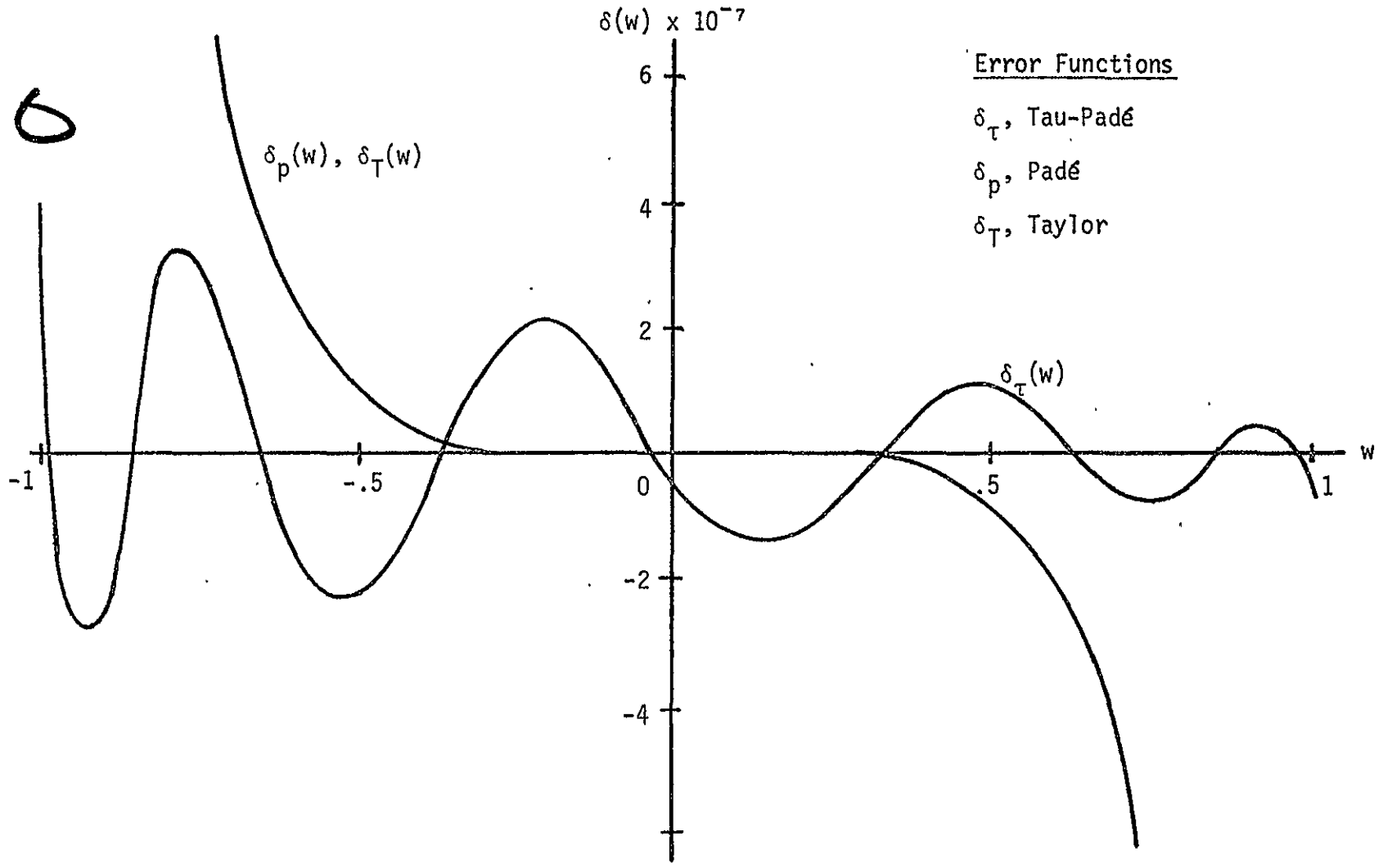


Figure 5.9a Error Curves for C

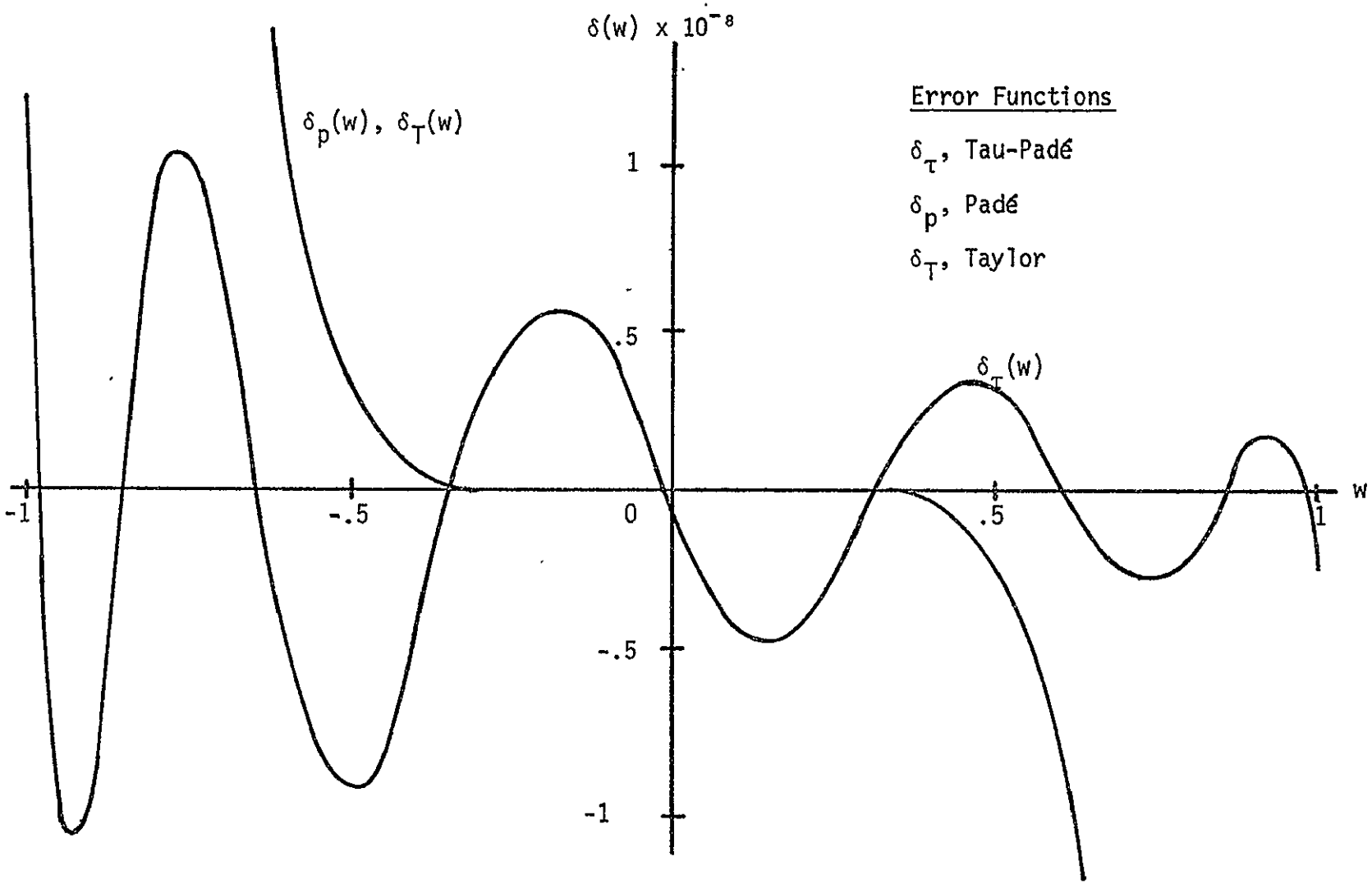


Figure 5.9b Error Curves for S

CHAPTER 6

EXPLICIT FORMS AND GENERALIZATIONS

The development of some explicit expressions for the coefficients and τ-variables for Tau-Padé approximations is outlined in this chapter. The resulting expressions are, in part, recursive in nature thus providing compact and easily handled formulas. Generalizations of the Padé and Tau-Padé algorithms are discussed briefly, noting that their utility is limited.

6.1 Explicit Expressions

Except for the simplist of problems, a computer must be used for solution of the Tau-Padé equations. With this in mind, efficiency may be gained in terms of computer time and storage by employing explicit formulas for some of the Tau-Padé parameters. The development of such formulas is illustrated by deriving the explicit forms for R_{114}. The extension to higher order forms will be obvious; explicit formulas for the parameters of $R_{mo\ell}$ and $R_{mn\ell}$ are given in the Appendix.

For R_{114} the Tau-Padé equations are

$$
\begin{aligned}
c_0 &= a_0 + \tau_3 s_{30} + \tau_4 s_{40} + \tau_5 s_{50} + \tau_6 s_{60} && \text{(a)}\\
c_1 + c_0 b_1 &= a_1 + \tau_3 s_{31} + \tau_4 s_{41} + \tau_5 s_{51} + \tau_6 s_{61} && \text{(b)}\\
c_2 + c_1 b_1 &= \tau_3 s_{32} + \tau_4 s_{42} + \tau_5 s_{52} + \tau_6 s_{62} && \text{(c)}\\
c_3 + c_2 b_1 &= \tau_3 s_{33} + \tau_4 s_{43} + \tau_5 s_{53} + \tau_6 s_{63} && \text{(d)}\quad (6.1)\\
c_4 + c_3 b_1 &= \tau_4 s_{44} + \tau_5 s_{54} + \tau_6 s_{64} && \text{(e)}\\
c_5 + c_4 b_1 &= \tau_5 s_{55} + \tau_6 s_{65} && \text{(f)}\\
c_6 + c_5 b_1 &= \tau_6 s_{66} && \text{(g)}
\end{aligned}
$$

As before, s_{kr} is the coefficient of x^r in T_k. The approach for solving these equations here is to use Gauss reduction. Hence (6.1g) is solved for τ_6,

$$\tau_6 = \frac{1}{s_{66}} (c_6 + c_5 b_1) \tag{6.2a}$$

Substituting into (6.1f) τ_5 is found as

$$\tau_5 = \frac{1}{s_{55}} \left\{ c_5 - \frac{s_{65}}{s_{66}} c_6 + \left(c_4 - \frac{s_{65}}{s_{66}} c_5\right) b_1 \right\} \tag{6.2b}$$

Similarly, expressions for the other τ's are found.

$$\begin{aligned} \tau_4 = \frac{1}{s_{44}} \Bigg\{ & c_4 - \frac{s_{54}}{s_{55}} c_5 + \left(\frac{s_{54}s_{65}}{s_{55}s_{66}} - \frac{s_{64}}{s_{66}} \right) c_6 \\ & + \left[c_3 - \frac{s_{54}}{s_{55}} c_4 + \left(\frac{s_{54}s_{65}}{s_{55}s_{66}} - \frac{s_{64}}{s_{66}} \right) c_5 \right] b_1 \Bigg\} \end{aligned} \tag{6.2c}$$

$$\begin{aligned} \tau_3 = \frac{1}{s_{33}} \Bigg\{ & c_3 - \frac{s_{43}}{s_{44}} c_4 + \left(\frac{s_{43}s_{54}}{s_{44}s_{55}} - \frac{s_{53}}{s_{55}} \right) c_5 \\ & + \left(- \frac{s_{43}s_{54}s_{65}}{s_{44}s_{55}s_{66}} + \frac{s_{43}s_{64}}{s_{44}s_{66}} + \frac{s_{53}s_{65}}{s_{55}s_{66}} - \frac{s_{63}}{s_{66}} \right) c_6 \\ & + \left[c_2 - \frac{s_{43}}{s_{44}} c_3 + \left(\frac{s_{43}s_{54}s_{65}}{s_{44}s_{55}s_{66}} - \frac{s_{53}}{s_{55}} \right) c_4 \right. \\ & \left. + \left(- \frac{s_{43}s_{54}s_{65}}{s_{44}s_{55}s_{66}} + \frac{s_{43}s_{64}}{s_{44}s_{66}} + \frac{s_{53}s_{65}}{s_{55}s_{66}} - \frac{s_{63}}{s_{66}} \right) c_5 \right] b_1 \Bigg\} \end{aligned} \tag{6.2d}$$

Substituting these expressions into (6.1c) yields

$$-c_2 + \frac{s_{32}}{s_{33}} c_3 + \left(\frac{s_{42}}{s_{44}} - \frac{s_{32}s_{43}}{s_{33}s_{44}} \right) c_4 + \left(\frac{s_{32}s_{43}s_{54}}{s_{33}s_{44}s_{55}} - \frac{s_{32}s_{53}}{s_{33}s_{55}} - \frac{s_{42}s_{54}}{s_{44}s_{55}} + \frac{s_{52}}{s_{55}} \right) c_5 + \left(- \frac{s_{32}s_{43}s_{54}s_{65}}{s_{33}s_{44}s_{55}s_{66}} + \frac{s_{32}s_{43}s_{64}}{s_{33}s_{44}s_{66}} + \frac{s_{32}s_{53}s_{65}}{s_{33}s_{55}s_{66}} - \frac{s_{32}s_{63}}{s_{33}s_{66}} + \frac{s_{42}s_{54}s_{65}}{s_{44}s_{55}s_{66}} - \frac{s_{64}s_{42}}{s_{66}s_{44}} - \frac{s_{52}s_{65}}{s_{55}s_{66}} + \frac{s_{62}}{s_{66}} \right) c_6 + \left[-c_1 + \frac{s_{32}}{s_{33}} c_2 + \left(\frac{s_{42}}{s_{44}} - \frac{s_{32}s_{43}}{s_{33}s_{44}} \right) c_3 + \left(\frac{s_{32}s_{43}s_{54}}{s_{33}s_{44}s_{55}} - \frac{s_{32}s_{53}}{s_{33}s_{55}} - \frac{s_{42}s_{54}}{s_{44}s_{55}} + \frac{s_{52}}{s_{55}} \right) c_4 + \left(- \frac{s_{32}s_{43}s_{54}s_{65}}{s_{33}s_{44}s_{55}s_{66}} + \frac{s_{32}s_{43}s_{64}}{s_{33}s_{44}s_{66}} + \frac{s_{32}s_{53}s_{65}}{s_{33}s_{55}s_{66}} - \frac{s_{32}s_{63}}{s_{33}s_{66}} + \frac{s_{42}s_{54}s_{65}}{s_{44}s_{55}s_{66}} - \frac{s_{64}s_{42}}{s_{66}s_{44}} - \frac{s_{52}s_{65}}{s_{55}s_{66}} + \frac{s_{62}}{s_{66}} \right) c_5 \right] b_1 = 0 \tag{6.2e}$$

This expression may then be solved for b_1. However, in its present form (6.2e) is extremely cumbersome. Examination of the terms yields a convenience which greatly facilitates the determination of the unknown coefficients. Judiciously grouping the rational forms of the s's results in the following recursion formulas:

$$\begin{aligned} D_1 &= \frac{s_{32}}{s_{33}} \\ D_2 &= \frac{1}{s_{44}} (s_{42} - D_1 s_{43}) \\ D_3 &= \frac{1}{s_{55}} (s_{52} - D_1 s_{53} - D_2 s_{54}) \\ D_4 &= \frac{1}{s_{66}} (s_{62} - D_1 s_{63} - D_2 s_{64} - D_3 s_{65}) \end{aligned} \tag{6.3}$$

Then in terms of these quantities, b_1 is simply

$$b_1 = -\left(\frac{-c_2 + D_1c_3 + D_2c_4 + D_3c_5 + D_4c_6}{-c_1 + D_1c_2 + D_2c_3 + D_3c_4 + D_4c_5}\right) \tag{6.4}$$

The τ-variables are then easily obtained by using equations (6.1d) - (6.1g). Similarly, a_0 and a_1 may be found immediately from (6.1a) and (6.1b). As the foregoing implies, the algebra involved in the developments is quite tedious for even the simple forms of $R_{mn\ell}$. By induction on the formulas for R_{104}, R_{114}, R_{214}, and R_{224}, those for the general case, $R_{mn\ell}$, were found and are given in the Appendix.

6.2 Generalization

In Chapter 3 the classical Padé method was presented. It was shown to be applicable to those functions possessing a Taylor expansion. In fact, the Padé method is applicable to any function which can be represented as a linear combination of certain functions, for example, Legendre polynomials. Cheney has shown this in (4). By way of illustration, let f(x) be represented as

$$f(x) = \sum_{k=0}^{\infty} c_k \, \phi_k(x) \tag{6.5}$$

where the ϕ_k are functions of the single variable, x. A rational approximation to f,

$$R_{mn} = \frac{a_0 \, \phi_0 + a_1 \, \phi_1 + \cdots + a_m \, \phi_m}{b_0 \, \phi_0 + b_1 \, \phi_1 + \cdots + b_n \, \phi_n} \tag{6.6}$$

may be obtained in a manner analogous to the standard Padé method as long as the ϕ_k satisfy the relation

$$\phi_i \, \phi_j = \sum_{k=0}^{i+j} A_{ijk} \, \phi_k \tag{6.7}$$

where the A_{ijk} are constant coefficients. Obviously this is the Padé method in general form. The classical Padé method is but a special case of this since then the ϕ_k are just the x^k. Equation (6.7) is satisfied since

$$x^i \, x^j = \sum_{k=0}^{i+j} A_{ijk} \, x^k = x^{i+j} \Rightarrow \begin{cases} A_{ijk} = 0, \; k \neq i+j \\ A_{ijk} = 1, \; k = i+j \end{cases} \tag{6.8}$$

An interesting development concerning the generalized Padé method was made by Maehly and is briefly presented by Snyder (18). Here the ϕ_k are Tchebycheff polynomials. The implementation of the algorithm is not a difficult task since the corresponding form of (6.7) is relatively simple:

$$T_i \, T_j = 1/2 \left\{ T_{i+j} + T_{|i-j|} \right\} \tag{6.9}$$

Generalizing the Tau-Padé equations is similar to the generalized Padé. However, in addition to the requirement that the ϕ_k satisfy (6.7), the Tchbycheff polynomials also must be expressed as a linear combination of the ϕ_k,

$$T_n(x) = \sum_{k=0}^{n} \xi_{nk} \, \phi_k \tag{6.10}$$

It is particularly interesting to consider the problem of obtaining the Tau-Padé approximant when f(x) is of the form (6.5) with $\phi_k = T_k$. In this case the procedure is nothing more than the application of Maehly's

Tchebycheff form of the generalized Padé method. The introduction of the τ-variables has no advantageous effect on determining the coefficients for the rational form. Thus the implication is that the Tau-Padé method, in effect, attempts to convert a power series into a "quasi-Tchebycheff" series (it will never actually make it since an infinity of terms would be required) while at the same time making use of the flexibility of rational forms.

As one might suspect, relations (6.7) and (6.10) are generally not readily available for ϕ_k's other than the simple powers of x or the Tchebycheff polynomials. Even when they are, the algebraic manipulations become overwhelming as shown by the form of (6.7) when the ϕ_k happen to be P_k, the Legendre polynomials (20):

$$P_m P_n = \sum_{r=0}^{m} \frac{A_{m-r} A_r A_{n-r}}{A_{n+m-r}} \left(\frac{2n + 2m - 4r + 1}{2n + 2m - 2r + 1} \right) P_{n+m-2r} , \tag{6.11}$$

$$A_m = \frac{1 \cdot 3 \cdot 5 \cdots (2m - 1)}{m!}$$

For these reasons, application of the generalized Padé or Tau-Padé method is usually not particularly attractive.

A natural question is to ask if the algorithm can be extended to handle functions of several variables. For example, consider a function of two variables represented by the following series:

$$f(x,y) = \sum_{k=0}^{\infty} c_k(y) \, \phi_k(x) \tag{6.12}$$

Since the c_k's are no longer constants, direct application of the Tau-Padé algorithm results in a rational approximation of the form

$$R_{mn\ell}(x,y) = \frac{\sum_{i=0}^{m} \sum_{j=0}^{m} a_i(y)\ \phi_j(x)}{\sum_{i=0}^{n} \sum_{j=0}^{n} b_i(y)\ \phi_j(x)} \tag{6.13}$$

This approach was attempted on the series associated with the gravitational potential of the Earth (zonal harmonics only), the corresponding form of (6.12) being

$$V(\phi,r) = \frac{\mu}{r} \sum_{k=0}^{\infty} J_k \left(\frac{R}{r}\right)^k P_k(\sin \phi) \tag{6.14}$$

where μ is the gravitational parameter of the Earth, r is the distance from the coordinate system origin, and ϕ is the geocentric latitude. Unfortunately, the amount of effort involved is extensive and no meaningful results have been obtained with this approach. Further investigation is necessary to determine its utility.

Probably one of the more notable generalizations of the Tau-Padé algorithm is the realization that a different error behavior may be obtained by merely using a set of weighting functions other than the Tchebycheff polynomials. Thus, using the Legendre polynomials, for example, would yield an approximation with different error characteristics.

CHAPTER 7

SUMMARY

This thesis is concerned with the problem of obtaining approximations to functions defined by a power series in a manner which allows facility in their handling and which yields near-optimality under the Tchebycheff norm. Rational functions are chosen as the approximating form because of their simplicity and flexibility, and are shown to be superior in many cases to the more common polynomial approximation. The results of this investigation offer a method which yields just such approximations.

The classical Padé method forms the basis of the technique developed in the investigation, utilizing the Taylor series representation of the function to be approximated. One of the earliest analytical schemes, it is also one of the simplest, merely requiring the solution of a system of linear equations.

The Tau method developed by Lanczos provides a tool for modifying the Padé method. It employs the important uniform error properties of the Tchebycheff polynomials to weight the coefficients in the Padé rational form to obtain near-optimal behavior of the associated error function. The combination of these two methods forms what, in this thesis, is referred to as the Tau-Padé method. It is illustrated by the approximation of numerous transcendental functions, and is shown to return the approximated function exactly when that function is of

rational form. Applications to the classical equation of Kepler offer additional insight into the method.

Some important conclusions are drawn concerning the use of the Tau-Padé method. The approach is found to be effective in obtaining a rational form which is near-optimal in its approximation of functions admitting a power series representation. Since polynomials are but special cases of rational forms, the method may be easily used to obtain near-optimal polynomial approximations. In generalizing the Tau-Padé method to functions defined by series other than power series, difficulties are usually encountered which hamper implementation of the method. However, when functions are represented by a Tchebycheff series, the generalized Tau-Padé equations reduce to a generalized Padé algorithm. From this the conclusion is drawn that the regular Tau-Padé method, in effect, attempts to "convert" a defining power series into a Tchebycheff series.

Two important restrictions are placed upon the Tau-Padé approach. First, the function to be approximated must, of course, be expressible in the form of a known power series. Second, because of this, the algorithm obviously may not be employed to obtain rational approximations from discrete data values. The specification of an interval over which the approximation is valid is an additional restriction but certainly not a severe one.

In conclusion, the Tau-Padé method offers the capability of approximating functions defined by a power series. The technique produces a near-optimal approximation in a non-iterative manner. It

provides an answer to the motivating question, "To what extent can such functions be accurately approximated?"

BIBLIOGRAPHY

1. Abramowitz, M., and I. A. Stegun (ed.). *Handbook of Mathematical Functions*, National Bureau of Standards Applied Mathematics Series No. 55. Washington: Government Printing Office, 1964.

2. Battin, R. H. *Astronautical Guidance*. New York: McGraw-Hill Book Company, 1964.

3. Boehm, B. W., "Functions Whose Best Rational Chebyshev Approximations Are Polynomials," *Numerische Mathematik* 6, 235-242 (1964).

4. Cheney, E. W. *Introduction to Approximation Theory*. New York: McGraw-Hill Book Company, 1966.

5. _____ and H. L. Loeb, "Two New Algorithms For Rational Approximation," *Numerische Mathematik* 3, 72-75 (1961).

6. _____, "On Rational Chebyshev Approximation," *Numerische Mathematik* 4, 124-127 (1962).

7. Eshbach, O. W. (ed.). *Handbook of Engineering Fundamentals*. New York: John Wiley and Sons, Inc., 1952.

8. Gottlieb, R. G., private communication, University of Texas, June 1970.

9. Irving, J., and N. Mullineux. *Mathematics in Physics and Engineering*. New York: Academic Press, Inc., 1959.

10. Kuo, Shan S. *Numerical Methods and Computers*. Reading: Addison-Wesley Publishing Company, 1965.

11. Lanczos, C. *Applied Analysis*. Englewood Cliffs: Prentice-Hall, Inc., 1956.

12. Maehly, H. J., "Methods for Fitting Rational Approximations, Part I: Telescoping Procedures for Continued Fractions," *Association for Computing Machinery Journal* 7, 150-162 (1960).

13. _____, "Methods for Fitting Rational Approximations, Parts II and III," *Association for Computing Machinery Journal* 10, 257-277 (1963).

14. Moulton, F. R. *Differential Equations*. New York: Dover Publications, Inc., 1958.

15. Roy, A. E. *The Foundations of Astrodynamics*. New York: The Macmillan Company, 1965.

16. Russell, P. E., private communication; University of Texas, April 1970.

17. Shanks, D., "Non-Linear Transformations of Divergent and Slowly Convergent Sequences," *Journal of Mathematical Physics* 34, 1-42 (1955).

18. Snyder, M. A. *Chebyshev Methods in Numerical Approximation*. Englewood Cliffs: Prentice-Hall, Inc., 1966.

19. Sokolnikoff, I. S., and R. M. Redheffer. *Mathematics of Physics and Modern Engineering*. New York: McGraw-Hill Book Company, 1958.

20. Whittaker, E. T., and G. N. Watson. *A First Course of Modern Analysis*. New York: Cambridge University Press, 1965.

21. Wynn, P., "The Rational Approximation of Functions Which Are Formally Defined by a Power Series," *Mathematics of Computation* 14, 147-186 (1960).

APPENDIX

EXPLICIT FORMULAS FOR THE TAU-PADÉ COEFFICIENTS

$$\left(f(x) = \sum_{k=0}^{\infty} c_k \, x^k \;, \quad R_{mn\ell}(x) = \frac{a_0 + a_1 x + \cdots + a_m \, x^m}{b_0 + b_1 x + \cdots + b_n \, x^n} \right)$$

$\underline{R_{mo\ell}}$:

$$a_k = c_k - \sum_{j=1}^{\ell} D_j \, c_{k+j} \;, \quad k = m$$

$$a_k = c_k - \sum_{j=1}^{\ell} \tau_{m+j} \, s_{m+j,k} \;, \quad k < m$$

$$D_j = \frac{1}{s_{m+j,m+j}} \left(s_{m+j,m} - \sum_{i=1}^{j-1} D_i \, s_{m+j,m+i} \right)$$

$j = 1, \cdots, \ell$; (no sum for $i > j-1$)

$$\tau_r = \frac{1}{s_{rr}} \left(c_r - \sum_{i=r+1}^{N} \tau_i \, s_{i,r} \right),$$

$r = m+1, \cdots, N$; (no sum for $i > N$)

$R_{mn\ell}$, $n > 0$:

$$a_k = c_k - \sum_{j=1}^{\ell} \tau_{m+n+j}\, s_{m+n+j,k}\,, \quad k = 0, \cdots, m$$

$$\tau_r = \frac{1}{s_{rr}} \left(c_r - \sum_{j=1}^{n} b_j\, c_{r-j} - \sum_{i=r+1}^{N} \tau_i\, s_{i,r} \right)$$

$r = m+n+1, \cdots, N$; (no sum for $i > N$)

$$D_{ij} = \frac{1}{s_{m+n+j,m+n+j}} \left(s_{m+n+j,i} - \sum_{k=1}^{j-1} D_{ik}\, s_{m+n+j,m+n+k} \right),$$

$j = 1, \cdots, \ell$; $i = m+1, \cdots, m+n$; (no sum for $k > j-1$)

$$W_{ik} = -c_{i-k} + \sum_{j=1}^{\ell} D_{ij}\, c_{m+n+j-k}\,,$$

$k = 0, \cdots, n$; $i = m+1, \cdots, m+n$

$$\underset{(n \times 1)}{\begin{bmatrix} b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{bmatrix}} = \underset{(n \times n)}{\begin{bmatrix} W_{m+1,1} & W_{m+1,2} & \cdots & W_{m+1,n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ W_{m+n,1} & W_{m+n,2} & \cdots & W_{m+n,n} \end{bmatrix}^{-1}} \underset{(n \times 1)}{\begin{bmatrix} W_{m+1,0} \\ \cdot \\ \cdot \\ \cdot \\ W_{m+n,0} \end{bmatrix}}$$