A SURVEY OF METHODS OF NUMERICAL APPROXIMATION TO FUNCTIONS

Submitted for the degree of M.Phil.
at the
University of Aston in Birmingham

RONALD ALFRED PICKERING

October, 1972

SUMMARY

This work undertakes a survey of methods of numerical approximation to functions. The functions considered are taken to be continuous within the range of approximation. Some consideration is given to the types of approximating functions in common use and the measurement of goodness of fit. It is seen that these two criteria together decide by what method the unknown coefficients are to be determined.

Some properties of orthogonal functions and continued fractions are presented. Methods of deriving interpolating functions are described. Approximations may often be based on series expansions and this is considered, with reference to Chebyshev series, asymptotic series and Padé approximants.

The next section deals with approximations derived when the measure of fit is chosen as one of the three Holder norms L_1, L_2 or L_∞. The L_1 problem is shown to be solved in some cases by treatment as an interpolation problem. The least-squares (L_2) problem is best treated using orthogonal polynomials. The minimax (L_∞) approximation is seen to be found only by means of an iterative process and is the best approach when finding rational function approximations.

The method of spline approximations is described. This is basically an interpolative approach, the practical method involves representing the function between the points of agreement, or knots, by cubic polynomials.

Finally a general summary covers the types of approximation considered. Some techniques, e.g. range reduction, are mentioned which help in certain cases with finding efficient approximations. An attempt is made to give a general strategy which can be adopted for finding a suitable approximation to a given function and which would be workable in all but exceptional cases.

CONTENTS

CHAPTER I

APPROXIMATION TO CONTINUOUS FUNCTIONS

Introduction

Numerical approximations to functions involves attempting to find a function f(x) which follows closely the behaviour of a given function y(x) in some region $a \leqslant x \leqslant b$. The approximation, once determined may be used instead of y(x), for example to evaluate $\int_a^b y(x)dx$ or to find $y(x_1)$, for some given value x_1 of the argument.

Why should it be found necessary to use f(x) instead of y(x) in such cases? Simply because f(x) can be chosen to be more "amenable" than y(x). It may be possible to choose f(x) so that it is relatively easy to integrate or differentiate or so that it is rapidly evaluated by an automatic machine. It is possible that y(x) itself is not explicitly known, when, for example, it is expressed as the solution of a differential equation.

In this work discussion is restricted to approximation to functions which are continuous in a single real variable.

Fundamental Considerations

There are three basic steps in finding an approximation to a given function. In the beginning, it is necessary to decide what form of function may be satisfactory as an approximation. Next, an expression must be chosen which can be used to give a measure of the "closeness" or goodness of fit of the approximation. Because of a geometric analogy, this is sometimes referred to as the distance between the two functions. Finally, having agreed upon the first two choices, it remains to derive and solve the system of equations which determines the coefficients of the approximation such that the distance function (or norm) is minimized.

There are no fixed rules governing what form and norm should be chosen. The choice will depend on a few general principles, mainly derived empirically and often constrained by what yields a practicable solution.

Traditionally, polynomials have been a first choice as approximating functions. In the past, this was strongly influenced by the lack of suitable computational facilities and with appropriate choice of norm, the coefficients of a polynomial could be found without a prohibitive amount of computation. Now, there is available virtually all the computing power we require, yet polynomial forms remain popular. One reason for this is the ease with which polynomials may be integrated or differentiated.

The choice of form and the distance function together determine the nature of the problem we are faced with when trying to evaluate the unknown coefficients. This problem is directly solved if it is a set of linear equations. This has meant that forms and norms have often been chosen so that the coefficients have been determined by linear equations. This, in the history of the subject, has led to the importance of least-squares approximations by polynomials. It is as well to point out that the linear problem can be ill-conditioned and that, in some cases, special care must be taken to avoid this. (See Chapter II)

Notwithstanding the advantages of polynomials, it is sometimes desirable to use other forms. These choices lead to non-linear problems to determine the unknowns. Such systems usually require more effort for their solution and we require some benefit to warrant the extra effort. This is usually in the form of a better degree of fit or perhaps the same degree of fit with fewer coefficients.

The second choice, that of the way of measuring the closeness of the approximation has a profound effect on the way in which the unknown parameters are determined. One point which is important when considering the goodness of fit is that generally, there are no a priori rules which determine the degree of the approximation to achieve a given order of closeness. Normally, we have to determine what is thought to be a suitable approximation and then decide if it is satisfactory. If this proves not to be the case, the process must be repeated with a higher degree approximation. Another alternative would be to divide up the range of fit and to determine separate approximations in each segment. In this case, some procedure is usually adopted to ensure that the approximations display some degree of smoothness at the joins of the segments.

Choice of Form of Approximation

There are limitations on the types of functions that are available for use in approximations. This may be due to the uses to which they are put or simply due to the practicability of deriving the unknown coefficients. Polynomials and trigonometric sums are often used because they display a "smoothness" of behaviour which is often closely matching that of the given functions. A smooth function can be thought of as one which displays an undulating rather than craggy nature. More precisely, smoothness implies that the function has continuous derivatives whose values remain relatively small. However, for this very reason, if the given function does not display characteristic polynomial behaviour, then a polynomial approximation may prove unsatisfactory. Such behaviour may take the form of a sharp "elbow" or perhaps a region of large slope whilst elsewhere the function may be relatively smooth. In such cases it may prove advantageous to use a different form of approximation, a natural choice being a rational function, that is, the ratio of two polynomials. In such a case, the determination of the coefficients is no longer a simple procedure. It may prove possible to avoid such a choice by carrying out a transformation of the independent variable. This can often be done fairly simply and the resultant function prove sufficiently smooth for a low-degree polynomial to give an adequate approximation.

Rational functions have another attractive feature. There is empirical evidence to suggest that for a great many problems, rational functions will give an approximation which has a smaller maximum error than polynomials with the same number of coefficients. For this reason, rational functions are often preferred as compact forms for use with automatic computers for function evaluation. In addition, rational functions may be converted to continued fractions which allows evaluation of the function economically in terms of the number of operations required.

Finally, the given function may possess special features that strongly suggest the use of special functions in the approximation. The most common of such functions are logarithmic or exponential terms. Special forms of this kind do not fit into general theory and have to be treated on their own merits.

The Measure of Goodness of Fit

When considering the distance of an approximation from a given function, it is natural to ask how close is it possible to get? In particular, if we choose polynomial form for the approximation, is it possible to increase the degree of the polynomial and thereby steadily reduce the error to any desired extent? Fortunately for the numerical analyst, Weierstrass established in a famous theorem that if $y(x)$ is a continuous function then it may be approximated to any degree of closeness by a polynomial. (A proof of this is given in Appendix A.1.) However, the theorem provides no hint as to how these polynomials may be derived.

Interpolation

Intuitively it may be felt that if the approximation has the same values as the original function at certain values of the argument, then it may be considered not to differ significantly at other points in the region spanned by the chosen arguments. In addition, the more points of agreement, the more reasonable the approximation should be. Historically, functions have often been defined in terms of tables. In such cases interpolation formulae have been used to evaluate the function at non-tabulated points using function values at equal intervals. Unfortunately it is not possible to ensure that approximations based on equal intervals have errors that reduce uniformly as the number of points increase. However, there is no reason to choose equally spaced points and it can be demonstrated that unequally-spaced points will be an advantage.

Other methods try to ensure a good fit by including information other than the function values. This usually consists of specifying the derivatives at certain points. Methods of this nature include the Hermite formula and cubic spline functions. If we include the function value and the value of its derivatives at only one single point then we have the Taylor Series.

The Lp Norms

When discussing interpolation methods, no specific mention was made of the measure of the goodness of fit. It will be seen in the relevant Chapters that the error may be estimated from the value of a certain high-order derivative, dependent on the degree of the approximation.

In the more general case, we require a "distance" function which is not dependent on the form of the approximation. This will then not only provide a measure of the goodness of fit, but will so characterize the problem as to lead the way to its solution. The measure that is adopted is the Lp, or Holder norm.

This is defined as

$$Lp\left[y(x) - f(x)\right] = \left[\int_a^b \left|y(x) - f(x)\right|^p \, dx\right]^{1/p} \quad p \geqslant 1 \qquad (1.1)$$

where y(x) and f(x) are the function and its approximation and $[a, b]$ is the range of fit.

The problem of approximation can now be defined as, having chosen the form of f(x), to determine its coefficients so that the expression on the right hand side of (1.1) is a minimum.

Only certain values of p are of practical importance.

(i) $p = 1$

$$L_1 = \int_a^b |y(x) - f(x)| \, dx \qquad (1.2)$$

Since $\left| \int_a^b [y(x) - f(x)] \, dx \right| \leqslant \int_a^b |y(x) - f(x)| \, dx$

it may seem reasonable to adopt the L_1 norm if $\int_a^b f(x)dx$ is to be used to represent $\int_a^b y(x)dx$.

(ii) $p = 2$

$$L_2 = \left[\int_a^b \{y(x) - f(x)\}^2 \, dx \right]^{1/2} \qquad (1.3)$$

This is the classical least-squares norm. In practice, the square root may be omitted without ambiguity.

(iii) $p = \infty$

It is possible to show that when $p \to \infty$, the Lp norm becomes

$$L_\infty = \max_{[a, b]} |y(x) - f(x)| \qquad (1.4)$$

(This is derived in Appendix A1.2). Because the object in each case is to determine the coefficients of $f(x)$ so that the distance function is minimised, the L_∞ norm is often referred to as the "minimax" norm.

The expressions (1.2), (1.3) and (1.4) impose different conditions on the approximation. The methods of deriving the coefficients are different in each case, and in general, we do not expect to find the coefficients of the approximations derived using the three norms to be the same.

Sometimes it is necessary or convenient to introduce a "weight-function" into the norm.

i.e. $$Lp = \left[\int_a^b w(x) |y(x) - f(x)|^p \, dx \right]^{1/p} \qquad (1.5)$$

where the weight-function $w(x)$ is a non-negative function of the argument in the range $[a, b]$. This function has the effect of giving more emphasis (or weight) to those errors in the regions where $w(x)$ is largest and vice-versa. It would appear simplest to take $w(x) = 1$, as this would give equal weight to all error values. However, there may be good reasons for other choices. If $w(x)$ is taken as $|y(x)^{-1}|$ then the norm will be based on the relative error rather than the absolute error. In other cases, $w(x)$ may take special forms so that special functions e.g. Chebyshev polynomials can be introduced into the approximation.

Chebyshev Sets

To complete this introduction, it is necessary to mention one concept which is important in the discussion of the existence of polynomial solutions in both L_1 and L_∞ approximations. This is the idea of Chebyshev sets.

Let us first consider the independence of a polynomial solution.

e.g. if $$f(a,x) = \sum_{i=0}^{n} a_i \phi_i(x) \qquad (1.6)$$

is an approximation where $\phi_i(x)$ is a (as yet undefined) polynomial of degree i, continuous in $[a,b]$ then we require that, except for isolated points in $[a,b]$,

$$f(a',x) \neq f(a'',x) \quad \text{unless } a' = a'' \qquad (1.7)$$

From (1.6) and (1.7)

$$\sum_{i=0}^{n} (a' - a'')_i \phi_i(x) \neq 0$$

or rearranging

$$\sum_{\substack{i=0 \\ i \neq j}}^{n} \frac{(a' - a'')_i}{(a' - a'')_j} \phi_i(x) \neq \phi_j(x) \qquad 0 \leqslant j \leqslant n \qquad (1.8)$$

This implies that the $\phi(x)$ we employ in the approximation $f(x)$ must be linearly independent.

However it is found that this is not sufficient of itself to ensure that the approximation exists and may be evaluated. For this to be true in the cases mentioned, it is necessary to demand a further property, which is the defining property of a Chebyshev set.

This can be stated in one of three equivalent ways. Let $\phi_i(x)$, $i = 0,1 \ldots n$ be polynomials forming the basis for $f(x)$, then

(i) No linear combination $\sum_{i=0}^{n} a\phi_i(x)$ of the $(n+1)$ functions $\phi_i(x)$ has more than n roots in $[a,b]$ unless it vanishes identically.

(ii) the determinant $\det \left\{ \phi_i(x_j) \right\}$ $i = 0,1, \ldots n$ cannot vanish if the x_j are $(n + 1)$ distinct points on $[a, b]$.

(iii) a unique linear expression of the form (1.6) can be found to interpolate any continuous function at $(n + 1)$ distinct points in $[a, b]$

These are three different and equivalent ways of expressing Haar's condition. Any set of functions satisfying Haar's condition is said to form a Chebyshev set.

We notice immediately that the first $(n + 1)$ powers of $x, [1, x, x^2 \ldots . x^n]$ form a Chebyshev set in any interval. Also, and not obviously, so do the first $(2n + 1)$ trigonometric functions $[1, \cos x, \sin x, \ldots . \cos 2nx, \sin 2nx]$ in the interval $[0, 2\pi]$.

It is fairly easy to derive a set of functions not forming a Chebyshev set, yet all functions used in polynomial approximations do form such sets.

No attempt has been made to elaborate many of the statements and concepts introduced in this first Chapter. The main Chapters of the work are devoted to such elaboration. Chapter II introduces orthogonal functions, which play an important role in methods of approximation, whilst Chapter III discusses the main features of continued fractions. The next two chapters are devoted to approximations derived by interpolation and from series expansions respectively. Chapters VI, VII and VIII are given to methods based on L_1, L_2 and L_∞ norms in turn and Chapter IX looks at cubic spline approximations. The final Chapter contains a general discussion and comparison of methods with some illustrative examples.

CHAPTER II

Orthogonal Polynomials

Introduction

In this chapter, we attempt to show how orthogonal polynomials arise naturally when considering problems of approximation. Reference is made to trigonometric functions as they occur in Fourier series. From the trigonometric functions are developed the Chebyshev polynomials. These play a major part in any discussion of methods of approximation. Some of their more notable properties are described.

Discrete Least-Squares Approximation

The 'least-squares' method has a well known application in curve fitting over a discrete set of given points

$(x_k y_k) \quad k = 0,1 \; ..... \; m$.

Let $f_n(x) = c_0 + c_1 x + c_2 x^2 + \; ...... \; + c_n x^n$ be the polynomial of approximation, then the coefficients c_j are chosen so that

$$S = \sum_{k=0}^{m} \left[w(x_k)\left\{y_k - f_n(x_k)\right\}^2 \right] \tag{2.1}$$

is a minimum ($w(x)$ is a positive weight function).

Now S can be made arbitrarily large by a suitably 'bad' choice of coefficients, so we expect that if an extreme value of S does exist, then it will be a minimum value.

The necessary condition for a minimum value of S is

$$\frac{\partial S}{\partial c_k} = 0 \qquad k = 0,1 \; ...... \; n$$

Now if $m > n + 1$, S will be non-zero, and the above condition leads to a set of linear equations defining the coefficients. These are termed the normal equations and are

$$\begin{aligned} s_0 c_0 + s_1 c_1 + \; .... \; + s_n c_n &= b_0 \\ s_1 c_0 + s_2 c_1 + \; .... \; + s_{n+1} c_n &= b_1 \\ \cdots\cdots\cdots\cdots\cdots\cdots & \;\; \cdots \\ s_n c_0 + s_{n+1} c_1 + \; .... \; + s_{2n} c_n &= b_n \end{aligned} \tag{2.2}$$

where $s_r = \sum_{k=0}^{m} x_k^r$ and $b_r = \sum_{k=0}^{m} x_k^r y_k$

The straightforward nature of this approach looks attractive, but two problems arise in practice.

Firstly, the matrix of coefficients in (2.2) can become ill-conditioned for even moderately large values of n. For example, if in (2.1) $w(x) = 1$ and x_k are equally spaced in $[0,1]$, then s_r in (2.2) becomes

$$s_r = \sum_{k=0}^{m} x_k^r \simeq m\int_0^1 x^r dx = \frac{m}{r+1} \quad \text{if m is large.}$$

So, removing the factor m, the matrix becomes

$$\begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots\cdots \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{6} & \\ \cdots\cdots & & & & \\ \cdots\cdots & & & & \end{bmatrix}$$

which is a well-known ill-conditioned matrix.

Secondly, if an approximation has been found of degree N say, and S proves insufficiently small, then to extend the approximation to degree N + 1 involves the solution of a completely new set of equations and all the previously computed coefficients will be changed.

These difficulties are overcome if we can express the approximating function in the form

$$f_n(x) = c_0\phi_0(x) + c_1\phi_1(x) + \cdots\cdots + c_n\phi_n(x) \qquad (2.3)$$

where the ϕ's are independent functions in the space defined by the points (x_k) and having the property

$$\sum_{k=0}^{m} w(x_k)\phi_i(x_k)\phi_j(x_k) = 0 \qquad \text{if } i \neq j \qquad (2.4)$$

$$\sum_{k=0}^{m} w(x_k)\phi_j^2(x_k) \neq 0$$

Functions having this property are termed orthogonal. In such cases the matrix of coefficients in (2.2) becomes diagonal and the coefficients in (2.3) are defined by:

$$C_j = \frac{\sum_{k=0}^{m} w(x_k)\phi_j(x_k)y_k}{\sum_{k=0}^{m} w(x_k)\phi_j^2(x_k)} \qquad j = 0,1, \ldots\ldots. n \qquad (2.5)$$

No longer is it necessary to invert an ill-conditioned matrix in order to evaluate the coefficients. Also, it can be seen from (2.5) that adding new terms to the approximation will not change the coefficients already evaluated.

The expression (2.1) for the error becomes

$$S = \sum_{k=0}^{m} w(x_k)\left[y_k^2 - \sum_{r=0}^{n} c_r^2\phi_r^2(x_k)\right] , \qquad (2.6)$$

and S may be found for a higher degree approximation by including an extra term $c_{n+1}\phi_{n+1}(x)$

The Continuous Case

The least-squares method can be applied to the problem of finding an easily computable function of the form (2.3) which approximates over a given finite interval to a continuous function y(x).

If the interval is taken as $[-1,1]$ we can write

$$S = \int_{-1}^{1} w(x)\left[y(x) - f_n(x)\right]^2 dx \qquad (2.7)$$

and the coefficients of $f_n(x)$ are chosen to minimise S. The condition for least S again produces a set of normal equations in which

$$s_{i+j} = \int_{-1}^{1} w(x)\phi_i(x)\phi_j(x)dx \qquad (2.8)$$

$$b_i \div \int_{-1}^{1} w(x)\phi_i(x)y(x)dx$$

and the orthogonality condition is

$$\int_{-1}^{1} w(x)\phi_i(x)\phi_j(x)dx = 0 \qquad \text{if } i \neq j$$

Hence, the off-diagonal elements of (2.8) are zero and the solution of the normal equations yields

$$c_r = \frac{\int_{-1}^{1} w(x)\phi_i(x)y(x)dx}{\int_{-1}^{1} w(x)\phi_i^2(x)dx} \qquad (2.9)$$

The weight function $w(x)$ is chosen to be non-negative in $[-1,1]$. One important property of a set of polynomials which form an orthogonal system is that any three consecutive polynomials are related by a recurrence of the form

$$\phi_{k+1}(x) = (A_k x + B_k)\phi_k(x) - C_k\phi_{k-1}(x)$$

(See J. R. Rice [13])

This will be used to develop a useful computational method for evaluating series whose terms are orthogonal polynomials.

Fourier Series

The best-known orthogonal system is the set of trigonometric functions cos x, cos 2x, ... , sin x, sin 2x, ... over the interval $[-\pi, \pi]$.

It is easily shown that

$$\int_{-\pi}^{\pi} \sin nx \sin mx \, dx = \int_{-\pi}^{\pi} \cos nx \cos mx = 0 \qquad \text{for } n \neq m$$

and

$$\int_{-\pi}^{\pi} \sin nx \cos mx \, dx = 0$$

So, if we represent a function in terms of an infinite series of trigonometric terms (a Fourier series) of the form

$$y(x) = \frac{a_o}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx) \qquad (2.10)$$

then

$$a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} y(t) \cos kt \, dt \qquad (2.10a)$$

$$b_k = \frac{1}{\pi} \int_{-\pi}^{\pi} y(t) \sin kt \, dt$$

Now if (2.10) be truncated at some point, what sort of approximation is obtained?

let

$$f_n(x) = \frac{a_o}{2} + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx) \qquad (2.11)$$

then, taking $w(x) = 1$, (2.7) becomes

$$S = \int_{-\pi}^{\pi} y^2(x)dx - 2\int_{-\pi}^{\pi}\left\{\frac{a_0}{2} + \sum_{k=1}^{n}(a_k\cos kx + b_k\sin kx)\right\} y(x)dx$$
$$+\pi\left\{\frac{a_0^2}{4} + a_1^2 + \ldots\ldots + a_n^2 + b_1^2 + \ldots\ldots\ b_n^2\right\}$$

The condition that S should be a minimum is $\frac{\partial s}{\partial a_j} = \frac{\partial s}{\partial b_j} = 0$

i.e. $-\int_{-\pi}^{\pi}\cos kx y(x)dx + \pi a_k = 0$

or $$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} y(x)\cos kx\ dx$$

similarly $$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} y(x)\sin kx\ dx$$

We notice that these are precisely the coefficients defined in (2.10a). That is to say, the truncated Fourier series is the best approximation in the least-squares sense for the interval $[-\pi, \pi]$

The error term for the truncated series is

$$S = \int_{-\pi}^{\pi} y^2(x)dx - (\frac{a_0^2}{4} + a_1^2 + \ldots\ldots a_n^2 + b_1^2 + \ldots\ldots +b_n^2)$$

It is natural to ask if the series defined in (2.10) is convergent if y(x) is continuous, does $f_n(x)$ defined in (2.11) approach y(x) as n increases? This problem is dealt with extensively in available literature (e.g. Lanczos 11) We note that the continuity of y(x) does not prove sufficient for convergence of the series. Sufficiency is expressed in the "Dirichlet conditions" which establish the desired smoothness of y(x). These conditions are not always necessary; Fejer's method of summation [13] can be used to compute a sequence which converges to y(x) when the only restriction on y(x) is that it is absolutely integrable.

One practical difficulty encountered in approximation by Fourier series arises from the fact that since the series is periodic, the function and its derivatives are expected to have the same values at the two end-points. At best, if the original function is not periodic, we might hope for the continuity of the function and its first derivative only at these end-points. In these circumstances, convergence of the series may prove slow. This problem may be avoided by using the following transformation.

Chebyshev Polynomials

Consider y(x) when we write $x = \cos\theta$

$$\text{then } y(x) = y(\cos\theta) = Y(\theta)$$

and the interval $[-1,1]$ in x becomes $[0,\pi]$ in θ. $Y(\theta)$ being a function of $\cos\theta$ will be an even function of θ and is periodic in θ.

Because $Y(\theta)$ is even, we can expand it as a series of cosine terms and the integral in (2.10) can be carried out over the positive half-range only.

$$\text{i.e.} \quad Y(\theta) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos k\theta$$

$$\text{where} \quad a_k = \frac{2}{\pi}\int_0^{\pi} Y(\theta)\cos k\theta \, d\theta$$

Now rewrite in terms of the original variable x

$$y(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k T_k(x)$$

$$a_k = \frac{2}{\pi}\int_{-1}^{1} \frac{T_k(x)\, y(x)\, dx}{\sqrt{1-x^2}} \qquad (2.12)$$

$$\text{where} \quad T_k(x) = \cos k\theta = \cos(k \cos^{-1} x)$$

We notice that from their relation to the cosine functions that the $T_k(x)$ are polynomials of degree k in x. They are orthogonal over the interval $[-1,1]$. These functions are called Chebyshev polynomials.

Chebyshev polynomials have many properties which have applications in function approximation. Some of these are derived in Appendix A 2.1. We list here the important features:

(i) They are orthogonal over the range $[-1,1]$ with respect to a weight function $\frac{1}{\sqrt{1-x^2}}$.

(ii) In the range $[-1,1]$, they have maximum and minimum values of ± 1 and $T_n(x)$ has extremes at exactly n+1 points.

(iii) They can be differentiated and integrated easily.

(iv) A truncated Chebyshev series has an error nearly proportional to the first neglected term (say $T_{n+1}(x)$). By virtue of the equal ripple property of $T_{n+1}(x)$ the error can be seen to be evenly distributed throughout the interval.

(v) A series of Chebyshev terms displays more rapid convergence than corresponding ~~power~~ Taylor series.

(vi) The Chebyshev polynomials prove to be orthogonal over certain discrete point sets with constant weights. This can be a useful feature when the integral in (2.12) does not prove analytically possible.

Chebyshev Polynomials of the Second Kind

In Appendix A2.1 it is shown that a recurrence relation for Chebyshev polynomials can be developed from the identity

$$\cos\,(n+1)\theta + \cos(n-1)\theta = 2\,\cos n\theta\,\cos\theta$$

i.e. $$T_{n+1}(x) + T_{n-1}(x) = 2xT_n(x)$$

but we also know that

$$\sin\,(n+1)\theta + \sin(n-1)\theta = 2\,\sin n\theta\,\cos\theta \qquad (2.13)$$

dividing by $\sin\theta$

$$\frac{\sin\,(n+1)\theta}{\sin\theta} + \frac{\sin\,(n-1)\theta}{\sin\theta} = 2\,\frac{\sin n\theta}{\sin\theta}\,\cos\theta$$

Consider $$U_n(\theta) = \frac{\sin(n+1)\theta}{\sin\theta}$$

then $$U_0(\theta) = 1$$

$$U_1(\theta) = 2\cos\theta$$

and (2.13) gives

$$U_n(\theta) = 2\cos\theta\,U_{n-1}(\theta) - U_{n-2}(\theta)$$

if we put $x = \cos\theta$,

$$U_0(x) = 1$$

$$U_1(x) = 2x$$

$$U_n(x) = 2x\,U_{n-1}(x) - U_{n-2}(x)$$

and clearly $U_n(x)$ is a polynomial of degree n in x.

These functions are called Chebyshev polynomials of the second kind. They have many analogous properties to those of the $T_n(x)$, although they do not possess the equal-ripple property of the latter. Some of these properties are derived in Appendix A2.2.

However, one useful feature of $U_n(x)$ is that its integral can be expressed explicitly in terms of the ordinary Chebyshev polynomials.

In later chapters it will be seen that the properties of Chebyshev polynomials can be used in many situations to obtain satisfactory approximating functions.

CHAPTER III

Continued Fractions

Introduction

The continued fraction form is represented by

$$f(x) = b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cdots\cdots\cdots\cdots + \cfrac{a_n}{b_n}}} \qquad \text{where} \quad \begin{array}{l} a = a(x) \\ b = b(x) \end{array}$$

or, more conveniently

$$f(x) = b_0 + \frac{a_1}{b_1 +}\;\frac{a_2}{b_2 +}\;\cdots\cdots\;\frac{a_n}{b_n} \qquad (3.1)$$

It can be noted that this expresses $f(x)$ as a rational function. The expression (3.1) is often a convenient computational form which, in many cases, can be used to obtain approximations to functions of higher accuracy than a polynomial expansion having the same number of coefficients. Howäver, problems in evaluation may arise. If at some stage a divisor is small, then the rounding error introduced could become unacceptably large.

If (3.1) is obtained by truncation of an infinite continued fraction, will the value of $f(x)$ converge as n tends to infinity. ~~more terms are added to the fraction~~? The convergence of fractions and their evaluation are discussed by Blanch [2.] Some of the results obtained are quoted here. In addition, a different method of computation is described and it is shown that the same criteria for convergence and truncation error estimates can be obtained for this method as for other methods. (Appendix A3.3)

Methods of Evaluation

Four methods of evaluation are described and a comparison made of their relative merits.

(i) The most obvious method is to use backward recurrence in (3.1)

i.e. generate $c_k = \dfrac{a_k}{b_k + c_{k+1}} < 0$ with $c_{n+1} = 0$ $\quad k = n, (n-1), \ldots\ldots 1$

then $f(x) = b_0 + c_1$

(ii) The second method uses forward recurrence.

let $f(x) = \frac{A_n}{B_n}$, then A_n and B_n can be found from the relation

$$Y_{j+1} = b_{j+1} Y_j + a_{j+1} Y_{j-1} \qquad j = 0,1,\ldots\ldots(n-1) \qquad (3.2)$$

where Y_j is either A_j or B_j given $A_{-1} = B_0 = 1$

$B_{-1} = 0, \; A_0 = b_0$

(This relation is established in Appendix A3.1)

(iii) A method which differs slightly from (ii) calculates a correction term which when added to each convergent $\frac{A_{j-1}}{B_{j-1}}$ gives $\frac{A_j}{B_j}$

It is shown in Appendix A3.1 that if

$$D_k = A_k B_{k-1} - B_k A_{k-1}$$

then
$$D_k = - a_k D_{k-1} \qquad (3.3)$$

and
$$\frac{A_n}{B_n} = \frac{A_{n-1}}{B_{n-1}} + \frac{D_n}{B_{n-1} B_n} \qquad (3.4)$$

At each stage D_n and B_n can be computed to obtain $\frac{A_n}{B_n}$

(iv) The last method differs in its approach in that successive convergents of the fraction are expressed as the partial sums of a series.

From (3.1) is evaluated

$$\rho_1 = \frac{a_1}{b_1}, \qquad 1 + \rho_2 = \frac{b_1 b_2}{b_1 b_2 + a_2}$$

$$1 + \rho_j = \frac{b_{j-1} b_j}{b_{j-1} b_j + a_j(1+\rho_{j-1})} \qquad j \geqslant 3$$

then $f(x) = b_0 + \sum_{i=1}^{n} U_i$

where $u_i = \rho_1 \rho_2 \cdots\cdots\cdots \rho_i$ (See Appendix A3.1)

Comparison of Methods of Evaluation

It remains to be decided which form is the most convenient for computation. All involve some division and could possibly suffer loss of significant figures if a divisor is small.Method (i) is best if the number of terms in the fraction is known beforehand, for it involves the least computational effort. However, if more terms need to be added to the fraction, the whole computation must be done again. The other forms, which all involve forward recurrence, can compute the next convergent with relative ease. Form (ii) does not directly compute the difference between successive convergents and since in some cases this can be used in the estimate of truncation error, it might be an advantage to use forms (iii) or (iv).

Table 3.1 compares the number of multiplications and divisions necessary to compute the next convergent by forward recurrence.

Method	Multiplications	Divisions
(ii)	4	1
(iii)	4	1
(iv)	3	1

Table 3.1

This shows that method (iv) requires the least effort of the forward schemes.

With regard to round-off errors, Blanch [2] found smaller error bounds for the backward scheme than for the forward schemes (ii) and (iii). General rules are difficult to formulate, for example the fraction

$$\frac{1}{2} + \frac{x-1}{10+}\;\frac{x-2}{-\frac{1}{2}+}\;\frac{x-3}{0+}\;\frac{x-4}{1}$$

could not be evaluated by machine using (i) or (iv) without suitable modification when $x = 4$ and for $x \simeq 4$ could produce intolerable rounding error. The best conclusion to be drawn is that whereas most continued fractions are well-conditioned, the only way to ascertain the condition of a particular example would be to test it in detail.

Convergence of Continued Fractions

In [2] the convergence of continued fractions is discussed making use of the relations described in methods (ii) and (iii). The results are summarized here together with similar results for the form (iv) which are obtained in Appendix A3.3.

However, in order that the number of fractions under consideration can be limited without loss of generality, use is made of the equivalence transformation proved in Appendix A3.2. This enables consideration to be conveniently restricted to the forms

$$b_o + \frac{1}{c_1 +}\ \frac{1}{c_2 +}\ \ldots\ldots\ldots \qquad \text{provided } a_k \neq 0$$

and $$b_o + \frac{d_1}{1\ +}\ \frac{d_2}{1\ +}\ \ldots\ldots\ldots \qquad \text{provided } b_k \neq 0$$

Summarizing the results obtained in [2]and in the Appendix A3.3;

A. If $F = \frac{a_1}{b_1 +}\ \frac{a_2}{b_2 +}\ \ldots\ldots\ldots\ldots$ where $a_k, b_k > 0$ (3.5)

then the even convergents approach a limit L_o, the odd convergents a limit L_1 such that $L_o \leqslant L_1$

In the summation form $\rho_j < 0$, for $j > 1$, and the terms of the series will alternate in sign.

B. If $F_1 = \frac{1}{b_1 +}\ \frac{1}{b_2 +}\ \ldots\ldots\ldots$ where $b_k > 0$ (3.6)

F_1 converges if and only if $\sum\limits_{k=1}^{\infty} b_k$ diverges

C. If $F_2 = \frac{1}{b_1 -}\ \frac{1}{b_2 -}\ \ldots\ldots\ldots$ where $b_k > 0$ (3.7)

a sufficient condition for the convergence of F_2 is $b_k \geqslant 2$ for all $k > N$. In this case $\rho_j > 0$ and the series used in the summation form has the same sign throughout.

D. If $F_3 = \frac{a_1}{1\ -}\ \frac{a_2}{1\ -}\ \ldots\ldots\ldots$ where $a_k > 0$ (3.8)

a sufficient condition for the convergence of F_3 is that $a_k \leqslant \frac{1}{4}$ for all $k > N$.

E. For fractions in which all the elements are positive, the value of the fraction will lie between the values of successive convergents, (ignoring rounding error).

F. For any fraction, let the truncation error be R_n

$$\text{i.e.} \quad R_n = F - \frac{A_n}{B_n}$$

$$\text{and let} \quad E_n = \frac{A_n}{B_n} - \frac{A_{n-1}}{B_{n-1}}$$

then for F_2 if $b_k \geqslant 2 + c$ where $\dot{c} > 0$

$$0 < |R_n| \leqslant \frac{d}{1-d} |E_n| \qquad \text{where } d = 1 + \tfrac{1}{2}c - \sqrt{(1+\tfrac{1}{2}c)^2 - 1} \qquad (3.9)$$

and for F_3, if $a_k \leqslant \frac{1}{4} - c$ where $c > 0$

$$0 < |R_n| \leqslant \frac{1}{(\frac{1}{2\sqrt{c}} - 1)} |E_n| \qquad (3.10)$$

In addition if ρ_j form a decreasing sequence for $j \geqslant n+1$,

$$\text{then} \quad |R_n| \leqslant \frac{|u_{n+1}|}{1 - |\rho_{n+1}|} ,$$

Examples

(1) $$F = \frac{1}{1+} \; \frac{1}{3+} \; \frac{1}{5+} \; \frac{1}{7+}$$

all the elements are positive and

$$\rho_1 = 1, \; \rho_2 = -\frac{1}{4}, \quad \rho_3 = -\frac{1}{21}, \quad \rho_4 = -\frac{4}{151}, \quad \rho_5 = -\frac{7}{460}$$

and all the values of ρ are negative apart from ρ_1. The convergents are the partial sums of the alternating series

$$1 - \frac{1}{4} + \frac{1}{84} - \frac{1}{3171} + \dots\dots\dots$$

and form the sequence

1.0000 0.76190476 0.76159420

0.7500 0.76158940 0.76159415

The odd convergents are increasing, the even convergents are decreasing and

$$0.76159415 < F < 0.76159420$$

(2) $$F = \frac{1}{1-} \; \frac{1}{3-} \; \frac{1}{5-} \; \frac{1}{7-}$$

$\rho_1 = 1, \; \rho_2 = \frac{1}{2}, \rho_3 = \frac{1}{9}, \; \rho_4 = \frac{2}{61}, \; \rho_5 = \frac{1}{60}$ which are all positive.

The series generated is

$$1 + \frac{1}{2} + \frac{1}{18} + \frac{1}{549} \; \dots\dots\dots$$

The convergents form the sequence

1.00 00, 1.5000, 1.555 56, 1.557 38 ....

but $u_6 = \frac{1}{32940}$ and applying (3.11)

$$|R_n| \leqslant \frac{1}{32940(1-\frac{1}{60})} \simeq 0.31 \times 10^{-4}$$

(The error is $\simeq 0.28 \times 10^{-4}$ at this point)

This estimate requires extra work in the evaluation of u_6 and ρ_5 .

Reconstructions of Fractions for More Rapid Convergence

Handscomb [7] quotes transformations which allow only the odd or even convergents to be computed when the original fraction is of the form

$$b_0 + \frac{c_1}{1+}\ \frac{c_2}{1+}\ \frac{c_3}{1+}\ \ldots\ldots, \tag{3.12}$$

Since this would halve the computational effort, it looks attractive. Although any fraction could first be transformed into the form (3.12) and then contracted, it would be helpful to apply the process to the more general form

$$b_0 + \frac{a_1}{b_1+}\ \frac{a_2}{b_2+}\ \ldots\ldots. \tag{3.13}$$

In appendix A3.5, the following forms are established for the odd and even parts of (3.13)

$$F_1 = b_0 + \frac{a_1 b_2}{(b_1 b_2 + a_2)} - \frac{\frac{b_4 a_3 a_2}{b_2}}{\left(a_4 + b_4 b_3 + \frac{b_4 a_3}{b_2}\right) -}\ \ldots. \tag{3.14}$$

$$F_2 = \frac{b_0 b_1 + a_1}{b_1} - \frac{\frac{b_3 a_2 a_1}{b_1}}{(a_3 b_1 + b_3 (b_1 b_2 + a_2) -}\ \frac{\frac{b_5 a_4 a_3}{b_3}}{\left(a_5 + b_5 b_4 + \frac{b_5 a_4}{b_3}\right) -} \tag{3.15}$$

Clearly, extra effort is required in evaluation of the partial numerators and partial denominators in either (3.14) and (3.15) compared with (3.13). Depending on the nature of the coefficients, this may outweigh any reduction in the amount of calculation required in computing the convergents.

Consider an example, the known expansion

$$\tan x = \frac{x}{1-}\quad \frac{x^2}{3-}\quad \frac{x^2}{5-}\quad \cdots\cdots\cdots \frac{x^2}{(2r-1)}\quad \cdots\cdots \qquad (3.16)$$

This can be written

$$\tan x = \frac{x}{1-}\quad \frac{\frac{x^2}{1.3}}{1-}\,\frac{\frac{x^2}{1.5}}{1-}\quad \cdots\cdots\cdots \quad \frac{\frac{x^2}{(2r-3)(2r-1)}}{1-}\quad \cdots\cdots$$

From (3.8), this is convergent if

$$\frac{x^2}{(2r-3)(2r-1)} \leqslant \frac{1}{4}$$

or

$$x^2 \leqslant (r-1)^2 - \frac{1}{4}$$

This will always be true for large enough r, for example, if $x = 3$, the coefficients satisfy the condition for convergence if r is greater than 5.

Using the transformation (3.15) or (3.16) we have

$$b_0 = 0 \qquad a_1 = x \qquad a_j = -x^2 \quad j > 1$$

$$b_j = 2j-1 \quad j \geqslant 1$$

hence,

$$a_1 b_2 = 3x \qquad\qquad a_2 b_1 b_2 = 3-x^2$$

$$\frac{b_4 a_3 a_2}{b_2} = \frac{7}{3}x^2 \qquad a_4 + b_4 b_3 + \frac{b_4 a_3}{b_2} = 35 - \frac{10x^2}{3} \qquad \text{etc.}$$

and

$$\tan x = \frac{3x}{(3-x^2)-}\quad \frac{\frac{7}{3}x^4}{(35-\frac{10}{3}x^2)-}\quad \frac{\frac{11}{7}x^4}{(99-\frac{18}{7}x^2)-}\quad \cdots\cdots$$

or, more easily for computational purposes

$$\tan x = \frac{\frac{3}{x}}{(\frac{3}{x^2}-1)-}\quad \frac{7}{(\frac{105}{x^2}-10)-}\quad \frac{33}{(\frac{693}{x^2}-18)-}\quad \cdots\cdots\cdots \qquad (3.17)$$

The convergents of (3.17) are the even convergents of (3.16).

If we note that the r^th^ term of (3.17) is

$$\frac{(4r-1)(4r-9)}{\left(\frac{(4r-5)(4r-3)(4r-1)}{x^2} - (8r-6)\right)}$$

then some of the reduction in computational effort in using (3.17) is lost due to the extra effort involved in evaluating the partial numerator and denominators. It may be noted that some of the terms in (3.17) could be evaluated using integer arithmetic, which could be a help in reducing rounding error.

A comparison is given in Table 3.2 of some results obtained from (3.16) and (3.17)

Value of x	Convergent No.	Original Form	Convergent No.	Contracted Form
1.00	4	1.557 3770	2	1.557 3770
	6	1.557 4077	3	1.557 4077
2.00	8	-2.185 0643	4	-2.185 0643
	10	-2.185 0399	5	-2.185 0399
3.00	8	-0.1425 4763	4	-0.1425 4763
	10	-0.1425 4654	5	-0.1425 4654

Table 3.2

Modification of the Summation Process When a Partial Denominator is Small

It has been shown that a continued fraction can be evaluated by summing the series

$$F = b_o + \sum_{j=1}^{n} u_j \qquad \text{where} \quad u_j = \rho_1 \rho_2 \cdots\cdots \rho_j$$

$$\text{and } 1+\rho_j = \frac{b_{j-1} b_j}{b_{j-1} b_j + a_j(1+\rho_{j-1})} \qquad j \geqslant 3$$

Now if b_j is very small, then $\rho_j \simeq -1$

$$u_n \simeq u_{n-1}(-1) = -u_{n-1}$$

also $1 + \rho_{j+1} = \dfrac{b_j b_{j+1}}{b_j b_{j+1} + a_{j+1}(1+\rho_j)}$ and considerable rounding error could occur in the calculation of ρ_{j+1}.

One way to avoid this possibility, assuming that it is due to only one isolated value of b_j is to avoid the computation of F_n and F_{n+1} and to jump from F_{n-1} to F_{n+2}

In appendix A3.6 the following expressions are developed

$$F_{n+2} = F_{n-1} - \frac{u_{n-1} a_n (b_{n+1} b_{n+2} + a_{n+2})(1 + \rho_{n-1})}{(b_{n+1} b_{n+2} + a_{n+2})(b_{n-1} b_n + a_n(1+\rho_{n-1})) + a_{n+1} b_{n-1} b_{n+2}}$$

$$U_{n+2} = \frac{-u_{n-1}\, a_n a_{n+1} a_{n+2}\, b_{n-1}\,(1+\rho_{n-1})}{\left[b_{n+1}(b_{n-1}b_n + a_n(1+\rho_{n-1})) + a_{n+1}b_{n-1}\right]\left[(b_{n+1}b_{n+2} + a_{n+2})(b_{n-1}b_n + a_n(1+\rho_{n-1})) + a_{n+1}b_{n-1}b_{n+2}\right]}$$

$$1+\rho_{n+2} = \frac{b_{n+1}\, b_{n+2}}{b_{n+1}b_{n+2} + a_{n+2}\left\{\dfrac{-a_{n+1}\, b_{n-1}}{b_{n+1}(b_{n-1}b_n + a_n(1+\rho_{n-1})) + a_{n+1}b_{n-1}}\right\}}$$

These expressions are valid for $n \geqslant 3$, with similar expressions for $n = 1$ and $n = 2$. The programme documented in Appendix A3 uses the summation form for the evaluation of a continued fraction and incorporates the above modifications. It was used in the following example.

Example

The fraction for tan x can be expressed in two ways to obtain the even and odd convergents respectively:

$$F_1 = \frac{\frac{3}{x}}{(\frac{3}{x^2}-1)-}\;\frac{\frac{7}{3}}{(\frac{35}{x^2}-\frac{10}{3})-}\;\frac{\frac{11}{7}}{(\frac{99}{x^2}-\frac{18}{7})-}\;\cdots\cdots$$

$$F_2 = \frac{\frac{1}{x}}{(\frac{1}{x^2}-\frac{1}{3})-}\;\frac{\frac{1}{1.3.3.5}}{(\frac{1}{x^2}-\frac{2}{3.7})-}\;\frac{\frac{1}{5.7.7.9}}{(\frac{1}{x^2}-\frac{2}{7.11})-}\;\cdots\cdots\cdots$$

It can be seen that $b_1 \simeq 0$ when $x \simeq \sqrt{3}$ and $b_2 \simeq 0$ when $x \simeq \sqrt{10.5}$. The following results were obtained using single precision floating point arithmetic and without the modification outlined above

x = 1.7320508

Convergent No.	F_1	F_2
5	-6.1480 5471 9	-6.1499 9224 2
6	-6.1480 5471 9	-6.1499 9224 2

x = 3.2403704

Convergent No.	F_1	F_2
6	0.0991 0224 103	0.0990 9549 136
7	0.0991 0224 103	0.0990 9549 136

The summation appears to converge, but the difference in the results indicate that the values obtained are unsatisfactory. With the modified programme the results become:

x = 1.7302508

Convergent No.	F_1	F_2
5	-6.1475 3345 4	-6.1475 3345 5
6	-6.1475 3345 4	-6.1475 3345 5
x = 3.2403704		
6	0.0991 0026 477	0.0991 0026 474
7	0.0991 0026 477	0.0991 0026 474

Here, it is noticed that F_1 and F_2 appear to converge and, in addition, agree in the values of the function. The discrepancy in the last figure could be accounted for by the fact that machine arithmetic is accurate to about 11 significant figures.

CHAPTER IV

Interpolating Functions

Introduction

One of the oldest problems in approximation is that in which we seek to express a given function f(x) in terms of a simpler function p(x) such that f(x) and p(x) and certain of their derivatives agree at given points x_i, i = 0,1 ....n. The p(x) is described as the interpolating function and can be used as an approximation to f(x) for values of x within the range defined by the x_i.

In this chapter, methods are described which define p(x) as a polynomial and as a rational function. A method of choosing nodes is given which is the best choice in terms of the minimum of the error norm.

Polynomial Interpolation Forms

An obvious choice for p(x) is the polynomial which takes the values of $f(x_i)$ at the nodes x_i. This is the Lagrange formula, given by

$$L_n(x) = \sum_{i=0}^{n} l_i(x)f(x_i)$$

$$\text{where } l_i(x) = \prod_{k=0}^{n} \frac{(x - x_k)}{(x_i - x_k)} \qquad k \neq i \quad i = 0,1,\ldots\ldots n \tag{4.1}$$

If we wish the polynomial to take the value of the function and its first derivative at the nodes, then we have the Hermite formula, given by

$$H_{2n+1}(x) = \sum_{i=0}^{n} l_i^2(x)\left[1 - 2l_i'(x_i)(x-x_i)\right] f(x_i) + \sum_{i=0}^{n} l_i^2(x_i)(x-x_i)f'(x_i) \tag{4.2}$$

We notice that whereas in (4.1) each node point leads to the introduction of a factor $(x-x_i)$, in (4.2) there is a factor $(x-x_i)^2$ to ensure the correspondence of both function and its first derivative. As the number of derivatives introduced into the constraint is increased, so more and more nodes can be thought to coalese. The ultimate would be if all nodes corresponded to one point in which case, at that point, there would be agreement of function value and its derivatives. This gives the well-known Taylor series representation, which is discussed in Chapter V.

Error in the Interpolation Formulæ

Consider the error in using the Lagrange formula. Since $L_n(x)$ agrees with $f(x)$ at $(n+1)$ chosen points, we assume an error of the form

$$f(x) - L_n(x) = C(x-x_0)(x-x_1) \ldots\ldots.(x-x_n)$$

Let $F(x)$ be a function of the form

$$F(x) = f(x) - L_n(x) - C(x-x_0)(x-x_1) \ldots\ldots.(x-x_n)$$

$F(x)$ will have zeros at $x_0, x \ldots\ldots x_n$ and by choosing a new argument x_{n+1}, between x_0 and x_n, with

$$C = \frac{f(x_{n+1}) - L_n(x_{n+1})}{(x_{n+1}-x_0)(x_{n+1}-x_1) \ldots\ldots(x_{n+1}-x_n)} \qquad (4.3)$$

then $F(x_{n+1}) = 0$ and $F(x)$ has at least $(n+2)$ zeros.

By Rolle's Theorem, $F'(x)$ will have at least $(n+1)$ zeros and continuing in this way $F^{(n+1)}(x)$ will have one zero, say x' between x_0 and x_n.

i.e. $0 = f^{(n+1)}(x') - C(n+1)!$ since $L_n(x)$ is of degree n only.

Substituting for C in (4.3) gives

$$f(x_{n+1}) - L_n(x_{n+1}) = \frac{f^{(n+1)}(x')}{(n+1)!}(x_{n+1}-x_0) \ldots\ldots (x_{n+1}-x_n)$$

Now x_{n+1} can be any point in the range x_0, x_n and in addition, the equation is valid at the $(n+1)$ nodes.

$$\therefore \quad f(x) - L_n(x) = \frac{f^{(n+1)}(x')}{(n+1)!} (x-x_0) \ldots\ldots(x-x_n), \qquad x_0 \leqslant x' \leqslant x_n \qquad (4.3a)$$

Similarly, it can be shown that the error in the Hermite formula is given by

$$f(x) - H_{2n+1}(x) = \frac{f^{(2n+2)}(x')}{(2n+2)!} (x-x_0)^2 (x-x_1)^2 \ldots\ldots.(x-x_n)^2 \qquad (4.4)$$

where $x_0 \leqslant x' \leqslant x_n$

It is assumed that all the required derivatives of $f(x)$ exist. The error expressions are functions of the degree of approximation and of the particular choice of nodes. It is desirable that

$$p_n(x) \rightarrow f(x) \qquad \text{as} \quad n \rightarrow \infty \qquad (4.5)$$

The question arises as to what choice of nodes will produce (4.5) or conversely, given an arbitrary choice of nodes (e.g. equally-spaced) is (4.5) generally true.

Convergence of Lagrange and Hermite Formulæ

Suppose that we wish to interpolate in the range $[-1,1]$ using a Lagrange formula. If $f(x)$ has bounded derivatives in the region, then (4.3a) requires that we minimize the expression

$$\max |(x-x_0)(x-x_1) \ldots\ldots (x-x_n)| \quad \text{in} \quad [-1,1] \qquad (4.6)$$

In Appendix A4.1 it is shown that the polynomial in (4.6) must be identical to $2^{-n}\, T_{n+1}(x)$, where $T_{n+1}(x)$ is the Chebyshev polynomial of degree (n+1). Hence, the nodes are given as the zeros of $T_{n+1}(x)$ i.e.

$$x_i = \cos\left[\frac{2_i + 1}{n + 1}\right]\frac{\pi}{2} \qquad i = 0,1, \ldots\ldots n$$

A similar result can be obtained for the Hermite polynomial. Again, if the derivatives of $f(x)$ are bounded, we can choose as the error norm to minimise the expression.

$$\left\{\max \left|\frac{f^{(2n+2)}(x)}{(2n+2)!}\right|\right\} \int_{-1}^{1} (x-x_0)^2\, (x-x_1)^2 \ldots\ldots (x-x_n)^2\, dx \qquad (4.7)$$

In Appendix A4.2 it is shown that the function under the integral sign must be orthogonal with respect to unit weight function over $[-1,1]$. Hence, the nodes must be the zeros of the Legendre Polynomial of appropriate degree $P_{n+1}(x)$.

If we turn our attention to the case of equally-spaced nodes, we find the results are discouraging. It has been demonstrated (e.g. by Runge) that even for a well-behaved function, approximations on equally spaced nodes can be shown to diverge as the number of nodes increases. This has been analysed by considering the behaviour of the given function in the complex plane surrounding the real region of approximation. It can be shown that if $f(z)$ has poles which are close to the real interval $[a,b]$, then it may prove impossible to find a sequence $L_n(x)$ to satisfy (4.5).

For a discussion of this problem and an example of the Runge phenomenon, see D. C. Handscomb [7], Chapter 3. A list of positive and negative results concerning the existence of interpolating functions is given in J. Todd [16] pp 146 - 149.

Newton's Interpolation Formula

Newton derived the interpolating polynomial through the given function values in a different form from that of Lagrange.

$$\text{Let } F(x) = f[x_0] + (x-x_0)f[x_0,x]$$

$$f[x_0,x] = f[x_0,x_1] + (x-x_1)f[x_0,x_1,x] \qquad (4.8)$$

$$f[x_0, \ldots\ldots x_{n-1},x] = f[x_0, \ldots\, x_{n-1},x_n] + (x-x_n)f[x_0,x_1, \ldots\, x_n,x]$$

Then the functions $f[x_0, \ldots x_i]$ are called divided differences and are defined by

$$f[x_0] = f(x_0) \qquad\qquad f[x_0,x_1] = \frac{f[x_1] - f[x_0]}{x_1 - x_0}$$

$$f[x_0, \ldots\, x_i] = \frac{f[x_1, \ldots\, x_i] - f[x_0, \ldots\, x_{i-1}]}{x_i - x_0} \qquad (4.8a)$$

Back-substitution in (4.8) leads to Newton's formula involving divided differences

$$F(x) = f[x_0] + (x-x_0)f[x_0,x_1] + (x-x_0)(x-x_1)f[x_0,x_1,x_2] + --- \qquad (4.9)$$

$$--- + (x-x_0) \ldots\ldots (x-x_{n-1})f[x_0, \ldots\, x_n] + E_n(x)$$

where $E_n(x) = (x-x_0) \ldots\, (x-x_n)f[x_0, \ldots\, x_n,x]$

If $E_n(x)$ is truncated, (4.9) defines a polynomial of degree n which will pass through (n+1) given points [F. Hildebrand [10]]. The coefficients in (4.9) are most conveniently evaluated by forming a divided difference table. e.g.

x	f(x)				
0	-1				
		1			
1	0		-1/5		
		3/5		0	
2	3/5		-1/5		1/85
		1/5		4/85	
3	4/5		-1/17		
		7/85			
4	15/17				

Interpolation formulæ may be arranged in a variety of ways, most conveniently of terms lying on a continuous path through the table. Using the two paths indicated, we have

$$F(x) = -1 + x + x(x-1)(-\tfrac{1}{5}) + x(x-1)(x-2)(x-3)(\tfrac{1}{85})$$

and

$$F(x) = \tfrac{3}{5} + (x-2)(\tfrac{1}{5}) + (x-2)(x-3)(-\tfrac{1}{5}) + (x-2)(x-3)(x-1)(\tfrac{4}{85})$$
$$+ (x-2)(x-3)(x-1)(x-4)(\tfrac{1}{85})$$

Approximation by Continued Fractions

The formula for interpolation given in (4.9) can be considered in the following recursive form

$$F(x) = u_0(x)$$
$$u_k(x) = u_k(x_k) + (x-x_k)u_{k+1}(x) \qquad k = (n-1), \;\ldots\ldots 1,0$$
$$\text{where} \qquad u_k(x) = f\left[x_0, x_1 \ldots\ldots x_{k-1}, x\right]$$

We now consider a similar recursive form

$$F(x) = v_0(x)$$
$$v_k(x) = v_k(x_k) + \frac{x-x_k}{v_{k+1}(x)} \qquad k = (n-1) \;\ldots\ldots 1,0 \quad (4.10)$$

The first few terms are

$$F(x) = v_0(x)$$
$$F(x) = v_0(x_0) + \frac{x-x_0}{v_1(x)}$$
$$F(x) = v_0(x_0) + \cfrac{x-x_0}{v_1(x_1) + \cfrac{x-x_1}{v_2(x)}}$$

If the $v_k(x)$ can be chosen correctly, then the function $F(x)$ can be made to take the values $f(x_i)$ at the nodes x_i.

In Appendix A4.3, we show that if

$$v_k(x) = \phi_k\left[x_0, x_1 \ldots\ldots x_{k-1}, x\right]$$

then (4.10) is an interpolating rational function where

$$\phi_1[x_0,x] = \frac{x - x_0}{\phi_0[x] - \phi_0[x_0]}$$

$$\phi_2[x_0,x_1,x] = \frac{x - x_1}{\phi_1[x_0,x] - \phi_1[x_0,x_1]}$$

$$\phi_k[x_0, \ldots x_{k-1},x] = \frac{x - x_{k-1}}{\phi_{k-1}[x_0, \ldots x_{k-2},x] - \phi_{k-1}[x_0, \ldots x_{k-2},x_{k-1}]} \qquad (4.10a)$$

Comparing (4.10a) with (4.8a), we see that the ϕ's are the inverted divided differences (or more simply, the inverted differences) of the Newton formula.

As for the divided differences, a table of inverted differences may be formed. However, one important difference exists between the two tables. In the inverted difference table, the order in which the points are introduced is important.

i.e. whereas

$$f[x_0,x_1,x_2,x_3] = f[x_2,x_3,x_1,x_0]$$ for divided differences

$$\phi_3[x_0,x_1,x_2,x_3] \neq \phi_3[x_2,x_2,x_1,x_0]$$ for inverted differences.

e.g.

x	$f(x)$	ϕ_1	ϕ_2	ϕ_3	ϕ_4
0	-1				
1	0	1			
2	$3/5$	$5/4$	4		
3	$4/5$	$5/3$	3	-1^*	
4	$15/17$	$17/8$	$8/3$	$-3/2$	-2

Then, using (4.10) we have the continued fraction

$$F(x) = -1 + \frac{x}{1 +}\;\frac{x-1}{4 +}\;\frac{x-2}{-1 +}\;\frac{x-3}{-2}$$

Now if interpolation is required in the middle of the range, the points starting at x = 2 may be introduced first and then work outwards.

x	$f(x)$	ϕ_1	ϕ_2	ϕ_3	ϕ_4
2	$3/5$				
3	$4/5$	5			
1	0	$5/3$	$3/5$		
4	$15/17$	$85/12$	$12/25$	-25	
0	-1	$5/4$	$4/5$	-5*	$-1/5$

and the interpolating fraction can be written as

$$F(x) = \frac{3}{5} + \frac{x-2}{5+}\ \frac{x-3}{3/5+}\ \frac{x-1}{-25+}\ \frac{x-4}{-1/5}$$

The entries marked with an asterisk are respectively $\phi_3[0,1,2,3]$ and $\phi_3[2,3,1,0]$ and it can be seen that they are not equal.

Reciprocal Difference Formula (Thiele Expansion)

When using interpolation formulae, we usually require the tabulated point nearest to the interpolation point to be used as the base point of the formula. In this way, the correction terms that are calculated should be small in magnitude and the effects of rounding error will be minimized. In the last section it was seen that reordering the points requires a complete recalculation of the inverted difference table. This can be avoided by using a slightly more complicated form of difference table.

The continued fraction interpolation so obtained is called a Thiele expansion. In Appendix A4.4 is derived the following form.

$$F(x) = f(x_0) + \frac{x - x_0}{\rho_1[x_0,x_1] +}\ \frac{x - x_1}{\rho_2[x_0,x_1,x_2] - f(x_0) +}\ \frac{x - x_3}{\rho_3[x_0,x_1,x_2,x_3] - \rho_1[x_0,x_1] +} \tag{4.11}$$

when the ρ's are termed the reciprocal differences defined by

$$\rho_k[x_0, \dots, x_k] = \frac{x_k - x_{k-1}}{\rho_{k-1}[x_0, \dots x_{k-2}, x_k] - \rho_{k-1}[x_0, \dots x_{k-2}, x_{k-1}]} + \rho_{k-2}[x_0, \dots x_{k-2}]$$

Again, the reciprocal differences may be arranged in the form of a table. By taking various continuous paths through this table, interpolation may be carried out anywhere in the range of the arguments.

Example

x	f(x)	ρ_1	ρ_2	ρ_3	ρ_4
0	-1				
		1			
1	0		3		
		5/3		0	
2	3/5		6/5		1
		5		-20	
3	4/5		29/25		
		85/4			
4	15/17				

Taking the leading diagonal, we obtain the expansion

$$F(x) = -1 + \frac{x}{1+} \; \frac{x-1}{4+} \; \frac{x-2}{-1+} \; \frac{x-3}{-2} \qquad (4.12)$$

Now, if the points are reordered, say for interpolation near x = 2, we can establish that

$$F(x) = \frac{3}{5} + \frac{x-2}{5+} \quad \frac{x-3}{3/5+} \quad \frac{x-1}{-25+} \quad \frac{x-4}{-1/5} \qquad (4.13)$$

Thiele's Expansion Involving Derivatives

An interesting case of the Thiele expansion is that when only one node is taken and the k^{th} convergent of the continued fraction agrees with the value $f(x_0)$ and (k-1) of its derivatives at x_0. In Appendix A4.5 we derive the form

$$F(x) = \phi_0(x_0) + \frac{x-x_0}{\phi_1(x_0)+} \quad \frac{x-x_0}{\phi_2(x_0)+} \quad \frac{x-x_0}{\phi_3(x_0)+} \qquad (4.14)$$

where $$\phi_{k+1}(x) = \frac{k+1}{\rho'_k(x)}$$

$$\rho_k(x) = \rho_{k-2}(x) + \phi_k(x)$$

with starting values

$$\rho_{-2}(x) = \rho_{-1}(x) = 0 \qquad \phi_0(x) = f(x)$$

This form can be used to provide formal expansions to continuously differentiable functions in the form of rational approximations. However, it may be that the function concerned is not expressible in the form (4.14). For example, if the function is symmetrical about x_0, the rational function derived from a truncated continued-fraction expansion could only contain even powers of $(x - x_0)$. In such cases some modification of the form is required.

In analogy with the Taylor series, such expansions are extremely good near to x_0, but rapidly become worse as we move away. In later Chapters, methods will be discussed of modifying the basic functions in order to find a more equal distribution of the error.

Comparison of Methods.

The various methods described in this Chapter may each be used to advantage in different circumstances. The Hermite formula is useful when the value of the derivative is prescribed at certain points. In cases where the derivative is unknown, formulæ involving only the ordinates must be used. The Lagrange form does not involve evaluation of a difference table, but estimation of the truncation error is not possible unless the given function is known analytically. The Newton form involves a difference table, but the truncation error may be estimated from the first neglected difference. Also, extra points are more easily incorporated in the Newton divided difference formula.

The use of continued fraction form is well-suited to problems of interpolation in the region of a point at which the given function becomes infinite. Continued fractions can be evaluated very quickly since only a few divisions are involved. However, particularly with the reciprocal difference form, derivation of the difference tables can be laborious. Also, it is possible that the particular continued fraction form that we seek does not exist. This problem can sometimes be overcome by a reordering of the points.

CHAPTER V

Approximations Derived from Series Expansions

Introduction

Some methods are described for deriving approximations by means of series expansions. The Padé table is shown to be one method of deriving a rational function approximation. As example is given of the use of a series of Chebyshev polynomials in the solution of a differential equation.

The Taylor Series

If a function and its first n derivatives are continuous in a region $[a,x]$, then the function can be represented as

$$f(x) = f(a) + \frac{f'(a)(x-a)}{1!} + \frac{f''(a)(x-a)^2}{2!} + \ldots\ldots + \frac{f^n(\xi)(x-a)^n}{n!} \qquad (5.1)$$

where $a \leqslant \xi \leqslant x$ and the superscript represents differentiaion.

An approximation to $f(x)$ is formed by omitting the last term in (5.1) and the truncated series will then agree with the value of the function and its first (n-1) derivatives at the single point $x = a$. The truncation error $e(x)$ is such that

$$|e(x)| \leqslant \left[\max_{[a,x]} \frac{f^n(x)}{n!}\right] (x-a)^n$$

The value of $|e(x)|$ will remain small in a region close to $x = a$, but will rapidly increase as we move away from this point. From a practical point of view, the infinite series obtained by allowing n to increase in (5.1) must not only be formally convergent, but the terms must decrease in magnitude rapidly enough to allow a reasonable point of truncation to be chosen. Also, the error estimate involving the n^{th} derivative may be difficult to evaluate. In such cases, reasonable error bounds are often found by other means. For example, the sum of the remainder terms may be estimated by comparison with a known series (e.g. a geometric series.) In the case of a series of alternating sign whose terms can be shown to be monotonically decreasing, we can bound the error using the first neglected term.

i.e. if $S = S_n - a_{n+1} + a_{n+2} - a_{n+3} + a_{n+4} - \dots\dots$

$$= S_n - (a_{n+1} - a_{n+2}) - (a_{n+3} - a_{n+4}) - \dots\dots$$

and since all the bracketed terms are positive

$$S < S_n$$

equally $S = S_n - a_{n+1} + (a_{n+2} - a_{n+3}) + (a_{n+4} - a_{n+5}) + \dots\dots$

and $S > S_n - a_{n+1}$

$$\therefore \quad S_n - a_{n+1} < S < S_n$$

A truncated power series can often be used as an effective approximation so long as the interval over which it is applied is small. Even with modern computing speed, and time is of no importance, if large numbers of terms have to be summed, then large accumulations of rounding error could occur, particularly if the first few terms are very large and considerable cancellation takes place. However, power series are easily integrated or differentiated term-by-term and there is usually a fairly reliable estimate of truncation error.

Asymptotic Series

If the behaviour of a function over a range involving large values of the argument is of concern, say $a \leq x < \infty$, then it is unlikely that a Taylor series will be practicable due to the large magnitude of the terms generated. (Even though a series may be formally convergent.)

If it is possible to define a series

$$f(x) = \sum_{n=0}^{\infty} \frac{a_n}{x^n}$$, then the series is said to be asymptotic

at infinity if for every n,

$$\lim_{x \to \infty} \left| x^n \left\{ f(x) - S_n(x) \right\} \right| = 0$$

where $$S_n(x) = \sum_{k=0}^{n} \frac{a_k}{x^k} \tag{5.2}$$

It ~~is~~ may be possible to use (5.2) as a reasonable approximation for large x even though it may not be convergent for finite values of the argument.

Asymptotic series possess the property that for a given value of x, the terms ~~first decrease in magnitude and then steadily~~ ultimately increase without bound. ~~Also, it is usually possible to show that~~ However, in some cases, the truncation error is less than the first omitted term. Hence, in practice it is often possible to find asymptotic series in which the first terms decrease fairly quickly and a reasonable approximation may be obtained with an early point of truncation. Notice that in contrast to convergent series, there is no automatic gain in including extra terms in the summation. Once terms begin to increase in magnitude, there is everything to be lost by including them.

Example 1

$$f(x) = \int_x^{\infty} e^{-t^2} dt$$

Integrating by parts $f(x) = \int_x^{\infty} \left(\frac{-1}{2t}\right)(-2te^{-t^2})dt$

$$= \left[\frac{-1}{2t} e^{-t^2}\right]_x^{\infty} - \int_x^{\infty} e^{-t^2}\left(\frac{1}{2t^2}\right) dt$$

$$= \frac{1}{2x} e^{-x^2} + \left[\frac{1}{4t^3} e^{-t^2}\right]_x^{\infty} + \int_x^{\infty} e^{-t^2}\left(\frac{3}{4t^4}\right) dt$$

$$= \frac{1}{2x} e^{-x^2} - \frac{1}{4x^3} e^{-x^2} - \left[\frac{1.3}{8t^5} e^{-t^2}\right]_x^{\infty} - \int_x^{\infty} e^{-t^2}\left(\frac{1.3.5}{8t^6}\right) dt$$

$$\therefore\ f(x) = e^{-x^2}\left\{\frac{1}{2x} - \frac{1}{4x^3} + \frac{1.3}{8x^5} + \dots\dots\dots\right\} + \frac{(-1)^{n-1}\, 1.3..(2n-1)}{2^n}\int_x^{\infty} e^{-t^2}\frac{dt}{t^{2n}}$$

Now if $|R_n| = \frac{1.3.5\ \dots.(2n-1)}{2^n}\int_x^{\infty} e^{-t^2}\frac{dt}{t^{2n}}$

$$= e^{-x^2} . \frac{1.3.5\ \dots.\ (2n-1)}{2^{n+1}\, x^{2n+1}} - \int_x^{\infty} \frac{1.3\ \dots.(2n-1)(2n+1)}{2^{n+1}} e^{-t^2}\frac{dt}{t^{2n+2}}$$

The first term on the right-hand side is the first neglected term of the series and the integral is positive and preceded by a negative sign. Hence

R_n will be smaller in magnitude than the first neglected term. If we put $x = 4$, the series becomes

$$f(4) = e^{-16}\left\{\frac{1}{8} - \frac{1}{256} + \frac{1.3}{8192} \cdots\cdots \right\}$$

As we proceed from one term to the next, the denominator is multiplied each time by 32. The terms will continue to decrease until the factors in the numerator exceed 32 when they will steadily increase without bound.

Example 2

Consider the derivation of an asymptotic series for the Bessel Function $J_n(x)$.

Now $J_n(x) = \sqrt{\frac{2}{\pi x}}\left\{P(x,n)\cos\left(x - \frac{n\pi}{2} - \frac{\pi}{4}\right) - Q(x,n)\sin\left(x - \frac{n\pi}{2} - \frac{\pi}{4}\right)\right.$

where

$$P(x,n) = \frac{1}{2\Gamma(n+\frac{1}{2})}\int_0^\infty e^{-u}u^{n-\frac{1}{2}}\left\{\left(1+\frac{iu}{2x}\right)^{n-\frac{1}{2}} + \left(1-\frac{iu}{2x}\right)^{n-\frac{1}{2}}\right\} du$$

$$Q(x,n) = \frac{1}{2\Gamma(n+\frac{1}{2})}\int_0^\infty e^{-u}u^{n-\frac{1}{2}}\left\{\left(1+\frac{iu}{2x}\right)^{n-\frac{1}{2}} - \left(1-\frac{iu}{2x}\right)^{n-\frac{1}{2}}\right\} du$$

[See G. N. Watson, A Treatise on the Theory of Bessel Functions.]

Consider the expansion

$$\left(1+\frac{iu}{2x}\right)^{n-\frac{1}{2}} = 1+\frac{(n-\frac{1}{2})}{1!}\left(\frac{iu}{2x}\right)+\frac{(n-\frac{1}{2})(n-\frac{3}{2})}{2!}\left(\frac{iu}{2x}\right)^2 + \cdots\cdot$$
$$+ \frac{(n-\frac{1}{2})\ \ldots(n-r+\frac{3}{2})}{(r-1)!}\left(\frac{iu}{2x}\right)^{r-1} + \frac{(n-\frac{1}{2})\ \ldots\ (n-r+\frac{1}{2})}{(r-1)!}\int_0^1\left(\frac{iu}{2x}\right)^r (1-t)^{r-1}\left(1+\frac{iut}{2x}\right)^{n-r-\frac{1}{2}} dt$$

[The remainder term can be derived from the more usual form $R(y) = \int_0^y \frac{(y-s)^{r-1}}{(r-1)!} f^r(y)ds$ by writing $y = \frac{iu}{2x}$, $s = \frac{iut}{2x}$ and $f(y) = (1+y)^{n-\frac{1}{2}}$]

Now for t in $[0,1]$, $\left|1+\frac{iut}{2x}\right| \geqslant 1$, hence if $r > n-\frac{1}{2}$ (5.3)

$$\left|\int_0^1 (1-t)^{r-1}\left(1+\frac{iut}{2x}\right)^{n-r-\frac{1}{2}} dt\right| \leqslant \left|\int_0^1 (1-t)^{r-1} dt\right| = \frac{1}{r}$$

$$\therefore \left(1+\frac{iu}{2x}\right)^{n-\frac{1}{2}} = 1 + \frac{(n-\frac{1}{2})}{1!}\left(\frac{iu}{2x}\right) + \frac{(n-\frac{1}{2})(n-\frac{3}{2})}{2!}\left(\frac{iu}{2x}\right)^2 + \cdots\cdots$$

$$\frac{(n-\frac{1}{2})\ \ldots(n-r+\frac{3}{2})}{(r-1)!}\left(\frac{iu}{2x}\right)^{r-1} + \xi_1\frac{(n-\frac{1}{2})\ \ldots(n-r+\frac{1}{2})}{r!}\left(\frac{iu}{2x}\right)^r$$

where $|\xi_1| \leqslant 1$

By a similar argument

$$\left(1 - \frac{iu}{2x}\right)^{n-\frac{1}{2}} = 1 - \frac{(n-\frac{1}{2})}{1!}\left(\frac{iu}{2x}\right) + \frac{(n-\frac{1}{2})(n-\frac{3}{2})}{2!}\left(\frac{iu}{2x}\right)^{2} + \ldots$$

$$+ \frac{(-1)^{r-1}(n-\frac{1}{2}) \ldots (n-\frac{1}{2}+\frac{3}{2})}{(r-1)!}\left(\frac{iu}{2x}\right)^{r-1} + \frac{\xi_2(n-\frac{1}{2}) \ldots (n-r+\frac{1}{2})}{r!}\left(\frac{iu}{2x}\right)^{r}$$

If the two expressions are added, terms of odd degree will cancel, whilst if they are subtracted, the even degree terms cancel.

Hence, writing $r = 2p$ gives

$$\left(1+\frac{iu}{2x}\right)^{n-\frac{1}{2}} + \left(1 - \frac{iu}{2x}\right)^{n-\frac{1}{2}} = 2 + 2\sum_{m=1}^{p-1} \frac{(-1)^{m}(n-\frac{1}{2})(n-\frac{3}{2}) \ldots (n-2m+\frac{1}{2})}{(2m)!}\left(\frac{u}{2x}\right)^{2m}$$

$$+ 2\frac{\xi_3(n-\frac{1}{2}) \ldots (n-2p+\frac{1}{2})}{(2p)!}\left(\frac{iu}{2x}\right)^{2p} \qquad |\xi_3| \leqslant 1 \qquad (5.4)$$

and from (5.3) $\quad 2p > n-\frac{1}{2}$

Equally, putting $r = 2p+1$

$$\left(1 + \frac{iu}{2x}\right)^{n-\frac{1}{2}} - \left(1 - \frac{iu}{2x}\right)^{n-\frac{1}{2}} = 2i\sum_{m=1}^{p} \frac{(-1)^{m-1}(n-\frac{1}{2}) \ldots (n-2m+\frac{3}{2})}{(2m-1)!}\left(\frac{u}{2x}\right)^{2m-1}$$

$$+ 2\frac{\xi_4(n-\frac{1}{2})(n-\frac{3}{2}) \ldots (n-2p-\frac{1}{2})}{(2p+1)!}\left(\frac{iu}{2x}\right)^{2p+1} \qquad (5.5)$$

and $2p + 1 > n-\frac{1}{2}$ or $2p > n-\frac{3}{2}$

Hence, substituting the series (5.4) and (5.5)

$$P(x,n) = \frac{1}{\Gamma(n+\frac{1}{2})}\int_0^{\infty} e^{-u}\, u^{n-\frac{1}{2}} \left\{1 + \sum_{m=1}^{p-1} \frac{(-1)^{m}(n-\frac{1}{2}) \ldots . (n-2m+\frac{1}{2})}{(2m)!}\left(\frac{u}{2x}\right)^{2m} + \right.$$

$$\left. \frac{\xi_3(n-\frac{1}{2}) \ldots (n-2p+\frac{1}{2})}{(2p)!}\left(\frac{u}{2x}\right)^{2p}\right\} dx$$

$$\text{Now} \quad \frac{1}{\Gamma(n+\frac{1}{2})}\int_0^{\infty} e^{-u}\, u^{n-\frac{1}{2}}\left(\frac{u}{2x}\right)^{2m} dx = \frac{1}{(2x)^{2m}\Gamma(n+\frac{1}{2})}\int_0^{\infty} e^{-u}\, u^{2m+n-\frac{1}{2}}\, du$$

$$= \frac{1}{(2x)^{2m}} \cdot \frac{\Gamma(2m+n+\frac{1}{2})}{\Gamma(n+\frac{1}{2})}$$

$$= \frac{1}{(2x)^{2m}}(2m+n-\frac{1}{2})(2m+n-\frac{3}{2}) \ldots (n+\frac{1}{2})$$

The m^{th} term in the integral for P(x,n) becomes

$$\frac{(-1)^m(n-\frac{1}{2})(n-\frac{3}{2}) \;\ldots\; (n-\overline{2m-\frac{1}{2}})(n+2m-\frac{1}{2})(n+2m-\frac{3}{2}) \;\ldots\;(n+\frac{1}{2})}{(2m)!\;(2x)^{2m}}$$

$$= \frac{(-1)^m(n^2-\frac{1}{2}^2)(n-\frac{3}{2}^2) \;\ldots\;(n^2-(2m-\frac{1}{2})^2)}{(2m)!\;(2x)^{2m}} \tag{5.6}$$

To consider the remainder term, we notice that

$$\left| \int_0^{\infty} \xi_1\, e^{-u}\, u^{n+2p-\frac{1}{2}}\, du \right| \leqslant \int_0^{\infty} e^{-u}\, u^{n+2p-\frac{1}{2}}\, du = \Gamma(n+2p+\tfrac{1}{2})$$

and if $2p > n-\frac{1}{2}$, as in (5.6)

$$|R| \leqslant \frac{(n^2-\frac{1}{2}^2)(n^2-\frac{3}{2}^2) \;\ldots\; (n^2-(2p-\frac{1}{2})^2)}{(2p)!\;(2x)^{2p}}$$

Hence, P(x,n) may be represented by the finite expansion

$$P(x,n) = 1+\sum_{m=1}^{p-1} \frac{(-1)^m(n^2-\frac{1}{2}^2)(n^2-\frac{3}{2}^2) \;\ldots\;(n^2-(2m-\frac{1}{2})^2)}{(2m)!\;(2x)^{2m}}$$

and the remainder will not exceed in absolute value the first neglected term.

In exactly the same manner, we find that

$$Q(x,n) = \sum_{m=1}^{p} \frac{(-1)^{m-1}(n^2-\frac{1}{2}^2)(n^2-\frac{3}{2}^2) \;\ldots\;(n^2(2m-\frac{3}{2})^2)}{(2m-1)!\;(2x)^{2m-1}}$$

where, if $2p > n-\frac{3}{2}$

$|R| \leqslant (p+1)^{th}$ term.

The Padé Table

It is possible to derive a sequence of approximations in the form of rational functions from the power series representation of the original function. (Handscomb[7])

Assume $f(x) = c_0 + c_1 x + c_2 x^2 + \ldots\ldots$ (5.3)

and that $R_{mn}(x)$ is a rational approximation of the form

$$R_{mn}(x) = \frac{a_0 + a_1 x + a_2 x^2 \ldots\ldots + a_m x^m}{1 + b_1 x + b_2 x^2 + \ldots\; b_n x^n} \tag{5.4}$$

where m and n are the degree (at most) of the numerator and denominator respectively. The coefficients in (5.4) are chosen so that $f(x) - R_{mn}(x)$ expressed as a power series has no term of degree less than $(m+n+1)$.

That is

$$(c_0 + c_1 x + \dots\dots c_{m+n} x^{m+n} + \dots\dots)(1 + b_1 x + \dots + b_n x^n) - (a_0 + a_1 x + \dots + a_m x^m) \qquad (5.5)$$

must be free of terms of degree less than (m + n + 1).

Equating coefficients to zero in (5.5) gives a set of (m+n+1) equations to determine the (m+n+1) unknowns in $R_{mn}(x)$.

The coefficients of x^{m+1} to x^{m+n} give a system of equations that can be solved for the b's, the remainder of the equations, involving both a and b can be solved once the b values are known

i.e.
$$\begin{bmatrix} c_{m+1-n} & c_{m+2-n} & \dots\dots & c_m & c_{m+1} \\ - & - & - & - & - \\ & & & & \\ c_m & c_{m+1} & \dots\dots & c_{m+n-1} & c_{m+n} \end{bmatrix} \begin{bmatrix} b_n \\ b_{n-1} \\ \vdots \\ b \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \qquad (5.6).$$

The matrix of coefficients $[c]$ has n rows and (n+1) columns. Since the right-hand-side is zero, it can be seen that (5.6) has a non-trivial solution if the determinant

$$\begin{vmatrix} c_{m-n+1} & c_{m-n+2} & \dots\dots\dots\dots & c_m \\ c_{m-n+2} & c_{m-n+3} & \dots\dots\dots & c_{m+1} \\ & & & \\ c_m & c_{m+1} & \dots\dots\dots\dots & c_{m+n-1} \end{vmatrix} \neq 0$$

The approximations (5.4) that are obtained in this way are called Padé approximants and are dependent on the choice of m and n. They can be arranged in the form of a table in which m and n are used to denote the row and column number respectively.

i.e.
$$\begin{array}{llll} R_{00} & R_{01} & R_{02} & \dots\dots\dots \\ R_{10} & R_{11} & R_{12} & \dots\dots\dots \\ R_{20} & R_{21} & R_{22} & \dots\dots\dots \\ \dots\dots & \dots\dots & \dots\dots & \dots\dots \end{array}$$

The advantage of using the rational Padé form is in its computational efficiency, since it can be expressed conveniently as a continued fraction. One disadvantage however, is that if the degree of approximation is to be increased, the new Padé approximant has to be evaluated right from the beginning.

We now consider two examples that illustrate the derivation of Padé approximants.

Examples

(i) $$\log_e(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} \;.........$$

The approximant

$$R_{22} = \frac{a_0 + a_1 x + a_2 x^2}{1 + b_1 x + b_2 x^2}$$ is found by making

$$(x - \tfrac{1}{2}x^2 + \tfrac{1}{3}x^3 - \tfrac{1}{4}x^4 \ldots.)(1 + b_1 x + b_2 x^2) - (a_0 + a_1 x + a_2 x^2) = 0(x^5)$$

Equating coefficients of x^j for $j = 0,1 \ldots 4$

$$a_0 = 0 \qquad\qquad b_2 - \tfrac{1}{2}b_1 + \tfrac{1}{3} = 0$$

$$a_1 = 1 \qquad\qquad -\tfrac{1}{2}b_2 + \tfrac{1}{3}b_1 + \tfrac{1}{4} = 0$$

$$b_1 - \tfrac{1}{2} = a_2$$

from which $$R_{22} = \frac{x + \frac{1}{2}x^2}{1 + x + \frac{1}{6}x^2} \tag{5.7}$$

The error $\log_e(1+x) - R_{22}(x)$ is tabulated in Table 5.1 and we notice that the error behaves in a similar manner to that of the truncated Taylor series, that is it is small close to the origin but rapidly increases on either side. Clearly the Padé approximant is not a suitable approximation near $x = -1$, but it is reasonable to suppose that this is due to the rapid descent of $\log_e(1+x)$ to $-\infty$ in this region.

(ii) Consider

$$\cosh x - \sin x = 1 - x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} - \frac{x^5}{120} + \;.......$$

Then the first few entries in the Padé table are

$$\begin{array}{lll} 1 & \dfrac{1}{1+x} & \dfrac{1}{1+x+\frac{1}{2}x^2} \\[2ex] 1-x & \dfrac{1-\frac{1}{2}x}{1+\frac{1}{2}x} & \dfrac{1-\frac{1}{2}x}{1+\frac{1}{2}x+\frac{5}{12}x^2} \end{array}$$

$1-x+\frac{1}{2}x^2$ $\qquad\qquad$ $\dfrac{1-\frac{4}{3}x+\frac{5}{6}x^2}{1-\frac{1}{3}x}$ $\qquad\qquad$ $\dfrac{1-\frac{12}{10}x+\frac{49}{60}x^2}{1-\frac{2}{10}x+\frac{1}{60}x^2}$

etc.

The error (coshx - sinx) - $R_{22}(x)$ is tabulated below and again it can be noticed that the error curve demonstrates the same pattern as in the previous example.

x	-1	-.8	-.6	-.4	-.2	0	.2	.4	.6	.8	1.0
$\log_e(1+x) - R_{22}(x)$	$-\infty$	-.04423	-.00325	-.00018	-.00001	0	.0000	.00002	.00012	.00038	.00084
$(\cosh x - \sin x) - R_{22}(x)$	.01747	.00576	.00138	.00019	.00002	0	.00000	-.00019	-.00155	-.00623	-.01923

Error in Padé Approximants

Table 5.1

It is possible that the rational approximation may be more appropriate over a wider range if it could be arranged for the error to be more equally distributed. A method which attempts this is now described.

Economisation of Rational Functions

The object is to take the original Padé approximant $R_{mn}(x)$ and perturb it so that the error is more equally distributed throughout the range and thus reduce the maximum error. The method described here is that derived by Ralston [12]. It involves taking a combination of Padé approximants in such a way that the predominant terms in the remainder form a Chebyshev polynomial. The form of the modification is derived in Appendix 5.1

(iii) In the approximation to $\log_e(1+x)$, since we have already noted that this form is unlikely to be a reasonable fit near $x = -1.0$ we consider the range $[-0.6,0.6]$.

Now we have $\quad \dfrac{P_0^0}{Q_0^0} = 0, \quad \dfrac{P_1^2}{Q_1^2} = \dfrac{x}{1+\frac{1}{2}x} \quad$ and $\quad \dfrac{P_2^4}{Q_2^4} = \dfrac{x+\frac{1}{2}x^2}{1+x+\frac{1}{6}x^2}$

The new approximation is

$$R_{22}^{*}(x) = \frac{P_2^4 + \delta_3 P_1^2 + \delta_1 P_0^0}{Q_2^4 + \delta_3 Q_1^2 + \delta_1 Q_0^0} \qquad \begin{array}{l} \text{since } t_4 = t_2 = t_0 = 0 \\ t_3 = -20, \quad t_1 = 5 \end{array}$$

$$\text{then } \delta_3 = \frac{1}{180} \cdot \frac{12}{1} \cdot \frac{(0.6)^2}{16}(-20) = -0.03$$

$$\delta_1 = \frac{1}{180} \cdot 1 \cdot \frac{(0.6)^4}{16}(5) = 0.00023$$

$$\text{hence } R_{22}^{*}(x) = \frac{x + \frac{1}{2}x^2 - 0.03[x]}{1+x+\frac{1}{6}x^2 - 0.03[1+\frac{1}{2}x] + 0.00023}$$

$$= \frac{0.97x + \frac{1}{2}x^2}{0.97023 + 0.98500x + \frac{1}{6}x^2} \qquad (5.8)$$

The error is then found to be

x	-.6	-.5	-.4	-.3	-.2	-.1	0
$\log(1+x) - R_{22}^{*}(x)$	-.001053	-.000033	.000040	-.000014	-.000036	-.000025	0
	.1	.2	.3	.4	.5	.6	
	.000019	.000024	.000014	.000001	-.000013	-.000009	

Figure 5.2(a) shows a comparison between the error produced by the modified and unmodified approximation. It is noticeable that the economised function still does not produce an ideal Chebyshev form of error oscillation.

(iv) In the approximation to cosh x - sin x, the range will be taken as $[-1.0, 1.0]$ i.e. in (5.1.1) a = 1 and $R_{00}^{0}(x)$, $R_{11}^{2}(x)$ and $R_{22}^{4}(x)$ will be as shown in the Padé table.

In this example $t_4 = t_2 = t_0 = 0$

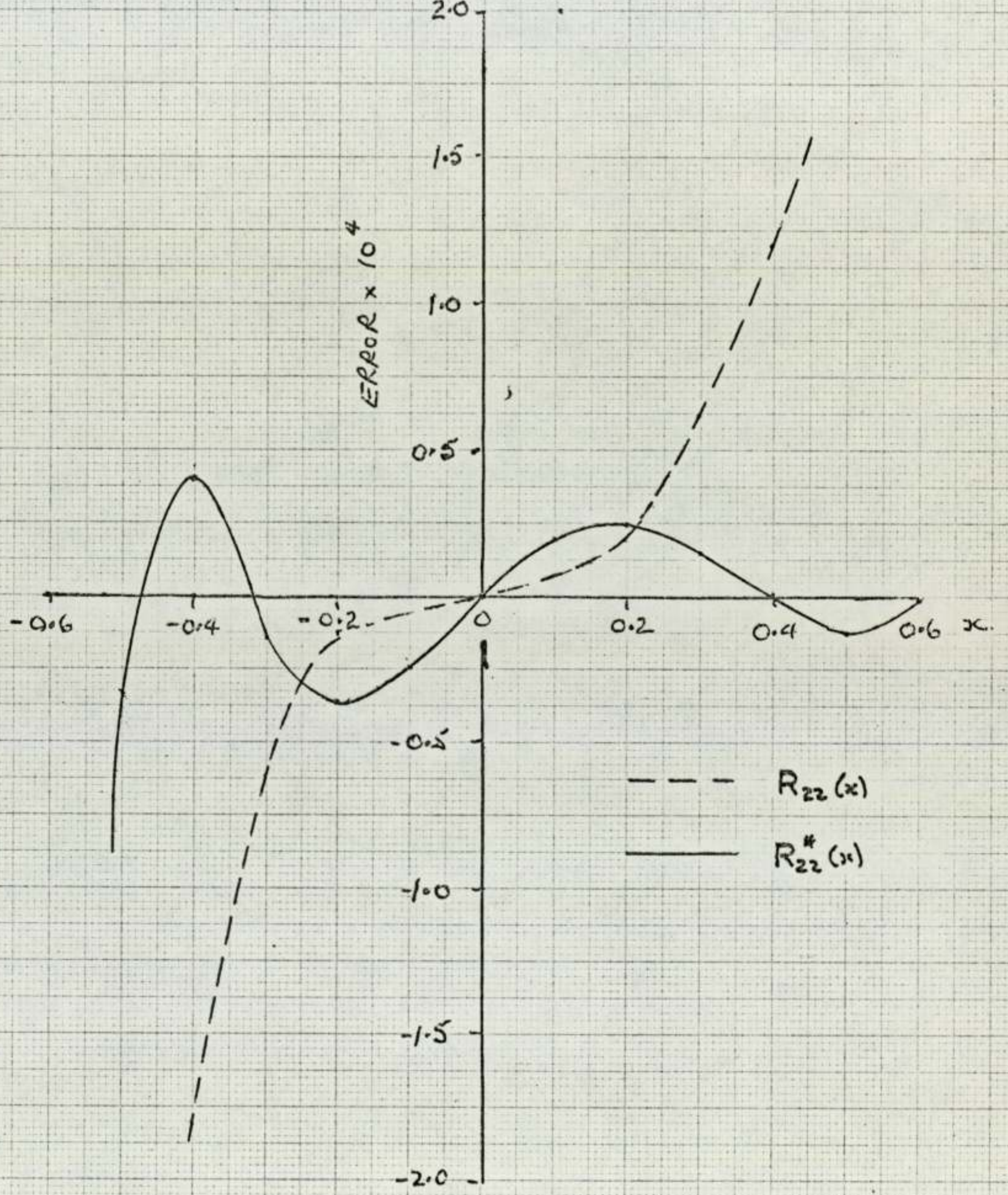
$t_3 = -20 \quad t_1 = 5$ as before

$$\text{and } \delta_3 = -\frac{13}{720} \cdot \frac{12}{5} \cdot \frac{1}{16} \cdot (-20) = \frac{13}{240} = 0.05417$$

$$\delta_1 = -\frac{13}{720}(-1) \cdot \frac{1}{16}(5) = \frac{13}{2304} = 0.00564$$

$$\text{hence } R^{*} = \frac{1.00000 - 1.30000x + 0.81667x^2 + 0.05417(1 - \frac{1}{2}x) + 0.00564(1)}{1.00000 - 0.30000x + 0.01667x^2 + 0.05417(1 + \frac{1}{2}x) + 0.00564(1)}$$

$$= \frac{1.05981 - 1.32708x + 0.81667x^2}{1.05981 - 0.27292x + 0.01667x^2} \qquad (5.9)$$



Approximation to $\log_e(1+x)$

Fig 5.2 (a)

The error $y(x) - R^*_{22}(x)$ is plotted in Figure 5.2(b) and it is obvious that the error distribution is not even approximately symmetrical about the horizontal axis. To investigate why this is so, we look at the nature of the modifying terms as given in Appendix A5.1.

The error in the economised form is given in this case by an expression of the form

$$y(x)-R^*_{22}(x) = \frac{Q_2^4 y(x) - P_2^4(x) + \delta_3[Q_1^2(x)y(x) - P_1^2(x)] + \delta_1[Q_0^0(x)y(x) - P_0^0(x)]}{Q_2^4(x) + \delta_3 Q_1^2(x) + \delta_1 Q_0^0(x)}$$

Now $Q_2^4 y(x)-P_2^4(x) = (1 - x + \frac{x^2}{2} + \frac{x^3}{6} \dots .)(1 - \frac{3x}{10} + \frac{1}{60} x^2) - (1 - \frac{13}{10} x + \frac{49x^2}{60})$

$$= - \frac{13}{720} x^5 + 0(x^6)$$

$$Q_1^2(x)y(x)-P_1^2(x) = (1 - x + \frac{x^2}{2} + \frac{x^3}{6} \dots)(1 + \tfrac{1}{2}x) - (1 - \tfrac{1}{2}x)$$

$$= \frac{5}{12} x^3 + \frac{x^4}{8} + \frac{x^5}{80} + 0(x^6)$$

$$Q_0^0(x)y(x)-P_0^0(x) = (1 - x + \frac{x^2}{2} + \frac{x^3}{6} \dots) 1 - 1$$

$$= -x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} - \frac{x^5}{120} + 0(x^6) \qquad (5.10)$$

Now if we take only the first terms of the remainders in (5.10) and substitute in the above expression, we get

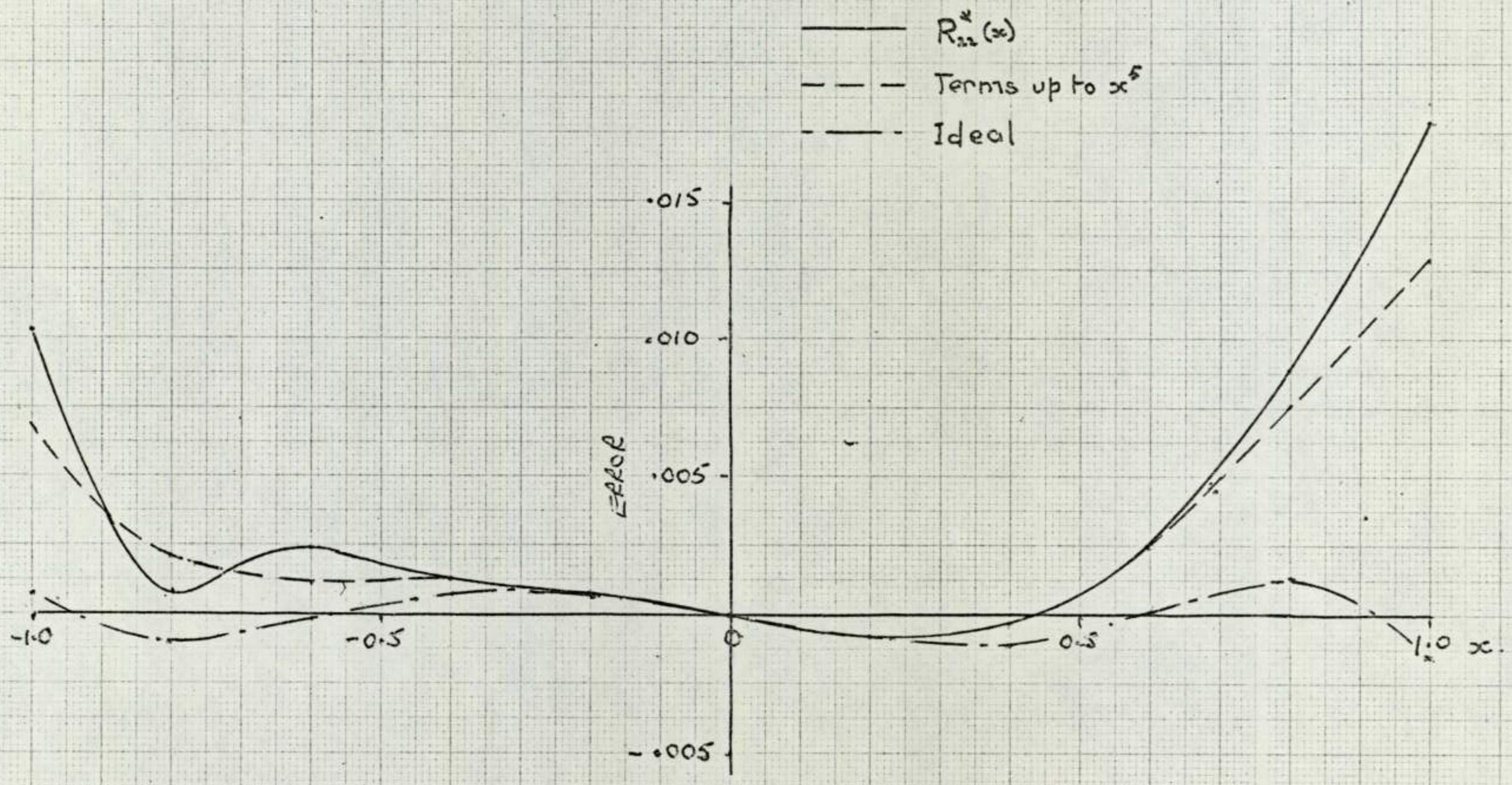
$$y(x)-R^*_{22}(x) = \frac{-0.01806x^5 + 0.02257x^3 - 0.00564x}{1.05981 - 0.27292x + 0.01667x^2} \qquad (5.11)$$

The expression in (5.11) is plotted in Figure 5.2(b) and it is clearly more like the shape that is desirable. Indeed, the numerator in (5.11) can be written

$$\frac{-0.01806}{16} [16x^5 - 20x^3 + 5x] = \frac{-0.01806}{16} T_5(x)$$

However, if <u>all</u> terms up to x^5 are retained in the remainders in (5.10), then

$$y(x)-R^*_{22}(x) = \frac{-0.01743x^5 + 0.00701x^4 + 0.02351x^3 + 0.00282x^2 - 0.00564x}{1.05981 - 0.27292x + 0.01667x^2} \qquad (5.12)$$



Approximation to $\cosh x - \sin x$

Fig 5.2 (b)

If the plot of (5.12) is compared with the actual error curve in figure 5.2(b), it can be seen that they are very similar. Hence it can be concluded that the presence of the extra terms in the error expressions account for the unsatisfactory shape of the error curve.

It seems that unless the range of approximation is kept relatively small, the form of the error produced by the economisation process may be far from ideal. In which case it is probably better to seek the true minimax approximation by some other method. (See Chapter VIII)

Expansion in a Series of Chebyshev Polynomials

A series in which the terms are Chebyshev polynomials offers several advantages when used as a means of approximation. In most circumstances, the coefficients in a Chebyshev series decrease rapidly in magnitude allowing early truncation of the series without incurring serious error. In particular, the truncated series has an error which is approximately equal to the first neglected term and hence may be nearly a function with the equal-error property. Since Chebyshev polynomials are only defined in the range $[-1,1]$, it may sometimes be necessary to make the substitution $x = \frac{1}{h}(z - a)$, which reduces $a - h \leq z \leq a + h$ to $-1 \leq x \leq 1$

One advantage of using Chebyshev polynomials is the relative ease with which they may be integrated or differentiated. This makes them particularly useful when the function under consideration can be expressed as the solution of a differential equation whose coefficients are polynomials in x. An example is given of this method of approach and two methods of estimating the error are compared.

Example

To find an approximation to y(x) given that

$$(3 + 2x)\, y' - y = 0 \quad \text{and } y(o) = 1 \tag{5.13}$$

Let the range in which the approximation is valid be $[-1,1]$ and assume a solution $p(x) = \frac{1}{2}a_0 + a_1 T_1(x) + a_2 T_2(x) + a_3 T_3(x)$ (5.14)

where $T_i(x)$ are Chebyshev polynomials.

Equation (5.13) is first integrated and then (5.14) is substituted for $y(x)$

$$\text{i.e.} \quad (3 + 2x)y - 3\int y\,dx = \text{const} \tag{5.15}$$

Now $\qquad 2xT_j = T_{j+1} + T_{j-1} \qquad j > 0 \quad \text{and } 2xT_0 = 2T_1$

and

$$\int a_j T_j dx = \frac{a_j}{2}\left[\frac{T_{j+1}}{j+1} - \frac{T_{j-1}}{j-1}\right], \qquad j \geqslant 2$$

$$= \frac{a_1}{4}\left[T_2 + T_0\right] \qquad j = 1$$

$$= a_0 T_1 \qquad j = 0$$

We can substitute (5.14) into (5.15) and using the above relationships, we obtain

$$\left(\frac{3a_0}{2} + \frac{a_1}{4}\right) + \left(-\frac{1}{2}a_0 + 3a_1 + \frac{5}{2}a_2\right)T_1(x) + \left(\frac{1}{4}a_1 + 3a_2 + \frac{7}{4}a_3\right)T_2(x)$$

$$+ \left(\frac{1}{2}a_2 + 3a_3\right)T_3(x) + \frac{5}{8}a_3 T_4(x) = \text{const.} \tag{5.16}$$

and the initial condition gives

$$\frac{1}{2}a_0 - a_2 = 1$$

Now all the conditions in (5.16) cannot be satisfied, so we choose to satisfy the initial condition and make the coefficients of T_1, T_2 and T_3 zero

$$\text{i.e.} \qquad \begin{aligned} \tfrac{1}{2}a_0 \qquad\qquad - a_2 \qquad &= 1 \\ -\tfrac{1}{2}a_0 + 3a_1 + \tfrac{5}{2}a_2 \qquad &= 0 \\ \tfrac{1}{4}a_1 + 3a_2 + \tfrac{7}{4}a_3 &= 0 \\ \tfrac{1}{2}a_2 + 3a_3 &= 0 \end{aligned} \tag{5.17}$$

Solving, we obtain

$$y(x) \simeq 0.967\,741\,94 + 0.349\,462\,36\,T_1(x) - 0.032\,258\,06\,T_2(x)$$
$$+ 0.005\,376\,35\,T_3(x) \tag{5.18}$$

Estimation of Error

In (5.16) we can satisfy all conditions (apart from the constant of integration) by introducing a term $\tau_4 T_4(x)$ on the right-hand side.[6]

we then have in addition to (5.17)

$$\frac{5}{8}a_3 = \tau_4 \quad \text{or} \quad \tau_4 = 0.003\ 360\ 22$$

Hence (5.14) satisfies exactly the system

$$(3 + 2x)p(x) - 3\int p(x)dx = \text{const.} + 0.003\ 360\ 22T_4(x) \qquad (5.19)$$

If $p(x) = y(x) + e(x)$

we find that substitution in (5.19) leads to $e(x)$ satisfying

$$(3 + 2x)\ e(x) - 3\int e(x)dx = 0.003\ 360\ 22T_4(x) \qquad (5.20)$$

with initial condition $e(o) = 0$

This equation has to be solved iteratively, in the form

$$(3 + 2x)e^{(r+1)}(x) = 0.0003\ 360\ 22T_4(x) + \alpha r + 3\int_0^x e^{(r)}(x)dx \qquad (5.21)$$

The αr is chosen to satisfy the initial condition and the integral is estimated by the trapezium rule.

Choosing $e^{(o)} = 0$

$$(3 + 2x)\ e^{(1)}(x) = 0.003\ 360\ 22\ T_4(x) + \alpha_0$$

when $x = 0$ $\qquad e^{(1)}(o) = \frac{1}{3}\left[.003\ 360\ 22 + \alpha_0\right]$

if $e^{(1)}(o) = 0 \qquad \alpha_0 = -0.003\ 360\ 22$

$$\therefore \quad (3 + 2x)\ e^{(1)}(x) = 0.003\ 360\ 22\left[T_4(x) - 1\right]$$

Then equation (5.21) can be used to find $e^{(2)}(x)$ and so on.

The results of the iterative process are tabulated below.

x	-1.0	-0.8	-0.6	-0.4	-0.2	0.0	0.2	0.4	0.6	0.8	1.0	
$e^{(1)}$	0	-4.42	-3.44	-1.64	-0.40	0	-0.30	-0.95	-1.47	-1.34	0	$\times 10^{-3}$
$e^{(2)}$	5.94	-1.12	-2.19	-1.31	-0.34	0	-0.32	-1.06	-1.76	-1.79	-0.49	$\times 10^{-3}$
$e^{(3)}$	1.20	-2.54	-2.52	-1.37	-0.36	0	-0.33	-1.09	-1.80	-1.88	-0.63	$\times 10^{-3}$
$e^{(4)}$	3.702	-2.061	-2.422	-1.358	-0.356	0	-0.332	-1.089	-1.867	-1.888	-0.650	$\times 10^{-3}$
actual error	3.495	-2.098	-2.468	-1.535	-0.369	0	-0.323	-1.076	-1.797	-1.891	-0.671	$\times 10^{-3}$

The actual error was obtained by comparing the series approximation with the exact solution of (5.15) i.e.

$$y = \sqrt{1 + \frac{2}{3}x}$$

Estimation of Error Using Neglected Coefficients

In Appendix A5.2 the method described by J. Oliver [18] is outlined for expressing the error in the solution to (5.15) in terms of the first few neglected coefficients. In our example, we wish to find

$$e(x) = \varepsilon_4(x)a_4 + \varepsilon_5(x)a_5 + \varepsilon_6(x)a_6$$

where $$\varepsilon_4(x) = \sum_{j=0}^{3}{}' \alpha_j^{(4)} T_j(x) - T_4(x)$$

$$\varepsilon_5(x) = \sum_{j=0}^{3} \alpha_j^{(5)} T_j(x) - T_5(x) \qquad \text{etc.}$$

It is shown in Appendix A5.2 that the $\alpha_j(i)$ satisfy equations with the same coefficients on the left-hand side as in (5.17) and different values on the right-hand side.

In the example above, we have

$$[A] \cdot \begin{bmatrix} \alpha_0^{(4)} \\ \alpha_1^{(4)} \\ \alpha_2^{(4)} \\ \alpha_3^{(4)} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \frac{3}{2} \end{bmatrix}$$

$$[A] \cdot \begin{bmatrix} \alpha_0^{(5)} \\ \alpha_1^{(5)} \\ \alpha_2^{(5)} \\ \alpha_3^{(5)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$[A] \cdot \begin{bmatrix} \alpha_0^{(6)} \\ \alpha_1^{(6)} \\ \alpha_2^{(6)} \\ \alpha_3^{(6)} \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

where $$[A] = \begin{bmatrix} \frac{1}{2} & 0 & -1 & 0 \\ -\frac{1}{2} & 3 & 5/2 & 0 \\ 0 & \frac{1}{4} & 3 & 7/4 \\ 0 & 0 & \frac{1}{2} & 3 \end{bmatrix}$$

from which we obtain

$\alpha_0^{(4)} = 1.258\ 06$

$\alpha_1^{(4)} = 0.518\ 82$

$\alpha_2^{(4)} = 0.370\ 97$

$\alpha_3^{(4)} = 0.561\ 83$

$\alpha^5 = 0$

$j = 0,1,2,3$

$\alpha_0^{(6)} = -1.935\ 48$

$\alpha_1^{(6)} = -0.341\ 13$

$\alpha_2^{(6)} = 0.032\ 26$

$\alpha_3^{(6)} = -0.005\ 38$

Since $e(x) \simeq a_4 \varepsilon_4(x) + a_5 \varepsilon_5(x) + a_6 \varepsilon_6(x)$

$e(x)$ will depend on reliable estimates being available for a_4, a_5 and a_6

In the figures tabulated below, the values used were

$$a_4 = -0.001\ 44$$
$$a_5 = 0.000\ 32$$
$$a_6 = -0.000\ 07$$

x	-1.0	-0.8	-0.6	-0.4	-0.2	0.0	0.2	0.4	0.6	0.8	1.0	
$a_4 \varepsilon_4(x)$	2.63	-1.66	-2.58	-1.84	-0.71	0	-0.09	-0.91	-1.96	-2.28	-0.49	$\times 10^{-3}$
$a_5 \varepsilon_5(x)$	0.32	-0.32	-0.02	0.28	0.27	0	-0.27	-0.28	0.02	0.32	-0.32	$\times 10^{-3}$
$a_6 \varepsilon_6(x)$	0.11	-0.004	0.11	0.11	0.04	0	0.05	0.13	0.14	0.03	0.16	$\times 10^{-3}$
$e(x)$	3.06	-1.98	-2.49	-1.45	-0.40	0	-0.31	-1.06	-1.80	-1.93	-0.65	$\times 10^{-3}$
actual error	3.50	-2.10	-2.47	-1.54	-0.37	0	-0.32	-1.08	-1.80	-1.89	-0.67	$\times 10^{-3}$

The error curves produced by the two methods of estimation are compared with the true error in fig. 5.3. In this case, both methods are shown to be reliable. One point is noticeable about the error, that is that it is far from symmetrically distributed about the axis, due to the initial condition producing zero error at the origin. This suggests that if the condition at the origin is relaxed a better error distribution may be achieved.[6]

<u>The Perturbed Condition</u>

Consider a solution of (5.15) of the form

$$p(x) = \tfrac{1}{2}a_0 + \sum_{j=1}^{4} a_j T_j(x)$$

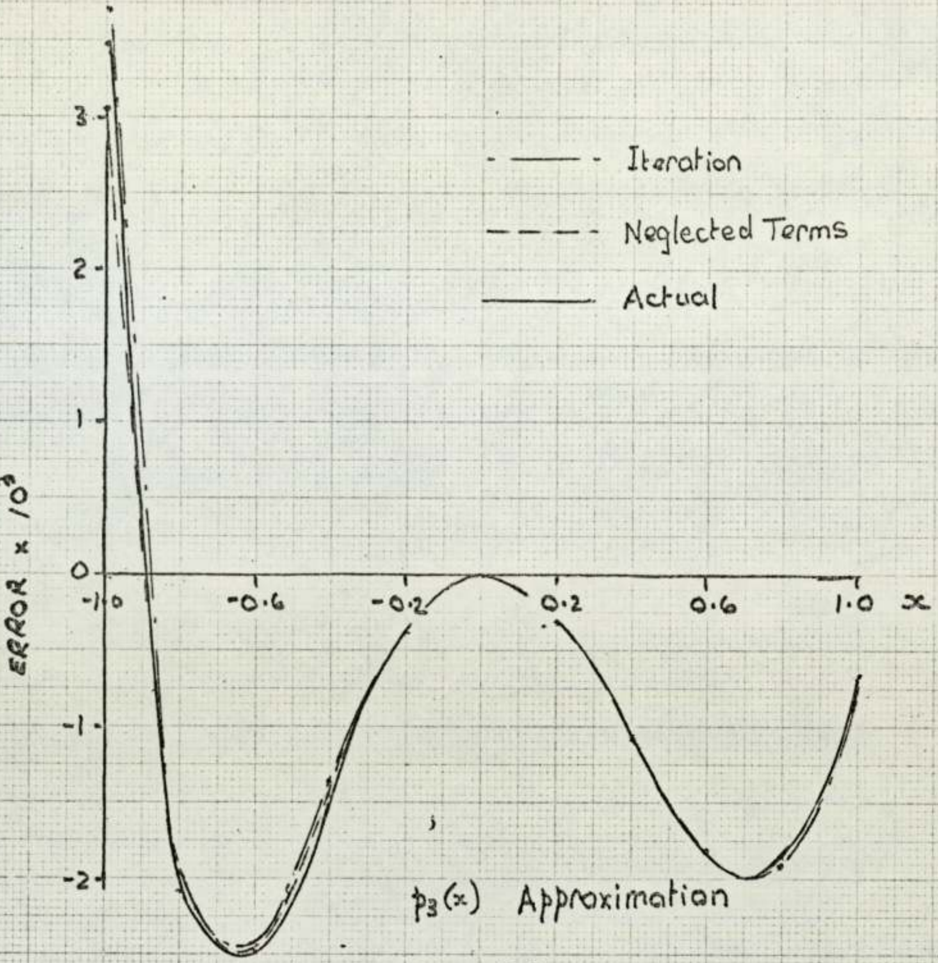
Introducing a term $\tau_6 T_6(x)$, we can solve the system

$$\begin{array}{rrrrrl}
\tfrac{1}{2}a_0 & & -a_2 & & +a_4 & = 1 \\
-\tfrac{1}{2}a_0 & +3a_1 & +\tfrac{5}{2}a_2 & & & = 0 \\
 & \tfrac{1}{4}a_1 & +3a_2 & +\tfrac{7}{4}a_3 & & = 0 \\
 & & \tfrac{1}{2}a_2 & +3a_3 & +\tfrac{3}{2}a_4 & = 0 \\
 & & & & \tfrac{7}{10}a_4 & = \tau_5
\end{array}$$

the original equation now becomes

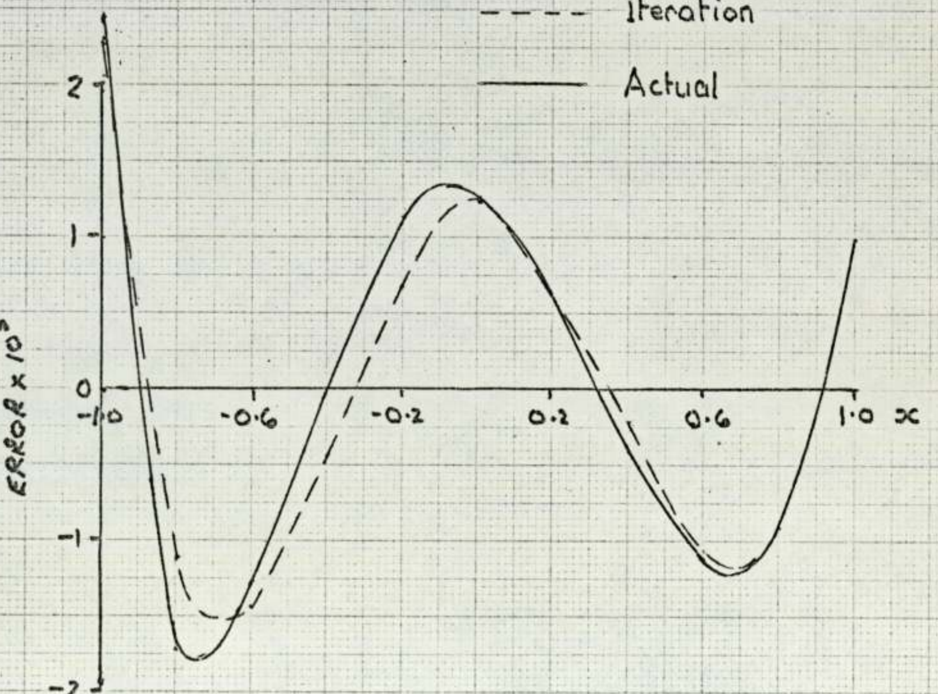
$$(3 + 2x)\ p(x) - 3\int p(x)\,dx \quad = \quad \text{const} + \tau_5 T_5(x) \tag{5.22}$$

and $p(x)$ is an exact solution of (5.22)



$p_3(x)$ Approximation

Fig 5.3



$\bar{p}_2(x)$ Approximation

Fig 5.4

Solving for the coefficients we get

$$a_0 = 1.937\ 076 \qquad a_2 = -0.032\ 729 \qquad a_4 = -0.001\ 269$$

$$a_1 + 0.350\ 120 \qquad a_3 = \ 0.006\ 089 \qquad \gamma_5 = 0.7a_4$$

We now take $\overline{p(x)}$ as the first four terms of $p(x)$

i.e. $(3 + 2x)\ (\overline{p(x)} + a_4T_4(x)) - 3\int(\overline{p(x)} + a_4T_4(x))\ dx = \text{const} + \gamma_5 T_5(x)$

or $(3 + 2x)\ \overline{p(x)} - 3\int \overline{p(x)}dx = \text{const}\, \gamma_5 T_5(x) - (3 + 2x)a_4T_4(x) + 3\int a_4 T_4(x)dx$ (5.23)

but $\quad (3 + 2x)y(x) - 3\int y(x)dx = \text{const}$

$\therefore$ if $\quad \overline{e}(x) = \overline{p}(x) - y(x)$, $e(x)$ satisfies the equation

$$(3+2x)\ \overline{e(x)} - 3\int\overline{e(x)}dx = \frac{7}{10}\ a_4T_5(x) - (3 + 2x)\ a_4T_4(x) + 3\int a_4T_4(x)\ dx \tag{5.24}$$

and if $p(x)$ satisfies the initial condition

$$\text{i.e. } p(o) = 1$$

$$\text{then } \overline{p(o)} + a_4 = 1$$

$$\text{also } \quad y(o) = 1$$

$$\therefore \quad \overline{e(o)} = -a_4 \tag{5.25}$$

As before (5.24) is solved iteratively in the form

$$(3 + 2x)\ \overset{(r)}{\overline{e}(x)} = a_4\left[\frac{7}{10}T_5(x) - (3+ 2x)T_4(x) + 3\int T_4(x)dx\right] + \alpha_{r-1} + 3\int \overset{(r-1)}{\overline{e}(x)}dx \tag{5.26}$$

with $\alpha_{(r-1)}$ chosen at each stage to satisfy (5.25)

Now set $\quad \overset{(o)}{\overline{e}(x)} = 0$

$$(3 + 2x)\overline{e}^{(1)} = a_4\left[\frac{7}{10}\ T_5(x) - (3 + 2x)T_4(x) + 3\int_0^x T_4(x)dx\right] + \alpha_0$$

when $x = 0 \quad \overline{e}(o) = -a_4 \qquad \therefore \alpha_0 = 0$

then $(3 + 2x)\overline{e}^{(2)} = a_4\left[\frac{7}{10}T_5(x) - (3 + 2x)T_4(x) + 3\int_0^x T_4(x)dx\right] + \alpha_1 + 3\int_0^x \overline{e}^{(1)}(x)\ dx$

whence $\alpha_1 = 0$ as before and $\overline{e}^{(2)}$ is obtained.

We tabulate below the results of the iteration

x	-1.0	-0.8	-0.6	-0.4	-0.2	0.0	0.2	0.4	0.6	0.8	1.0
$\bar{e}^{(1)}$	1.88	-1.62	-1.40	-0.14	0.92	1.27	0.47	-0.34	-1.18	-0.85	1.04 × 10^{-3}
$\bar{e}^{(2)}$	2.28	-1.28	-1.64	-0.14	0.67	1.27	0.61	-0.19	-1.15	-0.96	0.82 × 10^{-3}
$\bar{e}^{(3)}$	2.25	-1.14	-1.52	-0.47	0.70	1.27	0.64	-0.16	-1.11	-0.93	0.96 × 10^{-3}
$\bar{e}^{(4)}$	2.29	-1.09	-1.44	-0.44	0.69	1.27	0.64	-0.15	-1.10	-0.91	0.98 × 10^{-3}
actual error	2.45	-1.71	-1.27	-0.009	1.13	1.27	0.63	-0.37	-1.14	-0.95	1.02 × 10^{-3}

The estimated error curve is compared with the true error curve in fig 5.4.

Conclusion

The solution of this chosen problem has been satisfactory using a Chebyshev series with only very few terms. The two methods of estimating the error both gave reasonable estimates, although that involving the use of the neglected coefficients does require some accurate estimation of the unknown terms. The iterative scheme requires by far the most computation, but all the terms are known and no estimates are required. In this sense it is the more reliable method.

The error curve produced by perturbing the initial condition shows an improvement on that of the unperturbed solution and suggests that the extra degree of freedom afforded by this approach is of real benefit.

CHAPTER VI

Approximation in the L_1 Norm

Introduction

The problem considered here is finding an approximating function of the form

$$f(a,x) = \sum_{i=0}^{n} a_i \phi_i(x) \quad \text{in an interval } [x_1, x_2] \tag{6.1}$$

such that the integral

$$L_1(a) = \int_{x_1}^{x_2} |y(x) - f(a,x)| dx \tag{6.2}$$

is as small as possible.

We are dealing only with problems where $y(x)$ is continuous and $\phi_i(x)$ are chosen to have polynomial form. It is shown that the problem can be considered as one of interpolation which is sufficient in many cases to produce the best approximation. A programme is developed which uses this method when the $\phi_i(x)$ are chosen to be Chebyshev polynomials of the second kind.

The Condition for Best Approximation

It is necessary to find the condition which characterizes the best approximation in the L_1 sense. This requires the value of $L_1(a)$, defined in (6.2) to be a minimum. Now $L_1(a)$ can be considered as a function of the coefficients a_i of the approximation, so if we define a^* as the required optimum point, we require (see Rice [13])

$$\lim_{t \to 0} \left\{ \frac{L_1(a^* + ta) - L_1(a^*)}{t} \right\} = 0 \tag{6.3}$$

i.e., the derivatives of $L_1(a)$ at a^* are to be zero.

It is shown in Appendix A6.1 that (6.3) leads to the condition

$$\int_{x_1}^{x_2} \phi_i(x) \, \text{sign}\{y(x) - f(a^*,x)\} dx = 0 \qquad i=0,1, \dots n \tag{6.4}$$

where $\text{sign}(z) = \begin{cases} -1 & \text{if } z < 0 \\ 0 & \text{if } z = 0 \\ +1 & \text{if } z > 0 \end{cases}$

Now assume that it is possible to find a sign function $s(x)$, which takes one of the values ± 1, and which can be chosen so that

$$\int_{x_1}^{x_2} \phi_i(x)s(x)dx = 0 \qquad i = 0,1, \ldots\ldots n \tag{6.5}$$

It is possible to show that $s(x)$ must change sign at not less than $(n+1)$ points in $[x_1 , x_2]$, if the functions $\phi_i(x)$ form a Chebyshev set.

A chebyshev set may be defined as a sequence of functions $\phi_i(x)$, continuous over the chosen range, such that no linear combination $\sum_{j=0}^{n} \lambda_j\phi_j(x)$ has more than n roots inside $[x_1 ,x_2]$ unless it vanishes identically.

Assume that the chosen functions form a Chebyshev set. (This is certainly true in this Chapter.) If $s(x)$ has only n changes of sign, we can choose $f(a,x) = \sum_{j=0}^{n} \lambda_j\phi_j(x)$ to change sign at these points and no other, other. Then $f(a,x)s(x)$ will have a fixed sign throughout the range

$\therefore \int_{x_1}^{x_2} f(a,x)s(x)dx \neq 0$ and (6.5) cannot be true for all i, since we assume $f(a,x) = \sum_{j=0}^{n} \lambda_j\phi_j(x)$.

So $s(x)$ must have at least $(n+1)$ changes of sign. Now if we can find $s(x)$, we see from (6.4) and (6.5) that $f(a^*,x)$ could be determined from

$$\text{sign}\{y(x) - f(a^*,x)\} = s(x) \tag{6.6}$$

One way of satisfying (6.6) is to find the points x_k at which $s(x)$ changes sign in order to satisfy (6.5) and then to solve the interpolation problem.

$$y(x_k) = f(a,x_k) \tag{6.7}$$

If there are exactly $(n+1)$ points, x_k, then (6.7) determines $f(a,x)$ uniquely. Then if $y(x)$ and $f(a,x)$ do not agree at any other points in $[x_1 , x_2]$ we have found the required solution to (6.4). However, if the error curve has more than $(n+1)$ zeros, the solution to (6.7) will not give the required solution to (6.4). In this case, a descent method must be employed to adjust the coefficients to reduce the components of the derivative to zero (see Usow [17]) In this work, we consider only those cases where the L_1 problem is solved by interpolation. It is necessary to show that if the $\phi_i(x)$ are chosen to be polynomials, then the choice of interpolation points is fixed irrespective of the nature of the function y(x).

Choice of Interpolation Points

We consider the case where

$$\phi_r(x) = \sum_{j=0}^{r} b_j x^j \qquad \text{and } f(a,x) = \sum_{i=0}^{n} a_i \phi_i(x) \qquad (6.8)$$

It is necessary to find a set of points $[x_k]$ such that a sign function s(x) changes sign (n+1) times in the chosen interval $[x_1, x_2]$.

The range of approximation will be taken to be $[-1,1]$ and we consider the integral

$$I = \int_{-1}^{1} \sum_{j=0}^{r} b_j x^j s(x) dx \qquad r = 0,1 \ldots\ldots n \qquad (6.9)$$

By making the substitution $x = \cos\theta$, it is shown in Appendix A6.2 that if I is to be zero for all values of r,

then $s(\cos\theta) = \text{sign}\left[\sin(n+2)\theta\right]$

and that the points of interpolation for the functions defined in (6.8) are given by

$$x_k = \cos\left(\frac{k\pi}{n+2}\right) \qquad k = 1,2, \ldots..(n+1) \qquad (6.10)$$

which are the zeros of $U_{n+1}(x)$, the Chebyshev polynomial of the second kind.

Choice of Interpolating Functions

Experience with power series approximations suggests that the choice of x^j for the $\phi_j(x)$ may not be a good choice in terms of numerical stability. In the paragraph above, we see that Chebyshev polynomials have arisen naturally in the discussion. These polynomials can be integrated readily. Series of Chebyshev terms usually display repidly decreasing coefficients and can be truncated at an early stage without great loss of accuracy. In addition they are relatively easily summed using the appropriate recurrence relation.

Furthermore, if we use a finite series approximation

$$f(a,x) = \sum_{j=0}^{n} a_j U_j(x)$$

then $L_1(a) = \int_{-1}^{1} |y(x) - f(a,x)| \, dx \simeq \int_{-1}^{1} |a_{n+1} U_{n+1}(x)| \, dx$ (6.11)

[An expansion of the form $y(x) = \sum_{j=0}^{\infty} a_j U_j(x)$ is assumed.]

Now it can be shown (Todd [16] page 149) that the minimum value of $\int_{-1}^{1} |\overline{p_n}(x)| \, dx$ over all polynomials $\overline{p_n}(x)$ of degree n with leading coefficient unity is 2^{1-n} and is achieved when $\overline{p_n}(x) = \overline{U_n}(x)$ where $\overline{U_n}(x)$ is $U_n(x)$ divided by a suitable constant to give a unit leading coefficient.

In view of this and (6.11) it seems appropriate that an approximation based on a series of terms involving $U_j(x)$ should give a useful method of expressing the best L_1 form. In Appendix A6.3 is given details of the programme which attempts to find the best L_1 approximation by interpolation when the approximating function is a series of Chebyshev polynomials of the second kind.

It has been pointed out that the interpolation procedure does not provide the best L_1 approximation in all cases. No attempt is made in the programme to implement a descent method when interpolation fails to give the desired answer. It is felt that the L_1 approximation would have to be shown to possess some distinct advantage over other norms before the extra work involved in solving the optimization problem could be justified.

The Chapter is concluded with some examples of the use of the programme to obtain approximating functions. The last example is taken from Usow [17] and illustrates an example in which the interpolation method fails for certain values of n.

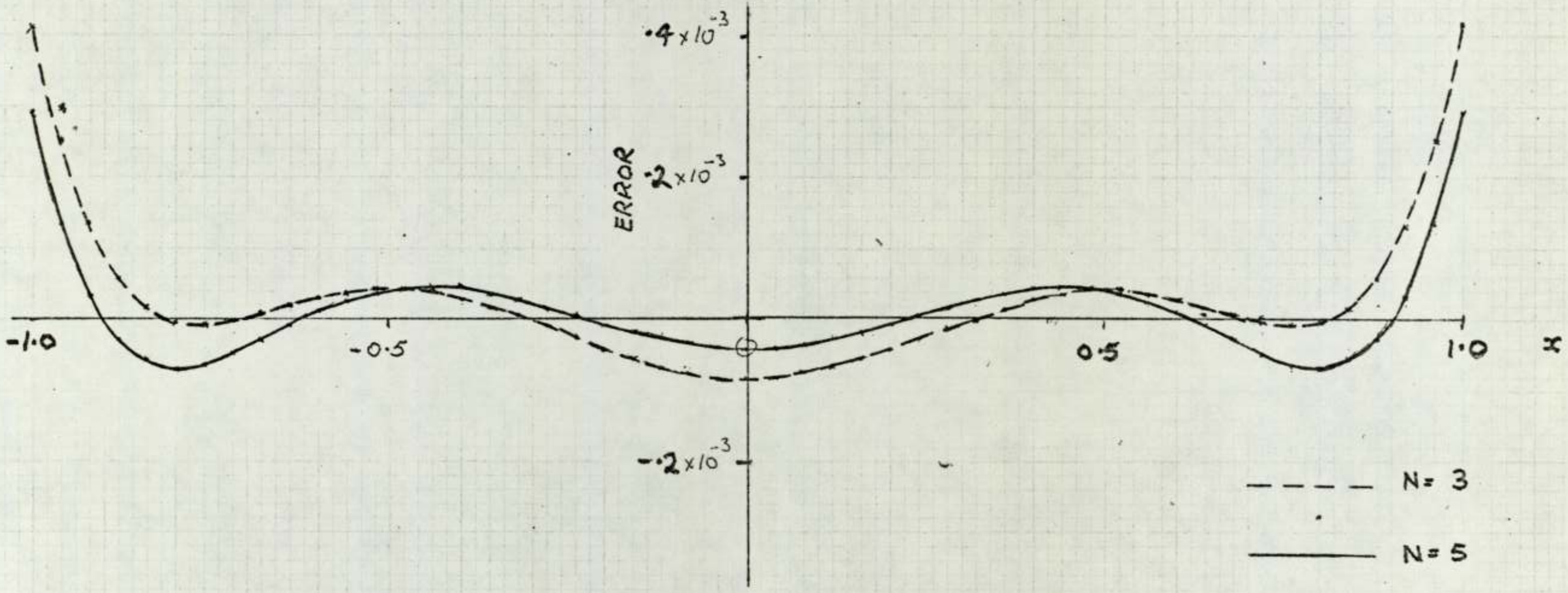
Example 1

Consider $y(x) = 0.92 \cosh x - \cos x$ and an approximation

$$f(x) = \sum_{i=0}^{3} a_i U_i(x)$$

In this case, the interpolation points are ± 0.3090, ± 0.8090 and the approximation is $f_3(x) = 0.15979\, U_0(x) + 0.23971\, U_2(x)$ the other coefficients being zero.

The number of zeros of the error curve turns out to be six, so the approximation is not best in the L_1 sense. For this reason, we now try an



Approximation to 0·92 cosh x - cos x

Fig 6·1

approximation of degree five and again solve the interpolation problem. The interpolation points are now ± 0.90097, ± 0.62349, ± 0.22252 giving $f_5(x) = 0.15979\ U_0(x) + 0.23975\ U_2(x) - 0.113 \times 10^{-6} U_4(x)$ and the error curve turns out to have the required six zeros. The two error curves are compared in figure 6.1. The values of the L_1 integral are

$$\int_{-1}^{1} |y(x) - f_3(x)|\ dx = 0.1058 \times 10^{-3}$$

$$\int_{-1}^{1} |y(x) - f_5(x)|\ dx = 0.8394 \times 10^{-4}$$

In the case of $f_5(x)$, the zeros of the error curve were located as being -0.90139, -0.62325, -0.22234, 0.22270, 0.62370, 0.90055.

This shows a maximum discrepancy with the given interpolation points of 0.00042.

As a further check, the integrals on the left-hand-side of equation (6.5) are calculated and found to give

[-0.000077, 0.00379, -0.00004, 0.00383, 0.00007, 0.00004]

These should strictly be zero if the approximation is optimum. Since the integrals are evaluated using the interpolated zeros of the error curve [Appendix A6.3], the discrepancy is reasonably accounted for by slight errors in positioning the zeros of $y(x) - f_5(x)$.

Example 2

$$\text{Consider}\quad y(t) = \begin{cases} e^{t} & 0 \leqslant t \leqslant 1 \\ e^{-t} - e^{-1} + e & 1 \leqslant t \leqslant 2 \\ e^{t} + e^{-2} - e^{-1} - e^{2} + e & 2 \leqslant t \leqslant 3 \end{cases}$$

The range is scaled to $[-1,1]$ by the transformation $x = \frac{2}{3}(t - \frac{3}{2})$. First, N is taken equal to 4 and interpolation points are taken as the zeros of $U_5(x)$.

Then $f_4(x) = 3.59554\ U_0(x) + 2.10396\ U_1(x) + 1.26684\ U_2(x)$
$+ 0.87025\ U_3(x) + 0.24483\ U_4(x)$

Interpolation points	-0.86603	-0.50000	0.0	0.50000	0.86603
Zeros of error curve	-0.86582	-0.50005	0.0	0.50000	0.86602

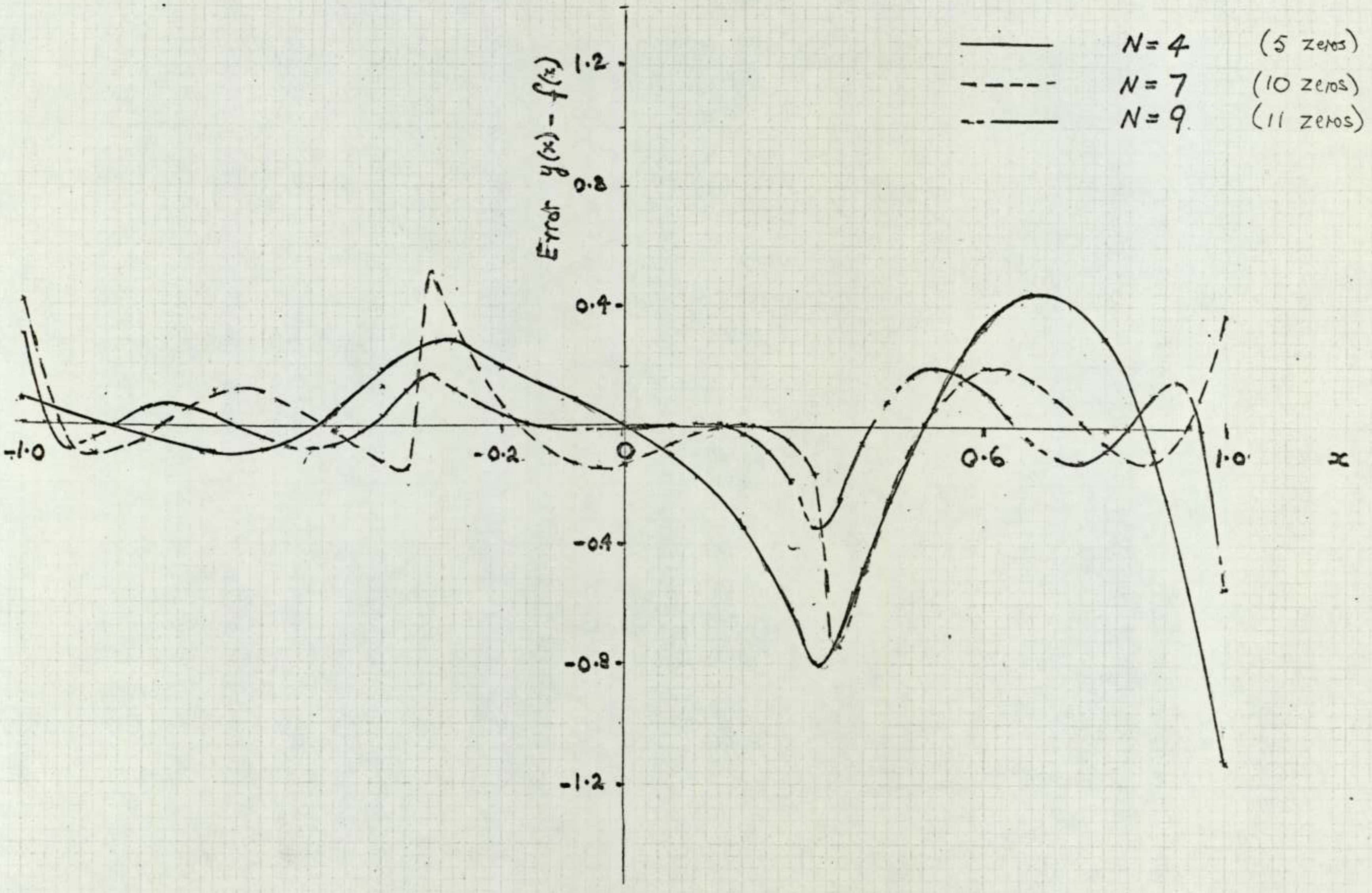

Error $y(x) - f(x)$
1·2
0·8
0·4
-0·4
-0·8
-1·2
-1·0
-0·2
0
0·6
1·0
x
$N = 4$ (5 zeros)
$N = 7$ (10 zeros)
$N = 9$ (11 zeros)

Fig 6.2.

The elements of the gradient vector are

$$[0.00041, \ -0.00094, \quad 0.00083, \quad -0.00053, \quad 0.00043]$$

$$\text{and} \int_{-1}^{1} |y(x) - f_4(x)| dx = 0.46088$$

The error curve for this approximation is plotted in figure 6.2

If now, N is taken as 7, we get

$$f_7(x) = 3.90317\ U_0(x) + 2.10546\ U_1(x) + 1.10182\ U_2(x) + 0.91162\ U_3(x) + $$
$$+ 0.38389\ U_4(x) - 0.17637\ U_5(x) - 0.16857\ U_6(x)$$
$$+ 0.03988\ U_7(x)$$

$$\text{and} \int_{-1}^{1} |y(x) - f_7(x)| dx = 0.25755$$

However, the error curve now turns out to have ten zeros instead of eight. It then appears that there might be a solution with N = 9, but in that case, eleven zeros are produced and the interpolation has not produced the best L_1 approximation. The approximation is

$$f_9(x) = 3.57331\ U_0(x) + 2.08090\ U_1(x) + 1.30196\ U_2(x) + 0.94899\ U_3(x)$$
$$+ 0.18620{}_4U\ (x) - 0.20406\ U_5(x) - 0.08562\ U_6(x) + 0.02599\ U_7(x)$$
$$+ 0.04525\ U_8(x) + 0.05775\ U_9(x)$$

$$\text{with} \int_{-1}^{1} |y(x) - f_9(x)| dx \ = \ 0.15487$$

The error curves for these two approximations are also plotted in figure 6.2. It may be noted that the discontinuities in the first derivatives at x = $\pm \frac{1}{3}$ [t = 1 or 2] causes the error curves to have quite sharp peaks at these points.

In conclusion, it is thought that the use of the L_1 norm to define an approximating function does not show any advantage over the more familiar L_2 or L_∞ norms. The idea of being able to express the problem as one of interpolation is attractive in its directness, but it does not always provide the best L_1 approximation. Expressing the approximation as a Chebyshev series provides a convenient computational form . It is difficult to appreciate the closeness of fit when the error norm is in the form of a definite integral. However, since computation of a fair number of points along the error curve is necessary for the numerical integration process, it is a simple matter to print out these points for reference purposes.

CHAPTER VII

Approximation in the L_2 Norm

Introduction

The concept of approximation in the L_2 norm (or least-squares approximation) was introduced in Chapter II. Some examples of this approach are now presented. A comparison is made of approximations expressed in series of Legendre and Chebyshev polynomials. Also the Chebyshev series obtained by using the orthogonality property over a discrete point set is compared with that where the coefficients are determined by integration. Reference is also made to the interpolation polynomial where the interpolation points are the zeros of a Chebyshev polynomial of suitable degree.

Orthogonal Polynomials

If we express the approximations as the sum of a finite number of orthogonal polynomials ϕ_i:

$$\text{i.e.} \quad f(x) = \sum_{i=0}^{n} a_i \phi_i(x) \tag{7.1}$$

the L_2 norm required that the integral

$$L_2 = \left[\int_a^b w(x)(y(x) - f(x))^2 dx\right]^{1/2} \tag{7.2}$$

shall be a minimum.

[Normally, the square root may be dispensed with, since the minimum of $(L_2)^2$ implies the minimum of L_2].

In Chapter II, it was seen that (7.2) leads to the expression for the coefficients in (7.1) given by

$$a_r = \frac{\int_a^b w(x)y(x)\phi(x) \quad dx}{\int_a^b w(x)\phi_r^2 (x) \quad dx} \tag{7.3}$$

It has been shown that if the weight-function $w(x)$ is taken as $(1-x^2)^{-\frac{1}{2}}$ and the range as $[-1,1]$, then (7.3) defines the coefficients of a series of Chebyshev polynomials. Would it not be computationally more convenient if $w(x)$ were chosen as unity? It can be shown that a set of polynomials with the necessary orthogonality are the Legendre polynomials given by

$$P_r(x) = \frac{1}{2^r r!}\frac{d^r}{dx^r}(x^2 - 1)^r \qquad (7.4)$$

In Appendix A7.1 some of the main properties of the Legendre polynomials are derived. In particular, it is seen that the coefficients in (7.1) become

$$a_r = \frac{2r+1}{2}\int_{-1}^{1} y(x)P_r(x)dx \qquad (7.5)$$

and the least-square error expression is

$$S = \int_{-1}^{1} y^2(x)dx - \sum_{r=0}^{n} \frac{2a_r^2}{2r+1} \qquad (7.6)$$

These can be compared with the similar expressions for the Chebyshev series, i.e.

$$\text{if} \quad f(x) = \sum_{r=0}^{n}{}' a_r T_r(x) \qquad \text{in} \quad [-1,1]$$

then

$$a_r = \frac{2}{\pi}\int_{-1}^{1}\frac{y(x)T_r(x)\ dx}{\sqrt{1-x^2}} \qquad (7.7)$$

$$\text{and} \quad S = \int_{-1}^{1}\frac{y^2(x)}{\sqrt{1-x^2}}dx - \frac{\pi}{2}\sum_{r=0}^{n}{}' a_r^2$$

[The prime indicates that the coefficient in the first term of the summation must be halved.]

Since the two types of orthogonal function have analogous properties, it might naturally be asked why choose the one with the awkward weight function? A significant reason lies in the shape of the error curve that is obtained in each case.

If the coefficients of the series approximation decrease fairly rapidly, then the first neglected term is a good indication of the truncation error. That is $e(x) \simeq a_{n+1}\phi_{n+1}(x)$ where $\phi_{n+1}(x)$ would be either the Legendre polynomial $P_{n+1}(x)$ or the Chebyshev polynomial $T_{n+1}(x)$.

The main difference between the two curves is that the Legendre polynomial oscillates with increasing amplitude towards the ends of the range, whereas the Chebyshev polynomial oscillates with equal amplitude throughout. Consequently, we expect a Chebyshev series to give very nearly a minimax error curve. Add to this that for practical purposes, integration can be replaced by summation over a discrete point set, with unit weight function, then the reasons for the preference of Chebyshev expansions can be appreciated. The following example illustrates the difference between the two error functions.

Example

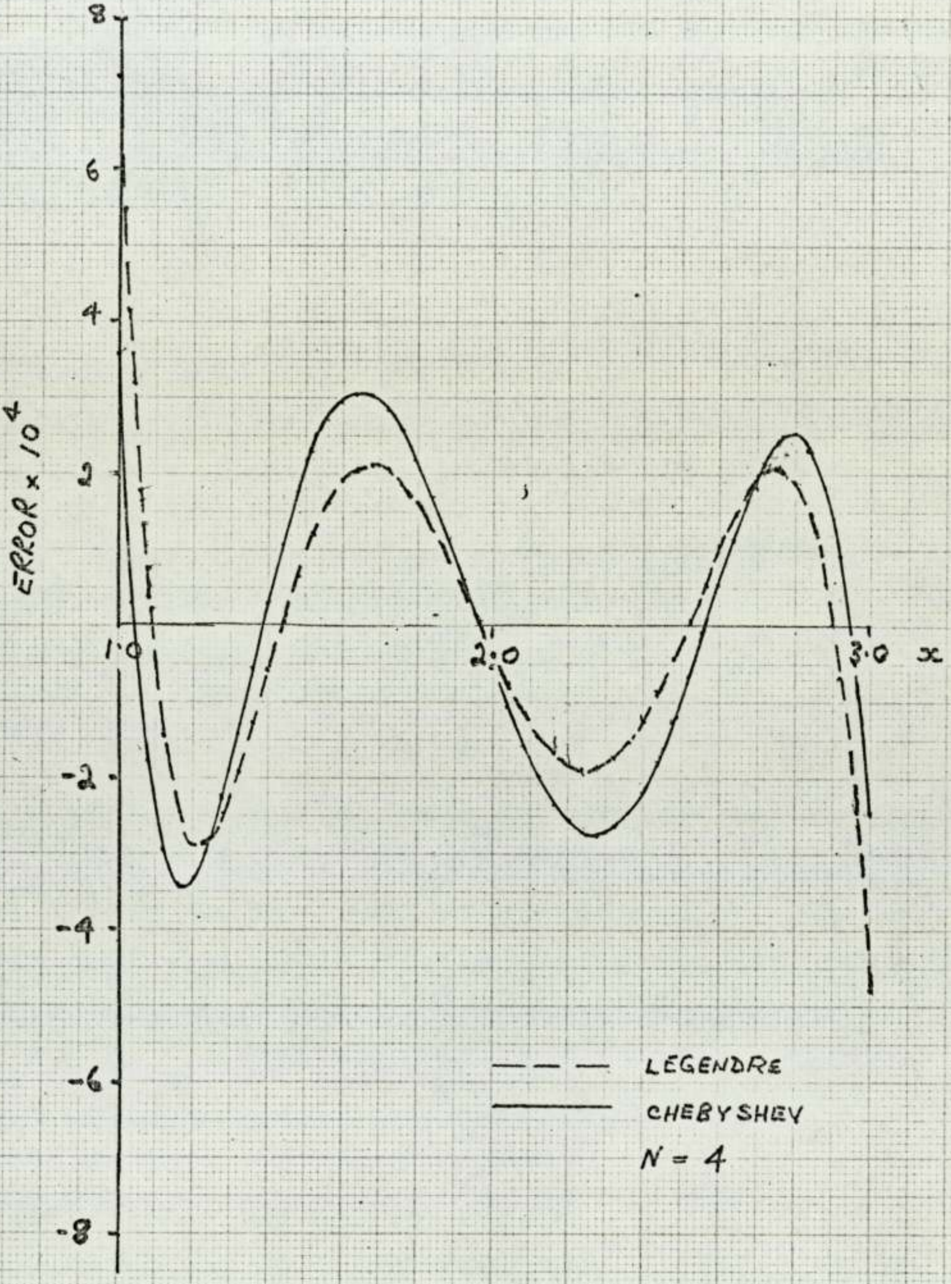
Let $y(x) = \sinh x \log_e(\tan h(\frac{x}{2}))$ and consider an approximation in the range $[1,3]$.

Making the transformation $z = x-2$ to make the range $[-1,1]$ and taking the degree of the approximation to be nine, we use (7.5) and (7.7) to obtain

$$\begin{aligned} f_L(z) = &-0.977\,539\,57\,P_0(z) - 0.036\,411\,50\,P_1(z) + 0.022\,185\,18\,P_2(z) \\ &-0.008\,565\,60\,P_3(z) + 0.002\,464\,54\,P_4(z) - 0.000\,580\,71\,P_5(z) \\ &+ 0.000\,120\,95\,P_6(z) - 0.000\,023\,94\,P_7(z) + 0.000\,004\,79\,P_8(z) \\ &- 0.000\,001\,00\,P_9(z) \end{aligned} \qquad (7.8)$$

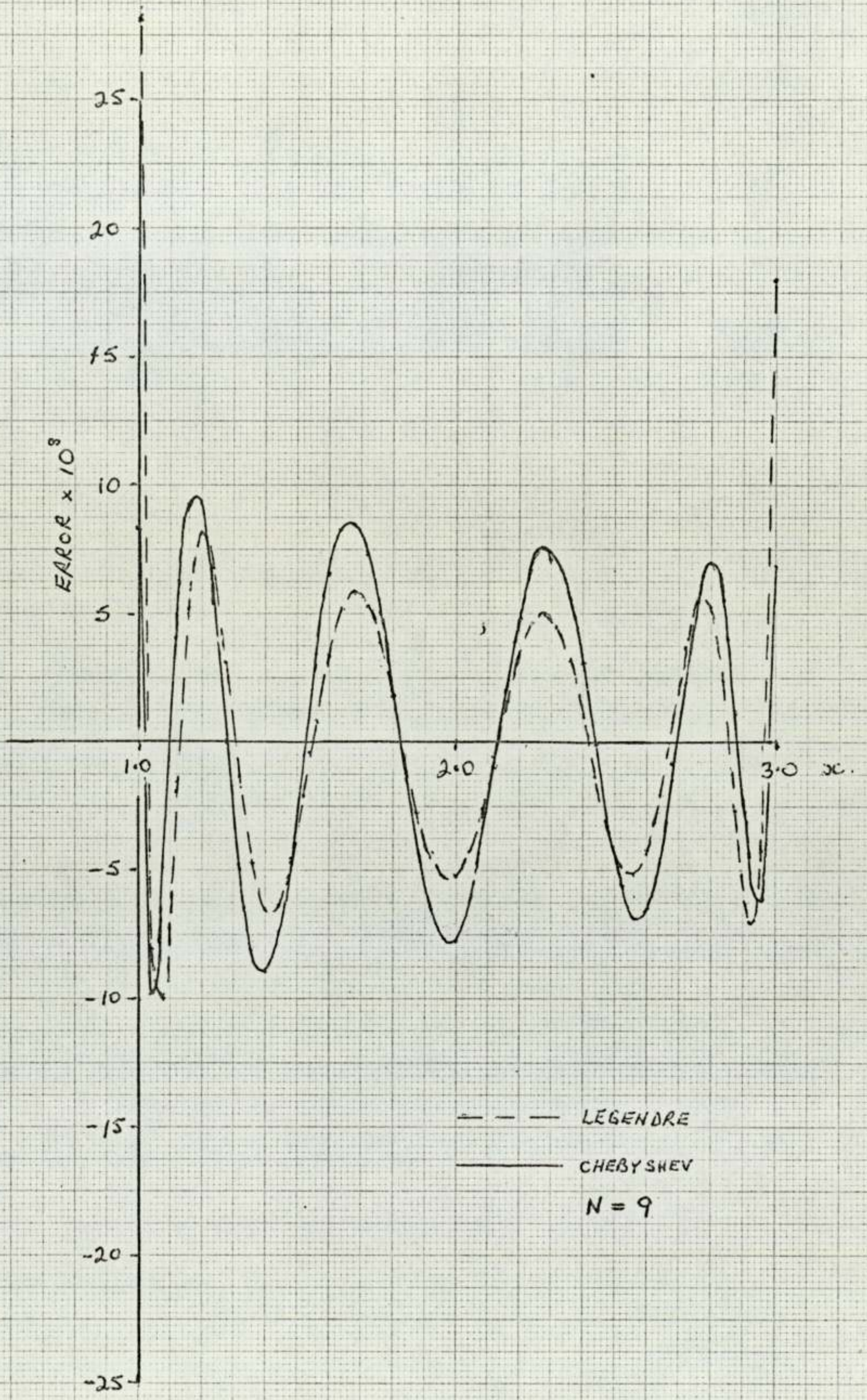
$$\begin{aligned} f_C(z) = &-0.971\,634\,52\,T_0(z) - 0.039\,763\,93\,T_1(z) + 0.017\,434\,62\,T_2(z) \\ &- 0.005\,516\,85\,T_3(z) + 0.001\,378\,40\,T_4(z) - 0.000\,291\,38\,T_5(z) \\ &+ 0.000\,055\,61\,T_6(z) - 0.000\,010\,23\,T_7(z) + 0.000\,001\,92\,T_8(z) \\ &- 0.000\,000\,38\,T_9(z) \end{aligned} \qquad (7.9)$$

It may be noted that due to the orthogonality properties of the polynomials, (7.8) and (7.9) provide approximations of lower degree simply by truncation at the appropriate point. Figures 7.1 and 7.2 show the errors for $f_L(x)$ and $f_C(x)$ for $n = 4$ and $n = 9$.



Approximation to $\sinh x \cdot \log_e(\tanh(\frac{x}{2}))$

Fig 7.1



Approximation to sinh x. $\log_e(\tanh(\frac{x}{2}))$

Fig 7.2

It was stated in Chapter II that there is a convenient computational method for the summation of a series of orthogonal functions. This is based on the property that they obey a three-term recurrence relation.

$$\text{i.e.} \quad \phi_n(x) = A_n\phi_{n-1}(x) + B_n\phi_{n-2}(x) \qquad (7.10)$$

It is shown in Appendix A7.2 that setting $b_{N+1} = 0$ and $b_N = a_N$ we form

$$b_k = a_k + A_{k+1}b_{k+1} + B_{k+2}b_{k+2}$$

$$\text{for } k = (N-1),(N-2) \; ...., \; 1$$

$$\text{Then} \quad \sum_{k=0}^{N} a_N\phi_N(x) = (a_0 + B_2b_2)\,\phi_0\,(x) + b_1\phi_1(x) \qquad (7.11)$$

and the summation is readily found without evaluating any of the polynomials apart from the trivial $\phi_1(x)$ and $\phi_0(x)$.

Indeed, for a series of Legendre polynomials, we have

$$A_{k+1} = \frac{2k+1}{k+1} \quad , \qquad B_{k+2} = -\frac{k+1}{k+2}$$

$$\text{with} \quad P_0(x) = 1 \qquad P_1(x) = x$$

and for a Chebyshev series

$$A_{k+1} = 2x \quad , \qquad B_{k+2} = -1$$

$$T_0(x) = 1 \qquad T_1(x) = x$$

This implies that the evaluation of series of orthogonal terms is no worse than ordinary polynomial evaluation in terms of the labour involved.

Determination of Coefficients by Summation

The coefficients of the Chebyshev series are defined by the integral in (7.7). Very rarely is it possible to formally integrate these expressions and some numerical technique must be employed, probably with the aid of a digital computer. If this is so, may it not be more convenient to use directly a method of summation based on the orthogonality of Chebyshev polynomials over a discrete point set?

i.e. $$\sum_{k=0}^{N}{}'' T_m(x_k)\, T_n(x_k) = \begin{cases} N & \text{if } m=n=0 \text{ or } m=n=N \\ \tfrac{1}{2}N & \text{if } m=n\neq 0 \text{ pr } N \\ 0 & \text{if } m\neq n \end{cases}$$

where $x_k = \frac{\cos \pi k}{N}$, $k = 0,1, \ldots .N$ and the double prime indicates that the first and last terms in the summation are halved.

$$\text{Let } y(x) = \sum_{s=0}^{\infty}{}' a_s\, T_s(x) \quad .$$

and consider $$b_r = \frac{2}{N} \sum_{k=0}^{N}{}'' y(x_k)\, T_r(x_k) \qquad (7.12)$$

then $$b_r = \frac{2}{N} \sum_{s=0}^{\infty}{}' a_s \sum_{k=0}^{N}{}'' T_s(x_k)\, T_r(x_k)$$

Now the orthogonality property dictates that the second summation is zero unless

$$T_s(x_k) = T_r(x_k) \qquad k = 0,1 \ldots\ldots N$$

i.e. $$\cos \frac{s\pi k}{N} = \cos \frac{r\pi k}{N}$$

whence $s = 2Np \pm r$ where $p = 0,1 \ldots.$

and $$b_r = a_r + a_{2N-r} + a_{2N+r} + a_{4N-r} + a_{4N+r} \quad \ldots..$$

A systematic procedure based on this formula can be found in Hayes

Hence, if N is sufficiently large and the coefficients of the series decrease reasonably quickly, b_r can be used as a very close approximation to a_r. From another point of view, if we replace x by $\cos\theta$ in (7.7) and use the trapezium rule over a set of equally-spaced points at intervals of π/N, we get exactly the equation (7.12). (Snyder, Chapter 3 [14])

Here is an example which compares the series obtained by evaluating the coefficients using (7.7) and (7.12)

Example

$$y(x) = \frac{x}{\left\{\sqrt{x^2+1} + x\right\}^3 \sqrt{x^2+1}}$$

Consider an approximation of degree nine in the range [-1,1.]

From (7.7) we obtain

$$\begin{aligned} f_1(x) = {} & -2.256\,975\,4\ T_0(x) + 4.000\,000\,0\ T_1(x) - 2.485\,803\,3\ T_2(x) \\ & + 0.999\,999\,98\ T_3(x) - 0.216\,797\,82\ T_4(x) - 0.000\,000\,02\ T_5(x) \\ & + 0.010\,793\,07\ T_6(x) - 0.000\,000\,02\ T_7(x) - 0.001\,081\,74\ T_8(x) \\ & - 0.000\,000\,02\ T_9(x) \end{aligned}$$

and from (7.12), using summation over 22 points,

$$\begin{aligned} f_2(x) = {} & -2.256\,975\,4\ T_0(x) + 4.000\,000\,0\ T_1(x) - 2.485\,803\,3\ T_2(x) \\ & -0.9995999\,98\ T_3(x) - 0.216\,797\,82\ T_4(x) - 0.000\,000\,02\ T_5(x) \\ & +0.010\,793\,07\ T_6(x) - 0.000\,000\,02\ T_7(x) - 0.001\,081\,74\ T_8(x) \\ & -0.000\,000\,02\ T_9(x) \end{aligned}$$

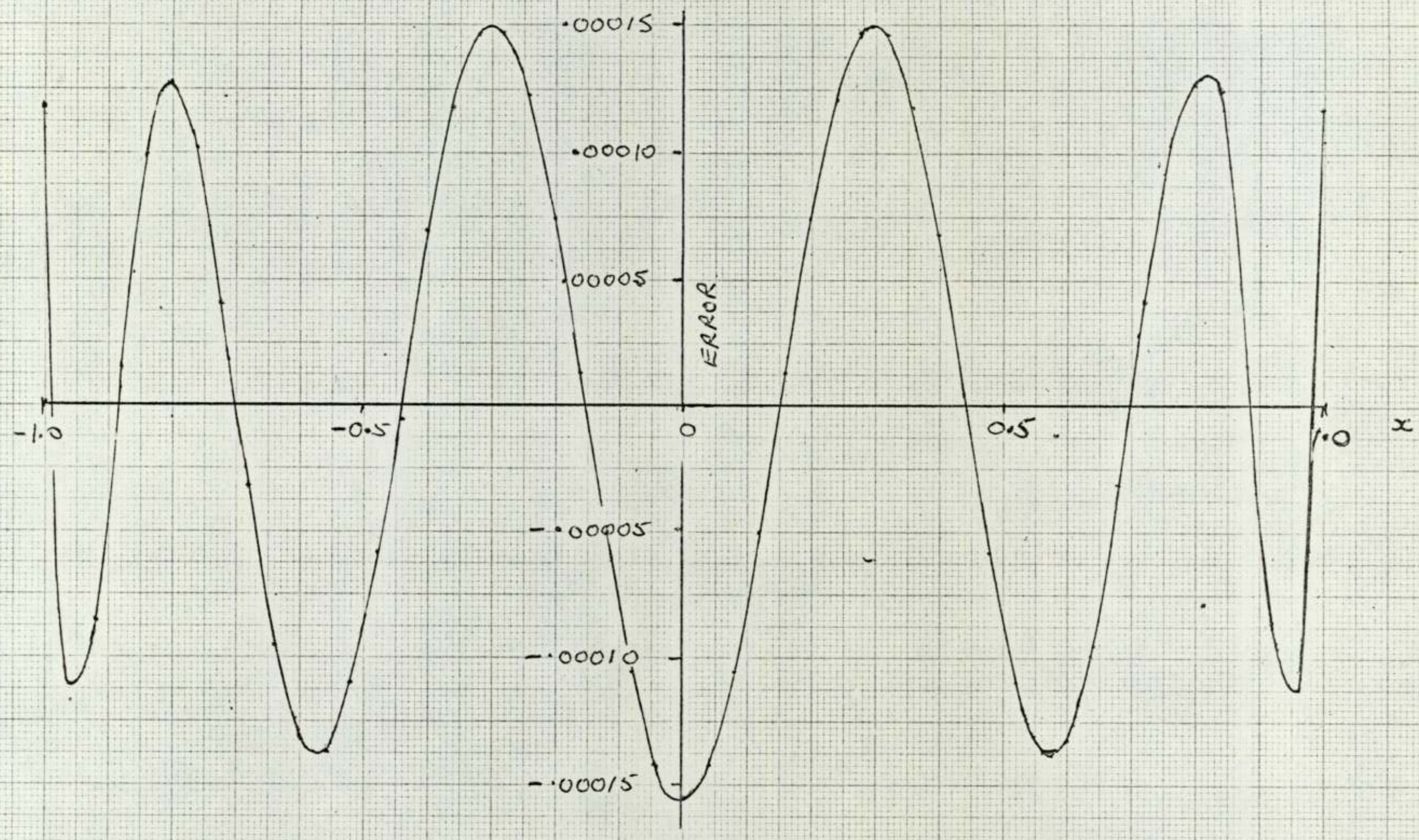
The error curve, which is the same in both cases to within 5×10^{-9} is shown in figure 7.3.

Interpolation Formula

By reference to figures 7.2 and 7.3 it can be seen that the number of zeros in the error curves correspond to the number of zeros of the first neglected polynomial term. It is of interest to consider the interpolation polynomial which takes as the interpolation points the zeros of the Chebyshev polynomial of appropriate degree. In Handscomb [7] it is shown that if $y(x)$ has no singularity on the real line $[a,b]$, then the interpolation formula with the above choice of points converges uniformly to $y(x)$.

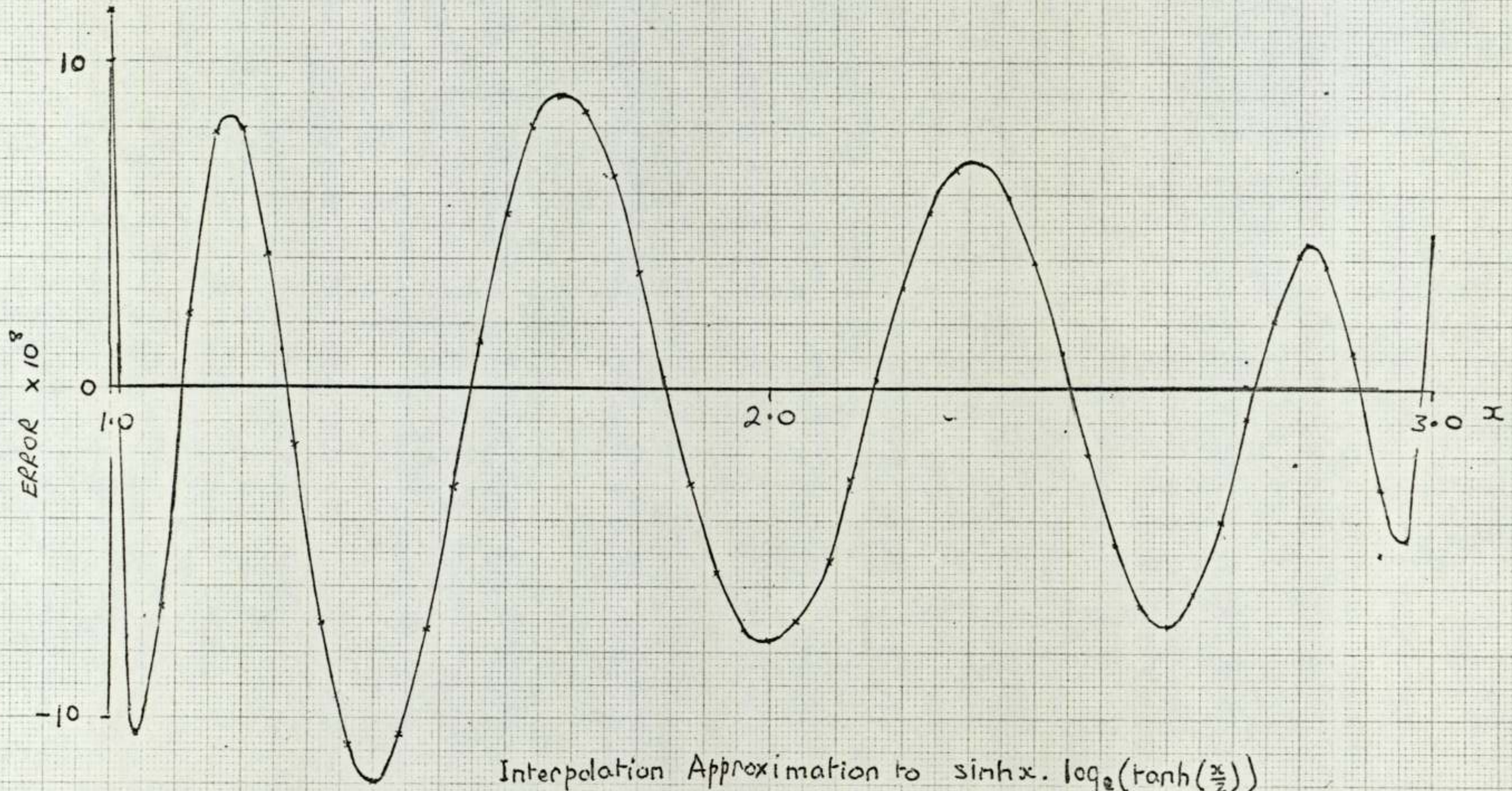
Figures 7.4 and 7.5 show the error curves when ninth-degree polynomials are used in approximating to the functions in the two previous examples. The points of agreement are taken as the zeros of the Chebyshev polynomial $T_{10}(x)$, (suitably transposed in the case with the range $[1,3]$).

It is seen that the interpolating function produces an error curve very similar to that of the Chebyshev series derived from the L_2 norm. In these examples, the Lagrangian interpolation formula was used, which since we are essentially using unequally-spaced points, may not be considered a convenient computational form. One method of overcoming this is to derive the interpolation function as a Chebyshev series (Hildebrand [10]). This



Approximation to $\dfrac{x}{\{\sqrt{x^2+1}+x\}^3\sqrt{x^2+1}}$

Fig 7.3



Interpolation Approximation to $\sinh x . \log_e(\tanh(\frac{x}{2}))$

Fig 7.4

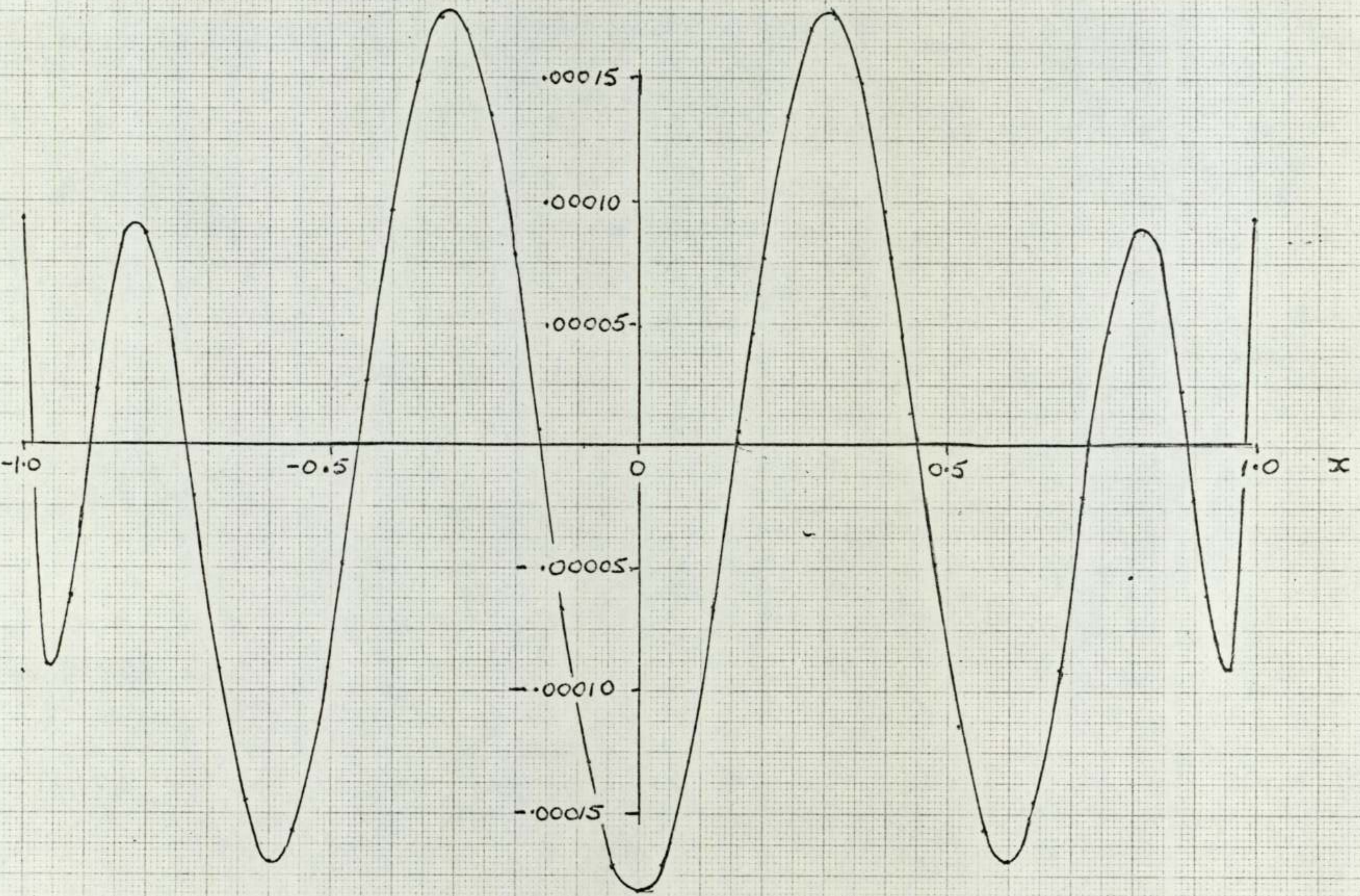


Fig 7.5 Interpolation Approximation to $\dfrac{x}{\{\sqrt{x^2+1}+x\}^3\sqrt{x^2+1}}$

can be carried out conveniently since these polynomials up to degree n are orthogonal over summation at the zeros of $T_{n+1}(x)$. However, the effort involved is virtually the same as that of finding the L_2 approximation by summation over discrete points. In the two examples considered, the latter approximation gives a smaller maximum error.

CHAPTER VIII

Approximation in the L_∞ Norm

Introduction

The L_∞ norm applied to the approximation $f(x)$ to the continuous function $y(x)$ in the range $[a,b]$ seeks to minimise

$$\lim_{p\to\infty}\left[\int_a^b w(x)\left\{y(x)-f(x)\right\}^p dx\right]^{1/p}$$

Taking the limit of this expression produces the condition that

$$\max_{x\in[a,b]} w(x)\left|y(x)-f(x)\right| \text{ is to be a minimum.}$$

This implies that the required approximation must produce the least possible (weighted) maximum error. For this reason, this norm is often referred to as the minimax norm. Another name commonly used is the Chebyshev norm.

The characterization of the minimax norm, which is that the maximum error must occur not less than a minimum number of times with alternating signs, leads to an iterative approach to finding the best approximation. The minimax problem can be solved for approximations which are the ratio of two polynomials if one accepts the additional complexity of the solution of non-linear equations.

Characterization of Best Approximation

Consider an approximation in $[a,b]$ of the form

$$f(x) = \sum_{i=0}^{n} a_i\phi_i(x) \qquad (8.1)$$

where $\phi_i(x)$ are polynomials of degree i. Then $f(x)$ will be a polynomial of degree n and we wish to determine a_i so that

$$\max_{[a,b]} \left|y(x)-f(x)\right| \text{ is a minimum.}$$

(Throughout this Chapter the weight function will be taken as unity. This choice minimizes the maximum value of the absolute error.) In Appendix A8.1 it is shown that if the $\phi_i(x)$ form a Chebyshev set, then a necessary and sufficient condition for (8.1) to be the best approximation of degree n is that the error curve $y(x) - f(x)$ achieves its extrema at not less than $(n+2)$ points in $[a,b]$ with alternating sign.

This important property provides a method of determining the required coefficients. In (8.1) there are $(n+1)$ unknown coefficients. There is also the actual value of the error extreme, say h. Hence there are $(n+2)$ unknowns for which the error extremes provide $(n+2)$ conditions. However, this method does not have the simplicity of (say) an interpolation problem, since we do not know in advance the points at which the error extremes occur. Consequently, methods of solution are essentially iterative. One method of approach is described below.

The Remes Algorithm

Two algorithms due to Remes [12] offer methods of solving the problem indicated in the last paragraph. The method described here is the second algorithm and proceeds as follows.

First, choose a reference of $(n+2)$ points $\{x_i\}$ in $[a,b]$ and then solve the $(n+2)$ equations

$$y(x_i) - f(x_i) = (-1)^i h \qquad i = 0,1, \ldots..(n+1) \qquad (8.2)$$

giving the coefficients $a_o \ldots.. a_n$ of $f(x)$ and the error $\pm$ h at the points $\{x_i\}$. When the error curve is now constructed, it is found that the chosen points of reference are not the points of maximum error. It is possible to locate the local extrema of the error curve and this can be done for not less than $(n+2)$ points with alternating signs and in such a way as to include the point of maximum error. These points are used as a new reference for a further solution of (8.2). The new error curve can again be scanned for the positions of the extrema. Proceeding in this way, we eventually find a set of points at which the error curve has extrema of equal magnitude and opposite sign. The approximation so found is the required minimax fit to the given function.

Example

Find an approximation of the form $f(x) = a_o + a_1 x + a_2 x^2$ in $[0,2]$ to the function e^x.

We require a starting reference of four points. A common choice are the extrema of the Chebyshev polynomial of appropriate degree. (suitably shifted in range)

In this case use the extrema of $T_3(z)$ in $[-1,1]$ and transpose to $[0,2]$, giving

x	0	0.5	1.50	2.00
e^x	1.0000	1.6487	4.4817	7.3891

Equations (8.2) become

$$\begin{aligned}
1.0000 - a_0 & & &= h \\
1.6487 - a_0 &- 0.5a_1 &- 0.25a_2 &= -h \\
4.4817 - a_0 &- 1.5a_1 &- 2.25a_2 &= h \\
7.3891 - a_0 &- 2.0a_1 &- 4.00a_2 &= -h
\end{aligned}$$

The solution to these equations gives

$$f^{(1)}(x) = 1.1205 + 0.0624x + 1.5058x^2$$

$$h^{(1)} = -0.1205$$

Figure 8.1 shows the error curve produced. Since the error curve is fairly rounded at the extrema, no special care is taken in locating the position of the extreme points. If linear interpolation is used to find the points where the first-order differences are zero in Table 8.2, we find the internal extremes at 0.571 and 1.546.

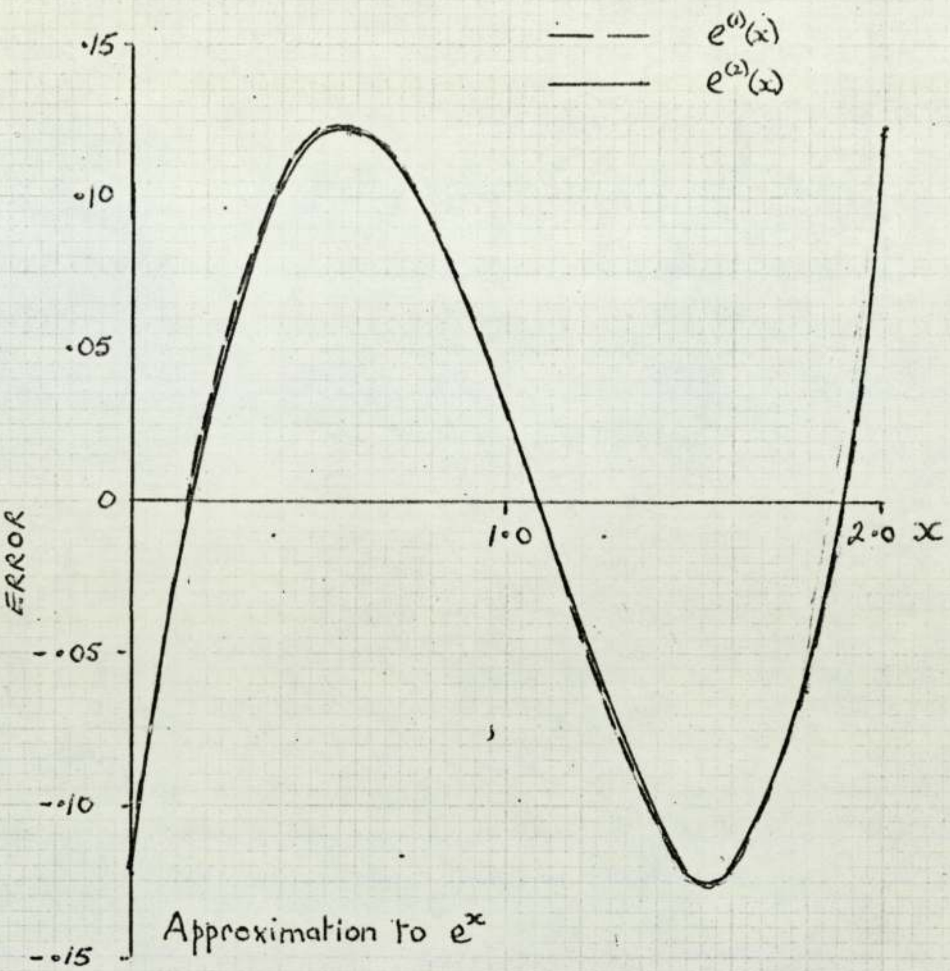
Using these and the end-points as a new reference, we have the new set of equations

$$\begin{aligned}
1.0000 - a_0 & & &= h \\
1.7700 - a_0 &- 0.571a_1 &- 0.3260a_2 &= -h \\
4.6927 - a_0 &- 1.546a_1 &- 2.3901a_2 &= h \\
7.3891 - a_0 &- 2.00a_1 &- 4.00a_2 &= -h
\end{aligned}$$

yielding $$f^{(2)}(x) = 1.1223 + 0.0604x + 1.5060x^2$$

$$h^{(2)} = -0.1223$$

Table 8.2 compares the errors in the two approximations $f^{(1)}(x)$ and $f^{(2)}(x)$. Linear interpolation in the first-order differences locates the



Approximation to e^x

Fig 8.1

internal extrema at 0.572 and 1.546. The error in $f^{(2)}(x)$ at these points is 0.1223 and -0.1225.

Hence, within the accuracy of the data used, the required approximation has been found.

x	0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
$e^{(1)}(x)$	-.1205	.0282	.1054	.1221	.0914	.0296	-.0437	-.1041	-.1221	-.0619	.1206
$e^{(2)}(x)$	-.1223	.0268	.1044	.1214	.0910	.0296	-.0433	-.1034	-.1213	-.0608	.1220

Error in Approximation to e^x

Table 8.2

Rational Function Approximation

The method of approach in minimax approximation lends itself to obtaining approximations in the form of rational functions, that is, the ratio of two polynomials.

If $$f(x) = \frac{\sum_{i=0}^{n} a_i x^i}{1 + \sum_{j=1}^{m} b_j x^j} \qquad (8.3)$$

then we notice that there are $(m+n+1)$ independent coefficients, as in this case, division throughout by b_0 has left the leading coefficient in the denominator equal to unity. In an analogous way to the linear case, we require (8.3) to produce an error curve with $(m+n+2)$ local extrema with alternating sign (Appendix A8.2). The approximation will be the required minimax fit when the extrema are of equal magnitude. Empirically, it is expected that for the same number of unknown coefficients, (8.3) will produce a smaller maximum error than (8.2). However, the equations that have to be solved for the coefficients in (8.3) are non-linear. It is possible, in trying to solve the problem iteratively to produce a solution with a pole in $f(x)$ where no such pole appears in the original function.

Below are discussed three methods of approach to the solution of the non-linear problem associated with rational approximations.

Method 1 Linearization

Let the approximation be written as $R(x) = \dfrac{P_n(x)}{Q_m(x)}$

Then, as before, we choose a reference $[x_i]$ $(i = 0,1, \ldots\ldots m+n+1)$ and solve the equations

$$y(x_i) - \frac{P_n(x_i)}{Q_m(x_i)} = (-1)^i h \qquad (8.4)$$

It is noticed that the non-linearity is introduced into (8.4) by the presence of the unknown h. Hence, we may choose a value of h and solve (8.4) as the linear system,

$$\left\{(-1)^i h - y(x_i)\right\} Q_m(x_i) + P_n(x_i) = 0 \qquad i = 0,1 \ldots\ldots (m+n+1) \qquad (8.4a)$$

The iterative approach requires to find both the positions of the local extrema and the required value of h that satisfies (8.4a).

Example

Consider an approximation of the form $\dfrac{P_1(x)}{Q_1(x)}$ to the integral $\displaystyle\int_0^x e^{-t^2} dt$ where $P_1(x)$ and $Q_1(x)$ are both linear functions.

$$\text{Now} \int_0^x e^{-t^2} dt = \frac{\sqrt{\pi}}{2} - \int_x^\infty e^{-t^2} dt$$

and the integral on the right can be integrated successively by parts to produce an asymptotic series (see Chapter V)

$$\text{i.e} \int_0^x e^{-t^2} dt = \frac{\sqrt{\pi}}{2} - e^{-x^2}\left\{\frac{1}{2x} - \frac{1}{4x^3} + \frac{1.3}{8x^5} - \frac{1.3.5}{16x^7} + \frac{1.3.5.7}{32x^9} \ldots\right\} \qquad (8.5)$$

This series may be used to evaluate the integral when x is relatively large. Hence, if we use the series when $x \geqslant 3$, the range of the rational approximation can be chosen as $[0,3]$.

Let $z = \frac{2}{3}(x - \frac{3}{2})$ and then the approximation

$$f(z) = \frac{a_0 + a_1 z}{1 + b_1 z} \quad \text{is in the range } [-1,1].$$

There are four unknowns in the problem, so the reference points are chosen as the points of extreme values of $T_3(z)$

i.e.

z	-1	$-\frac{1}{2}$	$\frac{1}{2}$	1
$y(z)$	0	0.630245	0.884929	0.886207

In order to linearize the equations, we choose h = -0.06 then substituting in (8.4) we have

$$a_0 - a_1 = -.06 + .06b_1$$

$$a_0 - 0.5a_1 = 0.690245(1 - 0.5b_1)$$

$$a_0 + 0.5a_1 = 0.824929(1 + 0.5b_1)$$

Since one unknown has been fixed, only three conditions can be satisfied and these have been arbitrarily chosen as the first three points in ascending order.

Solving these equations gives

$$a_0 = 0.786923 \qquad a_1 = 0.794654 \qquad b_1 = 0.871156$$

and the error when z = 1.0 is -0.04097

Figure 8.4 shows that the error curve is far from level and that the local extrema do not occur at the chosen reference points. As a next iteration, the reference points are chosen as -1.0, -0.75, 0.14, 1.0 and the value of h is chosen as before.

Proceeding in this way, choosing the extreme points as the new reference and at each stage making a suitable choice of h, the function f(z) is calculated to try and level the error curve. After the first three iterations, linear interpolation was used to try to find a suitable value of h to make the error at the fourth point (z = 1.0) equal to that at the other three chosen points. Table 8.3 shows the progress of the iteration process.

As a general remark it may be pointed out that linear interpolation was not a satisfactory method of locating the extreme point near z = -0.800. Linear interpolation of the first order differences of the computed error curve suggested that the maximum should occur slightly to the right of z = -0.800, whereas the error curve in the next iteration showed the error to be greatest at z = -0.800.

For example, in the last iteration, this method suggests that, interpolating between computed points spaced at 0.1 intervals, the maximum

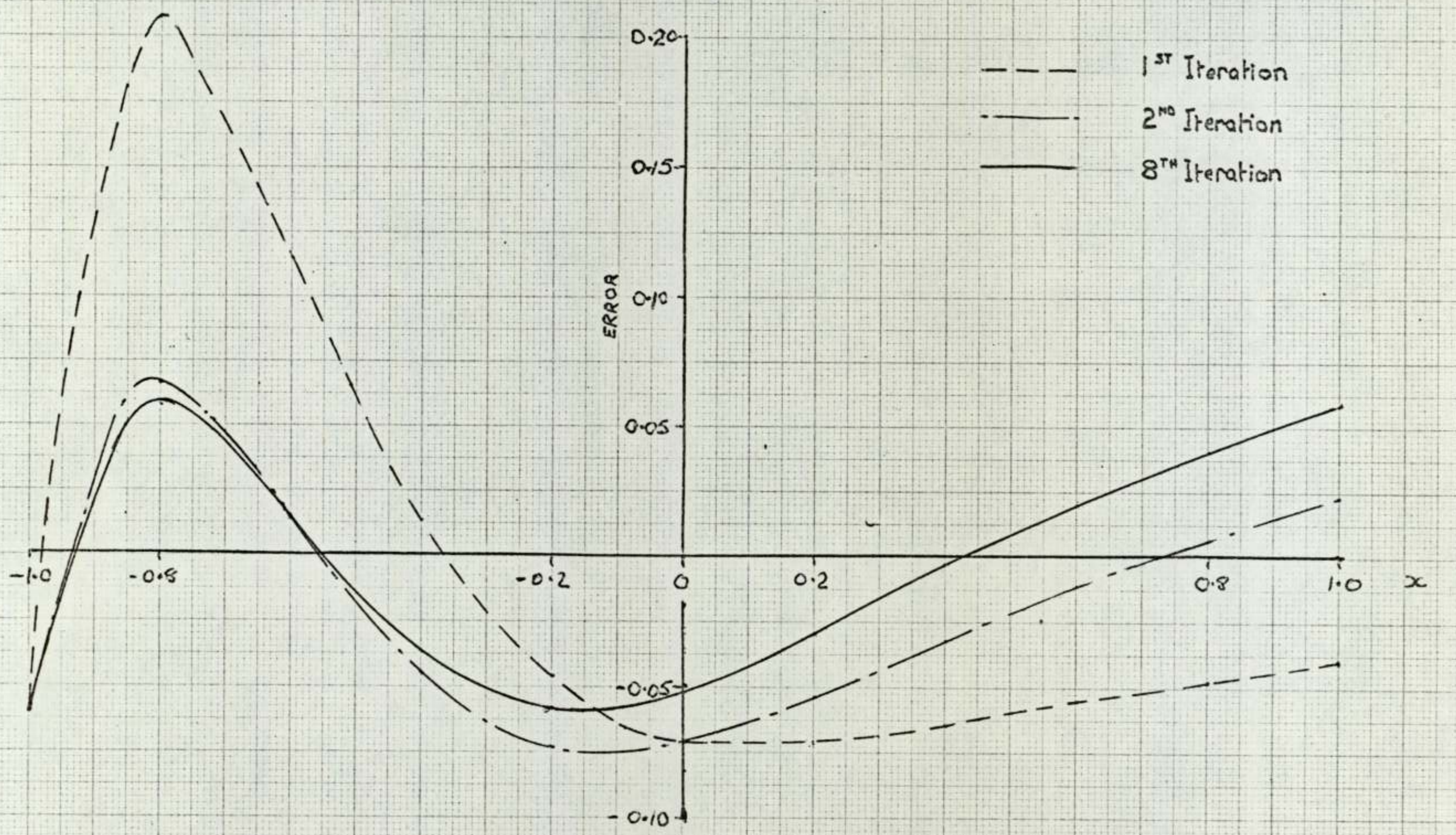

0·20
0·15
0·10
ERROR
0·05
-0·05
-0·10
-1·0
-0·8
-0·2
0
0·2
0·8
1·0
x
1ST Iteration
2ND Iteration
8TH Iteration

Fig 8.4

should be at $z^* = -0.771$. Actual calculation shows that

$$e(-0.800) = 0.0585$$

$$e(-0.795) = 0.0583$$

$$e(-0.790) = 0.0581$$

$$e(-0.700) = 0.0433$$

Iteration	Reference Points				h	e(1.0)
1	-1.0	-.5	.5	1.0	-.06	-.04097
2	-1.0	-.75	.14	1.0	-.06	.02270
3	-1.0	-.786	-.113	1.0	-.04	.11640
4	-1.0	-.768	-.231	1.0	-.0495	.07946
5	-1.0	-.756	-.172	1.0	-.055	.07275
6	-1.0	-.752	-.162	1.0	-.063	.03954
7	-1.0	-.800	-.162	1.0	-.058	.06029
8	-1.0	-.800	-.147	1.0	-.0588	.05780
9	-1.0	-.800	-.147	1.0	-.0585	.05873

Iteration	a_0	a_1	b_1
1	.786923	.794654	.871156
2	.785008	.800352	.744259
3	.828658	.842004	.666341
4	.816112	.830696	.705364
5	.813761	.829512	.713605
6	.797455	.813806	.740496
7	,805746	.822000	.719756
8	.804625	.820970	.722025
9	.805049	.821360	.721175

Approximation to $\int_0^x e^{-t^2} dt$

Table 8.3

Finally we illustrate the use of the series in (8.5). Putting x = 3 we have

$$\int_0^3 e^{-t^2}\,dt = 0.8862269 - e^{-9}\left\{\frac{1}{6} - \frac{1}{108} + \frac{3}{1944} - \frac{15}{34992} + \ldots\ldots\right\}$$

Taking only the first three terms of the expansion gives

$$f(3) \simeq 0.8862269 - 0.0001233\left[0.166667 - 0.009259 + 0.001543 \ldots.\right]$$

$$= 0.8862269 - 0.0000196$$

$$= 0.886207 \quad \text{to six significant figures.}$$

The error in f(3) to six significant figures is therefore zero. An error estimate may be obtained from the first neglected term of the series (See Chapter V)

i.e. $|e| < 0.0001233 \times \frac{15}{34992} \simeq 5 \times 10^{-8}$

Method 2 A Direct Method for Rational Functions

This method is due to Stoer [15]. In it we seek a rational approximation $R(x) = \frac{P_n(x)}{Q_m(x)}$ where the main feature of the method is that R(x) is expressed as a continued fraction. When the best approximation is found, there are at least $(n+m+2)$ points x_i in $[a,b]$ such that

$$y(x_i) - R(x_i) = (-1)^i h, \qquad i = 0,1 \ldots. m+n+1 \qquad (8.5)$$

where $|e|_{\max} = |h|$

The approach, as in the other methods, is an iterative one, but because of the continued fraction form of R(x), the interpolation property of Thiele's expansion is exploited to find the solution for h in (8.5) when the reference points x_i are given.

Let $R(x) = \frac{P_n(x)}{Q_m(x)}$ where $n \geqslant m$.

Then it is possible to write

$$R(x) = e_0 + e_1(x - x_0) + \ldots e_{n-m}(x-x_0)(x-x_1)\ldots.(x-x_{n-m-1})$$

$$+ e_{n-m+1}\frac{(x - x_0)(x - x_1)\ldots.(x - x_{n-m})}{\psi(x)}$$

where $\psi(x) = 1 + \frac{x - x_{n-m+1}}{e_{n-m+2} +} \cdots\cdots \frac{x - x_{n+m-1}}{e_{n+m}}$ (8.6)

The coefficients in (8.6) must be determined so that for a given choice of reference points $\left[x_i^{o}\right]$.

$$R(x_i^o) = y(x_i^o) - (-1)^i h^o \qquad i = 0,1, \ldots\ldots m+n+1 \qquad (8.7)$$

The polynomial part in (8.6) can be expressed in terms of divided differences and the continued fraction part in terms of reciprocal differences. (Ch. IV)

i.e. for divided differences

$$\begin{array}{lll}
R(x_0) = a_{00} & & \\
 & a_{11} & \\
R(x_1) = a_{10} & & \ddots \\
 & a_{21} & \\
R(x_2) = a_{20} & \vdots & \\
\vdots & & a_{n-m+1,\, n-m+1} \\
 & a_{n-m+1,1} & \\
R(x_{n-m+1}) = a_{n-m+1,0} & &
\end{array}$$

$$\text{where } a_{i,k} = \frac{a_{i,k-1} - a_{i-1,k-1}}{x_i - x_{i-k}}$$

$$\text{and } e_i = a_{i,i}$$

It is noted that if $m = 0$, then $P_n(x)$ passes through $(n+2)$ points and $a_{n+1,m+1} = 0$ which is the linear problem.

To obtain the coefficients in the continued fraction, we write (from 8.6)

$$\psi(x) = \frac{P_{n-m+1}(x) - P_{n-m}(x)}{R(x) - P_{n-m}(x)} \qquad (8.8)$$

where $P_k(x) = e_0 + e_1(x - x_0) + \ldots\ldots e_k(x - x_0)(x - x_1) \ldots\ldots (x - x_{k-1})$

The values of the right-hand side of (8.8) can be evaluated in terms of h^o, knowing that $R(x)$ must satisfy (8.7) at the reference points $\left[x_{n-m+1}, \ldots\ldots x_{n+m+1}\right]$

Hence, we have a table of reciprocal differences

$$\begin{array}{lll}
\psi(x_{n-m+1}) = c_{00} & & \\
 & c_{11} & \\
\psi(x_{n-m+1}) = c_{10} & & \ddots \\
\vdots & c_{21} & \\
\vdots & \vdots & c_{2m,2m} \\
\psi(x_{n+m+1}) = c_{2m,0} & &
\end{array}$$

$$\text{where } c_{i,k} = \frac{x_{n-m+1+i} - x_{n-m+1+i-k}}{c_{i,k-1} - c_{i-1,k-1}} + c_{i-1,k-2}$$

$$\text{and } e_{n-m+1+i} = c_{i,i} - c_{i-2,i-2} \qquad \left(i = 1, \ldots. 2m-1\right)$$

Now the function $R(x)$ will interpolate through the chosen points if the fraction terminates. This implies that $c_{2m,2m} = 0$ or

$$c_{2m-1,2m-1} = c_{2m,2m-1} \tag{8.9}$$

If (8.9) can be solved, then the continued fraction has the property (8.7). What is the nature of the equation in(8.9)? From (8.7) and (8.8) it can be seen that $R(x)$ and hence the terms in the reciprocal difference table will be functions of h^o. In fact (8.9) will represent the equality of two rational functions in h^o. In general, this equation cannot be solved directly and an iterative solution must be sought. This is the most difficult part of the process. Stoer gives an Algol programme which includes a solution of this problem by Newton's method, claiming that two iterations are usually sufficient to obtain the desired accuracy in h. As an alternative he suggests using the method of regula falsi. In the example given below, (8.9) yields a quadratic in h which can be solved directly for the value of smallest modulus.

Once this stage of the algorithm is solved, a new basis can be chosen from the extrema of the error curve and the process repeated until a satisfactory solution is found.

Example

To find an approximation to $\dfrac{1}{(1+x^2)^{\frac{1}{2}}}$ over the range $[0,1]$ in the form $\dfrac{P_1(x)}{Q_1(x)}$

Since n = m = 1, the continued fraction is $R(x) = e_0 + \dfrac{e_1(x-x_0)}{1+\dfrac{(x-x_1)}{e_2}}$

Choose as the initial reference, the extremes of the shifted Chebyshev polynomial $T_3{}^*(x)$, in the range $[0,1]$.

i.e.

	x	0.0	0.25	0.75	1.00
	y(x)	1.0000	0.9700	0.8000	0.7071
from (8.7)	R(x)	1+h	0.97-h	0.8+h	0.7071-h

The divided difference table becomes

$$R(0) = 1 + h$$

$$R(0.25) = .97 - h \qquad \frac{(.97-h)-(1+h)}{0.25}$$

$$\therefore\ e_0 = 1 + h \qquad e_1 = -0.1200 - 8h$$

Then
$$\psi(x) = \frac{e_1(x-x_0)}{R(x)-e_0} = \frac{(-0.1200-8h)x}{R(x)-(1+h)}$$

and the reciprocal difference table is

$$\psi(0.25) = 1.0000$$

$$\tfrac{1}{2}\left[\frac{1}{\dfrac{0.0900+6h}{0.20000} - 1.0000}\right]$$

$$\psi(0.75) = \frac{0.0900+6h}{0.2000}$$

$$\tfrac{1}{4}\left[\frac{1}{\dfrac{0.1200+8h}{0.2929+2h} - \dfrac{0.0900+6h}{0.2000}}\right]$$

$$\psi(1.00) = \frac{0.1200+8h}{0.2929+2h}$$

Hence, for the function R(x) to pass through the given four points, we have equality of the entries in the second column.

$$\text{i.e.} \qquad 0 = 36h^2 + 2.2122h - 0.0275$$

We take the numerically least solution as our value of h giving

$$h = 0.0106$$

whence $R(x_i) = 1.0106,\ 0.9594,\ 0.8106,\ 0.6965$

$$e_0 = 1.0106, \qquad e_1 = -0.2048$$

$$e_2 = -2.1551$$

and $$R(x) = 1.0106 - \frac{0.2048(x - 0.0)}{1 - \frac{(x - 0.25)}{2.1551}}$$

Figure 8.5 shows the error curve produced by R(x) and we choose as a new reference the points [0.0, 0.26, 0.73, 1.00].

i.e.				
x	0.0	0.26	0.73	1.00
$y(x)$	1.0000	0.9679	0.8077	0.7071

$$R(0) = 1.0000 + h \qquad R(0.26) = 0.9679 - h$$

$$a_{11} = \frac{0.9679 - h - (1+h)}{0.26 - 0.0} = -0.1235 - 7.6923\,h$$

$$\therefore \quad e_0 = 1.0000 + h \qquad e_1 = -0.1235 - 7.6923h$$

$$\psi(x) = \frac{(-0.1235 - 7.6923h)x}{R(x) - (1.0000 + h)}$$

The reciprocal difference table is

$$\psi(.26) = 1.0000$$

$$\frac{0.47}{.4688+29.201h-1.0000}$$

$$\psi(.73) = .4688+29.201h$$

$$\frac{0.27}{\frac{.1235 - 7.6923h}{.2929 + 2h} - (.4688 + 29.201h)}$$

$$\psi(1.00) = \frac{.1235-7.6923h}{.2929 + 2h}$$

This gives rise to the equation

$$0 = 15.768h^2 + 2.0224h - 0.0420$$

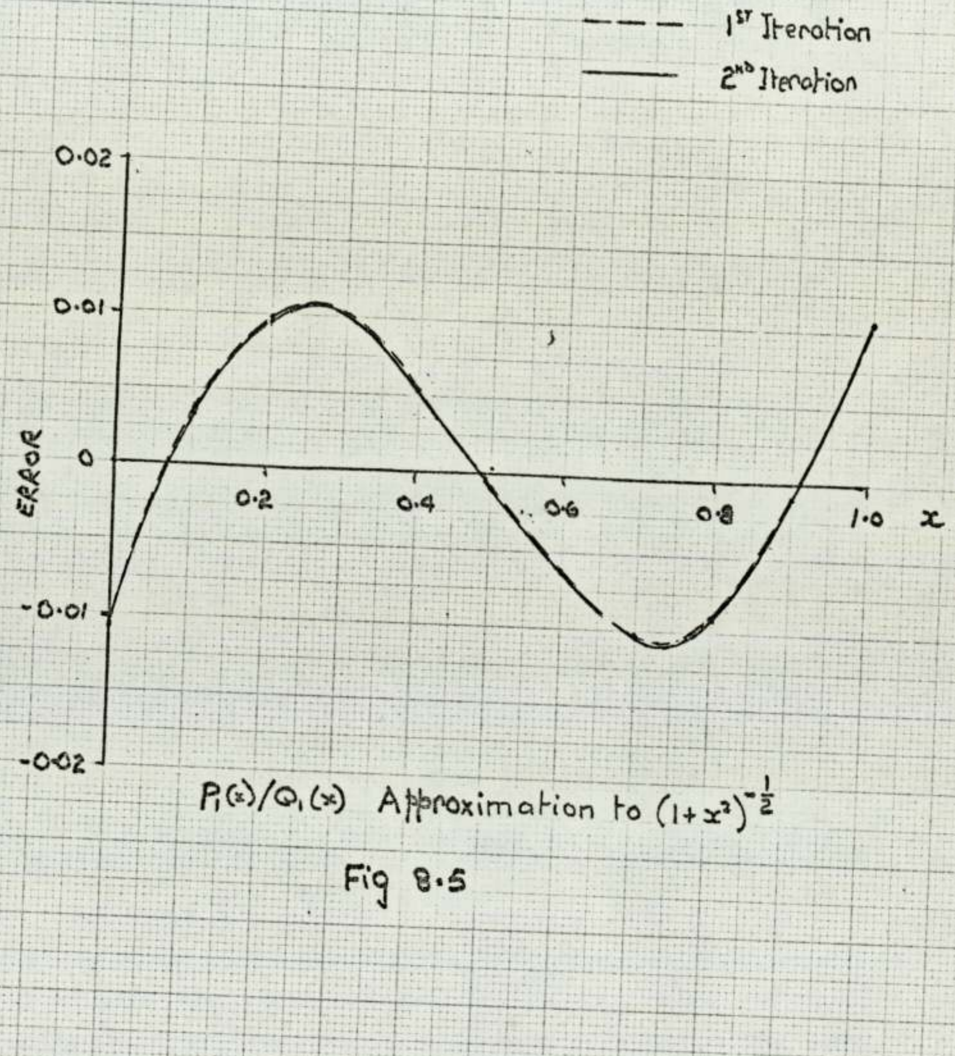
and taking the smallest root of this gives

$$h = 0.01067$$

hence $$e_0 = 1.0107 \qquad e_1 = -0.2056$$

$$e_2 = -2.1401$$

and $$R(x) = 1.0107 - \frac{0.2056(x - 0.0)}{1 - \frac{(x - 0.26)}{2.1401}}$$



$P_1(x)/Q_1(x)$ Approximation to $(1+x^2)^{-\frac{1}{2}}$

Fig 8·5

Again the error in $R(x)$ may be seen in figure (8.5). We notice that the actual error extrema must occur at points very close to the chosen reference points.

A tabulation of the computed error is given in Table 8.6.

x	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$y-R^{(1)}(x)$	-.0106	.0037	.0098	.0102	.0054	-.0003	-.0062	-.0102	-.0089	-.0010	.0106
$y-R^{(2)}(x)$	-.0107	.0035	.0097	.0101	.0053	-.0004	-.0063	-.0104	-.0090	-.0011	.0107

Error in Approximation to $(1+x^2)^{-\frac{1}{2}}$

Table 8.6

Method 3 Rational Approximation as an Eigenvalue Problem

If equation (8.4) is rearranged, we can write

$$-\left\{y(x_i) - (-1)^i h\right\} Q_m(x_i) + P_n(x_i) = 0 \qquad (8.10)$$

As before, an iterative method is applied to finding a reference

$$a \leqslant x_0 < x_1 \;\ldots\ldots\ldots < x_{m+n+1} \leqslant b$$

such that (8.10) is satisfied at these points and

$$\max_{a \leqslant x \leqslant b} \left| y(x) - \frac{P_n(x)}{Q_m(x)} \right| = h$$

The method employed here is due to Curtis and Osborne [4] and employs a method for solving eigenvalue problems developed by Osborne. Equation (8.4) can be thought of as an eigenvalue problem in which the maximum error h is the eigenvalue and the corresponding eigenvector is the column vector comprising the coefficients of the approximation.

In more detail, let $P_n(x) = \sum_{j=0}^{n} a_j x^j \qquad Q_m(x) = \sum_{j=0}^{m} b_j x^j$

then (8.10) can be written as

$$\left\{a_0 + a_1 x_i \ldots\ldots a_n x_i^n\right\} - \left\{y(x_i) - (-1)^i h\right\}\left\{b_0 + b_1 x_i + \ldots\ldots b_m x_i^m\right\} = 0$$

$$i = 0,1 \;\ldots\ldots\; n+m+1$$

or in matrix form

$$\left[Y \;\vdots\; -(F - hG)X\right].\left[C\right] = 0 \qquad (8.11)$$

where $y_{rs} = x_r^{s-1}$ $\qquad s = 1,2, \ldots\ldots n+1$

$$x_{rs} = x_r^{s-1} \qquad s = 1,2, \ldots\ldots m+1$$

$$F = \text{diagonal } \{y(x_r)\}$$

$$G = \text{diagonal } \{(-1)^r\}$$

and $r = 0,1, \ldots\ldots m+n+1$

and $[C]$ is the column vector $\left[a_0\ a_1 \ldots\ldots a_n\ b_0\ b_1 \ldots\ldots b_m\right]^T$

Appendix A8.3 sets out how the method seeks the solution to (8.11) iteratively. The method is summarized as follows.

Let $[x]^{(i)}$, $h^{(i)}$, $[c]^{(i)}$ be the values of the reference points, maximum error and coefficients respectively at some stage of the iteration. Then the algorithm becomes

$$\left[Y \,\vdots\, -(F - h^{(i)}G)X\right]^{(i)} \left[v^{(i+1)}\right] = \left[0 \,|\, GX\right]^{(i)} [c]^{(i)}$$

$$\left[Y \,\vdots\, -(F - h^{(i)}G).X\right]^{(i)} [c]^{(i+1)} = \left[0 \,\vdots\, GX\right]^{(i)} [v]^{(i+1)}$$

$$h^{(i+1)} = h^{(i)} - \frac{(v)^{(i+1)}_p}{(c)^{(i+1)}_p} \tag{8.12}$$

where $(v)p$ represents the element of maximum modulus (the p^{th}) in the vector $[v]$. In our case the coefficients are divided throughout by the coefficient of maximum modulus so that they are all numerically less or equal to one.

Equations (8.12) produce the new values of h and $[c]$. In order to determine the next reference set, a new error curve must be computed and the extrema found by interpolation.

Appendix A8.4 gives a listing of a computer programme which has been developed to exploit this method. Some examples of the use of this technique are now given.

Example 1

Consider an approximation to $y = \cos h^{-1}x$ in the range $[1,3]$. Since y is two-valued in the range, the positive value will be taken.

Also $\dfrac{dy}{dx} = \dfrac{1}{\sqrt{x^2 - 1}}$ and we see that when $x = 1$, $y = 0$ and $\dfrac{dy}{dx} = \infty$

This suggests that a rational approximation will be both difficult to find and probably unsatisfactory in terms of the size of the error near $x=1$. For this reason, the independent variable is chosen as

$$z = \sqrt{x^2 - 1}$$

and we choose a function $\frac{P_3(z)}{Q_3(z)}$ in the range $\left[0, \sqrt{8}\right]$.

As the initial basis, we choose the points of extreme value of the Chebyshev polynomial $T_6(u)$, suitably transposed into the range of approximation. Table 8.7 below shows the progress of the iteration in this case and Figure 8.8 is a plot of the final error curve.

Coefficients of $P_3(z)$

Iteration	a_0	a_1	a_2	a_3
1	.0000 54	.990 013	.889 981	.045 886
2	.000 157	.995 673	.737 218	.047 334
3	.000 142	.993 291	.781 573	.047 410
4	.000 144	.993 266	.786 585	.047 324
5	.000 144	.993 267	.786 532	.047 327

Coefficients of $Q_2(z)$ Max. Error

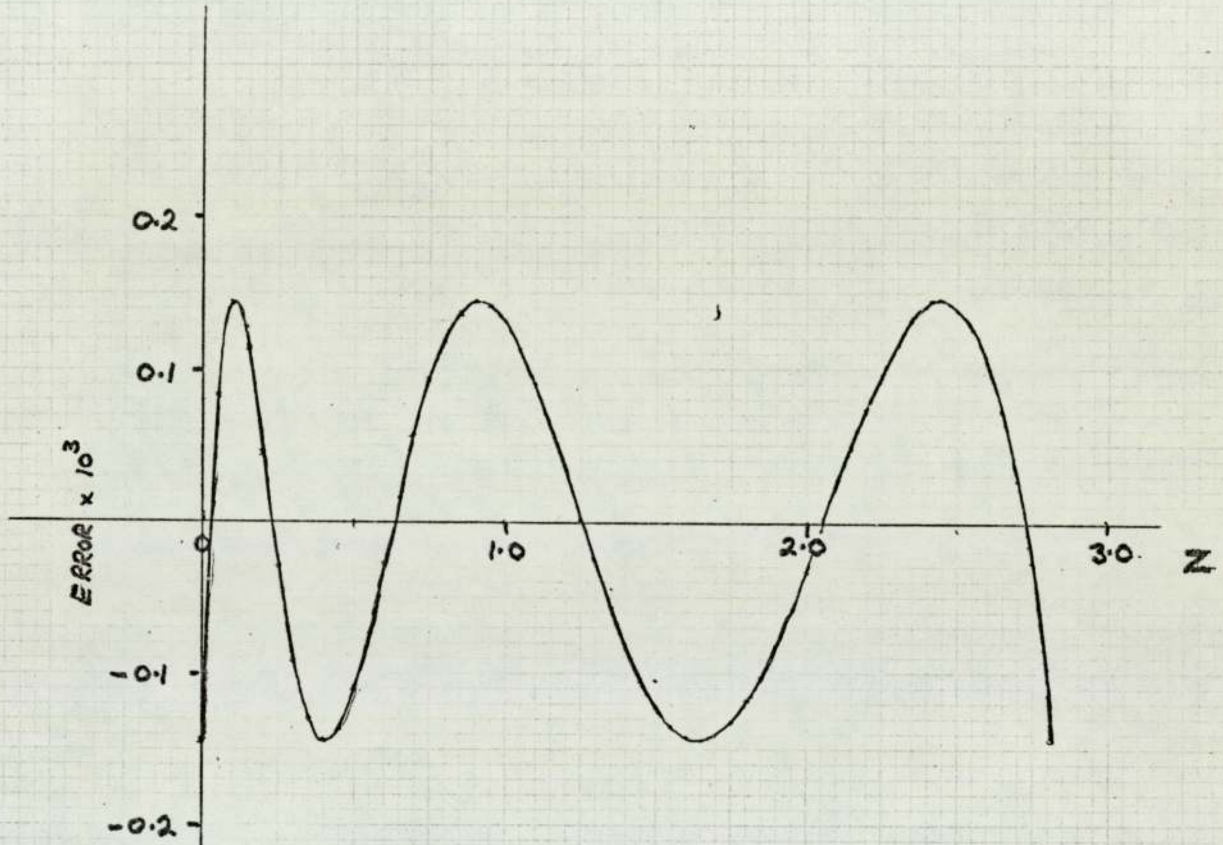
Iteration	b_0	b_1	b_2	$h \times 10^4$
1	1.00	.824 839	.360 433	.5437
2	1.00	.698 469	.327 877	1.5654
3	1.00	.733 833	.334 183	1.4226
4	1.00	.738 249	.335 321	1.4377
5	1.00	.738 202	.335 312	1.4385

Points of Extrema of Error Curve

Iteration	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	0.0	.286	.716	1.411	2.171	2.551	2.828
2	0.0	.120	.500	.970	1.720	2.550	2.828
3	0.0	.080	.350	.920	1.610	2.430	2.828
4	0.0	.110	.404	.901	1,640	2.447	2.828
5	0.0	.109	.407	.914	1.644	2.444	2.828

Approximation to $\cos h^{-1} x$ by Rational Function

Table 8.7



P_3/Q_2 Approximation to $\cosh^{-1} z$

Fig 8.8

Example 2

Let an approximation to $y(x) = 0.92 \cosh x - \cos x$ be of the form $f(x) = \frac{P_2(x)}{Q_2(x)}$ valid in the range $[-1,1]$.

Here, $n + m + 2 = 6$, so we choose the points at which $T_5(x)$ achieves its extreme values as the initial basis.

Figure 8.9 shows that the error curve has seven extrema instead of the expected six. However, we notice that since the original function is even, the odd terms of the approximation have zero coefficients. In that case, both $P_3(x)/Q_2(x)$ and $P_2(x)/Q_3(x)$ would have the same error curve as $P_2(x)/Q_2(x)$.

The following table summarizes the iteration process.

Coefficients of P (x) Max. Error

Iteration	a_0	a_1	a_2	$h \times 10^4$
1	-.079 833 6	0	.957 947 3	$-.29 \times 10^{-7}$
2	-.079 927 6	0	.958 597 2	-.7244
3	-.079 916 8	0	.958 556 6	-.8319
4	-.079 916 8	0	.958 557 0	-.8322

Coefficients of Q (x)

Iteration	b_0	b_1	b_2
1	1.00	0	-.001 385 3
2	1.00	0	-.000 670 8
3	1.00	0	-.000 692 6
4	1.00	0	-.000 692 0

Points of Extrema of Error Curve

Iteration	x_1	x_2	x_3	x_4	x_5	x_6	x_7
Initial values	-1.000	-.809	-.309	.309	.809	1.000	
1	-1.000	-.917	-.567	0.000	.567	.917	1.000
2	-1.000	-.864	-.491	0.000	.491	.864	1.000
3	-1.000	-.863	-.502	0.000	.502	.863	1.000
4	-1.000	-.863	-.499	0.000	.499	.863	1.000

Approximation to 0.92 cosh x - cos x by Rational Function

Table 8.10

Approximation to $0{\cdot}92 \cosh x - \cos x$

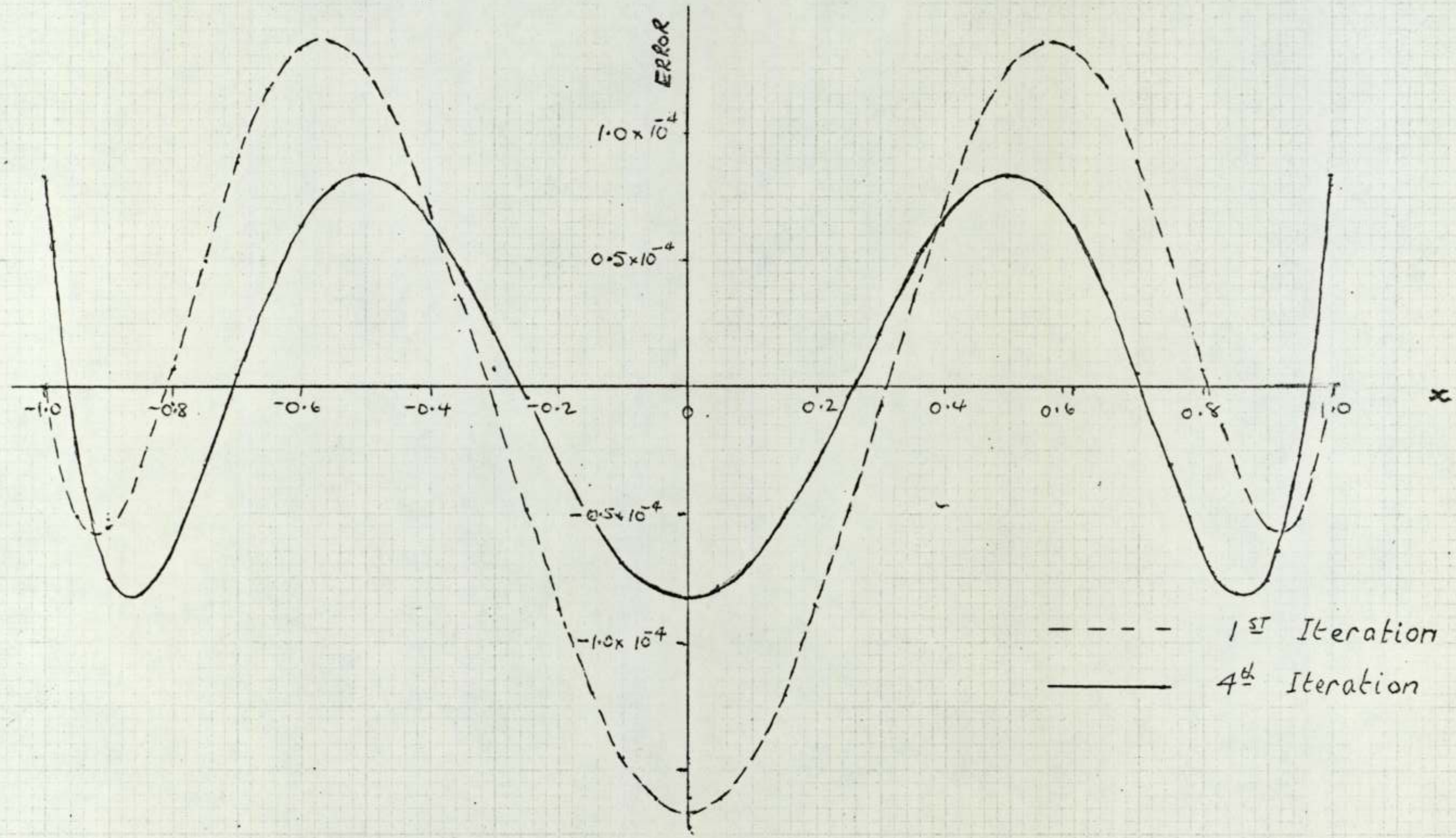


Fig 8.9

It is interesting to compare this result with that obtained for the L_1 approximation to the same function using an approximation which also produces an error curve with six zeros (See Chapter VI). The rational function approximation is seen to have a maximum error of approximately 0.83×10^{-4}, whereas the L_1 approximation has a maximum error of 0.29×10^{-3}. In addition, it can be seen that the profiles of the two error curves are quite different, the "equal oscillation" property being absent from the L_1 error. Over a restricted range, say $[-0.7, 0.7]$, the L_1 approximation has a smaller maximum error than the rational minimax approximation.

Example 3

In this example, a polynomial approximation of the form $P_4(x)$ is found to $y(x) = 2\pi \log_e\left(\frac{1+e^{-x}}{2}\right)$ in the range $[0,4]$.

Since six extrema might be expected, the initial reference was taken as the points of extreme value of $T_5(z)$ suitably scaled to the given range. The process is tabulated in table 8.11 below and the final error curve is plotted in figure 8.12.

* Coefficients of P (x)

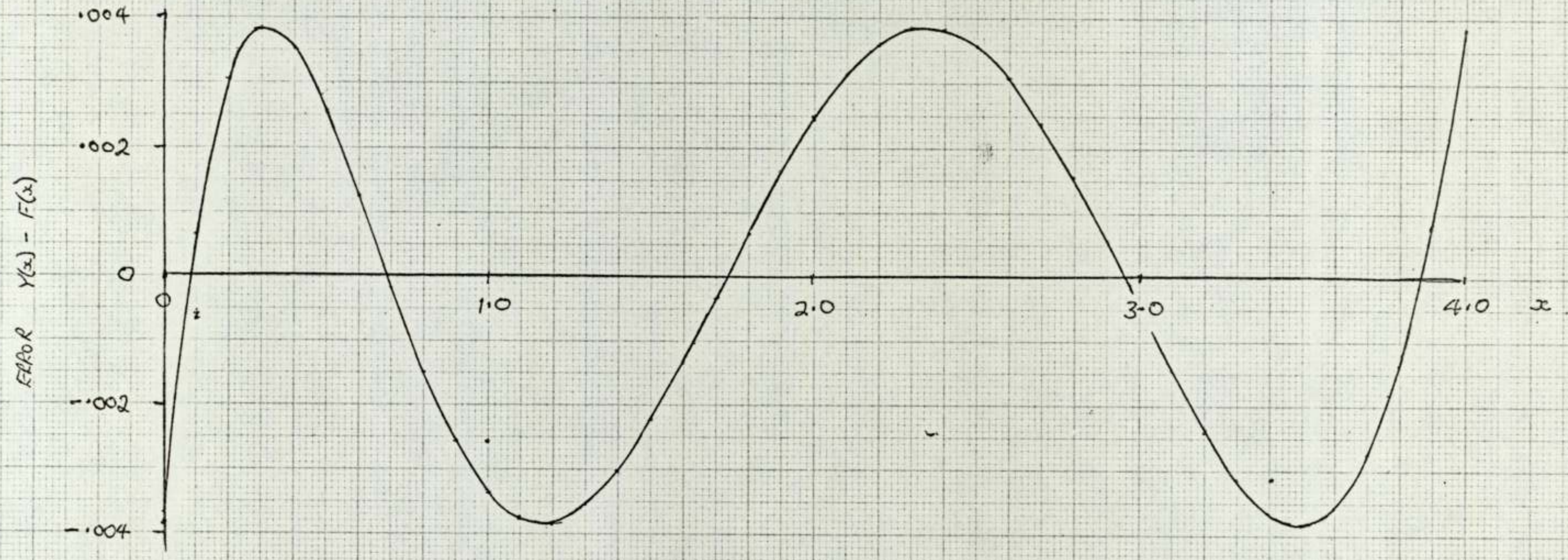
Iteration	a_0	a_1	a_2	a_3	a_4	$h \times 10^2$
1	.003 404 1	-3.187 141 0	.924 685 3	-.119 227 0	.005 361 6	.3404
2	.003 837 2	-3.199 788 9	.923 791 8	-.119 373 8	.005 417 0	.3837
3	.003 839 3	-3.199 810 7	.926 775 4	-.119 766 6	.005 418 3	.3839
4	.003 839 3	-3.199 810 7	.926 775 4	-.119 766 6	.005 418 3	.3839

Points of Extrema of the Error Curve

Iteration	x_1	x_2	x_3	x_4	x_5	x_6
1	0.0	0.310	1.155	2.365	3.508	4.000
2	0.0	0.322	1.169	2.345	3.488	4.000
3	0.0	0.322	1.169	2.344	3.488	4.000
4	0.0	0.322	1.169	2.344	3.488	4.000

Approximation to $2\pi\log_e \{\frac{1}{2}(1+e^{-x})\}$ by Polynomial

Table 8.11



$P_4(x)$ Approximation to $Y = 2\pi \log_e \{\frac{1}{2}(1 + e^{-x})\}$

Fig 8.12

* It may be noted that the coefficient a_1 is here numerically greater than one whereas it is stated that the programme scales all coefficients so that the maximum has a modulus equal to one. The reason for this is that the programme calculated the approximation in the rational form

$$P(x) = \frac{a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4}{b_0} \quad \text{and the values}$$

have been divided out to produce the more usual form.

Remarks on Minimax Approximation

Minimax approximation displays two characteristics that make it attractive. Firstly, the method of solution yields an explicit figure for the maximum error and secondly, because the error extrema are distributed throughout the range, the approximation, may be used with equal confidence anywhere within the range. Its disadvantage is the necessity for an iterative method of solution to find the best fit.

Since an iterative method has to be employed in any case, it is a natural extension to use the minimax criteria when finding a rational function approximation. Of the methods given for solving the non-linear problem, that involving linearization by choosing h seems least attractive because of the slowness of convergence. Stoer's method is the most complicated due to the inherent difficulty in manipulating continued fractions in comparison to polynomial forms. The method of Curtis and Osborne has worked well in those cases where it has been applied. It is also shown that the linear problem is solved by the same programme as the rational in which the degree of the denominator is zero.

It is concluded that the last method is the best one to adopt for finding minimax rational or polynomial approximations. If the final form is desired as a continued fraction, then it may be a better approach to use the programme provided by Stoer than to have to find a rational function and then convert to continued fraction.

CHAPTER IX

Cubic Spline Approximation

Introduction

The idea of spline approximation as a method of piecewise polynomial approximation is described and the equations are developed for the case when the polynomial is a cubic. In this case the approximation interpolates the given function at the joins or knots and smoothness of approximation is imparted by the ability to ensure continuity of the approximation and its first two derivatives at the knots. In this Chapter, a basic method of cubic spline approximation is programmed and some examples given of its application.

Concept of the Spline Approximation

Consider a set of n real values x_i in the range of approximation such that

$$a \leqslant x_1 < x_2 < \ldots\ldots < x_n \leqslant b \qquad (9.1)$$

An approximation is sought such that in each interval $\left[x_{j-1}, x_j\right]$ the approximating function is a low-degree polynomial. To ensure smoothness, the approximation and some of its derivatives are to be continuous at the interval joins, or knots, $x_2, x_3 \ldots\ldots, x_{n-1}$.

The simplest form of approximation is the polynomial of degree one, which is the broken line joining consecutive knots. In this case, no derivatives can be made continuous and the approximation is unsatisfactory. Consequently, polynomials of satisfactory form will be either quadratic or cubic.

In Appendix A9.1 it is demonstrated that splines of even degree display practical difficulties and the lowest degree spline which gives a useful approach is that of degree three. In this case, it is possible to prescribe that the first and second derivatives can be made continuous at the internal knots.

Defining Equations of Cubic Splines

The method described here follows Ahlberg, Nilson and Walsh [1]. Let M_j denote the value of the second derivative at the knot x_j. Now the second derivative of a cubic polynomial must be a linear function. In addition, the second derivative is to be continuous at the internal knots. Thus, we may write in $[x_{j-1}, x_j]$ that if f(x) is the required cubic polynomial

$$f''(x) = M_{j-1}\frac{(x_j - x)}{h_j} + M_j\frac{(x-x_{j-1})}{h_j} \qquad \text{where } h_j = x_j - x_{j-1}$$

This equation can be integrated twice and the two constants of integration evaluated from the fact that

$$f(x_{j-1}) = y_{j-1} \qquad \text{and } f(x_j) = y_j$$

where y(x) is the given function.

$$\text{i.e.} \quad f(x) = \frac{M_{j-1}(x_j-x)^3}{6h_j} + \frac{M_j(x-x_{j-1})^3}{6h_j} + \left(y_{j-1} - \frac{M_{j-1}h_j^2}{6}\right)\left(\frac{x_j-x}{h_j}\right)$$

$$+\left(y_j - \frac{M_j h_j^2}{6}\right)\left(\frac{x-x_{j-1}}{h_j}\right) \qquad \text{in } [x_{j-1}, x_j] \qquad (9.2)$$

Now the first derivatives of the splines must be continuous at the internal knots. From (9.2), differentiating and putting $x = x_j$ we have

$$f'(x_j -) = \frac{h_j}{6} M_{j-1} + \frac{h_j}{3} M_j + \frac{y_j - y_{j-1}}{h_j} \qquad (9.3a)$$

Equally, from the expression for f(x) in the interval $[x_j, x_{j+1}]$, we have

$$f'(x_j +) = -\frac{h_{j+1}}{3} M_j - \frac{h_{j+1}}{6} M_{j+1} + \frac{y_{j+1} - y_j}{h_{j+1}} \qquad (9.3b)$$

Hence, for continuity of the first derivative, (9.3a) and (9.3b) yield

$$\frac{h_j}{6} M_{j-1} + \frac{h_j + h_{j+1}}{3} M_j + \frac{h_{j+1}}{6} M_{j+1} = \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j} \qquad (9.4)$$

for j = 2,3 ....(n - 1)

Equations (9.4) provide (n - 2) conditions on the n unknowns, which are now the values M_j (j = 1,2 .... n). It is necessary, therefore, to impose two end conditions, one at each of the end points x_1 and x_n.

End Conditions

The two extra conditions which are applied at the ends of the range can be chosen in a variety of ways. The simplest choice is

$$M_1 = M_n = 0$$

which implies that the end-points are simple points of interpolation. This choice, although simple, tends to produce errors near the end points which are larger than for other choices. Another method is to prescribe the first derivative at the two end points. This may be useful in cases where the derivative is prescribed by the problem.

The end condition which has been adopted here is to put a restraint on the error at the mid-points of the two neighbouring intervals at the end of the range (Hayes Chapter 4 [9]). The errors at the mid-points of the extreme intervals are made equal to the errors at the mid-points of the two intervals next in line.

i.e. $$\left[e(x)\right]_{\frac{1}{2}(x_{t-1}+x_t)} = \left[e(x)\right]_{\frac{1}{2}(x_t + x_{t+1})} \qquad \text{for } t = 2 \text{ and } (n-1) \qquad (9.5)$$

Consideration is not given to the problem where the end-points of the range of approximation are outside the two outer chosen knots x_1 and x_n. In other words, it is assumed that the boundaries of the range are the first and last knot respectively.

Error Estimate

If constant knot spacing is employed, Curtis and Powell [5] have shown that an error estimate may be found in terms of the discontinuities of the third derivatives at the knots. They obtain the inequality

$$\max_{x_t \leqslant x \leqslant x_{t+1}} \left| y(x) - f(x) \right| \leqslant \frac{1}{384} \max\left(|D_t|, |D_{t+1}|\right) + O(h^6) \qquad (9.6)$$

$$\text{where } D_t = h^3 \left[f'''(x)\right]_{x_{t-}}^{x_{t+}} = h^4 y^{(iv)}(x_t) + O(h^8)$$

A sign that the errors may be large compared with the estimate may be given by the relative magnitude of the discontinuities in the third derivative.

From (9.2) we have

$$f'''(x) = \frac{1}{h_j}\left[M_j - M_{j-1}\right] \qquad \text{in } \left[x_{j-1}, x_j\right]$$

and the third derivative is easily calculated from the values of M_j.

Comments

The method of cubic spline approximation provides a straight forward approach to finding an approximation to a continuous function, exploiting the advantage in accuracy to be gained by reducing the range over which the function must provide a good fit. Its disadvantages lie mainly in the facts that a considerable amount of information must be stored, that is, the knots, the corresponding function values and the values of the second derivatives at the knots. Also equation (9.2) is rather cumbersome and rapid evaluation is not possible.

Computation of the Cubic Spline

The following assumptions are made

(a) The knot-spacing is constant throughout the range.

(b) The end points of the range were taken as the extreme knots.

(c) The end conditions are those of equation (9.5)

With (a), equations (9.4) can be written

$$\begin{bmatrix} \frac{1}{6}h & \frac{2}{3}h & \frac{1}{6}h & 0 & 0 & \cdots\cdots & \\ 0 & \frac{1}{6}h & \frac{2}{3}h & \frac{1}{6}h & 0 & \cdots\cdots & \\ 0 & 0 & \frac{1}{6}h & \frac{2}{3}h & \frac{1}{6}h & \cdots\cdots & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & & & & \frac{1}{6}h & \frac{2}{3}h & \frac{1}{6}h \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ \\ M_n \end{bmatrix} = \begin{bmatrix} \frac{1}{h}(y_3 - 2y_2 + y_1) \\ \frac{1}{h}(y_4 - 2y_3 + y_2) \\ \vdots \\ \\ \frac{1}{h}(y_n - 2y_{n-1} + y_{n-2}) \end{bmatrix} \qquad (9.7)$$

The end conditions (c) are

$$y(x_{1\frac{1}{2}}) - f(x_{1\frac{1}{2}}) = y(x_{2\frac{1}{2}}) - f(x_{2\frac{1}{2}})$$

but

$$f(x_{1\frac{1}{2}}) = -\frac{h^2}{16}M_1 + \frac{h^2}{16}M_2 + \tfrac{1}{2}(y_1 + y_2)$$

$$f(x_{2\frac{1}{2}}) = -\frac{h^2}{16}M_2 - \frac{h^2}{16}M_3 + \tfrac{1}{2}(y_2 + y_3)$$

$$\therefore \quad -\frac{h^2}{16}M_1 + \frac{h^2}{16}M_3 = y_{1\frac{1}{2}} - y_{2\frac{1}{2}} + \tfrac{1}{2}(y_3 - y_1) \qquad (9.7a)$$

Similarly

$$-\frac{h^2}{16}M_{n-2} + \frac{h^2}{16}M_n = y_{n-\frac{3}{2}} - y_{n-\frac{1}{2}} + \tfrac{1}{2}(y_n - y_{n-2})$$

These two equations added to (9.7) determine completely the values of M_j, the second derivatives at the knots.

The matrix of coefficients is tri-diagonal apart from the end conditions. The system is solved by the method proposed by Ahlbert, Nilson and Walsh[1] Chapter 2, based on successive elimination. The method may be summarized as follows:

Given
$$\begin{aligned}
b_1 x_1 + c_1 x_2 \quad &= d_1\\
a_2 x_1 + b_2 x_2 + c_2 x_3 \quad &= d_2\\
a_3 x_2 + b_3 x_2 + c_3 x_4 \quad &= d_3\\
\cdots\cdots\cdots \quad &\\
a_{n-1} x_{n-2} + b_{n-1} x_{n-1} + c_{n-1} x_n &= d_{n-1}\\
a_n x_{n-1} + b_n x_n &= d_n
\end{aligned}$$

we form
$$p_k = a_k q_{k-1} + b_k \qquad (q_0 = 0)$$
$$q_k = -c_k / p_k$$
$$u_k = (d_k - a_k u_{k-1}) / p_k \qquad (u_0 = 0)$$
for $k = 1, \ldots\ldots. n$

The required solution is then
$$x_k = q_k x_{k+1} + u_k \qquad k = 1, 2, \ldots.., (n-1)$$
$$x_n = u_n$$

and $x_n, x_{n-1}, \ldots.., x_1$ can be successively evaluated.

Two examples of the use of splines are now given.

Example 1

$$y(x) = \frac{e^{-x^2}}{1 + x^2} \qquad \text{in the range } [0,2].$$

The choice of interval between the knots was chosen initially as 0.5. The second derivatives calculated at the knots are used to find error estimates using (9.6). The actual maximum error in each interval was found by using quadratic interpolation between the three points spanning the extremum.

i.e. if $f_p = f_0 + p\mu\delta f_0 + \frac{p^2}{2!}\delta^2 f_0$

then $f_{max} = f_0 - \frac{(\delta f_{-\frac{1}{2}} + \delta f_{\frac{1}{2}})^2}{8(\delta^2 f_0)}$

The comparison of the estimated and computed results is seen in Table 9.1a. The process was repeated for step lengths of 0.25 and 0.125 and the corresponding results are in Table 9.1b and Table 9.1c.

It can be seen that the error estimate must, in certain instances, be treated with some caution. An indication that the estimate is too optimistic may be given by the magnitude of the discontinuities in the third derivative. If these are large, then the estimate is likely to prove inadequate.

$h = 0.5$

x(knots)	M	$f'''(x)$	$\left[f'''(x)\right]^{+}_{-}$	Error Estimate $\times 10^{+3}$	Max. Error $\times 10^{+3}$
0.0	-5.6544				
		12.651			6.80
0.5	0.6710		-11.034	3.60	
		1.616			5.0
1.0	1.4792		- 2.781	0.91	
		-1.165			1.8
1.5	0.3145		1.199	0.39	
		0.034			1.8
2.0	0.2087				

Table 9.1a

$h = 0.25$

x(knots)	M	$f'''(x)$	$\left[f'''(x)\right]^{+}_{-}$	Error Estimate $\times 10^{+4}$	Max. Error $\times 10^{+4}$
0.00	-4.6040				
		8.772			4.64
0.25	-2.4113		2.088	0.85	
		10.860			1.35
0.50	0.3035		-6.236	2.54	
		4.624			3.92
0.75	1.4596		-5.212	2.12	
		-0.588			0.887
1.00	1.3133		-1.340	0.55	
		-1.928			0.171
1.25	0.8312		0.316	0.13	
		-1.612			0.339
1.50	0.4281		0.652	0.27	
		-0.960			0.230
1.75	0.1978		0.508	0.21	
		-0.452			0.237
2.00	0.0753				

Table 9.1b

$h = 0.125$

x(knots)	M	$f'''(x)$	$[f'''(x)]^{+}_{-}$	Error Estimate $\times 10^{+5}$	Max Error $\times 10^{+5}$
0.0	-4.101				
		3.97			2.55
0.125	-3.605		5.42	2.86	
		9.39			2.52
0.25	-2.425		1.99	1.01	
		11.38			0.627
0.375	-1.002		-1.71	0.87	
		9.67			1.65
0.50	0.210		-3.34	1.69	
		6.33			1.94
0.625	1.003		-3.37	1.71	
		2.96			1.56
0.75	1.373		-2.52	1.28	
		0.44			1.004
0.875	1.428		0.62	0.316	
		1.06			0.511
1.00	1.295		0.70	0.356	
		1.76			0.169
1.125	1.075		0.14	0.071	
		1.90			0.057
1.25	0.837		-0.16	0.082	
		1.74			0.142
1.375	0.619		-0.30	0.153	
		1.44			0.171
1.50	0.439		-0.32	0.163	
		1.12			0.164
1.625	0.299		-0.30	0.153	
		0.82			0.141
1.75	0.196		-0.25	0.127	
		0.57			0.103
1.875	0.125		-0.19	0.097	
		0.38			0.104
2.00	0.077				

Table 9.1c

Example 2

Consider a spline function approximation to

$$y(x) = \cosh^{-1} x \qquad \text{in the range } [1,3] \text{ with knots}$$

equally spaced at intervals of 0.2. (The principal value of the function is considered.)

This is expected to prove a difficult problem, for when $x = 1$, $y = 0$ and yet all the derivatives are infinitely large. Apart from the problem of finding a cubic polynomial with a very large first derivative, the expression for D_t in (9.6), which involves the value of the fourth derivative, suggests that a reasonable error estimate will not be available. Table (9.2) shows the comparison between the error estimate and the actual maximum errors found by interpolation between points on the error curve calculated at intervals of 0.04.

$h = 0.2$

x(knots)	M	$f'''(x)$	$\left[f'''(x)\right]^{+}_{-}$	Error Estimate $\times 10^{+5}$	Max. Error $\times 10^{+5}$
1.0	-50.249				
		245.720			5991
1.2	-1.104		-250.055	520	
		-4.335			439.1
1.4	-1.991		11.115	23.0	
		6.780			120.0
1.6	-0.635		-6.455	1.35	
		0.325			30.3
1.8	-0.570		-0.680	1.42	
		1.005			9.03
2.0	-0.369		-0.625	1.30	
		0.380			1.95
2.2	-0.293		-0.060	0.125	
		0.320			0.802
2.4	-0.229		-0.110	0.229	
		0.210			0.088
2.6	-0.187		-0.055	0.110	
		0.155			0.114
2.8	-0.156		-0.040	0.083	
		0.115			0.118
3.0	-0.133				

Table 9.2

One point which is clearly illustrated in this particular example is the ability of spline functions to 'localize' the difficulties. Although the figures confirm what was thought about this particular function, nevertheless, in the range [2.2, 3.0], the approximation can be said to be quite reasonable. This means that the large errors experienced in the neighbourhood of x = 1.0, have been rapidly damped in moving away from the left-hand end of the range. This feature, which appears general in spline approximations, is one which makes spline functions an attractive approach.

CHAPTER X

GENERAL DISCUSSION

Introduction

Some attempt is made to summarize the points concerning the various methods mentioned and to compare their performance. It would be convenient if a direct answer could be given to the question that having been given a specific function, how would an approximation be found? A general approach is suggested but it is pointed out that several factors might affect what method is adopted.

Considerations Concerning Various Methods

Some of the features of the various methods described will be outlined. It seems natural to commence with interpolation forms. These are tremendously important in numerical analysis because of their use in integration and differentiation formulæ. One reason why their popularity has declined is that the difference tables often associated with them do not fit well into automatic machines. The Lagrange formula, which avoids difference tables is not a convenient expression to handle in its general form. Making use of equally-spaced nodes can improve this situation, but we have seen that it is often desirable to use points which are not equally-spaced for the best results. An alternative is to define the interpolating polynomial as a Chebyshev series, which results in about the same amount of labour as finding the coefficients of the least-squares approximation by summation over a discrete point set.

There is little doubt that interpolation form will continue to be used, particularly to derive integration formulæ , but for function evaluation it is possible to find functions of the same degree with rather better error curves.

The Hermite formula, in which the values of the function derivative are introduced into the approximation has not been demonstrated in an example. It is reasonable to suppose that this is a useful form in cases where it is important that the approximation reproduces the derivative of the function at points within the range.

Fourier series were discussed in Chapter II in relation to the least-squares approximation. This method of obtaining the approximation as a series of trigonometric terms is well known and widely applied in practical problems of curve-fitting. Unless the function under consideration is periodic in nature, a trigonometric series is likely to show slow convergence. For this reason, and the fact that trigonometric sums may not be rapidly evaluated, a Fourier series is best reserved for the approximation of periodic functions.

When considering approximations derived from series, it is worth remembering that a truncated Taylor series gives a small error if the range of fit is kept small. However, the rapid increase of the error towards the ends of the range has been noted. Also, because a series is formally convergent, does not guarantee its suitability for computation; for example consider the evaluation of e^x from its series expansion when $x = 10$. If a function has an asymptotic expansion, this can often be used profitably when the range of approximation is semi-infinite. It must be borne in mind that the truncation error cannot be made arbitrarily small so that normally a lower limit has to be set for the range over which the asymptotic series may be employed.

Padé approximants, being derived from series expansions have error distributions that look very like those of the series expansion. Being rational forms they can be expressed as continuted fractions for computational use. Modification of the Padé form to achieve a more equal distribution of the error is likely to work satisfactorily only if the range of approximation is kept reasonably small.

Of all polynomial forms, the one most likely to have the best practical use is an expansion in a series of Chebyshev polynomials. Because of the equal-oscillation form of these polynomials, Chebyshev series often possess truncation error curves which are very close to the minimum-maximum error condition. In addition, the coefficients of such series often decrease rapidly in magnitude, thus making it possible to truncate the series after only a few terms without incurring unreasonable error.

We now look at approximations in the three Holder norms, L_1, L_2 and L_∞. As was seen in Chapter VI, the L_1 approximation can often be found by a method of interpolation and that the truncation error is often nearly the form of a Chebyshev polynomial of the second kind. Now the interpolation method will not always provide the best L_1 approximation and in addition we expect the error curve to show the error increasing towards the ends of the range. For these reasons it is not considered that the L_1 norm is likely to be a normal choice in approximation problems. The one exception (Rice [13]) might be to provide an approximation to be used as an integrand, since the L_1 norm minimises the integral of the error modulus.

The L_2 norm is historically the earliest practical measure of approximation and retains its status even with the availability of massive computing power. The reason for this is that if the approximation is expressed as a series of orthogonal functions, the problem may be solved by a fast, direct and numerically stable method. Furthermore, if these orthogonal functions are chosen as Chebyshev polynomials, we expect the error curve to behave, in many cases, very like a minimax error curve. Add to this the ease of manipulation of polynomials and it would appear that any approximation problem might be solved satisfactorily by this method. The one drawback is that polynomials are essentially smooth functions and might not deal with problems where the original function has large derivatives or which may have regions of large curvature.

One way to deal with such problems would be to "stretch" the curve out by some simple transformation of the independent variable. The resultant improvement in the accuracy of a simple polynomial may be such as to outweight the extra burden of carrying out the transformation. Another approach is to adopt a rational function as an approximation. These functions lead to essentially non-linear problems for the derivation of the unknown parameters. The best approach with rational functions is to adopt the minimax norm and to determine the coefficients iteratively. Clearly, the equal-error distribution of the minimax approximation is an attractive property and would appear the best measure to adopt. There is no need, of course, to adopt a rational form when seeking a minimax approximation; polynomials being equally well suited to this type of solution. Indeed they may not be subject to the instability sometimes encountered when trying to find a rational approximation. However, empirical evidence demonstrates that for the same number of coefficients, a rational form will give a smaller maximum error than the corresponding polynomial. In order to exploit this feature fully, rational approximations are usually chosen in which the degrees of numerator and denominator are either equal or differ by one. Another point about minimax approximation is that the method of evaluating the coefficients also gives a specific figure for the maximum error. Lastly, concerning a rational function; when this is expressed as a continued fraction, it probably gives the most economical form for evaluation purposes.

On the contrary, considerable labour is involved in finding minimax approximations because of the iterative method involved. In particular, rational functions may give problems with stability, either through being unable to find sufficiently close starting values or through difficulties in locating the extrema of the error curve.

Spline function approximation is different again from the approach using the Holder norms. Clearly its closest affinity is with interpolation forms. For the purpose of function evaluation, the main problem is that the piecewise nature of the spline function is not a fast computational form. Satisfactory error estimates may not be available, although with an automatic machine it is not unreasonable to compute and print out the error curve using a fairly close mesh of points. End conditions present something of a problem. There is a general tendency with spline functions for the error to increase in magnitude towards the end of the range. Nevertheless, spline function approximations can be found directly by a convenient numerical process and their ability to "localize" perturbations produced by undesirable features is an attractive feature. This feature possibly makes them particularly suited to problems of curve fitting.

Examples of approximations found by some of the methods described are now given.

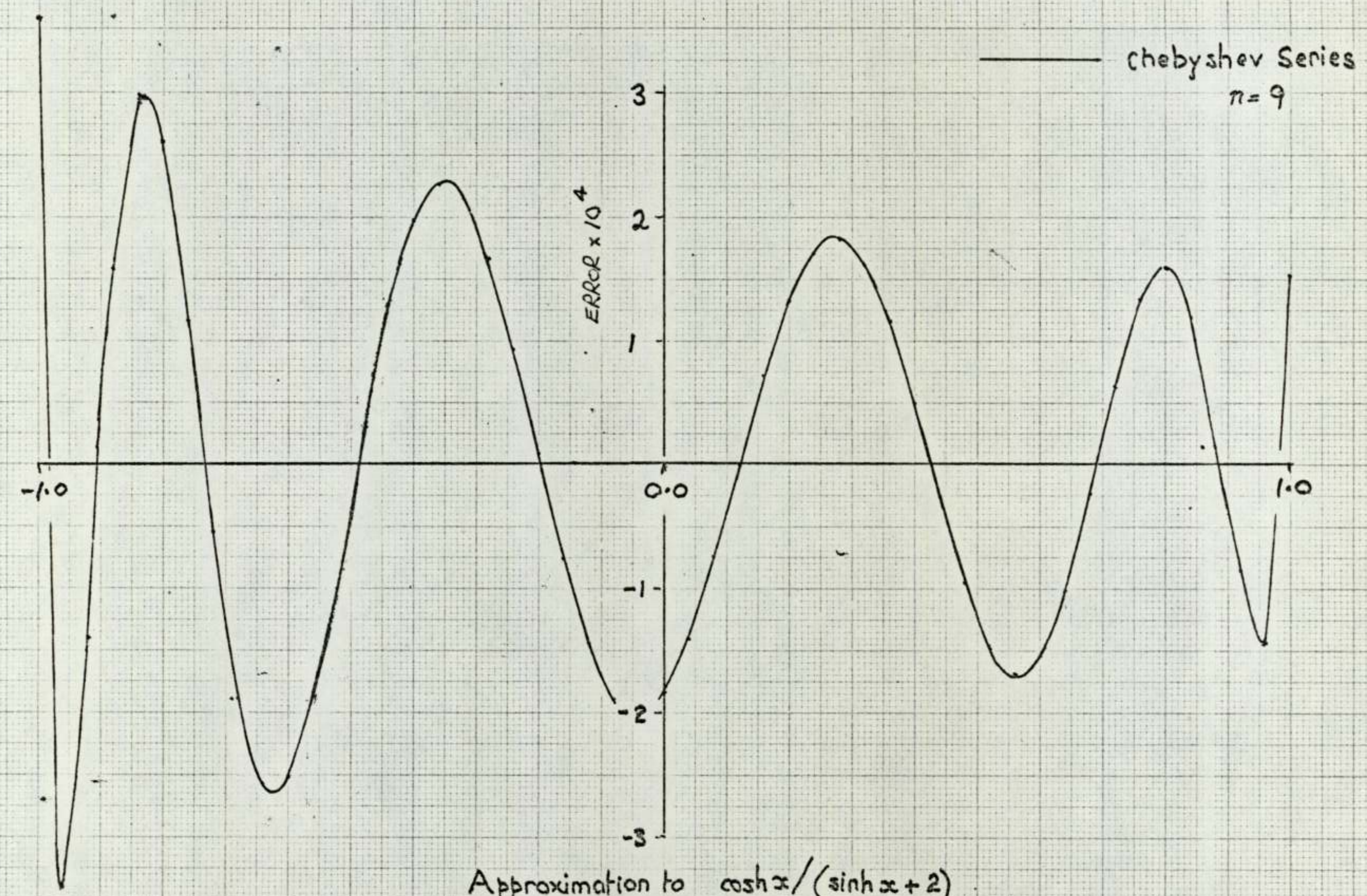
Example 1

In Chapter VII approximations were found to the function

$$\frac{x}{\left\{\sqrt{x^2+1}+x\right\}^{3}\sqrt{x^2+1}} \qquad \text{in the range } [-1,1]$$

Here a spline approximation is found using nine equally-spaced knots. In table 10.1, the details of the error are presented. It is interesting to note that with this particular function, the error is greatest in the middle of the range of approximation, which is contrary to expectation. It is noticeable that the discontinuities in the third derivative follow this trend and the error estimate, using the pxpression of Chapter IX is reliable.

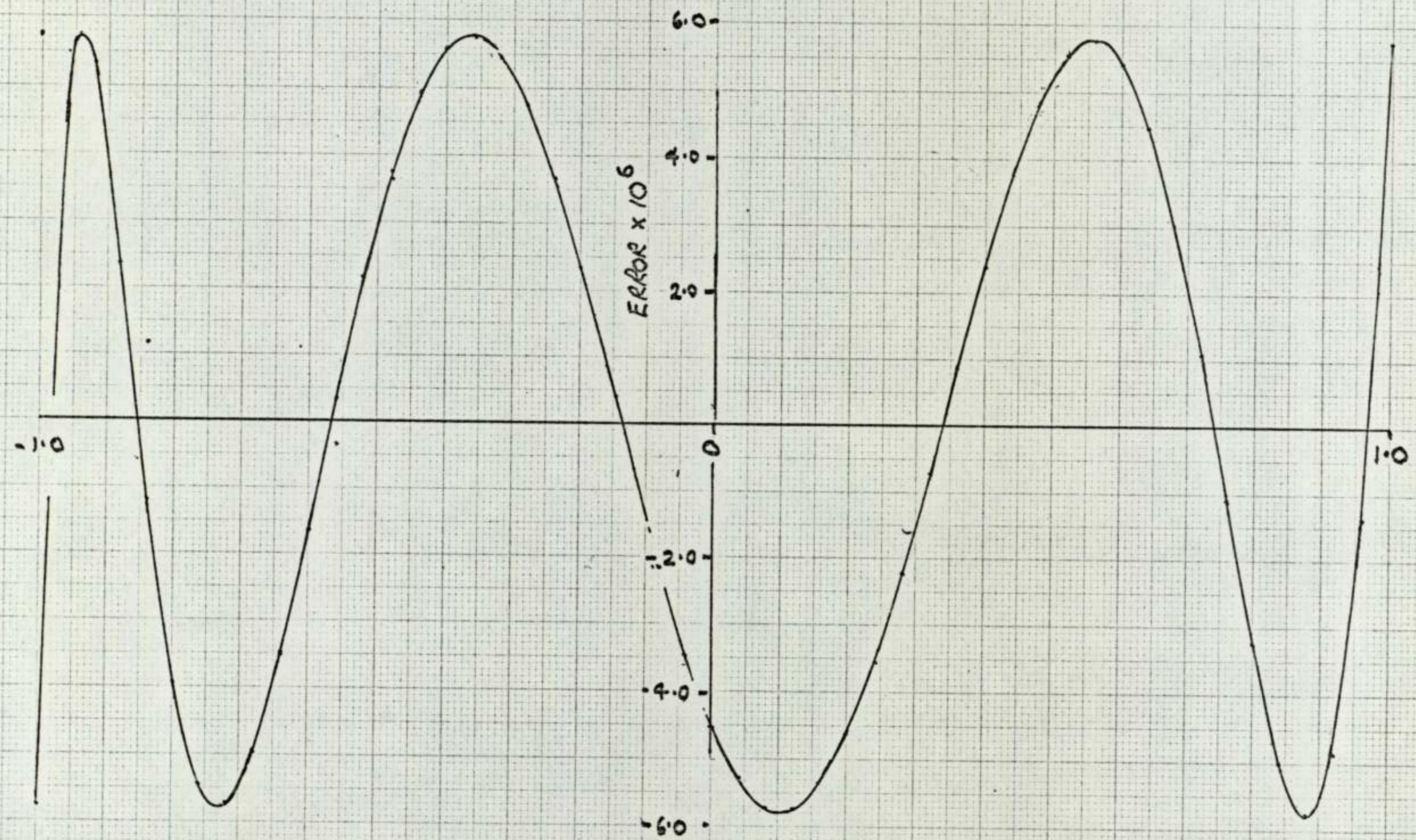
x	M	$[f''']^+_-$	Error Est. ($\times 10^5$)	Actual Error($\times 10^5$)
- 1.00	-47.88			
				5.56
- .75	-35.85	- 1.32	5.36	
				5.47
- .50	-24.15	- 4.40	17.90	
				27.64
- .25	-13.55	-10.94	44.40	
				64.11
0.00	- 5.68	-14.93	60.60	
				64.11
.25	- 1.55	-10.92	44.40	
				27.64
.50	- 0.146	- 4.45	18.10	
				5.47
.75	0.146	- 1.25	5.10	
				5.56
1.00	0.123			



Approximation to $\cosh x/(\sinh x + 2)$

Fig 10.2

P_3/Q_3 Minimax



Approximation to cosh x/(sinh x + 2)

Fig 10·3

Compared with the Chebyshev series approximation of degree nine, whose error curve has ten zeros, the maximum error is 0.00064 compared with 0.00015 for the series.

Example 2

Consider $\dfrac{\cosh x}{\sinh x + 2}$ in the range $[-1,1]$

First, a Chebyshev series of degree nine, the coefficients determined by summation over a set of discrete points, chosen as an orthogonal basis (See Chapter VII)

The approximation is

$$\begin{aligned} f_1(x) = {} & 0.787\,858 - 0.542\,979\; T_1(x) + 0.330\,826\; T_2(x) - 0.125\,213\; T_3(x) \\ & + 0.049\,993\; T_4(x) - 0.020\,329\; T_5(x) + 0.008\,163\; T_6(x) \\ & - 0.003\,283\; T_7(x) + 0.001\,322\; T_8(x) - 0.000\,532\; T_9(x) \end{aligned}$$

The error curve for $f_1(x)$ is plotted in figure 10.2. To show the advantage to be gained in using a rational function in terms of the magnitude of the maximum error, we can compare this with the approximation

$$f_2(x) = \frac{0.500\,005 - 0.074\,222x + 0.200\,353x^2 - 0.322\,750x^3}{1.000\,000 + 0.351\,506x - 0.173\,361x^2 + 0.043\,858x^3}$$

The error curve for $f_2(x)$ is plotted in figure 10.3 and we can see that although $f_2(x)$ has fewer independent coefficients, the maximum error is considerably less. The coefficients in $f_2(x)$ were determined by the iterative programme of Chapter VIII, which gave the maximum error $h = 0.576 \times 10^{-5}$

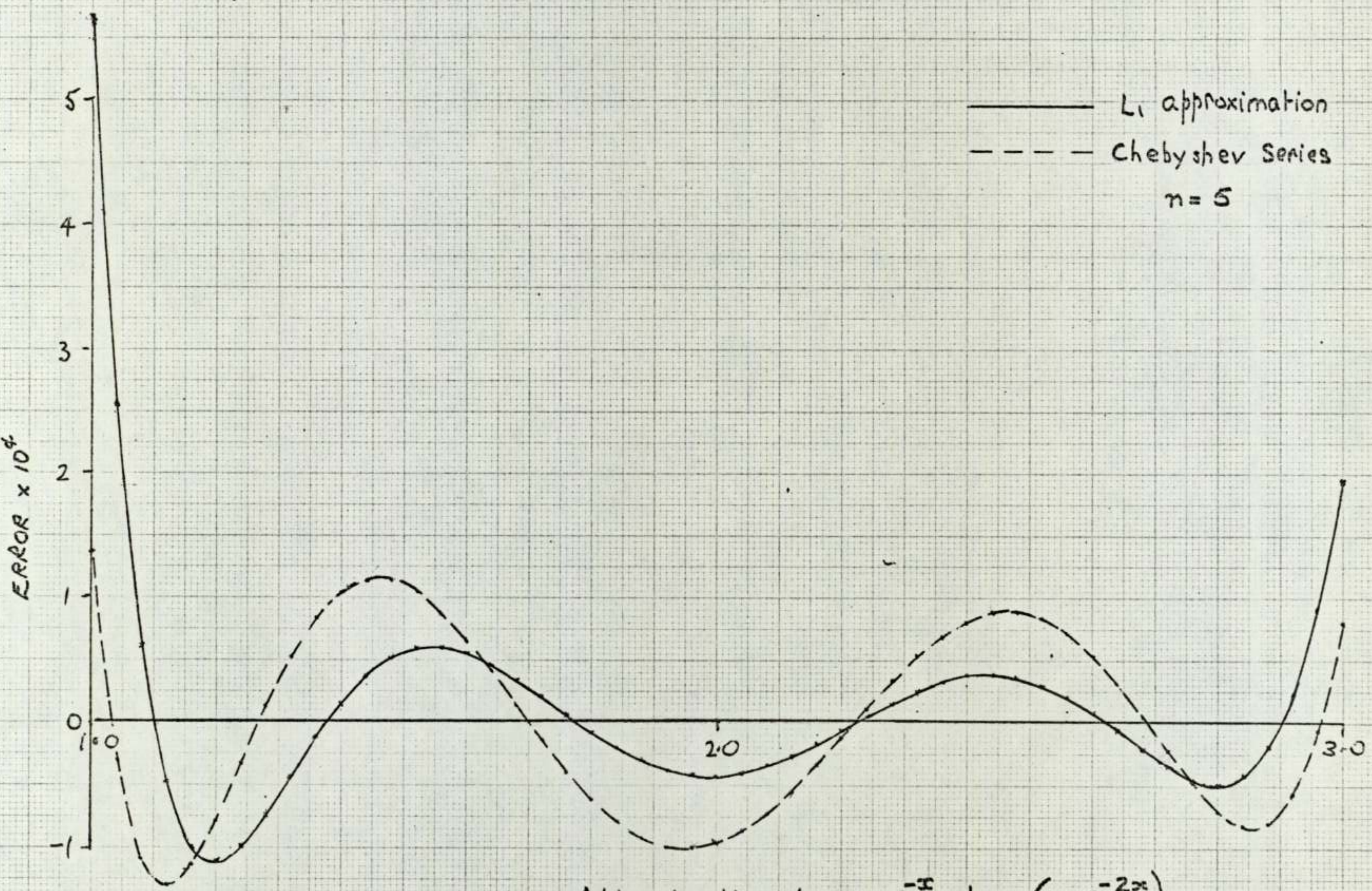
Example 3

As approximations to $x\,e^{-x} - \log_e(1 - e^{-2x})$ in the range $[1,3]$ we find a Chebyshev series and the L_1 approximation, both of degree 5. Using the method of Chapter VI, the L_1 fit is found in terms of the Chebyshev polynomials of the second kind.

i.e. $$\begin{aligned} f_1(z) = {} & 0.309\,999\,6 - 0.178\,046\,3\; T_1(z) + 0.021\,603\,3\; T_2(z) \\ & - 0.002\,309\,7\; T_3(z) + 0.000\,855\,9\; T_4(z) - 0.000\,340\,9\; T_5(z) \end{aligned}$$

where $z = x - 2$ and $-1 \leqslant z \leqslant 1$

$$\begin{aligned} f_2(z) = {} & 0.299\,197\,9 - 0.087\,868\,2\; U_1(z) + 0.010\,373\,5\; U_2(z) \\ & - 0.000\,983\,7\; U_3(z) + 0.000\,373\,8\; U_4(z) - 0.000\,144\,8\; U_5(z) \end{aligned}$$



Approximation to $xe^{-x} - \log_e(1 - e^{-2x})$

Fig 10.4

The error curves for these two expressions are plotted in figure 10.4. We notice that there are six zeros on the error curve and that, in this case, $f_2(z)$ is the best L_1 approximation of degree five. As might be expected, we notice that the L_1 fit gives errors of larger magnitude near the ends of the range and that the extremes are greater than for the ordinary Chebyshev series.

Example 4

In the range $[0,2]$, we now consider the function defined by

$$x^2y + (1-y)^2y = 1.0$$

The approximation is found in this case in three different forms. First, the Chebyshev series of degree eight was determined. Using this as the starting point, the minimax polynomial approximation of degree eight was found. Finally, to illustrate that the rational function with the same number of parameters gives a smaller maximum error, the rational approximation $P_4(x)/Q_4(x)$ was determined.

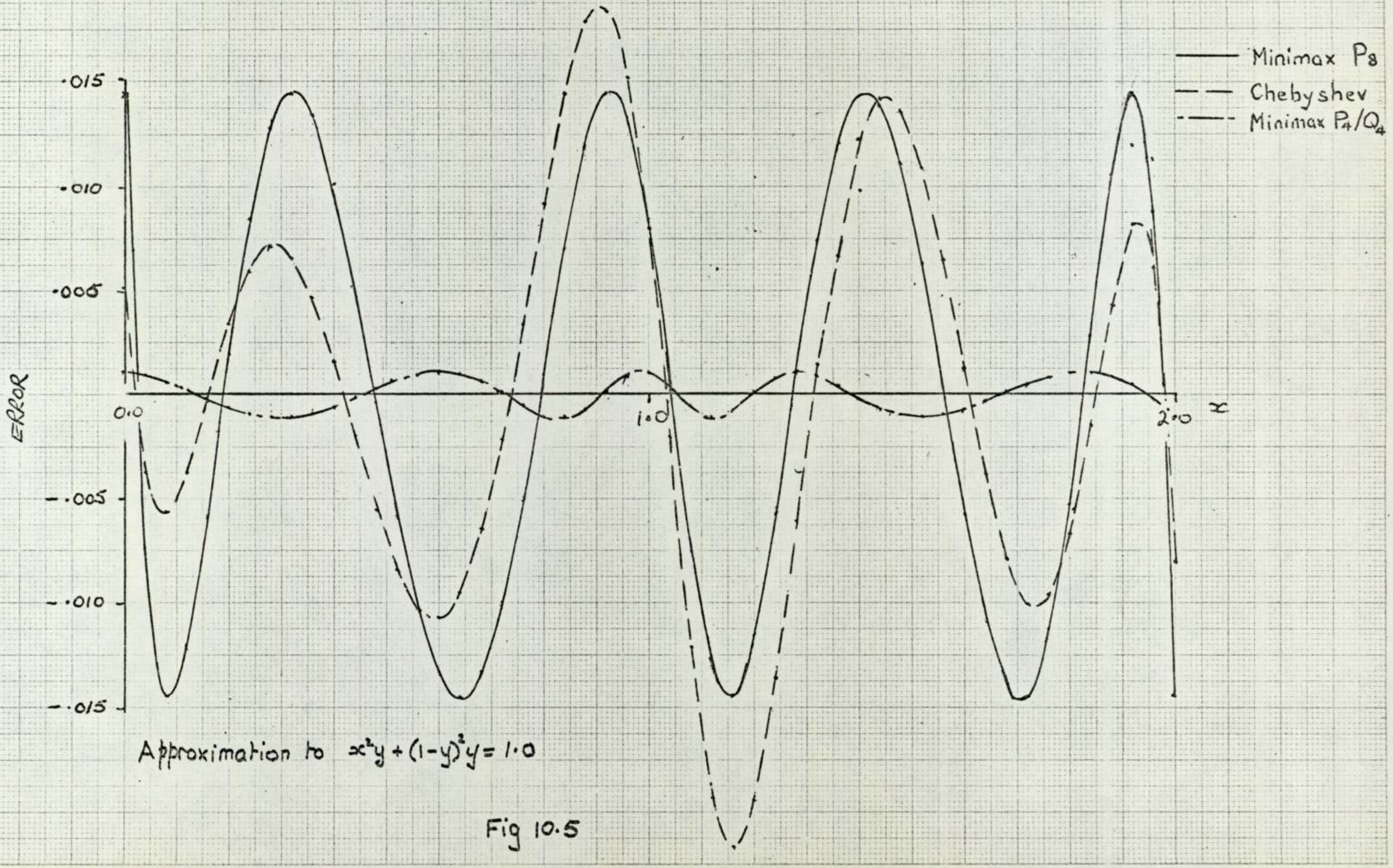
i.e. $$\begin{aligned} f_1(z) = {} & 0.992\,482 - 0.896\,245\,T_1(z) - 0.000\,133{}_2T\,(z) + 0.162\,341\,T_3(z) \\ & -0.005\,810\,T_4(z) - 0.049\,468\,T_5(z) - 0.002\,258\,T_6(z) \\ & +0.020\,943\,T_7(z) + 0.002\,920\,T_8(z) \qquad \text{where } z = x-1 \end{aligned}$$

$$\begin{aligned} f_2(x) = {} & 0.028\,738^{-1}\{0.050\,016 + 0.023\,032x - 0.222\,194x^2 + 0.675\,660x^3 \\ & - 1.000\,000x^4 + 0.677\,143x^5 - 0.174\,351x^6 - 0.008\,524x^7 \\ & + 0.007\,715x^8\} \end{aligned}$$

$$\begin{aligned} f_3(x) = {} & (-0.419\,969 + 1.000\,000x - 0.894\,927x^2 + 0.352\,376x^3 \\ & -0.054\,960x^4)/(-0.239\,461 + 0.570\,135x - 0.571\,611x^2 \\ & +0.308\,123x^3 - 0.084\,683x^4) \end{aligned}$$

The three error curves are compared in figure 10.5. The Chebyshev series is again seen to be close to the minimax error of the same degree, but the superiority of the rational function is clearly illustrated.

One final point, the rational function $f_3(x)$ can be represented by a terminating continued fraction form



Approximation to $x^2y + (1-y)^2y = 1 \cdot 0$

Fig 10.5

$$f_3(x) = 0.649\ 010 - \frac{1.799\ 66}{x - 0.200\ 596 +}\ \frac{1.926\ 69}{x - 1.274\ 93 +}\ \frac{0.089\ 059}{x - 1.094\ 418 +}$$

$$\frac{0.117\ 442}{x - 1.068\ 594}$$

In this form, the approximation can be evaluated with four divisions and no multiplications.

Example 5

Here we compare the approximations to

$$\frac{e^{x/4}}{\sqrt{1 + \frac{1}{4}(e^{x/2} - 1)}} \qquad \text{in the range } [0,4]$$

by a Chebyshev series of degree eight and interpolation through the zeros of $T_9(U)$, scaled to the given range. In figure 10.6 it can be seen that the two approximations give very nearly the same error distribution. However, the Lagrangian interpolation form, which was used by the computer programme to evaluate the figures plotted in figure 10.6, may not be considered an efficient computational form. One alternative is to express the interpolating function as a continued fraction, as in the manner of Chapter IV. The two approximations then are

$$\begin{aligned} f_1(z) &= 1.360\ 788\ 6 + 0.347\ 768\ 57\ T_1(z) - 0.017\ 788\ 06\ T_2(z) \\ &- 0.004\ 504\ 32\ T_3(z) + 0.000\ 353\ 49\ T_4(z) + 0.000\ 084\ 75\ T_5(z) \\ &- 0.000\ 006\ 87\ T_6(z) - 0.000\ 001\ 76\ T_7(z) + 0.000\ 000\ 13\ T_8(z) \end{aligned}$$

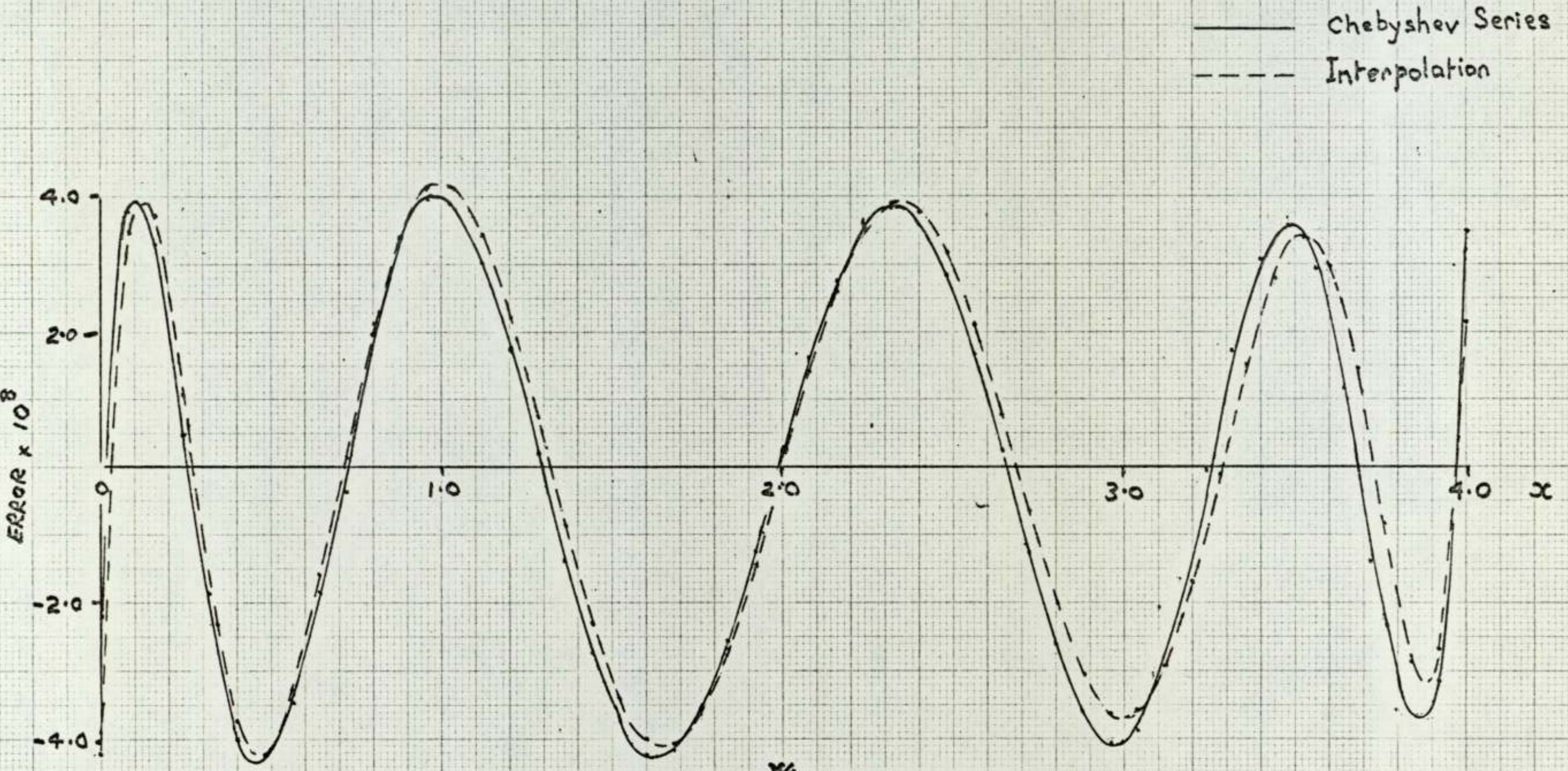
$$\text{where } z = \tfrac{1}{2}(x - 2)$$

$$f_2(x) = 1.005\ 702\ 5 + \frac{x - 0.030\ 384\ 4}{5.288\ 9142 +}\ \frac{x - 0.267\ 949\ 2}{-10.071\ 254\ 8 +}\ \frac{x - 0.714\ 424\ 8}{-0.080\ 126\ 9 +}$$

$$\frac{x - 1.315\ 959\ 6}{9.667\ 879\ 3 +}\ \frac{x - 2.000\ 000}{14.742\ 968\ 0 +}\ \frac{x - 2.684\ 040\ 4}{-1.951\ 823\ 6 +}$$

$$\frac{x - 3.285\ 575\ 2}{-1.788\ 073\ 8 +}\ \frac{x - 3.732\ 050\ 8}{4.309\ 930\ 6}$$

It must be pointed out that the coefficients in $f_2(x)$ were derived from a reciprocal difference table computed using a machine with a nine-digit quotient register. The accumulated rounding error is sufficient to make the fraction unsatisfactory for the computation of values accurate to the magnitude given in figure 10.6.



Approximation to $e^{x/4}/\sqrt{1+1/4(e^{x/2}-1)}$

Fig 10.6

General Approach to Function Approximation

When seeking an approximation, it is necessary to bear three things in mind; the use to which the approximation is to be put, the range of fit required and the accuracy that is demanded. If the approximation is to be the basis of a procedure that will be used innumerable times, it is worthwhile trying to find a fast computational form. On the other hand, if the function may be differentiated or integrated it would not be very helpful for it to be expressed say, as a continued fraction. When considering the degree of accuracy required, clearly time spent computing terms which are of no significance to the answer is wasted. However, when preparing a standard subroutine, it is often necessary to consider the most stringent demand and programme accordingly. The best method is to have several routines of varying accuracy and expect the user to choose the most appropriate [8].

When considering the type of approximation, the range of fit can be influential. Over a small range, a Taylor series or Padé approximant may prove more than adequate. However, since with these forms the error shows rapid increase towards the ends of the range, then for a moderately large range a method must be adopted which produces a more equal distribution of the error. For large ranges, it is essential to divide it up into two or more parts and adopt different approximations in the separate sections. The use of asymptotic series for regions in which the argument is large is an example of this. Another technique is to reduce the range over which the basic approximation must apply. For periodic functions, the range obviously need never exceed one complete cycle. It is possible to make use of the fact that the computer will work in floating point arithmetic to base 2, so that multiplication or division by 2 implies a change only in the integer exponent.

e.g. $\sqrt{x} = 2^k\sqrt{\frac{x}{2^{2k}}}$

and by using shifts involving changes in the exponent, the range may be reduced to $\frac{1}{4} \leqslant x \leqslant 1$

also $e^x = 2^{\frac{x}{\log_e 2}} = 2^{n+f}$ where n is integer,

$$= 2^n.2^f \qquad 0 \leqslant f < 1$$

and e^x can be computed by one division, a shift in the exponent and a function to find 2^f.

Equally $\log_e x = \log_e 2 . \log_2 x$

then if $x = 2^n.f$ where $\frac{1}{2} \leqslant f < 1$

$\log_2 x = n + \log_2 f$ and the range of approximation is reduced to $\left[\frac{1}{2},1\right]$ at the expehse of one multiplication and a shift of exponent.

Finally, is it possible to present a general method of approach to the problem of finding an approximation? Assuming that the function can be evaluated for any given argument, the process might proceed as follows.

Find the series of Chebyshev polynomials that gives the required degree of accuracy. If the error curve is to be levelled, this can be used as a starting point to find the minimax polynomial approximation.

The function that has been derived at this point may have rather a large number of terms and the next step could be taken to adopt a rational function form to reduce the number of coefficients involved for the same magnitude of error. However, the use to which the approximation is to be put may dictate that we prefer a few extra terms on the Chebyshev series to an awkward continued fraction.

As a final exercise, if no satisfaction has been found using the above methods, then subdivision of the range is appropriate. In this case, it is likely that different forms of approximation may be used in the different segments that make up the complete range.

Clearly, having proposed this approach it is possible to consider a host of reservations which are dependent on any special features that the original function might posses. However, some of these points should now have been mentioned and no attempt is made to reiterate them. The last words are given to Prof. P. J. Davis who said at Canterbury in 1967 (Hayes p. 162 [9])

"The comparison of numerical methods is like a comparison of cars. You must know what is in your pocket-book, how large your family is, what they like to do at weekends etc. Comparisons are hard to make, frequently hard to interpret but they ought to be undertaken."

Appendix A1

A.1. Proof of the Weierstrass Theorem

This proof is that given by Rice [13] and follows that of Lebesgue.

Theorem: If y(x) is continuous in [0,1] then there exists a polynomial Pn(y,x) such that for any $\mathcal{E}>0$

$$|Pn(y,x) - y(x)| < \mathcal{E} \qquad \text{for } 0 \leqslant x \leqslant 1$$

Proof: There is no loss of generality in choosing the range [0,1] since this can always be achieved by a suitable transformation.

The method involves two steps. In the first, it is established that y(x) can be approximated arbitrarily closely by a broken line b(x) and in the second that b(x) itself can be approximated by a polynomial.

For any $\mathcal{E}$, there exists $\delta > 0$, such that if $|x_2 - x_1| < \delta$

$$\text{then } |y(x_2) - y(x_1)| < \tfrac{1}{2}\mathcal{E} \qquad (1,1.1)$$

Now choose a set of points $x_1, x_2 \ldots\ldots x_m$, equally spaced in [0,1], so that $|x_j - x_{j-1}| < \delta$

Define the broken line $$b(x) = b_0 + \sum_{k=1}^{m} b_k\, \phi(x_k,x) \qquad (1,1.2)$$

$$\text{where} \qquad \phi(x_k,x) = x - x_k + |x - x_k|$$

The coefficients b_k in (1.1.2) can be chosen so that

$$b(x_k) = y(x_k) \qquad k = 1, \ldots\ldots, m$$

Then, it follows that if (1.1.1) holds,

$$|y(x) - b(x)| < \tfrac{1}{2}\mathcal{E} \text{ in } \quad 0 \leqslant x \leqslant 1 \qquad (1,1.3)$$

If we now show that each term $b_k\phi(x_k,x)$ in (1,1.2) can be approximated to within $\mathcal{E}/2m$, then the result will follow.

For $|x| \leqslant 1$ $\qquad |x| = \sqrt{1 - (1 - x^2)}$

or writing $u = (1 - x^2)$ $\qquad |x| = \sqrt{1 - u} = 1 - \tfrac{1}{2}u - \frac{1}{2^2 2!}u^2 - \frac{1.3}{2^3 3!}u^3 \ldots$

where the right-hand side is uniformly convergent in the region under consideration. Hence, there are polynomials in u (or $\overline{1 - x^2}$) which approximate to $|x|$ arbitrarily closely.

Hence it is possible to approximate to $|x - x_k|$ arbitrarily closely in $[0,1]$

$\therefore$ There exists polynomials $P_k(x)$ such that

$$|P_k(x) - b_k\phi(x_k,x)| < \varepsilon/2m$$

so $$|b(f,x) - \sum_{k=1}^{m} P_k(x)| < \varepsilon/2$$

$$\therefore \quad |y(x) - b(f,x) + b(f,x) - \sum_{k=1}^{m} P_k(x)| \leqslant \varepsilon/2 + \varepsilon/2$$

or $$|y(x) - \sum_{k=1}^{m} P_k(x)| \leqslant \varepsilon$$

Which completes the proof.

A1.2 Relationship between Lp and Minimax Approximations

~~We~~ wish to show that

$$\lim_{p \to \infty} Lp[y(x) - f(x)] = \max_{[a,b]}|y(x) - f(x)|$$

Let the approximation be of the form $f(x) = \sum_{i=1}^{n} a_i\phi_i(x)$ and without loss of generality we can choose the range of fit as $[0,1]$.

Then we shall denote the set of coefficients by $[Ap]$ such that $f(Ap,x)$ is the best approximation in the sense of the norm

$$Lp\left\{y(x) - f(Ap,x)\right\} = \left[\int_0^1 |y(x) - f(Ap,x)|^p \, dx\right]^{1/p}$$

being a minimum.

Assume that as $p \to \infty$, we can choose a subset of $[Ap]$ to form a sequence such that

$$\lim_{p\to\infty} [Ap] = [A_o]$$

Also, let $f(A_t,x)$ be the best minimax approximation to $y(x)$ in $[0,1]$

writing $$M_t = \max_{[0,1]} |y(x) - f(A_t,x)| \qquad (1,2.1)$$

$$M_o = \max_{[0,1]} |y(x) - f(A_o,x)| \qquad (1,2.2)$$

we require to prove that $M_o = M_t$

Assume contrariwise that $M_o > M_t$ and choose $\varepsilon = \dfrac{M_o - M_t}{M_t} > 0$

For some interval I in $[0,1]$, $|y(x) - f(A_o,x)| \geqslant \frac{2}{3}M_o + \frac{1}{3}M_t$

$$\text{but } M_o = M_t(1+\varepsilon)$$

$$\tfrac{2}{3}M_o + \tfrac{1}{3}M_t = M_t(1 + {}^{2\varepsilon}\!/_3)$$

$$\therefore \text{ in I } \quad |y(x) - f(A_o,x)| \geqslant M_t(1 + {}^{2\varepsilon}\!/_3) \qquad (1,2.3)$$

Now for some p_o, when $p \geqslant p_o$

$$|f(A_o,x) - f(A_p,x)| \leqslant \frac{\varepsilon M_t}{3}$$

Thus for $p \geqslant p_o$ in the interval I

$$|y(x) - f(Ap,x)| + |f(Ap,x) - f(A_o,x)| \geqslant |y(x) - f(Ap,x) + f(Ap,x) - f(A_o,x)|$$

hence $|y(x) - f(Ap,x)| + \dfrac{\varepsilon M_t}{3} \geqslant M_t(1 + \frac{2}{3}\varepsilon)$

$$\text{or } \quad |y(x) - f(Ap,x)| \geqslant M_t(1 + \frac{\varepsilon}{3}) \qquad (1,2.4)$$

If m is the length of the interval I

$$\left[\int_0^1 |y(x) - f(Ap,x)|^p \, dx\right]^{1/p} \geqslant \left[\int_I |y(x) - f(Ap,x)|^p \, dx\right]^{1/p}$$
$$\geqslant M_t(1 + \frac{\varepsilon}{3})m^{1/p} \qquad (1,2.5)$$

But $m^{1/p} \to 1$, as $p \to \infty$, and since

$$\left[\int_0^1 |y(x) - f(A_t,x)|^p \, dx\right]^{1/p} \leqslant M_t \qquad (1,2.6)$$

(1,2.5) and (1,2.6) imply that $f(Ap,x)$ cannot be the best Lp approximation hence M_o and M_t must be equal.

Appendix A2

A2.1 Properties of Chebyshev Polynomials of the First Kind

Recurrence of Relation

If $T_n(x) = \cos(n \cos^{-1} x)$

writing $\cos\theta = x$

then $T_n(\cos\theta) = \cos n\theta$ (2,1.1)

clearly $T_0(x) = 1$ and $T_1(x) = x$

but $\cos n\theta \cos m\theta = \frac{1}{2}\left[\cos(n+m)\theta + \cos(n-m)\theta\right]$

or $T_n(x)\, T_m(x) \doteqdot \frac{1}{2}\left[T_{n+m}(x) + T_{n-m}(x)\right]$ (2,1.2)

setting $m = 1$ and noting that $T_1(x) = x$, and rearranging

$$T_{n+1}(x) - 2x\, T_n(x) + T_{n-1}(x) = 0 \quad \text{when } n > 0 \qquad (2,1.3)$$

Integral of $T_n(x)$

Using the identity $\cos m\theta \sin\theta = \frac{1}{2}\left[\sin(m+1)\theta - \sin(m-1)\theta\right]$ and integrating with respect to θ

$$\int \cos m\theta \sin\theta \, d\theta = \frac{1}{2}\left[\frac{\cos(m+1)\theta}{m+1} - \frac{\cos(m-1)\theta}{m-1}\right] \quad m \nleq 1$$

or putting $x = \cos\theta$

$$\int T_m(x)\, dx = \frac{1}{2}\left[\frac{T_{m+1}(x)}{m+1} - \frac{T_{m-1}(x)}{m-1}\right] \quad m \nleq 1 \qquad (2,1.4)$$

and $\int T_0(x)\, dx \doteqdot T_1(x)$

$$\int T_1(x)\, dx = \frac{1}{4}\left[T_2(x) + T_0(x)\right]$$

Multiplication of Powers of x

Start with the trigonometric identity

$$\cos^n\theta = \frac{1}{2^{n-1}} \sum_{k \doteqdot 0}^{2k \leq n} {}^nC_k \cos(n-2k)\theta$$

Putting $x = \cos\theta$ yields

$$x^n = \frac{1}{2^{n-1}} \sum_{k=0}^{2k \leq n} {}^nC_k T_{(n-2k)}(x)$$

$$\therefore \quad x^n T_m(x) = \frac{1}{2^{n-1}} \left\{ \sum_{k=0}^{2k \leq n} {}^nC_k T_{(n-2k)}(x) \right\} T_m(x)$$

Then using (2,1.2)

$$T_{(n-2k)}(x)T_m(x) = \tfrac{1}{2}\left[T_{(m+n-2k)}(x) + T_{(m-n+2k)}(x)\right]$$

$$\therefore \quad x^nT_m(x) = \frac{1}{2^n}\sum_{k=0}^{2k\leqslant n} {}^nC_k\left\{T_{(m+n-2k)}(x) + T_{(m-n+2k)}(x)\right\}$$

Writing $n - j = k$, then the first term on the right-hand side becomes

$$\sum_{k=0}^{2k\leqslant n} {}^nC_kT_{(m+n-2k)}(x) = \sum_{k=0}^{2k\leqslant n} {}^nC_k \cos(m+n-2k)\theta$$

$$= \sum_{2j\geqslant n}^{n} {}^nC_{n-j}\cos(m-n+2j)\theta$$

$$= \sum_{2j\geqslant n}^{n} {}^nC_j \; T_{|(m-n+2j)|}(x) \qquad \text{since } {}^nC_{n-j} = {}^nC_j$$

Hence the right-hand side represents summation over 0 to n

$$\text{i.e.} \quad x^nT_m(x) = \frac{1}{2^n}\sum_{k=0}^{n} {}^nC_k \; T_{|(m-n+2k)|}(x) \qquad (2,1.5)$$

Zeros of $T_n(x)$

From (2,1.1), the zeros of $T_n(x)$ are the zeros of $\cos n\theta$ in $[0,n\pi]$

$$\text{i.e.} \quad n\theta = \left(\frac{\pi}{2}, \frac{3\pi}{2} \text{ etc.}\right)$$

$$\text{or} \quad x_k = \cos\left(\frac{2k-1}{2n}\right)\pi \qquad k = 1, 2, \ldots\ldots n \qquad (2,1.6)$$

There are n roots in $[-1,1]$ hence all the roots of $T_n(x)$ are real and lie in $[-1,1]$

Orthogonality

For the orthogonality property, we have

$$\int_{-1}^{1} \frac{T_n(x)\;T_m(x)}{\sqrt{1-x^2}}\,dx = \int_0^{\pi} \cos n\theta \cos m\theta \, d\theta = 0 \qquad \text{if } m \neq n$$

$$\int_{-1}^{1} \frac{T_0^2(x)}{\sqrt{1-x^2}}\,dx = \pi$$

$$\int_{-1}^{1} \frac{T_n^2(x)}{\sqrt{1-x^2}}\,dx = \frac{\pi}{2} \qquad n \neq 0 \qquad (2,1.7)$$

The first few Chebyshev polynomials $T_n(x)$ are

$$T_0(x) = 1$$

$$T_1(x) = x$$

$$T_2(x) = 2x^2 - 1$$

$$T_3(x) = 4x^3 - 3x$$

We notice that they are alternately odd and even functions and that they are not periodic in the argument x.

Orthogonality over Discrete Point Sets

The Chebyshev polynomials are orthogonal over certain discrete point sets when summation is used instead of integration.

i.e. $\sum_{k=0}^{N}{}'' \cos m\theta_k \cos n\theta_k = 0$ for $\theta_k = \frac{k\pi}{N}$ if $m \neq n$ $k = 0,1 \ldots\ldots N$

and $\sum_{k=0}^{N}{}'' \cos^2 n\theta_k = \frac{1}{2}N$ if $n \neq 0$ or N

$= N$ if $n = 0$ or N

The double prime indicates that the first and last terms of the summation are halved.

Then an approximation to y(x) has the form

$$f_m(x) = \sum_{r=0}^{m}{}' a_r T_r(x) \qquad (2,1.8)$$

$$\text{where} \quad a_r = \frac{2}{N} \sum_{k=0}^{N} y(x_k) T_r(x_k) \qquad x_k = \cos\left(\frac{k\pi}{N}\right) \qquad (2,1.9)$$

and $k = 0,1, \ldots\ldots N$

The single prime indicates that the first term is halved. The degree of the approximation m is less than N, otherwise an interpolation formula results.

Equally we find that

$$f_m(x) = \sum_{r=0}^{m}{}' b_r T_r(x) \quad \text{is an approximation to } y(x) \qquad (2,1.10)$$

$$\text{where} \quad a_r = \frac{2}{N+1} \sum_{k=0}^{N} y(x_k) T_r(x_k) \qquad (2,1.11)$$

and the discrete points $x_k = \cos\left(\frac{2k+1}{N+1}\right)\frac{\pi}{2}$ $k = 0,1, \ldots\ldots N$

A2.2 Properties of Chebyshev Polynomials of the Second Kind

Orthogonality

From $\int_0^{\pi} \sin m\theta \sin n\theta d\theta = 0 \qquad \text{if } m \neq n$

writing $U_{m-1}(x) = \dfrac{\sin m\theta}{\sin\theta} \qquad \text{where } x = \cos\theta$ (2,2.1)

then $\int_{-1}^{1} U_{m-1}(x)\, U_{n-1}(x)\sqrt{1-x^2}\, dx = 0 \qquad \text{if } m \neq n$

Also $\int_{-1}^{1} \sqrt{1-x^2}\, U_{m-1}^2(x)\, dx = \int_0^{\pi} \sin^2 m\theta\, d\theta = \frac{\pi}{2}$

and the functions $U_n(x)$ are seen to be orthogonal over $[-1,1]$ with respect to a weight function $\sqrt{1-x^2}$

Recurrence Relation

Starting with the identity

$$\frac{\sin(n+1)\theta}{\sin\theta} + \frac{\sin(n-1)\theta}{\sin\theta} = \frac{2\sin n\theta . \cos\theta}{\sin\theta}$$

we have

$$U_n(x) + U_{n-2}(x) = 2x\, U_{n-1}(x)$$

or $$U_n(x) = 2x\, U_{n-1}(x) - U_{n-2}(x) \qquad (2,2.2.)$$

Since from (2,2.1), we have that

$$U_0(x) = 1 \text{ and } U_1(x) = 2x$$

it follows from (2,2.2) that $U_n(x)$ is a polynomial of degree n in x

Integral of $U_n(x)$

Now $$T_n(x) = \cos(n\cos^{-1}x)$$

$$T_n'(x) = \frac{n\sin(n\cos^{-1}x)}{\sqrt{1-x^2}} = \frac{n\sin n\theta}{\sin\theta} \qquad \text{if } x = \cos\theta$$

or $$T_n'(x) = nU_{n-1}(x)$$

and $$\int U_{n-1}(x)dx = \frac{T_n(x)}{n} \qquad \text{if } n \geqslant 2 \qquad (2,2.3)$$

Relation between $T_n(x)$ and $U_n(x)$

Since $\sin(n+1)\theta - \sin(n-1)\theta = 2\cos n\theta \sin\theta$

then $\dfrac{\sin(n+1)\theta}{\sin\theta} - \dfrac{\sin(n-1)\theta}{\sin\theta} = 2\cos n\theta$

hence $U_n(x) - U_{n-2}(x) = 2T_n(x) \qquad \text{if } n > 1$ (2,2.4)

with $$U_1(x) = 2T_1(x)$$

$$U_0(x) = T_0(x)$$

Appendix A3.

A3.1 Recurrence Relations for Continued Fractions

(i) Let the value of the continued fraction be $\frac{A_n}{B_n}$ when it is truncated after the term $+\frac{a_n}{b_n}$ (termed the n^{th} convergent)

Then A and B can be generated by the recurrence

$$Y_{j+1} = b_{j+1} Y_j + a_{j+1} Y_{j-1} \qquad \text{for } j = 0,1 \ldots\ldots(n-1) \qquad (3.1.1)$$

with $A_{-1} = B_0 = 1, \quad A_0 = b_0, \quad B_{-1} = 0$

Proof If $f(x) = b_0 + \frac{a_1}{b_1+} \frac{a_2}{b_2+} \cdots\cdots\cdots \frac{a_n}{b_n}$ (3.1.2)

Then $\frac{A_0}{B_0} = \frac{b_0}{1} \qquad \frac{A_1}{B_1} = \frac{b_1 b_0 + a_1 .1}{b_1 .1 + 0} = b_0 + \frac{a_1}{b_1}$

Let (3.1.1) be true for n, then

$$\frac{A_{n+1}}{B_{n+1}} = b_0 + \frac{a_1}{b_1+} \cdots\cdots\cdots + \frac{a_n}{b_n + \frac{a_{n+1}}{b_{n+1}}}$$

$$= b_0 + \frac{a_1}{b_1+} \cdots\cdots\cdots \frac{a_n b_{n+1}}{b_n b_{n+1} + a_{n+1}}$$

$$\therefore \frac{A_{n+1}}{B_{n+1}} = \frac{(b_n b_{n+1} + a_{n+1})A_{n-1} + a_n b_{n+1} A_{n-2}}{(b_n b_{n+1} + a_{n+1})B_{n-1} + a_n b_{n+1} B_{n-2}}$$

$$= \frac{b_{n+1}(b_n A_{n-1} + a_n A_{n-2}) + a_{n+1} A_{n-1}}{b_{n+1}(b_n B_{n-1} + a_n B_{n-1}) + a_{n+1} B_{n-1}}$$

$$= \frac{b_{n+1} A_n + a_{n+1} A_{n-1}}{b_{n+1} B_n + a_{n+1} B_{n-1}}$$

Hence, by induction,(3.1.1) is true for all j.

(ii) The fraction (3,1.2) can also be evaluated by calculating the difference between one convergent and the next

Using the identity $D_k = A_k B_{k-1} - B_k A_{k-1}$

$$= (b_k A_{k-1} + a_k A_{k-2})B_{k-1} - (b_k B_{k-1} + a_k B_{k-2})A_{k-1}$$

$$= a_k (B_{k-1} A_{k-2} - A_{k-1} B_{k-2})$$

$$= - a_k D_{k-1} \qquad (3.1.3)$$

continuing in this way $D_j = (-1)^{j-1} a_j a_{j-1} \ldots\ldots a_1$

Now $$\frac{A_n}{B_n} - \frac{A_{n-1}}{B_{n-1}} = \frac{B_{n-1}A_n - B_nA_{n-1}}{B_nB_{n-1}}$$

$$= \frac{D_n}{B_{n-1}B_n}$$

hence $$\frac{A_n}{B_n} = \frac{A_{n-1}}{B_{n-1}} + \frac{(-a_nD_{n-1})}{B_{n-1}B_n} \qquad (3.1.4)$$

At each stage D_n and B_n need to be evaluated.

(iii) The evaluation of (3.1.2) can also be expressed as the summation of a series

from (3.14.) $f_n - f_{n-1} = \dfrac{D_n}{B_{n-1}B_n} = \dfrac{(-1)^{n-1}a_1a_2\ldots\ldots a_n}{B_{n-1}B_n}$

define $$\rho_j = \frac{-a_jB_{j-2}}{B_j} \quad \text{for } j \geqslant 2 \quad \rho_1 = \frac{a_1}{B_1} \qquad (3.1.5)$$

then $$f_n - f_{n-1} = \rho_1\rho_2 \cdots\cdots \rho_n$$

and $$1 + \rho_j = \frac{B_j - a_jB_{j-2}}{B_j} = \frac{b_jB_{j-1}}{B_j}$$

$$= \frac{b_j}{b_j + a_j \dfrac{B_{j-2}}{B_{j-1}}}$$

or $$1 + \rho_j = \frac{b_{j-1}b_j}{b_{j-1}b_j + a_j(1+\rho_{j-1})} . \quad \text{for } j > 2 \qquad (3.1.6)$$

and $\rho_1 = \dfrac{a_1}{b_1}$, $\quad 1 + \rho_2 = \dfrac{b_1b_2}{b_1b_2 + a_2}$

but $$f_n = f_0 + (f_1 - f_0) + (f_2 - f_1) \ldots\ldots + (f_n - f_{n-1})$$

$$= b_0 + \rho_1 + \rho_1\rho_2 + \ldots\ldots\ldots + \rho_1\rho_2 \cdots \rho_n$$

or $$f_n = b_0 + \sum_{i=1}^{n} u_i \qquad \text{where } u_i = \rho_1\rho_2 \cdots\cdots \rho_i$$

(3.1.7)

A3.2 The equivalence Transformation for Continued Fractions

A continued fraction is unchanged if some partial numberator a_j and partial denominator b_j, along with the immediately succeeding partial numerator are multiplied by the same non-zero constant.

Proof

Let a_j, b_j become ka_j, kb_j, then since $\frac{A_{j-1}}{B_{j-1}}$ is unaffected

$$\frac{A'_j}{B'_j} = \frac{kb_j A_{j-1} + ka_j A_{j-2}}{kb_j B_{j-1} + ka_j B_{j-2}} = \frac{A_j}{B_j}$$

$$\frac{A'_{j+1}}{B'_{j+1}} = \frac{b_{j+1}(kb_j A_{j-1} + ka_j A_{j-2}) + a_{j+1} A_{j-1}}{b_{j+1}(kb_j B_{j-1} + ka_j B_{j-2}) + a_{j+1} B_{j-1}} = \frac{A_{j+1}}{B_{j+1}} \quad \underline{\text{if}} \quad a_{j+1} \text{ becomes } ka_{j+1}$$

For the summation form

$$1 + \rho'_j = \frac{kb_{j-1} b_j}{kb_{j-1} b_j + ka_j(1 + \rho_{j-1})} = 1 + \rho_j$$

$$1 + \rho'_{j+1} = \frac{kb_j b_{j+1}}{kb_j b_{j+1} + a_{j+1}(1 + \rho_j)} = 1 + \rho_{j+1} \quad \underline{\text{if}} \quad a_{j+1} \text{ becomes } ka_{j+1}$$

Similarly, for the backward recurrence form, there will be a sub-calculatio calculation

$$+ \frac{a_j}{b_j + \frac{a_{j+1}}{q}} \qquad \text{where q is the value of the tail.}$$

If a_j, b_j and a_{j+1} are all multiplied by k, the value of this quotient will remain unaltered.

A3.3 The Convergence of Certain Continued Fractions

Certain criteria can be established for continued fractions evaluted in the form (3.1.7)

(i) Consider $F = b_0 + \frac{a_1}{b_1 +} \frac{a_2}{b_2 +} \ldots\ldots\ldots$ where all $a_k, b_k > 0$

For the series in (3.1.7) to converge, the ratio test gives

$$\lim_{n \to \infty} \left| \frac{u_n}{u_{n-1}} \right| = \lim_{n \to \infty} \left| \rho_n \right| < 1$$

But since all the coefficients are positive

$$1 + \rho_2 = \frac{b_1 b_2}{b_1 b_2 + a_2} < 1$$

$\therefore$ ρ_2 will be negative although $1 + \rho_2$ is positive.

Now $1+\rho_n = \dfrac{b_{n-1}b_n}{b_{n-1}b_n + a_n(1+\rho_{n-1})}$ $n \geqslant 3$

when $n = 3$, $1 + \rho_3$ will be positive and less than 1. Hence by induction the same is true for all n, and ρ_n is negative for $n \geqslant 2$.

$$\text{i.e.} \qquad = \frac{a_1}{b_1} > 0, \qquad \rho_n < 0 \qquad \text{for } n \geqslant 2 \tag{3.3.1}$$

This means that the series will alternate in sign and a necessary and sufficient condition for convergence is

$$\lim_{n \to \infty} u_n = 0$$

To show that even and odd convergents form monotonic sequences, we write

$$F_{2n} - F_{2n-2} = u_{2n-1} + u_{2n}$$

$$= (1+\rho_{2n})\rho_1\rho_2 \;\cdots\cdots\; \rho_{2n-1}$$

but all ρ_k are negative apart from ρ_1 and $|\rho_{2n}| < 1$

$\therefore$ $F_{2n} > F_{2n-2}$ and the even convergents form an increasing sequence.

Equally

$$F_{2n+1} - F_{2n-1} = (1+\rho_{2n+1})\rho_1\rho_2 \;\cdots\cdots\; \rho_{2n} < 0$$

$\therefore$ $F_{2n+1} < F_{2n-1}$ and the odd convergents form a decreasing sequence.

Finally

$F_{2n} - F_{2n-1} = \rho_1\rho_2 \;\cdots\cdots\; \rho_{2n} < 0$ since an odd number of ρ_j are negative.

$$\therefore \quad F_{2n} < F_{2n-1}$$

and if the even convergents approach a limit L_0 and the odd convergents a limit L_1, then

$$L_0 \leqslant L_1 \tag{3.3.2}$$

(ii) The fraction to be considered next is of the form

$$F_1 = \frac{1}{b_1+} \quad \frac{1}{b_2+} \quad \cdots\cdots\cdots \qquad \text{where } b_k > 0$$

let $|E_n| = |F_n - F_{n-1}| = |u_n| = |\rho_1\rho_2\cdots\cdots\rho_n|$

but from (3.1.5)

$$\rho_1 = \frac{1}{B_1} \qquad \rho_j = -\frac{B_{j-2}}{B_j}$$

$$\therefore \quad |E_n| = \frac{1}{B_{n-1}B_n}$$

Now $\qquad B_2 = b_2B_1 + B_0 \quad$ since all $a_j = 1$

$$= b_2b_1 + 1 < (1 + b_1)(1 + b_2)$$

let $\qquad B_k < (1 + b_1)(1 + b_2) \quad \ldots\ldots, \quad (1 + b_k)$

$$B_{k+1} = b_{k+1}B_k + B_{k-1}$$

$$< b_{k+1}(1 + b_1) \ldots\ldots (1 + b_k) + (1 + b_1) \ldots\ldots\ldots (1 + b_{k-1})$$

$$= (1 + b_1) \ldots\ldots\ldots\ldots (1 + b_{k-1})\left[b_{k+1}(1 + b_k) + 1\right]$$

$$B_{k+1} < (1 + b_1) \ldots\ldots\ldots\ldots (1 + b_{k-1})(1 + b_k)(1 + b_{k+1})$$

hence, by induction, we have

$$B_n < (1 + b_1)(1 + b_2) \ldots\ldots\ldots (1 + b_n)$$

and $$|E_n| < \frac{1}{[(1 + b_1) \ldots\ldots (1 + b_{k-1})]^2(1 + b_k)}$$

It can be shown [Knopp: Theory and Application of Infinite Series, Ch. 7] that $\prod_{k=1}^{\infty}(1 + b_k)$ converges if and only if $\sum_{k=1}^{\infty} b_k$ converges.

$\therefore$ If $\lim_{n\to\infty}|E_n| = 0$ then $\sum_{k=1}^{\infty} b_k$ must diverge. (3.3.3)

To prove (3.3.3) is a sufficient condition, let $\sum_{k=1}^{\infty} b_k$ be divergent

By repeated application of the recurrence relation

$$B_{2k+1} = b_{2k+1}B_{2k} + b_{2k-1}B_{2k-2} \ldots\ldots\ldots + b_1B_0$$

$$> (b_{2k+1} + b_{2k-1} + \ldots\ldots + b_1)\,B_0$$

since the B_k's form an increasing sequence.

Similarly $B_{2k} > (b_{2k} + b_{2k-2} + \ldots\ldots\ldots + b_2)B_1$

$$\therefore \; |E_{2k+1}| = \frac{1}{B_{2k}B_{2k+1}} < \frac{1}{(b_{2k+1} + b_{2k-1} + \ldots + b_1)(b_{2k} + b_{2k-2} \ldots b_2)B_1B_0}$$

and by virtue of the divergence of $\Sigma\, b_k$, the right-hand side of the inequality can be made as small as required.

(iii) Fractions are now considered in which the partial numerators are negative. By virtue of the equivalence transformation, it is sufficient to consider the two forms

$$F_2 = \frac{1}{b_1 -}\ \frac{1}{b_2 -} \ \ldots\ldots \qquad \text{where } b_k > 0$$

$$F_3 = \frac{a_1}{1-}\ \frac{a_2}{1-} \ \ldots\ldots \qquad \text{where } a_k > 0$$

In F_2
$$1 + \rho_j = \frac{b_{j-1} b_j}{b_{j-1} b_j - (1 + \rho_{j-1})}$$

$$\therefore \quad \frac{\rho_j}{1 + \rho_j} = \frac{1 + \rho_{j-1}}{b_{j-1} b_j}$$

Now if F_2 is convergent $|\rho_j| < 1$ for some $j \geqslant N$. So the right-hand side will be positive and so will the denominator on the left.

Hence the ρ_j are all positive for large enough j and all terms of the series for the evaluation of F_2 will have the same sign.

Equally for F_3
$$1 + \rho_j = \frac{1}{1 - a_j(1 + \rho_{j-1})}$$

$$\therefore \quad \rho_j = \frac{a_j(1 + \rho_{j-1})}{1 - a_j(1 + \rho_{j-1})}$$

$$\frac{\rho_j}{1+\rho_j} = \frac{a_j(1 + \rho_{j-1})}{1}$$

and, as before ρ_j will be positive for all $j \geqslant N$ and the series will be one-signed for large enough values of j.

Both series are convergent if $\lim\limits_{n \to \infty} |\rho_n| < 1$ by the ratio test. (3.3.4)

but for F_2 $$|\rho_n| = \left| \frac{1}{1 - \frac{1}{b_{n-1} b_n}(1+\rho_{n-1})} - 1 \right|$$

or $$0 < \frac{1}{1 - \frac{1(1+\rho_{n-1})}{b_{n-1} b_n}} < 2 \qquad \text{for } n \geqslant N$$

From R.H. inequality
$$1 < 2 - \frac{2(1 + \rho_{n-1})}{b_{n-1} b_n}$$

from (3.3.4) $$2(1+\rho_{n-1}) < b_{n-1} b_n$$

from (3.3.4) $$4< b_{n-1} b_n$$

$$2< b_n \qquad (3.3.5)$$

clearly, this will also satisfy the L.H. inequality.

Finally, if $b_k = 2$ in F_2

$$u_n = \rho_1\rho_2\cdots\rho_n = \frac{1}{B_{n-1}B_n} = \frac{1}{n(n+1)} \qquad \text{from (3.1.5)}$$

$$\therefore \quad F_2 = \frac{1}{2} + \frac{1}{6} + \frac{1}{12}$$

$$= (1 - \tfrac{1}{2})+(\tfrac{1}{2} - \tfrac{1}{3})+(\tfrac{1}{3} - \tfrac{1}{4}) + \ldots\ldots.$$

$$= \lim_{n\to\infty}(1- \frac{1}{n}) = 1$$

and F_2 is convergent if $b_k \geqslant 2$ for $k \geqslant N$ where N (3.3.6)

is some non-negative integer.

By a similar argument for F_3

$$0 < \frac{1}{1-a_n(1+\rho_{n-1})} < 2$$

from R.H. inequality

$$2a_n(1+\rho_{n-1})<1$$

the worst possible case would be $\rho=1$

$$\therefore \quad a_n < \frac{1}{4} \qquad \text{for } n \geqslant N$$

and this also satisfies the L.H. inequality.

If now, $a_n = \frac{1}{4}$

$$F_3 = \frac{\frac{1}{4}}{1 -}\;\frac{\frac{1}{4}}{1 -} \quad \ldots\ldots.$$

$$u_n = \frac{1}{B_{n-1}B_n} = \frac{1}{2n(n+1)}$$

$$\therefore \quad F_3 = \tfrac{1}{2}\left[\frac{1}{2} + \frac{1}{6} + \frac{1}{12} + \ldots\ldots..\right]$$

$$F_3 = \frac{1}{2}$$

and F_3 is convergent if $a_k \leqslant \frac{1}{4}$ for $k \geqslant N$ (3.3.7)

A3.4 Estimation of Truncation Error

(i) Let the truncation error in the summation form be

$$R_n = \sum_{j=n+1}^{\infty} u_j$$

$$= u_{n+1}(1 + \rho_{n+2} + \rho_{n+2}\rho_{n+3} + \cdots\cdots)$$

$$R_n \leqslant |u_{n+1}|\,(1 + |\rho_{n+2}| + |\rho_{n+2}\rho_{n+3}| + \cdots\cdots)$$

Now if $|\rho_{n+1}| > |\rho_{n+2}| > |\rho_{n+3}| \cdots\cdots$

$$R_n < |u_{n+1}|\,(1 + |\rho_{n+1}| + |\rho_{n+1}^2| + |\rho_{n+1}^3| + \cdots\cdots)$$

$$R_n < \frac{|u_{n+1}|}{1 - |\rho_{n+1}|} \tag{3.4.1}$$

(ii) For fractions of the form F_2 and F_3, it is necessary to quote the following results [see Blanch [2]].

If F_2 and, F_2' are fractions with terms b_k and b_k' such that $b_k' > b_k \geq 2$ for at least one value of k, then $F_2' < F_2$ (3.4.2)

If F_3 and F_3' are fractions with terms a_k and a_k' such that $0 < a_k' < a_k \leqslant \frac{1}{4}$, for at least one value of k, then $F_3' < F_2$ (3.4.3)

It is now possible to produce estimates for the truncation errors in F_2' and F_3'

For F_2', let

$$R_n = \frac{1}{b_{n+1}-}\;\frac{1}{b_{n+2}-}\;\cdots\cdots \quad \text{with } b_k \geqslant 2 + c, \quad c > 0$$

from (3.4.2)

$$R_n \leqslant \frac{1}{(2+c)-}\;\frac{1}{(2+c)-}\;\cdots\cdots$$

writing $$q = \frac{1}{(2+c)-}\;\frac{1}{(2+c)-}\;\cdots\cdots$$

or $$q = \frac{1}{(2+c)-q}$$

$$0 = q^2 - (2+c)q + 1$$

from which $$q = (1+\tfrac{1}{2}c) \pm \sqrt{(1+\tfrac{1}{2}c)^2 - 1}$$

If the positive sign is taken, q increases without bound as c increases

$$\therefore \quad q = (1+\tfrac{1}{2}c) - \sqrt{(1+\tfrac{1}{2}c)^2 - 1} < 1 \tag{3.4.4}$$

It is possible to write

$$F_2' = \frac{1}{b_1 -} \quad \frac{1}{b_2 -} \qquad \cdots\cdots\cdots \quad \frac{1}{b_n - \frac{R_n}{1}}$$

or $\quad F_2' = F_2'(n) + u_{n+1}$

where $\quad u_{n+1} = u_n \rho_{n+1}$ with $1 + \rho_{n+1} = \dfrac{b_n}{b_n - R_n(1+\rho_n)}$

$$\therefore \quad \rho_{n+1} = \frac{R_n}{\frac{b_n}{1+\rho_n} - R_n}$$

but $\quad 0 < \rho_n < 1$

$$\therefore \quad \rho_{n+1} \leq \frac{R_n}{1 - R_n}$$

$$\left| F_2' - F_{2(n)}' \right| \leqslant \frac{R_n}{1 - R_n} \left| u_n \right| \leqslant \frac{q}{1 - q} \left| \psi_n \right| \qquad (3.45)$$

Similarly if $a_k \leqslant \frac{1}{4} - c$ where $c > 0$, it can be shown that

$$\left| F_3' - F_3'(n) \right| \leqslant \frac{R_n}{\frac{1}{2} - R_n} \left| u_n \right| \leqslant \frac{p}{\frac{1}{2} - p} \left| u_n \right| \qquad (3.4.6)$$

where $\qquad p = \frac{1}{2} - \sqrt{c}$

A3.5 Contraction of a Continued Fraction

Using the recurrence relation, we can write

$$\begin{aligned} A_{2n} &= b_{2n} A_{2n-1} + a_{2n} A_{2n-2} \\ &= b_{2n} b_{2n-1} A_{2n-2} + a_{2n} A_{2n-2} + b_{2n} A_{2n-1} - b_{2n} b_{2n-1} A_{2n-2} \\ &= A_{2n-2} (a_{2n} + b_{2n} b_{2n-1}) + b_{2n}\, a_{2n-1}\, A_{2n-3} \end{aligned}$$

since $\quad A_{2n-1} = b_{2n-1} A_{2n-2} + a_{2n-1} + A_{2n-3}$

$$\therefore \quad A_{2n} = A_{2n-2} (a_{2n} + b_{2n} b_{2n-1}) + b_{2n} a_{2n-1} \left\{ \frac{1}{b_{2n-2}} A_{2n-2} - \frac{a_{2n-2}}{b_{2n-2}} A_{2n-4} \right\}$$

$$A_{2n} = A_{2n-2} \left\{ a_{2n} + b_{2n} b_{2n-1} + \frac{b_{2n} a_{2n-1}}{b_{2n-2}} \right\} - \frac{b_{2n} a_{2n-1} a_{2n-2}}{b_{2n-2}} A_{2n-4}$$

Equally $\quad B_{2n} = B_{2n-2} \left(a_{2n} + b_{2n} b_{2n-1} + \dfrac{b_{2n} a_{2n-1}}{b_{2n-2}} \right) - \dfrac{b_{2n} a_{2n-1} a_{2n-2}}{b_{2n-2}} B_{2n-4} \qquad (3.5.1$

Similarly $$A_{2n+1} = b_{2n+1} A_{2n} + a_{2n+1} A_{2n-1}$$

$$= b_{2n+1} b_{2n} A_{2n-1} + a_{2n+1} A_{2n-1} + b_{2n+1} A_{2n} - b_{2n+1} b_{2n} A_{2n-1}$$

$$= A_{2n-1} (a_{2n+1} + b_{2n+1} b_{2n}) + b_{2n+1} a_{2n} \left\{ \frac{1}{b_{2n-1}} A_{2n-1} - \frac{a_{2n-1}}{b_{2n-1}} A_{2n-3} \right\}$$

hence $$A_{2n+1} = A_{2n-1} \left(a_{2n+1} + b_{2n+1} b_{2n} + \frac{b_{2n+1} a_{2n}}{b_{2n-1}} \right) - \frac{b_{2n+1} a_{2n} a_{2n-1}}{b_{2n-1}} A_{2n-3}$$

and $$B_{2n+1} = B_{2n-1} \left(a_{2n+1} + b_{2n+1} b_{2n} + \frac{b_{2n+1} a_{2n}}{b_{2n-1}} \right) - \frac{b_{2n+1} a_{2n} a_{2n-1}}{b_{2n-1}} B_{2n-3} \tag{3.5.2}$$

(3.5.1) and (3.5.2) define recurrence relations which give convergents equal to the even and odd convergents of the original fraction with terms a_n and b_n.

A3.6 Modifications To Avoid Small Divisors in Summation Form

With the notation of (3.1.7)

$$f_{n+2} = f_{n-1} + u_n + u_{n+1} + u_{n+2}$$

$$= f_{n-1} + u_{n-1}\rho_n + u_{n-1}\rho_n\rho_{n+1} + u_{n-1}\rho_n\rho_{n+1}\rho_{n+2}$$

$$= f_{n-1} + u_{n-1}\rho_n(1 + \rho_{n+1}(1 + \rho_{n+2}))$$

$$= f_{n-1} + u_{n-1}\rho_n \left(1 + \frac{\rho_{n+1} b_{n+1} b_{n+2}}{b_{n+1} b_{n+2} + a_{n+2}(1+\rho_{n+1})} \right)$$

$$= f_{n-1} + u_{n-1} \frac{\rho_n(1+\rho_{n+1})(b_{n+1} b_{n+2} + a_{n+2})(b_n b_{n+1} + a_{n+1}(1+\rho_n))}{b_{n+1} b_{n+2}(b_n b_{n+1} + a_{n+1}(1+\rho_n)) + a_{n+2} b_n b_{n+1}}$$

$$= f_{n-1} + \frac{u_{n-1}\rho_n b_n b_{n+1}(b_{n+1} b_{n+2} + a_{n+2})}{(b_n b_{n+1}^2 b_{n+2} + b_{n+1} b_{n+2} a_{n+1}(1+\rho_n) + a_{n+2} b_n b_{n+1})}$$

$$= f_{n-1} + u_{n-1} \left[\frac{-a_n(b_{n+1} b_{n+2} + a_{n+2})(1+\rho_{n-1})b_n}{(b_n b_{n+1} b_{n+2} + a_{n+2} b_n)(b_{n-1} b_n + a_n(1+\rho_{n-1})) + b_{n-1} b_n b_{n+2} \cdot a_{n+1}} \right]$$

$$f_{n+2} = f_{n-1} - \frac{u_{n-1} a_n(b_{n+1} b_{n+2} + a_{n+2})(1 + \rho_{n-1})}{(b_{n+1} b_{n+2} + a_{n+2})(b_{n-1} b_n + a_n(1+\rho_{n-1})) + a_{n+1} b_{n-1} b_{n+2}} \tag{3.6.1}$$

Similarly

$$u_{n+2} = u_{n-1}\rho_n\rho_{n+1}\rho_{n+2} = u_{n-1}\rho_n\rho_{n+1}\left[\frac{-a_{n+2}(1+\rho_{n+1})}{b_{n+1}b_{n+2}+a_{n+2}(1+\rho_{n+1})}\right]$$

$$= u_{n-1}\rho_n\left[\frac{-a_{n+1}(1+\rho_n)}{b_nb_{n+1}+a_{n+1}(1+\rho_n)}\right]\left[\frac{-a_{n+2}b_n}{b_{n+2}(b_nb_{n+1}+a_{n+1}(1+\rho_n))+a_{n+2}b_n}\right]$$

$$= u_{n-1}\left[\frac{-a_n(1+\rho_{n-1})}{b_{n-1}b_n+a_n(1+\rho_{n-1})}\right]\left[\frac{-a_{n+1}b_{n-1}}{b_{n+1}(b_{n-1}b_n+a_n(1+\rho_{n-1}))+a_{n+1}b_{n-1}}\right]$$

$$\times\left[\frac{-a_{n+2}(b_{n-1}b_n+a_n(1+\rho_{n-1}))}{(b_{n+1}b_{n+2}+a_{n+2})(b_{n-1}b_n+a_n(1+\rho_{n-1}))+a_{n+1}b_{n-1}b_{n+2}}\right]$$

$$\therefore\quad u_{n+2} = \frac{-u_{n-1}a_na_{n+1}a_{n+2}b_{n-1}(1+\rho_{n-1})}{\left[b_{n+1}(b_{n-1}b_n+a_n(1+\rho_{n-1}))+a_{n+1}b_{n-1}\right]\left[(b_{n+1}b_{n+2}+a_{n+2})(b_{n-1}b_n+a_n(1+\rho_{n-1}))+a_{n+1}b_{n-1}b_{n+2}\right]} \qquad (3.62)$$

Finally

$$1+\rho_{n+2} = \frac{b_{n+1}b_{n+2}}{b_{n+1}b_{n+2}+a_{n+2}(1+\rho_{n+1})} \qquad (3.6.3)$$

where
$$\rho_{n+1} = \frac{-a_{n+1}b_{n-1}}{b_{n+1}(b_{n-1}b_n+a_n(1+\rho_{n-1}))+a_{n+1}b_{n-1}}$$

A3.7 Programme for the Evaluation of Continued Fractions by Summation

Below is listed a programme for the evaluation of a continued fraction using the summation form. Initially, a maximum figure is specified on the number of terms to be included in the summation. In addition, if the last term computed is less then 10^{-12}, the evaluation is terminated. The value of x for which the evaluation is desired is read from card. Termination of the programme occurs when a value of x is read which is greater than 90.0

A subroutine must be provided which will give the values of the coefficients a_n, b_n when x and n are provided by the main programme.

The modifications denoted in A3.6 are included and come into operation if any value of $^a/b$ is numerically greater than 10^4.

Output consists for each convergent of the coefficients a and b, the current value computed by the continued fraction, the last term added to the sum and the value of ρ.

The specimen output shows the computation of tan x using the contracted form F_1, used in one of the examples.

```
    READ(5,100)NMAX
100 FORMAT(I2)
108 READ(5,101)X
101 FORMAT(F12.8)
109 FORMAT(1H0,5X,29HEVALUATION OF FRACTION FOR X=,F12.8)
    IF(X.GT.90.0)GO TO 150
    WRITE(9,109)X
    WRITE(9,107)
107 FORMAT(1H ,5X,1HA,14X,1HB,6X,13HCONVERGENT NO,2X,9HVALUE OF ,
   18HFRACTION,4X,9HLAST TERM,5X,12HVALUE OF RHO)
    N=1
    CALL EVC(X,N,A,BZERO,B)
    IF(ABS(A/B).GE.1E+4)GO TO 111
    RHO=A/B
    C=RHO
    RM=RHO+BZERO
103 D=C/RM
    ERROR=ABS(D)
104 WRITE(9,148)A,B,N,RM,C,RHO
    IF(ERROR.LE.1E-12)GO TO 108
    N=N+1
    IF(N.GT.NMAX)GO TO 108
    AM1=A
    BM1=B
    RHO1=RHO
    CM1=C
    RM1=RM
    CALL EVC(X,N,A,BZERO,B)
    IF(N-2)105,105,106
105 RHO1=0.0
106 IF(ABS(A/B).GE.1E+4)GOTO113
    P=BM1*B
    DEN=P+A*(1+RHO1)
    RHO=P/DEN-1.0
    C=CM1*RHO
    RM=RM1+C
    GO TO 103
111 WRITE(9,149)A,B,N
149 FORMAT(1H ,E12.5,3X,E12.5,5X,I2,3X,25HPARTIAL DENOMINATOR SMALL)
148 FORMAT(1H ,E12.5,3X,E12.5,5X,I2,8X,E17.10,3(3X,E12.5))
    A1=A
    B1=B
    N=2
    CALL EVC(X,N,A2,BZERO,B2)
    WRITE(9,147)A2,B2,N
147 FORMAT(1H ,E12.5,3X,E12.5,5X,I2)
    N=3
    CALL EVC(X,N,A,BZERO,B)
    P=B1*B2+A2
    DEN=B*P+A*B1
    C=A1*A2*A/(P*DEN)
    RM= BZERO+A1*(B*B2+A)/DEN
    RHO=-A*B1/DEN
    ERROR=1.0
    GO TO 104
113 A1=A
    B1=B
    WRITE(9,149)A1,B1,N
    N=N+1
    CALL EVC(X,N,A2,BZERO,B2)
    P=A1*(1.0+RHO1)
    DEN=BM1*B1+P
```

```
      WRITE(9,147)A2,B2,N
      N=N+1
      CALL EVC(X,N,A,BZERO,B)
      C=-CM1*BM1*A2*A*P
      PROD=(B2*DEN+A2*BM1)*((B*B2+A)*DEN+BM1*B*A2)
      C=C/PROD
      RHO1=-A2*BM1/(B2*DEN+A2*BM1)
      P1=B2*B
      RHO=P1/(P1+A*(1.0+RHO1))-1.0
      P2=B*B2+A
      RM=RM1-CM1*P*P2/(P2*DEN+A2*BM1*B)
      ERROR=ABS((RM-RM1)/RM)
      GO TO 104
  150 STOP
      END


      SUBROUTINE EVC(Z,K,P,QZERO,Q)
      IF(K-1)200,200,201
  200 QZERO=0.0
      P=3.0/Z
      Q=3.0/Z**2-1.0
      RETURN
  201 L=4*K
      P=-(L-1.0)/(L-5.0)
      Q=(L-3.0)*(L-1.0)/Z**2-(2.0*L-6.0)/(L-5.0)
      RETURN
      END
```

```
     EVALUATION OF FRACTION FOR X=  1,00000000
      A             B          CONVERGENT NO  VALUE OF FRACTION     LAST TERM      VALUE OF RHO
 0,30000E 01    0,20000E 01        1           0,1500000000E 01    0,15000E 01     0,15000E 01
-0,23333E 01    0,31667E 02        2           0,1557377049E 01    0,57377E-01     0,38251E-01
-0,15714E 01    0,96429E 02        3           0,1557407722E 01    0,30673E-04     0,53459E-03
-0,13636E 01    0,19264E 03        4           0,1557407725E 01    0,22531E-08     0,73455E-04
-0,12667E 01    0,32073E 03        5           0,1557407725E 01    0,46195E-13     0,20503E-04

     EVALUATION OF FRACTION FOR X=  1,73205080
      A             B          CONVERGENT NO  VALUE OF FRACTION     LAST TERM      VALUE OF RHO
 0,17321E 01    0,87457E-08        1    PARTIAL DENOMINATOR SMALL
-0,23333E 01    0,83333E 01        2
-0,15714E 01    0,30429E 02        3          -0,6147560906E 01    0,38335E-01    -0,19357E-09
-0,13636E 01    0,62636E 02        4          -0,6147533459E 01    0,27447E-04     0,71598E-03
-0,12667E 01    0,10540E 03        5          -0,6147533454E 01    0,52709E-08     0,19204E-03
-0,12105E 01    0,15879E 03        6          -0,6147533454E 01    0,38134E-12     0,72348E-04

     EVALUATION OF FRACTION FOR X=  3,24037040
      A             B          CONVERGENT NO  VALUE OF FRACTION     LAST TERM      VALUE OF RHO
 0,92582E 00   -0,71429E 00        1          -0,1296148103E 01   -0,12961E 01    -0,12961E 01
-0,23333E 01   -0,10448E-06        2    PARTIAL DENOMINATOR SMALL
-0,15714E 01    0,68571E 01        3
-0,13636E 01    0,16208E 02        4           0,9909661685E-01    0,13076E-02     0,13372E-01
-0,12667E 01    0,28495E 02        5           0,9910026122E-01    0,36444E-05     0,27870E-02
-0,12105E 01    0,43789E 02        6           0,9910026477E-01    0,35488E-08     0,97379E-03
-0,11739E 01    0,62112E 02        7           0,9910026477E-01    0,15339E-11     0,43222E-03
-0,11481E 01    0,83471E 02        8           0,9910026477E-01    0,33991E-15     0,22160E-03

     EVALUATION OF FRACTION FOR X=  6,20483680
      A             B          CONVERGENT NO  VALUE OF FRACTION     LAST TERM      VALUE OF RHO
 0,48349E 00   -0,92208E 00        1          -0,5243524098E 00   -0,52435E 00    -0,52435E 00
-0,23333E 01   -0,24242E 01        2           0,1196113031E 02    0,12485E 02    -0,23811E 02
-0,15714E 01    0,19063E-07        3    PARTIAL DENOMINATOR SMALL
-0,13636E 01    0,27013E 01        4
-0,12667E 01    0,61229E 01        5          -0,7920992898E-01    0,32965E-01     0,79977E-01
-0,12105E 01    0,10335E 02        6          -0,7851452587E-01    0,69540E-03     0,21095E-01
-0,11739E 01    0,15359E 02        7          -0,7850923444E-01    0,52914E-05     0,76092E-02
-0,11481E 01    0,21203E 02        8          -0,7850921558E-01    0,18866E-07     0,35653E-02
-0,11290E 01    0,27871E 02        9          -0,7850921554E-01    0,36242E-10     0,19211E-02
-0,11143E 01    0,35366E 02       10          -0,7850921554E-01    0,41096E-13     0,11339E-02
```

Appendix A4

A4.1 Minimization of the Lagrangian Error Term

The error estimate for the (n+1)-point Lagrangian interpolation formula is

$$f(x) - L_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \; (x-x_0)(x-x_1) \ldots\ldots (x-x_n) \qquad (4.1.1)$$

The best choice of the x_i will be taken as that which minimises

$$\max|(x-x_0)(x-x\ ) \ldots\ldots\ x-x_n)|$$

This is so if the nodes are chosen as the zeros of the suitably-scaled Chebyshev polynomial $T_{n+1}(x)$.

Proof

Let $p_{n+1}(x)$ be a monic polynomial of degree (n+1) which has a smaller maximum deviation than the monic polynomial $2^{-n}T_{n+1}(x)$ in $[-1,1]$. Then $p_{n+1}(x) - 2^{-n}T_{n+1}(x)$ is a polynomial of degree n (at most) which must change sign between the (n+2) extrema of $T_{n+1}(x)$.

Hence $p_{n+1}(x) - 2^{-n}T_{n+1}(x)$ is of degree n with (n+1) zeros, which could only be true if it is identically zero throughout the region.

A4.2 Minimization of the Hermitian Error Term

The error estimate for the (n+1) - point Hermitian interpolation formula takes the form

$$f(x) - H_{2n+1}(x) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \; (x-x_0)^2 (x-x_1)^2 \ldots\ldots (x-x_n)^2 \qquad (4.2.1)$$

If we choose the error norm, as the L_1 norm i.e. we wish to minimise

$$\frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_{-1}^{1} (x-x_0)^2 \ldots\ldots (x-x_n)^2 dx \qquad (4.2.2)$$

where the interval of interpolation is $[-1,1]$.

This is so if the x_i are chosen as the zeros of the Legendre polynomial $P_{n+1}(x)$

Proof

$(x-x_0)(x-x_1) \;\ldots\; (x-x_n)$ is a monic polynomial and can be written as

$$\pi_{n+1}(x) = c_{n+1}P_{n+1}(x) + c_nP_n(x) + \ldots\ldots c_1P_1(x) + c_0$$

where $P_j(x)$ is the Legendre polynomial of degree j (4.2.3)

c_j are constants.

Then, by virtue of the orthogonality property of the Legendre polynomials

$$\int_{-1}^{1}\left[\pi_{n+1}(x)\right]^2 dx = \frac{2c_{n+1}^2}{2n+3} + 2\sum_{i=0}^{n} \frac{c_i^2}{2i+1} \qquad (4.2.4)$$

Clearly $c_{n+1} \neq 0$ otherwise the right-hand side of (4.2.3) is not of the required degree. Hence (4.2.4) is minimized by taking

$$c_n = c_{n-1} = c_{n-2} = \ldots\ldots\ldots = c_0 = 0$$

and $\pi_{n+1}(x) = c_{n+1}P_{n+1}(x)$ which is the suitably-scaled Legendre polynomial.

Hence, the nodes of the interpolation should be taken as the zeros of $P_{n+1}(x)$ in $[-1,1]$.

A4.3 Derivation of Interpolation Fraction Using Inverted Differences

Consider the sequence $f(x) = v_0(x)$

$$v_k(x) = v_k(x_k) = \frac{x-x_k}{v_{k+1}(x)} \qquad k = 0,1,2 \ldots \quad (4.3.1)$$

This leads to the continued fraction form

$$f(x) = v_0(x_0) + \frac{x-x_0}{v_1(x_1)+} \frac{x-x_1}{v_2(x_2)+} \frac{x-x_2}{v_3(x_3)+} \ldots \qquad (4.3.2)$$

If the fraction terminates after n divisions, the last term will be

$$v_n(x_n) + \frac{x-x_n}{v_{n+1}(x)}$$

If $x = x_k$, where $0 \leqslant k \leqslant n$, the fraction terminates before the last term and the value of $v_{n+1}(x)$ is of no consequence. If we remove the fraction $\frac{x-x_n}{v_{n+1}(x)}$, then (4.3.2) becomes a rational function which agrees with $f(x)$ at (n+1) points assuming that no divisor becomes zero for some $x = x_k$.

Introduce the notation

$$v_k(x) = \phi_k[x_0, x_1, \ldots\ldots x_{k-1}, x]$$ where ϕ_k will be shown to be the inverted difference of Chapter IV. Then (4.3.1.) can be written

$$v_{k+1}(x) = \frac{x-x_k}{v_k(x)-v_k(x_k)}$$ and we find recursively

that

$$\phi_0(x_k) = f(x_k)$$

$$v_1(x_1) = \phi_1[x_0,x_1] = \frac{x_1 - x_0}{\phi_0[x_1] - \phi_0[x_0]} = \frac{x_1 - x_0}{f(x_1) - f(x_0)}$$

$$v_2(x_2) = \phi_2[x_0,x_1,x_2] = \frac{x_2 - x_1}{\phi_1[x_0,x_2] - \phi_1[x_0,x_1]} \qquad \text{etc.}$$

and $$v_k(x_k) = \phi_k[x_0,x_1 \ldots . x_k] = \frac{x_k - x_{k-1}}{\phi_{k-1}[x_0 \ldots x_{k-2},x_k] - \phi_{k-1}[x_0 \ldots x_{k-2},x_{k-1}]} \tag{4.3.3}$$

We see that the coefficients $v(x)$ in (4.3.2) are the inverted differences derived in (4.3.3)

A4.4 Interpolation Formula Involving Reciprocal Differences (Thiele's Form)

We define a quantity ρ_k by the relation

$$\rho_k[x_0, \ldots . x_k] = \phi_k[x_0, \ldots .,x_k] + \phi_{k-2}[x_0, \ldots ,x_{k-2}] + \phi_{k-4}[x_0, \ldots x_{k-4}] + \ldots \tag{4.4.1}$$

where the ϕ's are defined in A4.3. The series is terminated by $\phi_0[x_0]$ if k is even and $\phi_1[x_0,x_1]$ if k is odd.

The quantity ρ_k is called the k^{th} reciprocal difference of f(x). In particular

$$\rho_0[x_0] = \phi_0[x_0] = f(x_0)$$

$$\rho_1[x_0,x_1] = \phi_1[x_0,x_1] = \frac{x_1 - x_0}{f(x_1) - f(x_0)}$$

An inductive argument will show that the ρ's are symmetrical in the arguments.

Now (4.4.1) implies that

$$\rho_k[x_0, \ldots .,x_k] - \rho_{k-2}[x_0, \ldots ,x_{k-2}] = \phi_k[x_0, \ldots ,x_k] \tag{4.4.2}$$

hence, using (4.3.3) we have

$$\rho_k[x_0, \ldots ,x_k] = \phi_k[x_0, \ldots x_k] + \rho_{k-2}[x_0, \ldots ,x_{k-2}]$$

$$= \frac{x_k - x_{k-1}}{\phi_{k-1}[x_0, \ldots x_{k-2},x_k] - \phi_{k-1}[x_0, \ldots x_{k-2},x_{k-1}]} + \rho_{k-2}[x_0, \ldots x_{k-1}]$$

$$\text{So } \rho_K[x_0,\cdots,x_K] = \frac{x_K - x_{K-1}}{\rho_{K-1}[x_0,\cdots,x_{K-2},x_K] - \rho_{K-1}[x_0,\cdots,x_{K-2},x_{K-1}]} + \rho_{K-2}[x_0,\cdots,x_{K-2}] \tag{4.4.3}$$

From (4.4.3) we can build up a table of differences which can be substituted in the fraction (4.3.2) giving the approximation

$$y(x) = f(x_0) + \frac{x - x_0}{\rho_1[x_1,x_0]+} \; \frac{x - x_1}{\rho_2[x_0,x_1,x_2] - f(x_0)+} \; \frac{x - x_2}{\rho_3[x_0,x_1,x_2,x_3] - \rho_1[x_0,x_1] + \cdots} \tag{4.4.4}$$

A4.5 Thiele's Expansion About a Single Point

We require the form of expansion in terms of a continued fraction when all the nodes in A.4.4 become coincident.

Equation (4.4.1) will tend to the form

$$y(x) = \phi_0(x_0) + \frac{x - x_0}{\phi(x_0) +} \; \frac{x - x_0}{\phi_2(x_0)+} \; \frac{x - x_0}{\phi_3(x_0)+} \; \cdots\cdots \tag{4.5.1}$$

where $\phi_k(x) = \lim\limits_{x_0 \ldots x_k \to x} \phi_k\left[x_0, \ldots, x_k\right]$

or using (4.4.2)

$$\phi_k(x) = \lim_{x_0 \ldots x_k \to x} \left\{\rho_k\left[x_0, \ldots, x_k\right] - \rho_{k-2}\left[x_0, \ldots x_{k-2}\right]\right\} \tag{4.5.2.}$$

In addition, since

$$\phi_k\left[x_0 \ldots x_{k-1}, x_k\right] = \frac{x_k - x_{k-1}}{\phi_{k-1}\left[x_0 \ldots x_{k-2}, x_k\right] - \phi_{k-1}\left[x_0 \ldots x_{k-2}, x_{k-1}\right]}$$

we have

$$\phi_k(x) = \lim_{x_k \to x} \frac{x_k - x}{\rho_{k-1}\left[x, \ldots x, x_k\right] - \rho_{k-1}\left[x \ldots x, x\right]} \tag{4.5.3}$$

But if this limit exists, it is given by

$$\left[\frac{1}{\dfrac{\partial\rho_{k-1}\left[x_0, \ldots\ldots x_{k-1}\right]}{\partial x_{k-1}}}\right]_{x_0, x_1 \ldots x_{k-1} \to x}$$

$$= \frac{k}{\dfrac{d\,\rho_{k-1}(x)}{dx}}$$

$$\therefore \quad \phi_k(x) = \frac{k}{\rho'_{k-1}(x)} \tag{4.5.4}$$

Also, from (4.5.2.), we have

$$\phi_k(x) = \rho_k(x) - \rho_{k-2}(x)$$

and so we have the recurrence relation

$$\rho_k(x) = \rho_{k-2}(x) + \phi_k(x) \quad ; \quad \phi_{k+1}(x) = \frac{k+1}{\rho'_k(x)} \tag{4.5.4}$$

with $\rho_{-2}(x) = \rho_{-1}(x) = 0$, $\phi_0(x) = f(x)$ which can be used to derive the constants in the expansion (4.5.1). This formula of expansion is very useful in finding continued fraction expansions for functions which are easily differentiated.

e.g. $f(x) = \tan^{-1}x$

$$\phi_0(x) = \tan^{-1}x \qquad\qquad \phi_0(1) = \frac{\pi}{4}$$

$$\rho_0(x) = \tan^{-1}x \qquad \rho'_0 = \frac{1}{x^2+1}$$

$$\phi_1(x) = 1+x^2 \qquad\qquad \phi_1(1) = 2$$

$$\rho_1(x) = 1+x^2 \qquad \rho'_1 = 2x$$

$$\phi_2(x) = \frac{2}{2x} \qquad\qquad \phi_2(1) = 1$$

$$\rho_2(x) = \tan^{-1}x + \frac{1}{x} \qquad \rho'_2 = \frac{1}{1+x^2} - \frac{1}{x^2}$$

$$\phi_3(x) = \frac{3}{\frac{1}{1+x^2} - \frac{1}{x^2}} \qquad\qquad \phi_3(1) = -6$$

$$\therefore \quad \tan^{-1}(x) = \frac{\pi}{4} + \frac{x-1}{2+}\;\frac{x-1}{1-}\;\frac{x-1}{6} \quad \text{...... etc.}$$

Appendix A5

A5.1 Ralston's Form of Economisation for a Rational Function

If $y(x)$ has a formal expansion

$$y(x) = c_0 + c_1 x + c_2 x^2 + \ldots .$$

then the Padé approximant $R^s_{jk}(x) = \dfrac{P^s_j(x)}{Q^s_k(x)}$

(where j and k are the degree of $P_j(x)$ and $Q_k(x)$ respectively and $j + k = s$)

is such that when expressed as a power series

$y(x) - R^s_{jk}(x)$ contains terms of degree $(s + 1)$ or higher.

Assume we require an approximation in the range $[-a,a]$ and write

$$x = az \qquad \text{where } -1 \leq z \leq 1 \quad ;$$

Then $\quad y(az) - R^s_{jk}(az) = d^{s+1}(az)^{s+1} + O((az)^{s+2})$

and
$$\lim_{a \to 0} \frac{f(az) - R^S{}_{jk}(az)}{a^{s+1}} = d^{s+1} z^{s+1} \qquad (5,1.1)$$

So, by appropriate choice of j and k it is possible to find Padé approximants with the property (5,1.1) for $s = 0,1,2, \ldots .$

If $R^N_{mn}(x)$ is the basic Padé approximant, we seek a modification

$$R^*_{mn}(x) = \frac{P^N_m(x) + \sum_{s=0}^{N-1} \delta_{s+1} P^s_j(x) + \delta_0}{Q^N_n(x) + \sum_{s=0}^{N-1} \delta_{s+1} Q^s_k(x)} \qquad (5,1.2)$$

For then

$$y(x) - R^*_{mn}(x) = \frac{Q^N_n(x)y(x) - P^N_m(x) + \sum_{s=0}^{N-1}\left[Q^s_k(x)y(x) - P^s_j(x)\right] - \delta_0}{Q^N_n(x) + \sum_{s=0}^{N-1} \delta_{s+1} Q^s_k(x)} \qquad (5,1.3)$$

The coefficients are now chosen so that for sufficiently small a, the right-hand-side of (5,1.3) will approximate to $\dfrac{d^{N+1}T_{N+1}(z)}{2^N}$ where

$T_{N+1}(x)$ is the Chebyshev polynomial of degree $(N + 1)$

$$\text{i.e.} \quad \delta_{s+1} = \frac{d^{N+1}}{d^{s+1}} \, \frac{a^{N-s}}{2^N} \, t_{s+1} \qquad s = 0,1, \ldots, (N-1)$$

$$\delta_0 = \frac{-d^{N+1}}{2^N} \, a^{N+1} t_0$$

where t_s is the coefficients of z^s in $T_{N+1}(z)$

Then, since $Q_k^s(0) = 1$ for all k, from (5.13) we have

$$\lim_{a \to o} \frac{y(az) - R^*_{mn}(az)}{a^{N+1}} = d^{N+1} \left[z^{N+1} + \sum_{s=0}^{N-1} \frac{t_{s+1}}{2^N} z^{s+1} + \frac{t_o}{2^N} \right]$$

$$= \frac{d^{N+1}}{2^N} T_{N+1}(z)$$

It may be noted that the Chebyshev polynomials are either even or odd functions and some of the δ_{s+1} are zero irrespective of the values of d^{s+1} in (5,1.1). The choice of j and k is not unique and may be chosen to satisfy $j + k = s$ in any fashion so long as $0 \leqslant j \leqslant m$ and $0 \leqslant k \leqslant n$ (except when $s = 0$ then $j = k = 0$).

A5.2 Estimation of Error in Truncated Chebyshev Series Solution

Let the nearly exact solution to the equation

$$(3 + 2x)y - 3\int ydx = \text{const} \quad \text{be given by} \tag{5.2.1}$$

$$y(x) = \sum_{r=0}^{6}{}' A_r T_r(x) \qquad \text{where } y(o) = 1 \tag{5,2.2}$$

Then substituting (5,2.2) into (5,2.1) and rearranging in terms of $T_j(x)$ gives

$$\left\{\frac{3}{2} A_o + \frac{1}{4} A_1\right\} + \left\{-\frac{1}{2} A_o + 3A_1 + \frac{5}{2} A_2\right\} T_1(x) + \left\{\frac{1}{4} A_1 + 3A_2 + \frac{7}{4} A_3\right\} T_2(x)$$
$$+ \left\{\frac{1}{2} A_2 + 3A_3 + \frac{3}{2} A_4\right\} T_3(x) + \left\{\frac{5}{8} A_3 + 3A_4 + \frac{11}{8} A_5\right\} T_4(x) + \left\{\frac{7}{10} A_4 + 3A_5 + \frac{13}{10} A_6\right\} T_5(x)$$
$$+ \left\{-\frac{1}{4} A_5 + 3A_6\right\} T_6(x) + \frac{11}{14} A_6 T_7(x) = \text{const.} \tag{5,2.3}$$

We compare this with the expression (5.16) obtained in Chapter V for the solution involving terms up to the third order

$$\left\{\frac{3}{2} a_o + \frac{1}{4} a_1\right\} + \left\{-\frac{1}{2} a_o + 3a_1 + \frac{5}{2} a_2\right\} T_1(x) + \left\{\frac{1}{4} a_1 + 3a_2 + \frac{7}{4} a_3\right\} T_2(x)$$
$$+ \left\{\frac{1}{2} a_2 + 3a_3\right\} T_3(x) + \frac{5}{8} a_4 T_4(x) = \text{const.} \tag{5,2.4}$$

We now subtract (5,2.3) from (5,2.4) and introduce the notation

$$\delta_s = a_s - A_s$$

$$\left\{\frac{3}{2}\,\delta_0 + \frac{1}{4}\,\delta_1\right\} + \left\{-\frac{1}{2}\,\delta_0 + 3\delta_1 + \frac{5}{2}\,\delta_2\right\} T_1(x) + \left\{\frac{1}{4}\,\delta_1 + 3\delta_2 + \frac{7}{4}\,\delta_3\right\} T_2(x)$$

$$+\left\{\frac{1}{2}\,\delta_2 + 3\delta_3 - \frac{3}{2}\,A_4\right\} T_3(x) + \ldots\ldots \qquad = 0 \qquad (5,2.5)$$

In a similar manner, we obtain from the initial condition

$$\tfrac{1}{2}\delta_0 - \delta_2 - A_4 + A_6 = 0 \qquad (5.2.6)$$

We now assume that the δ's can be expressed in terms of the first few neglected coefficients

$$\text{i.e. } \delta_s = \alpha_s^{(4)} A_4 + \alpha_s^{(5)} A_5 + \alpha_s^{(6)} A_6 \qquad (5,2.7)$$

The coefficients in (5.2.5) and (5.2.6) now become functions of A_4, A_5 and A_6. We must choose the α's to make as many terms in (5.2.5) to be zero as we can, together with satisfying (5.2.6)

Hence for terms in A_4 we must have

$$\begin{aligned}
\tfrac{1}{2}\alpha_0^{(4)} \qquad\qquad - \alpha_2^{(4)} \qquad &= 1 && [\text{Initial condition}\\
-\tfrac{1}{2}\alpha_0^{(4)} + 3\alpha_1^{(4)} + \tfrac{5}{2}\alpha_2^{(4)} \qquad &= 0 && [\text{coefficient of } T_1(x)\\
\tfrac{1}{4}\alpha_1^{(4)} + 3\alpha_2^{(4)} + \tfrac{7}{4}\alpha_3^{(4)} &= 0 && [\text{coefficient of } T_2(x)\\
\tfrac{1}{2}\alpha_2^{(4)} + 3\alpha_3^{(4)} &= \tfrac{3}{2} && [\text{coefficient of } T_3(x)
\end{aligned}$$

Comparing (5.2.4) and (5.2.5), together with the above we see that the solutions for the α's involves the same set of equations as for the original problem but with different right-hand sides.

Now the error is given by

$$e(x) \simeq \tfrac{1}{2}(a_0 - A_0) + (a_1 - A_1)T_1(x) + (a_2 - A_2)T_2(x)$$

$$+ (a_3 - A_3)T_3(x) - A_4T_4(x) - A_5T_5(x) - A_6T_6(x)$$

and since each of the differences has been expressed in terms of the neglected terms, from (5.2.7) we get

$$e(x) = A_4\,\varepsilon_4(x) + A_5\,\varepsilon_5(x) + A_6\,\varepsilon_6(x)$$

where

$$\varepsilon_j(x) = {\sum_{r=0}^{3}}' \alpha_r^{(j)} T_r(x) - T_j(x)$$

Appendix A6

A6.1 Expression for the Derivative of the L_1 Norm

Let the approximation be of the form $f(a,x) = \sum_{i=0}^{n} a_i \phi_i(x)$ (6,1.1)

and the error norm is

$$L_1 = \int_a^b |y(x) - f(a,x)|\, dx$$

so that L_1 is a function of the coefficients a_i in (6,1.1)

i.e. $L_1 = L_1(a)$

Then we seek the derivative of L_1 at some point a^*

i.e. we require an expression for $\lim_{t \to 0} \left\{ \frac{L_1(a^* + ta) - L_1(a^*)}{t} \right\}$ (6,1.2.)

we shall show that

$$\frac{d}{dt} L_1(a^*) = -\int_a^b f(a^*,x)\, \text{sign}\left\{y(x) - f(a^*,x)\right\} dx$$

where $\text{sign}(z) = \begin{cases} -1 & \text{if } z < 0 \\ 0 & \text{of } z = 0 \\ +1 & \text{if } z > 0 \end{cases}$

Proof

Define E as the set $E = \left\{x \mid \ |y(x) - f(a^*,x)| \leqslant \varepsilon\right\}$

Then if the domain $x = [a,b]$

$$\begin{aligned} L_1(a^* + ta) &= \int_X |y(x) - f(a^*,x) - tf(a,x)|\, dx \\ &= \int_{X-E} |y(x) - f(a^*,x) - tf(a,x)|\, dx + \\ &\qquad + \int_E |y(x) - f(a^*,x) - tf(a,x)|\, dx \end{aligned} \quad (6.1.3)$$

but

$$\begin{aligned} \int_{X-E} |y(x) - f(a^*,x) - tf(a,x)|\, dx &= \int_{X-E} [y(x) - f(a^*,x) - tf(a,x)]\, s_t(x)\, dx \\ &= \int_{X-E} |y(x) - f(a^*,x)|\, dx - t\int_{X-E} f(a,x)\, \text{sign}\left\{y(x) - f(a^*,x)\right\} dx + \psi(t) \end{aligned} \quad (6.1.4)$$

where

$$\psi(t) = \int_{X-E} [y(x) - f(a^*,x) - tf(a,x)]\,[s_t(x) - s(x)]\, dx$$

and $s_t(x) = \text{sign}[y(x) - f(a^*,x) - tf(a,x)]$

$s(x) = \text{sign}[y(x) - f(a^*,x)]$

Now, from (6.1.3) and (6.1.4.) we have

$$L_1(a^* + ta) = \int_{x-E} |y(x) - f(a^*,x)|\,dx - t\int_{x-E} f(a,x)\,\text{sign}\{y(x) - f(a^*,x)\}\,dx + \psi(t) + \int_E |y(x) - f(a^*,x) - tf(a,x)|\,dx \qquad (6.1.5)$$

but $L_1(a^*) = \int_x |y(x) - f(a^*,x)|\,dx$

$$\therefore \quad \int_{x-E} |y(x) - f(a^*,x)|\,dx = L_1(a^*) - \int_E |y(x) - f(a^*,x)|\,dx$$

Substituting in (6.1.5) and rearranging gives

$$\frac{L_1(a^* + ta) - L_1(a^*)}{t} + \int_{x-E} f(a,x)\,\text{sign}\{y(x) - f(a^*,x)\}\,dx$$
$$= \frac{\psi(t)}{t} + \frac{1}{t}\int_E |y(x) - f(a^*,x) - tf(a,x)|\,dx - \frac{1}{t}\int_E |y(x) - f(a^*,x)|\,dx \qquad (6.1.6)$$

Consider the values in E of the two right-hand integrals

$$|y(x) - f(a^*,x) - tf(a,x)| \leqslant |y(x) - f(a^*,x)| + t|f(a,x)|$$
$$\leqslant \varepsilon + tM \qquad \text{where } M = \max_{[x]} |f(a,x)|$$

$$\therefore \quad \int_E |y(x) - f(a^*,x) - tf(a,x)|\,dx \leqslant (\varepsilon + tM)\int_E dx$$

Equally $\int_E |y(x) - f(a^*,x)|\,dx \leqslant \varepsilon\int_E dx$

Also
$$\psi(t) = \int_{x-E} [y(x) - f(a^*,x) - tf(a,x)]\,[s_t(x) - s(x)]\,dx$$
$$\leqslant \int_{x-E} (\varepsilon + tM)\,[s_t(x) - s(x)]\,dx$$

Choosing $\varepsilon = tM$, we have

$$\frac{\psi(t)}{t} \leqslant \int_{x-E} 2M\,[s_t(x) - s(x)]\,dx$$

but $\lim_{t\to 0} s_t(x) = s(x) \quad \therefore \frac{\psi(t)}{t} \to 0 \quad$ as $t \to 0$

$\therefore$ from (6.1.6)

$$\lim_{t\to 0}\left| \frac{L_1(a^* + ta) - L_1(a^*)}{t} + \int_{x-E} f(a,x)\,\text{sign}\{y(x) - f(a^*,x)\}\,dx\right| \leqslant \left(\frac{2\varepsilon}{t} + M\right)\int_E dx$$

Also $\left(\frac{2\varepsilon}{t} + M\right)\int_E dx \leqslant 3M\int_E dx \qquad$ since $\varepsilon = tM$

As $t \to 0$, then $\varepsilon \to 0$ and E will become the set of points x for which $y(x) = f(a^*,x)$

So if $y(x)$ and $f(a^*,x)$ only agree at distinct points in x

$$\int_E dx \to 0 \qquad \text{as } t \to 0$$

Hence the right-hand side of (6,1.6) will then have limit zero as t→0 and

$$\lim_{t\to 0}\left\{\frac{L_1(a^* + ta) - L_1(a^*)}{t}\right\} = -\int_x f(a,x)\ \mathrm{sign}\left\{y(x) - f(a^*,x)\right\} dx \qquad (6,1.7)$$

The function f(a,x) will be of the form

$$f(a,x) = \sum_{i=0}^{n} a_i \phi_i(x)$$

and if we wish the derivative to be zero independently of the values of a_i from (6,1.7) when $L_1(a)$ is a minimum, we have

$$\int_x \phi_i(x)\mathrm{sign}\left[y(x) - f(a^*,x)\right] dx = 0 \qquad i = 0,1 \ \ldots\ldots\ n \qquad (6,1.8)$$

A fuller discussion of the characterisation and uniqueness of the L_1 approximation is given in Rice [13].

A6.2 Points of Interpolation in the Polynomial Case

Let the approximation be

$$f(a,x) = \sum_{i=0}^{n} a_i \phi_i(x) \qquad (6,2.1)$$

where the $\phi_i(x)$ have polynomial form

$$\phi_i(x) = \sum_{r=0}^{i} b_r x^r$$

Consider $$I = \int_{-1}^{1} \sum_{r=0}^{i} b_r x^r \ s(x)\ dx$$

Write x = cosθ

$$I = \int_{0}^{\pi} \sum_{r=0}^{i} b_r \cos^r\theta \ \ s(\cos\theta)\ \sin\theta d\theta$$

$$\text{but} \quad \cos\theta\ \sin\theta = \tfrac{1}{2}\sin 2\theta$$

$$\cos^2\theta\ \sin\theta = \tfrac{1}{4}\left\{\sin 3\theta + \sin\theta\right\}$$

$$\cos^3\theta\ \sin\theta = \tfrac{1}{8}\left\{\sin 4\theta + 2\sin 2\theta\right\} \qquad \text{etc}$$

so the integrand can be written as a sine series in multiples of θ

$$\text{i.e.} \quad \sum_{r=0}^{i} b_r \cos^r\theta\ \sin\theta = \sum_{r=0}^{i} c_r \ \sin(r+1)\theta$$

$$\text{and} \quad I = \int_0^{\pi} \sum_{r=0}^{i} c_r \sin(r+1)\theta\ s(\cos\theta)\ d\theta \qquad (6,2.2)$$

If the integral $\int_{-1}^{1} f(a,x)\ s(x)\ dx$ is to be zero, it follows from (6,2.1) and (6,2.2) that it is necessary that

$$\int_0^{\pi} \sin m\theta\ s(\cos\theta)\ d\theta \doteq 0 \qquad m = 1,2,\ \ldots\ldots\ n+1$$

Consider

$$J = \int_0^{\pi} \sin m\theta \, \text{sign}\left\{\sin(n+2)\theta\right\} d\theta$$

$$= \int_0^{\frac{\pi}{n+2}} \sin m\theta d\theta - \int_{\frac{\pi}{n+2}}^{\frac{2\pi}{n+2}} \sin m\theta d\theta + \quad + (-1)^{n+1} \int_{\frac{(n+1)\pi}{(n+2)}}^{\pi} \sin m\theta d\theta$$

then

$$J = \frac{1}{m}\left\{1 - 2\cos\frac{m\pi}{n+2} + 2\cos\frac{2m\pi}{n+2} - 2\cos\frac{3m\pi}{n+2} \ldots\ldots (-1)^n 2\cos\left(\frac{(n+1)m\pi}{n+2}\right) + (-1)^{n+1}\cos m\pi\right\}$$

It is possible to show by summation of this series that $J = 0$

$$\text{for } m = 1, 2, \ldots\ldots (n+1)$$

In other words, the sign function $s(x)$ which has the desired property when $f(a,x)$ has the form (6,2.1.) has zeros at the zeros of $\sin(n+2)\theta$ where $x = \cos\theta$.

The internal zeros of this sign function are given by

$$x_k = \cos\left(\frac{k\pi}{n+2}\right) \qquad k = 1, 2, \ldots\ldots (n+1) \qquad (6,2.3)$$

and we notice that these are the zero of the Chebyshev polynomial of the second kind. $U_{n+1}(x)$.

A6.3 Programme to Attempt L_1 Approximation by Interpolation

The programme attempts to find the approximation

$$f(x) = a_0U_0(x) + a_1U_1(x) + \ldots\ldots a_nU_n(x) \qquad \text{valid in } [-1,1] \qquad (6,3.1)$$

which minimises the L_1 norm

$$\int_{-1}^{1} |y(x) - f(x)| \, dx$$

This is done by solving the interpolation problem where $f(x)$ agrees with $y(x)$ at the zeros of the Chebyshev polynomial of the Second Kind $U_{n+1}(x)$. It has been shown that if the error curve changes sign only at the interpolation points, then $f(x)$ is the required L_1 approximation,

Input consists of the required degree of approximation N. The interpolation points are then derived from (6,2.3). The coefficients in (6,3.1.) are the solution of the set of linear equations

$$f(x_k) = y(x_k) \qquad k = 1, 2, \ldots\ldots (n+1) \qquad (6,3.2)$$

The values $y(x_k)$ are evaluated using a subroutine supplied by the user. This subroutine can also be used to take care of any transformation of the independent variable to reduce the range to the $[-1,1]$ employed in the main programme.

The equations (6,3.2) are solved using a standard subroutine for solving linear equations involving factorization of the matrix on the left-hand side into triangular form.

Having determined the approximating function the error is determined at fifty-one equally-spaced points in $[-1,1]$. These points are used to determine the error curve and the zeros of this curve are found by quadratic inverse interpolation between three adjacent points.

Additional output consists of the elements of the gradient vector

$$\int_{-1}^{1} U_j(x) \text{ sign}\left\{y(x) - f(x)\right\} dx \qquad j = 0,1 \ldots\ldots N \qquad (6,3.3)$$

The integrals in (6,3.3) are evaluated using the property

$$\int U_j(x)\, dx = \frac{T_{j+1}(x)}{j+1} \qquad \text{if } j \geqslant 2$$

$$= \tfrac{1}{2}(T_2(x) + T_0(x)) \qquad \text{if } j = 1$$

$$= T_1(x) \qquad \text{if } j = 0$$

If x_1, x_2, ... x_k are the points at which $y(x) - f(x)$ changes sign let $s = \text{sign}\{y(-1) - f(-1)\}$, then

$$\int_{-1}^{1} U_j(x) \text{ sign}\left\{y(x) - f(x)\right\} dx$$

$$= \frac{1}{j+1}\Big[- s\left[T_{j+1}(x)\right]_{-1} + 2\Big\{s\left[T_{j+1}(x)\right]_{x_1} - s\left[T_{j+1}(x)\right]_{x_2} + \ldots\ldots (-1)^{k-1}s\left[T_{j+1}(x)\right]_{x_k}\Big\} + (-1)^k s\left[T_{j+1}(x)\right]_1\Big]$$

The value of the error norm

$$L_1(x) = \int_{-1}^{1} |y(x) - f(x)|\, dx$$

is computed from the fifty-one points of the error curve using Simpson's Rule.

The printed output below shows the output when

$y(x) = e^x$ in the range $[0,2]$ when $N = 2$.

```
      MASTER UAPPL1
      DIMENSION X(15),RHS(15),CM(300),ERR(55),XE(55),A(20),AA(225)
      DIMENSION BB(15),REINT(20),Q(15)
      READ(5,100)N
  100 FORMAT(I2)
      DO9M=1,N+1
    9 X(N+2-M)=COS(M*3.1415927/(N+2))
      WRITE(9,107)(X(I),I=1,N+1)
      DO20I=1,N+1
      CALL F1(X(I),RHS(I))
      DO21J=1,N+1
      CALL UTERM(X(I),J-1,VALUE)
   21 CM(I+(N+1)*(J-1))=VALUE
   20 CONTINUE
      WRITE(9,97)
   97 FORMAT(1H0,20X,22HMATRIX OF COEFFICIENTS)
      DO22K=1,N+1
      WRITE(9,99)(CM(K+(N+1)*(J-1)),J=1,N+1)
   22 CONTINUE
   99 FORMAT(1H0,5X,10F10.5/1H ,5X,6F10.5)
      CALL F4ACSL(CM,RHS,N+1,(N+1)*(N+1),N+1,1,A,D,ID,IT,AA,BB,REINT
      WRITE(9,98)(A(J),J=1,N+1)
   98 FORMAT(30HOCOEFFICIENTS OF APPROXIMATION,/(1H ,7E16.8))
      STEP=0.04
      M=1
      XE(M)=-1.0
   19 CALL F1(XE(M),YVAL)
      CALL USUS(XE(M),A,N,APP)
      ERR(M)=YVAL-APP
      WRITE(9,106)XE(M),YVAL,APP,ERR(M)
  106 FORMAT(1H ,F5.2,5X,E13.6,5X,E13.6,5X,E13.6)
      M=M+1
      XE(M)=XE(M-1)+STEP
      IF(XE(M).LE.1.01)GOTO19
      M=1
      YONE=ABS(ERR(M))
      SUMOD=0.0
      DO130M=3,49,2
  130 SUMOD=SUMOD+ABS(ERR(M))
      SUMEV=0.0
      DO131M=2,50,2
  131 SUMEV=SUMEV+ABS(ERR(M))
      YEND=ABS(ERR(51))
      AR=0.04*(YONE+YEND+2.0*SUMOD+4.0*SUMEV)/3.0
      WRITE(9,102)AR
  102 FORMAT(1H0,25X,21HVALUE OF L1 INTEGRAL=,E13.6)
   17 M=1
      K=1
  299 IF(ERR(M))300,308,302
  300 IG1=-1
      GOTO303
  302 IG1= 1
  303 IF(ERR(M+1))304,309,306
  304 IG2=-1
      GOTO307
  306 IG2=1
  307 IF(IG1.EQ.IG2)GOTO310
      P=-ERR(M+1)/(ERR(M+1)-EPR(M))
      P=(-ERR(M+1)-0.5*P*(P+1)*(ERR(M+1)-2.*ERR(M)+ERR(M-1)))/
     1(ERR(M+1)-ERR(M))
      X(K)=XE(M+1)+P*0.04
      GOTO311
  308 X(K)=XE(M)
```

```
    GOTO311
309 X(K)=XE(M+1)
    M=M+1
311 NZS=K
    K=K+1
310 M=M+1
    IF(M.EQ.51)GOTO305
312 GOTO299
305 IF(K.LT.N+1)GOTO313
    WRITE(9,107)(X(J),J=1,NZS)
    K=1
 11 IF(ERR(1))31,32,32
 31 SGN=-1.0
    GOTO33
 32 SGN=1.0
 33 CALL EING(K,NZS,SGN,X,EVAL)
    WRITE(9,103)EVAL,K
 16 K=K+1
    IF(K.LE.N+1)GOTO11
    STOP
103 FORMAT(28H ELEMENT OF GRADIENT VECTOR=,E13.6,5X,
   18HELEMENT ,I2)
107 FORMAT(21H0ZEROS OF ERROR CURVE,/(1H ,9E13.5))
313 WRITE(9,105)
105 FORMAT(47H ERROR CURVE SCANNED INSUFFICIENT ZEROS LOCATED)
    STOP
    END


    SUBROUTINE USUS(Z,D,ND,SESUM)
    DIMENSION D(20)
    BNP1=0.0
    BN=D(ND+1)
    I=ND
 25 BVAL=D(I)+2.0*Z*BN-BNP1
    I=I-1
    IF(I.EQ.1)GOTO26
    BNP1=BN
    BN=BVAL
    GOTO25
 26 SESUM=D(1)-BN+2.0*Z*BVAL
    RETURN
    END


    SUBROUTINE UTERM(Z,NC,UV)
    ND=NC
    UZ=1.0
    U1=2.0*Z
    IF(ND-1)35,36,37
 35 UV=UZ
    RETURN
 36 UV=U1
    RETURN
 37 UV=2.0*Z*U1-UZ
    ND=ND-1
    IF(ND.EQ.1)RETURN
    UZ=U1
    U1=UV
    GOTO37
    END
```

```
      SUBROUTINE TCH(C,VAL,J)
      L=J
      IF(L-1)71,72,73
   71 VAL=1.0
      RETURN
   72 VAL=C
      RETURN
   73 TZ1=1.0
      TZ2=C
   74 VAL=2.0*C*TZ2-TZ1
      L=L-1
      IF(L.EQ.1)RETURN
      TZ1=TZ2
      TZ2=VAL
      GOTO74
      END


      SUBROUTINE EING(M,K1,SGN,Z,EVAL)
      DIMENSION Z(15)
      CALL TCH(-1.0,ORD,M)
      SUM=-SGN*ORD
      DO61J=1,K1
      CALL TCH(Z(J),ORD,M)
      SUM=SUM+SGN*ORD*2.0
   61 SGN=-SGN
      CALL TCH(1.0,ORD,M)
      EVAL= 1.0/M*(SUM+SGN*ORD)
      RETURN
      END


      SUBROUTINE F1(Z,FUNV)
      Y=Z+1.0
      FUNV=EXP(Y)
      RETURN
      END
```

```
ZEROS OF ERROR CURVE
 -0,70711E 00 -0,23196E-07  0,70711E 00

                      MATRIX OF COEFFICIENTS

          1,00000  -1,41421   1,00000

          1,00000  -0,00000  -1,00000

          1,00000   1,41421   1,00000

COEFFICIENTS OF APPROXIMATION
  0,30724629E 01  0,14752681E 01  0,35418102E 00
-1,00          0,100000E 01          0,118447E 01         -0,184470E 00
-0,96          0,104081E 01          0,119142E 01         -0,150609E 00
-0,92          0,108329E 01          0,120290E 01         -0,119617E 00
-0,88          0,112750E 01          0,121892E 01         -0,914242E-01
-0,84          0,117351E 01          0,123947E 01         -0,659610E-01
-0,80          0,122140E 01          0,126456E 01         -0,431535E-01
-0,76          0,127125E 01          0,129417E 01         -0,229249E-01
-0,72          0,132313E 01          0,132833E 01         -0,519566E-02
-0,68          0,137713E 01          0,136701E 01          0,101174E-01
-0,64          0,143333E 01          0,141023E 01          0,231006E-01
-0,60          0,149182E 01          0,145798E 01          0,338440E-01
-0,56          0,155271E 01          0,151027E 01          0,424410E-01
-0,52          0,161607E 01          0,156709E 01          0,489892E-01
-0,48          0,168203E 01          0,162844E 01          0,535900E-01
-0,44          0,175067E 01          0,169432E 01          0,563489E-01
-0,40          0,182212E 01          0,176474E 01          0,573756E-01
-0,36          0,189648E 01          0,183970E 01          0,567847E-01
-0,32          0,197388E 01          0,191918E 01          0,546950E-01
-0,28          0,205443E 01          0,200320E 01          0,512304E-01
-0,24          0,213828E 01          0,209176E 01          0,465198E-01
-0,20          0,222554E 01          0,218484E 01          0,406974E-01
-0,16          0,231637E 01          0,228246E 01          0,339028E-01
-0,12          0,241090E 01          0,238462E 01          0,262814E-01
-0,08          0,250929E 01          0,249131E 01          0,179844E-01
-0,04          0,261170E 01          0,260253E 01          0,916933E-02
 0,00          0,271828E 01          0,271828E 01         -0,541331E-08
 0,04          0,282922E 01          0,283857E 01         -0,935303E-02
 0,08          0,294468E 01          0,296339E 01         -0,187122E-01
 0,12          0,306485E 01          0,309275E 01         -0,278928E-01
 0,16          0,318993E 01          0,322664E 01         -0,367025E-01
 0,20          0,332012E 01          0,336506E 01         -0,449411F-01
 0,24          0,345561E 01          0,350801E 01         -0,524004E-01
 0,28          0,359664E 01          0,365550E 01         -0,588634E-01
 0,32          0,374342E 01          0,380753E 01         -0,641046E-01
 0,36          0,389619F 01          0,396408F 01         -0,678890E-01
 0,40          0,405520E 01          0,412517E 01         -0,699722E-01
 0,44          0,422070E 01          0,429080E 01         -0,700998E-01
 0,48          0,439295E 01          0,446095E 01         -0,680068E-01
 0,52          0,457223E 01          0,463564F 01         -0,634177E-01
 0,56          0,475882E 01          0,481487E 01         -0,560456E-01
 0,60          0,495303E 01          0,499862E 01         -0,455919E-01
 0,64          0,515517E 01          0,518692E 01         -0,317457E-01
 0,68          0,536556E 01          0,537974E 01         -0,141838E-01
 0,72          0,558453E 01          0,557710E 01          0,743074E-02
 0,76          0,581244E 01          0,577899E 01          0,334481E-01
 0,80          0,604965E 01          0,598541E 01          0,642332E-01
 0,84          0,629654E 01          0,619637E 01          0,100165E 00
 0,88          0,655350E 01          0,641186E 01          0,141640E 00
 0,92          0,682096E 01          0,663189E 01          0,189068E 00
 0,96          0,709933E 01          0,685645E 01          0,242877E 00
 1,00          0,738906E 01          0,708554E 01          0,303514E 00
```

```
                                        VALUE OF L1 INTEGRAL= 0.118896E 00

  ZEROS OF ERROR CURVE
   -0.70714E 00 -0.23140E-07  0.70709E 00
  ELEMENT OF GRADIENT VECTOR= 0.820615E-04       ELEMENT  1
  ELEMENT OF GRADIENT VECTOR=-0.469441E-04       ELEMENT  2
  ELEMENT OF GRADIENT VECTOR= 0.821580E-04       ELEMENT  3
```

Appendix A.7

A7.1 Legendre Polynomials

Consider a polynomial of $\phi_r(x)$ of degree r which is orthogonal to all x^k of inferior degree with respect to the interval $[-1,1]$ and unit weight function.

$$\text{i.e.} \quad \int_{-1}^{1} \phi_r(x)\, x^k \, dx = 0 \qquad k < r \qquad (7,1.1)$$

Following Hildebrand [10], let $\phi_r(x) = \dfrac{d^r}{dx^r} U_r(x)$ and integrate (7,1.1) r times by parts,

$$U_r^{(r-1)}(x)\, x^k - kU_r^{(r-2)}(x)\, x^{k-1} + \dots (-1)^{r-1} U_r(x) \left[x^k\right]^{(r-1)} = 0 \qquad (7,1.2$$

Now since $\phi_r(x)$ is to be a polynomial of degree r, its (r+1) th derivative must be zero.

$$\frac{d^{r+1}}{dx^{r+1}} \phi_r(x) = \frac{d^{2r+1}}{dx^{2r+1}} U_r(x) = 0 \qquad (7,1.3)$$

But for (7,1.2) to be satisfied for any x^k of degree less than x^r leads to 2r boundary conditions

$$U_r(\pm 1) = U_r'(\pm 1) = \dots = U_r^{(r-1)}(\pm 1) = 0 \qquad (7,1.4)$$

From (7,1.3) and (7,1.4) we get that

$$U_r(x) = C_r(x^2 - 1)^r \quad \text{and writing}$$

$C_r = \dfrac{1}{2^r r!}$ we have the Legendre polynomial

$$P_r(x) = \frac{1}{2^r r!} \frac{d^r}{dx^r} (x^2 - 1)^r \qquad (7,1.5)$$

Orthogonality Property

From the derivation we know that the Legendre polynomial is orthogonal to all polynomials of inferior degree.

Now consider $\int_{-1}^{1} P_r^2(x)\,dx = \int_{-1}^{1} \left[\frac{1}{2^r r!} \frac{(2r!)}{r!} x^r + \ldots \right] P_r(x)dx$

Because of the orthogonality property, terms involving powers of x of degree less than r make no contribution

$$\therefore \quad \int_{-1}^{1} P_r^2(x)dx = \frac{1}{2^r (r!)^2}(2r)! \int_{-1}^{1} x^r P_r(x)dx$$

$$= \frac{1}{2^r (r!)^2}(2r)! \int_{-1}^{1} x^r . \frac{1}{2^r(r!)} \frac{d^r}{dx^r} (x^2-1)^r dx$$

Integrating by parts, we have

$$\int_{-1}^{1} x^r \frac{d^r}{dx^r}(x^2-1)^r dx = \left[x^r \frac{d^{r-1}}{dx^{r-1}} (x^2-1)^r \right]_{-1}^{1} - \int_{-1}^{1} r\, x^{r-1} \frac{d^{r-1}}{dx^{r-1}} (x^2-1)^r dx$$

$$= \ldots\ldots\ldots$$

$$= 0 + (-1)^r\, r! \int_{-1}^{1} (x^2-1)^r dx$$

$$= 2r! \int_{0}^{\pi/2} \cos^{2r+1}\theta\, d\theta$$

$$= 2r! \frac{2r(2r-2) \ldots . 2}{(2r+1)(2r-1) \ldots . 3}$$

$$\therefore \quad \int_{-1}^{1} P_r^2(x)\,dx = \frac{2}{2r+1} \qquad (7,1.4)$$

Recursion Formula

Since $xP_n(x)$ is a polynomial of degree $(n+1)$, we may write

$$\int_{-1}^{1} xP_k(x)\, P_n(x)\, dx = \int_{-1}^{1} P_k(x) \sum_{i=0}^{n+1} c_i P_i(x)\, dx$$

$$= c_k \int_{-1}^{1} P_k^2(x)\, dx \qquad (7,1.5)$$

But $xP_k(x)$ is of degree $k+1$ and $P_n(x)$ is orthogonal to all polynomials of degree $n-1$ or less

$$\therefore \quad c_k = 0 \quad \text{for } k+1 < n \qquad \text{i.e. } k < n-1$$

and $xP_n(x) = c_{n+1}P_{n+1}(x) + c_nP_n(x) + c_{n-1}P_{n-1}(x)$ (7,1.6)

From (7,1.3), the coefficient of x^n in $P_n(x)$ is $\frac{(2n)!}{2^n(n!)^2}$

Equating coefficients of x^{n+1} in (7,1.6) gives

$$c_{n+1} = \frac{n+1}{2n+1}$$

Also, Legendre polynomials contain only either even or odd powers of x, hence, equating powers of x^n

$$c_n = 0$$

Now when k = n-1, consider the integrals in (7,1.5). Only the first term in $xP_{n-1}(x)$ need be retained on the left-hand-side.

$$\therefore \quad \frac{(2n-2)!}{2^{n-1}\left[(n-1)!\right]^2}\int_{-1}^{1} x^n P_n(x)\,dx = c_{n-1}\int_{-1}^{1} P_{n-1}^2(x)\,dx$$

whence $\quad c_{n-1} = \dfrac{n}{2n+1}$

and $\quad (n+1)\,P_{n+1}(x) = (2n+1)\,x\,P_n(x) - nP_{n-1}(x) \qquad (7,1.7)$

A7.2 Evaluation of Series of Orthogonal Terms

Orthogonal functions obey a recurrence relation of the form

$$\phi_n(x) = A_n\phi_{n-1}(x) + B_n\phi_{n-2}(x) \qquad (7,2.1)$$

Let

$$S = \sum_{k=0}^{N} a_k\phi_k(x)$$

$$= a_N\phi_N(x) + a_{N-1}\phi_{N-1}(x) + \ldots\ldots$$

$$= \left\{a_NA_N + a_{N-1}\right\}\phi_{N-1}(x) + \left\{a_{N-2} + B_Na_N\right\}\phi_{N-2}(x) + \ldots\ldots$$

$$= b_{N-1}\left\{A_{N-1}\phi_{N-2}(x) + B_{N-1}\phi_{N-3}(x)\right\} + \left\{a_{N-2} + B_Na_N\right\}\phi_{N-2}(x) + \ldots$$

where $b_{N-1} = A_Na_N + a_{N-1}$

$$\therefore \quad S = \left\{b_{N-1}A_{N-1} + a_{N-2} + B_Na_N\right\}\phi_{N-2}(x) + \left\{a_{N-3} + b_{N-1}B_{N-1}\right\}\phi_{N-3}(x)\ldots$$

$$= b_{N-2}\left\{A_{N-2}\phi_{N-3}(x) + B_{N-2}\phi_{N-4}(x)\right\} + \left\{a_{N-3} + b_{N-1}B_{N-1}\right\}\phi_{N-3}(x) +$$

where $b_{N-2} = b_{N-1}A_{N-1} + a_{N-2} + B_Na_N$

$$\therefore \quad S = b_{N-3}\phi_{N-3}(x) + \left\{b_{N-2}B_{N-2} + a_{N-4}\right\}\phi_{N-4}(x) + \ldots\ldots$$

where $b_{N-3} = b_{N-2}A_{N-2} + a_{N-3} + b_{N-1}B_{N-1}$

Hence, it can be seen that if we set $b_{N+1} = 0$, $b_N = a_N$, we can generate

$$b_k = a_k + b_{k+1} A_{k+1} + b_{k+2} B_{k+2} \qquad (7,2.2)$$

$$\text{for } k = (N-1), (N-2) \ldots\ldots, 1$$

At the end of the series, we then have

$$S = b_1 \phi_1 (x) + (a_0 + B_2 b_2) \phi_0 (x)$$

Now from (7,1.7), we see that for a series of Legendre polynomials

$$A_{k+1} = \frac{2k+1}{k+1}, \quad B_{k+2} = -\frac{k+1}{k+2}, \quad P_1(x) = x, \quad P_0(x) = 1$$

$$\text{and} \quad S = b_1 x + (a_0 - \tfrac{1}{2} b_2)$$

$$\text{where} \quad b_k = a_k + \frac{2k+1}{k+1} b_{k+1} - \frac{k+1}{k+2} b_{k+2} \qquad (7,2.3)$$

$$k = (N-1) \ldots 1$$

Similarly, for the Chebyshev series $\sum_{k=0}^{N}{}' a_k T_k(x)$

$$S = b_1 x + (\tfrac{1}{2} a_0 - b_2)$$

$$\text{where } b_k = a_k + 2x\, b_{k+1} - b_{k+2} \qquad (7,2.4)$$

$$k = (N-1) \ldots\ldots 1$$

Appendix A8

A8.1 Characterization of the Polynomial Minimax Approximation

Let the approximation be the form $f(x) = \sum_{i=0}^{n} a_i \phi_i(x)$ where $\phi_i(x)$ is a polynomial of degree i. Then if f(x) is the minimax approximation to y(x) in [a,b], the maximum error will be attained at not less than (n + 2) distinct points in [a,b] with alternating sign.

Proof

The proof follows that given in Handscomb [7] page 64. Assume the error e(x) = y(x) - f(x) is not identically zero, then it is possible to locate all points x_i at which the error reaches its maximum modulus. Let these be r in number.

Consider the sign changes in the list $e(x_1), e(x_2) \ldots\ldots e(x_r)$. If a sign-change occurs between x_j and x_{j+1}, define $p_j = \frac{1}{2}(x_j + x_{j+1})$.

Thus, there is a list of values

$a < p_1 < p_2 < \ldots\ldots < p_s < b$ if there are s sign changes in all.

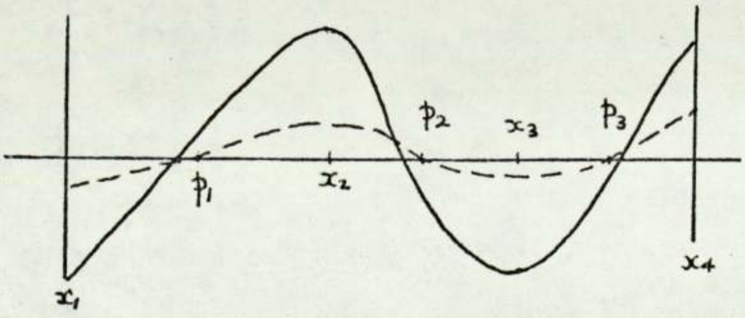
Now s cannot be greater than (r - 1), since e(x) can only change sign between extrema of opposite sign.

We shall now show that if $r < (n+2)$, then f(x) cannot be the minimax approximation.

Assume $r < (n+2)$ then r is at most equal to (n + 1) and since

$$s \leqslant r - 1 \quad \text{then} \quad s \leqslant n$$

Hence, we can find a unique polynomial p(x) of degree not greater than n, having zeros at the points $p_1, p_2 \ldots\ldots, p_s$. In addition p(x) can be chosen to have the sign of $e(x_0)$ at x_0.



Because of the choice of zeros, p(x) will have the same sign as e(x) at points where e(x) attains its extreme values. By adding a suitable multiple of p(x) to f(x), the extreme values could therefore be reduced and f(x) cannot be the minimax function.

Now p(x) may be considered as a more general combinationof the continuous functions ϕ_i.

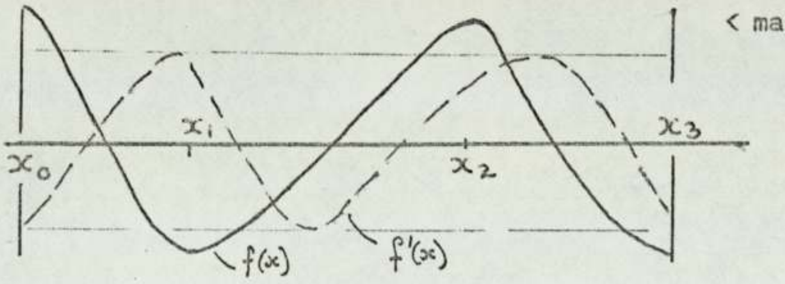
$$\text{i.e.} \quad p(x) = \sum_{i=0}^{n} c_i \phi_i(x) \qquad (8,1.1)$$

where some at least of the c_i are non-zero.

The above argument only holds if right-hand side of (8,1.1) has, at most, n roots in $[a,b]$. This defines the polynomials $\phi_i(x)$ as forming a Chebyshev set. (of which the powers of x are a particular case)

Hence, it has been shown that if the number of extrema of opposite sign is less than $(n+2)$, then f(x) is not the minimax function and the required result follows.

To prove sufficiency, let f(x) have $(n+2)$ extrema of equal magnitude and opposite sign. Let $f'(x)$ be a function such that $\max|y(x) - f'(x)| < \max|y(x) - f(x)|$



Let $[x_i]$ $i = 0,1 \ldots .(n+1)$ be the points at which f(x) achieves its extrema.

$$\text{Then sign } [f(x_i) - f'(x_i)] = - \text{sign}[f(x_{i+1}) - f'(x_{i+1})]$$

$$i = 0, \ldots .., n$$

This implies that $f(x) - f'(x)$ has $(n+1)$ zeros. But f(x) and $f'(x)$ are of degree n. Since they are also composed of polynomials forming a Chebyshev set, then their difference cannot be zero at $(n+1)$ points unless it is identically zero.

Hence f(x) is the required minimax function.

A8.2 Extension to Rational Function Minimax Approximation

Assume that the approximation has the form $R(x) = \dfrac{P_n(x)}{Q_m(x)}$ where n and m are the degree of the polynomials in the numerator and denominator respectively. Then analogous to the result in A8.1, $R_{nm}(x)$ will be a best minimax approximation if $y(x) - R_{nm}(x)$ has not less than $(n+m+2)$ extrema of equal magnitude and with alternating sign in $[a,b]$.

Proof

As before, we shall assume that the number of extrema is less than $(n+m+2)$ and show that this leads to a contradiction.

Let r be the number of points at which $e(x)$ reaches its extreme value. Consider the sign changes in the list

$$e(x_1),\ e(x_2)\ \ldots\ldots\ e(x_r)$$

If a sign change occurs between x_j and x_{j+1}, define $p_j = \frac{1}{2}(x_j + x_{j+1})$

There is now a set of points

$a < p_1 < p_2 \ldots\ldots < p_s < b$ if there are s sign changes in $e(x)$.

We shall assume that $r < n+m+2$ then since a sign change will only be listed between extrema of opposite sign,

$$r \ngtr n+m+1 \qquad \text{and } s \leqslant n+m$$

Define $\quad A(x) = (x - p_1)(x - p_2)\ \ldots\ldots(x - p_s)$

then $\quad A(x) = Q_m(x)a(x) - P_n(x)b(x)$

where $a(x)$ and $b(x)$ must be polynomials of degree not greater than n and m respectively.

Consider $\quad R'_{nm}(x) = \dfrac{P_n(x) - \beta a(x)}{Q_m(x) - \beta b(x)}$

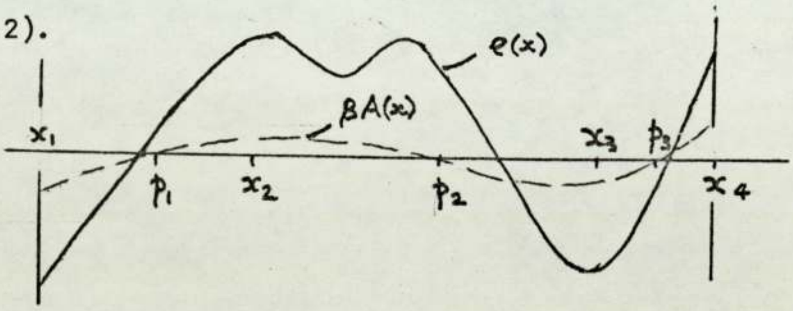
$$f(x) - R'_{nm}(x) = f(x) - R_{nm}(x) + \frac{\beta(Q_m(x)a(x) - P_n(x)b(x))}{Q_m(x)[Q_m(x) - \beta b(x)]} \qquad (8.2.1)$$

For a pole-free solution, $Q_m(x)$ must be one-signed in $[a,b]$ and be non-zero. Hence the denominator in the last term in (8,2.1) can be made one-signed by choosing β sufficiently small.

Now $\beta(Q_m(x)a(x) - P_n(x)b(x)) = \beta A(x)$

and by suitable choice of the sign of β, the last term can always be of opposite sign to $f(x) - R_{nm}(x)$ at the points of extreme value.

This would give $R'_{nm}(x)$ a smaller maximum error than $R_{nm}(x)$. Consequently if $R_{nm}(x)$ is to be the best minimax approximation, s must exceed $(n + m)$ and hence the number of extrema with alternating signs must be not less than $(n + m + 2)$.



A8.3 Iterative Scheme for the Construction of the Minimax Rational Approximating Function

Let $y(x)$ be a continuous function in $[a,b]$ and let the approximation be of the form $f(x) = P_n(x)/Q_m(x)$

where $P_n(x) = \sum_{j=0}^{n} a_j x^j$

$$Q_m(x) = \sum_{j=0}^{m} b_j x^j$$

It is required to find the set of reference points

$$a \leqslant x_0 \leqslant x_1 < \ldots\ldots\ldots < x_{n+m+1} \leqslant b$$

at which the error reaches its extreme value h

$$\text{i.e. } y(x_s) - \frac{P_n(x_s)}{Q_m(x_s)} = (-1)^s h \qquad s = 0,1, \ldots, n+m+1 \qquad (8,3.1)$$

$$\text{where } \max_{[a,b]} |y(x) - f(x)| = h$$

First, we rearrange (8,3.1) and then express in matrix form

$$\text{i.e. } P_n(x_s) - \{y(x_s) - (-1)^s h] Q(x_s) = 0$$

or more fully

$$a_0 + a_1 x_0 + \ldots\ldots + a_n x_0^n - \{y(x_0) - h\}\{b_0 + b_1 x_0 + \ldots\ldots + b_m x_0^m\} = 0$$

$$a_0 + a_1 x_1 + \ldots\ldots + a_n x_1^n - \{y(x_1) + h\}\{b_0 + b_1 x_1 + \ldots\ldots + b_m x_1^m\} = 0$$

$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$

$$a_0 + a_1 x_{n+m+1} + \ldots a_n x_{n+m+1}^n - \{y(x_{n+m+1}) - (-1)^{n+m+1} h\}\{b_0 + b_1 x_{n+m+1} \ldots b_m x_{n+m+1}^m\} = 0$$

This can be expressed in partitioned-matrix form as

$$\left[Y \,\vdots\, -(F - Gh)X\right] . [C] = 0 \qquad (8,3.2)$$

where
$$y_{rs} = x_r^{s-1} \qquad s = 1,2, \ldots..(n+1)$$

$$x_{rs} = x_r^{s-1} \qquad s = 1,2, \ldots..(m+1)$$

$$F = \text{diagonal } \{y(x_r)\}$$

$$G = \text{diagonal } \{(-1)^r\}$$

$$r = 0,1, \ldots..,(n+m+1)$$

and $[C]$ is the column vector $\left[a_0 \; a_1 \ldots\ldots a_n \; b_0 \; b_1 \ldots\ldots b_m\right]^T$

The method of Osborne [4] is applied to (8.3.2) resulting in an iterative scheme for the determination of $[C]$ and h.

Write (8.3.2) as $m(h).\bar{v} = 0$ (8.3.3)

Then if h_i $\bar{v}_i$ are approximate solutions of (8.3.3), we can write

$$\left[m(h_i) + \Delta h_i \frac{dm}{dh} + \ldots\right] \left[\bar{v}_i + \Delta \bar{v}_i\right] = 0$$

and retaining only first order small quantities,

$$m(h_i). \{\bar{v}_{i+1}\} = -\Delta h_i \frac{dm}{dh}(h_i) . \bar{v}_i$$

Now $-\Delta h_i$ will only act as a scale factor and may be removed

then $$m(h_i)\{\bar{v}_{i+1}\} = \frac{dm}{dh}(h_i) . \bar{v}_i \qquad (8.3.4)$$

Consider $$m(h) . \bar{v}_{i+1} = B(h).\bar{v}_i \qquad (8.3.5)$$

where B(h) is some function of h.

The solution of (8.3.3) will occur at the zeros of B(h), hence we may seek the solution by finding a solution to B(h) = 0. This is done by applying Newton's Method, for which is required an expression for $\frac{dB}{dh}$.

Differentiate (8.3.5) with respect to h.

$$\frac{dm}{dh} .\bar{v}_{i+1} + m.\frac{d\bar{v}_{i+1}}{dh} = \frac{dB}{dh}.\bar{v}_i$$

whence $$\frac{d\bar{v}_{i+1}}{dh} + m^{-1}\frac{dm}{dh}.\bar{v}_{i+1} = \frac{dB}{dh}.m^{-1}.\bar{v}_i \qquad (8.3.6)$$

Now the system is homegeneous, hence one element is independent of h.

(In our case, let the element of maximum modulus always be made equal to one.)

Let this element be the p^{th}.

Then $\left(\frac{d\bar{v}_{i+1}}{dh}\right)_p = 0$

and from (8.3.6) and (8.3.5) we have

$$\left(M^{-1} \cdot \frac{dM}{dh} \cdot \bar{v}_{i+1}\right)_p = \frac{1}{B} \; \frac{dB}{dh} \cdot \left(\bar{v}_{i+1}\right)_p$$

or

$$\frac{B}{\frac{dB}{dh}} = \frac{\left(\bar{v}_{i+1}\right)_p}{\left(M^{-1} \cdot \frac{dM}{dh} \cdot \bar{v}_{i+1}\right)_p} \qquad (8.3.7)$$

From (8.3.4), we may compute $\bar{v}_{i+1}$ by identifying $\bar{v}_i$ as the vector of coefficients $\bar{c}_i$ at the i^{th} stage of the iteration.

Also, since $\bar{v}_{i+1}$ is an approximate solution to the eigenvalue problem, we may use it in (8.3.4) to find a new vector of coefficients $\bar{c}_{i+1}$.

thus $\bar{c}_{i+1} = M^{-1}(h_i) \cdot \frac{dM(h_i)}{dh} \cdot \bar{v}_{i+1}$

If we choose from this equation the element in the p^{th} position, we have exactly the denominator in (8.3.7). Consequently, the process proceeds as follows:

Let h_i, $\bar{c}_i$ be the solution at the i^{th} stage, where $\bar{c}_i$ is scaled so that the largest element is equal to unity.

Solve $M(h_i) \cdot \bar{v}_{i+1} = \frac{dM}{dh}(h_i) \cdot \bar{c}_i$

$$M(h_i) \cdot \bar{c}_{i+1} = \frac{dM}{dh}(h_i) \cdot \bar{v}_{i+1}$$

and from (8.3.7) $h_{i+1} = h_i - \frac{\left(\bar{v}_{i+1}\right)_p}{\left(\bar{c}_{i+1}\right)_p}$ (8.3.8)

In terms of (8.3.2)

$$M(h) = \left[Y \mid -(F - Gh)X\right]$$

$$\frac{dM}{dh} = \left[0 \mid GX\right]$$

When new coefficients are determined, a new error curve can be computed and the points of extrema found by interpolation.

The whole process can then be repeated, using $\left[x_{i+1}\right]$, $\left[c_{i+1}\right]$ and h_{i+1} used as input to the next stage.

A8.4 Computer Programme for Rational Function Minimax Approximation

The programme listed below is an implementation of the scheme given in A8.3. Initial data consists of the degree of the polynomial in the numerator of the approximation, the degree of the polynomial in the denominator, the lower and upper bounds of the range of the approximation followed by the step-length to be used when computing points on the error curve. Two further fields are optional. Because the extrema of the error curve may be more closely packed near one end than the other, it is possible to change the step length at some point part-way through the error curve. This may be done by specifying the value of x at which the change is required to take place followed by the new step length. If these fields are left blank, this facility is ignored.

The programme then reads the values of x which are designated as the current reference and employs a user-provided subroutine to evaluate the given function at these points. The final input statement reads the current value of the error extreme value (h) followed by the coefficients of the approximation, with those of the numerator first and in ascending powers of x.

Output, after one iteration of the algorithm,consists of the current basis (for reference), the newly-computed value of the error extreme and the corresponding values of the coefficients of the rational function approximation. Finally, a table of values is printed defining the error curve. The left-hand column is the values of x at the required interval, the next column contains the value of the given function and this is followed by the value of the approximation. The right-hand column contains the error.

The specimen programme listed below shows the output of the fourth iteration when finding the $P_2(x)/Q_2(x)$ approximation to y = 0.92 cos hx -cos x (See example 2)

Sufficient storage has been allocated to allow the sum of the degrees of the numerator and denominator to be a maximum of thirteen.

```
      REAL MAXV,MAXC
      DIMENSION X(15),FN(15),C(15),A(225),B(225),V(15)
      DIMENSION RHS(15),AA(225),BB(225),REINT(15)
      READ(5,19)IP,IQ,XMIN,XMAX,DELX,X1,DELX1
   19 FORMAT(2I2,5F5.3)
      N=IP+IQ+2
      READ(5,18)(X(I),I=1,N)
      DO16I=1,N
      CALL F1(X(I),FN(I))
   16 CONTINUE
      WRITE(9,50)(X(I),I=1,N)
   50 FORMAT( 25H CURRENT REFERENCE POINTS/(1H ,12F9.5))
   18 FORMAT(10F8.5)
      READ(5,18)H,(C(I),I=1,N)
      DO20J=1,N
      A(J)=1.0
      DO29K=2,IP+1
   29 A(J+(K-1)*N)=X(J)**(K-1)
      A(J+(IP+1)*N)=-(FN(J)-(-1)**J*H)
      IF(IQ.EQ.0)GOTO20
      DO30K=IP+2,N-1
   30 A(J+K*N)=A(J+(IP+1)*N)*X(J)**(K-IP-1)
   20 CONTINUE
      DO21J=1,N
      DO31K=1,IP+1
   31 B(J+(K-1)*N)=0.0
      B(J+(IP+1)*N)=(-1)**J
      IF(IQ.EQ.0)GOTO21
      DO32K=IP+2,N-1
   32 B(J+K*N)=(-1)**J*X(J)**(K-IP-1)
   21 CONTINUE
      NA=N*N
      CALL FPMUMT(N,1,N,B(1),C(1),RHS(1),0,NRR)
      CALL F4ACSL(A,RHS,N,NA,N,1,V,D,ID,IT,AA,BB,REINT)
      CALL FPMUMT(N,1,N,B(1),V(1),RHS(1),0,NRR)
      CALL F4ACSL(A,RHS,N,NA,N,2,C,D,ID,IT,AA.BB,REINT)
      MAXC=0.0
      MAXV=0.0
      DO22J=1,N
      IF(ABS(C(J))-ABS(MAXC))22,22,23
   23 MAXC=C(J)
      MAXV=V(J)
   22 CONTINUE
      H=H-MAXV/MAXC
      DO24J=1,N
   24 C(J)=C(J)/MAXC
      WRITE(9,27)H,(C(J),J=1,IP+1)
   27 FORMAT(1H ,10HVALUE OF H,E17.8//16H COEFFICIENTS OF,
     110H NUMERATOR/(10X,6E17.8))
      WRITE(9,28)(C(J),J=IP+2,N)
   28 FORMAT(1H0,27HCOEFFICIENTS OF DENOMINATOR/10X,6E17.8)
      WRITE(9,40)
   40 FORMAT(1H0,10X,1HX,13X,4HFUNC,10X,6HAPPROX,16X,5HERROR)
      IF(X1)14,15,14
   14 XEND=X1
      GOTO 17
   15 XEND=XMAX
   17 Z=XMIN
   26 BC=0.0
      DO37I=1,IP+1
      K=IP-I+2
   37 BC=C(K)+Z*BC
```

```
      CC=0
      DO38I=1,IQ+1
      K=IQ+2-I
   38 CC=C(IP+K+1)+Z*CC
      VALUE=BC/CC
      CALL F1(Z,FUNV)
      ERROR=FUNV-VALUE
      WRITE(9,25)Z,VALUE,FUNV,ERROR
   25 FORMAT(1H ,3(5X,F10.6),10X,E15.6)
      Z=Z+DELX
      IF(Z.LE.XEND+DELX/2.0)GOTO26
      IF(Z.GT.XMAX)STOP
      XEND=XMAX
      DELX=DELX1
      Z=X1+DELX
      GOTO26
      END


      SUBROUTINE F1(Z,FVAL)
      FVAL=0.92*COSH(Z)-COS(Z)
      RETURN
      END
```

```
CURRENT REFERENCE POINTS
 -1.00000 -0,86270 -0,49920  0.00000  0,49920  0,86270
VALUE OF H  -0,83219411E-04

COEFFICIENTS OF NUMERATOR
           -0,79916781E-01  -0,67252399E-21   0,95855700E 00

COEFFICIENTS OF DENOMINATOR
            0,10000000E 01  -0,45852704E-21  -0,69200243E-03




     X               FUNC              APPROX                   ERROR
-1,000000          0,879249          0,879332               0,832194E-04
-0,950000          0,785672          0,785644              -0,273619E-04
-0,900000          0,696905          0,696830              -0,755113E-04
-0,850000          0,612947          0,612865              -0,818314E-04
-0,800000          0,533796          0,533733              -0,626648E-04
-0,750000          0,459450          0,459420              -0,306190E-04
-0,700000          0,389908          0,389913               0,493930E-05
-0,650000          0,325169          0,325206               0,374195E-04
-0,600000          0,265230          0,265292               0,625732E-04
-0,550000          0,210091          0,210169               0,780894E-04
-0,500000          0,159750          0,159833               0,832205E-04
-0,450000          0,114207          0,114285               0,784375E-04
-0,400000          0,073460          0,073526               0,651156E-04
-0,350000          0,037510          0,037555               0,452495E-04
-0,300000          0,006354          0,006375               0,211990E-04
-0,250000         -0,020008         -0,020012              -0,453621E-05
-0,200000         -0,041576         -0,041605              -0,295111E-04
-0,150000         -0,058350         -0,058402              -0,515005E-04
-0,100000         -0,070332         -0,070400              -0,686334E-04
-0,050000         -0,077521         -0,077600              -0,794986E-04
 0,000000         -0,079917         -0,080000              -0,832194E-04
 0,050000         -0,077521         -0,077600              -0,794986E-04
 0,100000         -0,070332         -0,070400              -0,686333E-04
 0,150000         -0,058350         -0,058402              -0,515005E-04
 0,200000         -0,041576         -0,041605              -0,295112E-04
 0,250000         -0,020008         -0,020012              -0,453621E-05
 0,300000          0,006354          0,006375               0,211990E-04
 0,350000          0,037510          0,037555               0,452495E-04
 0,400000          0,073460          0,073526               0,651156E-04
 0,450000          0,114207          0,114285               0,784375E-04
 0,500000          0,159750          0,159833               0,832205E-04
 0,550000          0,210091          0,210169               0,780894E-04
 0,600000          0,265230          0,265292               0,625732E-04
 0,650000          0,325169          0,325206               0,374195E-04
 0,700000          0,389908          0,389913               0,493928E-05
 0,750000          0,459450          0,459420              -0,306190E-04
 0,800000          0,533796          0,533733              -0,626648E-04
 0,850000          0,612947          0,612865              -0,818314E-04
 0,900000          0,696905          0,696830              -0,755114E-04
 0,950000          0,785672          0,785644              -0,273619E-04
 1,000000          0,879249          0,879332               0,832194E-04
```

Appendix A9

A9.1 Choice of Degree of Approximation

The approximating function is assumed to be a low degree polynomial which satisfies two conditions. Firstly, it is required that the approximation interpolates to the given function at the knots, secondly that some degree of smoothness is imparted by continuity of the spline and at least some of its derivatives at the internal knots.

The criteria of smoothness rules out the possibility of the broken line passing through the given interpolation points. The next possibility is the quadratic polynomial with continuity of the first derivative.

As illustration, let there be three knots $x_1 < x_2 < x_3$. A quadratic function in each of the two zones $[x_1\ x_3]$, $[x_2\ x_3]$ provides six unknown coefficients. The number of conditions imposed are

(i) Interpolation at three points (3)

(ii) Continuity of the approximation and its first derivative at x_2 (2

(iii) End conditions at x_1 and x_3 (2)

This makes seven conditions in all and it is seen that the even degree function cannot satisfy all the requirements. More generally, let the degree of the spline be taken as 2n and the number of knots as (N + 1). There are (2n + 1)N unknowns to determine, with the following constraints:

Continuity of derivatives of order 0,1,2, ...(2n - 1) at each interior points imposes 2n(N - 1) conditions

There are n end conditions at x_1 and x_{N+1} respectively

This leaves (2n + 1)N - 2n(N - 1) - 2n = N conditions for interpolation at (N + 1) knots, which is clearly impossible.

Now assume that the degree of the polynomial is (2n - 1). This time, there are 2nN unknowns.

Continuity of derivatives of order 0,1, ...(2n - 2) at (N - 1) points imposes (2n - 1)(N - 1) conditions.

End conditions impose $(n-1)$ constraints at x_1 and x_{N+1} respectively. This leaves $2nN - (2n-1)(N-1) - 2(n-1) = N+1$ which is exactly right for interpolation at $(N+1)$ points.

In particular, putting $n = 2$, we find that a cubic polynomial will interpolate at the chosen knots and give continuity of the function and its first two derivatives at the internal knots whilst requiring one end condition at x_1 and x_{N+1} respectively.

A9.2 Computer Programme for Cubic Spline Approximation

The input to the programme consists only of the values of $x = a$ and $x = b$, specifying the range of approximation $[a,b]$, together with h, the chosen distance between the knots. This is considered fixed throughout the range and since it is assumed that both a and b are to be knots, then $(b-a)$ must be an exact multiple of h.

In addition a subroutine must be provided to compute the value of the given function $y(x)$ for any value of x supplied by the main programme.

Output provided consists of the values of the second derivative of the spline at the knots, together with the function values, the approximation and the corresponding errors tabulated for values of the independent variable at intervals equal to one-fifth of the knot-spacing over the complete range of approximation.

In order to evaluate the spline function the machine has three arrays stored (i) the values of the knots (x_k) (ii) the corresponding function values $y(x_k)$, (iii) the computed values of the second derivative (M_k). Sufficient storage space has been allocated to allow a maximum number of fifty knots to be used in any one approximation.

The output listing is given below for the example when $y = \frac{e^{-x^2}}{1+x^2}$ and the knot-spacing $h = 0.5$.

```
    DIMENSION XKNOT(50),Y(50),RM(2500),C(50),SD(50),P(50)
    DIMENSION Q(50),U(50)
    WRITE(9,105)
    READ(5,19)XMIN,XMAX,H
 19 FORMAT(3F5,3)
    INDEX=(XMAX-XMIN)/H+1
    DO21I=1,INDFX
    XKNOT(I)=XMIN+(I-1)*H
    CALL F1(XKNOT(I),Y(I))
 21 CONTINUE
    WRITE(9,102)(XKNOT(J),J=1,INDEX)
    WRITE(9,103)(Y(J),J=1,INDEX)
    DO18I=1,INDEX
    DO18J=1,INDEX
 18 RM (I+INDEX*(J-1))=0.0
    DO17I=2,INDFX-1
    C(I)=(Y(I+1)-2,0*Y(I)+Y(I-1))/H
    RM(I+INDEX*(I-2))=H/6.0
    RM(I+INDEX*(I-1))=2,0*H/3.0
 17 RM(I+INDEX*(I))=H/6,0
    RM(1)=-H*H/16.0
    RM(INDEX*2+1)=-RM(1)
    XOH=XMIN+0.5*H
    XTH=XMIN+1.5*H
    CALL F1(XOH,FOH)
    CALL F1(XTH,FTH)
    C(1)=(Y(3)-Y(1))*0.5+FOH-FTH
    RM(INDEX+INDEX*(INDEX-1))=RM(INDEX*2+1)
    RM(INDEX+INDEX*(INDEX-3))=RM(1)
    XON=XMAX-0.5*H
    XTN=XMAX-1.5*H
    CALL F1(XON,FON)
    CALL F1(XTN,FTN)
    C(INDEX)=(Y(INDEX)-Y(INDEX-2))/2.0-FON+FTN
    P(1)=RM(1)
    U(1)=C(1)/P(1)
    Q(1)=-RM(INDEX*2+1)/P(1)
    P(2)=RM(INDEX+2)
    U(2)=(C(2)-RM(2)*C(1)/RM(1))/P(2)
    Q(2)=-(RM(INDEX*2+2)-RM(2)*RM(INDEX*2+1)/RM(1))/P(2)
    DO150J=3,INDEX-1
    P(J)=RM(INDFX*(J-1)+J)+RM(INDEX*(J-2)+J)*Q(J-1)
    U(J)=(C(J)-RM(INDEX*(J-2)+J)*U(J-1))/P(J)
150 Q(J)=-RM(INDEX*J+J)/P(J)
    P(INDEX)=RM(INDEX*(INDEX-2))*Q(INDEX-2)
    U(INDEX)=(C(INDFX)-RM(INDEX*(INDEX-2))*U(INDEX=2))/P(INDEX)
    Q(INDEX)=-RM(INDEX*INDEX)/P(INDEX)
    SD(INDEX)=(U(INDEX)-U(INDFX-1))/(Q(INDEX-1)-Q(INDEX))
    DO151J=1,INDEX-2
151 SD(INDEX-J)=SD(INDEX-J+1)*Q(INDEX-J)+U(INDEX-J)
    SD(1)=U(1)+Q(1)*SD(3)
    WRITE(9,104)(SD(J),J=1,INDEX)
102 FORMAT(1H0,20X,23HX-VALUES TAKEN AS KNOTS/1H0,7F16.4/
   1(1H ,7F16.4))
103 FORMAT(1H0,20X,30HFUNCTION VALUES AT GIVEN KNOTS/
   11H0,7E16.8/(1H ,7E16.8))
104 FORMAT(1H0,20X,27HVALUES OF SECOND DERIVATIVE/1H0,7F16.4/
   1(1H ,7F16.4))
105 FORMAT(1H0,30X,26HCUBIC SPLINE APPROXIMATION)
106 FORMAT(1H0,8X,1HX,12X,8HFUNCTION,8X,6HAPPROX,17X,5HERROR)
    WRITE(9,106)
```

```
      X=XMIN
      DELX=H/5.0
  101 DO32I=2,INDEX
      IF(X-XKNOT(I))31,31,32
   32 CONTINUE
   31 VAL1=XKNOT(I)-X
      VAL2=X-XKNOT(I-1)
      S=SD(I-1)*VAL1**3/(6.0*H)+SD(I)*VAL2**3/(6.0*H)+(Y(I-1)-H*H*
     1SD(I-1)/6.0)*VAL1/H+(Y(I)-H*H*SD(I)/6.0)*VAL2/H
      CALL F1(X,FUNV)
      ERROR=FUNV-S
      WRITE(9,25)X,FUNV,S,ERROR
   25 FORMAT(1H ,5X,F7.3,8X,F10.6,5X,F10.6,10X,E15.6)
      X=X+DELX
      IF(X.GT.(XMAX+DELX/2.0))STOP
      IF(X.GT.(XMAX-DELX/2.0))GO TO 22
      GOTO101
   22 I=INDEX
      GO TO 31
      END


      SUBROUTINE F1(Z,FVAL)
      FVAL=EXP(-Z*Z)/(1.0+Z*Z)
      RETURN
      END
```

```
                    CUBIC SPLINE APPROXIMATION

          X-VALUES TAKEN AS KNOTS

    0.0000         0.5000         1,0000         1,5000         2,0000

          FUNCTION VALUES AT GIVEN KNOTS

0.10000000E 01  0,62304063E 00  0,18393972E 00  0,32430531E-01  0,36631278E-02

          VALUES OF SECOND DERIVATIVE

   -5,6544         0.6710         1,4792         0,3145         0,2087

    X             FUNCTION        APPROX               ERROR
  0,000           1,000000        1.000000          0,727596E-11
  0,100           0,980247        0.987093         -0,684590E-02
  0,200           0,923836        0,930293         -0,645726E-02
  0,300           0,838469        0,842251         -0,378172E-02
  0,400           0,734607        0.735616         -0,100953E-02
  0,500           0,623041        0,623041         -0,727596E-11
  0,600           0,512997        0.515335         -0,233819E-02
  0,700           0,411159        0.415956         -0,479769E-02
  0,800           0,321520        0.326520         -0,499990E-02
  0,900           0,245778        0,248642         -0,286404E-02
  1,000           0,183940        0,183940          0,000000E 00
  1,100           0,134931        0,133372          0,155915E-02
  1,200           0,097102        0,095266          0,183528E-02
  1,300           0,068595        0,067294          0,130076E-02
  1,400           0,047587        0,047125          0,462218E-03
  1,500           0,032431        0,032431         -0,454747E-12
  1,600           0,021715        0,021234          0,481079E-03
  1,700           0,014287        0,012970          0,131678E-02
  1,800           0,009237        0,0C7428          0,180852E-02
  1,900           0,005868        0,004396          0,147166E-02
  2,000           0,003663        0,003663         -0,113687E-11
```

LIST OF REFERENCES

1. Ahlberg J. H., Nilson E.N. amd Walsh, J.L.: The Theory of Splines and Their Applications, Academic Press, 1967

2. Blanch G.; Numerical Evaluation of Continued Fractions, SIAM Review Volume 6, No.4, October 1964

3. Cody W.J.: A Survey of Practical Rational and Polynomial Approximation of Functions, SIAM Review, Volume 12, No.3, July, 1970

4. Curtis A.R. and Osborne M.R.: The Construction of Minimax Rational Approximations to Functions, Computer Journal Volume 9, 1966.

5. Curtis, A.R. and Powell, M.J.: Error Analysis for Equal-Interval Interpolation by Cubic Splines, UKAEA Harwell Report R5600, 1967

6. Fox, L and Parker, I.B.: Chebyshev Polynomials in Numerical Analysis, Oxford, 1968.

7. Handscomb, D.C.: Methods of Numerical Approximation, Pergammon 1966

8. Hart, J.F.; Cheyney, E.W. and others: Computer Approximations, Wiley 1968

9. Hayes, J.G.: Numerical Approximations to Functions and Data, Athlone Press, 1970

10. Hildebrand, F.B.: Introduction to Numerical Analysis, McGraw Hill, 1956

11. Lanczos, C.: Applied Analysis, Pitman 1957

12. Ralston, A.: A First Course in Numerical Analysis, McGraw Hill, 1965

13. Rice, J.R.: The Approximation of Functions Volume I, Addison Wesley 1964

14. Snyder, M.A.: Chebyshev Methods in Numerical Approximation, Prentice Hall, 1966

15. Stoer, J.: A Direct Method for Chebyshev Approximation by Rational Functions, J. Assoc. of Com. Machinery 1964

16. Todd, J.: Survey of Numerical Analysis, McGraw Hill, 1962

17. Usow, K.H.: On L_1 Approximation: Computation for Continuous Functions and Continuous Dependence, SIAM J of Num. Anal. Volume 4 1967

18. Oliver, J.: An Error Estimation Technique for the Solution of Ordinary Differential Equations in Chebyshev Series, Computer Journal Vol. 12 1969