CONTROL DATA® 6600 Computer System
Programming System/Library Functions

A STUDY OF MATHEMATICAL APPROXIMATIONS

FIRST EDITION
Introduction

In recognition of the advanced design techniques incorporated in the logic of the 6600, the development of the 6600 programming systems included a mathematical investigation of the library functions intended for the 6600. Previous methods of approximations were analyzed in terms of the 6600 and, where appropriate, new algorithms were programmed and tested using a CONTROL DATA 1604A in double-precision.

As a general goal, approximations of the required accuracy were sought which were as short as possible with the constraint that the range reduction mechanism be of reasonable complexity. In most cases, continued fraction forms appear best, although polynomial and rational forms were derived for many of the functions in order not to prejudge during the development phase the form most efficient in practice. Therefore, rather than limiting consideration to algorithms optimized for the 6600, this document includes a general discussion of methods along with presentation of a variety of algorithms selected to meet the general requirements. Consequently, the algorithms and the given coefficients, which were taken directly from the 1604A double precision results, are designed to achieve the stipulated accuracy but do not take into account the 6600 round-off effects.

Although the information derived here was gathered during the study of 6600 library function algorithms, the exact techniques implemented in the original library version of the 6600 systems or subsequent revisions thereto are not necessarily contained in this document.
# Table of Contents

INTRODUCTION ................................................................. 1

I. CHARACTERISTICS OF THE 6600 ........................................ 1

II. METHOD OF TESTING ....................................................... 2

III. GENERAL METHODS OF APPROXIMATION .............................. 3
     A. POLYNOMIAL APPROXIMATIONS .................................. 3
        1. Truncated Taylor-Maclaurin Power Series ...................... 3
        2. Chebyshev Expansions ........................................... 3
        3. Telescoped Polynomials ......................................... 4
        4. Best-Fit Polynomials ............................................ 4
     B. RATIONAL APPROXIMATIONS ....................................... 5
        1. Padé Approximants and Table .................................... 5
        2. Maehly’s Method .................................................. 6
        3. Best-Fit Rational Approximations ................................ 7
        4. Conversion of Rational Approximations to Continued Fractions 7
     C. CONTINUED FRACTIONS .............................................. 8
        1. Notation and Methods of Evaluating ............................ 8
        2. Error Estimate .................................................... 8
        3. Equivalent Continued Fractions .................................. 8
        4. Even and Odd Part Contractions ................................ 8
        5. Continued Fractions Equivalent to Series ...................... 9
        6. Functions Expressed as Continued Fractions ................... 9
        7. Telescoping Procedures for Continued Fractions (Maehly) .... 10
     D. RANGE REDUCTION .................................................... 11

IV. APPROXIMATIONS OBTAINED FOR LIBRARY FUNCTIONS ............... 12
     A. SQUARE ROOT ....................................................... 12
        1. Reduction of Range ............................................... 12
        2. Computation of $\sqrt{x}$ for $\frac{1}{4} \leq x < 1$ or $\frac{1}{4} < x \leq 1$ .......... 12
        3. Estimates for $y_1$ .............................................. 12
     B. CUBE ROOT .......................................................... 13
        1. Reduction of Range ............................................... 13
        2. Computation of $\sqrt[3]{x}$ for $\frac{1}{2} \leq x < 1$ ............... 13
        3. Estimates for $y_1$ .............................................. 13
     C. SIN u ............................................................... 14
        1. Range Reduction .................................................. 14
# Table of Contents (Continued)

2. Taylor-Maclaurin Series ................................. 14
3. Telescoped Polynomials .................................. 14
   a. \(|x| \leq \frac{\pi}{2}\) ............................... 14
   b. \(|x| \leq \frac{\pi}{6}\) .............................. 15
4. Padé Rational Approximation for Sin \(x\), \(|x| \leq \frac{\pi}{6}\) ........................................... 16
5. Comparison of Results .................................. 18

D. TAN \(u\) .................................................. 18
   1. Range Reduction .................................... 18
   2. Rational and Continued Fraction Forms Obtained from the
      Gaussian Continued Fraction ....................... 19
      a. \(|x| \leq \frac{\pi}{8}\) ................................. 19
      b. \(|x| \leq \frac{\pi}{4}\) ............................... 19
   3. Telescoped Rational and Continued Fraction Forms for \(|x| \leq \frac{\pi}{4}\) ......................... 19
   4. Comparison of Results ............................. 20

E. ARCTAN \(u\) ............................................. 21
   1. Range Reduction ................................... 21
   2. Rational and Continued Fraction Forms Obtained from the
      Gaussian Continued Fraction ....................... 21
      a. \(|x| \leq \sqrt{2} - 1\) ............................... 21
      b. \(|x| \leq \frac{\pi}{16}\) ............................... 21
   3. Telescoped Rational and Continued Fraction Forms for
      \(|x| \leq \frac{\pi}{16}\) ................................. 21
   4. Comparison of Results ............................. 22

F. ARCSIN \(u\) ............................................. 22
   1. Range Reduction ................................... 22
   2. Telescoped Polynomials for Arcsin \(x\), \(|x| \leq \frac{1}{2}\) ................................. 23
   3. Test Results ....................................... 24

G. EXPONENTIAL: \(e^u\) .................................... 24
   1. Range Reduction ................................... 24
   2. Taylor-Maclaurin Series .......................... 24
   3. Padé Rational .................................... 24
Table of Contents (Continued)

4. Rational and Continued Fraction Forms Obtained from Macon's
   Even Part of the Gaussian Continued Fraction for $e^u$ .......................... 24
5. Telescoped Rational and Continued Fraction Forms .................................. 25
6. Comparison of Results ........................................................................... 25

H. LOGARITHM: $\log_u u = \ln u$ ................................................................. 25
1. Reduction of Range .................................................................................. 25
2. Taylor-Maclaurin Series ......................................................................... 25
3. Telescoped Polynomial ........................................................................... 26
4. Rational and Continued Fraction Forms Obtained from the
   Gaussian Continued Fraction ....................................................................... 26
5. Telescoped Rational and Continued Fraction Forms ................................ 26
6. Comparison of Results ........................................................................... 27

V. REFERENCES ............................................................................................. 28

APPENDIX A ................................................................................................. 30
Formulae for conversion of a quotient of two nth order polynomials to
continued fraction form and for evaluating the resulting continued fraction.

1. COF 2: $n = 2$ ..................................................................................... 30
2. COF 3: $n = 3$ ..................................................................................... 30
3. COF 4: $n = 4$ ..................................................................................... 30

APPENDIX B ................................................................................................. 32
Gaussian continued fractions and their approximants.

1. Tan $x$ .................................................................................................... 32
2. Arctan $x$ ............................................................................................ 32
3. Exponential: $e^x$ .................................................................................. 33
4. Logarithm: $\log_e \frac{1+x}{1-x}$ .............................................................. 33

APPENDIX C ................................................................................................. 34
Constants
I. Characteristics of the 6600*

The CONTROL DATA 6600 is a large-scale, solid-state, general-purpose digital computing system composed of eleven independent computers. Ten of these are peripheral and control processors, each with a 12-bit 4096-word memory. The eleventh computer, the central processor, is a high-speed arithmetic device. The common element between these computers is a random-access central memory of 131.072 words (of length 60-bits) organized in 32 banks of 4096 words each. It is the central processor whose characteristics are to be considered in selecting appropriate algorithms.

The central processor has ten independent arithmetic and logical units which operate concurrently in the solution of a problem. In addition, it has 24 operating registers for functional units and 8 transistor registers for servicing functional units. A word length is 60 bits, 48 bits of which determine the integer coefficient, 11 bits the biased exponent, and 1 bit, the coefficient sign. Execution times for floating-point operations are as follows:

**Floating-point add:**
4 minor cycles = 4(100)/(10^8)SECS = 4(10^{-5})SECS

**Floating-point multiply:**
10 minor cycles = 10(100)/(10^8)SECS = 10^{-5} SECS

---

Floating-point divide:
29 minor cycles = 29(100)/(10^8)SECS
= 2.9(10^{-5})SECS

Hence, relative speeds are
M = 2.5A and D = 2.9M = 7.25A

Central processor instructions are sent automatically and in the original sequence to the instruction stack which holds up to 32 instructions. A branch to another area of the program voids the old instructions in the registers and brings in new ones. Branch orders of the type "GO TO K IF B1 < B2" require 6(10^{-5}) SECS and an additional 5(10^{-5}) SECS for a branch to an instruction which is out of the stack. High speed in the central processor depends upon minimizing memory references and waiting time for unrelated instructions and partial answers.

Since arithmetic computations are extremely efficient and several operations may be done simultaneously, algorithms which break down into independent blocks which can be computed in parallel are desirable. On the other hand, it is surmised that division of the interval of definition of a variable into many small sub-intervals, requiring the computer to do much testing and branching to other blocks of the program not in the stacker, is not very efficient for the 6600.

*(See Ref. 17)*
II. Method of Testing

Let $F(x)$ be the function to be considered as the correct one, and let $Y(x)$ be the approximation to $F(x)$ being tested. Define $A(x) = |F(x) - Y(x)|$ as the absolute error in $Y$ over some $x$ range, and $R(x) = \frac{A(x)}{F(x)}$ as the relative error in $Y$ over this range. For floating-point subroutines, accuracy is defined by the number of first correct significant digits, so that if

$$R \leq 5\left(10^{-\frac{n+1}{10}}\right)$$

the $n$ first significant digits are correct. For fixed-point subroutines, the absolute error measures accuracy. If

$$A \leq 5\left(10^{-\frac{n+1}{10}}\right)$$

then $n$ digits after the decimal point are correct. Since all approximations considered here are to be in floating-point, it shall be required that

$$R < 2^{-19} \sim 1.775(10^{-19})$$

which is the basic roundoff due to the size of the 6600 register.

The algorithms $Y(x)$ described here have been programmed on the 1604 in FORTRAN 63 using double-precision, and compared with double-precision library routines $F(x)$ supplied by Palo Alto. This provided a test for truncation error in the algorithm itself, but, of course, gave no effect of 6600 roundoff since the floating-point word length for the 6600 in single-precision is between those of the 1604 used in single and double precision modes.
III. General Methods of Approximation*

Three forms of approximations have been considered—(1) polynomials, (2) rationals, and (3) truncated continued fractions, together with techniques for improving convergence and for converting from one form to another.

The following sections attempt to give in capsule form the theoretical background for the approximations developed in later sections, though not all methods discussed were actually used in these approximations.

A. POLYNOMIAL APPROXIMATIONS

1. Truncated Taylor-Maclaurin Power Series

One of the most common approximations upon which many others are based is the Taylor-Maclaurin power series truncated to m terms,

\[ f(x) \approx \sum_{n=0}^{m} \frac{f^{(n)}(0)}{n!} x^n, \quad |x| < a, \text{ where } m \text{ is chosen sufficiently large to insure the desired accuracy.} \]

The absolute error is less than the maximum of the absolute value of the first neglected term over \(|x| < a\). Such approximations usually require too many terms to be used directly unless the interval \(|x| < a\) is so subdivided that the full number of terms used is only a small proportion of the time. (In this connection, see Ref. 13 on the subject of "Partitioned Polynomials"). More often, Taylor-Maclaurin series provide a starting point from which more efficient routines can be built.

2. Chebyshev Expansions (See Ref. 3 and 5)

Denote by \(T_n(x) \equiv \cos n\theta\) the Chebyshev polynomial of degree \(n\) in \(x = \cos \theta\). It is simply the polynomial in \(x\) obtained by expressing \(\cos n\theta\) in terms of \(\cos \theta\), then replacing \(\cos \theta\) by \(x\). Note that \(|x| \leq 1\) and that \(|T_n(x)| \leq 1\). The polynomial \(T_{n+1}(x)\) attains its greatest absolute value, one, in the interval \([-1,1]\) at \(n+2\) points (including endpoints) with alternating sign. These polynomials may be generated by the recursion

\[ T_{n+1}(x) = 2x T_n(x) - T_{n-1}(x) \]

with \(T_0(x) = 1\) and \(T_1(x) = x\).

Writing

\[ T_n(x) = C_n^0 + C_n^1 x + C_n^2 x^2 + \ldots + C_n^n x^n \]

the coefficients \(C_n^m\) are computed from

\[ C_n^m = 0 \quad \text{if } (n+m) \text{ is odd} \]

\[ C_n^m = \frac{2}{m} \left( \begin{array}{c} \binom{n+m}{m/2} - \binom{n+m-2}{m/2} \end{array} \right) (-1)^{n-m} \frac{2}{m} \]

if \((n+m)\) is even

Alternately, the shifted Chebyshev polynomials

\[ T_n(x) = T_n(2x-1), \]

obtained by the transformation \(x = \frac{1 + \cos \theta}{2}\), i.e., by replacing \(x\) by \((2x-1)\) in the polynomial \(T_n(x)\), can be used for the range \(0 \leq x \leq 1\). Coefficients for \(T_n(x)\) are found from

\[ C_n^m = 2^{2m-1} \left[ \frac{2}{m} \left( \binom{n+m}{m/2} - \binom{n+m-1}{m/2} \right) (-1)^{n-m} \right] \]

Tables of Chebyshev coefficients may be found in Refs. 3 and 13, and also in publications by the National Bureau of Standards. Reference 7 provides an excellent source of information for locating appropriate tables.

The Chebyshev expansion of \(f(ax)\) over the interval \(-a \leq ax \leq a\), truncated after the \(m\)th term, is written

\[ f(ax) \approx \frac{C_0(a)}{2} + \sum_{n=1}^{m} C_n(a) T_n(ax), \quad |x| \leq 1 \]

where

\[ C_n(a) = \frac{2}{\pi} \int_{-1}^{1} f(ax) T_n(x)(1-x^2)^{-1/2} \, dx \]

* (See Ref. 5, 4, 1, 2 and 18)
The truncation error is approximately $C_{n+1}(a) T_n(x)$. Since the coefficients in this expansion depend upon values of $f(ax)$ in the entire interval $(-a, a)$ rather than only upon values at zero, the approximation is more efficient and converges more rapidly than the Taylor-Maclaurin series as $n$ increases. However, the coefficients $C_n(a)$ are often extremely troublesome to compute accurately.

3. Telescoped Polynomials (See Ref. 3 and 5)

If $f(x)$ is an arbitrary polynomial of degree $n+1$, the “best” polynomial approximation of degree $n$ in $[-1, 1]$ is $P_n(x) = f(x) - a_{n+1} T_{n+1}(x)$ where $a_{n+1}$ is a constant chosen so that the coefficient of $x^{n+1}$ on the right-hand side vanishes. In the sense of the next Section (III.A.4) the polynomial $T_{n+1}(x)/2^n$ is the unique “best” approximation to zero in $[-1, 1]$ with leading term exactly

$$x^{n+1}; \text{or, } x^{n+1} = \frac{T_{n+1}(x)}{2^n}$$

is the unique “best” approximation to $x^{n+1}$ of degree $n$. This fact forms the basis for a telescoping procedure described in the following manner.

Let $f(x) \approx a_n + a_{n-1} x + a_{n-2} x^2 + \ldots + a_0 x^{n+1}, \ |x| \leq 1$ or $x \in [0, 1]$, represent $f(x)$ to the required accuracy in the interval. Usually $f(x)$ is represented by a truncated Taylor-Maclaurin power series. If the range of $x$ is not $[-1, 1]$ or $[0, 1]$, a simple transformation can be made to a variable with such a range.

$$T_{n+1}(x) \quad \frac{C_{n+1}}{C_{n+1}}$$

$$= x^{n+1} + t_0 x^n + t_{n-1} x^{n-2} + \ldots + t_1 x + t_0$$

where

$$t_i = C_{n+1}/C_{n+1}$$

Next try replacing $x^{n+1}$ by

$$x^{n+1} = \frac{T_{n+1}(x)}{C_{n+1}}$$

Next try replacing $x^{n+1}$ by

$$x^{n+1} = \frac{T_{n+1}(x)}{C_{n+1}}$$

Letting

$$a'_i = a_i - a_{n+1} t_i \quad \text{for } i = 0, n,$$

the result is

$$f(x) \approx a'_0 + a'_1 x + a'_2 x^2 + \ldots + a'_n x^n + \frac{a_{n+1} T_{n+1}(x)}{C_{n+1}}$$

Since

$$\left| T_{n+1}(x) \right| \leq 1,$$

we have

$$\left| \frac{a_{n+1} T_{n+1}(x)}{C_{n+1}} \right| \leq \left| \frac{a_{n+1}}{C_{n+1}} \right| = E_1.$$

Let $E_0$ be the original truncation error incurred by using terms of the Taylor series only through $(n+1)$, and let $E$ be the allowable error.

Then if $E_0 + E < E$, the term $a_{n+1} T_{n+1}(x)/C_{n+1}$ may be dropped, and the result is a polynomial approximation for $f(x)$ of degree $n$ with error less than $E$. The process may be repeated so long as $\Sigma E_i < E$. When using $T_{n+1}(x)$ for telescoping, the highest power of $x$ will decrease by two with each substitution, since $T_{n+1}(x)$ contains only alternate powers of $x$. $T_{n+1}(x)$, however, contains all powers of $x$ from 0 through $n+1$, so that its use in telescoping reduces the highest power of $x$ by one with each substitution. Note that telescoping a power series using Chebyshev polynomials is not equivalent to the Chebyshev expansion described in Section III.A.2. (See Ref. 5, p. 12 for an example).

4. Best-Fit Polynomials (See Ref. 22 and 4)

Let $f(x)$ be a given function continuous on the interval $[a, b]$, and let $g(x)$ be a given weight function, continuous and positive on $[a, b]$. That polynomial, $P_n(x)$ for which

$$\max_{[a, b]} \left| \frac{P_n(x) - f(x)}{g(x)} \right| = \text{Min}$$

is called the Chebyshev-approximant or best-fit polynomial of degree $n$ (in the sense of Chebyshev) with respect to the weight function $g(x)$. The weight function allows the option of minimizing either absolute or relative error by taking $g(x) = 1$ or $g(x) = |f(x)|$ respectively.
Let \( f(x) \) be an arbitrary single-valued function defined in the closed interval \([a, b]\) and let \( p_n(x) \) be a polynomial of degree \( n \) such that the deviation

\[
e_n(x) = f(x) - p_n(x)
\]

attains its greatest absolute value \( L \) at not less than \( n+2 \) distinct points in \([a, b]\) and is alternately \(+L\) and \(-L\) at the successive points. Then \( p_n(x) \) is the best polynomial approximation of degree \( n \) to \( f(x) \) in \([a, b]\) in the sense that the maximum value of \( |e_n(x)| \) is as small as possible.

These conditions imply a set of \((2n+2)\) equations for \( L \), the \((n+1)\) coefficients of \( p_n(x) \), and the \( n \) critical points, \( x_i \), at which the value \( \pm L \) is attained (the other two are endpoints). The assumption that \( e_n(x) \) is continuous on \([a, b]\) and is zero at each \( x_i \) is usually required. Hence the \( 2n+2 \) equations to be solved are

\[
e_n(x_i) = f(x_i) - p_n(x_i) = (-1)^i L
\]

for \( i = 0, n + 1 \)

\[
e_n'(x_i) = 0
\]

for \( i = 1, n \).

Their solution requires some sort of iterative procedure. (See Ref. 22).

If \( p_n(x) \) exists, it is unique. If \( f(x) \) is a continuous function in \([a, b]\), then there exists a unique polynomial of best approximation of given degree. If \( f(x) \) is a polynomial of degree \( n+1 \), the best polynomial approximation of degree \( n \) in \([-1, 1] \) is \( p_n(x) = f(x) - a_{n+1} T_{n+1}(x) \), where \( a_{n+1} \) is a constant chosen so that the coefficient of \( x^{n+1} \) on the right-hand side vanishes. No simple explicit expression is known for the best polynomial approximation of given degree to an arbitrary function \( f(x) \).

**B. RATIONAL APPROXIMATIONS—QUOTIENT OF TWO POLYNOMIALS**

1. Padé Approximants and Table

   (See Refs. 6, 5)

Given a power series \( P(x) = \sum_{i=0}^{\infty} C_i x^i \) and a pair of non-negative integers \((m, n)\), there exists a uniquely determined rational fraction, \( R_m^n(x) \), whose numerator and denominator are of degrees less than or equal to \( m \) and \( n \) respectively and whose expansion in ascending powers of \( x \) agrees term by term with \( P(x) \) for more terms than that of any other such rational fraction. \( R_m^n(x) \) is called a Padé approximant of \( P(x) \) and the associated table formed by putting \( R_m^n(x) \) in the \((n+1)\)st row and \((m+1)\)st column, \( m=0, 1, 2, \ldots \), is called a Padé table.

Let

\[
P(x) = \sum_{i=0}^{\infty} C_i x^i, \quad A_m(x) = \sum_{i=0}^{m} a_i x^i, \quad B_n(x) = \sum_{i=0}^{n} b_i x^i.
\]

Then there are \( m+n+2 \) coefficients \( a_i, b_i \) to be found.

Hence, if we write

\[
P(x) = \frac{A_m(x)}{B_n(x)} + \sum_{k=m+1}^{\infty} C_k x^k,
\]

\[
P(x) B_n(x) - A_m(x) B_n(x) = B_n(x) \sum_{k=n+m+1}^{\infty} C_k x^k \equiv \sum_{k=n+m+1}^{\infty} d_k x^k,
\]

or

\[
\left[ C_0 b_0 + (C_1 b_0 + C_2 b_1) x + \ldots + \sum_{i=0}^{k} C_{i-1} b_i x^i + \ldots + \sum_{i=0}^{n} C_{i-1} b_i x^n \right] +
\]

\[
\left[ \sum_{i=0}^{n+1} C_{i-1} b_i x^{i+1} + \ldots + \sum_{i=0}^{n} C_{k-1} b_i x^k + \ldots + \sum_{i=0}^{n} C_{n+m-1} b_i x^{n+m} \right] +
\]

\[
+ \ldots - [a_0 + a_1 x + \ldots + a_m x^m] = d_{n+m+1} x^{n+m+1} + \ldots.
\]
and set all coefficients of \( x \) through \( x^{m-1} \) to zero, the result is

\[
\sum_{k=0}^{n} C_{k+1} b_k - a_k = 0 \quad \text{for } k = 0, n
\]

\[
\sum_{k=0}^{n} C_{k+1} b_k - a_k = 0 \quad \text{for } k = n+1, n+m
\]

\[a_k = 0 \quad \text{for } k > m\]

which provide \( n+m+1 \) equations to be solved for \( a_i \) and \( b_i \). These equations will not contain \( a_i \)'s for \( k = m+1, m+n \), so will provide \( n \) equations for finding \( (n+1) \) \( b_i \)'s. Since one of the \( b_i \)'s must be arbitrary, choose \( b_n = 1 \).

Thus, to solve the above equations for the coefficients of the rational approximation

\[ R_n(x) \]

set

\[ b_n = 1 \]

and solve the \( n \) equations,

\[ \sum_{k=0}^{n} C_{k+1} b_k - a_k = 0 \quad \text{for } k = m+1, m+n \quad \text{for } b_i, i = 1, n.\]

Then solve the \( m+1 \) equations

\[ \sum_{k=0}^{m+n} C_{k+1} b_k = a_k, \quad k = 0, m \]

for \( a_i, i = 0, m \)

and

\[ R_n(x) = \left( \sum_{i=0}^{m} a_i x^i \right) \left/ \left( 1 + \sum_{i=1}^{n} b_i x^i \right) \right. \]

Padé approximants are most useful for \( m = n \) or \( m = n + 1 \). P(x) is normally taken to be the Taylor-Maclaurin expansion of the function to be approximated and \( n \) and \( m \) chosen so that the terms of the series through \( x^{m+n} \) yield an approximation with the accuracy desired. If \( m = n \), the resulting rational function is more accurate than the series through \( x^{m+n} \), i.e., the rational function actually agrees with more terms of the series than required. An error estimate is given in Ref. 5, p. 14, which requires the evaluation of the quotient of two determinants of orders \( (n+1)x(n+1) \) and \( nxn \) for the case \( m = n \). The elements of the determinants are \( C_i \)'s.

2. Maehly's Method (See Refs. 5 and 13)

Rational approximations may also be derived from Chebyshev expansions as shown by H. Maehly. If a function \( f(ax) \) in the interval \(-a \leq ax \leq a\) is expanded in its Chebyshev series,

\[ f(ax) = \sum_{i=0}^{\infty} C_i T_i(x), \quad (\text{See Section III.A.2.}) \]

and

\[ A_n(x) = \sum_{i=0}^{m} a_i T_i(x), \quad B_n(x) = \sum_{i=0}^{n} b_i T_i(x) \]

then the \( (n+m+1) \) unknowns \( a_i, b_i \) (\( b_n = 1 \)) are determined from

\[ \left( \sum_{i=0}^{m} b_i T_i(x) \right) \left( \sum_{i=0}^{n} C_i T_i(x) \right) - \left( \sum_{i=0}^{m} a_i T_i(x) \right) \]

\[ = \sum_{i=m+n+1}^{\infty} h_i T_i(x). \]

Using the relation \( T_{m+n}(x) + T_{m-n}(x) = 2T_m(x)/T_n(x) \), the following system of \( n+m+1 \) linear equations is obtained for the \( a_i \) and \( b_j \):

\[ a_0 = C_0 + \frac{1}{2} \sum_{i=1}^{n} b_i C_i \]

\[ a_1 = C_1 + \frac{1}{2} b_1 C_1 + \frac{1}{2} \sum_{j=1}^{n} b_1 (C_{j+1} + C_{j-1}), \quad j = 1, m+n \]

where \( a_i = 0 \) for \( j > m \) and \( b_j = 0 \) for \( j > n \). Note that it is necessary to find the coefficients \( C_i \) before Maehly's Method may be used. An error estimate is given in Ref. 5, p. 16, and a detailed example from Aresin x is given in Ref. 13, pp. 123-131.
3. Best-Fit Rational Approximations

(See Ref. 22)

Let

\[ A_m(x) = \sum_{i=0}^{m} a_i x^i \] be the numerator and

\[ B_n(x) = \sum_{i=0}^{n} b_i x^i \] the denominator of a rational

function \( R_N(x) = \frac{A_m(x)}{B_n(x)} \), \( N = n + m \).

In the same manner as was done for polynomials, a rational best-fit approximation of order \( N \) to the continuous function \( f(x) \) on the interval \([a, b]\) is defined as that rational function \( R_N^*(x) \) for which

\[
\max_{[a, b]} \frac{| R_N^*(x) - f(x) |}{g(x)} = \min.
\]

where \( g(x) \) is a given weight function, continuous and positive in \([a, b]\), e.g., \( g(x) = 1 \) or \( g(x) = |f(x)| \).

N. Achiesser has shown that \( R_N^*(x) \) is uniquely characterized by its error curve,

\[ \delta^*(x) = \frac{R_N^*(x) - f(x)}{g(x)} \]

assuming its maximum absolute value sufficiently often with alternating signs. Arguments \( x_i \) for which the maximum absolute value is assumed are called critical points.

An error curve has standard form if it meets the additional requirements that it has exactly \( N + 2 \) critical points, the first and last of which are endpoints of the interval, and has a continuous derivative with respect to \( x \) which vanishes at the critical points.

If an error curve has standard form, it is necessarily the optimal error curve corresponding to the best-fit rational \( R_N^*(x) \). The converse is not necessarily true—the optimal error curve need not have standard form.

\( \delta^*(x) \) in standard form yields \( 2N + 2 \) equations in the \( 2N + 2 \) unknowns \( x_i^*, i=1, N \); \( a_i^*, i=0, m \); \( b_i^*, i=1, n \), and the maximum error, \( \lambda \). Since one of the coefficients in \( R_N^*(x) \) is arbitrary, assume \( b_0 = 1 \).

\[
\delta(x^*_i) = (-1)^i \lambda, \ i = 0, N+1
\]

\[
\delta^*(x^*_i) = 0, \ i = 1, N
\]

This is a non-linear system of equations whose solution requires an iterative procedure. The term “direct” method is used to indicate that the coefficients of the best-fit rational are computed directly from the equations above, whereas an “indirect” method determines the corrections necessary to modify a fixed approximant (e.g., a Padé approximant) to obtain the best-fit rational. Details for several direct, indirect and combined methods, due largely to the late H. Maehly, may be found in Reference 22.

4. Conversion of Rational Approximations to Continued Fractions

Any rational fraction may be converted to an equivalent continued fraction. The continued fraction form evaluated from bottom to top (See Section III.C.I.) is nearly always more efficient (i.e., requires fewer operations) to evaluate than the rational form.

Assuming the degree \( m \) of the numerator is \( \leq \) the degree \( n \) of the denominator, set the rational function identically equal to the continued fraction:

\[
\frac{a_0 + a_1 z + a_2 z^2 + \ldots + a_m z^m}{b_0 + b_1 z + b_2 z^2 + \ldots + b_n z^n}
\]

\[ = C_0 + \frac{C_1}{(z+B_1) + \frac{C_2}{(z+B_2) + \ldots + \frac{C_n}{(z+B_n)}}} \]

Since the \( a_i \) and \( b_i \) are known, the \( C_i \) and \( B_i \) are found by equating coefficients of like powers of \( z \) after converting the right-hand side to a rational form. The resulting equations are non-linear, but easy to solve. Results for \( n = 2, 3, 4 \) are given in Appendix A as COF 2, COF 3 and COF 4 respectively.
C. CONTINUED FRACTIONS
(See References 6 and 8)

1. Notation and Methods of Evaluating

Let

\[ F = b_n + \frac{a_n}{b_{n-1} + \frac{a_{n-1}}{b_{n-2} + \frac{\ddots}{\ddots}}} \]

be a continued fraction.

F may also be written in either of the following two notations:

\[ F = b_0 + \frac{a_1}{b_1 + \frac{a_2}{b_2 + \frac{\ddots}{\ddots}}} \]

\[ F = b_0 + \frac{a_1}{b_1} + \frac{a_2}{b_2} + \frac{a_3}{b_3} + \ldots \]

Assume in the sequel that no \( b_i \), \( i \geq 1 \) is zero and that the continued fraction converges. \( F \) can be evaluated for \( n \) terms in several ways:

(a) Top to bottom

\[ A_n = b_n + a_n A_{n-1} \quad (n = 1, 2, \ldots) \]

\[ B_n = b_n + a_n A_{n-2} \]

where \( A_1 = 1, A_0 = b_0, B_1 = 0, B_0 = 1 \).

\( F_n = A_n / B_n \) is the \( n \)th approximant to \( F \).

(b) Bottom to top

\[ P_1 = a_n \]

\[ Q_1 = b_n \]

\[ P_i = a_{n-i} Q_{i-1} + P_{i-1} \quad (i = 2, n) \]

\[ Q_i = b_{n-i} Q_{i-1} + P_{i-1} \]

\( F_n = b_n + P_n / Q_n \)

2. Error Estimate

In determining the size of \( n \) required to give a desired accuracy, the following “determinant formula” is often helpful:

\[ \frac{A_n}{B_n} - \frac{A_{n-1}}{B_{n-1}} = (-1)^{n-1} \frac{a_1 a_2 \ldots a_n}{B_{n-1} B_n} \]

where \( B_{n-1} \) and \( B_n \) are defined in the previous Section 1. Note that this is merely an expression for the difference between the \((n-1)\)st and the \( n \)th approximants and says nothing about \( F - F_n \). No simple formula for the error \( F - F_n \) is known except for special cases.

3. Equivalent Continued Fractions

The continued fraction

\[ \bar{F} = C_0 b_0 + C_1 C_2 a_1 + C_2 C_3 a_2 + \ldots + C_{n-1} C_n a_n + \ldots \]

where \( C_i \neq 0 \), has approximant \( E_n / D_n \) where

\[ E_n = C_0 C_1 \ldots C_{n-a} \]

\[ D_n = C_1 \ldots C_n B_n \]

\[ D_0 = B_0 = 1 \]

and \( A_n / B_n \) is the \( n \)th approximant of \( F \) as defined in 1.

We have \( \frac{E_n}{D_n} = \frac{C_n A_n}{B_n} \)

and both converge or diverge together.

If \( C_n = 1 \), the two continued fractions \( F \) and \( \bar{F} \) are equivalent. This equivalence is useful for changing the form of a given continued fraction, say to one in which the numerators (or denominators) are all one.

4. Even and Odd Part Contractions

The even part of a continued fraction is the continued fraction whose approximants are \( A_{2n} / B_{2n} \); simi-
larly, the odd part is the continued fraction whose approximants are

\[
\frac{A_{2n+1}}{B_{2n+2}}, \quad n = 0, 1, \ldots
\]

The even and odd part contractions converge if the original fraction does and to the same value.

\[
F_{\text{even}} = b_0 + \frac{a_1 b_2}{(b_1 b_2 + a_2)} - \frac{a_2 a_3 b_4}{(b_2 b_3 + a_3) b_4 + b_2 a_4} - \frac{a_3 a_4 b_6}{(b_3 b_4 + a_4) b_6 + b_3 a_6} - \frac{a_4 a_5 b_8}{(b_4 b_5 + a_5) b_8 + b_4 a_8} - \ldots
\]

\[
F_{\text{odd}} = \frac{b_0 b_1 + a_1}{b_1} - \frac{a_2 a_1 b_2}{(b_1 b_2 + a_2) b_2 + b_1 a_2} - \frac{a_3 a_2 b_3}{(b_2 b_3 + a_3) b_3 + b_2 a_3} - \frac{a_4 a_3 b_4}{(b_3 b_4 + a_4) b_4 + b_3 a_4} - \ldots
\]

5. Continued Fractions Equivalent to Series

The series \( C_0 + C_1 + \ldots + C_n \ldots \) and the continued fraction

\[
C_0 + \frac{C_1}{1 - \frac{C_2/C_1}{1 + C_2/C_1} - \frac{C_3/C_2}{1 + C_3/C_2} - \ldots}
\]

are equivalent in the sense that

\[
C_n = \frac{A_n}{B_n} - \frac{A_{n-1}}{B_{n-1}}.
\]

Similarly, the power series \( C_0 + C_1 x + C_2 x^2 + \ldots \) and the continued fraction,

\[
C_0 + \frac{C_1 x}{1 - \frac{(C_2/C_1)x}{1 + (C_2/C_1)x} - \ldots}
\]

are equivalent.

Another method for obtaining a continued-fraction expansion from a function defined by a power series is the Quotient-Difference algorithm of Rutishauser described in References 18 and 19.

6. Functions Expressed as Continued Fractions

Each of the previous Sections 1 through 5 applies when the \( a_i \) or \( b_i \) are either constant or a function of \( x \). It is assumed that the continued fraction \( F(x) \) converges for \( |x| \leq \varepsilon \).

Especially useful are the continued fraction expansions of Gauss for the functions \( \tan x \), \( \arctan x \), \( e^x \), and \( \log_e \frac{1+x}{1-x} \). These are reproduced in Appendix B together with some of their successive approximants, \( A_n/B_n \).

As an illustration of the uses of some of the relations given in this Section C, consider the Gaussian continued fraction for \( e^x \) (Appendix B.3.a.):

\[
e^x = 1 + \frac{x}{1 - \frac{x}{2 + \frac{x}{3 - \frac{x}{2 + \frac{x}{5 - \frac{x}{2 + \frac{x}{7 - \ldots}}}}}}}
\]

The even part contraction of this continued fraction (III.C.4.) is

\[
e^x = 1 + \frac{2x}{(2-x) + \left[ (6+x)2-2x \right] + \ldots}
\]

\[
e^x = 1 + \frac{2x}{(2-x) + \frac{2x^2}{12 + \frac{4x^2}{20 + \frac{4x^2}{28 + \ldots}}}}
\]

which is equivalent by Section III.C.3. to

\[
e^x = 1 + \frac{2x}{(2-x) + \frac{x^2}{6 + \frac{x^2}{10 + \frac{x^2}{14 + \ldots}}}}
\]

This is now form 3.b. of Appendix B.
7. Telescoping Procedures for Continued Fractions (Maehly) (See Reference 21)

In IIIA.3, a method was described for telescoping a truncated power series by use of Chebyshev polynomials. Maehly has derived a method by which the \((n+1)st\) approximant of a continued fraction may be telescoped one step to a corrected \(n\)th approximant.

Let

\[
f(x) = \frac{a_0}{b_0} + \frac{a_1 x}{b_1} + \frac{a_2 x}{b_2} + \ldots
\]

be a convergent continued fraction representation of \(f(x)\) whose \((n+1)st\) approximant approximates \(f(x)\) to within the desired accuracy in the interval \(|x| \leq \varepsilon\). The \((n+1)st\) approximant is

\[
R_{n+1}(x) = \frac{a_0}{b_0} + \frac{a_1 x}{b_1} + \ldots + \frac{a_{n+1} x}{b_{n+1}} = \frac{A_{n+1}(x)}{B_{n+1}(x)}
\]

where \(A_{n+1}\) and \(B_{n+1}\) are defined recursively as

\[
\begin{align*}
A_{n+1} &= b_{n+1} A_n + a_{n+1} x A_{n-1} \\
B_{n+1} &= b_{n+1} B_n + a_{n+1} x B_{n-1} \\
A_0 &= a_0, \quad A_1 = a_0 b_1 \\
B_0 &= b_0, \quad B_1 = b_0 b_1 + a_1 x
\end{align*}
\]

Now we may alter either the \(a_k\) or the \(b_k\) or the \(A_n\) and \(B_n\). Formulae are given in Reference 21 for all of these cases, but only those for altering \(A_n\) and \(B_n\) to obtain a new \(R_n\) (called \(R_n^*\)) are given here.

\[
R_n^* = \frac{A_n^*}{B_n^*} = \frac{A_n + \gamma_0 x \sum_{k=2}^{n} \gamma_k A_{k-2}}{B_n + x \gamma_1 + x \sum_{k=2}^{n} \gamma_k B_{k-2}}
\]

where

\[
\gamma_k = -s_k (-\varepsilon)^{n+1-k} \prod_{i=k}^{n+1} \frac{a_i}{b_i}, \quad x = \varepsilon u, \quad |u| \leq 1,
\]

and

\[
S^{(n+1)}(u) = \frac{T_{n+1}(u)}{2^n} = \sum_{k=0}^{n+1} s_k u^k.
\]

\(T_{n+1}(u)\) is the Chebyshev polynomial of degree \((n+1)\).

For the uncorrected \(n\)th approximant, \(R_n(x)\), it is known that

\[
\lim_{\varepsilon \to 0} \frac{R_n(\varepsilon u) - f(\varepsilon u)}{\varepsilon^{n+1}} = C_{n+1}\varepsilon^{n+1}
\]

where

\[
C_{n+1} = (-1)^n \prod_{i=0}^{k} \frac{a_i}{b_i b_{i+1}}.
\]

The corresponding limit relation for \(R_n^*(x)\) is:

\[
\lim_{\varepsilon \to 0} \frac{R_n^*(\varepsilon u) - f(\varepsilon u)}{\varepsilon^{n+1}} = C_{n+1} S^{(n+1)}(u).
\]

Hence the corrected approximant \(R_n^*(x)\) yields an “almost best-fit” rational.

Since many of the functions of interest are “even” or “odd” functions, definitions and telescoping formulae will be given for these special cases. First replace \(x\) by \(x^2\) in the recursive definitions for \(A_{n+1}\) and \(B_{n+1}\).

**Even Functions:**

An even function, \(g(x)\), is one of the form

\[
g(x) = \frac{a_0}{b_0} + \frac{a_1 x^2}{b_1} + \ldots + \frac{a_n x^{2n}}{b_n} + \ldots \quad \text{for} \quad |x| \leq \varepsilon.
\]

Let

\[
x = \varepsilon u \quad \text{where} \quad |u| \leq 1\quad \text{and} \quad S^{(2n+2)}(u) = T_{2n+2}(u) = \sum_{k=0}^{n+1} s_{2k} u^{2k}, \quad \text{where} \quad T_{2n+2}(u)
\]

is the Chebyshev polynomial of degree \(2n+2\).

Define

\[
\gamma_k = -s_{2k} \varepsilon^{2(n+1-k)} \prod_{i=k}^{n+1} \frac{a_i}{b_i} \quad \text{for} \quad k = 0, n.
\]
Then
\[
R_n^* = \frac{A_n^*}{B_n^*} = \frac{A_n + \gamma_0 + x^2 \sum_{k=2}^{n} \gamma_k A_{k-2}}{B_n + x^2 \gamma_1 + x^2 \sum_{k=2}^{n} \gamma_k B_{k-2}}
\]
and
\[
\lim_{\epsilon \to 0} \frac{R_n^*(\epsilon u) - g(\epsilon u)}{\epsilon^{2n+2}} = C_{n+1} S^{(2n+2)}(u).
\]

**Odd Functions:**

An odd function, \(f(x)\), is one of the form
\[
f(x) = \frac{a_0 x}{|b_0|} + \frac{a_1 x^2}{|b_1|} + \ldots + \frac{a_n x^n}{|b_n|} + \ldots \text{ for } |x| \leq \epsilon.
\]

Let \(x = \epsilon u\) where \(|u| \leq 1\) and
\[
S^{(2n+3)}(u) = \frac{T_{2n+3}(u)}{2^{2n+2}} = \sum_{k=0}^{n+1} s_{2k+1} u^{2k+1},
\]
where \(T_{2n+3}(u)\) is the Chebyshev polynomial of degree \(2n+3\).

Define
\[
\gamma_k = -s_{2k+1} |x^{2(n-k)}| \prod_{l=k}^{n+1} \frac{a_l}{b_l} \text{ for } k=0,n.
\]

Then
\[
R_n^* = x \frac{A_n^*}{B_n^*} = \frac{A_n + \gamma_0 + x^2 \sum_{k=2}^{n} \gamma_k A_{k-2}}{B_n + x^2 \gamma_1 + x^2 \sum_{k=2}^{n} \gamma_k B_{k-2}}
\]
and
\[
\lim_{\epsilon \to 0} \frac{R_n^*(\epsilon u) - f(\epsilon u)}{2^{2n+3}} = C_{n+1} S^{(2n+3)}(u).
\]

**D. RANGE REDUCTION**

Use of a single approximation to a function \(f(x)\) over its entire range of definition is usually not feasible. Therefore, it is usual to subdivide the \(x\) range into intervals small enough to provide the desired accuracy with algorithms of reasonable length, but not into so many subintervals that the mechanism for deciding into which interval \(x\) falls and for acting accordingly becomes cumbersome and time consuming. Depending upon the characteristics of the computer, some balance must be struck between the number of intervals of subdivision and the number of terms (operations) in the algorithm.
IV. Approximations Obtained for Library Functions

The library functions considered are square root, cube root, sin u, tan u, arctan u, arcsin u, e^u and log u. Other functions may be computed in terms of these. Numerical coefficients are not given for all of the approximations tested, but those not given may be found in SSD Memos, Refs. 9 through 12.

A. SQUARE ROOT

(See References 5, 14, 15 and 16)

1. Reduction of Range

To find \sqrt{N}, N>0, first reduce N to the form \[ N=2^{m} \cdot x \] where \[ \frac{1}{4} \leq x < 1 \] (or \[ \frac{1}{4} < x \leq 1 \]) and m is zero or a positive or negative integer. If this representation is to be unique, only one of the two endpoints \[ \frac{1}{4}, 1 \] should be included in the range of x. For example \[ 16=2^4 \cdot 1=2^0 \left( \frac{1}{4} \right) \]. Hence \[ \sqrt{N}=2^{m} \cdot \sqrt{x} \].

2. Computation of \sqrt{x} for

\[ \frac{1}{4} \leq x < 1 \text{ or } \frac{1}{4} < x \leq 1 \]

\sqrt{x} is computed via a Newton-Raphson iteration starting with a first guess, \( y_1 \), for \( \sqrt{x} \). Successive approximations are found from

\[ y_{i+1} = \frac{1}{2} \left( y_i + \frac{x}{y_i} \right), i=1,2,\ldots \]

iterating until \( |y_{i+1} - y_i| < 2^{-19} \). The number of iterations required will depend upon the accuracy of the first guess, \( y_i \). For the various estimates for \( y_1 \) which follow, maximum and minimum number of iterations are given for values of N ranging from .1 to 10 at intervals of .1.

3. Estimates for \( y_1 \)

a. Best-fit rational derived by Machly (Ref. 14) for \( y_1 \) with max. relative error \(< 2.6(10^{-5})\).

\[ y_1 = a + \frac{b}{c+x} \]

\[ a = 3.090315520/\sqrt{2} \]

\[ b = -8.550050013/2/\sqrt{2} \]

\[ c = 3.090315520/2 \]

Max. number of iterations = 4

Min. number of iterations = 2

b. Padé approximation in continued fraction form (Ref. 5) with max. relative error \(< 2.3(10^{-4})\) for \( y_1 \).

\[ y_1 = \frac{25}{7} - \left[ \frac{5000}{343} \left( \frac{x + 15}{49} \right) \right. \]

\[ \left. \left( \frac{x + 235}{49} \right) \left( \frac{x + 15}{49} \right) - \frac{400}{2401} \right] \]

Max. number of iterations = 3

Min. number of iterations = 2

c. Padé approximation in continued fraction form with split range (Ref. 5) and max. relative error \(< 10^{-5}\) in \( y_1 \).

For \[ \frac{1}{4} \leq x \leq \frac{1}{2} \]

\[ y_1 = \frac{5 \sqrt{70}}{14} - \frac{500 \sqrt{70}}{49} \left( \frac{x + 3}{14} \right) \]

\[ \left( \frac{x + 47}{14} \right) \left( \frac{x + 3}{14} \right) - \frac{4}{49} \]

For \[ \frac{1}{2} \leq x < 1 \]

\[ y_1 = \frac{5 \sqrt{35}}{7} - \frac{200 \sqrt{35}}{49} \left( \frac{x + 3}{7} \right) \]

\[ \left( \frac{x + 47}{7} \right) \left( \frac{x + 3}{7} \right) - \frac{16}{49} \]

Max. number of iterations = 3

Min. number of iterations = 1

---

*See References 1, 2, 5, 13, 14, 15, 16 and 20.
The approximation c. gave no better results than b. and has the additional disadvantage of using a split range. Method a. needs one more iteration than b. so that a. takes the time-equivalent of 3 iterations +2D+1M+3A=3 iterations +20A, while b. takes 3 iterations +1D+2M+4A=3 iterations +16.25A, assuming all fractions in a. and b. are precomputed. Hence b. is slightly faster, but requires 3 more constants than a. Either a. or b. is preferable to c.

B. CUBE ROOT (References 16 and 20.d.)

1. Reduction of Range

To find \( \sqrt[3]{N} \) for \( N > 0 \), write \( N = 2^{2n-k} \cdot x \), where \( \frac{1}{2} \leq x < 1 \) and where \( k \) and \( n \) are either zero or integers with the same sign; \( k \) is restricted to the values \( k = 0, \pm 1, \pm 2 \). Then \( \sqrt[3]{N} = 2^{n/3} \sqrt[3]{x} \).

2. Computation of \( \sqrt[3]{x} \) for \( \frac{1}{2} \leq x < 1 \).

Find \( \sqrt[3]{x} \) by means of a Newton-Raphson iteration starting with a first approximation, \( y_1 \), for \( \sqrt[3]{x} \). Successive approximants are found from

\[
y_{i+1} = \frac{2}{3} \left( y_i + \frac{x}{2y_i^2} \right), \quad i = 1, 2, \ldots
\]

iterating until

\[ |y_{i+1} - y_i| < 2^{-49}. \]

The maximum and minimum number of iterations for arguments ranging from .1 to 10 at increments of .1 are given for each of the estimates for \( y_1 \) which follow.

3. Estimates for \( y_1 \)

a. Linear approximation for \( y_1 \) has absolute error of about \( 8(10^{-2}) \) for \( x = 1 \) (Reference 16).

\[
y_1 = A + Bx
\]

\[
A = .5914052048
\]

\[
B = .3319149488
\]

Maximum number of iterations = 5

Minimum number of iterations = 5

b. Rational approximation for \( y_1 \) with absolute error of about \( 9(10^{-5}) \) at \( x = 1 \) (Reference 20.d).

\[
y_1 = a_0 - \frac{a_1}{x + b_1}
\]

\[
a_0 = 1.78781
\]

\[
a_1 = 1.91548
\]

\[
b_1 = 1.42856
\]

Maximum number of iterations = 4

Minimum number of iterations = 3

c. Continued fraction approximations for \( y_1 \) with absolute error of about \( 9(10^{-5}) \) at \( x = 1 \) (Reference 20.d).

\[
y_1 = a_0 - \frac{a_3}{(x + b_1) - (x + b_2)}
\]

\[
a_0 = 2.502926
\]

\[
a_1 = 8.045125
\]

\[
b_1 = 4.612244
\]

\[
a_2 = .3598496
\]

\[
b_2 = .3877552
\]

Maximum number of iterations = 3

Minimum number of iterations = 3

Method a. takes 2 iterations more than c., while b. takes only one more. Hence, in addition to the three iterations required by all three methods, the time-equivalent of

\[
7M + 2D + 3A = 35A \quad \text{is needed for a.,}
\]

\[
3M + 2D + 3A = 25A \quad \text{for b., and}
\]

\[
2M + 1D + 4A = 16.25A \quad \text{for c.}
\]

Thus, even if no operations could be done in parallel, c. is more efficient than b., which is more
efficient than a. Of course, c. would be even more effective if parallel operations are done.

C. \text{SIN } u

1. Range Reduction

Assume first that the argument, \( u \), has been reduced to the range \( -\frac{\pi}{2} \leq u \leq \frac{\pi}{2} \). Then two ranges are considered,

\[
| x | \leq \frac{\pi}{2} \text{ and } | x | \leq \frac{\pi}{6}
\]

a. \( | x | \leq \frac{\pi}{2} \)

For \( | u | \leq \frac{\pi}{2} \), let \( x = u \) and \( \sin u = \sin x \)

b. \( | x | \leq \frac{\pi}{6} \)

For \( | u | \leq \frac{\pi}{6} \), let \( x = u \) and \( \sin u = \sin x \)

For \( | u | > \frac{\pi}{6} \),

let \( x = \frac{u}{3} \) and \( \sin u = \sin x \) \( (3 - 4 \sin^2 x) \).

2. Taylor-Maclaurin Series (See III.A.1.)

\[
\sin x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!}
\]

For \( | x | \leq \frac{\pi}{2} \) the terms through \( n = 9 \) provide an approximation with absolute and relative errors \(<2^{-49}\), while for \( | x | \leq \frac{\pi}{6} \), \( n = 6 \) is sufficient. These two polynomials form the basis for much better approximations to be developed in the next sections.

R.E. \( \leq \frac{x^{2n+1}}{(2n+1)! \sin x} \leq \frac{\left(\frac{\pi}{2}\right)^{2n+1}}{(2n+1)!} \) for \( | x | \leq \frac{\pi}{2} \).

R.E. \( \leq \frac{\left(\frac{\pi}{6}\right)^{2n+1}}{(2n+1)!} \) for \( | x | \leq \frac{\pi}{6} \).

so that for

\[
n = 10, \quad \frac{\left(\frac{\pi}{2}\right)^{21}}{211} \sim 2.6(10^{-15}) < 2^{-49}
\]

and for

\[
n = 7, \quad \frac{\left(\frac{\pi}{6}\right)^{15}}{151} \sim 10^{-16.33} < 2^{-49}.
\]

3. Telescopred Polynomials (See III.A.3.)

a. \( | x | \leq \frac{\pi}{2} \)

\[
\sin x \sim \sum_{n=0}^{9} \frac{(-1)^n x^{2n+1}}{(2n+1)!}
\]

with error \(<2.6(10^{-16})\).

Let \( x = \frac{\pi}{2} y \), \( | y | \leq 1 \).

Then

\[
\sin x \sim \sum_{n=0}^{9} \frac{(-1)^n \left(\frac{\pi}{2}\right)^{2n+1}}{(2n+1)!} y^{2n+1}
\]

\[
= \sum_{n=0}^{8} \frac{(-1)^n \left(\frac{\pi}{2}\right)^{2n+1}}{(2n+1)!} y^{2n+1} - \frac{\pi}{19} \quad y^{19}
\]

Substitute for \( y^{19} \) in terms of the Chebyshev polynomial \( T_{19}(y) \),

\[
y^{19} = \left[ \frac{19}{4} y^{17} - \frac{19}{2} y^{15} + \ldots + \frac{19}{2^{11}} y + \frac{T_{19}(y)}{2^{18}} \right]
\]

\[
= \sum_{k=0}^{8} C_{19}^{(2k+1)} y^{2k+1} + \frac{T_{19}(y)}{2^{18}} .
\]

Now

\[
\frac{\left(\frac{\pi}{2}\right)^{19}}{19!} \frac{T_{19}(y)}{2^{18}} \leq \frac{\left(\frac{\pi}{2}\right)^{19}}{19!2^{18}} \sim 1.66(10^{-19})
\]
and this error added to the original error of 2.6(10^{-10}) is still <2^{-69}, so that the term in T_{16}(y) may be dropped. The result is

\[
\sin x \sim \sum_{n=0}^\infty \left[ (-1)^n \left( \frac{\pi}{2} \right) \frac{2n-1}{(2n+1)!} - \frac{\pi}{19!} C_{19}^{(19)} \right] y^{2n+1}
\]

where \( C_{2n+1}^{(19)} \) is the coefficient of \( y^{2n+1} \) in the Chebyshev polynomial of degree 19. Similarly, after substitution for \( y^{19} \) in terms of \( T_{19}(y) \), it is found that the coefficient of \( T_{19}(y) \) is about 9(10^{-17}) with the total error still less than 2^{-69}, so that the term containing \( T_{19}(y) \) can be dropped. A further substitution for \( y^{19} \) results in a coefficient for \( T_{19}(y) \) of about 7(10^{-19}) which is too large to be dropped. Thus we must be satisfied with substitutions for \( y^{19} \) and \( y^{17} \) only, reducing the degree of the polynomial in \( y \) from 19 to 15.

Now, transforming back to the variable \( x = 2x/\pi \), the final result is

\[
\sin x = \sum_{n=0}^\infty C_{2n+1}^{(19)} x^{2n+1} = x \sum_{n=0}^\infty C_{2n+1}^{(19)} z^n \quad \text{where} \quad z = x^2
\]

and

\[
B = \left( \frac{\pi}{2} \right)^2
\]

\[
A = \left( 1 - \frac{B}{72} \right) / 16!
\]

\[
C_0 = \frac{1}{9} - \frac{143B^5}{1812^{10}} - \frac{55B^4A}{2^4} = 2.75573 15528 52388 14908 \quad E-06
\]

\[
C_{11} = -\frac{1}{11!} + \frac{91B^4}{1812^{10}} + \frac{13B^4A}{2^4} = -2.50518 24652 10732 20541 \quad E-08
\]

\[
C_{13} = \frac{1}{13!} - \frac{35B^3}{1812^{10}} - \frac{7B^3A}{2^4} = 1.60466 21504 47864 08126 \quad E-10
\]

\[
C_{15} = -\frac{1}{15!} + \frac{B^2}{1812} + \frac{BA}{2^4} = -7.35769 03984 39792 89177 \quad E-13
\]

\[
|x| \leq \frac{\pi}{6}
\]

In like manner, the polynomial approximation

\[
\sin x \sim \sum_{n=0}^\infty (-1)^n x^{2n+1} / (2n+1)!
\]

can be telescoped to \( \sin x \sim x \sum_{n=0}^\infty d_{n+1} x^n \),

where \( z = x^2 \) and

\[
B = \left( \frac{\pi}{6} \right)^2
\]

\[
d_0 = 1 - \frac{B}{12!2^{12}} = 9.9999 99999 99999 783585 \quad E-01
\]

\[
d_3 = -\frac{1}{3!} + \frac{7B^3}{12!2^{10}} = -1.6666 66666 66477 96113 \quad E-01
\]

\[
d_5 = \frac{1}{5!} - \frac{7B^4}{12!2^8} = 8.3333 33333 26814 74112 \quad E-03
\]

\[
d_7 = -\frac{1}{7!} + \frac{3B^5}{12!2^6} = -1.98412 69813 94935 49426 \quad E-04
\]

\[
d_9 = -\frac{1}{9!} + \frac{3B^6}{12!2^4} = -1.98412 26903 46738 823484 \quad E-04
\]
\[ d_0 = \frac{1}{9!} - \frac{5B^2}{12!2^4} \]
\[ = 2.75568288724422163687 \quad \text{E-06} \]
\[ d_{11} = -\frac{1}{11!} + \frac{B}{12!2^4} \]
\[ = -2.49090213488806521004 \quad \text{E-08} \]

The remaining algorithms for \( \sin x \) are all for the range

\[ |x| \leq \frac{\pi}{6}. \]

4. **Padé Rational Approximations for \( \sin x \)**

\[ |x| \leq \frac{\pi}{6} \text{ (See III.B.1.)} \]

Write

\[ \frac{\sin x}{x} = \sum_{i=0}^{5} C_i z^i \text{ where } z = x^2 \text{ and } C_i = \frac{(-1)^i}{(2i+1)!}. \]

From IV.C.2, it is known that this polynomial gives the desired accuracy for \( |x| \leq \frac{\pi}{6} \). Thus, taking \( n=m=3 \) in the formulas of III.B.1.,

\[ \sum_{i=0}^{1} C_{k+1} b_i = 0 \quad \text{for } k=4, 5, 6 \]
\[ \sum_{i=0}^{1} C_{k-1} b_i = a_k \quad \text{for } k=0, 1, 2, 3 \]

and \( b_6 = 1 \).

there result the equations

\[ C_1 b_3 + C_3 b_2 + C_5 b_1 = -C_4 \]
\[ C_3 b_3 + C_5 b_2 + C_7 b_1 = -C_5 \]
\[ C_5 b_3 + C_7 b_2 + C_9 b_1 = -C_6 \]

for \( b_3, b_2 \) and \( b_1 \),

and then the equations

\[ a_0 = C_0 \]
\[ a_1 = C_1 + C_0 b_1 \]
\[ a_2 = C_2 + C_1 b_1 + C_0 b_2 \]
\[ a_3 = C_3 + C_2 b_1 + C_1 b_2 + C_0 b_3 \]

for \( a_0, a_1, a_2 \) and \( a_3 \). The solution is

\[ a_0 = 1 \]
\[ a_1 = -325,523/2,283,996 \]
\[ a_2 = 34,911/7,613,320 \]
\[ a_3 = -479,249/11,511,339,840 \]

\[ b_0 = 1 \]
\[ b_1 = 18,381/761,332 \]
\[ b_2 = 1,261/4,567,992 \]
\[ b_3 = 2,623/1,644,477,120 \]

so that

\[ \sin x = \sum_{k=0}^{5} \frac{a_k + a_{k+1} z + a_{k+2} z^2 + a_{k+3} z^3}{b_k + b_{k+1} z + b_{k+2} z^2 + b_{k+3} z^3}, \quad z = x^2. \]

Using COF 3 (Appendix A.2,a.) to find the coefficients \( C_0, C_1, C_2, C_3 \) and \( B_1, B_2 \) and \( B_3 \), this rational form is now converted to continued fraction form

\[ \sin x = \sum \frac{C_0 + C_1 (z+B_1) + C_2 (z+B_2) + C_3 (z+B_3)}{(z+B_1) + (z+B_2) + (z+B_3)} \]

\[ = \sum \frac{p}{q} \]

16
and evaluated from bottom to top as indicated in Appendix A.2.b.

\[
C_0 = -2.6101 46506 18158 052394 \quad E01
\]

\[
C_1 = 7.3922 21532 39111 816487 \quad E03
\]

\[
C_2 = 7.3906 24698 46447 182801 \quad E03
\]

\[
C_3 = 2.4085 74343 97122 268316 \quad E03
\]

\[
B_1 = 1.3171 07143 04415 748255 \quad E02
\]

\[
B_2 = -2.5069 20716 84564 146366 \quad E00
\]

\[
B_3 = 4.3867 21136 85869 652633 \quad E01
\]

Since \( \sin x < x \) and \( \frac{\sin x}{x} = C_0 + \frac{P_z}{Q_z} \), the right-hand side must yield a number less than one. In fact, when \( |x| \leq \frac{\pi}{6} \), the right-hand side should be between .955 and 1. To obtain such a number by adding \( \frac{P_z}{Q_z} \) to \( C_0 = -26. \ldots \) could result in the loss of at least one significant digit. Possible remedies are to go back to the original rational form, or to introduce a parameter \( \xi \) and an extra term \( \xi x \) into the computation (Ref. 5, p. 25) by writing

\[
\frac{\sin x}{x} + \xi x = \sum_{i=0}^{n} C_i z^i,
\]

where

\[
\overline{C_i} = \frac{(-1)^i}{(2i+1)!} \quad \text{for } i = 0, 2, 3, 4, 5, 6
\]

and

\[
\overline{C_i} = \xi - \frac{1}{3i}.
\]

Solving the same system of equations as before, but replacing \( C_i \) by \( \overline{C_i} \), the result in terms of \( \xi \) is

\[
b_1 = \frac{2623}{1.235,520(1331-3990\xi)}
\]

\[
b_2 = \frac{23}{326,040} + \frac{2448b_3}{19}
\]

\[
b_1 = \frac{1}{110} + 72b_2 - 3024b_3
\]

\[
b_0 = 1
\]

\[
a_0 = 1
\]

\[
a_1 = \left( \xi - \frac{1}{6} \right) + b_1
\]

\[
a_2 = \frac{1}{120} + \left( \xi - \frac{1}{6} \right) b_1 + b_2
\]

\[
a_3 = \frac{-1}{5040} + \frac{b_1}{120} + \left( \xi - \frac{1}{6} \right) b_2 + b_3
\]

After finding \( a_i \) and \( b_i \) for a particular value of \( \xi \), the corresponding values \( C_i \) and \( B_i \) must be recomputed from COF 3. Then

\[
\sin x = x \left[ C_0 - \xi x + \frac{P_z}{Q_z} \right].
\]

Values of \( \xi \) tried were \( \xi = .078, .0785, 2.532 \) and \( 2.533 \), and the corresponding values of \( C_0 \) were about \( .0907, .2534, .3399 \) and \( .0145 \) respectively.

\( \xi = 0 \) is, of course, the case originally computed.
5. Comparison of Results

1604 runs in double-precision using arguments from $u=0^\circ$ through $u=90^\circ$ at $1^\circ$ intervals yielded the following results:

<table>
<thead>
<tr>
<th>Method</th>
<th>Maximum Absolute Error</th>
<th>Maximum Relative Error</th>
<th>Max. Error of $N$ in the $k$th Significant Digit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Taylor, $</td>
<td>x</td>
<td>\leq \frac{\pi}{2}$</td>
<td>2.56($10^{-16}$) at 90$^\circ$</td>
</tr>
<tr>
<td>3a. Cheby, $</td>
<td>x</td>
<td>\leq \frac{\pi}{2}$</td>
<td>3.45($10^{-16}$) at 90$^\circ$</td>
</tr>
<tr>
<td>3b. Cheby, $</td>
<td>x</td>
<td>\leq \frac{\pi}{6}$</td>
<td>3.78($10^{-17}$) at 30$^\circ$</td>
</tr>
<tr>
<td>4. Padé Rational, $</td>
<td>x</td>
<td>\leq \frac{\pi}{6}$</td>
<td>3.57($10^{-17}$) at 30$^\circ$</td>
</tr>
<tr>
<td>Cont’d Fraction, $\xi=0$</td>
<td>3.57($10^{-17}$) at 30$^\circ$</td>
<td>7.13($10^{-17}$) at 30$^\circ$</td>
<td>3.57 in 17th at 30$^\circ$</td>
</tr>
<tr>
<td>” ” $\xi=.078$</td>
<td>5.12($10^{-17}$) at 30$^\circ$</td>
<td>1.02($10^{-16}$) at 30$^\circ$</td>
<td>5.12 in 17th at 30$^\circ$</td>
</tr>
<tr>
<td>” ” $\xi=.0785$</td>
<td>5.14($10^{-17}$) at 30$^\circ$</td>
<td>1.03($10^{-16}$) at 30$^\circ$</td>
<td>5.14 in 17th at 30$^\circ$</td>
</tr>
<tr>
<td>” ” $\xi=2.532$</td>
<td>2.3304($10^{-17}$) at 30$^\circ$</td>
<td>4.661($10^{-17}$) at 30$^\circ$</td>
<td>2.33 in 17th at 30$^\circ$</td>
</tr>
<tr>
<td>” ” $\xi=2.533$</td>
<td>2.3301($10^{-17}$) at 30$^\circ$</td>
<td>4.660($10^{-17}$) at 30$^\circ$</td>
<td>2.33 in 17th at 30$^\circ$</td>
</tr>
</tbody>
</table>

In these double precision runs, results for a rational function and its corresponding continued fraction were nearly identical.

D. TAN $u$

1. Range Reduction

Two ranges are considered, $|x| \leq \frac{\pi}{8}$ and $|x| \leq \frac{\pi}{4}$, assuming first that the argument $u$ has already been reduced to the interval $|u| < \frac{\pi}{2}$.

a. $|x| \leq \frac{\pi}{8}$

For each $u$ find $k$ and $x$ so that $u = k\left(\frac{\pi}{8}\right) + x$, where $k = 0, 1, 2, 3$ and $0 \leq x \leq \frac{\pi}{8}$.

Let $B = \tan k\left(\frac{\pi}{8}\right)$, i.e.,

$$\tan\left(\frac{\pi}{8}\right) = \sqrt{2} - 1$$

b. $|x| \leq \frac{\pi}{4}$

For

$$|u| \leq \frac{\pi}{4}$$, let $x = u$ and $\tan u = \tan x$.

For

$$|u| > \frac{\pi}{4}$$, let $x = \frac{\pi}{2} - u$ and $\tan u = \frac{1}{\tan x}$.
2. Rational and Continued Fraction Forms Obtained from the Gaussian Continued Fraction (Appendix B.1)

a. For \(|x| \leq \frac{\pi}{8}\)

The appropriate approximant from Appendix B.1 is

\[
\tan x = x \left( \frac{A_0}{B_0} \right).
\]

The corresponding continued fraction form obtained by using COF 3 of Appendix A.2.a. to find \(C_i\) and \(B_j\) is

\[
\tan x = x \left[ \frac{C_0}{(z+B_1)} + \frac{C_1}{(z+B_2)} + \frac{C_2}{(z+B_3)} \right]
\]

\[
= x \left[ C_0 + \frac{P_2}{Q_2} \right], \quad z = x^2
\]

which is evaluated via Appendix A.2.b.

Thus, by converting the original 6 or 7 level Gaussian continued fraction to rational form, and then back to another continued fraction form using COF 3, it has been reduced to one of three levels.

b. For \(|x| \leq \frac{\pi}{4}\)

Both \(\tan x = x \frac{A_7}{B_7}\) and \(\tan x = x \frac{A_8}{B_8}\)

were tested in rational form. Since the former proved accurate enough, only that continued fraction form was tried. Coefficients were found from Appendix A.3.a., and the continued fraction was evaluated as indicated in Appendix A.3.b.

3. Telescoped Rational and Continued Fraction Forms for \(|x| \leq \frac{\pi}{4}\), (See III.C.7.)

From the preceding Section IV.D.2.b., we have the rational approximation

\[
\tan x = x \left[ \frac{A_7}{B_7} \right]
\]

\[
= x \left[ \frac{a_1 + a_2 z + a_3 z^2 + a_4 z^3}{b_0 + b_1 z + b_2 z^2 + b_3 z^3 + b_4 z^4} \right],
\]

where \(z = x^2\).

Telescoping one step would reduce the degree of the denominator from 4 to 3 and the corresponding continued fraction from a 4 to a 3-level one.

Consequently we shall telescope \(x \frac{A_7}{B_7}\), or correct \(x \frac{A_8}{B_8}\), using the results of III.C.7, for an “odd” function.

Take

\[
n = 6, \quad \epsilon = \frac{\pi}{4}, \quad r_0 = \frac{a_0}{b_0} = 1,
\]

and \(r_i = \frac{a_i}{b_i} = \frac{-1}{(2i+1)}\) for \(i = 1, 7\).

Since

\[
S^{(18)}(u) = \frac{T_{13}(u)}{2^{14}} = u^{15} - \frac{15u^{13}}{4} + \frac{45u^{11}}{8}
\]

\[
- \frac{275u^{9}}{2^6} + \frac{225u^{7}}{2^7} - \frac{189u^{5}}{2^8} + \frac{35u^{3}}{2^9} - \frac{15u}{2^{11}}.
\]

\[
s_1 = -15/2^{14}
\]

\[
s_3 = 35/2^{10}
\]

\[
s_5 = -189/2^{9}
\]

\[
s_7 = 225/2^{7}
\]

\[
s_9 = -275/2^{6}
\]

\[
s_{11} = 45/8
\]

\[
s_{13} = -15/4
\]

and

\[
\gamma_k = -s_{2k+1} \int e^{(r-\beta)} \prod_{i=k}^{r} r_1 \text{ for } k = 0, 6.
\]
Then

\[ R^*_x = x \left[ \frac{A_0 + \gamma_x + x^2(\gamma_2 A_0 + \gamma_4 A_2 + \gamma_6 A_4) + \gamma_6 A_6}{B_0 + \gamma_x x^2 + x^2(\gamma_2 B_0 + \gamma_4 B_2 + \gamma_6 B_4 + \gamma_6 B_6)} \right]. \]

After combining, the resulting \(a_1\) and \(b_1\) are

\[ a_0 = 135,135 + \gamma_x \]
\[ a_1 = -17,325 + (\gamma_2 + 3 \gamma_4 + 15 \gamma_6 + 105 \gamma_8 + 945 \gamma_x) \]
\[ a_2 = 375 - \gamma_4 - 10 \gamma_6 - 105 \gamma_8 \]
\[ a_3 = -1 + \gamma_6 \]
\[ b_0 = 135,135 \]
\[ b_1 = -62,370 + \gamma_x + (\gamma_2 + 3 \gamma_4 + 15 \gamma_6 + 105 \gamma_8 + 945 \gamma_x) \]
\[ b_2 = 3150 - \gamma_4 - 6 \gamma_6 - 45 \gamma_8 - 420 \gamma_x \]
\[ b_3 = -28 + \gamma_6 + 15 \gamma_8 \]

and

\[ \tan x = x \left[ \frac{a_0 + a_1 z + a_2 z^2 + a_3 z^3}{b_0 + b_1 z + b_2 z^2 + b_3 z^3} \right], \]
\[ z = x^2, \quad |x| \leq \frac{\pi}{4}. \]

Again, coefficients and evaluation of the corresponding continued fraction form are obtained using COF 3, Appendix A.2.a.&b.

\[ C_0 = 3.5911 01496 97721 76037 74655 \quad E-02 \]
\[ C_1 = -9.4381 65598 19183 41369 40110 \quad E00 \]
\[ C_2 = -1.4096 32418 00227 61516 62209 \quad E03 \]
\[ C_3 = -1.5692 00421 75952 56069 73336 \quad E02 \]
\[ B_1 = -5.5204 04171 66464 89417 73271 \quad E01 \]
\[ B_2 = -4.0981 70874 59656 10393 42606 \quad E01 \]
\[ B_2 = -1.5783 03284 55044 64639 80047 \quad E01 \]

**4. Comparison of Results**

The following results were obtained from 1604 double-precision runs using arguments from \(u = 0^\circ\) through \(u = 89^\circ\) at \(1^\circ\) intervals.

<table>
<thead>
<tr>
<th>Method</th>
<th>Maximum Absolute Error</th>
<th>Maximum Relative Error</th>
<th>Max. Error of (N) in the (k)th Significant Digit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2a. Rational &amp; Cont'd Fraction, (</td>
<td>x</td>
<td>\leq \frac{\pi}{8})</td>
<td>(4.89 (10^{-15})) at (89^\circ)</td>
</tr>
<tr>
<td>2b. Rational &amp; Cont'd Fraction, (</td>
<td>x</td>
<td>\leq \frac{\pi}{4})</td>
<td>(4.54 (10^{-15})) at (45^\circ)</td>
</tr>
<tr>
<td>3. Telescoped form of 2b(1),</td>
<td>(4.70 (10^{-16})) at (45^\circ)</td>
<td>(4.70 (10^{-16})) at (45^\circ)</td>
<td>(4.70) in 16th at (45^\circ)</td>
</tr>
</tbody>
</table>
E. ARCTAN u

1. Range Reduction

The argument $0 < u < \infty$ is reduced to one of two ranges:

a. $|x| \leq \sqrt{2} - 1$

For $0 \leq u \leq 1$, let $x = \frac{u - (\sqrt{2} - 1)}{1 + u (\sqrt{2} - 1)}$

and arctan $u = \frac{\pi}{8} + \arctan x$.

For $1 < u < \infty$, let $x = \frac{1 - u (\sqrt{2} - 1)}{u + (\sqrt{2} - 1)}$

and arctan $u = \frac{3\pi}{8} - \arctan x$.

b. $|x| \leq \tan \frac{\pi}{16}$

For $0 \leq u \leq 1$, set $y = u$, $A = 0$, $B = 1$.

For $1 < u < \infty$, set $y = \frac{1}{u}$, $A = \frac{\pi}{2}$, $B = -1$.

Then

for $0 \leq y \leq \sqrt{2} - 1$, set $y_1 = \tan \frac{\pi}{16}$, $\alpha = \frac{\pi}{16}$

and

for $\sqrt{2} - 1 < y \leq 1$, set $y_1 = \tan \frac{3\pi}{16}$, $\alpha = \frac{3\pi}{16}$.

Then

$$x = \frac{y - y_1}{1 + y \cdot y_1},$$

and arctan $u = A + B (\alpha + \arctan x)$.

2. Rational and Continued Fraction Forms Obtained from the Gaussian Continued Fraction (Appendix B.2.)

a. $|x| \leq \sqrt{2} - 1$

The appropriate approximant is

$$\arctan x = x \frac{A_{10}}{B_{10}}$$

where $A_{10}/B_{10}$ is the quotient of two fifth-order polynomials in $z$, $z = x^2$. It could be converted to a 5-level continued fraction, but was tested only in its rational form because it seemed less useful than the methods for the range $|x| \leq \tan \frac{\pi}{16}$.

b. $|x| \leq \tan \frac{\pi}{16}$

The approximants $\arctan x = xA_n/B_n$ and $\arctan x = xA_n/B_n$ were tried in both rational and continued fraction form. Coefficients for the continued fraction forms were computed from COF 3 and COF 4 respectively (Appendix A.2 and 3). The approximation $\arctan x = xA_n/B_n$ was not quite accurate enough, hence $A_n/B_n$ was telescoped (or $A_n/B_n$, corrected) as described in the next section to an approximation of the same order as $A_n/B_n$.

3. Telescoped Rational and Continued Fraction Forms for $|x| \leq \tan \frac{\pi}{16}$

Using the formulae in III.C.7. for an odd function and the approximants $A_n/B_n$ in Appendix B.2., and taking $n = 6$, $e = \tan \frac{\pi}{16}$ and $S^{(12)}(u) = T_{13}(u)2^{-11}$, the coefficients in the new $R_n$ are:

$$a_0 = 1.3513 \ 49999 \ 99999 \ 99825 \ 84405 \ E05$$
$$a_1 = 1.7196 \ 2403 \ 9358 \ 3535 \ 64629 \ E05$$
$$a_2 = 5.2490 \ 4316 \ 37362 \ 32796 \ 35437 \ E04$$
$$a_3 = 2.2180 \ 98888 \ 44607 \ 11614 \ 67914 \ E03$$

$$b_0 = 1.3513 \ 50000 \ 00000 \ 00000 \ 00000 \ E05$$
$$b_1 = 2.1700 \ 7403 \ 93585 \ 74205 \ 67287 \ E05$$
$$b_2 = 9.7799 \ 30329 \ 54139 \ 12080 \ 84660 \ E04$$
$$b_3 = 1.0721 \ 37452 \ 05929 \ 68736 \ 47196 \ E04$$

and

$$\arctan x = x \left[ \frac{a_0 + a_1 z + a_2 z^2 + a_3 z^3}{b_0 + b_1 z + b_2 z^2 + b_3 z^3} \right],$$

$$z = x^2, \quad |x| \leq \tan \frac{\pi}{16}.$$
Coefficients of the corresponding continued fraction found from COF 3, Appendix A.2.a. are:

\[
\begin{align*}
C_0 &= 2.0688 \quad 56828 \quad 18530 \quad 47509 \quad 55460 \quad \text{E-01} \\
C_1 &= 3.0086 \quad 82092 \quad 05174 \quad 87448 \quad 28121 \quad \text{E00} \\
C_2 &= -3.4976 \quad 10177 \quad 36154 \quad 25858 \quad 60195 \quad \text{E00} \\
C_3 &= -1.3433 \quad 64284 \quad 54181 \quad 78822 \quad 14637 \quad \text{E-01} \\
B_1 &= 5.1827 \quad 26637 \quad 17441 \quad 95978 \quad 52947 \quad \text{E00} \\
B_2 &= 2.6194 \quad 66421 \quad 36919 \quad 73145 \quad 76466 \quad \text{E00} \\
B_3 &= 1.3197 \quad 06666 \quad 86630 \quad 28901 \quad 33958 \quad \text{E00}
\end{align*}
\]

then

\[
\arctan x = x \left[ C_0 + \frac{P_2}{Q_2} \right] ,
\]

as evaluated from Appendix A.2.b.

4. Comparison of Results

For arguments of the form \( u = \tan y \) with \( y \) ranging from \( y = 1^\circ \) through \( y = 89^\circ \) at intervals of \( 1^\circ \), results of machine runs on the 1604 in double-precision are given in the table following.

<table>
<thead>
<tr>
<th>Method</th>
<th>Maximum Absolute Error</th>
<th>Maximum Relative Error</th>
<th>Max. Error of ( N ) in the ( k )th Significant Digit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2a. Rational &amp; Cont'd Fraction, (</td>
<td>x</td>
<td>\leq \sqrt{2} - 1 )</td>
<td>Arctan ( x = xA_{10}/B_{10} ) ( 2.28 \times 10^{-10} ) at tan ( 45^\circ )</td>
</tr>
<tr>
<td>2b. Rational &amp; Cont'd Fraction, (</td>
<td>x</td>
<td>\leq \tan \frac{\pi}{16} )</td>
<td>1) Arctan ( x = xA_6/B_6 ) ( 2.42 \times 10^{-15} ) at tan ( 45^\circ )</td>
</tr>
<tr>
<td></td>
<td>2) Arctan ( x = xA_7/B_7 ) ( 2.36 \times 10^{-17} ) at tan ( 45^\circ )</td>
<td>( 2.75 \times 10^{-16} ) at tan ( 1^\circ )</td>
<td>4.80 in 17th at tan ( 1^\circ )</td>
</tr>
<tr>
<td></td>
<td>3. Telescoped form of ( 2b(2),</td>
<td>x</td>
<td>\leq \tan \frac{\pi}{16} )</td>
</tr>
</tbody>
</table>

F. ARCSIN \( u \)

1. Range Reductions

No algorithm is practical for the entire range \( |u| \leq 1 \). Arcsin \( u \) can be computed in terms of arctan \( u \), or the range \( |u| \leq 1 \) can be reduced to \( |x| \leq \frac{1}{2} \) and an algorithm for arcsin \( x \) applied in this range. In either case, it is necessary to take a square root; in the latter case, the range can be so adjusted that the square root operation is performed only a small proportion of the time.

a. Arcsin \( u = \arctan x \) where \( x = \frac{u}{\sqrt{1-u^2}} \) requires the square root operation all the time.

b. For

\[
\begin{align*}
0 \leq u \leq \frac{1}{2}, \text{ set } x = u \text{ and arcsin } u &= \arcsin x \\
\frac{1}{2} < u \leq 1, \text{ set } x &= \sqrt{1-u} \\
\text{and arcsin } u &= \frac{\pi}{2} - 2 \arcsin x.
\end{align*}
\]
This reduction makes use of the square root half the time.

c. For

\[
\begin{aligned}
0 & \leq u \leq \frac{1}{2}, \text{ set } x = u, A = 0, B = 1 \\
\frac{1}{2} & < u \leq \frac{\sqrt{3}}{2} \sim .866, \\
\text{set } x &= 2u^2 - 1, A = \frac{\pi}{4}, B = \frac{1}{2} \\
\frac{\sqrt{3}}{2} & < u \leq 1, \\
\text{set } x &= \sqrt{\frac{1-u}{2}}, A = \frac{\pi}{2}, B = -2
\end{aligned}
\]

Then \( \arcsin u = A + B \arcsin x \), so that a square root is used \(.134\) of the time.

d. For

\[
\begin{aligned}
0 & \leq u \leq \frac{1}{2}, \text{ set } x = u, A = 0, B = 1 \\
\frac{1}{2} & < u \leq \frac{\sqrt{3}}{2} \sim .866, \\
\text{set } x &= 2u^2 - 1, A = \frac{\pi}{4}, B = \frac{1}{2} \\
\frac{\sqrt{3}}{2} & < u \leq \frac{1}{2} \sqrt{2 + \sqrt{3}} \sim .965, \\
\text{set } x &= 8u^4 - 8u^2 + 1, A = \frac{3\pi}{8}, B = \frac{1}{4} \\
.965 & < u \leq 1, \\
\text{set } x &= \sqrt{\frac{1-u}{2}}, A = \frac{\pi}{2}, B = -2
\end{aligned}
\]

Then \( \arcsin u = A + B \arcsin x \) and the square root is needed \(.035\) of the time.

All of the range reductions except a. depend upon an algorithm for \( \arcsin x \) where \( |x| \leq \frac{1}{2} \). Only one will be developed.

2. Telescoped Polynomials for \( \arcsin x \),

\[
|x| \leq \frac{1}{2}
\]

\[
\frac{\arcsin x}{x} = 1 + \frac{x^2}{6} + \frac{1 \cdot 3x^4}{2 \cdot 4 \cdot 5} + \frac{1 \cdot 3 \cdot 5x^6}{2 \cdot 4 \cdot 6 \cdot 7} + \ldots + \frac{1 \cdot 3 \cdot \ldots \cdot 39x^{4n}}{2 \cdot 4 \cdot 6 \ldots \cdot 40 \cdot 41}
\]

is the truncated series expansion of \( \frac{\arcsin x}{x} \) with error less than \( 6.47 (10^{-16}) < 2^{-40} \).

Let

\[
z = x^2 = \frac{y}{4} \text{ so that } 0 \leq y \leq 1 \text{ when } 0 \leq x^2 \leq \frac{1}{4}.
\]

Then

\[
\frac{\arcsin x}{x} = C_0 + C_1 y + C_2 y^2 + \ldots + C_{20} y^{20},
\]

where

\[
C_0 = 1 \text{ and } C_n = \frac{(2n-1)^2}{8n(2n+1)} C_{n-1} \text{ for } n = 1, 20.
\]

Using the shifted Chebyshev polynomials (see III.A.2. and 3.) this polynomial in \( y \) of degree 20 can be telescoped to one of degree 11. When the transformation from the variable \( y \) back to variable \( x \) is made, the final coefficients are:

\[
\begin{array}{cccc}
d_0 & = & 9.9999 & 99999 & 99999 & 78634 & 63136 & E-01 \\
d_1 & = & 1.6666 & 66666 & 66910 & 24987 & 33835 & E-01 \\
d_2 & = & 7.4999 & 999995 & 41191 & 18650 & 66303 & E-02 \\
d_3 & = & 4.4642 & 86051 & 93986 & 28433 & 58075 & E-02 \\
d_4 & = & 3.0381 & 81646 & 51631 & 62166 & 80726 & E-02 \\
d_5 & = & 2.2375 & 00912 & 35718 & 53117 & 84246 & E-02 \\
d_6 & = & 1.7312 & 76426 & 25238 & 60583 & 99121 & E-02 \\
d_7 & = & 1.4331 & 24507 & 67065 & 51847 & 69121 & E-02 \\
d_8 & = & 9.3428 & 65551 & 28506 & 27072 & 55181 & E-03 \\
d_9 & = & 1.8356 & 67090 & 64025 & 76498 & 65645 & E-02 \\
d_{10} & = & -1.1862 & 23970 & 78013 & 60943 & 70754 & E-02 \\
d_{11} & = & 3.1627 & 12225 & 71360 & 72001 & 51992 & E-02
\end{array}
\]
so that
\[
\arcsin x = x + \left[ d_0 + d_1 x + d_2 x^2 + \ldots + d_{11} x^{11} \right],
\]
with \( z = x^2 \).

Note that one of the coefficients, \( d_{10} \), changed sign! This would suggest that perhaps the telescoping had proceeded too far. Such was not the case as results in the next section show.

3. Test Results
The telescoped polynomial was used in each of the ranges \( b, c, d \) for values of \( u \) from \( u = 0 \) through \( u = 1 \) at intervals of \( .01 \) with these results—

Max. Absolute Error: \( 5.44(10^{-1}) \) at \( u = .5 \)
Max. Relative Error: \( 1.04(10^{-1}) \) at \( u = .5 \)
Max. Error of \( N \) in the
Kth Significant Digit: \( 5.44 \) in 16th at \( u = .5 \)

G. EXPONENTIAL: \( e^u \)

1. Range Reduction
Write \( e^u = 2^w e^x \) where \( n \) and \( x \) are found as follows:

Let \( y = \frac{u}{\log_2} \) and \( n = \left[ y + \frac{1}{2} \right] \) = the integral part of \( y + \frac{1}{2} \). The minus sign holds if \( u < 0 \), hence \( y < 0 \).

(Alternatively, one could compute \( e^{1n} \) and take the reciprocal when \( u \) is negative.) Let \( w = y - n \) and \( x = w \log_2 \). Then \( e^x \) may be computed from one of the algorithms which follow for \( |x| < \frac{\log_2}{2} \). If \( u = 0 \), \( e^u \) should be set to 1.

2. Taylor-Maclaurin Series

\[
e^x = \sum_{n=0}^{12} \frac{x^n}{n!}
\]

with

truncation error \( \frac{x^{13}}{13!} < \frac{(\log_2)^{13}}{2^{13} 13!} \sim 1.07(10^{-15}) \)

which is less than \( 2^{-19} \) for \( |x| \leq \frac{\log_2}{2} \).

3. Padé Rational

(Diagonal of the Padé table) Ref. 18

\[
e^x \sim \frac{P_n(x)}{P_n(-x)}
\]

where

\[
P_n(x) = \frac{n!}{(2n)!} \left[ \sum_{j=0}^{n} \frac{(2n-j)!}{j!(n-j)!} x^j \right]
\]

\( n = 5 \) and \( n = 6 \) were tested.

4. Rational and Continued Fraction Forms Obtained from Macon’s Even Part of the Gaussian Continued Fraction for \( e^x \)

(Appendix B.3.b.)

\[
e^x = \frac{S+x}{S-x} \text{ where } S = 2 + F
\]

Approximants tested for \( F \) were

\[
F = x^2 \frac{A_3}{B_3}, \quad F = x^2 \frac{A_4}{B_4} \text{ and } F = x^2 \frac{A_5}{B_5}
\]
in both rational and continued fraction form. Coefficients for the continued fraction forms were computed from COF 3, COF 2 and COF 2 respectively (Appendix A). The case \( F = x^2 \frac{A_5}{B_5} \) proved accurate enough.

Thus,

\[
e^x = \frac{S+x}{S-x}
\]

where

\[
S = 2 + z \left[ \frac{2520 + 28z}{15120 + 420z + z^2} \right], \quad z = x^2
\]
or, in continued fraction form,

\[
S = 2 + z \left[ \frac{P_1}{Q_1} \right] \text{ where } P_1 = C_1 (z + B_3)
\]

\[
Q_1 = (z + B_4) (z + B_5) + C_2
\]
and
\[ C_1 = 28, \quad B_1 = 330. \]
\[ C_2 = -14,580, \quad B_2 = 90. \]

Note that in this case the continued fraction form is not much shorter than the rational form.

5. Telescoped Rational and Continued Fraction Forms

Using formulae from III.C.7. for an even function, an unsuccessful attempt was made to telescope
\[
\frac{F}{x^2} = \frac{A_3}{B_3}.
\]

It would have reduced the degree of the denominator from second degree to first degree in \( z \).

6. Comparison of Results

Results of 1604 double precision tests using arguments from \( u = -9.9 \) to \( u = 10.0 \) at intervals of .1 are:

<table>
<thead>
<tr>
<th>Method</th>
<th>Maximum Absolute Error</th>
<th>Maximum Relative Error</th>
<th>Max. Error of ( N ) in the Kth Significant Digit</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Padé ( n=5 )</td>
<td>3.35(10^-12)</td>
<td>8.26(10^-16)</td>
<td>6.53 in 16th</td>
</tr>
<tr>
<td>Padé ( n=6 )</td>
<td>5.12(10^-16)</td>
<td>1.72(10^-19)</td>
<td>1.33 in 19th</td>
</tr>
<tr>
<td>4. Rational and Cont’d Fraction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a) ( F=x^2A_3/B_4 )</td>
<td>5.76(10^-20)</td>
<td>2.63(10^-23)</td>
<td>1.95 in 23rd</td>
</tr>
<tr>
<td>b) ( F=x^2A_3/B_4 )</td>
<td>5.12(10^-16)</td>
<td>1.72(10^-19)</td>
<td>1.33 in 19th</td>
</tr>
<tr>
<td>c) ( F=x^2A_3/B_4 )</td>
<td>3.35(10^-12)</td>
<td>8.26(10^-16)</td>
<td>6.53 in 16th</td>
</tr>
<tr>
<td>5. Telescoped form of 4.c.</td>
<td>3.33(10^-8)</td>
<td>3.07(10^-12)</td>
<td>2.63 in 12th</td>
</tr>
</tbody>
</table>

at \( u=10 \).  at \( u=\pm 5.2 \)  at \( u=4.5 \)

H. LOGARITHM: \( \log u = \ln u \)

1. Reduction of Range

Write \( u=2^m \cdot n \) where \( \frac{1}{2} \leq m < 1 \), and \( n \) may be zero or a positive or negative integer.

Let
\[
x = \frac{m - \sqrt{2}/2}{m + \sqrt{2}/2}
\]
and compute \( \ln \left( \frac{1+x}{1-x} \right) \)
from the algorithm.

Then
\[
\ln u = \left(n - \frac{1}{2}\right) \ln 2 + \ln \left( \frac{1+x}{1-x} \right).
\]

For \( u=1 \), \( \ln u \) should be set to zero.

The approximations to follow are for \( \ln \left( \frac{1+x}{1-x} \right) \)
where \( |x| < 3-2\sqrt{2} \).

2. Taylor-Maclaurin Series

\[
\log \frac{1+x}{1-x} = 2x \left[ 1 + \frac{x^2}{3} + \frac{x^4}{5} + \frac{x^6}{7} + \ldots + \frac{x^{20}}{21} \right]
\]
with error
\[
< \frac{2x^{23}}{23} < \frac{2^{24} 10^{-23}}{23} \sim 7.3(10^{-16})
\]
for
\[ |x| < 3 - 2\sqrt{2}. \]

Actually one more term could be dropped.

3. Telescoped Polynomial

Let \( z = x^2 = a^2y \) where \( a = 3 - 2\sqrt{2} \) and \( 0 \leq y \leq 1 \) in the truncated Taylor series preceding.

\[
\log \frac{1+x}{1-x} = 2x \left[ C_0 + C_1 z + C_2 z^2 + \ldots + C_n z^n \right],
\]

\( z = x^2 \)

\[
C_0 = 1.0000 \quad 00000 \quad 00000 \quad 01720 \quad 16224 \quad E00
\]
\[
C_1 = 3.3333 \quad 33333 \quad 32761 \quad 81768 \quad 85283 \quad E01
\]
\[
C_2 = 2.0000 \quad 00003 \quad 09807 \quad 78908 \quad 99307 \quad E01
\]
\[
C_3 = 1.4285 \quad 70799 \quad 46082 \quad 73472 \quad 61398 \quad E01
\]
\[
C_4 = 1.1111 \quad 71831 \quad 83715 \quad 43428 \quad 06719 \quad E01
\]
\[
C_5 = 9.0809 \quad 35658 \quad 17935 \quad 37172 \quad 14254 \quad E02
\]
\[
C_6 = 8.4191 \quad 86575 \quad 86305 \quad 31375 \quad 34817 \quad E02
\]

4. Rational and Continued Fraction Forms Obtained from the Gaussian Continued Fraction (Appendix B.4.)

Approximations considered were

a. \( \log \frac{1+x}{1-x} = \frac{A_0}{B_0} \)

b. \( \log \frac{1+x}{1-x} = \frac{A_1}{B_1} \)

c. \( \log \frac{1+x}{1-x} = \frac{A_3}{B_3} \)

in both rational and continued fraction form using COF 3, COF 4 and COF 4 respectively (Appendix A) for computing coefficients of the continued fraction forms.

Coefficients for the continued fraction form of a.

\[
C_0 = \quad 4.1795 \quad 91836 \quad 73499 \quad 38775 \quad 51020 \quad E01
\]
\[
C_1 = -5.9412 \quad 24489 \quad 79591 \quad 83673 \quad 46939 \quad E00
\]
\[
C_2 = -3.3502 \quad 52481 \quad 31135 \quad 23355 \quad 48517 \quad E00
\]
\[
C_3 = -1.2872 \quad 09952 \quad 96610 \quad 95132 \quad 66527 \quad E01
\]
\[
B_0 = -5.1029 \quad 95328 \quad 38691 \quad 94833 \quad 74554 \quad E00
\]
\[
B_2 = -2.5841 \quad 78755 \quad 04759 \quad 66008 \quad 33351 \quad E00
\]
\[
B_3 = -1.3128 \quad 25916 \quad 56548 \quad 39157 \quad 92095 \quad E00
\]

5. Telescoped Rational and Continued Fraction Forms

The approximation 4,a., \( \log \frac{1+x}{1-x} = \frac{A_3}{B_3} \) was telescoped and provided rational and continued fraction forms still of sufficient accuracy. The formulas of III.C.7 for an odd function were used with \( n = 5 \) and \( x = 3 - 2\sqrt{2} \).

\[
\log \frac{1+x}{1-x} = x \left[ \frac{a_0 + a_1 z + a_2 z^2}{b_0 + b_1 z + b_2 z^2 + b_3 z^3} \right], \quad z = x^2
\]

\[
a_0 = \quad 2.0789 \quad 99999 \quad 99999 \quad 84154 \quad 93231 \quad E04
\]
\[
a_1 = -2.1545 \quad 27006 \quad 88655 \quad 98004 \quad 53920 \quad E04
\]
\[
a_2 = \quad 4.2329 \quad 18706 \quad 18926 \quad 27409 \quad 32222 \quad E03
\]
\[
b_0 = \quad 1.0395 \quad 00000 \quad 00000 \quad 00000 \quad 00000 \quad E04
\]
\[
b_1 = -1.4237 \quad 63503 \quad 44403 \quad 34724 \quad 39577 \quad E04
\]
\[ b_2 = 4.7788 \ 37699 \ 95350 \ 61419 \ 58903 \ \text{E}03 \]
\[ b_3 = -2.3041 \ 91303 \ 93980 \ 93764 \ 71785 \ \text{E}02 \]

From COF 3 (Appendix A.2.), the continued fraction form is

\[
\log \frac{1+x}{1-x} = x \left( \frac{P_z}{Q_z} \right)
\]

6. Comparison of Results

Letting \( u \) range from \( u = .1 \) through \( u = 10 \), at intervals of .1, largest errors observed in 1604 double-precision runs are recorded below:

<table>
<thead>
<tr>
<th>Method</th>
<th>Maximum Absolute Error</th>
<th>Maximum Relative Error</th>
<th>Max. Error of ( N ) in the ( K )th Significant Digit</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Chebyshev teles. polynomial</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. degree 7 in ( z )</td>
<td>( 2.95 \times 10^{-18} )</td>
<td>( 4.25 \times 10^{-18} )</td>
<td>2.95 in 18th</td>
</tr>
<tr>
<td>b. degree 6 in ( z )</td>
<td>( 6.24 \times 10^{-17} )</td>
<td>( 9.00 \times 10^{-17} )</td>
<td>6.24 in 17th</td>
</tr>
<tr>
<td>4. Rational &amp; Cont'd Fractions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. ( xA_n/B_n )</td>
<td>( 6.81 \times 10^{-16} )</td>
<td>( 9.88 \times 10^{-16} )</td>
<td>6.81 in 16th</td>
</tr>
<tr>
<td>b. ( xA_n/B_n )</td>
<td>( 5.11 \times 10^{-18} )</td>
<td>( 7.57 \times 10^{-18} )</td>
<td>5.11 in 18th</td>
</tr>
<tr>
<td>c. ( xA_n/B_n )</td>
<td>( 3.83 \times 10^{-26} )</td>
<td>( 5.52 \times 10^{-26} )</td>
<td>3.83 in 20th</td>
</tr>
<tr>
<td>5. Telescoped form of 4a.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( 7.12 \times 10^{-16} )</td>
<td>( 1.03 \times 10^{-15} )</td>
<td>7.12 in 16th</td>
</tr>
</tbody>
</table>

\( u = .5, 2, 4, 8 \)
\( u = .5, 2 \)
\( u = .5, 2 \)
V. References

BOOKS

MIMEOGRAPHED NOTES

INTERNAL MEMOS
System Sciences Division, Control Data Corporation
11. J. Westlake, "Algorithms for Log_x, Cube Root, Aresin x and Improved Algorithms for e^x, Tan x, Arctan x and Sin x", 4/16/64.
12. J. Westlake, "Telescoped Continued Fractions for e^x and log_x", 4/29/64.

CONTROL DATA PUBLICATIONS
PAPERS FROM TECHNICAL JOURNALS


(a) $e^x$ for $-\infty < x < +\infty$, Vol. 1, No. 2, April 1957, pp. 110-115

(b) Arctan $N$ for $-\infty < N < +\infty$, Vol. 2, No. 1, Jan. 1958, pp. 43-53

(c) Arccos $N$ for $0 < N < 1$, Vol. 2, No. 3, July 1958, pp. 218-222

(d) Sin $N$, Cos $N$, and $\sqrt{N}$, Vol. 3, No. 2, April 1959, pp. 147-152


A. Appendix

Formulae for Conversion of a Quotient of Two nth order Polynomials to Continued Fraction Form and for Evaluating the Resulting Continued Fraction

1. COF 2: n = 2
   a. Conversion formulae
      
      \[
      F_z = \frac{a_0 + a_2 z + a_2 z^2}{b_0 + b_1 z + b_2 z^2} = C_0 + \frac{C_1}{(z + B_1) + \frac{C_2}{(z + B_2)} + \frac{C_3}{(z + B_3)}}
      \]

      \[
      C_0 = a_2 b_2
      \]

      \[
      \alpha_0 = a_0 - C_0 b_0
      \]

      \[
      \alpha_1 = a_1 - C_0 b_1
      \]

      \[
      C_1 = \alpha_1 / b_2
      \]

      \[
      B_2 = (C_1 b_1 - \alpha_0) / \alpha_1
      \]

      \[
      C_2 = (C_1 b_0 - \alpha_0 B_1) / \alpha_1
      \]

      \[
      B_3 = \alpha_0 / \alpha_1
      \]

   b. Evaluation
      
      \[
      P_1 = C_1 (z + B_3)
      \]

      \[
      Q_1 = (z + B_3) (z + B_3) + C_3
      \]

      \[
      P_2 = C_1 Q_1
      \]

      \[
      Q_2 = (z + B_1) Q_1 + P_1
      \]

      \[
      F_3 = C_0 + \frac{P_2}{Q_2}
      \]

2. COF 3: n = 3
   a. Conversion formulae
      
      \[
      F_z = \frac{a_0 + a_2 z + a_2 z^2 + a_3 z^3 + a_3 z^4}{b_0 + b_1 z + b_2 z^2 + b_3 z^3 + b_4 z^4} = C_0 + \frac{C_1}{(z + B_1) + \frac{C_2}{(z + B_2) + \frac{C_3}{(z + B_3) + \frac{C_4}{(z + B_4)}}}}
      \]

      \[
      C_0 = a_3 / b_3
      \]

      \[
      \alpha_0 = a_0 - C_0 b_0
      \]

      \[
      \alpha_1 = a_1 - C_0 b_1
      \]

      \[
      \alpha_2 = a_2 - C_0 b_2
      \]

      \[
      \alpha_3 = a_3 - C_0 b_3
      \]

      \[
      C_1 = \alpha_3 / b_4
      \]

      \[
      B_1 = (b_2 C_1 - \alpha_2) / \alpha_3
      \]

      \[
      C_2 = (b_2 C_1 - \alpha_2 - \alpha_3 B_1) / \alpha_3
      \]

      \[
      R = (\alpha_1 B_2 + a_0 - b_1 C_1)
      \]

      \[
      S = (a_2 B_1 - b_2 C_1)
      \]

      \[
      W = a_3 C_2
      \]

      \[
      B_2 = (a_3 C_2 + R) / W
      \]

      \[
      C_3 = (B_2 R + a_3 C_2 + S) / W
      \]

3. COF 4: n = 4
   a. Conversion formulae
      
      \[
      F_z = \frac{a_0 + a_2 z + a_2 z^2 + a_3 z^3 + a_4 z^4}{b_0 + b_1 z + b_2 z^2 + b_3 z^3 + b_4 z^4} = C_0 + \frac{C_1}{(z + B_1) + \frac{C_2}{(z + B_2) + \frac{C_3}{(z + B_3) + \frac{C_4}{(z + B_4)}}}}
      \]

      \[
      C_0 = a_4 / b_4
      \]

      \[
      \alpha_0 = a_0 - C_0 b_0
      \]

      \[
      \alpha_1 = a_1 - C_0 b_1
      \]

      \[
      \alpha_2 = a_2 - C_0 b_2
      \]

      \[
      \alpha_3 = a_3 - C_0 b_3
      \]

      \[
      C_1 = \alpha_3 / b_4
      \]

      \[
      B_1 = (b_2 C_1 - \alpha_2) / \alpha_3
      \]

      \[
      C_2 = (b_2 C_1 - \alpha_2 - \alpha_3 B_1) / \alpha_3
      \]

      \[
      R = (\alpha_1 B_2 + a_0 - b_1 C_1)
      \]

      \[
      S = (a_2 B_1 - b_2 C_1)
      \]

      \[
      W = a_3 C_2
      \]

      \[
      B_2 = (a_3 C_2 + R) / W
      \]

      \[
      C_3 = (B_2 R + a_3 C_2 + S) / W
      \]
\[ V = C_3 W \]
\[ B_2 = (C_2 R + \alpha C_2 + B_2 S) / (-V) \]
\[ C_4 = \left[ \alpha_0 C_2 B_2 + S(B_2 B_2 + C_2) \right] / (-V) \]
\[ B_4 = (B_2 S + \alpha_0 C_2) / V \]

b. Evaluation

\[ P_1 = C_5 (z + B_4) \]

\[ Q_1 = (z + B_3) (z + B_4) + C_4 \]
\[ P_2 = C_3 Q_1 \]
\[ Q_2 = (z + B_2) Q_1 + P_1 \]
\[ P_3 = C_3 Q_2 \]
\[ Q_3 = (z + B_1) Q_2 + P_2 \]
\[ F_4 = C_0 + \frac{P_3}{Q_4} \]
Appendix

Gaussian Continued Fractions and Their Approximants

1. Tan x

\[
\frac{\tan x}{x} = \frac{1}{1 - \frac{x^2}{3 - \frac{x^2}{5 - \frac{x^2}{7 - \frac{x^2}{9 - \frac{x^2}{11 - \frac{x^2}{13 - \frac{x^2}{15 - \frac{x^2}{17 - \frac{A_n}{B_n}}}}}}}}}
\]

- \[A_0 = 1\]
- \[A_1 = 3\]
- \[A_2 = 15 - x^2\]
- \[A_3 = 105 - 10x^2\]
- \[A_4 = 945 - 105x^2 + x^4\]
- \[A_5 = 10,935 - 1260x^2 + 21x^4\]
- \[A_6 = 135,135 - 17,325x^2 + 378x^4 - x^6\]
- \[A_7 = 2,027,025 - 270,270x^2 + 6990x^4 - 36x^6\]
- \[A_8 = 34,459,425 - 4,729,725x^2 + 135,135x^4 - 990x^6 + x^8\]
- \[B_0 = 1\]
- \[B_1 = 3 - x^2\]
- \[B_2 = 15 - 6x^2\]
- \[B_3 = 105 - 45x^2 + x^4\]
- \[B_4 = 945 - 420x^2 + 15x^4\]
- \[B_5 = 10,935 - 4725x^2 + 210x^4 - x^6\]
- \[B_6 = 135,135 - 62,370x^2 + 3150x^4 - 28x^6\]
- \[B_7 = 2,027,025 - 945,945x^2 + 51,975x^4 - 630x^6 + x^8\]
- \[B_8 = 34,459,425 - 16,216,200x^2 + 945,945x^4 - 13,860x^6 + 45x^8\]

2. Arctan x

\[
\frac{\arctan x}{x} = \frac{1}{1 + \frac{x^2}{3 + \frac{4x^2}{5 + \frac{9x^2}{7 + \frac{16x^2}{9 + \frac{25x^2}{11 + \frac{36x^2}{13 + \frac{49x^2}{15 + \frac{64x^2}{17 + \frac{81x^2}{19 + \frac{100x^2}{21 - \frac{A_n}{B_n}}}}}}}}}}}
\]

- \[A_0 = 1\]
- \[A_1 = 3\]
- \[A_2 = 15 + 4x^2\]
- \[A_3 = 105 + 55x^2\]
- \[A_4 = 945 + 735x^2 + 64x^4\]
- \[A_5 = 10,935 + 10,710x^2 + 2079x^4\]
- \[A_6 = 135,135 + 173,250x^2 + 53,467x^4 + 2,304x^6\]
- \[A_7 = 2,027,025 + 3,108,105x^2 + 1,327,995x^4 + 136,491x^6\]
- \[A_8 = 34,459,425 + 61,486,425x^2 + 33,648,615x^4 + 5,742,495x^6 + 147,456x^8\]
- \[A_9 = 654,729,075 + 1,332,431,100x^2 + 891,080,190x^4 + 216,602,100x^6 + 13,552,575x^8\]
- \[A_{10} = 13,749,310,575 + 31,426,995,600x^2 + 24,861,326,490x^4 + 7,913,505,600x^6 + 865,153,575x^8 + 14,745,600x^{10}\]
- \[B_0 = 1\]
- \[B_1 = 3 + x^2\]
- \[B_2 = 15 + 9x^2\]
- \[B_3 = 105 + 90x^2 + 9x^4\]
- \[B_4 = 945 + 1050x^2 + 225x^4\]
- \[B_5 = 10,935 + 14,175x^2 + 4725x^4 + 225x^6\]
- \[B_6 = 135,135 + 218,295x^2 + 99,225x^4 + 11,025x^6\]
- \[B_7 = 2,027,025 + 3,783,780x^2 + 2,182,950x^4 + 396,900x^6 + 11,025x^8\]
- \[B_8 = 34,459,425 + 72,972,900x^2 + 51,081,030x^4 + 13,097,700x^6 + 883,025x^8\]
- \[B_{10} = 654,729,075 + 1,550,674,125x^2 + 1,277,025,750x^4 + 425,675,250x^6 + 49,116,375x^8 + 893,025x^{10}\]
\[ B_{10} = 13,749,310,575 + 36,010,099,125x^2 \\
y + 34,114,830,750x^4 + 14,047,283,250x^6 \\
+ 2,341,213,875x^8 + 108,056,025x^{10} \]

3. Exponential: \( \text{e}^x \)

a. Gaussian Continued Fraction

\[ e^x = 1 + \frac{x}{1 - \frac{x}{2 + \frac{x}{3 - \frac{x}{2 + \frac{x}{5 - \frac{x}{2 + \frac{x}{7 - \ldots}}}}}} } \]

b. Macon — even part contraction of (a)

\[ e^x = 1 + \frac{2x}{2 - x + \frac{x^2}{6 + \frac{x^2}{10 + \frac{x^2}{14 + \frac{x^2}{18 + \frac{x^2}{22 + \frac{x^2}{26}}}}}} = \frac{S + x}{S - x} \] where \( S = 2 + F \)

and \( F = \frac{x^2}{6 + \frac{x^2}{10 + \frac{x^2}{14 + \frac{x^2}{18 + \frac{x^2}{22 + \frac{x^2}{26}}}}}} \)

or \[ \frac{F}{x^2} = 1 + \frac{x^2}{6 + \frac{x^2}{10 + \frac{x^2}{14 + \frac{x^2}{18 + \frac{x^2}{22 + \frac{x^2}{26}}}}}} = \frac{x^2}{26} = \frac{A_n}{B_n} \]

\[ A_0 = 1 \\
A_1 = 10 \\
A_2 = 140 + x^2 \\
A_3 = 2520 + 28x^2 \\
A_4 = 55,440 + 756x^2 + x^4 \\
A_5 = 1,441,440 + 22,176x^2 + 54x^4 \\
B_0 = 6 \\
B_1 = 60 + x^2 \\
B_2 = 840 + 20x^2 \]

4. Logarithm: \( \log_e \left( \frac{1+x}{1-x} \right) \)

\[ \frac{25x^2}{11} - \frac{36x^2}{13} - \frac{49x^2}{15} - \frac{64x^2}{17} \sim \frac{A_n}{B_n} \]

where \( A_n = 2 \\
A_1 = 6 \\
A_2 = 30 - 8x^2 \\
A_3 = 210 - 110x^2 \\
A_4 = 1800 - 1470x^2 + 128x^4 \\
A_5 = 20,790 - 21,420x^2 + 4158x^4 \\
A_6 = 270,270 - 346,500x^2 + 106,974x^4 - 4608x^6 \\
A_7 = 4,054,050 - 6,216,210x^2 + 2,654,190x^4 \\
- 272,862x^6 \\
A_8 = 68,918,850 - 122,972,850x^2 + 67,297,230x^4 \\
- 11,484,990x^6 + 294,912x^8 \]

\[ B_0 = 1 \\
B_1 = 3 - x^2 \\
B_2 = 15 - 9x^2 \\
B_3 = 105 - 90x^2 + 9x^4 \\
B_4 = 945 - 1050x^2 + 225x^4 \\
B_5 = 10,395 - 14,175x^2 + 4725x^4 - 225x^6 \\
B_6 = 135,135 - 218,295x^2 + 99,225x^4 - 11,025x^6 \\
B_7 = 2,027,025 - 3,783,780x^2 + 2,182,950x^4 - 396,900x^6 \\
+ 11,025x^8 \\
B_8 = 34,459,425 - 72,972,900x^2 + 51,081,090x^4 \\
- 13,097,700x^6 + 893,025x^8 \]

33
Appendix

Constants

\[\begin{align*}
\pi &= 3.14159 \quad 26535 \quad 89793 \quad 23846 \quad 26433 \\
&\quad 83279 \quad 50288 \\
e &= 2.71828 \quad 18284 \quad 59045 \quad 23536 \quad 02874 \\
&\quad 71352 \quad 66249 \\
\sqrt{2} &= 1.41421 \quad 55623 \quad 73965 \quad 04880 \quad 16867 \\
&\quad 24209 \quad 69807 \\
\sqrt{3} &= 1.73205 \quad 50807 \quad 68889 \quad 79335 \quad 27446 \\
\log_{10} 10 &= 2.30258 \quad 50929 \quad 94045 \quad 68401 \quad 79914 \\
&\quad 54684 \quad 39420 \\
\log_{10} 2 &= 0.69314 \quad 71805 \quad 59944 \quad 30941 \quad 72321 \\
&\quad 21458 \quad 17656 \\
\log_{10} e &= 0.43429 \quad 44819 \quad 03251 \quad 82765 \quad 11289 \\
&\quad 18916 \quad 60508 \\
\log_{10} 2 &= 0.30103 \quad 99956 \quad 63981 \quad 19521 \quad 37388 \\
&\quad 94724 \quad 49362 \\
\sqrt{2} &= 1.25992 \quad 10498 \quad 94873 \quad 16476 \quad 7211 \\
\sqrt{3} &= 5.9160 \quad 79783 \quad 09961 \quad 60425 \quad 67328 \\
\sqrt{70} &= 8.3666 \quad 00265 \quad 34075 \quad 54797 \quad 81720 \\
\tan \frac{\pi}{16} &= 0.19891 \quad 23673 \quad 79658 \quad 00691 \quad 15976 \\
\tan \frac{3\pi}{16} &= 0.68817 \quad 88379 \quad 19298 \quad 91999 \quad 77577
\end{align*}\]
CONTROL DATA SALES OFFICES

ALAMOGORDO • ALBUQUERQUE • ATLANTA • BEVERLY HILLS • BOSTON
CAPE CANAVERAL • CHICAGO • CLEVELAND • COLORADO SPRINGS • DALLAS
DAYTON • DENVER • DETROIT • DOWNEY, CALIF. • HONOLULU • HOUSTON
HUNTSVILLE • ITHACA • KANSAS CITY, KAN. • MINNEAPOLIS • NEWARK
NEW ORLEANS • NEW YORK CITY • OAKLAND • OMAHA • PHILADELPHIA
PHOENIX • PITTSBURGH • SACRAMENTO • SALT LAKE CITY • SAN BERNARDINO
SAN DIEGO • SAN FRANCISCO • SEATTLE • WASHINGTON, D.C.

INTERNATIONAL OFFICES

FRANKFURT, GERMANY • HAMBURG, GERMANY • STUTTGART, GERMANY
LUCERNE, SWITZERLAND • ZURICH, SWITZERLAND • MELBOURNE, AUSTRALIA
CANBERRA, AUSTRALIA • ATHENS, GREECE • LONDON, ENGLAND • OSLO, NORWAY
PARIS, FRANCE • STOCKHOLM, SWEDEN • MEXICO CITY, MEXICO, (REGAL ELECTRONICA DE MEXICO, S.A.) • OTTAWA, CANADA, (COMPUTING DEVICES OF CANADA, LIMITED) • TOKYO, JAPAN, (C. ITOH ELECTRONIC COMPUTING SERVICE CO., LTD.)

CONTROL DATA CORPORATION
8100 34th AVENUE SOUTH, MINNEAPOLIS, MINNESOTA 55440

© 1984, Control Data Corporation
Printed in the United States of America