601 145 00

6600

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

Table of Contents (Continued)

Table of Contents (Continued)

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 $= 4(100)(10^{-9})$SECS $= 4(10^{-7})$SECS

Floating-point multiply:

10 minor cycles $= 10(100)(10^{-9})$SECS $= 10^{-6}$ SECS

Floating-point divide:

29 minor cycles $= 29(100)(10^{-9})$SECS

$= 2.9(10^{-6})$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 $B_i < B_j$" require $6(10^{-7})$ SECS and an additional $5(10^{-7})$ 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 $R(x) = \left| \frac{A(x)}{F(x)} \right|$ 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

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

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

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

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

$$R < 2^{-49} \sim 1.775(10^{-15}),$$

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,

$$f(x) \sim \sum_{n=0}^{m} \frac{f^{(n)}(0)\, x^n}{n!}, \quad |x| < a, \text{ where m is}$$

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 is used 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

$$C_n^m = 0 \quad \textit{if } (n+m) \textit{ is odd}$$

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

if (n+m) is even

Alternately, the shifted Chebyshev polynomials

$$T_n^*(x) = T_n(2x-1),$$

obtained by the transformation $x = \dfrac{1+\cos\theta}{2}$, 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

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

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 mth term, is written

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

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 $C_{m+1}(a)$ $T_{m+1}(x)$. 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 $f(x)\cong a_0+a_1x+a_2x^2+\ldots+a_{n+1}x^{n+1}$, $|x|\le 1$ or $x\varepsilon\,[0,1]$, 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.

Now

$$\frac{T_{n+1}(x)}{C^{n+1}_{n+1}} \text{ or } \left(\frac{T^*_{n+1}(x)}{C^{n+1}_{n+1}}\right)$$
$$=x^{n+1}+t_nx^n+t_{n-1}x^{n-1}+\ldots+t_1x+t_0$$

where

$$t_m=C^m_{n+1}/C^{n+1}_{n+1} \quad .$$

Next try replacing x^{n+1} by

$$x^{n+1}-\frac{T_{n+1}(x)}{C^{n+1}_{n+1}}, \text{ i.e., substitute}$$

$$x^{n+1}=\frac{T_{n+1}(x)}{C^{n+1}_{n+1}}-t_nx^n-t_{n-1}x^{n-1}-\ldots-t_1x-t_0.$$

Letting

$$a_i'=a_i-a_{n+1}t_i \text{ for } i=0, n, \text{ the result is}$$

$$f(x)\cong a_0'+a_1'x+a_2'x^2+\ldots+a_n'x^n+\frac{a_{n+1}T_{n+1}(x)}{C^{n+1}_{n+1}} \quad .$$

Since

$|T_{n+1}(x)|\le 1$, we have

$$\left|\frac{a_{n+1}T_{n+1}(x)}{C^{n+1}_{n+1}}\right|\le\left|\frac{a_{n+1}}{C^{n+1}_{n+1}}\right|=E_1 \; .$$

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 $E_0+E_1<E$, the term $a_{n+1}\,T_{n+1}(x)/C^{n+1}_{n+1}$

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 $\sum_{i=0} E_i<E$. 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

$$\max_{[a,b]}\frac{|P^*_n(x)-f(x)|}{g(x)}=\text{Min}$$

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 $g(x)\equiv 1$ or $g(x)\equiv|f(x)|$ 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

$$\varepsilon_n(x) = f(x) - p_n(x)$$

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 $|\varepsilon_n(x)|$ 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 $\varepsilon_n'(x)$ is continuous on $[a, b]$ and is zero at each x_i is usually required. Hence the $2n+2$ equations to be solved are

$$\varepsilon_n(x_i) = f(x_i) - p_n(x_i) = (-1)^i L$$

$$\text{for } i = 0, n+1$$

$$\varepsilon_n'(x_i) = 0$$

$$\text{for } i = 1, n\ .$$

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, $R_n^m(x)$, 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. $R_n^m(x)$ is called a Padé approximant of $P(x)$ and the associated table formed by putting $R_n^m(x)$ in the $(n+1)$st. row and $(m+1)$st. column, n, $m = 0, 1, 2, \ldots$, is called a Padé table.

Let

$$P(x) = \sum_{i=0}^{\infty} C_i x^i,\ A_m(x) = \sum_{i=0}^{m} a_i x^i,$$

$$B_n(x) = \sum_{i=0}^{n} b_i x^i.$$

Then there are $m+n+2$ coefficients a_i, b_i to be found.

Hence, if we write

$$P(x) \equiv \frac{A_m(x)}{B_n(x)} + \sum_{k=n+m+1}^{\infty} C_k x^k$$

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

or

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

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

$$+ \ldots - [a_0 + a_1 x + \ldots + a_m x^m] \equiv d_{n+m+1} x^{n+m+1} + \ldots .$$

and set all coefficients of x through x^{n+m} to zero, the result is

$$\sum_{i=0}^{k} C_{k-i}b_i - a_k = 0 \text{ for } k=0, n$$

$$\sum_{i=0}^{n} C_{k-i}b_i - a_k = 0 \text{ for } k=n+1, n+m$$

$$a_k = 0 \text{ for } k>m$$

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 $b_0=1$.

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

$$R_n^m(x),$$

set

$$b_0=1$$

and solve the n equations,

$$\sum_{i=0}^{n} C_{k-i}b_i = 0 \text{ for } k=m+1, m+n \text{ for } b_i, i=1, n.$$

Then solve the $(m+1)$ equations

$$\sum_{i=0}^{\min(k,n)} C_{k-i}b_i = a_k, \; k=0, m$$

for a_i, $i=0$, m

and

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

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^{n+m} yield an approximation with the accuracy desired. If $m=n$, the resulting rational function is *more* accurate than the series through x^{n+m}; 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,

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

and

$$A_m(x) = \sum_{i=0}^{m} a_i T_i(x), \; B_n(x) = \sum_{i=0}^{n} b_i T_i(x)$$

then the $(n+m+1)$ unknowns a_i, b_i $(b_0=1)$ are determined from

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

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

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_i:

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

$$a_i = C_j + \frac{C_0 b_j}{2} + \frac{1}{2}\sum_{i=1}^{n} b_i \left(C_{j+i} + C_{|j-i|}\right), j=1, m+n$$

where $a_j=0$ 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 Arcsin 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 $R_N(x) = A_m(x)/B_n(x)$, $N = n+m$.

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

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

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

N. Achieser 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 end-points 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, λ. 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.1.) is nearly always more efficient (i.e., requires fewer operations) to evaluate than the rational form.

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

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

$$\equiv C_0 + \frac{C_1}{(z+B_1)+} \frac{C_2}{(z+B_2)+} \ldots \frac{C_n}{(z+B_n)} \cdot$$

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

$$F = b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \dots}}}$$

be a continued fraction.

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

$$F = b_0 + \frac{a_1}{b_1 +}\,\frac{a_2}{b_2 +}\,\frac{a_3}{b_3 +} \dots$$

$$F = b_0 + \frac{a_1 \rfloor}{\lfloor b_1} + \frac{a_2 \rfloor}{\lfloor b_2} + \frac{a_3 \rfloor}{\lfloor b_3} + \dots$$

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*

$$\left.\begin{aligned} A_n &= b_n\, A_{n-1} + a_n\, A_{n-2} \\ B_n &= b_n\, B_{n-1} + a_n\, B_{n-2} \end{aligned}\right\} n = 1, 2, \dots$$

where $A_{-1} = 1$, $A_0 = b_0$, $B_{-1} = 0$, $B_0 = 1$.

$F_n = A_n/B_n$ is the nth approximant to F.

(*b*) *Bottom to top*

$$P_1 = a_n$$

$$Q_1 = b_n$$

$$\left.\begin{aligned} P_i &= a_{n+1-i}\, Q_{i-1} \\ Q_i &= b_{n+1-i}\, Q_{i-1} + P_{i-1} \end{aligned}\right\} i = 2, n$$

$$F_n = b_0 + \frac{P_n}{Q_n}$$

2. Error Estimate

In determining the size of n required to give a desired accuracy, the following "determinant formula" is often helpful:

$$\frac{A_n}{B_n} - \frac{A_{n-1}}{B_{n-1}} = (-1)^{n-1}\,\frac{a_1\, a_2 \dots a_n}{B_{n-1}\, B_n}$$

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 nth 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

$$\overline{F} = C_0 b_0 + \frac{C_0 C_1 a_1}{C_1 b_1 +}\,\frac{C_1 C_2 a_2}{C_2 b_2 +} \dots \frac{C_{n-1} C_n a_n}{C_n b_n +} \dots,$$

where $C_i \neq 0$, has approximant E_n/D_n where

$$E_n = C_0 C_1 \dots C_n A_n$$

$$D_n = C_1 \dots C_n B_n$$

$$D_0 = B_0 = 1$$

and A_n/B_n is the nth approximant of F as defined in 1.

We have $\dfrac{E_n}{D_n} = \dfrac{C_0 A_n}{B_n}$

and both converge or diverge together.

If $C_0 = 1$, the two continued fractions F and $\overline{F}$ 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 $\dfrac{A_{2n}}{B_{2n}}$; similarly, the *odd* part is the continued fraction whose approximants are

$$\frac{A_{2n+1}}{B_{2n+1}}, n=0, 1, \ldots \;.$$

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

$$F_{even}=b_0+\frac{a_1b_2}{(b_1b_2+a_2)-}\;\frac{a_2a_3b_4}{(b_2b_3+a_3)b_4+b_2a_4-}$$

$$\frac{a_4a_5b_2b_6}{(b_4b_5+a_5)b_6+b_4a_6-}\;\frac{a_6a_7b_4b_8}{(b_6b_7+a_7)b_8+b_6a_8-}\ldots$$

$$F_{odd}=\frac{b_0b_1+a_1}{b_1}-\frac{a_1a_2b_3/b_1}{(b_1b_2+a_2)b_3+b_1a_3-}$$

$$\frac{a_3a_4b_1b_5}{(b_3b_4+a_4)b_5+b_3a_5}-\frac{a_5a_6b_3b_7}{(b_5b_6+a_6)b_7+b_5a_7}-\ldots$$

5. Continued Fractions Equivalent to Series

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

$$C_0+\frac{C_1}{1-}\;\frac{C_2/C_1}{(1+C_2/C_1)}-$$

$$\frac{C_3/C_2}{(1+C_3/C_2)-}\;\frac{C_n/C_{n-1}}{(1+C_n/C_{n-1})}-$$

are equivalent in the sense that

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

Similarly, the power series $C_0+C_1x+C_2x^2+\ldots$ and the continued fraction,

$$C_0+\frac{C_1x}{1-}\;\frac{(C_2/C_1)x}{1+(C_2/C_1)x}-\ldots$$

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.):

$$e^x=1+\frac{x}{1-}\;\frac{x}{2+}\;\frac{x}{3-}\;\frac{x}{2+}\;\frac{x}{5-}\;\frac{x}{2+}\;\frac{x}{7-}\cdots$$

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

$$e^x=1+\frac{2x}{(2-x)+}\;\frac{2x^2}{[(6+x)2-2x]+}$$

$$\frac{4x^2}{[(10+x)2-2x]+}\;\frac{4x^2}{[(14+x)2-2x]+}\cdots$$

$$e^x=1+\frac{2x}{(2-x)+}\;\frac{2x^2}{12+}\;\frac{4x^2}{20+}\;\frac{4x^2}{28+}\cdots$$

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

$$e^x=1+\frac{2x}{(2-x)+}\;\frac{x^2}{6+}\;\frac{x^2}{10+}\;\frac{x^2}{14+}$$

This is now form 3.b. of Appendix B.

7. Telescoping Procedures for Continued Fractions (Maehly) (See Reference 21)

In III.A.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 nth approximant.

Let

$$f(x)= \frac{\alpha_0|}{|b_0} + \frac{\alpha_1 x|}{|b_1} + \frac{\alpha_2 x|}{|b_2} + \ldots$$

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

$$R_{n+1}(x)= \frac{\alpha_0|}{|b_0} + \frac{\alpha_1 x|}{|b_1} + \ldots + \frac{\alpha_{n+1} x|}{|b_{n+1}} = \frac{A_{n+1}(x)}{B_{n+1}(x)}$$

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

$$\left.\begin{aligned} A_{n+1} &= b_{n+1}A_n + \alpha_{n+1} x\, A_{n-1} \\ B_{n+1} &= b_{n+1}B_n + \alpha_{n+1} x\, B_{n-1} \end{aligned}\right\} \; n \geq 1$$

$$A_0 = \alpha_0,\ A_1 = \alpha_0 b_1$$

$$B_0 = b_0,\ B_1 = b_0 b_1 + \alpha_1 x$$

Now we may alter either the α_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

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

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 nth approximant, $R_n(x)$, it is known that

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

where

$$C_{n+1} = (-1)^n \frac{\alpha_0}{b_0} \prod_{i=0}^{n} \frac{\alpha_{i+1}}{b_i b_{i+1}}.$$

The corresponding limit relation for R^*_n 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

$$g(x)= \frac{\alpha_0|}{|b_0} + \frac{\alpha_1 x^2|}{|b_1} + \ldots + \frac{\alpha_n x^2|}{|b_n} + \ldots \text{ for } |x| \leq \varepsilon.$$

Let

$x = \varepsilon u$ where $|u| \leq 1$ and

$$S^{(2n+2)}(u) = \frac{T_{2n+2}(u)}{2^{2n+1}} = \sum_{k=0}^{n+1} s_{2k} u^{2k}, \text{ where } T_{2n+2}(u)$$

is the Chebyshev polynomial of degree 2n+2.

Define

$$\gamma_k = -|s_{2k}|\, \varepsilon^{2(n+1-k)} \prod_{i=k}^{n+1} \frac{\alpha_i}{b_i} \text{ for } k=0, n.$$

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

$$\lim_{\varepsilon \to 0} \frac{R_n^*(\varepsilon u) - g(\varepsilon u)}{\varepsilon^{2n+2}} = C_{n+1} S^{(2n+2)}(u).$$

Odd Functions:

An odd function, f(x), is one of the form

$$f(x) = \frac{\alpha_0 x|}{|b_0} + \frac{\alpha_1 x^2|}{|b_1} + \ldots + \frac{\alpha_n x^2|}{|b_n} + \ldots \text{ for } |x| \leq \varepsilon.$$

Let $x = \varepsilon u$ 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

$$\gamma_k = -|s_{2k+1}| \varepsilon^{2(n+1-k)} \prod_{i=k}^{n+1} \frac{\alpha_i}{b_i} \text{ for } k = 0, n.$$

Then

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

and

$$\lim_{\varepsilon \to 0} \frac{R_n^*(\varepsilon u) - f(\varepsilon u)}{2^{2n+3}} = C_{n+1} S^{(2n+3)}(u).$$

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 $\log_e u$. 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 $N=2^{2m}\cdot x$ 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 end-points $\frac{1}{4}$, 1 should be included in the range of x. For example $16=2^4\cdot 1=2^6\left(\frac{1}{4}\right)$.

Hence $\sqrt{N}=2^m\cdot\sqrt{x}$.

2. Computation of $\sqrt{x}$ for $\frac{1}{4}\leq x<1$ 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 $|y_{i+1}-y_i|<2^{-49}$. The number of iterations required will depend upon the accuracy of the first guess, y_1. 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.

*See References 1, 2, 5, 13, 14, 15, 16 and 20.

3. Estimates for y_1

a. Best-fit rational derived by Maehly (Ref. 14) for y_1 with max. relative error $<2.6(10^{-3})$.

$$y_1 = a + \frac{b}{c+x}$$

$a = 3.090315520/\sqrt{2}$

$b = -8.550050013/2\sqrt{2}$

$c = 3.090315520/2$

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.

$$y_1 = \frac{25}{7} - \left[\frac{\frac{5000}{343}\left(x+\frac{15}{49}\right)}{\left(x+\frac{235}{49}\right)\left(x+\frac{15}{49}\right)-\frac{400}{2401}}\right]$$

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 $\frac{1}{4}\leq x\leq\frac{1}{2}$

$$y_1 = \frac{5\sqrt{70}}{14} - \frac{\frac{50\sqrt{70}}{49}\left(x+\frac{3}{14}\right)}{\left(x+\frac{47}{14}\right)\left(x+\frac{3}{14}\right)-\frac{4}{49}}$$

For $\frac{1}{2}\leq x<1$

$$y_1 = \frac{5\sqrt{35}}{7} - \frac{\frac{200\sqrt{35}}{49}\left(x+\frac{3}{7}\right)}{\left(x+\frac{47}{7}\right)\left(x+\frac{3}{7}\right)-\frac{16}{49}}$$

Max. number of iterations = 3

Min. number of iterations = 1

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 20d.)

1. Reduction of Range

To find $\sqrt[3]{N}$ for $N>0$, write $N=2^{3n+k}\cdot x$, 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 $k=0, \pm1, \pm2$. Then $\sqrt[3]{N}=2^n\, 2^{k/3}\, \sqrt[3]{x}$.

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

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

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 $9(10^{-4})$ 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 $9(10^{-6})$ at $x=1$ (Reference 20.d.).

$$y_1 = a_0 - \frac{a_1}{(x+b_1)-}\,\frac{a_2}{(x+b_2)}$$
$$= a_0 - \frac{a_1(x+b_2)}{(x+b_1)(x+b_2)-a_2}$$

$a_0 = 2.502926$

$a_1 = 8.045125$

$b_1 = 4.612244$

$a_2 = .3598496$

$b_2 = .3877552$

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	is needed for a.,
3M+2D+3A=25A	for b., and
2M+1D+4A=16.25A	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. 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.)

$$\text{Sin } x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!}$$

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.

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

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

so that for

$$n=10, \frac{\left(\frac{\pi}{2}\right)^{21}}{21!} \sim 2.6(10^{-16}) < 2^{-49}$$

and for

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

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

a. $|x| \leq \frac{\pi}{2}$

$$\text{Sin } x \sim \sum_{n=0}^{9} \frac{(-1)^n x^{2n+1}}{(2n+1)!} \text{ with error} < 2.6(10^{-16}).$$

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

Then

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

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

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

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

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

Now

$$\left| \frac{\left(\frac{\pi}{2}\right)^{19}}{19!} \frac{T_{19}(y)}{2^{18}} \right| \leq \frac{\left(\frac{\pi}{2}\right)^{19}}{19!\,2^{18}} \sim 1.66(10^{-19}),$$

and this error added to the original error of $2.6(10^{-16})$ is still $<2^{-49}$, so that the term in $T_{19}(y)$ may be dropped. The result is

$$\sin x \sim \sum_{n=0}^{8} \left[\frac{(-1)^n \left(\frac{\pi}{2}\right)^{2n+1}}{(2n+1)!} - \frac{\left(\frac{\pi}{2}\right)^{19}}{19!} C^{(19)}_{2n+1} \right] y^{2n+1}$$

where $C^{(19)}_{2n+1}$ is the coefficient of y^{2n+1} in the Chebyshev polynomial of degree 19. Similarly, after substitution for y^{17} in terms of $T_{17}(y)$, it is found that the coefficient of $T_{17}(y)$ is about $9(10^{-17})$ with the total error still less than 2^{-49}, so that the term containing $T_{17}(y)$ can be dropped. A further substitution for y^{15} results in a coefficient for $T_{15}(y)$ of about $7(10^{-15})$ 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 by setting $y = 2x/\pi$, the final result is

$$\sin x = \sum_{n=0}^{7} C_{2n+1} x^{2n+1} = x \sum_{n=0}^{7} C_{2n+1} z^n \quad \text{where } z = x^2$$

and

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

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

$$C_1 = 1 - \frac{B^9}{18!2^{18}} - \frac{B^8 A}{2^{16}}$$
$$= \quad 9.99999\ 99999\ 99990\ 30428 \quad \text{E-01}$$

$$C_3 = -\frac{1}{3!} + \frac{15B^8}{18!2^{16}} + \frac{3B^7 A}{2^{12}}$$
$$= -1.66666\ 66666\ 66477\ 96113 \quad \text{E-01}$$

$$C_5 = \frac{1}{5!} - \frac{33B^7}{18!2^{13}} - \frac{21B^6 A}{2^{11}}$$
$$= \quad 8.33333\ 33332\ 26184\ 74112 \quad \text{E-03}$$

$$C_7 = -\frac{1}{7!} + \frac{33B^6}{18!2^{10}} + \frac{33B^5 A}{2^9}$$
$$= -1.98412\ 69813\ 94935\ 49426 \quad \text{E-04}$$

$$C_9 = \frac{1}{9!} - \frac{143B^5}{18!2^{10}} - \frac{55B^4 A}{2^8}$$
$$= \quad 2.75573\ 15528\ 52388\ 14908 \quad \text{E-06}$$

$$C_{11} = -\frac{1}{11!} + \frac{91B^4}{18!2^8} + \frac{13B^3 A}{2^5}$$
$$= -2.50518\ 24652\ 10732\ 20541 \quad \text{E-08}$$

$$C_{13} = \frac{1}{13!} - \frac{35B^3}{18!2^6} - \frac{7B^2 A}{2^4}$$
$$= \quad 1.60466\ 21504\ 47864\ 08126 \quad \text{E-10}$$

$$C_{15} = -\frac{1}{15!} + \frac{B^2}{18!2} + \frac{BA}{2^2}$$
$$= -7.35769\ 03984\ 39792\ 89177 \quad \text{E-13}$$

b. $|x| \leq \frac{\pi}{6}$

In like manner, the polynomial approximation

$$\sin x \sim \sum_{n=0}^{6} \frac{(-1)^n x^{2n+1}}{(2n+1)!}$$

can be telescoped to $\sin x \sim x \sum_{n=0}^{5} d_{2n+1} z^n$,

where $z = x^2$ and

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

$$d_1 = 1 - \frac{B^6}{12!2^{12}}$$
$$= \quad 9.9999\ 99999\ 99999\ 783585 \quad \text{E-01}$$

$$d_3 = -\frac{1}{3!} + \frac{7B^5}{12!2^{10}}$$
$$= -1.6666\ 66666\ 66644\ 563864 \quad \text{E-01}$$

$$d_5 = \frac{1}{5!} - \frac{7B^4}{12!2^7}$$
$$= \quad 8.3333\ 33332\ 68836\ 248631 \quad \text{E-03}$$

$$d_7 = -\frac{1}{7!} + \frac{3B^3}{12!2^4}$$
$$= -1.9841\ 26903\ 46738\ 823484 \quad \text{E-04}$$

$$d_9 = \frac{1}{9!} - \frac{5B^2}{12!2^4}$$

$$= \quad 2.7556\ 82887\ 24422\ 163687 \quad E\text{-}06$$

$$d_{11} = -\frac{1}{11!} + \frac{B}{12!2^2}$$

$$= -2.4909\ 02134\ 88806\ 521004 \quad E\text{-}08$$

The remaining algorithms for sin x are all for the range

$$|x| \leq \frac{\pi}{6}.$$

4. Padé Rational Approximations for sin x, $|x| \leq \frac{\pi}{6}$ (See III.B.1.)

Write

$$\frac{\sin x}{x} = \sum_{i=0}^{6} C_i z^i \text{ where } z = x^2 \text{ and } C_i = \frac{(-1)^i}{(2i+1)!}.$$

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.,

$$\sum_{i=0}^{3} C_{k-i}\, b_i = 0 \quad \text{for } k = 4, 5, 6$$

$$\sum_{i=0}^{k} C_{k-i}\, b_i = a_k \quad \text{for } k = 0, 1, 2, 3$$

and $b_0 = 1$,

there result the equations

$$C_1b_3 + C_2b_2 + C_3b_1 = -C_4$$

$$C_2b_3 + C_3b_2 + C_4b_1 = -C_5$$

$$C_3b_3 + C_4b_2 + C_5b_1 = -C_6$$

for b_1, b_2 and b_3,

and then the equations

$$a_0 = C_0$$

$$a_1 = C_1 + C_0b_1$$

$$a_2 = C_2 + C_1b_1 + C_0b_2$$

$$a_3 = C_3 + C_2b_1 + C_1b_2 + C_0b_3$$

for a_0, a_1, a_2 and a_3. The solution is

$$a_0 = 1$$

$$a_1 = -325{,}523/2{,}283{,}996$$

$$a_2 = \quad 34{,}911/7{,}613{,}320$$

$$a_3 = -479{,}249/11{,}511{,}339{,}840$$

$$b_0 = 1$$

$$b_1 = 18{,}381/761{,}332$$

$$b_2 = 1{,}261/4{,}567{,}992$$

$$b_3 = 2{,}623/1{,}644{,}477{,}120$$

so that

$$\sin x = x\left[\frac{a_0 + a_1z + a_2z^2 + a_3z^3}{b_0 + b_1z + b_2z^2 + b_3z^3}\right],\ z = x^2.$$

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

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

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

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

$C_0 = -2.6101\ 46506\ 18158\ 052394$ E01

$C_1 = 7.3922\ 21532\ 39111\ 816487$ E03

$C_2 = 7.3906\ 24698\ 46447\ 182801$ E03

$C_3 = 2.4085\ 74343\ 97122\ 268316$ E03

$B_1 = 1.3171\ 07143\ 04415\ 748255$ E02

$B_2 = -2.5089\ 20716\ 84564\ 146366$ E00

$B_3 = 4.3867\ 21136\ 85869\ 652633$ E01

Since $\sin x < x$ and $\frac{\sin x}{x} = C_0 + \frac{P_2}{Q_2}$, 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

$$\frac{P_2}{Q_2} \text{ to } C_0 = -26. \ldots$$

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 ξ and an extra term ξz into the computation (Ref. 5, p. 25) by writing

$$\frac{\sin x}{x} + \xi z = \sum_{i=0}^{6} \overline{C}_i z^i,$$

where

$$\overline{C}_i = \frac{(-1)^i}{(2i+1)!} \text{ for } i=0, 2, 3, 4, 5, 6$$

and

$$\overline{C}_1 = \xi - \frac{1}{3!}.$$

Solving the same system of equations as before, but replacing C_i by $\overline{C}_i$, the result in terms of ξ is

$$b_3 = \frac{2623}{1{,}235{,}520(1331\text{-}3990\xi)}$$

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

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

$$b_0 = 1$$

$$a_0 = 1$$

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

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

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

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

$$\sin x = x\left[C_0 - \xi z + \frac{P_2}{Q_2}\right].$$

Values of ξ 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° through u=90° at 1° intervals yielded the following results:

Method	*Maximum Absolute Error*	*Maximum Relative Error*	*Max. Error of N in the kth Significant Digit*
2. Taylor, $\lvert x \rvert \leq \frac{\pi}{2}$	$2.56(10^{-16})$ at 90°	$2.56(10^{-16})$ at 90°	2.56 in 16th at 90°
3a. Cheby, $\lvert x \rvert \leq \frac{\pi}{2}$	$3.45(10^{-16})$ at 90°	$9.64(10^{-16})$ at 1°	3.45 in 16th at 90°
3b. Cheby, $\lvert x \rvert \leq \frac{\pi}{6}$	$3.78(10^{-17})$ at 30°	$2.10(10^{-16})$ at 1°	8.59 in 17th at 4°
4. Padé Rational, $\lvert x \rvert \leq \frac{\pi}{6}$	$3.57(10^{-17})$ at 30°	$7.13(10^{-17})$ at 30°	3.57 in 17th at 30°
Cont'd Fraction, $\xi=0$	$3.57(10^{-17})$ at 30°	$7.13(10^{-17})$ at 30°	3.57 in 17th at 30°
" " $\xi=.078$	$5.12(10^{-17})$ at 30°	$1.02(10^{-16})$ at 30°	5.12 in 17th at 30°
" " $\xi=.0785$	$5.14(10^{-17})$ at 30°	$1.03(10^{-16})$ at 30°	5.14 in 17th at 30°
" " $\xi=2.532$	$2.3304(10^{-17})$ at 30°	$4.661(10^{-17})$ at 30°	2.33 in 17th at 30°
" " $\xi=2.533$	$2.3301(10^{-17})$ at 30°	$4.660(10^{-17})$ at 30°	2.33 in 17th at 30°

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, $\lvert x \rvert \leq \frac{\pi}{8}$ and $\lvert x \rvert \leq \frac{\pi}{4}$, assuming first that the argument u has already been reduced to the interval $\lvert u \rvert < \frac{\pi}{2}$.

a. $\lvert x \rvert \leq \frac{\pi}{8}$

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.,

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

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

$$\tan 3\left(\frac{\pi}{8}\right) = \sqrt{2} + 1$$

Then $\tan u = \frac{B + \tan x}{1 - B\tan x}$.

b. $\lvert x \rvert \leq \frac{\pi}{4}$

For

$\lvert u \rvert \leq \frac{\pi}{4}$, let $x = u$ and $\tan u = \tan x$

For

$\lvert u \rvert > \frac{\pi}{4}$, 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

$$\tan x = x\left(\frac{A_6}{B_6}\right).$$

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

$$\tan x = x\left[C_0 + \frac{C_1}{(z+B_1)+}\,\frac{C_2}{(z+B_2)+}\,\frac{C_3}{(z+B_3)}\right]$$

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

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

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

$$= x\left[\frac{a_0+a_1z+a_2z^2+a_3z^3}{b_0+b_1z+b_2z^2+b_3z^3+b_4z^4}\right],$$

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 $x\frac{A_6}{B_6}$, using the results of III.C.7. for an "odd" function.

Take

$$n = 6,\ \varepsilon = \frac{\pi}{4},\ r_0 = \frac{\alpha_0}{b_0} = 1,$$

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

Since

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

$$-\frac{275u^9}{2^6} + \frac{225u^7}{2^7} - \frac{189u^5}{2^9} + \frac{35u^3}{2^{10}} - \frac{15u}{2^{14}}.$$

$s_1 = -15/2^{14}$

$s_3 = 35/2^{10}$

$s_5 = -189/2^9$

$s_7 = 225/2^7$

$s_9 = -275/2^6$

$s_{11} = 45/8$

$s_{13} = -15/4$

and

$$\gamma_k = -|s_{2k+1}|\,\varepsilon^{2(7-k)}\prod_{i=k}^{7} r_i \text{ for } k = 0, 6.$$

Then

$$R_6^* = x \left[\frac{A_6+\gamma_0+x^2(\gamma_2A_0+\gamma_3A_1+\gamma_4A_2+\gamma_5A_3+\gamma_6A_6)}{B_6+\gamma_1x^2+x^2(\gamma_2B_0+\gamma_3B_1+\gamma_4B_2+\gamma_5B_3+\gamma_6B_6)} \right] .$$

After combining, the resulting a_i and b_i are

$a_0 = \quad 135{,}135+\gamma_0$

$\quad = \quad 1.3513\ 50000\ 00000\ 01534\ 86631\quad E05$

$a_1 = -17{,}325+(\gamma_2+3\gamma_3+15\gamma_4+105\gamma_5+945\gamma_6)$

$\quad = -1.7336\ 10607\ 38165\ 56878\ 55239\quad E04$

$a_2 = \quad 378-\gamma_4-10\gamma_5-105\gamma_6$

$\quad = \quad 3.7923\ 56370\ 39100\ 52361\ 14363\quad E02$

$a_3 = -1+\gamma_6$

$\quad = -1.0118\ 62505\ 28977\ 08637\ 24561\quad E00$

$b_0 = \quad 135{,}135$

$b_1 = -62{,}370+\gamma_1+(\gamma_2+3\gamma_3+15\gamma_4+105\gamma_5+945\gamma_6)$

$\quad = -6.2381\ 10607\ 38156\ 27938\ 77667\quad E04$

$b_2 = \quad 3150-\gamma_3-6\gamma_4-45\gamma_5-420\gamma_6$

$\quad = \quad 3.1549\ 37661\ 62835\ 53263\ 58151\quad E03$

$b_3 = -28+\gamma_5+15\gamma_6$

$\quad = -2.8176\ 93975\ 34850\ 64078\ 19239\quad E01$

and

$$\tan x = x \left[\frac{a_0+a_1z+a_2z^2+a_3z^3}{b_0+b_1z+b_2z^2+b_3z^3} \right] ,$$

$$z = x^2,\ |x| \le \frac{\pi}{4} .$$

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

$C_0 = \quad 3.5911\ 01496\ 97721\ 76037\ 74655\quad E\text{-}02$

$C_1 = -9.4381\ 65598\ 19183\ 41369\ 40110\quad E00$

$C_2 = -1.4096\ 32418\ 00227\ 61516\ 62209\quad E03$

$C_3 = -1.5692\ 00421\ 75952\ 56069\ 73336\quad E02$

$B_1 = -5.5204\ 04171\ 66464\ 89417\ 73271\quad E01$

$B_2 = -4.0981\ 70874\ 59656\ 10393\ 42606\quad E01$

$B_3 = -1.5783\ 03284\ 85044\ 64639\ 80047\quad E01$

4. Comparison of Results

The following results were obtained from 1604 double-precision runs using arguments from u=0° through u=89° at 1° intervals.

Method	*Maximum Absolute Error*	*Maximum Relative Error*	*Max. Error of N in the kth Significant Digit*
2a. Rational & Cont'd Fraction, $\lvert x \rvert \le \frac{\pi}{8}$			
$\tan x = xA_6/B_6$	$4.89(10^{-15})$ at 89°	$8.53(10^{-17})$ at 89°	4.89 in 17th at 89°
2b. Rational & Cont'd Fraction, $\lvert x \rvert \le \frac{\pi}{4}$			
1) $\tan x = xA_7/B_7$	$4.54(10^{-16})$ at 45°	$4.54(10^{-16})$ at 45°	4.54 in 16th at 45°
2) $\tan x = xA_8/B_8$	$8.71(10^{-19})$ at 45°	$8.71(10^{-19})$ at 45°	8.71 in 19th at 45°
3. Telescoped form of 2b(1), $\lvert x \rvert \le \frac{\pi}{4}$	$4.70(10^{-16})$ at 45°	$4.70(10^{-16})$ at 45°	4.70 in 16th at 45°

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 $x = \dfrac{u-(\sqrt{2}-1)}{1+u(\sqrt{2}-1)}$

and $\arctan u = \dfrac{\pi}{8} + \arctan x$ ·

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

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

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

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

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

Then

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

and

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

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 $|x| \leq \tan \dfrac{\pi}{16}$.

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

The approximants $\arctan x = xA_6/B_6$ and $\arctan x = xA_7/B_7$ 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 $\arctan x = xA_6/B_6$ was not quite accurate enough, hence A_7/B_7 was telescoped (or A_6/B_6 corrected) as described in the next section to an approximation of the same order as A_6/B_6.

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

Using the formulae in III.C.7. for an odd function and the approximants A_n, B_n in Appendix B.2., and taking $n=6$, $\varepsilon = \tan \dfrac{\pi}{16}$ and $S^{(15)}(u) = T_{15}(u)2^{-14}$, the coefficients in the new R^*_6 are:

a_0 = 1.3513 49999 99999 99825 84406 E05

a_1 = 1.7196 24603 93687 38533 64289 E05

a_2 = 5.2490 48316 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 74603 93685 74205 67287 E05

b_2 = 9.7799 30329 54139 12080 84660 E04

b_3 = 1.0721 37452 05929 68736 47196 E04

and

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

$$z = x^2, \; |x| \leq \tan \frac{\pi}{16}.$$

Coefficients of the corresponding continued fraction found from COF 3, Appendix A.2.a. are:

$C_0 = \ \ 2.0688\ 56828\ 18530\ 47509\ 55450 \quad E\text{-}01$

$C_1 = \ \ 3.0086\ 82092\ 05174\ 87448\ 28121 \quad E00$

$C_2 = -3.4976\ 10177\ 36154\ 25858\ 60195 \quad E00$

$C_3 = -1.3433\ 64284\ 54181\ 78822\ 14637 \quad E\text{-}01$

$B_1 = \ \ 5.1827\ 26637\ 17441\ 95978\ 52947 \quad E00$

$B_2 = \ \ 2.6194\ 66421\ 36919\ 73145\ 76466 \quad E00$

$B_3 = \ \ 1.3197\ 06666\ 86630\ 28901\ 33958 \quad E00$

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 $y = 1°$ through $y = 89°$ at intervals of 1°, results of machine runs on the 1604 in double-precision are given in the table following.

Method	*Maximum Absolute Error*	*Maximum Relative Error*	*Max. Error of N in the kth Significant Digit*
2a. Rational & Cont'd Fraction, $\lvert x \rvert \leq \sqrt{2} - 1$			
Arctan $x = xA_{10}/B_{10}$	$2.28(10^{-16})$ at tan 45°	$4.47(10^{-15})$ at tan 1°	7.80 in 16th at tan 1°
2b. Rational & Cont'd Fraction, $\lvert x \rvert \leq \tan \frac{\pi}{16}$			
1) Arctan $x = xA_6/B_6$	$2.42(10^{-15})$ at tan 45°	$3.40(10^{-14})$ at tan 1°	5.94 in 15th at tan 1°
2) Arctan $x = xA_7/B_7$	$2.36(10^{-17})$ at tan 45°	$2.75(10^{-16})$ at tan 1°	4.80 in 17th at tan 1°
3. Telescoped form of 2b(2), $\lvert x \rvert \leq \tan \frac{\pi}{16}$	$2.38(10^{-17})$ at tan 45°	$2.68(10^{-16})$ at tan 1°	4.68 in 17th at tan 1°

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 $x = \dfrac{u}{\sqrt{1-u^2}}$ requires the square root operation all the time.

b. For

$$\left[\begin{array}{l} 0 \leq u \leq \frac{1}{2}, \text{ set } x = u \text{ and } \arcsin u = \arcsin x \\ \frac{1}{2} < u \leq 1, \text{ set } x = \sqrt{\frac{1-u}{2}} \\ \text{and } \arcsin u = \frac{\pi}{2} - 2 \arcsin x. \end{array}\right.$$

This reduction makes use of the square root half the time.

c. For

$$\begin{cases} 0 \le u \le \frac{1}{2}, \text{ set } x = u,\ A = 0,\ B = 1 \\ \frac{1}{2} < u \le \frac{\sqrt{3}}{2} \sim .866, \\ \text{set } x = 2u^2 - 1,\ A = \frac{\pi}{4},\ B = \frac{1}{2} \\ \frac{\sqrt{3}}{2} < u \le 1, \\ \text{set } x = \sqrt{\frac{1-u}{2}},\ A = \frac{\pi}{2},\ B = -2 \end{cases}$$

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

d. For

$$\begin{cases} 0 \le u \le \frac{1}{2}, \text{ set } x = u,\ A = 0,\ B = 1 \\ \frac{1}{2} < u \le \frac{\sqrt{3}}{2} \sim .866, \\ \text{set } x = 2u^2 - 1,\ A = \frac{\pi}{4},\ B = \frac{1}{2} \\ \frac{\sqrt{3}}{2} < u \le \frac{1}{2}\sqrt{2+\sqrt{3}} \sim .965, \\ \text{set } x = 8u^4 - 8u^2 + 1,\ A = \frac{3\pi}{8},\ B = \frac{1}{4} \\ .965 < u \le 1, \\ \text{set } x = \sqrt{\frac{1-u}{2}},\ A = \frac{\pi}{2},\ B = -2 \end{cases}$$

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 $|x| \le \frac{1}{2}$. Only one will be developed.

2. Telescoped Polynomials for Arcsin x, $|x| \le \frac{1}{2}$

$$\frac{\arcsin x}{x} = 1 + \frac{x^2}{6} + \frac{1 \cdot 3x^4}{2 \cdot 4 \cdot 5} + \frac{1 \cdot 3 \cdot 5x^6}{2 \cdot 4 \cdot 6 \cdot 7} + \dots + \frac{1 \cdot 3 \cdot 5 \dots 39x^{40}}{2 \cdot 4 \cdot 6 \dots 40 \cdot 41}$$

is the truncated series expansion of $\frac{\arcsin x}{x}$ with error less than $6.47(10^{-16}) < 2^{-49}$.

Let

$$z = x^2 = \frac{y}{4} \text{ so that } 0 \le y \le 1 \text{ when } 0 \le x^2 \le \frac{1}{4}.$$

Then

$$\frac{\arcsin x}{x} = C_0 + C_1 y + C_2 y^2 + \dots + C_{20} y^{20},$$

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:

d_0	=	9.9999 99999 99999 78634 63136	E-01
d_1	=	1.6666 66666 66910 24987 33835	E-01
d_2	=	7.4999 99995 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 55117 84246	E-02
d_6	=	1.7312 76426 25238 66058 99121	E-02
d_7	=	1.4331 24507 67095 51847 69121	E-02
d_8	=	9.3428 06551 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

so that

$\arcsin x = x\ [d_0 + d_1 z + d_2 z^2 + \ldots + d_{11} z^{11}]$,

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=.01 through u=1.00 at intervals of .01 with these results—

Max. Absolute Error:	$5.44(10^{-16})$ at u=.5
Max. Relative Error:	$1.04(10^{-15})$ 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 $e^u = 2^n e^x$ where n and x are found as follows:

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

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

$$\text{truncation error} < \frac{x^{13}}{13!} < \frac{(\log_e 2)^{13}}{2^{13}\,13!} \sim 1.07(10^{-15})$$

which is less than 2^{-49} for $|x| \leq \frac{\log_e 2}{2}$.

3. Padé Rational

(Diagonal of the Padé table) Ref. 18

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

where

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

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^u

(Appendix B.3.b.)

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

Approximants tested for F were

$$F = x^2\frac{A_5}{B_5},\ F = x^2\frac{A_4}{B_4} \text{ and } F = x^2\frac{A_3}{B_3}$$

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 $F = x^2\ A_3/B_3$ proved accurate enough.

Thus,

$$e^x = \frac{S+x}{S-x}$$

where

$$S = 2 + z\left[\frac{2520 + 28z}{15{,}120 + 420z + z^2}\right],\ z = x^2$$

or, in continued fraction form,

$$S = 2 + z\left[\frac{P_1}{Q_1}\right] \text{ where } P_1 = C_1(z+B_2)$$

$$Q_1 = (z+B_1)(z+B_2) + C_2$$

and

$C_1 = 28. \qquad B_1 = 330.$

$C_2 = -14{,}580 \qquad B_2 = 90.$

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:

Method	*Maximum Absolute Error*	*Maximum Relative Error*	*Max. Error of N in the Kth Significant Digit*
3. Padé n=5	$3.35(10^{-12})$	$8.26(10^{-16})$	6.53 in 16th
Padé n=6	$5.12(10^{-16})$	$1.72(10^{-19})$	1.33 in 19th
4. Rational and Cont'd Fraction			
a) $F=x^2A_5/B_5$	$5.76(10^{-20})$	$2.63(10^{-23})$	1.95 in 23rd
b) $F=x^2A_4/B_4$	$5.12(10^{-16})$	$1.72(10^{-19})$	1.33 in 19th
c) $F=x^2A_3/B_3$	$3.35(10^{-12})$	$8.26(10^{-16})$	6.53 in 16th
5. Telescoped form of 4.c.	$3.33(10^{-8})$	$3.07(10^{-12})$	2.63 in 12th
	at u=10.	at u=±5.2	at u=4.5

H. LOGARITHM: $\mathrm{Log}_e u = \ln u$

1. Reduction of Range

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

Let

$$x = \frac{m-\sqrt{2}/2}{m+\sqrt{2}/2} \text{ and compute } \ln\left(\frac{1+x}{1-x}\right)$$

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

$$< \frac{2x^{23}}{23} < \frac{2^{24}\,10^{-23}}{23} \sim 7.3(10^{-18})$$

for

$$|x| < 3 - 2\sqrt{2}.$$

Actually one more term could be dropped.

3. Telescoped Polynomial

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

$$\frac{\log \frac{1+x}{1-x}}{2x} = 1 + \frac{a^2 y}{3} + \frac{a^4 y^2}{5} + \ldots + \frac{a^{20} y^{10}}{21} .$$

Shifted Chebyshev polynomials are used and substitutions made for y^{10}, y^9, y^8 and y^7 after which the coefficients are converted back to coefficients of z by the substitution $y = z/a^2$. The result is

$$\log \frac{1+x}{1-x} = 2x\,[C_0 + C_1 z + C_2 z^2 + \ldots + C_6 z^6] ,$$

$$z = x^2$$

C_0 = 1.0000 00000 00000 01720 16224 E00

C_1 = 3.3333 33333 32761 81768 85283 E-01

C_2 = 2.0000 00003 09807 78908 99307 E-01

C_3 = 1.4285 70799 46082 73472 61398 E-01

C_4 = 1.1111 71831 83715 43428 06719 E-01

C_5 = 9.0609 35658 17935 37172 14254 E-02

C_6 = 8.4191 86575 86305 31375 34817 E-02

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

Approximations considered were

a. $\log \frac{1+x}{1-x} = x\frac{A_6}{B_6}$

b. $\log \frac{1+x}{1-x} = x\frac{A_7}{B_7}$

c. $\log \frac{1+x}{1-x} = x\frac{A_8}{B_8}$

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. are:

C_0 = 4.1795 91836 73469 38775 51020 E-01

C_1 = −5.9412 24489 79591 83673 46939 E00

C_2 = −3.3502 52481 31135 23355 48171 E00

C_3 = −1.2872 09952 96610 95132 66527 E-01

B_1 = −5.1029 95328 38691 94833 74554 E00

B_2 = −2.5841 78755 04759 66008 33351 E00

B_3 = −1.3128 25916 56548 39157 92095 E00

$$\log \frac{1+x}{1-x} = x\left[C_0 + \frac{P_2}{Q_2}\right].$$

5. Telescoped Rational and Continued Fraction Forms

The approximation 4.a., $\log \frac{1+x}{1-x} = x\frac{A_6}{B_6}$ 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 $\varepsilon = 3 - 2\sqrt{2}$.

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

a_0 = 2.0789 99999 99999 84154 93231 E04

a_1 = −2.1545 27006 88655 98004 53920 E04

a_2 = 4.2239 18706 18926 27409 32222 E03

b_0 = 1.0395 00000 00000 00000 00000 E04

b_1 = −1.4237 63503 44403 34724 39577 E04

$b_2 = \ \ 4.7788\ 37699\ 95350\ 61419\ 58903 \quad E03$

$b_3 = -2.3041\ 91303\ 93980\ 93764\ 71785 \quad E02$

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

$$\log \frac{1+x}{1-x} = x\left(\frac{P_2}{Q_2}\right)$$

$C_1 = -1.8331\ 45841\ 21857\ 31138\ 11787 \quad E01$

$C_2 = -2.2902\ 78600\ 16831\ 17096\ 16207 \quad E01$

$C_3 = -2.3867\ 37346\ 87530\ 69616\ 24548 \quad E\text{-}01$

$B_1 = -1.5638\ 98338\ 99084\ 36432\ 27360 \quad E01$

$B_2 = -3.7096\ 29801\ 61962\ 30733\ 12768 \quad E00$

$B_3 = -1.3911\ 47833\ 44601\ 98894\ 76038 \quad E00$

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:

Method	*Maximum Absolute Error*	*Maximum Relative Error*	*Max. Error of N in the Kth Significant Digit*
3. Chebyshev teles. polynomial			
a. degree 7 in z	$2.95(10^{-18})$	$4.25(10^{-18})$	2.95 in 18th
b. degree 6 in z	$6.24(10^{-17})$	$9.00(10^{-17})$	6.24 in 17th
4. Rational & Cont'd Fractions			
a. xA_6/B_6	$6.81(10^{-16})$	$9.83(10^{-16})$	6.81 in 16th
b. xA_7/B_7	$5.11(10^{-18})$	$7.37(10^{-18})$	5.11 in 18th
c. xA_8/B_8	$3.83(10^{-20})$	$5.52(10^{-20})$	3.83 in 20th
5. Telescoped form of 4.a.	$7.12(10^{-16})$	$1.03(10^{-15})$	7.12 in 16th
	u=.5, 2, 4, 8	u=.5, 2	u=.5, 2

V. References

BOOKS

1. C. Hastings, "Approximations for Digital Computers", Princeton University Press, 1955.
2. G. N. Lance, "Numerical Methods for High-Speed Computers", Illiffe and Sons Ltd., 1960.
3. C. Lanczos, "Applied Analysis", Prentice-Hall, New York, 1956.
4. R. E. Langer, "On Numerical Approximation", University of Wisconsin Press, 1959.
5. A. Ralston and H. S. Wilf, "Mathematical Methods for Digital Computers", John Wiley & Sons, 1960.
6. H. S. Wall, "Analytic Theory of Continued Fractions", D. Van Nostrand Company, New York, 1948.
7. A. Fletcher, J. C. P. Miller, L. Rosenhead and L. J. Comrie, "An Index of Mathematical Tables", Addison-Wesley Pub. Co., 1962.

MIMEOGRAPHED NOTES

8. E. Frank, "Lectures on the Theory of Continued Fractions", sponsored by O.N.R., Numerical Analysis Research, U.C.L.A., 1957.

INTERNAL MEMOS

System Sciences Division, Control Data Corporation

9. J. Westlake, "Algorithms for Tan x", 3/2/64.
10. J. Westlake, "Algorithms for Square Root, Arctan x, e^x and Sin x", 3/17/64.
11. J. Westlake, "Algorithms for $Log_e u$, Cube Root, Arcsin x and Improved Algorithms for e^u, Tan x, Arctan x and Sin x", 4/16/64.
12. J. Westlake, "Telescoped Continued Fractions for e^u and $log_e u$", 4/29/64.

CONTROL DATA PUBLICATIONS

13. R. E. Smith, G. A. Heuer and D. J. Kiel, "Mathematical Approximations", Control Data Technical Report No. 52, April 1963.
14. H. J. Maehly, "Approximations for the Control Data 1604", March 1960.
15. Fortran Systems for the Control Data 1604 Computer, Computer Division Publication 087A, 1961.
16. Library Functions for the Control Data 3600, Programming Systems Bulletin, November 1963.
17. Control Data 6600 Computer System Reference Manual, First Edition, August 1963.

PAPERS FROM TECHNICAL JOURNALS

18. E. W. Cheney and T. H. Southard, "A Survey of Methods for Rational Approximation", SIAM Review, July 1963, Vol. 5, No. 3, pp. 219-231.

19. P. Henrici, "The Quotient-Difference Algorithm", National Bureau of Standards Appl. Math. Series No. 49 (1958), pp. 23-46.

20. E. G. Kogbetliantz, "Computation of____________Using an Electronic Computer", I.B.M. J. Research and Devel.:

    (a) e^n for $-\infty < N < +\infty$, 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) Arcsin N for $0 < N < 1$, Vol. 2, No. 3, July 1958, pp. 218-222

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

21. H. J. Maehly, "Methods for Fitting Rational Approximations, Part I: Telescoping Procedures for Continued Fractions", J. Assoc. for Computing Mach., Vol. 7, No. 2, April 1960, pp. 150-162.

22. H. J. Maehly (prepared posthumously by Christoph Witzgall), "Methods for Fitting Rational Approximations, Parts II and III", J. Assoc. for Computing Mach., Vol. 10, No. 3, July 1963, pp. 257-277.

23. H. L. Loeb, "Algorithms for Chebyshev Approximations Using the Ratio of Linear Forms", J. Soc. Indust. Appl. Math. 8 (1960), pp. 458-465.

24. N. Macon and M. Baskerville, "On the Generation of Errors in the Digital Evaluation of Continued Fractions", J. Assoc. Computing Mach. Vol. 3 (1956), pp. 199-202.

25. F. D. Murnaghan and J. W. Wrench, Jr., "The Determination of the Chebyshev Approximating Polynomial for a Differentiable Function", M.T.A.C. 13 (1959), pp. 185-193.

26. A. M. Ostrowski, "Note on a Logarithm Algorithm", M.T.A.C. Vol. 9 (1955), pp. 65-68.

27. D. Shanks, "Non-Linear Transformations of Divergent and Slowly Convergent Sequences", J. of Math. and Phys. 34 (1955), pp. 1-42.

28. D. Teichroew, "Use of Continued Fractions in High-Speed Computing", M.T.A.C., Vol. 6 (1952), pp. 127-133.

29. P. Wynn, "The Rational Approximation of Functions Which are Formally Defined by a Power Series Expansion", Math. of Computation 14 (1960), pp. 147-186.

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*

$$\text{Let } F_2 \equiv \frac{a_0+a_1z+a_2z^2}{b_0+b_1z+b_2z^2}$$

$$\equiv C_0 + \frac{C_1}{(z+B_1)+}\frac{C_2}{(z+B_2)}$$

$$C_0 = a_2b_2$$

$$\alpha_0 = a_0 - C_0b_0$$

$$\alpha_1 = a_1 - C_0b_1$$

$$C_1 = \alpha_1/b_2$$

$$B_1 = (C_1b_1 - \alpha_0)/\alpha_1$$

$$C_2 = (C_1b_0 - \alpha_0B_1)/\alpha_1$$

$$B_2 = \alpha_0/\alpha_1$$

b. *Evaluation*

$$P_1 = C_1(z+B_2)$$

$$Q_1 = (z+B_1)(z+B_2) + C_2$$

$$F_2 = C_0 + \frac{P_1}{Q_1}$$

2. COF 3: n = 3

a. *Conversion formulae*

$$\text{Let } F_3 \equiv \frac{a_0+a_1z+a_2z^2+a_3z^3}{b_0+b_1z+b_2z^2+b_3z^3}$$

$$\equiv C_0 + \frac{C_1}{(z+B_1)+}\frac{C_2}{(z+B_2)+}\frac{C_3}{(z+B_3)}$$

$$C_0 = a_3/b_3$$

$$\alpha_0 = a_0 - C_0b_0$$

$$\alpha_1 = a_1 - C_0b_1$$

$$\alpha_2 = a_2 - C_0b_2$$

$$C_1 = \alpha_2/b_3$$

$$B_1 = (b_2C_1 - \alpha_1)/\alpha_2$$

$$T = b_0C_1 - \alpha_0B_1$$

$$C_2 = (b_1C_1 - \alpha_1B_1 - \alpha_0)/\alpha_2$$

$$W = \alpha_2C_2$$

$$B_2 = (\alpha_1C_2 - T)/W$$

$$C_3 = (-T \cdot B_2 + \alpha_0C_2)/W$$

$$B_3 = T/W$$

b. *Evaluation*

$$P_1 = C_2(z+B_3)$$

$$Q_1 = (z+B_2)(z+B_3) + C_3$$

$$P_2 = C_1Q_1$$

$$Q_2 = (z+B_1)Q_1 + P_1$$

$$F_3 = C_0 + \frac{P_2}{Q_2}$$

3. COF 4: n = 4

a. *Conversion formulae*

$$\text{Let } F_4 \equiv \frac{a_0+a_1z+a_2z^2+a_3z^3+a_4z^4}{b_0+b_1z+b_2z^2+b_3z^3+b_4z^4}$$

$$\equiv C_0 +$$

$$\frac{C_1}{(z+B_1)+}\frac{C_2}{(z+B_2)+}\frac{C_3}{(z+B_3)+}\frac{C_4}{(z+B_4)}$$

$$C_0 = a_4/b_4$$

$$\alpha_0 = a_0 - C_0b_0$$

$$\alpha_1 = a_1 - C_0b_1$$

$$\alpha_2 = a_2 - C_0b_2$$

$$\alpha_3 = a_3 - C_0b_3$$

$$C_1 = \alpha_3/b_4$$

$$B_1 = (b_3C_1 - \alpha_2)/\alpha_3$$

$$C_2 = (b_2C_1 - \alpha_1 - \alpha_2B_1)/\alpha_3$$

$$R = (\alpha_1B_1 + \alpha_0 - b_1C_1)$$

$$S = (\alpha_0B_1 - b_0C_1)$$

$$W = \alpha_3C_2$$

$$B_2 = (\alpha_2C_2 + R)/W$$

$$C_3 = (B_2R + \alpha_1C_2 + S)/W$$

$$V = C_3W$$

$$B_3 = (C_3R + \alpha_0C_2 + B_2S)/(-V)$$

$$C_4 = [\alpha_0C_2B_3 + S(B_2B_3 + C_3)]/(-V)$$

$$B_4 = (B_2S + \alpha_0C_2)/V$$

b. *Evaluation*

$$P_1 = C_3(z + B_4)$$

$$Q_1 = (z + B_3)(z + B_4) + C_4$$

$$P_2 = C_2Q_1$$

$$Q_2 = (z + B_2)Q_1 + P_1$$

$$P_3 = C_1Q_2$$

$$Q_3 = (z + B_1)Q_2 + P_2$$

$$F_4 = C_0 + \frac{P_3}{Q_3}$$

B. Appendix

Gaussian Continued Fractions and Their Approximants

1. Tan x

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

$A_0 = 1$

$A_1 = 3$

$A_2 = 15 - x^2$

$A_3 = 105 - 10x^2$

$A_4 = 945 - 105x^2 + x^4$

$A_5 = 10{,}395 - 1260x^2 + 21x^4$

$A_6 = 135{,}135 - 17{,}325x^2 + 378x^4 - x^6$

$A_7 = 2{,}027{,}025 - 270{,}270x^2 + 6930x^4 - 36x^6$

$A_8 = 34{,}459{,}425 - 4{,}729{,}725x^2 + 135{,}135x^4 - 990x^6 + x^8$

$B_0 = 1$

$B_1 = 3 - x^2$

$B_2 = 15 - 6x^2$

$B_3 = 105 - 45x^2 + x^4$

$B_4 = 945 - 420x^2 + 15x^4$

$B_5 = 10{,}395 - 4725x^2 + 210x^4 - x^6$

$B_6 = 135{,}135 - 62{,}370x^2 + 3150x^4 - 28x^6$

$B_7 = 2{,}027{,}025 - 945{,}945x^2 + 51{,}975x^4 - 630x^6 + x^8$

$B_8 = 34{,}459{,}425 - 16{,}216{,}200x^2 + 945{,}945x^4 - 13{,}860x^6 + 45x^8$

2. Arctan x

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

$A_0 = 1$

$A_1 = 3$

$A_2 = 15 + 4x^2$

$A_3 = 105 + 55x^2$

$A_4 = 945 + 735x^2 + 64x^4$

$A_5 = 10{,}395 + 10{,}710x^2 + 2079x^4$

$A_6 = 135{,}135 + 173{,}250x^2 + 53{,}487x^4 + 2{,}304x^6$

$A_7 = 2{,}027{,}025 + 3{,}108{,}105x^2 + 1{,}327{,}095x^4 + 136{,}431x^6$

$A_8 = 34{,}459{,}425 + 61{,}486{,}425x^2 + 33{,}648{,}615x^4 + 5{,}742{,}495x^6 + 147{,}456x^8$

$A_9 = 654{,}729{,}075 + 1{,}332{,}431{,}100x^2 + 891{,}080{,}190x^4 + 216{,}602{,}100x^6 + 13{,}852{,}575x^8$

$A_{10} = 13{,}749{,}310{,}575 + 31{,}426{,}995{,}600x^2 + 24{,}861{,}326{,}490x^4 + 7{,}913{,}505{,}600x^6 + 865{,}153{,}575x^8 + 14{,}745{,}600x^{10}$

$B_0 = 1$

$B_1 = 3 + x^2$

$B_2 = 15 + 9x^2$

$B_3 = 105 + 90x^2 + 9x^4$

$B_4 = 945 + 1050x^2 + 225x^4$

$B_5 = 10{,}395 + 14{,}175x^2 + 4725x^4 + 225x^6$

$B_6 = 135{,}135 + 218{,}295x^2 + 99{,}225x^4 + 11{,}025x^6$

$B_7 = 2{,}027{,}025 + 3{,}783{,}780x^2 + 2{,}182{,}950x^4 + 396{,}900x^6 + 11{,}025x^8$

$B_8 = 34{,}459{,}425 + 72{,}972{,}900x^2 + 51{,}081{,}030x^4 + 13{,}097{,}700x^6 + 893{,}025x^8$

$B_9 = 654{,}729{,}075 + 1{,}550{,}674{,}125x^2 + 1{,}277{,}025{,}750x^4 + 425{,}675{,}250x^6 + 49{,}116{,}375x^8 + 893{,}025x^{10}$

$B_{10}=13{,}749{,}310{,}575+36{,}010{,}099{,}125x^2$
$+34{,}114{,}830{,}750x^4+14{,}047{,}283{,}250x^6$
$+2{,}341{,}213{,}875x^8+108{,}056{,}025x^{10}$

3. Exponential: e^x

a. *Gaussian Continued Fraction*

$$e^x=1+\frac{x}{1-}\;\frac{x}{2+}\;\frac{x}{3-}\;\frac{x}{2+}\;\frac{x}{5-}\;\frac{x}{2+}\;\frac{x}{7-}\quad\cdots$$

b. *Macon — even part contraction of (a)*

$$e^x=1+\frac{2x}{2-x+}\;\frac{x^2}{6+}\;\frac{x^2}{10+}\;\frac{x^2}{14+}\;\frac{x^2}{18+}\;\frac{x^2}{22+}\;\frac{x^2}{26}$$

$$e^x=\frac{(2+F)+x}{(2+F)-x}=\frac{S+x}{S-x}\qquad\text{where } S=2+F$$

$$\text{and } F=\frac{x^2}{6+}\;\frac{x^2}{10+}\;\frac{x^2}{14+}\;\frac{x^2}{18+}\;\frac{x^2}{22+}\;\frac{x^2}{26}$$

$$\text{or}\quad\frac{F}{x^2}=\frac{1}{6+}\;\frac{x^2}{10+}\;\frac{x^2}{14+}\;\frac{x^2}{18+}\;\frac{x^2}{22+}\;\frac{x^2}{26}\sim\frac{A_n}{B_n}$$

$A_0=1$

$A_1=10$

$A_2=140+x^2$

$A_3=2520+28x^2$

$A_4=55{,}440+756x^2+x^4$

$A_5=1{,}441{,}440+22{,}176x^2+54x^4$

$B_0=6$

$B_1=60+x^2$

$B_2=840+20x^2$

$B_3=15{,}120+420x^2+x^4$

$B_4=332{,}640+10{,}080x^2+42x^4$

$B_5=8{,}648{,}640+277{,}200x^2+1512x^4+x^6$

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

$$\frac{\log_e\left(\frac{1+x}{1-x}\right)}{x}=\frac{2}{1-}\;\frac{x^2}{3-}\;\frac{4x^2}{5-}\;\frac{9x^2}{7-}\;\frac{16x^2}{9-}\;\frac{25x^2}{11-}\;\frac{36x^2}{13-}\;\frac{49x^2}{15-}\;\frac{64x^2}{17}\sim\frac{A_n}{B_n}$$

$A_0=2$

$A_1=6$

$A_2=30-8x^2$

$A_3=210-110x^2$

$A_4=1890-1470x^2+128x^4$

$A_5=20{,}790-21{,}420x^2+4158x^4$

$A_6=270{,}270-346{,}500x^2+106{,}974x^4-4608x^6$

$A_7=4{,}054{,}050-6{,}216{,}210x^2+2{,}654{,}190x^4$
$-272{,}862x^6$

$A_8=68{,}918{,}850-122{,}972{,}850x^2+67{,}297{,}230x^4$
$-11{,}484{,}990x^6+294{,}912x^8$

$B_0=1$

$B_1=3-x^2$

$B_2=15-9x^2$

$B^3=105-90x^2+9x^4$

$B_4=945-1050x^2+225x^4$

$B_5=10{,}395-14{,}175x^2+4725x^4-225x^6$

$B_6=135{,}135-218{,}295x^2+99{,}225x^4-11{,}025x^6$

$B_7=2{,}027{,}025-3{,}783{,}780x^2+2{,}182{,}950x^4-396{,}900x^6$
$+11{,}025x^8$

$B_8=34{,}459{,}425-72{,}972{,}900x^2+51{,}081{,}030x^4$
$-13{,}097{,}700x^6+893{,}025x^8$

C. Appendix

Constants

π	= 3.14159 26535 89793 23846 26433 83279 50288
e	= 2.71828 18284 59045 23536 02874 71352 66249
$\sqrt{2}$	= 1.41421 35623 73095 04880 16887 24209 69807
$\sqrt{3}$	= 1.7320 50807 56887 72935 27446
$\log_e 10$	= 2.30258 50929 94045 68401 79914 54684 36420
$\log_e 2$	= 0.69314 71805 59945 30941 72321 21458 17656
$\log_{10} e$	= 0.43429 44819 03251 82765 11289 18916 60508
$\log_{10} 2$	= 0.30102 99956 63981 19521 37388 94724 49302
$\sqrt[3]{2}$	= 1.25992 10498 94873 16476 7211
$\sqrt{35}$	= 5.9160 79783 09961 60425 67328
$\sqrt{70}$	= 8.3666 00265 34075 54797 81720
$\tan \frac{\pi}{16}$	= 0.19891 23673 79658 00691 15976
$\tan \frac{3\pi}{16}$	= 0.66817 86379 19298 91999 77577

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

Pub. No. 60114500