NATIONAL AERONAUTICS AND SPACE ADMINISTRATION



*Technical Report 32-1324*

# *Computational Methods for Mathematical Functions*

*H. C. Thacher, Jr.*
University of Notre Dame



JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA

May 15, 1970

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

*Technical Report 32-1324*

# *Computational Methods for Mathematical Functions*

*H. C. Thacher, Jr.*
University of Notre Dame

JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA

May 15, 1970

Prepared Under Contract No. NAS 7-100
National Aeronautics and Space Administration

# Preface

The work described in this report was performed by the Systems Division of the Jet Propulsion Laboratory.

# Foreword

In the summer of 1967, Dr. Henry Thacher, Jr., Professor of Computing Science at the University of Notre Dame, conducted a series of lectures at JPL on computational methods for mathematical functions. This report contains notes for this series of lectures. Section I gives an up-to-date survey of the most important literature in this field of specialty. Section II discuses infinite expansions in general, reviewing such topics as convergence criteria and truncation error analysis. Section III discusses the relative advantages of series representation and gives examples of error appraisal for infinite series. Section IV discusses power series in particular. Finally, Sections V and VI are devoted to continued fractions and rational functions corresponding to a given power series.

Those of us who attended Professor Thacher's lectures were impressed by his clarity of presentation and are certain that these lectures will be of interest and help to scientists and engineers who are interested in computation.

Acknowledgment is made to Matt Sweeney for editorial assistance.

Edward W. Ng

# Contents

## Contents (contd)



## Tables



## Figures

## Abstract

This report, adapted from lecture notes by the present author, is an expository monograph on analyses useful to problems of computing mathematical functions. Certain well-known types of infinite expansions are surveyed. These include convergent and asymptotic series in general and power series in particular; as well as continued fractions and rational functions. Discussions of these methods emphasize numerical properties such as truncation error analysis, transformation for acceleration of convergence, and recurrence relations amenable to computation. Examples of applications are given for most methods discussed, and exercises are suggested for the reader to apply the various ideas proposed.

# Computational Methods for Mathematical Functions

## I. Introduction

The applications of mathematical analysis to specific problems of physics, engineering, and other areas often require more specific knowledge of the properties of the functions that arise than is customary in pure mathematical investigation. Because of the frequency with which they arise, certain functions, particularly solutions of Laplace's equation, of the diffusion equation, and of the wave equation, have been studied intensively, and have been given the name "special functions," or "special functions of mathematical physics." The expert in special functions is apt to be interested far more in the characteristic properties of individual functions than are most modern analysts.

Since many applications require numerical values at one stage or another, considerable effort has been devoted to the numerical computation and tabulation of special functions. Numerical tables have been compiled and published for many of the most important functions, and before the days of automatic computation, much of the skill of the applied mathematician was devoted to expressing the solutions to problems as simple combinations of tabulated functions. With present equipment, the value of precomputed tables has diminished. It is usually more economical to generate values, either of the function actually required, or of appropriate standard functions, as needed. In either case, however, it is helpful to be familiar with a variety of methods for evaluating functions, and it is the purpose of this monograph to supply some of this information. Many of these methods were originally developed for computing the standard special functions, and we will draw on these calculations for illustrations and exercises. This monograph may thus serve as an introduction to the computational properties of the major special functions.

It is clearly impossible in a work of this scope to include all the properties of all the special functions that have been studied, or even of the most important ones. The literature on special functions is not only voluminous, but widely scattered both in time and space. The most comprehensive survey is the three volumes of Erdelyi, Magnus, Oberhettinger, and Tricomi (Ref. 1). The companion volumes, Erdelyi, Magnus, Oberhettinger, and Tricomi (Ref. 2) also contain much valuable information

on integrals and integral representations of special functions. Among the shorter works which deserve mention for their broad coverage of special functions are the classic Whittaker and Watson (Ref. 3), and the briefer treatments by Rainville (Ref. 4), Magnus and Oberhettinger (Ref. 5), Sneddon (Ref. 6), and Hochstadt (Ref. 7).

Valuable information is also to be found in volumes of tables. The collections of Jahnke and Emde (Ref. 8), Jahnke, Emde and Lösch (Ref. 9), and Abramowitz and Stegun (Ref. 10), hereafter referred to as AMS 55, give collections of the most important formulas, and references to the specialized literature in addition to tables of numerical values. AMS 55 also includes sample calculations illustrating many important methods of evaluating functions.

The introductions to basic tables of individual functions ordinarily describe the methods by which the values were computed and frequently include other useful background information. An exhaustive bibliography of such tables is contained in Fletcher, Miller, Rosenhead, and Comrie (Ref. 11). Similar information for tables of statistical functions appears in Greenwood and Hartley (Ref. 12). More recent tables are reviewed in the journal, *Mathematics of Computation*, which also contains research publications on computation of special functions.

Many of the more important special functions are the subjects of individual monographs. An outstanding example is Watson (Ref. 13), which covers a far wider range than its title would suggest. Integrals of Bessel functions, and much more, are discussed in Luke (Ref. 14), while Slater (Ref. 15) and Slater (Ref. 16) consider the broad classes of confluent hypergeometric and generalized hypergeometric functions. References to other monographs may be found in the bibliographies of the appropriate chapters of Abramowitz and Stegun (Ref. 10).

## II. Infinite Expansions

The various infinite expansions, series, infinite products, continued fractions, and so on, rank high among methods of defining special functions, and among characteristic properties of functions defined in other ways. A major advantage of this mode of definition is that it suggests, at least in principle, an explicit method of numerical evaluation. To serve as an adequate definition of a function, the expansion must converge, at least over part of the domain—for regions outside the domain of convergence the definition may be extended by analytic continuation. For computing function values, on the other hand, convergence is neither sufficient nor necessary: Many convergent expansions approach their limits too slowly to be usable and may involve intolerable cancellation errors as well, while early approximants of divergent expansions may differ from the function being evaluated by much less than the acceptable error, even though the ultimate divergence eventually becomes apparent.

In this section we will consider properties common to most types of expansions, deferring to later chapters discussion of topics of more restricted applicability.

With each infinite expansion $A(x)$, we may associate a sequence of approximants, $\alpha_0, \alpha_1, \cdots, \alpha_k, \cdots$, corresponding to truncation of the expansion after $0, 1, \cdots, k, \cdots$ operations. In infinite series, these approximants are the partial sums; in infinite products, the partial products; for continued fractions, the successive convergents; and so on. In general, these approximants will be real- or complex-valued functions defined over the domain of definition $D$, although for some expansions, including continued fractions, a finite number of approximants may fail to exist. The independent variable $x$ may be a single real or complex variable, or may be a real or complex vector of finite dimension.

If $f(x)$ is a function for which $A(x)$ is a formal expansion, we define the absolute truncation error $R_n(x)$ of the expansion by:

$$R_n(x) = f(x) - \alpha_{n-1}(x) \tag{1}$$

and the relative truncation error by

$$\bar{R}_n(x) = \frac{R_n(x)}{f(x)} = 1 - \frac{\alpha_{n-1}(x)}{f(x)} \tag{2}$$

The relative error is of more interest for computations on floating point computers.

### A. Convergence

An infinite expansion is said to *converge* to $f(x)$ for a particular value of $x$ if

$$\lim_{n\to\infty} \alpha_n(x) = f(x)$$

that is if, for any real $\epsilon > 0$, there exists a finite integer $N(\epsilon, x)$ such that for all $n \geqq N(\epsilon, x)$,

$$|R_n(x)| < \epsilon \tag{3}$$

For infinite series and products, the concept of *absolute convergence* plays a significant role in justifying rearrangements and other formal manipulations. Let the series and product be

$$\left.\begin{aligned} S &\equiv \sum_{k=0}^{\infty} \phi_k(x) \\ P &\equiv \prod_{k=0}^{\infty} [1 + \phi_k(x)] \end{aligned}\right\} \tag{4}$$

and let

$$\left.\begin{aligned} S_a &\equiv \sum_{k=0}^{\infty} |\phi_k(x)| \\ P_a &\equiv \prod_{k=0}^{\infty} [1 + |\phi_k(x)|] \end{aligned}\right\} \tag{5}$$

Then $S$ is absolutely convergent if and only if $S_a$ converges, and $P$ is absolutely convergent if and only if $P_a$ converges. It is well known that absolute convergence implies convergence for both series and products. For other types of infinite expansions, such as continued fractions, rearrangement is ordinarily meaningless, and the analogue of absolute convergence is unimportant. Unfortunately, the term is occasionally applied to behavior which is not analogous. For example, Khovanskii (Ref. 17), following Pringsheim, calls a continued fraction absolutely convergent if

$$T_m \equiv \frac{a_m}{b_m^+} \frac{a_{m+1}}{b_{m+1}^+} \cdots \tag{6}$$

converges for $m = 0, 1, 2, \cdots$, and conditionally convergent if $T_m$ converges only for sufficiently large $m$.

For infinite expansions of functions, the concept of *uniform convergence* is important. An expansion converges uniformly to a function $f(x)$ in a domain $D$ if and only if for any real $\epsilon > 0$, there exists a finite integer $N(\epsilon)$, and for *all* $x$ in $D$,

$$|R_n(x)| < \epsilon \tag{7}$$

Uniform convergence implies convergence of the expansion for every $x$ in $D$. It also implies that, if the approximates are continuous, the limit function $f(x)$ is continuous throughout $D$. It is thus a very strong condition.

The definition of uniform convergence suggests that we are measuring the distance between two functions by the largest difference between their values throughout the whole domain. Although this is a reasonable measure, it is not the only possible one. We can define convergence of sequences and expansions in any norm by the requirement that

$$\lim_{n \to \infty} \|R_n(x)\| = 0$$

where the norm need only satisfy the conditions:

$$\left.\begin{aligned} \|f\| &= 0 \qquad \text{if and only if } f \equiv 0 \\ \|f + g\| &\leq \|f\| + \|g\| \\ \|cf\| &= |c|\,\|f\| \qquad \text{for } c \text{ a scalar constant} \end{aligned}\right\} \tag{8}$$

Among the common norms are the $L_p$ norms:

$$\|f\| = \left[\int_D w(x)\,|f(x)|^p\,dx\right]^{1/p} \tag{9}$$

with $w(x)$ a fixed, nonnegative weight function, and $p \geq 1$. Convergence in the sense that

$$\lim_{n \to \infty} \|R_n(x)\| = 0$$

for a norm other than the max norm is often referred to as *convergence in the mean* or *convergence in norm.* Uniform convergence is clearly a sufficient condition for convergence in the mean, and if $f(x)$ and the approximants are all continuous in $D$, it is also necessary. However, a continuous expansion can converge in the mean to a limit function with a countable number of discontinuities where uniform convergence is clearly impossible.

### B. Asymptotic Expansions

Asymptotic expansions play an important role in computation, since they are often highly efficient. The most familiar asymptotic expansions are series and are usually asymptotic for large values of some variable. We will follow Erdelyi (Ref. 18) in giving a somewhat more general treatment, although we will still restrict ourselves to expansions which are asymptotic with respect to a single complex variable, $z$, confined to a fixed domain, $D$. Other independent variables will be treated as fixed parameters and will be denoted collectively by $p$.

A sequence of functions, $\{\phi_n(z, p)\}$, will be said to form an *asymptotic sequence* for $z \to z_0$ in $D$ if, for each $n$, $\phi_n(z, p)$ is defined for all $z$ in $D$, and, for each real

$\epsilon > 0$ there exists a $\xi$ such that,

$$|\phi_{n+1}(z,p)| < \epsilon\,|\phi_n(z,p)| \tag{10}$$

for all $z$ in $D$ satisfying $|z - z_0| < \xi$. If a $\xi$ exists such that Eq. (10) holds for some $\xi$ and for all $z$ in $D$ satisfying $|z| > \xi$, then the sequence is said to be an asymptotic sequence for $z \to \infty$ in $D$. It follows from this definition that the elements of an asymptotic sequence all approach zero as $z \to z_0$. Some simple examples are: $z^{-n}$ as $z \to \infty$ and $(z - z_0)^n$ as $z \to z_0$.

An expansion,

$$A(z,p) = \{\alpha_n(z,p), \qquad n = 1, 2, \cdots\} \tag{11}$$

will be said to be an *asymptotic expansion* to $N$ approximants for the function $f(z,p)$ as $z \to z_0$ in the domain $D$ if $f(z,p)$ and the sequence

$$\{[\alpha_n(z,p) - \alpha_{n-1}(z,p)], n = 2, 3, \cdots, N\}$$

are defined for all $z$ in $D$, if the sequence

$$\{[\alpha_n(z,p) - \alpha_{n-1}(z,p)], n = 2, 3, \cdots, N\}$$

is an asymptotic sequence for $z \to z_0$, and if, for any $\epsilon > 0$, a $\xi$ can be found such that

$$|f(z,p) - \alpha_N(z,p)| < \epsilon\,|\alpha_N(z,p) - \alpha_{N-1}(z,p)| \tag{12}$$

for all $z$ in $D$ satisfying $|z - z_0| < \xi$.

A given function $f(z,p)$ may have several asymptotic expansions for a given domain $D$ and center $z_0$:

$$f(z,p) \sim A^{(k)}(z,p) = \{\alpha_n^{(k)}(a,p)\} \qquad (k = 1, 2, \cdots, K) \tag{13}$$

However, these will be distinct only if, for each $n$, the elements

$$[\alpha_n^{(k)}(z,p) - \alpha_{n-1}^{(k)}(z,p)]$$

are linearly independent. Thus, the coefficients $a_j$ in an asymptotic series

$$f(z,p) \sim \sum_{j=1}^{N} a_j\,\phi_j(z,p) \qquad z \to z_0 \text{ in } D \tag{14}$$

are uniquely determined by specifying $f(z,p)$, and the asymptotic sequence $\phi_j(z,p)$, and in fact are given recursively by

$$a_j = \lim_{z \to z_0} \frac{f(z,p) - \sum_{i=1}^{j-1} a_i\,\phi_i(z,p)}{\phi_j(z,p)} \qquad j = 1, 2, \cdots, N \tag{15}$$

Conversely, an asymptotic expansion never suffices, by itself, to determine a function $f(z,p)$ uniquely, since if $g(z)$ is any function which vanishes sufficiently rapidly at $z_0$, $f(z,p) + cg(z)$ will have exactly the same asymptotic expansion as $f(z,p)$ for all constants $c$. For example, if $D$ is the positive real axis, and $f(z,p)$ has the asymptotic power series expansion, for $z \to \infty$,

$$f(z,p) \sim \Sigma a_j\,z^{-j} \tag{16}$$

then, $f(z,p) + ce^{-z}$ also has this asymptotic power series expansion.

Asymptotic expansions may have the property that, for a given value of $\epsilon$, a value of $\xi$ can be found for which Eq. (12) is satisfied for all $n$. In analogy to the terminology for convergent expansions, such expansions are said to be uniformly asymptotic with respect to $n$. Similarly, functions depending on parameters may have asymptotic expansions that are uniform with respect to some or all of the parameters, at least within a particular domain of parameter space.

Under our definitions, asymptotic expansions may be, but need not be, convergent. Divergent asymptotic expansions are, however, of such great importance that the term asymptotic is often loosely used in contrast to convergent.

### C. Truncation Error Analysis

As we have pointed out, convergence is of concern primarily when using infinite expansions to characterize functions and has no essential relation to the value of the expansion for computation. For our interests, methods of appraising the truncation error for approximants of finite order will be much more pertinent than methods of investigating convergence.

Truncation errors may be estimated in several ways. It may be possible to find a closed expression for the

error, as an integral or infinite series, for example. This closed expression may then be bounded by appropriate methods. Even if such an *a priori* bound cannot be obtained, it may be possible to demonstrate some relation between the error and numerical results obtained during the calculation. For example, if all the elements of a continued fraction are positive, successive convergents are alternately greater than and less than the limit, and the error can be no greater than the difference between the convergent at which the expansion is terminated, and the first convergent which is neglected. The techniques available for obtaining bounds of these types depend quite strongly upon the particular form of expansion, and will be described at appropriate places in later sections. It frequently happens that an expansion is being studied as the basis of an alternate method of calculation, and that another method of evaluating the function with adequate accuracy is also available for reference. In this case, an empirical investigation of truncation error can be made by computing the differences between the values computed by the truncated expansion and by the reference method throughout the range. Even if the domain of the reference method does not include the entire domain of the new expansion, reasonably reliable estimates of the truncation error are often possible if the two domains overlap significantly, and further analysis may permit the estimates to be made rigorous. An important special case is when truncation error is estimated by comparing computed values with check values from published tables. The use of independent reference function values has the advantage that it not only allows the estimation of truncation errors, but also guards to some extent against programming blunders.

Although the analysis of convergence of an expansion may be difficult and delicate compared with the experimental investigation of truncation error, the result of such a study is far simpler than that of a thorough exploration of truncation. The former consists simply of a delineation of the domain in which the expansion converges. Analysis of truncation error produces far more information, the entire set of domains within which the truncation error of the $n$th approximant is less than each of a fairly large number of specified tolerances. For use with digital computers, the most important tolerances are of the form

$$\rho_s = \frac{1}{2}\beta^{-s} \tag{17}$$

where $\beta$ is the radix of the computer and $s$ may range from 0 up to the maximum number of digits to be considered.

This information can be presented in several ways. Ordinarily, the quantity of interest is $n_{x,s}$, the number of the first approximant for which the truncation error at $x$ is less than $\rho_n$. For many infinite expansions of functions of a single real variable (but not for many of the more effective ones, such as expansions in series of Chebyshev polynomials) $n_{x,s}$ is an increasing function of $|x - x_0|$, where $x_0$ is the "center" of the expansion. For each value of $s$, it is then possible to construct a critical table giving the end points of the largest interval $[x_{-n,s}, x_{n,s}]$ for which the absolute error satisfies:

$$|R_n(x)| \leq \rho_s \qquad x_{-n,s} \leq x \leq x_{n,s} \tag{18}$$

or, for floating point applications when $|f(x)|$ changes significantly, for which the relative error satisfies:

$$|\bar{R}_n(x)| \leq \rho_s \qquad x_{-n,s} \leq x \leq x_{n,s} \tag{19}$$

If several values of $s$ must be considered, either for use in constructing approximations of variable precision or to present results pertinent to several computers, one can construct a double-entry table giving $n_{x,s}$ as a function of the independent variables $x$ and $s$. Such a tabulation involves some loss of efficiency, since $x_{n,s}$ will not ordinarily be independent of $s$, so that the table will no longer be a critical one. Wynn (Refs. 19 and 20) presents tables of this sort for twenty-eight of the better known continued fraction expansions.

Except under exceptional circumstances, table lookup is not as efficient for evaluating functions as is the use of an appropriate analytical approximation. Wynn (Ref. 21) describes a linear programming algorithm for replacing the double-entry table by a polynomial in $s$ and $x$ and illustrates the method by giving one-sided linear and quadratic approximations to the profile for the continued fraction expansion of $\ln(1 + x)$. Unfortunately, this approach does not seem to have been followed up. Ad hoc formulas for expansions of specific functions, or for the associated recurrences have also been developed. Gautschi (Refs. 22 and 23, p. 51) gives a formula for an effective starting point for computing Bessel functions of the first kind by backward recurrence, and also describes an alternate approach due to Kahan. The method can also be extended to regular Coulomb wave functions (Refs. 23, p. 68, and 24). Unfortunately, Gautschi's formulas do not uniformly overestimate the order of the approximant, so that a higher order approximant must also be computed to validate the result.

The results of a study of the truncation error of an expansion of a function of a single real variable may also be presented as a three-dimensional graph, with coordinates giving the order of the approximation, the error (most conveniently on a logarithmic scale), and the value of the independent variable. A contour plot, giving $R_n(x)$ as a function of $x$ for various values of $n$ is also quite clear and is more useful for quantitative purposes. Examples of both these presentations will be found in the next section.

When the study of truncation error is extended to functions of more than one independent real variable, or to functions of one or more complex variables, the results are even harder to summarize. The tabular form requires at least a triple-entry table, while the complete graphical presentation requires four or more dimensions. One-sided approximation in three or more independent variables has not yet been explored at all. In some cases, particularly for functions of a single complex variable, it may be possible to describe the domains within which $\|R_n\| < \rho_s$ by a single parameter. Thus, for truncated power series, the domains are often nearly circles, which can be characterized by their radii. For some continued fractions, the boundaries of the domains are families of confocal ellipses, and for others, families of parabolas. If such a parameter can be found, the complexity of the representation can be reduced to more manageable terms. Another way of reducing the complexity is to abandon the attempt to represent all levels of precision, and to present merely the domains within which a given order of approximant is required to attain a single stated accuracy. O'Shea and Thacher (Ref. 25) performed this task for the modified complex error function,

$$w(z) = \frac{i}{\pi}\int_{-\infty}^{\infty}\frac{e^{-t^2}}{z-t}\,dt = e^{-z^2}\left(1+\frac{2i}{(\pi)^{1/2}}\int_0^z e^{+z^2}\,dt\right)$$
$$= e^{-z^2}\operatorname{erfc}(-iz) \tag{20}$$

They present plots showing the regions of the $z$-plane in which various convergents of two different continued fractions suffice to give 5-decimal accuracy.

The number of expansions for which the truncation error has been studied in detail is far smaller than the number for which the domain of convergence has been established. It would therefore be helpful to have some rule connecting the domain of convergence with the rate. Although we shall see many counter examples, the rule of thumb that the expansion with the larger domain of convergence is apt to be more rapidly convergent near the center of its range holds often enough to give such expansions some priority among the candidates.

### D. Expansions for the Gamma Function, an Example

As an example of an experimental truncation error analysis, we may consider two different expansions for the gamma function. This function, a generalization of the factorial, is defined for $Re z > 0$ by

$$\Gamma(z) = (z-1)! = \int_0^\infty t^{z-1}e^{-t}\,dt \tag{21}$$

It is analytic throughout the complex plane, except for simple poles at $z = -n\,(n = 0, 1, 2, \cdots)$. The existence of these poles is reflected in Euler's infinite product form:

$$\frac{1}{\Gamma(z)} = z\,e^{\gamma z}\prod_{k=1}^{\infty}\left[\left(1+\frac{z}{k}\right)e^{-z/k}\right] \tag{22}$$

which converges throughout the finite complex plane. Here $\gamma$ is Euler's constant:

$$\gamma = \lim_{n\to\infty}\left[\sum_{k=1}^{n}\frac{1}{k} - \log n\right] \cong 0.5772156649\cdots \tag{23}$$

In contrast, we may consider Stirling's series for $\ln\Gamma(z)$:

$$\ln\Gamma(z) \sim \left(z-\frac{1}{2}\right)\ln z - z + \frac{1}{2}\ln 2\pi$$
$$+\sum_{k=1}^{\infty}\frac{B_{2k}}{2k(2k-1)\,z^{2k-1}} \tag{24}$$

where $B_{2k}$ denotes the $2k$th Bernoulli number. This series diverges for all $z$, but is, nevertheless, extremely valuable for computation, since it is asymptotic for $z\to\infty$ in the domain $|\arg z| < \pi$, and the error for a given number of terms decreases very rapidly for even relatively small values of $|z|$.

Although the minimum error attainable for each $x$ is strictly limited by the divergent nature of Stirling's series, and is relatively large for small $|z|$, and for $z$ near the negative real axis, this difficulty can be overcome by using

the basic recurrence,

$$\Gamma(z+1) = z\Gamma(z) \tag{25}$$

To evaluate $\Gamma(z)$ to any desired accuracy, one may choose an $n$ large enough so that Stirling's formula will give the required relative accuracy for $\Gamma(z+n)$, and then use Eq. (25) to compute, successively, $\Gamma(z+n-1)$, $\Gamma(z+n-2)$, $\cdots$, $\Gamma(z)$.

For $z$ in the left half plane, the reflection formula

$$\Gamma(z)\Gamma(-z) = -\frac{\pi}{z}\csc \pi z \tag{26}$$

is also applicable, provided, of course, that $z$ is not close enough to an integer to cause problems with overflow.

### E. Exercises

The exponential function can be computed using the Maclaurin series

$$e^{-x} = \sum_{k=0}^{\infty} \frac{(-x)^k}{k!} \tag{27}$$

or by the continued fraction,

$$e^{-x} = \frac{1}{1+}\,\frac{x}{1-}\,\frac{x}{2+}\,\frac{x}{3-}\,\frac{x}{2+}\,\frac{x}{5-}\,\frac{x}{2+}\,\frac{x}{7-}\,\frac{x}{2+}\,\frac{x}{9-}\,\cdots \tag{28}$$

Both expansions converge throughout the finite complex plane. Tabulate the relative errors of the first 10 approximants of these expansions for $x = 2^n$ $[n = 4(1)4]$.

Note that the continued fraction

$$f = \frac{p_1}{q_1+}\,\frac{p_2}{q_2+}\,\frac{p_3}{q_3+}\,\cdots\,+\frac{p_n}{q_n} \tag{29}$$

can be evaluated by setting

$$r_n = \frac{p_n}{q_n} \tag{30}$$

and, for $k = n-1, n-2, \cdots, 1$,

$$r_k = \frac{p_k}{(q_k + r_{k+1})} \tag{31}$$

where $f = r_1$.

## III. Series

Infinite series are by far the most familiar form of infinite expansion, and, although they are not always the most effective for computation, they often play a significant role in the development of the other forms of expansion. It is thus appropriate to begin by discussing the properties of series in general, and of the more important special types. This section is devoted to characteristics that are common to almost all forms of series. Later sections discuss the more important special types.

### A. Definition

An expansion

$$A(z) = \alpha_n(z), \qquad n = 0, 1, 2, \cdots \tag{32}$$

is a series if a sequence of functions

$$\{\alpha_n(z), n = 0, 1, 2, \cdots\}$$

defined on a domain $D$ can be found such that

$$\alpha_n(z) = \sum_{k=0}^{n} a_k \phi_k(z) \tag{33}$$

The most important types of series are those for which the sequence $\{\phi(z)\}$ is chosen in advance, and the $\{\phi_n(z)\}$ are linearly independent over the domain $D$. In such series, the coefficients $a_k$ are uniquely defined. For a series to be useful in defining a function, it is necessary that the sequence $\alpha_n(z)$ converge; for it to be usable in computation, one must have at least a bound on the errors of the $\alpha_n(z)$.

### B. Advantages of Series Representations

The various series expansions have numerous advantages, which account for their wide popularity. They are by far the most familiar form of expansion, and almost every textbook of analysis above the elementary calculus level devotes a significant amount of space to the properties, derivation, and manipulation of the commoner forms. In addition to the treatments in more general works, there are several monographs which summarize many of the extensive theoretical and practical results available in the literature. Among the more comprehensive we may cite Knopp (Ref. 26) and Bromwich (Ref. 27), while the smaller volume of Hirschman (Ref. 28) presents many useful results in concise form. Many useful formal techniques for manipulating and summing series are given by Schwatt (Ref. 29), while Jolley (Ref. 30) and

Mangulis (Ref. 31) contain extensive collections of series with known sums.

Another major advantage of series over other forms of expansion in their linearity. This facilitates not only such operations as addition, subtraction, and multiplication by scalars, but also allows term-by-term application of linear operations such as integration and differentiation. Such formal operations must, of course, be justified either by considerations of convergence, or more appropriately for our viewpoint, by analysis of the remainder after a finite number of terms. This justification also generally turns out to be easier for series than for other types of expansion.

Series expansions are also advantageous because of the variety of general methods by which they may be derived for functions defined in other ways. Taylor's theorem allows one to write down a power series expansion for any function whose derivatives can be evaluated at a particular point, while formal series solutions to linear differential equations are also readily produced. The coefficients in Fourier series are expressible as integrals by Euler's formula.

A final advantage of series is that any other expansion can be represented as a series, either by special conversion algorithms or, at least, by the trivial formulation

$$\phi_0(z) = \alpha_0(z) \qquad \phi_n(z) = \alpha_n(z) - \alpha_{n-1}(z) \quad (n = 1, 2, 3, \cdots) \tag{34}$$

As we shall see, algorithms exist or can be constructed for the reverse transformation to many other forms of expansion. Thus, the development of many other expansions proceeds through series at one stage or another of its history.

### C. Appraisal of Errors

To use a series expansion for computation, we must be able to appraise the truncation error. In many cases, an error expression which may be used for this purpose appears as a by-product of the derivation; such special error bounds are essential for divergent series since the series itself does not define the function being expanded uniquely. For convergent series, however, it is often possible and convenient to bound the truncation error directly, and such bounds can even be used to establish convergence.

The remainder after $n$ terms of a convergent series,

$$f(x) = \sum_{k=0}^{\infty} a_k \phi_k(x) \tag{35}$$

is given by

$$R_n(x) \equiv f(x) - \sum_{k=0}^{n-1} a_k \phi_k(x) = \sum_{k=n}^{\infty} a_k \phi_k(x) \tag{36}$$

Our task is to produce convenient bounds for the last series.

If $R_n(x)$ is complex, a single bound is not possible. We may either choose to produce separate bounds for the real and imaginary parts of $R_n(x)$:

$$\mathcal{R}_e[R_n(x)] = \sum_{k=n}^{\infty} \mathcal{R}_e[a_k \phi_k(x)]$$

$$\mathcal{I}_m[R_n(x)] = \sum_{k=n}^{\infty} \mathcal{I}_m[\phi_k \phi_k(x)] \tag{37}$$

or merely to bound the magnitude of $R_n(x)$:

$$|R_n(x)| \leq \sum_{k=n}^{\infty} |a_k| \, |\phi_k(x)| \tag{38}$$

In either case, the problem reduces to finding a bound for a series of real terms, and, for Eq. (38), of a series of real nonnegative terms. The bound (Eq. 38) is, however, applicable only if the original series is absolutely convergent.

If the series is to be useful for computation, the rate of convergence must ordinarily be relatively high. This allows us to avoid the more delicate methods of bounding and to be content with relatively crude techniques.

One of the most widely applicable methods of bounding remainders of series is analogous to the comparison test for establishing convergence. Let

$$C(x) = \sum_{k=0}^{\infty} c_k(x) \tag{39}$$

be a convergent series with sum $C(x)$ which is either known, or more readily bounded than the original series. If, for all $k$ and some specified set of $x$,

$$a_{n+k} \phi_{n+k}(x) \leq c_k(x) \tag{40}$$

then, for this set of $x$,

$$R_n(x) \leqq C(x) \tag{41}$$

A lower bound on $R_n(x)$ may be obtained by finding a series for which the inequality (Eq. 40) is reversed.

To use the comparison bound effectively, the $c_k(x)$ should not be significantly larger than $a_{n+k}\,\phi_k(x)$, at least for the largest terms in the remainder. Thus, a set of comparison series with known sums is a decided convenience. Extensive collections are contained in Jolley (Ref. 30), and Mangulis (Ref. 31), while many of the handbooks, such as Abramowitz and Stegun (Ref. 10), Gradshteyn and Ryzhik (Ref. 32), and Dwight (Ref. 33), also contain useful results. Among the most valuable comparison series are the binomial series:

$$(1+s)^r = \sum_{k=0}^{\infty} \frac{r!}{k!\,(r-k)!}\, s^k \qquad (|s|<|) \tag{42}$$

with its special case, the geometric series ($r=-1$)

$$\frac{1}{1-s} = \sum_{k=0}^{\infty} s^k \qquad (|s|<|) \tag{43}$$

and the Riemann zeta function series:

$$\xi(s) = \sum_{k=1}^{\infty} k^{-s} \qquad (Re\, s > |) \tag{44}$$

A table of $\xi(s)$ for integer $s$ appears in AMS 55 (Table 23.3 of Ref. 10), while fractional values may be found in Dwight (Ref. 34).

The sum of the comparison series need not be known exactly for the comparison bound to be applicable; it suffices that it can be bounded reasonably sharply. It is therefore useful to develop other methods of bounding series.

Bounds for monotone series of nonnegative terms can be obtained by a modification of the integral test for convergence. Suppose that in Eq. (39), $c_k(x) \geqq c_{k+1}(x) \geqq 0$ for $k \geqq 0$, and suppose that we can find a monotone nonincreasing function of $s$, $f(s,x)$, defined for $s \geqq -1$, and which, for $k \geqq 0$ satisfies $f(k,x) = c_k(x)$. Then

$$\int_0^{\infty} f(s,x)\,ds \leqq C(x) \leqq \int_{-1}^{\infty} f(s,x)\,ds \tag{45}$$

These bounds follow quite simply by observing that, by the mean value theorem,

$$\int_{k-1}^{k} f(s,x)\,ds = f(\xi,x)$$
$$\int_{k}^{k+1} f(s,x)\,ds = f(\zeta,x) \tag{46}$$

with $k-1 \leqq \xi \leqq k \leqq \zeta \leqq k+1$. Then, in view of the monotonicity of $f(s,x)$,

$$\int_k^{k+1} f(s,x)\,ds \leqq f(k,x) = c_k(x) \leqq \int_{k-1}^{k} f(s,x)\,ds \tag{47}$$

The bounds (Eq. 45) follow by summing these inequalities for $k = 0, 1, 2, \cdots$.

In a significant number of cases, the truncation error may be expressed in terms of the last included, or first neglected, term. There are two cases in which this criterion is applicable with minimal special analysis. The first is when the series converges sufficiently rapidly. Specifically, suppose that, for some fixed $n$ and $x$, and for all $k \geqq 0$, an $\alpha$ between 0 and 1 can be found for which the terms in Eq. (35) all satisfy the inequality:

$$|a_{n+k}\,\phi_{n+k}(x)| \leqq |a_{n-1}\,\phi_{n-1}(x)|\,\alpha^{k+1} \tag{48}$$

Then,

$$|R_n(x)| \leqq \sum_{k=0}^{\infty} |a_{n+k}\phi_{n+k}(x)| \leqq |a_{n-1}\phi_{n-1}(x)| \sum_{k=0}^{\infty} \alpha^{k+1} \tag{49}$$

The last sum is simply the harmonic series, Eq. (43), and so we obtain the bound:

$$|R_n(x)| \leqq |a_{n-1}\phi_{n-1}(x)|\,\frac{\alpha}{1-\alpha} \tag{50}$$

In particular, if $\alpha < \frac{1}{2}$, the truncation error is smaller in magnitude than the last term retained.

The second bound applies to alternating series. Suppose that for fixed $x$,

$$\lim_{k\to\infty} |a_k\phi_k(x)| = 0$$

and that an $n$ can be found such that, for all $k \geqq 0$,

$$|a_{n+k}\phi_{n+k}(x)| \geqq |a_{n+k+1}\phi_{n+k+1}(x)| \tag{51}$$

and

$$\frac{a_{n+k}\,\phi_{n+k}(x)}{|a_{n+k}\phi_{n+k}(x)|} = (-1)^k \frac{a_n\phi_n(x)}{|a_n\phi_n(x)|} \tag{52}$$

(These conditions are necessary and sufficient for convergence of the series at $x$.) Then the truncation error is of the same sign as, and no greater in magnitude than, the first neglected term:

$$0 \leq \frac{R_n(x)}{a_n\phi_n(x)} \leq 1 \tag{53}$$

Convergence is not necessary for this bound to hold; it can also be justified for some alternating divergent asymptotic series, including Stirling's series (Eq. 38), but for such cases the validity must be established individually on an ad hoc basis. The justification is frequently based on the observation that if it can be shown, using some other representation, that successive remainders alternate in sign, then the truncation error must be of the same sign as, and smaller in magnitude than, the first neglected term. Since the sign of the remainder is often considerably easier to establish than more quantitative bounds, this simple criterion is widely applicable.

Two observations should be made on the computational use of these error bounds. The first concerns efficiency. Even a simple bound such as Eq. (50) or Eq. (53) requires a significant amount of computation. An algorithm that computes successive terms of a series, testing at each term whether the remainder is tolerably small may well be less efficient than one which evaluates the series by summing the number of terms given by even a crude overestimate of the number necessary.

The second warning concerns the loss of significance which may occur in summing series with terms of variable sign, including, in particular, alternating series to which the bound (Eq. 53) applies. In many such series the magnitude $l$ of the largest term is considerably larger than the magnitude $s$ of the sum. Under these circumstances, working with $d$-digit floating-point arithmetic, the relative error in $s$ due to roundoff in the largest term will be of the order of $(l/s)\times 10^{-d}$. For example, consider the Maclaurin series for $e^{-x}$. For $x = -11$, the largest term is $11^{10}/10! = 0.71477\times 10^4$, while $e^{-11} = 0.16702\times 10^{-4}$. Thus in 8-decimal floating-point arithmetic, the sum is hardly significant. Thus, in spite of the attractive error bounds available for alternating series, they should be carefully scrutinized for cancellation error, and used only where this can be shown to be insignificant.

### D. Example of Truncation Error Bounding

To illustrate the application of these methods of bounding the remainder, we may consider the Fourier series (Tolstov, Ref. 35, p. 149, Eq. 16)

$$\int_0^x \ln\tan\frac{x}{2}\,dx = -2\sum_{k=0}^{\infty}\frac{\sin(2k+1)x}{(2k+1)^2} \qquad (0 \leq x \leq \pi) \tag{54}$$

For this series, we have

$$R_n(x) = -2\sum_{k=n}^{\infty}\frac{\sin(2k+1)x}{(2k+1)^2} \tag{55}$$

and so, using Eq. (35),

$$|R_n(x)| \leq 2\sum_{k=n}^{\infty}(2k+1)^{-2}\,|\sin(2k+1)x| \tag{56}$$

Since $|\sin(2k+1)x| \leq 1$, we consider the comparison series

$$C_n = \sum_{k=n}^{\infty}(2k+1)^{-2} = \sum_{k=0}^{\infty}(2k+2n+1)^{-2} \tag{57}$$

From Table 23.3 of AMS 55, which gives values of

$$\sum_{k=0}^{\infty}(2k+1)^{-n}$$

for $n = 1(1)42$, we find that $C_0 = 1.2337005501\cdots$. Rather than evaluate $C_n$ by subtracting the first $n$ terms from $C_0$, we will bound $C_n$ by Eq. (45). Letting

$$f(s) = (2s+2n+1)^{-2} \tag{58}$$

so that

$$\int f(s)\,ds = -\frac{1}{2(2s+2n+1)} \tag{59}$$

we obtain the bounds:

$$\int_{-1}^{\infty} f(s)\,ds = \frac{1}{2(2n-1)} \geq C_n \geq \int_0^{\infty} f(s)\,ds = \frac{1}{2(2n+1)} \tag{60}$$

and

$$|R_n(x)| \leq 2C_n \leq \frac{1}{2(2n-1)} \tag{61}$$

**Table 1. Remainders of series $\sum_{k=0}^{\infty} (2k+1)^{-2}$**

| $n$ | $S_n$ | $C_n$ | $\frac{1}{2(2n+1)}$ | $\frac{1}{2(2n-1)}$ |
|---|---|---|---|---|
| 1 | 1.00000 00000 | 0.23370 05501 | 0.16666 66667 | 0.50000 00000 |
| 2 | 1.11111 11111 | 0.12258 94390 | 0.10000 00000 | 0.16666 66667 |
| 3 | 1.15111 11111 | 0.08258 94390 | 0.07142 85714 | 0.10000 00000 |
| 4 | 1.17151 92744 | 0.06218 12757 | 0.05555 55556 | 0.07412 85714 |
| 5 | 1.18386 49535 | 0.04983 55966 | 0.09545 45455 | 0.05555 55556 |
| 6 | 1.19212 94163 | 0.04157 11388 | 0.03846 15385 | 0.04545 45455 |
| 7 | 1.19804 65761 | 0.03565 39740 | 0.03333 33333 | 0.03846 15385 |
| 8 | 1.20249 10205 | 0.03120 95296 | 0.02941 17647 | 0.03333 33333 |
| 9 | 1.20595 12281 | 0.02774 93220 | 0.02631 57895 | 0.02941 17647 |
| 10 | 1.20872 13112 | 0.02497 92389 | 0.02380 95238 | 0.02631 57895 |

To illustrate the sharpness of the bounds (Eq. 60), we present in Table 1 the first ten partial sums of $C_0$,

$$S_n = \sum_{k=0}^{n-1} (2k+1)^{-2} \tag{62}$$

the remainders, $C_n$, computed by $C_n = C_0 - S_n$, and the upper and lower bounds of Eq. (60).

Even though the terms in the series decrease as $k^{-2}$, the rate of convergence is clearly far too slow for practical computation. To obtain guaranteed 8-decimal accuracy, about $5 \times 10^7$ terms would be needed. The fact that we have both upper and lower bounds on the remainder, and that both are positive, suggests a simple way of improving the accuracy. If we add to $S_n$ the average of the upper and lower bounds on $C_n$, $\frac{1}{4}n$, the error in the result will be less than one half the difference in the bounds, $\frac{1}{2}(4n^2 - 1)$:

$$C_0 = S_n + \frac{1}{4}n + \epsilon \qquad |\epsilon| \leq \frac{1}{8n^2 - 2} \tag{63}$$

With this elementary correction, we can obtain 8-decimal accuracy with 5000 terms and even 10 terms will give a result with an error rigorously less than $1.3 \times 10^{-3}$ in magnitude. The actual error turns out to be about $2 \times 10^{-5}$.

Thus a very simple transformation produces a highly significant improvement in the rate of convergence. The next section will be devoted to a discussion of other transformations which can be used to improve convergence.

### E. Elementary Transformations of Series

The discussion following our last example demonstrated the possibility of converting a series that is completely useless for computation into a usable one by a relatively minor modification of the method of computing successive approximants. A variety of devices are known for this purpose, and in this section we will consider some useful transformations which essentially convert one series into another with more desirable properties. More powerful methods, which transform a series into some other form of expansion, will be considered in later chapters.

If we are relying on the sum of our series as the definition of the function to be computed, the transformation should not alter the value of this sum. It is convenient, also, that the transformation preserve convergence, if it exists, since otherwise expressions for the truncation error must be carried through the entire calculation.

The convergence-preserving property of a transformation can often be demonstrated by applying the general rearrangement theorem of A. Markoff, a proof for which appears in Knopp (Ref. 26, p. 242):

Let a convergent series

$$T = \sum_{k=0}^{\infty} t^{(k)}$$

be given, with each of its terms expressed as a convergent series:

$$t^{(k)} = \sum_{j=0}^{\infty} a_j^{(k)} \qquad (k = 0, 1, 2, \cdots) \tag{64}$$

Let the sums

$$\sum_{k=0}^{\infty} a_j^{(k)}$$

be convergent for $j = 0, 1, 2, \cdots$, with sums, $sj$, so that the remainders,

$$R_m^{(k)} = \sum_{j=m}^{\infty} a_j^{(k)} \qquad (m \geqq 0) \tag{65}$$

with fixed $m$ also form a convergent series, with sum $R_m$, and suppose, finally that

$$\lim_{m \to \infty} R_m = 0$$

Then the sum of the $s_j$ also forms a convergent series, and

$$T = \sum_{j=0}^{\infty} s_j = \sum_{k=0}^{\infty} t^{(k)} \tag{66}$$

Let us consider the case where the $t^{(k)}$ are finite sums,

$$t^{(k)} = \sum_{j=0}^{n} \alpha_j c_j^{(k)} \tag{67}$$

so that there is no question of convergence. Suppose, too, that the series

$$s_j = \alpha_j \sum_{k=0}^{\infty} c_j^{(k)} \tag{68}$$

converge. Then the remainders $R_m^{(k)}$ also converge, and, in fact are identically zero for $m > n$. We may thus conclude that

$$T = \sum_{k=0}^{\infty} \sum_{j=0}^{n} \alpha_j c_j^{(k)} = \sum_{j=0}^{n} \alpha_j \sum_{k=0}^{\infty} c_j^{(k)} \tag{69}$$

In words, the terms of a linear combination of convergent series may be expressed as the same linear combination of corresponding terms of the component series.

A simple application of this result, but one which may be highly effective when judiciously applied, is known as *Kummer's transformation*. It consists merely of subtracting corresponding terms of an appropriate series with known sum from the series to be evaluated. Let

$$S(x) = \sum_{k=0}^{\infty} a_k \phi_k(x) \tag{70}$$

be the series to be summed, and suppose that we can find a series

$$S'(x) = \sum_{k=0}^{\infty} a'_k \phi'_k(x) \tag{71}$$

for which we know the sum, $S'(x)$, and for which

$$\lim_{k \to \infty} \frac{a_k \phi_k(x)}{a'_k \phi'_k(x)} = \gamma(x) \tag{72}$$

with $\gamma(x)$ also known. Then the series

$$S(x) - \gamma(x) S'(x) = \sum_{k=0}^{\infty} [a_k \phi_k(x) - \gamma(x) a'_k \phi'_k(x)] \tag{73}$$

converges, and converges more rapidly than the original series.

As an example, we may consider the series $C_0$ of the last section. Writing it in the form

$$C_0 = \sum_{k=0}^{\infty} \frac{1}{4k^2 + 4k + 1} \tag{74}$$

we see that for large $k$, the terms approach those of the series

$$C'_0 = \sum_{k=1}^{\infty} \frac{1}{4k(k+1)} = \frac{1}{4} \sum_{k=1}^{\infty} \left( \frac{1}{k} - \frac{1}{k+1} \right)$$

$$= \frac{1}{4} \left( \sum_{k=1}^{\infty} \frac{1}{k} - \sum_{k=2}^{\infty} \frac{1}{k} \right) = \frac{1}{4} \tag{75}$$

Thus

$$C_0 - \frac{1}{4} = 1 + \sum_{k=1}^{\infty} \left( \frac{1}{4k^2 + 4k + 1} - \frac{1}{4k^2 + 4k} \right)$$

$$= 1 - \sum_{k=1}^{\infty} \frac{1}{4k(k+1)(2k+1)^2} \tag{76}$$

The improved convergence of the transformed series is clear. The terms decrease as $1/k^4$, and ten terms of the series give almost 5-decimal accuracy.

It is perfectly permissible to apply Kummer's formula repeatedly. Although this is equivalent to a single application with a more complicated reference series, it is usually easier to construct a series with the desired properties a step at a time. Knopp (Ref. 26, pp. 260–262) gives several instructive examples of Kummer's transformation, and includes a general method for repeated application to the series $C_0$.

Use of summation devices can occasionally lead to numerical difficulties, even when the transformation is an identity for exact arithmetic. As an example, consider the double sum

$$f(z,h)=\sum_{j=0}^{\infty}\left[\sum_{k=j+1}^{\infty}\frac{(-h)^k}{k!}\right]\frac{(j+1)!}{z^j} \tag{77}$$

which appears in the converging factor for the exponential integral. The inner sum can be recognized as the tails of the exponential series, and one might be tempted to replace it by the finite expression:

$$T_j=\sum_{k=j+1}^{\infty}\frac{(-h)^k}{k!}=e^{-h}-\sum_{k=0}^{j}\frac{(-h)^j}{j!} \tag{78}$$

Observe that since the leading term in $T_j$ is

$$\frac{(-h)^{j+1}}{(j+1)!}$$

the series (77) converges, at least for $|h|<1$. When the second expression for $T_j$ is evaluated, however, the absolute value of the error is at least as large as the error in evaluating $e^{-h}$, which, for $|h|$ small, is of the order of $2^{-t}$ for $t$-bit floating point arithmetic. For double precision, on a 36-bit machine, this error is of the order of $10^{-15}$. Nevertheless, this minuscule error was sufficient to completely destroy numerical convergence of the series.

It is, however, possible (and profitable) to reduce the doubly infinite series to a series of finite sums by interchanging the order of summation:

$$\begin{aligned}f(z,h)&=\sum_{k=1}^{\infty}\sum_{j=0}^{k-1}\frac{(j+1)!}{z^j}\,\frac{(-h)^k}{k!}\\&=z\sum_{k=1}^{\infty}\sum_{j=1}^{k}\frac{j!}{z^j}\,\frac{(-h)^k}{k!}\end{aligned} \tag{79}$$

### F. Analytical Transformations of Series

It is often useful to transform series by the analytical operations of differentiation and integration. Writing the series as a finite sum with a remainder,

$$f(x)=\sum_{k=0}^{n-1}a_k\phi_k(x)+R_n(x) \tag{80}$$

it follows from the linear property of integration and differentiation that

$$\int_a^b f(x)\,dx=\sum_{k=0}^{n-1}a_k\int_a^b\phi_k(x)\,dx+\int_a^b R_n(x)\,dx \tag{81}$$

and

$$f'(x)=\sum_{k=0}^{n-1}a_k\phi_k'(x)+R_n'(x) \tag{82}$$

whenever the indicated operations are meaningful (i.e., whenever $f(x)$ and the $\phi_k(x)$ are integrable, or differentiable).

The utility of this simple process depends upon the possibility of appraising the integral or derivative of the remainder, and is thus strongly dependent on the form in which $R_n(x)$ is given. If $R_n(x)$ can be expressed as an integral, as is often possible, the remainder in Eq. (81) becomes a double integral, and that in Eq. (82) the derivative of an integral. These can often be reduced and simplified by the familiar rules of calculus. Considerably more difficulty is encountered when the remainder is expressed in terms of some function of an undetermined intermediate value, such as, for example, the Cauchy or Lagrange form of the remainder for power series. In this case, the intermediate value depends upon $x$ in an undeterminable fashion, and the effect of integration or differentiation cannot be specified.

Turning now to infinite series, we consider the conditions under which the series of derivatives or integrals of successive terms of a convergent series is convergent, and converges to the derivative or integral of the sum of the original series. If

$$f(x)=\sum_{k=0}^{\infty}a_k\phi_k(x) \tag{83}$$

is a *uniformly convergent* series, and if the $\phi_k(x)$ are continuous functions of $x$ in some domain $D$ of the complex plane, then

$$\int_c f(x)\,dx=\sum_{k=0}^{\infty}a_k\int_c\phi_k(x)\,dx \tag{84}$$

for any path $C$ in the domain $D$. The series (84) is convergent. Similarly,

$$f'(x) = \sum_{k=0}^{\infty} a_k \phi_k'(x) \tag{85}$$

provided the series on the right converges uniformly, and that the series (83) converges.

Among the applications of term-by-term integration and differentiation of series we may mention the derivation of series from known series for the derivative or integral, conversely, obtaining closed expressions for the sums of series as derivatives or integrals of known series, and the production of series solutions of differential and integral equations.

Most of these applications will be discussed in full detail at appropriate places in our later development. It may, however, be of interest to illustrate the use of these techniques by deriving the series (54) for

$$f(x) = \int_0^x \ln \tan\left(\frac{t}{2}\right) dt \tag{86}$$

from the simpler series

$$\ln\left(2 \sin \frac{t}{2}\right) = -\sum_{k=1}^{\infty} \frac{\cos kt}{k} \quad (0 < t < 2\pi) \tag{87}$$

$$\ln\left(2 \cos \frac{t}{2}\right) = -\sum_{k=1}^{\infty} (-1)^k \frac{\cos kt}{k} \quad (-\pi < t < \pi) \tag{88}$$

which converge uniformly over the indicated intervals.

Subtracting Eq. (88) from Eq. (87) we have, for $0 < t < \pi$, the uniformly convergent series

$$\ln \tan\left(\frac{t}{2}\right) = -\sum_{k=1}^{\infty} [1 - (-1)^k] \frac{\cos kt}{k} \tag{89}$$

or, since all terms with even $k$ vanish,

$$\ln \tan\left(\frac{t}{2}\right) = -2 \sum_{k=0}^{\infty} \frac{\cos (2k+1)t}{2k+1} \tag{90}$$

Now, integrating between $\epsilon$ and $x$

$$\int_\epsilon^x \ln \tan\left(\frac{t}{2}\right) dt = -2 \sum_{k=0}^{\infty} \int_\epsilon^x \frac{\cos (2k+1)t}{2k+1} dt \tag{91}$$

$$= -2 \sum_{k=0}^{\infty} \frac{\sin (2k+1)t}{(2k+1)^2} \Bigg|_\epsilon^x \tag{92}$$

or,

$$\int_\epsilon^x \ln \tan\left(\frac{t}{2}\right) dt = -2 \left\{ \sum_{k=0}^{\infty} \frac{\sin (2k+1)x}{(2k+1)^2} - \sum_{k=0}^{\infty} \frac{\sin (2k+1)\epsilon}{(2k+1)^2} \right\} \tag{93}$$

Since Eq. (90) converges uniformly for $0 < t < \pi$, Eq. (93) converges for $0 < \epsilon < x < \pi$. Although the integrand is unbounded as $t$ approaches 0, and as $t$ approaches $\pi$, it can be shown that the integral itself converges.

### G. Linear Summation Processes

The question of the meaning, if any, to be attached to the sum of a divergent series has attracted the interest of mathematicians since the time of Euler. A summary of the principal results on the problem appears in the monograph by Hardy (Ref. 36), while briefer treatments are given in Chapter 13 of Knopp (Ref. 26), and Chapter 5 of Hirschman (Ref. 28). In recent times, the approach to the summability problem has been to consider a particular type of transformation, and if it converts a particular divergent series into a convergent one, to define the value of the divergent series (under the particular summation transformation) as the sum of the resulting convergent series.

It should be remarked that the summability problem is quite distinct from the main problem which concerns us. The divergent series with which we are primarily concerned are formal expansions, often asymptotic, of a function that is perfectly well defined in some other way. Our task is not merely to find a transformation that induces convergence, and thus allows us to assign some value to the series, but to find one that will improve convergence toward the predetermined value.

To prevent utter chaos, the transformations studied in summability theory are ordinarily required to be regular,

that is, to transform all convergent series into convergent series. To be significant to the summability problem, a summation process must also be effective, that is, it must induce convergence in at least one divergent series. Unfortunately the properties of preserving or inducing convergence imply very little about the effect of the transformation upon the rate of convergence, and there seem to be few general a priori rules for estimating the computational value of applying a particular transformation to a specified class of series. One must usually resort either to experimental computations, or to detailed analysis for each particular series.

Omitting, for the time being, explicit designation of any independent variable, let

$$f = \sum_{j=0}^{\infty} \phi_j \tag{94}$$

denote the series, convergent or divergent, to be transformed, and let

$$Tf = \sum_{i=0}^{\infty} \Psi_i \tag{95}$$

designate the resulting transformed series. We will, in this section, restrict our attention to linear transformations for which, if $f$ and $g$ are any two series, and $\alpha$ and $\beta$ are scalars,

$$T(\alpha f + \beta g) = \alpha Tf + \beta Tg \tag{96}$$

Each such operator is equivalent to an infinite matrix, $\Theta$ say, with elements $\theta_{ij}$, and the effect of the transformation may be given by

$$\Psi_i = \sum_{i=0}^{\infty} \theta_{ij}\phi_j \tag{97}$$

Since these are the equations for multiplication of a vector by a matrix, we may associate with $f$ and $Tf$ the column vectors $\phi$ and $\Psi$, the elements of which are, respectively, $[\phi_i]$ and $[\Psi_i]$, and represent the transformation by

$$\Psi = \Theta\phi \tag{98}$$

Two other interpretations of the transformation $T$ are instructive and widely used. Under the first, $T$ is thought of as producing, not the individual terms of the transformed series, but the sequence of partial sums. Thus, $T$ may also be specified by a matrix $\Lambda$, with elements $\lambda_{i,j}$, and with

$$Tf = \lim_{i\to\infty} \sum_{j=0}^{\infty} \lambda_{i,j}\phi_j \tag{99}$$

The coefficients $\lambda_{i,j}$ are related to the $\theta_{i,j}$ by

$$\lambda_{i,j} = \sum_{k=0}^{i} \theta_{k,j}; \left[\theta_{i,j} = \begin{cases} \lambda_{i,j} - \lambda_{i-1,j} & (i > 0) \\ \lambda_{i,j} & (i = 0) \end{cases}\right] \tag{100}$$

Finally, $T$ may be interpreted as a sequence-to-sequence transformation, mapping the partial sums of $f$ into the partial sums of $Tf$. This interpretation may be specified by an array $M$, with elements $\mu_{i,j}$, and with

$$Tf = \lim_{i\to\infty} \sum_{j=0}^{\infty} \mu_{i,j} \sum_{k=0}^{j} \phi_k \tag{101}$$

The $\mu_{i,j}$ are related to the coefficients for the other interpretations by

$$\mu_{i,j} = \lambda_{i,j} - \lambda_{i,j+1} = \sum_{k=0}^{i} (\theta_{k,j} - \theta_{k,j+1}) \tag{102}$$

Conditions that a linear series-to-series transformation should preserve absolute convergence follow easily from Markoff's general rearrangement theorem given in the last section. The somewhat more stringent conditions required for regularity can be found in Hardy (Ref. 36, Th. 2, p. 43). Since the linear transformations we shall discuss are all regular, we will omit the details.

The most familiar linear summation operator is the Cesaro arithmetic mean operator $C(1)$, for which

$$\theta_{i,j} = \begin{cases} 1 & i = j = 0 \\ \dfrac{j}{i(i+1)} & 1 \leq j \leq i \\ 0 & j = 0 \text{ or } i < j \end{cases} \tag{103}$$

$$\lambda_{i,j} = \begin{cases} 1 - \dfrac{j}{i+1} & 0 \leq j \leq i \\ 0 & i < j \end{cases} \tag{104}$$

and

$$\mu_{i,j} = \begin{cases} \dfrac{1}{i+1} & 0 \leq j \leq i \\ 0 & i < j \end{cases} \tag{105}$$

As can be seen from Eq. (105), the $i$th partial sum of the transformed series is the arithmetic mean of the first $i$ partial sums of the original series. The Cesaro transformation thus tends to damp series which fail to converge because the partial sums oscillate too rapidly. As an example of its value we may observe (Fejer's theorem; see Knopp, Ref. 26, Th. 280, p. 494), if $f(x)$ is periodic, with period $2\pi$, and if

$$\int_{-\pi}^{\pi} |f(t)|\,dt$$

exists, then the formal Fourier series,

$$f(x) \sim \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)\,dt = \sum_{k=1}^{\infty}\left\{\left[\frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos ktdt\right]\cos kx + \left[\frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin ktdt\right]\sin kx\right\} \qquad (106)$$

is summable by the $C(1)$ process for all $x$. It has the value $f(x)$ at points where $f(x)$ is continuous, and the value $[f(x-0)+f(x+0)]/2$ at points where $f(x)$ has a jump discontinuity.

As is true for most of the other common linear summation processes, Cesaro summation is effective only for alternating series. It may readily be shown that if all the remainders are of the same sign, Cesaro summation will decrease the rate of convergence, even though it does not actually destroy it. To illustrate the behavior of Cesaro summation we show, in Table 2, the remainders at $x = 1$ of the partial sums, and of the $C(1)$ means for the following alternating series:

$$\frac{1}{1+x} = \sum_{k=0}^{\infty}(-x)^k \qquad (1\times 1<1) \qquad (107)$$

$$\ln(1+x) = \sum_{k=0}^{\infty}\frac{(-x)^{k+1}}{k+1} \qquad (-1<x\leq 1) \qquad (108)$$

$$-\frac{1}{2}\int_0^{\pi x/2}\ln\tan\left(\frac{t}{2}\right)dt = \sum_{k=0}^{\infty}\frac{\sin(2k+1)\pi x}{(2k+1)^2} \qquad (0\leq x\leq 2) \qquad (109)$$

The values of these functions at $x = 1$ are, respectively, ½, 0.693147, and 0.915966.

Cesaro summation is clearly effective for the first series, which is divergent for $x = 1$. It reduces the error of the conditionally convergent series for $\ln 2$ by a factor of about 3. For the third series, which has a fairly rapid initial rate of convergence, the Cesaro averages actually increase the error significantly. This is because the averaging process overweights the early, relatively inaccurate,

**Table 2. Remainders of direct and C (1) summation of alternating series: x = 1**

| f / k | $\frac{1}{1+x}$ | | $\ln(1+x)$ | | $-\frac{1}{2}\int_0^{\pi x/2}\ln\tan\left(\frac{t}{2}\right)dt$ | | $-\frac{1}{2}\int_0^{\pi x/2}\ln\tan\left(\frac{t}{2}\right)dt - S_{99}$ |
|---|---|---|---|---|---|---|---|
| | $R_k$ | C (1) $R_k$ | $R_k$ | C (1) $R_k$ | $R_k$ | C (1) $R_k$ | C (1) $R_k$ |
| 0 | −0.500000 | −0.500000 | −0.306853 | −0.306853 | −0.084034 | −0.084034 | |
| 1 | +0.500000 | 0.000000 | +0.193147 | −0.056853 | +0.027077 | −0.028479 | |
| 2 | −0.500000 | −0.166667 | −0.140186 | −0.084631 | −0.012923 | −0.023194 | |
| 3 | +0.500000 | 0.000000 | +0.109814 | −0.036019 | +0.007485 | −0.015599 | |
| 4 | −0.500000 | −0.100000 | −0.090186 | −0.046853 | −0.004861 | −0.013451 | |
| 5 | +0.500000 | 0.000000 | +0.076481 | −0.023519 | +0.003404 | −0.010642 | |
| 6 | −0.500000 | −0.071429 | −0.066377 | −0.032023 | −0.002514 | −0.009481 | |
| 7 | +0.500000 | 0.000000 | +0.058623 | −0.020692 | +0.001931 | −0.008054 | |
| 8 | −0.500000 | −0.055556 | −0.052488 | −0.024225 | −0.001529 | −0.007329 | |
| 9 | +0.500000 | 0.000000 | +0.047513 | −0.017051 | +0.001241 | −0.006472 | |
| 10 | −0.500000 | −0.045455 | −0.043397 | −0.019446 | −0.001027 | −0.005977 | −0.001027 |
| 11 | +0.500000 | 0.000000 | +0.039937 | −0.014498 | +0.000864 | −0.005407 | −0.000082 |
| 12 | −0.500000 | 0.038462 | −0.036986 | −0.016228 | −0.000736 | −0.005048 | −0.000300 |
| 13 | +0.500000 | 0.000000 | +0.034442 | −0.012608 | +0.000635 | −0.004642 | −0.000066 |
| 14 | −0.500000 | −0.033333 | −0.032225 | −0.013916 | −0.000554 | −0.004370 | −0.000164 |
| 15 | +0.500000 | 0.000000 | +0.030275 | −0.011154 | +0.000487 | −0.004066 | −0.000055 |

partial sums. A considerable increase in precision may be obtained by summing the first few terms directly, and only applying the averaging process to the slowly convergent tails. The effect of summing the first ten terms of Eq. (92) directly, and only applying the $C(1)$ process to the remainder of the series can be seen in the last, incomplete, column of the table.

### H. The Euler Transformation

The Euler transformation is undoubtedly the most valuable of the linear summation processes for computational purposes. We will present it first in a somewhat generalized form due to Knopp. A more extensive treatment than ours, emphasizing the convergence-inducing aspects, may be found in Chapters 8 and 9 of Hardy (Ref. 36). The treatment in Knopp (Ref. 26, pp. 509–518) is limited almost completely to the $E(1)$ transformation.

Let the series to be summed be

$$f(x) = \sum_{k=0}^{\infty} a_k x^{k+1} \tag{110}$$

and suppose that the series converges, at least for sufficiently small $x$. If we introduce the new variable $y$ by

$$\begin{aligned} x &= \frac{y}{(1-qy)} \\ y &= \frac{x}{(1+qx)} \end{aligned} \tag{111}$$

with $q > 0$, then

$$f(x) = \sum_{k=0}^{\infty} a_k \left(\frac{y}{1-qy}\right)^{k+1} = \sum_{k=0}^{\infty} a_k \sum_{j=k}^{\infty} \binom{j}{k} q^{j-k} y^{j+1} \tag{112}$$

using the common generating function for the binomial coefficients (AMS 55, Sec. 24.1.1)

$$(1-t)^{-(m+1)} = \sum_{k=m}^{\infty} \binom{k}{m} t^{k-m} \qquad (|t|<|) \tag{113}$$

Changing the order of summation, we have

$$f(x) = \sum_{j=0}^{\infty} y^{j+1} \sum_{k=0}^{j} \binom{j}{k} q^{j-k} a_k \tag{114}$$

or, letting

$$a_j^{(q)} = \frac{1}{(q+1)^{j+1}} \sum_{k=0}^{j} \binom{j}{k} q^{j-k} a_k \tag{115}$$

$$f(x) = \sum_{j=0}^{\infty} a_j^{(q)} \{(q+1)y\}^{j+1} \tag{116}$$

For $x = 1$, this reduces to

$$f(1) = \sum_{k=0}^{\infty} a_k = \sum_{j=0}^{\infty} a_j^{(q)} \tag{117}$$

If the second summation converges, the original series is said to be summable by the $E(q)$ process. It is shown in Hardy (Ref. 36) that this process is regular, that its power is an increasing function of $q$, and that the operators obey the multiplication law:

$$E(q)\,[E(r)f] = E(q + r + qr)f \tag{118}$$

We will concentrate our attention on the original Euler transformation, the $E(1)$ transformation, for which

$$a_j^{(1)} = \frac{1}{2^{j+1}} \sum_{k=0}^{j} \binom{j}{k} a_k \tag{119}$$

If we write

$$a_k = (-1)^k c_k \tag{120}$$

and use the well-known expression for the forward differences of a sequence of numbers,

$$\Delta^j c_n = \sum_{k=0}^{j} \binom{j}{k} (-1)^k c_{n+k} \tag{121}$$

we can write the $E(1)$ transformation in the more familiar form:

$$E(1) \sum_{k=0}^{\infty} (-1)^k c_k = \sum_{j=0}^{\infty} (-1) \frac{j \Delta^j c_0}{2^{j+1}} \tag{122}$$

For automatic computation, a formulation ascribed to van Wijngaarden, and described in Modern Computing

Methods (Ref. 37, p. 125) is efficient and economical of storage. Let $M$ denote the forward mean operator:

$$\left.\begin{aligned} Ma_n &= \frac{1}{2}(a_n + a_{n+1}) \\ M^{j+1}a_n &= \frac{1}{2}(M^j a_n + M^j a_{n+1}) \end{aligned}\right\} \tag{123}$$

Then

$$E(1)f = E(1)\sum_{k=0}^{\infty} a_k = \frac{1}{2}\sum_{j=0}^{\infty} M^j a_0 \tag{124}$$

A convenient algorithm for the Euler transformation based on this formulation is used as an example in the Algol Report, Naur, et al. (Ref. 38).

For accelerating the convergence of power series, and, by the substitution $x = e^{i\theta}$, of Fourier series, the form

$$E(1)\sum_{k=0}^{\infty} a_k x^k = \frac{1}{1-x}\sum_{j=0}^{\infty} \Delta^j a_0 \left(\frac{x}{1-x}\right)^j \tag{125}$$

which follows simply from Eq. (112) by the substitutions $q = 1$, $x = -x$, is often convenient.

Like most transformations, the Euler transformation has a somewhat variable effect on the rate of convergence. It is apt to be less effective, or even deleterious, when applied to series that converge rapidly to begin with. In contrast to the Cesaro transformation, however, alternating series that converge rapidly enough for the Euler transformation to be ineffective usually converge rapidly enough in the original form to be usable for computation.

As a simple example of the Euler transformation, we may consider the series

$$\frac{1}{1+x} = \sum_{k=0}^{\infty} (-1)^k x^k \tag{126}$$

which the Cesaro summation process summed, with convergence $1/n$ for $x = 1$. It is easily seen that

$$\Delta^j x^k = x^k (x-1)^j$$

and so, using Eq. (120), we have, at least formally,

$$\frac{1}{1+x} = \sum_{j=0}^{\infty} (-1)\frac{j(x-1)^j}{2^{j+1}} \tag{127}$$

By the ratio test, this series converges for $|1 - x| < 2$. For $x = 1$, a single term gives the correct result, but for $x < 1/3$, the magnitudes of successive terms decrease more slowly than do those of the original series.

The use of the Euler transformation for inducing convergence suggests that its utility is not limited to convergent series. Rosser (Ref. 39) has pointed out the advantages of the Euler transformation for alternating divergent asymptotic series. Since asymptotic series are often characterized by an initial region of rapid convergence, in which the Euler transformation is of little value, and may even be harmful, Rosser recommends applying the transformation only to the divergent part of the series, while summing the convergent section directly. The optimum strategy would be to sum directly up to the point where application of the transformation increases the rate of convergence enough to compensate for the effort of using it. This, however, is a poorly defined point, and depends strongly upon the computation equipment, so it is questionable whether the increase in efficiency would justify the additional analysis.

If the original series diverges strongly enough, a single application of the Euler transformation may not suffice to induce rapid convergence, or to induce convergence at all. For such cases, the transformation may be applied repeatedly, summing each transformed series until divergence becomes apparent, and then transforming the tails.

We may illustrate the application of Euler's transformation in summing asymptotic series by one of Rosser's examples. The exponential integrals, $E_n(x)$, may be defined, for $\mathcal{R}_e\, x > 0$, by the integral

$$E_n(x) = \int_1^{\infty} e^{-xt} t^{-n}\, dt \qquad (n = 0, 1, 2, \cdots) \tag{128}$$

By successive integration by parts, we find

$$xe^x E_n(x) = \sum_{k=0}^{\nu-1} (-1)^k \frac{(n+k-1)!}{(n-1)!\,x^k} + \frac{(-1)^\nu (n+\nu-1)!}{(n-1)!\,x^{\nu-1}} e^x E_{n+1}(x) \tag{129}$$

Since, for $x$ real and greater than 0,

$$\frac{1}{x+n} \leq e^x E_n(x) \leq \frac{1}{x+n-1} \leq \frac{1}{x} \qquad (130)$$

the error term is unbounded for large $n$, and the series diverges for all $x$. It is, however, asymptotic for $|x| \to \infty$ for $|arg\, x| < 3\pi/2$. It is also apparent that, for real $x$, the truncation error is of the same size as the first neglected term, and smaller than it in magnitude.

We will attempt to use this series to compute $5e^5 E_1(5)$. By another method, Miller and Hurst (Ref. 40) obtain the value 0.852110881423911. The magnitudes of the terms $c_k = k!/5^k$ are given in the second column of Table 3. The smallest term is $c_4 = 0.03840$, and the sum through $c_3$ is 0.83200. Thus, the minimum error obtainable by direct summation of the asymptotic series is about 4%, as can be confirmed by comparison with the exact value. The third column gives the values of $\Delta^k c_4$, computed by Rosser using some special devices to minimize the amount of multiple-precision calculation, and the fourth gives the magnitudes of the terms in the alternating series produced by transforming the tails. Again, the initial convergence is excellent, but the $c'_k$ with $k > 7$ increase in magnitude. The sum of the terms through $c'_6$ is 0.0201604800, and, adding this correction to the original sum, we obtain 0.8521604800, agreeing now to almost 5 decimals. The fifth and sixth columns give the differences of $c'_7$, and the magnitudes of the terms of the Euler transform of the tails of the second series. The sum of this third series is $-0.0000496492$. Adding this correction to the cumulative sum, we obtain the value 0.8521108308, differing from the correct value by only $5.06 \times 10^{-8}$.

In summing a divergent asymptotic series by the Euler transformation, or indeed by any summation process, it is necessary to verify that the sum represents the desired function, and not some one of the multitude of other functions having the same asymptotic expansion. In our example, the agreement with an independent calculation to almost 8 decimals is strong evidence for the validity of the process. In this case, moreover, Rosser was able to confirm the rigor of the process by judicious rearrangement of the remainder. Detailed analysis is required, however, for each asymptotic series.

The Euler transformation, in either the form of Eq. (122), or the form of Eq. (124), is computationally stable when the calculations are carried out to a fixed number of decimal places. The maximum buildup of roundoff error in forming the differences is more than compensated by the divisions by powers of 2, while the formation of the means is also stable. In summing strongly divergent series, however, multiple precision will be needed to maintain the necessary number of decimal places because of the rapid growth of the coefficients.

**Table 3. Evaluation of $5e^5 E_1(5)$ using Euler transformation on $e^x E_1(x) \sim \sum_{k=0}^{\infty} (-1)^k c_k(x)$; $c_k = \frac{k!}{x^k}$**

| $k$ | $c_k$ | $\Delta^{k-4} c_4$ | $c'_{k-4} = \Delta^{k-4} \frac{c_4}{2^{k-3}}$ | $\Delta^{k-11} c'_{11}$ | $\frac{c''_{k-11} = \Delta^{k-11} c''_{11}}{2^{k-10}}$ |
|---|---|---|---|---|---|
| 0 | 1.00000 00000 | | | | |
| 1 | 0.20000 00000 | | | | |
| 2 | 0.08000 00000 | | | | |
| 3 | 0.04800 00000 | | | | |
| 4 | 0.03840 00000 | 0.03840 00000 | 0.01920 00000 | | |
| 5 | 0.03840 00000 | 0.00000 00000 | 0.00000 00000 | | |
| 6 | 0.04608 00000 | 0.00768 00000 | 0.00096 00000 | | |
| 7 | 0.06451 20000 | 0.00307 20000 | 0.00019 20000 | | |
| 8 | 0.10321 92000 | 0.00645 12000 | 0.00020 16000 | | |
| 9 | 0.18579 45600 | 0.00761 85600 | 0.00011 90400 | | |
| 10 | 0.37158 91200 | 0.01406 97600 | 0.00010 99200 | | |
| 11 | 0.81749 60640 | 0.02602 59840 | 0.00010 16640 | 0.00010 16640 | 0.00005 08320 |
| 12 | 1.96199 05536 | 0.05613 40416 | 0.00010 96368 | 0.00000 79728 | 0.00000 19932 |
| 13 | 5.10117 54394 | 0.13145 60410 | 0.00012 83750 | 0.00001 07654 | 0.00000 13457 |
| 14 | 14.28329 12303 | 0.33766 21486 | 0.00016 48741 | 0.00000 69955 | 0.00000 04372 |
| 15 | 42.84987 36909 | 0.93823 63791 | 0.00022 90616 | 0.00000 29320 | 0.00000 00916 |
| 16 | 137.11959 58109 | 2.80697 67610 | 0.00034 26485 | 0.00000 88515 | 0.00000 01383 |
| 17 | 466.20662 57571 | 8.98851 15364 | 0.00054 86152 | 0.00000 06344 | 0.00000 00050 |
| 18 | 1678.34385 27256 | 30.66826 95734 | 0.00093 59213 | 0.00001 45895 | 0.00000 00570 |
| 19 | 6377.70664 03573 | 111.03898 71074 | 0.00169 43205 | 0.00000 02917 | 0.00000 00006 |

The useful domain of the Euler transformation is limited to alternating series, since otherwise the magnitude of the differences increases rapidly. It may, however, be possible to convert a series of terms of uniform sign to an alternating series by a preliminary transformation, and then to apply the Euler transformation. One such device, also due to van Wijngaarden, is described in Modern Computing Methods (Ref. 37, p. 126).

Letting

$$f = \sum_{k=1}^{\infty} a_k \tag{131}$$

and, for $k = 1, 2, 3, \cdots$,

$$c_k = \sum_{j=0}^{\infty} 2^j a_{2^j k} \tag{132}$$

then

$$f = \sum_{k=1}^{\infty} (-1)^{k+1} c_k \tag{133}$$

provided either the series (Eq. 131) converges and $|a_k| \geqq |a_{k+1}|$ for $k = 1, 2, 3, \cdots$, or that a $K$ and a $c$ greater than 0 exist such that

$$|a_k| < K k^{-c-1} \quad (k = 1, 2, 3, \cdots)$$

The formal validity of the transformation can be justified by observing that

$$2c_{2k} = \sum_{j=0}^{\infty} 2^{j+1} a_{2^{j+1} k} = \sum_{j=0}^{\infty} 2^j a_{2^j k} - a_k = c_k - a_k \tag{134}$$

Hence

$$\sum_{k=1}^{\infty} a_k = \sum_{k=1}^{\infty} (c_k - 2c_{2k}) = \sum_{k=1}^{\infty} (-1)^{k+1} c_k \tag{135}$$

The identity (Eq. 134) can also be used to facilitate the computation of the even terms in the transformed series.

### I. The Euler–Maclaurin[1] Summation Formulas

In our discussion of bounding the truncation error of a series, we saw that useful upper and lower bounds for finite or infinite sums can be obtained in the form of integrals. The Euler–Maclaurin summation formulas represent an extension of this approach, and express the difference between the sum and the integral as a (usually divergent) series depending upon the derivatives of the integrand at the two ends of the interval.

Derivations and discussions of the Euler–Maclaurin summation formula can be found in many places, including Knopp (Ref. 26, pp. 522–527) and Hardy (Ref. 36, pp. 318–331). One of the most useful presentations for our needs is to be found in Steffensen (Ref. 41, pp. 129–138), and almost all the results we need may be found there.

Let $B_n$ denote the $n$th Bernoulli number, and $B_n(x)$ the $n$th Bernoulli polynomial. Let

$$\hat{B}_n(x) = B_n(x - [x]) \tag{136}$$

where $[x]$ is the largest integer contained in $x$. Let $f(x)$ have $m$ continuous derivatives on the interval $(0, n)$. Then, the general Euler–Maclaurin formula is

$$\sum_{k=0}^{n-1} f(k+\theta) = \int_0^n f(t)\,dt + \sum_{k=1}^{m} \frac{B_k(\theta)}{k!} \left[f^{(k-1)}(n) - f^{(k-1)}(0)\right] - \int_0^u \frac{B_m(\theta - t)}{m!} f^{(m)}(t)\,dt \tag{137}$$

for any $\theta$ between 0 and 1, and any finite $n$.

If both the sum and the integral

$$\left.\begin{aligned} S &= \sum_{k=0}^{\infty} f(k+\theta) \\ J &= \int_0^{\infty} f(t)\,dt \end{aligned}\right\} \tag{138}$$

[1]There are significant differences in notation and nomenclature associated with these formulas. Knopp (Ref. 26, p. 523) gives complete priority to Euler, and refers to the whole family as Euler summation formulas. Other authors use the name Euler–Maclaurin summation formula for the commonest special case, and the name Euler summation formula for the generalized forms. To avoid confusion with the multitude of other useful results named after Euler we will, without implying or denying codiscovery, refer to all formulas in this class as Euler–Maclaurin formulas.

converge and if

$$\lim_{n\to\infty} f^{(k-1)}(n) = 0 \qquad (k = 1, 2, \cdots, m) \tag{139}$$

then Eq. (137) holds for $n \to \infty$.

Setting $\theta = 0$ in Eq. (137), and observing that, for $k = 1, 2, \cdots, B_k(0) = B_k$, that $B_{2k+1} = 0$, while $B_1 = -\frac{1}{2}$, we obtain the most familiar special case:

$$\sum_{k=0}^{n-1} f(k) = \int_0^n f(t)\,dt - \frac{1}{2}[f(n) - f(0)] + \sum_{k=1}^{[m/2]} \frac{B_{2k}}{(2k)!}[f^{(2k-1)}(n) - f^{(2k-1)}(0)] + R_m \tag{140}$$

with

$$R_m = -\int_0^n \frac{\hat{B}_m(-t)}{m!} f^{(m)}(t)\,dt \tag{141}$$

If $m$ is even, say $m = 2r$, it may be shown (Steffensen, Ref. 41, pp. 132–133) using special properties of the Bernoulli polynomials, that

$$R_{2r} = -\frac{B_{2r}}{(2r)!}[f^{(2r-1)}(n) - f^{(2r-1)}(0)] + \frac{n B_{2r} f^{(2r)}(\tau)}{(2r)!} \qquad (0 \leq \tau \leq n) \tag{142}$$

The first term simply cancels the last term in the sum in Eq. (140), and so

$$\sum_{k=0}^{n-1} f(k) = \int_0^n f(t)\,dt - \frac{1}{2}[f(n) - f(0)] + \sum_{k=1}^{r-1} \frac{B_{2k}}{(2k)!}[f^{(2k-1)}(n) - f^{(2k-1)}(0)] + \tilde{R}_r \tag{143}$$

where, now,

$$\tilde{R}_r = \frac{nB_{2r}}{(2r)!} f^{(2r)}(\tau) \qquad (0 \leq \tau \leq n) \tag{144}$$

This expression can be rather inconvenient, particularly if $n$ is large and $f^{(2r)}(t)$ varies widely over $0 \leq t \leq n$. Simple bounds can, however, be obtained if $f^{(2r)}(t)$ does not change sign on $0 \leq t \leq n$. In this case, it can be shown that $\tilde{R}_r$ is of the same sign as the first omitted term in the series, and less than twice its magnitude (provided the term does not vanish). If, in addition, $f^{(2r+2)}(t)$, then, because of the alternation of sign of the Bernoulli numbers, $\tilde{R}_r$ and $\tilde{R}_{r+1}$ are of opposite sign. This implies that under these circumstances, $\tilde{R}_r$ is of the same sign as the first neglected term, and is less than it in magnitude provided the term does not vanish. Unlike the expression (Eq. 144) the bounds in terms of the first neglected term are valid, and convenient as $n \to \infty$.

In addition to its use as a summation formula, Eq. (143) may also be considered as a corrected form of the trapezoidal quadrature rule. The fact that the correction terms (but not necessarily $R_r$) are small if the odd derivatives are small at the ends of the interval, and vanish completely if they are equal there (e.g., if a periodic function is being integrated over an integral number of periods) will turn out to be very useful later.

An Euler–Maclaurin summation formula corresponding to the midpoint quadrature rule can also be developed. Letting $\theta = \frac{1}{2}$, and $m = 2r$ in Eq. (137), we can obtain, after some manipulation,

$$\sum_{k=0}^{n-1} f\left(k + \frac{1}{2}\right) = \int_0^n f(t)\,dt + \sum_{k=1}^{r-1} \frac{B_{2k}\left(\frac{1}{2}\right)}{(2k)!}[f^{(2k-1)}(n) - f^{(2k-1)}(0)] + \hat{R}_r \tag{145}$$

with

$$\hat{R}_r = n \frac{B_{2r}\left(\frac{1}{2}\right)}{(2r)!} f^{(2r)}(\tau) \qquad (0 \leq \tau \leq n) \qquad (146)$$

The formula can also be expressed in terms of the Bernoulli numbers, instead of the values of the Bernoulli polynomials at $x = \frac{1}{2}$, by the identity

$$B_k\left(\frac{1}{2}\right) = -\left(1 - \frac{1}{2^{(k-1)}}\right) B_k \qquad (k = 0, 1, 2, \cdots) \qquad (147)$$

Bounds analogous to those for $\tilde{R}_r$ can be found on $\hat{R}_r$ in terms of the first neglected term in the series. If it is only known that $f^{(2r)}(t)$ does not change sign on $0 \leq t \leq n$, then $\hat{R}_r$ is of the same sign as the first neglected term, and less than *three* times its magnitude, while if $f^{(2r)}(t)$ and $f^{(2r+2)}(t)$ do not change sign, and are both of the same sign on $0 \leq t \leq n$, the magnitude of the error is less than the first neglected term, both, of course, provided the term does not vanish.

Because of the rapid growth of the magnitude of the Bernoulli numbers, the series on the right side of Eq. (143) or Eq. (145) diverges for all but a very restricted class of functions $f$. In many cases, however, the error after a finite number of terms is tolerably small, or can be made so, and thus the Euler–Maclaurin summation formulas are among the most useful methods of summing series of positive terms. It is also possible, for certain functions, to sum the series using convergence-inducing transformations. Results of this type are given by Hardy (Ref. 36, pp. 341–345).

As a first illustration of the usefulness of the Euler–Maclaurin summation formulas, we will study its application to the gamma function, which we defined in Section II-D. For any $z$ with real part greater than 1, let $z = x + n$, with $n$ a positive integer less than $\mathscr{R}_e(z)$. Let

$$\left.\begin{aligned} f(t) &= \ln(x + t) \\ f^{(j)}(t) &= -(-1)^j (j-1)! (x+t)^{-j} \end{aligned}\right\} \qquad (148)$$

Then by the recurrence (Eq. 25),

$$\ln \Gamma(z) = \ln \Gamma(x) + \sum_{k=0}^{n-1} \ln(x + k) = \ln \Gamma(x) + \sum_{k=0}^{n-1} f(k) \qquad (149)$$

Applying Eq. (140),

$$\ln \Gamma(z) = \ln \Gamma(x) + \int_0^n \ln(x + t)\, dt - \frac{1}{2}[\ln(x + n) - \ln(x)] + \sum_{k=1}^{[m/2]} \frac{B_{2k}}{(2k)!}\left[\frac{(2k-2)!}{(x+n)^{2k-1}} - \frac{(2k-2)}{x^{2k-1}}\right] + R_m(x, n) \qquad (150)$$

Evaluating the integral,

$$\ln \Gamma(z) = \ln \Gamma(x) + (x + n)\ln(x + n) - (x - n) - x \ln(x + x) - \frac{1}{2}\ln(x + n) + \frac{1}{2}\ln x + \sum_{k=1}^{[m/2]} \frac{B_{2k}}{2k(2k-1)(x+n)^{2k-1}} - \sum_{k=1}^{[m/2]} \frac{B_{2k}}{2k(2k-1)x^{2k-1}} + R_m(x, n) \qquad (151)$$

Reintroducing the variable $z$, and collecting terms,

$$\ln \Gamma(z) = \left(z - \frac{1}{2}\right)\ln z - z + \sum_{k=1}^{[m/2]} \frac{B_{2k}}{2k(2k-1)z^{2k-1}} + \left\{\ln \Gamma(x) - \left(x - \frac{1}{2}\right)\ln(x + x) - \sum_{k=1}^{[m/2]} \frac{B_{2k}}{2k(2k-1)z^{2k-1}}\right\} + R_m(x, z) \qquad (152)$$

The first part of Eq. (152) is Stirling's formula, except for the constant term, $\frac{1}{2}\ln 2\pi$. It may be demonstrated by a variety of arguments (e.g., Knopp, Ref. 26, pp. 530–539; Hardy, Ref. 36, pp. 334–335) that the terms in braces and the remainder may be combined to give:

$$\ln\Gamma(z) = \left(z - \frac{1}{2}\right)\ln z - z + \sum_{k=1}^{r} \frac{B_{2k}}{2k(2k-1)} + \frac{1}{2}\ln 2\pi + \frac{1}{2k+1}\int_0^\infty \frac{B_{2k+1}(-t)}{(z+t)^{2k+1}}\,dt \tag{153}$$

### J. Evaluation of the Hurwitz Zeta Function by Euler–Maclaurin Summation

The Hurwitz zeta function, $\zeta(s,a)$, may be defined, for $\mathcal{R}_e\, s > 1$, and for $a \neq 0, -1, -2, \cdots$ by the series

$$\zeta(s,a) = \sum_{k=0}^{\infty} \frac{1}{(k+a)^s} \tag{154}$$

The Riemann zeta function, $\zeta(s)$, which we have mentioned as a useful comparison series, is the special case $\zeta(s,1)$. Its major properties are summarized in Whittaker and Watson (Ref. 3, pp. 265–280), and in Erdelyi, Magnus, Oberhettinger and Tricomi (Ref. 1, Vol. 1, pp. 24–27, 32–35). The monographs by Titchmarsh (Refs. 42 and 43) are devoted mainly to the Riemann special case, which has received by far the most study. The zeta functions are closely connected with the Bernoulli polynomials and numbers, and with the gamma function. Much study has also been motivated by the relation of the Riemann zeta function to the distribution of prime numbers.

Among the computationally significant properties of the zeta function are the recurrence:

$$\zeta(s,a+n) = \zeta(s,a) - \sum_{k=0}^{n-1} \frac{1}{(k+a)^s} \tag{155}$$

and the special values:

$$\zeta(-m,a) - = \frac{B_{m+1}(a)}{(m+1)} \qquad (m = 0,1,2,\cdots) \tag{156}$$

where $B_k(x)$ denotes the $k$th Bernoulli polynomial, and

$$\zeta(2m,1) = \zeta(2m) = \frac{(-1)^{m+1}(2\pi)^{2m} B_{2m}}{2(2m)!} \qquad (m = 1,2,\cdots) \tag{157}$$

Since it is apparent from Eq. (154) that

$$\lim_{\mathcal{R}_e(s)\to\infty} \zeta(s,1) = 1$$

the last result shows that the magnitude of the even Bernoulli numbers grows asymptotically as

$$\frac{2(2m)!}{(2\pi)^{2m}}$$

Although the series (Eq. 154) converges for $Re(s) > 1$, the rate is unsatisfactorily slow unless $Re(s) >> 1$. We will therefore develop an alternative method of computation based on the Euler–Maclaurin summation formula. Letting $f(t) = (t+a)^{-s}$, so that

$$f^{(j)}(t) = \frac{(-1)^j \Gamma(s+j)}{\Gamma(s)(t+a)^{s+j}} \tag{158}$$

we obtain, using Eq. (123) with $m = 2r+1$,

$$\sum_{k=0}^{n-1} \frac{1}{(k+a)^s} = \int_0^n \frac{dt}{(t+a)^s} - \frac{1}{2}\left[\frac{1}{(n+a)^s} - \frac{1}{a^s}\right] - \sum_{k=1}^{r} \frac{B_{2k}}{(2k)!}\frac{\Gamma(s+2k-1)}{\Gamma(s)}\left[\frac{1}{(n+a)^{s+2k-1}} - \frac{1}{a^{s+2k-1}}\right] + \frac{\Gamma(s+2r+1)}{\Gamma(s)\Gamma(2r+2)}\int_0^n \frac{\hat{B}_{2r+1}(-t)}{(t+a)^{s+2r+1}}\,dt \tag{159}$$

Both the integrals and the sum converge as $n \to \infty$, and so we can write:

$$\zeta(s,a) = \frac{1}{2a^s} + \frac{1}{(s-1)a^{s-1}} + \sum_{k=1}^{r} \frac{B_{2k}\Gamma(s+2k-1)}{(2k)!\,\Gamma(s)\,a^{s+2k-1}} + \frac{\Gamma(s+2r+1)}{\Gamma(s)\Gamma(2r+2)}\int_0^\infty \frac{B_{2r+1}(-t)}{(t+a)^{s+2r+1}}\,dt \tag{160}$$

Since all the derivatives of $f(t)$ are positive on $0 \leq t \leq \infty$, the remainder in truncating this series before any term is smaller than it in magnitude, and of the same sign.

We will use this formula to compute an 8-decimal value of $\zeta(2,1)$, which, from Eq. (157), is equal to $\pi^2/6 = (1.64493\,4067$ to 9 decimals). With this value of $s$, Eq. (160) becomes

$$\zeta(2,a) = \frac{1}{a} + \frac{1}{2a^2} + \sum_{k=1}^{r} \frac{B_{2k}}{a^{2k+1}} + \widetilde{R}_r \qquad (161)$$

The values of the terms $B_{2k}/a^{2k+1}$ are given in Table 4 for $a = 1, 2, 3,$ and 4. The term of minimum magnitude for $a = 1$ is the third, $+0.0238 \cdots$. Thus, only one-decimal accuracy can be obtained by applying the Euler–Maclaurin sum formula directly to the series. In fact, the series diverges so rapidly that a single application of the Euler transformation (Exercise 3.10-5) only adds one additional decimal. With larger values of $a$, however, the performance increases rapidly. For $a = 2$, the smallest term, the sixth, is 0.00003 5604; for $a = 3$, the ninth term is only $+0.000000047$, while for $a = 4$, the tenth term is $-0.0000000001$, and the magnitudes are still decreasing. Thus, we may obtain the desired 8-decimal accuracy by computing $\zeta(2,1)$. An alternative interpretation of this procedure is to say that we sum the first three, relatively rapidly converging, terms directly, and apply the transformation only to the tails. Except for the difficulty of programming a suitable decision rule, this strategy has much to recommend it for all forms of acceleration of convergence.

### K. Exercises

(1) Use the integral bound to obtain upper and lower bounds for the truncation error of the series (Eq. 76).

(2) The elliptic integral of the first kind is defined by

$$F(\phi|m) = \int_0^{\phi} \frac{d\theta}{[-m\sin^2\theta]^{1/2}} = \int_0^{\arcsin\phi} \frac{dt}{[(1-t^2)(1-mt^2)]^{1/2}} \qquad (162)$$

where $m$ is known as the parameter. When $\phi = \pi/2$, the integral is known as the complete elliptic integral of the first kind, and is denoted by $K(m)$. The quantity

$$m_1 = 1 - m \qquad (163)$$

is called the complementary parameter. The complete elliptic integral of the complementary parameter is denoted by $k'(m)$:

$$K^{1\prime}(m) = K(m_1) = K(1-m) = \int_0^{\pi/2} \frac{d\theta}{[10m, \sin^2\theta]^{1/2}} \qquad (164)$$

For fixed $m$, the value of $\phi$ for which $F(\phi|m) = u$ is known as the amplitude of $u$, $am(u|m)$. The basic

**Table 4. Computation of $\zeta(2,1) = 1.6449340668$ by Euler–Maclaurin sum formula**

| $k$ | $\frac{B_{2k}}{1^{2k+1}}$ | $\frac{B_{2k}}{2^{2k+1}}$ | $\frac{B_{2k}}{3^{2k+1}}$ | $\frac{B_{2k}}{4^{2k+1}}$ |
|---|---|---|---|---|
| 1 | +0.16666 66667 | +0.02083 33333 | +0.00617 28395 | +0.00260 41667 |
| 2 | −0.03333 33333 | −0.00104 16667 | −0.00013 71742 | −0.00003 25521 |
| 3 | +0.02380 95238* | +0.00018 60119 | +0.00001 08868 | +0.00000 14532 |
| 4 | −0.03333 33333 | −0.00006 51042 | −0.00000 16935 | −0.00000 01272 |
| 5 | +0.07575 75758 | +0.00003 69910 | +0.00000 04277 | +0.00000 00181 |
| 6 | −0.25311 35531 | −0.00003 08977* | −0.00000 01588 | −0.00000 00038 |
| 7 | +1.16666 66667 | +0.00003 56038 | +0.00000 00813 | +0.00000 00011 |
| 8 | −7.09215 68627 | −0.00005 41089 | −0.00000 00549 | −0.00000 00004 |
| 9 | +54.97117 79449 | +0.00010 48492 | +0.00000 00473* | +0.00000 00002 |
| 10 | −529.12424 24242 | −0.00025 23061 | −0.00000 00506 | −0.00000 00001 |
| $1/a$ | 1.00000 00000 | 0.50000 00000 | 0.33333 33333 | 0.25000 00000 |
| $1/2a^2$ | 0.50000 00000 | 0.12500 00000 | 0.05555 55556 | 0.03125 00000 |
| $\sum_{k=1}^{a-1} k^{-2}$ | 0.00000 00000 | 1.00000 00000 | 1.25000 00000 | 1.36111 11111 |
| $\Sigma^*$ | 1.63333 33334 | 1.64494 95653 | 1.64493 40428 | 1.64493 40668 |
| Error | +0.01160 07335 | −0.00001 54985 | +0.00000 00240 | 0.00000 00000 |

*The smallest term is indicated by an asterisk. The sum $\Sigma^*$ excludes this and the following terms.

Jacobian elliptic functions are defined by

$$\mathrm{sn}\,(u|m) = \sin \mathrm{am}\,(u|m) \tag{165}$$

$$\mathrm{cn}\,(u|m) = \cos \mathrm{am}\,(u|m) \tag{166}$$

$$\mathrm{dn}\,(u|m) = \{1 - m\,[\mathrm{sn}\,(u|m)]^2\}^{1/2} \tag{167}$$

These functions are doubly periodic in the complex plane, with periods $4K(m)$ and $4iK'(m)$.

Instead of the independent variables $u$ and $m$, it is often convenient to use the nome

$$q = e^{-\pi}\frac{K'(m)}{K(m)} \tag{168}$$

and the argument

$$v = \frac{\pi u}{2K(m)} \tag{169}$$

In terms of these variables, the Jacobian elliptic functions have the series expansions:

$$\mathrm{sn}\,(u|m) = \frac{2\pi}{m^{1/2}K(m)}\sum_{j=0}^{\infty}\frac{q^{j+1/2}}{1-q^{2j+1}}\sin(2j+1)v \tag{170}$$

$$\mathrm{cn}\,(u|m) = \frac{2\pi}{m^{1/2}K(m)}\sum_{j=0}^{\infty}\frac{q^{j+1/2}}{1+q^{2j+1}}\cos(2j+1)v \tag{171}$$

$$\mathrm{dn}\,(u|m) = \frac{\pi}{2K(m)} + \frac{2\pi}{K(m)}\sum_{j=1}^{\infty}\frac{q^{j}}{1+q^{2j}}\cos 2jv \tag{172}$$

Find bounds (in terms of $q$) for the error incurred in truncating each of these series after $n$th term.

(3) The error function, $\mathrm{erf}\,(x)$, is defined by the integral:

$$\mathrm{erf}\,(x) = \frac{2}{(\pi)^{1/2}}\int_0^x e^{-t^2}\,dt \tag{173}$$

and the complementary error function by

$$\mathrm{erfc}\,(x) = 1 - \mathrm{erf}\,(x) = \frac{2}{\pi^{1/2}}\int_x^{\infty} e^{-t^2}\,dt \tag{174}$$

For $x$ real and nonnegative,

$$0 \leq \mathrm{erf}\,(x) \leq 1 \qquad 0 \leq \mathrm{erfc}\,(x) \leq 1 \tag{175}$$

It can be shown by successive integration by parts that, for any $n \geq 0$

$$(\pi)^{1/2}xe^{x^2}\,\mathrm{erfc}\,(x) = \sum_{k=0}^{n-1}\frac{(2k)!\,(-1)^k}{2^k k!\,(2x^2)\,k} + \frac{1}{2}\frac{(-1)^n(2n)!}{2^{2n}n!}e^{x^2}\int_x^{\infty}\frac{e^{-t^2}}{t^{2m}}\,dt \tag{176}$$

It can be seen by examining the remainder that the series is asymptotic as $x \to \infty$, but diverges for all $x$.

(a) Show that for real $x$, the error in truncating the series before any particular term is of the same sign as that term, and smaller than it in magnitude.

(b) Evaluate $2(\pi)^{1/2}e^4\,\mathrm{erfc}\,(2)$ to 4 decimals using the series (Eq. 176) and the Euler transformation, repeated if necessary. The value to 7 decimals is 0.9053542.

(4) (Modern Computing Methods, Ref. 37, p. 126.) Use van Wijngaarten's transformation (Eqs. 115 and 116), to show that for the Riemann zeta function

$$\zeta(s) = \sum_{k=1}^{\infty}k^{-s} = \frac{1}{1-2^{1-s}}\sum_{k=1}^{\infty}(-1)^{k+1}k^{-s}\,(Re\,s > 1) \tag{177}$$

(5) Apply the Euler transformation to the terms of the Euler–Maclaurin transformation of the series for $\zeta(2,1)$ given in Table 4.

## IV. Power Series

Power series are by far the most widely known form of series expansion, both because of their analytical importance, and because of the ease with which they can

be manipulated. The analytical importance stems, to a large extent, from the fact that the existence of a convergent series expansion for a function in nonnegative powers of $(x - x_0)$ is a necessary and sufficient condition for the function to be analytic throughout the circle of convergence of the series. The ease of manipulation is largely due to the fact that the powers of $(x - x_0)$ form the simplest set of basis functions, $\phi_k(x)$ which is closed under multiplication.

Although we will concentrate our attention on power series in a single complex variable (power series in several variables are straightforward enough, but discouragingly complicated), we will include in the class of power series all expansions of the form

$$f(x) = -\sum_{k=-\infty}^{\infty} c_k (x - x_0)^k \tag{178}$$

We thus allow series which may be identified with Laurent expansions about a pole or essential singularity, as well as with conventional Taylor expansions. In many cases, however, the $c_k$ will all vanish for $k$ outside certain limits, e.g., for $k < 0$. The point $x_0$ is generally known as the center of the expansion.

In this section, we will be concerned with methods of obtaining power series expansion, including expressions for the truncation error, with methods of manipulating them, and finally, with assessing the utility of power series for computing function values.

### A. Taylor's Theorem

Taylor's theorem is undoubtedly the most familiar method of deriving ascending power series expansions. Although it is often clumsier than other ways of obtaining the same result, it provides a useful expression for the remainder, particularly when $x$ and $x_0$ are restricted to real values. Moreover, if $x$ is analytic in a region including $x$ and $x_0$, its (necessarily convergent) expansion in nonnegative powers of $x - x_0$ is uniquely determined, so that the remainder expression holds for a convergent power series obtained by any other means.

Let $c$ be a connected region in the complex plane containing $x$ and $x_0$, and let $f(x)$ and its first $n$ derivative be continuous on $c$. Let

$$R_n(x) = \frac{1}{(n-1)!} \int_{x_0}^{x} f^{(n)}(t)\,(x - t)^{n-1}\,dt \tag{179}$$

Then, by repeated integration by parts, it can be shown that

$$f(x) = \sum_{k=0}^{n-1} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k + R_n(x) \tag{180}$$

We will refer to this result as Taylor's formula. It holds regardless of whether or not the infinite series converges, provided only that the specified derivatives exist, and are continuous.

Since $(x - t)^{n-1}$ is of uniform sign on $(x_0, x)$, we may apply the mean value theorem to obtain Lagrange's form of the remainder:

$$R_n(x) = \frac{f^{(n)}(\xi)}{n!} (x - x_0)^n \qquad 0 \leq \frac{\xi - x_0}{x - x_0} \leq 1 \tag{181}$$

Applying the theorem differently

$$R_n(x) = \frac{f^{(n)}(\xi)\,(x - \xi)^{n-1}}{(n-1)!} \int_{x_0}^{x} dt$$

$$= \frac{f^{(n)}(\eta)\,(x - \eta)^{n-p}\,(x - x_0)^p}{(n-1)!} \qquad 0 \leq \frac{\xi - x_0}{x - x_0} \leq 1 \tag{182}$$

which is Canchy's form of the remainder. More generally, for any integer $p$, with $1 \leq p \leq n$,

$$R_n(x) = \frac{f^{(n)}(\eta)\,(x - \eta)^{n-p}}{(n-1)!} \int_{x_0}^{x} (x - t)^{p-1}\,dt$$

$$= \frac{f^{(n)}(\eta)\,(x - \eta)^{n-p}\,(x - x_0)^p}{p\,(n-1)!} \qquad 0 \leq \frac{\xi - x_0}{x - x_0} \leq 1 \tag{183}$$

The major inconvenience in using Taylor's formula to obtain power series expansions is the difficulty in determining the required derivatives. Even though these exist, they often become extremely complicated, and tedious to determine as the order increases, although systems for doing formal symbolic manipulation on the computer may do much to alleviate the problem. Schwatt (Ref. 29, pp. 1–4) devotes considerable attention to this question, and derives general expressions for the derivatives of any order of a number of special types of function, but the task of determining closed expressions for the higher derivatives of all but a very restricted set of functions is ordinarily impractical. For this reason, we will turn

next to a consideration of formal methods for manipulating power series.

### B. Algebraic Operations on Power Series

It is often more convenient to obtain power series expansions by suitable manipulation of known series than by other methods, such as direct application of Taylor's series. In this section we will outline the formal relations corresponding to the operations of addition, subtraction, multiplication, and division of power series, and indicate the conditions under which these formal operations preserve rigorous significance.

Let

$$\left.\begin{aligned} f(x) &\sim \sum_{k=0}^{\infty} a_k (x - x_0)^k \\ g(x) &\sim \sum_{k=0}^{\infty} b_k (x - x_0)^k \end{aligned}\right\} \tag{184}$$

denote power series corresponding, at least formally, to the functions $f(x)$ and $g(x)$, and let $\alpha$ and $\beta$ be constants. Then, formally,

$$\alpha f(x) + \beta g(x) \sim \sum_{k=0}^{\infty} (\alpha a_k + \beta b_k)(x - x_0)^k \tag{185}$$

By the discussion following Markoff's main rearrangement theorem, the correspondence is equality, rather than formal equivalence, if the series $f(x)$ and $g(x)$ are absolutely convergent. This condition is actually too stringent. If the series for $f(x)$ and $g(x)$ converge even conditionally, the sum (185) will converge to $\alpha f(x) + \beta g(x)$. Even more, if the series for $f(x)$ and $g(x)$ are asymptotic expansions to $N$ terms, the series (185) will be an asymptotic expansion of $\alpha f(x) + \beta g(x)$ to $N$ terms.

The terms in the formal product of two series can be arranged in a variety of ways. For power series, one of the most useful is the Cauchy product

$$f(x)\, g(x) \sim \sum_{k=0}^{\infty} \left\{ \sum_{j=0}^{k} a_j b_{k-j} \right\} (x - x_0)^k \tag{186}$$

The formal equivalence may be replaced by equality if the series for $f(x)$ and $g(x)$ converge absolutely. If the series for $f(x)$ and $g(x)$ only converge conditionally, the product series may diverge, although it is summable by the $C(1)$ process to the product $f(x)\, g(x)$. If the series $f(x)$ and $g(x)$ are asymptotically convergent to $N$ terms, the product series is also asymptotically convergent to $N$ terms.

The coefficients for the quotient of two power series cannot be expressed in as convenient a form. However, if we let

$$\frac{f(x)}{g(x)} \sim \sum_{k=0}^{\infty} c_k (x - x_0)^k \tag{187}$$

then, whenever $g(x)$ does not vanish in the interval $[x_0, x]$, we can write:

$$f(x) \sim \sum_{k=0}^{\infty} a_k (x - x_0)^k \sim g(x) \sum_{k=0}^{\infty} c_k (x - x_0)^k \sim \left\{ \sum_{k=0}^{\infty} b_k (x - x_0)^k \right\} \left\{ \sum_{k=0}^{\infty} c_k (x - y_0)^k \right\} \tag{188}$$

Using the Cauchy product

$$\sum_{k=0}^{\infty} a_k (x - x_0)^k \sim \sum_{k=0}^{\infty} \left\{ \sum_{j=0}^{k} c_j b_{k-j} \right\} (x - x_0)^k \tag{189}$$

Equating coefficients of equal powers of $(x - x_0)^k$, we obtain, for $k = 0, 1, 2, \cdots$,

$$\sum_{j=0}^{k} c_j b_{k-j} = a_k \tag{190}$$

and, since if $b_0 = 0$, $g(x_0)$ would vanish, contrary to our assumption, we obtain the recursive formula for the $c_k$:

$$c_k = \frac{1}{b_0} \left\{ a_k - \sum_{j=0}^{k-1} b_{k-j} c_j \right\} \tag{191}$$

The series (187) converges for a neighborhood of $x_0$ if the series for $f(x)$ converges absolutely.

To illustrate the utility of algebraic manipulation of power series we may use the function:

$$\frac{x}{e^x - 1} = \frac{x}{\sum_{k=1}^{\infty} \frac{x^k}{k!}} = \frac{1}{\sum_{k=0}^{\infty} \frac{x^k}{(k+1)!}} \sim \sum_{k=0}^{\infty} \frac{B_k}{k!} x^k \tag{192}$$

We write the right side as we do because the $B_k$ which appear there are, in fact, the Bernoulli numbers which we have already met several times. With $b_k = 1/(k+1)!$ and $a_0 = 1$, $a_k = 0$, $k > 0$, Eq. (191) becomes:

$$\frac{B_0}{0!} = 1 \frac{B_k}{k!} = -\sum_{j=0}^{k-1} \frac{1}{(k+1-j)!} \frac{B_j}{j!} \tag{193}$$

which may be rearranged in the neater form:

$$B_k = \frac{1}{k+1} \sum_{j=0}^{k-1} \binom{k+1}{j} B_j \tag{194}$$

Starting with $B_0 = 1$, we find, successively

$$\left.\begin{aligned} B_1 &= -\frac{1}{2}(B_0) = -\frac{1}{2} \\ B_2 &= -\frac{1}{3}\{B_0 + 3B_1\} = -\frac{1}{3}\left\{1 - \frac{3}{2}\right\} = \frac{1}{6} \\ B_3 &= -\frac{1}{4}\{B_0 + 4B_1 + 6B_2\} = -\frac{1}{4}\left\{1 - \frac{4}{2} + \frac{6}{6}\right\} = 0 \\ B_4 &= -\frac{1}{5}\{B_0 + 5B_1 + 10B_2 + 10B_3\} = -\frac{1}{5}\left\{1 - \frac{5}{2} + \frac{10}{6}\right\} = -\frac{1}{30} \end{aligned}\right\} \tag{195}$$

and so on.

We can use this result to obtain a power series for the cotangent. By Euler's formula,

$$\left.\begin{aligned} \cos x &= \frac{1}{2}(e^{ix} + e^{-ix}) \\ \sin x &= \frac{1}{2i}(e^{ix} - e^{-ix}) \end{aligned}\right\} \tag{196}$$

we find

$$\cot x = \frac{\cos x}{\sin x} = i\left(\frac{e^{ix} + e^{-ix}}{e^{ix} - e^{-ix}}\right) = i\left(\frac{e^{2ix} + 1}{e^{2ix} - 1}\right) \tag{197}$$

The last expression may be written as

$$\cot x = \frac{1}{2x}\left\{\frac{2ix}{e^{2ix} - 1} + \frac{2ixe^{2ix}}{e^{2ix} - 1}\right\}$$

$$= \frac{1}{2x}\left\{\frac{2ix}{e^{2ix} - 1} + \frac{-2ix}{e^{-2ix} - 1}\right\} \tag{198}$$

Each of the last two terms may be expanded by Eq. (192), and we obtain:

$$\cot x = \frac{1}{2x}\left\{\sum_{k=0}^{\infty} \frac{B_k}{k!}(2ix)^k + \sum_{k=0}^{\infty} \frac{B_k}{k!}(-1)^k (2ix)^k\right\} \tag{199}$$

Adding term by term, and simplifying

$$\cot x = \frac{1}{2x} \sum_{k=0}^{\infty} \frac{B_{2k} 2^{2k}}{(2k)!}(-1)^k x^{2k} \tag{200}$$

which is the desired series.

### C. Reversion of Power Series

The power series expansion of the inverse of a function with a known convergent power series expansion can be obtained by the process known as reversion. The problem of finding the expansion of $(x - x_0)$ in powers of $(y - y_0)$ where

$$y - y_0 = \sum_{k=1}^{\infty} a_k (x - x_0)^k \tag{201}$$

is discussed in Knopp (Ref. 26, pp. 184–188). Explicit expressions for the first seven coefficients in the series

$$x - x_0 = \sum_{k=1}^{\infty} c_k (y - y_0)^k \tag{202}$$

in terms of the $a_k$ are given in AMS 55 (Ref. 10, Eq. 3.6.25), while van Orstrand (Ref. 44) gives the first thirteen coefficients. We will treat the slightly more general problem of expressing the solution, $Y$, of the equation

$$F(Y) = G(X) \tag{203}$$

as a power series in $X - X_0$, where $F$ and $G$ are given as power series, and the root $Y_0$ of the equation for $X = X_0$ is known. The method is given by Thacher (Ref. 45) and a complete algorithm in Thacher (Ref. 46).

For the power series to define the functions $F$ and $G$, they must converge in the neighborhoods of $Y_0$ and $X_0$. We will further assume that the derivative $F'(Y_0)$ does not vanish. Under these circumstances, the variable changes:

$$\left.\begin{aligned} y &= Y - Y_0 \qquad & f(y) &= \frac{F(y + Y_0)}{F'(Y_0)} \\ x &= X - X_0 \qquad & g(x) &= \frac{G(x + X_0)}{F'(Y_0)} \end{aligned}\right\} \tag{204}$$

transform Eq. (186) to

$$f(y) = g(x) \tag{205}$$

where $f$ and $g$ have the convergent power series expansions:

$$\left.\begin{aligned} f(y) &= y + \sum_{j=2}^{\infty} b_j y^j \\ g(x) &= \sum_{k=1}^{\infty} a_k x^k \end{aligned}\right\} \tag{206}$$

Let us denote the coefficient of $x^k$ in the power series expansion of $y^j$ by $c_{k,j}$, so that

$$y^j = \sum_{k=j}^{\infty} c_{k,j} x^k \tag{207}$$

The required coefficients are the $c_{k,1}$. The form of the transformed equation requires that $y(0) = 0$, so that the $c_{k,j}$ with $k < j$ all vanish, while the existence and continuity of $g(x)$ in a neighborhood of $x = 0$ are insured by the implicit function theorem.

By the Cauchy product,

$$\begin{aligned} y^{j+1} = y\,y^j &= \left(\sum_{k=1}^{\infty} c_{k,1} x^k\right)\left(\sum_{k=j}^{\infty} c_{k,j} x^k\right) \\ &= \sum_{k=j+1}^{\infty} \left(\sum_{i=1}^{k-j} c_{i,1}\, c_{k-i,j}\right) x^k \end{aligned} \tag{208}$$

The $c_{k,j}$ thus obey the recurrence

$$c_{k,j+1} = \sum_{i=1}^{k-j} c_{i,1}\, c_{k-i,j} \qquad (k > j \geq 1) \tag{209}$$

But combining Eqs. (205)–(207), we find

$$\sum_{k=1}^{\infty} c_{k,1} x^k + \sum_{j=2}^{\infty} \sum_{k-j}^{\infty} b_j\, c_{k,j}\, x^j = \sum_{k=1}^{\infty} a_k x^k \tag{210}$$

Interchanging the order of summation in the double series, and equating coefficients of equal powers of $x$, we find for $k = 1, 2, \cdots$,

$$c_{k,1} = a_k - \sum_{j=2}^{k} b_j\, c_{k,j} \tag{211}$$

For $k = 1$, the summation vanishes, in accordance with the usual rule on finite sums with upper limit less than the lower limit, and so $c_{1,1} = a_1$. The coefficient $c_{k,j+1}$ in Eq. (209) is expressed in terms of coefficients with first subscript less than $k$, and so, for $k = 2, 3, \cdots$, we may use this recurrence to compute $c_{k,j}$ for $j = 2, 3, \cdots, k$. We now have all the data needed to compute $c_{k,1}$ using Eq. (211), after which we may proceed to the next larger value of $k$.

As an example, let us derive a power series for the solution of the equation

$$Y^Y = X \tag{212}$$

knowing the particular solution $Y_0 = X_0 = 1$. Letting $y = Y - 1$, $x = X - 1$, and taking logarithms, we find

$$\begin{aligned} (1 + y)\ln(1 + y) = y + \sum_{j=2}^{\infty} \frac{(-1)^j}{j(j-1)} y^j &= \ln(1 + x) \\ &= \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} x^k \end{aligned} \tag{213}$$

Using Eqs. (209) and (211), we obtain the partial matrix of $c_{k,j}$ presented in Table 5. Because of the branch point of $Y(X)$ at $X = e^{-1/e} = 0.6922 \cdots$, the series converges only within a circle of radius $0.3077 \cdots$. However, the convergence can be improved by various transformations to give useful results. Even for $X = 3$, two applications of the Euler transformation to the first 7 terms of the series give $y = 0.960$, compared to the correct value of 1.000.

### D. Derivation of Power Series by Analytic Manipulation

Power series are particularly well adapted to term-by-term differentiation and integration, since the derivative and integral of a single power are again single powers with simple constant multipliers. Thus, when the function to be expanded is the derivative or integral of a function with a known series expansion, one may obtain a formal series for the new function by differentiating or integrating the individual terms of the original series, and each coefficient of the new series will depend on only one of the original coefficients. The validity of this formal procedure must be justified by the general criteria of Section 3-F.

As a simple illustration of this approach, let us consider the inverse sine,

$$\arcsin x = \int_0^x (1 - t^2)^{-1/2}\, dt \tag{214}$$

By the binomial theorem,

$$(1 - t^2)^{-1/2} = \sum_{k=0}^{\infty} \binom{-1/2}{k} (-t^2)^k$$

$$= 1 + \sum_{k=1}^{\infty} \prod_{j=0}^{k-1} \left(-\frac{1}{2} - j\right) \frac{(-1)^k}{k!} t^{2k} \tag{215}$$

which can be written as:

$$(1 - t^2)^{-1/2} = 1 + \sum_{k=1}^{\infty} \prod_{j=0}^{k-1} \frac{2j+1}{2j+2} t^{2k} \tag{216}$$

Now integrating

$$\arcsin x = \int_0^x dt + \sum_{k=1}^{\infty} \prod_{j=0}^{k-1} \frac{2j+1}{2j+2} \int_0^x t^{2k}\, dt$$

$$= x + \sum_{k=1}^{\infty} \prod_{j=0}^{k-1} \frac{(2j+1)}{2j+2} \frac{t^{2k+1}}{2k+1} \tag{217}$$

The relative case of this derivation compared to the use of Taylor's theorem will be readily apparent to anyone who attempts to determine the high-order derivatives of $\arcsin x$ at $x = 0$.

A variation of this approach is particularly convenient when the function is known to satisfy a linear differential or integral equation with polynomial coefficients. This consists of assuming the existence of a power series with coefficients which are as yet undetermined. Term-by-term differentiation and integration of this series then yields formal power series for the derivatives and integrals of the function. If we introduce these series into the functional equation, and collect equal powers of the independent variable, we obtain an infinite set of linear equations, one for each power of the independent variable. If the system of equations is inadequate to determine the coefficients uniquely, the initial conditions will ordinarily supply additional constraints to make the problem determinate.

To illustrate this procedure, let us determine the binomial expansion. Let

$$f(x) = (1 + x)^r \tag{218}$$

Differentiating,

$$f'(x) = r(1 + x)^{r-1} = \frac{rf(x)}{1 + x} \tag{219}$$

so that $f(x)$ satisfies the ordinary differential equation:

$$(1 + x) f'(x) - rf(x) = 0 \tag{220}$$

Assume, subject to later verification, that $f(x)$ has the power series expansion, convergent uniformly in some neighborhood of the origin

$$f(x) = \sum_{k=0}^{\infty} a_k x^k \tag{221}$$

Then

$$f'(x) = \sum_{k=0}^{\infty} k a_k x^{k-1} \tag{222}$$

and our differential equation becomes:

$$\sum_{k=0}^{\infty} k a_k x^{k-1} + \sum_{k=0}^{\infty} k a_k x^k - \sum_{k=0}^{\infty} r a_k x^k = 0 \tag{223}$$

which simplifies to

$$\sum_{k=0}^{\infty} [(k+1) a_{k+1} + (k - r) a_k] x^k = 0 \tag{224}$$

**Table 5. Reversion of series for $(1 + y)\ln(1 + y) = \ln(1 + x)$**

| $k$ \ $j$ | $C_{k,j}$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | 1 | | | | | | |
| 2 | $-1$ | 1 | | | | | |
| 3 | $\frac{3}{2}$ | $-2$ | 1 | | | | |
| 4 | $-\frac{17}{6}$ | 4 | $-3$ | 1 | | | |
| 5 | $\frac{37}{6}$ | $-\frac{26}{3}$ | $\frac{15}{2}$ | $-4$ | 1 | | |
| 6 | $-\frac{1759}{120}$ | $\frac{81}{4}$ | $-\frac{37}{2}$ | 12 | $-5$ | 1 | |
| 7 | $\frac{13279}{360}$ | $-\frac{1003}{20}$ | $\frac{187}{4}$ | $-\frac{100}{3}$ | $\frac{35}{2}$ | $-6$ | 1 |

Since this equation must hold uniformly for some neighborhood of the origin, the coefficient of each power of $x$ must vanish individually, and the coefficients must satisfy the recurrences:

$$(k+1)\,a_{k+1} + (k-r)\,a_k = 0 \tag{225}$$

or

$$a_{k+1} = \left(\frac{k-r}{k+1}\right) a_k \qquad (k = 0, 1, 2, \cdots) \tag{226}$$

This recurrence will produce a set of coefficients satisfying Eq. (220) for arbitrary $a_0$. To determine the appropriate set, we must use additional information. It is apparent from Eq. (218) that $f(0) = 1$ for all $r$. Thus the appropriate value of $a_0$ is 1. We may now observe that since

$$\lim_{k\to\infty} \frac{k-r}{k+1} = 1$$

the series (221) converges uniformly by the ratio test for $|x| < 1$.

The solution to Eq. (226) can be expressed explicitly:

$$a_{k+1} = \frac{\prod_{j=0}^{k} (r-j)}{\prod_{j=0}^{k} (j+1)}\, a_0 = \frac{r!}{(r-k-1)!\,(k+1)!} \tag{227}$$

which is, of course, the well-known expression for the binomial coefficients. However, for computational purposes, this is not particularly useful, since it is ordinarily easier to generate the coefficients from the recurrence than from the closed form.

When the recurrence involves three or more terms, as will often be the case, the stability of the recurrence must be considered. The problem is entirely analogous to that involved in evaluating functions from their recurrence, which we will consider later.

It is often very convenient to proceed by way of the differential equation when it is necessary to change the dependent or independent variable in a power series expansion. Independent variable changes are easily made in the differential equation using the chain rule:

$$\frac{d}{dx} f[z(x)] = \frac{df}{dz} \cdot \frac{dz}{dx} \tag{228}$$

subject, of course, to the requirement that $z$ be differentiable. After eliminating the original variable from the coefficients of the differential equation, the method of undetermined coefficients then gives a recurrence from which the coefficients of the new series may readily be computed.

As an example, we may consider Dawson's function,

$$F(x) \equiv e^{-x^2} \int_0^x e^{t^2}\, dt \tag{229}$$

Except for the factor $i(\pi)^{1/2} e^{-x^2}/2$, Dawson's function is the same as the error function of pure imaginary argument. It occurs in computing resonance absorption both of light and of neutrons.

Differentiating Eq. (229), we obtain

$$\frac{dF}{dx} = -2xe^{-x^2} \int_0^x e^{t^2}\, dt + e^{-x^2} e^{x^2} = -2xF(x) + 1 \tag{230}$$

Now let us expand this function as a power series in $z = \alpha + \beta x$, where $\alpha$ and $\beta$ are parameters. Applying the chain rule and simplifying, we find:

$$\beta^2 \frac{dF}{dz} + 2(z - \alpha) F = \beta \tag{231}$$

Now letting $F(z)$ and $dF/dz$ have the formal power series expansions

$$F(z) = \sum_{k=0}^{\infty} c_k z^k \qquad \frac{dF}{dz} = \sum_{k=0}^{\infty} (k+1) c_{k+1} z^k \tag{232}$$

our differential equation becomes:

$$\beta^2 \sum_{k=0}^{\infty} (k+1) c_{k+1} z^k + 2(z - \alpha) \sum_{k=0}^{\infty} c_k z^k = \beta \tag{233}$$

or,

$$\beta^2 c_1 - 2\alpha c_0 + \sum_{k=1}^{\infty} [\beta^2 (k+1) c_{k+1} - 2\alpha c_k + 2c_{k-1}] z^k = \beta \tag{234}$$

The $c_k$ must thus satisfy the equations:

$$\beta^2 c_1 - 2\alpha c_0 = \beta \tag{235}$$

$$c_{k+1} = \frac{2}{\beta^2 (k+1)} (\alpha c_k - c_{k-1}) \qquad (k = 1, 2, 3, \cdots) \tag{236}$$

As might be expected from the fact that we have neglected any boundary conditions on our differential equation, these conditions are inadequate to determine the $c_k$ completely. A solution can be constructed for any value of $c_0$. To fix $c_0$, we observe that $z(\alpha) = 0$, and hence that $c_0 = F(\alpha)$. Thus, if we have an accurate value of $F(x)$ at the center of the expansion, we may construct the whole expansion. In particular, for $\alpha = 0$, it is clear from Eq. (229) that $F(0) = 0$, and, setting $\beta = 1$, Eqs. (235) and (236) become:

$$c_1 = 1 \qquad c_{k+1} = \frac{-2}{k+1} c_{k-1} \tag{237}$$

Only odd terms appear in the expansion, with

$$c_{2k+1} = (-1)^k \frac{2}{3} \cdot \frac{2}{5} \cdots \frac{2}{2k+1} = (-1)^k \frac{2^{2k} (k+1)!}{(2k+2)!} \tag{238}$$

The same results might, of course, be obtained using Taylor's theorem, and evaluating the derivatives by Eq. (231).

The linear transformation, $z = \alpha + \beta x$, which we have used as an example, represents merely a change of variable and scale. More general transformations, such as the linear fractional transformation, $z = (\alpha + \beta x)/(\gamma + \gamma x)$, may be introduced by the same method, although with somewhat more complicated algebra. Even the linear transformation should not be neglected, since it can result in significant economies in computation. It should be observed, however, that the accuracy of the transformed series may be quite sensitive to the accuracy of the initial values. Thus extreme care, including multiple precision calculations may be required for computing $F(\alpha)$. Since this task need only be performed once, this does not reduce the advantages of the transformation significantly.

### E. Singularities and the Convergence of Power Series

The convergence of power series can, of course, be studied by all the methods applicable to series in general. A far more instructive approach is possible, however, for ascending power series, using the properties of the corresponding function in the complex plane. A power series

$$f(x) = \sum_{k=0}^{\infty} c_k (x - x_0)^k \tag{239}$$

can be shown, by methods of the theory of functions of a complex variable, to converge uniformly in a region of the complex plane specified by $|x - x_0| < R$, that is, in a circle of radius $R$ with center at $x_0$. Further, the maximum $R$ for which this holds is the distance from $x_0$ to the nearest singularity of $f(x)$ in the complex $x$-plane. A knowledge of the singularities of $f(x)$ is thus of considerable assistance in choosing an effective power series expansion. Although, as we have repeatedly emphasized, convergence is neither a necessary nor a sufficient condition for the computational value of an expansion, ascending power series rarely converge usefully near the edge of their circle of convergence. In choosing a power series for computing a function over a specified region, it is thus desirable that the region lie as close to the center of the circle of convergence as possible.

One obvious way of accomplishing this goal is to transform the independent variable so that the center of the expansion lies near the center of the region. The linear transformation $z = \alpha + \beta x$, discussed in the last section,

is the simplest such transformation. To illustrate its effectiveness, suppose that we wish to compute $\ln x$ on the interval $1 \leq x \leq 2$, with a precision of 3 decimals. The familiar series

$$\ln x = -\sum_{k=1}^{\infty} \frac{(1-x)^k}{k} \qquad (0 < x \leq 2) \qquad (240)$$

converges over the desired domain, but for $x = 2$, 1000 terms are necessary for the required precision. The poor convergence of this series is due to the branch point singularity of $\ln x$ at the origin. The point $x = 2$ lies on the boundary of the circle of convergence, as shown in Fig. 1a, while the center of the circle of convergence lies at one end of the domain of interest.

We are led to investigate transformations of the independent variable which will improve this situation. Two approaches are possible. We may find transformations which reduce the size of the domain relative to the circle of convergence, and transformations which tend to center the domain within the circle. A transformation of the first type is the change of variable $z = 1 - 1/x$, which leads to the series

$$\ln x = -\ln(1-z) = \sum_{k=1}^{\infty} \frac{z^k}{k} = \sum_{k=1}^{\infty} \frac{\left(1 - \frac{1}{x}\right)^k}{k} \quad (|z| < 1) \qquad (241)$$

This series converges for $1/2 \leq x < \infty$. Moreover, the desired domain runs from $z = 0$ to $z = 1/2$, as shown in Fig. 1b, and even for $x = 2$, only 7 terms are required to give the specified precision.

The transformation $z = (5x - 8)/(2x - 8)$ centers the domain of interest within the circle of convergence without reducing its size, as can be seen from Fig. 1c. This transformation leads to the series

$$\ln x = \ln \frac{\frac{8}{5} - \frac{8z}{5}}{1 - \frac{2z}{5}} = \ln \frac{8}{5} + \ln(1-z) - \ln\left(1 - \frac{2z}{5}\right) \qquad (242)$$

$$= \ln \frac{8}{5} + \sum_{k=1}^{\infty} \frac{\left(\frac{2}{5}\right)^k - 1}{k} \left(\frac{5x-8}{2x-8}\right)^k \qquad (243)$$

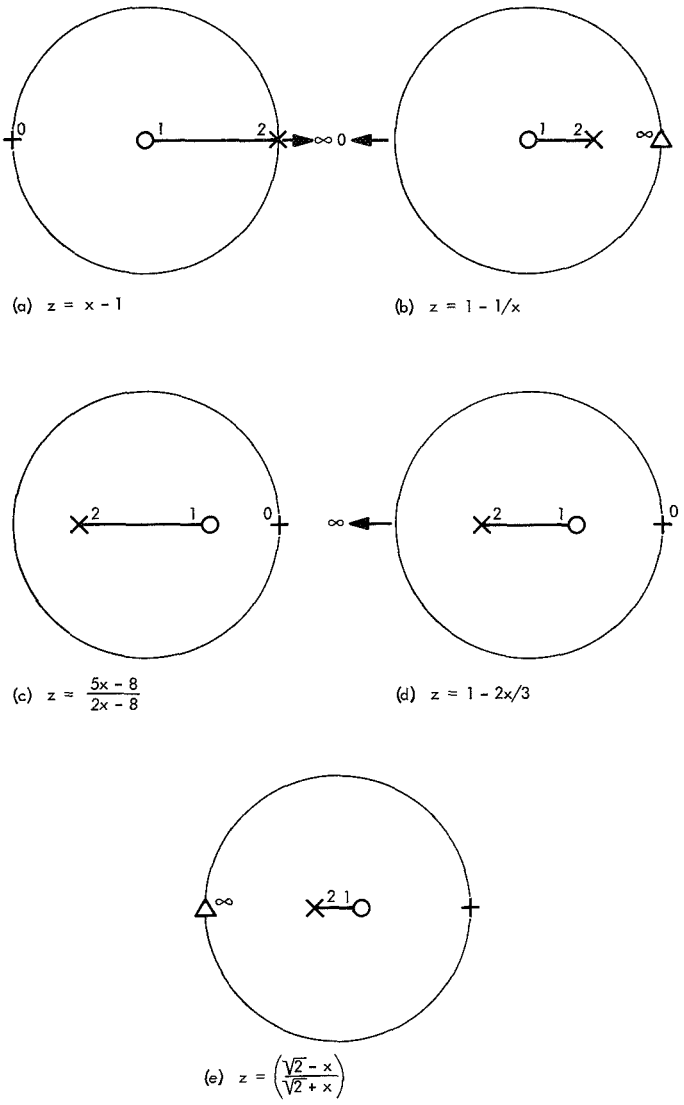


**Fig. 1. Circles of convergence**

This series converges for $0 < x < 16/7$, and, like Eq. (241), requires only 7 terms to attain an error of less than 0.001.

An even simpler transformation, $z = 1 - 2x/3$, both centers the domain, and reduces its relative size, as can be seen in Fig. 1d. The corresponding series

$$\ln x = \ln \frac{3}{2} - \sum_{k=1}^{\infty} \frac{\left(1 - \frac{2x}{3}\right)^k}{k} \qquad (244)$$

converges for $0 < x \leq 3$, and requires only 5 terms to reduce the error to less than 0.001.

On considering Fig. 1, it is apparent that the convergence of the series (240), (243), and (244) is limited by the singularity at the origin, while the limitation for Eq. (241) is due to the singularity at $\infty$. A transformation which leads to a limitation by both these singularities might be expected to be particularly effective. The transformation $z = (1 - \beta x)/(1 + \beta x)$ accomplishes this goal, mapping the origin onto $z = 1$, and the point at infinity onto $z = -1$. If, in addition, we give $\beta$ the value $(2)^{1/2}/2$, the transformation also centers the image of $1 \leq x \leq 2$ in the circle of convergence, as shown in Fig. 1e. The series we obtain,

$$\ln x = \ln (2)^{1/2} - \sum_{k=0}^{\infty} \frac{2}{2k+1} \left( \frac{(2)^{1/2} - x}{(2)^{1/2} + x} \right)^{2k+1} \tag{245}$$

converges for $0 < x < \infty$, and converges extremely rapidly for the domain we specified. Terms through $z^3$ give an error of less than 0.0001, while terms through $z^9$ give full 8-decimal accuracy. In addition, because of the symmetry, we have the additional benefit that all the even powers of $z$ drop out, so that only half the number of terms appear.

A transformation of the dependent variable may also reduce the difficulties encountered near a singularity. The change consists of expressing $f$, the function to be computed as either the sum or the product of a function, $\phi$, say, with same type of singularity, and a function, $g$, which is analytic over a wider domain than $f$. The device is only useful, of course, if $\phi$ may be computed by some method which does not depend upon its being analytic. Logarithms, roots, and exponentials have this property.

There are many examples of useful transformations of this kind. Both the subtraction and factoring of singularities may be illustrated by considering the exponential integral, $E_1(x)$, defined by

$$E_1(x) \equiv \int_x^{\infty} \frac{e^{-t}}{t} dt \qquad (|\arg x| < \pi) \tag{246}$$

This function has a singularity (actually, as we shall see, a logarithmic branch point) at the origin, and an essential singularity at $\infty$. For any $\epsilon > 0$, we may write the integral in the form

$$\begin{aligned} E_1(x) &= -\int_{\epsilon}^{x} \frac{e^{-t}}{t} dt + \int_{\epsilon}^{\infty} \frac{e^{-t}}{t} dt \\ &= \int_{\epsilon}^{x} \frac{1 - e^{-t}}{t} dt - \int_{\epsilon}^{x} \frac{dt}{t} + \int_{\epsilon}^{\infty} \frac{e^{-t}}{t} dt \end{aligned} \tag{247}$$

$$\begin{aligned} &= \int_{\epsilon}^{x} \frac{1 - e^{-t}}{t} dt - \ln x - \int_{\epsilon}^{1} \frac{dt}{t} \\ &\quad + \int_{\epsilon}^{1} \frac{e^{-t}}{t} dt + \int_{1}^{\infty} \frac{e^{-t}}{t} dt \end{aligned} \tag{248}$$

Letting $t = 1/u$ in the last integral, and combining terms, we find

$$\begin{aligned} E_1(x) = &\int_{\epsilon}^{x} \frac{1 - e^{-t}}{t} dt - \ln x \\ &- \int_{\epsilon}^{1} \frac{1 - e^{-t} - e^{-1/t}}{t} dt + \int_{0}^{\epsilon} \frac{e^{-1/t}}{t} dt \end{aligned}$$

As $\epsilon \to 0$, the second integral converges, and in fact (Ref. 3, p. 243, Ex. 4) converges to Euler's constant, $\gamma$. The last integral, of course, vanishes, and we have the result:

$$E_1(x) = \int_0^{x} \frac{1 - e^{-t}}{t} dt - \ln x - \gamma \tag{249}$$

The remaining integral is frequently referred to as $Ein(x)$. It is readily expanded in powers of $x$, by expanding the exponential in Maclaurin series and integrating term-by-term. We find

$$Ein(x) = -\sum_{k=1}^{\infty} \frac{(-x)^k}{k!\,k} \tag{250}$$

Since each term of this series is smaller than the corresponding term of the exponential series, which converges for all finite $x$, $Ein(x)$ has no singularities in the finite complex plane.

It is not possible to eliminate the singularity at $\infty$ completely, but a considerable improvement in the behavior may be obtained by writing

$$E_1(x) = e^{-x} \int_x^{\infty} \frac{e^{-(t-x)}}{t} dt = e^{-x} \int_0^{\infty} \frac{e^{-u}\,du}{u + x} = \frac{e^{-x}}{x} \int_0^{\infty} \frac{e^{-u}\,du}{1 + \frac{u}{x}} \tag{251}$$

Now for any $n$ and all $y$,

$$\frac{1}{1+y} \equiv \sum_{k=0}^{n-1} (-y)^k + \frac{(-y)^n}{1-y} \tag{252}$$

Introducing this identity in Eq. (251) and interchanging summation and integration, which is justified since the

integrals converge,

$$E_1(x) = \frac{e^{-x}}{x}\left\{\sum_{k=0}^{n-1}\frac{\int_0^\infty u^k e^{-u}\,du}{(-x)^k} + \frac{1}{(-x)^n}\int_0^\infty \frac{u^n e^{-u}}{1+\frac{u}{x}}\,du\right\} \tag{253}$$

Using the standard definite integral

$$\int_0^\infty u^k e^{-u}\,du = \Gamma(k+1) = k! \tag{254}$$

this reduces to

$$E_1(x) = \frac{e^{-x}}{x}\left\{\sum_{k=0}^{n-1}\frac{k!}{(-x)^k} + \frac{n!}{(-x)^n}\frac{1}{n!}\int_0^\infty \frac{u^n e^{-u}}{1+\frac{u}{x}}\,du\right\} \tag{255}$$

The ratio of the magnitudes of successive terms in the sum is $(k+1)/x$, and is greater than 1 for all $k \geqq x$. Thus, the series obtained by letting $n \to \infty$ in Eq. (255) diverges for all $x$. However, for $x > 0$, the converging factor

$$C_n(x) \equiv \frac{1}{n!}\int_0^\infty \frac{u^n e^{-u}}{1+\frac{u}{x}}\,du \tag{256}$$

is positive for all $n$, and since $1 + u/x > 1$,

$$C_n(x) \leqq \frac{1}{n!}\int_0^\infty u^n e^{-u}\,du = 1 \tag{257}$$

Thus, the error in neglecting the remainder in Eq. (255) is of the same sign as $n!/(-x)^n$, and smaller in magnitude. For fixed $n$, the error thus vanishes as $x \to \infty$, and our series, although divergent everywhere, is a valid asymptotic representation of $E_1(x)$. A brief calculation reveals, also, that $n!/n^n$ decreases rapidly with $n$, so that for $n = 20$, almost $8D$ accuracy is obtained, and for $n = 30$, almost $12D$.

### F. Laurent Series

Although it is usually advantageous to remove singularities by factorization or subtraction, as outlined in the last section, it is also possible to construct power series expansions which take account of singularities at or near the center of expansion by including both positive and negative powers. The basis of such expansions is Laurent's theorem, which is proved in most text books on complex analysis (e.g., Ref. 3, p. 100), or Hille (Ref. 47, pp. 209–211). This theorem may be stated as follows:

Let $f(x)$ be analytic in the annulus

$$\rho_j < r \leqq |x - x_0| \leqq R < \rho_{j+1} \tag{258}$$

Then, for any point in the interior of this annulus, $f(x)$ has the absolutely convergent power series expansion

$$f(x) = \sum_{k=-\infty}^{\infty} c_k (x - x_0)^k \tag{259}$$

where the $c_k$ may be represented by the contour integrals

$$c_k = \frac{1}{2\pi i}\oint \frac{f(x)}{(x-x_0)^{k+1}}\,dx \tag{260}$$

The contour of integration must lie entirely within the annulus, and circle the point $x_0$ once in the positive direction.

We will refer to the (open) annulus $\rho_j < |x - x_0| < \rho_{j+1}$ as the $j$th annulus of convergence for $f(x)$ if $\rho_j$ has the smallest possible value such that $f(x)$ is analytic throughout the annulus, and $\rho_{j+1}$ has the largest possible value.

Then the circles $|x - x_0| = \rho_j$ and $|x - x_0| = \rho_{j+1}$ must each contain at least one singularity of $f(x)$. The series for $f(x)$ is unique for each annulus of convergence, although different annuli will normally have different associated series.

Leaving aside questions of speed of convergence, which will normally favor removal of singularities, Laurent series have several other computational defects. First, the requirement that $f(x)$ be analytic in an entire annulus prevents their use for most multiple valued functions. Second, the contour integrals which define the coefficients are not, ordinarily, easy to evaluate numerically. In fact, a major application of Laurent's theorem is to the evaluation of contour integrals which may be interpreted as coefficients of a Laurent series. If the function may be expanded by some other technique, the uniqueness of the Laurent expansion allows us to give values to the integral appearing in Eq. (260). This second drawback may, however, be less serious than has previously been considered in view of recent investigations by Lyness (Ref. 48) and collaborators, which have demonstrated that numerical evaluation of contour integrals is a practical method of estimating many quantities of interest in numerical analysis.

## V. Continued Fractions

In spite of their convenience, familiarity, and other advantages, the flexibility of infinite series in representing functions is severely limited. The convergence of power series is limited by the nearest singularity to the center of convergence, while the singularities which can appear in the domain in which functions are adequately represented by series of other types are limited to those appearing in the basis functions. Continued fractions avoid many of these difficulties. Not only do they provide globally convergent representations of meromorphic functions, but the regions of divergence for functions with branch points or essential singularities is typically significantly smaller than for series. Further, the numerical behavior of continued fractions is often better than that of the corresponding series.

Unfortunately, the theory of continued fractions is considerably more complicated than that of series, and is far less familiar to most mathematicians and scientists. The major reference works on these expansions are Khovanskii (Ref. 17), Wall (Ref. 49), and Perron (Ref. 50), while the review by Blanch (Ref. 51) contains a summary of the major computationally important properties, as well as a careful discussion of computational pitfalls.

### A. Definitions and Notation

The rational expression

$$f_n = q_0 + \cfrac{p_1}{q_1 + \cfrac{p_2}{q_2 + \cfrac{\ddots}{\ddots + \cfrac{p_n}{q_n}}}} \tag{261}$$

where the sets $\{p_k\}, \{q_k\}$ are real or complex quantities, possibly functions of one or more independent variables, is known as a *finite continued fraction.* To save space, we will hereafter write such fractions in the form:

$$f_n = q_0 + \frac{p_1}{q_1 +}\,\frac{p_2}{q_2 +}\,\cdots\,\frac{p_n}{+\,q_n} \tag{262}$$

The $\{p_k\}$ are known as the *partial numerators*, and the $\{q_k\}$ as the *partial denominators* of the continued fraction, while the $\{p_k\}$ and $\{q_k\}$ are referred to collectively as its *elements.* The fractions $p_k/q_k$ are called *partial quotients.*

For finite $n$, we may reduce this expression to a simple ratio, by setting $r_n = q_n$, and computing, successively, $r_{n-1}, r_{n-2}, \cdots, r_0$ by the recurrence:

$$r_{k-1} = q_{k-1} + \frac{p_k}{r_k} \tag{263}$$

Then, $f_n = r_0$. Thus:

$$f_0 = q_0 \tag{264}$$

$$f_1 = q_0 + \frac{p_1}{q_1} = \frac{q_1 q_0 + p_1}{q_1} \tag{265}$$

$$f_2 = q_0 + \frac{p_1}{q_1 +}\,\frac{p_2}{q_2} = q_0 + \frac{p_1}{\dfrac{(q_2 q_1 + p_2)}{q_2}}$$

$$= \frac{(q_2 q_1 + p_2) q_0 + q_2 p_1}{(q_2 q_1 + p_2)} \tag{266}$$

and so on. If the algorithm is carried out numerically, it will fail if any of the $r_k$ other than $r_0$ happens to vanish. In many cases, however, the difficulty is only formal,

and disappears if the manipulations are performed symbolically before substituting numerical values for the elements.

If the sets of elements $\{p_k\}$ and $\{q_k\}$ are (denumerably) infinite, we may define the *infinite continued fraction*

$$f = q_0 + \frac{p_1}{q_1 +} \frac{p_2}{q_2 +} \cdots \tag{267}$$

as the limit of the sequence of *convergents*, $f_n$, obtained by truncating the expansion after the $n$th partial quotient. The $n$th convergent may be expressed as a simple ratio

$$f_n = \frac{P_n}{Q_n} \tag{268}$$

where the $n$th *numerator*, $P_n$, and the $n$th *denominator*, $Q_n$, are polynomials in the elements.

The convergence properties of continued fractions are determined by the behavior of the sequence of convergents, following the general terminology of Section II-A. Establishing the convergence or divergence of a continued fraction may be a very difficult task. Some important special cases where it is possible will be given in this section.

### B. Forward Recurrence for Numerators and Denominators

In the last section, we saw that a finite continued fraction can always be converted to a simple ratio by starting at the highest order partial quotient and applying the recurrence (263). If the ratio corresponding to a different convergent is required using this backward recurrence method, the whole process must be repeated. It is, however, possible to develop successive numerators, $P_k$, and denominators, $Q_k$, in increasing order.

We observe, first of all, that the first three convergents given by Eqs. (264)–(266) can be written in the form:

$$f_0 = \frac{P_0}{Q_0} = \frac{q_0(1) + 1(0)}{q_0(0) + 1(1)} \tag{269}$$

$$f_1 = \frac{P_1}{Q_1} = \frac{q_1(P_0) + p_1(1)}{q_1(Q_0) + p_1(0)} \tag{270}$$

$$f_2 = \frac{P_2}{Q_2} = \frac{q_2(P_1) + p_2(P_0)}{q_2(Q_1) + p_2(Q_0)} \tag{271}$$

If, formally, we set $P_{-1} = 1$, $Q_{-1} = 0$, these expressions suggest that successive numerators and denominators obey the recurrence:

$$\left.\begin{aligned} P_k &= q_k P_{k-1} + p_k P_{k-2} \\ Q_k &= q_k Q_{k-1} + p_k Q_{k-2} \end{aligned}\right\} \tag{272}$$

We shall prove that this is in fact so by induction.

The recurrence (272) clearly holds for $k = 2$. Now, let us assume that it holds for $k = n$. The convergent $f_{n+1}$ can be considered as the $n$th convergent of the continued fraction

$$f'_n = q_0 + \frac{p_1}{q_1 +} \frac{p_2}{q_2 +} \cdots \frac{p_n}{+ q_n + \left(\dfrac{p_{n+1}}{q_{n+1}}\right)} = \frac{P'_n}{Q'_n} \tag{273}$$

This fraction differs from $f_n$ only in the $n$th partial denominator, which is $q_n + (p_{n+1}/q_{n+1})$ instead of merely $q_n$. The $(n-1)$th and $(n-2)$th numerators and denominators of $f'_n$ are only of order $n$, so we can write by Eq. (272)

$$\begin{aligned} P'_n &= \left(q_n + \frac{p_{n+1}}{q_{n+1}}\right) P_{n-1} + p_n P_{n-2} \\ &= q_n P_{n-1} + p_n P_{n-2} + \frac{p_{n+1}}{q_{n+1}} P_{n-1} \end{aligned} \tag{274}$$

or

$$P'_n = P_n + \frac{p_{n+1}}{q_{n+1}} P_{n-1} \tag{275}$$

and, similarly,

$$Q'_n = Q_n + \frac{p_{n+1}}{q_{n+1}} Q_{n-1} \tag{276}$$

Thus,

$$\frac{P'_n}{Q'_n} = \frac{P_{n+1}}{Q_{n+1}} = \frac{P_n + \left(\dfrac{p_{n+1}}{q_{n+1}}\right) P_{n-1}}{Q_n + \left(\dfrac{p_{n+1}}{q_{n+1}}\right) Q_{n-1}} = \frac{q_{n+1} P_n + p_{n+1} P_{n-1}}{q_{n+1} Q_n + p_{n+1} Q_{n-1}} \tag{277}$$

The recurrence thus holds for $k = n + 1$, and thus, by induction, for all $k$.

In using Eq. (272), individual numerator-denominator pairs may not be simplified by removing common factors,

although a factor that appears simultaneously in numerator and denominator of *two* successive convergents may be deleted before continuing with the recurrence. A particular application of this principle occurs when, in numerical application of the recurrence, the magnitudes of the numerators or denominators become so extreme that overflow or underflow threatens. Under these circumstances, $P_{k-2}$, $P_{k-1}$, $Q_{k-2}$ and $Q_{k-1}$ may all be rescaled simultaneously before computing $P_k$ and $Q_k$.

Computing the $n$th convergent of a continued fraction by the forward recurrence requires about $4n$ multiplications, $2n$ additions, and one division, compared to about $n$ divisions and $n$ additions for the backward recurrence. For evaluation of a single convergent, it thus requires almost twice as many operations. On most modern computers, however, the ratio of the times for multiplication and division is considerably smaller than 1:1, and may be as small as 1:7. The same observation holds for complex arithmetic, where a complex division requires almost twice as many real operations as a complex multiplication. Thus even for a single convergent, the forward recurrence may be more economical in spite of the larger number of operations. It has unquestionable superiority if several adjacent convergents are required for the same continued fraction.

The forward recurrence formula is important as an analytical tool, as well as for numerical computation. We will now use it to obtain a useful expression for the difference between two successive convergents. Let

$$t_n = f_n - f_{n-1} = \frac{P_n}{Q_n} - \frac{P_{n-1}}{Q_{n-1}} = \frac{P_n Q_{n-1} - Q_n P_{n-1}}{Q_n Q_{n-1}} \tag{278}$$

Then, using Eq. (272),

$$\begin{aligned} Q_n Q_{n-1} t_n &= (q_n P_{n-1} + p_n P_{n-2}) Q_{n-1} \\ &\quad - (q_n Q_{n-1} + p_n Q_{n-2}) P_{n-1} \end{aligned} \tag{279}$$

or

$$\begin{aligned} Q_n Q_{n-1} t_n &= -p_n (P_{n-1} Q_{n-2} - Q_{n-1} P_{n-2}) \\ &= -p_n Q_{n-1} Q_{n-2} t_{n-1} \end{aligned} \tag{280}$$

Applying this recurrence $n$ times, we find that

$$\begin{aligned} Q_n Q_{n-1} t_n &= (-1)^n \left( \prod_{k=1}^{n} p_k \right) (P_0 Q_{-1} - Q_0 P_{-1}) \\ &= (-1)^{n+1} \left( \prod_{k=1}^{n} p_k \right) \end{aligned} \tag{281}$$

We have thus converted our finite continued fraction to the finite sum,

$$f_n = q_0 + \sum_{k=1}^{n} \frac{(-1)^{k+1}}{Q_k Q_{k-1}} \left( \prod_{j=1}^{k} p_j \right) \tag{282}$$

This representation is often useful for analysis of convergence and truncation error. Moreover, although the evaluation of a sequence of convergents by this formula requires 4 multiplications, 2 additions, and 1 division per convergent, significantly more than either the forward or the backward recurrence, Maehly[2] has pointed out that it has definite computational advantages. In the first place, the explicit generation of the successive increments to $f$ facilitates monitoring convergence, particularly for those fractions for which the truncation error alternates in sign. A second advantage that occurs when multiple precision values must be computed is that this formula makes it unnecessary to maintain maximum precision throughout the calculation. As the magnitude of $t_n$ decreases, the precision of the calculations may be successively relaxed. Full precision must be maintained throughout the forward recurrence, while it is difficult to determine when precision should be increased using the backward recurrence.

The recurrence may also be used to form the elements of a new continued fraction with convergents equal to every other convergent of a given fraction. We may write for the original fraction,

$$\left.\begin{aligned} P_n &= q_n P_{n-1} + p_n P_{n-2} \\ P_{n-1} &= q_{n-1} P_{n-2} + p_{n-1} P_{n-3} \end{aligned}\right\} \tag{283}$$

and so

$$\begin{aligned} P_n &= q_n (q_{n-1} P_{n-2} + p_{n-1} P_{n-3}) + p_n P_{n-2} \\ &= (q_n q_{n-1} + p_n) P_{n-2} P + q_n p_{n-1} P_{n-3} \end{aligned} \tag{284}$$

Now, using

$$P_{n-2} = q_{n-2} P_{n-3} + p_{n-2} P_{n-4} \tag{285}$$

we may eliminate $P_{n-3}$, and obtain

$$P_n = (q_n q_{n-1} + p_n) + \frac{q_n p_{n-1}}{q_{n-2}} P_{n-2} - \left( \frac{q_n p_{n-1} p_{n-2}}{q_{n-2}} \right) P_{n-4} \tag{286}$$

[2]H. J. Maehly, unpublished manuscript, 1961.

and similarly for the denominators

$$Q_n = (q_n q_{n-1} + p_n) + \frac{q_n p_{n-1}}{q_{n-2}} Q_{n-2} - \left(\frac{q_n p_{n-1} p_{n-2}}{q_{n-2}}\right) Q_{n-4} \tag{287}$$

But this is the recurrence for a continued fraction $f'$, with elements

$$\left.\begin{aligned} q'_0 &= q_0 \\ p'_1 &= p_1 q_2 \\ q'_1 &= q_1 q_2 + p_2 \end{aligned}\right\} \tag{288}$$

and, for $k > 1$

$$\left.\begin{aligned} p'_k &= -\frac{q_{2k} p_{2k-1} p_{2k-2}}{q_{2k-2}} \\ q'_k &= \frac{q_{2k} q_{2k-1} 8_{2k-2} + p_{2k} q_{2k-2} + q_{2k} p_{2k-1}}{q_{2k-2}} \end{aligned}\right\} \tag{289}$$

Successive convergents of this fraction are equal to the even convergents of $f$.

Letting $n = 2k + 1$ in Eqs. (286) and (287), we obtain the elements of the fraction, $f''$, the convergents of which are equal to the odd convergents of $f$:

$$q''_0 = \frac{q_0 q_1 + p_1}{q_1} \tag{290}$$

and, for $k \geqq 1$

$$\left.\begin{aligned} p''_k &= -\frac{q_{2k+1} p_{2k} p_{2k-1}}{q_{2k-1}} \\ q''_k &= \frac{q_{2k+1} q_{2k} q_{2k-1} + q_{2k-1} p_{2k+1} + q_{2k+1} p_{2k}}{q_{2k-1}} \end{aligned}\right\} \tag{291}$$

The two continued fractions $f'$ and $f''$ are known, respectively, as the even and odd parts of the fraction $f$.

Construction of the even or odd part of a continued fraction is the simplest example of the process known as contraction. More extreme transformations of this type are described by Perron (Ref. 50, pp. 197–203), who also discusses the inverse process, expansion. Since the sequence of convergents of a contracted continued fraction is a subsequence of the convergents of the original fraction, a contracted fraction may well converge when the original diverges. Under these circumstances, however, the identity of the limit must be verified.

Since the even and odd parts of a continued fraction converge essentially twice as fast as the original, contraction can lead to definite computational economy, and should be seriously considered when using continued fractions for extended computation.

### C. Equivalence Transformations and Canonical Forms

A minor inconvenience in working with continued fractions is the variety of sets of elements which may represent the same function, which makes it far from obvious when two fractions are, in fact, identical. Although it is possible to specify canonical forms for the elements, this is not always desirable because of the added complexity which may appear and obscure the rules for forming the elements. Transformations which change the elements of a continued fraction without altering the sequence of its convergents are thus far more useful than the corresponding transformations for series.

A simple but powerful transformation of this type is based on the observation that, for any $k$ and any $\alpha_k \neq 0$, the fraction

$$f = q_0 + \frac{p_1}{q_1 +} \cdots + \frac{p_k}{q_k +} \frac{p_{k+1}}{q_{k+1} +} \cdots \tag{292}$$

is identically equal to

$$f = q_0 + \frac{p_1}{q_1 +} \cdots + \frac{\alpha_k p_k}{\alpha_k q_k +} \frac{\alpha_k p_{k+1}}{q_{k+1} +} \cdots \tag{293}$$

A transformation of this type is known as an *equivalence transformation.*

Equivalence transformations may be applied repeatedly. In particular, letting $\alpha_k = 1/q_k$, for $k = 1, 2, 3, \cdots$, we may transform Eq. (292) into an equivalent continued fraction with unit partial denominators:

$$f = q_0 + \frac{\frac{p_1}{q_1}}{1 +} \frac{\frac{p_2}{q_1 q_2}}{1 +} \cdots + \frac{\frac{p_k}{q_{k-1} q_k}}{1 +} \cdots \tag{294}$$

A particular advantage of this form is the simplification which it introduces into the elements of the contracted fractions, Eqs. (289) and (290).

The transformation to a fraction with unit partial numerators is somewhat more complicated. Since the factor $\alpha_k$ appears in both the $k$th and $(k+1)$th partial numerators, after taking $\alpha_1 = 1/p_1$, we must have, for $k > 1$

$$\alpha_k = \frac{1}{p_k}\alpha_{k-1} \tag{295}$$

Thus, the required fraction is

$$f = q_0 + \cfrac{1}{\cfrac{q_1}{p_1}} \cfrac{1}{+\alpha_2 q_2 +} \cdots + \cfrac{1}{\alpha_k q_k + \cdots} \tag{296}$$

with, for $k = 1, 2, \cdots$,

$$\alpha_{2k} = \prod_{j=1}^{k} \frac{p_{2j-1}}{p_{2j}}$$

$$\alpha_{2k+1} = \frac{1}{p_1}\prod_{j=1}^{k} \frac{p_{2j}}{p_{2j+1}} \tag{297}$$

Continued fractions of the form (296) are sometimes called *ordinary continued fractions.*

Other transformations which do not affect the sequence of convergents are also possible. Khovanskii (Ref. 17, pp. 19–23) discusses the transformation of Eq. (292) to the equivalent form:

$$f = \cfrac{p_0}{1 +} \cfrac{p_1'}{q_1' +} \cfrac{p_2'}{q_2' +} \cdots + \cfrac{p_k'}{q_k' + \cdots} \tag{298}$$

Since the formulas for the elements are rather complicated, we will not reproduce them here.

The number of possible forms for a continued fraction is greatly increased when the elements become functions of an independent variable, $x$. There are three particular forms which deserve mention because of their relation to other forms of functional representation. The first, called by Wall (Ref. 49, pp. 399–410) the *corresponding* or *C-fraction* form may be represented as

$$f(x) = q_0 + \cfrac{p_1 x^{n_1}}{1 +} \cfrac{p_2 x^{n_2}}{1 +} \cdots + \cfrac{p_k x^{n_k}}{1 + \cdots} \tag{299}$$

with the $p_k$ nonzero constants, and the $n_k$ positive integers. Fractions of this type may be constructed which have convergents equal to the partial sums of any power series.

The form

$$f(x) = \cfrac{1}{q_1 x +} \cfrac{1}{q_2 +} \cdots + \cfrac{1}{q_{2k-1}x +} \cfrac{1}{q_{2k} + \cdots} \tag{300}$$

or, under an obvious equivalence transformation,

$$f(x) = \cfrac{p_1}{x +} \cfrac{p_2}{1 +} \cdots + \cfrac{p_{2k-1}}{x +} \cfrac{p_{2k}}{1 + \cdots} \tag{301}$$

is known as the Stieltjes, or S-form, after its originator. If the $q_k$ (and thus the $p_k$) are real and positive, Eqs. (236) and (237) is called an *S-fraction,* or *Stieltjes fraction.*[3] S-fractions are the simplest continued fraction representations of functions defined as Laplace transforms.

Fractions of the form

$$f(x) = \cfrac{p_1}{q_1 + x +} \cfrac{p_2}{q_2 + x +} \cdots + \cfrac{p_k}{q_k + x + \cdots} \tag{302}$$

are called (Wall, Ref. 49, p. 103) *Jacobi fractions* or *J-fractions.* A $J$-fraction with all the $p_k$ and $q_k$ real, and with $p_1 > 0$, $p_k < 0$ $(k = 2, 3, \cdots)$ is referred to as a Grommer fraction (Perron, Ref. 51, p. 377). The $J$-fraction form is computationally attractive, since it allows the introduction of two independent parameters for each convergent computed. Unfortunately, serious loss of significance can occur in evaluating some fractions of this form, so careful checking is desirable. A further drawback is that it is not possible to construct either $J$- or S-fractions that correspond to all possible power series.

No simple formulas are known for direct conversion among the $J$-, $C$-, and S-fraction forms. For finite continued fractions, the most effective method is to reduce any one of the forms to a ratio of polynomials, using the recurrence (272), and then to generate the desired form by repeated division.

### D. The Continued Fraction for the Inverse Tangent

The inverse tangent, arctan $(x)$, can be shown to have the continued fraction expansion,

$$\arctan(x) = \cfrac{x}{1 +} \cfrac{x^2}{3 +} \cfrac{4x^2}{5 +} \cdots + \cfrac{(k-1)^2 x^2}{(2k-1) + \cdots} \tag{303}$$

[3]This nomenclature agrees with Perron (Ref. 50, p. 377) and Wall (Ref. 49, p. 118). Henrici (Ref. 52, p. 170) applies the term Stieltjes fraction to a $C$-fraction of the form of Eq. (299) with $n_1 = 0$, and $n_k = 1$ $(k > 1)$, and with all $p_k$ real and positive.

In contrast to the power series expansion, which converges only for $|x|<1$, and for $x^2 \neq -1$, and to the expansion in inverse powers of $x$, which converges only for $|x|>1$, the continued fraction converges over the entire finite complex plane, with the exception of the sections of the imaginary axis with $|x| \geqq 1$. We shall use this fraction to illustrate some of the results of the last two sections.

The elements of this fraction are:

$$\left.\begin{array}{lll} p_1 = x & p_k = (k-1)^2 x^2 & (k>1) \\ q_0 = 0 & q_k = (2k-1) & (k \geqq 1) \end{array}\right\} \tag{304}$$

To illustrate the various methods of evaluation, we will use Eq. (303) to compute $\arctan(1) = \pi/4 = 0.78539\,81634$. Intermediate results for the computation in eight significant decimal arithmetic using backward recurrence, forward recurrence, and the series (282) are given in Table 6. The intermediate results for the backward recurrence (263) appear in the fourth column. The fifth, sixth, and seventh columns give the $k$th numerators, denominators and convergents for the forward recurrence (272). The last four columns give the computation of the convergents as sums, using Eq. (262). Only the first three terms, and the accumulation were carried out in 10$D$ arithmetic. The remainder of the calculations used only eight significant digits. Nevertheless, the error of the 15th convergent is less than $10^{-10}$.

By an equivalence transformation with $\alpha_k = 1/(2k-1)$, we obtain the $C$-fraction:

$$\arctan x = \cfrac{x}{1+}\,\cfrac{\frac{x^2}{3}}{1+}\,\cfrac{\frac{4x^2}{15}}{1+}\cdots + \cfrac{(k-1)^2 x^2 (2k-1)(2k-3)}{1+}\cdots \tag{305}$$

The expressions (289) for the elements of the even part of a continued fraction simplify, in the case of unit partial denominators, to:

$$p'_1 = p_1 \qquad p'_k = -p_{2k-1}p_{2k-2} \qquad (k>1) \tag{306}$$

$$q'_0 = q_0,\ q'_1 = 1 + p_2 \qquad q'_k = 1 + p_{2k} + p_{2k-1} \qquad (k>1) \tag{307}$$

For the arctangent, after considerable algebraic manipulation, we find

$$\left.\begin{array}{l} \arctan(x) = \cfrac{x}{1+\frac{x^2}{3}-1+\left(\frac{11}{21}\right)x^2-\cdots} \\ \qquad\qquad \cfrac{\left(\frac{4}{45}\right)x^4}{} \\ \qquad\qquad \cfrac{\frac{(2k-2)^2(2k-3)^2}{(4k-3)(4k-5)^2(4k-7)}x^4}{} \\ \qquad -1+\cfrac{8k^2-12k+3}{(4k-1)(4k-5)}x^2-\cdots \end{array}\right\} \tag{308}$$

**Table 6. Evaluation of arctan (1) = $\pi/4$ = 0.78539 816340 by continued fraction, Eq. 303**

| $k$ | $p_k$ | $q_k$ | Backward recurrence[a] (Eq. 263) $n=10$, $r_k$ | Foward recurrence[b] $P_k$ | $Q_k$ | $f_k$ | Series[c] $Q_k$ $Q_{k-1}$ | $\prod_{j=1}^{k} p_j$ | $t_k$ | $\sum_{j=0}^{k} t_j$ |
|---|---|---|---|---|---|---|---|---|---|---|
| −1 | | | | 1.0000000(0) | 0.0000000(0) | | | | | |
| 0 | | 0 | 0.7853 9813 | 0.0000000(0) | 1.0000000(0) | 0.0000000(0) | | | 0.00000 00000 | 0.00000 00000 |
| 1 | 1 | 1 | 1.2732 396 | 1.0000000(0) | 1.0000000(0) | 1.00000000 | 1.0000000000(0) | 1.0000000000(0) | 1.00000 00000 | 1.00000 00000 |
| 2 | 1 | 3 | 3.6597 918 | 3.0000000(0) | 4.0000000(0) | 0.75000000 | 4.0000000000(0) | 1.0000000000(0) | −0.25000 00000 | 0.75000 00000 |
| 3 | 4 | 5 | 6.0625 189 | 1.9000000(1) | 2.4000000(1) | 0.79166667 | 9.6000000000(1) | 4.0000000000(0) | +0.04166 66667 | 0.79166 66667 |
| 4 | 9 | 7 | 8.4704 372 | 1.6000000(2) | 2.0400000(2) | 0.78431373 | 4.89600000(3) | 3.6000000000(1) | −0.00735 29412 | 0.78431 37255 |
| 5 | 16 | 9 | 10.8811 178 | 1.7440000(3) | 2.2200000(3) | 0.78558559 | 4.5288000(5) | 5.7600000(2) | +0.00127 18601 | 0.78558 55856 |
| 6 | 25 | 11 | 13.2899 706 | 2.3184000(4) | 2.9520000(4) | 0.78536585 | 6.5534400(7) | 1.4400000(4) | −0.00021 97319 | 0.78536 58537 |
| 7 | 36 | 13 | 15.7207 257 | 3.6417600(5) | 4.6368000(5) | 0.78540373 | 1.3687834(10) | 5.1840000(5) | +0.00003 78730 | 0.78540 37267 |
| 8 | 49 | 15 | 18.0099 010 | 6.5986560(6) | 8.4016800(6) | 0.78539721 | 3.8956910(12) | 2.5401600(7) | −0.00000 65204 | 0.78539 72063 |
| 9 | 64 | 17 | 21.2631 579 | 1.3548442(8) | 1.7250408(8) | 0.78539835 | 1.4493241(15) | 1.6257024(9) | +0.00000 11217 | 0.78539 72063 |
| 10 | 81 | 19 | 19.0000 000 | 3.1086951(9) | 3.9581136(9) | 0.78539815 | 6.8279075(17) | 1.3168189(11) | −0.00000 01929 | 0.78539 81351 |
| 11 | 100 | 21 | | | 1.0037079(11) | | 3.9727899(20) | 1.3168189(13) | +0.00000 00331 | 0.78539 81682 |
| 12 | 121 | 23 | | | 2.7874600(12) | | 2.7977956(23) | 1.5933509(15) | −0.00000 00057 | 0.78539 81625 |
| 13 | 144 | 25 | | | 8.4139894(13) | | 2.34536589(26) | 2.2944253(17) | +0.00000 00010 | 0.78539 81635 |
| 14 | 169 | 27 | | | 2.7428579(15) | | 2.3078377(29) | 3.8775788(19) | −0.00000 00002 | 0.78539 81633 |
| 15 | 196 | 29 | | | 9.6034298(16) | | 2.6340843(32) | 7.6000544(21) | +0.00000 00000 | 0.78539 81633 |

[a] Eq. (263), $n = 0$
[b] Eq. (272)
[c] Eq. (282)

For $x \neq 0$, this may be transformed to the Jacobi form:

$$x \arctan(x) = \cfrac{1}{x^{-2} + \cfrac{1}{3} - \cfrac{\frac{4}{45}}{x^{-2} + \cfrac{11}{21} - \cfrac{\frac{(2k-2)^2(2k-3)^2}{(4k-3)(4k-5)^2(4k-7)}}{-x^{-2} + \cfrac{8k^2-12k+3}{(4k-1)(4k-5)} - \cdots}}} \tag{309}$$

If the coefficients can be evaluated in advance and stored, this form offers significant economy over the form (308), or the original form (Eq. 303). After evaluating $x^{-2}$ [$x^4$ and $x^2$ for Eq. (308), and $x^2$ for Eq. (303)], each backward recurrence step requires only 2 additions and 1 division, compared to 2 additions, 2 multiplications, and 1 division for Eq. (244), and 2 additions, 2 multiplications, and 2 divisions for a double step of Eq. (303).

If the coefficients cannot be prestored, a further equivalence transformation will eliminate the two divisions per step required to compute them. Letting

$$\alpha_k = (4k-1)(4k-3)(4k-5) \tag{310}$$

Eq. (309) becomes:

$$x \arctan(x) = \cfrac{3}{3x^{-2} + 1 - \cfrac{7}{105x^{-2} + 5 - \cdots - \cfrac{(4k-1)(2k-2)^2(2k-3)^2(4k-9)}{-(4k-1)(4k-3)(4k-5)x^{-2} + (4k-3)(8k^2-12k+3) -}}} \tag{311}$$

This form requires 11 additions, 10 multiplications, and 1 division per backward recurrence step, compared to 11 additions, 8 multiplications, and 3 divisions for Eq. (309).

### E. Truncation Error and Convergence

We turn, now, to the error incurred when a convergent infinite continued fraction is terminated after a finite number of partial quotients. Let, as usual,

$$f = q_0 + \cfrac{p_1}{q_1 + \cfrac{p_2}{q_2 + \cdots}} \tag{312}$$

and let the truncated expansion be

$$f_n = q_0 + \cfrac{p_1}{q_1 + \cfrac{p_2}{q_2 + \cdots + \cfrac{p_n}{q_n}}} \tag{313}$$

Let the "tail" be

$$\theta_n = \cfrac{p_{n+1}}{q_{n+1} + \cfrac{p_{n+2}}{q_{n+2} + \cdots}} \tag{314}$$

We wish first to find an expression for the remainder,

$$R_{n+1} \equiv f - f_n \tag{315}$$

in terms of $\theta_n$ and other computable quantities. Writing Eq. (312) in terms of the tail, we obtain the formally finite fraction,

$$f = q_0 + \cfrac{p_1}{q_1 + \cfrac{p_2}{q_2 + \cdots + \cfrac{p_n}{q_n + \theta_n}}} \tag{316}$$

Except for the last partial denominator, this fraction is identical to Eq. (313), and has identical convergents except for the last. Denoting the $n$th convergent of Eq. (316) by $P_n/Q_n = f$, we have, by Eqs. (278) and (281),

$$f = \frac{P_{n-1}}{Q_{n-1}} + \frac{(-1)^{n+1}}{Q_n Q_{n-1}} \prod_{k=1}^{n} p_k \tag{317}$$

while

$$f_n = \frac{P_{n-1}}{Q_{n-1}} + \frac{(-1)^{n+1}}{Q_n Q_{n-1}} \prod_{k=1}^{n} p_k \tag{318}$$

and

$$R_{n+1} = \frac{(-1)^{n+1}}{Q_{n-1}} \left( \prod_{k=1}^{n} p_k \right) \left( \frac{1}{Q_n} - \frac{1}{Q_n} \right) \tag{319}$$

But

$$Q_n = (q_n + \theta_n) Q_{n-1} + P_n Q_{n-2} = \theta_n Q_{n-1} + Q_n \tag{320}$$

and so

$$\frac{1}{Q_n} - \frac{1}{Q_n} = -\frac{\theta_n Q_{n-1}}{Q_n Q_n} \tag{321}$$

Thus, again using Eq. (281),

$$R_{n+1} = (-1)^n \left( \prod_{k=1}^{n} p_k \right) \frac{\theta_n}{Q_n \widetilde{Q}_n} = -t_n \frac{\theta_n Q_{n-1}}{\theta_n Q_{n-1} + Q_n} \tag{322}$$

and we have expressed the truncation error in terms of the tails, the values of the last two convergents, and the values of the last two denominators.

More convenient bounds can be obtained if we restrict the values of the elements. Let us first consider what can be deduced if the elements are all real and positive. (Blanch, Ref. 51, p. 38) calls continued fractions of this type fractions of Class I.) It follows from the forward recurrence (272) that all the $P_k$ and $Q_k$ are then positive, and that $\theta_n$ is positive for all $n$. Hence,

$$0 \leq \frac{\theta_n Q_{n-1}}{(\theta_n Q_{n-1} + Q_n)} \leq 1 \tag{323}$$

and

$$R_{n+1} = -\xi t_n \qquad (0 \leq \xi \leq 1) \tag{324}$$

The truncation error is no greater in magnitude than, and opposite in sign to the difference between, the last two convergents. With all the elements positive, it is easily seen from Eq. (281) that the $t_n$, and hence the $R_{n+1}$, alternate in sign with increasing $n$. The value of $f$ thus always lies between $f_{n-1}$ and $f_n$.

The restriction to positive elements also simplifies the investigation of convergence. It follows immediately from Eq. (281) that, for all $n$,

$$f_{2n-2} < f_{2n} < f_{2n+1} < f_{2n-1} < f_{2n-3} \tag{325}$$

regardless of convergence. Since $\{f_{2k}\}$ is a monotone increasing sequence bounded above by $f_1$, and $\{f_{2k+1}\}$ is a monotone decreasing sequence bounded below by $f_0$, each sequence must individually approach a limit, $f_{2n} \to f_l$, and $f_{2n+1} \to f_u$ say, with $f_l \leq f_u$. The continued fraction converges if, and only if $f_l = f_u$.

Several theorems connect the convergence of a continued fraction with the divergence of various series of combinations of the elements. In particular, Khovanskii (Ref. 17, pp. 42–45) shows that a necessary and sufficient condition for the convergence of the continued fraction (312) with all elements positive is the divergence of any one of the series:

$$\left.\begin{aligned} S_1 &= \sum_{k=1}^{\infty} \left( \prod_{j=1}^{k} \frac{p_{2j-1}}{p_{2j}} \right) q_{2k} \\ S_2 &= \sum_{k=1}^{\infty} \frac{1}{p_1} \left( \prod_{j=1}^{k} \frac{p_{2j}}{p_{2j+1}} \right) q_{2k+1} \\ S_3 &= \sum_{k=2}^{\infty} \left( \frac{q_{k-1} q_k}{p_k} \right)^{1/2} \end{aligned}\right\} \tag{326}$$

With the trivial exception of the first partial numerator, the elements of the continued fraction (303) for $\arctan x$ are positive for all real $x$ The data of Table 6 illustrates the inequalities (325) as well as the error bound (324). To prove convergence, we may use the series $S_1$ of Eq. (326), which becomes

$$\begin{aligned} S_1 &= \sum_{k=1}^{\infty} P_1 \left( \prod_{j=1}^{k-1} \frac{p_{2j+1}}{p_{2j}} \right) \frac{q_{2k}}{p_{2k}} \\ &= x \sum_{k=1}^{\infty} \left( \prod_{j=1}^{k-1} \frac{(2j)^2 x^2}{(2j-1)^2 x^2} \right) \frac{4k-1}{(2k-1)^2 x^2} \end{aligned} \tag{327}$$

Each factor in the product is greater than 1, and so we can write

$$\begin{aligned} S_1 &\geq \frac{1}{x} \sum_{k=1}^{\infty} \frac{4k-1}{(2k-1)^2} \\ &= \frac{1}{x} \sum_{k=1}^{\infty} \left( 1 + \frac{3k-1}{(2k-1)^2} \right) \frac{1}{k} \geq \frac{1}{x} \sum_{k=1}^{\infty} \frac{1}{k} \end{aligned} \tag{328}$$

Since the divergence of the last series is well known, we have established the convergence of Eq. (303) for all real $x$.

Convergence of continued fractions is easily established, and bounds on their magnitudes determined if the elements obey certain inequalities. Suppose that by equivalence transformations, we can convert the fraction either to the form:

$$f_1 = \frac{p_1 e^{i\phi_1}}{1 -} \frac{p_2 e^{i\phi_2}}{1 -} \frac{p_3 e^{i\phi_3}}{1 -} \cdots \tag{329}$$

with all the $p_k$ real, and satisfying the inequalities $0 < p_k \leq 1/4$, or to the form:

$$f_2 = \frac{1}{q_1 e^{i\phi_1} -} \frac{1}{q_2 e^{i\phi_2} -} \frac{1}{q_3 e^{i\phi_3} -} \cdots \tag{330}$$

with all the $q_k$ real and satisfying the inequality $2 \leq q_k$. Then (Blanch, Ref. 51, pp. 392–394), the fraction converges, and

$$|f_j| < \bar{f}_j \qquad (j = 1, 2) \tag{331}$$

where $\bar{f}_j$ is the fraction corresponding to $f_j$ with all the $\phi_k = 0$. Moreover, the values of $\bar{f}_j$ satisfies the inequalities

$$p_1 < \bar{f}_1 \leq \frac{1}{2} \qquad \frac{1}{q_1} < \bar{f}_2 \leq 1 \tag{332}$$

We may use these results to establish the convergence, and bound the tails of the continued fraction (305) for arctan $(x)$ within the unit circle. For this fraction, the tails are given by

$$-\theta_n = \frac{\frac{n^2 (ix)^2}{(2n-1)(2n+1)}}{1 -} \frac{\frac{(n+1)^2 (ix)^2}{(2n+1)(2n+3)}}{1 -} \cdots \tag{333}$$

and, with $\phi_k = \pi$,

$$p_k = \frac{(n+k)^2 |x|^2}{4(n+k)^2 - 1} = \frac{1}{4}\left[1 + \frac{1}{4(n+k)^2 - 1}\right] |x|^2 \tag{334}$$

For any $x$ of magnitude less than 1, we may select $n \geq \frac{1}{2}(1 - |x|^2)^{\frac{1}{2}}$ and thus ensure that

$$\left[1 + \frac{1}{4(n+k)^2 - 1}\right] |x|^2 \leq 1 \tag{335}$$

and so $0 < p_k \leq 1/4$ for all positive $k$. This establishes the convergence of $\theta_n$, and thus of Eq. (305) for all $x$ of magnitude less than 1, and also the upper bound,

$$|\theta_n| \leq \frac{2n^2 |x|^2}{(4n^2 - 1)} \tag{336}$$

Sharper bounds can sometimes be obtained for fractions of the form $\bar{f}_1$ or $\bar{f}_2$ by a pair of comparison theorems (Blanch, Ref. 51, pp. 391–392). Let

$$\bar{f}'_1 = \frac{p'_1}{1 -} \frac{p'_2}{1 -} \cdots \tag{337}$$

be a fraction of known value, with

$$0 < p'_k \leq \frac{1}{4} \tag{338}$$

for $k = 1, 2, 3, \cdots$. Then,

$$\bar{f}'_1 \leq \bar{f}_1 \tag{339}$$

and equality holds only if $p'_k = p_k$ for all $k$. Similarly, if a fraction

$$\bar{f}'_2 = \frac{1}{q'_1 -} \frac{1}{q'_2 -} \cdots \tag{340}$$

is a fraction of known value with

$$2 \leq q_k \leq q'_k \tag{341}$$

for $k = 1, 2, 3, \cdots$, then

$$\bar{f}'_2 \leq \bar{f}_2 \tag{342}$$

and equality exists only if all elements of the two fractions are equal.

A useful set of comparison fractions for use with these theorems are the fractions in which all the partial numerators and partial denominators are equal. (Continued fractions of this type are called periodic. Fractions in which the numerators and denominators approach fixed limits are called periodic in the limit.) Consider the fraction

$$\bar{f}'_1 = \frac{p}{1 -} \frac{p}{1 -} \frac{p}{1 -} \cdots \tag{343}$$

which converges, as we have seen, for $0 < p \leq 1/4$. The tail, $\theta_0$, of $\bar{f}'_1$, is equal to $\bar{f}'_1$, and so, for $0 < p \leq 1/4$,

$$\bar{f}'_1 = \frac{p}{(1 - \bar{f}'_1)} \tag{344}$$

or

$$\bar{f}'^2_1 - \bar{f}'_1 + p = 0 \tag{345}$$

Thus

$$\bar{f}'_1 = \frac{1}{2}[1 - (1 - 4p)^{\frac{1}{2}}] \tag{346}$$

since the larger root does not approach 0 as $p \to 0$. Similarly, for $q \geqq 2$,

$$\bar{f}_2' = \frac{1}{q-}\frac{1}{q-}\frac{1}{q-}\cdots = \frac{q}{2}\left\{1-\left[1-\left(\frac{4}{q^2}\right)\right]^{1/2}\right\} \tag{347}$$

A number of valuable convergence criteria for continued fractions which are periodic, or periodic in the limit, are established in Khovanskii (Ref. 17, pp. 58–75). Two of the most useful apply to fractions of the form:

$$f(x) = \frac{p_1}{1+}\frac{p_2 x}{1+}\frac{p_3 x}{1+}\cdots+\frac{p_k x}{1+}\cdots \tag{348}$$

where $x$ and $p_k$ may now be complex, and

$$\lim_{k\to\infty} p_k = p$$

Fractions of this form converge uniformly in a large domain, with the possible exception of a set of isolated poles. If $p = 0$, this domain is the entire finite complex plane. If $p \neq 0$, it is the entire complex plane, with the exception of a neighborhood of the segment between $-\frac{1}{4}p$ and $\infty$ of the ray from the origin passing through the point $-\frac{1}{4}p$. For $p$ real, this branch cut will be a section of the real axis.

Bounds on the truncation error in terms of $t_n$, the difference between the last two convergents of the truncated fraction, are particularly convenient for practical computation. For convergent fractions with all positive elements, the expression (324) is satisfactory: the true value always lies between the values of the last two convergents. For fractions of the form (329) or (330), the truncation error may be many times greater than $t_n$, if the $p_k$ are all close enough to $\frac{1}{4}$, or the $q_k$ to 2. If, however, the magnitudes of the partial numerators in Eq. (329) are bounded away from $\frac{1}{4}$, or the magnitude of the partial denominators in Eq. (330) are bounded away from 2, we may obtain useful bounds using the following results of Blanch (Ref. 51, Th. 8).

In Eq. (329) let the magnitudes of the partial numerators $p_k$ satisfy,

$$0 < p_k \leqq \frac{1}{4} - \gamma \qquad \left(0 < \gamma < \frac{1}{4}\right) \tag{349}$$

for fixed $\gamma$ and all $k$. Then

$$0 < |R_{n+1}| \leqq \left(\frac{1}{2(\gamma)^{1/2}} - 1\right)|t_n| \tag{350}$$

Similarly, in Eq. (330), let the magnitudes of the partial denominators $q_k$ satisfy

$$2 + \gamma \leqq q_k \qquad (0 < \gamma) \tag{351}$$

for fixed $\gamma$ and all $k$. Then

$$0 < |R_{n+1}| \leqq \left\{\frac{\left[\left(1+\frac{\gamma}{2}\right)^2 - 1\right]^{1/2}}{\gamma} - \frac{1}{2}\right\}|t_n| \tag{352}$$

The values of these factors for various values of $\gamma$ are shown in Fig. 2 and in Table 7.

An interesting geometrical bound for the truncation error of $S$-fractions in the complex plane recently discovered by Henrici and Pfluger, is described by Henrici (Ref. 53, pp. 47–48). Let $f(x)$ be a convergent Stieltjes fraction of the form (301), and let its $k$th convergent be $f_k(x)$. Consistent with the convention on initial values for the forward recurrence, set $f_{-1}(x) = \infty$, and $f_0(x) = 0$.

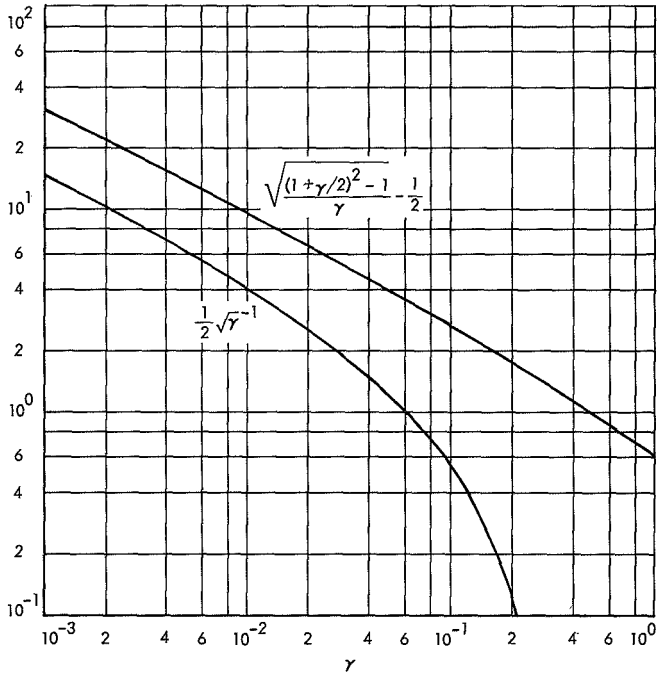


**Fig. 2. Coefficients for error bounds**

**Table 7. Truncation error bounding factors**

| $\gamma$ | $(\gamma)^{1/2}$ | $0.5\,\gamma^{-1/2} - 1$ | $(1 + 0.5\,\gamma)^2 - 1$ | $[(1 + 0.5\,\gamma)^2 - 1]^{1/2}$ | $[(1 + 0.5\,\gamma)^2 - 1]^{1/2}/\gamma - 0.5 - \frac{1}{2}$ |
|---|---|---|---|---|---|
| $1 \times 10^{-6}$ | 0.001 | 499.000 | — | — | — |
| $2 \times 10^{-6}$ | 0.001414214 | 352.554 | — | — | — |
| $4 \times 10^{-6}$ | 0.00200000 | 249.000 | — | — | — |
| $8 \times 10^{-6}$ | 0.002828427 | 175.777 | — | — | — |
| $10^{-5}$ | 0.003162277 | 157.114 | — | — | — |
| 0.001 | 0.03162277 | 14.812 | 0.00100025 | 0.031626729 | 31.526729 |
| 0.002 | 0.044721360 | 10.180 | 0.00200100 | 0.044732538 | 21.8663 |
| 0.004 | 0.063245553 | 6.9057 | 0.00400400 | 0.063277168 | 15.3193 |
| 0.008 | 0.089442719 | 4.5902 | 0.00801600 | 0.089532117 | 10.6915 |
| 0.0100 | 0.10000000 | 4.0000000 | 0.01002500 | 0.10012492 | 9.5125 |
| 0.0200 | 0.14142136 | 2.5355 | 0.02010000 | 0.14177447 | 6.5887 |
| 0.04 | 0.20000000 | 1.500000 | 0.04040000 | 0.20099751 | 4.5249 |
| 0.08 | 0.28284271 | 0.7677670 | 0.08160000 | 0.28565714 | 3.0707 |
| 0.10 | 0.31622772 | 0.5811391 | 0.10250000 | 0.32015621 | 2.7015621 |
| 0.20 | 0.44721360 | 0.1180340 | 0.2100000 | 0.45825757 | 1.7913 |
| 0.25 | 0.50000000 | 0.000000 | 0.265625 | 0.51538820 | 1.5616 |
| 0.40 | — | — | 0.44000000 | 0.66332496 | 1.1583 |
| 0.80 | — | — | 0.96000000 | 0.97979589 | 0.72474 |
| 1.00 | — | — | 1.25000000 | 1.1180340 | 0.6180340 |

Then, provided the $p_k(x)$ are all strictly positive, all $f_k(x)$ are distinct, each set of three consecutive convergents, $f_{k-1}(x)$, $f_k(x)$, $f_{k+1}(x)$ determines a unique circle in the complex plane. Let $c_k$ be the arc of this circle which begins at $f_{k-1}(x)$ and passes through $f_{k+1}(x)$ to $f_k(x)$. Henrici and Pfluger's result is that, for each $k$, the value of $f(x)$ lies in the lens shaped region bounded by $c_k$ and the portion of $c_{k-1}$ lying between $f_{k-1}(x)$ and $f_k(x)$.

For $x$ real, the $c_k$ degenerate into segments of the real axis, and for $x$ positive, Henrici and Pfluger's rule reduces to the rule that the value of a convergent continued fraction with all elements positive lies between the values of any pair of successive convergents. If the convergents form a monotone sequence, as can happen with $x$ real and negative, the requirement that $f_{k+1}(x)$ lie in the interior of the arc $c_k$ means that the $c_k$ are the entire real axis with the exception of the interval $[f_{k-1}(x), f_k(x)]$. The bound is not, in this case, particularly useful.

### F. Derivation of Continued Fractions from Derivatives at a Single Point

It is now time to consider methods of determining the elements of continued fractions corresponding to a given function. The appropriate method depends, of course, on the way in which the function is defined. We begin with functions which are defined locally by specifying the values of the function and its derivatives at a point $a$. The resulting formula, which is analogous to Taylor's formula for power series, is generally referred to as Thiele's formula. As might be expected, the method is considerably more complicated than the corresponding method for series, and it is often more convenient to carry out the process in two steps, first expressing the function as a Taylor series, and then converting the series to a continued fraction using one of the algorithms of the next section.

The customary derivations of Thiele's formula (e.g., Ref. 54, pp. 119–121, and Ref. 55, pp. 426–438) begin by considering rational interpolation by reciprocal differences, and then develop the corresponding formulas for derivatives by allowing the interpolation points to coincide. Since we have not yet discussed rational interpolation, we will follow a somewhat different approach, the basic idea of which goes back to unpublished work of J. W. Tukey.

Denoting, as usual, the $n$th convergent of our continued fraction by

$$f_n(x) = \frac{P_n(x)}{Q_n(x)} = q_0 + \frac{x-a}{q_1 +} \frac{x-a}{q_2 +} \cdots + \frac{x-a}{q_n} \tag{353}$$

we consider the auxiliary function

$$T_n(x) \equiv Q_n(x) f(x) - P_n(x) = Q_n(x)\,[f(x) - f_n(x)] \tag{354}$$

where $f(x)$ is the function to be expanded.

Differentiating Eq. (354) $\mu$ times, using Leibtnitz' rule, we find

$$T_n^{(\mu)}(x) = \sum_{k=0}^{\mu} \binom{\mu}{k} Q_n^{(\mu-k)}(x) \left[f^{(k)}(x) - f_n^{(k)}(x)\right] \tag{355}$$

Thus, the vanishing of $T_n^{(\mu)}(a)$ is a necessary condition that $f^{(k)}(a) = f_n^{(k)}(a)$, $(k = 0, 1, 2, \cdots, \mu)$. Moreover, provided that $Q_n(a) \neq 0$, the vanishing of

$$T_n^{(k)}(a), \qquad (k = 0, 1, 2, \cdots, \mu)$$

is also sufficient.

If $P_n(x)$ and $Q_n(x)$ obey the same linear recurrence, $T_n(x)$ will obey it also. In particular, since $P_n(x)$ and $Q_n(x)$ are the $n$th numerator and denominator of the continued fraction (353), by (272),

$$T_n(x) = q_n T_{n-1}(x) + (x - a) T_{n-2}(x) \tag{356}$$

our task is to choose the (constant) partial denominators, $q_k$, so that $T_n^{(\mu)}(a) = 0$ $(\mu = 0, 1, 2, \cdots, n)$. Differentiating Eq. (356) $\mu$ times, we find

$$T_n^{(\mu)}(x) = q_n T_{n-1}^{(\mu)}(x) + (x - a) T_{n-2}^{(\mu)}(x) + \mu T_{n-2}^{(\mu-1)}(x) \tag{357}$$

Now let us assume that $q_0, q_1, \cdots, q_{n-1}$ have been chosen so that $T_{n-1}^{(\mu)}(a) = 0$ $(\mu = 0, 1, 2, \cdots, n-1)$ and that $T_{n-2}^{(\mu)}(a) = 0$ $(\mu = 0, 1, 2, \cdots, n-2)$. Then, whatever choice of $q_n$ we make, $T_n^{(\mu)}(a)$ will certainly vanish for $\mu \leqq n-2$, and also for $\mu = n-1$, since even if $T_{n-2}^{(n-1)}(a) \neq 0$, the factor $x - a$ eliminates this term. Our choice of $q_n$ is thus entirely determined by the condition that

$$T_n^{(n)}(a) = q_n T_{n-1}^{(n)}(a) + n T_{n-2}^{(n-1)}(a) = 0 \tag{358}$$

If, then, we set

$$q_n = -\frac{n T_{n-2}^{(n-1)}(a)}{T_{n-1}^{(n)}(a)} \tag{359}$$

we will, provided $Q_n(a) \neq 0$, have insured that

$$f_n^{(k)}(a) = f^{(k)}(a)$$

for $k = 0, 1, 2, \cdots, n$. We can then proceed to determine the next partial denominator of our continued fraction.

The algorithm may be summarized as follows: Using the usual convention on initial values for continued fractions, set $P_{-1}(x) = 1$, $Q_{-1}(x) = 0$, $P_0(x) = q_0 = f(a)$, and $Q_0(x) = 1$, so that

$$T_{-1}(x) = -1 \qquad T_0(x) = f(x) - f(a) \tag{360}$$

and the derivatives of the test function at $x = a$ (the only point at which we require them) are

$$T_{-1}^{(\mu)}(a) = 0 \qquad T_0^{(\mu)}(a) = f^{(\mu)}(a) \qquad (\mu = 1, 2, 3, \cdots) \tag{361}$$

Then, for $n = 1, 2, 3, \cdots$, compute $q_n$ by Eq. (359) and $T_n^{(\mu)}(a)$ (for $\mu > n$) by Eq. (357) which, for $x = a$, simplifies to

$$T_n^{(\mu)}(a) = q_n T_{n-1}^{(\mu)}(a) + \mu T_{n-2}^{(\mu-1)}(a) \tag{362}$$

To illustrate the procedure, let us expand $e^x$ as a continued fraction about $a = 0$. The $f^{(k)}(a)$ are all equal to 1, and so

$$T_{-1}(0) = -1, \qquad T_0(0) = 0 \tag{363}$$

$$T_{-1}^{(\mu)}(0) = 0 \qquad T_0^{(\mu)}(0) = 1 \qquad (\mu = 1, 2, 3, \cdots) \tag{364}$$

Thus

$$q_1 = -1 \frac{(-1)}{1} = 1 \tag{365}$$

and, for $\mu > 1$,

$$T_1^{(\mu)}(0) = T_0^{(\mu)}(0) + \mu T_{-1}^{(\mu-1)}(0) = 1 \tag{366}$$

Thus,

$$q_2 = \frac{-2T_0^{(1)}(0)}{T_1^{(2)}(0)} = -2 \tag{367}$$

and,

$$T_2^{(\mu)}(0) = -2T_1^{(\mu)}(0) + \mu T_0^{(\mu-1)}(0) = \mu - 2 \tag{368}$$

Now let us assume that, for $\mu > k \geqq 1$, the derivatives of the test function are given by:

$$T^{(\mu)}_{2k-1}(0) = -(-1)^k k \prod_{j=1}^{k-1} (\mu - 2k + j) \tag{369}$$

$$T^{(\mu)}_{2k}(0) = \prod_{j=0}^{k-1} (\mu - 2k + j) \tag{370}$$

and the partial denominators, correspondingly, by

$$q_{2k-1} = -(-1)^k (2k - 1) \tag{371}$$

$$q_{2k} = (-1)^k 2 \tag{372}$$

We have already shown that this is true for $k = 1$. Now let us assume that it is true for some specified value of $k$. Then

$$q_{2k+1} = -(2k+1) \frac{T^{(2k)}_{2k-1}(0)}{T^{(2k+1)}_{2k}(0)}$$

$$= -(2k+1) \frac{\left[-(-1)^k k \prod_{j=1}^{k-1} (2k - 2k + j)\right]}{\prod_{j=0}^{k-1} (2k + 1 - 2k + j)} \tag{373}$$

$$= (-1)^k (2k+1) = -(-1)^{k+1} [2(k+1) - 1] \tag{374}$$

Using Eqs. (369), (370), and (371) in Eq. (362), we find, after some algebra,

$$T^{(\mu)}_{2(k+1)-1}(0) = -(-1)^{k+1} (k+1) \prod_{j=1}^{(k+1)-1} [\mu - 2(k+1) + j] \tag{375}$$

which yields, with Eqs. (370) and (359),

$$q_{2(k+1)} = (-1)^{k+1} 2 \tag{376}$$

Again using Eq. (362), we find after reduction

$$T^{(\mu)}_{2(k+1)}(0) = \prod_{j=0}^{(k+1)-1} [\mu - 2(k+1) + j] \qquad (\mu > k+1) \tag{377}$$

Since Eqs. (374)–(377) are the same as Eqs. (369)–(372) with $k$ replaced by $k + 1$, our expressions are valid for all $k$.

Thus,

$$e^x = 1 + \overbrace{\frac{x}{1+}\frac{x}{-2+}}^{k=1} \overbrace{\frac{x}{-3+}\frac{x}{2+}}^{k=2} \frac{x}{5+} \frac{x}{-2+} \frac{x}{-7+} \cdots \tag{378}$$

$$= 1 + \frac{x}{1-}\frac{x}{2+}\frac{x}{3-}\frac{x}{2+}\frac{x}{5-}\frac{x}{7+}\cdots \tag{379}$$

### G. Derivation of Continued Fractions from Recurrences

Although Thiele's theorem is applicable, in principle, to any function for which the appropriate reciprocal differences exist, it is often too clumsy to be useful, and other methods, depending upon special characteristics of the function, turn out to be more convenient. Many important families of special functions can be shown to obey a linear homogeneous three-term recurrence, or difference equation, and in this section we will describe the application of such relations in deriving continued fractions. The direct use of recurrences in numerical computation has recently been reviewed by Gautschi (Ref. 23).

We consider, then, a sequence of numbers

$$f^{(k)}, k = 0, 1, 2, \cdots,$$

obeying the recurrence:

$$f^{(k+1)} + q_k f^{(k)} - p_k f^{(k-1)} = 0 \tag{380}$$

where the $p_k$ and $q_k$ are independent of $f$, but may be functions of $k$, as well as of one or more independent variables which we refrain from indicating explicitly.

Provided that $f^{(k)} \neq 0$, we may rearrange Eq. (380) to

$$\left[\frac{f^{(k+1)}}{f^{(k)}}\right] + q_k = -p_k \left[\frac{f^{(k-1)}}{f^{(k)}}\right] \tag{381}$$

or, providing in addition that $f^{(k-1)} \neq 0$ and $p_k \neq 0$,

$$\frac{f^{(k)}}{f^{(k-1)}} = \frac{p_k}{q_k + \left[\frac{f^{(k+1)}}{f^{(k)}}\right]} \tag{382}$$

If, further, $f^{(k+1)} \neq 0$, $p_{k+1} \neq 0$, we can write

$$\frac{f^{(k)}}{f^{(k+1)}} = \frac{p_k}{q_k +} \frac{p_{k+1}}{q_{k+1} + \left[\frac{f^{(k+2)}}{f^{(k+1)}}\right]} \tag{383}$$

and in general, provided $f^{(k-1)} \neq 0$ and $f^{(k+j)} \neq 0$, $p^{(k+j)} \neq 0$ for $j = 0, 1, 2, \cdots, n$,

$$\frac{f^{(k)}}{f^{(k-1)}} = \frac{p_k}{q_k +} \frac{p_{k+1}}{q_{k+1} +} \cdots \frac{p_{k+n}}{+ q_{k+n} + \left[\frac{f^{(k+n+1)}}{f^{(k+n)}}\right]} \tag{384}$$

We have thus obtained a finite continued fraction for the ratio of two consecutive values of any solution of the difference equation (289) with an expression for the tail,

$$\theta_n = \frac{f^{(k+n+1)}}{f^{(k+n)}} \tag{385}$$

which may be useful for bounding the truncation error.

Unfortunately, the infinite continued fraction obtained by letting $n \to \infty$ in Eq. (384) does not necessarily converge. Moreover, a homogeneous three-term difference equation like Eq. (380) always has two linearly independent solutions, and the question arises: if the infinite fraction converges, what linear combination of these solutions does $f^{(k)}$ represent?

To answer these questions, we must anticipate and introduce the concept of a minimal solution of a difference equation. A solution $f^{(k)}$, of a difference equation, is said to be minimal, or distinguished, if

$$\lim_{k \to \infty} \frac{f^{(k)}}{g^{(k)}} = 0 \tag{386}$$

when $g^{(k)}$ is any other solution of the equation which is not merely proportional to $f^{(k)}$. A minimal solution does not necessarily exist, but if one does, it is unique. The questions of the convergence and value of a continued fraction may be rephrased in terms of the existence and identity of the minimal solution to the corresponding difference equations by the following theorem of Pincherle, which is proved in Gautschi (Ref. 23, Th. 1.1):

The infinite continued fractions

$$\theta_{k-1} = \frac{p_k}{q_k +} \frac{p_{k+1}}{q_{k+1} +} \cdots \qquad (k = 1, 2, 3, \cdots) \tag{387}$$

converge if, and only if, the difference equation

$$f^{(k+1)} = -q_k f^{(k)} + p_k f^{(k-1)} \tag{388}$$

has a minimal solution, $f^{(k)}$, with $f^{(0)} \neq 0$. If Eq. (387) converges, then,

$$\theta_k = \frac{f^{(k)}}{f^{(k-1)}} \tag{389}$$

for $k = 1, 2, 3, \cdots$, provided $f^{(k-1)} \neq 0$.

The analysis of certain difference equations to determine the existence of a minimal solution is considered in Gautschi (Ref. 23). Although in general the task is of the same order of difficulty as establishing the convergence of the corresponding continued fraction, its solution may be obvious in particular special cases.

### H. The Hypergeometric Function

To illustrate the use of recurrences in deriving continued fractions, we may consider the Gauss hypergeometric function, which is defined in the neighborhood of the origin by the power series:

$${}_2F_1(a, b; c; x) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{(c)_k} \frac{x^k}{k!} \tag{390}$$

where $(\alpha)_0 = 1$, and

$$(\alpha)_n \equiv \alpha(\alpha + 1) \cdots (\alpha + n - 1) = \frac{\Gamma(\alpha + n)}{\Gamma(\alpha)} \qquad (n \geqq 1) \tag{391}$$

The notation of Eq. (391) is known as Pochhammer's symbol, and sometimes as the ascending factorial.

Observing that

$$\lim_{b \to \infty} \frac{(b)_k}{b^k} = \lim_{b \to \infty} \prod_{j=1}^{k} \frac{(b + j - 1)}{b} = \lim_{b \to \infty} \prod_{j=1}^{k} \left[1 + \frac{(j-1)}{b}\right] = 1 \tag{392}$$

we see that

$$\lim_{b \to \infty} {}_2F_1\left(a, b; c; \frac{x}{b}\right) = \sum_{k=0}^{\infty} \frac{(a)_k}{(c)_k} \frac{x^k}{k!} \equiv {}_1F_1(a; c; x) \tag{393}$$

The function ${}_1F_1(a; c; x)$ is called the *confluent hypergeometric function*, or Kummer's function.

The hypergeometric functions are discussed, to varying levels of detail, in most books on special functions, and are also the subject of several monographs. Slater (Ref. 16) devotes her first chapter to the Gauss function before proceeding to more general series of the same form, while Slater (Ref. 15) is entirely dedicated to the confluent hypergeometric function. All the elementary functions and many of the more important higher special functions can be expressed in terms of the Gauss and confluent hypergeometric functions. A table of functions which are special cases of the Gauss function may be found in Erdelyi, Magnus, Oberhettinger, and Tricomi (Ref. 1, Vol. 1, p. 87), while a similar table for the confluent hypergeometric function appears on p. 509 of AMS 55 (Ref. 10).

It is immediately apparent from Eq. (390) that the Gauss hypergeometric function is symmetric with respect to its first two parameters:

$$ {}_2F_1(a,b;c;x) = {}_2F_1(b,a;c;x) \tag{394}$$

and also that

$$ {}_2F_1(0,b;c;x) = {}_2F_1(a,0;c;x) = 1 \tag{395}$$

By comparing the coefficients of equal powers of $x$, it can be shown that, the Gauss hypergeometric function satisfies, among others, the difference equation,

$$ {}_2F_1(a,b;c;x) = {}_2F_1(a,b+1;c+1;x) - \frac{a(c-b)}{c(c+1)} x \, {}_2\frac{d}{dx^2} \times F_1(a+1,b+1;c+2;x) \tag{396}$$

and the differential equation

$$ x(1-x)\frac{d^2}{dx^2} {}_2F_1(a,b;c;x) + [c-(a+b+1)x]\frac{d}{dx} 2F_1(a,b;c;x) + ab\,{}_2F_1(a,b;c;x) = 0 \tag{397}$$

with initial conditions

$$ {}_2F_1(a,b;c;0) = 1 \qquad \frac{d}{dx} {}_2F_1(a,b;c;0) = \frac{ab}{c} \tag{398}$$

We shall now use the recurrence (396) to obtain a continued fraction for the ratio of two hypergeometric functions first discovered by Gauss. If

$$ {}_2F_1(a,b+1;c+1;x) \neq 0$$

we may write

$$ \frac{{}_2F_1(a,b;c;x)}{{}_2F_1(a,b+1;c+1;x)} = 1 - \frac{a(c-b)x}{c(c+1)} \frac{{}_2F_1(a+1,b+1;c+2;x)}{{}_2F_1(a;b+1;c+1;x)} \tag{399}$$

or, providing ${}_2F_1(a,b;c;x) \neq 0$,

$$ \frac{{}_2F_1(a,b+1;c+1;x)}{{}_2F_1(a,b;c;x)} = \frac{1}{1 - \dfrac{a(c-b)}{c(c+1)} x \dfrac{{}_2F_1(a+1,b+1;c+2;x)}{{}_2F_1(b+1,a;c+1;x)}} \tag{400}$$

But, because of the symmetry

$$ \frac{{}_2F_1(a+1,b+1;c+2;x)}{{}_2F_1(a,b+1;c+1;x)} = \frac{{}_2F_1(b+1,a+1;c+2;x)}{{}_2F_1(a;b+1;c+1;x)} \tag{401}$$

and this is of the form of the left side of Eq. (400) with $a$ replaced by $b+1$, $b$ by $a$, and $c$ by $c+1$. Thus,

$$ \frac{{}_2F_1(a,b+1;c+1;x)}{{}_2F_1(a,b;c;x)} = \frac{1}{1 - \dfrac{\dfrac{a(c-b)}{c(c+1)} x}{1 - \dfrac{(b+1)(c-a+1)}{(c+1)(c+2)} x \dfrac{{}_2F_1(b+2,a+1;c+3;x)}{{}_2F_1(b+1,a+1;c+2;x)}}} \tag{402}$$

or, using symmetry again,

$$\frac{{}_2F_1(a,b+1;c+1;x)}{{}_2F_1(a,b;c;x)} = \frac{1}{1-\dfrac{\dfrac{a(c-b)}{c(c+1)}x}{1-\dfrac{(b+1)(c-a+1)}{(c+1)(c+2)}x\dfrac{{}_2F_1(a+1,b+2;c+3;x)}{{}_2F_1(a+1,b+1;c+2;x)}}} \tag{403}$$

Applying Eq. (403) repeatedly, with $a = a + j$, $b = b + j$, and $c = c + 2j$, we obtain the continued fraction:

$$\frac{{}_2F_1(a,b+1;c+1;x)}{{}_2F_1(a,b;c;x)} = \frac{1}{1+}\,\frac{p_2x}{1+}\,\cdots\,+\frac{p_nx}{1+\theta_n} \tag{404}$$

where

$$p_{2k} = -\frac{(a+k-1)(c+k-1-b)}{(c+2k-1)(c+2k-2)}$$

$$p_{2k+1} = -\frac{(b+k)(c-a+k)}{(c+2k)(c+2k-1)} \tag{405}$$

and

$$\theta_{2k} = p_{2k+1}x\frac{{}_2F_1(a+k;b+k+1;c+2k+1;x)}{{}_2F_1(a+k,b+k;c+2k;x)} \tag{406}$$

$$\theta_{2k+1} = p_{2k+2}x\frac{{}_2F_1(b+k+1,a+k+1;c+2k+2;x)}{{}_2F_1(b+k+1,a+k;c+2k+1;x)} \tag{407}$$

Since,

$$\lim_{k\to\infty} p_k = -\frac{1}{4} \tag{408}$$

this fraction is periodic in the limit, of the form of Eq. (348) and converges uniformly in the entire complex plane, with the exception of a branch cut along the real axis, from $+1$ to $\infty$ and with the possible exception of a set of isolated poles.

Letting $x = z/a$ in Eq. (404), and taking the limit as $a \to \infty$, we obtain a continued fraction for the ratio of two confluent hypergeometric functions:

$$\frac{{}_1F_1(b+1;c+1;z)}{{}_1F_1(b;c;z)} = \frac{1}{1+}\,\frac{p_2z}{1+}\,\frac{p_3z}{1+}\,\cdots \tag{409}$$

where

$$p_{2k} = \lim_{a\to\infty} -\frac{(a+k+1)(c+k-1-b)}{(c+2k-1)(c+2k-2)a}$$

$$= -\frac{(c+k-1-b)}{(c+2k-1)(c+2k-2)} \tag{410}$$

$$p_{2k+1} = \lim_{a\to\infty} -\frac{(b+k)(c-a-k)}{(c+2k)(c+2k-1)a}$$

$$= \frac{b+k}{(c+2k)(c+2k-1)} \tag{411}$$

Here,

$$\lim_{k\to\infty} p_k = 0 \tag{412}$$

and so the fraction converges uniformly throughout the finite complex plane, except, possibly, for a set of isolated poles.

Observing that the Pochhammer symbol obeys the identities

$$(1)_k = k! \qquad (a)_k = \frac{a(a+1)_k}{(a+k)} \tag{413}$$

so that

$$\frac{1}{(2k+1)} = \frac{\left(\frac{1}{2}\right)_k (1)_k}{\left(\frac{3}{2}\right)_k k!} \tag{414}$$

we find, for $|x| \leq 1, x^2 \neq -1$

$$\arctan(x) = \sum_{k=0}^{\infty}\frac{(-1)^k x^{2k+1}}{2k+1} = x\sum_{k=0}^{\infty}\frac{(-x^2)^k}{2k+1}$$

$$= x\,{}_2F_1\left(\frac{1}{2},1;\frac{3}{2};-x^2\right) \tag{415}$$

Since, by Eq. (395)

$$ {}_2F_1\left(\frac{1}{2}, 0; \frac{3}{2}; -x^2\right) = 1 \tag{416} $$

$$ \arctan(x) = x \frac{{}_2F_1\left(\frac{1}{2}, 1; \frac{3}{2}; -x^2\right)}{{}_2F_1\left(\frac{1}{2}, 0; \frac{3}{2}; -x^2\right)} \tag{417} $$

and, using Eqs. (404) and (405), we obtain the continued fraction (305). Our general result on convergence of the Gauss continued fraction now shows that Eq. (305) converges except for the line $-x^2$ real and no smaller than 1, i.e., for $x$ pure imaginary and no smaller than 1 in magnitude.

Other examples of important functions which can be expressed as ratios of Gauss hypergeometric functions, or of confluent hypergeometric functions, and thus expanded in continued fractions appear in the exercises.

### I. Exercises

1. Gautschi (Ref. 23, p. 29) shows that, if we define the sequence of numbers $\rho_k$ by

$$ f_n = q_0 + \sum_{k=1}^{n} \prod_{j=1}^{k} \rho_j \tag{418} $$

then $\rho_1$ and $\rho_2$ are given by

$$ \rho_1 = \frac{p_1}{q_1} \qquad 1 + \rho_2 = \frac{1}{\left(1 + \frac{p_2}{q_1 q_2}\right)} \tag{419} $$

and, for $k \geqq 1$,

$$ 1 + \rho_{k+1} = \frac{1}{1 + \left(\frac{p_{k+1}}{q_k q_{k+1}}\right)(1 + \rho_k)} \tag{420} $$

Note that an algorithm based on this recurrence is less likely to overflow than one based on Eq. (282).

2. The continued fraction

$$ f = \frac{1}{2 -} \frac{1}{2 -} \frac{1}{2 -} \cdots = 1 \tag{421} $$

Compute the first ten convergents of this fraction, tabulating, for $k = 1\,(1)\,10$, $f_k$, the value of the convergent, $t_k = f_k - f_{k-1}$, the truncation error, $R_{k+1}$, and the ratio $R_{k+1}/t_k$.

Do the same for the fraction

$$ f = \frac{1}{3 -} \frac{1}{3 -} \frac{1}{3 -} \cdots = \frac{3}{2}\left[1 - \left(\frac{5}{9}\right)^{1/2}\right] = 0.3819660112 \tag{422} $$

3. Show that the Hurwitz zeta function, $\zeta(s, a)$, which is defined in Section III-J, satisfies the three-term recurrence:

$$ (a+1)^s \zeta(s, a+2) - [(a+1)^s + a^s]\,\zeta(s, a+1) + a^s \zeta(s, a) = 0 \tag{423} $$

and thus, letting

$$ \alpha_k = \left(\frac{a+k-1}{a+k}\right)^s \tag{424} $$

$$ \zeta(s, a+k+1) - (1+\alpha_k)\,\zeta(s, a+k) + \alpha_k \zeta(s, a+k-1) = 0 \tag{425} $$

Use this result to derive the continued fraction,

$$ \frac{\zeta(s, a+k)}{\zeta(s, a+k-1)} = \frac{\alpha_k}{(1+\alpha_k) -} \frac{\alpha_{k+1}}{(1+\alpha_k)(1+\alpha_{k+1}) -} \cdots \frac{\alpha_{k+n}}{(1+\alpha_{k+n-1})(1+\alpha_{k+n}) -} \cdots \tag{426} $$

Investigate the convergence of this continued fraction as a function both of $s$ and of $a$.

Observing the identity

$$ \zeta(s, a) = \frac{1}{a^s\left[1 - \frac{\zeta(s, a+1)}{\zeta(s, a)}\right]} \tag{427} $$

use Eq. (426) to compute $\zeta(2, 1)$ and $\zeta(3, 1)$. Compare your results for various convergents of Eq. (426) with the partial sums of the series (154).

## VI. The Padé Table and the *qd* Algorithm

The scope of the methods of constructing continued fractions discussed in the last chapter is not nearly so

broad as that of the methods available for expanding functions in series. For this reason, we will be interested in rational functions which correspond to a given series in the sense that the initial terms of their series expansion agree with the initial terms of the given series. In this chapter we will consider methods of constructing such rational functions, and some of their most important properties.

### A. The Padé Table

Let

$$f(x) = \sum_{k=0}^{\infty} a_k x^k \tag{428}$$

be a given formal power series (not necessarily convergent) with $a_0 \neq 0$, and let

$$R_{m,n}(x) \equiv \frac{P_m^{(n)}(x)}{Q_n^{(m)}(x)} \equiv \frac{\sum_{k=0}^{m} p_k^{(m,n)} x^k}{\sum_{k=0}^{n} q_k^{(m,n)} x^k} \tag{429}$$

denote a rational function with at least one of the $q_k^{(m,n)}$ different from zero. The $m+n+2$ coefficients in Eq. (429) may be chosen so that the formal expansion of $R_{m,n}(x)$ in ascending powers of $x$ agrees with the power series (482) at least through the term $a_{m+n}x^{m+n}$, i.e., so that

$$f(x) - R_{m,n}(x) = x^{m+n+1}\rho_{m,n} \tag{430}$$

where $\rho_{m,n}$ is a series in nonnegative powers of $x$. The rational function defined in this way is unique, up to common factors of numerator and denominator, since if any two rationals, $R_{m,n}(x)$ and $\bar{R}_{m,n}(x)$, say, satisfy

$$\left.\begin{aligned} f(x) - \frac{P_m^{(n)}(x)}{Q_m^{(n)}(x)} &= x^{m+n+1}\rho_{m,n} \\ f(x) - \frac{\bar{P}_m^{(n)}(x)}{\bar{Q}_n^{(m)}(x)} &= x^{m+n+1}\bar{\rho}_{m,n} \end{aligned}\right\} \tag{431}$$

then, subtracting,

$$\frac{P_m^{(n)}(x)}{Q_n^{(m)}(x)} - \frac{\bar{P}_m^{(n)}(x)}{\bar{Q}_n^{(m)}(x)} = x^{m+n+1}(\bar{\rho}_{m,n} - \rho_{m,n}) \tag{432}$$

or

$$P_m^{(n)}(x)\,\bar{Q}_n^{(m)}(x) - \bar{P}_m^{(n)}(x)\,Q_n^{(m)}(x) = x^{m+n+1}(\bar{\rho}_{m,n} - \rho_{m,n})\,Q_n^{(m)}(x)\,\bar{Q}_n^{(m)}(x) \tag{433}$$

Now the left side is a polynomial of degree $n+m$, at most, while the right side is of degree $m+n+1$ at least. This can only happen if both sides vanish, i.e., if

$$R_{m,n}(x) \equiv \bar{R}_{m,n}(x) \tag{434}$$

The various $R_{m,n}(x)$ satisfying Eq. (430) for a given power series may be arranged in a square table, with all the rows having equal values of $m$, and all the columns, equal values of $n$. The first row thus consists of the partial sums of $f(x)$, and the first column of the reciprocals of the partial sums of the series expansion of $1/f(x)$. A table of this sort is known as the Padé table of the formal series (428).

If each entry in the Padé table is distinct, the table, and the corresponding series, are said to be hypernormal. If two or more entries are identical, the table is said to be abnormal, or *degenerate*: A degenerate table will arise if, for some $(m,n)$, the constant term of $\rho_{m,n}$ in Eq. (430) vanishes, since then we may write

$$f(x) - R_{m,n}(x) = x^{m+n+2}\rho_{m+1,n} = x^{m+n+2}\rho_{m,n+1} \tag{435}$$

and these are the conditions which must be satisfied by $R_{m+1,n}(x)$ and $R_{m,n+1}(x)$. Since we have just shown that the elements of the Padé table are unique, the table is degenerate. An element, $R_{m,n}(x)$ is said to be normal if it appears only as the $(m,n)th$ entry of the table. It should be observed that degeneracy in the Padé table, although it complicates the theory, is desirable from a practical standpoint, since it implies that element with minimum $m+n$ is actually more efficient than would be expected.

A variety of methods for finding the elements of the Padé table will be described in later sections. A straightforward, although laborious, one is to reduce the problem to the solution of a set of linear algebraic equations. Writing Eq. (430) in the equivalent form:

$$Q_n^{(m)}(x)\,f(x) - P_n^{(m)}(x) = x^{m+n+1}\,Q_n^{(m)}(x)\,\rho_{m,n} \tag{436}$$

and introducing the explicit forms for $Q_n^{(m)}(x)$, $f(x)$, and $\rho_n^{(m)}(x)$, we obtain:

$$\sum_{k=0}^{\infty}\sum_{j=0}^{n} a_{k-j}\, q_j^{(m,n)}\, x^k - \sum_{k=0}^{m} P_k^{(m,n)}\, x^k = x^{m+n+1}\, Q_n^{(m)}(x)\, \rho_{m,n} \tag{437}$$

where, for convenience, we have set $a_j = 0$ for $j$ negative.

By hypothesis, $Q_n^{(m)}(x)$ has at most $n$ zeros, and thus, if Eq. (437) is to hold for all $x$, the coefficients of each power of $x$ must vanish independently. We thus obtain:

$$\sum_{j=0}^{n} a_{k-j}\, q_j^{(m,n)} = P_k^{(m,n)} \qquad (k = 0, 1, 2, \cdots, m) \tag{438}$$

and

$$\sum_{j=0}^{n} a_{k-j}\, q_j^{(m,n)} = 0 \qquad (k = m+1, m+2, \cdots, m+n) \tag{439}$$

The system (439) consists of $n$ homogeneous linear equations for the $n+1$ unknown $q_j^{(m,n)}$ and thus always has a nontrivial solution. Once this solution has been found, the $p_k^{(m,n)}$ may be determined by simple substitution in Eq. (438).

This formulation leads to an important, if somewhat clumsy, criterion for the normality of an element of the Padé table. Let us introduce the Hankel determinants, $H_\nu^{(\mu)}$, of the formal power series (428) by:

$$H_\nu^{(\mu)} \equiv \begin{vmatrix} c_\mu, c_{\mu+1}, \cdots, c_{\mu+\nu-1} \\ c_{\mu+1}, c_{\mu+2}, \cdots, c_{\mu+\nu} \\ \cdot \quad \cdot \quad\quad \cdot \\ \cdot \quad \cdot \quad\quad \cdot \\ \cdot \quad \cdot \quad\quad \cdot \\ c_{\mu+\nu-1}, c_{\mu+\nu}, \cdots, c_{\mu+2\nu-2} \end{vmatrix} \qquad \begin{matrix} (\mu \geqq 0) \\ (\nu > 0) \end{matrix} \tag{440}$$

with the additional conventions that, for negative $\mu$, (Eq. 440) is to hold with $c_{-j} = 0, j = 1, 2, 3, \cdots$, while $H_0^{(\mu)} = 1$.

Let us denote one of the elements of the table that is equal to $R_{\bar{n}}^{(\bar{m})}$ *by* $R_n^{(m)}$. If $\bar{m} + \bar{n} > m + n$, then Eq. (439) must hold for $k = m + n + 1$ as well as for the previous values, and the $q^{(m,n)}$ must satisfy an $(n+1) \times (n+1)$ system of linear homogeneous equations. A necessary and sufficient condition for this is the vanishing of the determinant of the coefficients, which is the Hankel determinant $H_{n+1}^{(m-n+1)}$. If $\bar{m} + \bar{n} \leqq m + n$, and $m$ and $\bar{m}$ and $n$ and $\bar{n}$ are not both equal, then either $p_m^{(m,n)}$ or $q_n^{(m,n)}$ or both must vanish. In the first case, the last equation of the set (438) is also homogeneous, and may be taken with the set (439) to obtain an $(n+1) \times (n+1)$ system. The condition that this system have a nontrivial solution is that the determinant $H_{n+1}^{(m-n)}$ be different from zero. In the second case, we may apply the theorem (Bocher, Ref. 56, p. 47) that all solutions of a system of $n$ homogeneous equations in $n+1$ unknowns are proportional to the $n \times n$ determinants, with alternating signs, obtained by deleting first the first column, then the second, and so on from the matrix of coefficients. The determinant corresponding to $q_n^{(m,n)}$ is $H_n^{(m-n+1)}$, which must, accordingly, vanish in this case. In the last case, there must also be a solution (obtained by multiplying numerator and denominator by $x$) to the system (438, 439) with $q_0^{(m,n)} = p_0^{(m,n)} = 0$. From Eq. (438) with $k = 0$, the vanishing of $q_0^{(m,n)}$ with $a_0 \neq 0$ is necessary, and sufficient for the vanishing of $p_0^{(m,n)}$. The theorem used in the second case now implies that $q_0^{(m,n)}$ must be proportional to the determinant $H_n^{(m-n+2)}$, which must, accordingly, vanish for degeneracy. Collecting all these results, we find that the vanishing of one or more of the four Hankel determinants $H_n^{(m-n+2)}$, $H_n^{(m-n+1)}$, $H_{n+1}^{(m-n+1)}$, or $H_{n+1}^{(m-n)}$ is a necessary and sufficient condition for the element $R_n^{(m)}$ of the Padé table to be degenerate.

A necessary and sufficient condition that the entire Padé table be hypernormal is that none of the $H_k^{(n)}, k \geqq 0, n > -k$ vanish. An alternative form of the last criterion (Wall, Ref. 49, p. 379) is that $H_k^{(n)} \neq 0$ and that $\bar{H}_k^{(n)} \neq 0\, (n \geqq 0, k \geqq 0)$ where $\bar{H}_k^{(n)}$ is the Hankel determinant of the reciprocal series

$$\frac{1}{f(x)} = \frac{1}{\sum_{k=0}^{\infty} a_k x^k} = \sum_{k=0}^{\infty} \bar{a}_k x^k \tag{441}$$

Terminology in this area is somewhat variable. We follow Henrici (Ref. 52, pp. 162–163) in calling a series hypernormal if all elements of the Padé table are nondegenerate. We shall also follow Henrici in calling a series *normal* if none of the $H_k^{(n)}$ with nonnegative $n$ vanish. This is equivalent to requiring that all the entries on and above the main diagonal of the Padé table be nondegenerate. Older authors, including Wall (Ref. 49) and Perron (Ref. 50) use the term normal in place of our hypernormal.

### B. The *qd* Array

As we have pointed out, the coefficients of the numerator and denominator of any element of the Padé table of a power series can be obtained by solving a system of $n + m + 1$ linear equations. Since the equations are of a rather special form, it is possible to avoid much of the labor associated with their solution by using a special scheme known as the quotient-difference, or *qd* algorithm. This algorithm, first developed by Rutishauser (Refs. 57–59), can be applied to a wide variety of numerical tasks. The review by Henrici (Ref. 52) summarizes most of the information pertinent to our interests.

The first step in applying the *qd* algorithm is to construct the so-called *qd* array corresponding to the given series. Let

$$f(x) = \sum_{k=0}^{\infty} a_k x^k \tag{442}$$

be a given power series, and now compute, for $n = 0, 1, 2, \cdots$,

$$q_1^{(n)} = \frac{a_{n+1}}{a_n} \tag{443}$$

Setting $e_0^{(n)} = 0$, we now form, successively the quantities $e_1^{(n)}, q_2^{(n)}, e_2^{(n)}, \cdots$, by alternating the recurrences:

$$\left.\begin{aligned} e_k^{(n)} &= e_k^{(n+1)} + [q_k^{(n+1)} - q_k^{(n)}] \\ q_{k+1}^{(n)} &= \frac{q_k^{(n+1)} e_k^{(n+1)}}{e_k^{(n)}} \end{aligned}\right\} \tag{444}$$

For hand computation, the results may conveniently be arranged in a pattern similar to the conventional difference table:

$$\begin{array}{lll} e_0^{(0)} = 0 & & \\ & q_1^{(0)} = \dfrac{a_1}{a_0} & \\ e_0^{(1)} = 0 & & \\ & q_1^{(1)} = \dfrac{a_2}{a_1} & \\ e_0^{(2)} = 0 & & \\ & q_1^{(2)} = \dfrac{a_3}{a_2} & \\ e_0^{(3)} = 0 & & \\ & q_1^{(3)} = \dfrac{a_4}{a_3} & \\ e_0^{(4)} = 0 & & \\ & q_1^{(4)} = \dfrac{a_5}{a_4} & \\ e_0^{(5)} = 0 & & \end{array}$$

$$\begin{array}{lll} & q_2^{(0)} & \\ e_1^{(1)} & & e_2^{(0)} \\ & q_2^{(1)} & \\ & & e_2^{(1)} \\ & q_2^{(2)} & \\ e_1^{(3)} & & \end{array} \qquad \begin{array}{l} \\ \\ q_3^{(0)} \\ \\ \\ \end{array}$$

The recurrences then connect the quantities at the vertices of a rhombus in this array, as indicated above. The *qd* algorithm is therefore often referred to as a rhombus algorithm. Other rhombus algorithms, with different rules connecting the quantities at the vertices, are also useful. Some of the more important of these will be discussed later in this chapter.

The complete *qd* array cannot be constructed if any of the $e_k^{(n)}$ vanish, since the following $q_{j+1}^{(n)}$, and all quantities depending on it, are then undefined. A necessary and sufficient condition that this difficulty should not arise is that the Hankel determinants, $H_k^{(n)}$, should not vanish for any $n \geqq 0$ and $k \geqq 0$. This is the condition that all elements of the Padé table of $f(x)$ which lie on or above the main diagonal should be normal. If $H_k^{(n)} \neq 0$ for all $n \geqq 0$, and for all $k$, $0 \leqq k \leqq K$, the array cannot be continued beyond the column $e_K^{(n)}$.

For normal series (hypernormal in Henrici's terminology), there is a useful relation between the elements of the *qd* array for $f$, and the elements, $\bar{q}_k^{(n)}, \bar{e}_k^{(n)}$, of the reciprocal series.

$$\bar{f}(x) = \frac{1}{f(x)} = \frac{1}{\sum_{k=0}^{\infty} a_k x^k} = \sum_{k=0}^{\infty} \bar{a}_k x^k \tag{445}$$

It is:

$$\bar{q}_1^{(0)} = -q_1^{(0)} \tag{446}$$

and otherwise

$$\left.\begin{aligned} \bar{q}_k^{(n)} &= e_{n+k-1}^{(1-n)} \\ \bar{e}_k^{(n)} &= q_{n+k}^{(1-n)} \end{aligned}\right\} \tag{447}$$

$$\left.\begin{aligned} q_k^{(n)} &= \bar{e}_{n+k-1}^{(1-n)} \\ e_k^{(n)} &= \bar{q}_{n+k}^{(1-n)} \end{aligned}\right\} \tag{448}$$

where the elements with negative subscripts obey the same recurrences as those with positive subscripts. The extended *qd* array may thus be written:

$$\begin{array}{llll} & q_1^{(0)} = -\bar{q}_1^{(0)} & q_2^{(-1)} = \bar{e}_0^{(2)} = 0 & q_3^{(-2)} = \bar{e}_0^{(3)} = 0 \\ e_0^{(1)} = 0 & e_1^{(0)} = \bar{q}_1^{(1)} & e_2^{(-1)} = \bar{q}_1^{(2)} & \\ & q_1^{(1)} = \bar{e}_1^{(0)} & q_2^{(0)} = \bar{e}_1^{(1)} & q_3^{(-1)} = \bar{e}_1^{(2)} \\ e_0^{(2)} = \bar{q}_2^{(-1)} = 0 & e_1^{(1)} = \bar{q}_2^{(0)} & e_2^{(0)} = \bar{q}_2^{(1)} & \\ & q_1^{(2)} = \bar{e}_2^{(-1)} & q_2^{(1)} = \bar{e}_2^{(0)} & q_3^{(0)} = \bar{e}_2^{(1)} \\ e_0^{(3)} = \bar{q}_3^{(-2)} = 0 & e_1^{(2)} = \bar{q}_3^{(-1)} & e_2^{(1)} = \bar{q}_3^{(0)} & \end{array}$$

Except for the element $\bar{q}_1^{(0)}$, the array of the reciprocal series can be obtained by simply reflecting the array in the diagonal $n = 1/2$, as indicated in the diagram.

When the coefficients of both $f$ and $\bar{f}$ are available, these relations furnish a useful check, as well as an alternate, and frequently more stable method in constructing the array. In addition, as we shall see they are necessary for computing the subdiagonal elements of the Padé table.

For our purposes, the most important property of the $qd$ array is the connection which it permits between power series and continued fractions. Let us consider the set of formal series,

$$f^{(\nu)}(x) = \sum_{k=0}^{\infty} a_{k+\nu} x^k \qquad (\nu = 0, 1, 2, \cdots) \tag{449}$$

so that $f^{(0)}(x)$ is the $f(x)$ of Eq. (442), and suppose that the diagonal and subdiagonal elements of the $qd$ array of $f^{(0)}(x)$ exist, i.e., that $f^{(0)}(x)$ is normal. Then, for all $\nu$, the series $f^{(\nu)}(x)$ corresponds to the continued fraction

$$f^{(\nu)} = \frac{a_\nu}{1-} \frac{q_1^{(\nu)} x}{1-} \frac{e_1^{(\nu)} x}{1-} \frac{q_2^{(\nu)} x}{1-} \frac{e_2^{(\nu)} x}{1-} \cdots \tag{450}$$

in the sense that the expansion of the continued fraction as a formal power series is identically equal to the series (449). Moreover, such a correspondence exists for all $\nu$ only if $f^{(0)}(x)$ is normal.

We can write, for any $\nu$,

$$f(x) = \sum_{k=0}^{\nu-1} a_k x^k + x^\nu f^{(\nu)}(x) \tag{451}$$

Observing that the $2n$th convergent of the continued fraction can be reduced to the ratio of two polynomials each of degree $n$, and the $(2n+1)$th convergent to the ratio of a polynomial of degree $n$ to one of degree $n+1$, we see that successive convergents of Eq. (451) are alternately ratios of a polynomial of degree $n+\nu$ to one of degree $n$, and of a polynomial of degree $n+\nu$ to one of degree $n+1$. The approximants of Eq. (451) are thus the entries in the $\nu$th superdiagonal of the Padé table, and of the diagonal immediately below it.

The $qd$ algorithm may also be applied to series in negative powers of $x$ of the form:

$$F^{(\nu)}(x) = \frac{1}{x} f^{(\nu)}\left(\frac{1}{x}\right) = \sum_{k=0}^{\infty} a_{k+\nu} x^{-(k+1)} \qquad (\nu = 0, 1, 2, \cdots) \tag{452}$$

The continued fractions corresponding to these series are given by the S-forms:

$$F^{(\nu)}(x) = \frac{a_\nu}{x-} \frac{q^{(\nu)}}{1-} \frac{e_1^{(\nu)}}{1-} \frac{q_2^{(\nu)}}{1} \cdots \qquad (\nu = 0, 1, 2, \cdots) \tag{453}$$

as can be verified by change of variable in Eq. (450) followed by an equivalence transformation. A necessary and sufficient condition for this expression is again the normality of the series for $f(x)$. The even and odd parts of this continued fraction are the $J$-forms:

$$G^{(\nu)}(x) = \frac{a_\nu}{x - q_1^{(\nu)} -} \frac{q_1^{(\nu)} e_1^{(\nu)}}{x - e_1^{(\nu)} - q_2^{(\nu)} -} \frac{q_2^{(\nu)} e_2^{(\nu)}}{x - e_2^{(\nu)} - q_3^{(\nu)} -} - \cdots - \tag{454}$$

and

$$U^{(\nu)}(x) = \frac{a_\nu}{x}\left\{1 + \frac{q_1^{(\nu)}}{x - q_1^{(\nu)} - e_1^{(\nu)} -} \frac{e_1^{(\nu)} q_2^{(\nu)}}{x - q_2^{(\nu)} - e_2^{(\nu)} -} \frac{e_2^{(\nu)} q_3^{(\nu)}}{x - q_3^{(\nu)} - e_3^{(\nu)} -} \cdots\right\} \tag{455}$$

As usual, these forms offer significant savings in effort over the S-form.

The $f^{(\nu)}(x)$ and $F^{(\nu)}(x)$ may instructively be interpreted in terms of the converging factors for power series discussed in Section IV. It will be recalled that these factors are defined by:

$$f(x) = \sum_{k=0}^{\nu-1} a_k x^k + \theta_\nu(x)\, a_\nu x^\nu \tag{456}$$

so that

$$\theta_v(x) = \frac{f^{(v)}(x)}{a_v} \tag{457}$$

and similarly for series in inverse powers of $x$. If the initial rate of convergence of the series is high, so that $|a_v x^v| << |f(x)|$, the relative precision to which $f^{(v)}(x)/a_v$ must be evaluated will be considerably less than the basic precision for $f(x)$. For this reason, some authors recommend direct summation of the first few terms (up to the minimum term for a divergent asymptotic series) before introducing the $qd$ continued fraction. For machine computation, however, the additional programming complication, and the need for retaining the entire $qd$ array may not be worth the slight gain in efficiency except in special cases.

### C. Continued Fractions for the Exponential Function

To illustrate the properties of the $qd$ array, and its use in constructing continued fractions and elements of the Padé table, we will consider the series for the exponential,

$$e^x = \sum_{k=0}^{\infty} \frac{1}{k!} x^k \tag{458}$$

and its reciprocal series

$$\frac{1}{e^x} = e^{-x} = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} x^k \tag{459}$$

We have, then,

$$q_1^{(n)} = \frac{a_{n+1}}{a_n} = \frac{1}{(n+1)!}\frac{n!}{1} = \frac{1}{n+1} \tag{460}$$

while

$$\bar{q}_1^{(n)} = \frac{\bar{a}_{n+1}}{a_n} = \frac{(-1)^{k+1}}{(n+1)!}\frac{n!}{(-1)^n} = -\frac{1}{n+1} = e_n^{(1-n)} \qquad (n \neq 0) \tag{461}$$

Next,

$$e_1^{(n)} = e_0^{(n+1)} + [q_1^{(n+1)} - q_1^{(n)}] = 0 + \left[\frac{1}{n+2} - \frac{1}{n+1}\right] = \frac{-1}{(n+1)(n+2)} \tag{462}$$

while

$$\bar{e}_1^{(n)} = \frac{1}{(n+1)(n+2)} = q_{n+1}^{(1-n)} \tag{463}$$

In particular, setting $n = 0$ in Eqs. (460) and (461), we confirm that $q_1^{(0)}$ is, in fact $-\bar{q}^{(0)}$, while setting $n = 1$ in Eqs. (460) and (463) $\bar{q}_1^{(1)}$ actually does turn out to equal $e_1^{(0)}$ and $e_1^{(1)}$ to equal $q_1^{(0)}$. Continuing, we find

$$q_2^n = q_1^{(n+1)}\frac{e_1^{(n+1)}}{e_1^{(n)}} = \left[\frac{1}{(n+2)}\right]\left[\frac{-1}{(n+2)(n+3)}\right]\left[\frac{(n+1)(n+2)}{-1}\right] = \frac{n+1}{(n+2)(n+3)} \tag{464}$$

$$e_2^{(n)} = \frac{-2}{(n+3)(n+4)} \tag{465}$$

and

$$\left.\begin{aligned} q_3^{(n)} &= \frac{(n+2)}{(n+4)(n+5)} \\ e_3^{(n)} &= \frac{-3}{(n+5)(n+6)} \end{aligned}\right\} \tag{466}$$

About now, a pattern begins to emerge, and we see that the results may be expressed in the forms:

$$\left.\begin{aligned} q_k^{(n)} &= \frac{n+k-1}{(n+2k-1)(n+2k-2)} \\ e_k^{(n)} &= \frac{-k}{(n+2k)(n+2k-1)} \end{aligned}\right\} \tag{467}$$

That these formulas do, in fact, hold for all $k$ is easily confirmed by induction, assuming Eq. (467) and computing first $q_{k+1}^{(n)}$ and then $e_{k+1}^{(n)}$.

In similar fashion, or more simply by using Eq. (463), we can demonstrate that the elements of the array for the reciprocal series are given by

$$\left.\begin{aligned} \bar{q}_k^{(n)} &= \frac{-(n+k-1)}{(n+2k-1)(n+2k-2)} \\ \bar{e}_k^{(n)} &= \frac{k}{(n+2k)(n+2k-1)} \end{aligned}\right\} \tag{468}$$

We thus find the two families of expansions of $e^x$ involving continued fractions:

$$e^x = \sum_{k=0}^{n-1} \frac{1}{k!}x^k + \frac{x^n}{n!}\left[\frac{1}{1-}\frac{\frac{1}{n+1}x}{1-}\frac{\frac{-1}{(n+1)(n+2)}x}{1-}\cdots-\frac{\frac{(n+k-1)}{(n+2k-1)(n+2k-2)}x}{1-}\frac{\frac{k}{(n+2k)(n+2k-1)}x}{1-}\cdots\right] \tag{469}$$

$$e^x = (e^{-x})^{-1} = \left\{\sum_{k=0}^{n-1}\frac{(-1)^k x^k}{k!} + \frac{(-1)^n x^n}{n!}\left[\frac{1}{1-}\frac{-\frac{1}{(n+1)}x}{1-}\frac{\frac{1}{(n+1)(n+2)}x}{1-}\cdots\right.\right.$$
$$\left.\left.-\frac{\frac{(n+k-1)}{(n+2k-1)(n+2k-2)}x}{1-}\frac{\frac{k}{(n+2k)(n+2k-1)}x}{1}\cdots\right]\right\}^{-1} \tag{470}$$

Writing Eq. (469) in the form:

$$\frac{n!\,f^{(n)}(x)}{x^n} = \frac{n!}{x^n}\left[e^x - \sum_{k=0}^{n-1}\frac{x^k}{k!}\right] = \frac{1}{1+}\frac{P_2 x}{1+}\frac{P_3 x}{1+}\cdots \tag{471}$$

with

$$\left.\begin{aligned} P_{2k} &= -q_k^{(n)} = \frac{(n+k-1)}{(n+2k-1)(n+2k-2)} \\ P_{2k+1} &= -e_k^{(n)} = \frac{k}{(n+2k)(n+2k-1)} \end{aligned}\right\} \tag{472}$$

and comparing with Eq. (409), we see that $n!f^{(n)}/x^n$ has the same continued fraction expansion as the ratio of confluent hypergeometric functions,

$$\frac{{}_1F_1(1;n+1;x)}{{}_1F_1(0;n;x)}$$

Since ${}_1F_1(0;n;x) = 1$, we obtain the identity:

$${}_1F_1(1;n+1;x) = \frac{n!}{x^n}\left[e^x - \sum_{k=0}^{n-1}\frac{x^k}{k!}\right] \tag{473}$$

which is also apparent from inspection of the power series. An alternate derivation of Eq. (469) could thus be based on the Gauss continued fraction for the ratio of two hypergeometric functions.

In this example, we were fortunate in being able to recognize the algebraic form of the elements of the $qd$ array. This is rarely possible in practice, although the advantages of an explicit general form, both from the analytical and from the numerical standpoints, justify a serious attempt to find one in almost all cases. In most problems, however, it will be necessary to construct the $qd$ array numerically. The initial segment of the array for $e^x$ is shown in Table 8. The calculations were carried out by applying the rhombus rules (444) to the tabular

**Table 8. Numerical *qd* array for $e^x$**

| $n-k$ | $e_0^{(n)}$ | $q_1^{(n)}$ | $e_1^{(n)}$ | $q_2^{(n)}$ | $e_2^{(n)}$ | $q_3^{(n)}$ | $e_3^{(n)}$ | $q_4^{(n)}$ | $e_4^{(n)}$ | $q_5^{(n)}$ | $e_5^{(n)}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | 0.100000(+1) | | | | | | | | | |
| | 0.000000 | | −0.500000(+0) | | | | | | | | |
| 2 | | 0.500000(+0) | | 0.166667(+0) | | | | | | | |
| | 0.000000 | | −0.166667(+0) | | −0.166668(+0) | | | | | | |
| 3 | | 0.333333(+0) | | 0.166666(+0) | | 0.999968(−1) | | | | | |
| | 0.000000 | | −0.833330(−1) | | −0.999980(−1) | | −0.999878(−1) | | | | |
| 4 | | 0.250000(+0) | | 0.150001(+0) | | 0.100006(+0) | | 0.714643(−1) | | | |
| | 0.000000 | | −0.500000(−1) | | −0.666690(−1) | | −0.714513(−1) | | −0.715848(−1) | | |
| 5 | | 0.200000(+0) | | 0.133332(+0) | | 0.952237(−1) | | 0.713308(−1) | | 0.550166(−1) | |
| | 0.000000 | | −0.333330(−1) | | −0.476140(−1) | | −0.535232(−1) | | −0.552125(−1) | | −0.535131(−1) |
| 6 | | 0.166667(+0) | | 0.119051(+0) | | 0.893145(−1) | | 0.696415(−1) | | 0.567160(−1) | |
| | 0.000000 | | −0.238100(−1) | | −0.357210(−1) | | −0.417338(−1) | | −0.449640(−1) | | |
| 7 | | 0.142857(+0) | | 0.107140(+0) | | 0.833017(−1) | | 0.664103(−1) | | | |
| | 0.000000 | | −0.178570(−1) | | −0.277732(−1) | | −0.332713(−1) | | | | |
| 8 | | 0.125000(+0) | | 0.972238(+1) | | 0.778036(−1) | | | | | |
| | 0.000000 | | −0.138890(−1) | | −0.222256(−1) | | | | | | |
| 9 | | 0.111111(+0) | | 0.888872(+1) | | | | | | | |
| | 0.000000 | | −0.111110(−1) | | | | | | | | |
| 10 | | 0.100000(+0) | | | | | | | | | |

values, starting with $e_0^{(n)}$ and $q_1^{(n)}$. Each computed value was rounded to six significant decimals, and the rounded value was used in later calculations. The underlined digits differ from the correct values by more than one unit. It is clear that the construction of the *qd* array in this way is mildly unstable.

Fortunately, however, when the continued fraction is converging reasonably well, inaccuracies in the partial numerators appear to have relatively little effect upon the values of the convergents, although the effects on the individual numerators and denominators is more serious. Thus, the convergents of exp (1) using the partial numerators from Table 8 differed by no more than $4 \times 10^{-7}$ from those using the correct values, when evaluated in 7-decimal arithmetic. On the other hand, $P_{11}$ and $Q_{11}$ differed in the third decimal. Experience with other functions has confirmed this behavior, and with adequate cross checking, the continued fractions derived from the numerical *qd* array appear to be acceptable methods of evaluating functions. Since the array needs only to be constructed once, double, triple, or higher precision floating point arithmetic, or multiple-word rational arithmetic may be used for this task, if available, without serious cost.

This example also illustrates the sad fact that conversion to a continued fraction is not always advantageous. The original exponential series converges as fast as the continued fraction, and is far easier to evaluate. This is somewhat unusual, however, and even for fairly rapidly convergent series it is usually worth the effort to explore the possibilities of even faster convergence, and better numerical stability, offered by the corresponding continued fraction.

## References


1. Erdelyi, A., Magnus, W., Oberhettinger, F., and Tricomi, F. G., *Higher Transcendental Functions*, 3 Vol. (Vol. 3, 1955). McGraw-Hill Book Co., New York, N.Y., 1953.
2. Erdelyi, A., Magnus, W., Oberhettinger, F., and Tricomi, F. G., *Tables of Integral Transforms*, 2 Vol. McGraw-Hill Book Co., New York, N.Y., 1954.
3. Whittaker, E. T., and Watson, G. N., *A Course of Modern Analysis*, 4th ed. Cambridge University Press, Cambridge, 1927.
4. Magnus, W., and Oberhettinger, F., *Formulas and Theorems for the Special Functions of Mathematical Physics*, J. Werner (Trans.). Chelsea, New York, N.Y., 1949.
5. Rainville, E. D., *Special Functions*. MacMillan, New York, N.Y., 1960.
6. Sneddon, I. N., *Special Functions of Mathematical Physics*. Oliver and Boyd, Edinburgh, 1956.
7. Hochstadt, H., *Special Functions of Mathematical Physics*. Holt, Rinehart and Winston, New York, N.Y., 1966.
8. Jahnke, E. and Emde, F., *Tables of Functions with Formulae and Curves*. Dover, New York, N.Y., 1945.
9. Jahnke, E., Emde, F., and Losch, F., *Tables of Higher Functions*. McGraw-Hill Book Co., New York, N.Y., 1960.
10. Abramowitz, M., and Stegun, I., (eds.), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards Applied Mathematics Series Vol. 55, U. S. Government Printing Office, Washington, D.C., 1964.

## References (contd)


11. Fletcher, A., Miller, J. C. P., Rosenhead, L., and Comrie, L. J., *An Index of Mathematical Tables*, 2 Vol., 2d ed. Addison-Wesley Publishing Co., Reading, Mass, 1962.

12. Greenwood, A. J., and Hartley, H. O., *Guide to Tables in Mathematical Statistics*. Princeton University Press, Princeton, N.J., 1962.

13. Watson, G. N., *A Treatise on the Theory of Bessel Functions*, 2d. ed. Cambridge University Press, Cambridge, 1944.

14. Luke, Y., *Integrals of Bessel Functions*. McGraw-Hill Book Co., New York, N.Y., 1962.

15. Slater, L. J., *Confluent Hypergeometric Functions*. Cambridge University Press, Cambridge, 1960.

16. Slater, L. J., *Generalized Hypergeometric Functions*. Cambridge University Press, Cambridge, 1966.

17. Khovanskii, A. N., *The Application of Continued Fractions and Their Generalizations to Problems in Approximation Theory* (P. Wynn, Trans.). Noordhoff, Groningen, 1963.

18. Erdelyi, A., *Asymptotic Expansions*. Dover, New York, N.Y., 1956.

19. Wynn, P., "The Numerical Efficiency of Certain Continued Fraction Expansions. Ia," *Proc. Koninkl. Akad. Wettensch. Amsterdam, 65A*, 127–137, 1962.

20. Wynn, P., "The Numerical Efficiency of Certain Continued Fraction Expansions. Ib," *Proc. Koninkl. Akad. Wettensch. Amsterdam, 65A*, 138–148, 1962.

21. Wynn, P., "Numerical Efficiency Profile Functions," *Proc. Koninkl. Nederl. Akad. Wettensch, Amsterdam, 65A*, 118–126, 1962.

22. Gautschi, W., "Algorithm 236—Bessel Functions of the First Kind," *Commun. Assoc. for Comput. Mach. 7*, 143–174, 1964.

23. Gautschi, W., "Computational Aspects of Three-Term Recurrence Relations," *SIAM Rev. 9*, 24–82, 1967.

24. Gautschi, W., "Algorithm 292—Regular Coulomb Wave Functions," *Commun. Assoc. Comput. Mach. 9*, 793–795, 1966.

25. O'Shea, D. M., and Thacher, H. C., Jr., "Computation of Resonance Line Shape Functions," *Trans. Am. Nuclear Soc. 6*, 36–37, 1963.

26. Knopp, K., *Theory and Applications of Infinite Series*, R. C. H. Young (Trans.). Blackie and Son, Ltd., London, 1928.

27. Bromwich, T. J. I'A., *Theory of Infinite Series* (2d. ed.). MacMillan, London, 1926.

28. Hirschman, I. I., Jr., *Infinite Series*. Holt, Rinehart and Winston, New York, N.Y., 1926.

29. Schwatt, I. J., *An Introduction to the Operations with Series*, University of Pennsylvania Press, Philadelphia, 2d ed. Chelsea Publishing Co., New York, N.Y., 1924.

## References (contd)


30. Jolley, L. B. W., *Summation of Series*, Dover, New York, N.Y., 1961.

31. Mangulis, V., *Handbook of Series for Scientists and Engineers.* Academic Press, New York, N.Y., 1965.

32. Gradshteyn, I. S., and Ryzhik, I. M., *Table of Integrals, Series, and Products* (4th ed. prepared by Ya. U. Geronimus and M. Yu. Tseytlin, translated by Scripta Technica Inc., translation edited by A. Jeffrey). Academic Press, New York, N.Y., 1965.

33. Dwight, H. B., *Tables of Integrals and Other Mathematical Data*, 4th ed. Macmillan, New York, N.Y., 1961.

34. Dwight, H. B., *Mathematical Tables of Elementary and Some Higher Mathematical Functions*, 2d ed. Dover, New York, N.Y., 1958.

35. Tolstov, G., *Fourier Series* (R. A. Silverman, Trans.). Prentice-Hall Inc., Englewood Cliffs, N.J., 1962.

36. Hardy, G. H., *Divergent Series.* Oxford University Press, Oxford, 1949.

37. Modern Computing Methods, prepared by the staff of the Mathematics Division, National Physical Laboratory. 2d. ed., N.P.L. Notes on Applied Science, Her Majesty's Stationery Office, London, 1960.

38. Naur, P., et al., "Report on the Algorithmic Language ALGOL 60," *Commun. ACM. 3*, 299–314, 1960.

39. Rosser, J. B., "Transformations to Speed the Convergence of Series," *J. Res. Nat'l. Bur. Stds. 46*, 56–64, 1951.

40. Miller, J., and Hurst, R. P., "Simplified Calculation of the Exponential Integral," *Math. Tables and other Aids to Comp. 12*, 187–193, 1958.

41. Steffensen, J. F.,, *Interpolation*, 2d. ed. Chelsea, New York, N.Y., 1950.

42. Titchmarch, E. C., *The Zeta Function of Riemann.* Cambridge University Press, Cambridge, 1930.

43. Titchmarsh, E. C., *The Theory of the Riemann Zeta-Function.* Oxford University Press, Oxford, 1951.

44. van Orstrand, C. E., "Reversion of Power Series," *Phil. Mag.* (6) 19, 366, 1910.

45. Thacher, H. C., Jr., "Solution of Transcendental Equations by Series Reversion," *Comm. ACM 9*, 10–11, 1966.

46. Thacher, H. C., Jr., "Algorithm 237 SERREV," *Comm. ACM 9*, 11, 1966.

47. Hille, E., *Analytic Function Theory*, Vol. I. Ginn and Co., New York, N.Y., 1959.

48. Lyness, J. N., "Numerical algorithms based on the theory of complex variables," *Proceedings of the 22d National Conference*, ACM, Thompson, Washington, D.C., 1967.

49. Wall, H. S., *Analytic Theory of Continued Fractions.* D. Van Nostrand, New York, N.Y., 1948.

## References (contd)


50. Perron, O., *Die Lehre van den Kettenbrüchen*, 2d. ed. Reprint (n.d.). Chelsea Publishing Co., New York, N.Y., 1929.

51. Blanch, G., "Numerical Evaluation of Continued Fractions," *SIAM Rev. 6*, 383–421, 1964.

52. Henrici, P., "Some Applications of the Quotient-Difference Algorithm," *Proc. Symp. Appl. Math. 15*, 159–183, 1963.

53. Henrici, P., "Error Bounds for Computations for Continued Fractions," *Error in Digital Computation* (L. B. Rall, Ed.), Vol. 2, pp. 39–53. John Wiley and Sons, New York, N.Y., 1965.

54. Milne-Thomson, L. M., *The Calculus of Finite Differences*. Macmillan and Co., London, 1933.

55. Nörlund, Niels Erik, *Vorlesungen iiker Differenzenrechnung*. Chelsea Publishing Co., New York, N.Y., 1954.

56. Bocher, M., *Introduction to Higher Algebra*. Macmillan, New York, N.Y., 1907.

57. Rutishauser, H., "Der Quotienten-Differenzen-Algorithmus," *Zeit. fur angewan. Math. und Physik 5*, 233–251, 1954.

58. Rutishauser, H., "Anwendungen des Quotienten-Differenzen-Algorithmus," *Zeit. fur angewan. Math. und Physik*, 5, 496–508, 1954.

59. Rutishauser, H., *Der Quotienten-Differenzen-Algorithmus*. Birkhauser, Basle, 1957.





NASA — JPL — Coml., L.A., Calif.