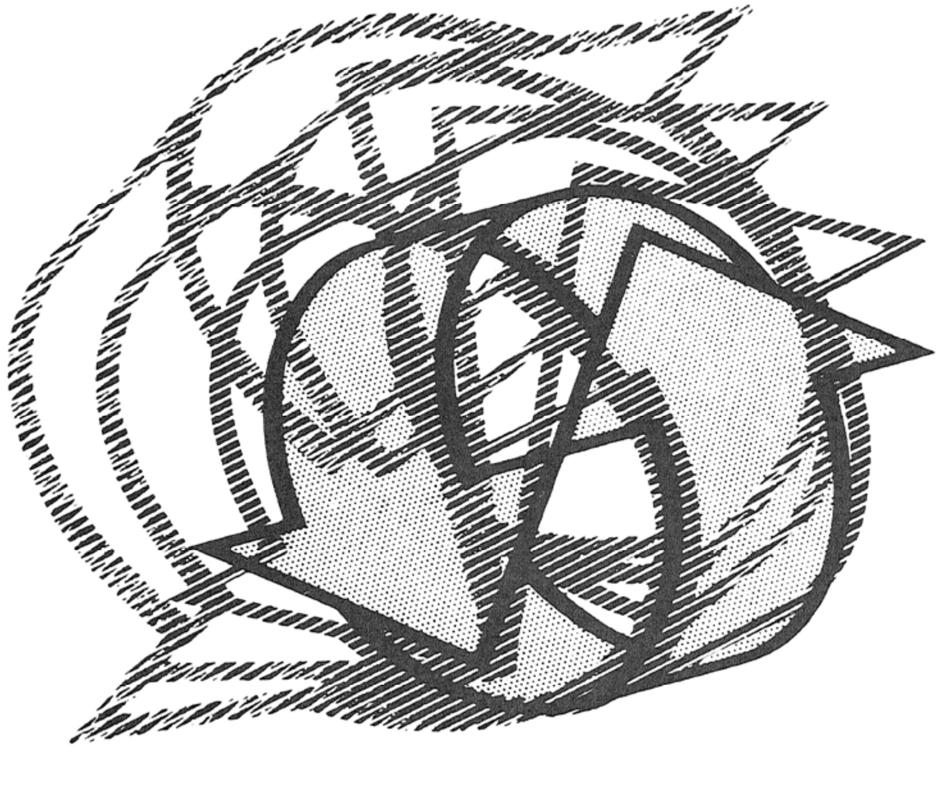
CWI Tract 10

Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

Calculation of special functions: the gamma function, the exponential integrals and error-like functions

C.G. van der Laan, N.M. Temme



Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

The CWI is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

CWI Tract 10

Calculation of special functions: the gamma function, the exponential integrals and error-like functions

C.G. van der Laan, N.M. Temme

Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

1980 Mathematics Subject Classification: 65D20, 65-04, 33-04.
ISBN 90 6196 277 3

CONTENTS

INTRODUCTION

The main scope of these notes is to review and to discuss several aspects of implementations for the numerical computation of special functions. In this tract we consider functions which are related to the Euler gamma function, the exponential integrals and the error functions. For each of these groups we give

1. definitions, analytic properties and fundamental formulas;
2. algorithms, implementations, error analysis, references to tabulated coefficients, and testing aspects.

We have limited ourselves to discuss the most important implementations, although we aimed at giving a complete survey. With respect to testing we have enumerated the techniques in use; no systematic testing has been done, although occasionally weak points or expensive methods have been observed.

We feel that these notes fill up a gap in the existing literature, and we consider them as an addition to the *Handbook of Special Functions* (Abramowitz & Stegun) and to the various books of Luke. Furthermore we mention in this respect Hart: *Computer Approximations* and Lyusternik et al.: *Handbook for Computing Elementary Functions*.

At the beginning of this project we intended to include more groups of functions, such as elliptic integrals, incomplete gamma functions and Bessel functions. However, the present notes grew out and the other groups are intended for a possible subsequent volume. Much depends on the need for it. We invite the readers to inform us on this point. Also, is the present form and set-up all right? Reactions are very welcome and will give us the motivation to continue, or to stop.

The first two chapters contain general information on the computation of special functions. The first one gives an annotated introduction to the literature and to several program libraries. Some local program libraries are surveyed when sufficient information happened to be available to us. No systematic search is made in checking more Computer Centres.

The second chapter gives a theoretical background on error analysis, recurrence relations, continued fractions and generalized hypergeometric functions.

This tract originated from regular meetings of the Working Group Approximation of Functions, i.e., the Dutch group on the subject. We kindly acknowledge and appreciate the contributed sections of our colleagues Dr. R.M.M. Mattheij of the University of Nijmegen (section II. 3.2: *The general aspects of three term recurrence relations*), and Drs. J.P. Hollenberg of the University of Groningen (section II. 4: *Continued fractions*). Furthermore, we like to thank the members of the working group for their much appreciated comments and criticisms and for the patience for waiting on this final version.

I. INTRODUCTION TO THE LITERATURE AND SOFTWARE

This chapter provides an introduction to the literature on the computation of (special) functions and to the available software. We give an annotated selection of relevant books and papers on the subject, which includes papers on general aspects of software and software engineering.

1. LITERATURE

ABRAMOWITZ, M. & STEGUN, I.A. (1964), *Handbook of mathematical functions with formulas, graphs and mathematical tables,* Nat. Bur. Standards Appl. Math. Series, 55, U.S. Government Printing Office, Washington, D.C.

A standard for general properties. A good starting point for classifying special functions, standard notation and definition. Contains many analytical properties, not always the most useful properties for numerical computation. No algorithms are provided for the evaluation, although now and then a polynomial or rational approximation is given. Tables are included with detailed information on how to use them in order to obtain values which are not tabulated. The numerical data are useful for the occasional (desk) calculator, which is actual because of the growing popularity of hand-hold calculators and the break-through of personal computers. It contains material up to 1960.

ABRAMOWITZ, M. (1954), *On the practical evaluation of integrals.* 175-190, in: Ph.J. Davis, Ph. Rabinowitz, (1967), Numerical Integration, Blaisdell.

Important with respect to the recognition of special functions in integrals with parameters. Examples concern: a disguised erfc, reduction to a known form, evaluating by a limiting procedure, use of functional relationships and termwise integration, extraction of singular part, reduction to a differential equation, Laplace transformation, saddle point approximation, inversion of order of integration.

ACTON, F.S. (1970), *Numerical methods that work,* Harper & Row Publishers.

In order to find approximations to a function and to choose between them on experimental grounds, trial and error is examplified in chapter 1. A priori transformation of an approximation problem in order to remove singularities is treated in chapter 15 by means of: substraction of the singular part, substitution of trigonometric functions, and substitution of Jacobian elliptic functions. Complex arguments are not considered.

BAUER F.L. (1973), *Software and software engineering*, SIAM Rev. 15, 469-480.

The development of software, in the past, now, and in the future, is discussed, exposing the weaknesses at that moment. Suggestions are given to overcome the 'software crisis'.

BAUER, F.L. (1980), *A trend for the next ten years of software engineering,* in: H. Freeman, P.M. Lewis (eds.): Software Engineering.

A sequel to Bauer (1973) where the correctness preserving transformation technique is emphasized. The CIP-project is treated as an example.

BOEHM, B.W. et al. (1978), *Characteristics of software quality,* North-Holland Publishing Company.

The aspects inherent to software quality are discussed.

BOEHM, B.W. (1981), *Software engineering economics,* Prentice Hall.

Given quality criteria of software, ways to produce software in an economic way are discussed.

BRENT, R.P. (1980), *Unrestricted algorithms for elementary and special functions*, in: S.H. Lavington (ed.) Information Processing 80, 613-619. North Holland Publishing Company.

Unrestricted algorithms which are useful for the computation of elementary and special functions when the required precision is not known in advance are described. Discussed are: the evaluation of power series, asymptotic expansions, continued fractions, recurrence relations, Newton iteration, contour integration and transformation of power series into a better conditioned form.

CLENSHAW, C.W. & F.W.J. OLVER (1980), *An unrestricted algorithm for the exponential function,* SIAM J. Numer. Anal., 17, 2, 310-331.

An algorithm is presented for the computation of the exponential function of real argument. There are no restrictions on the range of the argument or on the precision that may be demanded in the results.

BULIRSCH, R. & J. STOER, (1968), *Darstellung von Funktionen in Rechenautomaten,* 352-446 in: R. Sauer & I. Szabó, (eds.), *Mathematische Hilfsmittel des Ingenieurs,* Teil III, Springer Verlag.

Chebyshev polynomials (a lot of practical information), continued fractions, elliptic integrals, Fourier analysis, and Bessel functions are treated.

CODY, W.J. (1969), *Performance testing of function subroutines,* Proc. Spring Joint Computer, Conf., 34, 759-763 AFIPS Press, Montvale, N.J.

A general approach with respect to testing is given. Experience with testing of special functions is exposed and the test methods for special functions are enumerated.

CODY, W.J. (1970), *A survey of practical rational and polynomial approximation of functions,* SIAM Rev., 12, 400-423.

A good introduction to min-max approximations. Is partly overruled by GAUTSCHI (1975).

CODY, W.J., (1974), *The construction of numerical subroutine libraries,* SIAM Rev., 16, 30-46.

How to develop and maintain a collection of optimal numerical software, with respect to machine pecularities.

CODY, W.J. (1975a), *The FUNPACK package of special functions subroutines,* ACM Trans. Math. Software, 1, 13-25.

The design criteria of FUNPACK are exposed. FUNPACK is highly machine dependent, but from the start it is implemented for three lines of (large scale) computers: IBM 360-370, CDC 6000-7000, UNIVAC 1108. Important with respect to advanced software engineering.

CODY, W.J. (1975b), *An overview for software development for special functions,* in: G.A. Watson (ed.), *Numerical Analysis,* Lecture Notes in Mathematics 506, 38-48, Springer Verlag.

Again important. It treats various number representations and their influence on accuracy - transmitted error and generated error - as well as how to do best. The concept of wobbling word length is introduced. Examples are given with respect to the gamma function.

CODY, W.J. & W. WAITE (1980), *Software manual for the elementary functions,* Prentice Hall.

Algorithms and test programs for the functions, SQRT, ALOG, ALOG1O, EXP, **, SIN,COS,TAN,COT,ASIN,ACOS,ATAN,ATAN2,SINH,COSH are discussed. The test programs are available in machine readable form from the authors and IMSL.

A must for every one who does not trust the elementary function implementations in use or anyone who intends to provide some. Arithmetic pecularities of computers and their consequences for designing optimal special function software are treated in a simple and coherent way.

CODY, W.J. (1980a), *Basic concepts for computational software,* 1-23. In: P.C. Messina and A. Murli (eds.): Problems and Methodologies in Mathematical Software production. Lecture Notes in Computer Science 142, Springer Verlag.

The relation of numerical mathematics and software engineering is given for the area of approximation of functions. Arithmetic pecularities which ought not to occur are given in their simplest form. The software attributes reliability, robustness and (trans-)portability are discussed. As an illustration an implementation for $|z|$ is derived under account of the discussed criteria.

CODY, W.J. (1980b), *Implementation and testing of function software,* 24-47 in: P.C. Messina and A. Murli (eds.), Problems and Methodologies in Mathematical Software Production, Lecture Notes in Computer Science 142, Springer Verlag.

An overview of proven techniques for preparing and testing function software. The elefunt collection of CODY & WAITE (1980) is treated as an example.

CODY, W.J. (1981), *Funpack - a package of special function subroutines* TM-385 Applied Mathematics Division, Argonne National Laboratory.

The package includes subroutines to evaluate certain Bessel functions, complete elliptic integrals, exponential integrals, Dawson's integral, and the psi-function. The paper reconstructs the events and decisions leading to FUNPACK. It concludes with: "We also feel that special function programs can now be written more portable than FUNPACK without sacrificing quality."

CALGO: Collected Algorithms of the ACM.

Nowadays the background of the algorithms, and how to use them, are published in TOMS, with the complete listing of the code on microfiche. Most of the implementations are in PFORT, a subset of FORTRAN. The implementations are refereed before publication. The implementations, supplemented with remarks and certifications, are issued in ACM's looseleaf service CALGO. The machine readable versions of the algorithms can be obtained via IMSL. CALGO provides an index to the above implementations as well as implementations published elsewhere.

DITKIN, V.A., K.A. KARPOV & M.K. KERIMOV (1981), *The computation of special functions,* USSR Comput. Maths. Math. Phys., 20, 3-12.

Gives a review of methods for computing special functions, with the accent on methods used when tabulating the functions. An extensive list of references includes a lot of Russian contributions on table making.

FORD, B. (1978), *Parametrization of the environment for transportable numerical software*. ACM Trans. Math. Software, 4,2, 100-103.

In order to obtain better transportable FORTRAN 66 software the IFIP Working Group 2.5 on mathematical software defined parameters for:

a. static arithmetic characteristics (i.e. radix, mantissa length, relative precision, overflow threshold, underflow treshold, symmetric range);

b. basic input-output characteristics (i.e. standard input unit, standard output unit, standard error message unit, number of characters per record of the standard input unit, number of characters per record of the standard output unit);

c. miscellaneous characteristics (i.e. number of characters per word, page size, number of decimal digits).

REMARK. Some of the suggestions in b) and c) are catered for in FORTRAN 77 (e.g. standard units are default, character data type is provided).

FOX, P.A., A.D. HALL & N.L. SCHRYER (1978), *The PORT mathematical subroutine library,* ACM Trans. Math. Software, 4,2, 104-126.

A significant portable program library in the PFORT subset of FORTRAN 66.

FULLERTON, L.W. (1977), *Portable special function routines*, 452-483 in: W. Cowell (ed.): *Portability of numerical software,* Lecture Notes in Computer Science, 57.

Design criteria for his portable FNLIB are given and judged against CODY's and SCHONFELDER's approach. In fact FORTRAN equivalents of those in the 'handbook special functions' are treated.

FULLERTON, L.W. (1980), *A bibliography on the evaluation of mathematical functions*. CSTR 86, Bell Laboratories.

Over 250 references on the evaluation of mathematical software have been collected in this annotated bibliography. Because it includes a permuted index, one may easily find articles about specific functions. The collection has been compiled with two groups of users in mind: Those who frequently consult with scientists and engineers, and those who are developers of mathematical software and who need to examine past work before writing programs. Papers of a highly theoretical nature have been excluded from this bibliography.

GAUTSCHI, W. (1967), *Computational aspects of three-term recurrence relations,* SIAM Rev. 9, 24-82.

How to work with three-term homogeneous recurrence relations is exposed and illustrated with examples on: Bessel functions, incomplete gamma/ beta functions, Legendre functions, Coulomb functions, repeated integrals of the error function, Fourier coefficients, a Sturm-Liouville problem.

GAUTSCHI, W. (1972), *Zur Numerik rekurrenter Relationen,* Computing 9, 107-126.

Systems of linear first-order recurrence relations as well as higher order scalar recurrence relations are analyzed with respect to numerical stability. Examples of severe numerical instability are presented involving scalar first- and second-order recurrence relations. Devices for counteracting instability are indicated.

GAUTSCHI, W. (1975), *Computational methods in special functions - a survey,* 1-98 in: R. Askey (ed.) *Theory and applications of special functions,* Academic Press.

Emphasis is put on methods for computing approximations such as: best rational approximation, truncated Chebyshev expansion, Taylor series and asymptotic expansions, Padé and continued fraction approximations, representation and evaluation of approximations, linear recurrence relations, nonlinear recurrence algorithms for elliptic integrals and elliptic functions. A final paragraph is devoted to software for special functions (NATS, NAG and others).

HANDBOOK SERIES SPECIAL FUNCTIONS

This project has been started by Numerische Mathematik in a similar spirit as the series on Linear Algebra and Approximation. Published are:

Clenshaw, C.W. c.s. (1963), *Algorithms for special functions I,* 4, 403-419.

Miller, G.F. (1965), *Algorithms for special functions II,* 7, 194-196.

Bulirsch, R. (1965), *Numerical calculation of elliptic integrals and elliptic functions,* 7, 78-90.

Bulirsch, R. (1965), *Numerical calculation of elliptic integrals and elliptic functions II,* 7, 353-354.

Bulirsch, R. (1967), *Numerical calculation of the Sine, Cosine and Fresnel integrals*, 9, 380-385.

Bulirsch, R. (1969), *Numerical calculation of elliptic integrals and elliptic functions III*, 13, 305-315.

The series is not continued after these publications. However, see also Temme, N.M. (1983), *The numerical computation of the confluent hypergeometric function* U(a,b,z), 41, 63-82.

HART, J.F. et al.(1968), *Computer approximations*, John Wiley.

A good basis for developing an implementation of a special function. Design phase, general methods, choice and application of approximation, description and use of tables as well as examples are discussed. Provided in appendices are: tables of constants, conversion routines, some decimal and octal constants as well as a bibliography on published approximations.

HENRICI, P. (1977), *Computational analysis with the HP-25 pocket calculator*, John Wiley.

Shows what kind of numerical analysis can be done on a hand-hold calculator. Algorithms are given for (incomplete) gamma function, error function, complete elliptic integrals, Bessel functions (integer and arbitrary order, of the first and second kind), Riemann zeta function.

HENRICI, P. (1974,1977), *Applied and computational complex analysis*, John Wiley.

I. *Power series, integration, conformal mapping, location of zeros.*

II. *Special functions, integral transformations, asymptotics, continued fractions.*

Basic material for those who apply mathematical analysis in order to obtain the most suitable representation of a function for computation (Volume III in the series has been published but it is not related to approximation of functions).

HOUSEHOLDER, A.S. (1953), *Principles of numerical analysis*, McGrawHill.

The first chapter *The art of computation* is still of value.
In the chapters on approximation the mathematical description of the problems is still relevant, while the treated algorithms are overruled by more recent ones.

IEEE P754/82 - 10.0(1982), *A proposed standard for binary floating-point arithmetic*.

A nearly final proposal towards standardization. See also KAHAN (1983).

LUKE, Y.L. (1969), *The special functions and their approximations,* Academic Press, 2 Vols.

Volume I develops the ${}_2F_1$, ${}_1F_1$, ${}_pF_q$ and the G-functions.
Volume II is mainly concerned with approximations of these functions with particular emphasis on expansion in series of Chebyshev polynomials of the first kind, and with the approximations of these functions by the ratio of two polynomials. Tables of coefficients are given.

LUKE, Y.L. (1975), *Mathematical functions and their approximations,* Academic Press.

The author himself classified the book as a supplement to Abramowitz and Stegun. Approximations for ${}_pF_q$-named functions via analytical and numerical methods (so, no Mathieu-like functions). Chebyshev and Padé expansions are provided as well as (recursion) recipes for the computation of the coefficients of these expansions. Surveys numerical data in literature. Contains theorems, no proofs. More attractive for numerical oriented people than LUKE (1969). Emphasis is put on how to choose an expansion such that the problem is practically solvable.

LUKE, Y.L. (1977), *Algorithms for the computation of mathematical functions,* Academic Press.

As a sequel to the previous books FORTRAN programs are given in order to calculate the coefficients of the approximations.

LYUSTERNIK, L.A. et al.(1965), *Handbook for computing elementary functions,* Pergamon press.

Provides basic formulae and some coefficients for approximating elementary functions.

OLVER, F.J.W. (1978), *A new approach to error arithmetic,* SIAM J. Numer. Anal, 15,2, 368-393.

By modification of the standard definition of relative error, a form of error arithmetic is developed that is well-suited to floating-point computations. Rules are given for conversion from interval analysis to the new approach, and vice versa, both for real and complex variables. Illustrative applications include accumulation of products, quotients, sums and inner products, and the evaluation of polynomials. Also included are some new error bounds for basic operations in floating-point arithmetic.

PARTSCH, H.& R. STEINBRUGGEN (1981), *A comprehensive survey on program transformation systems,* TUM report I 8108, München.

The important aspect of transformation of software is surveyed around the CIP-L project of the Technical University of Munich.

RIVLIN, T.J. (1974), *The Chebyshev polynomials,* John Wiley.

A survey of the most important properties of Chebyshev polynomials are given along with applications with respect to interpolation, approximation, integration, and ergodic theory.

SCHONFELDER, J.L. (1976), *The production of special function routines for a multi-machine library,* Software-Practice and Experience, 6, 71-82.

The design of the special function chapter of the NAG program library is discussed.

STEGUN, I.A. & ZUCKER, R., *Automatic computing methods for special functions.*

So far four articles have been published in the Journal of Research of the National Bureau of Standards B:

Part I. *Error, probability and related functions,* 74B, 211-224, 1970.

Part II. *The exponential integral* $E_n(x)$, 78B, 199-216, 1974.

Part III. *The sine, cosine, exponential integrals and related functions,* 80B, 291-311, 1976.

Part IV. *Complex error function, Fresnel integrals, and other related functions,* 81, 661-686, 1981.

Contains FORTRAN routines. Variable precision and multi-machine approach.

TEMME, N.M. (1976), *Speciale functies*, 179-206 in: J.C.P. Bus, (red.) *Colloquium numerieke programmatuur*, deel 1b, MC-Syllabus 29.1b, Mathematisch Centrum, Amsterdam. (Dutch).

About integrals which can be recast as special functions and applications of routines, among others Bessel function routines.

VANDEVENDER, W.H. & K.H. HASKELL (1982), *The SLATEC Mathematical subroutine Library*, SIGNUM, 17,3, 16-21.

A report is given of a cooperative effort to create a mathematical subroutine library characterized by portability, good numerical technology, good documentation, robustness and quality assurance. The result is a portable FORTRAN mathematical subroutine library of over 130,000 lines of code, with on-line documentation and help facilities.

2. SOFTWARE

The construction of multi-machine program libraries gave rise to discussions on several software engineering aspects, such as

. machine parametrization (FORD (1978)),
. reliable arithmetic (IEEE P754/82-10.0),
. computer aided design, computer-surveyed and intuition-controlled programming (SCHONFELDER (1976), PARTSCH and STEINBRUEGGEN (1981)),
. multi-machine testing (CODY (1969a), SCHONFELDER (1976), CODY and WAITE (1980))
. (trans-)portability (FOX et al. (1977)).

In our opinion one should design an algorithm in a design language and transform it by correctness preserving transformation software into a portable computer language. It should be possible to use the resulting implementations in any user language instead of to transliterate portable computer language implementations into various user languages, e.g. PASCAL, ALGOL 68 or Ada.

2.1. Multi-machine program libraries

The considered libraries are: NATS (FUNPACK), IMSL, NAG, PORT and SLATEC. Summarized are: the target computers, the contents with respect to special functions, and the design philosophy.

FUNPACK (release II, 1976)

- designed for: IBM 360/370, CD 6000/7000, UNIVAC 1108/1110 and written in FORTRAN 66.
  contains implementations for:
  - exponential integrals: Ei, E_1, e^{-x} Ei(x),
  - psi function: $\psi = \Gamma'/\Gamma$,
  - Dawson integral: D,
  - Bessel functions: J_0, J_1, Y_ν,
  - Modified Bessel functions: I_0, I_1, K_0, K_1,
  - complete elliptic integrals: E,K,
  - as well as routines for error handling.
- design criteria:
  - modular, subroutine based structure (no multiple entry points),
  - robustness (error handling can be overruled by the user),
  - ultimate accuracy and efficiency,
  - not portable under ultimate accuracy and efficiency requirement,
  - accuracy profile testing and field validation.

<u>IMSL</u> (edition 9, 1982)

- available in FORTRAN for three categories of computers:
  - supercomputers (CRAY 1, CYBER 200);
  - mainframes and upper mini's (roughly 15 machine ranges)
  - mini's (e.g. DEC PDP 11).
- contains implementations for:
  - various probability functions and their inverses,
  - various special functions of mathematical physics and some inverses, some also in double precision or for complex arguments.
    The complete list is too extensive to reproduce here.
- design criteria:
  - to provide a general reliable and robust mathematical and statistical library,
  - high performance,
  - converter portable.

REMARKS.

1. The error handling routine is called UERSET with input parameters: ier and name;
   name contains the name of the subroutine where the error is detected and ier denotes either:

- a hard failure (ier > 128),
- a warning with fix error (128 ≥ ier > 64),
- a warning error (64 ≥ ier > 32) or
- an undefined error (32 ≥ ier).

More detailed information of the error is given in the documentation of the specific routine: name.

NAG (mark 10, 1983)

NAG provides program libraries in FORTRAN, ALGOL 60 and ALGOL 68. Here we concentrate on the FORTRAN library.

- available on a wide range of computers.
- contains with respect to special functions implementations for:

- circular function,	tan
- inverse circular sine and cosine,	arcsin, arccos
- hyperbolic sine, cosine, tangent and their inverses,	
- gamma and log gamma function,	Γ, ln Γ
- exponential integrals,	E_1,Ci,Si
- error function and probability functions,	erf,erfc,D,P,Q
- Fresnel integrals,	S,C
- Bessel and Airy functions (plus scaling),	$A_i,B_i,J_0,J_1,Y_0,Y_1,$
- modified Bessel functions,	I_0,I_1,K_0,K_1
- elliptic integrals,	R_C,R_F,R_D,R_J

- design criteria:
  - to provide a general, reliable and robust mathematical and statistical library in a few major languages,
  - high performance and for special functions a uniform approximation method via Chebyshev series,
  - processor portable.

REMARKS.

1. Error handling is done via the function PO1AAF.
2. The special function implementations have two parameters: the argument and an integer ifail.

ifail: entry 0, hard failure mechanism is used
1, soft failure
exit 0, no errors.
≠ 0, an error occurred; the value indicates the error as given in the documentation.

3. In the documentation the condition of the function is displayed in clear graphs.

PORT (version 1, 1977)

PORT is a general portable program library written in the PFORT-subset of FORTRAN

. available on various machines, as the name indicates.

. contains special function implementations for

- tangent, inverse cosine and sine (single and double precision),
- hyperbolic sine, cosine and their inverses as well as the inverse hyperbolic tangent (single and double precision),
- complex double precision exponent and logarithm,
- Bessel functions: $J_k(z)$,
- modified Bessel functions: $I_k(z)$.

. design criteria:

- to provide a general, reliable and robust mathematical library as transportable as possible via parametrization of the environment: parameter values are provided for various machines,
- dynamic storage allocation is simulated via an array in common,
- centralised error handling.

REMARKS.

1. Error handling is done via the principal error routine SETERR:
   - just remember the error (recovery mode),
   - print and stop,
   - print, dump and stop.

   The status of the recovery mode can be handled via ENTSRC.
2. The programs do not contain (in their calling sequences) a parameter to indicate, on a return from a subprogram, whether an error has occurred. an error number can be retrieved via the function NERROR. Error messages are enumerated in the documentation and provided via SETERR, where the first 72 characters of the messages are printed.

SLATEC (version 1, 1982)

SLATEC stands for the cooperation of Sandia, Los Alamos, Air Force Weapons Laboratory, Technical Exchange Committee. The computing centers of Sandia National Laboratory, Lawrence Livermore National Laboratory and the National Bureau of Standards joined the project, VANDEVENDER & HASKELL (1982).

With respect to special functions the library consists of FNLIB (FULLERTON (1977)), FUNPACK and exponential and Bessel functions from AMOSLIB of SANDIA Laboratories.

. design criteria:

- FORTRAN 66 portability based on the PFORT-verifier,
- good numerical technology, programming style and documentation,
- reliable and robust,
- uniform processing of error conditions.

2.2. Local program libraries

From a historical point of view the multi-machine program libraries emerged from local activities, e.g.: IMSL from SSP, FUNPACK (the NATS activity) from Argonne National Laboratory, NAG from NPL and HARWELL. Below we summarise some local program libraries with respect to their special functions chapter.

ARGONNE

- Apart from FUNPACK they have made available on their IBM 360/370:

circular functions	sin, cos, tan, cotan,
inverse circular functions	arcsin, arccos
hyperbolic functions	sinh, cosh,
exponential integrals	FUNPACK, also complex
gamma function	FUNPACK, γ, Γ,ln Γ, χ^2
error function	FUNPACK, erf, erfc,
Bessel functions	J_r, Y_r,
modified Bessel functions	$e^{-x}I_\nu$, e^xK_ν
Coulomb wave function	F_L, G_L
Coulomb phase shift	σ_L
Legendre functions	Q_n^m, dQ_n^m/dz
angular momentum coefficients	
zeta function	ζ, $\zeta-1$,
elliptic integrals	E,K.

CERN library (March, 1976)

- Available among others on CD-CYBER and contains FORTRAN routines, often also in double precision, for:

exponential integrals	E_1,Ei, Si, Ci,
gamma function ($\mathbb{R}$,$\mathbb{C}$)	Γ, ln Γ, ψ, quotient Γ functions

error functions ($\mathbb{C}$), probability functions	erf, D, P, Q
Fresnel integrals	S, C,
Bessel functions ($\mathbb{R}$, $\mathbb{C}$)	J_r, Y_r,
modified Bessel functions	I_ν, K_ν,
Coulomb wave functions	
Legendre functions	
Θ-functions, Jacobi elliptic function	
complete elliptic integrals	E, K,
Whittaker functions	$M_{k,m}$,
Fermi-Dirac function	
Struve functions	H_0, H_1.

HARWELL (August, 1977)

- Available on IBM 360/370 and contains FORTRAN routines for:

exponential integrals	E_1
gamma function	Γ, B,
error function ($\mathbb{C}$)	erf, D,
Fresnel integrals	C, S,
complete elliptic integrals	E, K, Π,
incomplete elliptic integrals	1st and 2nd kind
Bessel functions	J_0, J_1, Y_0, Y_1,
modified Bessel functions	I_0, K_0, I_1, K_1,
spherical Bessel functions	j_n,
Kelvin functions	ber, bei, ker, kei, ber', bei', ker', kei',
angular momentum coefficients	

NUMAL (see HEMKER (1981))

- Written in ALGOL 60 for CD-CYBER (elsewhere converted into FORTRAN under supervision of P. Wynn) and contains implementations for:

inverse circular functions	arcsin, arccos,
exponential integrals	E_i, E_1, $E_n(x)$, α_n, Si, Ci,
gamma function	Γ, ln Γ, γ, B, B_x,
error function	erf, erfc,
Fresnel integrals	C, S,
Bessel functions	J_r, Y_r,
modified Bessel functions	I_ν, K_ν,

spherical Bessel functions	j, y,
Airy functions (also: zeroes of)	Ai, Bi, Ai', Bi'.

Implementations for the probability functions: binomial, χ^2, F, hypergeometric, normal, Smirnov, Students T, non-central T, Poisson, and their inverses are provided in the Statistical library STATAL.

2.3. Published software

The Index to the Collected Algorithms of the ACM contains references to software published in roughly ten journals. The ACM publishes software in their Transactions-series, where Transactions on Mathematical Software (TOMS) is of special concern for us. Software published in TOMS is validated, and available in machine-readable form from the ACM distribution service.

Software related to special functions published in TOMS, up to 1983, is listed below.

TOMS	Algorithm number	Item
1.4	498	Airy functions using Chebyshev series approximations
3.1	511	CDC 6600 subroutines IBESS and JBESS for Bessel functions $I_\nu(x)$ and $J_\nu(x)$, $x \geq 0, \nu \geq 0$
3.3	518	Incomplete Bessel function I_0: The Von Mises' distribution
3.3	521	Repeated integrals of the coerror function
5.4	542	Incomplete Gamma function
6.3	556	Exponential integrals
7.2	571	Statistics for Von Mises' and Fisher's distribution of directions: $I_1(x)/I_0(x)$, $I_{1.5}(x)/I_{0.5}(x)$
7.3	577	Algorithms for incomplete elliptic integrals
9.2	597	Sequence of modified Bessel functions of the first kind
9.2	599	Sampling from Gamma and Poisson distributions
9.4	609	A portable FORTRAN subroutine for the Bickley functions
9.4	610	A portable FORTRAN subroutine for derivatives of the psi function

REMARKS.

.If one intends to use software published in TOMS, we advise to look in the loose-leaf collection of the ACM for additional REMARKS or CERTIFICATIONS, done by the 'scientific community' after the implementation has been published. On the other hand if one uses software published in TOMS and detects some flaws it is worthwhile for the 'scientific community' to contribute a REMARK or CERTIFICATION.

.An index to program collections is also provided by Guide to Available Mathematical Software (GAMS). It is intended for the National Bureau of Standards Staff and it gives an overview with respect to: NBS Core Math. Libraries, Mathware and the libraries IMSL, NAG and PORT.

.A general bibliography on numerical software is published by EINARSSON (1977) with an update of chapter 16 in EINARSSON (1979).

.(Added in print) IMSL has available a new FORTRAN library, the SFUN/ LIBRARY for evaluating the following special functions: elementary functions, trigonometric and hyperbolic functions, exponential integrals, gamma functions, error functions and Bessel functions. It will be available initially for FORTRAN 77 compilers on IBM, VAX, DEC 10/20, CDC and DG 32-bit Eclipse.

II. GENERAL ASPECTS OF COMPUTING FUNCTIONS

In this chapter we will discuss certain topics that play a fundamental role in the subsequent chapters. In section 1 we mention some aspects of error analysis for the computation of functions, in section 2 we classify algorithms and describe the general structure of implementations. Section 3 deals with recurrence relations where the first order recurrence relation is treated in detail. Two-term recurrence relations are treated from a numerical algebraic as well as from a pragmatic point of view, where pecularities of recurrence relations, arising from computing special functions, are exposed. Section 4 gives an introduction to continued fractions and section 5 pays attention to some basic properties of hypergeometric functions.

1. ERROR ANALYSIS

In this section we point out that a user needs only to consider carefully the effect of perturbation of the argument of a function if the designer of an implementation takes care of sufficiently accurate and well-conditioned approximations and benign computational processes.

In discussing the sources of error in the computation of functions we will consider:

a. the effect of perturbation of the argument;

b. the effect of approximation of a function by more elementary functions;

c. the effect of finite precision arithmetic.

Generally speaking, the designer of a function routine takes care of (b) and (c) while a user has to be aware of (a). In order to understand this and to be aware of the assumptions, we will pose the problem and quantify the qualitative aspects (a), (b) and (c).

The problem is: given

$\tilde{z}$, an approximation of z,

Af, a well-conditioned approximation of f depending on a set of coefficients $\{a_k\}$ and an argument z,

Af_c, a benign implementation of Af,

then the value of $Af_c(\{a_k\};\tilde{z})$ and an estimate of

$$(1.1)\qquad |f(z)-Af_c(\{a_k\};\tilde{z})| \text{ or } |f(z)-Af_c(\{a_k\};\tilde{z})|/|f(z)|$$

are desired. In the sections 1.2 and 1.3 we will return to well-conditioned approximations and benign implementations.

To estimate (1.1) we consider

$$(1.2)\qquad |f(z)-Af_c(\{a_k\};\tilde{z})| \le |f(z)-f(\tilde{z})| + |f(\tilde{z})-Af(\{a_k\};\tilde{z})| + \\ + |Af(\{a_k\};\tilde{z}) - Af_c(\{a_k\};\tilde{z})|.$$

The terms in the upper bound correspond to the qualitative aspects (a), (b) and (c); they will be treated in the sections 1.1, 1.2 and 1.3, respectively.

1.1. Perturbation of the argument

For a function holomorphic within γ, $\gamma = \{t \mid |t-z| = r\}$, we have the Taylor formula

$$(1.3)\qquad f(\tilde{z}) - f(z) = (\tilde{z}-z)f'(z) + \frac{(\tilde{z}-z)^2}{2\pi i}\oint_\gamma \frac{f(t)}{(t-\tilde{z})^2(t-z)}\,dt.$$

In first order we obtain for the absolute and relative errors the well-known estimates

$$(1.4)\qquad |f(\tilde{z}) - f(z)| \simeq |\tilde{z}-z|\,|f'(z)|$$

$$|f(\tilde{z}) - f(z)|/|f(z)| \simeq |\tilde{z}-z|\,|f'(z)/f(z)|,$$

with the relative error amplification

$$\frac{|f(\tilde{z})-f(z)|/|f(z)|}{|\tilde{z}-z|/|z|} \simeq \left|\frac{zf'(z)}{f(z)}\right| .$$

EXAMPLE. Suppose $z = 100$, to three correct significant figures. Then the relative error in $f(z) = e^z$ is .5, or 50%. Hence the value of $f(z)$ has no significant figures.

REMARK. In order to estimate the errors, an estimate of $|f'(z)|$ must be available. The NAG library provides a graph of $|f'(x)|$, for x in the relevant parts of $\mathbb{R}$.

1.2. Approximation of a function by more elementary functions

By approximation of a function we mean replacing the mapping

$$f: z \to f(z) \tag{1.5}$$

by

$$Af: \{a_k\}, z \to Af(\{a_k\};z). \tag{1.6}$$

Choices are to be made with respect to:

- approximating form and size (e.g. polynomial or rational form),
- representation of approximating form (e.g. representation of a polynomial as a sum of Chebyshev polynomials or powers of the independent variable). This has to be done such that for some prescribed ε:

$$|f(z) - Af(\{a_k\};z)| < \varepsilon \text{ or } |f(z) - Af(\{a_k\};z)|/|f(z)| < \varepsilon, \tag{1.7}$$

the so-called residual or truncation error, and

$$Af \text{ is well-conditioned with respect to } \{a_k\}. \tag{1.8}$$

As a measure of condition of a representation with respect to $\{a_k\}$ we introduce the *condition function* C as the 1-norm of the vector of the relative derivatives of Af with respect to the parameters, i.e.,

$$C(\{a_k\};z) := \sum_k \left| \frac{a_k}{Af(\{a_i\};z)} \frac{\partial Af(\{a_i\};z)}{\partial a_k} \right|. \tag{1.9}$$

The maximum of the condition function times $|Af(\{a_k\};z)|$ over all relevant z is taken as the *condition number* κ. (We suppose that for these definitions $Af(\{a_k\};z)$ is bounded away from zero in the z-domain.)

Different approximations to f may yield different condition numbers. If we have two approximations, say

$$\mathrm{Af}(\{a_k\};z) \quad \text{and} \quad \mathrm{Af}(\{b_k\};z),$$

within the same z-domain, we can compare the condition numbers, say κ_a and κ_b. If $\kappa_a < \kappa_b$ then we call Af *better conditioned with respect to* $\{a_k\}$ *than to* $\{b_k\}$. The best conditioned approximation of a number of approximations is characterized by the lowest condition number. A well-conditioned form is characterized by a sufficiently low condition number, which possibly reflects a compromise between accuracy, efficiency and portability.

In polynomial approximation, on $[-1,1]$, the condition number of the power sum representation, $P_n(x) = \Sigma\, a_k x^k$, equals the condition number of the Chebyshev representation, $P_n(x) = \Sigma b_k T_k(x)$, if the coefficients $\{a_k\}$, and hence $\{b_k\}$, are strictly alternating or of the same sign (NEWBERY (1974)).

The condition function (1.9) may be used for representations in terms of an infinite set $\{a_k\}$. As an example consider the expansions

$$e^{-x} = \sum_{n=0}^{\infty} (-1)^n x^n/n!, \quad e^{-x} = 1/\sum_{n=0}^{\infty} x^n/n!, \qquad x \in [0,x_0],$$

where x_0 is a positive number. The condition functions for these representations are e^{2x} and 1, respectively, and the condition numbers are e^{x_0} and 1, respectively.

1.3. Finite precision arithmetic

Given an approximation $\mathrm{Af}(\{a_k\};z)$, a well-conditioned computational problem, we must take into account the aspects of finite precision arithmetic, in particular in view of different computational processes. Our approach is inspired by BAUER (1974), who considers computational graphs in computations. For instance, the evaluation of a^2-b^2 may be performed by either of the processes $(a-b)(a+b)$ and $(a^2) - (b^2)$, yielding two different computational graphs. Another example in point is the evaluation of a polynomial by using Horner's rule. It gives a different computational graph than the process that computes the polynomial straightforwardly. We write $\mathrm{Af}_c(\{a_k\};z)$ if the approximation Af is computed according to a given computational graph c. In a graph several intermediate results arise, giving intermediate rounding errors. Loosely speaking, we call a process (a computational graph) *benign* if the effect of intermediate rounding errors does not spoil the computational aim. Intermediate results obtained by multiplication or division need not to be considered.

As in the previous subsection it is possible to give a more rigorous definition of the concept benign, as was done for the condition function C of an approximation Af. First we introduce the *condition function of a computational process* Af_c as the 1-norm of the vector of the relative derivatives of $Af(\{a_k\};z)$ with respect to the intermediate results. Then we introduce the *condition number* κ_c of the computational graph Af_c as the maximum of the condition function of Af_c times $|Af_c(\{a_k\},z)|$ over all relevant z. The computational graph is called *benign* if the condition number of the computational graph Af_c is smaller than the condition number of the computational problem Af.

REMARK. In numerical linear algebra the concept of growth of intermediate results in a computation is used in order to decide upon which algorithm is best with respect to error propagation.

EXAMPLE 1. Consider the evaluation of the polynomial

$$Af(\{a_k\};x) = \sum_{k=0}^{3} a_k x^k$$

as an approximation for a function f. The condition function of the approximation is

$$C(\{a_k\};x) = \sum_{k=0}^{3} |a_k x^k| / |\sum_{k=0}^{3} a_k x^k|$$

and the condition function of the computational graph based on Horner's rule is given by

$$(|a_3x^3+a_2x^2| + |a_3x^3+a_2x^2+a_1x|)/|\sum_{k=0}^{3} a_k x^k| .$$

EXAMPLE 2. The evaluation of a^2-b^2. The condition function of this problem is

$$(|a \frac{\partial(a^2-b^2)}{\partial a}| + |b \frac{\partial(a^2-b^2)}{\partial b}|)/|a^2-b^2| = 2 \frac{|a^2|+|b^2|}{|a^2-b^2|} .$$

The evaluation may be performed by the processes $(a-b)(a+b)$ and $(a^2) - (b^2)$ with condition functions

$$2 \quad \text{and} \quad (|a^2| + |b^2|)/|a^2-b^2|,$$

which are both smaller than the condition function of the problem.

REMARK. The above ideas are candidates for automatization. Work in this direction is indicated by BAUER (1974) and realized by MILLER (1975), MILLER & SPOONER (1978) with respect to the behaviour of absolute errors and LARSON et al. (1983) with respect to the behaviour of relative error.

2. SOFTWARE

In this section we classify algorithms and describe the general structure of implementations.

2.1. Algorithms and implementations

In mathematical software for function approximations algorithms may be classified according to the input parameters:

(2.1) - the argument;

(2.2) - the argument and the precision.

In the *(nearly) maximum precision class* (2.1) the approximation $Af(\{a_k\};z)$ is determined such that the approximation error (1.7) is less than the machine accuracy ε. Commonly a uniform approximation is predetermined based on an error bound for the worst case; when several approximation approaches are combined - say Taylor series and asymptotic series - uniform approximations are commonly used for each approach. We call this a nearly class because the resulting error (1.1), in general slightly exceeds the machine precision due to finite precision arithmetic, even for exact z. In the *variable precision class* (2.2) approximations are used for which the approximation error (1.7) is easily available. Often, (especially when the approximating error alternates) the differences of the n-th and the (n+1)-st approximation majorates the approximation error; this property is commonly used as a stopping criterion. However, when dealing with monotonic convergence or with finite precision arithmetic this criterion may not be reliable.

The implementations based upon (2.1) are generally structured in:

(2.4) - checking of argument z;

(2.5) - selection of appropriate $\{a_k\}$ as a function of z (segment or domain) and initialization;

(2.6) - evaluation of appropriate approximation $Af_c(\{a_k\};z)$.

The implementations based upon (2.2) are generally structured in:

(2.7) - checking of argument z and the precision δ,

(2.8) - evaluation of $Af_c(\{a_k\};z)$ such that an approximation of the truncation error is less than δ.

Two elaborations of module (2.8) are in use:

(2.8.1) - selection of appropriate $\{a_k\}$ as a function of z (segment or domain) and δ (precision); one could think of evaluation of part of a finite Chebyshev sum - for example as determined by the procedure Set (CLENSHAW, c.s. (1963)) - or one could think of evaluation of the appropriate minimax approximations while the coefficients of respectively the finite Chebyshev sum or the minimax approximations are included in the program for the precision range.

(2.8.2) - no selection of appropriate $\{a_k\}$ as a function of z is made a priori.

Further refinement of the modules with respect to portability may be achieved with either a special target computer in mind (advantage may be taken of, or measures may be taken against, some machine-environmental-peculiarities; this approach was common in local program libraries) or for a range of computers (standards and subsets are used; NAG approach). Of course one could think of a range of computers as a sum of special target computers (NATS approach).

2.2. Testing

In our opinion testing of software is verifying by a human being the correctness of different design stages of an implementation. RUTISHAUSER (1976) distinguishes for the creation of mathematical software the design stages:

- . formal algorithm: a description of the principal flow of a calculation;
- . naive program: an unambigious definition of the calculation process is given, where program correctness is empirically obtained via checking of a limited number of argument values;
- . strict program: apart from round-off error effects the program is proven to be correct;
- . numerically safe: the errors in the results are within proven bounds.

The various states of a program can be placed in the total activity of mathematical problem solving in the following way.

↓ numerical mathematics - applied mathematics ↑

Starting point	Tasks	Region of competence
mathematical problem		analysis
	discretisation	
discrete mathematical problem		algebra
	developing numerical method	
formal algorithm		numerical calculation in exact arithmetic
	taking care of finite precision arithmetic	
naive program		numerical calculation in finite precision arithmetic
strict program		sequential safety
strict program with a priori or a posteriori error bounds		numerical safety

Nowadays test activities, at least with respect to approximation of functions, deal with the 'naive program'-level. On this level the technique is automated by generation, via possibly different algorithms, of multi-length tables by CODY (1973,1975b) and SCHONFELDER (1976). Consistency tests are treated by NEWBERY & LEIGH (1971).

The creation of strict programs via pre- and postconditions and Hoare-like loop invariants has not been done in the considered software.

The creation of numerically safe programs has not yet emerged, while first order error bounds are provided in the documentation of some considered implementations.

With respect to error bounds one could think of a first order estimate and a rigorous estimate, where the latter is generally pessimistic. During the checking of obtained values one could classify the errors into the classes red, orange, green. Where red indicates a true error because it exceeds the rigorous bounds, orange indicates a possible error because it is within the rigorous bounds but exceeds the first order bounds, and green indicates an acceptable error because it is within the first order bounds.

In our discussion of some special function implementations we will concentrate our efforts on the 'naive program'-level

- are the used approximations accurate enough?
- are the used stopping criteria provable correct?
- is the program readible; does it look correct?
- can we classify the implementations as 'good in principle'?

Only after positive answers on the above questions by an initiated worker one can consider to

either

perform the costly job of stringent tests in the sense of Cody and Schonfelder

or

proof the program correctness and to provide bounds for the numerical errors.

Only in the last case a numerically safe program is obtained while for practical purposes the former approach is sufficient.

3. LINEAR RECURRENCE RELATIONS

The behaviour of linear scalar recurrence relations in finite precision arithmetic is described in terms of first order matrix-vector recursions. We shall treat two-term recurrence relations (1×1-matrix) and three-term relations (2×2-matrix) separately. Our main tool is the concept of stability of the problem: amplification of perturbations of input data into the answer. For this class of problems it is convenient to consider rounding errors as perturbations of the input data. The amplification is quantified by the earlier introduced condition (§1.2.). Wherever appropriate we make use of geometric concepts in order to abstract from details and to strengthen the intuition. Furthermore, we shall make use of general knowledge of the solution when it concerns special functions. A half page introduction with practical information about stability directions for a few classical examples is given in ABRAMOWITZ & STEGUN (1964, p.XII); see also §3.3. A state of the art survey is given in GAUTSCHI (1975).

3.1. First order inhomogeneous scalar recurrence relations

A thorough treatment of the stability, with emphasis on the effect of perturbations of the initial value, is given by GAUTSCHI (1972a). His graphs of ρ_n make clear whether we must prefer the forward to the backward recurrence or consider starting somewhere in between, eventually as a function of the (real) argument of the approximated function. We shall introduce Gautschi's ρ_n as part of the condition of the problem; this quantity reflects the stability due to the initial value neglecting other effects. Moreover, we shall introduce a new quantity σ_n, which reflects the stability due to the initial value and the inhomogeneous terms while other effects are neglected. The examination of the stability of a recursion can be done by demonstrating that ρ_n or σ_n are large, an instable recursion, and if not by proving that the condition is small.

Introduction

As an introduction we shall talk about

$$(3.1.1)\qquad \begin{aligned} &y_{j+1} = a_j y_j + b_j, \qquad j = 0,1,\ldots \\ &y_0 \text{ given.} \end{aligned}$$

Recurrence relations of this type play a role, for example, in calculations of the incomplete gamma function with the exponential integrals as a special case. Sometimes recurrence relations are used in the forward direction, as in (3.1.1), and sometimes they are used in the backward direction:

$$(3.1.2)\qquad \begin{aligned} &y_j = (y_{j+1}-b_j)/a_j, \qquad j = n,n-1,\ldots \\ &y_n \quad \text{given,} \end{aligned}$$

provided of course that $a_j \neq 0$. We like to stress that mathematically the same values $\{y_j\}_0^n$ are defined, but that the algorithms differ, especially in finite precision arithmetic.

In order to decide a priori upon which algorithm is to be preferred in finite precision arithmetic, we will derive macroscopic quantities which govern the stability of linear first-order inhomogeneous recurrence relations.

The formulation of the problem

The recurrence relation (3.1.1) may be stated as: given

$$y_0,\ \{a_k\}_{k=0}^{n-1},\quad \{b_k\}_{k=0}^{n-1}\ ;$$

obtain

$$(3.1.3)\qquad f_n = \left(\prod_{j=0}^{n-1} a_j\right) y_0 + \sum_{j=0}^{n-1} \left(\prod_{k=j+1}^{n-1} a_k\right) b_j$$

in finite precision arithmetic as accurate as possible. (This formula may be derived from (3.1.1) by the variation of parameters technique (HAMMING (1971)). The first term of the right hand side equals the solution of the homogeneous problem; the second term is a particular solution).

Stability

The stability of the problem may be characterized by the condition as introduced in formula (1.9), i.e.,

$$(3.1.4)\qquad c^n = c^n_{y_0} + c^n_{b_k} + c^n_{a_k}$$

with

$$(3.1.4a)\qquad c^n_{y_0} = \left|\frac{y_0}{f_n}\frac{\partial f_n}{\partial y_0}\right| = \left|\prod_{j=0}^{n-1} a_j\right| |y_0/f_n|, \text{ due to the initial value,}$$

$$(3.1.4b)\qquad c^n_{b_k} = \sum_{k=0}^{n-1}\left|\frac{b_k}{f_n}\frac{\partial f_n}{\partial b_k}\right| = \sum_{k=0}^{n-1}\left|\prod_{j=k+1}^{n-1} a_j\right| |b_k/f_n|,$$

due to the inhomogeneous terms,

$$(3.1.4c)\qquad c^n_{a_k} = \sum_{k=0}^{n-1}\left|\frac{a_k}{f_n}\frac{\partial f_n}{\partial a_k}\right| = \sum_{k=0}^{n-1}\left|\left(\prod_{j=0}^{n-1} a_j\right)y_0 + \sum_{j=0}^{k-1}\left(\prod_{i=j+1}^{n-1} a_i\right)b_j\right|/|f_n|,$$

due to the coefficients $\{a_k\}$.

The condition may macroscopically be represented in terms of the solution of the homogeneous recurrence: $f_n^{(h)}$, as

$$(3.1.5)\qquad c^n = |f_n^{(h)}/f_n|\left\{1 + \sum_{k=0}^{n-1}\left\{|b_k/f_{k+1}^{(h)}| + |1 + \sum_{j=0}^{k-1} b_j/f_{j+1}^{(h)}|\right\}\right\}.$$

The absolute value of the quotient of the homogeneous solution (with the same initial value) and the inhomogeneous solution is Gautschi's ρ_n. So from the perturbation point of view ρ_n reflects the stability of f_n due to a perturbation in the initial value, say g: $\rho_n = c_g^n$. In the sequel we will use ρ_n as a symbol to denote the relative amplification of a perturbation of the initial value into the answer f_n, given a particular recurrence relation, independent of whether we call it a forward or a backward recurrence. If we consider ρ_n large - a so called ill-conditioned *initial-value* problem - we may pose another problem, for example by recurring in the opposite direction; the latter generally has a different ρ_n. For the calculation of f_n by (3.1.1) - we call this forward - we have for ρ_n of the forward problem (superscripted by f)

$$\rho_n^f = \left|\left(\prod_{j=0}^{n-1} a_j\right)y_0/f_n\right|,$$

while for the calculation of f_n by (3.1.2) by starting at y_{n+k} - we call this backward (superscripted by b) - we arrive at

$$\rho_n^b = \left|\left(\prod_{j=n}^{n+k-1} a_j^{-1}\right) y_{n+k}/f_n\right|.$$

From $\{\rho_j^f\}_{j=n}^{n+k}$ we may obtain ρ_n^b, the backward amplification factor, as

$$\rho_n^b = \rho_n^f/\rho_{n+k}^f .$$

We appreciate Gautschi's graphs of ρ_n, given a particular recurrence relation, because from these we may obtain by the above formula the ρ's of the recurrence relation in the opposite direction.

Another representation of c^n is obtained if we use the solution of the absolute recurrence: $f_n^{(a)}$ (with input parameters $|f_0|,\{|a_k|\},\{|b_k|\}$), in (3.1.4)

$$(3.1.6) \qquad c^n = f_n^{(a)}/|f_n| + |f_n^{(h)}/f_n| \sum_{k=0}^{n-1} |1+\sum_{j=0}^{k-1} b_j/f_{j+1}^{(h)}|,$$

with $f_j^{(h)}$ again not zero, of course.
In the sequel we will denote the quotient of the solution of the absolute recurrence and the solution of the given recurrence by σ_n, i.e.

$$(3.1.6a) \qquad \sigma_n = f_n^{(a)}/|f_n|.$$

From the perturbation point of view σ_n reflects the stability of f_n due to perturbations of the initial value as well as perturbations of the inhomogeneous terms, because the sum of the right-hand sides of (3.1.4a) and (3.1.4b) equals $f_n^{(a)}/|f_n|$; so σ_n is the symbol to denote the amplification of perturbations of the initial value and of the inhomogeneous terms into the answer f_n. This amplification is realistic when all perturbations are roughly equal. From these representations we easily obtain the inequalities

$$(3.1.7) \qquad c^n \geq |f_n^{(a)}/f_n| \geq |f_n^{(h)}/f_n|.$$

If ρ_n or σ_n is large we have an ill-conditioned problem. Usually this is stated in a geometrical sense (see also section II.3.2): an ill-conditioned *initial-value problem* is characterized by the dominance of $f_n^{(h)}$ over f_n (ρ_n is large); an ill-conditioned *inhomogeneous* (*initial-value*) *problem* is characterized by the dominance of $f_n^{(a)}$ over f_n (σ_n is large). The latter is not generally known, e.g. it is not mentioned in section II.3.2. Finally, we will consider a problem suitably conditioned when c^n is tolerable; this introduces the context.

For an absolute recurrence relation, say the absolute version of (3.1.3), an attainable upper bound for the condition is given by

(3.1.8) $\quad c^n \le n+1.$

Derivation: (3.1.4a) + (3.1.4b) contributes 1
(3.1.4c) contributes at most n-times the contributions of
(3.1.4a) and (3.1.4b); a well-conditioned problem .

REMARK. With matrix recurrence relations the relative values of the solution of an absolute recurrence are of importance.

For the problem of evaluating a polynomial as a power sum: the condition may, after confluence of all a_j into x, conveniently be bounded below by

(3.1.9) $\quad c^n \ge \{f_n^{(a)} + |x\, df_n(x)/dx|\}/|f_n|,$

where we recognize the contribution of the derivative; for this particular recurrence we have for the absolute recurrence and for σ_n

$$f_n^{(a)} = \sum_{k=0}^{n-1} |b_{n-k}\; x^k|$$

$$\sigma_n = \sum_{k=0}^{n-1} |b_{n-k}\; x^k| / |\sum_{k=0}^{n-1} b_{n-k}\; x^k|.$$

So, σ_n equals the 1-norm of the relative derivatives with respect to the coefficients in the power sum representation.

REMARKS.

1. We like to stress that so far we have considered the stability of the problem and not yet any particular computational graph nor the effects of finite precision arithmetic. The condition of the problem gives (first-order) information about the effect of perturbation of the input data - initial value and recurrence coefficients - into the solution; finite precision arithmetic or bad algorithms can only make things worse. Even if the input data are exact representable in the machine and there are no measurement errors in the picture, the above introduced concepts are still of interest. Namely for the class of problems defined by the recurrence algorithms, the intermediate rounding errors due to finite precision arithmetic can be considered in a natural way as perturbations of the initial data; backward analysis is easily applied.

For example, the recurrence relation (3.1.1) in finite precision arithmetic (the tilde denotes finite precision operations)

$$\begin{aligned} &y_{j+1} = a_j \mathbin{\tilde{*}} y_j \mathbin{\tilde{+}} b_j, \quad j = 0,1,\ldots \\ &y_0 \quad \text{given,} \end{aligned} \tag{3.1.10}$$

may be stated (in first order) as (the tilde denotes the perturbated coefficients which yield the same result as (3.1.10))

$$\begin{aligned} &y_{j+1} = \tilde{a}_j * y_j + \tilde{b}_j, \quad j = 0,1,\ldots \\ &y_0 \quad \text{given} \end{aligned} \tag{3.1.11}$$

in exact arithmetic, with $a_j = \tilde{a}_j(1+\delta+\delta_1)$, $b_j = \tilde{b}_j(1+\delta_1)$, where we assumed for the machine operators (with tilde): $x \mathbin{\tilde{*}} y = x * y(1+\delta)$, $x \mathbin{\tilde{+}} y = (x+y)(1+\delta_1)$ with $\max(|\delta|,|\delta_1|) \le \varepsilon$ (= machine precision).
So the contributions (3.1.4b) and (3.1.4c) govern also the effect of intermediate rounding errors. Commonly, (3.1.11) is replaced by

$$y_{j+1} = a_j * y_j + b_j + \Delta_j,$$

with Δ_j the local error. We think our approach for this class of problems simpler, because, the local errors are absorbed in the recurrence coefficients and so we only have to look at the effects of perturbations of the input data; i.e., we only have to concentrate on the *condition of the computational problem* and not on the condition of the computational graph. The perturbations of the recurrence coefficients due to finite precision arithmetic is of the order of the machine precision: the recurrence algorithm is benign.

2. One can ask whether the positive recursion with bound for the condition (3.1.8) is well-conditioned or not. Such pin-point questions can easily be circumvented by going back to the perturbation idea; the condition is only a macroscopic tool. The condition as a 1-norm is a suitable tool when all perturbations, initial or due to interpretation of rounding errors, are of equal order of magnitude. For rounding errors in recurrence relations this is the case, so we think this norm convenient. For

example (3.1.9) only states that perturbations are linearly amplified. Whether this is tolerable or not depends upon the circumstances. One has to decide upon this oneself given the particular situation; generally a linear amplification bound is considered harmless.

3. An example of a stable initial value problem but an unstable inhomogeneous problem, is given by

$$(\ldots((\delta)-K_1)+K_2)-K_3) +\ldots+ K_{2n}, \quad K_i \sim K > 0,$$

where $\rho_{2n} = 1$ and $\sigma_{2n} \sim 2nK/\delta$; σ_{2n} can be made as large as we please.

4. If the contribution to the condition is mainly due to (3.1.4a), and we judge this intolerable, we can look for another problem formulation: a terminal value problem for example; i.e., (3.1.2). In closed form the solution may be represented by

$$(3.1.12) \qquad f_n = \left(\prod_{j=n}^{n+k-1} a_j^{-1}\right) y_{n+k} - \sum_{i=n}^{n+k-1} \left(\prod_{j=i}^{n+k-1} a_j^{-1}\right) b_i,$$

with k suitable chosen. This is the so-called backward recurrence; y_{n+k} as starting value must be known or a perturbation of it is harmless in f_n, so that we can take nearby values: (asymptotic) estimates or crudely 0. As a special class of problems we have the absolute recurrence relations with their stable properties. Absolute recurrences may just be given or recognized as the opposite recurrence from a recurrence relation with all $\{a_k\}$ and $\{y_k\}$ positive and all $\{b_k\}$ negative. An error in the starting value of an absolute (or positive) recurrence is damped because the (inhomogeneous) solution dominates the solution of the homogeneous recurrence relation. The effectiveness of this damping determines k: fast damping induces a small k; slow damping needs a large k.

5. In estimating the condition of a problem defined by a recurrence relation with non-constant coefficients, we can sometimes - for the so defined slowly-varying recursions - consider the general problem as a perturbation of a problem with constant coefficients. On the other hand some recurrence relations have variable coefficients which exhibit a high regularity; for these recurrence relations one can look for - and we will in the sequel - handsome representations of (3.1.4a), (3.1.4b) or (3.1.4c).

6. When dealing with a polynomial it is convenient to have a tool which can be used in order to decide upon its representation. In this remark we restrict ourselves to the problem of evaluation of the representations:

$$\text{a power sum} \quad , \; P_n(x) = \sum_{k=0}^{n} a_k \, x^k ;$$

$$\text{a Chebyshev sum} \quad , \; P_n(x) = \sum_{k=0}^{n} b_k \, T_k(x) .$$

Our tool is: the representation with smallest 1-norm of the coefficient vector is best. (Indeed, the relative perturbations are amplified by $\sum_{k=0}^{n} |a_k \, x^k|$ or $\sum_{k=0}^{n} |b_k \, T_k(x)|$, which in turn are uniformly bounded by the 1-norms.) With this tool we easily understand Newbery's (1974) experimental result (see §1.2) as well as some spread results:

CLENSHAW (1962):

$$J_0(x) \sim \sum_{k=0}^{12}{}' \, b_{2k} \, T_{2k}(x/8) = \sum_{k=0}^{12} a_{2k}(x/8)^{2k},$$

where the Chebyshev sum representation is to be preferred because

$$\|b_{2k}\|_1 = 1 < \|a_{2k}\|_1 = 427.$$

HART c.s. (1968):

$$T_6(x) = 32x^6 - 48x^4 + 18x^2 - 1,$$

where the explicit power sum representation is not to be preferred, because

$$\|1\|_1 < \|a_k\| = 99.$$

RUTISHAUSER (1968); $(T_k^*(x) = T_k(2x-1))$:

$$1 - 13.7x + 67.5x^2 - 153x^3 + 162x^4 - 64.8x^5 =$$

$$= -(.522T_1^*(x) + .352T_3^*(x) + .126T_5^*(x)),$$

where the shifted Chebyshev sum representation is to be preferred, because

$$\|b_k\|_1 = 1 < \|a_k\| = 462.$$

NEWBERY (1974): If $\{a_k\}$ are of the same sign or strictly alternating then

$\|b_k\|_1 = \|a_k\|_1$; no preference, so for efficiency reasons the power sum can be used.

GAUTSCHI (1972b) introduced the condition number of the coordinate map associating to each polynomial its coefficients with respect to a system of orthogonal polynomials. Let

$$M_n\colon \mathbb{R}^n \to P_{n-1},$$

i.e., with $(u_0,\dots,u_{n-1})$ we associate $\sum_{k=0}^{n-1} u_k\, p_k(x)$, with $\{p_k\}_{k=0}^{\infty}$ a set of orthogonal polynomials. Then

$$\mathrm{cond}_\infty\, M_n = \|M_n\|_\infty \|M_n^{-1}\|_\infty.$$

From the perturbation point of view we have

$$\frac{\|\Delta P_{n-1}\|_\infty}{\|P_{n-1}\|_\infty} \geq \frac{1}{\mathrm{cond}_\infty M_n} \frac{\|\Delta u\|_\infty}{\|u\|_\infty} .$$

We did not follow Gautschi's approach because it concentrates on uniform results for a class of problems, while we are more concerned with tools for particular problems which do reflect the (known) qualitative behaviour. Gautschi's ideas are worked out in GAUTSCHI (1972b, for orthogonal polynomials; 1979a, for polynomials in power form; 1978, for polynomials).

7. The second term in (3.1.9) is inherent in the polynomial and cannot be minimized. We like to remark, however, that a perturbation of the argument of the function, which is approximated by a polynomial, was already considered (see §1.1). But, because of confluence of all a_k into x we are not surprised to see again the derivative of the approximating function - the polynomial - in the stability of the problem of evaluating the approximation.

8. For the homogeneous recurrence relation the condition is

$$c^n = n+1.$$

For this simple case the effect of relative perturbations of the input data δf_0 and $\{\delta a_k\}$, is in first order easily given by

$$|\delta f_n| = |\delta f_0 + \sum_{k=0}^{n-1} \delta a_k| \le c^n * \varepsilon,$$

where $\varepsilon = \max\{\delta f_0, \delta a_k\}$. We see at once that the bound is attained if all relative perturbations of the input data are the same. Rounding errors behave not that systematic. In order to get a more realistic estimate of the effect of rounding errors we could think of an effective machine precision or introduce an effective condition notion.

EXAMPLES.

1. The following example about evaluating a polynomial is theoretical. It is constructed in order to elucidate the use of different algorithms or properly speaking: to contrast the forward problem with the backward problem. Let

$$\begin{aligned} y_{j+1} &= 2 * y_j - 1, \quad j = 0,1,2,\dots,n-1 \\ y_0 &= 1 - 2^{-2n} + \varepsilon. \end{aligned} \tag{3.1.13}$$

The solution is given by

$$f_j = 2^j y_0 - \sum_{k=0}^{j-1} 2^k = (y_0-1)2^j+1, \quad j = 0,1,2,\dots,n. \tag{3.1.14}$$

The backward formulation is given by

$$\begin{aligned} y_j &= (y_{j+1}+1)/2, \quad j = n+k-1,\dots,n+1,n \\ y_{n+k} &= (\varepsilon-2^{-2n})2^{n+k}+1; \end{aligned} \tag{3.1.15}$$

we have taken k = n, so the (terminal) starting value is

$$y_{2n} = \varepsilon 2^{2n}.$$

In table 3.1 we have enumerated the results of: the direct method (3.1.14), $f_j^{(d)}$; the forward problem (3.1.13), $f_j^{(f)}$; and the backward problem (3.1.15), $f_j^{(b)}$. We have taken $n = k = 10$, ε = machine precision, and relative perturbated starting values: $\tilde{y}_0 = 1 - 2^{-2n}$ and $\tilde{y}_{2n} = \varepsilon 2^{2n}(1-\varepsilon)$. The erroneous digits with respect to the direct method are underlined.

j	direct: $f_j^{(d)}$	forward: $f_j^{(f)}$	backward: $f_j^{(b)}$
0	.99999 90463 2569	.99999 90463 256$\underline{8}$	.99999 90463 2569
1	.99999 80926 5138	.99999 80926 513$\underline{7}$	.99999 80926 5138
2	.99999 61853 0276	.99999 61853 027$\underline{3}$	.99999 61853 0276
3	.99999 23706 0553	.99999 23706 05$\underline{47}$	.99999 23706 055$\underline{2}$
4	.99998 47412 1105	.99998 47412 10$\underline{94}$	.99998 47412 1105
5	.99996 94824 2210	.99996 94824 21$\underline{88}$	.99996 94824 2210
6	.99993 89648 4420	.99993 89648 4$\underline{375}$	.99993 89648 4420
7	.99987 79296 8841	.99987 79296 8$\underline{750}$	.99987 79296 8841
8	.99975 58593 7682	.99975 58593 7$\underline{500}$	.99975 58593 7682
9	.99951 17187 5364	.99951 17187 5$\underline{000}$	.99951 17187 536$\underline{3}$
10	.99902 34375 0728	.99902 34375 0$\underline{000}$	.99902 34375 072$\underline{7}$
11	.99804 68750 1455	.99804 68750 $\underline{0000}$	.99804 68750 1455
12	.99609 37500 2910	.99609 37500 $\underline{0000}$	.99609 37500 2910
13	.99218 75000 5821	.99218 75000 $\underline{0000}$	.99218 75000 582$\underline{0}$
14	.98437 50001 1642	.98437 50000 $\underline{0000}$	.98437 50001 164$\underline{1}$
15	.96875 00002 3283	.96875 0000$\underline{0}$ $\underline{0000}$	.96875 00002 3283
16	.93750 00004 6566	.93750 0000$\underline{0}$ $\underline{0000}$	.93750 00004 6566
17	.87500 00009 3132	.87500 0000$\underline{0}$ $\underline{0000}$	.87500 00009 3132
18	.75000 00018 6265	.75000 000$\underline{00}$ $\underline{0000}$	.75000 00018 626$\underline{4}$
19	.50000 00037 2529	.50000 000$\underline{00}$ $\underline{0000}$	.50000 00037 2529
20	.00000 00074 5058	.00000 000$\underline{00}$ $\underline{0000}$	.00000 00074 5058

Table 3.1

Discussion

The contribution to the condition of f_j due to a perturbation in y_0 is

$$(3.1.16) \qquad \rho_j \sim 2^j, \qquad j = 0,1,2,\ldots,n,$$

so we expected the forward recursion to grow erroneously. A perturbation of $h_0 = y_0 - 1$, the linearly transformed initial value, is amplified by

$$(3.1.17) \qquad \rho_j \sim 2^{j-2n}.$$

(Note the difference in the condition due to the simple change of variable!) The backward problem is a positive recurrence, so we expected it benign;

an error in f_{n+k} is damped by $2^{-(n+k-j)}$ in f_j.

For the problem of evaluating a polynomial other algorithms can be considered. TRAUB & SHAW (1974) introduced a family of splitting algorithms for the power sum representation.

In stead of

$$(3.1.18) \qquad P_n(x) = (\dots((a_n x+a_{n-1})x+a_{n-2})x+\dots+a_1)x + a_0$$

they considered

$$(3.1.19) \qquad P_n(x) = (\dots((a_n x^q+a_{n-1}x^{q-1}+\dots+a_{n-q})x^{q+1} + \dots$$
$$+ (a_{2q+1}x^q+a_{2q}x^{q-1}+\dots+a_{q+1}))x^{q+1} +$$
$$+ (a_q x^q+a_{q-1}x^{q-1}+\dots+a_0),$$

with q+1 a divisor of n+1. The advantage of this approach is that the linear amplification factor, say n, can be reduced to the sum of factors of n+1. Furthermore, this approach is also advantageous when all derivatives are needed because the number of multiplications is of order $O(n)$, while the complete Horner is of $O(n^2)$. The problem of summation of numbers may be considered as a special case of polynomial evaluation. BABUSKA (1972) reported the benign nature, ${}^2\log n$, of the repeated splitting-summation computational graph. However, his example

$$S_n = \sum_{k=1}^{n} 1/k$$

should have been compared with the "backward" process

$$(\dots((1/n+1/(n-1)) + 1/(n-2))+\dots+1/2)+1.$$

The conditions of S_n with 1 as starting value, c_1^n – i.e. summation of decreasing terms – and with 1/n as starting value, $c_{1/n}^n$ – i.e. summation of increasing terms – behave as

$$(c_1^n = 1 + \{\sum_{j=2}^{n} \sum_{k=1}^{j} 1/k\}/S_n) >$$

$$(c_{1/n}^n = 1 + \{\sum_{j=2}^{n} \sum_{k=0}^{j-1} 1/(n-k)\}/S_n) \sim 1+n/\log n.$$

This illustrates the general rule of thumb: keep intermediate results small. Another elaboration of this general rule is the summation technique of HAMMING (1971): order the positive terms and negative terms; merge these rows by keeping the intermediate results as close to zero as possible.

2. This example is given by GAUTSCHI (1972a).. Let

$$f_n = n!\ (e^x - e_n(x)), \quad n = 0,1,2,.. \tag{3.1.20}$$

with

$$e_n = \sum_{k=0}^{n} x^k/k!. \tag{3.1.21}$$

Gautschi enumerates illustratively for $x = 1$ the horrible results obtained by the forward recurrence

$$\begin{aligned} y_j &= j * y_{j-1} - x^j, \qquad j = 1,2,... \\ y_0 &= e^x - 1. \end{aligned} \tag{3.1.22}$$

The condition of this initial value problem is bounded below by

$$\rho_n = \left| \frac{e^x - 1}{e^x - e_n(x)} \right| . \tag{3.1.23}$$

For x away from zero we see immediately $\lim_{n\to\infty} \rho_n = \infty$; an unstable initial value problem.

The recurrence relation (3.1.22) is easily posed backwards

$$y_{j-1} = (y_j + x^j)/j, \tag{3.1.24}$$

which for $x > 0$ is an absolute recurrence and so a benign problem. Moreover, a perturbation in the (terminal) starting value y_t is damped by $j!/t!$ in f_j. As an illustration we have depicted ρ_n for $x = 5, 10, 15, 20, 25, 30$. See Figure 1.

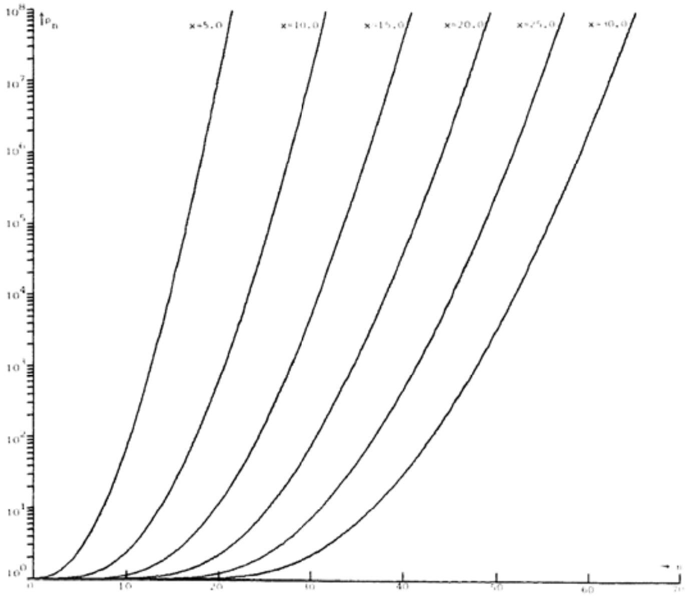


Figure 1. ρ_n of (3.1.23)

3. Calculation of exponential integrals

The exponential integrals

$$E_n(z) = \int_1^\infty t^{-n} e^{-zt} \, dt \tag{3.1.25}$$

obey the recurrence relation

$$\begin{aligned} &y_{k+1} = -z/k \; y_k + e^{-z}/k, \quad k = 1,2,3,\ldots \\ &y_1 \quad \text{given.} \end{aligned} \tag{3.1.26}$$

The contribution to the condition due to perturbations in the initial value is given by

$$\rho_n = \left| \frac{z^n E_1(z)}{n! E_{n+1}(z)} \right| \sim |z|^{n-1} |n+z+1| / n! . \tag{3.1.27}$$

For some $x \in \mathbb{R}^+$ GAUTSCHI (1972a) depicted the graphs as given in Figure 2.

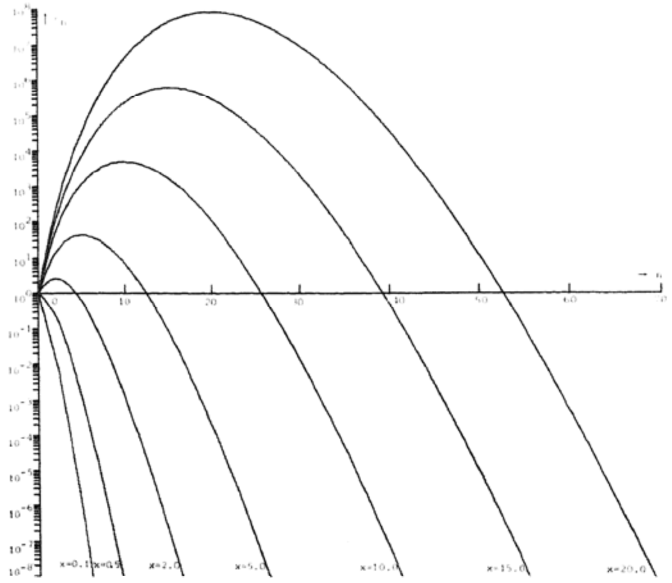


Figure 2. ρ_n of (3.1.27)

The stability for $z \in \mathbb{C}$ is similar, because ρ_n is approximately a function of $|z|$. So the above graphs may be seen as iso-$|z|$-curves. The curves have a maximum for $\ell = [|z|]$. The graphs suggest to start at y_ℓ and recur down the ρ_n-hill on either side. So we obtain from (3.1.26) either of the problems

$$(3.1.28a) \quad y_{k+1} = -z/k \; y_k + e^{-z}/k, \qquad k = \ell, \ell+1, \ldots, n$$

$$(3.1.28b) \quad y_k = -k/z \; y_{k+1} + e^{-z}/z, \qquad k = \ell-1, \ell-2, \ldots, 1$$

with

$$y_\ell \text{ given}, \qquad\qquad \ell = [|z|].$$

After some calculations we arrived at the bound for the condition

$$c^n \leq \tfrac{1}{2}|n-\ell| * (|n-\ell|+1).$$

A class closely related to the exponential integral is

$$\alpha_n = \int_1^\infty t^n e^{-zt} dt, \qquad \mathrm{Re}\ z > 0,\ n = 0,1,2,\ldots$$

The integrals obey the recurrence relation

$$\alpha_{k+1} = ((k+1)\alpha_k + e^{-z})/z, \qquad k = 0,1,2,\ldots$$

$$\alpha_0 = e^{-z}/z.$$

For $z \in \mathbb{R}^+$ this is a positive recurrence relation and thus a benign problem. For $\{z \mid \mathrm{Re} > 0,\ z \in \mathbb{C}\}$ we obtained with respect to the condition

$$\sigma_n = e_n(|z|)/|e_n(z)|.$$

Namely,

$$\alpha_n = e^{-z} n!\ e_n(z)/z^{n+1}$$

and the solution of the recurrence relation with absolute values equals

$$\alpha_n^{(a)} = |e^{-z}| n!\ e_n(|z|)/|z|^{n+1},$$

with

$$e_n(z) = 1 + z + z^2/2! + \ldots + z^n/n!.$$

The limit

$$\lim_{n\to\infty} \sigma_n = e^{|z|}/e^x, \qquad z = x+iy,$$

is a growing function of y, so we expect the recursion to become unstable when $|y|$ increases. As an illustration we have depicted σ_n as iso-$|\mathrm{im}\ z|$-graphs with Re z = 1 and Im z = 1,5,10,15,20,25,30. See Fig.3.

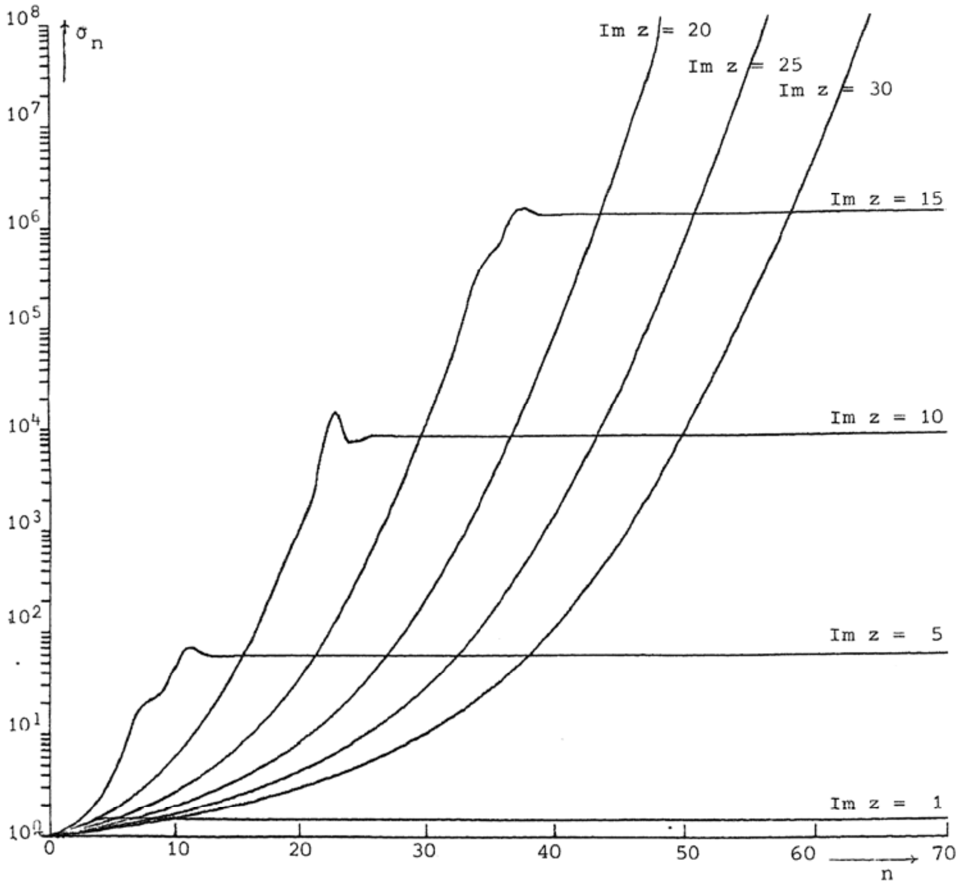


Figure 3. σ_n for the computation of α_n

The integrals

$$\beta_n = \int_{-1}^{1} t^n e^{-zt} dt$$

obey the recurrence relation

$$\beta_{k+1} = ((k+1)\beta_k - (-1)^k e^z - e^{-z})/z, \quad k = 0,1,...$$

$$\beta_0 = (e^z - e^{-z})/z.$$

In closed form the solution can be represented by

$$\beta_n = \frac{n!}{z^{n+1}} (e^z e_n(-z) - e^{-z} e_n(z)).$$

The solution of the recurrence relation with absolute values can be bounded below by

$$\beta_n^{(a)} \geq \frac{n!}{|z|^{n+1}} 2 \sinh(x)\ e_n(|z|), \qquad z = x + iy.$$

As a consequence σ_n can be bounded below by

$$\sigma_n \geq 2|\sinh(\mathrm{Re}\ z)| \frac{e_n(|z|)}{|e^z e_n(-z) - e^{-z} e_n(z)|}. \tag{3.1.29}$$

For the second factor in the lower bound we have

$$\frac{e_n(|z|)}{|e^z e_n(-z) - e^{-z} e_n(z)|} \sim \frac{(n+1)!}{|z|^{n+1}} e^{|z|} \quad \text{for } n \to \infty; \tag{3.1.30}$$

so the recurrence is (ultimately) unstable. As an illustration we have depicted (3.1.30) for $|z| = 1,2,3,5,7$ with Re $z = 1$, as iso-$|z|$-graphs, in Figure 4.

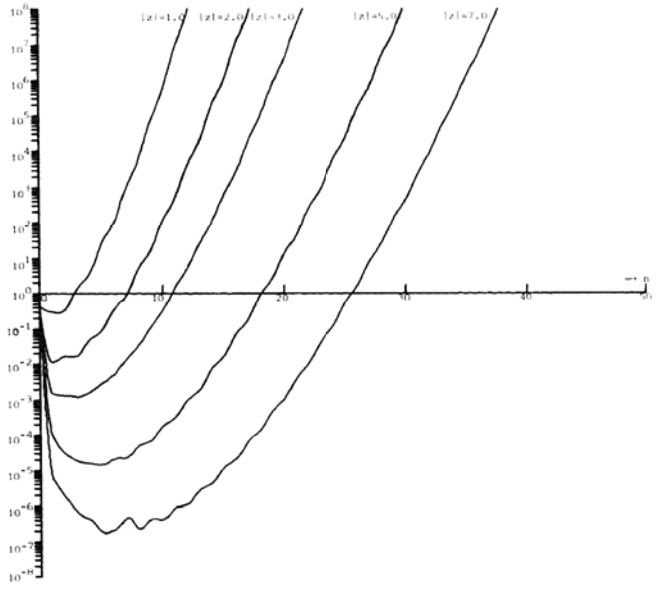


Figure 4. Graphs of (3.1.30]

REMARK. In STEGUN & ABRAMOWITZ (1956) it is suggested that forward recursion is stable if the function is increasing as the index increases. According to this principle the calculation of $E_n(x)$ for small arguments should be stable in the backward direction. This is not the case.

4. Let

$$f(a,b) = [\Gamma(a) - \Gamma(a+b)]/b, \quad a > 0, \quad b \geq 0.$$

The computation of f(a,b) is straightforward if b is bounded away from zero. If b is small, however, the above representation of f is not stable. For an application of f(a,b) we refer to GAUTSCHI (1979b), where it is needed in the computation of the incomplete gamma functions. Gautschi computes f(1,b) by using a Taylor expansion of the gamma function. Here we analyse the recursion, of which $y_0 = f(a,b)$,

$$(3.1.31) \qquad \begin{aligned} y_k &= (y_{k+1}+\Gamma(a+b+k))/(a+k), \quad k = N-1, N-2, \ldots, 0 \\ y_N &= \frac{\Gamma(a+b+N)}{b}\left(\frac{\Gamma(a+N)}{\Gamma(a+b+N)} -1\right). \end{aligned}$$

The starting value may accurately be obtained by (III 2.12) or its modifications. The stability with respect to the initial value, y_N, is given by

$$\rho_0 = \left|\frac{y_N}{y_0}\frac{\partial y_0}{\partial y_N}\right| = \left|\frac{\Gamma(a)}{\Gamma(a+N)}\frac{\Gamma(a+N)-\Gamma(a+b+N)}{b}\Big/\frac{\Gamma(a)-\Gamma(a+b)}{b}\right|.$$

(This may be obtained via (3.1.12) with $a_j = a + j$ and $b_j = -\Gamma(a+b+j)$.) For small b we arrive at

$$\rho_0 \sim |\psi(a+N)/\psi(a)| = \mathcal{O}(\log N), \qquad N \to \infty.$$

The stability which also accounts for the inhomogeneous terms, for $b \to 0$, is given by

$$\sigma_0 \sim \rho_0 + N/|\psi(a)|.$$

The conclusion is that (3.1.31) is a mildly unstable inhomogeneous problem.

3.2. General aspects of three-term recurrence relations

We give a survey of problems and methods involved with recursions, with the emphasis on three-term recurrence relations. Stability of solutions is discussed and some algorithms are given for the computation of minimal (or dominated) solutions.

3.2.1. Introduction

Recursions play an important role in special functions. Of course, the three term recurrence relation is a well-known tool for calculating functions of mathematical physics, such as Bessel functions. But also processes like determining partial sums of a series or evaluating polynomials with Horner's scheme, exploit recursions.

In this section we consider several aspects of recursions which are in particular important from a computational point of view. The general *second order scalar recursion* (or difference equation) has the form

$$(1.1) \qquad \xi_{i+1} = a_i\xi_i + b_i\xi_{i-1} + c_i, \qquad i \geq 1.$$

This recurrence relation is called *homogeneous* if $\forall_i c_i = 0$ and *inhomogeneous* otherwise. A *solution* of (1.1), i.e. a sequence $\{\xi_0,\xi_1,\ldots\}$ satisfying (1.1) for all i, will be denoted by $\{\xi_i\}$.

In order to be able to study more general recursions we introduce *matrix vector recursions,* viz.

$$(1.2) \qquad x_{i+1} = A_i x_i + r_i, \qquad i \geq 0$$

where $\forall_i x_i \in \mathbb{R}^n$ (for some fixed integer n) and $\forall_i A_i$ is a square matrix. As for the scalar case $\{x_i\}$ will denote a solution of (1.2).

In §3.2.2 we shall consider the constant scalar recursion, which can be used as a kind of model problem. In order to get insight into the problems which are involved in the numerical computation of solutions, it is very useful to study the growth of solutions of (1.1) or (1.2), which is therefor the subject of §3.2.3. Armed with such information it will be possible to understand the effect of (rounding) errors made during the recursion, as will be shown in §3.2.4. As it will turn out that straightforward use of (1.1) or (1.2), for i.e., the initial value problem, is unstable for certain solutions (which are of great interest), other methods have to be used. In §3.2.5

we shall give a brief discussion of several such methods.

For general papers dealing with this subject, see e.g. GAUTSCHI (1967, 1972a, 1975), MATTHEIJ (1977) & OLIVER (1968a).

3.2.2. The scalar second order constant recursion

Consider the recursion

$$(2.1)\qquad \xi_{i+1} = a\xi_i + b\xi_{i-1}, \qquad i \geq 1.$$

As is known (NÖRLUND (1924, p.295) the general solution can be found using the so-called *characteristic equation*, given by

$$(2.2)\qquad \tau^2 = a\tau + b.$$

Let (2.2) have roots α and β with $|\alpha| < |\beta|$, then the general solution of (2.1) is

$$(2.3)\qquad \xi_i = p\alpha^i + q\beta^i, \qquad p,q \in \mathbb{R}\,.$$

Obviously the solutions $\{\phi_i\} = \{\alpha^i\}$ and $\{\psi_i\} = \{\beta^i\}$ constitute a basis for the two dimensional solution space. We have

$$(2.4)\qquad \lim_{i\to\infty} \frac{\phi_i}{\psi_i} = 0.$$

Therefore $\{\psi_i\}$ is called a *dominant* solution and $\{\phi_i\}$ a *dominated* (or minimal) solution, cf. GAUTSCHI (1967).

It is immediately clear that any solution of (2.1) written in the form (2.3) and with $q \neq 0$ will dominate $\{\phi_i\}$. If we use the recursion (2.1) in practice, we inevitably make rounding errors. A complete and detailed analysis of their effects on the computed solution is a tedious and laborious task. However, investigating the effect of a single rounding error, made during the computation, at stage j say, is often satisfactory to get insight into the well - posedness of the problem. For simplicity we take $j = 0$; so assume ξ_0 is perturbed by a quantity ε_0. Denote the solution of (2.1) with initial values $\bar{\xi}_0 = \xi_0 + \varepsilon_0$ and $\bar{\xi}_1 = \xi_1$ by $\{\bar{\xi}_i\}$, then we clearly have

$$(2.5)\qquad \bar{\xi}_{i+1} - \xi_{i+1} = a(\bar{\xi}_i-\xi_i)+b(\bar{\xi}_{i-1}-\xi_{i-1}).$$

Hence the error $\{\bar{\xi}_i - \xi_i\}$ is propagated as a solution of (2.1), so we have

$$\bar{\xi}_i - \xi_i = \bar{p}\phi_i + \bar{q}\psi_i. \tag{2.6}$$

Substituting $\bar{\xi}_0 - \xi_0 = \delta_0$ and $\bar{\xi}_1 - \xi_1 = 0$ it can be seen that $\bar{q} \neq 0$. This means that the perturbation ε_0 generates a dominant solution.

A similar statement also holds for perturbations made at other stages. Therefore (2.1) is not an appropriate recursion to compute a *dominated* solution, at least if relative precision is desired. In order to show that (2.1) is suitable for computing a *dominant* solution, we have to take contaminations of errors into account. Thinking of a computer with floating point arithmetic, however, the rounding errors generally are relatively small with respect to a *computed* iterand (i.e. if no serious cancellation occurs) and therefore only generate small additional components of $\{\psi_i\}$; whence the total relative error will remain small.

The previous analysis also applies to inhomogeneous recursions. Consider

$$\xi_{i+1} = a\xi_i + b\xi_{i-1} + c_i, \qquad i \geq 1. \tag{2.7}$$

Let $\{\chi_i\}$ be a particular solution of (2.7), then the general solution will be given by

$$\xi_i = \phi_i + \psi_i + \chi_i. \tag{2.8}$$

Perturbing ξ_0 as above, we see that the difference between the computed $\bar{\xi}_i$ and ξ_i itself also obeys (2.5) for all i and that the perturbations are propagated as solution of the *homogeneous* part! Therefore (2.7) can be suitable for the computation of $\{\xi_i\}$, if $\{\xi_i\}$ is not dominated by $\{\phi_i\}$ or $\{\psi_i\}$.

3.3.3. General linear recursions; estimating the growth of solutions

The three-term recursion of the previous section can also elegantly be described using some linear algebra. Define

$$x_i = \begin{pmatrix} \xi_i \\ \xi_{i+1} \end{pmatrix} \tag{3.1}$$

and

$$(3.2) \qquad A = \begin{pmatrix} 0 & 1 \\ b & a \end{pmatrix}.$$

Then (2.1) can be written as

$$(3.3) \qquad x_{i+1} = Ax_i, \qquad i \geq 0.$$

Using this relation, we see that (2.1) is mathematically equivalent to power iteration with the matrix A and initial vector x_0. It is known that x_i will asymptotically have the direction of the subdominant eigenvector; the latter problem, however, is known to be numerically unstable. Now if the coefficients are varying (cf. (1.1)) then we can define

$$(3.4) \qquad A_i = \begin{pmatrix} 0 & 1 \\ b_i & a_i \end{pmatrix}.$$

It will not be surprising perhaps that if the coefficients are only mildly varying, there also exists a solution of which the iterands are directionally close to successive dominant eigenvectors of the A_i and likewise a solution close to successive subdominant eigenvectors, cf. MATTHEY (1975), VAN DER SLUIS (1976).

The special form of the A_i - viz. the companion matrix - is of no importance of course. More generally, if the A_i are *slowly varying* n-th order matrices then it can be shown that under some conditions there exist solutions whose directions are close to successive eigenvectors corresponding to a certain eigenvalue of the A_i; cf. MATTHEIJ(1976), VAN DER SLUIS (1976), SCHÄFKE (1965). We give a qualitative formulation below. Consider the recursion

$$(3.5) \qquad x_{i+1} = A_i x_i, \qquad i \geq 0.$$

PROPERTY 3.6. For each i let $\lambda_i(1),\ldots,\lambda_i(n)$ denote the eigenvalues of A_i with $|\lambda_i(1)|>\ldots>|\lambda_i(n)|$ and $e_i(1),\ldots,e_i(n)$ the corresponding eigenvectors. If for each j and all i, $\lambda_i(j)$ is close to $\lambda_{i+1}(j)$ and sufficiently separated from $\lambda_{i+1}(\ell)$, $\ell \neq j$, and a similar statement holds for the directions of the eigenvectors, then there exists solutions $\{x_i(1)\},\ldots,\{x_i(n)\}$with $x_i(j) \approx \lambda_i(j)\ldots\lambda_0(j)e_i(j)$.

For the solutions $\{x_i(j)\}$ of (3.5) we have

PROPERTY 3.7. For each j and ℓ for which $1 \le j < \ell \le n$ we have

$$\lim_{i\to\infty} \frac{\|x_i(\ell)\|}{\|x_i(j)\|} = 0,$$

i.e., $\{x_i(j)\}$ dominates $\{x_i(\ell)\}$.

The solutions $\{x_i(1)\},\dots,\{x_i(n)\}$ in 3.6 constitute a basis of the solution space, and are called a *fundamental system*. It is often convenient to think of such a basis in terms of eigenvalues and eigenvectors. The requirements of 3.6 may be weakened such that only a separation between $\lambda_i(1),\dots,\lambda_i(k)$ on one hand and $\lambda_i(k+1),\dots,\lambda_i(n)$ on the other hand, and likewise of the corresponding invariant subspaces of $\mathbb{R}^n$, is assumed. The solution space can then be divided in a subspace whose elements dominate the elements of the complementary subspace (cf. MATTHEIJ(1980)).

For the inhomogeneous recursion

$$(3.8) \qquad x_{i+1} = A_i x_i + r_i,$$

the general solution is lying in a linear variety to be found from a fundamental system of the homogeneous part, in matrix notation $\{\phi_i\}$ (i.e., successive columns of the Φ_i constitute a solution of (3.5)) on one hand and some particular solution of (3.8), $\{y_i\}$ say, on the other hand, so

$$(3.9) \qquad x_i = \Phi_i v + y_i, \qquad v \in \mathbb{R}^n \text{ a constant vector.}$$

By considering all possible v in (3.9) we can try to find out if there is any particular solution which has a growth character different from any complementary solution (i.e. of the homogeneous part). Of course this depends on the r_i. We give a simple first order example.

Consider

$$(3.10) \qquad x_{i+1} = \frac{1}{3} x_i + r_i, \qquad i = 0,1,\dots .$$

The solution of the homogeneous part equals $\{(\frac{1}{3})^i\}$, apart from a constant factor being the initial value. The general solution of (3.10) is given by

$$(3.11)\qquad x_i = \sum_{j=1}^{i} (\tfrac{1}{3})^{i-j}\, r_{j-1} + x_0(\tfrac{1}{3})^{i}.$$

If e.g. $r_i \equiv 1$, then the homogeneous solution is dominated by an particular solution. But if e.g. $r_i = (\frac{1}{9})^i$ then with $x_0 = -\frac{9}{2}$, x_i will be equal to $-\frac{9}{2}(\frac{1}{9})^i$, which is therefore dominated by $\{(\frac{1}{3})^i\}$.

In order to find out what the growth character is of such a more or less "pure" particular solution the following trick may be helpful in cases where solutions can be expected with an exponential growth type (as in the slowly varying case above) (cf. MATTHEIJ (1977,§4)). In relation to (3.8) define

$$(3.12)\qquad \eta_0 = \frac{\|r_0\|^2}{\|r_1\|},$$

$$(3.13)\qquad \left(\begin{array}{c} x_1 \\ \hline \eta_1 \end{array}\right) = \left(\begin{array}{c|c} A_1 & \dfrac{\|r_1\|}{\|r_0\|^2}\, r_0 \\ \hline \emptyset & \dfrac{\|r_1\|}{\|r_0\|} \end{array}\right) \left(\begin{array}{c} x_0 \\ \hline \eta_0 \end{array}\right),$$

$$(3.14)\qquad \left(\begin{array}{c} x_{i+1} \\ \hline \eta_{i+1} \end{array}\right) = \left(\begin{array}{c|c} A_i & \dfrac{\|r_i\|}{\|r_{i-1}\|}\, r_i \\ \hline \emptyset & \dfrac{\|r_i\|}{\|r_{i-1}\|} \end{array}\right) \left(\begin{array}{c} x_i \\ \hline \eta_i \end{array}\right)$$

The recursion (3.13), (3.14) is of order n+1. The corresponding matrices have the eigenvalues of A_i plus an eigenvalue equalling the factor to which $\|r_i\|$ increases with respect to $\|r_{i-1}\|$. If this recursion is slowly varying then there certainly are solutions corresponding to the eigenvalues as indicated in the beginning of this section. Note that the additional eigenvalue in the examples above equals 1 and $\frac{1}{9}$ respectively, which nicely corresponds to the results above.

So far we have tried to bring some ordering in the solution space. Generally it will be very difficult for a *certain* solution of which e.g. only x_0 is given, to find out whether it is a dominated solution or not, at least theoretically. For computational methods cf. MATTHEIJ (1982,.§6). However even though a certain solution may be classifiable from purely mathematical point of view, as a dominant one, it may have a subdominant character if one just looks at the first few iterands (cf. (3.9) with such a choice of v, having only very small coordinates corresponding to the dominant solutions). Although such situations may seem pathological one should be warned since any numerical procedure that is only suitable either for dominant or for dominated solutions, will fail then.

REMARK. With respect to special functions one often knows the behaviour of the solutions of the related recurrence relations.

3.2.4. The effect of errors made during the recursion

We saw in §3.2.2 for the second order scalar recursion that (rounding) errors are propagated as solutions of the homogeneous part of the recursion, in first order. In the general case the situation is the same. For the *stability* of the recursion we may distinguish between *absolute* and *relative stability*, by which we mean that the effects of small perturbations are not large or not large with respect to the solution respectively. A more precise definition would require a specification of "small" and "large". However, it is not unusual to have such a more or less qualitative notion only, and it is quite suited for our (limited) purposes. In order to find out whether or not the recursion is good natured we either have to investigate the solutions of the homogeneous part absolutely, or in relation to the desired solution. Absolute stability then means that the solutions of the homogeneous part are bounded (have growth factors not exceeding 1), cf. stability theory for discretizations of O.D.E. Relative stability then implies that no complementary solution dominates the desired solution, or even nicer, any complementary solution is dominated so that errors are damped out relatively. We shall give some examples.

EXAMPLE 4.1. The recursion (3.10) is absolutely stable since the solutions of the homogeneous part damp out. If $r_i \equiv 1$, then it is also relatively stable for any particular solution. However, if $r_i = (\frac{1}{9})^i$, then the recursion is not relatively stable for the dominated solution $\{-\frac{9}{2}(\frac{1}{9})^i\}$.

EXAMPLE 4.2. "Summation of a strongly decaying series." Let $S = \Sigma_0^\infty a_i$; assume that S is of order unity. We are interested in $S(N) = \Sigma_0^N a_i$, which is a sufficient approximation to S (both in absolute and relative sense). Consider the following two algorithms

$$(4.3) \qquad s_0 = 0, \quad s_{i+1} = s_i + a_i, \quad S(N) = s_{N+1}$$

$$(4.4)\qquad T_0 = 0,\quad T_{i+1} = T_i + a_{N-i},\quad S(N) = T_{N+1}.$$

For the desired solution of (4.3), viz. $\{S_i\}_{i=0}^N$, we have $S_i \approx 1$, whereas for the solution of (4.4), viz. $\{T_i\}_{i=0}^N$, we see that T_i is strongly increasing. In both cases the solutions of the homogeneous part are $\{1\}$. Hence the stability properties of (4.3) make it preferable to use (4.4). This once more explains why one should sum up such a series with the smallest term (cf. also II.3.1).

For a more quantitative analysis one has to add up all effects of the rounding errors and their contaminations. This may be a laborious job. However, the order of the error (and this is usually sufficient for a practical user) is often predictable. If we denote the relative computer accuracy by ξ, then the relative error in the computed x_i (for the stable case) is of the order $\xi\ \Sigma_{j=0}^{i} \max(\|A_j\|,\|r_j\|)$. If the A_i and r_i resp. do not differ too much in norm for i varying we therefore may say that the rounding error is realistically estimated by $\max_j(\|A_j\|,\|r_j\|)i\xi$.

3.2.5. Methods to approximate solutions for which the initial value problem is not stable

We shall restrict ourselves, for shortness sake, to *relative stability* questions from now on. Those who are interested in absolute stability can easily adapt the subsequent results using the remarks made about this subject in §3.2.4. Another reason for considering the relative case especially is that the slowly varying recursions that we introduced in §3.2.3 have solutions of exponential type, which makes *relative precision* in the approximations to a more natural question.

Now assume that a solution $\{x_i\}$ satisfying

$$(5.1)\qquad x_{i+1} = A_i x_i + r_i,$$

and x_0 given, is dominated by *complementary* solutions (i.e. of the homogeneous part). We shall consider several algorithms for the computing (or rather approximating) such a solution.

3.2.5.1. Miller's algorithm

We return to the tree term recursion, cf. §3.2.2. As is known many special functions obey such a relation, which however often is unstable for increasing index. An extensive study can be found in GAUTSCHI (1967). The classical approach to overcome this is Miller's algorithm, named after MILLER (1952), who first introduced backward recursion for the computation of Bessel functions. The basic idea can easily be demonstrated with the help of the constant recursion of §3.2.2. From (2.1) we obtain (if $b \neq 0$)

$$\xi_{i-1} = -\frac{a}{b}\,\xi_i + \frac{1}{b}\,\xi_{i+1}, \tag{5.2}$$

or in matrix notation

$$\begin{pmatrix}\xi_{i-1}\\ \xi_i\end{pmatrix} = A^{-1}\begin{pmatrix}\xi_i\\ \xi_{i+1}\end{pmatrix}. \tag{5.3}$$

For N suitably large and

$$\begin{pmatrix}\xi_N\\ \xi_{N+1}\end{pmatrix} \neq \kappa\begin{pmatrix}1\\ \beta\end{pmatrix}, \quad (\kappa \in \mathbb{R})$$

we would find a $\binom{\xi_0}{\xi_1}$ which has almost the direction of the subdominant eigenvectors of A, viz. $\binom{1}{\alpha}$. Mathematically this *inverse iteration* is equivalent to backward recursion. Again we have a counter part for the variable case by considering dominated and dominant solutions rather than sequences of iterates of the eigenvectors. If some suitable "end" vector $x_N^{(N)}$, say, has a nonzero component of the dominated solution then backward recursion implies a relative decrease of the undesired component as $i \to 0$. Historically one used to take $x_N^{(N)} = \binom{1}{0}$. We shall write out the results for the recursion in (2.1) (cf.§3.2.2) and $\{\phi_i\}$ the solution to be determined. So assume that a sequence $\{\xi_i^{(N)}\}_{i=0}^{N+1}$ is computed satisfying

$$\xi_{N+1}^{(N)} = 0; \qquad \xi_N^{(N)} = 1. \tag{5.4}$$

We find (cf. (2.3))

$$\xi_i^{(N)} = \frac{\psi_{N+1}\phi_i - \phi_{N+1}\psi_i}{\psi_{N+1}\phi_N - \phi_{N+1}\psi_N} = p_N\phi_i + q_N\psi_i, \text{ say.} \tag{5.5}$$

Since $|q_N/p_N| = |\phi_{N+1}/\psi_{N+1}|$ it follows that, at least for large (N-i), $\xi_i^{(N)}$ is almost proportional to ϕ_i. Hence $\xi_i^{(N)}/p_N$ would be a nice approximant for ϕ_i (the better the larger (N-i) is). The quantity p_N itself is hard to determine. But if a relation of the form

$$(5.6) \qquad \sum_{i=0}^{\infty} \mu_i \phi_i = 1,$$

is given (possibly $\mu_i = 0$ for $i > 0$, so ϕ_0 is given), then a satisfactory approximant for p_N is given by

$$(5.7) \qquad \sum_{i=0}^{N} \mu_i \xi_i^{(N)}.$$

For more detailed analysis of the error see GAUTSCHI (1967), MATTHEIJ & VAN DER SLUIS (1976), OLVER (1967a), ZAHAR (1977). From the geometrical interpretation of this algorithm as inverse iteration, it is immediately clear that one can often fasten the convergence by choosing a better approximation for the direction of $\binom{\phi_N}{\phi_{N+1}}$ than just $\binom{1}{0}$ (cf. MATTHEIJ & VAN DER SLUIS (1976), OLVER & SOOKNE (1972)); therefore one needs estimates for this dominated solution. For more general situations than this trivial constant case one can consult the cited literature. From §3.2.3 it follows that a good guess will also be $\binom{1}{\alpha_N}$ where α_N is the absolutely smallest eigenvalue of A_N.

The generalization of this algorithm for the matrix vector and/or higher dimensional cases is similar. Success is only assured if the desired solution is dominated by all solutions of a well determined (n-1) dimensional solution space (i.e., loosely speaking, where no dominant solutions are directionally close to the dominated one). The computed sequence has to be normalized and this may be done with a similar relation as (5.6) (cf. MATTHEIJ & VAN DER SLUIS, (1976)). The algorithm can also fruitfully be applied to inhomogeneous recursions if all solutions of the homogeneous part are dominant. Since the desired particular solution is unique then, no normalization of the computed sequence is necessary.

The choice of N depends on the accuracy required. We shall investigate the relative error, $\tau_i^{(N)}$ say, in our example. Suppose $\mu_0 = \frac{1}{\phi_0}$; $\mu_i = 0$, $i > 0$.

We obtain:

$$(5.8)\qquad \tau_i^{(N)} = \frac{\phi_i - (\xi_i^{(N)}\phi_0)/\xi_0^{(N)}}{\phi_i} = \frac{\psi_i/\phi_i - \psi_0/\phi_0}{\psi_{N+1}/\phi_{N+1} - \psi_0/\phi_0} \approx \left(\frac{\phi_{N+1}}{\phi_i}\right)\Big/\left(\frac{\psi_{N+1}}{\psi_i}\right).$$

The last estimate in (5.8) equals $(\frac{\alpha}{\beta})^{N-i}$.

Hence we see - in agreement with power method theory - that the relative error almost decreases with a factor $\frac{\alpha}{\beta}$ at each iteration step. For slowly varying recursions (§3.2.3) we have a similar error behaviour, viz.

$$\tau_i^{(N)} \approx \prod_{j=i}^{N} \frac{\lambda_j(n)}{\lambda_j(n-1)} .$$

Again, knowledge of the order of magnitude of the solutions of the recursion is very useful to estimate the error cf. (5.8).

As far as rounding errors concerns it has been shown in MATTHEIJ & VAN DER SLUIS (1976) that *the relative error in* x_i *is almost proportional to* i and not to (N-i) or N, cf. §3.2.4. In the inhomogeneous case the error even is independent of the number of steps. Anyway the actual choice of N has no influence on the accuracy of x_i with respect to rounding errors.

3.2.5.2. Olver's algorithm

If the recursion is third order or second order inhomogeneous or even higher order there may be a situation where both forward recursion and backward recursion (Miller's algorithm) will be unstable. Viz. if the desired (possibly particular) solution is dominated by some solution of the homogeneous part and dominates some other complementary solution in turn. A well-known scalar example is given by the recursion for the Struve function $H_i(x)$ (cf. ABRAMOWITZ & STEGUN (1964, p.496))

$$(5.9)\qquad H_{i+1}(x) = \frac{2i}{x} H_i(x) - H_{i-1}(x) + \frac{(\frac{1}{2}x)^i}{\sqrt{\pi}\,\Gamma(i+\frac{3}{2})} .$$

The homogeneous part of (5.9) is also satisfied by the Bessel functions of the first and second kind, viz. $\{J_i(x)\}$ and $\{Y_i(x)\}$ respectively.

An efficient algorithm for stable computation of "intermediate" solutions like $\{H_i(x)\}$ was developed by OLVER (1967b);we shall deduce it in such a way that generalizations may be easily understood, (cf. MATTHEIJ(1977)). Consider the general second order scalar recursion

$$\xi_{i+1} = a_i\xi_i + b_i\xi_{i-1} + c_i. \tag{5.10}$$

Define a solution $\{\rho_i\}$ of the homogeneous part of (5.10) by

$$\rho_0 = 0; \quad \rho_1 = 1. \tag{5.11}$$

By substituting

$$\xi_i\rho_{i+1} - \xi_{i+1}\rho_i = \eta_i, \tag{5.12}$$

we find a first order recursion for $\{\eta_i\}$:

$$\eta_i = -b_i\eta_{i-1} - c_i\rho_i. \tag{5.13}$$

This recursion can be derived using Abel's transformation trick (NÖRLUND(1924 p.289)); in his paper Olver employs a somewhat unconventional elimination method for a system of equations that was found by considering recurrences for $\xi_0,\ldots,\xi_N$ and imposing boundary values. The recursion (5.13) is used in *forward direction* whereas after choosing an end value $\xi_{N+1}^{(N)}$ as approximation to ξ_N, a sequence of approximating values $\{\xi_i^{(N)}\}$ (to $\{\xi_i\}$) is computed by

$$\xi_i^{(N)} = (\eta_i + \xi_{i+1}^{(N)}\,\rho_i)/\rho_{i+1}, \tag{5.14}$$

(cf. (5.12)), i.e., in *backward direction*.

In order to understand why this is a fruitful approach it may be helpful to remark that the substitution (5.12) and the result (5.13) in fact are equivalent to *reducing the order* of a recursion when some solution (of the homogeneous part) is known. If this reduction solution was a dominant one, then there is hope that after the order reduction the transformed subdominant solution will become dominant, in particular the desired solution $\{\eta_i\}$; hence forward recursion for $\{\eta_i\}$ is expected to be stable. We shall work this out later. Assume that $\{\rho_i\}$ is a dominant solution (this is true, except for the singular case that the direction of the first iterand of the dominated solution (of the companion matrix vector recursion) has the direction of $\binom{0}{1}$) and let $\{\sigma_i\}$ be a solution of the homogeneous part dominated by $\{\xi_i\}$. Let the solution of the homogeneous part of (5.13) be defined by

(5.15) $$\zeta_i = \sigma_i\rho_{i+1} - \sigma_{i+1}\rho_i \ .$$

Then

(5.16) $$\frac{\zeta_i}{\eta_i} = \frac{\sigma_i}{\xi_i}\left[\frac{\rho_{i+1}/\rho_i-\sigma_{i+1}/\sigma_i}{\rho_{i+1}/\rho_i-\xi_{i+1}/\xi_i}\right] .$$

The factor between brackets in (5.16) will be bounded if for all i the growth factors ρ_{i+1}/ρ_i are larger than ξ_{i+1}/ξ_i (which is reasonable since $\{\rho_i\}$ dominates $\{\xi_i\}$). Hence $\{\eta_i\}$ dominates $\{\zeta_i\}$.

For the approximant $\{\xi_i^{(N)}\}$ we find

(5.17) $$\frac{\xi_i^{(N)}}{\rho_i} - \frac{\xi_{i+1}^{(N)}}{\rho_{i+1}} = \frac{\eta_i}{\rho_i\rho_{i+1}} \ .$$

Hence

(5.18) $$\xi_i^{(N)} = \rho_i \sum_i^N \frac{\eta_j}{\rho_j\rho_{j+1}} + \frac{\rho_i\xi_{N+1}^{(N)}}{\rho_{N+1}} ,$$

which also holds without the superscript (N). On account of the dominance we therefore have by a limit argument

(5.19) $$\xi_i = \rho_i \sum_i^\infty \frac{\eta_j}{\rho_j\rho_{j+1}} \ .$$

If the solutions are of exponential type (e.g. growing like eigenvalues of a suitable associated matrix cf. §3.2.3), then (5.19) will be of geometrical type and thus has fast convergence. In fact we then have

(5.20) $$\frac{\eta_j}{\rho_j\rho_{j+1}} \approx \frac{\xi_j}{\rho_j}\kappa_j ,$$

where κ_j is almost independent of j.

Comparing (5.18) and (5.19) we find for the relative error

$$(5.21) \qquad \left|\frac{\xi_i-\xi_i^{(N)}}{\xi_i}\right| = \frac{\left|\Sigma_{N+1}^{\infty} \dfrac{\eta_j}{\rho_j\rho_{j+1}} - \dfrac{\xi_{N+1}^{(N)}}{\rho_{N+1}}\right|}{\left|\Sigma_i^{\infty} \dfrac{\eta_j}{\rho_j\rho_{j+1}}\right|} \approx \left|\frac{\xi_{N+1}\rho_i}{\xi_i\rho_{N+1}} - \frac{\xi_{N+1}^{(N)}}{\xi_i}\,\frac{\rho_i}{\rho_{N+1}}\right| .$$

On the other hand if $\xi_{N+1}^{(N)} = 0$ we can simply estimate (5.21) using computed values of $\{\eta_i\}$ and $\{\rho_i\}$ by

$$(5.22) \qquad \left|\frac{\xi_i-\xi_i^{(N)}}{\xi_i}\right| \approx \left|\frac{\eta_{N+1}}{\rho_{N+1}\rho_{N+2}}\right| \Big/ \left|\frac{\eta_i}{\rho_i\rho_{i+1}}\right| .$$

If one wishes to approximate $\xi_0,\dots,\xi_p$ say, then (5.22) provides for an easily accessible criterion to estimate the value of N, viz. by recurring forwards with (5.13) until for a certain $N > p$ and all $i \le p$, (5.22) is smalller than the required tolerance. Note that the algorithm can also be used to approximate the dominated solution of a homogeneous three term recurrence relation.

3.2.5.3. More general algorithms for approximating "intermediate" solutions

Above we have remarked that Olver's method was basically equivalent to classical order reduction. Hence a generalization to higher order recursions is straightforward. However, repeated use of such order reduction might deteriorate the conditioning of the problem, whereas possible convergence is hard to prove. On account of Olver's derivation, viz. via a kind of LU decomposition of an associated large (and sparse!) system, some authors (cf. OLIVER (1968b)) have proposed generalizations based on linear algebraic methods. A less attractive feature of such an approach is that a fairly simple problem is translated into a usually more complicated algebraic problem, with questions like pivotting, equilibration and loosing sparseness.

A more general method, in some way also a generalization of Olver's, was proposed by MATTHEIJ (1977, 1982).It deals with matrix vector recursions: Suppose the solution $\{x_i\}$ of (1.2) is *dominated* by solutions of the homogeneous recursion, that constitute a well defined k-dimensional subspace S_1, say whereas $\{\phi_i\}$ is *not dominated* by the solutions in the complementary space S_2 say. Let T_0 be an n-th order nonsingular matrix, such that its first k columns span a subspace of $\mathbb{R}^n$ that has an empty intersection with the subspace spanned by initial values of elements of S_2. (this is a harmless assumption and is comparable to the condition necessary for successful use of the QR algorithm). Given the recurrence (1.2) i.e.

$$x_{i+1} = A_i x_i + r_i$$

we can obtain a transformed decoupled recursion

(5.23) $$y_{i+1} = V_i y_i + s_i$$

with $\{T_i\}$ a sequence of nonsingular matrices chosen such that

(5.24) $$V_i = T_{i+1}^{-1} A_i T_i, \quad \text{is block triangular}$$

and

(5.25) $$s_i = T_{i+1}^{-1} r_i,$$

(5.26) $$y_i = T_i^{-1} x_i.$$

Partitioning the vectors into the first k and the last (n-k) coordinates and the matrices V_i correspondingly, we find

(5.27a) $$y_{i+1}^1 = V_i^{11} y_i^1 + V_i^{12} y_i^2 + s_i^1 \quad \updownarrow k$$

(5.27b) $$y_{i+1}^2 = V_i^2 y_i^2 + s_i^2 \quad \updownarrow (n-k).$$

It can now be shown (cf. MATTHEIJ (1980)) that the solutions of the homogeneous part of (5.27a), viz. the recursion

(5.28) $$z_{i+1} = V_i^{11} z_i,$$

have a growth character corresponding to the solutions $\in S_1$, whereas the solutions of the homogeneous part of (5.27b), viz. the recursion

(5.29) $$u_{i+1} = V_i^{22} u_i,$$

grow like solutions ϵS_2.

Thus it turns out that a stable computation of $\{y_i^2\}$ has to be done in forward direction, whereas $\{y_i^1\}$ has to be computed in backward direction; the latter can be performed after $\{y_i^2\}$ has been calculated. Of course we have to find some approximation to y_N^1, say for N large enough, before we can start the backward algorithm (cf. Miller's and Olver's algorithm). We shall give an idea of the thus introduced truncation error for the case of a slowly varying homogeneous recursion (cf. §3.2.3); so let $\|x_{i+1}\|/\|x_i\| \approx \lambda_i(k+1)$. Denote the sequence of approximants of $\{y_i^1\}$ by $\{y_i^1(N)\}$. Then the relative truncation error in y_i is given by

$$(5.30) \qquad \frac{\|y_i^1(N)-y_i^1\|}{\|y_i\|} = \frac{\|(\prod_i^{N-1} B_j)^{-1}(y_N^1(N)-y_N^1)\|}{\|y_i\|} \approx \left|\prod_{j=i}^{N-1} \frac{\lambda_j(k+1)}{\lambda_j(k)}\right| \frac{\|y_N^1(N)-y_N^1\|}{\|y_N\|} \; .$$

If we know a good approximation of y_N^1 then this should be used of course. If we do not have such an approximation at our disposal we may choose $y_N^1(N) = 0$. We remark that the error found in (5.30) again resembles the power method like results in §§3.2.5.1-2. As in Olver's algorithm we may use computed quantities to estimate the error and thus equip the algorithm with a self search device for determining an N necessary to obtain a certain relative precision: indeed, a good estimator for $\|(\prod_{j=i}^{N-1} B_j)^{-1}\|$ is given by the inverse of the product of the absolutely smallest eigenvalues of the B_j, whereas y_i and y_i^1 can be estimated by y_i^2. For refinements see MATTHEIJ (1982,§5).

It can be shown that the relative rounding error in the computed result is proportional to i - as was also found in Miller's algorithm - and even independent of i for inhomogeneous recursions with solutions of the homogeneous part, that are sufficiently dominant and dominated resp.

A straightforward way to determine these $\{T_i\}$ and $\{V_i\}$ is by using orthogonal matrices. The factorization step (5.23) is then performed via QR-decomposition, which can be performed by Householder's or Given's method (cf. WILKINSON (1965)). As a by-product the matrices V_i^{11} will be upper triangular, which means that their eigenvalues are known (necessary to estimate their norms) and moreover that inversion of the B_i - which is necessary for the backward recursion - is simple and stable alike. Finally backtransformation of the computed sequence $\{y_i^1(N)/y_i^2\}$ is simple, because inversion of an orthogonal matrix is equivalent to transposing.

The algorithm of Olver in §3.2.5.2 can be considered as a special kind of triangularizing the corresponding companion matrix vector recursion, but not via orthogonal matrices. For the general scalar problem the application of the triangularization method to the companion matrix recursion using orthogonal matrices, is likely to disturb the sparseness. Hence it is worthwhile to investigate whether there are special choices for the T_i which preserve the scalar character of the recursion. A more detailed description is still under construction.

3.2.6 Conclusion

In the previous sections we have tried to give a survey of problems and methods involved with recursion. We did not go too much into details; the interested reader can consult the papers indicated in the references. There are many more related subjects, some of them are treated elsewhere in this tract. Examples are the summation of series of dominated solution (see the excellent method given in the papers by DEUFLHARD (1976,1977)) or the problems encountered when there is no dominance phenomenon, and the rounding errors may become important.

3.3. Three-term recursions; some practical points of view

In this section we again consider three-term recursions. In the previous section general aspects of these recursions were considered. For applications, especially for special functions, it is worth-while to have information how to use the algorithms for practical problems.

3.3.1. On the growth of solutions of three-term difference equations

Let us consider the recursion

$$(1.1)\qquad y_{n+1} + a_n y_n + b_n y_{n-1} = 0, \qquad n = 1,2,\ldots,$$

where a_n, b_n are given sequences of real or complex numbers, $b_n \neq 0$. The general solution of (1.1) can be written as a linear combination of any pair f_n, g_n linearly independent solutions, that is

$$(1.2)\qquad y_n = Af_n + Bg_n$$

where A and B are complex numbers not depending on n. We are interested in the special case that the pair f_n, g_n has the property

$$(1.3)\qquad \lim_{n\to\infty} f_n/g_n = 0.$$

Any solution (1.2) with $B \neq 0$ then satisfies $f_n/y_n \to 0$, $n \to \infty$. If $B = 0$ in (1.2) y_n is called a minimal solution of (1.1), if $B \neq 0$ it is called a dominant solution. If we have two initial values y_0, y_1 of (1.1) and f_0, f_1, g_0, g_1 are known, then we can compute A and B, viz.

$$A = \frac{g_1 y_0 - g_0 y_1}{f_0 g_1 - f_1 g_0}, \qquad B = \frac{y_0 f_1 - y_1 f_0}{g_0 f_1 - g_1 f_0}.$$

The denominators are non-zero if f_n, g_n are linearly independent. When we prescribe that the initial values y_0, y_1 are intended for a minimal solution, then $B = 0$. It follows that in that case just one initial value can be prescribed, the remaining one follows from the relation $y_0 f_1 = y_1 f_0$. In computations this leads to well known instabilities for the evaluation of minimal solutions. If our initial conditions y_0, y_1 do not fulfil exactly the condition $B = 0$, then the computed solution (1.2) behaves ultimately as a dominant solution (even when computing with infinite precision), although we intended to compute a minimal one.

For applications it is important to know whether a given recursion (1.1) has dominant and minimal solutions. Sometimes this can be concluded from the asymptotic behaviour of the numbers a_n, b_n in (1.1). The following theorem is quoted from GAUTSCHI (1967). For a proof the reader may consult the references given there.

THEOREM. *Let* a_n, b_n *have the asymptotic behaviour*

$$a_n \sim an^{\alpha}, \quad b_n \sim bn^{\beta}, \quad ab \neq 0, \quad \alpha,\beta \textit{ real}, \quad n \to \infty$$

and let t_1, t_2 *be the zeros of the characteristic polynomial* $\Phi(t) = t^2 + at + b$, $|t_1| \geq |t_2|$.

(i) *If* $\alpha > \frac{1}{2}\beta$ *then the difference equation* (1.1) *has two linearly independent solutions* $y_{n,1}$ *and* $y_{n,2}$, *for which*

$$\frac{y_{n+1,1}}{y_{n,1}} \sim -an^{\alpha}, \quad \frac{y_{n+1,2}}{y_{n,2}} \sim -\frac{b}{a}n^{\beta-\alpha}, \quad n \to \infty.$$

(ii) *If* $\alpha = \frac{1}{2}\beta$ *then* (1.1) *has two linearly independent solutions* $y_{n,1}, y_{n,2}$ *for which*

$$\frac{y_{n+1,1}}{y_{n,1}} \sim t_1 n^{\alpha}, \quad \frac{y_{n+1,2}}{y_{n,2}} \sim t_2 n^{\alpha}, \qquad n \to \infty,$$

provided $|t_1| > |t_2|$. *If* $|t_1| = |t_2|$ *then*

$$\limsup_{n\to\infty} [\,|y_n|(n!)^{-\alpha}]^{1/n} = |t_1|$$

for all nontrivial solutions of (1.1).

(iii) *If* $\alpha < \frac{1}{2}\beta$ *then*

$$\limsup_{n\to\infty} [\,|y_n|(n!)^{-\beta/2}]^{1/n} = |b|^{1/2}$$

for all nontrivial solutions of (1.1).

In both case (i) and the first part of case (ii) $f_n = y_{n,2}$ is a minimal solution of (1.1). Furthermore, in the first part of case (ii)

$$\lim_{n\to\infty} \frac{y_{n+1}}{n^{\alpha} y_n} = t_r, \quad r = 1, \text{ or } r = 2,$$

where r = 2 for the minimal solution, and r = 1 for any other solution. To see this we remark that from (i) we derive

$$\frac{y_{n+1,2}}{y_{n+1,1}} \Big/ \frac{y_{n,2}}{y_{n,1}} \sim \frac{b}{a^2} n^{\beta-2\alpha}, \quad n \to \infty,$$

which tends to zero since $\beta-2\alpha < 0$. Hence the sequence $\{y_{n,2}/y_{n,1}\}$ tends to zero. In the first part of (ii) we have

$$\frac{y_{n+1,2}}{y_{n+1,1}} \Big/ \frac{y_{n,2}}{y_{n,1}} \sim t_2/t_1, \quad n \to \infty.$$

Since $|t_1| > |t_2|$, we again conclude that $\{y_{n,2}/y_{n,1}\}$ tends to zero.

The second part of case (ii) of the theorem and case (iii) give no information about dominant and/or minimal solutions. As will become clear from the examples below, we need extra information of the solutions of (1.1) in these cases.

Some insight in the above theorem can be obtained from the companion matrix vector recursion:

$$\begin{pmatrix} y_{k+1} \\ y_k \end{pmatrix} = A_k \begin{pmatrix} y_k \\ y_{k-1} \end{pmatrix}, \quad A_k = \begin{pmatrix} -a_k & -b_k \\ 1 & 0 \end{pmatrix}.$$

The eigensystem of A_k is given by

$$\text{eigenvectors: } E_k = \left(e_k^+ \,\middle|\, e_k^- \right) = \begin{pmatrix} \lambda_k^+ & \lambda_k^- \\ 1 & 1 \end{pmatrix}$$

$$\text{eigenvalues: } \Lambda_k = \begin{pmatrix} \lambda_k^+ & \\ & \lambda_k^- \end{pmatrix}.$$

The quotients of the elements of each eigenvector behave as λ_k^+ and λ_k^-, respectively. The eigenvalues with $a_k = ak^{\alpha}$ and $b_k = bk^{\beta}$ are given by

$$\lambda_k^{\pm} = (-ak^{\alpha} \pm ak^{\alpha} \sqrt{1-4bk^{\beta-2\alpha}/a^2})/2$$

and behave, for k large, as depicted in the following table.

roots situation	$\lvert\lambda_k^+\rvert$	$\lvert\lambda_k^-\rvert$
$\beta < 2\alpha$	$\lvert b/a\rvert k^{\beta-\alpha}$	$\lvert a\rvert k^{\alpha}$
$\beta = 2\alpha$	$\lvert\lambda_1^+\rvert k^{\alpha}$	$\lvert\lambda_1^-\rvert k^{\alpha}$
$\beta > 2\alpha$	$\sqrt{\lvert b\rvert}k^{\beta/2}$	$\sqrt{\lvert b\rvert}k^{\beta/2}$

If we assume

$$E_{k+1}^{-1}E_k \sim I$$

which is the case with the above specified coefficients then two independent solutions of the matrix vector recursion are given by the eigenvectors e_k^+ and e_k^-. The quotient of successive elements of the independent solutions behave as given in the above table for the eigenvalues as a function of the relation between β and 2α.

Examples

1. Bessel functions.

Recursion: $$y_{n+1} - \frac{2n}{z}y_n + y_{n-1} = 0.$$

Solutions: $$f_n = J_n(z), \quad g_n = Y_n(z), \quad z \neq 0.$$

Case of theorem: (i), $a = -\frac{2}{z}$, $\alpha = 1$,
$b = 1$, $\beta = 0$.

Conclusion of theorem: $$\frac{g_{n+1}}{g_n} \sim \frac{2n}{z}, \quad \frac{f_{n+1}}{f_n} \sim \frac{z}{2n}.$$

Known asymptotic behaviour: $$f_n \sim (2\pi n)^{-\frac{1}{2}}\left(\frac{ez}{2n}\right)^n, \qquad n \to \infty.$$
$$g_n \sim -(\pi n/2)^{-\frac{1}{2}}\left(\frac{ez}{2n}\right)^{-n}$$

2. Legendre functions.

a) Recursion with respect to the order

Recursion: $$y_{m+1} + 2mz(z^2-1)^{-\frac{1}{2}}y_m + (m+\alpha)(m-\alpha-1)y_{m-1} = 0.$$

Solutions: $$f_m = P_\alpha^m(z), \quad g_m = Q_\alpha^m(z), \quad \operatorname{Re} z > 0,$$
$$\alpha \in \mathbb{C}, \quad \alpha \neq -1,-2,\ldots \quad z \notin (0,1].$$

Case of theorem: (ii), $a = 2z(z^2-1)^{-\frac{1}{2}}$, $\alpha = 1$
$b = 1$, $\beta = 2$

$$t_1 = -[(z+1)/(z-1)]^{\frac{1}{2}}, \quad t_2 = t_1^{-1}$$

$$|t_1| > 1 > |t_2|.$$

Conclusion of theorem: $\lim\limits_{m\to\infty} \dfrac{f_{m+1}}{mf_m} = t_2$, $\lim\limits_{m\to\infty} \dfrac{g_{m+1}}{mg_m} = t_1$.

b) Recursion with respect to the degree

Recursion: $$y_{n+1} - z\,\frac{2n+2a+1}{n+a-m+1}\,y_n + \frac{n+a+m}{n+a-m+1}\,y_{n-1} = 0.$$

Solutions: $$f_n = Q_{a+n}^m(z), \quad g_n = P_{a+n}^m(z), \quad \operatorname{Re} z > 0.$$

Case of theorem: (ii), $a = -2z$, $\alpha = 0$
$b = 1$, $\beta = 0$

$$t_1 = z + (z^2-1)^{\frac{1}{2}}, \quad t_2 = t_1^{-1}$$

$$|t_1| > 1 > |t_2|.$$

Conclusion of theorem: $\lim\limits_{n\to\infty} f_{n+1}/f_n = t_2$, $\lim\limits_{n\to\infty} g_{n+1}/g_n = t_1$.

3. Coulomb wave functions

Recursion: $$L[(L+1)^2+\eta^2]^{\frac{1}{2}}y_{L+1} - (2L+1)[\eta + L(L+1)/\rho]y_L$$
$$+ (L+1)[L^2+\eta^2]^{\frac{1}{2}}y_{L-1} = 0, \qquad L = 1,2,\ldots$$

Solutions: $$f_L = F_L(\eta,\rho), \quad g_L = G_L(\eta,\rho), \quad \eta \in \mathbb{R}, \quad \rho > 0.$$

Case of theorem: (i), $a = -\frac{2}{\rho}$, $\alpha = 1$
$b = 1$, $\beta = 1$.

Conclusion of theorem: $g_{L+1}/g_L \sim \frac{2L}{\rho}$, $f_{L+1}/f_L \sim \frac{\rho}{2L}$, $L \to \infty$.

Known asymptotic behaviour: $f_L \sim C_L(\eta)\rho^{L+1}$

$$\left.\begin{aligned} g_L &\sim [2LC_L(\eta)\rho^L]^{-1} \\ C_L(\eta) &\sim 2^{-\frac{1}{2}}e^{-\pi\eta/2}\Big[\frac{e}{2(L+1)}\Big]^{L+1}. \end{aligned}\right. \qquad L \to \infty$$

4. Incomplete beta functions

Recursion: $$y_{n+1} - \Big(1 + \frac{n+p+q-1}{n+p}\,x\Big)y_n + \frac{n+p+q-1}{n+p}\,x\,y_{n-1} = 0.$$

Solutions: $f_n = I_x(p+n,q)$, $\quad g_n = 1$, $\quad 0 \le x < 1$.

Case of theorem: (ii) $a = -(1+x)$, $\quad \alpha = 0$

$b = x$, $\quad \beta = 0$

$t_1 = 1$, $\quad t_2 = x$.

Conclusion of theorem: $\lim_{n\to\infty} \dfrac{f_{n+1}}{f_n} = x.$

Known asymptotic behaviour: $f_n \sim (1-x)^{q-1}n^{q-1}x^{p+n}/\Gamma(q)$.

5. Repeated integrals of the error function

Recursion: $$y_{n+1} + \frac{z}{n+1}\,y_n - \frac{1}{2(n+1)}\,y_{n-1} = 0.$$

Solutions: $f_n = e^{z^2}i^n\mathrm{erfc}\,z$, $\quad g_n = (-1)^n e^{z^2} i^n\mathrm{erfc}(-z)$,

$$i^n\mathrm{erfc}\,z = \int_z^\infty i^{n-1}\mathrm{erfc}\,t\,dt, \quad i^0\mathrm{erfc}\,z = \mathrm{erfc}\,z$$

$$i^{-1}\mathrm{erfc}\,z = 2\pi^{-\frac{1}{2}}e^{-z^2}, \quad z \in \mathbb{C}.$$

Case of theorem: (iii) $a = z$, $\quad \alpha = -1$

$b = -\frac{1}{2}$ $\quad \beta = -1.$

Conclusion of theorem: $\limsup_{n\to\infty}[|y_n|(n!)^{\frac{1}{2}}]^{1/n} = 2^{-\frac{1}{2}}$

for both $y_n = f_n$ and $y_n = g_n$

Known asymptotic behaviour: $i^n\mathrm{erfc}\,z \sim 2^{-n}e^{-\frac{1}{2}z^2 - z\sqrt{2n}}/\Gamma(\frac{n}{2}+1)$ hence

$$(-1)^n\frac{f_n}{g_n} \sim e^{-2z\sqrt{2n}}, \quad n \to \infty.$$

6. Confluent hypergeometric functions $U(a,b,z)$, $M(a,b,z)$

a) Recursion with respect to a

Recursion: $(n+a+1-b)y_{n+1} + (b-z-2a-2n)y_n + (a+n-1)y_{n-1} = 0.$

Solutions: $f_n = \frac{\Gamma(a+n)}{\Gamma(a)} U(a+n,b,z),$

$g_n = \frac{\Gamma(a+n)}{\Gamma(1+a+n-b)} M(a+n,b,z).$

Case of theorem: (ii) $a = -2, \quad \alpha = 0$

$b = 1, \quad \beta = 0$

$t_1 = t_2 = 1.$

Conclusion of theorem: $\limsup_{n \to \infty} |y_n|^{1/n} = 1$

for both $y_n = f_n$ and $y_n = g_n$.

Known asymptotic behaviour: $f_n \sim c_1 n^{\frac{1}{2}b-\frac{3}{4}} e^{-2\sqrt{nz}},$

$g_n \sim c_2 n^{\frac{1}{2}b-\frac{3}{4}} e^{+2\sqrt{nz}},$

hence $\frac{f_n}{g_n} \sim c_3 e^{-4\sqrt{nz}},$

c_i not depending on n.

b) Recursion with respect to b

Recursion: $zy_{n+1} + (1-b-n-z)y_n + (b+n-a-1)y_{n-1} = 0.$

Solutions: $f_n = \frac{\Gamma(b+n-a)}{\Gamma(b+n)} M(a,b+n,z),$

$g_n = U(a,b+n,z).$

Case of theorem: (i), $a = -1/z, \quad \alpha = 1$

$b = 1/z, \quad \beta = 1.$

Conclusion of theorem: $\frac{g_{n+1}}{g_n} \sim n/z, \quad \frac{f_{n+1}}{f_n} \sim 1.$

Known asymptotic behaviour: $f_n \sim n^{-a}$

$g_n \sim z^{1-b-n}\Gamma(b+n-1)/\Gamma(a).$

7. Jacobi polynomials

Recursion:

$$(2n+2)(n+\alpha+\beta+1)(2n+\alpha+\beta)y_{n+1} =$$
$$(2n+\alpha+\beta+1)\{(2n+\alpha+\beta+2)(2n+\alpha+\beta)x + \alpha^2 - \beta^2\}y_n$$
$$-2(n+\alpha)(n+\beta)(2n+\alpha+\beta+2)y_{n-1}$$

Solutions: $g_n = P_n^{(\alpha,\beta)}(x)$, $f_n = Q_n^{(\alpha,\beta)}(x)$, $x \in \mathbb{C}$.

Case of theorem: (ii) $a = -2x$, $\alpha = 0$

$b = 1$, $\beta = 0$

$$t_1 = x + \sqrt{x^2-1}, \quad t_2 = x - \sqrt{x^2-1}$$
$$|t_1| = |t_2| = 1 \text{ if } x \in [-1,1]$$
$$|t_1| > 1, \ |t_2| < 1 \text{ if } x \notin [-1,1]$$

Conclusion of theorem: $x \in [-1,1]$: $\limsup_{n \to \infty} |y_n|^{1/n} = 1$

for both $y_n = P_n^{(\alpha,\beta)}(x)$ and

$y_n = Q_n^{(\alpha,\beta)}(x)$,

$$x \notin [-1,1]: \ \frac{P_{n+1}^{(\alpha,\beta)}(x)}{P_n^{(\alpha,\beta)}(x)} \sim t_1, \quad \frac{Q_{n+1}^{(\alpha,\beta)}(x)}{Q_n^{(\alpha,\beta)}(x)} \sim t_2$$

Known asymptotic behaviour: $x \in (-1,1)$, $x = \cos\theta$, $0 < \theta < \pi$

$$P_n^{(\alpha,\beta)}(x) \sim [2/(\pi n \sin\theta)]^{\frac{1}{2}} \cos[(n+\tfrac{1}{2})\theta - \pi/4]$$

$$x \notin [-1,1], \quad P_n^{(\alpha,\beta)}(x) \sim n^{-\frac{1}{2}}\phi(x)t_1^n,$$

$$Q_n^{(\alpha,\beta)} \sim n^{-\frac{1}{2}}\psi(x)t_2^n$$

where ϕ and ψ are independent of n.

Examples 1 through 5 are extensively treated in GAUTSCHI (1967). For ALGOL 60 implementations of the algorithms see Gautschi's references. Example 6 is considered (with ALGOL 60 algorithm) in TEMME (1983).

Information on the Jacobi polynomials $P_n^{(\alpha,\beta)}(x)$ and of $Q_n^{(\alpha,\beta)}(x)$, Jacobi's function of the second kind, can be found in SZEGÖ (1974). The Jacobi polynomials contain as special cases the important Chebyshev and Legendre polynomials. It follows that for x not lying in the interval of orthogonality the polynomials $P_n^{(\alpha,\beta)}(x)$ can be safely recurred from initial values $P_0^{(\alpha,\beta)}(x) = 1$ and $P_1^{(\alpha,\beta)}(x) = \frac{1}{2}(\alpha,\beta) + \frac{1}{2}(\alpha+\beta+2)x$. For $x \in [-1,1]$ (for numerical applications the most important case) the theorem is inconclusive and also our formulas give no information. The point is that $Q_n^{(\alpha,\beta)}(x)$ is usually not considered for $x \in [-1,1]$. It is a many-valued function of x and it can be made single-valued and regular in the complex plane by cutting the plane along the segment $[-1,1]$. If $x \in [-1,1]$, the values $Q_n^{(\alpha,\beta)}(x \pm i0)$ are not equal.

For the case $\alpha = \beta = 0$, SZEGŐ (1974, p. 224) gives the result

$$Q_n^{(0,0)}(\cos\theta \pm i0) \sim \left(\frac{\pi}{2n\sin\theta}\right)^{\frac{1}{2}} e^{\pm i[(n+\frac{1}{2})\theta+\pi/4]}, \qquad n \to \infty,$$

where $0 < \theta < \pi$. It follows from his analysis that the result for the general case can be obtained using the same method. It gives the same behaviour as for $Q_n^{(0,0)}(\cos\theta \pm i0)$ except for a constant factor depending on α, β and θ, but not on n. Thence we conclude that the asymptotic behaviour of $P_n^{(\alpha,\beta)}(x)$ and $Q_n^{(\alpha,\beta)}(x\pm i0)$ is the same, apart from a shift in the phase of the oscillatory part of the functions. It follows that for $x \in [-1,1]$ the Jacobi polynomial $P_n^{(\alpha,\beta)}(x)$ is not a minimal solution of the recursion given in Example 7. Rounding errors become important in this case when using the recursion relation for computing successive Jacobi polynomials.

Other examples for the use of backward recurrence relations can be found in CLENSHAW (1962) and CLENSHAW & PICKEN (1966), where the method is used to generate coefficients for the expansion of many special functions in series of Chebyshev polynomials of the first kind (in these cases higher order recursions are involved).

3.3.2. The Miller algorithm

Let us suppose we want to compute the minimal solution $\{f_n\}$ of the recursion (1.1) with the normalizing relation

$$(2.1) \qquad \sum_{n=0}^{\infty} \lambda_n f_n = s, \quad s \neq 0$$

where s and λ_n are given numbers. Of course a finite number, viz. $f_0,\ldots,f_N$, will be considered where $N \geq 0$. As mentioned in earlier sections, Miller's algorithm is based on choosing $\nu > N$ and computing a solution $\{y_n^{(\nu)}\}$ of (1.1) with initial values

$$y_{\nu+1}^{(\nu)} = 0, \qquad y_{\nu}^{(\nu)} = 1$$

(where the numbers 0 and 1 may be replaced by any other pair if at least one of the numbers is not equal to zero; in some cases a different choice may speed up the convergence). It follows (cf. (5.5) on p.55) that for $0 \le n \le \nu-1$

$$(2.2) \qquad y_n^{(\nu)} = \frac{g_{\nu+1}f_n - f_{\nu+1}g_n}{g_{\nu+1}f_\nu - f_{\nu+1}g_\nu} = p_\nu f_n + q_\nu g_n, \text{ say.}$$

Hence $y_n^{(\nu)}/p_\nu = f_n - f_{\nu+1}/g_{\nu+1}\; g_n$ and from (1.3) we derive that for $0 \le n \le N$

$$\lim_{\nu\to\infty} y_n^{(\nu)}/p_\nu = f_n .$$

It follows that, if ν is large enough, f_n can be computed from $y_n^{(\nu)}$ and p_ν. The latter is not known, in general, and we proceed using (2.1). We compute

$$(2.3) \qquad s^{(\nu)} = \sum_{n=0}^{\nu} \lambda_n y_n^{(\nu)}, \qquad f_n^{(\nu)} = \frac{s}{s^{(\nu)}}\, y_n^{(\nu)},$$

then we have for the relative error in f_n (if $f_n \neq 0$)

$$(2.4) \qquad \frac{f_n^{(\nu)} - f_n}{f_n} = \frac{s/s^{(\nu)}\; y_n^{(\nu)} - f_n}{f_n} = \frac{s(p_\nu + q_\nu g_n/f_n) - s^{(\nu)}}{s^{(\nu)}}$$

$$= \frac{\sigma_\nu - \rho_{\nu+1}/\rho_n + \tau_\nu}{1 - \sigma_\nu - \tau_\nu}$$

with

$$(2.5) \qquad \sigma_\nu = \frac{1}{s}\sum_{m=\nu+1}^{\infty} \lambda_m f_m, \quad \rho_n = f_n/g_n, \quad \tau_\nu = \frac{\rho_{\nu+1}}{s}\sum_{m=0}^{\nu} \lambda_m g_m .$$

On account of (1.3) and the convergence of (2.1) it follows that the left-hand side of (2.4) tends to zero (for $\nu \to \infty$) if and only if τ_ν tends to zero. Also, (2.4) gives information on the relative error when the quantities σ_ν, ρ_n, ρ_ν and τ_ν can be estimated.

To facilitate the error analysis, the quantities σ_ν and τ_ν, representing sums, are replaced by the possibly most relevant terms in these sums, viz.

$$\sigma_\nu \doteq \frac{1}{s}\lambda_{\nu+1}f_{\nu+1}, \qquad \tau_\nu \doteq \frac{\rho_{\nu+1}}{s}\lambda_\nu g_\nu .$$

Then the relative error (2.4) is written as

$$(2.6) \qquad \frac{f_n^{(\nu)}-f_n}{f_n} \doteq \frac{1}{s}\lambda_{\nu+1}f_{\nu+1} + \frac{f_{\nu+1}}{g_{\nu+1}}\frac{\lambda_\nu g_\nu}{s} - \frac{f_{\nu+1}}{g_{\nu+1}}\frac{g_n}{f_n}$$

$$\doteq \frac{1}{s}\lambda_{\nu+1}f_{\nu+1} - \frac{f_{\nu+1}}{g_{\nu+1}}\frac{g_n}{f_n}, \qquad (n = 0,\dots,N).$$

For obtaining an a priori estimate of ν, which makes the right-hand side of (2.6) smaller than a given quantity $\varepsilon > 0$, GAUTSCHI (1967) considered only the case $n = N$ (taking into account (1.3) this is a reasonable step). In the examples in his paper he replaced the values $f_{\nu+1}$, $g_{\nu+1}$, f_n and g_n by asymptotic approximations. Then, using an inversion process, he obtained a first estimate of ν. By computing successive values $f_n^{(\nu)}, f_n^{(\nu+5)}, \dots$ $(n = 0,1,\dots,N)$ the values of $f_n^{(\nu+5j)}$ are accepted if they agree with $f_n^{(\nu+5(j-1))}$ within the prescribed relative accuracy. An unpleasant feature of this procedure is that computing time is wasted if either the first estimate of ν is much too low or much too high. Another difficulty is a slight uncertainty associated with the acceptance criterion: mere numerical agreement of solutions computed with two different values of ν(ν and $\nu+5$) does not guarantee their accuracy. In §3.3.4 we describe a different procedure for obtaining estimates of ν.

3.3.3. Gautschi's modification of the Miller algorithm

In GAUTSCHI (1967) the computation of $f_n^{(\nu)}$, $n = 0,\dots,N$, follows a different scheme. It is based on the ratios (we suppose throughout that $f_n \neq 0$)

$$(3.1) \qquad r_n = f_{n+1}/f_n$$

and it originates from continued fractions for these ratios of minimal solutions of three term recursions. From (1.1) it follows that the r_n satisfy the non-linear recursion

$$(3.2) \qquad r_{n-1} = \frac{-b_n}{a_n+r_n}, \qquad n \geq 1$$

and substituting for r_n a similar equation a continued fraction arises. For the partial sums of (2.1) we introduce

$$s_n = \frac{1}{f_n} \sum_{m=n+1}^{\infty} \lambda_m f_m, \tag{3.3}$$

hence for s_n we have the recursion

$$s_{n-1} = r_{n-1}(\lambda_n + s_n), \qquad n \geq 1. \tag{3.4}$$

If r_ν and s_ν are known for some $\nu > N$ the ratios r_n and the partial sums can be obtained from (3.2) and (3.4) respectively, applying these recursions for $n = \nu, \nu-1, \ldots, 1$. In particular we have

$$s_0 = \frac{1}{f_0} \sum_{m=1}^{\infty} \lambda_m f_m = \frac{1}{f_0}(s - \lambda_0 f_0)$$

and so

$$f_0 = s/(\lambda_0 + s_0). \tag{3.5}$$

This gives the initial value of the desired solution. The remaining values follow from $f_n = r_{n-1} f_{n-1}$, $n = 1, \ldots, N$.

In the actual algorithm the quantities r_ν and s_ν for starting the recursions (3.2) and (3.4) are taken equal to zero. The infinite continued fractions r_n and the infinite series s_n are thus replaced by truncated fractions and truncated series $(n < \nu)$. In fact two sequences $\{r_n^{(\nu)}\}$, $\{s_n^{(\nu)}\}$ $(0 \leq n \leq \nu)$ are defined according to the recursion scheme

$$\begin{aligned} & r_\nu^{(\nu)} = 0, \quad r_{n-1}^{(\nu)} = -b_n/(a_n + r_n^{(\nu)}) \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad n = \nu, \ldots, 1 \\ & s^{(\nu)} = 0, \quad s_{n-1}^{(\nu)} = r_{n-1}^{(\nu)}(\lambda_n + s_n^{(\nu)}) \\ & f_0^{(\nu)} = s/(\lambda_0 + s_0^{(\nu)}), \quad f_n^{(\nu)} = r_{n-1}^{(\nu)} f_{n-1}^{(\nu)}, \qquad n = 1, \ldots, N. \end{aligned} \tag{3.6}$$

It can be verified that the quantities $f_n^{(\nu)}$ obtained in this way are the same (mathematically, perhaps not numerically) as those in (2.3) and, as a consequence, the relative errors are as in (2.4).

While algorithm (3.6) and Miller's algorithm (resulting in the computation of (2.3)) are mathematically equivalent, they have different computational characteristics. In many cases, e.g., the quantities $y_n^{(\nu)}$ of (2.2) grow rapidly as ν increases (especially those for small n), and may cause "overflow" on a digital computer. In contrast with this, the quantities $r_n^{(\nu)}$ in (3.6) converge to a finite limit as $\nu \to \infty$, and so does $s_n^{(\nu)}$ if the algorithm converges at all.

When applying the Miller algorithm or Gautschi's version (3.6) of it, one should take care of two points. The first is (it is important for both versions) to take a normalization (2.1) in which no cancellation of leading digits occurs when summing it numerically. Sometimes one has some choice in the selection of (2.1). Consider, for instance, for the computation of the modified Bessel functions the two series

$$e^z = I_0(z) + 2I_1(z) + 2I_2(z) + \dots$$

$$e^{-z} = I_0(z) - 2I_1(z) + 2I_2(z) - \dots$$

for $z \in \mathbb{C}$. For $\operatorname{Re} z \to \infty$ we have $I_n(z) \sim e^z/(2\pi z)^{\frac{1}{2}}$. It follows that the condition function (see §II.1.2) of the first series is much smaller than that of the second one (1 and e^{2z}, respectively, for real positive z).

A second point is that we assumed $f_n \neq 0$. In Gautschi's algorithm this assumption is very important, in Miller's original algorithm it can be dropped. Zero-values of f_n can occur, for instance, in the case of ordinary Bessel functions with

$$a_n = \frac{-2n}{z}, \quad b_n = 1, \quad f_n = J_n(z).$$

Although exact values of zeros of $J_n(z)$ are not representable on the computer (except for $z = 0$) the algorithm may break down in this event. Consider the first elements $r_n^{(\nu)}$ computed according to (3.6):

$$r_\nu^{(\nu)} = 0, \quad r_{\nu-1}^{(\nu)} = \frac{z}{2\nu}, \quad r_{\nu-2}^{(\nu)} = \frac{2\nu z}{4\nu(\nu-1)-z^2} \ .$$

The number ν, the starting value of Miller's algorithm, is (for this case) larger than $|z|$ (GAUTSCHI (1967, p. 51). Hence, the value of $r_{\nu-2}^{(\nu)}$ is well-defined. Values of $r_n^{(\nu)}$ $(n < \nu-2)$ may become undefined, owing to a vanishing denominator. Computer programs must be protected against this phenomenon. According to Gautschi (see the discussion on p. 42 of his paper) the presence of zeros need be of no great concern for the computed values $f_n^{(\nu)}$ in the final step of algorithm (3.6).

3.3.4. Olver's algorithm

This algorithm is already mentioned on p.57 in the previous subsection II.3.2. Here we consider a few practical aspects of it and we will indicate how it can be used in combination with Gautschi's algorithm. In the latter the estimate of ν, see our remarks at the end of §3.3.2, is not very satisfactory, whereas Olver's version is rather attractive for the estimation of ν. We only consider the homogeneous recursion (1.1); in OLVER (1967b) also the inhomogeneous case is treated. The combination of the algorithms of Gautschi and Olver is discussed in OLVER & SOOKNE (1972), where it is applied to the well-used example of the Bessel functions. For the sake of completeness we summarize Olver's algorithm.

Let the given difference equation be denoted by (1.1). We compute a solution $\{p_n\}$ defined by

$$p_0 = 0, \quad p_1 = 1, \quad p_{n+1} = -a_n p_n - b_n p_{n-1} \qquad (n \geq 1).$$

Furthermore we introduce sequences $\{e_n\}$ and $\{E_n\}$ with $e_0 = s$ (see (2.1)) and $e_n = b_n e_{n-1}$ $(n \geq 1)$, and E_n defined as the (necessarily convergent) series

$$E_n = \sum_{m=n}^{\infty} \frac{e_m}{p_m p_{m+1}}, \qquad n \geq 1; \tag{4.1}$$

the process fails if, and only if, one of the numbers p_n vanishes. The above given quantities are used to compute a minimal solution $\{y_n\}$ of (1.1) with initial value $y_0 = s$.

PROPOSITION. *The sequence* $\{y_n\}$ *given by*

$$y_n = p_n E_n = p_n \sum_{m=n}^{\infty} \frac{e_m}{p_m p_{m+1}}, \qquad n \geq 0, \tag{4.2}$$

where for $n = 0$ (4.2) *is to be interpreted as* $y_0 = s$, *is a minimal solution of* (1.1).

PROOF. Substituting (4.2) in (1.1) gives (for $n \geq 1$)

$$p_{n+1}E_{n+1} + a_n p_n E_n + b_n p_{n-1}E_{n-1}$$
$$= p_{n+1}E_{n+1} + a_n p_n (E_{n+1} + \frac{e_n}{p_n p_{n+1}}) + b_n p_{n-1} (\frac{e_{n-1}}{p_{n-1}p_n} + \frac{e_n}{p_n p_{n+1}} + E_{n+1}).$$

Since $\{p_n\}$ is a solution of (1.1) it is easily verified that this expression vanishes identically. From its construction it follows that y_n is a *minimal* solution. □

In Olver's algorithm the wanted solution $\{y_n\}$ is approximated by a finite part of the series in (4.2), viz.

$$(4.3) \qquad y_n^{(\nu)} = p_n \sum_{m=n}^{\nu-1} \frac{e_m}{p_m p_{m+1}}, \qquad 0 \leq n \leq \nu-1,$$

with, again, the assumption $y_0^{(\nu)} = s$. It is easily verified that $\{y_n^{(\nu)}\}$ is also a solution of (1.1) (for $0 \leq n \leq \nu$) with "boundary values"

$$(4.4) \qquad y_0^{(\nu)} = s, \qquad y_\nu^{(\nu)} = 0.$$

The truncation errors and the relative errors are given by

$$(4.5) \qquad y_n - y_n^{(\nu)} = p_n E_\nu, \qquad \frac{y_n - y_n^{(\nu)}}{y_n} = \frac{E_\nu}{E_n},$$

both defined for $n \leq \nu$, but only of interest for $0 \leq n \leq N$.

The value of ν plays the same role as in Miller's algorithm and in Gautschi's version of it, i.e., it is used for starting the backward process for computing the solution $\{y_n^{(\nu)}\}$ of which the values for $n = 0,\nu$ are given in (4.4) and the remaining follow from

$$(4.6) \qquad p_{n+1}y_n^{(\nu)} - p_n y_{n+1}^{(\nu)} = e_n$$

applied successively for $n = \nu-1,\nu-2,\ldots,1$. Here the quantities E_i are used to decide whether the error is satisfactorily small. If the infinite series (4.1) are replaced by their first terms then the second of (4.5) reduces to

$$(4.7) \qquad \frac{y_n - y_n^{(\nu)}}{y_n} \doteq \frac{e_\nu}{e_n} \frac{p_n p_{n+1}}{p_\nu p_{\nu+1}}, \qquad 0 \leq n \leq N.$$

Thus, the relative error is easily computed (approximately) by the quantities p_i and e_i. The right-hand side of (4.7) is computed for $\nu = N+1, N+2, \ldots$ until they fall below the desired relative accuracy. Since the E_i in (4.5) are replaced by approximations it is not proved that a value of ν accepted in this way is a correct value. To make the choice more rigorous one may use bounds for the solution $\{p_n\}$ in order to obtain upper bounds for $|E_\nu/E_n|$. Theorems and examples in OLVER (1967c) may be useful in this connection.

For a full understanding of Olver's method we remark that (4.6) can be conceived as a first order recursion for $y_n^{(\nu)}$. As observed on p.58 in the previous subsection II.3.2 the original recursion (1.1) is reduced in order: the difficult problem for the second order recursion is reduced to a perhaps less difficult problem for first order recursion. In this connection the theory of subsection II.3.1 may be important.

The algorithm for the computation of $y_n^{(\nu)}$ is not always well-conditioned. This may be analysed by using the results of II.3.1. In some cases instabilities occur due to a loss in accuracy in the formation of the sequence $\{p_n\}$ (initially p_n may be like a multiple of the minimal solution, although it increases ultimately in proportion to the dominant solution). Therefore we remark that the two values (4.4) can also be used to compute $y_n^{(\nu)}$ with the help of Gautschi's algorithm (3.6) with the simple normalization $y_0^{(\nu)} = s$ (i.e., $\lambda_0 = 1$, $\lambda_m = 0$, $m \geq 1$). In OLVER & SOOKNE (1972) this device is followed for the computation of ordinary Bessel functions.

It remains to give information for the computation of a minimal solution of (1.1) in the case of a general normalization (2.1).

One could reason as follows (however it will be a false reasoning). Suppose we have computed (within a given accuracy) a minimal solution $\{y_n\}$ of (1.1) with initial condition $y_0 = s$ as a simple normalization. As mentioned in §3.3.1 any other minimal solution $\{f_n\}$ (satisfying f.i. a general normalization relation (2.1)) is a multiple of y_n. That is, by using (2.1), we infer that

$$(4.8) \qquad f_n = \frac{s}{t}\, y_n, \qquad t = \sum_{n=0}^{\infty} \lambda_n y_n .$$

For computations we suppose that in this series and in (2.1) the symbol ∞ is replaced by ν. Then, for ν we have two conditions

(i) to make the second of (4.5) or (4.7) small enough,

(ii) to make both infinite series

$$\sum_{m=\nu+1}^{\infty} \lambda_m y_m \qquad \text{and} \qquad \sum_{m=\nu+1}^{\infty} \lambda_m f_m$$

small enough.

Moreover we suppose that the y_n are replaced by $y_n^{(\nu)}$, of which the computation is described earlier in this part.

This reasoning is used in OLVER & SOOKNE (1972, p. 945) (in fact only condition (i) is mentioned) and we will show, as is done properly in OLVER (1967b, §11), how to obtain a correct condition on ν. The point is that the computed y_n (i.e., $y_n^{(\nu)}$) is not an exact minimal solution, since it is computed with two conditions given in (4.4). In §3.3.1 we remarked that for a minimal solution one and only one value can be prescribed.

Let $\{f_n\}$ be the wanted minimal solution of (1.1) (to be computed for $n = 0,1,\ldots,N$) with normalization (2.1). Let $\{y_n^{(\nu)}\}$ be computed as above with condition (4.4), and $\{y_n\}$ the exact minimal solution of (1.1) with $y_0 = s$. Then we have (compare (4.5))

$$(4.9) \qquad y_n - y_n^{(\nu)} = p_n E_\nu, \qquad 0 \le n \le \nu.$$

Using (4.8) we obtain

$$(4.10) \qquad f_n = \frac{s}{t}(y_n^{(\nu)} + p_n E_\nu) = \frac{s}{t_\nu + T_\nu}\,(y_n^{(\nu)} + p_n E_\nu)$$

with

$$t_\nu = \sum_{n=0}^{\nu} \lambda_n y_n^{(\nu)}, \qquad T_\nu = E_\nu \sum_{n=0}^{\nu} \lambda_n p_n + \sum_{n=\nu+1}^{\infty} \lambda_n y_n .$$

In (4.10), t_ν and $y_n^{(\nu)}$ are known whenever a choice of ν is made. The small quantities T_ν and $p_n E_\nu$ are not known. We approximate f_n of (4.10) by $f_n^{(\nu)}$ defined by

$$(4.11) \qquad f_n^{(\nu)} = \frac{s}{t_\nu}\, y_n^{(\nu)}, \qquad n = 0,1,\ldots,N.$$

Then the relative error in this approximation is obtained by using (4.10) and (4.11), that is,

$$\frac{f_n-f_n^{(\nu)}}{f_n^{(\nu)}} = \frac{-sy_n^{(\nu)}T_\nu/t_\nu+sp_nE_\nu}{f_n^{(\nu)}(t_\nu+T_\nu)} .$$

To the first order of small quantities we have $t_\nu + T_\nu \doteq t_\nu$, $p_nE_\nu/y_n^{(\nu)} \doteq p_nE_\nu/y_n = E_\nu/E_n$ (see (4.2)). We obtain for the relative error approximately

$$\text{(4.12)} \qquad \frac{f_n-f_n^{(\nu)}}{f_n^{(\nu)}} \doteq E_\nu/E_n - T_\nu/t , \qquad 0 \le n \le N < \nu.$$

The first part corresponds with condition (i) on page 79. The second part, which does not depend on n, is connected with condition (ii). It is clear that it contains more than the series mentioned there. Actually we have

$$\text{(4.13)} \qquad T_\nu/t_\nu = \frac{E_\nu \sum_{n=0}^{\nu} \lambda_n p_n + \sum_{n=\nu+1}^{\infty} \lambda_n p_n E_n}{\sum_{n=0}^{\nu} \lambda_n y_n^{(\nu)}} .$$

If more information on p_n and E_n and the remaining quantities is available this expression can be estimated further. For the present discussion the only possible step is to replace the series by their most relevant terms, viz. $T_\nu/t_\nu \doteq (\lambda_\nu p_\nu E_\nu + \lambda_{\nu+1}p_{\nu+1}E_{\nu+1})/(\lambda_0 s)$. Using the first term of (4.1) we obtain

$$\text{(4.14)} \qquad T_\nu/t_\nu \doteq (\lambda_\nu e_\nu/p_{\nu+1} + \lambda_{\nu+1}e_{\nu+1}/p_{\nu+2})/(\lambda_0 s)$$

and this expression is easily computed.

CONCLUSION

Although Olver's algorithm gives a better control on error analysis than the Miller algorithm, in the final stage of the above analysis approximations are used. In general one has to use such approximations for obtaining the starting value ν of the backward approximation process. For special cases bounds for p_i and E_i may be constructed in order to obtain more rigorous and possibly strict error bounds. We believe, however, that the choice of ν based on testing the smallness of (4.7) and (4.14) is more reliable than the estimations based on asymptotic expressions (as mentioned in §3.3.2), whenever these are available.

4. CONTINUED FRACTIONS

In this chapter we discuss continued fractions. In section 4.1 through 4.4 some basic theory about continued fractions is introduced. In sections 4.5 through 4.7 we treat the approximation of (infinite) continued fractions and the evaluation of these approximations. This material can be found scattered in the literature. New is an estimation for the condition in section 4.8. In section 4.9 we give some examples.

4.1. Introduction

In this section we introduce continued fractions and establish some notations and definitions.

A mathematical function can often be represented by a *continued fraction*. A continued fraction is defined as an ordered pair $((\{a_n\},\{b_n\}),\{c_n\})$, where $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ are complex numbers with all $a_k \neq 0$ and where $\{c_n\}$ is a sequence in the extended complex plane defined as follows:

$$(4.1.1)\qquad \begin{cases} c_k = S_k(0), \quad k = 1,2,\ldots \quad \text{where} \\ S_0(w) = w, \quad S_k(w) = S_{k-1}(s_k(w)) \quad k = 1,2,\ldots \quad \text{and} \\ s_k(w) = a_k/(b_k+w), \quad k = 1,2,\ldots\ . \end{cases}$$

The *continued-fraction algorithm* is the function Φ which assigns to each pair $(\{a_n\},\{b_n\})$ the sequence $\{c_n\}$.

The prescriptions to perform the operations may be denoted by

$$\cfrac{a_1}{b_1+\cfrac{a_2}{b_2+\cfrac{a_3}{b_3+\ldots}}}\ ,$$

typographically this is not convenient, so we write

$$\frac{a_1\rvert}{\lvert b_1} + \frac{a_2\rvert}{\lvert b_2} + \frac{a_3\rvert}{\lvert b_3} + \ldots \quad \text{or} \quad \frac{a_1}{b_1+}\,\frac{a_2}{b_2+}\,\frac{a_3}{b_3+}\,\ldots\ .$$

We also use

$$(4.1.2) \qquad \overset{\infty}{\underset{i=1}{\Phi}} \frac{a_i}{b_i} \; ,$$

in analogy to series and the Σ-symbol. From this continued fraction we call

a_k the *k-th partial numerator,*

b_k the *k-th partial denominator,*

$c_k = \overset{k}{\underset{i=1}{\Phi}} \frac{a_i}{b_i}$ the *k-th convergent.*

A continued fraction is said to *converge* if the sequence $\{c_n\}$ converges. The *value,* c, of the continued fraction is the limit of $\{c_n\}$.

The analytic behaviour of continued fractions is treated in WALL (1948), in PERRON (1950) and KHOVANSKII (1956). More recent views on the matter and the applications of continued fractions in numerical analysis are found in HENRICI (1977a) and JONES & THRON (1980). This chapter leans heavily upon the last book. Recent conference proceedings are JONES, THRON & WAADELAND (1982).

4.2. Some examples

In this section we demonstrate some methods to construct a continued fraction.

In order to construct a simple example, which will be useful further on, we look at the quadratic equation

$$(4.2.1) \qquad z^2 - bz - a = 0 \; ,$$

where the roots, z_1 and z_2, satisfy the two equations

$$\begin{cases} z_1 + z_2 = b \\ z_1 z_2 = -a \; . \end{cases}$$

Eliminating z_2 we get

$$z_1 = -a/(b-z_1) \, .$$

It is easily verified that, if we take in (4.1.1)

$$s_k(w) = s(w) = a/(b+w),$$

the continued fraction, thus defined, has the convergents

$$c_k = s(s(\ldots(s(0))\ldots)$$

and the limit, c, (if existing) of $\{c_n\}$ has the property

$$c = s(c).$$

This enables us to write

$$(4.2.2) \qquad z_1 = -c = -\mathop{\Phi}_{i=1}^{\infty} a/b \ .$$

As an illustration we take a = b = 1 and find for the golden ratio, r,

$$(4.2.3) \qquad z_1 = (\sqrt{5-1})/2 = r = \mathop{\Phi}_{i=1}^{\infty} 1/1 \ .$$

In a similar way, due to Gauss, we find for the quotient of two hypergeometric series a continued fraction

$$\frac{F(a,b+1;c+1;z)}{F(a,b;c;z)} = \cfrac{1}{1-\frac{a(c-b)z}{c(c+1)}\cdot\frac{F(b+1,a+1;c+2;z)}{F(b+1,a;c+1;z)}} \qquad \text{or}$$

$$(4.2.4) \qquad \begin{cases} \dfrac{F(a,b;c;z)}{F(a,b+1;c+1;z)} = 1 + \mathop{\Phi}\limits_{i=1}^{\infty} \dfrac{-d_i z}{1} \\ \left.\begin{array}{l} d_{2k} = \dfrac{(b+k)(c-a+k)}{(c+2k)(c+2k-1)} \\ d_{2k+1} = \dfrac{(a+k)(c-b+k)}{(c+2k+1)(c+2k)} \end{array}\right\} \quad k = 0,1,\ldots \ . \end{cases}$$

A more formal treatment of the convergence of this continued fraction and various applications is given in JONES & THRON (1980, §6.1.1).

4.3. Some relations

In this section we investigate the connection between continued fractions and other parts of mathematical analysis, in order to be able to use the theory developed elsewhere.

A well known and useful part of elementary continued fraction theory is that if we introduce two sequences $\{p_k\}$ and $\{a_k\}$, which are defined by

$$\text{(4.3.1)} \qquad \left.\begin{array}{l} p_{-1} = 1; \quad p_0 = 0; \quad p_k = b_k p_{k-1} + a_k p_{k-2} \\ q_{-1} = 0; \quad q_0 = 1; \quad q_k = b_k q_{k-1} + a_k q_{k-2} \end{array}\right\} \qquad k = 1,2,\ldots ,$$

it can be proved that for c_k, the k-th convergent of $\mathop{\Phi}_{i=1}^{\infty} a_i/b_i$, holds

$$\text{(4.3.2)} \qquad c_k = p_k/q_k \qquad k = 1,2,\ldots .$$

For the continued fraction (4.2.3) we get $c_k = F_{k-1}/F_k$, where $\{F_n\}$ are the Fibonacci numbers.

More important is that we have connected continued fractions with recurrence relations, see also JONES & THRON (1980, §5.2). Three-term recurrence relations (like (4.3.1)) are surveyed by GAUTSCHI (1967).

In order to link continued fractions with series we take (4.3.1) and (4.3.2) and we get

$$c_i - c_{i-1} = (-1)^{i+1} \frac{\prod_{j=1}^{i} a_j}{q_{i-1} q_i} .$$

From this and $c_n = \Sigma_{j=1}^{n} (c_j - c_{j-1})$ we get

$$\text{(4.3.3)} \qquad \mathop{\Phi}_{i=1}^{n} \frac{a_i}{b_i} = \sum_{i=1}^{n} (-1)^{i+1} \frac{\prod_{j=1}^{i} a_j}{q_i q_{i-1}} .$$

Conversely, there is the identity of Euler

$$\text{(4.3.4)} \qquad \sum_{i=0}^{n} d_i x^i = \frac{d_0}{1 + \mathop{\Phi}_{i=1}^{n} \dfrac{-(d_i/d_{i-1})x}{1+(d_i/d_{i-1})x}} .$$

A very much related and fertile part of the mathematical theory is touched upon when we consider a continued fraction as a function, c, of a parameter, x, with (formal) power series expansion $\Sigma\gamma_i x^i$. We approximate c(x) by a rational function

$$P_n(x)/Q_m(x) \quad ,$$

where P_n is a polynomial in x of degree at most n and Q_m of degree at most m. We can choose the approximation so that

$$c(x)Q_m(x) + P_n(x)$$

has a (formal) power series expansion $\Sigma\ \delta_i x_i$, in which $\delta_i = 0,\ 0 \le i \le n+m$. If we impose a normalization condition and if we require that P_n and Q_m have no common factors, we can prove that P_n and Q_m are unique.

Frobenius conceived P_n/Q_m as an element of matrix and Padé developed the theory; we say that P_n/Q_m occupies the position (n,m) of the *Padé table*. It can be proved (see JONES & THRON (1980, Theorem 5.19)) that the convergents of a continued fraction c(x) occupy the stair step sequence

$$P_0/Q_0,\ \ P_1/Q_0,\ \ P_1/Q_1,\ \ P_2/Q_1,\ \ P_2/Q_2,\ \ \ldots$$

of the Padé table of the power series $\Sigma\ \gamma_i x^i$ of c(x). Theory about Padé tables can be found in BAKER (1975), GILEWICZ (1978) and JONES, THRON & WAADELAND (1982, §5.5). Recent conference proceedings are CABANNES (1976) and WUYTACK (1979). Connections between Padé tables and numerical analysis are surveyed in WUYTACK (1976). A bibliography of Padé approximations and related matters like continued fractions is BREZINSKI (1976).

4.4. Some transformations

In this section we point out some ways to simplify a continued fraction, without changing its value.

Different sequences $\{a_n\}$ and $\{b_n\}$ can lead to the same sequence of convergents. To establish some transformations of $\{a_n\}$ and $\{b_n\}$ we conceive $\{c_n\}$ as

$$(4.4.1)\qquad \begin{cases} c_k = T_k(0), \quad k = 1,2,\ldots \qquad \text{where} \\ T_0(w) = w, \quad T_k(w) = T_{k-1}(t_k(w)), \quad k = 1,2,\ldots \quad . \end{cases}$$

In order to confirm (4.1.1) t_n must have the form

$$(4.4.2)\qquad t_k(w) = \begin{cases} \dfrac{T_{k-1}^{-1}(c_k) + x_k w}{1+y_k w} & \text{if} \quad T_{k-1}^{-1}(c_k) < \infty \\ z_k/w + x_k & \text{else,} \end{cases}$$

where $\{x_n\}$, $\{y_n\}$ and $\{z_n\}$ can be arbitrarily chosen.
Also it can be shown that if we take for t_k the transform

$$t_k(w) = \frac{\alpha_k + \gamma_k(w)}{\beta_k + \delta_k(w)}, \quad k = 1,2,\ldots ,$$

we can construct a continued fraction in such a way that $c_k = T_k(0)$. For proof and details see JONES & THRON (1980, §2.4) or THRON & WAADELAND (1982). These two theoretical results have the practical implication (especially because the arbitrary construction of t_k) that we have a certain degree of freedom in the choice of $\{a_n\}$ and $\{b_n\}$. In fact we see immediately that we can safely write instead of (4.1.2)

$$\frac{a_1 r_1|}{|b_1 r_1} + \frac{a_2 r_1 r_2|}{|b_2 r_2} + \frac{a_3 r_2 r_3|}{|b_3 r_3} + \ldots ,$$

which leads, with a suitable choice for $\{r_i\}$, to a continued fraction like

$$\overset{\infty}{\underset{i=1}{\Phi}} 1/\beta_i \quad \text{or} \quad \overset{\infty}{\underset{i=1}{\Phi}} \alpha_i/1 .$$

Bernoulli found for the problem to construct a continued fraction with known convergents

$$a_1 = c_1 \; ; \quad b_1 = 1 \quad \text{and (supposing } c_0 = 0)$$

$$a_n = \frac{c_{n-1} - c_n}{c_{n-1} - c_{n-2}} ; \quad b_n = \frac{c_n - c_{n-2}}{c_{n-1} - c_{n-2}} \quad \text{for} \quad n = 2,3,\ldots \quad ,$$

which is, essentially, the simplest form of (4.4.2).
In order to obtain another practical result we take for t_k

$$t_k(w) = S_{2k-1}(S_{2k}(w)) \ .$$

Evaluating this and constructing a new continued fraction we get two new sequences $\{a_n^*\}$ and $\{b_n^*\}$ with convergents $\{c_{2n}\}$. This continued fraction is called an even *contraction* of the original one. Of course many other contractions are possible, for example odd contractions.
The theory gives for the sequences $\{a_n^*\}$ and $\{b_n^*\}$

$$a_1^* = a_1 b_1 \ , \quad b_1^* = a_2 + b_1 b_2$$

$$\left.\begin{aligned} a_k^* &= (-a_{2k-2}\ a_{2k-1}\ b_{2k})/b_{2k-2} \\ b_k^* &= \frac{a_{2k-1}b_{2k}+a_{2k}b_{2k-2}+b_{2k-2}b_{2k-1}b_{2k}}{b_{2k-2}} \end{aligned}\right\} \ k = 2,3,\ldots \ ,$$

which applied to (4.2.3) gives

$$r = (\sqrt{5-1})/2 = 1/(2+\overset{\infty}{\underset{i=2}{\Phi}}-1/3) \ .$$

The advantage of contractions is of course the possibility to get better approximations from the same computing effort.

4.5. Convergence of continued fractions

In this section we investigate the convergence of c_k, *the* k-th *convergent of a continued fraction.*

To obtain an easy-to-use result we suppose that in (4.1.2) $a_i > 0$ and $b_i > 0$ for $i = 1,2,\ldots$. Then the convergents show the following pattern:

$$(4.5.1) \qquad 0 < c_2 \le \ldots \le c_{2k} \le \ldots \ \ldots \le c_{2k+1} \le \ldots \le c_1 \ .$$

This alternating behaviour is illustrated by the convergents of (4.2.3):

$$
\begin{aligned}
c_1 &= 1 \\
c_2 &= .5 \\
c_3 &= .66\ldots \\
c_4 &= .6 \\
c_5 &= .625 \\
c_6 &= .615\ldots \\
c_7 &= .619\ldots \\
&\vdots \\
r = c &= .61803\ldots = (\sqrt{5-1})/2 \; .
\end{aligned}
$$

In this case we see that convergence means that both

$$\lim_{n\to\infty} c_{2n} \quad \text{and} \quad \lim_{n\to\infty} c_{2n+1}$$

exist and are finite and equal.

Convergence theory is a fascinating topic in continued fractions but the reader has to turn to the theoretical works mentioned in §4.1. A survey of recent results is given in THRON (1974). We will just mention three important theorems.

Worpitzky's theorem states that

$$c_n = \overset{n}{\underset{i=1}{\Phi}} \frac{\alpha_i}{1}$$

converges to a value c for n → ∞ if

$$|\alpha_i| \le 1/4 \quad \text{for} \quad i = 2,3,\ldots \quad ,$$

moreover we have

$$|c| < 1/2 \quad \text{and} \quad |c_n - c| \le \frac{1}{2n+1} \; .$$

This convergence region can be generalized to a parabolic one. See also JONES & THRON (1980, §4.4).

The well-known theorem of Seidel states that (4.1.2) with $a_i > 0$ and $b_i > 0$ for $i = 1,2,\ldots$ converges iff at least one of the series

$$\sum_{i=1}^{\infty} \frac{\prod_{k=1}^{i} a_{2k-1}}{\prod_{k=1}^{i} a_{2k}} b_{2i} \quad \text{and} \quad \sum_{i=1}^{\infty} \frac{\prod_{k=1}^{i} a_{2k}}{\prod_{k=0}^{i} a_{2k+1}} b_{2i+1}$$

is divergent. Note how easily this is applied to the continued fractio (4.2.3).

One of the theorems of Van Vleck states that

$$c_n = \frac{a_0|}{|1} + \frac{a_1 z|}{|1} + \frac{a_2 z|}{|1} + \dots + \frac{a_n z|}{|1}$$

converges to a function $f(z)$ if

$$\lim_{n\to\infty} a_n = a < \infty .$$

if $a = 0$ this convergence is at any closed region that contains no pol of f.
If $a \neq 0$ this convergence is for all z outside a cut $\{z \mid |z| > |\tfrac{1}{4}a|,$ $\arg(z) = \arg(-\tfrac{1}{4}a)\}$ from $-\tfrac{1}{4}a$ to ∞ in the direction of $-\tfrac{1}{4}a$ and outside t poles of f.

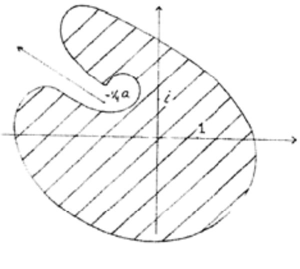


Figure 1.

Having ensured the convergence of c_n to c, one is, for practical sons, interested in the speed of this convergence. For a survey paper FIELD (1977). In order to use results about the convergence of series construct a sequence $\{\sigma_i\}$, so that

$$(4.5.2) \qquad \overset{n}{\underset{i=1}{\Phi}} \frac{a_i}{b_i} = \sum_{i=1}^{n} \prod_{j=1}^{i} (\sigma_j - 1) \ .$$

From (4.3.3) we develop this sequence as follows

$$(4.5.3)\qquad \begin{cases} \sigma_1 = \dfrac{a_1+b_1}{b_1}\ ; \qquad \sigma_2 = \dfrac{1}{1+a_2/b_1b_2} \\ \sigma_k = \dfrac{1}{a+(a_{k+1}/b_kb_{k+1})\sigma_k} \qquad \text{for } k = 2,3,\dots\ . \end{cases}$$

GAUTSCHI & SLAVIK (1978) apply this to compose the speed of convergence of two continued fraction with the same value.

For a simple theorem we look at a continued fraction like (4.1.2) with $\lim_{n\to\infty} a_n = a$ and $\lim_{n\to\infty} b_n = b$. We can prove that for a certain N

$$\mathop{\Phi}\limits_{i=N}^{\infty} a_i/b_i$$

converges to z_1 with a *geometric convergence rate* $|z_2/z_1|$ $(|z_1|>|z_2|)$, where z_1 and z_2 are the roots of (4.2.1), see also SAUER & SZABO (1968). For a brief explanation and more examples of the geometric convergence rate see GAUTSCHI (1983).

To consider an example we look at

$$(4.5.4)\qquad \frac{\ln(1+x)}{x}$$

and we construct a continued fraction of it, using (4.2.4)

$$\frac{\ln(1+x)}{x} = \frac{\ln(1+x)/x}{1} = \frac{F(1,1;2;-x)}{F(1,0;1;-x)}\ .$$

So we have

$$\frac{\ln(1+x)}{x} = \frac{1|}{|1} + \frac{x/2|}{|1} + \dots + \frac{kx/(4k-2)|}{|1} + \frac{kx/(4k+2)|}{|1} + \dots$$

and $a = x/4$; $b = 1$. The continued fraction converges for $x > -1$ like a geometric series with quotients

$$\frac{1-\sqrt{1+x}}{1+\sqrt{1+x}}\ .$$

The Taylor-series for (4.5.4) is

$$1 - x/2 + x^2/3 - x^3/4 + \dots\ ,$$

which converges for $|x| < 1$ with a geometric convergence rate x.
So we see that the continued fraction has a greater region of convergence and since $|1-\sqrt{1+x}|/|1+\sqrt{1+x}| < |x|$, (x complex, $|x|<1, x\neq 0$), it converges faster that the Taylor series (for $|x|<1$).

Moreover we can accelerate the convergence if we calculate the sequence $\{c_n^*\}$ with $c_k{*} = S_k(t_k)$, with t_k (instead of zero as in 4.1.1) a suitable approximation of the "tail", see JACOBSEN & WAADELAND (1982). In THRON & WAADELAND (1980) there is taken $t_n = \Phi_{i=n}^{\infty}\ a/b$, which is minus a root of equation (4.2.1). It is shown there and in JONES, THRON & WAADELAND (1982, §8.4) that $\lim_{n\to\infty}|c_n^*-c_n| = 0$ and an upperbound for $|c_n^*-c_n|$ is given. All these approximations and the function itself are displayed in Figure 2.

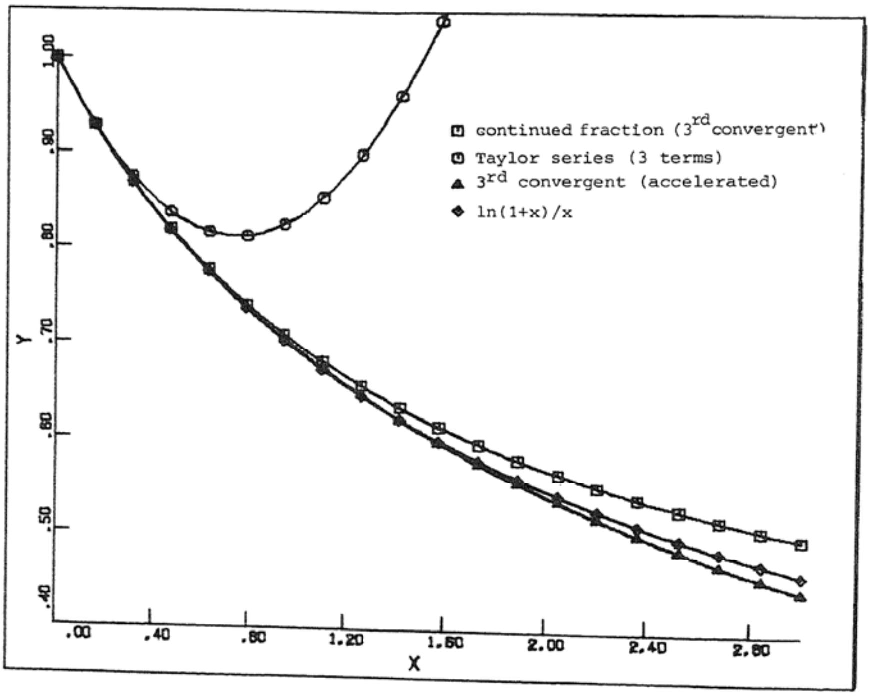


Figure 2. Approximations of $\ell n(1+x)/x$

4.6. Truncation errors

In this section we estimate the "errors" made while approximating the value of a continued fraction.

An easy case to find an estimation or bound for the difference between the value of a (convergent) continued fraction and a convergent is (4.1.2) with $a_i > 0$ and $b_i > 0$ for $i = 1,2,\ldots$. Here it is clear from (4.5.1) that the truncation error is smaller than the difference between two consecutive convergents. In general this is, however, not true.

To get further insight we look at a simple example

$$\frac{1|}{|2} - \frac{1|}{|2} - \frac{1|}{|2} - \ldots . \tag{4.6.1}$$

This continued fraction has, see (4.2.1) and (4.2.2), the value 1, being a root of

$$z^2 + z - 2 = 0.$$

For the convergents of (4.6.1) we get $c_k = \frac{k}{k+1}$. So the truncation error is now

$$1 - c_n = \frac{1}{n+1} ,$$

however

$$c_n - c_{n-1} = \frac{1}{n(n+1)} .$$

It should be noted that in this example the difference between two consecutive convergents is of a lower order of magnitude than the truncation error. This remark can be generalized as follows:

Consider the continued fraction $\underset{i=1}{\overset{\infty}{\Phi}} \alpha_i/1$ and $\underset{i=1}{\overset{\infty}{\Phi}} 1/\beta_i$ and let

$$|\alpha_i| \le 1/4 - \varepsilon \quad \text{and} \quad |\beta_i| \ge 2(1+\varepsilon) \ , \ \varepsilon > 0, \quad \text{for} \quad i = 1,2,\ldots .$$

(So both are convergent). For these continued fractions we can prove

$$0 < |c-c_k| \le \frac{1-2\sqrt{\varepsilon}}{2\sqrt{\epsilon}} |c_k-c_{k-1}|$$

for the first one and

$$0 < |c-c_k| \le (\sqrt{\frac{1}{4}+\frac{1}{2\varepsilon}}-\frac{1}{2}) |c_k-c_{k-1}|$$

for the second one.
For the proof itself and more details see BLANCH (1964) or JONES & THRON (1971) or JONES & THRON (1980, §8.3).
For more specialized results see FIELD & JONES (1972), GRAGG (1968), HENRICI & PFLUGER (1966), JEFFERSON (1969), JONES & SNELL (1969) and GILL (1982).

4.7. A special type of continued fractions

In this section we treat Stieltjes fractions and the way to construct them.

Frequently one considers continued fractions with partial numerators and denominators of a special form (g-fractions or T-fractions, for example). For the representation of functions the z-fraction can be usefull. This is a continued fraction of the form

$$(4.7.1) \qquad \frac{e_0|}{|z} - \frac{f_1|}{|1} - \frac{e_1|}{|z} - \frac{f_2|}{|1} - \dots$$

and is, of course, a function of z. The partial numerators and denominators are complex numbers different from zero. To every z-fraction, there corresponds exactly one formal power series in z^{-1}

$$(4.7.2) \qquad \sum_{i=0}^{\infty} \frac{d_i}{z^{i+1}}$$

such that the expansion of the n-th approximant of (4.7.1) in powers of z^{-1} agrees with (4.7.2) through the term $d_{n-1}\ z^{-n}$ for $n = 1,2,\dots$.

Theory about existence and convergence can be found in JONES & THRON (1980, §7.1). A more detailed review can also be found in SAUER & SZABO (1968). Inclusion regions (depending on the convergents) for the value of a Stieltjes fraction and bounds for the truncation error are given in

HENRICI & PFLUGER (1966).

Now we look at the problem of calculating $\{f_i\}$ and $\{e_i\}$, given the sequence $\{d_i\}$, i.e., how to determine (4.7.1) from (4.7.2). The following algorithm, called the Q(uotient) D(ifference) algorithm is due to Rutishauser. It is obvious that $e_0 = d_0$, then let us introduce the formal series

$$F_k(z) = \sum_{i=0}^{\infty} \frac{d_{i+k}}{z^{i+1}}$$

and suppose that $F_k(z)$ and

$$c_k(z) = \frac{d_k|}{|z} - \frac{f_{1,k}|}{|1} - \frac{e_{1,k}|}{|z} - \frac{f_{2,k}|}{|1} - \frac{e_{2,k}|}{|z} - \dots$$

are corresponding (often denoted as $F_k(z) \sim c_k(z)$). From contractions we get continued fractions for $zF_k(z) - d_k$ and $F_{k+1}(z)$, which must be equal and so we get the recursion relations

$$e_{\ell-1,k+1} + f_{\ell,k+1} = f_{\ell,k} + e_{\ell,k};\ e_{0,k} = 0$$

(this is used for the calculation of $\{e_{i,j}\}$)

and

$$f_{\ell,k+1}e_{\ell,k+1} = e_{\ell,k}f_{+1,k}; \quad f_{1,k} = \frac{d_{k+1}}{d_k}$$

(this is used for the calculation of $\{f_{i,j}\}$).

It can be shown that no zero divisors can occur. So the scheme is well defined. However the numerical stability can be poor. GARGANTINI & HENRICI (1967) explain this and they construct a stable form of the algorithm.

4.8. Evaluation

In this section we discuss methods for calculating a convergent of a continued fraction and their consequences.

A convergent can be numerically evaluated in basicly three ways. The backward computation of, say, $\overset{n}{\underset{i=1}{\Phi}} \frac{a_i}{b_i}$ can be done as follows:

$$d_{n+1} := 0;\ \ d_i := a_i/(b_i+d_{i+1}) \qquad \text{for } i = n,n-1,\dots,1.$$

This can be conceived as a Miller algorithm and yields one convergent, $c_n = d_1$. A recursive operator for this method is programmed in VAN WIJNGAARDEN (1976).

As can be seen from (4.3.1) and (4.3.2) one can use two three-term recurrence relations to construct a sequence of convergents. A relation like (4.3.1) can be considered as a triangular system of n equations and n unknows. In MIKLOŠKO (1977) this is used to speed up the computation. A third method can be derived from (4.3.3) as follows (see also (4.5.2) and (4.5.3)):

$$u_1 := 1; \quad v_1 := c_1 := a_1/b_1$$

$$\left.\begin{aligned} u_{k+1} &:= \cfrac{1}{1+\cfrac{a_{k+1}}{b_k b_{k+1}} - u_k} \\ v_{k+1} &:= v_k(u_{k+1}-1) \\ c_{k+1} &:= c_k + v_{k+1} \end{aligned}\right\} \qquad \text{for } k = 1,2,\ldots\ .$$

While evaluating a continued fraction we have to deal with two types of errors. If we don't have an exact value for the tail, we have to cope with the truncation error. Besides that, there can be errors due to finite arithmethic. For bounds of these errors see BLANCH (1964), JONES & THRON (1974), MIKLOŠKO (1976) or JONES & THRON (1980, §10.1). The work of Blanch seems to indicate that the backward algorithm is numerically more stable than the forward algorithm.

We will look at the condition (see section II.1). Suppose we want to approximate the value of $\Phi_{i=1}^{\infty} \frac{\alpha i}{1}$, this can only make sense if the continued fraction converges. In this case, see section 4.5, convergence is ensured if $|\alpha_i| \le \frac{1}{4}$ for $i = 1,2,\ldots$. For establishing the condition of the n-th convergent, c_n, we need to evaluate the expression

$$\frac{\alpha_i}{c_n} \frac{\partial c_n}{\partial \alpha_i} \qquad i = 0,1,\ldots,n-1,$$

and for notation purposes we will write for a certain fixed n:

$$c_n = \frac{\alpha_1}{r_1}\ ,\ r_i = 1 + \frac{\alpha_{i+1}}{r_{i+1}} \qquad \text{for} \quad i = 1,2,\ldots,n-1,\ r_n = 1.$$

We now have

$$\frac{\partial c_n}{\partial \alpha_i} = \frac{\partial c_n}{\partial r_1} \left(\prod_{k=1}^{i-2} \frac{\partial r_k}{\partial r_{k+1}} \right) \frac{\partial r_{i-1}}{\partial \alpha_i} = - \frac{\alpha_1}{r_1^2} \left(\prod_{k=1}^{i-2} - \frac{\alpha_{k+1}}{r_{k+1}^2} \right) \frac{1}{r_i} ,$$

which, with $\alpha_{i+1} = r_{i+1}(r_i-1)$ gives

$$\frac{\alpha_i}{c_n} \frac{\partial c_n}{\partial \alpha_i} = (-1)^{i-1} \prod_{k=1}^{i-1} \frac{r_k-1}{r_k} .$$

To establish a bound for the condition, we have to estimate $|(r_k-1)/r_k|$. Therefore suppose

$$(4.8.1) \qquad |r_k-1| \le (n-k)/(2(n-k)+2) ,$$

which is trivial for k = n. We then have

$$|r_{k-1}-1| = |\frac{\alpha_k}{r_k}| \le \frac{1/4}{1-\frac{(n-k)}{2(n-k)+2}} = \frac{n-k+1}{2(n-k+1)+2} ,$$

because we have imposed $|\alpha_i| \le 1/4$ for $i = 1,2,\dots,n$. So (4.8.1) is true by induction and can be used to derive

$$|\frac{r_{k-1}}{r_k}| < \frac{n-k}{n-k+2} .$$

Using this we get

$$|\frac{\alpha_i}{c_n}\frac{\partial c_n}{\partial \alpha_i}| \le \frac{(n-i+2)(n-i+1)}{n(n+1)} \qquad \text{for} \quad i = 1,2,\dots,n$$

and the estimation for the condition of c_n is now

$$\sum_{i=1}^{n} |\frac{\alpha_i \partial c_n}{c_n \partial \alpha_i}| \le \frac{\sum_{k=1}^{n} k(k+1)}{n(n+1)} = (n+2)/3 .$$

This can be considered small. To illustrate this the condition of three representations of ln(1+x)/x (this function is plotted in section 4.5) is plotted in Figure 3.

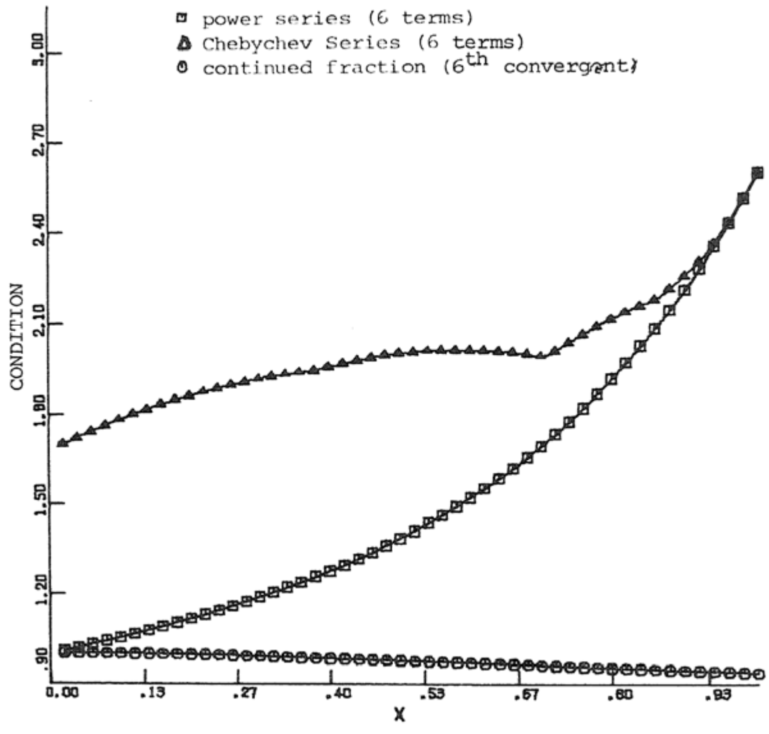


Figure 3. Condition of approximations for $\ell n(1+x)/x$

4.9. Examples of special functions

In this section we construct a continued fraction for the error function, the gamma function and for confluent hypergeometric functions.

The complementary error function

$$\mathrm{erfc}(z) = \frac{2}{\sqrt{\pi}} \int_z^\infty e^{-t^2} dt$$

can be written as a Stieltjes fraction, see section 4.7. To do this we proceed as follows:
From the asymptotic expansion

$$\frac{\sqrt{\pi}}{2}\,\mathrm{erfc}(z) \sim ze^{-z^2}\left(\frac{1}{2z^2} - \frac{1}{(2z^2)^2} + \frac{1.3}{(2z^2)^3} - \ldots\right)$$

and the QD algorithm we get

$$\sqrt{\pi}\,\frac{e^{z^2}}{z}\,\mathrm{erfc}(z) = \frac{1\rvert}{\lvert z^2} + \frac{1/2\rvert}{\lvert 1} + \frac{2/2\rvert}{\lvert z^2} + \frac{3/2\rvert}{\lvert 1} + \frac{4/2\rvert}{\lvert z^2} + \ldots .$$

See also ABRAMOWITZ & STEGUN (1964) formula 7.1.14.

The error function itself

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}\,dt$$

can be written as

$$\frac{\sqrt{\pi}}{2}\,\mathrm{erf}(z) = z\sum_{i=0}^{\infty} d_i (z^2)^i$$

with

$$d_i = \frac{(-1)^i}{(2i+1)!} .$$

According to (4.3.3) we get

$$\sqrt{\pi}\,\frac{\mathrm{erf}(z)}{2z} = \frac{1\rvert}{\lvert 1} + \frac{z^2/3\rvert}{\lvert 1} - \frac{z^2/30\rvert}{\lvert 1} + \frac{39z^2/70\rvert}{\lvert 1} - \frac{739z^2/1638\rvert}{\lvert 1} + \ldots .$$

To get a continued fraction for the Γ-function we look at ABRAMOWITZ & STEGUN (1964), formula 6.1.40, which is

$$\ln\Gamma(z) - (z-1/2)\ln z + z - \tfrac{1}{2}\ln(2\pi) \sim$$

$$z\sum_{m=1}^{\infty}\frac{B_{2m}}{2m(2m-1)(z^2)^m}\,; \qquad \{B_i\}\text{ are the Bernoulli numbers.}$$

Using the methods described in section 4.7 we get the corresponding Stieltjes fraction

$$z\left(\frac{B_2/2\rvert}{\lvert z^2} - \frac{f_1\rvert}{\lvert 1} - \frac{e_1\rvert}{\lvert z^2} - \frac{f_2\rvert}{\lvert 1} - \ldots\right).$$

Using the QD algorithm we get (compare ABRAMOWITZ & STEGUN (1964), formula 6.1.48 and see also CHAR (1980))

$\{f_i\}$ =
{-1/30, -195/371, -29944523/19733142,
-294045 27905795295658/97692 14287853155785,
-2637081256939 77190019319929 45645578779349/
527124426791 79808019665536 49147604697542,...}

and

$\{e_i\}$ =
{-53/210, -22999/22737, -109535241009/48264275462,
-45 53770304201134 32210116914702/11 30841289236750 14537885725485,
-152537490709 05480988163889 74729859908667 53853122697839/
24274291553 10512843829739 81089021953653 73879212227720,...}.

For the confluent hypergeometric function we consider the function

$$U^{\nu,\rho}(x) = \int_0^\infty e^{-tx} t^{\nu-1} (1+t)^{-\rho} dt;$$

$$\text{Re } x > 0; \text{ Re } \nu > 0; \rho \in \mathbb{C}.$$

In terms of confluent hypergeometric functions we have (in the notation of ABRAMOWITZ & STEGUN (1964)):

$$U^{\nu,\rho}(x) = \Gamma(\nu) U(\nu, \nu+1-\rho, x).$$

By specifying ν and ρ we obtain error functions, incomplete gamma functions, etc. Taking in the above integral $\nu+1$ instead of ν, a partial integration yields

$$xU^{\nu+1,\rho}(x) = \nu U^{\nu,\rho}(x) - \rho U^{\nu+1,\rho+1}(x)$$

from which follows

$$\frac{U^{\nu+1,\rho}(x)}{U^{\nu,\rho}(x)} = \cfrac{\nu}{x + \cfrac{\rho}{U^{\nu+1,\rho}(x)/U^{\nu+1,\rho+1}(x)}} \quad .$$

Furthermore, writing in the integral

$$t^{\nu+1}(1+t)^{-\rho-1} = t^{\nu}(1+t)^{-\rho} - t^{\nu}(1+t)^{-\rho-1},$$

we obtain

$$U^{\nu+2,\rho+1}(x) = U^{\nu+1,\rho}(x) - U^{\nu+1,\rho+1}(x) \quad .$$

Combining the above relations we obtain finally

$$\frac{U^{\nu+1,\rho}(x)}{U^{\nu,\rho}(x)} = \cfrac{\nu}{\cfrac{U^{\nu+2,\rho+1}(x)}{U^{\nu+1,\rho+1}(x)}}$$

which yields a continued fraction for the quotient $U^{\nu+1,\rho}(x)/U^{\nu,\rho}(x)$. If x is positive and ρ and ν are real, then, from a certain index i we have positive numerators and denominators. If $\rho = 1$ we have

$$U^{\nu,1}(x) = \Gamma(\nu)e^{x}\Gamma(1-\nu,x),$$

where $\Gamma(a,x)$ is the incomplete gamma function.
Hence

$$\frac{U^{\nu,1}(x)}{U^{\nu-1,1}(x)} = \frac{(\nu-1)\Gamma(1-\nu,x)}{\Gamma(2-\nu,x)} \quad .$$

From the well-known relation

$$\Gamma(a+1,x) = a\Gamma(a,x) + x^{a}e^{-x}$$

we thus obtain

$$\frac{U^{\nu,1}(x)}{U^{\nu-1,1}(x)} = -\ 1 + \frac{e^{-x}x^{1-\nu}}{\Gamma(2-\nu,x)}$$

which gives, using the above continued fraction

$$\Gamma(a,x) = \frac{e^{-x}x^{a}|}{|\ x} + \frac{1-a|}{|\ 1} + \frac{1|}{|x} + \frac{2-a|}{|\ 1} + \ \dots \ .$$

The case $a = \frac{1}{2}$ gives the complementary error function, the case $a = 1-n$, $n = 0,1,\dots,$ gives exponential integrals. The expansion for $\Gamma(a,x)$ converges for all $a \in \mathbb{C}$ and for all $x \neq 0$, $|\arg(x)| < \pi$. In GAUTSCHI & SLAVIK (1978) a different approach for functions like $U^{\nu,\rho}(x)$ is used, based upon the methods described in section 4.7.

5. HYPERGEOMETRIC FUNCTIONS

It is nearly impossible to study special functions without the notion of hypergeometric functions. In this section we give a short introduction to this subject, in order to be able to describe interrelations between special functions considered in later chapters. For a more complete and more rigorous introduction the reader should consult the literature, for instance RAINVILLE (1960) (a very readable book on special functions) or LUKE (1969) (with much more information, especially on expansions which are useful for numerical computations).

The usual definition is through power series, giving the ${}_pF_q$-functions. This is done in section 5.1 (Gauss-functions ${}_2F_1$). These classes can be extended considerably by using a Mellin-Barnes contour integral; this approach is described in section 5.3. In section 5.4 we give some useful expansions, for instance, in terms of Chebyshev polynomials.

Our attitude is to be careful with general forms of special functions. From our own experience and from the good examples in the literature, we know that a basic knowledge of these functions can be very convenient. For computations, the general setting of ${}_pF_q$ and Meijer's G-function is rather useless when too many parameters are involved. Already the well-studied case of Bessel functions (belonging to the ${}_0F_1$-functions) with two complex parameters may yield serious problems for certain combinations of these parameters.

Not all interesting special functions are of hypergeometric type. A completely different class, with as prototype Mathieu's functions, is described by ARSCOTT (1981) as the Land beyond Bessel, or as "higher" special functions. They can not be defined by simple power series or integrals.

5.1. Gauss' hypergeometric function ${}_2F_1$

Here we introduce the best-known hypergeometric function ${}_2F_1(a,b;c;z)$ *by means of its power series expansion. We give the relevant properties and some relations with other special functions, for instance with orthogonal polynomials.*

The usual definition for Gauss' hypergeometric function is

$$(1.1)\qquad {}_2F_1(a,b;c;z) = \sum_{n=0}^{\infty} \frac{(a)_n(b)_n}{(c)_n n!} z^n,$$

where $|z| < 1$, $c \neq 0, -1, -2, \ldots$. In (1.1) Pochhammer's symbol is used for the shifted factorial

$$(\alpha)_0 = 1, \quad (\alpha)_n = \alpha(\alpha+1)\ldots(\alpha+n-1), \quad n \geq 1$$

or $(\alpha)_n = \Gamma(\alpha+n)/\Gamma(\alpha)$, where Γ is Euler's gamma function.

From the ratio test it follows that (1.1) has the disc $|z| < 1$ as its domain of convergence. From well-known properties of the gamma function, for instance

$$(1.2)\qquad \Gamma(\alpha+n)/\Gamma(\beta+n) \sim n^{\alpha-\beta}, \qquad n \to \infty,$$

it follows that

$$(1.3)\qquad \frac{(a)_n(b)_n}{(c)_n n!} \sim n^{a+b-c-1}, \qquad n \to \infty,$$

so long as none of a,b,c is zero or a negative integer. Hence, a sufficient condition for absolute convergence of (1.1) on $|z| = 1$ is $\mathrm{Re}(c-a-b) > 0$.

For $a = 1$, $b = c$, (1.1) reduces to $1/(1-z)$, which explains the name hypergeometric. Other examples in terms of elementary functions are

$$(1.4)\qquad \begin{aligned} {}_2F_1(a,b;b;z) &= (1-z)^{-a}, \\ {}_2F_1(1,1;2;z) &= z^{-1}\ln(1-z)^{-1}. \end{aligned}$$

A more extensive list is given in ABRAMOWITZ & STEGUN (1964, p.556).

The ${}_2F_1$-function is symmetric in a and b. When one of these parameters is a negative integer, say $a = -m$, then (1.1) is a polynomial. This follows from

$$(-m)_n = \begin{cases} (-1)^n m(m-1)\ldots(m-n+1) & n \leq m \\ 0 & n > m \end{cases}$$

or $(-m)_n = (-1)^n m!/\Gamma(m-n+1)$. Some particular cases are

	Name polynomial
${}_2F_1(-n,n;\frac{1}{2};x) = T_n(1-2x)$	Chebyshev
${}_2F_1(-n,n+1;1;x) = P_n(1-2x)$	Legendre
${}_2F_1(-n,n+2\alpha;\alpha+\frac{1}{2};x) = \frac{n!}{(2\alpha)_n} C_n^{(\alpha)}(1-2x)$	Gegenbauer
${}_2F_1(-n,n+\alpha+\beta+1;\alpha+1;x) = \frac{n!}{(\alpha+1)_n} P_n^{(\alpha,\beta)}(1-2x)$	Jacobi

An extensive theory is based on the differential equation

$$(1.5) \qquad z(1-z)y''(z)+[c-(a+b+1)z]y'(z) - aby(z) = 0,$$

resulting into the well-known transformation formulas. They express functions of argument z into combinations of functions with argument $z^{\pm 1}$, $(1-z)^{\pm 1}$, $[z/(z-1)]^{\pm 1}$, giving interesting relations for numerical computations. See ABRAMOWITZ & STEGUN (1964, p.559). The convergence of the series (1.1) is rather poor when $|z|$ is close to unity. The transformation formulas can always give a reduction to $|z| \le \frac{1}{2}$, although some combinations of parameters may yield rather complicated expressions.

The ${}_2F_1$-functions include as further special cases

Legendre functions
incomplete beta functions
elliptic integrals.

The last two cases easily follow from the integral representation

$$(1.6) \qquad {}_2F_1(a,b;c;z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a}dt,$$

which is valid when Re c > Re b > 0. The integral gives a one valued analytic function in the z-plane cut along the real axis from 1 to $+\infty$. Hence, (1.6) gives the analytic continuation of (1.1) in the case that Re c > Re b > 0. The relation between the right-hand sides of (1.1) and (1.6) is easily verified by expanding $(1-tz)^{-a}$ in a binomial series and using the integral representation for the beta function $B(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$, that is,

$$(1.7) \qquad B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt, \text{ Re } x > 0, \text{ Re } y > 0.$$

When Re(c-a-b) > 0, (1.7) gives the limiting form for (1.6)

$$(1.8) \qquad {}_2F_1(a,b;c;1) = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)}, \quad c \neq 0,-1,-2,\dots .$$

There is an extensive list of contiguous relations for ${}_2F_1$-functions, expressing ${}_2F_1(a,b;c;z)$ in terms of ${}_2F_1(a+\alpha,b+\beta;c+\gamma;z)$, where α,β,γ equal 0,1,-1 (in all possible combinations). Also, derivatives can play a role here. A simple example obtained from (1.1) is

$$\frac{d}{dz}\,{}_2F_1(a,b;c;z) = \frac{ab}{c}\,{}_2F_1(a+1,b+1;c+1;z).$$

For more examples we refer to ABRAMOWITZ & STEGUN (1964, p.557,558).

5.2. A generalization of the ${}_2F_1$

We introduce the generalization ${}_pF_q$ by means of the power series. A compact notation is used and examples are given for special functions.

We consider a generalization of Gauss' hypergeometric function by writing

$$(2.1) \qquad {}_pF_q(\alpha_p;\rho_q;z) = \sum_{k=0}^{\infty} [(\alpha_p)_k/(\rho_q)_k]\,\frac{z^k}{k!},$$

where α_p is interpreted as $\alpha_1,\dots,\alpha_p$ and $(\alpha_p)_k$ as $\Pi_{h=1}^{p}(\alpha_h)_k$; the same for ρ_q and $(\rho_q)_k$. To distinguish between nominator and denominator parameters, a notation is used of the form

$${}_pF_q\left({\alpha_p \atop \rho_q}\middle| z\right).$$

No denominator parameter ρ_h is allowed to be zero or a negative integer. If any numerator parameter α_h in (2.1) is zero or a negative integer, the series terminates.

With respect to convergence we have the following possibilities

(a) if $p \leq q$, the series converges for all finite z;

(b) if $p = q+1$, the series converges for $|z| < 1$, and diverges for $|z| > 1$;

(c) if $p > q+1$, the series diverges for $z \neq 0$.

If the series terminates the conclusions in (b) and (c) do not apply. If $p = q+1$, the series is absolutely convergent on the circle $|z| = 1$ if

$$\mathrm{Re}\left(\sum_{h=1}^{q} \rho_h - \sum_{h=1}^{p} \alpha_h\right) > 0.$$

When the series is not convergent it may have a meaning as an asymptotic expansion.

We permit p or q, or both, to be zero. Then the parameters α_h or ρ_h are absent. For example, the first of (1.4) is

$$_1F_0(a;;z) = (1-z)^{-a}$$

and

$$_0F_0(;;z) = e^z.$$

An important class of functions, with many examples as special functions of mathematical physics, is governed by the case $p = q = 1$. It gives the Kummer or Whittaker functions, which are known as degenerate or confluent hypergeometric functions. See Chapter 13 in ABRAMOWITZ & STEGUN (1964) or Chapter IV in LUKE (1969). This class includes Bessel functions, incomplete gamma functions (and the special cases the exponential integrals, sine- and cosine-integrals, error functions and Fresnel integrals), Laguerre polynomials, Hermite polynomials, Coulomb wave functions and parabolic cylinder functions.

The adjective "confluent" originates from the limiting process

$$(2.2) \qquad \lim_{b\to\infty} {}_2F_1(a,b;c;z/b).$$

The limit is $_1F_1(a;c;z)$, as follows from elementary analysis. The $_2F_1$-function with variable z/b has a differential equation (see (1.5)) with singular points $0,b,\infty$; the limiting form defines an entire function with a singularity at $z = \infty$, which is a confluence with those at b and ∞.

In LUKE (1969, Chapter VI) a lot of named special functions are expressed as $_pF_q$'s, including the examples mentioned above with $p = q = 1$.

5.3. The G-function

Only the basic ideas behind the role of the G-function are considered here. For a good understanding of the theory of hypergeometric functions it is important to know about it. Interested readers should consult the literature.

As mentioned in the previous section, the definition (2.1) is useless for the case $p > q+1$. In that case the series diverges except when $z = 0$. It may be interpreted as an asymptotic expansion, however. Which function is a natural candidate to have that expansion as an asymptotic series?

When $p = q+1$, (2.1) defines a function for $|z| < 1$; what is the analytic continuation of this function beyond $|z| = 1$?

These, and many more, questions can be answered when we introduce the G-function. It appears that representations in terms of series may be rather restrictive in defining special functions, whereas a definition in terms of a contour integral in the complex plane may be much more flexible.

To introduce the G-function let us first consider the integral

$$(3.1) \qquad I(z) := \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} z^s\Gamma(1+s)\Gamma(-s)ds, \qquad -1 < c < 0,$$

where the many-valued function z^s is defined by $z^s = \exp(s(\ln|z|+i \arg z))$, with $|\arg z| < \pi$. The product of gamma functions can be replaced by $\Gamma(1+s)\Gamma(-s) = -\pi/\sin(\pi s)$, from which information on convergence and other analytical aspects (residues, for instance) can be obtained. The contour of integration can be shifted to the right, across the poles at $s = 0,1,2,\ldots$. It easily follows that the infinite series of residues converges when $|z| < 1$ and that

$$(3.2) \qquad I(z) = \sum_{n=0}^{\infty} (-1)^n z^n = \frac{1}{1+z}, \qquad |z| < 1.$$

On the other hand, by shifting the contour to the left and picking up the residues at $s = -1,-2,-3,\ldots$, we obtain

$$(3.3) \qquad I(z) = -\sum_{n=1}^{\infty} (-1)^n z^{-n} = \frac{1}{1+z}\,, \qquad |z| > 1.$$

Hence, the contour integral (3.1) contains both series representations in

(3.2) and (3.3), the first for $|z| < 1$, the latter for $|z| > 1$. Observe that the series are examples of hypergeometric series. (In this example there is a lot of symmetry between the cases $|z| < 1$, $|z| > 1$ since we have the equation $I(z^{-1}) = 1 - I(z)$, which easily follows from (3.1) without knowledge of $I(z) = 1/(1+z)$.)

A second example is governed by

$$(3.4) \qquad F(z) = \frac{\Gamma(c)}{2\pi i\Gamma(a)\Gamma(b)} \int_L \frac{\Gamma(a+s)\Gamma(b+s)\Gamma(-s)}{\Gamma(c+s)} z^s ds$$

which contains (3.1) as a special case ($b=c, a=1$). The contour runs from $-i\infty$ to $+i\infty$ and separates the poles of $\Gamma(a+s)\Gamma(b+s)$ (at $s = -a-n$, $s = -b-m$, $n,m = 0,1,\ldots$) from those of $\Gamma(-s)$ (at $s = 0,1,\ldots$). We suppose that $|\arg z| < \pi$ and a,b,c are not equal to $0,-1,-2,\ldots$. In general, the contour cannot be a vertical line, but it meanders in order to separate the poles of the gamma functions. Shifting it to the right we obtain an infinite series of residues, which converges to

$$(3.5) \qquad F(z) = {}_2F_1(a,b;c;-z), \quad |z| < 1.$$

A shift to the left will result in two series of hypergeometric type; when $a-b$ is not equal to an integer we obtain one of the transformation formulas

$$\begin{aligned}{}_2F_1(a,b;c;-z) &= \frac{\Gamma(c)\Gamma(b-a)}{\Gamma(b)\Gamma(c-a)} z^{-a} \; {}_2F_1(a,1-c+a;1-b+a;-\tfrac{1}{z}) \\ &+ \frac{\Gamma(c)\Gamma(a-b)}{\Gamma(a)\Gamma(c-b)} z^{-b} \; {}_2F_1(b,1-c+b;1-a+b;-\tfrac{1}{z}),\end{aligned}$$

$|\arg z| < \pi$; the two residue series converge of course when $|z| > 1$. Again it follows that the contour integral (3.4) contains series expansions for $|z| < 1$ as well as for $|z| > 1$.

We could write down a similar representation for the ${}_pF_q$, just be extension of numerator and denominator parameters in (3.4). Such an integral has a meaning when $p \geq q$, for a restricted domain of $\arg z$.

The G-function includes the ${}_pF_q$ as a special case and is defined as

$$(3.7) \qquad G^{m,n}_{p,q}\left(z\middle|{}^{a_p}_{b_q}\right) = \frac{1}{2\pi i} \int_L z^s \; B(s) ds$$

$$B(s) = \frac{\Gamma(b_1-s)\ldots\Gamma(b_m-s)\Gamma(1-a_1+s)\ldots\Gamma(1-a_n+s)}{\Gamma(1-b_{m+1}+s)\ldots\Gamma(1-b_q+s)\Gamma(a_{n+1}-s)\ldots\Gamma(a_p-s)},$$

$0 \le m \le q$, $0 \le n \le p$. L separates the poles of $\Gamma(b_j - s)$ $(j = 1,\ldots,m)$ from those of $\Gamma(1-a_j+s)$ $(j = 1,\ldots,n)$; a- and b-poles should not coincide. Further information on L is found in, for instance, LUKE (1969, Vol. 1, p.144). In the same reference we find a list with named special functions in terms of the G-function (p.225). For instance, we have in compact notation as in (2.1)

$$ {}_pF_q\left({\alpha_p \atop \rho_q}\Big| z\right) = \frac{\Gamma(\rho_q)}{\Gamma(\alpha_p)}\, G^{1,p}_{p,q+1}\left(-z\Big|{1-\alpha_p \atop 0,1-\rho_q}\right) $$

for $p \le q$, $z \in \mathbb{C}$, or $p = q+1$, $|z| < 1$.

The G-function contains also functions related to the generalized hypergeometric function ${}_pF_q$. For instance, the second solution of the differential equation for ${}_1F_1(a;c;z)$, which, in general, is singular at $z = 0$, whereas ${}_1F_1$ is entire in z.

Although the definition of the G-function is quite complicated when many parameters are involved, the basic idea is rather simple and well understood via the trivial example (3.1), or via (3.4) and (3.5). Observe that (3.7) has the form of the inversion of the Mellin transform; hence the Mellin transform of the G-function is (under several conditions) a combination of gamma functions. Other integral transforms for many special functions follow also from those for the G-function. Generally speaking, representation (3.7) is a convenient starting point for manipulations with special functions of hypergeometric type. The recent monograph of MARICHEV (1983) may be very helpful for obtaining transforms of special functions.

5.4. Expansions for hypergeometric functions

We give the construction of a continued fraction for the ${}_2F_1$-functions and some Chebyshev expansions for the ${}_2F_1$-function and a confluent hypergeometric function.

The power series gives a good starting point for computing the ${}_pF_q$-functions in the neighborhood of $z = 0$. For the ${}_2F_1$-functions several transformation formulas are available in order to reduce computation to $|z| < \frac{1}{2}$. The coefficients are easily generated during computations, so no pretabulated coefficients are needed. Still there is a need for other types of expansions. In the continued fraction approach, the coefficients are, again, easily constructed; in the Chebyshev expansions a more ingenious algorithm based on recursions can be used.

5.4.1. Continued fraction for ${}_2F_1(a,1;c+1;z)$

Gauss first showed that one can construct a continued fraction for a ratio of ${}_2F_1$-s; when b=1 it results into a fraction for a single ${}_2F_1$. One can easily verify for the terms in (1.1) that

$$\frac{(a)_n(b+1)_n}{(c+1)_n n!} - \frac{(a)_n(b)_n}{(c)_n n!} = \frac{a(c-b)}{c(c+1)}\frac{(a+1)_{n-1}(b+1)_{n-1}}{(c+2)_{n-1}(n-1)!},$$

from which we conclude that

$$(4.1) \qquad {}_2F_1(a,b+1;c+1;z) - {}_2F_1(a,b;c;z) = \frac{za(c-b)}{c(c+1)}\,{}_2F_1(a+1,b+1;c+2;z).$$

This can be rewritten as

$$\frac{F(a,b+1,c+1)}{F(a,b,c)} = 1/[1-\frac{za(c-b)}{c(c+1)}\frac{F(a+1,b+1,c+2)}{F(a,b+1,c+1)}]$$

where we used an obvious short-hand notation for the ${}_2F_1$. Similarly, by an interchange of symbols (recall the symmetry in a and b)

$$\frac{F(a+1,b+1,c+2)}{F(a,b+1,c+1)} = 1/[1-\frac{z(b+1)(c+1-a)}{(c+1)(c+2)}\frac{F(a+1,b+2,c+3)}{F(a+1,b+1,c+2)}]$$

and replacing this in the former we find a relation between the ratio F(a,b+1,c+1)/F(a,b,c) and the ratio with a,b,c replaced by a+1,b+1,c+2, respectively. The case b=0 is of particular interest since in that case ${}_2F_1(a,0;c;z) = 1$. Then the set up of the continued fraction reads

$${}_2F_1(a,1;c+1;z) = \cfrac{1}{1-\cfrac{za/(c+1)}{1-\frac{z(c-a+1)}{(c+1)(c+2)}\frac{{}_2F_1(a+1,2;c+3;z)}{{}_2F_1(a+1;1;c+2;z)}}}$$

with an obvious extension to the general form. Note that for a = c = 1 we have the second of (1.4). The incomplete beta function can also be written in terms of ${}_2F_1$ with second parameter equal to unity (consider, for instance, in (1.6) the transformation $t = (z-\tau)/[z(1-\tau)]$).

5.4.2. Chebyshev expansions

We give a few examples of Chebyshev expansions of hypergeometric functions. In LUKE (1969) a considerable collection of expansions is included; in LUKE (1977) algorithms are found in order to obtain coefficients of the (Chebyshev) expansions.

Luke's general approach is to expand wide classes of hypergeometric functions in terms of other hypergeometric functions (polynomials, Bessel functions) which are rather easily computed. The coefficients again are of hypergeometric type, and the numerical problem to compute the coefficients is not always trivial. A general approach here is to use a recursion relation for the coefficients. A special algorithm (based on Miller's algorithm, see III.3) is needed, since forward recursion is not stable. An example is

$$(4.2)\qquad {}_2F_1(a,b;c;z) = \sum_{n=0}^{\infty} c_n(w)T_n^*(z/w), \qquad z/w \in [0,1],$$

where T_n^* is the shifted Chebyshev polynomial. The coefficients are

$$(4.3)\qquad c_n(w) = \frac{\varepsilon_n(a)_n(b)_n w^n}{2^{2n}(c)_n n!}\, {}_3F_2\left(\begin{matrix}a+n, b+n, \frac{1}{2}+n\\ c+n, 1+2n\end{matrix}\middle| w\right)$$

and c_n satisfies the recursion

$$(4.4)\qquad c_n = \alpha_n c_{n+1} + \beta_n c_{n+2} + \gamma_n c_{n+3},$$

where $\alpha_n, \beta_n, \gamma_n$ are given in LUKE (1977,Ch.4). The factor ε_n equals $\frac{1}{2}$ (when n=0) and 1 (when $n > 0$). Luke gives a detailed analysis on the computation of the coefficients c_n. Observe that each coefficient is more complicated than the wanted ${}_2F_1$ in (4.2). The backward recursion scheme does not use any accurate initial c_n-value, however.

A second, and more interesting example is the expansion of the Kummer U function in LUKE (1969, II, p.25). The U function is related to the ${}_1F_1$-function. It is the irregular (at z=0) solution of Kummer's equation $zy'' + (c-z)y' - ay = 0$, of which $y(z) = {}_1F_1(a;c;z)$ is a regular, entire solution (ABRAMOWITZ & STEGUN (1964, Ch.13). It includes many named functions as special cases. The expansion reads

$$(4.5)\qquad (\omega z)^a U(a;c;\omega z) = \sum_{n=0}^{\infty} c_n(z)T_n^*(1/\omega),$$

$\omega \geq 1$, and the parameter z can be used to cover a wide range of complex values, but it should be bounded away from zero. In fact, (4.5) is an expansion "around infinity", whereas (4.2) is useful near the origin. The coefficients c_n obey a recursion as in (4.4). In this case a representation in terms of the G-function is possible.

5.4.3. Representations for $|z| > 1$

It follows from (3.7) that

$$G^{m,n}_{p,q}\left(z\middle|{a_p \atop b_q}\right) = G^{n,m}_{q,p}\left(z^{-1}\middle|{1-b_q \atop 1-a_p}\right). \tag{4.6}$$

This important relation can be used to obtain representations for $|z| > 1$, for instance for the ${}_2F_1$-functions. The representation (3.6) is a special case, although some combinations of the parameters must be excluded: $a-b$ not an integer. When $a-b \in \mathbb{Z}$ a more complicated relation holds, involving logarithms of z. In the case of ${}_2F_1$-functions, convergent expansions result from (4.6). In general, i.e., for general p,q, the functional equation (4.6) may yield series which have a meaning as asymptotic expansions.

III. THE GAMMA FUNCTION AND RELATED FUNCTIONS

0. HISTORY

The gamma function is the most plausible generalization of the factorial function. Euler was confronted with this matter when an apparently simple problem was proposed to him. It was expected that n! (this notation was not yet used then), was expressible in elementary algebraic quantities. Just as the triangular numbers $T_n = 1+2+...+n$ can be expressed as $T_n = \frac{1}{2}n(n+1)$. In Euler's days, one paid much attention to these questions. First, because such a formula enables one to compute T_n or n! immediately, secondly because it gives the possibility for interpolating: $T_n = \frac{1}{2}n(n+1)$ also has a sense for non-integer values of n.

In 1729, Euler proved that for n! such a simple formula did not exist; or, there was no formula with a finite number of algebraic evaluations. At the same time he turned up with the formula

$$(0.0) \qquad n! = \int_0^1 (-\ln x)^n \, dx,$$

of which indeed the right hand is defined for real positive values of n. Nowadays, the above integral is often presented differently, and in Legendre's notation $\Gamma(n+1) = n!$, the gamma function is defined as follows

$$(0.1) \qquad \Gamma(z) = \int_0^\infty e^{-t}\, t^{z-1}\, dt, \quad \text{Re } z > 0.$$

Immediately we have the fundamental property

$$(0.2) \qquad \Gamma(z+1) = z\Gamma(z).$$

It is easily understood that in several ways the factorial function can be generalized to a function defined for positive real numbers. What makes Euler's choice such a plausible one? After the event it appeared that (0.1) frequently occurs and, in a "natural" way emerges in many problems. But this does not answer the question. However, it is possible to formulate criteria into which Euler's choice fits exactly. In this way, the gamma function is accepted and incorporated in the Bourbaki-works. There the definition is:

> the gamma function $\Gamma: \mathbb{R}^+ \to \mathbb{R}^+$ is the function f with $f(1) = 1$ and that satisfies for $x > 0$:
> - $f(x) > 0$
> - $f(x+1) = xf(x)$
> - f is logarithmic convex (i.e., ln f is convex).

For the equivalence between this definition and those of section 1 the reader is referred to BOURBAKI (1951) or ARTIN (1964).

A striking property is that the gamma function cannot satisfy a differential equation with algebraic coefficients (Hölder's result). This makes the gamma function a function of completely different type of transcendency than other special functions, such as Bessel functions, Legendre functions, etc.. While the difference equation (0.2) is so simple!

More elaborate information on the functions of this chapter can be found in LUKE (1975), WHITTAKER & WATSON (1927) (an important book for classical results and methods in analysis and special functions), ARTIN (1964) (a little monograph with emphasis on convexity properties and elementary analysis; a classic), HOCHSTADT (1971) (recommended for lessons on special functions), and NG (1975) (a survey and evaluation of software for the complex gamma function).

1. DEFINITIONS AND ANALYTICAL BEHAVIOUR

In section 1.1, *we give the Euler and Weierstrass representations of the gamma function as well as a graph of* $|\Gamma(z)|$. *In section* 1.2 *we introduce the* ψ *and polygamma functions by series representations.*

In section 1.3 *some integral representations of the beta function and its relation to the gamma function are given. In sections* 1.3 *and* 1.4 *the occurrence of the gamma function in other special functions is mentioned.*

1.1 Gamma function

Apart from the definition given in section 0 the following three definitions are usually considered.

$$(1.1) \quad \text{(Euler)} \qquad \Gamma(z) = \int_0^\infty e^{-t}\, t^{z-1}\, dt, \ \mathrm{Re}\, z > 0.$$

$$(1.2) \quad \text{(Euler)} \qquad \Gamma(z) = \lim_{n\to\infty} \frac{n!\; n^z}{z(z+1)\ldots(z+n)},\ z \neq 0,-1,-2,\ldots\ .$$

$$(1.3) \quad \text{(Weierstrass)} \quad 1/\Gamma(z) = z\, e^{\gamma z} \prod_{n=1}^{\infty} (1+z/n)\; e^{-z/n},$$

with $\gamma = 0.57721\ldots$, Euler's constant. The equivalence of these definitions is proved in HOCHSTADT (1971).

From (1.3) many other results follow. It is the most manageable definition. It readily follows from (1.3) that $1/\Gamma$ has zeros for $z = 0,-1,-2,\ldots$ and that it is nowhere singular. In an analytical sense, $1/\Gamma$ is easier to cope with than Γ itself; the latter does have singularities. This difference is reflected in numerical approximations. For $1/\Gamma$ approximations are usually more favourable (less terms in a series for obtaining a given precision). Of course (1.3) is useless for direct computations, although it is so powerful from an analytical point of view. (A simple numerical consideration learns us that for a relative accuracy of ε, about $\frac{1}{2}z^2/\varepsilon$ factors in (1.3) are needed.)

The decomposition (Prym) into incomplete gamma functions

$$\Gamma(z) = \int_0^1 e^{-t}\, t^{z-1}\, dt + \int_1^\infty e^{-t}\, t^{z-1}\, dt =$$
$$= \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\,(z+n)} + \int_1^\infty e^{-t}\, t^{z-1}\, dt$$

gives insight in analytical aspects of Γ. The last integral is an entire function of z, while the series gives information about the singularities of Γ. It follows that

$$\lim_{z\to -n} (z+n)\,\Gamma(z) = \frac{(-1)^n}{n!},$$

That is to say, Γ has in $-n$, $n = 0,1,2,\ldots$, a pole of the first order with residue $(-1)^n/n!$.

After this introductory matter the graph of Γ is easily drawn. See Figure 1. We also give the landscape of $|\Gamma(z)|$ for complex values of z. See Figure 2 (from JAHNKE & EMDE (1945)).

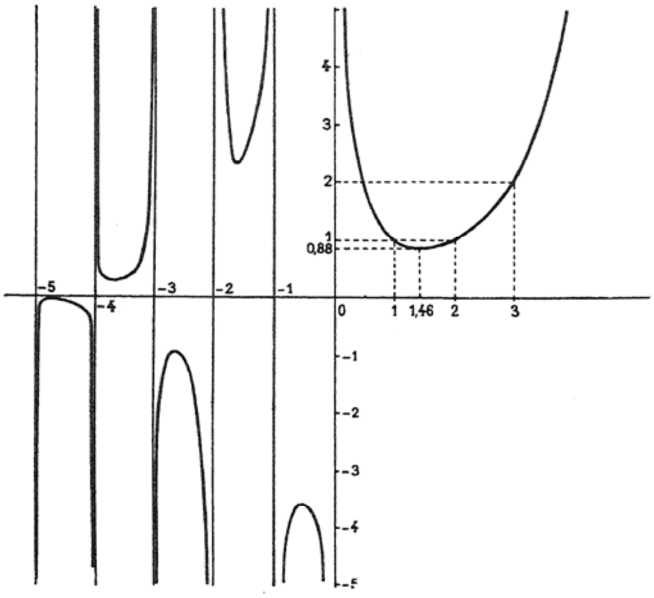


Figure 1. Graph of Γ(x), x real

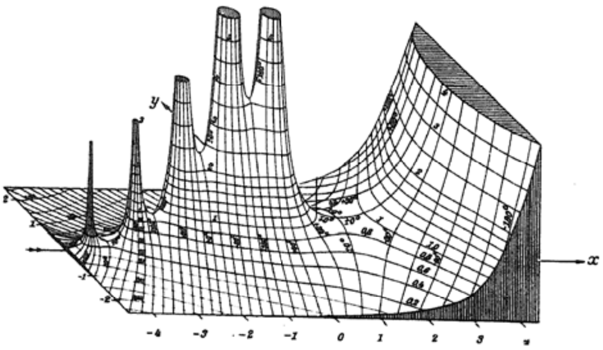


Figure 2. |Γ(z)| for complex z

An important relation for range reduction is

$$(1.4) \qquad \frac{1}{\Gamma(1+z)\,\Gamma(1-z)} = \frac{\sin \pi z}{\pi z},$$

which is called the reflection formula. It is easily proved by using (1.3), which yields at the right-hand of (1.4): $\Pi(1-z^2/n^2)$. This is connected with the factorization of the sine function.

We conclude this subsection with some integral representations which follow immediately from the above results. With the methods of function theory we obtain Hankel's formula

$$(1.5) \qquad \Gamma(z) = \frac{i}{2\sin \pi z} \int e^{-\zeta}(-\zeta)^{z-1}\, d\zeta, \quad z \notin \mathbb{Z}$$

where the contour of integration is drawn in Figure 3. By using (1.4) one obtains

$$(1.6) \qquad 1/\Gamma(z) = \frac{1}{2\pi i} \int e^{t}\, t^{-z}\, dt,$$

with contour as in Figure 4. The branch cuts of $(-\zeta)^{z-1}$ and t^{-z} in (1.5) and (1.6), run as usually from 0 to ∞ and from 0 to $-\infty$, respectively. The integral in (1.6) is valid for all $z \in \mathbb{C}$ and is very useful for analytic manipulations.

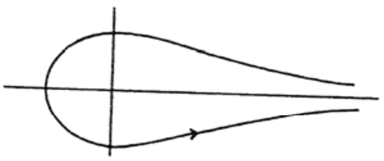


Figure 3

Contour for (1.5)

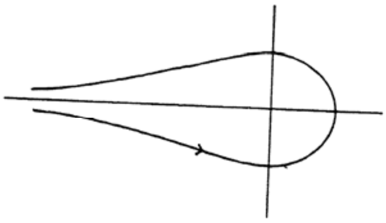


Figure 4

Contour for (1.6)

1.2 <u>Psi function and polygamma functions</u>

From (1.3) we derive

$$(1.7) \qquad \frac{d}{dz} \ln \Gamma(z+1) = -\gamma + \frac{z}{1(z+1)} + \frac{z}{2(z+2)} + \cdots \quad , \quad z \neq -1,-2,\ldots .$$

The ψ-function is defined by

$$(1.8) \qquad \psi(z) = \frac{d}{dz} \ln \Gamma(z) = \Gamma'(z)/\Gamma(z).$$

From (1.7) we obtain the well-known series representation

$$(1.9) \qquad \psi(z) = -\gamma - \frac{1}{z} + \sum_{n=1}^{\infty} \frac{z}{n(z+n)} .$$

The higher order derivatives of (1.8) are the polygamma functions $\psi^{(k)}$. Repeated differentiation of (1.9) leads to ever better converging series

$$(1.10) \qquad \psi'(z) = \sum_{n=0}^{\infty} (z+n)^{-2} , \quad \psi^{(k)}(z) = (-1)^{k+1} k! \sum_{n=0}^{\infty} (z+n)^{-k-1}.$$

The integral

$$(1.11) \qquad \psi(z) = -\gamma + \int_0^1 \frac{1-t^{z-1}}{1-t} dt, \quad \text{Re } z > -1,$$

is verified by expanding the denominator of the integrand and by comparing the result with (1.9). The series in (1.9) and (1.10) converge for all $z \in \mathbb{C}$, $z \neq 0,-1,-2,\ldots$. By using (0.2) and the corresponding recursion $\psi(z+1) = \psi(z) + 1/z$, $\psi^{(k)}$ can be examined in these exceptional points.

1.3. Beta function

The beta function is for Re $p > 0$, Re $q > 0$ defined by

$$(1.12) \qquad B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1} dt = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}.$$

We have $B(p,q) = B(q,p)$. Other forms are

$$B(p,q) = \int_0^1 (1+t)^{-p-q}(t^{p-1}+t^{q-1}) dt = \int_0^{\infty} t^{p-1}(1+t)^{-p-q} dt.$$

Integrals with circular functions expressing beta functions are

$$(1.13) \qquad \int_0^{\frac{1}{2}\pi} (\sin t)^{2p-1}(\cos t)^{2q-1} dt = \tfrac{1}{2} B(p,q),$$

$$(1.14) \qquad \int_0^{\pi} (\sin t)^{\alpha} e^{i\beta t} dt = \frac{2^{-\alpha}\pi e^{i\pi\beta/2}}{(\alpha+1)B[\frac{1}{2}(\alpha+\beta)+1, \frac{1}{2}(\alpha-\beta)+1]}, \quad \text{Re } \alpha > -1.$$

With hyperbolic functions one encounters

$$(1.15)\qquad \int_0^\infty \cosh(2yt)\ \cosh^{-2x}(t)\, dt = 2^{2x-2}\ B(x-y,x+y),\ \text{Re } x > |\text{Re } y|,$$

$$(1.16)\qquad \int_0^\infty \sinh^{\alpha}(t)\ \cosh^{-\beta}(t)\, dt = \tfrac{1}{2}B[(1+\alpha)/2,(\beta-\alpha)/2],$$

$$\text{Re}(\beta-\alpha) > 0,\ \text{Re } \alpha > -1.$$

1.4 Coulomb phase shift

The Coulomb functions $F_L(\eta,\rho)$ and $G_L(\eta,\rho)$, with ρ as argument, η as parameter and L the order (integer), are solutions of the differential equation

$$\frac{d^2w}{d\rho^2} + [1-\frac{2\eta}{\rho} - \frac{L(L+1)}{\rho^2}]\ w = 0.$$

This equation is used in the description of physical problems involving collisions and scattering of charged particles. The gamma function appears in the formulas for the asymptotic behaviour of F_L and G_L as $\rho \to \infty$:

$$F_L(\eta,\rho) \sim \sin\chi,\ G_L(\eta,\rho) \sim \cos\chi,$$

with $\chi = \rho - \eta \ln 2\rho - \tfrac{1}{2}\pi L + \sigma_L(\eta)$, containing the Coulomb phase shift

$$\sigma_L(\eta) = \arg[\Gamma(L+1+i\eta)] = \text{Im}[\ln \Gamma(L+1+i\eta)].$$

1.5 Relation with other special functions

The gamma function is frequently used in formulas for many other special functions, especially those of hypergeometric type, cf. Ch.II.5. As an example we give the series expansion of the Bessel function

$$J_\nu(z) = \frac{(z/2)^\nu}{\Gamma(\nu+1)}\ (1-\frac{z^2/4}{(\nu+1)} + \frac{(z^2/4)^2}{2!(\nu+1)(\nu+2)} + \ldots).$$

In the ALGOL 60 procedures for the computation of the Bessel functions, GAUTSCHI (1964b) used a gamma function algorithm; it was not used for summing the above series, but for an algorithm based on recursion relations.

2. FUNDAMENTAL FORMULAS

In section 2.1 we discuss asymptotic expansions for the gamma function and the psi function; rational expressions are mentioned in section 2.1.3. In section 2.2 we give Chebyshev expansions. In section 2.3 formulas for analytic continuation are given.

2.1 Expansions

2.1.1 Asymptotic expansions of the gamma function

The following representation of $\ln\Gamma(z)$ is of fundamental importance for deriving expansions for large values of $|z|$:

$$(2.1)\qquad \ln\Gamma(z) = (z-\tfrac{1}{2})\ln z - z + \tfrac{1}{2}\ln(2\pi) + S(z).$$

$S(z)$ (for large $|z|$) gives a small correction with respect to the remaining terms of the right-hand side. These terms yield the well-known Stirling formula

$$(2.2)\qquad \Gamma(z) \sim z^z e^{-z}(2\pi/z)^{\frac{1}{2}}, \quad z \to \infty.$$

$S(z)$ can be written as a Laplace integral

$$(2.3)\qquad S(z) = \int_0^\infty e^{-zt} f(t)\,dt, \quad f(t) = [(e^t-1)^{-1} + \tfrac{1}{2} - 1/t]/t.$$

For an elementary and elegant proof of this representation see LAUWERIER (1974, p.30). A different representation is

$$(2.4)\qquad S(z) = 2\cdot\int_0^\infty \frac{\arctan(t/z)}{e^{2\pi t}-1}\,dt;$$

(2.3) and (2.4) are called Binet's integrals. In both equations we assume that $\operatorname{Re} z > 0$. The proof of (2.4) (and of (2.3)) can be found in WHITTAKER & WATSON (1927).

More information on S is obtained by, for instance, expanding f of (2.3) in powers of t. This well-known technique for the asymptotic expansion of Laplace integrals is outlined in LAUWERIER (1974). Here f is an even function, and we write

$$(2.5)\qquad f(t) = \sum_{n=0}^{N-1} a_n t^{2n} + t^{2N} f_N(t), \; N = 1,2,\ldots,$$

with $a_n = B_{2n+2}/(2n+2)!$. B_m are the Bernoulli numbers, which are special cases of the Bernoulli polynomials $B_n(x)$ appearing in the expansion

$$(2.6)\qquad \frac{te^{xt}}{e^t-1} = \sum_{n=0}^{\infty} B_n(x)\,\frac{t^n}{n!}, \quad |t| < 2\pi.$$

The Bernoulli numbers B_n are given by $B_n = B_n(0)$. The first few are $B_0 = 1$, $B_1 = -\frac{1}{2}$, $B_2 = \frac{1}{6}$, $B_4 = -\frac{1}{30}$, $B_6 = \frac{1}{42}$, $(B_3 = B_5 = \ldots = 0)$. For more information see ABRAMOWITZ & STEGUN (1964, p.804) and LUKE (1975, Ch.1).

From the definition of f it is concluded that f_N is bounded on $[0,\infty)$; that is, there are assignable numbers $M_N = \sup_{t\geq 0} |f_N(t)|$. So, we can write

$$(2.7)\qquad S(z) = \sum_{n=0}^{N-1} \frac{B_{2n+2}}{(2n+1)(2n+2)z^{2n+1}} + E_N(z)$$

and for E_N we have for every z with Re z > 0 the estimation

$$|E_N(z)| \leq \left|\int_0^\infty t^{2N} f_N(t)\, e^{-zt}\, dt\right| \leq \frac{M_N(2N)!}{(\text{Re } z)^{2N+1}}.$$

This bound for E_N tells us the following: for given $\varepsilon > 0$ and $N(= 1,2,\ldots)$, we can choose z, Re z > 0, such that $|E_N(z)| < \varepsilon$. For fixed N, $|E_N(z)|$ becomes smaller according as Re z increases.

<u>REMARK</u>. It is *not* concluded that for fixed z, $|E_N(z)|$ becomes smaller according as N increases.

The numbers M_N are not easily evaluated, and so, this method does not give much information for numerical application. From (2.4) more insight is gained in this respect. The interested reader is referred to WHITTAKER & WATSON (1927, p. 251). The result is

$$(2.8)\qquad |E_N(z)| \leq \frac{B_{2N+2}K(z)}{(2N+1)(2N+2)}\,|z|^{-2N-1},$$

where

$$K(z) = \sup_{u\geq 0} |z^2/(z^2+u^2)|.$$

If $|\arg z| < \frac{1}{4}\pi$, then $K(z) = 1$. For real positive z, $E_N(z)$ is less in absolute value than the first term neglected in (2.7) and it has the same sign. These results are used in numerical algorithms. For the use of error bounds for complex z see NG (1975) and also LUKE (1975, p.7).

From (2.8) it follows that $E_N(z) = \mathcal{O}(z^{-2N-1})$ for Re z → +∞. In the terminology of asymptotic analysis we call (2.7) an asymptotic expansion. Inserting the values of the first Bernoulli numbers we arrive at the representation (Stirling's series)

$$(2.9)\qquad \ln \Gamma(z) = (z-\tfrac{1}{2}) \ln z - z + \tfrac{1}{2} \ln (2\pi) + \frac{1}{12z} - \frac{1}{360z^3} + \frac{1}{1260z^5} - \frac{1}{1680z^7} + \mathcal{O}(z^{-9}).$$

For numerical applications (2.9) is very important. The error bound in (2.8) gives a good criterion for selecting N and the range of z, especially when $|\arg z| < \pi/4$.

Stirling's series is valid, however, for $|\arg z| < \pi$, but for Re $z < 0$ its usefulness detoriates as z approaches the negative reals.

By exponentiation of (2.9) we obtain expansions for Γ or 1/Γ. The result is

$$(2.10)\qquad \begin{aligned} \Gamma(z) &\sim e^{-z}z^{z}\,(2\pi/z)^{\frac{1}{2}}\,\left(1+\frac{1}{12z}+\frac{1}{288z^2}-\frac{139}{51840z^3}+\ldots\right) \\ 1/\Gamma(z) &\sim e^{z}z^{-z}\,(2\pi/z)^{-\frac{1}{2}}\,\left(1-\frac{1}{12z}+\frac{1}{288z^2}+\frac{139}{51840z^3}+\ldots\right), \end{aligned}$$

again for $|\arg z| < \pi$, $z \to \infty$. We remark that the series in (2.10) contain powers of z^{-1}, whereas (2.9) is essentially in powers of z^{-2}; thus (2.9) is more efficient than (2.10). Of course, expansions for Γ and 1/Γ can be obtained directly from their integral representation. Numerical values of more coefficients in (2.10) are given in WRENCH (1968) and SPIRA (1971), together with more useful information on numerical aspects of the gamma function.

Writing in (2.3) for z the value z+α and expanding $f(t)\, e^{-\alpha t}$ in powers of t, we obtain (with (2.6)) the expansion

$$(2.11)\qquad \ln \Gamma(z+\alpha) = (z+\alpha-\tfrac{1}{2}) \ln z - z + \tfrac{1}{2} \ln (2\pi) + \sum_{m=1}^{N} (-1)^{m+1} \frac{B_{m+1}(\alpha)}{m(m+1)\,z^m} + \mathcal{O}(z^{-N-1}),$$

for $z \to \infty$, $|\arg z| < \pi/2$. Combining expansions of this type yields

$$(2.12) \qquad \frac{\Gamma(z+a)}{\Gamma(z+b)} = z^{a-b}\ \{1+\frac{(a-b)(a+b-1)}{2z} + \mathcal{O}(z^{-2})\}.$$

From LUKE (1975) it follows that, again, the general term in this expansion is a Bernoulli polynomial. Luke also considers an interesting modification of (2.12) due to Fields but details will not be given here. See LUKE (1975, p.11).

In applications we are often confronted with quotients of gamma functions: $\Gamma(x)/\Gamma(y)$. If x and y are both large it is recommended not to compute $\Gamma(x)$ and $\Gamma(y)$ separately, for the computer's range of the reals is (especially for this function) rather limited. It is better to use representations such as (2.12). On the other hand, it is possible to avoid overflow by writing $\Gamma(x)/\Gamma(y) = \exp(\ln \Gamma(x) - \ln \Gamma(y))$, but the subtraction may cause a loss of significant digits. Here (5.1) on p.136 may be useful.

2.1.2 Expansions for the psi and polygamma functions

By formal differentiation of (2.11) we obtain asymptotic expansions (with $m \geq 0$; $c_0 = -\ln z$, $c_m = (m-1)!/z^m (m\geq 1)$)

$$(2.13) \qquad \psi^{(m)}(z) = (-1)^{m-1}\Big[c_m + \frac{m!}{2z^{m+1}} + \sum_{k=1}^{N-1} B_{2k}\ \frac{(2k+m-1)!}{(2k)!z^{2k+m}} + \mathcal{O}(z^{-2N-m})\Big],$$

$N = 1,2,\ldots$, $z \to \infty$ in $|\arg z| < \pi$. On the other hand we have (1.9) and (1.10). Direct application of these formulas is not efficient for computations, but they may be transformed, e.g. by using the Euler-MacLaurin summation formula. But asymptotic methods based on (2.13) and range reduction (see section 2.3) may result in more efficient algorithms.

In order to demonstrate the Euler-MacLaurin method we give more details. The theory can be found in, for instance, LAUWERIER (1974) and KNOPP (1964). Suppose, we want to evaluate series of the form $\sum_{i=0}^{\infty} f(i)$, where f is a function defined for non-negative real numbers. The Euler-MacLaurin method can be used by choosing a positive integer n and by computing the partial sum $\sum_{i=0}^{n-1} f(i)$ directly. The remainder is written as follows. For $k = 1,2,3,\ldots$ we have

$$(2.14) \qquad \sum_{i=n}^{\infty} f(i) = \tfrac{1}{2} f(n) + \int_n^{\infty} f(x)\, dx - \sum_{i=1}^{k} \frac{B_{2i}}{(2i)!} f^{(2i-1)}(n) + R_k,$$

$$R_k = \frac{1}{(2k+1)!}\int_n^{\infty} f^{(2k+1)}(x)\ P_{2k+1}\ (x)\ dx,$$

where $P_{2k+1}(x)$ is the periodic continuation of the Bernoulli polynomials $B_n(x)$ with respect to $[0,1]$. That is, $P_k(x) = B_k(x)$ for $x \in [0,1]$ and $P_k(x+j) = P_k(x)$ for all integers j. For the validity of (2.14) we suppose that the first 2k+1 derivatives of f exist on $[0,\infty)$, that $f^{(j)}(\infty) = 0$, $j = 0,1,\ldots,2k+1$ and that the integrals occurring in (2.14) exist.

For the polygamma functions $\psi^{(m)}$ with series expansions (1.10) we take $f(x) = (z+x)^{-m-1}$, and we suppose that $m \geq 1$. The integral $\int_n^\infty f(x)\,dx$ is easily evaluated (this is of importance for the applicability of the method). Moreover the derivatives of f are available. Let us give the result for k = 3:

$$(2.15)\qquad \psi^{(m)}(z) = (-1)^{m+1}\, m!\, \sum_{i=0}^{n-1} (z+i)^{-m-1} + $$
$$+ (-1)^{m+1}(z+n)^{-m}\left[(m-1)!+\frac{m!}{2(z+n)}+\frac{(m+1)!}{12(z+n)^2}-\frac{(m+3)!}{720(z+n)^4}+\frac{(m+5)!}{30240(z+n)^6}\right]+$$
$$+ (-1)^{m+1} m!\, R_3.$$

By using well-known estimates for the Bernoulli polynomials (see ABRAMOWITZ & STEGUN (1964, p.805)), viz.

$$(2.16)\qquad |P_{2n+1}(x)| < \frac{2(2n+1)!}{(2\pi)^{2n+1}}\,\frac{1}{1-2^{-2n}},\quad x \geq 0,$$

we obtain a bound for R_3. For z = 1, m = 1 and n = 10 we obtain $|R_3| \leq 3.56 \times 10^{-10}$. (Observe that a check can be made by using $\psi^{(1)}(1)=\pi^2/6$.)

2.1.3 Rational approximation of ψ

LUKE (1975) gives a rational approximation of the form

$$(2.17)\qquad \psi(z) + \gamma = \frac{2(z-1)}{z}\,\frac{A_n(z)}{B_n(z)} + S_n(z),\quad \operatorname{Re} z > 0,$$

where A_n and B_n are polynomials. They satisfy a fourth order recursion relation. From estimations of $S_n(z)$ given by Luke we expect that (2.17) gives an efficient algorithm.

2.2 Chebyshev series

An expansion of ψ in terms of Chebyshev polynomials is given by Wimp (see HART c.s. (1968, section 6.6)).

$$\psi(x+a) = 2 \sum_{k=0}^{\infty}{}' c_k(a)T_k(x), \quad -1 < x \le 1, \quad a > 1.$$

with

$$c_k(a) = - \sum_{j=0}^{\infty} \frac{\{[(j+a)^2-1]^{\frac{1}{2}}-(j+a)\}^k}{[(j+a)^2-1]^{\frac{1}{2}}}, \quad k \ge 1.$$

Integration with respect to a will give a Chebyshev series for ln $\Gamma(x)$. See also in this connection NÉMETH (1967), of which the results are quoted in LUKE (1975, p.4).

2.3 Range reduction

Important relations are

$$(2.18) \qquad \Gamma(z+1) = z\Gamma(z) \qquad \text{(recursion)},$$

$$(2.19) \qquad \Gamma(\bar{z}) = \overline{\Gamma(z)} \qquad \text{(conjugation)},$$

$$(2.20) \qquad \frac{1}{\Gamma(1-z)\Gamma(1+z)} = \frac{\sin \pi z}{\pi z} \qquad \text{(reflection), see (1.4).}$$

Straightforward application of (2.20) may involve some pitfalls, which can be avoided by a proper representation of the quantities, as indicated by KUKI (1972). For example, if $z = x+iy$, $x < 0$, $y < 0$, he writes

$$\log \Gamma(z) = \log(2\pi) + \pi y - i\pi[x-\tfrac{1}{2}] - \log H(z) - \log \Gamma(1-z),$$

with

$$(2.21) \qquad H(z) = -(1+e^{2\pi y}) \tanh \pi y + e^{2\pi y}(2 \sin^2 \pi\tilde{x} + i \sin 2\pi\tilde{x}), \tilde{x} = x-[x+\tfrac{1}{2}].$$

By using (1.2), Gauss' duplication relation can be proved. It is given by

$$(2.22) \qquad \Gamma(2z) = \Gamma(z)\Gamma(z+\tfrac{1}{2})2^{2z-1} \pi^{-\frac{1}{2}},$$

with generalization

$$\Gamma(mz) = \Gamma(z)\Gamma(z+\frac{1}{m}) \Gamma(z+\frac{2}{m}) \ldots \Gamma(z+\frac{m-1}{m}) m^{mz-\frac{1}{2}}(2\pi)^{\frac{1}{2}(1-m)},$$

where $m = 2,3,4,\ldots$.

For the psi and polygamma functions analogous formulas exist (see ABRAMOWITZ & STEGUN (1964)).

3. ALGORITHMS AND IMPLEMENTATIONS

In this section we give information on available algorithms and software for the gamma function and the related functions. We discuss nearly maximum precision implementations in section 3.1 and variable precision implementations in section 3.2. Finally, known implementations are listed.

3.1. Nearly maximum precision

The computation of the logarithm of the gamma function may be done by computation of (2.7), for some strip parallel to the imaginary axis, followed by (2.18), (2.19) or (2.20) or some combination. A survey of the approaches and activities is given by NG(1975); we select the approach of KUKI(1972) as an illustration. He partioned the first quadrant of the complex plane by the curve

$$x = t(y) = \max\{.1, \min(10, 10\sqrt{2} - |y|)\} \tag{3.1}$$

where $z = x + iy$, $x,y \in \mathbb{R}$. The used algorithm for $Af(\{a_k\};z)$ reads

$$A\ln\Gamma(z) \quad , \quad \text{for } x > t(y),\ y > 0 \tag{3.2}$$

$$A\ln\Gamma(z+k) - \ln \prod_{j=0}^{k-1}(z+j) \quad , \quad \text{for } x < t(y),\ x > 0,\ y > 0 \tag{3.3}$$

$$\text{with } t(y) \le x+k < t(y)+1$$

$$\ln\pi - \ln\sin\pi z - A\ln\Gamma(1-z), \quad \text{for } x < 0,\ y < 0 \tag{3.4}$$

$$\overline{A\ln\Gamma(\bar{z})} \quad , \quad \text{for } x < 0,\ y > 0 \tag{3.5}$$

$$\text{or } x > 0,\ y < 0,$$

where $A\ln\Gamma$ is the following approximation of $\ln\Gamma$:

$$A\ln\Gamma(z) = (z-\tfrac{1}{2})\ln z - z + \tfrac{1}{2}\ln 2\pi + S_N(z) \tag{3.6}$$

with $S_N(z)$ the first series of (2.7), i.e.,

$$(3.7)\qquad S_N(z) = \sum_{k=1}^{N} B_{2k} z^{-2k+1}/[2k(2k-1)].$$

In order to make the subtraction in (3.3) harmless, Kuki considered

$$(3.8)\qquad (\mathrm{Aln}\Gamma(z+k) - k\ln(z+k)) - \ln\{ \prod_{j=0}^{k-1} (z+j)/(z+k) \}.$$

To avoid cancellations the subtraction in the first term of (3.8) is done analytically by combining kln(z+k) with other terms in (3.6). The continuous branch of the second term is chosen; because the imaginary part is between 0 and 4.7 the principle value of the logarithm is augmented by $2\pi i$ when appropriate. In (3.4) the reflection formula (2.20) is used. When the gamma function is desired, the reflection formula may be used more directly by writing

$$(3.9)\qquad \frac{1}{\Gamma(z)} = \frac{\sin\pi z}{\pi}\,\Gamma(1-z) = (\sin\pi x \cosh\pi y + i \cos\pi x \sinh\pi y)\Gamma(1-z)/\pi$$

where the sinh function, with good relative precision, is to be used, and $\Gamma(1-z)$ may be obtained from the log gamma by exponentiation. The use of the reflection formula may be minimized by using a complex sine and, for log gamma, a complex logarithm (Spira's approach, see NG(1975,p.64)). The behaviour near the poles must be considered during the computation of log H(z) (see (2.21)); if the perturbation of the argument

$$\Delta z = |z-\tilde{z}|$$

is such that

$$|H(z)| - \exp(2\pi y)2\pi\Delta z \le 0,$$

then z is considered as a singularity. Ng proposed to deliver the largest positive number representable in the machine in this case; Kuki assigned this value to the imaginary part as well.

In IMSL the implementation for the log gamma is based on the work of CODY & HILLSTRÖM(1967) and the reflection formula. Cody c.s. approximated the approximation: from the Stirling series and the recurrence relation they provided minimax approximations. The used algorithm for the log gamma, AlnΓ, reads

$$-\ell n\ x + R^{0}_{n,m}(x+1) \qquad , \quad 0 < x < .5$$

$$(x-1)R^{1}_{n,m}(x) \qquad , \quad .5 \le x < 1.5$$

$$(x-2)R^{2}_{n,m}(x) \qquad , \quad 1.5 \le x \le 4$$

$$R^{3}_{n,m}(x) \qquad , \quad 4 < x \le 12$$

$$(x-\tfrac{1}{2})\ln x - x + \tfrac{1}{2}\ln 2\pi + x^{-1} R^{4}_{n,m}(1/x^{2}), \qquad x > 12,$$

where $R^{i}_{n,m}(x) = P^{i}_{n}(x)/Q^{i}_{m}(x)$, a ratio of polynomials. The partitioning of the interval has been chosen such that for modest values of n and m the maximal errors in each subinterval are nearly the same. The reflection formula is used in the form

$$A\ln|\Gamma(x)| = \ln\pi - \ln|\sin\pi x| - \ln|\Gamma(1-x)|.$$

The computational problem for the gamma function, $A\Gamma$, is for the IMSL implementation based on HART c.s. (1968) as follows:

$$\pi/(\sin\pi x\ A\Gamma(1-x)), \qquad x < 0$$

$$A\Gamma(x+k)/\prod_{j=0}^{k-1}(x+j)\ , \qquad 0 < x < 2,\ k \in \mathrm{IN},\ 2 \le x+k < 3$$

$$R^{5}_{n,m}(x) \qquad , \quad 2 \le x \le 3$$

$$A\Gamma(x-k)\prod_{j=0}^{k-1}(x+j-k), \qquad 3 < x \le 12,\ k \in \mathrm{IN},\ 2 < x-k \le 3$$

$$\exp(A\ln\Gamma(x)) \qquad , \quad 12 < x.$$

Attention has been paid to the argument reduction of the sine.

3.2. Variable precision

CLENSHAW c.s.(1963) considered a variable precision implementation based on the Chebyshev expansion of $1/\Gamma(1+x)$ as follows. The computational problem is

(3.10) $A1/\Gamma(x)$; for $0 \le x \le 1$,

with recursion for the remaining x-values, where

$$A1/\Gamma(x) = \sum_{k=0}^{N} c_k T_k(2x-1)$$

and $N \le 14$ such that for the desired precision δ

$$|c_{N+1}| < \delta \le |c_N|.$$

The first neglected term in the Chebyshev series represents the approximation error. The poles are handled by an error jump. ANTONINO & SCHWACHHEIM(1967) published an implementation with arbitrary precision; this must be understood in the sense of nearly machine independent. On every machine they obtain nearly maximum precision. The algorithm is based on the observation that the first neglected term in the Stirling series for the log gamma majorates the approximation error for $|\arg z| \le \pi/4$ (LUCAS & TERRIL(1971)). LUCAS c.s.(1971) used a similar approach for evaluation of the gamma function for complex arguments. The implementation S14HAA/F (NAG) is based on (3.10) with fixed N and $|x| < 50$, in order to prevent overflow for a CD CYBER.

Without giving any further details we give a selection of implementations known to us.

Gamma function and log gamma function for real argument:

MGAMMA/MLGAMA	IMSL
S14ABA/F	NAG
GAMMA	CALGO 309, ANTONINO c.s(1967)
GAMMA/LOG GAMMA	NUMAL
GAMMA/LOG GAM	MSL
GAMMA/ALOGAM/DGAMMA	CERN
GAMMA	CALGO 221, GAUTSCHI(1964a)

Gamma function and log gamma function for complex argument:

CDLGAM	CALGO 421, KUKI(1972)
CGAMMA	CALGO 404, LUCAS c.s.(1971)
CGAMMA/CLOGAM	CERN , KOLBIG(1972).

Psi function:

psi	FUNPACK
CDIGAM	CERN , KOLBIG(1972)
POLYGAMMA	CALGO 349, MEDEIROS c.s.(1969).
PSIFN/DPSIFN	CALGO 610, AMOS (1983)

Ratio of complex gamma functions:

CRAGAM	CERN.

Coulomb phase shift:

COULOMB	CALGO 300, GUNN(1967).

REMARKS.

1. We have omitted the early publications in CALGO because we consider them overruled.
2. The reader may not conclude that we agree with the methods in the above implementations. It falls outside the scope of this tract to give full certifications for all algorithms.
3. The CERN algorithm for the computation of the ratio of two gamma functions is based on straightforward application of $\Gamma(x)/\Gamma(y) = \exp[\ln\Gamma(x) - \ln\Gamma(y)]$. As mentioned earlier, for large x and y we recommend to use (2.12) or modifications of this expansion (see LUKE(1975,p.11)).
4. For applications an implementation of the "tamed" function $\Gamma(z)e^{z}z^{-z}$ and of S(z) defined in (2.1) would be useful.

4. SOME ASPECTS OF ERROR ANALYSIS

In this section we consider some aspects of error analysis in connection with the concepts introduced in section II.1.

For large and intermediate values of $|z|$ the error caused by perturbation of the argument is significant in the evaluation of $\Gamma(z)$. For the relative error in z, ε_z, we have the estimate

$$(4.1) \qquad |[\Gamma(\tilde{z}) - \Gamma(z)]/\Gamma(z)| \simeq |\varepsilon_z z||\psi(z)|$$

with $\psi(z) \sim \ln z$, $z \to \infty$, $|\arg z| < \pi$. For $z = 100$, $z\psi(z) = 460.0\ldots.$ so that 2 or 3 figures may be lost. For small z, for instance in the interval [1,2], the relative error is slightly damped.

The amplification factor for the relative error, viz. $zf'(z)/f(z)$, for $\ln\Gamma(z)$ and $\psi(z)$ approaches 1 and 0, respectively, for $|z| \to \infty$. Hence, the relative error in the computations is not larger than in z (for large $|z|$). For $\ln \Gamma(z)$ the intrinsic (absolute) error is given by

$$(4.2) \qquad |\ln\Gamma(\tilde{z}) - \ln\Gamma(z)| \simeq |\tilde{z}-z|\psi(z).$$

In order to obtain an estimate of the intrinsic error, KUKI (1972) used a practical variant of (4.2) during the computation cf the logarithm of the gamma function. This estimate is composed of quantities available during the calculation.

The intrinsic error estimate $|\psi(z)|\Delta z$ is approximated for

$$\begin{array}{lll}
x > t(y),\ y > 0 & \text{as} & |\ln(z)|\Delta z \\
x < t(y),\ x > 0,\ y > 0 & \text{as} & \left|2 + \dfrac{1}{|z| - \Delta z}\right|\Delta z, \quad |z| \text{ small} \\
 & & |\ln(z+n)|\Delta z \qquad , \quad |z| \text{ large} \\
x < 0 \quad ,\ y < 0 & \text{as} & \dfrac{\exp(2\pi y)2\pi\Delta z}{|H(z)| - \exp(2\pi y)2\pi\Delta z}
\end{array}$$

with $\Delta z = |\tilde{z}-z|$, $z = x + iy$ and $H(z)$ defined by (2.21).

For real parameters some examples will be given on condition numbers as introduced in section II.1.

EXAMPLE 1. (Condition of part of the truncated Stirling series)
When computing the gamma function the Stirling series is commonly used, say on $[a,\infty]$, where a is appropriately chosen. From (2.9) we select the sum

$$S(x) = \frac{1}{12} - \frac{1}{360x^2} + \frac{1}{1260x^4} - \frac{1}{1680x^6} \qquad ,x \in [a,\infty) \tag{4.3}$$

with representations

$$\begin{aligned} P_3(w) &= \frac{1}{12} - \frac{w}{360} + \frac{w^2}{1260} - \frac{w^3}{1680} = \sum_k a_k w^k \\ &= \sum_{k=0}^{3} c_k T_k(2a^2w-1) \qquad , w \in [0,1/a^2). \end{aligned} \tag{4.4}$$

The condition numbers of (4.3) and (4.4) are equal; for $a = 10$ we obtain $\kappa \approx 1$. So we prefer the power sum representation.

EXAMPLE 2. (Condition of polynomials in rational approximation of $\ln \Gamma$)
CODY c.s. (1967) gives for $.5 \leq x \leq 1.5$, among others, the approximation

$$\ln \Gamma(x) \approx (x-1)\{(2.02x^2-2.74x-2.61)/(x^2+3.97x-.80)\}. \tag{4.5}$$

Representation of the numerator as $\sum_{k=0}^{2} b_k T_k(2(x-1))$ yields a slightly better conditioned representation; the denominator is better conditioned as a power sum.

EXAMPLE 3.
KUKI (1972) estimated for log gamma the rounding error by the value of the dominant term of the Stirling series, to be multiplied by a factor because of neglected smaller contributions, as

$$|(z-\tfrac{1}{2})\ln(z+k)|\ \varepsilon \qquad \text{in case of (3.2)} \tag{4.6}$$

$$(3.15) + |\mathrm{Re}\ \ln \prod_{j=0}^{k-1} (z+j)/(z+k)|\ \varepsilon \qquad \text{in case of (3.3)} \tag{4.7}$$

$$\{|(\ln 2\pi + \pi y) - i\pi[x-\tfrac{1}{2}]| + |\mathrm{Re}(\ln H(z))|\}\varepsilon \qquad \text{in case of (3.4).} \tag{4.8}$$

To (4.8) the effect of evaluation of $\ln \Gamma(1-z)$ must be added by appropriate use of either (4.6) or (4.7); ε is of the order of machine accuracy. We can understand this, because an error bound of a sum is proportional to the sum of the moduli of the terms, as is well known. In our approach we consider each term of the Stirling series as a parameter a_i; the condition number κ represents the sum of the moduli of the terms.

5. TABULATED COEFFICIENTS

In this section we summarize approximations with published coefficients. For more information see LUKE (1975,p.21).

LUKE (1975)

$1/\Gamma(z+1)$	$\sum a_n z^n$	a_n	20 d	$\lvert z\rvert < \infty$
$\Gamma(z+3)$	$\sum a_n z^n$	a_n	20 d	$\lvert z\rvert < 3$
$\Gamma(x+1)$	$\sum a_n T_n^*(x)$	a_n	20 d	$0 \le x \le 1$
$1/\Gamma(x+1)$	$\sum a_n T_n^*(x)$	a_n	20 d	$0 \le x \le 1$
$\Gamma(x+3)$	$\sum a_n T_n^*(x)$	a_n	20 d	$0 \le x \le 1$
$\ln\,\Gamma(x+3)$	$\sum a_n T_n^*(x)$	a_n	20 d	$0 \le x \le 1$
$\psi^{(m)}(x+3)$	$\sum a_n^{(m)} T^*(x)$	$a_n^{(m)}$	20 d	$0 \le x \le 1$
$(m=0,1,\ldots,6)$				
$S(x)$ (see (2.7))	$\sum a_n T_{2n}(1/x)$	a_n	15 d	$x \ge 1$

HART c.s.(1968)

$\Gamma(x)$	rational approximations up to 22 d on [2,3]
$S(x)$	rational approximations up to 22 d on [8,1000] and [12,1000]

CODY c.s.(1967)

$\ln\,\Gamma(x)$	rational approximations up to 22 d on [0,12]

CODY c.s.(1970)

$\sigma_0(\eta)$ (see(1.4))	rational approximations up to 22 d on $(-\infty,\infty)$.

REMARKS.

1. The approximations for Γ in HART c.s.(1968) are not on [0,1] (as stated there) but on [2,3].

2. T_n and T_n^* denote the Chebyshev polynomials of the first kind usually considered on [-1,1] and [0,1]; $T_n^*(x) = T_n(2x-1)$.

3. In CODY c.s.(1967) ln Γ(x) is represented as

$$(x-1) * \text{rational function on } [\tfrac{1}{2},1\tfrac{1}{2}]$$

$$(x-2) * \text{rational function on } [1\tfrac{1}{2},4]$$

in order to preserve accuracy near the zeros 1 and 2. We favour the approximation

$$x * \text{rational function on } [-\tfrac{1}{2},\tfrac{1}{2}]$$

for ln Γ(1+x); the computation near x = 2 is reduced to the problem near 1 by

$$\ln \Gamma(2+y) = \ln \Gamma(1+y) + \ln(1+y)$$

with y = x - 2. An algorithm for accurate avaluation of ln(1+x) for small x is needed; no known library provides this function (see, however, KAHAN (1983)). An efficient algorithm for ln(1+x) may be based on the expansion

$$\ln(1+x) = 4 \sum_{k=1}^{\infty} \frac{p^{2k+1}}{2k+1} T_{2k+1} \left(\frac{1+p^2}{2p} \frac{x}{2+x}\right), \tag{5.1}$$

$$0 < p < 1, \frac{-4p}{(p+1)^2} < x < \frac{4p}{(p-1)^2}$$

(cf. LYUSTERNIK c.s.(1965)). From some analysis it follows that we can take p = 1/7, yielding the x-interval [-7/16,7/9] for safe evaluation of ln(1+x). In order to obtain relative accuracy near x = 0, the odd Chebyshev series should be evaluated, for instance, by using Clenshaw's algorithm given in CLENSHAW (1962).

6. TESTING

When testing one can think of verification of the coding and an accurate performance profile. In both cases one needs:

known values (tables, previous or multiple precision programs), mathematical relationships.

We agree with NG(1975) to use for testing the duplication formula (2.22), because the algorithms do not use it; we do not agree with HART c.s. (1968) because the recurrence relation is generally used in the algorithms. Arguments may be selected in different ways; arguments near singularities or other difficult values must be incorporated in the test set.

Known values of the gamma and related functions are published by LUKE(1970) and ABRAMOWITZ & STEGUN(1964). The latter contains references to published tables; in this connection see also FLETCHER c.s.(1962). For automatic table comparison NG(1975) constructed a reference subprogram which computed the complex gamma function in extended precision using a package of subroutines in 70-bit (about 21 decimal) arithmetic, composed by Lawson c.s. of JPL. SCHONFELDER(1976) used a package MLARITHA (ALGOL 68) to produce multi-length function values with which the multi-machine library routines are compared. Background information about testing of functions is given in NEWBERY & LEIGH(1971) and CODY(1969). The NATS approach is discussed in CODY(1975a); the NAG approach in SCHONFELDER(1976). Further we mention CODY(1973).

7. APPLICATIONS

In this section we mention some analytical applications of the functions of this chapter.

7.1. Summation of rational series by means of polygamma functions

An infinite series whose general term is a rational function in the index may always be reduced to a finite series of psi and polygamma functions.

EXAMPLE. (ABRAMOWITZ & STEGUN (1964,p.265))

$$S = \sum_{n=1}^{\infty} \frac{1}{(n+1)\ (2n+1)\ (4n+1)} = \sum_{n=1}^{\infty} u_n,$$

with

$$u_n = \frac{1/3}{n+1} - \frac{1}{n+1/2} + \frac{2/3}{n+1/4} =$$

$$= \frac{1}{3}\ \left(\frac{1}{n+1} - \frac{1}{n}\right) - \left(\frac{1}{n+1/2} - \frac{1}{n}\right) + \frac{2}{3}\ \left(\frac{1}{n+1/4} - \frac{1}{n}\right).$$

Application of (1.9) yields

$$S = -\frac{1}{3}\psi(2) + \psi(1\tfrac{1}{2}) - \frac{2}{3}\psi(1\tfrac{1}{4}).$$

For alternating series we can use the relation

$$\sum_{n=0}^{\infty} \frac{(-1)^n}{n+z} = \tfrac{1}{2}\psi(\frac{1+z}{2}) - \tfrac{1}{2}\psi(\frac{z}{2}).$$

7.2. Substitution of factorials by their integral representations

Sometimes it is useful to replace in series factorials or expressions of gamma functions by their integral representations.

EXAMPLE.

For $|x| < 4$ let

$$f(x) = \sum_{n=0}^{\infty} \frac{(n!)^2}{(2n+1)!} x^n .$$

Then by use of (1.12) we obtain

$$f(x) = \sum_{n=0}^{\infty} x^n \int_0^1 \{t(1-t)\}^n dt = \int_0^1 \frac{dt}{1 - xt(1-t)} =$$

$$= \frac{4}{\sqrt{4x-x^2}} \arctan(\frac{x}{\sqrt{4x-x^2}}) = \frac{4}{\sqrt{4x-x^2}} \arcsin(\sqrt{x/4}).$$

This result could also have been obtained by use of hypergeometric functions.

In order to facilitate the use of this technique we enumerate the integral representations of some expressions of factorials; for more relations see DINGLE(1973). For notational convenience we use $n!, (n-\tfrac{1}{2})!, \ldots$ instead of $\Gamma(n+1)$, $\Gamma(n+\tfrac{1}{2}), \ldots$.

$n!$	$\int_0^\infty e^{-t} t^n dt$
$1/n!$	$1/(2\pi i) \int e^t t^{-n-1} dt$ (see(1.6))
$(\tfrac{1}{2}n)!$	$2\int_0^\infty t^{n+1} e^{-t^2} dt$
$n!/(\tfrac{1}{2}n)!$	$2\pi^{-\frac{1}{2}}\int_0^\infty (2t)^n e^{-t^2} dt$
$(n-\tfrac{1}{2})!/n!$	$2\pi^{-\frac{1}{2}}\int_0^{\frac{1}{2}\pi} \sin^{2n} t dt = 2\pi^{-\frac{1}{2}}\int_0^{\frac{1}{2}\pi} \cos^{2n} t\, dt$

$n!/(n-m)!$	$(\partial/\partial t)^m t^n\big\|_{t=1}$
$(n-\frac{1}{2})!/n!n$	$4\pi^{-\frac{1}{2}}\int_0^1 (1-t^2)^{n-1} t \arcsin t\, dt$
$(n+\alpha)!(n+\beta)!$	$4\int_0^\infty t^{2n+\alpha+\beta+1} K_{\alpha-\beta}(2t)\, dt$ (Bessel function)
$(n+\alpha)!/(n+\beta)!$	$\frac{1}{(\beta-\alpha-1)!}\int_0^1 t^{n+\alpha}(1-t)^{\beta-\alpha-1} dt =$
	$\frac{2^\alpha(\alpha-\beta)!}{\pi}\int_{-\infty}^\infty (\frac{2}{1+it})^n \frac{dt}{(1+it)^{\beta+1}(1-it)^{\alpha-\beta+1}}$
$[(\frac{1}{2}n-\frac{1}{2})!]^2/n!$	$2\int_0^{\frac{1}{2}\pi} (\frac{1}{2}\sin t)^n dt = 2\int_{-\infty}^\infty \frac{dt}{(2\cosh t)^{n+1}}$
$(n!)^2/(2n+1)!$	$\int_0^1 [t(1-t)]^n dt$
$n!/[(\frac{1}{2}n)!]^2$	$2\pi^{-1}\int_0^{\frac{1}{2}\pi} (2\cos t)^n dt$
$(2n)!/(n!)^2$	$2\pi^{-1}\int_1^\infty (4/t^2)^n \frac{dt}{t(t^2-1)^{\frac{1}{2}}}$

REMARK.

Observe that a factorial form is transformed into a power form; the validity of interchanging summation and integration must be verified.

7.3. Laplace transforms as psi functions

Laplace transforms of the functions

$$\frac{\cosh\beta t}{\cosh\gamma t}, \frac{\sinh\beta t}{\sinh\gamma t}, \frac{\sinh\beta t}{\cosh\gamma t}$$

are representable in terms of psi functions; see OBERHETTINGER & BADII (1973) for an extensive table of Laplace transforms.

IV. EXPONENTIAL INTEGRALS AND RELATED FUNCTIONS

In section 1 we give the definitions and some relations between the functions of this chapter. In section 2 attention is paid to expansions of these functions. Taylor and asymptotic expansions are derived in section 2.1. Chebyshev expansions and continued fractions are mentioned in 2.2 and 2.3, with reference to earlier given results for hypergeometric functions.

1. DEFINITIONS and ANALYTICAL BEHAVIOUR

Many results for the functions in this chapter follow from the more general hypergeometric functions, of which some results are given in II.5. Especially, results for Chebyshev expansions follow easily from the expansions of confluent hypergeometric functions. For a first introduction, however, some specific properties and results of the exponential integrals are easier understood by considering special cases instead of the wider class of hypergeometric functions. At the end of this section we give the relations with these functions.

1.1. The exponential integrals

The function we start with is the well-known exponential integral

$$(1.1)\qquad E_1(x) = \int_x^\infty \frac{e^{-t}}{t}\,dt,$$

which we consider temporarily for $x > 0$. We cannot express it in a finite number of elementary functions. For $x = 0$ it is not defined and we first give a representation from which the behaviour near $x = 0$ is easily understood.

Let us consider the auxiliary function

$$f_\nu(x) = \int_x^\infty \frac{e^{-t}}{t^\nu}\, dt, \qquad x > 0,\ \nu \le 1,$$

which for $\nu = 1$ coincides with E_1. If $\nu < 1$ we can write

$$f_\nu(x) = \int_0^\infty e^{-t}t^{-\nu}\, dt - \int_0^x e^{-t}t^{-\nu}\, dt$$

$$= \Gamma(1-\nu) - \frac{1}{1-\nu}\, x^{1-\nu} + \int_0^x \frac{1-e^{-t}}{t^\nu}\, dt.$$

If we now try to substitute $\nu = 1$ we must carry out a limiting process. The integral is well-defined for $\nu = 1$, but,however, the remaining terms are not. By writing

$$g(\nu) = \Gamma(1-\nu) - \frac{1}{1-\nu}\, x^{1-\nu} = \frac{\Gamma(2-\nu)-x^{1-\nu}}{1-\nu}, \qquad \nu < 1$$

and applying l'Hôpital's rule we obtain

$$\lim_{\nu\to 1} g(\nu) = \Gamma'(1) - \ln x = -\gamma - \ln x$$

(see (1.7) of Chapter III). Hence, it follows that

$$E_1(x) = -\gamma - \ln x + \int_0^x \frac{1-e^{-t}}{t}\, dt, \qquad x > 0. \tag{1.2}$$

This formula enables us to consider E_1 for complex values of its argument. It appears that the singularity of E_1 at 0 is described by the logarithm, which is a many-valued function. The integral in (1.2) represents an entire function of x. Hence, E_1 is a many-valued function of which the principal branch can be defined by

$$E_1(z) = -\gamma - \ln z + \int_0^z \frac{1-e^{-t}}{t}\, dt, \qquad z \ne 0,\ |\arg z| < \pi, \tag{1.3}$$

where the logarithm has its principal branch (real for positive z). The analytic continuation for other values of the phase of z is given by

$$E_1(xe^{\pm\pi i}) = -\gamma - \ln x \mp \pi i + \int_0^{-x} \frac{1-e^{-t}}{t}\, dt, \qquad x > 0, \tag{1.4}$$

$$E_1(ze^{2n\pi i}) = E_1(z) - 2n\pi i, \qquad n = \pm 1, \pm 2, \ldots,\ z \ne 0.$$

<u>REMARK</u>. Using Cauchy's theorem and (1.1) for complex values of z, viz.

$$(1.5) \qquad E_1(z) = \int_z^\infty \frac{e^{-t}}{t}\,dt, \qquad z \neq 0,\ |\arg z| < \pi$$

the relations in (1.4) can also be understood. (The path in (1.5) should avoid the negative reals and the origin.) Increasing the phase of z in (1.5) beyond the range $(-\pi,\pi)$ gives an integral of the type (1.5) plus an integral over a closed circuit around $t = 0$, which can be evaluated by computing the residue at $t = 0$.

The following exponential integral is also used:

$$(1.6) \qquad \mathrm{Ei}(x) = -\fint_{-x}^{\infty} \frac{e^{-t}}{t}\,dt = \fint_{-\infty}^{x} \frac{e^{t}}{t}\,dt, \qquad x \in \mathbb{R},\ x \neq 0,$$

where the symbol $\fint$ is used to mean the Cauchy principal value of the integral, e.g.,

$$\mathrm{Ei}(x) = \lim_{\varepsilon \downarrow 0} \Big\{ \int_{-\infty}^{-\varepsilon} \frac{e^{t}}{t}\,dt + \int_{\varepsilon}^{x} \frac{e^{t}}{t}\,dt \Big\}, \qquad x > 0.$$

(If $x < 0$ then the integrals in (1.6) need not to be interpreted as principal value integrals). Ei is real for real x and it is usually not considered for complex values of its argument.

From the first integral in (1.6) it easily follows that

$$(1.7) \qquad \mathrm{Ei}(-x) = -E_1(x), \qquad x > 0.$$

For negative x this relation does not hold, as Ei is real for $x < 0$ whereas E_1 is not (this follows from the first of (1.4)). For $x > 0$ we have from (1.6)

$$\mathrm{Ei}(x) = -\fint_{-x}^{x} \frac{e^{-t}}{t}\,dt - E_1(x) = -\int_{-x}^{x} \frac{e^{-t}-1}{t}\,dt - E_1(x) =$$

$$= \int_0^x \frac{1-e^{-t}}{t}\,dt - E_1(x) - \int_0^{-x} \frac{1-e^{-t}}{t}\,dt =$$

$$= \gamma + \ln x - \gamma - \ln x - E_1(xe^{\pm\pi i}) \mp \pi i,$$

where we use $\fint_{-x}^{x} \frac{dt}{t} = 0$, (1.2) and the first of (1.4). Hence, for $x > 0$ we have

$$(1.8)\qquad \mathrm{Ei}(x) = -E_1(xe^{\pm\pi i}) \mp \pi i$$

or

$$(1.9)\qquad \mathrm{Ei}(x) = -\frac{1}{2}[E_1(xe^{\pi i}) + E_1(xe^{-\pi i})],$$

which are the modifications of (1.7) for $x < 0$. Combining the above results, we obtain

$$(1.10)\qquad \mathrm{Ei}(x) = \gamma + \ln|x| + \int_0^x \frac{e^t-1}{t}\,dt, \qquad x \in \mathbb{R},\ x \neq 0.$$

In Figure 1 the graphs of E_1 and Ei are shown.

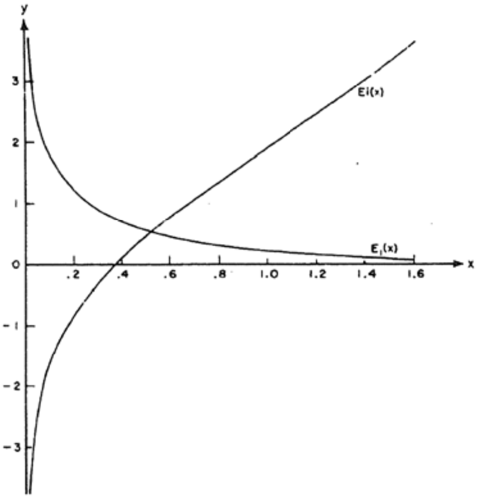


Figure 1. Graphs of Ei(x) and $E_1(x)$

Generalizations of E_1 are defined by

$$(1.11)\qquad E_\nu(z) = \int_1^\infty \frac{e^{-tz}}{t^\nu}\,dt, \qquad \mathrm{Re}\ z > 0,$$

where ν may be any complex number. Generally one encounters E_ν for integer values of ν, especially $\nu = n = 1,2,\ldots$. For arbitrarily ν-values E_ν is rather considered as an incomplete gamma function. E_n has, just as E_1, a logarithmic singularity at $z = 0$. There is also a branch point at infinity.

One may think it peculiar, this definition of E_ν. From (1.5) one should expect $\int_z^\infty t^{-\nu}e^{-t}dt$. The present definition, however, defines E_n as a repeated integral of E_1, as follows from $E_n'(z) = -E_{n-1}(z)$. Hence

$$E_n(z) = \int_z^\infty E_{n-1}(t)dt = \int_z^\infty \ldots \int_t^\infty \frac{e^{-t}}{t}\,(dt)^n,$$

from which follows (the proof is left to the reader)

$$(1.12) \qquad E_n(z) = \frac{e^{-z}}{(n-1)!}\int_0^\infty \frac{e^{-t}t^{n-1}}{t+z}\,dt, \qquad n = 1,2,\ldots\,.$$

In this formula we can take $|\arg z| < \pi$.

By partial integration in (1.11), or writing $t^{n-1} = t^{n-2}[(t+z)-z]$ in (1.12) we obtain the recursion

$$(1.13) \qquad nE_{n+1}(z) = e^{-z} - zE_n(z), \qquad n = 1,2,\ldots\,.$$

For numerical computations this relation is very important. See GAUTSCHI (1961a) and GAUTSCHI (1973). For stability aspects of this recursion see also II.3. A variant recursion is given by ACTON (1974).

The logarithmic integral li(x) is defined by

$$(1.14) \qquad \mathrm{li}(x) = \int_0^x \frac{dt}{\ln t} = \mathrm{Ei}(\ln x), \qquad x > 0.$$

The functions $\alpha_n(z)$, defined by

$$\alpha_n(z) = \int_1^\infty t^n e^{-zt}dt, \qquad \mathrm{Re}\, z > 0,\ n = 0,1,2,\ldots,$$

are special cases of $E_\nu(z)$. We have the recursion

$$z\alpha_n(z) = e^{-z} + n\alpha_{n-1}(z), \qquad n = 1,2,3,\ldots\,.$$

For stability aspects see again II.3. For $z \in \mathbb{R}^+$ this is a positive recursion and therefore stable.

1.2. The sine and cosine integrals

If we consider (1.3) with z replaced by $ze^{\frac{1}{2}\pi i}$ and if we separate, for real z, the real and imaginary parts we obtain functions connected with the sine and cosine integrals:

$$E_1(ze^{\frac{1}{2}\pi i}) = -\gamma - \frac{1}{2}\pi i - \ln z + \int_0^{iz} \frac{1-e^{-t}}{t}\,dt =$$

$$= -\gamma - \ln z + \int_0^1 \frac{1-\cos zt}{t}\,dt + i[-\frac{1}{2}\pi + \int_0^1 \frac{\sin zt}{t}\,dt].$$

The definitions for the sine and cosine integrals are

$$\mathrm{Si}(z) = \int_0^z \frac{\sin t}{t}\,dt$$

$$(1.15) \qquad \mathrm{Ci}(z) = \gamma + \ln z + \int_0^z \frac{\cos t-1}{t}\,dt, \qquad |\arg z| < \pi.$$

On the other hand, we have using (1.5)

$$(1.16) \qquad E_1(iz) = \int_{iz}^{\infty} \frac{e^{-t}}{t}\,dt = \int_1^{\infty} \frac{e^{-izt}}{t}\,dt$$

$$= \int_0^{\infty} \frac{\cos(z+t)}{z+t}\,dt - i\int_0^{\infty} \frac{\sin(z+t)}{z+t}\,dt.$$

Hence, combining the above results we obtain

$$(1.17) \qquad \begin{aligned} \mathrm{Si}(z) &= \frac{1}{2}\pi - f(z)\cos z - g(z)\sin z \\ \mathrm{Ci}(z) &= \qquad\quad f(z)\sin z - g(z)\cos z, \end{aligned}$$

where,

$$(1.18) \qquad \begin{aligned} f(z) &= \int_0^{\infty} \frac{\sin t}{t+z}\,dt \\ g(z) &= \int_0^{\infty} \frac{\cos t}{t+z}\,dt \end{aligned} \qquad z \neq 0,\ |\arg z| < \pi.$$

REMARK. In the second integral of (1.16) we integrate from 1 to + ∞. The integral exists for Re iz > 0. In the first integral we integrate from iz to + ∞, avoiding the non-positive real t-values. By using the principle of analytic continuation the restriction Re iz > 0 may be dropped; we proceed with the z-values indicated in (1.18).

In Figure 2 the graphs of Si and Ci are shown.

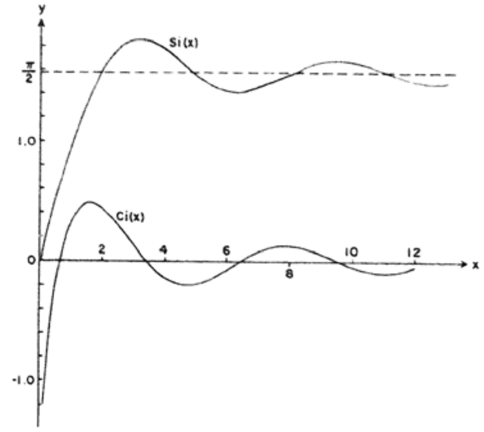


Figure 2. Graphs of Si(x) and Ci(x)

For large $|z|$, the functions in (1.18) are slowly varying. The oscillatory or exponential behaviour of Si and Ci is fully described by the behaviour of the circular functions in (1.17). We give another representation of f and g.

Writing

$$(1.19)\qquad g(z) + if(z) = \int_0^\infty \frac{e^{it}}{t+z}\,dt$$

and integrating $\int_{L_R} \frac{e^{i\tau}}{\tau+z}\,d\tau$ around the contour $L_R = A_R \cup B_R \cup C_R$, where

$$A_R = \{t \mid 0 \le t \le R\},\; B_R = \{it \mid 0 \le t \le R\},$$

$$C_R = \{\tau \mid |\tau| = R,\; 0 \le \arg\tau \le \frac{1}{2}\pi\}$$

for positive R, and letting $R \to \infty$, we obtain for Re $z > 0$

$$g(z) + if(z) = i\int_0^\infty \frac{e^{-t}}{it+z}\,dt = \int_0^\infty e^{-zt}\,\frac{t+i}{t^2+1}\,dt.$$

Hence, we have by writing $f(z) = f_1(x,y) + if_2(x,y)$, $g(z) = g_1(x,y) + ig_2(x,y)$, $z = x + iy$ and by separating the real and imaginary parts in the last integral

$$g_1(x,y) - f_2(x,y) = \int_0^\infty \frac{e^{-xt}}{t^2+1}(t \cos yt + \sin yt)dt$$

$$g_2(x,y) + f_1(x,y) = \int_0^\infty \frac{e^{-xt}}{t^2+1}(\cos yt - t \sin yt)dt.$$

Since g_1 and f_1 are even functions of y, whereas g_2 and f_2 are odd (use (1.18) in order to verify this) we can solve the above equations for f_1, f_2, g_1, g_2 and we obtain

$$(1.20) \qquad f(z) = \int_0^\infty \frac{e^{-zt}}{t^2+1}dt, \quad g(z) = \int_0^\infty \frac{te^{-zt}}{t^2+1}dt, \qquad \operatorname{Re} z > 0.$$

Remark that if $\operatorname{Re} z > 0$ the above integrals exist and the point $\tau = -z$ lies outside the contour L_R.

Apart from the function Si defined in (1.15) the function si, given by

$$(1.21) \qquad \operatorname{si}(z) = \operatorname{Si}(z) - \frac{1}{2}\pi = -\int_z^\infty \frac{\sin t}{t}dt$$

is used. Furthermore we have the representation

$$(1.22) \qquad \operatorname{Ci}(z) = -\int_z^\infty \frac{\cos t}{t}dt, \qquad |\arg z| < \pi,$$

where t avoids the non-positive reals.

Finally, we mention the "hyperbolic analogues" of (1.15)

$$(1.23) \qquad \begin{aligned} \operatorname{Shi}(z) &= \int_0^z \frac{\sinh t}{t}dt \\ \operatorname{Chi}(z) &= \gamma + \ln z + \int_0^z \frac{\cosh t - 1}{t} \qquad |\arg z| < \pi. \end{aligned}$$

1.3. Relations with hypergeometric functions

The confluent hypergeometric function U (ABRAMOWITZ & STEGUN (1964, Ch. 13)) can be used for these functions. We have

$$(1.24)\qquad \begin{cases} E_1(z) = e^{-z}U(1,1,z) \\ E_1(z) + \gamma + \ln z = z\ {}_2F_2(1,1;2,2;z), \\ E_n(z) = e^{-z}z^{n-1}U(n,n,z) = e^{-z}U(1,2-n,z). \end{cases}$$

In terms of incomplete gamma functions we have

$$(1.25)\qquad E_\nu(z) = z^{\nu-1}\Gamma(1-\nu,z), \qquad \nu \in \mathbb{C}.$$

The functions f and g of section 1.2 follow from

$$(1.26)\qquad U(1,1,ze^{\mp\frac{1}{2}\pi i}) = g(z) \pm if(z).$$

2. FUNDAMENTAL FORMULAS

2.1. Expansions based on Taylor series and asymptotic series

2.1.1. Taylor expansions

By expanding the exponential function in the integrals of (1.3) and (1.10) we obtain the representations

$$(2.1)\qquad \begin{aligned} E_1(z) &= -\gamma - \ln z - \sum_{n=1}^{\infty} \frac{(-z)^n}{n\,n!}, \qquad z \neq 0,\ |\arg z| < \pi \\ Ei(x) &= \gamma + \ln|x| + \sum_{n=1}^{\infty} \frac{x^n}{n\,n!}, \qquad x \neq 0,\ x \in \mathbb{R}, \end{aligned}$$

in which the series converge for all finite values of z and x. Similar expansions can be obtained for the functions Si, Ci, Shi, Chi.

With induction with respect to n we obtain, using (1.13), for n = 1,2,...

$$(2.2)\qquad E_n(z) = \frac{(-z)^{n-1}}{(n-1)!}[-\ln z + \psi(n)] - \sum_{\substack{m=0 \\ m\neq n-1}}^{\infty} \frac{(-z)^m}{(m-n+1)m!},$$

where $z \neq 0$, $|\arg z| < \pi$. For $\psi(n)$ see III.1.9.

As remarked earlier these expansions converge for all finite values of the argument x or z. However, the applicability for numerical purposes is rather limited. This will become clear when we have considered the behaviour of E_1 and Ei for large values of their argument.

2.1.2. Asymptotic expansions

By repeatedly using (1.13) we obtain

$$E_1(z) = \frac{e^{-z}}{z}\,[1 - e^{z}E_2(z)] = \frac{e^{-z}}{z}\,[1 - \frac{1}{z} + \frac{2}{z}\,e^{z}E_3(z)],$$

and so for $n = 0,1,2,\ldots$

$$E_1(z) = \frac{e^{-z}}{z}\,[1 - \frac{1}{z} + \frac{2!}{z^2} - \frac{3!}{z^3} + \ldots + (-1)^n\,\frac{n!}{z^n} + R_n(z)] \tag{2.3}$$

where

$$\begin{aligned} R_n(z) &= (-1)^{n+1}(n+1)!\; z^{-n}e^{z}E_{n+2}(z) = \\ &= (-1)^{n+1}(n+1)!\; z^{-n}e^{z}\int_1^\infty \frac{e^{-zt}}{t^{n+2}}\,dt. \end{aligned} \tag{2.4}$$

For real $z = x$ we have

$$x^{-n}e^{x}E_{n+2}(x) = x^{-n}e^{x}\int_1^\infty t^{-n-2}e^{-xt}dt \le x^{-n}e^{x}\int_1^\infty e^{-xt}dt = x^{-n-1}.$$

Hence,

$$R_n(x) = (-1)^{n+1}(n+1)!x^{-n-1}\theta_n(x), \qquad 0 \le \theta_n(x) \le 1, \tag{2.5}$$

which says that the remainder in (2.3) is less in absolute value than the first term neglected and has the same sign.

For complex values of $z = x + iy$ it easily follows that $R_n(z) = = \mathcal{O}(R_n(x)) = \mathcal{O}(x^{-n-1})$, $x \to \infty$. Hence, we can conclude that (2.3) gives an asymptotic expansion of $E_1(z)$ for $z \to \infty$, $|\arg z| < \frac{1}{2}\pi$. By using other representations of $R_n(z)$, more information can be obtained.

Let us write (2.4) in the form

$$R_n(z) = (-1)^{n+1}(n+1)!\; ze^{z}\int_z^\infty \frac{e^{-u}}{u^{n+2}}\,du. \tag{2.6}$$

If we put $u = z + \rho = x + iy + \rho$, where x,y,ρ are real, then

$$R_n(z) = (-1)^{n+1}(n+1)!\; z\int_0^\infty \frac{e^{-\rho}d\rho}{[(x+\rho)+iy]^{n+2}}\,, \tag{2.7}$$

which can be estimated as

$$(2.8)\qquad \begin{aligned} |R_n(z)| &\le S_n = (n+1)!\ (x^2+y^2)^{-\frac{1}{2}(n+2)} && \text{if } x \ge 0,\\ |R_n(z)| &\le T_n = (n+1)!\ |y|^{-n-2} && \text{if } x \le 0. \end{aligned}$$

From these estimates the asymptotic character of the expansion (2.3) follows for $z \to \infty$, $|\arg z| < \pi$, with $|y| \to \infty$ for $x < 0$. It can be shown that the domain of arg z can be extended beyond the bounds $\pm\pi$ to $\pm 3\pi/2$. However, numerical bounds for the remainder are not easily obtained outside $(-\pi,\pi)$.

The way of constructing the asymptotic expansion (2.3) seems rather ad hoc. By using (1.12) and expanding $1/(t+z)$ in powers of t we can use a general method from asymptotic analysis for integrals of the Laplace transform type. (See III.2.1 how this is applied in relation with the gamma function).

The asymptotic expansion (2.3) can be used for the computation of $E_1(z)$ for large values of $|z|$. In order to obtain information for the range of applicability we reason as follows. For a given $z = x + iy$ the remainder $R_n(z)$ in (2.3) is considered for $n = 0,1,2,\ldots$. We remark that $|R_n(z)|$ decreases until n reaches a certain value n_0 (depending on z). From (we suppose $x > 0$)

$$\frac{S_{n+1}}{S_n} = \frac{n+2}{(x^2+y^2)^{\frac{1}{2}}}$$

we infer that S_n decreases until n exceeds the value $(x^2+y^2)^{\frac{1}{2}}$. Hence to obtain the least value of $|R_n(z)|$, it is plausible to take $n \doteq (x^2+y^2)^{\frac{1}{2}} = |z|$. For this value of n, we find, using Stirling's formula that S_n is about $(2\pi/n)^{\frac{1}{2}}e^{-n} = (2\pi/|z|)^{\frac{1}{2}}e^{-|z|}$. If this value is smaller than the desired accuracy, the given value of $|z|$ is large enough in order to use (2.3), otherwise $|z|$ is too small.

Other techniques from analysis may be used for (2.3) if $|z|$ is too small. We mention the Euler transformation (see LAUWERIER (1974) or OLVER (1974)) or techniques for a further expansion of the remainder in an asymptotic expansion. For this last aspect, the asymptotic expansion of $E_1(z)$ is often used as an example in the literature (see, for instance, BERG (1977) or OLVER (1974, p. 523)).

For small values of $|z|$ the first expansion in (2.1) can be considered. However, the condition function (see II.2.1) is exponentially increasing for increasing $x = \text{Re } z$. Intuitively one can verify this immediately. For large x, $E_1(x)$ behaves as e^{-x}/x. In order to obtain such small function values, serious cancellation takes place in summing the quantities in the first formula of (2.1).

For Ei the asymptotic expansion for $x \to -\infty$ and for $x \to +\infty$ can be given. The first case follows from (1.7) and (2.3). For $x \to \infty$ we have for $N = 0,1,2,\ldots$

$$\mathrm{Ei}(x) = \frac{e^x}{x}\left[\sum_{n=0}^{N} \frac{n!}{x^n} + \mathcal{O}(x^{-N-1})\right]. \tag{2.9}$$

Remark that it results from formal substitution of $z = -x$ in (2.3) and using (1.7) for $x < 0$. A proof of (2.9) follows from (1.9) and the fact that (2.3) holds for $\arg z = \pm\, \pi i$ (which is not proved here). A direct proof may be obtained by observing that

$$\mathrm{Ei}(x) = \int_1^x \frac{e^t}{t}\, dt + \mathcal{O}(1), \qquad x \to +\infty$$

which follows from (1.10). By partial integration of this integral we arrive at (2.9).

For E_n we have for $n = 1,2,\ldots$

$$E_n(z) = \frac{e^{-z}}{z}\left[\sum_{m=0}^{N} \frac{(-1)^m}{z^m}\frac{\Gamma(n+m)}{\Gamma(n)} + \mathcal{O}(z^{-N-1})\right] \tag{2.10}$$

for $n = 1,2,\ldots$, $N = 0,1,2,\ldots$, $z \to \infty$, $|\arg z| < 3\pi/2$. The remainder in (2.10) can be expressed in terms of the exponential integral $E_{n+N+1}(z)$. In this expansion n is fixed, i.e., it is not supposed that n grows with z. An expansion of $E_n(x)$ valid for $x + n \to \infty$ is given in Gautschi (1959), together with numerical bounds for the remainder.

The asymptotic behaviour of Si and Ci is well described by (1.17) and the expansions of f and g, which follow. We use (1.20).

Let us write

$$\frac{1}{1+t^2} = \sum_{n=0}^{N-1} (-t^2)^n + (-1)^N \frac{t^{2N}}{1+t^2}\,.$$

Then for $\mathbb{R}\mathrm{e}\; z > 0$

$$f(z) = \frac{1}{z}\sum_{n=0}^{N-1} \frac{(2n)!}{(-z^2)^n} + (-1)^N \int_0^\infty \frac{e^{-zt}t^{2N}}{1+t^2}\, dt$$

(2.11)

$$g(z) = \frac{1}{z^2}\sum_{n=0}^{N-1}\frac{(2n+1)!}{(-z^2)^n} + (-1)^N\int_0^\infty \frac{e^{-zt}t^{2N+1}}{1+t^2}\,dt.$$

Bounds for the remainders in these expansions follow from replacing $1/(1+t^2)$ by 1.

2.2. Chebyshev expansions

2.2.1. Expansions near the origin

Coefficients in the Chebyshev series for small argument values can be expressed in terms of Bessel functions. From LUKE (1969, Vol. II, p. 41-42) we obtain (with $\varepsilon_0 = 1$, $\varepsilon_k = 2$, $k \geq 1$)

$$\int_0^{ax} t^{-1}(e^t-1)dt = \sum_{n=0}^\infty E_nT_n(x), \qquad -1 \leq x \leq 1,$$

$$E_0 = -2\sum_{r=1}^\infty (-)^r v_r I_{2r}(a),$$

$$v_1 = \frac{1}{2}, \quad v_r = \sum_{k=1}^{r-1}(1/k) + (1/2r), \qquad r \geq 2,$$

$$E_n = (2/n)\sum_{k=0}^\infty \varepsilon_k(-)^k I_{n+2k}(a), \qquad n \geq 1.$$

$$\int_0^{ax} t^{-1}(1-e^{-t})dt = \sum_{n=0}^\infty A_n(a)T_n^*(x), \qquad 0 \leq x \leq 1,$$

$$A_0(a) = 2e^{-a/2}\sum_{k=0}^\infty u_k I_{k+1}(a/2),$$

$$u_0 = 1, \quad u_k = 2\sum_{r=1}^k (1/r) + 1/(k+1), \qquad k \geq 1,$$

$$A_n(a) = \frac{2(-)^{n+1}e^{-a/2}}{n}\left[I_n(\tfrac{a}{2}) + 2\sum_{k=0}^\infty I_{n+k+1}(\tfrac{a}{2})\right],$$

$$A_n(a) = \frac{2(-)^{n+1}}{n}\,[1 - e^{-a/2}\{I_0(\tfrac{a}{2}) + 2\sum_{k=1}^{n-1} I_k(\tfrac{a}{2}) + I_n(\tfrac{a}{2})\}].$$

$$\int_0^{ax} t^{-1}(1-\cos t)dt = \sum_{n=0}^{\infty} A_n T_{2n}(x), \qquad -1 \le x \le 1,$$

$$A_0 = 2\sum_{r=1}^{\infty} v_r J_{2r}(a),$$

$$v_1 = \frac{1}{2}, \quad v_r = \sum_{k=1}^{r-1} (1/k) + (1/2r), \qquad r \ge 2,$$

$$A_n = \frac{(-)^{n-1}}{n}\sum_{k=0}^{\infty} \varepsilon_k J_{2n+2k}(a), \qquad n \ge 1,$$

$$\int_0^{ax} t^{-1}\sin t\, dt = \sum_{n=0}^{\infty} B_n T_{2n+1}(x), \qquad -1 \le x \le 1,$$

$$B_n = \frac{2(-)^n}{2n+1}\sum_{k=0}^{\infty} \varepsilon_k J_{2n+2k+1}(a).$$

$$\int_0^{ax} t^{-1}(1-e^{it})dt = \sum_{n=0}^{\infty} C_n(a)T_n^*(x), \qquad 0 \le x \le 1,$$

$$C_n(a) = A_n(ia), \quad A_n(a) \text{ as in } (*).$$

These expansions follow from more general expansions for confluent hypergeometric functions. In order to show the relation with the Bessel functions we give some information on the construction of the coefficients.

Let us consider the well-known expansion (see ABRAMOWITZ & STEGUN (1964, p. 361))

$$\cos \lambda x = 2 \sum_{k=0}^{\infty}{}' (-1)^k J_{2k}(\lambda) T_{2k}(x)$$

where J_{2k} is a Bessel function. We suppose $-1 \le x \le 1$ and λ positive real, although it may be complex. Integrating with respect to λ we obtain

$$\int_0^\mu \cos \lambda x \, d\lambda = \frac{\sin \mu x}{x} = 2 \sum_{k=0}^{\infty}{}' (-1)^k e_{2k} T_{2k}(x),$$

with

$$e_n = \int_0^\mu J_n(\lambda) \, d\lambda.$$

From ABRAMOWITZ & STEGUN (1964, p. 480) it follows that $e_n = e_{n+2} + 2J_{n+1}(\mu)$. Integrating with respect to x gives

$$\int_0^t \frac{\sin \mu x}{x} \, dx = 2 \sum_{k=0}^{\infty}{}' (-1)^k e_{2k} \int_0^t T_{2k}(x) dx.$$

Using

$$\int_0^t T_0(x) dx = T_1(t)$$

$$\int_0^t T_{2k}(x) dx = \frac{1}{2} \left[\frac{T_{2k+1}(t)}{2k+1} - \frac{T_{2k-1}(t)}{2k-1}\right], \qquad k > 0$$

we obtain

$$\int_0^t \frac{\sin \mu x}{x} \, dx = \sum_{n=0}^{\infty} f_{2n+1} T_{2n+1}(t)$$

$$f_{2n+1} = \frac{(-1)^n}{(2n+1)} (e_{2n} + e_{2n+2}).$$

In this way the expansion for Si is obtained:

$$\text{(2.12)} \qquad \text{Si}(x) = \sum_{n=0}^{\infty} f_{2n+1} T_{2n+1}(x/\mu) \qquad -\mu \le x \le \mu,$$

where f_{2n+1} depends on μ. This free parameter is included in order to choose an interval for the approximation. The expansions on p. 153–154 all contain a free parameter a. If a is chosen too large the approximations may become ill-conditioned. In BULIRSCH (1967) coefficients for Si and Ci are given for the x-interval [-16,16]. For x near the end points of this interval the given expansions yield inaccurate results.

2.2.2. Expansions near infinity

For the sine and cosine integrals it is preferred to obtain expansions for f and g. From (1.26) and the expansion (4.5) on p.112 we obtain

$$iz[g(z) - if(z)] = \sum_{n=0}^{\infty} C_n(i\mu)T_n^*(\mu/z) \tag{2.13}$$

where $C_n(z)$ can be obtained from a recurrence relation.

For $E_1(z) = e^z\, U(1,1,z)$ such an expansion is also available. A direct approach can be given by observing that

$$y(x) = ze^zE_1(z), \qquad z = 1/x$$

satisfies the differential equation

$$x^2y' + (1+x)y = 1.$$

Substitution of

$$y(x) = \sum_{j=0}^{\infty} c_j(\lambda)T_j^*(x/\lambda) \tag{2.14}$$

gives the following recurrence relation

$$(2+\lambda)c_0(\lambda) + 3\lambda c_1(\lambda) + (3\lambda-2)c_2(\lambda) + \lambda c_3(\lambda) = 4$$

$$\begin{aligned}(k-1)\lambda c_{k-2}(\lambda) &+ 2[2+\lambda(2k-1)]c_{k-1}(\lambda) + 6\lambda k c_k(\lambda) + \\ &+ 2[(2k+1)\lambda - 2k]c_{k+1}(\lambda) + \lambda(k+1)c_{k+2}(\lambda) = 0,\end{aligned}$$

$$(k \geq 2).$$

For $\lambda = 1$ this result is al o given in FOX & PARKER (1968). It is a special case of (4.5) on p.112 and it converges for wide ranges of x and λ. Let us rewrite (2.14) as follows:

$$\omega z\, e^{\omega z}E_1(\omega z) = \sum_{j=0}^{\infty} c_j(1/z)T_j^*(1/\omega), \qquad \omega \geq 1. \tag{2.15}$$

Then from LUKE (1969,II.p.25) it follows that this expansion converges for all $z \neq 0$, $|\arg z| < 3\pi/2$.

The coefficients $c_j(1/5)$, giving an expansion of $E_1(x)$ for $x \geq 5$, are given in LUKE (1969, Vol. II, p. 322) and LUKE (1975, p. 105).

2.3. Continued fractions

We mention the important fraction

$$E_n(z) = e^{-z}\left(\frac{1}{z+}\,\frac{n}{1+}\,\frac{1}{z+}\,\frac{n+1}{1+}\,\frac{2}{z+}\,\cdots\right)$$

for $|\arg z| < \pi$, which converges better with increasing $|z|$. This fraction again follows from the results for hypergeometric functions (see (1.24) and p.100 ff).

From the above fraction the even and odd contraction can be obtained. Let

$$e^{z}E_n(z) = \underset{k=1}{\overset{\infty}{\Phi}}\,\frac{a_k}{b_k}$$

then the coefficients of the contractions are given in the following table.

coefficients contraction	a_1	b_1	$a_k, k=2,3,\ldots$	$b_k, k = 2,3,\ldots$
even	1	$z+n$	$-(k-1)(n+k-2)$	$z+n+2k-2$
odd	n	$z+n+1$	$(k-1)(n+k-1)$	$z+n+2k-1$

For $\dot{z} \in \mathbb{R}^+$ the sequence of convergents of the even contraction is increasing, and the sequence of convergents of the odd contraction is decreasing (see II.4.5).

3. ALGORITHMS AND IMPLEMENTATIONS

Discussed are special cases and known implementations of exponential type integrals. In contrast with the gamma function no efficient universal routine is available. In our opinion this is hardly possible because the class of exponential type integrals contains various special cases.

3.1. Exponential integral for real positive argument and integer order: $E_n(x)$

The computation of $E_n(x)$ (see (1.11)) can be based on a combination of the following representations:

$$E_n(x) = (-x)^{n-1}/(n-1)!\ \{\psi(n)-\ln(x)\} - \sum_{\substack{m=0\\ m\neq n-1}}^{\infty} (-x)^m/((m-n+1)m!)\,; \tag{3.1}$$

$$E_n(x) = e^{-x}\left(\frac{1}{x+}\,\frac{n}{1+}\,\frac{1}{x+}\,\frac{n+1}{1+}\,\frac{2}{x+}\,\cdots\right) \tag{3.2}$$

and the even and odd contractions (see 2.3);

$$E_n(x+h) = \sum_{j=0}^{\infty} \frac{(-h)^j}{j!}\,E_{n-j}(x) \quad \text{(Taylor expansion);} \tag{3.3}$$

$$E_n(x) = (e^{-x}-nE_{n+1}(x))/x \qquad \text{(recurrence relation);} \tag{3.4}$$

$$E_n(x) \sim \frac{e^{-x}}{x(n-1)!}\sum_{k=0}^{\infty} (-1)^k (k+n-1)!/x^k \qquad \text{(asymptotic expansion).} \tag{3.5}$$

3.1.1. The implementation of Stegun and Zucker (1974)

They implemented in ANSI FORTRAN 66 in double precision the exponential integral with the parameters

input :	rn	- the order ($\in \mathbb{N}$)
	x	- the argument, $x \geq 0$
output:	enx	- the value of $E_n(x)$
	expenx	- the value of $e^x E_n(x)$
	ier	- the integer error indicator.

In the routine the following machine dependent parameters have to be initialized:

rinf - the largest number x such that x and -x belong to the system of real (computer) numbers;

rmaxi - the largest integer i such that all integers in the range [-i,i] belong to the system of integer (computer) numbers;

nbm the number of binary digits of the mantissa.

Moreover, the constant of Euler, γ, must be initialized for the particular machine (in the paper 35 digits of γ are listed).
The computational problem used for $E_n(x)$, resp. $e^x E_n(x)$, reads

n=0	rinf	, for $0 \le x \le 1/\text{rinf}$;
	e^{-x}/x resp. $1/x$,	, for $1/\text{rinf} < x$;
$n \in \mathbb{N}$	min(1/(n-1), rinf)	, for $x = 0$;
	an a posteriori finite part of the series (3.1)	, for $0 < x \le 1$;
	an a posteriori finite part of the even contraction (3.2)	, for $1 < x < \text{rinf}-n$;
	0	, for $\text{rinf}-n < x$.

For $x \in (0,1]$ the series (3.1) is evaluated in the forward direction where the (n-1)-st term, apart from a factor, consists of ln(x) and $\psi(n)$; the evaluation is terminated when

$$|\text{TM/SUM}| < \text{TOLER}$$

with TOLER a tolerance variable and TM the value of the most recent term.
The even contraction of the continued fraction is evaluated in the forward direction (see II.4.8); the evaluation is terminated if either

$$1 - c_{i-1}/c_i \le 0 \quad \text{or} \quad 1 - c_{i-1}/c_i \le \text{TOLER},$$

with c_i the i-th convergent of the even contraction. The error indicator, IERR, is set to

0 : no error detected
1 : n < 0 or x < 0 (the value -rinf is delivered for $E_n(x)$ resp. $e^x E_n(x)$

2 : $0 < n <$ RMAXI and $n \notin \mathbb{N}$ (the value rinf is delivered for $E_n(x)$ resp. $e^x E_n(x)$).

REMARKS.

- In the note on the parameters for transportable numerical software of IFIP WG-2.5 the parameters RINF, RMAXI and NBM are called SOVFLO, IOFLO and SDIGIT, respectively.
- The summation of the alternating almost monotonically decreasing series is handled with care: the (n-1)-st term does not majorate in general the remainder of the series and therefore the termination test is not applied to this term.
- We consider the used termination criterion of the evaluation of the continued fraction not correct. The reason why and a counter example is given below.

The implemented even contraction is a continued fraction of the second class of BLANCH (1964).

After an equivalence transformation the continued fraction reads

$$c = \underset{k=1}{\overset{\infty}{\Phi}} \frac{\alpha_k}{1}, \qquad \alpha_1 = 1/(x+n), \alpha_k = \frac{-(k-1)(n+k-2)}{(x+n+2k-2)(x+n+2k-4)}.$$

The converge behaviour is given by the fraction

$$c = \underset{k=1}{\overset{\infty}{\Phi}} \frac{-.25}{1} = \frac{1}{2}, \quad c_k = \underset{j=1}{\overset{k}{\Phi}} \frac{-.25}{1} = -\frac{1}{2}k/(k+1).$$

For this continued fraction we have

$$1 - c_{k-1}/c_k < 1 - c_k/c$$

and therefore we do not consider the implemented criterion

$$1 - c_{k-1}/c_k < \text{TOLER}$$

foolproof, because the left hand side is a factor $(k+1)/k^2$ smaller than the actual truncation error.

The counter example below demonstrates that the truncation error majorates the prescribed tolerance for $x = 1$, i.e.

$$1 - c_{k-1}/c_k < \text{TOLER} < 1 - c_k/c.$$

$$\ldots$$
$$c_8 = 859580/1441729$$
$$c_9 = 748420/1255151$$
$$\ldots$$
$$c_{14} = 9591325648580/16083557845279$$

with $c_9^2/c_8 - c_{14} < 0$ (!).
Because of $c > c_{14} > c_9^2/c_8$, which in our case is equivalent to,

$$1 - c_8/c_9 < 1 - c_9/c,$$

we have that the tested quantity (left-hand side of inequality) is not an upper bound for the relative truncation error (right-hand side of inequality).

3.1.2. The implementation of GAUTSCHI (1973) and AMOS (1980)

Gautschi implemented in ALGOL 60: $f_n(x) = e^x E_n(x)$, $n = 1,2,\ldots N$, and $x > 0$. The parameters are

input : x - the argument, $x > 0$;
nmax - the number of exponential integrals: $f_1(x),\ldots,f_{nmax}(x)$;
d - the accuracy requirement: the required number of significant decimal digits;
output : f - an array for the values of $f_1(x),\ldots,f_{nmax}(x)$.

In the implementation no measures against overflow or underflow are taken; Euler's constant is initialized to 24 digits.
The used computational problem reads

an a posteriori finite part of the series (3.1) and the forward recurrence relation for $f_2(x),\ldots,f_{nmax}(x)$, for $0 < x \le 1$;

an a posteriori finite part of the even and odd contraction of the continued fraction (3.2) for $f_m(x)$, with m the integer closest to x,

and the recurrence relation
in the backward direction for
$f_1(x),\ldots,f_{m-1}(x)$ and in the
forward direction for $f_{m+1}(x),\ldots$
$f_{nmax}(x)$, for $1 < x$.

The series is evaluated in the forward direction; the summation is terminated if either

$$s_o - s_e \le eps * s$$

or s_o or s_e cease to behave monotonically, with

s_o the partial sum with an odd number of terms,
s_e the partial sum with an even number of terms,
s the arithmetic mean of s_o and s_e,
eps the precision 10^{-d}.

The even and odd contraction of (3.2) are evaluated by means of the evaluation of the sum representation of the convergents where the quotient of the terms obey a nonlinear two-terms recurrence relation (see II-4.5.3); the evaluation is terminated if either the successive convergents of the even or the odd contraction cease to behave monotonically (Rutishauser's device) or

$$w_o - w_e \le eps * (w_o+w_e)/2$$

with w_o and w_e the convergents of the odd and even contraction, respectively. The error handling consists of testing for $x \le 0$, $nmax \le 0$, and, when appropriate, calling a procedure RECOVERY (to be supplied by the user) followed by a return to the user program.

REMARKS.

- No range for the argument is indicated where the implementation is free of overflow and underflow; so we do not consider it robust.
- In order to prevent infinite loops in cases where d, the required accuracy, is specified unreasonable large for a particular computer, we propose to test $_{10}{}^{-d}$ against the machine precision (which is named and callable in nowadays languages) times a small factor (5 or 10), instead of using Rutishauser's device, which is stronger because it requires the monotonic behaviour of the *calculated* quantities.

- The implementation can be considered as belonging to the variable precision class. (See II.1).

An improved FORTRAN variant of the ALGOL 60 implementation is due to AMOS (1980). Implemented is the computation of

$$E_{N+k}(x) \ , \quad x \geq 0, \quad N \geq 1, \ k = 0, 1, \ldots, M-1.$$

The used algorithm reads

recursion starting with $E_n(x)$, with n, the integer closest to x within the constraint $N \leq n \leq N+M-1$

$E_n(x)$ is calculated via

the power series for $0 \leq x \leq 2$

the confluent hypergeometric function $U(n,n,x)$ for $2 < x < 0$.

$U(n,n,x)$ is computed by recurrence relations via the Miller algorithm for $U(n+k,n,x)$, $k = 0,1,\ldots$, with a normalizing relation derived from the two-term recurrence relation satisfied by $E_n(x)$ and $E_{n+1}(x)$. Truncation error bounds are derived and used in error tests in EXPINT. Exponential scaling is also provided as a subroutine option. The improvement concerns savings in the recurrence when N is large and $x \leq 1$ or when x is large and N+M-1 is small.

3.1.3. The implementation in NUMAL

The implementations:

ENX delivers $E_k(x)$, $k = n_1,\ldots,n_2$, $x > 0$

NONEXPENX delivers $e^x E_k(x)$, $k = n_1,\ldots,n_2$, $x > 0$

are in ALGOL 60 for a CD CYBER, and are heavily based on Gautschi's; EI is based on CODY & THACHER (1968,1969).

The computational problem used in ENX and NONEXPENX is:

$E_1(x) = -E_i(-x)$
and the forward recurrence
for $E_{n_1},\ldots,E_{n_2}$, for $0 \leq x \leq 1.5$

an a posteriori finite part
of the Taylor expansion of
$E_m(x)$, with m the integer
closest to x, and the recurrence relation with $E_m(x)$
as initial value , for $1.5 < x < 10.5$

an a posteriori finite part of the even and odd continued fraction (3.2) for $e^x E_m(x)$ and the recurrence relation with $E_m(x)$ as initial value, analogous to Gautschi , for $10.5 \le x$.

REMARKS.

- $E_1(x)$ may efficiently be obtained from a call to E_i $(E_1(x) = -E_i(-x))$.
- The disadvantage of the slow convergence of the continued fraction near $x = 1$ is replaced by the unproven used termination criterion of the Taylor series. The termination criterion is to handle an alternating series and a positive series; for the alternating series we agree with the used criterion (but use names for the machine precision), while for the positive case we propose

  ... <u>while</u> |w1| > 10* machine precision <u>do</u> ...

  Namely, for the positive series the first neglected term is a measure for the relative truncation error.
- The implementation is not portable and not robust. (machine parameters are not named and there is no error handling).
- The implementation can be considered as belonging to the variable precision class.

3.1.4. The implementation in NAG

The implementation S13AA delivers $E_1(x)$. The computational problem reads

$$\sum_j{}' a_j T_j(t) - \ln x \quad \text{with } t = \tfrac{1}{4}x - 1 \qquad , \text{ for } 0 < x \le 4$$

$$e^{-x}/x \sum_j{}' b_j T_j(t) \quad \text{with } t = (11.25-x)/(3.25+x) \qquad , \text{ for } 4 < x < xhi$$

$$0 \qquad , \text{ for } xhi \le x$$

where xhi is machine dependent.

The only parameters are the argument and an error indicator. The latter is set to

0 - no error detected

1 - $x \le 0$ (E_1 is set to zero).

REMARKS.

- The method is based on CLENSHAW c.s. (1963; see II); the bilinear transformation is modified because of the faster convergence of the Chebyshev series in this new variable.
- The required terms of the Chebyshev sums are a priori converted to a power sum; the evaluation of the latter is more efficient while the same accuracy is obtained. (A 'streamlined' form is not considered)
- The implementation is available in NAGF (mark 7), NAGA and NAGB (mark 2).
- THACHER (1965) obtained experimentally a faster convergent series for $0 < x \le 4$ by the transformation $t = \beta x/[1+(\beta-.25)x]$, $\beta = .2915$

3.1.5. The implementation in NATS

The implementation EONE which calls a poly-algorithm is written in ANSI FORTRAN 66. The only parameter of EONE is the argument. The computational problem is

$$\begin{array}{ll} P_{lm}(x) - \ln x \quad , & \text{for } 0 < x \le 1 \\ e^{-x}Q_{lm}(1/x) \quad , & \text{for } 1 < x \le 4 \\ e^{-x}/x\ \{1 + R_{lm}(1/x)/x\}, & \text{for } 4 < x, \end{array}$$

where P,Q,R denote rational functions with l the degree of the numerator and m the degree of the denominator. The coefficients for the various domains and a diversity of relative precisions (down to 10^{-21}) are given in CODY & THACHER (1968).

The error handling is done by a call to the general FUNPACK routine MONERR; $x \le 0$ is signalized.

REMARK.

. The paper of Cody & Thacher also entailed the software provided by PACIOREK (1970), IMSL and CERN.

3.2. Exponential integral for imaginary argument: $E_1(ix)$

According to (1.14a) and (1.15) we have

$$E_1(ix) = -Ci(x) + i(Si(x) - \pi/2)$$

with Ci(x) the cosine integral and Si(x) the sine integral. The computation of Ci(x) and Si(x) can be based on a combination of the following representations as special cases of (3.1) and (3.2):

$$(3.6)\qquad E_1(ix) = -(\gamma+\ln x + \sum_{m=1}^{\infty} (-1)^m x^{2m}/((2m)(2m)!) + i\left(\sum_{m=0}^{\infty} (-1)^m x^{2m+1}/((2m+1)(2m+1)!)-\pi/2\right);$$

$$(3.7)\qquad E_1(ix) = e^{-ix} \mathop{\Phi}_{k=1}^{\infty} \frac{a_k}{b_k} ,$$

where the coefficients are given in the following table.

coefficients / representation	a_1	b_1	a_k, k = 2,3,...	b_k, k = 2,3,...
cont. fraction	1	ix	k÷2	ix, k is odd 1, k is even
even contraction	1	ix+1	$-(k-1)^2$	ix+2k-1
odd contraction	1	ix+2	(k-1)k	ix+2k

(3.8) The formulas (1.17).

3.2.1. The implementation of Stegun and Zucker (1976)

They implemented in ANSI FORTRAN 66 in double precision a poly-algorithm for the sine, cosine, exponential integrals and related functions with the parameters

input: ic - an indicator for the desired integrals

ic	functions to be computed
1	Si, Ci
2	E_i, $e^{-x}E_i$
3	E_i, $e^{-x}E_i$, Shi, Chi
4	Si, Ci, E_i, $e^{-x}E_i$, Shi, Chi

x - the argument (> 0 for ic = 2)

output: the appropriate function values are returned in si,ci,ei,exnei, shi,chi; moreover, the variables cii,shii are used (they have sense for negative arguments).

In the routine the following machine dependent parameters have to be initialized

rinf - the largest number x such that x and -x belong to the system of real (computer) numbers;

nbm - the number of binary digits of the mantissa;

ulsc - the maximum value for the argument in order to obtain reliable results;

γ, $\pi/2$, π, log 2.

The computational problems used for Ci and Si read

-rinf resp. 0 , for $x = 0$

an a posteriori finite part of the series (3.6) , for $0 < x \leq 2$

an a posteriori finite part of the even contraction (3.7) , for $2 < x <$ ulsc

0 resp. $\pi/2$, for ulsc $\leq x$

Ci(-x)-iπ resp. -Si(-x) , for $x < 0$.

The matching of the various methods is based on the results of an experimentally obtained efficiency profile; the results of the experiments are provided in the paper. The series is evaluated in the forward direction; the evaluation is terminated when

$$|\mathrm{TM}/\mathrm{SUM}| < \mathrm{TOLER},$$

with TOLER a tolerance variable, TM the value of the most recent term and SUM the calculated partial sum of Si or Ci, respectively.
The even contraction of the (complex) continued fraction is evaluated in the forward direction (see II.4.8); the evaluation is terminated if either

$$|1 - c_{i-1}/c_i|^2 \leq \mathrm{TOLER}^2$$

or

$$|1 - c_{i-2}/c_{i-1}|^2 \leq |1 - c_{i-1}/c_i|^2$$

with c_i the i-th convergent of the even contraction. The error indicator, IERR, is set to

0 - no error detected

1 - if ic $\neq$ 1 and $x < 0$ then ei and exnei contain invalid results.

REMARKS.

- The even contraction of the continued fraction (3.7) can be represented by

$$c = \mathop{\Phi}_{k=1}^{\infty} \frac{\alpha_k}{1}, \quad \alpha_1 = 1/(1+ix), \quad \alpha_k = \frac{-(k-1)^2}{(2k-1+ix)(2k-3+ix)} .$$

  This fraction is of the third class of BLANCH (1964). We doubt the correctness of the stopping criterion, because

$$\lim_{k\to\infty} \alpha_k = -.25$$

  and therefore theorem 8 of Blanch (the remainder is in absolute sense estimated above by a factor times the difference of successive convergents) does not apply.
- The chosen value of the imaginary part of Ci and Chi, for $x < 0$, is $-\pi$. We should omit the delivery of this value, because it is not universal to deliver $-\pi$ ($+\pi$ could also have been chosen; it can be reflected into the choice of the value of the logarithm for negative values) and in absence of the type DOUBLE COMPLEX the parameter list is confusing.
- For the naming of the machine parameters see the earlier remark in 3.1.1.
- Although in the introduction of the paper the easy-to-modify criterion of the implementation is explicitly mentioned as an aim, we feel that the authors did not succeed with respect to this point.

3.2.2. The implementation in NAG

The implementations S13AC and S13AD deliver Ci respectively Si. The computational problem is heavily based on BULIRSCH (1967) and given in the following table.

Si(x)	Ci(x)	argument range
$x * \sum_k' a_k T_k(t)$	$\ln(x) + \sum_k' b_k T_k(t)$	$0 < x \le 16$ with $t = 2(x/16)^2 - 1$
$\frac{\pi}{2} - \frac{f(t)\cos x}{x} - \frac{g(t)\sin x}{x^2}$	$\frac{f^c(t)\cos x}{x} - \frac{g^c(t)\sin x}{x^2}$	$16 < x < xhi$ with $t = 2(16/x)^2 - 1$
$\frac{\pi}{2}$	0	$xhi \le x$
$-$ Si(-x)	not provided $(\mathrm{Ci}(x) = \mathrm{Ci}(-x) \mp i\pi)$	$x < 0$

The functions f, g, f^c, g^c are represented by Chebyshev expansions; xhi is machine dependent.

The only parameters are the argument and an error indicator. The error indicator is not used in S13AD (Si); in S13AC (Ci) an error return is given with ifail = 1 when $x \le 0$.

3.3. Exponential integral for negative argument: $E_1(xe^{\pm i\pi})$

From relation (1.8) we have

$$E_1(xe^{\pm i\pi}) = - \mathrm{Ei}(x) \mp i\pi.$$

In literature implementations for Ei(x) are provided. The computation of Ei(x) can be based on a matching of (2.1) and (2.9).

3.3.1. The implementation of Stegun and Zucker (1976)

The general description is given in 3.2.1. The used computational problem reads

the series (2.1) , for $0 < x \le$ aell

the asymptotic expansion (2.9) , for aell $< x$,

with aell a machine dependent parameter. The error indicator, IERR, is set to

0 - no error detected

1 - negative argument; ei and exnei contain invalid results.

3.3.2. The implementation in NATS

The implementation EI, which calls a poly-algorithm is written in ANSI FORTRAN 66. The only parameter of EI is the argument. The computational problem with experimentally chosen representations reads

$$
\begin{array}{ll}
-E_1(-x) & \text{, for } x < 0 \\
\ln(x/x_0)+(x-x_0)\sum_k p_k T_k^*(x/6)/ & \text{, for } 0 < x \le 6 \text{ and } x_0 \text{ the zero of} \\
\qquad\qquad\qquad\quad \sum_k q_k T_k^*(x/6) & \quad \mathrm{Ei}(x) \\
\frac{e^x}{x}\left\{\alpha_0 + \underset{k}{\Phi}\, \frac{\beta_{k-1}}{x+\alpha_k}\right\} & \text{, for } 6 < x \le 12 \\
\frac{e^x}{x}\left\{\alpha'_0 + \underset{k}{\Phi}\, \frac{\beta'_{k-1}}{x+\alpha'_k}\right\} & \text{, for } 12 < x \le 24 \\
\frac{e^x}{x}\left\{1+ \frac{1}{x}(\alpha''_0 + \underset{k}{\Phi}\, \frac{\beta''_{k-1}}{x+\alpha''_k})\right\} & \text{, for } 24 < x < \mathrm{xmax} \\
\mathrm{xinf} & \text{, for } \mathrm{xmax} < x.
\end{array}
$$

The error handling is done by a call to the general routine MONERR; x = 0 and overflow are signalized.

REMARKS.

- The coefficients for the various above mentioned functions and for a diversity of relative precisions (down to 10^{-21}) are given by CODY & THACHER (1969) and obtained via rational minimax approximation.
- The paper of CODY & THACHER also entailed the software provided by PACIOREK (1970), IMSL, NUMAL and CERN.

3.4. Exponential integral for general z: $E_n(z)$

The only implementation known to us is BEAM (1960), which is based on the evaluation of the continued fraction. In relation with (1.24), LUKE (1977) gives routines for the determination of the Chebyshev coefficients of the confluent hypergeometric function U for *real arguments* (see his ch.VIII) and also routines for the expansion of exponential type integrals in series of Chebyshev polynomials. When the coefficients b_k, c_k of the following expansions

$$f(x) = \sum_{k=0}^{\infty} b_k T_k^*(x/\lambda), \quad 0 \le x \le \lambda$$

$$F(x) = \sum_{k=0}^{\infty} c_k T_k^*(\lambda/x), \quad 0 < \lambda \le x$$

are given, the routines yield the coefficients g_k, h_k in

$$g(x) = e^{-ax}x^{u-1} \int_0^x e^{at}t^u f(t)\,dt = \sum_{k=0}^{\infty} g_k T_k^*(x/\lambda), \quad 0 \le x \le \lambda$$

or

$$G(x) = e^{bx}x^{-u} \int_x^{\infty} e^{-bt}t^u F(t)\,dt = \sum_{k=0}^{\infty} h_k T_k^*(\lambda/x), \quad 0 < \lambda \le x$$

Re $b > 0$ (see LUKE (1977, Ch.XI)).
These routines may be used as starting point for the complex case.

On the other hand, one can use via (1.24) the expansion (4.5) on p.112, where the coefficients are functions of the argument z. The coefficients obey a recurrence relation (p.27 in LUKE (1969)) and asymptotic estimates are given (p.28 in LUKE (1969)). For the calculation of these coefficients a proper use of the Miller algorithm should be made. A concise treatment is given in LUKE (1976).

TODD (1954) recommends the Laguerre quadrature method in preference to the asymptotic expansion for the terms I_1, I_2 in

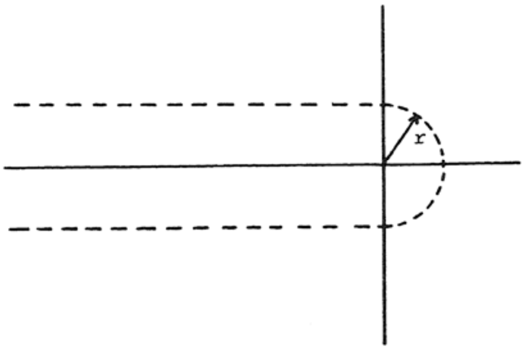
$$e^z E_1(z) = I_1 - iI_2$$

with

$$I_1 = \int_0^{\infty} e^{-\rho} \frac{x+\rho}{(x+\rho)^2+y^2}\,d\rho$$

$$I_2 = \int_0^{\infty} e^{-\rho} \frac{y}{(x+\rho)^2+y^2}\,d\rho, \quad z = x + iy.$$

Bounds for the approximating error are given; their iso-bound curves look like



For z outside 'the finger' the approximating error is less than

$$(n!)^2/r^{2n+1}.$$

Implementations for obtaining the (Laguerre) Gaussian weights and abscissae are published by GAUTSCHI (1968a,1968b and absorbed in NUMAL) and GOLUB & WELSCH (1969); the ALGOL 68 library of NAG contains implementations in order to access tabulated Gaussian weights and abscissae, so a limited number are direct available. In STROUD & SECREST (1966) FORTRAN-routines are provided.

REMARK. Todd compares the quadrature method with the asymptotic expansion; the above given estimate of the error for the quadrature method is smaller than the error estimate for the asymptotic expansion (2.8). As shown by the experiments of STEGUN & ZUCKER (1974,1976) - i.e., the continued fraction is more efficient than the asymptotic expansion along the real and imaginary axis - it is interesting to compare the Laguerre quadrature method with the continued fraction for reasonable (large) z. Consider Gautschi's remark in relation with w(z) on this matter (Chapter V).

4. SOME ASPECT OF ERROR ANALYSIS

In this section we consider: the effect of perturbation of the argument, the recurrence relation for $E_n(z)$ *and the recurrence relation for* $E_n(x)$, $n > 0$ *in order to obtain* $E_m(x)$, $m < 0$.

4.1. The effect of perturbation of the argument

From

$$\frac{d}{dz} E_n(z) = -E_{n-1}(z)$$

we have for the relative error amplification (in first order)

$$\left|\frac{z}{E_n(z)} \frac{d}{dz} E_n(z)\right| = \left|z \frac{E_{n-1}(z)}{E_n(z)}\right| \ .$$

For the special case $E_1(z)$ we have for $x \in \mathbb{R}^+$

argument	amplification / $\|x\|$	asymptotical amplification for $x \to \infty$
x	$e^{-x}/E_1(x)$	x
$-x$	$e^x/\|Ei(x)\|$	e^x
ix	$1/\|E_1(ix)\|$	x

REMARKS.

- The error amplification is inversely proportional to the modulus of the delivered function value.
- The documentation of $E_1(x)$ in the NAG manual contains a graph of $e^{-x}/E_1(x)$.
- It is curious, that when the amplification factor is so simple in terms of the result, as is the case here, it is not even mentioned in the documentation of most program libraries, or used in the routines.

4.2. The recurrence relation for $E_n(z)$, $n > 0$

In II.3.1 we considered the recurrence relation and derived a quadratic bound for the condition if we start the recursion for $n = [|z|]$ and recur down the ρ_n-hill.

REMARK. From (1.11) it follows that $x > 0$, $k > m$ imply $E_k(x) < E_m(x)$. So it is tempting to conclude that backward recursion is stable. This is no rule of thumb, as suggested by STEGUN & ABRAMOWITZ (1956): "...However, in the case of the function $y_n(x)$ there will be no loss of accuracy since this function is an increasing function of n for fixed n...". The point is here that (1.13) is not *positive* in the backward form (i.e., the coefficients are not all positive). A recursion, which is positive in the direction of increasing function values, is stable.

4.3. The recurrence relation for $E_n(x)$, $n > 0$

This recurrence is used in NUMAL in order to obtain the derivatives of $E_n(n)$, $n = 2,3,\ldots,10$, because for $x \in (1.5,10.5)$ the Taylor expansion

$$E_n(y+h) = \sum_{k=0}^{\infty} \frac{(-h)^k}{k!} E_{n-k}(y), \qquad y = [x+.5]$$

is used, where the $E_{n-k}(y)$ are obtained from the recurrence

$$E_{j-1}(y) = (e^{-y}-(j-1)E_j(y))/y, \qquad j = y,y-1,\ldots .$$

For E_j with positive index, we have a recurrence down the ρ-hill, while for a negative index we have the recurrence for α_n, a positive recurrence, and so a benign problem. (see II.3.1 with respect to the recurrence of α_n).

5. TABULATED COEFFICIENTS

BULIRSCH (1967)

Si(x), Ci(x) coefficients of the Chebyshev expansions up to 17d on $|x| < 16$, $16 \leq |x|$.

CLENSHAW, MILLER & WOODGER (1963)

Ei(x) coefficients of the Chebyshev expansions up to 17d on $x^2 < 16$, $x \leq -4$.

(In CLENSHAW (1962) the above coefficients are provided for the same intervals up to 21d).

CODY & THACHER (1968)

$E_1(x)$ rational approximations up to 22d on $(0,1],[1,4],[4,\infty)$; continued fraction expansion up to 25s for small x.

CODY & THACHER (1969)

Ei(x) rational approximations up to 22d on $(0,6],[6,12],[12,24],$ $[24,\infty)$;
zero of Ei(x) to 30d.

LUKE (1969) (main reference, with much more information; see also LUKE (1975))

$\{E_1(z)+\ln z+\gamma\}/z$ main diagonal Padé approximations for $n = 2,\ldots,10$ up to 20d and the corresponding approximation errors for $z = 1,i,-1$.

$Si(z)/z$ main diagonal Padé approximations for $n = 2,\ldots,10$ up to 20d and the corresponding approximation errors for $z = 1,2,\ldots,10$

$4(\gamma+\ln z-Ci(z))/z^2$ main diagonal Padé approximations for $n = 2,4,\ldots,10$ up to 20d and the corresponding approximation errors for $z = 1,2,\ldots,10$

$(\gamma+\ln z-Ei(z))/z$ diagonal of rational approximation array for $n = 0,1,\ldots,10$ up to 13d and the corresponding approximation errors for $z = r,ir,-r$ with $r = 1,2,\ldots,10$

$E_1(x), x > 5$ coefficients of the Chebyshev expansion (p.322),

$Ci(x), |x| < 8$ coefficients of the Chebyshev expansion (p.325)

$Si(x), |x| < 8$ coefficients of the Chebyshev expansion (p.325)

$ze^zE_1(z)$ diagonal of rational approximation array for $n = 0,1,\ldots,10$ up to 13d and the corresponding approximation errors for a variety of z.

6. TESTING

STEGUN & ZUCKER (1974,1976) compared the results of their programs with published values. Further checks were obtained by comparing with other (less efficient) methods or algorithms e.g.: overlapping the power series with either the asymptotic expansion or the continued fraction, using various forms of the continued fraction, numerical quadrature.
The numerical accuracy was ascertained by comparing multi-precision results with analogous results of the single and double precision implementations. Their test argument values are $0,10^j(10^j)\,10^{j+1}$, $j = -2(1)$ jend, with jend appropriately chosen. A driver program was published in the 1974 paper.

REMARK.

- Error bounds for the evaluation of the approximation in finite precision are not provided in the paper; the correct coding as well as the numerical accuracy are evidenced by experimentation. The implementations belong to the naive program class.

- The truncation error is of the order of the machine precision because the implementation is of the variable precision class.

GAUTSCHI (1973) compared the results of his procedure with results in 40s obtained by Thacher via a desk calculator, by - we assume - the same algorithm; to be precise: in the latter comparison the 40s results were taken as yard stick, and compared with the results from a FORTRAN double precision version of the published ALGOL 60 implementation. So, another piece of software was tested with respect to accuracy.

The implementation in NUMAL was based on the assumptions: Gautschi's algorithm and implementation as well as CODY & THACHER (1968,1969) coefficients are correct. Diverse tests showed that the coding, the algorithms and the coefficients were correct.

The NATS-implementations were tested on random arguments: a so called accuracy profile.

The NAG-implementations were tested with automatic portable test software.

7. APPLICATIONS

This paragraph illustrates the paraphrase: pitfalls in computation or why a math book and even a program library is not enough. Program libraries will not - and we think will never - be sufficient for solving problems, because a huge (infinite) number of problems are thinkable which can be expressed in basic special functions, and all those problems can't be incorporated in the program library. So mathematical skill, numerical insight and programming technique remain necessary, though on a less extended scale.

ABRAMOWITZ & STEGUN (1964) tabulate from 5.1.28 up to and including 5.1.44 some integrals which are expressible in exponential integrals. TEMME (1976) mentions the special case (see also ACTON (1970))

$$\int_0^\infty \frac{\sin xt}{t2+1}\,dt = \tfrac{1}{2}\{e^{-x}\mathrm{Ei}(x) - e^{x}\mathrm{Ei}(-x)\},\ x > 0.$$

For small x cancellation occurs so another representation is desired, e.g.

$$-\sinh(x)\ \ (\gamma + \ln x\ ) + \tfrac{1}{2}\{e^{-x}\sum_{n=1}^{\infty}\frac{x^n}{nn!} - e^x\sum_{n=1}^{\infty}\frac{(-1)^n x^n}{nn!}\}.$$

KADLEC (1976) encountered in transport problems through a scattering medium definite integrals of the type

$$\int_0^1 R_{m,n}(x)f(x)dx$$

with R_{mn} a rational function ($m\leq n$) with real coefficients and (known) real poles, and $f(x)$ one of the functions $e^{-\gamma/x}$, $\ln|rx+s|$, $\ln|rx+s|.e^{-\gamma x}$. As an example we mention

$$\int_0^1 \frac{e^{-\gamma/x}}{ax+b}\,dx = (E_1(\gamma) - e^{\gamma a/b}E_1(\gamma(1+a/b)))/a,\ \gamma > 0,\ a \neq 0.$$

For small a cancellation occurs and another representation is necessary, e.g.

$$E_1(\gamma)\{\frac{1-e^{\gamma a/b}}{a}\} - \gamma/b\sum_{k=1}^{\infty}(-\gamma a/b)^{k-1}/k!E_{1-k}(\gamma).$$

The exponential integrals with negative index can be obtained - in a stable way - via recursion (see 4.2). The subtraction in the first term must be performed with high relative precision, e.g., via

$$-\frac{\gamma e^{\delta}\ \sinh\ \delta}{2b\ \ \delta},\ \delta = \gamma a/(2b).$$

REMARK. The above denominator, ax+b, is typical, because the rational function can be split into partial fractions.

ACTON (1970,ch.4,ex15) asks for the evaluation of

$$\int_0^x \frac{1-e^{-t}}{t}\sin t\ dt,\qquad \text{for } x = .2(.1)2.$$

A representation of this integral is (see 5.1.36 ABRAMOWITZ & STEGUN)

$$\mathrm{Si}(x) - (\pi/4+\mathrm{Im}\ E_1((1+i)x)).$$

For the required argument values we may look in ABRAMOWITZ & STEGUN for tabulated values; in table 5.1 Si is given, while E_1 for complex argument is not sufficiently given in table 5.7.
LUKE (1969, Ch.XVII, table 64.1) enumerates Padé approximations for

$$E(z) = \{E_1(z) + \ln z + \gamma\}/z.$$

So we could program the rational function and obtain the desired values of E_1, on e.g. a HP. (By the way, HENRICI (1977b) did not publish a program for $E_1(z)$).
If we compute $E_1(z)$ in an environment where program libraries are available, then we could obtain Si via a library; at the moment $E_1(z)$ is not available in current program libraries so the programming, testing etc. for this function remains. Apart from LUKE's approximation one could program the series (2.1) combined with the continued fraction given in paragraph (2.3). Properties such as the monotonic behaviour of the even and odd contraction do not hold for general z, so GAUTSCHI's implementation can't be transliterated for complex argument.

BEAM (1960) implemented in ALGOL 60 the Legendre continued fraction via algorithm (4.5.3) given in II.4. Because of the conciseness of the implementation and the used algorithm we consider BEAM's implementation useful in absence of faster algorithms. The programming with type complex can make the implementation more concise.
Another approach could be to integrate the Taylor series of the integrand. The resulting series can be transformed into a continued fraction. So the problem is then reduced to the evaluation of a continued fraction.

Finally, we like to remark that this application appeared in the context of numerical integration and that a straightforward application of a quadrature routine would easily yield the result provided the subtraction is handled with care.

As an example of where exponential integrals can be used in the computation of more advanced special functions we mention the implementation in FORTRAN of the computation of the Bickley functions, repeated integrals of $K_0(x)$ Bessel function, in terms of exponential integrals $E_n(x)$ due to AMOS (1983) and published as CALGO 609.

V. ERROR FUNCTIONS AND RELATED FUNCTIONS

In this chapter we consider the error functions erf, erfc and some related functions as w(z), Dawson's integral and the Fresnel integrals. In section 1 we give the definitions and interrelations in addition to the main analytical aspects of these functions. In section 2 important expansions are considered: power series, Chebyshev series, continued fractions and some other expansions. In section 3 the algorithms and implementations are discussed and section 4 is devoted to aspects of error analysis. A selection of tabulated coefficients for several types of expansions is given in section 5. In section 6 the testing of some implementations is enumerated. In section 7 some applications are given.

1. DEFINITIONS AND ANALYTIC BEHAVIOUR

The most well-known representations are given. Some less obvious relations are proved. Especially the relation between Fresnel integrals and the basic function w(z) is considered, together with relations for the functions f(z) and g(z) which describe the asymptotic behaviour of the Fresnel integrals.

1.1. The error function

The error function erf(z) is defined for all (finite) complex values of z by the integral

$$\text{(1.1)} \qquad \operatorname{erf}(z) = 2\pi^{-\frac{1}{2}} \int_0^z e^{-t^2}\, dt.$$

Its complement with respect to 1 is denoted by

$$\text{(1.2)} \qquad \operatorname{erfc}(z) = 1 - \operatorname{erf}(z) = 2\pi^{-\frac{1}{2}} \int_z^\infty e^{-t^2}\, dt.$$

In statistics we often see a slightly different function, P(z), called the normal or Gaussian probability function, and its complementary function Q(z) = 1 - P(z), given by

$$\text{(1.3)}\qquad \begin{aligned} P(z) &= (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{z} e^{-\frac{1}{2}t^2}\, dt = \tfrac{1}{2}\mathrm{erfc}(-z/\sqrt{2}) \\ Q(z) &= (2\pi)^{-\frac{1}{2}} \int_{z}^{\infty} e^{-\frac{1}{2}t^2}\, dt = \tfrac{1}{2}\mathrm{erfc}(z/\sqrt{2}). \end{aligned}$$

This leads to

$$\text{(1.4)}\qquad \mathrm{erf}(z) = 2P(z\sqrt{2}) - 1,\ \mathrm{erfc}(z) = 2Q(z\sqrt{2}).$$

We furthermore introduce the functions

$$\text{(1.5)}\qquad \begin{aligned} w(z) &= e^{-z^2}\,\mathrm{erfc}(-iz) \\ F(z) &= e^{-z^2}\int_0^z e^{t^2}\, dt. \end{aligned}$$

The function F is called Dawson's integral and w(z) is known in physics as the plasma dispersion function.

The Fresnel integrals (with applications in optics) are defined by

$$\text{(1.6)}\qquad C(z) = \int_0^z \cos\tfrac{1}{2}\pi t^2\, dt,\ S(z) = \int_0^z \sin\tfrac{1}{2}\pi t^2\, dt;$$

they can be expressed in terms of erfc and of w as will be done in the next subsection. For representing the Frensel integrals for large values of $|z|$ it is useful to introduce, for Re $z^2 > 0$,

$$\text{(1.7)}\qquad \begin{aligned} f(z) &= \frac{1}{\pi\sqrt{2}}\int_0^\infty \frac{e^{-\frac{1}{2}\pi z^2 t}}{1+t^2}\, t^{-\frac{1}{2}}\, dt, \\ g(z) &= \frac{1}{\pi\sqrt{2}}\int_0^\infty \frac{e^{-\frac{1}{2}\pi z^2 t}}{1+t^2}\, t^{\frac{1}{2}}\, dt. \end{aligned}$$

The relation between these functions and C and S will also be given below.

Finally, we introduce the repeated integrals of the error function, i.e., we define

$$\text{(1.8)}\qquad i^n\,\mathrm{erfc}(z) = \int_z^\infty i^{n-1}\,\mathrm{erfc}(t)\, dt,\ n = 0,1,2,\ldots$$

with

$$i^0 \operatorname{erfc}(z) = \operatorname{erfc}(z).$$

The operator i^n can be used with negative values of n, in which case it acts as a differential operator. For instance, we have

$$i^{-1} \operatorname{erfc}(z) = 2\pi^{-\frac{1}{2}} e^{-z^2}.$$

These functions often occur in physics and chemistry, notably in problems involving heat and mass transfer. They are defined for all finite values of z.

A representation as a single integral is given by

$$i^n \operatorname{erfc}(z) = \frac{2\pi^{-\frac{1}{2}}}{n!} \int_z^\infty (t-z)^n e^{-t^2} dt,$$

from which a recursion formula with respect to n easily follows. For $n = 1,2,...$ we obtain

$$(1.9) \qquad i^n \operatorname{erfc}(z) = -\frac{z}{n} i^{n-1} \operatorname{erfc}(z) + \frac{1}{2n} i^{n-2} \operatorname{erfc}(z).$$

The functions introduced here are special cases of confluent hypergeometric functions, ABRAMOWITZ & STEGUN (1964, Ch.13). For instance

$$i^n \operatorname{erfc}(z) = \pi^{-\frac{1}{2}} 2^{-n} e^{-z^2} U(\tfrac{1}{2}n+\tfrac{1}{2},\tfrac{1}{2},z^2),$$

with as special case the error function for $n = 0$. In terms of parabolic cylinder functions we have

$$i^n \operatorname{erfc}(z) = e^{-\frac{1}{2}z^2} (2^{n-1}\pi)^{-\frac{1}{2}} D_{-n-1}(z\sqrt{2}).$$

The second solution of the difference equation for erfc can be obtained from the known solutions of the difference equation for the U-function. The solutions are useful in order to decide upon how to use the recurrence relation.

1.2. Some further relations for the error functions

First we give another integral representation for erfc. We have

$$(1.10)\qquad \operatorname{erfc}(z) = \frac{2}{\pi} e^{-z^2} \int_0^\infty e^{-z^2t^2} \frac{dt}{t^2+1},$$

where we suppose that the integral converges; i.e., we suppose temporarily that Re $z^2 \geq 0$. We prove (1.10) by differentiating with respect to z. Thus we obtain for the-right hand side (denoted by $\phi(z)$)

$$\phi'(z) = -\frac{4z}{\pi} e^{-z^2} \int_0^\infty e^{-z^2t^2}\, dt = -\frac{2}{\sqrt{\pi}} e^{-z^2}.$$

Integrating this relation and using $\phi(0) = 1$, we obtain indeed (1.2).

From (1.10) we obtain furthermore

$$(1.11)\qquad \operatorname{erfc}(z) = \frac{2z}{\pi} e^{-z^2} \int_0^\infty \frac{e^{-\tau^2}}{\tau^2+z^2}\, d\tau,$$

where the domain of arg z^2 can be extended to $(-\pi,\pi)$. When we introduce $\zeta = iz$, with $0 < \arg\zeta < \pi$, we obtain from (1.11) and (1.5)

$$\operatorname{erfc}(-i\zeta) = e^{\zeta^2} w(\zeta) = \frac{2\zeta}{i\pi} e^{\zeta^2} \int_0^\infty \frac{e^{-\tau^2}}{\tau^2-\zeta^2}\, d\tau.$$

Writing

$$\frac{\zeta}{\tau^2-\zeta^2} = \tfrac{1}{2}\left[\frac{1}{\tau-\zeta} - \frac{1}{\tau+\zeta}\right],$$

we obtain

$$w(\zeta) = \frac{\zeta}{i\pi} \int_{-\infty}^\infty \frac{e^{-\tau^2}}{\tau^2-\zeta^2}\, d\tau = \frac{1}{2\pi i} \int_{-\infty}^\infty e^{-\tau^2} \left[\frac{1}{\tau-\zeta} - \frac{1}{\tau+\zeta}\right] d\tau$$

from which we obtain

$$(1.12)\qquad w(z) = \frac{i}{\pi} \int_{-\infty}^\infty \frac{e^{-t^2}}{z-t}\, dt,\ \operatorname{Im} z > 0.$$

This formula tells us that for Im $z > 0$ the function w is the Hilbert transform of the Gaussian or normal density function. For Im $z < 0$ we can use one of the relations

$$(1.13)\qquad \mathrm{erf}(-z) = -\mathrm{erf}(z),\ \mathrm{erfc}(-z) = 2-\mathrm{erfc}(z)$$

which easily follow from (1.1) and (1.2). In combination with (1.5), (1.13) yields

$$(1.14)\qquad w(-z) = 2e^{-z^2} - w(z).$$

Of course, this relation follows in a more direct way by using residue theory when z in (1.12) crosses the real axis. Other symmetry relations are

$$(1.15)\qquad \mathrm{erf}(\overline{z}) = \overline{\mathrm{erf}(z)},\ w(\overline{z}) = \overline{w(-z)}.$$

From (1.14) and (1.15) it follows that for the evaluation of $w(z)$ we can restrict $z = x + iy$ to the quarter plane $x \geq 0$, $y \geq 0$. The function $w(z)$ can be considered as the basic function. For Dawson's integral of (1.5) we have

$$(1.16)\qquad F(z) = \frac{i}{2}\sqrt{\pi}\left[e^{-z^2} - w(z)\right]$$

and the Fresnel integrals introduced in (1.6) satisfy

$$(1.17)\qquad \begin{aligned} C(z) \pm i\, S(z) &= \int_0^z e^{\pm\frac{1}{2}i\pi t^2}dt = \frac{1\pm i}{2}\,\mathrm{erf}\,\left[\tfrac{1}{2}\sqrt{\pi}(1\mp i)z\right] \\ &= \frac{1\pm i}{2}\left[1-e^{\zeta_\pm^2}w(\pm\zeta_\pm)\right],\ \zeta_\pm = \tfrac{1}{2}\sqrt{\pi}(1\pm i)z. \end{aligned}$$

The representations (1.16) and (1.17) are subject to cancellation of leading digits when they are used in computations for small values of z.

Next we will show how f and g of (1.7) are related to w, C and S. From the definitions and (1.11) we obtain for Re $z > 0$

$$(1.18)\qquad \begin{aligned} g(z)\pm if(z) &= \pi^{-1}2^{-\frac{1}{2}}\int_0^\infty e^{-\frac{1}{2}\pi z^2 t}\, t^{-\frac{1}{2}}\,\frac{dt}{t\mp i} \\ &= \pi^{-1}2^{\frac{1}{2}}\int_0^\infty e^{-\frac{1}{2}\pi z^2\tau^2}\,\frac{d\tau}{\tau^2\mp i} = \frac{1\pm i}{2}\,w(\pm\zeta_\pm) \end{aligned}$$

with $\zeta_\pm$ given in (1.17). Combining (1.17) and (1.18) we have (using $\zeta_+^2 = -\zeta_-^2$)

$$(1.19)\qquad \begin{aligned} C(z) + iS(z) &= \frac{1+i}{2} - e^{\frac{1}{2}i\pi z^2}\left[g(z) + if(z)\right], \\ C(z) - iS(z) &= \frac{1-i}{2} - e^{-\frac{1}{2}i\pi z^2}\left[g(z) - if(z)\right], \end{aligned}$$

from which we obtain

$$(1.20)\qquad \begin{aligned} C(z) &= \tfrac{1}{2} - \cos \tfrac{1}{2}\pi z^2\, g(z) + \sin \tfrac{1}{2}\pi z^2\, f(z) \\ S(z) &= \tfrac{1}{2} - \sin \tfrac{1}{2}\pi z^2\, g(z) - \cos \tfrac{1}{2}\pi z^2\, f(z). \end{aligned}$$

Inverting (1.20) we can express f and g in terms of C and S:

$$(1.21)\qquad \begin{aligned} f(z) &= \left[\tfrac{1}{2} - S(z)\right] \cos \tfrac{1}{2}\pi z^2 - \left[\tfrac{1}{2} - C(z)\right] \sin \tfrac{1}{2}\pi z^2, \\ g(z) &= \left[\tfrac{1}{2} - C(z)\right] \cos \tfrac{1}{2}\pi z^2 + \left[\tfrac{1}{2} - S(z)\right] \sin \tfrac{1}{2}\pi z^2. \end{aligned}$$

Since C and S are entire functions, it follows that f and g are entire functions as well.

The oscillatory behaviour of C(z) and S(z) is fully described by the circular functions in (1.20). For large $|z|$ the functions f and g are slowly varying. In the next section we give more information on the asymptotic expansions of f and g, from which the asymptotic representations of S and C are obtained by using (1.20).

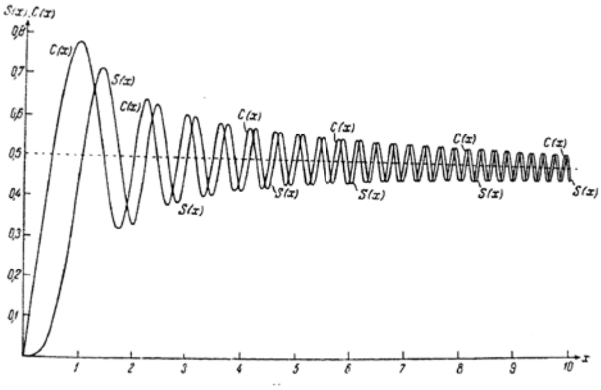


Figure 1. Graphs of S(x) and C(x), x ≥ 0.

To summarize the relations between the functions introduced here we express them in terms of the function $w(z)$.

$$\begin{aligned}
\operatorname{erf}(z) &= 1 - e^{-z^2}\, w(iz) \\
\operatorname{erfc}(z) &= e^{-z^2}\, w(iz) \\
i^n\operatorname{erfc}(z) &= (-i)^n\, 2^{-n}\, e^{-z^2}\, w^{(n)}(iz)/n! \\
P(z) &= \tfrac{1}{2} e^{-z^2}\, w(-iz/\sqrt{2}) \\
Q(z) &= \tfrac{1}{2} e^{-z^2}\, w(iz/\sqrt{2}) \\
F(z) &= \tfrac{1}{2} i\sqrt{\pi}\,\left[e^{-z^2} - w(z)\right] \\
C(z) \pm iS(z) &= \frac{1\pm i}{2}\left[1 - e^{\zeta_\pm^2}\, w(\pm\zeta_\pm)\right] \\
f(z) \pm ig(z) &= \frac{1\pm i}{2}\, w(\pm\zeta_\pm) \qquad\qquad \zeta_\pm = \tfrac{1}{2}\sqrt{\pi}\,(1\pm i)\,z
\end{aligned} \tag{1.22}$$

In the third line, $w^{(n)}(iz)$ means $\dfrac{d^n}{du^n}w(u)$ evaluated at $u = iz$.

2. FUNDAMENTAL FORMULAS

2.1. Expansions based on Taylor series and on asymptotic series

2.1.1. Taylor expansions

The expansion

$$\operatorname{erf}(z) = 2\pi^{-\frac{1}{2}} \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{n!\,(2n+1)} \tag{2.1}$$

is obtained by expanding the exponential function in (1.1). For large values of z this alternating series may be unsuitable. By transforming the integral (1.1) via $t \to z\sqrt{1-t}$ we obtain

$$\operatorname{erf}(z) = \frac{ze^{-z^2}}{\sqrt{\pi}} \int_0^1 e^{tz^2} \frac{dt}{\sqrt{1-t}}.$$

By expanding $\exp(tz^2)$ we obtain by using the integral representation for the beta function

$$\operatorname{erf}(z) = e^{-z^2} \sum_{n=0}^{\infty} \frac{z^{2n+1}}{\Gamma(n+3/2)}, \tag{2.2}$$

a series with positive terms if $z > 0$.

A Taylor series for $w(z)$ is given by

$$(2.3) \qquad w(z) = \sum_{n=0}^{\infty} \frac{(iz)^n}{\Gamma(1+n/2)} .$$

It is obtained by combining (1.5), (1.2) and (2.3). For C(z), S(z) and F(z) power series easily follow from the above ones or from their integral representations.

2.1.2. Asymptotic expansions

The starting point here is the integral for erfc(z) in (1.2). We transform it into

$$(2.4) \qquad \operatorname{erfc}(z) = \frac{e^{-z^2}}{z\sqrt{\pi}} \int_0^{\infty} e^{-\tau} \frac{d\tau}{\sqrt{1+\tau/z^2}}.$$

In this representation we suppose that $|\arg z| < \pi/2$. For $n = 0,1,\ldots$ we obtain by partial integration

$$(2.5) \qquad \operatorname{erfc}(z) = \frac{e^{-z^2}}{z\pi} \left[\sum_{m=0}^{n-1} (-1)^m \Gamma(m+\tfrac{1}{2}) z^{-2m} + (-1)^n \Gamma(n+\tfrac{1}{2}) z^{-2n} \theta_n(z)\right]$$

with

$$(2.6) \qquad \theta_n(z) = \int_0^{\infty} e^{-\tau}(1+\tau/z^2)^{-n-\frac{1}{2}} \, d\tau.$$

Suppose now that $z \in \mathbb{C}$ is such that

$$(2.7) \qquad |1+\tau/z^2| \geq 1 \text{ for all } \tau \geq 0.$$

Then $|\theta_n(z)| \leq 1$ and we conclude that for the values of z satisfying (2.7) the absolute value of the remainder in the asymptotic expansion (2.5), taking n terms, is not larger than the absolute value of the first neglected term. For real z, it has the sign of this term.

To describe the values of z satisfying (2.4), we remark that the equation

$$|1+\zeta| = 1$$

in the ζ-plane is satisfied by the points on the circle $(u+1)^2 + v^2 = 1$, where $\zeta = u + iv$. It follows that (2.7) is satisfied by those z-values satisfying $|\arg z^2| \leq \pi/2, z \neq 0$. By using complex values of τ in (2.6), we can give bounds for a wider z-domain. To show this we write

$$(2.8)\qquad z^2 = re^{i\phi},\ r > 0,\ -\pi < \phi \le -\pi/2 \text{ or } \pi/2 \le \phi < \pi.$$

Then we consider the integral

$$(2.9)\qquad \int_{C_R} e^{-\tau}(1+\tau/z^2)^{-n-\frac{1}{2}}d\tau$$

where the contour in the complex τ-plane is shown in the following picture. R is a positive number, $\theta = \phi-(\text{sign}\,\phi)\pi/2$.

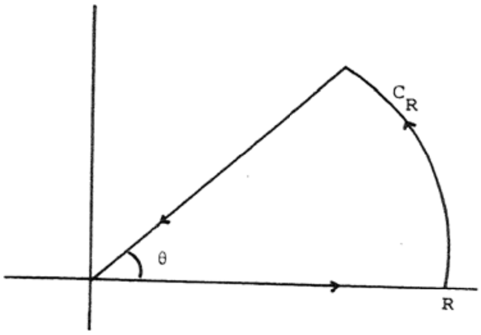


Figure 2. Contour of integration for (2.9).

From (2.8) it follows that $|\theta| < \pi/2$. The singularity $\tau = -z^2$ of the integrand in (2.9) lies outside the contour C_R. Hence, by using Cauchy's theorem, (2.9) vanishes. Furthermore, the contribution along the circular arc of C_R vanishes in the limit $R \to \infty$. It follows that (2.6) can be written as

$$\theta_n(z) = \int_0^{\infty e^{i\theta}} e^{-\tau}(1+\tau/z^2)^{-n-\frac{1}{2}}d\tau,$$

where on the path of integration arg $\tau = \theta$. Again we can use (2.7), now for $\tau = \rho e^{i\theta}$. Since arg $\tau/z^2 = \pm\pi/2$, (2.7) holds true for the considered values of τ and z^2. The bound for $\theta_n(z)$ thus becomes

$$|\theta_n(z)| \le \int_0^\infty e^{-\tau\cos\theta}d\tau = 1/\cos\theta = 1/|\sin\phi|.$$

Resumé. The remainder $\theta_n(z)$ in the expansion (2.5) is bounded as follows:

(2.10) $$|\theta_n(z)| \le \begin{cases} 1 & \text{if } |\phi| \le \pi/2 \\ 1/|\sin\phi| & \text{if } -\pi < \phi \le -\pi/2 \text{ or } \pi/2 \le \phi < \pi \end{cases}$$

where $\phi = \arg z^2$, $n = 0,1,2,\ldots$.

Remark. By refining the above methods it can be shown that (2.5) gives an asymptotic expansion for the range $|\arg z| < 3\pi/4$. Note that for $|\arg z| \ge \pi/2$ the reflection formula $\operatorname{erfc}(-z) = 2 - \operatorname{erfc}(z)$ can be used.

For f and g introduced in (1.7) we can also obtain asymptotic expansions. By writing

$$1/(1+t^2) = \sum_{m=0}^{n-1} (-t^2)^m + (-1)^n t^{2n}/(1+t^2)$$

we obtain upon substituting this in the integrals of (1.7)

(2.11) $$\begin{aligned} f(z) &= \frac{1}{\pi\sqrt{2}} \sum_{m=0}^{n-1} \frac{(-1)^m \Gamma(2m+\frac12)}{(\frac12\pi z^2)^{2m+\frac12}} + \frac{(-1)^n}{\pi\sqrt{2}} \int_0^\infty \frac{e^{-\frac12\pi z^2 t} t^{2n-\frac12}}{1+t^2}\, dt, \\ g(z) &= \frac{1}{\pi\sqrt{2}} \sum_{m=0}^{n-1} \frac{(-1)^m \Gamma(2m+3/2)}{(\frac12\pi z^2)^{2m+3/2}} + \frac{(-1)^n}{\pi\sqrt{2}} \int_0^\infty \frac{e^{-\frac12\pi z^2 t} t^{2n+\frac12}}{1+t^2}\, dt, \end{aligned}$$

where $\operatorname{Re} z^2 > 0$. Bounds for the remainders in these expansions follow from replacing $1/(1+t^2)$ by 1. Bounds for larger z-domains are obtained by using more refined estimates. It can be proved that (2.11) gives asymptotic expansions for $z \to \infty$, $|\arg z| < \pi/2$.

2.2. Chebyshev expansions

LUKE (1969) gives several Chebyshev expansions for erf(x), in which case the coefficients can be expressed in terms of Bessel functions. For example

$$e^{a^2x^2} \operatorname{erf}(ax) = \sqrt{\pi}\, e^{\frac12 a^2} \sum_{n=0}^{\infty} I_{n+\frac12}(\tfrac12 a^2) T_{2n+1}(x),$$

where $-1 \le x \le 1$ and $a \in \mathbb{C}$. Tabulated coefficients are given by LUKE (1969) for erf(x), erfc(x) and Dawson's integral (p. 323/324, vol II). SCHONFELDER (1978) gives for erf(x) and erfc(x) coefficients in Chebyshev expansions which enable computation with accuracy of about 10^{-30}. A modification of SCHONFELDER (1978) is considered in SHEPHERD & LAFRAMBOISE (1981).

In various Chebyshev expansions the (real) independent variable is transformed in order to obtain a faster convergent series. (See SCRATON (1970), LOCHER (1975) for the choice of the transformation from a theoretical point of view, while THACHER (1965) and SCHONFELDER (1978) determined experimentally the parameters of the transformation).

HUMMER (1964) has given an elegant method for the calculation of the coefficients of the expansion of Dawson's integral in Chebyshev polynomials. (The method also applies, with respect to the Chebyshev series of y(x), to

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} e^{-x^2} y(x)$$

because y(x) obeys the differential equation $y'(x) = 2xy(x)+1$.)

Let

$$F(kx) = \sum_{n=0}^{\infty} a_n(k) T_{2n+1}(x) \qquad x \in [-1,1]$$

$$a_n(k) = \frac{2}{\pi} \int_0^{\pi} F(k \cos\theta) \cos(2n+1)\theta \, d\theta .$$

Integration by parts and using the differential equation yields

$$a_n(k) = \frac{k^2}{2(2n+1)} [a_{n+1}(k) - a_{n-1}(k)], \qquad n = 1,2,\dots .$$

The formula for the coefficients is given by

$$a_n(k) = \sum_{r=0}^{n} \frac{(n+r)!}{r!(n-r)!} k^{-2r-1} \{(-1)^{r+n} - e^{-k^2}\}.$$

2.3. Continued fractions

The well-known continued fraction

$$\sqrt{\pi} e^{z^2} z^{-1} \operatorname{erfc}(z) = \frac{1|}{|z^2} + \frac{\frac{1}{2}|}{|1} + \frac{\frac{2}{2}|}{|z^2} + \frac{\frac{3}{2}|}{|1} + \frac{\frac{4}{2}|}{|z^2} + \cdots \tag{2.12}$$

as given in II.4.9 converges for Re $z > 0$. Contraction gives a more efficient fraction.

McCABE (1974) discussed continued fractions for Dawson's integral, which by transforming z can also be used for erf(z) or erfc(z).

2.4. Other expansions

STRECOK (1968) has given expansions of erf(x) based on the Poisson summation formula

$$\sum_{m=-\infty}^{\infty} e^{-K(m+T)^2} = (\pi/K)^{\frac{1}{2}} \sum_{n=-\infty}^{\infty} e^{-KT^2 + (KT+in\pi)^2/K}.$$

One of his expansions is

$$\operatorname{erf}(x) = \frac{2}{\pi}\,[x/5+\sum_{n=1}^{37} n^{-1}e^{-(n/5)^2}\sin(2nx/5)], \qquad |x| \le 5\pi/2,$$

with accuracy of about 10^{-24}.

By using the Gauss-Hermite quadrature rule it is easy to derive from (1.12)

$$(2.13) \qquad w(z) = \frac{i}{\pi} \lim_{n\to\infty} \sum_{k=1}^{n} \frac{H_k^{(n)}}{z-x_k^{(n)}}, \qquad \operatorname{Im} z > 0,$$

where $x_k^{(n)}$ and $H_k^{(n)}$ are the zeros and weight factors of the Hermite polynomials.

The trapezoidal integration rule applied on (1.12) gives for any $h > 0$

$$(2.14) \qquad w(z) = \frac{ih}{\pi} \sum_{n=-\infty}^{\infty} \frac{e^{-n^2h^2}}{z-nh} + E_h(z) \begin{array}{lll} + \; P & \text{if} & y < \pi/h \\ + \; \frac{1}{2}P & \text{if} & y = \pi/h \\ + \; 0 & \text{if} & y > \pi/h \end{array}$$

where $z = x + iy$, $P = 2e^{-z^2}/[1-\exp(-2\pi iz/h)]$ and $E_h(z) = \mathcal{O}(e^{-\pi^2/h^2})$, uniformly for $z \in \mathbb{C}$. See for instance LUKE (1969, vol II, p. 214) and MATTA & REICHEL (1971). When z is close to a multiple of h, say $z \doteq n_0h$ for some integer n_0, then a limiting process has to be used in order to compute

$$\frac{ih}{\pi}\,\frac{e^{-n_0^2h^2}}{z-n_0h} + P.$$

However, by choosing a smaller value of h, it is always possible to avoid such cases. The quadrature rule (2.13) asks for the values of $x_k^{(n)}$ and $H_k^{(n)}$, whereas (2.14) can be applied without pre-tabulated constants. The error term for

(2.14) is satisfactory and the series converges very fast. For complex values of z,(2.14) is an excellent starting point for a reliable and efficient algorithm. For application of (2.14) to functions related to the error function see MATTA & REICHEL (1971). Some expansions in this reference are subject to loss of accuracy. However by using elementary analytical operations this always can be settled.

3. ALGORITHMS AND IMPLEMENTATIONS

In this section algorithms and implementations are considered for: the error function (erf), the complementary error function (erfc) and repeated integrals (i^nerfc), the probability functions (P and Q), w of z (w), Dawson's integral (F) and the Fresnel integrals (C,S).

3.1. w(z) considered as basic module

The functions to be considered can be expressed in terms of w(z) (see (1.22)).For the important special case $z \in \mathbb{R}^+$, it follows that w(z) is needed for $z \in \mathbb{R}^+$, $z = ix$ with $x \in \mathbb{R}^+$ and $z = \sqrt{\pi}/2(1+i)x$ with $x \in \mathbb{R}^+$.

REMARK. In those cases where a subtraction occurs care has to be taken for small values of z in order to preserve sufficient relative precision.

3.2. Implementations for w(z)

In order to calculate w(z) for all $z \in \mathbb{C}$ we only need to consider $z \in Q_1$ with $Q_1 = \{z \mid \mathrm{Re}\ z \geq 0,\ \mathrm{Im}\ z \geq 0\}$, because of the symmetry relations (1.14) and (1.15).

3.2.1. The implementation of GAUTSCHI (1970a,b)

The aim of Gautschi is : *To propose a single algorithm which is uniformly effective for all complex arguments. Current practice attempts to achieve the desired economy by adopting different methods in different regions of the complex plane.*

The computational procedure is

$$(3.1) \qquad \sigma_N^{[\mu]}(z,h) = \sum_{k=0}^{N} (2h)^k U_k^{[\mu]}(z+ih)$$

with N, μ, dependent on z and h, chosen such that

$$|w(z) - \sigma_N^{[\mu]}(z,h) \;| < \varepsilon \text{ (an a priori fixed precision)}$$

where

$$U_k^{[\mu]}(z) = \frac{2}{\sqrt{\pi}} \prod_{j=-1}^{k-1} r_j^{[\mu]}(z)$$

and

$$r_\mu^{[\mu]}(z) = 0, \; r_{k-1}^{[\mu]}(z) = .5/(-iz+(k+1)r_k^{[\mu]}).$$

(h, N and μ are chosen empirically, such that the routine is as efficient as possible).

REMARKS. The source text is given in ALGOL 60 and therefore the complex arithmetic is handled by real data types and by programming the operations in real arithmetic. FORTRAN transcriptions of this code should have cleaned up this unnecessary coding, unless double precision is aimed at in FORTRAN 77; Several libraries have not done this in their single precision FORTRAN version. It is desirable and more comprehensive to use complex data types and complex arithmetic in the FORTRAN code.
The implementation is designed for 10 digits absolute accuracy. For machines with a different accuracy the implementation should be adapted. For machines with a higher accuracy, x_0, g_0, h_0 and N should be increased, while μ should be decreased; the amounts must be determined experimentally. Furthermore, it is desirable to use names for the constants: 4.29(=y_0), 5.33(=x_0), 1.6(=h_0), 6(=N_0), 23(=N_1), 9(=μ_0), 21(=μ_1), 1.128...(=$2/\sqrt{\pi}$), and to provide tabulated optimal values for (some of) these parameters for various (machines) accuracies, in order to facilitate transportability.

In Gautschi's paper it is indicated how to adapt the constants in order to meet 14 digits accuracy near the origin. Strangely enough several libraries copied just the original version of the algorithm, although they generally aim at 14 digits accuracy.
A warning in the paper is given when the algorithm is used for w(-z) via the symmetry relations, because of loss of relative precision near the zeros of w(-z); an expansion is needed for that region.
The programming should have taken into account the representation

$$\sigma_N^{[\mu]}(z,h) = \frac{2}{\sqrt{\pi}}(r_{-1}(1+(2h)r_0(1+\ldots+(2h)r_{N-2}(1+(2h)r_{N-1}))\ldots),$$

for h > 0, as a modification of formula (3.12) as given in the paper.

3.2.2. The recurrence relations of ACTON (1974)

Acton considered a.o. recurrence relations for the integrals,

$$I_n(c) = \int_0^\infty \frac{e^{-ct^2}}{(1+t^2)}\left(\frac{t^2}{1+t^2}\right)^n dt$$

$$J_n(c) = \int_0^\infty e^{-ct^2}\left(\frac{t^2}{1+t^2}\right)^n dt, \qquad \text{Re } c > 0.$$

I_0 is proportional to erfc (see 1.10); because of the relation between erfc and w (see (1.22)), we have

$$w(z) = \frac{2}{\pi} I_0(-z^2),$$

so the recurrence relations of ACTON can be used to compute w(z).

I_n and J_n obey the recurrence relations

$$(2n-1)I_{n-1} = 2n\, I_n + 2c\, J_n$$

$$J_{n-1} = I_{n-1} + J_n.$$

Starting from $(\tilde{I}_N, \tilde{J}_N) = (1,0)$, N sufficiently large, I_0 is approximated by

$$I_0^a = \frac{\tilde{I}_0}{\tilde{J}_0} J_0,$$

where $\tilde{I}_0$, $\tilde{J}_0$ are obtained from the recursion and $J_0 = .5\sqrt{\pi/c}$. In section 4.2. an explanation of ACTON's technique as well as improved starting values are given.

3.2.3. The approximation of OLDHAM (1968)

OLDHAM provides an approximation aimed at small machines for

$$x\,e^{x^2}\ \mathrm{erfc}(x) = xw(ix)\,,$$

which is not robust for x → 0^+ because of overflow.

3.3. The error function and the complementary error function: erf and erfc

3.3.1. erf and erfc for complex argument

STEGUN and ZUCKER (1981) have provided an implementation for the erroı function of a complex argument, as an extension of their earlier work – STEGUN and ZUCKER (1970) – where an implementation for real argument is giv

Their main concern was to provide accuracy over the entire domain of definition while the methods employed were selected in order to ensure efficiency, portability and ease of programming and modification. If one supplies approximate values for the maximum machine value, minimum machine value, the upper bound of the sine, cosine routine, and the upper bound to the acceptable relative error and gives the square root of π to the require number of significant figures, the detailed methods are designed to work fc computations ranging from very low precision to multi-precision. The algor: used in the first quadrant is a combination of: the Taylor series for erf, the asymptotic expansion and the even contraction of a continued fraction representation for erfc.

3.3.2. erf and erfc for real argument

3.3.2.1. The implementation of STEGUN & ZUCKER (1970)

They implemented in ANSI FORTRAN 66 in double precision the error function with parameters:

```
input  : x    - the independent variable
output : erf  - the value of the error function at x
         erfc - the value of the complementary error function at x.
```

In the routine the machine dependent parameters:

NBC - the number of binary digits in the characteristic of a floating point number;

NBM - accuracy desired or maximum number of binary digits in the mantissa of a floating point number.

The used computational procedure reads

the series (2.2) for erf x, with $0 < x \le 1$,

the even contraction of the continued fraction (2.12) for erfc x with $1 < x \le \mathrm{ULCF}$,

erf x = 1 and erfc x = 0 for x > ULCF,

the symmetry relations for $x < 0$,

where ULCF is a machine dependent constant.

The summation of the series (2.2) is terminated if the next term as an estimate of the remainder is smaller than $2^{-\mathrm{NBM}}$, the relative truncation error.

The even contraction of the continued fraction is evaluated in the forward direction (see II. 4.8); the evaluation is terminated if either

$$|1 - c_i/c_{i-1}| \le \mathrm{TOLER}$$

or

$$c_{i-1} > c_i,$$

with c_i the i-th convergent of the even contraction.

REMARKS.

- In the note on the parameters for transportable numerical software of IFIP WG-2.5 the parameter NBM is called SDIGIT. For the parameter NBC SRANGE can be used.
- We doubt the used termination criterion of the evaluation of the continued fraction because of the following. After an equivalence transformation the continued fraction reads

$$c = \mathop{\Phi}_{k=1}^{\infty} \frac{d_k}{1}, \qquad d_1 = \frac{2x}{2x^2+1}, \qquad d_k = \frac{-(2k-3)(2k-4)}{(2x^2+4(k-1)-3)(2x^2+4k-3)}.$$

The convergence behaviour resembles the continued fraction

$$c = \underset{k=1}{\overset{\infty}{\Phi}} \frac{-.25}{1} = -.5, \qquad \text{with the k-th convergent } c_k = -.5k/(k+1).$$

For the last fraction we have

$$|1 - c_k/c_{k-1}| = 1/(k^2-1), \qquad \text{the estimated truncation error,}$$

$$|1 - c_k/c| = 1/(k+1), \qquad \text{the truncation error;}$$

so

$$|1 - c_k/c_{k-1}| < |1 - c_k/c|.$$

From the above, we consider the implemented criterion

$$|1 - c_k/c_{k-1}| < \text{TOLER}$$

not correct, because it is possible that the actual truncation error exceeds TOLER:

$$|1 - c_k/c_{k-1}| < \text{TOLER} < |1 - c_k/c| .$$

(We expect that again a counter example can be obtained for the actual continued fraction as was the case for the exponential integral, where we demonstrated for the used continued fraction

$$|1 - c_{k-1}/c_k| < \text{TOLER} < |1 - c_k/c| .$$

Furthermore, it is remarkable that here the quantity

$$1 - c_k/c_{k-1}$$

is considered, while in their publication about exponential integrals $1 - c_{k-1}/c_k$ is considered. However, for practical purposes the stopping criterion may be sufficient; by introducing a factor it can become robust).

3.3.2.2. <u>Some implementations from CALGO</u>

CALGO 180, 181 (THACHER) for erf and erfc

These ALGOL 60 implementations are intended for large arguments. The algorithm used is the evaluation of the (Lagrange) continued fraction (2.12) for the complementary error function via a method due to Maehly.

This method for the evaluation of the even convergents of a continued fraction is a variant of (4.33 in II. 4).

Given the continued fraction

$$\overset{\infty}{\underset{k=1}{\Phi}} \frac{a_k}{b_k}$$

and the recursion

$$q_k = b_k q_{k-1} + a_k q_{k-2}, \; q_{-1} = 0, \; q_0 = 1$$

then the even convergents, c_{2k}, are given by

$$c_{2k} = \sum_{\ell=1}^{k} \Big(\prod_{j=1}^{2\ell-1} a_j \Big) b_{2\ell} / (q_{2\ell} q_{2\ell-2}) .$$

For the case of erf and erfc overflow was reported near x = 1, and because the use was intended for large x no attention was paid to scaling.

Because scaling, in the method due to Maehly, can be of general importance, one can introduce scaling factors $\{f_k\}$, as follows.

$$c_{2k} = \sum_{\ell=1}^{k} \Big(\prod_{j=1}^{2\ell-1} a_j^* \Big) b_{2\ell}^*$$

with

$$a_j^* = f_{j-1} f_j a_j , \qquad b_j^* = f_j b_j ,$$

$$f_j = 1/(a_j f_{j-1} + b_j) , \qquad j = 1,2,\ldots, \; f_0 = 1 .$$

This representation can be derived from the representation of the even convergents of the equivalent continued fraction

$$\overset{\infty}{\underset{k=1}{\Phi}} \frac{f_{k-1} f_k a_k}{f_k b_k}$$

where f_k is chosen such that $q_j^* = 1$, i.e. the partial denominators equal one.

CALGO 209 (IBBETSON) for normal distribution function erfc

The used algorithm in ALGOL 60 is Horner's rule for the evaluation of some peculiar polynomial approximation, while for $|x| \geq 6$ the value 1 is taken. The number of decimals in the coefficients suggest a more accurate approximation, which is misleading.

REMARK

This implementation is overruled by the more concise approximation 7.1.26 in Abramowitz and Stegun, which also yields ±7 digits accuracy. Furthermore, one could consult HENRICI (1977b), when an approximation for a pocket calculator is desired.

CALGO 272 (Mac LAREN) for normal distribution function erfc

This ALGOL 60 implementation combines the Taylor series expansion of erf around suitable points for small x and the asymptotic expansion for erfc for large x. "Small", "large" and the required precision can be adapted by a modification of the named variables B, N and EPS, respectively.

REMARKS

Although this implementation is characterized by the parameters B, N and EPS, we consider it not worthwhile to implement this procedure on a machine with a larger machine precision than 2_{10}-8, especially in view of IMSL, NAG and published coefficients of rational approximations.

CALGO 304 (HILL & JOYCE) for normal curve integral erfc

The ALGOL 60 implementation combines the series expansion (2.2) of erf and the (Lagrange) continued fraction (2.12) of erfc. The continued fraction is evaluated by the forward algorithm (see 4.3.1. in II. 4). In the remarks the odd contraction of the (Lagrange) continued fraction is proposed as a faster algorithm. Furthermore intermediate overflow can occur and so scaling must be introduced, as proposed in a remark by Holmgren.

3.3.2.3 The representation of MATTA & REICHEL (1971)

For real argument values their series representation reduces to the trapezoidal integration rule (2.14).

3.3.2.4. The implementation in IMSL and CERN

The implementations with multiple entry points can deliver either the value of the error function, or the value of the complemented error function or the value of the normal distribution function.
The computational problem is

$$
\begin{array}{ll}
xR_{km}(x^2), & \text{for } 0.0 \le x < 0.477 \\
e^{-x^2} R_{km}(x), & \text{for } 0.477 \le x \le 4.0 \\
\frac{e^{-x^2}}{x} \{\frac{1}{\sqrt{\pi}} + \frac{1}{x^2} R_{km}(\frac{1}{x^2})\}, & \text{for } x > 4.0,
\end{array}
$$

where $R_{km}(x)$ are rational function of degree k in the numerator and m in the denominator. For negative values of the argument the problem is reduced to the above computational problem via the symmetry relations (1.13). The approximations and the coefficients for the various domains and a variety of relative precisions (down to 10^{-19}) are given in CODY (1969b).

The implementations in NUMAL

These implementations are variants of the IMSL and CERN implementations and also based on the approximations and coefficients given by CODY (1969b). The difference concerns the language - ALGOL 60 - and that an auxiliary procedure is provided, which delivers the intermediate result

$$e^{x^2} \operatorname{erfc}(x).$$

3.3.2.5. The implementations in NAG

The implementation S15AE delivers erf(x). The computational problem is

$$
\begin{array}{lll}
x\sum_k{}' a_k T_k(t), \quad t = T_2(x/2), & & \text{for } |x| \le 2 \\
\operatorname{sgn} x \; \{1 - \frac{e^{-x^2}}{|x|\sqrt{\pi}} \sum_k{}' b_k T_k(t)\}, & t = \frac{x-7}{x-3}, & \text{for } 2 < |x| < \text{xhi} \\
\operatorname{sgn} x, & & \text{for xhi} \le |x|.
\end{array}
$$

The implementation S15AD delivers erfc(x). The computational problem is

$$\begin{array}{ll} e^{-x^2} \sum_{k=0}' a_k T_k(t), \quad t = (x-3.75)/(x+3.75), & \text{for } 0 \le x \le xhi \\ 0, & \text{for } xhi < x. \end{array}$$

For x < 0 the symmetry relations are used.

REMARKS

- xhi is machine dependent. The number of terms of the Chebyshev series, N, as a function of the number of decimal digits in the machine precision, d, is given by the following tables, for erf and erfc respectively.

erf	$\lvert x \rvert \le 2$	$2 \le \lvert x \rvert \le xhi$
N	8 12 15 17	8 11 15 18
d	8 12 14 18	8 12 14 18

erfc	$0 \le xhi$
N	12 18 21 23 26
d	8 12 14 16 18

- The only parameters are the argument and the error indicator; the latter is included for consistency reasons with other routines and has no meaning in these routines.
- The ALGOL 68 library provides operators apart from the routines.
- The coefficients of the Chebyshev approximations are published by SCHONFELDER (1978). For the error function for $2 \le |x| \le xhi$, different coefficients in the Möbius transformation are used in the paper and in the library up to mark 7.
- The original version of this algorithm as published by CLENSHAW c.s. (1963) is implemented in the library of the Boeing company, NEWBERY (1971). This version needs more terms in the Chebyshev approximation.

3.3.2.6. The implementation for the TEXAS INSTRUMENTS 58/59

Implemented is formula 7.1.26 of ABRAMOWITZ & STEGUN (1964, p. 299).

3.3.2.7. The implementation for the HEWLETT PACKARD 67/97

For $x \leq 3$ the series expansion (2.2) is used for erf. For $x > 3$ erf(x) is computed via the asymptotic expansion (2.5) for erfc(x). HENRICI (1977b) used for $x \leq 2$ the series expansion (2.2) and for $x > 2$ erf(x) is computed via the continued fraction for erfc(x).

3.4. The repeated integral of the error function: i^nerfc

For complex values of the argument no implementation is known to us. In GAUTSCHI (1977a) stability for the recurrence relation is reported in the forward direction for Re $z > 0$ and in the backward direction for Im $z < 0$.

3.4.1. The implementation of GAUTSCHI (1977a): i^nerfc x for $x \in \mathbb{R}$

For $x \in \mathbb{R}^+$ this implementation is an improvement in efficiency over the algorithms which use backward recurrence, based on relation (1.9), alone. The repeated integral of the error function is for $x > 0$ a (weakly) minimal solution of the recurrence relation in the forward direction, while for $x < 0$ it is a dominant solution. The minimal solution is computed for $x > 0$ whenever this can be done within the desired correct significant decimal digits specified by the user, because for small x the quotient of the dominant and minimal solution is not too large. In order to start the forward recursion the Taylor series is used. The backward recursion is elaborated via the continued fraction variant of the Miller algorithm, where the starting index is first estimated via GAUTSCHI (1961b) and refined, if necessary, by increasing the index by 10 and comparing the results with the previous results.

REMARKS

- To the authors opinion the series (2.2) should have been used instead of the Taylor series. (An experiment on the HP-97 showed that up to $x = 5$ no more terms are needed while a more accurate answer is obtained. Furthermore for larger values of x care has to be taken with the alternating series with respect to underflow and overflow as well as the stopping criterion).

- The used truncation error criterion for the termination of the summation of the Taylor series is not robust, because the series is not an absolutely monotonely decreasing series and therefore the first neglected term does not majorate in absolute value the remainder. Alternating series with in absolute value monotonically decreasing terms, can be obtained by the repeated expansion as used in CALGO 123.
- The repeated integral of the error function can be expressed as a confluent hypergeometric function (see below 1.9), and calculated by the algorithms given in TEMME (1983). For small positive values of the argument x, efficiency is obtained by the use of asymptotic expansions. The used backward recurrence relation has a larger domain of stability and is a modification of (1.9) by introducing the derivative. For the special case of the repeated integral of the error function this is not necessary.

3.5. The probability functions: P and Q

These functions are strongly related to the error function. In literature Q is also called: *normalarea* (BERGSON (1966)) and *normaltail* (ADAMS (1969)). These implementations are overruled by respectively HILL & JOYCE (1967) and CODY's work (1969b) and the entailed implementations e.g. in IMSL, GAUTSCHI's (1970a) implementation of w(z) and the formula (1.22) provide a general, modular starting point for an implementation. On the other hand the algorithms and implementations in TEMME (1983) for the confluent hypergeometric function could be used as basis.

In NUMAL the relations between P and Q and erf and erfc is indicated in the documentation, and so the calculation is referred to the implementation of erf and erfc.

In NAG explicit routine headings are provided for P and Q (S15AB and S15AC) which call erfc (S15AD).

3.6. Dawson's integral: F

The material given below is strongly related to the error function, namely, $F(z) = ie^{-z^2}\mathrm{erf}(-z)$. Although in (1.22) the relation between w and F is given, it is advised to circumvent 'the subtraction' only once, while the above formula couples F and erf.

The computation of F(z) can be based on various series representations for the error function (see (2.1), (2.2)). Besides these series representations various continued fractions are used in literature:

- the continued fraction associated with the Taylor series:

$$\frac{z}{1+}\,\frac{2z^2}{3-}\,\frac{4z^2}{5+}\,\frac{6z^2}{7-}\ \ldots;$$

- the asymptotic series:

$$\frac{1}{2z}+\frac{1}{2^2z^3}+\frac{1}{2^3z^5}+\ \ldots;$$

- the continued fraction associated with the asymptotic series;

$$\frac{1}{2z-}\,\frac{2}{2z-}\,\frac{4}{2z-}\,\frac{6}{2z-}\ \ldots;$$

- the continued fraction

$$\frac{z}{1+2z^2-}\,\frac{4z^2}{3+2z^2-}\,\frac{8z^2}{5+2z^2-}\,\frac{12z^2}{7+2z^2-}\ldots\,.$$

(The last fraction converges reasonably fast for small as well as large argument values $z \in \mathbb{R}^+$. Truncation error estimates were provided by MC CABE (1974); references for the truncation error for the other continued fractions are mentioned). DIJKSTRA (1977) considered a similar continued fraction for the generalization of Dawson's integral. (Truncation errors were derived for $z \in \mathbb{C}$ and estimated for $z \in \mathbb{R}^+$ in the paper; the published continued fraction restricted to Dawson's integral is equivalent to Mc Cabe's version).

Furthermore, the following representations are mentioned in literature.

- the recurrence relation of ACTON (1974), see section 3.2.2;
- the series representations of MATTA & REICHEL (1971), see (2.9);
- the Chebyshev expansions of LUKE (1969, chapter IX).

3.6.1. Dawson's integral for complex argument: F(z)

No specific implementation is known to us.

3.6.1.1. The representation of MATTA & REICHEL (1971)

Matta & Reichel represent $F(z)$, $z = x + iy$, by

$$\frac{\sqrt{\pi}}{2}\{ie^{-z^2} + K(y,x) - i\,H(y,x)\}$$

where the series representations of H and K are given in the paper. This representation is based on (2.14).

3.6.2. Dawson's integral for real arguments: F(x)

3.6.2.1. The implementations in FUNPACK, IMSL and CERNLIB

These implementations are based on CODY c.s. (1970). The computational problem for F(x) is:

$$\begin{array}{ll} x\,R_{\ell m}(x^2), & |x| \le 2.5 \\ \frac{1}{x}\,R_{\ell m}(x^{-2}), & 2.5 \le |x| \le 3.5 \;\&\; 3.5 \le |x| \le 5 \\ \frac{1}{2x}\,\{1 + x^{-2}R_{\ell m}(x^{-2})\}, & 5 \le |x|, \end{array}$$

With $R_{\ell m}$ rational functions with numerator of degree ℓ and denominator m. For an accuracy of 15 digits, the degree of the polynomials in the rational functions are: $(\ell,m) = (8,8)$, $(7,7)$, $(7,7)$ and $(6,6)$, respectively.

REMARK

The coefficients of the equivalent Jacobi fraction are published, except for $|x| \le 2.5$.

3.6.2.2. The implementation in NAG

The computational problem is

$$\begin{array}{lll} x\,\sum{}' a_k T_k(t), & t = T_2(x/4), & |x| \le 4 \\ x^{-1}\sum{}' b_k\, T_k(t), & t = T_2(4/x), & 4 \le |x|. \end{array}$$

For an accuracy of 15 digits the degree of the polynomials are 28 and 24, respectively. The Chebyshev sums are represented as power sums in the routines for efficiency reasons.

REMARKS

- The coefficients could have been derived exactly, by the technique of HUMMER (1964), which comes down to partial integration of the integral representation of the coefficients and substitution of $F'(x) = 1-2xF(x)$.
- LUKE (1969) has published coefficients of the expansion in odd degree Chebyshev polynomials with transition point $x = 3$.

3.6.2.3. The representation of MATTA & REICHEL (1971)

For real argument values their representations reduce to the trapezoidal integration rule (2.14).

3.7. The Fresnel integrals: C and S

These functions are closely related to $w(z)$ as indicated in (1.22). The separate functions can be expressed in terms of w by writing

$$\begin{pmatrix} C(z) \\ S(z) \end{pmatrix} = \tfrac{1}{2}\left\{\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \xi^2 \begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix} \begin{pmatrix} A(\xi) \\ B(\xi) \end{pmatrix}\right\}$$

with

$$A(\xi) = (w(i\xi)-w(\xi))/2$$
$$B(\xi) = (w(i\xi)+w(\xi))/2$$

and

$$\xi = \sqrt{\pi}/2(1+i)z.$$

(This relation can be derived from (1.22) and the symmetry relations

$$C(iz) = iC(z), \qquad S(iz) = -iS(z)).$$

REMARK

The subtraction of $w(i\xi)$ and $w(\xi)$ has to be handled with care.

On the other hand for large values of z the formulas (1.20) can be used, where $f(z)$ and $g(z)$ need to be evaluated. These functions can be evaluated by:

- the asymptotic expansions (2.11);
- the recurrence relations of ACTON (1974), especially for $z \in \mathbb{R}$;
- rational approximations of CODY (1968), for $z \in \mathbb{R}$;
- the representation given in MATTA & REICHEL (1971).

3.7.1. The Fresnel integrals for complex argument: C(z) and S(z)

No general implementation is known to us; so far no practical demand seems to exist.

3.7.2. The Fresnel integrals for real arguments: C(x) and S(x)

3.7.2.1. The implementation in NUMAL

This implementation is based upon CODY (1968). The computational problem is

$$C(x) \simeq x\,R_{\ell m}(x^4), \qquad |x| \le 1.2 \quad \text{and} \quad 1.2 \le |x| \le 1.6$$
$$S(x) \simeq x^3 R_{\ell m}(x^4), \qquad |x| \le 1.2 \quad \text{and} \quad 1.2 \le |x| \le 1.6$$

and for the remaining argument values thc representation (1.20) is used, where the functions f and g are approximated by

$$\begin{aligned} f(x) &\simeq R_{\ell m}(x^{-4})/x, && \text{for } 1.6 \le |x| \le 1.9 \text{ and } 1.9 \le |x| \le 2.4 \\ &\simeq (1/\pi + R_{\ell m}(x^{-4})/x^4)/x, && \text{for } 2.4 \le |x| \\ g(x) &\simeq R_{\ell m}(x^{-4})/x^3, && \text{for } 1.6 \le |x| \le 1.9 \text{ and } 1.9 \le |x| \le 2.4 \\ &\simeq (1/\pi^2 + R_{\ell m}(x^{-4})/x^4)/x^3, && \text{for } 2.4 \le |x|. \end{aligned}$$

For an accuracy of 15 digits, the degree of the polynomials in the rational functions are given in the following table

	$\|x\| < 1.2$	$1.2 \le \|x\| \le 1.6$		$1.6 \le \|x\| \le 1.9$	$1.9 \le \|x\| \le 2.4$	$2.4 \le \|x\|$
C ℓ	4	5	f ℓ	4	5	5
m	4	5	m	4	5	5
S ℓ	4	5	g ℓ	5	5	6
m	4	5	m	5	5	6

3.7.2.2. The implementations in NAG and CERNLIB

These implementations are based upon BULIRSCH (1967), and NEMETH (1965) for the recurrence relations for the coefficients. The computational problem for $x^2 < 9$ is:

$$C(x) \simeq x\sum c_r T_{2r}(x^2/9)$$
$$S(x) \simeq \frac{x^3}{9}\sum c_r T_{2r}(x^2/9).$$

For the remaining argument values representation (1.20) is used, with

$$f(x) \simeq \sum c_r T_{2r}(9/x^2)/x$$

$$g(x) \simeq (9/x^3)\sum d_r T_{2r}(9/x^2).$$

For an accuracy of 15 digits the degree of the polynomials are given in the following table.

	$x^2 < 9$		$x^2 \ge 9$
degree for C	21	degree for f	14
degree for S	20	degree for g	17

3.7.2.3. The recurrence relations of ACTON (1974) for f and g

Acton considered, among others, recurrence relations for the integrals

$$I_n(c) = \int_0^\infty \frac{e^{-ct}\sqrt{t}}{1+t^2}\left(\frac{t^2}{1+t^2}\right)^n dt$$

$$J_n(c) = \int_0^\infty \frac{e^{-ct}}{(1+t^2)\sqrt{t}}\left(\frac{t^2}{1+t^2}\right)^n dt$$

$$K_n(c) = \int_0^\infty \frac{e^{-ct}}{\sqrt{t}}\left(\frac{t^2}{1+t^2}\right)^n dt.$$

Because

$$f(x) = J_0(-\pi/2\ x^2)/(\pi\sqrt{2})\ ,$$

$$g(x) = I_0(-\pi/2\ x^2)/(\pi\sqrt{2})\ ,$$

the recurrence relations as given by Acton, can be used to compute f and g, and hence C and S. In section 4.2 an explanation of Acton's technique is given

3.7.2.4. <u>The representation of MATTA & REICHEL (1971) for f and g</u>

Matta & Reichel represent f and g as defined in (1.20) by

$$f(x) = \sqrt{2}/(2\pi x)\ \{\ H(x\sqrt{\pi}/2)\ +\ K(x\sqrt{\pi}/2)\}$$

$$g(x) = \sqrt{2}/(2\pi x)\ \{\ H(x\sqrt{\pi}/2)\ -\ K(x\sqrt{\pi}/2)\}$$

with series representations for H and K given in their paper.

<u>REMARK</u>

Cancellation occurs if first H and K are evaluated and then their sum and difference; one must first represent the sum and difference of H and K analytically as infinite series.

4. SOME ASPECTS OF ERROR ANALYSIS

In this section we consider: the effect of perturbation of the argument, the reccurence relations of Acton.

4.1. The effect of perturbation of the argument

The relative error amplification of a function f due to a perturbation Δz of the argument z, is defined by

$$\frac{|\ f(z+\Delta z)\ -\ f(z)\ |\ /\ |\ f(z)\ |}{|\ \Delta z/z\ |},\quad z \neq 0,\ f(z) \neq 0.$$

In the following the relative error amplification is given for the functions related to the error function.

function	first derivative	relative error amplification
$\mathrm{erf}(z)$	$2/\sqrt{\pi}\ e^{-z^2}$	$\lvert z/(e^{z^2}\int_0^z e^{-t^2}dt)\rvert$
$\mathrm{erfc}(z)$	$-2/\sqrt{\pi}\ e^{-z^2}$	$\lvert z/(e^{z^2}\int_z^\infty e^{-t^2}dt)\rvert$
$i^n\mathrm{erfc}(z)$	$-i^{n-1}\mathrm{erfc}(z)$	$\lvert zi^{n-1}\mathrm{erfc}(z)/i^n\mathrm{erfc}(z)\rvert$
$P(z)$	$1/\sqrt{2\pi}\ e^{-z^2/2}$	$\lvert z/(\sqrt{2}\ e^{z^2/2}(\int_0^{z/2} e^{-t^2}dt+\sqrt{\pi/2}))\rvert$
$Q(z)$	$-1/\sqrt{2\pi}\ e^{-z^2/2}$	$\lvert z/(\sqrt{2}\ e^{z^2/2}(\int_0^{z/2} e^{-t^2}dt-\sqrt{\pi/2}))\rvert$
$F(z)$	$1\ -\ 2z\,F(z)$	$\lvert z/F(z)\ -\ 2z^2\rvert$
$C(z)$	$\cos\frac{\pi}{2}z^2$	$\lvert z\cos\frac{\pi}{2}z^2/\ C(z)\rvert$
$S(z)$	$\sin\frac{\pi}{2}z^2$	$\lvert z\sin\frac{\pi}{2}z^2/S(z)\rvert$

4.2. The recurrence relations of ACTON (1974)

Acton considered the class of integrals

$$F(c)\ =\ \int_0^\infty \frac{\exp(-cX)}{t^{k/2}(1+Y)}dt,$$

with $X = t$ or t^2, $Y = t$ or t^2, $k = 0, \pm 1$, $c > 0$. For the calculation of these integrals via recursion the class of integrals is extended to

$$F_n(c)\ =\int_0^\infty \frac{\exp(-cX)}{t^{k/2}(1+Y)}\left(\frac{Y}{1+Y}\right)^n dt.$$

Moreover, the sum function (and sometimes repeated sum function)

$$G_n(c) = \int_0^\infty \frac{\exp(-cX)}{t^{k/2}} \left(\frac{Y}{1+Y}\right)^n dt$$

is used, with $G_0(c)$ known. Note that

$$F_n(c) = G_n(c) - G_{n+1}(c)$$

or

$$G_n(c) = G_0(c) - \textstyle\sum_{k=0}^{n-1} F_k(c).$$

From these functions a homogeneous matrix vector recursion

$$(4.1) \qquad y(k-1) = A(k)y(k), \quad y(k) = (y_1(k), \ldots, y_p(k))$$

is derived, with $A(k) \geq 0$. The order p equals 2,3 or 4. $F(c)$ equals $y_1(0)$ (unknown) and $y_m(0)$ is given for a certain value of m, $1 \leq m \leq p$.

4.2.1. The Miller/Acton algorithm

The M/A algorithm reads:

- take n sufficiently large; rule of thumb $n = [150/c]$,
- take as starting vector a unit vector as given in the paper of Acton,
- calculate $\tilde{y}(0)$ via the recursion,
- calculate $y_i(0) = \tilde{y}_i(0)\, y_m(0)/\tilde{y}_m(0)$ for the desired components.

In the following we will explain the algorithm where the assumptions will be stated explicitly.

Suppose,

- the eigensystem of $B = \Pi_{k=\ell+1}^{n} A(k)$ is given by $\{E,\Lambda\}$, i.e. $BE = E\Lambda$ and each eigenvector is scaled such that the i-th component equals one,
- $y(n) = Ev$,

then for every r-th component of the vector $y(\ell)$ we have

$$(4.2) \qquad y_r(\ell) = (By(n))_r = (BEv)_r = E_{rj}\Lambda_{jj}v_j(1+\varepsilon_r)$$

with

$$\varepsilon_r = (\sum_{k\neq j} E_{rk}\Lambda_{kk}v_k)/(E_{rj}\Lambda_{jj}v_j),$$

and j free for choice. From (4.2) we obtain

$$(4.3) \qquad y_i(\ell) = y_m(\ell)\,\frac{E_{ij}(1+\varepsilon_i)}{E_{mj}(1+\varepsilon_m)}\,.$$

Suppose furthermore, that for an arbitrary starting vector w a vector s exists with the property

$$w = Es,$$

then

$$(Bw)_r = E_{rj}\Lambda_{jj}s_j(1+\delta_r)$$

with

$$\delta_r = \Big(\sum_{k\neq j} E_{rk}\Lambda_{kk}s_k\Big)/(E_{rj}\Lambda_{jj}s_j)\,.$$

Substitution of

$$\frac{E_{ij}}{E_{mj}} = \frac{(Bw)_i(1+\delta_m)}{(Bw)_m(1+\delta_i)}$$

in (4.3) yields

$$(4.4) \qquad y_i(\ell) = y_i^a(\ell)\;\frac{(1+\varepsilon_i)(1+\delta_m)}{(1+\varepsilon_m)(1+\delta_i)} \quad , \text{ with } y_i^a(\ell) = y_m(\ell)\,\frac{(Bw)_i}{(Bw)_m}\,.$$

By assuming ε_i, ε_m, δ_i, δ_m negligible and taking $\ell = 0$ and w a unit vector the M/A algorithm is obtained from (4.4) with $\widetilde{y}(0) = Bw$, i.e., $y_i^a(\ell)$.

REMARKS

- $\delta = 0$ if $w = E_j$, the dominant eigenvector of B.
- $\varepsilon = 0$ if $y(n) = E_j$, the dominant eigenvector of B.

4.2.2. Estimating the starting index

If we start with $w = E_j$ then the relative truncation error of the M/A algorithm is in first order given by

$$\frac{|\,y_i(0) - y_i^a(0)\,|}{|\,y_i(0)\,|} = |\,\varepsilon_i + \varepsilon_m\,|.$$

Let for each A(k) the eigensystem be given by $\{E(k), \Lambda(k)\}$ with E(k) nonsingular, and let M(k) be defined by

$$E(k)M(k) = E(k+1)$$

then

$$B = E(1) \prod_{k=1}^{n} (\Lambda(k)M(k))E^{-1}(n+1).$$

Furthermore, if $M(k) = I$ in first order (a so-called slowly varying recursion) then the eigensystem of B is given by $\{E(p), \Lambda(k)\}$, $p \in [1,n]$. If we furthermore assume that the magnitude of ε_i is determined by the subdominant eigenvector with index s say, and $v_s/v_j = 1$ then

$$\varepsilon_i = \prod_{k=1}^{n} \Lambda(k)_{ss}/\Lambda(k)_{jj}.$$

The truncation error is governed by ε_i and ε_m; if we concentrate on ε_i then the starting index can be estimated, given a desired relative accuracy eps, by

$$(4.5) \qquad \{n \mid \prod_{k=1}^{n} \mid \Lambda(k)_{ss}/\Lambda(k)_{jj} \mid \le \text{eps} < \prod_{k=1}^{n-1} \mid \Lambda(k)_{ss}/\Lambda(k)_{jj} \mid\}.$$

<u>REMARKS</u>

- Although the above theory orginated from the desire to understand and to apply Acton's recurrence relations for the parameter $c \in \mathbb{C}$, Re $c > 0$, it is possibly of use in other matrix vector relations, such as those which result from three-term recurrence relations by considering the companion matrix, or the matrix vector recursions for a function and its derivative, e.g. as used by TEMME (1983) for the calculation of the confluent hypergeometric function.
- Acton proposed to start with a unit vector; we propose to start with $E_j(n)$ the dominant eigenvector of $A(n)$.
- Intermediate scaling must be implemented to circumvent overflow. For a matrix of order two a direct formulation, with implicit scaling, can be obtained by the recurrence relation for the quotients $y_1(k)/y_2(k)$, where $y_1(0)$ is to be calculated.

5. TABULATED COEFFICIENTS

For a more complete list, especially through 1969, see LUKE (1969).

BULIRSCH (1967)

C(x), S(x) coefficients of the Chebyshev expansions up to 17d on $x^2 < 9$, $9 \leq x^2$.

CLENSHAW, MILLER, WOODGER (1963)

erf(x) coefficients of the Chebyshev expansions up to 16d on $x^2 < 16$.

erfc(x) coefficients of the Chebyshev expansions up to 19d on $x \geq 4$.

(In CLENSHAW (1962) the above coefficients are provided for the same intervals up to 20d).

CODY (1968)

C(x), S(x) rational approximations up to 19d on [0,1.2], [1.2,1.6], [1.6,1.9], [1.9,2.4] and $[2.4,\infty)$.

CODY (1969b)

erf(x), erfc(x) rational approximations up to 22d on [0,.5], [.46875,4.0] and $[4,\infty)$.

CODY, PACIOREK, THACHER (1970)

F(x) rational approximations up to 22d on [0,2.5], [2.5,3.5], [3.5,5], $[5,\infty)$.

HART c.s. (1968)

erfc(x) various rational approximations.

HUMMER (1964)

F(x) coefficients of the Chebyshev expansions up to 16d on [0,5]; recurrence relations for the coefficients are also provided.

LUKE (1969)

The Chebyshev expansion of erf(ax) and the Fresnel integrals is given in chapter IX, 9.3.

Coefficients of the Chebyshev expansion up to 20d are given for: $\sqrt{\pi}/2\mathrm{erf}(x)$, $|x| < 3$ (table 22); $e^{x^2}F(x)$, $|x| < 3$ (table 22) and $x > 3$ (table 23); $\sqrt{\pi}/2\mathrm{erfc}(x)$, $x \geq 3$ (table 23); $\sqrt{2\pi}\,C(\sqrt{2/\pi}x)$, $0 \leq x \leq 8$ (table 24); $\sqrt{2\pi}\,S(\sqrt{2/\pi}x)$, $0 \leq x \leq 8$ (table 24).

Main diagonal Padé approximations up to 22d and the approximation errors for z = 1, i, - 1, are given for erf(z), $e^{z^2}F(z)$, C(z) + iS(z) in table 65.1 and 65.2.

NEMETH (1965)

C(x), S(x) coefficients of the Chebyshev expansions up to 12d on [0,8]; recurrence relations for the coefficients are also provided.

SCHONFELDER (1978)

erf(x), erfc(x) coefficients of the Chebyshev expansion up to 30d on [0,4], [4,∞).

SHEPHERD & LAFRAMBOISE (1981)

erf(x) coefficients of the Chebyshev expansion up to 22d on [0,∞).

6. TESTING

The testing with respect to accuracy of algorithm 363, w(z) implemented by GAUTSCHI (1970b) in ALGOL 60, consisted of:

- a comparison with a 14d implementation of the same algorithm,
- a comparison with tabulated values.

In both cases the aimed 10 decimal accuray has been obtained. In the certification of algorithm 363 Kölbig performed more elaborate tests on a FORTRAN transcription of the algorithm extended to the complex plane.

STEGUN & ZUCKER (1970) compared their implementation of erf with those obtained by using various polynomials or rational approximations. Furthermore, special values were checked by asymptotic expansions and numerical integration; moreover, single precision results were checked against double precision results. They reported that in all cases the obtained accuracy agreed within NBM-(I+3) binary digits, where I is the number of binary digits representing the integer part of x^2. (Because of the reduction of the argument x^2 of the exponential function I binary digits are 'lost').

The correctness of the coefficients for functions related to the error function as published by CODY (1968,1969b) and CODY, PACIOREK & THACHER (1970) was verified by comparison of subroutines based on the coefficients and the 'master routines' using 5000 pseudo-random arguments.

The NAG implementations have been tested with automatic portable test software (see introduction).

GAUTSCHI (1977) compared his FORTRAN implementation with:

- a mixed-precision variant (single and double),
- numerical tables,
- numerical integration,
- results from various asymptotic expansions.

7. APPLICATIONS

We mention a few examples only. Applications in statistics and in physics are abundant. For a list of integrals of error functions see NG & GELLER (1969).

7.1. Inverse error function

Implementations are available in IMSL, NUMAL and CERNLIB. In principle an implementation for erf, erfc and a zerofinder will do, because

$$\text{inverf:} \qquad y \to \{x \mid y = \text{erf } x\}.$$

This general approach has the following pitfalls

a. For y away from zero - x large - the problem is sensitive for perturbations in y, because

$$\frac{dx}{dy} = \frac{\sqrt{\pi}}{2} e^{x^2}.$$

For large x, change of the dependent variable into $\bar{y} = 1-y$ yields erfc x - $\bar{y}$ = 0. For the relative perturbation we have

$$\frac{\bar{y}}{x}\frac{dx}{d\bar{y}} = \frac{\text{erfc } x}{x}\frac{\sqrt{\pi}}{2} e^{x^2} \sim \frac{1}{2\,x^2}, \; x \to \infty.$$

Therefore, in terms of $\bar{y}$ the problem is better conditioned. To circumvent underflow one should consider (BLAIR et al. (1976))

$$\xi(-\ell n \text{ erfc } x)^{\frac{1}{2}} - 1 = 0, \; \xi = (-\ell n\,\bar{y})^{-\frac{1}{2}}.$$

b. Efficiency considerations.

In the zerofinder use of the cheap derivative must be made. Therefore the Newton-Raphson algorithm or another routine where use is made of derivatives must be considered. Furthermore, the problem can be approximated by considering a truncated asymptotic series or continued fraction approximation of erfc x. On the other hand if erfc is approximated by a series expansion,a priori inversion of the series can be considered:

$$\text{erfc}\ x - \bar{y} = 0 \longrightarrow P_n(x) \rightarrow x = Q_m(\bar{y}),$$

where $Q_m(\bar{y})$ can be represented in terms of Chebyshev polynomials (see STRECOK (1968)). In IMSL the inverse function inverf is implemented via near minimax rational function approximations. BLAIR et al. (1976) have published the coefficients, in ± 23d, of the rational approximation and the algorithm by which these coefficients were obtained

REMARK

The implementation in CERNLIB did not circumvent pitfall a), so for large x values the algorithm is ill-conditioned.

7.2. Cornu's spiral

In the theory of Fresnel diffraction in optics the intensity of the light behind a slit is governed by Cornu's spiral. The coordinates of any point (x,y) on Cornu's spiral are given by the Fresnel integrals

$$x = \int_0^v \cos \pi t^2/2\ dt$$

$$y = \int_0^v \sin \pi t^2/2\,dt,$$

where t is the lenght along the spiral. The lenght v determines the width of the diffraction slit. The intensity of the diffracted light is equal to $x^2 + y^2$, which needs the evaluation of the integrals.

7.3. Integrals in terms of the error or related functions

Sometimes it needs some skill to recognize integrals. In Abramowitz and Stegun the integral

$$\int_0^\infty e^{-xt}\sin(t^2)/t\,dt, \qquad x > 0$$

is expressed in C and S as

$$\pi/2((.5-C(y))^2 + (.5-S(y))^2), \qquad y = x/\sqrt{2\pi}.$$

However, for large values of x cancellation occurs, because C(x) and S(x) behave as $.5 + 0(1/x)$ for large x. After some manipulation the integral can be expressed in the function f and g, as given in (1.20), as

$$\pi/2(f^2(y)+g^2(y)).$$

If only the function w is available the last expression is easily expressed in terms of w via (1.22).

7.4. Error integral in computing literature

In FORSYTHE, MALCOLM & MOLER (1977) the exercises of each chapter start with a problem related to the error function.

REFERENCES

ABRAMOWITZ, M. (1954), *On the practical evaluation of integrals,* 175-190 in: Ph.J. Davis and Ph.Rabinowitz (1967), Numerical Integration, Blaisdell.

ABRAMOWITZ, M., STEGUN, I.A. (1964), *Handbook of mathematical functions with formulas, graphs and mathematical tables,* Nat. Bur. Standards Appl. Series, 55, U.S. Government Printing Office, Washington, D.C.

ACTON, F.S. (1970), *Numerical methods that work,* Harper & Row.

ACTON, F.S. (1974), *Recurrence relations for the Fresnel integral* $\int_0^\infty t^{-\frac{1}{2}}(1+t^2)^{-1}\exp(-ct)dt$ *and similar integrals,* Comm. ACM **17**, 480-481.

ADAMS, A.G. (1969), Algorithm 39, *Areas under the normal curve,* Computer J. **12**, 197-198.

AMOS, D.E. (1980), *Computation of exponential integrals,* ACM Trans. Math. Softw. **6**, 365-377.

AMOS, D.E. (1983a), Algorithm 609, *A portable FORTRAN subroutine for the Bickley functions* $Ki_n(x)$, ACM Trans. Math. Softw. **9**, 480-493.

AMOS, D.E. (1983b), Algorithm 610, *A portable FORTRAN subroutine for the derivation of the Psi function,* ACM Trans. Math. Softw. **9**, 494-502.

ANTONINO, M.S.F., G. SCHWACHHEIM (1967), *Gamma function with arbitrary precision,* CALGO 309, Comm. ACM **10**, 511.

ARSCOTT, F.M. (1981), *The land beyond Bessel: A surey of higher special functions,* 26-45 in: Ordinary and partial differential equations, W.N. Everitt and B.D. Sleeman (eds.), Lecture notes in mathematics 846, Springer Verlag.

ARTIN, E. (1964), *The gamma function,* Holt, Rinehart and Winston.

BABUSKA, I. (1972), *Numerical stability in problems of linear algebra,* SIAM J. Numer. Anal. **9**, 53-77.

BAKER, G.A. (1975), *Essentials of Padé approximants,* Academic Press.

BAUER, F.L. (1973), *Software and software engineering,* SIAM Rev. **15**, 469-480.

BAUER, F.L. (1974), *Computational graphs and rounding error,* SIAM J. Numer. Anal. **11**, 87-96.

BAUER, F.L. (1980), *A trend for the next ten years of software engineering,* 1-25 in: H. Freeman, P.M. Lewis (eds.), Software Engineering, Academic Press.

BEAM, A. (1960), *Complex exponential integral,* CALGO 14, Comm. ACM **3**, 406.

BERG, L. (1977), *Zur Abschätzung des Restgliedes in der asymptotischen Entwicklung des Exponential integrals,* Computing **18**, 361-363.

BERGSON, A. (1966), Algorithm 13, *NORMAL AREA.* Computer J. **9**, 322-323.

BLAIR, J.M., C.A. EDWARDS, J.H. JOHNSON (1976), *Rational Chebyshev approximations for the inverse of the error function,* Math. Comp. **30**, 827-830.

BLANCH, G. (1964), *Numerical evaluation of continued fractions,* SIAM Rev. **6**, 383-421.

BOEHM, B.W. et al. (1978), *Characteristics of software quality,* North-Holland Publishing Company.

BOEHM, B.W. (1981), *Software engineering economics,* Prentice Hall.

BOURBAKI, N. (1951), *Eléments de mathématique,* Livre IV, Ch. VII, *La fonction gamma,* Paris.

BRENT, R.P. (1980), *Unrestricted algorithms for elementary and special functions,* 613-619 in: S.H. Lavington (ed.), Information Processing 80, North Holland Publishing Company.

BREZINSKI, C. (1976), *A bibliography of Padé approximation and some related matters,* 245-267 in CABANNES (1976).

BULIRSCH, R. (1967), *Numerical calculation of the sine, cosine and Fresnel integrals,* Numer. Math. **9**, 380-385.

BULIRSCH, R., J. STOER (1968), *Darstellung von Funktionen in Rechenautomaten,* 352-446 in: R. Sauer, I. Szabó, (eds.), *Mathematische Hilfsmittel des Ingenieurs* III, Springer Verlag.

CHAR, B. (1980), *On Stieltjes' continued fraction for the gamma function,* Math. of Comp. **34**, 547-551.

CABANNES, H. (ed.) (1976), *Padé approximants method and its applications to mechanics,* Lecture Notes in Physics 47, Springer Verlag.

CHIARELLA C., A. REICHEL (1968), *On the evaluation of integrals related to the error function,* Math. Comp. **22,** 137-143.

CLENSHAW, C.W. (1962), *Chebyshev series for mathematical functions,* Nat. Phys. Lab. Math. tables, Vol. V, H.M. Stationary Office, London.

CLENSHAW, C.W., G.F. MILLER, M. WOODGER (1963), *Algorithms for special functions I,* Numer. Math. **4,** 403-419.

CLENSHAW, C.W., S.M. PICKEN (1966) *Chebyshev series for Bessel functions of fractional order,* Nat. Phys. Lab. Math. Tables, Vol. VIII, H.M. Stationary Office, London.

CLENSHAW, C.W., F.W.J. OLVER (1980), *An unrestricted algorithm for the exponential function,* SIAM J. Numer. Anal. **17,** 310-331.

CODY, W.J. (1968), *Chebyshev approximations for the Fresnel integrals,* Math. Comp. **22,** 450-453.

CODY, W.J. (1969a), *Performance testing of function subroutines,* 759-763 in: Proc. Spring Joint Computer, Conf., **34** AFIPS Press, Montvale, N.J.

CODY, W.J. (1969b), *Rational Chebyshev approximations for the error function,* Math. Comp. **23,** 631-637.

CODY, W.J. (1970), *A survey of practical rational and polynomial approximation of functions,* SIAM Rev. **12,** 400-423.

CODY, W.J. (1973), *The evaluation of mathematical software,* 121-135 in: *Program test methods* (W.C. Hetzel, ed.), Proc. of symposium on computer program test methods, Prentice Hall.

CODY, W.J. (1974), *The construction of numerical subroutine libraries,* SIAM Rev. **16,** 30-46.

CODY, W.J. (1975a), *The FUNPACK package of special functions subroutines,* ACM Trans. Math. Softw. **1,** 13-25.

CODY, W.J. (1975b), *An overview for software development for special functions,* 38-48 in: G.A. Watson (ed.), *Numerical Analysis,* 506, Springer Verlag.

CODY, W.J. (1980a), *Basic concepts for computational software,* 1-23 in: P.C. Messina and A. Murli (eds.): Problems and Methodologies in Mathematical Software production, Lecture Notes in Computer Science 142, Springer Verlag.

CODY, W.J. (1980b), *Implementation and testing of function software,* 24-47 in: P.C. Messina and A. Murli (eds.), Problems and Methodologies in Mathematical Software Production, Lecture Notes in Computer Science 142, Springer Verlag.

CODY, W.J. (1981), *FUNPACK - a package of special function subroutines,* TM-385 Applied Mathematics Division, Argonne National Laboratory.

CODY, W.J., K.J. HILLSTRÖM (1967), *Chebyshev approximations for the natural logarithm of the gamma function,* Math. Comp. **21,** 198-203.

CODY, W.J., K.J. HILLSTRÖM (1970), *Chebyshev approximation for the Coulomb phase shift,* Math. Comp. **24,** 671-678.

CODY, W.J., K.A. PACIOREK, H.C. THACHER Jr. (1970), *Chebyshev approximations for Dawson's integral,* Math. Comp. **24,** 171-178.

CODY, W.J., H.C. THACHER Jr. (1968), *Rational approximations for the exponential integral $E_1(x)$,* Math. Comp. **22,** 641-649.

CODY, W.J., H.C. THACHER Jr. (1969), *Chebyshev approximation for the exponential integral $Ei(x)$,* Math. Comp. **23,** 289-303.

CODY, W.J., W. WAITE (1980), *Software manual for the elementary functions,* Prentice Hall.

DAVIS, Ph.J. (1959), *Leonhard Euler's integral: a historical profile of the gamma function,* Am. Math. Monthly **66,** 849-869.

DEUFLHARD, P. (1976), *On algorithms for the summation of certain special functions,* Computing **17,** 37-48.

DEUFLHARD, P. (1977), *A summation technique for minimal solutions of linear homogeneous difference equations,* Computing **18,** 1-13.

DIJKSTRA, D. (1977), *A continued fraction expansion for a generalization of Dawson's integral,* Math. Comp. **31,** 503-510.

DINGLE, R.B. (1973), *Asymptotic expansions: their derivation and interpretapretation,* Academic Press.

DITKIN, V.A., K.A. KARPOV, M.K. KERIMOV (1981), *The computation of special functions,* USSR Comput. Maths. Math. Phys. **20,** 3-12.

EINARSSON, B. (1977), *Bibliography on numerical software,* IEEE repository, IEEE Computer Society 5855, Naples Plaza, suite 301, Long Beach, California 90803.

EINARSSON, B. (1979), *Bibliography on the evaluation of numerical software.* Journ. Comp. Appl. Math. **5,** 145-159.

FETTIS, H.E. (1974), *A stable algorithm for computing the inverse error function in the 'Tail-end' region,* Math. Comp. **28,** 585-587.

FIELD, D.A. (1977), *Estimates of the speed of convergence of continued fraction expansions of functions,* Math. Comp. **31,** 495-502.

FIELD, D.A., W.B. JONES (1972), *A priori estimates for truncation error of continued fractions* $K(1/b_n)$, Numer. Math. **19,** 283-302.

FLETCHER, A., J.C.P. MILLER, L. ROSENHEAD, L.J. COMRIE (1962), *An index of mathematical tables,* (2nd. ed.), Addison-Wesley.

FONG, K.W., T.H. JEFFERSON, T. SUYEHIRO (1984), *Formal conventions of the SLATEC Library,* SIGNUM **19** (1), 17-22.

FORD, B. (1978), *Parameterization of the environment for transportable numerical software,* ACM Trans. Math. Softw. **4,** 100-103.

FORSYTHE, G.E., M.A. MALCOLM, C.B. MOLER (1977), *Computer methods for mathematical computations,* Prentice Hall.

FOX, L., I.B. PARKER (1968), *Chebyshev polynomials in numerical analysis,* Oxford Univ. Press.

FOX, P.A., A.D. HALL, N.L. SCHRYER (1978), *The PORT mathematical subroutine library,* ACM Trans Math. Softw. **4,** 104-126.

FULLERTON, L.W. (1977), *Portable special function routines,* 452-483 in: W. Cowell (ed.): *Portability of numerical software,* Lecture notes in computer science 57, Springer Verlag.

FULLERTON, L.W. (1980), *A bibliography on the evaluation of mathematical functions.* CSTR 86, Bell Laboratories, Murray Hill, N.J.

GAUTSCHI, W. (1959), *Exponential integral* $\int_1^\infty e^{-xt} t^{-n} dt$ *for large values of* n, J. Res. Nat. Bur. Standards, **62,** 123-125.

GAUTSCHI, W. (1961a), *Recursive computation of certain integrals,* J. ACM **8,** 21-40.

GAUTSCHI, W. (1961b), *Recursive computation of the repeated integrals of the error function,* Math. Comp. **15,** 227-232.

GAUTSCHI, W. (1964a), *Gamma function,* CALGO 221, COMM. ACM **7,** 143.

GAUTSCHI, W. (1964b), *Bessel functions of the first kind,* CALGO 236, Comm. ACM **7,** 479-480.

GAUTSCHI, W. (1967), *Computational aspects of three-term recurrence relations,* SIAM Rev. **9,** 24-82.

GAUTSCHI, W. (1968a), *Construction of Gauss-Christoffel quadrature formulas,* Math. Comp. **22,** 251-270.

GAUTSCHI, W. (1968b), *Gaussian quadrature formulas,* CALGO 331, Comm. ACM **11,** 432-436.

GAUTSCHI, W. (1970a), *Efficient computation of the complex error function,* SIAM J. Numer. Anal. **7,** 187-198.

GAUTSCHI, W. (1970b), *Complex error function,* CALGO 363, Comm. ACM **12,** 280.

GAUTSCHI, W. (1972), *Zur Numerik rekurrenter Relationen,* Computing **9,** 107-126.

GAUTSCHI, W. (1972b), *The condition of orthogonal polynomials,* Math. Comp. **26,** 923-924.

Gautschi, W. (1973), *Exponential integrals,* Calgo 471, Comm. ACM **16,** 761-763.

Gautschi, W. (1975), *Computational methods in special functions - a survey,* 1-98 in: R. Askey (ed.), Theory and applications of special functions, Academic Press.

Gautschi, W. (1977a), *Evaluation of the repeated integrals of the coerror function,* ACM Trans. Math. Softw. **3,** 240-252.

Gautschi, W. (1977b), *Repeated integrals of the coerror function,* Calgo 521, ACM Trans. Math. Softw. **3,** 301-302.

Gautschi, W. (1978), *Questions of numerical condition related to polynomials;* 45-72 in: Boor, D. de, G.H. Golub (eds.), Recent advances in numerical analysis, Academic Press.

Gautschi, W. (1979a), *The condition of polynomials in power form,* Math. Comp. **33,** 343-352.

Gautschi, W. (1979b), *A computational procedure for incomplete gamma functions,* ACM Trans. Math. Softw. **5,** 466-481.

Gautschi, W. (1983), *On the convergence behavior of continued fractions with real elements,* Math. Comp. **40,** 337-342.

Gautschi, W., J. Slavik (1978), *On the computation of modified Bessel function ratios,* Math. Comp. **32,** 685-875.

Gargantini, I., P. Henrici (1967), *A continued fraction algorithm for the computation of higher trancendental functions in the complex plane,* Math. Comp. **21,** 18-29.

Gilewicz, J. (1978), *Approximants de Padé,* Lecture notes in mathematics 667, Springer Verlag.

Gill, J. (1982), *Truncation error analysis for continued fractions* $K(a_n / 1)$, where $\sqrt{|a_n|} + \sqrt{|a_{n-1}|} < 1$, 67-70 in: Jones, Thron, Waadeland (1982).

Golub, G.H., J.H. Welsch (1969), *Calculation of Gauss quadrature rules,* Math. Comp. **23,** 221-230.

Gragg, W.B. (1968), *Truncation error bounds for g-fractions,* Numer. Math. **11,** 370-379.

Gunn, J.H. (1967), *Coulomb wave functions,* Calgo 300, Comm. ACM **10,** 244.

Hamming, R.W. (1969), *One man's view of computer science,* J. ACM, **16,** 3-12.

Hamming, R.W. (1971), *Introduction to applied numerical analysis,* McGraw-Hill.

Hamming, R.W. (1973), *Numerical methods for scientists and engineers,* McGraw-Hill.

Hart, J.F. et al. (1968), *Computer approximations,* John Wiley.

Hemker, P.W. (ed.) (1984), numal, *Numerical procedures in Algol 60,* MC Syllabus 47.1 - 47.7, Mathematisch Centrum, Amsterdam.

Henrici, P. (1974,1977a), *Applied and computational complex analysis,* John Wiley

I. *Power series, integration, conformal mapping, location of zeros.*

II. *Special functions, integral transformations, asymptotics, continued fractions.*

Henrici, P. (1977b), *Computational analysis with the HP-25 pocket calculator,* John Wiley.

Henrici, P. (1982), *Essentials of numerical analysis (with pocket calculator demonstrations), and its solution manual,* John Wiley.

Henrici, P., P. Pfluger (1966), *Truncation error estimates for Stieltjes fractions,* Numer. Math. **9,** 120-138.

Hill, I.D., S.A. Joyče (1967), *Normal curve integral,* Calgo 304, Comm. ACM **10,** 374-376.

Hochstadt, H. (1971), *The functions of mathematical physics,* John Wiley.

Householder, A.S. (1953), *Principles of numerical analysis,* McGraw-Hill.

Hummer, D.G. (1964), *Expansion of Dawson's function in a series of Chebyshev polynomials,* Math. Comp. **18,** 317-319.

Jacobsen, L., H. Waadeland (1982), *Some useful formulas involving tails of continued fractions,* 99-105 in: Jones, Thron, Waadeland (1982).

Jahnke, E., F. Emde (1945), *Tables of functions,* 4th. ed., Dover Publications.

Jefferson, T.H. (1969), *Truncation error estimates for T-fractions,* SIAM J. Numer. Anal. **6,** 359-364.

Jones, R.E., D.K. Kahaner (1983), *XERROR, the SLATEC error handling package,* Softw. Pract. and Exp. **13,** 251-257.

Jones, W.B., R.I. Snell (1969), *Truncation error bounds for continued fractions,* SIAM J. Numer. Anal. **6,** 210-221.

Jones, W.B., W.J. Thron (1971), *A posteriori bounds for the truncation error of continued fractions,* SIAM J. Numer. Anal. **8,** 693-705.

Jones, W.B., W.J. Thron (1974), *Numerical stability in evaluation of continued fractions,* Math. of Comp. **28,** 795-810.

Jones, W.B., W.J. Thron (1980), *Continued fractions: Analytic theory and applications,* Encyclopedia of mathematics and it Applications, No 11, Addison-Wesley.

Jones W.B., W.J. Thron, H. Waadeland (1982), *Analytic theory of continued fractions,* Lecture notes in Mathematics 932, Springer Verlag.

Kadlec, J. (1976), *On the evaluation of some integrals occurring in scattering problems,* Math. Comp. **30,** 263-277.

Kahan, W. (1983), *Mathematics written in sand - the HP-15c, Intel 8087, etc.,* Proceedings of American Statistical Association, 12-26, Meetings of the ASA-ENAR-WNAR-IMS-SSC, Toronto, August 1983.

Khovanskii, A.N. (1963), *The application of continued fractions and their generalizations to problems in approximation theory* (Translated by P. Wynn), Noordhoff.

Knopp, K. (1964). *Theorie und Anwendung der unendlichen Reihen,* Springer Verlag.

Kölbig, K.S. (1972), *Programs for computing the logarithm of the gamma function and the digamma function for complex argument,* Computer Phys. Comm. **4,** 221-226.

Kuki, H. (1972), *Complex gamma function with error control,* Calgo 421, Comm. ACM **15,** 262-267, 271-272.

Lanczos, C. (1964), *A precision approximation of the gamma function,* J. SIAM Numer. Anal., ser B **1,** 86-96.

Larson, J., A. Sameh (1978), *Efficient calculation of the effects of roundoff errors.* ACM Trans. Math. Softw. **4,** 228-236.

Lauwerier, H.A. (1974), *Asymptotic analysis,* MC Tract 54, Mathematical Centre, Amsterdam.

Locher, F. (1975), *Konvergenzbeschleunigung von Čebyšev-Entwicklungen,* Computing **15,** 235-246.

Lucas, C.W.Jr., C.W. Terril (1971), *Complex gamma function,* Calgo 404, Comm. ACM **14,** 48-49.

Luke, Y.L. (1969), *The special functions and their approximations* (2 Vols.), Academic Press.

Luke, Y.L. (1970), *Evaluation of the gamma function by means of Padé approximation,* SIAM J. Math. Anal. **1,** 266-281.

Luke, Y.L. (1975), *Mathematical functions and their approximations,* Academic Press.

Luke, Y.L. (1976), *On the expansion of the exponential type integrals in series of Chebyshev polynomials,* 180-200 in: Law, A.G., B.N. Sahney (eds), Theory of approximation with applications, Academic Press.

Luke, Y.L. (1977), *Algorithms for the computation of mathematical functions,* Academic Press.

Lyusternik, L.A., O.A. Chervonenkis, A.R. Yanpolskii (1965), *Handbook for computing elementary functions,* Pergamonn Press.

Matta, F., A. Reichel (1971), *Uniform computation of the error function and other related functions,* Math. Comp. **25,** 339-344.

Mattheij, R.M.M. (1975), *Accurate estimates of solutions of second order recursions,* Lin. Algebra Applics **12,** 29-34.

Mattheij, R.M.M. (1977), *Estimating and determining solutions of matrix vector recursions,* Thesis, Utrecht.

Mattheij, R.M.M. (1980), *Characterizations of dominant and dominated solutions of linear recursions,* Numer. Math. **35,** 421-442.

Mattheij, R.M.M. (1982), *Stable computation of dominated solutions of linear recursions,* BIT **22,** 79-93.

Mattheij, R.M.M., A. van der Sluis (1976), *Error estimates for Miller's algorithm,* Num. Math. **20,** 61-78.

McCabe, J.H. (1974), *A continued fraction expansion, with a truncation error estimate, for Dawson's integral,* Math. Comp. **28,** 811-816.

Medeiros, A.D. de, G. Schwachheim (1969), *Polygamma functions with arbitrary precision,* Calgo 349, Comm. ACM **12,** 213-214.

Mikloško, J., (1976), *Investigation of algorithms for numerical computation of continued fractions,* USSR Comput. Maths. Math. Phys. **16** (4), 1-12.

Mikloško, J. (1977), *A fast algorithm for repeated computation of linear recurrence relations,* BIT **17,** 430-436.

Miller, J.C.P. (1952), *Bessel functions,* Part II in: *Mathematical Tables* vol. *X*, Baas, Cambridge University Press.

Miller, W. (1975), *Software for round-off analysis I.* ACM Trans. Math. Softw. **1,** 108-128.

Miller, W., D. Spooner (1978), *Software for round-off analysis II.* ACM Trans. Math. Softw. **4,** 369-387.

Nemeth, G. (1965) *Chebyshev expansions for Fresnel integrals,* Numer. Math. **7,** 310-312.

Nemeth, G. (1967), *A Stirling-sor Csebisev sorfejtése,* Matem. Lapok **18,** 329-333.

Newbery, A.C.R. (1971), *The Boeing library and handbook of mathematical routines,* 153-169 in: Rice J.R. (ed.), Mathematical Sofware, Academic Press.

Newbery, A.C.R., A.P. Leigh (1971), *Consistency test for elementary functions,* AFIPS **39,** 419-422.

Newbery, A.C.R. (1974), *Error analysis for polynomial evaluation,* Math. Comp. **28,** 789-793.

Ng. E.W. (1975), *A comparison of computational methods and algorithms for the complex gamma function,* ACM Trans. Math. Softw. **1,** 56-70.

Ng. E.W., M. Geller (1969), *A table of integrals of the error functions,* J. Res. Nat. Bur. Standards **73B,** 1-20. For additions and corrections, see the same journal, 75B, 149-164, 1971.

Nörlund, N.E. (1924), *Vorlesungen über Differenzenrechnung,* Springer Verlag.

Oberhettinger, F., L. Badii (1973), *Tables of Laplace transforms,* Springer Verlag.

Oldham, K.B. (1968), *Approximations for the $x \exp x^2$ erfc x function,* Math. Comp. **22,** 454.

Oliver, J. (1968a), *The numerical solution of linear recurrence relations,* Num. Math. **11,** 349-360.

Oliver, J. (1968b), *An extension of Olver's error estimation technique for linear recurrence relations,* Num. Math. **12,** 459-467.

Olver, F.W.J. (1967a), *Error analysis of Miller's recurrence algorithms,* Math. Comp. **18,** 65-74.

Olver, F.W.J. (1967b), *Numerical solution of second order linear difference equations,* J. Res. NBS **71B,** 11-129.

Olver, F.W.J. (1967c), *Bounds for the solutions of second order linear difference equations,* J. Res. NBS **71B,** 161-166.

Olver, F.W.J. (1974), *Asymptotics and special functions,* Academic Press.

Olver, F.J.W. (1978), *A new approach to error arithmetic,* SIAM J. Numer. Anal. **15,** 368-393.

Olver, F.W.J., D.J. Sookne (1972), *Note on backward recurrence algorithms,* Math. Comp. **26,** 941-947.

Paciorek, K.A. (1970), *Exponential integral $E_i(x)$,* Calgo 385, Comm. ACM **13, 444-445.**

Partsch, H., R. Steinbruggen (1981), *A comprehensive survey on program transformation systems,* TUM report I 8108, München.

Perron, O. (1950), *Die Lehre van den Kettenbrüchen,* Chelsea.

Rainville, E.D. (1960), *Special functions,* Macmillan.

Rice, J.R. (1964), *On the L_∞ Walsh arrays for $\Gamma(x)$ and erfc (x),* Math. Comp. **18,** 617-626.

Rivlin, T.J. (1974), *The Chebyshev polynomials,* John Wiley.

Rutishauser, H. (1968), *Zum Problematik der Nullstellenbestimmung bei Polynomen;* 281-295 in: Dejon, W. (eds.), Constructive aspects of the fundamental theorem of algebra, Wiley-Interscience.

Rutishauser, H. (1976), *Vorlesungen über numerische Mathematik,* 2 Vols., Birkhäuser.

Sauer, R., I. Szabo (1968), *Mathematische Hilfsmittel des Ingenieurs III,* Springer.

Schaback, R., K. Scherer, (1976), *Approximation theory,* Lecture notes in mathematics 556, Springer Verlag.

Schäfke, F.W. (1965), *Lösungstypen van Differenzen und Summengleichungen in normierten Abelschen Gruppen,* Math. Z. **88,** 81-104.

Schonfelder, J.L. (1976), *The production of special function routines for a multi-machine library,* Software-Practice and Experience **6,** 71-82.

Schonfelder, J.L. (1978), *Chebyshev expansions for the error and related functions,* Math. Comp. **32,** 1232-1240.

Scraton, R.E. (1970), *A method for improving the convergence of Chebyshev series,* Comp. J. **13,** 202-203.

Shepherd, N.M., J.G. Laframboise (1981), *Chebyshev approximation of* $(1+2x)exp(x^2)erfc\ x$ *in* $0 \leq x < \infty$, Math. Comp. **36,** 249-253.

Sluis, A. van der (1976), *Estimating the solutions of slowly varying recursions,* SIAM J. Math. Anal. **7,** 662-695.

Spira, R. (1971), *Calculating of the gamma function by Stirling's formula,* Math. Comp. **25,** 317-322.

Stegun, I.A., M. Abramowitz (1956), *Pitfalls in computation,* J. Soc. Indust. Appl. Math. **4,** 207-219.

Stegun, I.A., Zucker, R. (1970, 1974, 1976, 1981), *Automatic computing methods for special functions,* J. Res. Nat. Bur. Stand.

Error, probability and related functions, **74B,** 221-224, 1970

The exponential integral $E_n(x)$, **78B,** 199-216, 1974.

The sine, cosine, exponential integrals and related functions, **80B,** 291-311, 1976.

Complex error function, Fresnel integrals, and other related functions, **86,** 661-686, 1981.

Strecok, A.J. (1968), *On the calculation of the inverse of the error function,* Math. Comp. **22,** 144-158.

Stroud, A.H., D. Secrest (1966), *Gaussian quadrature formulas,* Englewood Cliffs N.J. Prentice-Hall.

Szegö, G. (1974), *Orthogonal polynomials,* Colloq. Publ. Vol. 23, Amer. Math. Soc., Providence, Rhode Island.

Temme, N.M. (1976), *Speciale functies,* 179-206 in: J.C.P. Bus (red.) *Colloquium numerieke programmatuur,* MC-Syllabus 29.1b, Mathematisch Centrum, Amsterdam. (Dutch).

Temme, N.M. (1983): *The numerical computation of the confluent hypergeometric function* $U(a,b,z)$ Numer. Math. **41,** 63-82.

Thacher, H.C. (1965), *Independent variable transformations in approximation,* 567-577 in: Proceedings IFIP-conference, Vol.2.

Thron, W.J. (1974), *A survey of recent convergence results for continued fractions,* Rocky Mountain J. of Math. **4,** 273-282.

Thron, W.J., H. Waadeland (1980), *Accelerating convergence of limit periodic continued fractions* $K(a_n / 1)$, Numer. Math. **34,** 155-170.

Thron, W.J., H. Waadeland (1982), *On certain transformations of continued fractions,* 225-240 in: Jones, Thron, Waadeland (1982).

Todd, J. (1954), *Evaluation of the exponential integral for large complex arguments,* J. Res. Nat. Bur. Stand. **52,** 313-317.

Traub, J.F., M. Shaw (1974), *On the number of multiplications for the evaluation of a polynomial and* *some of its derivatives,* J. ACM **21,** 161-167.

VANDEVENDER, W.H., K.H. HASKELL (1982), *The SLATEC mathematical subroutine library,* SIGNUM **17** (3), 16-21.

WALL, H.S., (1948), *Analytic theory of continued fractions,* Van Nostrand.

WHITTAKER, E.T., G.N. WATSON (1927), *A course in modern analysis,* Cambridge University Press.

WIJNGAARDEN, A VAN, et al. (1976), *Revised report on the algorithmic language ALGOL 68,* Mathematical Centre, Amsterdam.

WRENCH, J.W.Jr. (1968), *Concerning two series for the gamma function,* Math. Comp. **22,** 617-626.

WILKINSON, J.H. (1965), *The algebraic eigenvalue problem,* Clarendon.

WUYTACK, L. (1976), *Applicatons of Padé approximation in numerical analysis,* 453-463 in: CABANNES (1976).

WUYTACK, L. (ed.), (1979), *Padé approximation and its applications,* Lecture notes in Mathematics 765, Springer Verlag.

ZAHAR, R.V.M. (1977), *A mathematical analysis of Miller's algorithm,* Numer. Math. **27,** 427-447.

INDEX