Contemporary Mathematicians

Claude Brezinski
Ahmed Sameh
Editors

Walter Gautschi

Selected Works with Commentaries

Volume 1

Birkhäuser

Contemporary Mathematicians

Joseph P.S. Kung
University of North Texas, USA

Editor

For further volumes:
http://www.springer.com/series/4817

Claude Brezinski • Ahmed Sameh
Editors

Walter Gautschi, Volume 1

Selected Works with Commentaries

Birkhäuser

Editors
Claude Brezinski
U.F.R. de Mathématiques
Université des Sciences et Technologies
de Lille
Villeneuve d'Ascq, France

Ahmed Sameh
Department of Computer Science
Purdue University
West Lafayette, IN, USA

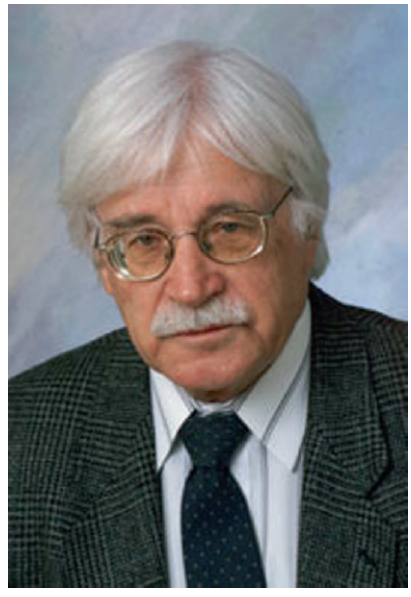
ISBN 978-1-4614-7033-5 ISBN 978-1-4614-7034-2 (eBook)
DOI 10.1007/978-1-4614-7034-2
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013949491

Mathematics Subject Classification (2010): 01Axx, 65Dxx, 65Lxx, 65Qxx, 65Yxx

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.birkhauser-science.com)



Walter Gautschi, 2001

Contents

List of Contributors

Walter Van Assche
Department of Mathematics
KU Leuven, Heverlee, Belgium

John C. Butcher
Department of Mathematics
The University of Auckland
Auckland, New Zealand

Martin Gander
Section de Mathématiques
Université de Genève
Genève, Switzerland

Nick Higham
School of Mathematics
The University of Manchester
Manchester, UK

Jacob Korevaar
Korteveg de Vries Institut
University of Amsterdam
Amsterdam, The Netherlands

Lisa Lorentzen
Institutt for Matematiske
Fag NTNU
Trondheim, Norway

Gradimir Milovanović
Matematički Institut SANU
Beograd, Serbia

Giovanni Monegato
Dipartimento di Matematica
Politecnico di Torino
Torino, Italy

Lothar Reichel
Department of Mathematical Sciences
Kent State University
Kent, OH, USA

Javier Segura
Departamento de Matematicas
Estadistica y Computacion
Universidad de Cantabria
Santander, Spain

Miodrag M. Spalević
Department of Mathematics
University of Belgrad
Belgrade, Serbia

Gerhard Wanner
Section de Mathématiques
Université de Genève
Genève, Switzerland

1

Preface

Claude Brezinski and Ahmed Sameh

Walter Gautschi is a world-renowned numerical analyst whose research contributions cover a wide range of topics including numerical conditioning, special functions, interpolation and approximation, orthogonal polynomials, quadrature, linear recurrence relations, ordinary differential equations, and history of mathematics. His contributions have had a significant impact on the field, and his papers are widely cited. Walter has published 3 books, 34 book chapters, 160 refereed journal papers, 7 refereed papers in conference proceedings, translated 3 books, and edited 5 conference proceedings. His papers are characterized by their clarity of exposition and will remain excellent resources for researchers in the field. Walter has 4820 citations in Google Scholar and 174,000 citations in Google. His two books: *Numerical analysis — an introduction*, published by Birkhäuser, and *Orthogonal polynomials — computation and approximation*, published by Oxford University Press, have set a high standard for graduate textbooks in their respective subjects.

Walter's 65th birthday was celebrated by a conference held in his honor in December 1993 at Purdue University, attended by leaders in the field, such as Richard Askey, Carl de Boor, John Butcher, Ward Cheney, Paul Erdős, Gene Golub, Bill Gragg, Arieh Iserles, Charles Micchelli, Frank Olver, John Rice, Ted Rivlin, Ed Saff, Frank Stenger, Richard Varga, Jet Wimp, among others. The proceedings of this conference were published by Birkhäuser in 1994. Since then, Walter has added significantly to his contributions to warrant this publication (also by Birkhäuser) of his selected works together with commentaries by foremost experts in the respective areas of Walter's contributions. Volume 1 collects papers on numerical conditioning, special functions, interpolation and approximation; Volume 2 those on orthogonal polynomials — on the real line and on the semicircle —, and quadrature — of Chebyshev, Gauss, and Kronrod type —; and Volume 3 papers on linear recurrence relations, ordinary differential eqations, computer algorithms and software packages, history and biography, and miscellaneous topics.

C. Brezinski and A. Sameh (eds.), *Walter Gautschi, Volume 1: Selected Works with Commentaries*, Contemporary Mathematicians, DOI 10.1007/978-1-4614-7034-2_1,

The papers included are chosen by Walter, and the editors wish to thank the publishers of Walter's papers for permission to reprint them here. The editors also express their gratitude to the commentators for their excellent reviews and prompt response.

Finally, we wish to thank Birkhäuser for their wonderful cooperation to produce these volumes, thereby preserving and making easily accessible Walter's contribution to Computational Mathematics. We also thank Professor Michela Redivo-Zaglia of the University of Padua for lending a hand to one of the editors with Birkhäuser's latex style in the early phase of the work.

We present these volumes, honoring Walter and the memory of his late brother Werner, as a tribute to Walter — an inspiring and valued colleague. We are proud to call him a great friend.

Claude Brezinski
Ahmed Sameh

December 17, 2012

Part I

Walter Gautschi

In the article of Section 3, numbers in brackets refer to the numbered list of papers in Section 4, those in boldface type to papers included in these selected works.

2

Biography of Walter Gautschi

Claude Brezinski and Ahmed Sameh

Walter Gautschi was born on December 11, 1927 in Basel, Switzerland, together with his twin brother Werner. He attended primary and secondary schools in Basel, graduating in 1947 from the Mathematisch-Naturwissenschaftlichen Gymnasium. He then enrolled at the University of Basel to study mathematics as the primary subject, with physics, physical chemistry, and actuarial mathematics as secondary subjects. In the early 1950s he became an assistant of Professor Alexander M. Ostrowski, obtaining a Ph. D. in 1953 under his supervision with a thesis on graphical integration of ordinary differential equations. He then received a two-year fellowship for study abroad from the Janggen-Poehn foundation in St. Gallen, of which he spent the first year at the *Istituto Nazionale per le Applicazioni del Calcolo* in Rome, founded and directed by Mauro Picone, and a second year at the Harvard Computation Laboratory. It was at the Harvard Computation Laboratory where he got his first hands-on experience with electronic computers, programming (in machine code) on Professor Aiken's MARK III computer. In 1956, under a contract with the American University, he joined the staff of the Computation Laboratory at the National Bureau of Standards in Washington, D. C. (now the National Institute of Standards and Technology). There, his major project was the preparation of two chapters of the *Handbook of Mathematical Functions* edited by Milton Abramowitz and Irene A. Stegun. Abramowitz introduced Walter to the work of J. C. P. Miller on backward recurrence, which became one of the early areas of emphasis in Walter's research. Because of employment difficulties related to Walter's Swiss citizenship, he had to leave the Bureau in 1959 and he joined Alston Householder's Mathematics Panel at the Oak Ridge National Laboratory. Through contacts with chemists at the laboratory, he became interested in the numerical aspects of Gaussian quadrature and orthogonal polynomials, which was to become one of the principal areas of Walter's research contributions. During the four years at the Oak Ridge laboratory he was twice invited to lecture at the Michigan University Engineering Summer Conferences then organized by Robert C. F. Bartels.

C. Brezinski and A. Sameh (eds.), *Walter Gautschi, Volume 1: Selected Works with Commentaries*, Contemporary Mathematicians, DOI 10.1007/978-1-4614-7034-2_2,

In 1960, after the untimely death of Walter's twin brother Werner in 1959, he married his widow, Erika Wüst, and adopted their son Thomas, born only after Werner's death. The marriage brought forth three more children, Theresa, Doris, and Caroline, born respectively in 1961, 1965, and 1969.

In 1963, Walter started his academic career, accepting a professorship jointly at the then (1962) newly established Department of Computer Sciences and the Department of Mathematics at Purdue University. It was to become a life-long association, interrupted only by sabbatical years, 1970–1971 as a Fulbright scholar at the Technical University of Munich, and 1976–1977 at the University of Wisconsin. Walter regularly taught the beginning graduate course on Numerical Analysis, an advanced course on the numerical solution of ordinary differential equations, and occasionally courses on numerical linear algebra and optimization. Notes prepared over the years on the first two of these courses, and also notes prepared for summer courses taught repeatedly in Perugia, Italy, in the 1970s, led in 1997 to the publication of his book on Numerical Analysis by Birkhäuser Boston. A second edition of this book appeared in 2012. Another book, that grew out of seminars held on the constructive aspects and applications of orthogonal polynomials, was published by Oxford University Press in 2004.

Throughout his academic career, Walter participated and lectured at numerous national and international meetings and was a frequent visitor at other academic institutions, notably the Polytechnics of Milan and Turin, the University of Padua, the ETH in Zurich, and his alma mater, the University of Basel. For many years he was also a consultant at Argonne National Laboratory.

In 2001, Walter was elected a Foreign and Corresponding Member of two European Academies, respectively the Bavarian Academy of Sciences in Munich and the Turin Academy of Sciences (once the Royal Society). He was also named a SIAM Fellow in 2012.

From 1966 to 1999, Walter was a member of the Editorial Committee of *Mathematics of Computation* and its Managing Editor from 1984 to 1995. His meticulous attention to details was legendary. Other journals for which he served as an Associate Editor are *Numerische Mathematik*, 1971 to the present (Honorary Editor since 1991), the *SIAM Journal on Mathematical Analysis*, 1970–1973, and *Calcolo*, 1975–1987. In addition, in 1981–1983, Walter served as a Special Editor of *Linear Algebra and its Applications*. On the 50th anniversary of *Mathematics of Computation*, Walter edited an AMS proceedings volume entitled *A half-century of computational mathematics*, and he was co-editor of a number of other proceedings volumes. He was also active as a translator, translating (jointly with R. Bartels and C. Witzgall) the text *Numerische Mathematik* by J. Stoer, preparing an annotated translation of H. Rutishauser's *Vorlesungen über numerische Mathematik*, and (jointly with his wife Erika) an English translation of E. A. Fellmann's *Leonhard Euler*.

Walter officially retired from Purdue University in 2000 with the title of Professor Emeritus, but both his research and lecturing activities continued unabatedly ever since.

For more details on Walter's life, and especially his early research activities, see also Walter Gautschi's "Reflections and recollections" in *Approximation and Computation — a festschrift in honor of Walter Gautschi* (R. V. M. Zahar, ed.), pp. xvii–xxxv, Birkhäuser, Boston, 1994.

3

A brief summary of my scientific work and highlights of my career

Walter Gautschi

I have worked in a number of different areas of (mostly computational) mathematics. They are organized here in thirteen sections. For the sake of brevity, when referring to joint papers, coauthors are not identified explicitly.

1. *Numerical conditioning.* The general theme here is to analyze the sensitivity of a problem to small perturbations in the data. This has been an area of continued interest to me, given my predilection to fundamental issues.

An example of this is the extensive work on the condition of Vandermonde and Vandermonde-like matrices. The former [**16**, **19**, **34**, **51**, **52**, **62**, **110**] are shown to be always ill-conditioned, exponentially so or worse, if the nodes are real. They are usually well-conditioned if the nodes are complex. A noteworthy example [**120**] is the $n \times n$ Vandermonde matrix whose nodes are the first n members of an infinite sequence of complex numbers on the unit circle, for example the Van der Corput sequence. The (spectral) condition number is then shown to be bounded by $\sqrt{2n}$. In the case of (real) Vandermonde-like matrices whose entries are not powers of the nodes, but orthogonal polynomials evaluated at the nodes, the matter depends on the Christoffel numbers, or Christoffel function (evaluated at the nodes) of the underlying measure, more precisely, on the ratio of their arithmetic and harmonic means [**83**]. Another interesting problem treated very recently pertains to optimally scaled and optimally conditioned Vandermonde and Vandermonde-like matrices [**200**]. For a survey, see also [**118**].

Other instances of work in this area are the condition of polynomial bases [**43**, **66**], the condition of algebraic equations [**45**], and most notably, the condition of moment maps in the theory of orthogonal polynomials and related quadratures [**40**, **81**, **98**].

2. *Special functions.* My contributions to this subject are four-fold: numerical evaluation, inequalities, asymptotics, and expository work.

C. Brezinski and A. Sameh (eds.), *Walter Gautschi, Volume 1: Selected Works with Commentaries*, Contemporary Mathematicians, DOI 10.1007/978-1-4614-7034-2_3,

In the first of these categories, the influential work [**29**] should be mentioned on computational aspects of three-term recurrence relations. This centers around the concept of minimal solution of three-term recurrence relations and related algorithms involving continued fractions. The latter have been successfully applied to the computation of many special functions, such as Bessel functions [23], Legendre functions [24], Coulomb wave functions [28, 33, **35**], incomplete beta and gamma functions [22, 158], repeated integrals of the error function [**13**, 59, 60], and Stieltjes transforms of orthogonal polynomials [**75**]. Special mention deserves an efficient algorithm developed for computing the complex error function [36, **39**] (a Stieltjes transform of the Hermite weight function), which relies on similar ideas and which has found widespread use in the physics and nuclear engineering communities. In the paper [63], attention is drawn to a continued fraction of Perron as a useful alternative to the more customary Gauss-type continued fraction for evaluating ratios of modified Bessel functions of a real argument. A variety of techniques, including Taylor series and continued fraction expansions, are employed in the calculation of incomplete gamma functions [**68**, 69, 70]. Applying Gaussian quadrature led to useful procedures for computing hypergeometric and confluent hypergeometric functions [**168**], Bessel and Airy functions [**169**], modified Bessel functions of complex orders [**178**], and Kontorovich–Lebedev integral transforms [**181**]. High-precision nonstandard Gaussian quadrature rules are also employed to compute certain integrals involving the Lambert W-function [**199**].

With regard to inequalities, the two-sided inequalities for gamma function ratios [**9**], published in 1959, have been most widely noted (and now bear my name), although they were obtained in the context of more general two-sided inequalities for the incomplete gamma function. Of a quite different nature are the harmonic mean inequalities for the gamma function [**47**, **48**], obtained in the 1970s. In [**72**], classical inequalities of Laguerre for the largest zero of Jacobi, Laguerre, and Hermite polynomials are sharpened. Beginning in 2007, in a series of papers [**182**, **190**, **191**, **192**, **203**], a number of far-reaching conjectures are set forth regarding inequalities for zeros of Jacobi polynomials, all based on extensive numerical computation. Bernstein's inequality for Jacobi polynomials is analyzed in [**193**] with regard to sharpness and extended to larger domains of the Jacobi parameters. The computational work therein also suggests a numerical value for the best constant in the Erdélyi–Magnus–Nevai conjecture on orthonormal Jacobi polynomials.

There is one short paper on asymptotics [**10**] generalizing an asymptotic formula of G. Blanch for exponential integrals.

Most important among my expository work on special functions are the two chapters [20, 21] on the exponential integrals and the error function in the famous handbook of Abramowitz and Stegun.

3. *Interpolation and approximation.* An early paper [11] deals with bivariate linear interpolation of an analytic function in the complex plane and the respective error committed.

According to a classical result of Erdős and Turán, Lagrange interpolation of any continuous function on $[-1, 1]$ at the n zeros of an orthogonal polynomial of degree n converges in the mean as $n \to \infty$. Does the same conclusion hold if one inserts $n+1$ additional points in a well-specified manner (similar to Kronrod's method in the theory of quadrature)? This is explored in [**132**] with mixed success: the answer is conjectured to be "yes" for Jacobi polynomials $P_n^{(\alpha,\beta)}$ with parameters α, β suitably restricted, but is proved to be "no" for Chebyshev polynomials of the first, third, and fourth kind (the answer being trivially "yes" for Chebyshev polynomials of the second kind). For quadrature convergence in the sense of Erdős and Turán, however, the answer is "yes" for all four Chebyshev polynomials, as is proved in [**147**]. Under an additional interlacing condition on the interpolation points, we also established necessary and sufficient conditions for quadrature convergence to hold and conjectured them to be satisfied for Jacobi polynomials with parameters $|\alpha| \le \frac{1}{2}$, $|\beta| \le \frac{1}{2}$.

Inspired by work in physics, I became interested in approximating a function in such a way that as many of its moments as possible are preserved. I began by considering functions f on $\mathbb{R}_+$ and approximation by piecewise constant functions with both the location and height of their jumps being freely variable [**89**]. The problem was generalized in [**100**] to approximation on $\mathbb{R}_+$ by spline functions of fixed degree and variable (positive) knots. Interestingly, under appropriate conditions the problem has a unique solution expressible in terms of the nodes and weights of a Gaussian quadrature formula relative to a weight function which depends on f. Unique existence is always assured if f is completely monotonic on $\mathbb{R}_+$. Analogous problems on a finite interval can also be solved [**102**] and involve generalized Gauss–Radau and Gauss–Lobatto formulae. For a summary of this work and related work by others, see [131].

Other approximation-theoretic problems considered pertain to continued fractions [61, **87**, 127], Padé approximation [**86**], Fourier analysis [**41**], and the summation of slowly convergent series [**93**, **124**, **125**, **175**].

4. *Orthogonal polynomials on the real line.* The constructive theory of orthogonal polynomials is an area of work for which I am probably best known. (I have been called Mr. Orthogonal Polynomials by some of my colleagues!) I was the first to take up the problem of computationally generating orthogonal polynomials relative to essentially arbitrary weight functions or measures. While the solution via moments, in principle, is classically known, it is problematic computationally because of severe ill-conditioning. The major effort, indeed, was to carefully analyze the degree of ill-conditioning and to find methods that successfully surmount this ill-conditioning. The approach I have taken was to either replace moments by so-called modified moments (an idea that had been floating around at the time) and study the condition number of the relevant moment map; or else, to discretize the underlying inner product and take the corresponding discrete orthogonal polynomials to approximate the desired ones. The former approach led to two algo-

rithms, one based on Cholesky decomposition [**40**], and another, more efficient one [**81**, §2.4], given the name modified Chebyshev algorithm, because I could trace its origin to an 1859 memoir of Chebyshev dealing with ordinary moments of a discrete measure. The second appoach, often more effective, is entirely original with me [**31**]. It led to what I called a discretized Stieltjes procedure [**81**, §2.2], since Stieltjes in 1884 briefly alluded to an algorithm of this kind (without discretization). A Fortran program implementing the method has been published in [32]. Both algorithms are extended in [**145**] to Sobolev orthogonal polynomials, which are orthogonal with respect to an inner product also containing derivatives and accompanying measures. They are applied in [**153**] to illustrate theoretical results about the asymptotic distribution of zeros of Sobolev orthogonal polynomials and their derivatives. Very special Sobolev orthogonal polynomials involving a derivative of fixed order with an associated one-point atomic measure are discussed in [**151**] along with their zeros.

The algorithms thus developed, sometimes in conjunction with analytic or symbolic variable-precision tools, have been used to generate (recursion coefficients of) orthogonal polynomials with special, sometimes unusual, weight functions, for example the reciprocal gamma function [**80**], weight functions of interest in theoretical chemistry that are supported on two separate intervals [**90**], Einstein and Fermi functions [**93**], Freud and half-range Hermite weight functions [**195**], refinable [**161**] and densely oscillating, or rapidly exponentially decaying, weight functions [**176**], and sub-range Jacobi weight functions [**205**].

Other important algorithms studied pertain to modifications of the weight function, for example multiplying it by a positive rational function [**77**], [**179**, §2.6]. A notable special case is multiplication by the square of the respective orthogonal polynomial, which gives rise to what in [134] are called induced orthogonal polynomials. They are relevant, e.g., in the problem of extended interpolation mentioned in §3. Repeated modifications by linear divisors are studied in [**206**] and applied to generate special Gaussian quadrature rules for dealing with nearby poles. Some of these algorithms can also be used to "neutralize" singularities other than poles [**207**].

5. *Orthogonal polynomials on the semicircle.* An entirely new kind of (complex) orthogonal polynomials was introduced in 1985: polynomials orthogonal on the semicircle [**95**]. The novelty here is the non-Hermitian nature of the underlying inner product. Yet, many properties of these new polynomials, and also of the respective zeros, resemble properties known for classical orthogonal polynomials with positive definite or Hermitian weight functions. This was further developed in a number of papers, [**97**, **104**, **113**] and summarized in [116].

6. *Chebyshev quadrature.* The majority of my papers is dedicated to problems of quadrature. My early work in this area, suggested by a visitor (Hiroki Yanagiwara) from Japan, deals extensively with weighted Chebyshev and Chebyshev-type

formulae. The former are weighted quadrature rules with equal (real) coefficients, distinct (real) nodes, and polynomial degree of exactness equal to the number of nodes. From a celebrated result of Bernstein (relative to constant weight functions) one can expect such quadrature rules to exist only for a finite, typically small, number of nodes. A severe case in point is exhibited in [**50**]. In all remaining instances one can try to find substitute formulae by relaxing the exactness condition in one way or another. This is the kind of problem studied by me and co-workers in the mid-1970s [**46**, **50**, **53**, **57**, **58**]. A historical summary is provided in [**55**].

7. *Kronrod and other quadratures.* Gauss–Kronrod formulae give rise to intriguing problems of existence, that is, of determining if and when all nodes are real and distinct. This has been studied, using algebraic tools, for Jacobi weight functions in [**109**]. Other instances of such formulae, [**111**, **114**], involve weight functions of Bernstein–Szegő type, i.e., Chebyshev weights of any of the four kinds divided by a quadratic polynomial which remains positive on $[-1, 1]$, or weights whose orthogonal polynomials have a three-term recurrence relation with ultimately constant coefficients [148]. Computing Gauss–Kronrod formulae is a topic discussed in [**99**, **108**]. For a review up to about 1987, see [**107**].

A convergence result for interpolatory quadrature rules with Chebyshev nodes (already studied by Fejér), when applied to improper integrals having monotonic singularities at the endpoints ± 1, is proved in [**30**]. The result is of interest in the generation of orthogonal polynomials by the discretized Stieltjes procedure (cf. §4).

Evaluating the Hilbert transform (a Cauchy principal value integral) of the classical Jacobi, Laguerre, and Hermite measures and of the respective orthogonal polynomials is discussed in [**103**, **166**]. The latter satisfy the same three-term recurrence relation as the one for the orthogonal polynomials themselves, but exhibit a phenomenon of pseudostability (cf. §9). A case of computing singular integrals is studied in [105].

In [**160**], a new look is taken at adaptive quadrature employing, among other devices, a 4-point Gauss–Lobatto formula and two successive Kronrod extensions thereof. The procedure has been incorporated into one of the Matlab quadrature routines, `quadl`.

A challenging integral involving an integrand that is densely oscillating near one of the endpoints of the interval of integration, with amplitudes tending to infinity, is evaluated in [188] by elementary means.

8. *Gauss-type quadrature.* The larger part of my work on numerical quadrature, however, concerns Gauss-type quadrature rules and, apart from [**31**, **40**, **65**], began to appear around 1981 after my long historical essay [**74**] on Gauss-Christoffel quadrature rules, written on the occasion of Christoffel's 150th anniversary of birth. The work can be divided into five parts: (i) geometric properties, (ii) explicit formulae and computation, (iii) validation, (iv) error estimation for analytic functions, and (v) polynomial/rational formulae.

(i) In 1961, P. J. Davis and Philip Rabinowitz proved that the classical Gauss–Jacobi formula has weights which, when suitably normalized and plotted over the corresponding nodes, come to lie on the upper half of the unit circle, asymptotically for large orders. In 2006, I have shown [**180**] that this pretty "circle theorem" is true for a much larger class of weight functions, essentially the Szegő class, not only for Gauss formulae, but also for Gauss–Radau, Gauss–Lobatto, and, under more restictive conditions, even for Gauss–Kronrod formulae.

(ii) There is a large number of papers dealing with the numerical calculation not only of Gaussian formulae (which essentially amounts to the numerical generation of the respective orthogonal polynomials – see §4 – followed by an eigenvalue/vector computation involving the Jacobi matrix of the orthogonal polynomials), but also of ordinary [**163**, **164**] and generalized [**173**], [**194**] Gauss–Radau and Gauss–Lobatto formlae, especially of very high order, as well as Gauss–Turán formulae [**154,** 211] (which involve derivative values of the function to be integrated up to some even order). For Gauss–Radau and Gauss–Lobatto formulae with double endpoints and Chebyshev weight functions of all four kinds, explicit formulae for the boundary weights are derived in [**126**].

(iii) The problem of validation, considered in [**84**], consists in assessing *a posteriori* the accuracy of the nodes and weights of a Gaussian quadrature formula, once computed in one way or another. In view of the severe ill-conditioning mentioned in §4, this is a nontrivial problem.

(iv) The remainder term of weighted Gaussian quadrature formulae over a finite interval applied to analytic functions can be estimated by contour integration techniques. This is the subject of the frequently cited work [**85**] and of [**119**]. Additional work on this topic is done in [**121**, **123**], where the same techniques are applied to Gauss–Radau and Gauss–Lobatto quadratures.

(v) Quadrature formulae with polynomial degrees of exactness are of limited use when the function to be integrated has poles, especially poles near the interval of integration. In such cases, it is more meaningful to include among the functions that are integrated exactly also rational functions having the same, or at least the more important, poles. It turns out that such polynomial/rational n-point quadrature formulae (that exactly integrate m rational functions, $0 < m \leq 2n$, with prescribed poles of given multiplicities and polynomials of degree $2n - m - 1$) can be constructed in terms of classical (polynomial) Gauss formulae with modified weight functions and hence can be computed by methods described earlier. This is discussed, and illustrated by a number of examples, in [**137**], and implemented in a computer algorithm in [159]. Integrals over half-infinite intervals and exact for special rational functions are considered in [**128**]. Polynomial/rational versions of other quadrature rules, specifically Gauss–Kronrod, Gauss–Turán rules, and quadrature procedures for Cauchy principal value integrals, are developed in [162] and (favorably) compared with the polynomial counterparts. For an updated summary that includes also estimates of the remainder term and an additional

example, see [**167**]. Applications to Fermi–Dirac and Bose–Einstein integrals and comparisons with results in the physics literature are made in [**136**].

The theory and algorithms described in §§4–5 and §§7–8 and various applications thereof, are the subject of a monograph [B3] published in 2004 by Oxford University Press. For additional surveys, see also [65, 92, **94**, 112, 115, 117, **122**, 130, 140, 157].

9. *Linear difference equations.* As already mentioned in §2, linear homogeneous difference equations of order two (i.e., three-term recurrence relations) are an important tool for computing special functions. So are inhomogeneous difference equations of order one (for example, see [12, **26**, 27, **37**, 38, 44]). In [42, **150**], the numerical stability of initial and boundary value problems for such difference equations is discussed systematically using the concept of amplification factors. Special attention is given to a phenomenon of "pseudostability" (stability in theory, but instability in practice). Its adverse effects on computing are illustrated in [**135**] in the case of discrete orthogonal polynomials when computed by their three-term recurrence relation (cf. also [**150**, §3.4.2]).

10. *Ordinary differential equations.* I acquired an interest in this topic early on, already during my doctoral thesis work. I obtained error bounds [5] for special Runge–Kutta methods developed by Zurmühl for single differential equations of arbitrary order, following work of Bieberbach on the classical Runge–Kutta method for first-order differential equations. In 1961 I developed numerical methods based on trigonometric rather than algebraic polynomials [**14**], anticipating methods later called "exponentially fitted". Only recently, they have attracted renewed interest in the context of oscillatory second-order differential equations and also in time integration schemes for Maxwell equations in three dimensions. An early expository account of the theory of one-step and multistep methods is given in [15], which for the first time includes Dahlquist's theory of stability and convergence of linear multistep methods. Later in 1975, in the paper [**54**] dedicated to Mauro Picone, it is proposed to estimate the global error (not the local error, as is usually done) of one-step methods by integrating numerically the variational differential equation along with the main differential equation. For multistep methods, this is done in my book [B2, §6.3.5 of the 2d edition]. Asymptotic estimates are derived in [56] for certain coefficients of interest in Adams, Störmer, and Cowell multistep methods. The last paper on differential equations [**73**] appeared in 1980. Within a class of stable multistep methods it determined the method which has minimum coefficient in the asymptotic formula for the global error.

11. *Software.* Much of my work on computing special functions is supported by pieces of software, initially written in Algol and Fortran and published in separate algorithms, and later written in Matlab and placed on my homepage. I also wrote major software packages in support of my work on orthogonal polynomials and quadrature: the Fortran package `ORTHPOL` [**141**] and the

Matlab package `OPQ` [174, **179**, `http://www.cs.purdue.edu/ archives/2002/wxg /codes/OPQ.html`]. Some of the routines in the latter package have been rewritten in symbolic Matlab and collected in the package `SOPQ` [`http://www.cs.purdue.edu /archives/2002/wxg/codes/SOPQ.html`], and can therefore be run in variable-precision arithmetic. This, incidentally, provides another approach to overcome the ill-conditioning mentioned in §4: simply do the computation with as many digits as are required to compensate for the loss of accuracy caused by ill-conditioning.

12. *History and biography.* Every so often, I took time out to review the history of a subject I had been working on myself. This led to a number of special-topic surveys, for example on computational methods in special functions [**49**], advances in Chebyshev quadrature [**55**], questions of numerical condition related to polynomials [**64**], Gauss-Christoffel quadrature formulae [**74**], Gauss-Kronrod quadrature [**107**], remainder estimates for analytic functions [**129**], applications and computation of orthogonal polynomials [**146**], the incomplete gamma function since Tricomi [**155**] (written on the occasion of Tricomi's 100th anniversary of birth), and the interplay between classical analysis and numerical linear algebra [**170**], a tribute to Gene H. Golub. There are also appreciations of the work, and sometimes the life, of individual personalities, for example Yudell L. Luke [**91**] (an obituary), Philip Rabinowitz [**143**], Luigi Gatteschi [**144**, **189**], Alexander M. Ostrowski [82], [**196**] (published on the occasion of the 100th anniversary of the Swiss Mathematical Society), Joseph-Louis Lagrange [210], and, above all, Leonhard Euler [**187**]. Two of my articles deal specifically with Euler's handling of slowly convergent series [**183**] and with Euler's curious attempt [**186**] (communicated in a 1734 letter to his friend Daniel Bernoulli) to interpolate the common logarithm from the known values $\log 10^k = k$, $k = 0, 1, 2, 3, \ldots$.

13. *Miscellanea.* Additional work not easily subsumed under any of the categories above concerns a proof, under weaker assumptions, of a necessary condition of Picone in the calculus of variation [7] and an extension thereof to double integrals [6], families of algebraic test equations [**71**], the error behavior in optimal relaxation methods [78], monotonicity and complete monotonicity properties related to the successive remainder terms of the exponential series [79], an algorithm for simultaneous orthogonal transformation of several positive definite matrices to nearly diagonal form [**96**], summation procedures [**175**] for evaluating the interesting Hardy–Littlewood function $H(x) = \sum_{k=1}^{\infty} \sin(x/k)/k$, and the analytic smoothing of the discrete spiral of Theodorus [**197**].

———ooOOoo———

Apart from my election in 2001 to two prominent European Academies – the prestigious Bavarian Academy of Sciences and Humanities, founded in 1759, and the

Turin Academy of Sciences, once the Royal Academy, founded in 1761 by Lagrange and others – and the designation of SIAM Fellow in 2012, there are two highlights in my career that stand out. The first is my collaboration in 1984 with Louis de Branges on the proof of the Bieberbach conjecture [**101**]. De Branges knew that the validity of the Bieberbach conjecture for the nth coefficient hinges on the validity of a system of n inequalities involving integrals of Jacobi polynomials (in fact ${}_3F_2$ hypergeometric functions). I was able to use my software package [**141**] for orthogonal polynomials to verify computationally that these inequalities are indeed valid for all n up to 30. More importantly, I made a now famous telephone call to Richard Askey, which led to the incredible discovery that these inequalities are true not only for $n \leq 30$ but for all n, being a special case of results proved several years earlier by Askey and Gasper. That finished off de Branges's proof of the Bieberbach conjecture. The second highlight was the Euler lecture I was invited to give as part of the Euler 300th anniversary year before an audience of some 3,000 attendees of the ICIAM 2007 Congress in Zürich. An expanded version of the lecture has been published in [**187**], and a preliminary rendition thereof given, and recorded, at Purdue University; the video is made available with permission at `springer.com` (type in the ISBN of Vol. 1, 978-1-4614-7033-5, and click on `EulerLect.avi`).

4

Publications

Walter Gautschi

Books

B1. (with H. Bavinck and G. M. Willems) *Colloquium approximatietheorie*, MC Syllabus 14, Mathematisch Centrum Amsterdam, 1971.
B2. *Numerical analysis: an introduction*, Birkhäuser, Boston, 1997. [2d edition, 2012]
B3. *Orthogonal polynomials: computation and approximation*, Oxford University Press, Oxford, 2004.

Proceedings Edited

P1. (with G. Allasia, L. Gatteschi, and G. Monegato) *International conference on special functions: theory and computation*, Rend. Sem. Mat. Univ. Politec. Torino, Special Issue, Università e Politecnico di Torino, Turin, 1985.
P2. *Mathematics of Computation 1943–1993*: *a half-century of computational mathematics*, Proc. Sympos. Appl. Math. 48, American Mathematical Society, Providence, RI, 1994 (xx + 643 pages).
P3. (with G. H. Golub and G. Opfer) *Applications and computation of orthogonal polynomials*, Internat. Ser. Numer. Math. 131, Birkhäuser, Basel, 1999 (xiv + 268 pages).
P4. (with F. Marcellán and L. Reichel) *Numerical analysis 2000*, Vol. 5: *Quadrature and orthogonal polynomials*, J. Comput. Appl. Math. 127, nos. 1–2 (2001).
P5. (with G. Mastroianni and Th. M. Rassias) *Approximation and computation: in honor of Gradimir V. Milovanović*, Springer Optim. Appl. 42, Springer, Dordrecht, 2011.

C. Brezinski and A. Sameh (eds.), *Walter Gautschi, Volume 1: Selected Works with Commentaries*, Contemporary Mathematicians, DOI 10.1007/978-1-4614-7034-2_4,

Translations

T1. (with R. Bartels and C. Witzgall) *Introduction to numerical analysis* by J. Stoer and R. Bulirsch, translated from German, Springer, New York, 1980. [2d ed., Texts in Appl. Math., v. 12, Springer, New York, 1993.]
T2. *Lectures on numerical mathematics* by H. Rutishauser, annotated translation from German, Birkhäuser, Boston, 1990.
T3. (with M. Mattmüller) *Consideration of some series which are distinguished by special properties*, Memoir E190 by Leonhard Euler, translated from Latin, `http://math.dartmouth.edu/~euler`
T4. (with E. Gautschi) *Leonhard Euler* by E .A. Fellmann, translated from German, Birkhäuser, Basel, 2007.

Publications

1951

1. *Ein Analogon zu Grammels Methode der graphischen Integration gewöhnlicher Differentialgleichungen*, Z. Angew. Math. Mech. 31, 242–243.

1953

2. *Fehlerabschätzungen für die graphischen Integrationsverfahren von Grammel und Meissner–Ludwig*, Verh. Naturforsch. Ges. Basel 64, 401–435.

1954

3. *Über die zeichnerischen Ungenauigkeiten und die zweckmässige Bemessung der Schrittlänge beim graphischen Integrationsverfahren von Meissner–Ludwig*, Verh. Naturforsch. Ges. Basel 65, 49–66.
4. *Über eine Klasse von linearen Systemen mit konstanten Koeffizienten*, Comment. Math. Helv. 28, 186–196.

1955

5. *Über den Fehler des Runge–Kutta-Verfahrens für die numerische Integration gewöhnlicher Differentialgleichungen n-ter Ordnung*, Z. Angew. Math. Phys. 6, 456–461.

1956

6. *Una estensione agli integrali doppi di una condizione di Picone necessaria per un estremo*, Atti Accad. Naz. Lincei. Rend. Cl. Sci. Fis. Mat. Nat. (8) 20, 283–289.
7. *Bemerkung zu einer notwendigen Bedingung von Picone in der Variationsrechnung*, Comment. Math. Helv. 31, 1–4.

8. (with F. Malmborg) *Calculations related to the improved free-volume-theory of liquids (AF Problem 116)*, Harvard Computation Laboratory, Problem Report 100, VI-1–VI-41.

1959

9. *Some elementary inequalities relating to the gamma and incomplete gamma function*, J. Math. and Phys. 38, 77–81.
10. *Exponential integral* $\int_1^\infty e^{-xt}t^{-n}dt$ *for large values of n*, J. Res. Nat. Bur. Standards 62, 123–125.
11. *Note on bivariate linear interpolation for analytic functions*, Math. Tables Aids Comput. 13, 91–96.

1961

12. *Recursive computation of certain integrals*, J. Assoc. Comput. Mach. 8, 21–40.
13. *Recursive computation of the repeated integrals of the error function*, Math. Comp. 15, 227–232.
14. *Numerical integration of ordinary differential equations based on trigonometric polynomials*, Numer. Math. 3, 381–397.

1962

15. (with H. A. Antosiewicz) *Numerical methods in ordinary differential equations*, Ch. 9 in *Survey of numerical analysis* (J. Todd, ed.), 314–346, McGraw-Hill, New York.
16. *On inverses of Vandermonde and confluent Vandermonde matrices*, Numer. Math. 4, 117–123.
17. *Diffusion functions for small argument*, SIAM Rev. 4, 227–229.

1963

18. *Instability of linear second-order difference equations*, in *Proc. IFIP Congress 62* (C. M. Popplewell, ed.), 207, North-Holland, Amsterdam.
19. *On inverses of Vandermonde and confluent Vandermonde matrices II*, Numer. Math. 5, 425–430.

1964

20. (with W. F. Cahill) *Exponential integral and related functions*, Ch. 5 in *Handbook of mathematical functions* (M. Abramowitz and I. A. Stegun, eds.), 227–251, Nat. Bur. Standards Appl. Math. Ser. 55. [Russian translation by V. A. Ditkin and L. N. Karmazina, in *Spravočnik po special'nym funkciyam*, 55–79, Nauka, Moscow, 1979.]
21. *Error function and Fresnel integrals*, Ch. 7 in *Handbook of mathematical functions* (M. Abramowitz and I. A. Stegun, eds.), 295–329, Nat. Bur. Standards Appl. Math. Ser. 55.[Russian translation by V. A. Ditkin and L. N. Karmazina, in *Spravočnik po special'nym funkciyam*, 119–152, Nauka, Moscow, 1979.]

22. *Algorithm 221 — Gamma function*, and *Algorithm 222 — Incomplete beta function ratios*, Comm. ACM 7, 143–144; Certification of Algorithm 222, *ibid.*, 244.

1965

23. *Algorithm 236 — Bessel functions of the first kind*, Comm. ACM 7, 479–480; Certification of Algorithm 236, *ibid.* 8, 105–106.
24. *Algorithm 259 — Legendre functions for arguments larger than one*, Comm. ACM 8, 488–492.

1966

25. *Computation of transcendental functions by recurrence relations*, in *Proc. IFIP Congress 65*, v. 2 (W. A. Kalenich, ed.), 485–486, Spartan Books, Washington, D. C.
26. *Computation of successive derivatives of* $f(z)/z$, Math. Comp. 20, 209–214.
27. *Algorithm 282 — Derivatives of* e^x/x, $\cos(x)/x$, *and* $\sin(x)/x$, Comm. ACM 9, 272.
28. *Algorithm 292 — Regular Coulomb wave functions*, Comm. ACM 9, 793–795.

1967

29. *Computational aspects of three-term recurrence relations*, SIAM Rev. 9, 24–82.
30. *Numerical quadrature in the presence of a singularity*, SIAM J. Numer. Anal. 4, 357–362.

1968

31. *Construction of Gauss–Christoffel quadrature formulas*, Math. Comp. 22, 251–270.
32. *Algorithm 331 — Gaussian quadrature formulas*, Comm. ACM 11, 432–436.

1969

33. *Remark on Algorithm 292*, Comm. ACM 12, 280.
34. *On the condition of a matrix arising in the numerical inversion of the Laplace transform*, Math. Comp. 23, 109–118.
35. *An application of minimal solutions of three-term recurrences to Coulomb wave functions*, Aequationes Math. 2, 171–176; abstract, *ibid.* 1 (1968), 208.
36. *Algorithm 363 — Complex error function*, Comm. ACM 12, 635.

1970

37. (with B. J. Klein) *Recursive computation of certain derivatives — a study of error propagation*, Comm. ACM 13, 7–9.
38. (with B. J. Klein) *Remark on Algorithm 282*, Comm. ACM 13, 53–54.
39. *Efficient computation of the complex error function*, SIAM J. Numer. Anal. 7, 187–198.

40. *On the construction of Gaussian quadrature rules from modified moments*, Math. Comp. 24, 245–260.

1972

41. *Attenuation factors in practical Fourier analysis*, Numer. Math. 18, 373–400.
42. *Zur Numerik rekurrenter Relationen*, Computing 9, 107–126. [English translation in: Aerospace Research Laboratories, Report ARL 73-0005, February 1973.]
43. *The condition of orthogonal polynomials*, Math. Comp. 26, 923–924.

1973

44. *Algorithm 471 — Exponential integrals*, Comm. ACM 16, 761–763.
45. *On the condition of algebraic equations*, Numer. Math. 21, 405–424.

1974

46. (with H. Yanagiwara) *On Chebyshev-type quadratures*, Math. Comp. 28, 125–134.
47. *A harmonic mean inequality for the gamma function*, SIAM J. Math. Anal. 5, 278–281.
48. *Some mean value inequalities for the gamma function*, SIAM J. Math. Anal. 5, 282–292.

1975

49. *Computational methods in special functions — a survey*, in *Theory and applications of special functions* (R. A. Askey, ed.), 1–98, Math. Res. Center, Univ. Wisconsin Publ. 35, Academic Press, New York.
50. *Nonexistence of Chebyshev-type quadratures on infinite intervals*, Math. Comp. 29, 93–99.
51. *Norm estimates for inverses of Vandermonde matrices*, Numer. Math. 23, 337–347.
52. *Optimally conditioned Vandermonde matrices*, Numer. Math. 24, 1–12.
53. (with L. A. Anderson) *Optimal weighted Chebyshev-type quadrature formulas*, Calcolo 12, 211–248.
54. *Stime dell'errore globale nei metodi "one-step" per equazioni differenziali ordinarie*, Rend. Mat. (2) 8, 601–617.

1976

55. *Advances in Chebyshev quadrature*, in *Numerical analysis* (G. A. Watson, ed.), 100–121, Lecture Notes Math. 506, Springer, Berlin.
56. *Comportement asymptotique des coefficients dans les formules d'intégration d'Adams, de Störmer et de Cowell*, C. R. Acad. Sci. Paris Sér. A-B 283, A787–A788.

57. *Qualche contributo recente sul problema di Chebyshev nella teoria dell'integrazione numerica*, Rend. Sem. Mat. Univ. e Politec. Torino 35, 39–44.

1977

58. (with G. Monegato) *On optimal Chebyshev-type quadratures*, Numer. Math. 28, 59–67.
59. *Evaluation of the repeated integrals of the coerror function*, ACM Trans. Math. Software 3, 240–252.
60. *Algorithm 521 — Repeated integrals of the coerror function*, ACM Trans. Math. Software 3, 301–302.
61. *Anomalous convergence of a continued fraction for ratios of Kummer functions*, Math. Comp. 31, 994–999.

1978

62. *On inverses of Vandermonde and confluent Vandermonde matrices III*, Numer. Math. 29, 445–450.
63. (with J. Slavik) *On the computation of modified Bessel function ratios*, Math. Comp. 32, 865–875.
64. *Questions of numerical condition related to polynomials*, in *Symposium on recent advances in numerical analysis* (C. de Boor and G. H. Golub, eds.), 45–72, Academic Press, New York. [Revised and reprinted in MAA Studies in Mathematics 24: *Studies in numerical analysis* (G. H. Golub, ed.), 140–177, Math. Assoc. America, Washington, DC, 1984.]

1979

65. *On generating Gaussian quadrature rules*, in *Numerische Integration* (G. Hämmerlin, ed.), 147–154, Internat. Ser. Numer. Math. 45, Birkhäuser, Basel.
66. *The condition of polynomials in power form*, Math. Comp. 33, 343–352.
67. *On the preceding paper "A Legendre polynomial integral" by James L. Blue*, Math. Comp. 33, 742–743.
68. *A computational procedure for incomplete gamma functions*, ACM Trans. Math. Software 5, 466–481.
69. *Algorithm 542 — Incomplete gamma functions*, ACM Trans. Math. Software 5, 482–489.
70. *Un procedimento di calcolo per le funzioni gamma incomplete*, Rend. Sem. Mat. Univ. e Politec. Torino 37, 1–9.
71. *Families of algebraic test equations*, Calcolo 16, 383–398.

1980

72. (with F. Costabile) *Stime per difetto per gli zeri più grandi dei polinomi ortogonali*, Boll. Un. Mat. Ital. (5) 17A, 516–522.
73. (with M. Montrone) *Metodi multistep con minimo coefficiente dell'errore globale*, Calcolo 17, 67–75.

1981

74. *A survey of Gauss–Christoffel quadrature formulae*, in *E. B. Christoffel — the influence of his work in mathematics and the physical sciences* (P. L. Butzer and F. Fehér, eds.), 72–147, Birkhäuser, Basel.
75. *Minimal solutions of three-term recurrence relations and orthogonal polynomials*, Math. Comp. 36, 547–554.
76. *Recognition of Christoffel's work on quadrature during and after his lifetime*, in *E. B. Christoffel — the influence of his work in mathematics and the physical sciences* (P. L. Butzer and F. Fehér, eds.), 724–727, Birkhäuser, Basel.

1982

77. *An algorithmic implementation of the generalized Christoffel theorem*, in *Numerical integration* (G. Hämmerlin, ed.), 89–106, Internat. Ser. Numer. Math. 57, Birkhäuser, Basel.
78. (with R. E. Lynch) *Error behavior in optimal relaxation methods*, Z. Angew. Math. Phys. 33, 24–35.
79. *A note on the successive remainders of the exponential series*, Elem. Math. 37, 46–49.
80. *Polynomials orthogonal with respect to the reciprocal gamma function*, BIT 22, 387–389.
81. *On generating orthogonal polynomials*, SIAM J. Sci. Statist. Comput. 3, 289–317.

1983

82. *To Alexander M. Ostrowski on his ninetieth birthday*, Linear Algebra Appl. 52/53, xi–xiv.
83. *The condition of Vandermonde-like matrices involving orthogonal polynomials*, Linear Algebra Appl. 52/53, 293–300.
84. *How and how not to check Gaussian quadrature formulae*, BIT 23, 209–216.
85. (with R. S. Varga) *Error bounds for Gaussian quadrature of analytic functions*, SIAM J. Numer. Anal. 20, 1170–1186.
86. *On Padé approximants associated with Hamburger series*, Calcolo 20, 111–127.
87. *On the convergence behavior of continued fractions with real elements*, Math. Comp. 40, 337–342.

1984

88. (with G. V. Milovanović) *On a class of complex polynomials having all zeros in a half circle*, in *Numerical methods and approximation theory* (G. V. Milovanović, ed.), 49–53, Faculty of Electronic Engineering, Univ. Niš, Niš.
89. *Discrete approximations to spherically symmetric distributions*, Numer. Math. 44, 53–60.
90. *On some orthogonal polynomials of interest in theoretical chemistry*, BIT 24, 473–483.
91. (with J. Wimp) *In memoriam: Yudell L. Luke, June 26, 1918 – May 6, 1983*, Math. Comp. 43, 349–352.

1985

92. *Some new applications of orthogonal polynomials*, in *Polynômes orthogonaux et applications* (C. Brezinski, A. Draux, A. P. Magnus, P. Maroni and A. Ronveaux, eds.), 63–73, Lecture Notes Math. 1171, Springer, Berlin.
93. (with G. V. Milovanović) *Gaussian quadrature involving Einstein and Fermi functions with an application to summation of series*, Math. Comp. 44, 177–190. Supplement, *ibid.*, S1–S11.
94. *Orthogonal polynomials — constructive theory and applications*, J. Comput. Appl. Math. 12/13, 61–76.
95. (with G. V. Milovanović) *Polynomials orthogonal on the semicircle*, Rend. Sem. Mat. Univ. e Politec. Torino, Special Issue, 179–185.

1986

96. (with B. Flury) *An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form*, SIAM J. Sci. Statist. Comput. 7, 169–184.
97. (with G. V. Milovanović) *Polynomials orthogonal on the semicircle*, J. Approx. Theory 46, 230–250.
98. *On the sensitivity of orthogonal polynomials to perturbations in the moments*, Numer. Math. 48, 369–382.
99. (with F. Caliò and E. Marchetti) *On computing Gauss–Kronrod quadrature formulae*, Math. Comp. 47, 639–650.
100. (with G. V. Milovanović) *Spline approximations to spherically symmetric distributions*, Numer. Math. 49, 111–121.
101. *Reminiscences of my involvement in de Branges's proof of the Bieberbach conjecture*, in *The Bieberbach conjecture* (A. Baernstein II, D. Drasin, P. Duren, and A. Marden, eds.), 205–211, Proc. Symp. on the Occasion of the Proof, Math. Surveys Monographs 21, American Mathematical Society, Providence, RI.

1987

102. (with M. Frontini and G. V. Milovanović) *Moment-preserving spline approximation on finite intervals*, Numer. Math. 50, 503–518.
103. (with J. Wimp) *Computing the Hilbert transform of a Jacobi weight function*, BIT 27, 203–215.
104. (with H. J. Landau and G. V. Milovanović) *Polynomials orthogonal on the semicircle II*, Constructive Approx. 3, 389–404.
105. (with M. A. Kovačević and G. V. Milovanović) *The numerical evaluation of singular integrals with coth-kernel*, BIT 27, 389–402.
106. *A conjectured inequality for Hermite interpolation at the zeros of Jacobi polynomials*, Problem 87-7, SIAM Rev. 29, 297–298.

1988

107. *Gauss–Kronrod quadrature — a survey*, in *Numerical methods and approximation theory III* (G. V. Milovanović, ed.), 39–66, Faculty of Electronic Engineering, Univ. Niš, Niš.
108. (with S. E. Notaris) *Newton's method and Gauss–Kronrod quadrature*, in *Numerical integration III* (H. Brass and G. Hämmerlin, eds.), 60–71, Internat. Ser. Numer. Math. 85, Birkhäuser, Basel.
109. (with S. E. Notaris) *An algebraic study of Gauss–Kronrod quadrature formulae for Jacobi weight functions*, Math. Comp. 51, 231–248.
110. (with G. Inglese) *Lower bounds for the condition number of Vandermonde matrices*, Numer. Math. 52, 241–250.
111. (with T. J. Rivlin) *A family of Gauss–Kronrod quadrature formulae*, Math. Comp. 51, 749–754.

1989

112. *Orthogonality — conventional and unconventional — in numerical analysis*, in *Computation and control* (K. Bowers and J. Lund, eds.), 63–95, Progress in Systems and Control Theory 1, Birkhäuser, Boston.
113. *On the zeros of polynomials orthogonal on the semicircle*, SIAM J. Math. Anal. 20, 738–743.
114. (with S. E. Notaris) *Gauss–Kronrod quadrature formulae for weight functions of Bernstein–Szegö type*, J. Comput. Appl. Math. 25, 199–224. [Erratum, *ibid.* 27 (1989), 429.]

1990

115. *Some applications and numerical methods for orthogonal polynomials*, in *Numerical analysis and mathematical modelling* (A. Wakulicz, ed.), 7–19, Banach Center Publications 24, PWN Polish Scientific Publishers, Warsaw.
116. *Orthogonal polynomials on the semicircle*, in *Numerical analysis and mathematical modelling* (A. Wakulicz, ed.), 21–27, Banach Center Publications 24, PWN Polish Scientific Publishers, Warsaw.

117. *Computational aspects of orthogonal polynomials*, in *Orthogonal polynomials* (P. Nevai, ed.), 181–216, NATO ASI Series, Series C: Mathematical and Physical Sciences 294, Kluwer, Dordrecht.
118. *How (un)stable are Vandermonde systems?*, in *Asymptotic and computational analysis* (R. Wong, ed.), 193–210, Lecture Notes Pure Appl. Math. 124, Dekker, New York.
119. (with E. Tychopoulos and R.S. Varga) *A note on the contour integral representation of the remainder term for a Gauss–Chebyshev quadrature rule*, SIAM J. Numer. Anal. 27, 219–224.
120. (with A. Córdova and S. Ruscheweyh) *Vandermonde matrices on the circle: spectral properties and conditioning*, Numer. Math. 57, 577–591.
121. (with S. Li) *The remainder term for analytic functions of Gauss–Radau and Gauss–Lobatto quadrature rules with multiple end points*, J. Comput. Appl. Math. 33, 315–329.

1991

122. *Computational problems and applications of orthogonal polynomials*, in *Orthogonal polynomials and their applications* (C. Brezinski, L. Gori and A. Ronveaux, eds.), 61–71, IMACS Annals Comput. Appl. Math. 9, Baltzer, Basel.
123. *On the remainder term for analytic functions of Gauss–Lobatto and Gauss–Radau quadratures*, Rocky Mountain J. Math. 21, 209–226.
124. *A class of slowly convergent series and their summation by Gaussian quadrature*, Math. Comp. 57, 309–324.
125. *On certain slowly convergent series occurring in plate contact problems*, Math. Comp. 57, 325–338.
126. (with S. Li) *Gauss–Radau and Gauss–Lobatto quadratures with double end points*, J. Comput. Appl. Math. 34, 343–360.
127. *On the paper "A continued fraction approximation of the modified Bessel function $I_1(t)$" by P.R. Parthasarathy and N. Balakrishnan*, Appl. Math. Letters 4, 47–51.
128. *Quadrature formulae on half-infinite intervals*, BIT 31, 438–446.

1992

129. *Remainder estimates for analytic functions*, in *Numerical integration* (T. O. Espelid and A. Genz, eds.), 133–145, NATO ASI Series, Series C: Mathematical and Physical Sciences 357, Kluwer, Dordrecht.
130. *Applications and computation of orthogonal polynomials*, Proc. Eighteenth South African Sympos. Numer. Math. (S. Abelman, ed.), 47–71, Department of Computer Science, University of Natal, Durban.
131. *Spline approximation and quadrature formulae*, Atti Sem. Mat. Fis. Univ. Modena 40, 169–182.
132. *On mean convergence of extended Lagrange interpolation*, J. Comput. Appl. Math. 43, 19–35.

1993

133. *The spiral of Theodorus, special functions, and numerical analysis*, Supplement A in *Spirals: from Theodorus to chaos* by P. J. Davis, 67–87, A K Peters, Boston.
134. (with S. Li) *A set of orthogonal polynomials induced by a given orthogonal polynomial*, Aequationes Math. 46, 174–198.
135. *Is the recurrence relation for orthogonal polynomials always stable?*, BIT 33, 277–284.
136. *On the computation of generalized Fermi–Dirac and Bose–Einstein integrals*, Comput. Phys. Comm. 74, 233–238.
137. *Gauss-type quadrature rules for rational functions*, in *Numerical integration IV* (H. Brass and G. Hämmerlin, eds.), 111–130, Internat. Ser. Numer. Math. 112, Birkhäuser, Basel.
138. (with S. E. Notaris) *Problem 6*, in *Numerical integration IV* (H. Brass and G. Hämmerlin, eds.), 379–380, Internat. Ser. Numer. Math. 112, Birkhäuser, Basel.

1994

139. *Summation of slowly convergent series*, in *Numerical analysis and mathematical modelling* (A. Wakulicz, ed.), 7–18, Banach Center Publications 29, PWN Polish Scientific Publishers, Warsaw.
140. *Applications and computation of orthogonal polynomials*, in *Advances in computational mathematics: New Delhi, India* (H. P. Dikshit and C. A. Micchelli, eds.), Series in Approximations and Decompositions 4, World Scientific, Singapore.
141. *Algorithm 726: ORTHPOL — a package of routines for generating orthogonal polynomials and Gauss-type quadrature rules*, ACM Trans. Math. Software 20, 21–62; Remark on Algorithm 726, *ibid.* 24 (1998), 355.
142. *Reflections and recollections*, in *Approximation and computation: a festschrift in honor of Walter Gautschi* (R. V. M. Zahar, ed.), xvii–xxxv, Internat. Ser. Numer. Math. 119, Birkhäuser, Basel.

1995

143. *The work of Philip Rabinowitz on numerical integration*, Numer. Algorithms 9, 199–222.
144. *Luigi Gatteschi's work on special functions and numerical analysis*, in *special functions* (G. Allasia, ed.), *Annals Numer. Math.* 2, 3–19.
145. (with M. Zhang) *Computing orthogonal polynomials in Sobolev spaces*, Numer. Math. 71, 159–183.

1996

146. *Orthogonal polynomials: applications and computation*, Acta Numerica 1996 (A. Iserles, ed.), 45–119, Cambridge University Press, Cambridge.
147. (with S. Li) *On quadrature convergence of extended Lagrange interpolation*, Math. Comp. 65, 1249–1256.
148. (with S. E. Notaris) *Stieltjes polynomials and related quadrature formulae for a class of weight functions*, Math. Comp. 65, 1257–1268.

1997

149. (with J. Waldvogel) *Contour plots of analytic functions*, Ch. 25 in *Solving problems in scientific computing using Maple and Matlab* (W. Gander and J. Hřebíček, eds.), 3d ed., 359–372, Springer, Berlin. [Chinese translation by China Higher Education Press and Springer, 1999; Portuguese translation of 3d ed. by Editora Edgard Blücher Ltda, São Paulo, 2001; Russian translation of 4th ed. by Vassamedia, Minsk, Belarus, 2005.]
150. *The computation of special functions by linear difference equations*, in *Advances in difference equations* (S. Elaydi, I. Győri, and G. Ladas, eds.), 213–243, Gordon and Breach, Amsterdam.
151. On the computation of special Sobolev-type orthogonal polynomials, in *The heritage of P. L. Chebyshev: a festschrift in honor of the 70th birthday of T. J. Rivlin* (C. A. Micchelli, ed.), Ann. Numer. Math. 4, 329–342.
152. *Moments in quadrature problems*, in *Approximation theory and applications: a collection of papers to commemorate the Cornelius Lanczos centennial* (E. L. Ortiz and T. J. Rivlin, eds.), Comput. Math. Appl. 33, 105–118.
153. (with A. B. J. Kuijlaars) *Zeros and critical points of Sobolev orthogonal polynomials*, J. Approx. Theory 91, 117–137.
154. (with G. V. Milovanović) *s-orthogonality and construction of Gauss–Turán-type quadrature formulae*, J. Comput. Appl. Math. 86, 205–218.

1998

155. *The incomplete gamma functions since Tricomi*, in *Tricomi's ideas and contemporary applied mathematics*, 203–237, Atti Convegni Lincei 147, Accademia Nazionale dei Lincei, Roma.
156. *Ostrowski and the Ostrowski prize*, Math. Intelligencer 20, 32–34. [Revised and translated into German, Uni Nova 87 (2000), 60–62, Universität Basel.]

1999

157. *Orthogonal polynomials and quadrature*, Electron. Trans. Numer. Anal. 9, 65–76.
158. *A note on the recursive calculation of incomplete gamma functions*, ACM Trans. Math. Software 25, 101–107.
159. *Algorithm 793: GQRAT — Gauss quadrature for rational functions*, ACM Trans. Math. Software 25, 213–239.

2000

160. (with W. Gander) *Adaptive quadrature — revisited*, BIT 40, 84–101.
161. (with L. Gori and F. Pitolli) *Gauss quadrature for refinable weight functions*, Appl. Comput. Harmon. Anal. 8, 249–257.
162. (with L. Gori and M. L. Lo Cascio) *Quadrature rules for rational functions*, Numer. Math. 86, 617–633.
163. *High–order Gauss–Lobatto formulae*, Numer. Algorithms 25, 213–222.
164. *Gauss–Radau formulae for Jacobi and Laguerre weight functions*, Math. Comput. Simulation 54, 403–412. [Reprinted in *Computational Science, Mathematics and Software* (R. F. Boisvert and E. Houstis, eds.), 237–248, Purdue University Press, West Lafayette, IN, 2002.]

2001

165. *Remark: "Barycentric formulae for cardinal (SINC-) interpolants by Jean-Paul Berrut,"* Numer. Math. 87, 791–792.
166. (with J. Waldvogel) *Computing the Hilbert transform of the generalized Laguerre and Hermite weight functions*, BIT 41, 490–503.
167. *The use of rational functions in numerical quadrature*, J. Comput. Appl. Math. 133, 111–126.

2002

168. *Gauss quadrature approximations to hypergeometric and confluent hypergeometric functions*, J. Comput. Appl. Math. 139, 173–187.
169. *Computation of Bessel and Airy functions and of related Gaussian quadrature formulae*, BIT 42, 110–118.
170. *The interplay between classical analysis and (numerical) linear algebra — a tribute to Gene H. Golub*, Electron. Trans. Numer. Anal. 13, 119–147.
171. *Alessandro M. Ostrowski (1893–1986). La sua vita e le opere*, Boll. Docenti Matem. 45, 9–19.

2003

172. (with F. E. Harris and N. M. Temme) *Expansions of the exponential integral in incomplete gamma functions*, Appl. Math. Lett. 16, 1095–1099.

2004

173. *Generalized Gauss–Radau and Gauss–Lobatto formulae*, BIT 44, 711–720.

2005

174. *Orthogonal polynomials (in Matlab)*, J. Comput. Appl. Math. 178, 215–234.
175. *The Hardy–Littlewood function: an exercise in slowly convergent series*, J. Comput. Appl. Math. 179, 249–254.

176. *Computing polynomials orthogonal with respect to densely oscillating and exponentially decaying weight functions and related integrals*, J. Comput. Appl. Math. 184, 493–504.
177. *A historical note on Gauss–Kronrod quadrature*, Numer. Math. 100, 483–484.
178. *Numerical quadrature computation of the Macdonald function for complex orders*, BIT Numer. Math. 45, 593–603.

2006

179. *Orthogonal polynomials, quadrature, and approximation: computational methods and software (in Matlab)*, in *Orthogonal polynomials and special functions — computation and applications* (F. Marcellán and W. Van Assche, eds.), 1–77, Lecture Notes Math. 1883, Springer, Berlin.
180. *The circle theorem and related theorems for Gauss-type quadrature rules*, Electron. Trans. Numer. Anal. 25, 129–137.
181. *Computing the Kontorovich–Lebedev integral transforms and their inverses*, BIT Numer. Math. 46, 21–40.

2007

182. (with P. Leopardi) *Conjectured inequalities for Jacobi polynomials and their largest zeros*, Numer. Algorithms 45, 217–230.
183. *Leonhard Eulers Umgang mit langsam konvergenten Reihen*, Elem. Math. 62, 174–183.
184. *Commentary, by Walter Gautschi,* in *Milestones in matrix computation: selected works of Gene H. Golub, with commentaries* (R. H. Chan, Ch. Greif, and D. P. O'Leary, eds.), Ch. 22, 345–358, Oxford University Press, New York.
185. *A guided tour through my bibliography*, Numer. Algorithms 45, 11–35.

2008

186. *On Euler's attempt to compute logarithms by interpolation: a commentary to his letter of February 16, 1734 to Daniel Bernoulli*, J. Comput. Appl. Math. 219, 408–415.
187. *Leonhard Euler: his life, the man, and his work*, SIAM Rev. 50, 3–33. [Also published in *ICIAM 07,* 6^{th} *International Congress on Industrial and Applied Mathematics, Zürich, Switzerland, 16–20 July 2007* (R. Jeltsch and G. Wanner, eds.), 447–483, European Mathematical Society, Zürich, 2009. Chinese translation in *Mathematical Advance in Translation* (2008)(2–3).]
188. *The numerical evaluation of a "challenging" integral*, Numer. Algorithms 49, 187–194.
189. (with C. Giordano) *Luigi Gatteschi's work on asymptotics of special functions and their zeros*, Numer. Algorithms 49, 11–31.
190. *On a conjectured inequality for the largest zero of Jacobi polynomials*, Numer. Algorithms 49, 195–198.

2009

191. *On conjectured inequalities for zeros of Jacobi polynomials*, Numer. Algorithms 50, 93–96.
192. *New conjectured inequalities for zeros of Jacobi polynomials*, Numer. Algorithms 50, 293–296.
193. *How sharp is Bernstein's inequality for Jacobi polynomials?*, Electr. Trans. Numer. Anal. 36, 1–8.
194. *High-order generalized Gauss–Radau and Gauss–Lobatto formulae for Jacobi and Laguerre weight functions*, Numer. Algorithms 51, 143–149.
195. *Variable-precision recurrence coefficients for nonstandard orthogonal polynomials*, Numer. Algorithms 52, 409–418.

2010

196. *Alexander M. Ostrowski (1893–1986): his life, work, and students*, in `math.ch-/100` *Swiss Mathematical Society 1910–2010* (B. Colbois, C. Riedtmann, and V. Schroeder, eds.), 257–278, European Mathematical Society, Zürich.
197. *The spiral of Theodorus, numerical analysis, and special functions*, J. Comput. Appl. Math. 235, 1042–1052.
198. *Gauss quadrature routines for two classes of logarithmic weight functions*, Numer. Algorithms 55, 265–277.

2011

199. *The Lambert W-functions and some of their integrals: a case study of high-precision computation*, Numer. Algorithms 57, 27–34.
200. *Optimally scaled and optimally conditioned Vandermonde and Vandermonde-like matrices*, BIT Numer. Math. 51, 103–125.
201. *My collaboration with Gradimir V. Milovanović*, in *Approximation and computation — in honor of Gradimir V. Milovanović* (W. Gautschi, G. Mastroianni, and Th. M. Rassias, eds.), 33–43, Springer Optim. Appl. 42, Springer, Dordrecht.
202. *Experimental mathematics involving orthogonal polynomials*, in *Approximation and computation — in honor of Gradimir V. Milovanović* (W. Gautschi, G. Mastroianni, and Th. M. Rassias, eds.), 117–134, Springer Optim. Appl. 42, Springer, Dordrecht.
203. *Remark on "New conjectured inequalities for zeros of Jacobi polynomials" by Walter Gautschi, Numer. Algorithms 50, 293–296 (2009)*, Numer. Algorithms 57, 511.

2012

204. *Numerical integration over the square in the presence of algebraic/logarithmic singularities with an application to aerodynamics*, Numer. Algorithms 61, 275–290.
205. *Sub-range Jacobi polynomials*, Numer. Algorithms 61, 649–657.

2013

206. *Repeated modifications of orthogonal polynomials by linear divisors*, Numer. Algorithms 63, 369–383.
207. *Neutralizing nearby singularities in numerical quadrature*, Numer. Algorithms, DOI 10.1007/s11075-012-9672-9.
208. *A brief summary of my scientific work and highlights of my career*, in *Walter Gautschi — selected works with commentaries, Vol. 1* (C. Brezinski and A. Sameh, eds.), 9–17, Birkhäuser, Boston, MA.

2014

209. *Kommentar (Interpolation des Logarithmus) zum Brief Leonhard Eulers an Daniel Bernoulli vom 16.(27.)2.1734*, in *Briefwechsel Eulers mit Daniel Bernoulli*, Opera Omnia IVA/3, Birkhäuser, Basel, to appear.
210. *Interpolation before and after Lagrange*, Rend. Semin. Mat. Univ. Politec. Torino, to appear.
211. *High-precision Gauss–Turán quadrature rules for Laguerre and Hermite weight functions*, Numer. Algorithms, to appear.

Part II

Commentaries

In all commentaries, reference numbers preceded by "GA" refer to the numbers in the list of Gautschi's publications; see Section 4. Numbers in boldface type indicate that the respective papers are included in these selected works.

5

Numerical conditioning

Nicholas J. Higham

A theme running through Gautschi's work is numerical conditioning. His many papers on this topic fall broadly into two categories: those on conditioning of Vandermonde matrices and those on conditioning of polynomials.

5.1. Conditioning of Vandermonde matrices

A Vandermonde matrix has the form

$$V_n = V(x_1, x_2, \dots, x_n) = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \\ \vdots & \vdots & & \vdots \\ x_1^{n-1} & x_2^{n-1} & \dots & x_n^{n-1} \end{bmatrix} \in \mathbb{C}^{n\times n},$$

where $x_1, \dots, x_n \in \mathbb{C}$. It is worth noting that in Gautschi's papers the nodes x_i are always indexed from 1, as here, whereas they tend to be indexed from 0 in papers concerned with numerical solution of Vandermonde systems. Vandermonde matrices have long been of interest in linear algebra and numerical analysis because of the explicit formula for the determinant, $\prod_{1\le j<i\le n}(x_i - x_j)$, the fact that the inverse can be obtained from explicit formulae (see Traub [7, Sec. 14] for a short historical survey), and the general ill conditioning of Vandermonde matrices, all of which make them useful in classroom exercises and as test matrices for computational algorithms.

In a long sequence of papers starting in 1962, Gautschi investigated the conditioning of Vandermonde matrices, obtaining upper and lower bounds as well as results on the optimal placement of the x_i to minimize the condition number. The condition number in question is the matrix condition number with respect to inversion in the ∞-norm: for nonsingular $A \in \mathbb{C}^{n\times n}$, $\kappa_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty$, where $\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^n |a_{ij}|$. The original motivation for this work came from

C. Brezinski and A. Sameh (eds.), *Walter Gautschi, Volume 1: Selected Works with Commentaries*, Contemporary Mathematicians, DOI 10.1007/978-1-4614-7034-2_5,

experiences in computing Gaussian quadrature rules from moments of the weight function, in which possibly ill-conditioned confluent Vandermonde matrices arise.

The Vandermonde matrix V_n is nonsingular precisely when the nodes x_i are distinct. When some of the nodes coincide the appropriate form of V_n for practical applications such as Hermite interpolation is as follows: the x_i are ordered so that equal points are contiguous and if x_i is repeated k times then V_n has columns comprising $[1, x, x^2, \ldots, x^{n-1}]$ and its first $k-1$ derivatives, all evaluated at x_i. Gautschi's papers focus on the cases $k = 1$ and $k = 2$.

In [GA**16**], Gautschi obtains the upper bound

$$\|V_n^{-1}\|_\infty \le \max_i \prod_{j \ne i} \frac{1 + |x_j|}{|x_i - x_j|}, \tag{5.1}$$

and shows that there is equality when $x_j = |x_j| e^{i\theta}$ for all j with a fixed θ, so in particular when $x_j \ge 0$ for all j. A bound for the confluent case is also given, and a slightly sharper bound was obtained the following year in [GA**19**]. The third paper [GA**62**] in this "On inverses ..." series appeared in 1978 and gives the lower bound

$$\|V_n^{-1}\|_\infty \ge \max_i \prod_{j \ne i} \frac{\max(1, |x_j|)}{|x_i - x_j|}, \tag{5.2}$$

which differs from the upper bound in (5.1) by at most a factor 2^{n-1}. A practical application of these early results is in [GA34], where they are applied to a Vandermonde system arising in numerical inversion of the Laplace transform.

In [GA**51**], Gautschi specializes to the case where the nodes are located symmetrically with respect to the origin. In particular, he shows that while for nodes equispaced on $[0, 1]$,

$$\kappa_\infty(V_n) \sim \frac{1}{\pi} e^{-\frac{\pi}{4}} e^{\frac{n}{4}(\pi + 2\log 2)} \approx \frac{1}{\pi} e^{-\frac{\pi}{4}} (3.1)^n,$$

for the Chebyshev points $x_i = \cos(\frac{2i-1}{2n}\pi)$ the rate of growth is much slower:

$$\kappa_\infty(V_n) \sim \frac{3^{3/4}}{4} (1 + \sqrt{2})^n.$$

A natural question is how to choose the nodes to minimize the condition number. This is considered in [GA**52**], where some characterizations of the optimal nodes are obtained and optimal configurations either symmetric about the origin or nonnegative are computed explicitly for small n.

The 1988 paper [GA**110**] returns to lower bounds, showing that for nonnegative nodes, $\kappa_\infty(V_n) > 2^{n-1}$ for $n \ge 2$, while for real nodes symmetric about the origin, $\kappa_\infty(V_n) > 2^{n/2}$ for $n > 2$. Even larger lower bounds for the 2-norm condition number were subsequently obtained by Beckermann [1]:

$$\kappa_2(V_n) \geq \left(\frac{2}{n}\right)^{1/2} (1+\sqrt{2})^{n-2}, \quad \kappa_2(V_n) \geq \frac{1}{2n^{1/2}} [(1+\sqrt{2})^{2(n-1)} + (1+\sqrt{2})^{-2(n-1)}],$$

for arbitrary nodes and nonnegative nodes, respectively.

These exponential lower bounds are alarming, but they do not necessarily rule out the use of Vandermonde matrices in practice. One of the reasons is that there exist specialized algorithms for solving Vandermonde-systems whose accuracy is not dependent on the condition number κ, and which in some cases can be proved to be highly accurate. The first such algorithm is an $O(n^2)$ operation algorithm for solving $V_n x = b$ of Björck and Pereyra [3], whose error analysis was given by Higham [5]. There is now a long list of generalizations of this algorithm in various directions, of which we mention just Demmel and Koev [4] and Bella et al. [2]; various other algorithms up to 2002 are described or cited in the chapter "Vandermonde systems" in [6].

Another important observation is that the exponential lower bounds are for real nodes. For complex nodes, V_n can be much better conditioned. Indeed V_n is $n^{1/2}$ times a *unitary* matrix when the x_i are the roots of unity. Moreover, it is shown in [GA**120**] that when the nodes are $e^{2\pi i c_j}$, with the c_j from the Van der Corput sequence, then $\kappa_2(V_n) < (2n)^{1/2}$.

The matrix V_n corresponds to a monomial basis for the space of polynomials of degree up to $n-1$. Other bases can be chosen—in particular, ones built from polynomials that satisfy a three-term recurrence (and in particular, orthogonal polynomials). The latter *Vandermonde-like* matrices can be much better conditioned than Vandermonde matrices, as shown in [GA**83**].

Gautschi gives an excellent summary of his work on Vandermonde matrices up to 1990 in [GA**118**]. In his most recent contribution on Vandermonde matrices [GA**200**], he refines his earlier results and computations.

5.2. Conditioning of polynomials

The papers [GA**43**, GA**45**, GA**64**, GA**66**, GA**98**] are concerned with several aspects of the conditioning of polynomials: the conditioning of the bases, the conditioning of zeros of the polynomial, and the conditioning of the problem of generating an orthogonal polynomial from moments of the weight function. For this last topic, see also Section 11.2.1, Vol. 2.

References

[1] Bernhard Beckermann. The condition number of real Vandermonde, Krylov and positive definite Hankel matrices. *Numer. Math.*, 85(4):553–577, 2000.

[2] T. Bella, Y. Eidelman, I. Gohberg, I. Koltracht, and V. Olshevsky. A fast Björck–Pereyra-type algorithm for solving Hessenberg-quasiseparable-Vandermonde systems. *SIAM J. Matrix Anal. Appl.*, 31(2):790–815, 2009.

[3] Åke Björck and Victor Pereyra. Solution of Vandermonde systems of equations. *Math. Comp.*, 24(112):893–903, 1970.
[4] James Demmel and Plamen Koev. The accurate and efficient solution of a totally positive generalized Vandermonde linear system. *SIAM J. Matrix Anal. Appl.*, 27(1): 142–152, 2005.
[5] Nicholas J. Higham. Error analysis of the Björck–Pereyra algorithms for solving Vandermonde systems. *Numer. Math.*, 50(5):613–632, 1987.
[6] Nicholas J. Higham. *Accuracy and stability of numerical algorithms*. Second edition, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002, xxx+680 pp. ISBN 0-89871-521-0.
[7] J. F. Traub. Associated polynomials and uniform methods for the solution of linear problems. *SIAM Rev.*, 8(3):277–301, 1966.

6

Special functions

Javier Segura

The collection of papers by Walter Gautschi dealing with special functions has, of course, connections with other sections in these volumes. First, we have to mention the article [GA**29**], which is included in Section 21, Vol. 3, dedicated to difference equations, where the conditioning of three-term recurrence relations is analyzed and methods of computation using recurrence relations are developed; for a more recent review, see [GA**150**]. Recurrence relations are basic tools for computing special functions, particularly functions of hypergeometric type. In [GA**29**], also the relation between the existence of a minimal solution for the recurrence and the convergence of the associated continued fraction is discussed. Reference [GA**29**] is a pioneering and influential paper in the field of special functions, and it is a highly cited paper (316 citations as of now).

Before one starts using recurrence relations for computing special functions, the main question to be examined is the asymptotic conditioning of the recurrence. This is one of the subjects developed in [GA**29**] for three-term recurrence relations. However, as Gautschi points out in [GA**61**], transitory effects may take place in some recurrences which cause the associated continued fraction to seemingly converge, but to a wrong value. This behavior was revisited recently in [6], where additional examples of such transitory behavior were found. An important conclusion to be drawn is that the study of asymptotic conditioning of the recurrence is not sufficient to ensure a stable use of recurrences, and that one should also pay attention to possible transitory effects.

Gautschi has developed a number of algorithms for the computation of special functions by Gaussian quadrature; these algorithms benefit from his pioneering work on the computation of Gauss quadratures (see, for instance, [GAB3] and Section 15.1, Vol. 2). For additional computer algorithms, see Section 23, Vol. 3. These algorithms have had a great influence on software developers for special functions.

C. Brezinski and A. Sameh (eds.), *Walter Gautschi, Volume 1: Selected Works with Commentaries*, Contemporary Mathematicians, DOI 10.1007/978-1-4614-7034-2_6,

In addition to the articles dealing with methods of computation, Gautschi published a number of papers dealing with inequalities for special functions and their zeros. The best-known, and most cited, of all is [GA**9**], containing the famous Gautschi's inequality among other interesting results.

Below we summarize the contributions contained in the papers of this section. Apart from [GA**61**], which, as already mentioned, is a paper dealing with the conditioning of some recurrence relations, we are dividing the contributions in two categories: computation of special functions, and inequalities for special functions and their zeros.

6.1. Computation of special functions

The methods for computing special functions are varied. Basic methods include the use of series (convergent or asymptotic), recurrence relations, and continued fractions. For computing a function efficiently, one usually needs to combine several of these methods. Gautschi's contributions have a wide scope, but probably the best-known are those regarding the computation of incomplete gamma functions and related functions (exponential integrals, error function). In these papers, all of the above-mentioned methods find applications.

Numerical quadratures have not been widely used for computing special functions, even though this is also an interesting possibility. An advantage of numerical quadrature is that, when a suitable integral representation is at hand, quadrature may provide methods with a large range of validity or even, in some cases, may be the standalone method. Gautschi's main interest in this area is the application of Gaussian quadrature.

6.1.1. Exponential integrals, incomplete gamma functions, and the error function

In this subsection we consider six contributions dealing with the computation of incomplete gamma functions and related functions.

In the first, [GA**10**], Gautschi considers an expansion for the exponential integral

$$\int_1^{+\infty} e^{-xt} t^{-n} dt$$

valid for large positive n and arbitrary $x > 0$, complete with error bounds. The expansion generalizes the four-term expansion derived by G. Blanch in [4].

In [GA**13**], methods for computing the repeated integrals of the error function,

$$\mathrm{i}^n\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^{+\infty} \frac{(t-x)^n}{n!} e^{-t^2} dt,$$

are considered. These are particular cases of parabolic cylinder functions (solutions of $y''(x) - (x^2/4 + a)y(x) = 0$) up to a trivial change of variables, and they satisfy a three-term recurrence relation

$$y_{n+1} + a_n(x)y_n + b_n(x)y_{n-1} = 0.$$

It is shown that the functions are a minimal solution of the recurrence, which means that $\lim_{n\to+\infty} \mathrm{i}^n\mathrm{erfc}(x)/y_n(x) = 0$ for any other linearly independent solution $y_n(x)$. The function can therefore be computed by a backward application of the recurrence (Miller's method). The paper analyzes in detail the convergence of the method and, by bounding the error, values of n are given with which to start the backward recurrence. This is a first instance of techniques developed in more detail, and with a larger scope, in the later paper [GA**29**].

A method for computing the complex error function $w(z) = \exp(-z^2)\mathrm{erfc}(-iz)$ is developed in [GA**39**] (see also the accompanying software [GA36]). This function, also called Faddeeva or Voigt function, is important in many physical applications. It is not surprising that this article received a high number of citations (155 at this time), many of them coming from papers in physics. From a mathematical point of view, it is interesting to observe that, contrary to what is common in the evaluation of special functions (and particularly for complex variables), a single numerical scheme can be used to compute the function for all values of the variable. A central role is played by a Laplace continued fraction, convergent for complex (nonreal) z, which is equivalent to Gauss–Hermite quadrature for

$$\frac{i}{\pi}\int_{-\infty}^{+\infty} \frac{e^{-t^2}}{z-t}dt$$

and which provides asymptotic approximations as $z \to \infty$ (also on the real line). This is combined with the use of Taylor series, which can be computed recursively. The resulting procedure is such that, as $|z|$ becomes large, it turns into the computation of the Laplace continued fraction.

Reference [GA**68**] (see also [GA69] for an algorithm) describes methods for computing the incomplete gamma functions

$$P(a,x) = \frac{1}{\Gamma(a)}\int_0^x t^{-a}e^{-t}dt, \quad Q(a,x) = \frac{1}{\Gamma(a)}\int_x^{+\infty} t^{-a}e^{-t}dt,$$

for moderate real values of a and x. These functions are important in many applications, particularly in statistics; they are related (see, e.g., [12, Ch. 8]) to the central accumulated χ^2 distributions by $P(\chi^2|\nu) = P(\nu/2, \chi^2/2)$, $Q(\chi^2|\nu) = Q(\nu/2, \chi^2/2)$. The method of computation combines the use of Taylor series with a continued fraction for $Q(a,x)$ together with the use of recursion. As mentioned in [GA**68**], the range of computation can be enlarged by using the asymptotic approximations of Temme [17]. The methods in [GA**68**] continue to be benchmark methods for

computing incomplete gamma functions, as can be seen, for instance, from [7]. Incidentally, the commentator at this moment is working (in collaboration with A. Gil and N. M. Temme) on numerical methods for the computation and inversion of central and noncentral χ^2 distributions, and for the central case, Gautschi's ideas are as useful today as they were 34 years ago. By some mistake, this reference is not included in the ISI database. However, a search in Google Scholar reveals 56 citations for [GA**68**] and 28 citations for the companion paper [GA69].

Finally, we have the review paper [GA**155**], which not only reviews methods of computation for incomplete gamma functions, but also discusses applications, asymptotics, inequalities, and more. This review paper is an important source of information for these functions and, according to Google Scholar, has been cited 66 times.

6.1.2. Computing special functions by Gaussian quadrature

Gautschi has developed a variety of methods for computing special functions based on Gaussian quadrature, both classical (like Gauss–Jacobi or Gauss–Laguerre) and nonclassical.

In [GA**168**], Gautschi considers the computation of Gauss and confluent hypergeometric functions using Gauss–Jacobi quadrature. In the confluent case, the starting point is the integral representation, valid for $b > a > 0$ and z real or complex,

$$M(a,b,z) = \frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)} \int_0^1 e^{zt}(1-t)^{b-a-1}t^{a-1}dt.$$

For this parameter range ($b > a > 0$), we thus have an integral with a Jacobi weight $(1-t)^{b-a-1}t^{a-1}$, and Gauss–Jacobi quadrature appears to be a natural method of evaluation. The paper analyzes the respective error and compares the estimated number of nodes required for a given accuracy with the number experimentally observed. In the Gauss hypergeometric case, similar ideas apply for the computation of ${}_2\mathrm{F}_1(a,b;c;z)$ when $c > b > 0$. This gives methods of evaluation valid for wide ranges of the parameters, subject to the restrictions imposed in the respective integral representations. Without doubt, these methods should be considered when constructing numerical algorithms for hypergeometric functions and are a viable alternative to methods based on Taylor series and connection formulas (see, for instance, [10, Sections 2.3.1–2.3.3]).

In [GA**169**], Gautschi considers the computation of Gauss quadrature rules with nonstandard weights, namely $K_\nu(x)$ (modified Bessel function) and $\mathrm{Ai}(x)$ (Airy function). For this purpose, the coefficients of the recurrence relation for the associated orthogonal polynomials are numerically evaluated using the Stieltjes–Gautschi procedure (cf. Section 11.2.2, Vol. 2), which in turn requires the evaluation of a number of integrals of the same type (that is, with modified Bessel functions or Airy functions as weight functions). For this, one first needs a numerical method for computing the Bessel and Airy functions and then an alternative quadrature to evaluate

the integrals in question. Starting with the latter, the approximation consists in dividing the interval of integration in three subintervals where different Gaussian quadratures are used (for instance, a combination of Gauss–Jacobi, Gauss–Legendre and Gauss–Laguerre for the case of the modified Bessel weight). With regard to methods of computing the modified Bessel function and the Airy function, integral representations are used which are suitable for Gauss–Laguerre quadrature. For instance,

$$\mathrm{Ai}(x) = \frac{1}{\sqrt{\pi}} \frac{\zeta^{-1/6} e^{-\zeta}}{(48)^{1/6}\Gamma(\frac{5}{6})} \int_0^\infty \left(2 + \frac{t}{\zeta}\right)^{-1/6} t^{-1/6} e^{-t} dt, \quad \zeta = \frac{2}{3} x^{\frac{3}{2}},$$

can be dealt with using Gauss–Laguerre quadrature with parameter $\alpha = -1/6$. For modified Bessel functions $K_\nu(x)$ similar representations are available (the Airy case is related to the Bessel case with $\nu = 1/3$). These methods of computation can be extended to the complex plane. In particular, for the Airy function they can be used for $|\arg(z)| < 2\pi/3$, with a modification when $\arg(z) \to \pm 2\pi/3$ to avoid a singularity close to the real axis; with this modification, the method was used in the algorithm [9] for intermediate values of $|z|$.

The references [GA**178**, GA**199**] contain two more examples of applying classical and nonclassical Gaussian quadrature rules to the computation of special functions. In the former, methods for the computation of modified Bessel functions of complex orders, $K_{\alpha+i\beta}(x)$, are discussed; a nonclassical Gaussian quadrature with a double exponential weight is chosen, which is suggested by the integral representations used for computing these functions. In the latter, the computation of some integrals of the Lambert W-functions are considered, which are computed by a combination of Gauss–Legendre quadrature and a nonclassical Gaussian quadrature with weight function $x^{\alpha-1}[\ln(1/x)]^{\alpha-\beta+1}$ on $[0, 1/e]$.

6.2. Inequalities

Gautschi's work also addresses questions involving inequalities and bounds for special functions and their zeros: inequalities for orthogonal polynomials and their zeros, and inequalities for gamma functions.

6.2.1. Orthogonal polynomials and their zeros

In [GA**72**] a method is developed for finding lower bounds for the largest zeros of orthogonal polynomials.

The idea is nice and simple and relies on two facts. The first is that $r_1 > r_2 > \ldots > r_n > 0$ and $u_k = \sum_{i=1}^n c_i r_i^k$ with c_i positive, letting $\sigma_k = u_{k+1}/u_k$, imply $\sigma_k < \sigma_{k+1} < r_1$ and $\lim_{k\to\infty} \sigma_k = r_1$. The second comes from applying the error formula for Gaussian quadrature.

If c_i, $i = 1, \ldots, n$, are the weights of an n-point Gaussian quadrature on the interval $[a, b]$ $(b > a > 0)$ with nonnegative integrable weight $w(x)$, r_i the zeros of the associated orthogonal polynomial of degree n, and denoting the moments by $m_k = \int_a^b x^k w(x)dx$, one has $m_{2n-1} = u_{2n-1}$ because Gaussian quadrature has degree of exactness $2n-1$ while, by a well-known expression for the error of Gaussian quadrature,

$$m_{2n} = u_{2n} + \int_a^b p_n(x)^2 w(x)dx,$$

with $p_n(x)$ the monic orthogonal polynomial with zeros r_i, $i = 1, ..., n$.

Combining these facts yields

$$r_1 > \frac{u_{2n}}{u_{2n-1}} = \frac{1}{m_{2n-1}} \left[m_{2n} - \int_a^b p_n(x)^2 w(x)dx \right].$$

For classical orthogonal polynomials, computing these quantities is a simple matter, and the paper shows that the idea is effective inasmuch as the bounds so obtained improve those that follow from Laguerre's theorem [16, p. 119].

Recent work of Gautschi deals with inequalities satisfied by the zeros of orthogonal polynomials ([GA**182**, GA**190**, GA**191**, GA**192**, GA**203**]). These papers combine analysis with experimental approaches in order to suggest new conjectured inequalities. The paper [GA**193**] also has an experimental flavor, where Gautschi investigates the sharpness of Bernstein's inequality for Jacobi polynomials $P_n^{(\alpha,\beta)}(\cos\theta)$ [5], valid for $|\alpha| \leq 1/2$ and $|\beta| \leq 1/2$; it also investigates the validity of the inequality in larger domains $(\alpha \geq -1/2, \beta \geq -1/2)$ concluding that it remains valid but with a somewhat larger constant. The results also lend support to the Erdélyi-Magnus-Nevai conjecture,

$$\max_{x\in[-1,1]} (1-x)^{\alpha+1/2}(1+x)^{\beta+1/2} \hat{P}_n^{(\alpha,\beta)}(x) = \mathcal{O}(\max[1, (\alpha^2+\beta^2)^{1/4}]),$$

where $\hat{P}_n^{(\alpha,\beta)}(x)$ is the normalized Jacobi polynomial, and suggest the best constant implied in the O-term.

6.2.2. Gamma functions

Gautschi proved a number of inequalities for gamma and related functions.

In the paper [GA**9**], two-sided inequalities for incomplete gamma functions are established. Related inequalities for the gamma function are named after Gautschi. These inequalities have received a great deal of attention in the mathematical community. Again, for some reason, the reference is not included in the ISI database, but, according to Google Scholar, the paper has been cited 109 times. A good number of papers exist improving or extending in some way Gautschi's inequality, which in its best-known form can be written as

$$x^{1-s} < \frac{\Gamma(x+1)}{\Gamma(x+s)} < (x+1)^{1-s}, \quad x > 0,\, 0 < s < 1.$$

It is not possible to cite all papers dealing with improvements or extension of these attractive inequalities, but, just to mention a few of them, we cite [13, 15, 14, 3, 2, 8].

The papers [GA**47**, GA**48**] deal with harmonic mean inequalities for gamma functions, for example the interesting inequality

$$\frac{2}{1/\Gamma(x) + 1/\Gamma(1/x)} \geq 1, \quad x > 0.$$

This type of inequalities also has attracted considerable attention; see for instance [1, 11].

References

[1] Horst Alzer. A harmonic mean inequality for the gamma function. *J. Comput. Appl. Math.*, 87(2):195–198, 1997.

[2] Horst Alzer. On some inequalities for the incomplete gamma function. *Math. Comp.*, 66(218):771–778, 1997.

[3] Joaquin Bustoz and Mourad E. H. Ismail. On gamma function inequalities. *Math. Comp.*, 47(176):659–667, 1986.

[4] G. Blanch. An asymptotic expansion for $E_n(x) = \int_1^\infty (e^{-xu}/u^n)du$. *NBS Applied Math. Series*, 37:61, 1954.

[6] Yunshyong Chow, L. Gatteschi, and R. Wong. A Bernstein-type inequality for the Jacobi polynomial. *Proc. Amer. Math. Soc.*, 121(3):703–709, 1994.

[6] Alfredo Deaño and Javier Segura. Transitory minimal solutions of hypergeometric recursions and pseudoconvergence of associated continued fractions. *Math. Comp.*, 76(258):879–901, 2007.

[7] Armido R. DiDonato and Alfred H. Morris, Jr. Computation of the incomplete gamma function ratios and their inverse. *ACM Trans. Math. Software*, 12(4):377–393, 1986.

[8] Neven Elezović, Carla Giordano, and Josip Pečarić. The best bounds in Gautschi's inequality. *Math. Inequal. Appl.*, 3(2):239–252, 2000.

[9] Amparo Gil, Javier Segura, and Nico M. Temme. Algorithm 819: AIZ, BIZ: two Fortran 77 routines for the computation of complex Airy functions. *ACM Trans. Math. Software*, 28(3):325–336, 2002.

[10] Amparo Gil, Javier Segura, and Nico M. Temme. *Numerical methods for special functions*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007. xiv+417 pp. ISBN: 978-0-898716-34-4.

[11] C. Giordano and A. Laforgia. Inequalities and monotonicity properties for the gamma function. In: Proceedings of the Fifth International Symposium on Orthogonal Polynomials, Special Functions and their Applications (Patras, 1999), *J. Comput. Appl. Math.*, 133(1–2):387–396, 2001.

[12] Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous univariate distributions*. Vol. 1, Second edition, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, New York, 1994, xxii+756 pp. ISBN: 0-471-58495-9.

[13] D. Kershaw. Some extensions of W. Gautschi's inequalities for the gamma function. *Math. Comp.*, 41(164):607–611, 1983.

[14] Andrea Laforgia. Further inequalities for the gamma function. *Math. Comp.*, 42(166): 597–600, 1984.

[15] Lee Lorch. Inequalities for ultraspherical polynomials and the gamma function. *J. Approx. Theory*, 40(2):115–120, 1984.

[16] Gábor Szegő. *Orthogonal polynomials*. American Mathematical Society, Providence, R.I., fourth edition, 1975. American Mathematical Society, Colloquium Publications, Vol. XXIII, xiii+432 pp.

[17] N. M. Temme. On the computation of the incomplete gamma functions for large values of the parameters. In *Algorithms for approximation (Shrivenham, 1985)*, 479–489, *Inst. Math. Appl. Conf. Ser. New Ser.*, 10, Oxford Univ. Press, New York, 1987.

7

Interpolation and approximation

Miodrag M. Spalević

In the papers collected here, Walter Gautschi makes vital contributions to the theory of interpolation and approximation. He considers attenuation factors in practical Fourier analysis, Padé approximants associated with Hamburger series, the convergence behavior of continued fractions with real elements, moment-preserving spline approximations, and the convergence of extended Lagrange interpolation. Further, he uses numerical computations to examine the validity of mathematical conjectures regarding zeros of Jacobi polynomials and weighted Newton–Cotes quadrature formulae.

7.1. Attenuation factors in practical Fourier analysis

Gautschi has written one contribution on this topic, the wonderful paper [GA**41**]. It is of great value in computational science and engineering, showing how the Fourier coefficients of a variety of interpolants and approximants to periodic data can be computed as economically as the Fast Fourier Transform; in other words, as economically as computing the Fourier coefficients of trigonometric interpolants. Regretfully, too few mathematicians, scientists and engineers have taken note of this work. The current Google Scholar citation count of 52 is respectable, but still modest. In this regard, the paper ranks only in the 27th position among Gautschi's articles.

Assume N real values f_μ $(\mu = 0, 1, \ldots, N-1)$ are given. These can be extended periodically by setting $f_{\mu+\ell N} := f_\mu$ $(\forall \ell \in \mathbf{Z})$, and can be viewed as function values at the equispaced points $x_\mu := 2\pi\mu/N$ of a 2π-periodic function f defined on $\mathbb{R}$. If $N = 2s + 1$ is odd, the discrete Fourier transform (DFT) yields the N Fourier coefficients,

$$\hat{c}_n := \frac{1}{N} \sum_{\mu=0}^{N-1} f_\mu e^{-inx_\mu} \qquad (-s \le n \le s)$$

C. Brezinski and A. Sameh (eds.), *Walter Gautschi, Volume 1: Selected Works with Commentaries*, Contemporary Mathematicians, DOI 10.1007/978-1-4614-7034-2_7,

of a unique trigonometric interpolant of degree s. Similarly, if $N = 2s$ is even, one obtains the coefficients of a suitably normalized unique interpolant of degree s. Moreover, if N is a power of 2, or, more generally, if N has only a few different small prime factors, the DFT can be implemented efficiently with the Fast Fourier Transform (FFT). Yet, the trigonometric interpolant is not always a useful approximant of the data because of the Gibbs phenomenon. A periodic spline interpolant, or some other approximant φ, may be preferable. These will typically have an infinite number of Fourier coefficients c_n that tend to 0 as $|n| \to \infty$, while $\hat{c}_n$ is N-periodic if evaluated for all n. For a broken-line interpolant it was pointed out already in 1898 by Oumoff, and for spline interpolants in 1928 by Eagle, that

$$c_n = \tau_n \hat{c}_n \qquad (\forall n) \tag{7.1}$$

with factors τ_n that depend only on the family of interpolants and not on the data. They are called *attenuation factors*. Also, of course, $|\tau_n| \to 0$ as $|n| \to \infty$.

In the paper [GA**41**], Gautschi develops a general theory of attenuation factors and some of their generalizations. The fundamental theorem states that, under mild conditions, Eqn. (7.1) holds if and only if the operator $P : \{f_\mu\} \mapsto \varphi$ is linear and translation invariant. A further basic result characterizes the structure of τ_n for families of interpolating functions. Numerous examples of suitable families and their attenuation factors are treated in detail, as are some more general cases in which Eqn. (7.1) must be modified. Further, as usual, Gautschi includes an informative historical account. Altogether, this is a highly interesting, most comprehensive, and eminently applicable paper.

Applications include the solution of boundary integral equations like those for numerical conformal mappings [8, 17] and of periodic Fredholm integral equations [3].

Regarding extensions of the theory, we mention Locher's detailed treatment [13] of attenuation factors for families of interpolants generated by translates of a single function, and Gutknecht's extension [9] of the attenuation factor theory to the multivariate case, including the treatment of box splines, which were of current interest. Further work on the multivariate case is due to ter Morsche [21] and Steidl [19].

7.2. Padé approximants associated with Hamburger series

Gautschi's paper on this topic is [GA**86**]. A Hamburger series is a formal power series in which the coefficients are moments of a bounded nondecreasing function λ defined on the real axis and having infinitely many points of increase; in particular, it is called a Stieltjes series if λ is supported on the nonnegative real axis. It is well known that the Padé approximants associated with a Stieltjes or Hamburger series are closely related to orthogonal polynomials relative to the measure $d\lambda$ or the measures $d\lambda_j(t) = t^j d\lambda(t)$. The latter are positive definite for a Stieltjes series and also for a Hamburger series if j is even. It is this connection with orthogonal

polynomials that is utilized in order to study the following three aspects of Padé approximants associated with Hamburger series:

(i) Conditions under which all entries of the Padé table are normal, i.e., each entry appears only once in the table.

(ii) In the case $d\lambda(t) = w(t)dt$, where $w(t)$ is a nonnegative weight function on a symmetric interval $[-a, a]$, $a > 0$, continuous on the open interval $(-a, a)$ and such that all moments μ_k exist with $\mu_0 > 0$, there is an in-depth analysis of the Padé approximants $f[n-1, n]$ down the first subdiagonal of the Padé table. Specifically, it is shown that their power series expansion coefficients — Gauss quadrature sums equal to, or approximating, the moments μ_k of $d\lambda$ — satisfy certain monotonicity properties under appropriate assumptions on w, properties that are known to be true unrestrictedly in the case of Stieltjes series.

(iii) Several methods for computing the Padé approximants, exploiting their connection with Gaussian quadrature. With the main concern being numerical stability, methods based on moments are avoided. Instead, assuming the recursion coefficients of the underlying orthogonal polynomials are known, one can generate not only the moments, but, more importantly, also the nodes and weights of the respective Gauss–Christoffel quadrature schemes. The same data, moreover, can be used to generate the recursion coefficients of $t^{2j}d\lambda(t)$, which are needed to compute the Padé approximants $f[n-1+2j, n]$. Three different methods are discussed to accomplish this: a numerical implementation of Christoffel's theorem, the modified Chebyshev algorithm (cf. Section 11.2.1, Vol. 2), and the QR algorithm.

Interestingly, the results obtained in (ii) have an important application regarding the sign of the remainder term $R_n(t^k)$ of Gaussian quadratures. This in turn can be used to obtain error bounds for Gauss quadrature of analytic functions, either by contour integration on circular contours in the complex plane, or by Hilbert space methods. The former was applied by Gautschi and others (cf. [GA**129**] and Section 15.3, Vol. 2), while the latter initially by Akrivis and others, and later by Notaris (cf. [16]).

7.3. Convergence behavior of continued fractions with real elements

When a sequence (e.g. the partial sums of a series or the convergents of a continued fraction), or an iterative method converges, one is most often interested in their asymptotic behavior. While this aspect is quite important, it does not give a clue as to when to stop the sequence or the iterative method. In the paper [GA**87**], Walter Gautschi discusses transient convergence rates of continued fractions whose convergence is guaranteed by Worpitzky's theorem. He wrote: "Properties of monotone behavior [of this rate] significantly add to the understanding of the quality of convergence". A series is said to be equivalent to a continued fraction if its partial sums are equal to the convergents of the continued fraction. The transient convergence

rates of a continued fraction can be derived from the successive terms of the equivalent series. Hence, results relating the behavior of the partial numerators of the continued fraction to its transient convergence rates are obtained. Gautschi also suggests a stopping criterion and illustrates its use by a numerical example.

7.4. Moment-preserving spline approximation

The moments of a function often have some very important physical meaning which should be preserved when one tries to approximate the function.

In [GA**89**] Gautschi discusses numerically stable methods for computing approximations to spherically symmetric distributions in $\mathbb{R}^d$, hence to functions of one variable, the radial distance $r \in \mathbb{R}_+$, with d being a parameter. An approach popular among physicists is to approximate the function by discrete functions, either linear combinations of Dirac delta functions or Heaviside step functions. For the Maxwell velocity distribution, Laframboise and Stauffer [12] and Calder, Laframboise and Stauffer [5] construct such approximations which are optimal in the sense of matching as many initial moments as possible. The resulting equations are solved in [12] by Prony's method and in [5] by a reduction to an eigenvalue problem involving Hankel matrices. Both methods are classical, but they are prone to severe ill-conditioning. This is corrected in [GA**89**] by applying a method based on Gaussian quadrature. For example, in the case of approximation by step functions, the moment-matching problem leads to a system of equations which determines a certain Gauss–Christoffel quadrature formula relative to a nonnegative measure depending on f and d. This Gauss formula can be constructed by numerically stable procedures (cf. Section 15.1, Vol. 2). In order to demonstrate the proposed methods, Gautschi generates numerical data required for approximating the distributions of Maxwell, Bose–Einstein, and Fermi–Dirac.

Subsequently, Gautschi (jointly with Milovanović) in [GA**100**] extends that work to approximation by splines of arbitrary degree m with n (variable) knots. The spline to be constructed is to match the first $2n$ moments of f. Under suitable assumptions on f, they show that the problem has a unique solution if and only if a certain n-point Gaussian quadrature formula exists corresponding to a (possibly nondefinite) moment functional or measure depending on f, d, and m. Existence and uniqueness is assured if f is completely monotonic on $[0, \infty)$.

Pointwise convergence of the given approximation process depends on a convergence property of the Gauss–Christoffel formula, since the error of the spline approximation can be expressed in terms of the remainder term of the respective Gauss–Christoffel formula applied to a certain function depending on d and m. Examples are given, including distributions from statistical mechanics.

Continuing the work in [GA**89**, GA**100**], Gautschi (jointly with Frontini and Milovanović) in [GA**102**]) discusses the analogous approximation problem on a *finite* interval, standardized to be the interval $[0, 1]$. In this case, the interpretation of

the independent variable as a radial distance is no longer meaningful, and the functions $f = f(t)$ are now simply functions of a real variable t on the interval $[0, 1]$. Additional constraints on the derivatives of the approximation at an endpoint of $[0, 1]$ may also be imposed. It is shown that, if the approximations exist, they can be represented in terms of the nodes and weights of generalized Gauss–Lobatto and Gauss–Radau formulae corresponding to appropriate moment functionals or measures. Pointwise convergence as $n \to \infty$, with fixed $m > 0$, is proved for functions f that are completely monotonic on $[0, 1]$ and is illustrated by numerical examples.

This work of Walter Gautschi inspired other researchers to investigate the possibility of similarly approximating a function f by spline functions of degree m and defects d_i; this turns out to be related to quadrature formulae with multiple knots (see [1, 7, 11, 15, 18]). An alternative approach to moment-preserving approximations can be found in [4].

7.5. Convergence of extended Lagrange interpolation

The problem suggested in [GA**132**, GA**147**], reminiscent of Kronrod's problem in numerical integration (cf. Section 14.1, Vol. 2), is to study convergence as $n \to \infty$ of polynomials interpolating a continuous function at the zeros of an nth-degree orthogonal polynomial and at $n+1$ additional points suitably interspaced. This is an interesting follow-up to the convergence theory of Erdős and Turán (involving only the n zeros of the orthogonal polynomial), but is still wide open for more definitive answers.

7.6. Experimental mathematics involving orthogonal polynomials

Numerical computation is an important tool for investigating mathematical ideas and examining the validity of mathematical conjectures. This is explored in [GA**202**] in the context of Jacobi polynomials and quadrature formulae.

7.6.1. Jacobi polynomials

There is a well-known result regarding the convergence of the zeros of the nth-degree Jacobi polynomial $P_n^{(\alpha,\beta)}$, $\alpha > -1$, $\beta > -1$, to the corresponding zeros of the Bessel function J_α as $n \to \infty$ (cf. [20, Theorem 8.1.2]). Domains in the (α, β)-plane are investigated in which convergence is monotone, which can be expressed by an inequality between the zeros of $P_n^{(\alpha,\beta)}$ and $P_{n+1}^{(\alpha,\beta)}$. This work is initially undertaken for the largest zero, and then extended to all zeros with a subsequent investigation of a modified inequality.

The well-known inequality of Bernstein for Legendre polynomials (cf. [2]) was generalized by Chow, Gatteschi and Wong to Jacobi polynomials $P_n^{(\alpha,\beta)}$ in the domain $|\alpha| \leq 1/2,\ |\beta| \leq 1/2$ (cf. [6]). Defining the sharpness of this inequality in a suitable manner, the degree of sharpness is examined both in the original domain of validity of the inequality and also, for a suitably modified inequality, in the larger domains $-1/2 < \alpha < s,\ -1/2 < \beta < s$ where $s > 1/2$.

7.6.2. Quadrature formulae

It is customary to call an interpolatory quadrature formula positive if all its weights are positive. This positivity property is examined in the following cases:

(a) The $(2n-1)$-point weighted Newton–Cotes formulae for the weight function $w(t) = (1-t)^{\alpha+1/2}(1+t)^{\beta+1/2},\ \alpha > -1,\ \beta > -1$, on $[-1,1]$, with the nodes being the zeros of the Jacobi polynomials $P_n^{(\alpha,\beta)}$ and $P_{n-1}^{(\alpha+1,\beta+1)}$. Positivity of these formulae is known when $\alpha = \beta = -1/2$ and was conjectured by Milovanović to hold for arbitrary $\alpha > -1$, $\beta > -1$ (cf. [14, Section 5.1.2]). Gautschi tested this conjecture numerically for many choices of α, β, n, and confirmed it in all cases.

(b) The generalized Gauss–Radau and Gauss–Lobatto formulae, which are quadrature formulae of Gauss type in which one or both boundary points have arbitrary multiplicity $r \geq 2$ (those with $r = 1$ being the usual Gauss–Radau and Gauss–Lobatto formulae). The positivity property in this case, conjectured by Gautschi [GA**173**, §§2.2,3.2], has subsequently been proved by Joulak and Beckermann (cf. [10]) even for the most general Gauss–Radau and Gauss–Lobatto formulae having not necessarily equal multiplicities at the boundary points.

7.7. Exotic weight functions

In [GA**181**], Gautschi considers the problem of computing integral transforms whose kernels are modified Bessel functions of complex order. Their integrands exhibit super-exponential decay at infinity, or dense oscillation at zero. Both types of behavior are captured in appropriate weight functions, the former in $w(t) = \exp(-\mathrm{e}^t),\ 0 \leq t < \infty$, and the latter in $w(t) = 1 + \sin(\beta \ln(1/t) + \gamma),\ 0 < t \leq 1$. In order to develop relevant quadrature formulae, one has to generate the respective orthogonal polynomials, which is accomplished via the Stieltjes–Gautschi procedure (cf. Section 11.2.2, Vol. 2) and the classical Chebyshev algorithm in symbolic variable-precision computation.

Acknowledgements. I would like to express my thanks to Martin Gutknecht, Sotirios Notaris, and the editors, for their help in writing this commentary.

References

[1] Ana Maria Acu. Moment preserving spline approximation on finite intervals and Chakalov–Popoviciu quadratures. *Acta Univ. Apulensis Math. Inform.*, 13:37–56, 2007.

[2] Serge Bernstein. Sur les polynômes orthogonaux relatifs à un segment fini. *J. de Math., IX. Sér.*, 10:219–286, 1931.

[3] Jean-Paul Berrut and Michèle Reifenberg. Numerical solution of periodic Fredholm integral equations by means of attenuation factors. *J. Integral Equations Appl.*, 9(1):1–20, 1997.

[4] Borislav Bojanov and Laura Gori. Moment preserving approximations. *Math. Balkanica (N. S.)*, 13(3–4):385–398, 1999.

[5] A. C. Calder, J. G. Laframboise, and A. D. Stauffer. Optimum step-function approximation of the Maxwell distribution (unpublished).

[6] Yunshyong Chow, L. Gatteschi, and R. Wong. A Bernstein-type inequality for the Jacobi polynomial. *Proc. Amer. Math. Soc.*, 121(3):703–709, 1994.

[7] Marco Frontini and Gradimir V. Milovanović. Moment-preserving spline approximation on finite intervals and Turán quadratures. *Facta. Univ. Ser. Math. Inform.*, 4:45–56, 1989.

[8] Martin H. Gutknecht. The evaluation of the conjugate function of a periodic spline on a uniform mesh. *J. Comput. Appl. Math.*, 16(2):181–201, 1986.

[9] Martin H. Gutknecht. Attenuation factors in multivariate Fourier analysis. *Numer. Math.*, 51(6):615–629, 1987.

[10] Hédi Joulak and Bernhard Beckermann. On Gautschi's conjecture for generalized Gauss–Radau and Gauss–Lobatto formulae. *J. Comput. Appl. Math.*, 233(3): 768–774, 2009.

[11] M. A. Kovačević and G. V. Milovanović. Spline approximation and generalized Turán quadratures. *Portugal. Math.*, 53(3):355–366, 1996.

[12] J. G. Laframboise and A. D. Stauffer. Optimum discrete approximation of the Maxwell distribution. *AIAA Journal*, 7(3):520–523, 1969.

[13] F. Locher. Interpolation on uniform meshes by the translates of one function and related attenuation factors. *Math. Comp.*, 37(156):403–416, 1981.

[14] Giuseppe Mastroianni and Gradimir V. Milovanović. *Interpolation processes. Basic theory and applications.* Springer Monographs in Mathematics, Springer, Berlin, 2008, xiv+444 pp. ISBN: 978-3-540-68346-9.

[15] Gradimir V. Milovanović and Milan A. Kovačević. Moment-preserving spline approximation and Turán quadratures. In *Numerical mathematics (Singapore, 1988)*, 357–365, Internat. Schriftenreihe Numer. Math., 86, Birkhäuser, Basel, 1988.

[16] Sotirios E. Notaris. The error norm of quadrature formulae. *Numer. Algorithms*, 60(4):555–578, 2012.

[17] Michèle Reifenberg and Jean-Paul Berrut. Numerical solution of boundary integral equations by means of attenuation factors. *IMA J. Numer. Anal.*, 20(1):25–46, 2000.

[18] Miodrag M. Spalević. Calculation of Chakalov–Popoviciu quadratures of Radau and Lobatto type. *ANZIAM J.*, 43(3):429–447, 2002.

[19] Gabriele Steidl. On multivariate attenuation factors. *Numer. Algorithms*, 9(3–4): 245–261, 1995.

[20] Gábor Szegő. *Orthogonal polynomials*. Fourth edition. Colloquium Publications, Vol.XXIII, American Mathematical Society, Providence, R.I., 1975, xiii+432 pp.
[21] H. G. ter Morsche. Attenuation factors and multivariate periodic spline interpolation. In *Topics in multivariate approximation (Santiago, 1986)*, 165–174, Academic Press, Boston, MA, 1987.

Part III

Reprints

8

Papers on Numerical Conditioning

---

16 On inverses of Vandermonde and confluent Vandermonde matrices, Numer. Math. 4, 117–123 (1962)

19 On inverses of Vandermonde and confluent Vandermonde matrices. II, Numer. Math. 5, 425–430 (1963)

43 The condition of orthogonal polynomials, Math. Comp. 26, 923–924 (1972)

45 On the condition of algebraic equations, Numer. Math. 21, 405–424 (1973)

51 Norm estimates for inverses of Vandermonde matrices, Numer. Math. 23, 337–347 (1975)

62 On inverses of Vandermonde and confluent Vandermonde matrices. III, Numer. Math. 29, 445–450 (1978)

64 Questions of numerical condition related to polynomials, in *Symposium on recent advances in numerical analysis* (C. de Boor and G. H. Golub, eds.), 45–72, Academic Press, New York, 1978. [Revised and reprinted in *MAA Studies in Mathematics 24: Studies in numerical analysis* (G. H. Golub, ed.), 140–177, Math. Assoc. America, Washington, DC, 1984.]

66 The condition of polynomials in power form, Math. Comp. 33, 343–352 (1979)

83 The condition of Vandermonde-like matrices involving orthogonal polynomials, Linear Algebra Appl. 52/53, 293–300 (1983)

110 (with G. Inglese) Lower bounds for the condition number of Vandermonde matrices, Numer. Math. 52, 241–250 (1988)

118 How (un)stable are Vandermonde systems?, in *Asymptotic and computational analysis* (R. Wong, ed.), 193–210, Lecture Notes Pure Appl. Math. 124, Dekker, New York, 1990

120 (with A. Córdova and S. Ruscheweyh) Vandermonde matrices on the circle: spectral properties and conditioning, Numer. Math. 57, 577–591 (1990)

200 Optimally scaled and optimally conditioned Vandermonde and Vandermonde-like matrices, BIT Numer. Math. 51, 103–125 (2011)

8.1. [16] "On inverses of Vandermonde and confluent Vandermonde matrices"

---

[16] "On inverses of Vandermonde and confluent Vandermonde matrices," *Numer. Math.* **4**, 117–123 (1962).

Numerische Mathematik 4, 117—123 (1962)

On inverses of Vandermonde and confluent Vandermonde matrices

By

WALTER GAUTSCHI*

1. Introduction

A Vandermonde matrix of order n is a matrix of the form

$$V_n = V_n(x_1, x_2, \ldots, x_n) = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \\ \cdot & \cdot & \cdots & \cdot \\ x_1^{n-1} & x_2^{n-1} & \ldots & x_n^{n-1} \end{pmatrix} \quad (n > 1), \tag{1.1}$$

where x_ν are real or complex numbers. By a confluence of the l-th column into the k-th column we mean the following limit operation: Replace in the l-th column x_l by $x_k + \varepsilon$ and subtract from it the k-th column; divide this new l-th column by ε and then let $\varepsilon \to 0$.

If the resulting matrix is denoted by $U_{n,kl}$ we have

$$U_{n,kl} = \begin{pmatrix} 1 & \ldots & 1 & 0 & 1 & \ldots & 1 \\ x_1 & \ldots & x_{l-1} & 1 & x_{l+1} & \ldots & x_n \\ x_1^2 & \ldots & x_{l-1}^2 & 2x_k & x_{l+1}^2 & \ldots & x_n^2 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ x_1^{n-1} & \ldots & x_{l-1}^{n-1} & (n-1)x_k^{n-2} & x_{l+1}^{n-1} & \ldots & x_n^{n-1} \end{pmatrix}. \tag{1.2}$$

In other words, $U_{n,kl}$ is the same matrix as V_n except for the l-th column, which is the derivative of the k-th column.

A matrix that is obtained from (1.1) by one or more confluences of columns is called a confluent Vandermonde matrix. The following, for example, is a confluent Vandermonde matrix of order $2n$, obtained by confluences of the columns $n+1$ into 1, $n+2$ into 2, ..., $2n$ into n:

$$U_{2n} = \begin{pmatrix} 1 & \ldots & 1 & 0 & \ldots & 0 \\ x_1 & \ldots & x_n & 1 & \ldots & 1 \\ \vdots & & \vdots & \vdots & & \vdots \\ x_1^{2n-1} & \ldots & x_n^{2n-1} & (2n-1)x_1^{2n-2} & \ldots & (2n-1)x_n^{2n-2} \end{pmatrix}. \tag{1.3}$$

* Oak Ridge National Laboratory, operated by Union Carbide Corporation for the U.S. Atomic Energy Commission, Oak Ridge, Tennessee.

The purpose of this paper is to estimate the norm of inverses of Vandermonde and confluent Vandermonde matrices. Such estimates are expected to be useful in various questions of numerical analysis. In the construction of Gauss-type quadrature formulas, for example, norm estimates of the inverse of the matrix (1.3) may be used to assess the errors in the zeros and weight factors from those in the moments.

It will be convenient to adopt the following matrix norm,

$$\|A\| = \max_{1 \leq \nu \leq n} \sum_{\mu=1}^{n} |a_{\nu\mu}|, \qquad A = (a_{\nu\mu}). \tag{1.4}$$

The use of this particular norm is no real restriction since for any other norm $\|A\|_1$, one has $m\|A\| \leq \|A\|_1 \leq M\|A\|$ with positive constants m, M depending only on n, and not on A (see [3], Satz IV).

2. Preliminaries

We denote by σ_m the m-th elementary symmetric function in the n variables $x_1, x_2, \ldots, x_n$,

$$\sigma_m = \sigma_m(x_1, \ldots, x_n) = \sum x_{\nu_1} x_{\nu_2} \ldots x_{\nu_m} \qquad (1 \leq m \leq n), \qquad \sigma_0 = 1.$$

Lemma. *We have*

$$1 + |\sigma_1| + |\sigma_2| + \cdots + |\sigma_n| \leq \prod_{\nu=1}^{n} (1 + |x_\nu|), \tag{2.1}$$

where equality holds if and only if all x_ν are located on the same ray through the origin, that is, if and only if

$$x_\nu = |x_\nu| e^{i\varphi} \qquad (\nu = 1, 2, \ldots, n). \tag{2.2}$$

Proof. Let $p(x) = \prod_{\nu=1}^{n} (x - x_\nu)$. Then

$$p(x) = \sum_{m=0}^{n} (-1)^m \sigma_m x^{n-m}.$$

In particular,

$$p(-1) = (-1)^n \sum_{m=0}^{n} \sigma_m. \tag{2.3}$$

On the other hand, by definition,

$$p(-1) = (-1)^n \prod_{\nu=1}^{n} (1 + x_\nu). \tag{2.4}$$

We distinguish three cases.

Case I. All $x_\nu \geq 0$. Then all $\sigma_m \geq 0$, and from (2.3) and (2.4) we find

$$\sum_{m=0}^{n} |\sigma_m| = \sum_{m=0}^{n} \sigma_m = (-1)^n p(-1) = \prod_{\nu=1}^{n} (1 + x_\nu) = \prod_{\nu=1}^{n} (1 + |x_\nu|).$$

This proves (2.1) with equality sign.

Case II. All x_ν satisfy (2.2). Then $\sigma_m(x_1, \ldots, x_n) = e^{im\varphi}\sigma_m(|x_1|, \ldots, |x_n|)$, and

$$\sum_{m=0}^{n} |\sigma_m| = \sum_{m=0}^{n} \sigma_m(|x_1|, \ldots, |x_n|) = \prod_{\nu=1}^{n} (1+|x_\nu|)$$

by the result of Case I.

Case III. There is at least one pair of variables, say (x_1, x_2), such that $x_1 x_2 \neq 0$, $\arg x_1 \neq \arg x_2$. Then

$$\begin{aligned} |\sigma_1| &= |x_1 + x_2 + \cdots + x_n| \leqq |x_1 + x_2| + |x_3| + \cdots + |x_n| \\ &< |x_1| + |x_2| + |x_3| + \cdots + |x_n|, \end{aligned}$$

that is,

$$|\sigma_1(x_1, \ldots, x_n)| < \sigma_1(|x_1|, \ldots, |x_n|).$$

Since also

$$|\sigma_m(x_1, \ldots, x_n)| \leqq \sigma_m(|x_1|, \ldots, |x_n|) \qquad (m > 1),$$

we find, using again the result of Case I,

$$\sum_{m=0}^{n} |\sigma_m(x_1, \ldots, x_n)| < \sum_{m=0}^{n} \sigma_m(|x_1|, \ldots, |x_n|) = \prod_{\nu=1}^{n} (1+|x_\nu|).$$

This proves (2.1) with strict inequality, and the lemma is completely proved.

Later we also use the notation σ_m^λ to denote the m-th elementary symmetric function in the $n-1$ variables x_ν with x_λ missing,

$$\sigma_m^\lambda = \sigma_m(x_1, \ldots, x_{\lambda-1}, x_{\lambda+1}, \ldots, x_n).$$

By the symmetry of σ_m we have for $\lambda < \mu$

$$\begin{aligned} &\sigma_m^\lambda(x_1, \ldots, x_{\lambda-1}, x_{\lambda+1}, \ldots, x_{\mu-1}, t, x_{\mu+1}, \ldots, x_n) \\ &\quad = \sigma_m^\mu(x_1, \ldots, x_{\lambda-1}, t, x_{\lambda+1}, \ldots, x_{\mu-1}, x_{\mu+1}, \ldots, x_n). \end{aligned} \tag{2.5}$$

3. Inverse of Vandermonde matrix

We prove now

Theorem 1. *Let $x_\nu \neq x_\mu$ for $\nu \neq \mu$. Then, with the matrix norm defined in* (1.4), *we have*

$$\|V_n^{-1}\| \leqq \max_{1 \leqq \lambda \leqq n} \prod_{\substack{\nu=1 \\ \nu \neq \lambda}}^{n} \frac{1+|x_\nu|}{|x_\nu - x_\lambda|}. \tag{3.1}$$

If the x_ν satisfy (2.2), *then* (3.1) *is actually an equality.*

Proof. Let $V_n^{-1} = (v_{\lambda\mu})$. It is well known (see [2, p. 306], or [1]) that

$$v_{\lambda\mu} = (-1)^{\mu-1} \frac{\sigma_{n-\mu}^\lambda}{\prod_{\nu \neq \lambda} (x_\nu - x_\lambda)}. \tag{3.2}$$

Therefore,

$$\sum_{\mu=1}^{n} |v_{\lambda\mu}| = \frac{\sum_{\mu=1}^{n} |\sigma_{n-\mu}^\lambda|}{\prod_{\nu \neq \lambda} |x_\nu - x_\lambda|} \qquad (\lambda = 1, 2, \ldots, n).$$

Theorem 1 now follows immediately from the lemma in section 2.

We note that the last statement in Theorem 1 cannot be reversed, that is, if (3.1) holds with equality sign then it does not necessarily follow that all x_ν lie on the same ray through the origin. This is shown by the example $n=3$, $x_1=8$, $x_2=2$, $x_3=-1$, for which

$$V_3=\begin{pmatrix} 1 & 1 & 1 \\ 8 & 2 & -1 \\ 64 & 4 & 1 \end{pmatrix}, \qquad V_3^{-1}=\begin{pmatrix} -2/54 & -1/54 & 1/54 \\ 8/18 & 7/18 & -1/18 \\ 16/27 & -10/27 & 1/27 \end{pmatrix}.$$

Here, $\|V_3^{-1}\|=\max(4/54, 16/18, 1)=1$, and the bound on the right of (3.1) equals $\max(1/9, 1, 1)=1$, so that (3.1) is in fact an equality, even though $x_1 x_3<0$.

4. Inverses of confluent Vandermonde matrices

In this section we establish norm estimates for the inverses of the confluent matrices $U_{n,kl}$, U_{2n} defined in (1.2) and (1.3), respectively. At the same time explicit expressions are derived for the elements of $U_{n,kl}^{-1}$.

Theorem 2. *Let $x_\nu \neq x_\mu$ for $\nu \neq \mu$ $(\nu, \mu=1, 2, \ldots, l-1, l+1, \ldots, n)$, and let*

$$a_\lambda=\begin{cases} \dfrac{1+|x_k|}{|x_k-x_\lambda|} & (\lambda \neq k, l) \\ \max\left[1+|x_k|,\ 1+(1+|x_k|)\displaystyle\sum_{\substack{\nu=1 \\ \nu\neq k,\,l}}^{n}\frac{1}{|x_\nu-x_k|}\right] & (\lambda=k). \end{cases} \tag{4.1}$$

Then, with the matrix norm defined in (1.4), *we have*

$$\|U_{n,kl}^{-1}\| \leqq \max_{\substack{1\leqq\lambda\leqq n \\ \lambda\neq l}} a_\lambda \prod_{\substack{\nu=1 \\ \nu\neq\lambda,\,l}}^{n}\frac{1+|x_\nu|}{|x_\nu-x_\lambda|}. \tag{4.2}$$

Proof. Assume for the sake of definiteness that $k<l$. Let us introduce the "perturbed" Vandermonde matrix

$$V_{n,kl}(\varepsilon)=V_n(x_1, \ldots, x_{l-1}, x_k+\varepsilon, x_{l+1}, \ldots, x_n),$$

and the auxiliary matrix

$$E_{kl}(\varepsilon)=\begin{pmatrix} 1 & \ldots & 0 & \ldots & 0 & \ldots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \ldots & 1 & \ldots & -\varepsilon^{-1} & \ldots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \ldots & 0 & \ldots & \varepsilon^{-1} & \ldots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \ldots & 0 & \ldots & 0 & \ldots & 1 \end{pmatrix} \begin{matrix} \\ \\ k\text{-th row} \\ \\ l\text{-th row} \\ \\ \\ \end{matrix} \tag{4.3}$$

$$\begin{matrix} & & k\text{-th} & & l\text{-th} & & \\ & & \text{column} & & \text{column} & & \end{matrix}$$

Then it is not difficult to see, that by definition of confluence,

$$U_{n,kl}=\lim_{\varepsilon\to 0} V_{n,kl}(\varepsilon)\, E_{kl}(\varepsilon).$$

From this we get

$$U_{n,kl}^{-1} = \lim_{\varepsilon \to 0} E_{kl}^{-1}(\varepsilon)\, V_{n,kl}^{-1}(\varepsilon), \tag{4.4}$$

provided that the limit on the right-hand side exists.

Inverting (4.3) we have

$$E_{kl}^{-1}(\varepsilon) = \begin{pmatrix} 1 & \dots & 0 & \dots & 0 & \dots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \dots & 1 & \dots & 1 & \dots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \dots & 0 & \dots & \varepsilon & \dots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \dots & 0 & \dots & 0 & \dots & 1 \end{pmatrix} \begin{matrix} \\ \\ k\text{-th row} \\ \\ l\text{-th row} \\ \\ \\ \end{matrix} .$$

$$\begin{matrix} k\text{-th} & & l\text{-th} \\ \text{column} & & \text{column} \end{matrix}$$

Therefore, if $V_{n,kl}^{-1}(\varepsilon) = [v_{\lambda\mu}(\varepsilon)]$, we find

$$E_{kl}^{-1}(\varepsilon)\, V_{n,kl}^{-1}(\varepsilon) = \begin{pmatrix} v_{11}(\varepsilon) & \dots & v_{1n}(\varepsilon) \\ \vdots & & \vdots \\ v_{k1}(\varepsilon) + v_{l1}(\varepsilon) & \dots & v_{kn}(\varepsilon) + v_{ln}(\varepsilon) \\ \vdots & & \vdots \\ \varepsilon\, v_{l1}(\varepsilon) & \dots & \varepsilon\, v_{ln}(\varepsilon) \\ \vdots & & \vdots \\ v_{n1}(\varepsilon) & \dots & v_{nn}(\varepsilon) \end{pmatrix}.$$

Since $V_{n,kl}$ is a Vandermonde matrix, the elements of its inverse are given by (3.2), that is

$$v_{\lambda\mu}(\varepsilon) = (-1)^{\mu-1} \frac{\sigma_{n-\mu}^{\lambda}}{\prod\limits_{\nu \neq \lambda} (x_\nu - x_\lambda)}. \tag{4.5}$$

It is understood here, that x_l, wherever it occurs, is to be replaced by $x_k + \varepsilon$. If $\lambda \neq k, l$ the expression in (4.5) has a well defined limit, as $\varepsilon \to 0$, namely

$$\lim_{\varepsilon \to 0} v_{\lambda\mu}(\varepsilon) = (-1)^{\mu-1} \frac{\sigma_{n-\mu}^{\lambda}(x_1, \dots, x_{l-1}, x_k, x_{l+1}, \dots, x_n)}{(x_k - x_\lambda) \prod\limits_{\nu \neq \lambda, l} (x_\nu - x_\lambda)} \qquad (\lambda \neq k, l). \tag{4.6}$$

If $\lambda = k$, we have, using (2.5), with $\lambda = k$, $\mu = l$, $t = x_k + \varepsilon$,

$$\begin{aligned} v_{k\mu}(\varepsilon) &= (-1)^{\mu-1} \frac{\sigma_{n-\mu}^{k}(x_1, \dots, x_{l-1}, x_k + \varepsilon, x_{l+1}, \dots, x_n)}{\varepsilon \prod\limits_{\nu \neq k, l} (x_\nu - x_k)} \\ &= \frac{(-1)^{\mu-1}}{\varepsilon} \frac{\sigma_{n-\mu}^{l}(x_1, \dots, x_{k-1}, x_k + \varepsilon, x_{k+1}, \dots, x_n)}{\prod\limits_{\nu \neq k, l} (x_\nu - x_k)}. \end{aligned} \tag{4.7}$$

If, finally, $\lambda = l$ then

$$v_{l\mu}(\varepsilon) = \frac{(-1)^{\mu}}{\varepsilon} \frac{\sigma_{n-\mu}^{l}(x_1, \dots, x_{k-1}, x_k, x_{k+1}, \dots, x_n)}{\prod\limits_{\nu \neq k, l} (x_\nu - x_k - \varepsilon)}. \tag{4.8}$$

The sum of the two expressions in (4.7) and (4.8) is seen to have the form $(-1)^{\mu-1}\varepsilon^{-1}[\sigma(x+\varepsilon)\pi^{-1}(x)-\sigma(x)\pi^{-1}(x+\varepsilon)]$ where σ, π stand for the numerator and denominator functions, both considered as functions of $x=x_k$. Since

$$\frac{\sigma(x+\varepsilon)}{\pi(x)}-\frac{\sigma(x)}{\pi(x+\varepsilon)}=\varepsilon\frac{\frac{d}{dx}[\sigma(x)\,\pi(x)]}{\pi^2(x)}+o(\varepsilon)\qquad(\varepsilon\to 0)$$

we obtain

$$\lim_{\varepsilon\to 0}[v_{k\mu}(\varepsilon)+v_{l\mu}(\varepsilon)]=(-1)^{\mu-1}\frac{\frac{\partial}{\partial x_k}\left[\sigma^l_{n-\mu}\prod_{\nu\neq k,l}(x_\nu-x_k)\right]}{\prod_{\nu\neq k,l}(x_\nu-x_k)^2}. \tag{4.9}$$

Let us carry out the differentiation in the numerator. We first observe that

$$\frac{\partial}{\partial x_k}\sigma^l_{n-\mu}=\sigma^{l,k}_{n-\mu-1}\qquad(\mu=1,2,\ldots,n),\qquad \sigma^{l,k}_{-1}=0,$$

where $\sigma^{l,k}_m$ denotes the m-th elementary symmetric function in the $n-2$ variables x_ν with both x_l and x_k missing. Next we note that

$$\frac{\frac{\partial}{\partial x_k}\prod_{\nu\neq k,l}(x_\nu-x_k)}{\prod_{\nu\neq k,l}(x_\nu-x_k)}=-\sum_{\nu\neq k,l}\frac{1}{x_\nu-x_k}.$$

Therefore, we obtain from (4.9)

$$\lim_{\varepsilon\to 0}[v_{k\mu}(\varepsilon)+v_{l\mu}(\varepsilon)]=\frac{(-1)^{\mu-1}}{\prod_{\nu\neq k,l}(x_\nu-x_k)}\left\{\sigma^{l,k}_{n-\mu-1}-\sigma^l_{n-\mu}\sum_{\nu\neq k,l}\frac{1}{x_\nu-x_k}\right\}. \tag{4.10}$$

Finally, from (4.8) we see that

$$\lim_{\varepsilon\to 0}\varepsilon\,v_{l\mu}(\varepsilon)=(-1)^{\mu}\frac{\sigma^l_{n-\mu}}{\prod_{\nu\neq k,l}(x_\nu-x_k)}. \tag{4.11}$$

The relations (4.6), (4.10) and (4.11) now show not only that the limiting matrix in (4.4), and thus $U^{-1}_{n,kl}$, exists, but they also give explicit expressions for the elements $u_{\lambda\mu}$ of $U^{-1}_{n,kl}$. From these, and from the lemma in section 2 we conclude

$$\sum_{\mu=1}^{n}|u_{\lambda\mu}|\leqq\frac{1+|x_k|}{|x_k-x_\lambda|}\prod_{\nu\neq\lambda,l}\frac{1+|x_\nu|}{|x_\nu-x_\lambda|}\qquad(\lambda\neq k,l),$$

$$\sum_{\mu=1}^{n}|u_{k\mu}|\leqq\left\{1+(1+|x_k|)\sum_{\nu\neq k,l}\frac{1}{|x_\nu-x_k|}\right\}\prod_{\nu\neq k,l}\frac{1+|x_\nu|}{|x_\nu-x_k|},$$

$$\sum_{\mu=1}^{n}|u_{l\mu}|\leqq(1+|x_k|)\prod_{\nu\neq k,l}\frac{1+|x_\nu|}{|x_\nu-x_k|},$$

which is equivalent to (4.1), (4.2). Theorem 2 is proved.

The argument in the proof of Theorem 2 can be applied repeatedly to deal with matrices that are derived from a Vandermonde matrix by more than one confluence of columns. One so obtains, for example, the following

Theorem 3. *Let* $x_\nu \neq x_\mu$ *for* $\nu \neq \mu$ $(\nu, \mu = 1, 2, \ldots, n)$, *and let*

$$b_\lambda = \max\left[1 + |x_\lambda|,\ 1 + 2(1 + |x_\lambda|) \sum_{\substack{\nu=1 \\ \nu \neq \lambda}}^{n} \frac{1}{|x_\nu - x_\lambda|}\right]. \tag{4.12}$$

Then, with the matrix norm defined in (1.4), *we have*

$$\|U_{2n}^{-1}\| \leqq \max_{1 \leqq \lambda \leqq n} b_\lambda \left(\prod_{\substack{\nu=1 \\ \nu \neq \lambda}}^{n} \frac{1 + |x_\nu|}{|x_\nu - x_\lambda|}\right)^2. \tag{4.13}$$

References

[1] MACON, N., and A. SPITZBART: Inverses of Vandermonde matrices. Amer. Math. Monthly **65**, 95—100 (1958).

[2] MUIR, TH.: The theory of determinants, vol. I. Reprinted in Dover Publications, New York, 1960.

[3] OSTROWSKI, A.: Über Normen von Matrizen. Math. Z. **63**, 2—18 (1955).

Oak Ridge National Laboratory
Mathematics Panel
Post Office Box X
Oak Ridge, Tennessee

(Received October 18, 1961)

8.2. [19] "On inverses of Vandermonde and confluent Vandermonde matrices. II"

---

[19] "On inverses of Vandermonde and confluent Vandermonde matrices. II," *Numer. Math.* **5**, 425–430 (1963).

Numerische Mathematik 5, 425–430 (1963)

On inverses of Vandermonde and confluent Vandermonde matrices. II

By
WALTER GAUTSCHI*

1. Introduction

In a previous paper of the same title [1], we were concerned with estimating the maximum row sum norm of inverses of Vandermonde and confluent Vandermonde matrices. In particular, we considered the matrix

$$U_{2n}=\begin{pmatrix} 1 & 1 & \dots & 1 & 0 & 0 & \dots & 0 \\ x_1 & x_2 & \dots & x_n & 1 & 1 & \dots & 1 \\ x_1^2 & x_2^2 & \dots & x_n^2 & 2x_1 & 2x_2 & \dots & 2x_n \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_1^{2n-1} & x_2^{2n-1} & \dots & x_n^{2n-1} & (2n-1)x_1^{2n-2} & (2n-1)x_2^{2n-2} & \dots & (2n-1)x_n^{2n-2} \end{pmatrix},$$

of interest in the construction of Gaussian quadrature formulas and Hermite interpolation, and obtained an upper bound for $\|U_{2n}^{-1}\|$ by a process of repeated confluences. In the present paper we wish to present an alternative derivation of a similar, slightly sharper bound, and state conditions under which this bound is attained. We shall prove, in fact, the following

Theorem. *Let $x_1, x_2, \dots, x_n$ be mutually distinct real or complex numbers. With $\|\cdot\|$ denoting the maximum row sum norm, we have*

$$\|U_{2n}^{-1}\| \leqq \max_{1\leqq\lambda\leqq n} b_\lambda \prod_{\substack{\nu=1\\ \nu\neq\lambda}}^{n} \left(\frac{1+|x_\nu|}{|x_\lambda - x_\nu|}\right)^2, \tag{1.1}$$

where b_λ is the larger of the two quantities

$$b_\lambda^{(1)} = 1+|x_\lambda|, \qquad b_\lambda^{(2)} = \left|1+2x_\lambda \sum_{\nu\neq\lambda} 1/(x_\lambda - x_\nu)\right| + 2\left|\sum_{\nu\neq\lambda} 1/(x_\lambda - x_\nu)\right|. \tag{1.2}$$

If all x_ν are located on the same ray through the origin, i.e., if

$$x_\nu = |x_\nu|\, e^{i\varphi} \qquad (\nu = 1, 2, \dots, n), \tag{1.3}$$

and if in addition

$$\left\{1+2|x_\lambda| \sum_{\nu\neq\lambda} 1/(|x_\lambda| - |x_\nu|)\right\} \sum_{\nu\neq\lambda} 1/(|x_\lambda| - |x_\nu|) \geqq 0 \quad \text{for all } \lambda, \tag{1.4}$$

then (1.1) holds with equality sign.

Remark. As will be apparent from the proof of the theorem, condition (1.4) can be weakened inasmuch as it need only hold for those values λ' of λ which

* Oak Ridge National Laboratory, operated by Union Carbide Corporation for the U. S. Atomic Energy Commission, Oak Ridge, Tennessee. Now at Purdue University, Lafayette, Indiana.

furnish the maximum on the right in (1.1), and for which $b_{\lambda'}^{(2)} > b_{\lambda'}^{(1)}$. The set of such values λ' may be empty.

The second expression in (1.2) may be simplified, but enlarged, if one observes that

$$b_\lambda^{(2)} \leqq 1 + 2(1 + |x_\lambda|)\Big|\sum_{\nu \neq \lambda} 1/(x_\lambda - x_\nu)\Big| \leqq 1 + 2(1 + |x_\lambda|)\sum_{\nu \neq \lambda} 1/|x_\lambda - x_\nu|.$$

Replacing $b_\lambda^{(2)}$ in (1.2) by the last member of these inequalities, one is led back to our previous result in [1]. Also, in the case of (1.3) and (1.4), we may write

$$b_\lambda^{(2)} = \Big|1 + 2(1 + |x_\lambda|)\sum_{\nu \neq \lambda} 1/(|x_\lambda| - |x_\nu|)\Big|.$$

In the following section 2 we state two preliminary results, which will be used in the proof of the theorem, given in section 3. Our proof is sufficiently constructive to suggest a simple algorithm for the calculation of the elements of U_{2n}^{-1}. This is discussed in section 4, where also an ALGOL procedure for computing U_{2n}^{-1} is presented.

2. Two lemmas

We denote by σ_m the m-th elementary symmetric function in the n variables $x_1, x_2, \ldots, x_n$,

$$\sigma_m = \sigma_m(x_1, x_2, \ldots, x_n) = \sum x_{\nu_1} x_{\nu_2} \ldots x_{\nu_m} \qquad (1 \leqq m \leqq n), \qquad \sigma_0 = 1,$$

and define $\sigma_m = 0$ if $m > n$. We set

$$p_n(x) = \prod_{\nu=1}^{n} (x - x_\nu).$$

Lemma 1. *Let* $\tau_\mu = \tau_\mu(x_1, x_2, \ldots, x_n)$ *be defined by*

$$p_n^2(x) = \sum_{\mu=0}^{2n} (-1)^\mu \tau_\mu x^{2n-\mu}. \tag{2.1}$$

Then

$$\sum_{\mu=0}^{2n} |\tau_\mu| \leqq \prod_{\nu=1}^{n} (1 + |x_\nu|)^2, \tag{2.2}$$

and equality holds if and only if all x_ν *satisfy* (1.3).

Proof. Clearly,

$$\tau_\mu = \sigma_0\sigma_\mu + \sigma_1\sigma_{\mu-1} + \cdots + \sigma_\mu\sigma_0.$$

We distinguish three cases.

Case I. All $x_\nu \geqq 0$. Then all $\tau_\mu \geqq 0$, and

$$\sum_{\mu=0}^{2n} |\tau_\mu| = \sum_{\mu=0}^{2n} \tau_\mu = p_n^2(-1) = \prod_{\nu=1}^{n} (1 + x_\nu)^2 = \prod_{\nu=1}^{n} (1 + |x_\nu|)^2,$$

which proves equality in (2.2).

Case II. All x_ν satisfy (1.3). Then $\tau_\mu(x_1, \ldots, x_n) = e^{i\mu\varphi}\tau_\mu(|x_1|, \ldots, |x_n|)$, and

$$\sum_{\mu=0}^{2n} |\tau_\mu| = \sum_{\mu=0}^{2n} \tau_\mu(|x_1|, \ldots, |x_n|) = \prod_{\nu=1}^{n} (1 + |x_\nu|)^2$$

by the result of Case I.

Case III. Any other set of x_ν not covered by Case I or Case II. Then there exists at least one pair of variables, say (x_1, x_2), such that $|x_1+x_2|<|x_1|+|x_2|$. Consequently,

$$|\tau_1(x_1,\ldots,x_n)| = 2|x_1+x_2+\cdots+x_n| < 2(|x_1|+|x_2|+\cdots+|x_n|)$$
$$= \tau_1(|x_1|,\ldots,|x_n|).$$

Since also

$$|\tau_\mu(x_1, x_2, \ldots, x_n)| \leqq \tau_\mu(|x_1|,|x_2|,\ldots,|x_n|) \qquad (\mu>1),$$

we conclude, using again the result of Case I,

$$\sum_{\mu=0}^{2n}|\tau_\mu(x_1,\ldots,x_n)| < \sum_{\mu=0}^{2n}\tau_\mu(|x_1|,\ldots,|x_n|) = \prod_{\nu=1}^{n}(1+|x_\nu|)^2.$$

This establishes (2.2) with strict inequality, and completes the proof of Lemma 1.

Lemma 2. *Let*

$$(s-tx)p_n^2(x) = \sum_{\mu=0}^{2n+1} c_\mu x^{2n-\mu+1}. \tag{2.3}$$

Then

$$\sum_{\mu=0}^{2n+1}|c_\mu| \leqq (|s|+|t|)\prod_{\nu=1}^{n}(1+|x_\nu|)^2. \tag{2.4}$$

If all x_ν satisfy (1.3), *and if*

$$s=|s|\,e^{i\chi}, \qquad t=|t|\,e^{i\psi}, \qquad \chi=\varphi+\psi \pmod{2\pi}, \tag{2.5}$$

then (2.4) *holds with equality sign.*

Proof. Define

$$\tau_{-1}=\tau_{2n+1}=0. \tag{2.6}$$

From (2.1) we obtain

$$(s-tx)p_n^2(x) = \sum_{\mu=0}^{2n+1}(-1)^{\mu-1}(s\,\tau_{\mu-1}+t\,\tau_\mu)\,x^{2n-\mu+1},$$

so that

$$c_\mu = (-1)^{\mu-1}(s\,\tau_{\mu-1}+t\,\tau_\mu). \tag{2.7}$$

Consequently,

$$\sum_{\mu=0}^{2n+1}|c_\mu| = \sum_{\mu=0}^{2n+1}|s\,\tau_{\mu-1}+t\,\tau_\mu| \leqq |s|\sum_{\mu=0}^{2n+1}|\tau_{\mu-1}| + |t|\sum_{\mu=0}^{2n+1}|\tau_\mu|,$$

and by Lemma 1, and (2.6),

$$\sum_{\mu=0}^{2n+1}|c_\mu| \leqq (|s|+|t|)\prod_{\nu=1}^{n}(1+|x_\nu|)^2.$$

If (1.3) and (2.5) hold, then

$$c_\mu = (-1)^{\mu-1}e^{i(\psi+\mu\varphi)}\{|s|\,\tau_{\mu-1}(|x_1|,\ldots,|x_n|)+|t|\,\tau_\mu(|x_1|,\ldots,|x_n|)\}$$
$$(\mu=0,1,\ldots,2n+1),$$

and so

$$\sum_{\mu=0}^{2n+1}|c_\mu| = (|s|+|t|)\prod_{\nu=1}^{n}(1+|x_\nu|)^2,$$

again by Lemma 1, and (2.6). Lemma 2 is proved.

3. Proof of the theorem

Let $P_\lambda(x)$ and $Q_\lambda(x)$ denote the fundamental Hermite interpolation polynomials, relative to the nodes $x_1, x_2, \ldots, x_n$, so that

$$(3.1)\quad P_\lambda(x_\nu)=\delta_{\nu\lambda},\quad P_\lambda'(x_\nu)=0,\quad Q_\lambda(x_\nu)=0,\quad Q_\lambda'(x_\nu)=\delta_{\nu\lambda}\ (\nu,\lambda=1,2,\ldots,n)$$

where $\delta_{\nu\lambda}$ is the Kronecker delta. As is well-known, we have

$$(3.2)\qquad P_\lambda(x)=l_\lambda^2(x)\,[1-2l_\lambda'(x_\lambda)\,(x-x_\lambda)],\qquad Q_\lambda(x)=l_\lambda^2(x)\,(x-x_\lambda),$$

where $l_\lambda(x)$ are the fundamental Lagrange interpolation polynomials,

$$(3.3)\qquad l_\lambda(x)=\prod_{\substack{\nu=1\\ \nu\neq\lambda}}^{n}\frac{x-x_\nu}{x_\lambda-x_\nu}.$$

Let

$$(3.4)\qquad P_\lambda(x)=\sum_{\mu=1}^{2n}a_{\lambda\mu}x^{\mu-1},\qquad Q_\lambda(x)=\sum_{\mu=1}^{2n}b_{\lambda\mu}x^{\mu-1}.$$

Consider now the linear system of equations whose matrix is U_{2n},

$$(3.5)\quad \begin{cases} u_1+u_2+\cdots+u_n & =v_1\\ x_1u_1+x_2u_2+\cdots+x_nu_n+u_{n+1}+u_{n+2}+\cdots+u_{2n} & =v_2\\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ x_1^{2n-1}u_1+x_2^{2n-1}u_2+\cdots+x_n^{2n-1}u_n+ & \\ \quad+(2n-1)\,[x_1^{2n-2}u_{n+1}+x_2^{2n-2}u_{n+2}+\cdots+x_n^{2n-2}u_{2n}] & =v_{2n}. \end{cases}$$

If the μ-th equation is multiplied by $a_{\lambda\mu}$, for $\mu=1, 2, \ldots, 2n$, and the resulting equations are added, one obtains in view of (3.1), (3.4)

$$(3.6)\qquad u_\lambda=\sum_{\mu=1}^{2n}a_{\lambda\mu}v_\mu\qquad(1\leqq\lambda\leqq n).$$

Similarly, multiplying by $b_{\lambda\mu}$, and adding, one obtains

$$(3.7)\qquad u_{n+\lambda}=\sum_{\mu=1}^{2n}b_{\lambda\mu}v_\mu\qquad(1\leqq\lambda\leqq n).$$

Thus, (3.6) and (3.7) solve (3.5), so that

$$(3.8)\qquad U_{2n}^{-1}=\begin{pmatrix}A\\ B\end{pmatrix},$$

where A is the $n\times 2n$-matrix of the coefficients $a_{\lambda\mu}$, and B the $n\times 2n$-matrix of the coefficients $b_{\lambda\mu}$ in (3.4).

From (3.2) and (3.3) it is seen that P_λ is an expression of the form (2.3), with n replaced by $n-1$, and

$$(3.9)\quad s=\frac{1+2x_\lambda l_\lambda'(x_\lambda)}{\prod_{\nu\neq\lambda}(x_\lambda-x_\nu)^2},\qquad t=\frac{2l_\lambda'(x_\lambda)}{\prod_{\nu\neq\lambda}(x_\lambda-x_\nu)^2},\qquad p_{n-1}(x)=\prod_{\nu\neq\lambda}(x-x_\nu).$$

Hence, by Lemma 2,

$$(3.10)\qquad \sum_{\mu=1}^{2n}|a_{\lambda\mu}|\leqq\left(|1+2x_\lambda l_\lambda'(x_\lambda)|+2|l_\lambda'(x_\lambda)|\right)\prod_{\nu\neq\lambda}\left(\frac{1+|x_\nu|}{|x_\lambda-x_\nu|}\right)^2.$$

Similarly, $Q_\lambda(x)$ is an expression of the form (2.3), with

$$s = -\frac{x_\lambda}{\prod_{\nu\neq\lambda}(x_\lambda - x_\nu)^2}, \qquad t = -\frac{1}{\prod_{\nu\neq\lambda}(x_\lambda - x_\nu)^2}, \tag{3.11}$$

and p_{n-1} as in (3.9). Thus again, by Lemma 2,

$$\sum_{\mu=1}^{2n} |b_{\lambda\mu}| \leqq (1+|x_\lambda|) \prod_{\nu\neq\lambda} \left(\frac{1+|x_\nu|}{|x_\lambda - x_\nu|}\right)^2. \tag{3.12}$$

Observing that

$$l_\lambda'(x_\lambda) = \sum_{\nu\neq\lambda} \frac{1}{x_\lambda - x_\nu},$$

the assertions (1.1), (1.2) now follow immediately from (3.8), (3.10), and (3.12).

Assuming now (1.3) and (1.4) to be true, we have for the quantities s and t in (3.9),

$$s = \pm|s|\, e^{-2i\varphi}, \qquad t = \pm|t|\, e^{-3i\varphi},$$

where the signs are both plus, or both minus. Similarly for s and t in (3.11), where

$$s = -|s|\, e^{-i\varphi}, \qquad t = -|t|\, e^{-2i\varphi}.$$

In either case, the conditions (2.5) of Lemma 2 are satisfied, and so with (2.4), also (3.10) and (3.12), hold with equality sign. The theorem is thus completely proved.

4. Computation of U_{2n}^{-1}

Formula (3.8) may serve as a basis for calculating the elements of U_{2n}^{-1}. In view of (3.2), (3.4), and (2.7), the only nontrivial computation required is that of the coefficients τ_μ defined in (2.1). These can be obtained recursively as follows.

Denote, more precisely than before, $\tau_\mu = \tau_{\mu,n}$, so that

$$p_m^2(x) = \sum_{\mu=0}^{2m} (-1)^\mu \tau_{\mu,m} x^{2m-\mu}.$$

Since

$$p_{m+1}^2(x) = (x^2 - 2x_{m+1}x + x_{m+1}^2)\, p_m^2(x),$$

we obtain by equating the coefficients of equal powers of x on the left and right,

$$\tau_{\mu,m+1} = \tau_{\mu,m} + 2x_{m+1}\tau_{\mu-1,m} + x_{m+1}^2 \tau_{\mu-2,m} \qquad (\mu = 0, 1, \ldots, 2m+2). \tag{4.1}$$

We assume here, that $\tau_{0,0} = 1$, and $\tau_{\mu,m} = 0$ whenever $\mu < 0$ or $\mu > 2m$. The quantities $\tau_{\mu,n}$, $\mu = 0, 1, \ldots, 2n$, may thus be obtained by applying (4.1) in turn with $m = 0, 1, \ldots, n-1$. With this in mind, the ALGOL procedure below is readily understood.

```
procedure Uinverse (n, x, v); value n; integer n; array x, v;
comment Given x_ν = x[ν], ν = 1, 2, ..., n, this procedure generates v_λμ = v[λ, μ],
        λ, μ = 1, 2, ..., 2n, where v_λμ are the elements of U_2n^-1. It is assumed
        that n ≧ 1, and that the x_ν are mutually distinct real numbers;
begin integer lambda, nu, mu, m; real sum, product, d, s, t, sgn;
    array x0[0:n-1], tau[-1:2×n-1], tau1[1:2×n-2];
    for lambda:=1 step 1 until n do
```

```
    begin
        for nu:=1 step 1 until n-1 do x0[nu]:=x[if nu<lambda
                then nu else nu+1];
        for mu:=-1 step 1 until 2×n-1 do tau[mu]:=0;
        tau[0]:=1;
        for m:=0 step 1 until n-2 do
        begin
            for mu:=1 step 1 until 2×m+2 do
                tau1[mu]:=tau[mu]+2×x0[m+1]
                        ×tau[mu-1]+x0[m+1]↑2×tau[mu-2];
            for mu:=1 step 1 until 2×m+2 do tau[mu]:=tau1[mu]
        end;

        sum:=0; product:=1;
        for nu:=1 step 1 until n-1 do
        begin
            d:=x[lambda]-x0[nu];
            sum:=sum+1/d;
            product:=d×product
        end;

        product:=product↑2;
        s:=(1+2×x[lambda]×sum)/product; t:=2×sum/product;
        sgn:=1;
        for mu:=1 step 1 until 2×n do
        begin
            v[lambda, mu]:=sgn×(s×tau[2×n-mu-1]
                              +t×tau[2×n-mu]);
            sgn:=-sgn
        end;

        s:=-x[lambda]/product; t:=-1/product;
        for mu:=1 step 1 until 2×n do
        begin
            v[n+lambda, mu]:=sgn×(s×tau[2×n-mu-1]
                                +t×tau[2×n-mu]);
            sgn:=-sgn
        end
    end
end Uinverse
```

Reference

[1] GAUTSCHI, W.: On inverses of Vandermonde and confluent Vandermonde matrices. Numer. Math. **4**, 117–123 (1962).

Computer Sciences Center
Purdue University
West Lafayette, Indiana

(Received August 12, 1963)

8.3. [43] "The Condition of Orthogonal Polynomials"

---

[43] "The Condition of Orthogonal Polynomials," *Math. Comp.* **26**, 923–924 (1972).

MATHEMATICS OF COMPUTATION, VOLUME 26, NUMBER 120, OCTOBER 1972

The Condition of Orthogonal Polynomials

By Walter Gautschi

Abstract. An estimate is given for the condition number of the coordinate map associating to each polynomial its coefficients with respect to a system of orthogonal polynomials.

Let $w(x) \geqq 0$ be a weight function on the finite interval $[a, b]$, and $\{p_k(x)\}_{k=0}^{\infty}$ the associated orthogonal polynomials. We consider the linear parametrization map M_n: $\mathbf{R}^n \to \mathbf{P}_{n-1}$ which associates to each (real) vector $u^T = [u_0, u_1, \cdots, u_{n-1}] \in \mathbf{R}^n$ the (real) polynomial $p(x) = \sum_{k=0}^{n-1} u_k p_k(x) \in \mathbf{P}_{n-1}$. The object of this note is to estimate the condition

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

of the map M_n, the infinity norms in $\mathbf{R}^n$ being defined by $||u||_\infty = \max_{0 \leqq k \leqq n-1} |u_k|$, and in $\mathbf{P}_{n-1}$ by $||p||_\infty = \max_{a \leqq x \leqq b} |p(x)|$. Letting

$$\mu_0 = \int_a^b w(x)\, dx, \qquad h_k = \int_a^b p_k^2(x) w(x)\, dx, \qquad k = 0, 1, 2, \cdots,$$

we show in fact that

$$\operatorname{cond}_\infty M_n \leqq \max_{0 \leqq k \leqq n-1} \left(\frac{\mu_0}{h_k}\right)^{1/2} \max_{a \leqq x \leqq b} \sum_{k=0}^{n-1} |p_k(x)|. \tag{1}$$

For Chebyshev polynomials $p_k(x) = T_k(x)$ on $[-1, 1]$, e.g., this gives

$$\operatorname{cond}_\infty M_n \leqq 2^{1/2} n \qquad (p_k = T_k),$$

while for Legendre polynomials $p_k(x) = P_k(x)$ on $[-1, 1]$ one gets

$$\operatorname{cond}_\infty M_n \leqq n(2n - 1)^{1/2} \qquad (p_k = P_k).$$

In order to prove (1), we first observe that, for any $u \in \mathbf{R}^n$,

$$||M_n u||_\infty = \left|\left| \sum_{k=0}^{n-1} u_k p_k(x) \right|\right|_\infty \leqq ||u||_\infty \max_{a \leqq x \leqq b} \sum_{k=0}^{n-1} |p_k(x)|,$$

so that

$$||M_n||_\infty \leqq \max_{a \leqq x \leqq b} \sum_{k=0}^{n-1} |p_k(x)|. \tag{2}$$

On the other hand, if $M_n^{-1} p = u$, then, by orthogonality,

Received December 27, 1971.

AMS 1970 *subject classifications.* Primary 33A65; Secondary 65G05.

Key words and phrases. Orthogonal polynomials, parametrization, conditioning, Chebyshev polynomials, Legendre polynomials.

$$u_k = \frac{1}{h_k}\int_a^b p(x)p_k(x)w(x)\,dx, \qquad k = 0, 1, \cdots, n-1.$$

Therefore, using the Schwarz inequality,

$$\begin{aligned}|u_k| &\leqq \frac{1}{h_k}\int_a^b |p(x)|\,(w(x))^{1/2}\cdot|p_k(x)|\,(w(x))^{1/2}\,dx\\ &\leqq \frac{1}{h_k}\left(\int_a^b p^2(x)w(x)\,dx\int_a^b p_k^2(x)w(x)\,dx\right)^{1/2}\\ &\leqq \frac{1}{h_k}\left(||p||_\infty^2\int_a^b w(x)\,dx\cdot h_k\right)^{1/2} = ||p||_\infty(\mu_0/h_k)^{1/2}.\end{aligned}$$

It follows that, for all $p \in \mathbf{P}_{n-1}$,

$$||M_n^{-1}p||_\infty \leqq ||p||_\infty \max_{0\leq k\leq n-1}(\mu_0/h_k)^{1/2},$$

so that

$$||M_n^{-1}||_\infty \leqq \max_{0\leq k\leq n-1}(\mu_0/h_k)^{1/2}. \tag{3}$$

Combining (2) and (3) gives the desired result (1).

In terms of the orthonormal polynomials $\pi_k(x) = h_k^{-1/2}p_k(x)$, we may write (1) in the form

$$\operatorname{cond}_\infty M_n \leqq \max_{0\leq k\leq n-1}(\mu_0/h_k)^{1/2}\max_{a\leq x\leq b}\sum_{k=0}^{n-1}h_k^{1/2}\,|\pi_k(x)|. \tag{1'}$$

If we let $h = \min_{0\leq k\leq n-1} h_k$, we see that the bound in (1') is larger than or equal to

$$(\mu_0/h)^{1/2}\max_{a\leq x\leq b}\sum_{k=0}^{n-1}h^{1/2}\,|\pi_k(x)| = \mu_0^{1/2}\max_{a\leq x\leq b}\sum_{k=0}^{n-1}|\pi_k(x)|,$$

so that, among all possible normalizations, the one with $h_0 = h_1 = \cdots = h_{n-1}$ gives the best bound in (1).

Acknowledgment. The author is indebted to the referee for the observation made in the last sentence of this note.

Department of Computer Sciences
Purdue University
Lafayette, Indiana 47907

8.4. [45] “On the Condition of Algebraic Equations”

---

[45] “On the Condition of Algebraic Equations,” *Numer. Math.* **21**, 405–424 (1973).

Numer. Math. 21, 405–424 (1973)

On the Condition of Algebraic Equations*

Walter Gautschi**

Received March 12, 1973

Summary. Given an algebraic equation, in which the polynomial in question is expressed in terms of any set of basis polynomials, we study the sensitivity of the roots with respect to small perturbations in the coefficients of the equation. The degree of sensitivity of each root is measured by an appropriate condition number. We analyze this condition number first in the case where the basis polynomials are the powers, and then, in less detail, in the case where the basis is a set of orthogonal polynomials. Several examples are treated, allowing for a comparative study.

1. Introduction

Our object in this paper is to study the sensitivity of the roots of an algebraic equation with respect to small perturbations in the equation. We shall consistently consider *relative* perturbations (i.e., small *percentage* changes), both in the equation and in its roots. The former are measured in terms of relative changes in the *nonzero* coefficients. We assume thereby that the polynomial in question is represented linearly in terms of a system of basis polynomials. We shall concentrate on two particular bases: the successive powers, and orthogonal polynomials. One of our motivations for this work in fact was a desire to learn more about the influence of parametrization upon the condition of the roots. From a series of papers by Specht [5] it is known that different parametrizations of the polynomial in question lead to different kinds of information concerning the location of its zeros, in particular, concerning regions which contain all of the zeros. Thus, for equations in power form the regions obtained are circles about the origin, while for equations in orthogonal polynomial form the regions are infinite strips along the real axis. It is to be expected, and in fact will be confirmed, that the condition of the roots, too, may depend drastically on the particular parametrization adopted.

In Section 2 we define, and comment on, an appropriate condition number for the roots of an algebraic equation. Section 3 takes up equations in power form. We obtain relatively simple bounds on the condition number of a particular root in terms of all the roots of the equation. The bounds are sharp for configurations of roots in which all are lying on a semiray through the origin or symmetrically on either side of a straight line through the origin. The results are applied in Section 4 to a number of examples, some of which were considered previously by Wilkinson [6]. It is possible, in these examples, to obtain precise asymptotic information (for large degrees n) concerning the condition of the roots. In Section 5 we turn our attention to equations written in terms of orthogonal polynomials. Precise results

* Work performed in part at the U.S.A.F. Aerospace Research Laboratories under contract F33615-71-C-1463 with Technology Incorporated.
** Department of Computer Sciences, Purdue University, Lafayette, Indiana.

are now much harder to come by. We content ourselves, essentially, with deriving an upper bound for the condition number, which is often quite realistic, but occasionally (particularly in the case of complex roots) can be off by many orders of magnitude. To gain further insights into the matter we felt it desirable to compute the condition number by "brute force". Procedures for accomplishing this for any combination of roots and orthogonal polynomials are described in Section 6. Results of these computation, when applied to the examples of Section 4, are discussed in Section 7.

2. Condition Number for Algebraic Equations

Let $f(x)$ be a polynomial of exact degree n in whose zeros we are interested. Given a system of polynomials $\{p_r\}$ such that

$$\text{degree } (p_r) = r, \quad r = 0, 1, 2, \ldots, \tag{2.1}$$

there is a unique representation of f in the form

$$f(x) = \sum_{r=0}^{n} a_r p_r(x). \tag{2.2}$$

We shall assume that the leading coefficient is unity,

$$a_n = 1. \tag{2.3}$$

This can always be achieved by a suitable scaling of f, which does not affect its zeros. We also write

$$f(x) = f(a, x),$$

where a denotes the (real or complex) coefficient vector

$$a^T = [a_0, a_1, \ldots, a_{n-1}].$$

Let now $\mathring{\xi}$ be a simple (real or complex) zero of f, corresponding to the coefficient vector $a = \mathring{a}$,

$$f(\mathring{a}, \mathring{\xi}) = 0, \quad \frac{\partial f}{\partial x}(\mathring{a}, \mathring{\xi}) \neq 0. \tag{2.4}$$

The equation

$$f(a, x) = 0, \tag{2.5}$$

by the implicit function theorem, can then be solved in a certain neighbourhood $U(\mathring{a})$ of $\mathring{a}$, giving rise to a unique smooth solution

$$x = \xi(a), \quad a \in U(\mathring{a}), \tag{2.6}$$

for which

$$\lim_{a \to \mathring{a}} \xi(a) = \mathring{\xi}. \tag{2.7}$$

Definition 2.1. *The k-th condition number c_k, $k = 0, 1, \ldots, n-1$, of the (simple) root $\mathring{\xi}$ of* (2.4) *is defined by*

$$c_k = \lim_{\substack{a_k \to \mathring{a}_k \\ a_l = \mathring{a}_l \text{ for } l \neq k}} \left| \frac{\xi(a) - \mathring{\xi}}{\mathring{\xi}} \Big/ \frac{a_k - \mathring{a}_k}{\mathring{a}_k} \right|, \tag{2.8}$$

where $\xi(a)$ is the root of (2.5) *satisfying* (2.7) *and $\mathring{a}_k \neq 0$. If $\mathring{a}_k = 0$ we define $c_k = 0$.*

Definition 2.1 is meaningful only for a nonvanishing root, $\mathring{\xi} \neq 0$. Moreover, $c_k = 0$ in the case $\mathring{a}_k = 0$ is the correct limit value of c_k as $\mathring{a}_k \to 0$ [cf. (2.11) below].

The k-th condition number c_k measures the amount of error magnification due to a perturbation of one coefficient, $\mathring{a}_k$. It is desirable to introduce a single condition number indicating the extent of error magnification as *all* (nonvanishing) coefficients are perturbed. This can conveniently be done by means of the condition vector

$$c^T = [c_0, c_1, \ldots, c_{n-1}] \tag{2.9}$$

and an appropriate vector norm $\|\cdot\|$.

Definition 2.2. *The condition number $\varkappa(\mathring{\xi})$ of the root $\mathring{\xi}$ is defined by*

$$\varkappa(\mathring{\xi}) = \|c\|, \tag{2.10}$$

where c is the condition vector (2.9).

We write $\varkappa_\infty(\mathring{\xi}) = \|c\|_\infty$, $\varkappa_1(\mathring{\xi}) = \|c\|_1$, etc., for special choices of the vector norm. Since $\varkappa_\infty \leqq \varkappa_1$, we give preference to the L_1-norm $\|\cdot\|_1$.

Observing from Definition 2.1 that

$$c_k = \left| \frac{\mathring{a}_k}{\mathring{\xi}} \left[\frac{\partial \xi}{\partial a_k} \right]_{a = \mathring{a}} \right|, \qquad k = 0, 1, \ldots, n-1, \tag{2.11}$$

and computing the partial derivative in (2.11) by differentiating the identity

$$f(a, \xi(a)) = 0, \qquad a \in U(\mathring{a}),$$

partially with respect to a_k, one obtains the following known result [6, p. 38ff.].

Theorem 2.1. $\varkappa_1(\mathring{\xi}) = \dfrac{1}{|\mathring{\xi} f'(\mathring{\xi})|} \sum\limits_{k=0}^{n-1} |\mathring{a}_k p_k(\mathring{\xi})|$.

Corollary. *If $\tilde{p}_n(x) = l_n^{-1} p_n(x) = x^n + \ldots$, and correspondingly $\tilde{f}(x) = l_n^{-1} f(x) = x^n + \ldots$, then*

$$\varkappa_1(\mathring{\xi}) \geqq \left| \frac{\tilde{p}_n(\mathring{\xi})}{\mathring{\xi} \tilde{f}'(\mathring{\xi})} \right|. \tag{2.12}$$

Proof. Since $\sum\limits_{k=0}^{n-1} \mathring{a}_k p_k(\mathring{\xi}) + p_n(\mathring{\xi}) = f(\mathring{\xi}) = 0$, we obtain from Theorem 2.1

$$\varkappa_1(\mathring{\xi}) \geqq \frac{1}{|\mathring{\xi} f'(\mathring{\xi})|} \left| \sum_{k=0}^{n-1} \mathring{a}_k p_k(\mathring{\xi}) \right| = \left| \frac{p_n(\mathring{\xi})}{\mathring{\xi} f'(\mathring{\xi})} \right|,$$

which is equivalent to (2.12).

We note the following properties of the condition number, which are easily established:

(i) $\varkappa_1(\mathring{\xi})$ does not depend on the particular way the polynomials $\{p_r\}$ are normalized.

(ii) The scaling $a_n = 1$ adopted in (2.2) does not substantially influence the condition of the zeros of f. In fact, if $f^*(a^*, x)$ is another scaling of f, and $\varkappa_1^*(\mathring{\xi})$ the corre-

sponding condition number, defined analogously to (2.8) and (2.10), then

$$\kappa_1(\mathring{\xi}) \leqq \kappa_1^*(\mathring{\xi}) \leqq 2\kappa_1(\mathring{\xi}).$$

(iii) If $\mathring{\xi} \neq 0$ is a multiple zero of f, then $\kappa_1(\mathring{\xi}) = \infty$.

The result of Theorem 2.1 may be interpreted in terms of two other condition numbers relating to the *evaluation of* $f(x)$ *at* $x = \mathring{\xi}$. We have indeed

$$f(a, \mathring{\xi}) - f(\mathring{a}, \mathring{\xi}) = \sum_{k=0}^{n-1}{}' \mathring{a}_k p_k(\mathring{\xi}) \frac{a_k - \mathring{a}_k}{\mathring{a}_k},$$

where prime indicates summation over nonvanishing $\mathring{a}_k$. The relation describes the influence of relative perturbations in the coefficients $\mathring{a}_k$ upon the value of f at $x = \mathring{\xi}$. Since $f(\mathring{a}, \mathring{\xi}) = 0$, we must settle for the *absolute* error in f. An appropriate condition number for this type of sensitivity is given by

$$\kappa_{f,\,\mathrm{coeff}}(\mathring{\xi}) = \sum_{k=0}^{n-1} |\mathring{a}_k p_k(\mathring{\xi})|.$$

Similarly, we may introduce the condition number

$$\kappa_{f,\,\arg}(\mathring{\xi}) = |\mathring{\xi} f'(\mathring{\xi})|$$

describing the sensitivity of the value of f at $\mathring{\xi}$ with respect to relative perturbation of $\mathring{\xi}$. Then

$$\kappa_1(\mathring{\xi}) = \frac{\kappa_{f,\,\mathrm{coeff}}(\mathring{\xi})}{\kappa_{f,\,\arg}(\mathring{\xi})}.$$

3. Equations in Power Form

We now take $p_r(x) = x^r$, and thus consider

$$f(x) = \sum_{r=0}^{n} a_r x^r, \qquad a_n = 1. \tag{3.1}$$

From Theorem 2.1 we get

$$\kappa_1(\xi) = \frac{1}{|\xi f'(\xi)|} \sum_{k=0}^{n-1} |a_k|\,|\xi|^k, \tag{3.2}$$

where ξ is a simple zero of f.

It is easily seen from (3.2) that $\kappa_1(\xi)$ *is invariant with respect to scaling of the independent variable.* In other words, if a new variable x^* is introduced by means of $x = \omega x^*$, where $\omega \neq 0$ is arbitrary complex, carrying $f(x)$ into $f^*(x^*) = \omega^{-n} f(\omega x^*)$, then the condition number $\kappa_1^*(\xi^*)$ for the zero ξ^* of the transformed polynomial f^* is the same as the condition number $\kappa_1(\xi)$ for $\xi = \omega\xi^*$. Note, however, that the condition number is *not* invariant with respect to translation.

Denoting the zeros of f by $\xi_1, \xi_2, \ldots, \xi_n$ we now wish to express the condition number of the (simple) root ξ_μ in terms of all the roots. Clearly

$$a_k = (-1)^{n-k} \sigma_{n-k}(\xi_1, \xi_2, \ldots, \xi_n), \qquad k = 0, 1, \ldots, n-1,$$

where $\sigma_l(\xi_1, \xi_2, \ldots, \xi_n)$ denote the elementary symmetric functions in the variables $\xi_1, \xi_2, \ldots, \xi_n$. Therefore, by (3.2),

$$\varkappa_1(\xi_\mu) = \frac{1}{|\xi_\mu f'(\xi_\mu)|} \sum_{l=1}^{n} |\sigma_l(\xi_1, \xi_2, \ldots, \xi_n)| \, |\xi_\mu|^{n-l}. \tag{3.3}$$

In order to further estimate this number we need the following auxiliary result.

Lemma. *Let $\xi > 0$, and let $\sigma_r(\xi_1, \xi_2, \ldots, \xi_n)$, $r = 0, 1, \ldots, n$, denote the elementary symmetric functions in n variables, where $\sigma_0 = 1$. Then*

$$\sum_{r=0}^{n} |\sigma_r(\xi_1, \xi_2, \ldots, \xi_n)| \, \xi^{n-r} \leqq \prod_{\nu=1}^{n} (\xi + |\xi_\nu|), \tag{3.4}$$

where equality holds if and only if all ξ_ν are located on the same ray emanating from the origin, i.e., $\xi_\nu = |\xi_\nu| e^{i\phi}$, $\nu = 1, 2, \ldots, n$.

Proof. We have shown previously [3] that

$$\sum_{r=0}^{n} |\sigma_r(x_1, x_2, \ldots, x_n)| \leqq \prod_{\nu=1}^{n} (1 + |x_\nu|), \tag{3.5}$$

with equality holding precisely when $x_\nu = |x_\nu| e^{i\phi}$, all ν. Letting $x_\nu = \xi_\nu/\xi$, and observing that $\sigma_r(x_1, x_2, \ldots, x_n) = \xi^{-r} \sigma_r(\xi_1, \xi_2, \ldots, \xi_n)$, we get from (3.5)

$$\sum_{r=0}^{n} |\sigma_r(\xi_1, \xi_2, \ldots, \xi_n)| \xi^{-r} \leqq \prod_{\nu=1}^{n} \left(1 + \frac{|\xi_\nu|}{\xi}\right),$$

which, upon multiplying through by ξ^n, establishes the lemma.

Theorem 3.1. *For the condition number $\varkappa_1(\xi_\mu)$ in (3.3) we have*

$$\varkappa_1(\xi_\mu) \leqq \frac{2 \prod_{\substack{\nu=1 \\ \nu \neq \mu}}^{n} \left(1 + \left|\frac{\xi_\nu}{\xi_\mu}\right|\right) - 1}{\prod_{\substack{\nu=1 \\ \nu \neq \mu}}^{n} \left|1 - \frac{\xi_\nu}{\xi_\mu}\right|}, \tag{3.6}$$

where equality holds if and only if $\xi_\nu = |\xi_\nu| e^{i\phi}$, all ν.

Proof. Using (3.4) with $\xi = |\xi_\mu|$, we get

$$\begin{aligned} \sum_{l=1}^{n} |\sigma_l(\xi_1, \xi_2, \ldots, \xi_n)| \, |\xi_\mu|^{n-l} &\leqq \prod_{\nu=1}^{n} (|\xi_\mu| + |\xi_\nu|) - |\xi_\mu|^n \\ &= 2|\xi_\mu| \prod_{\nu \neq \mu} (|\xi_\mu| + |\xi_\nu|) - |\xi_\mu|^n \\ &= |\xi_\mu|^n \left\{ 2 \prod_{\nu \neq \mu} \left(1 + \left|\frac{\xi_\nu}{\xi_\mu}\right|\right) - 1 \right\}. \end{aligned}$$

Since

$$|\xi_\mu f'(\xi_\mu)| = |\xi_\mu \prod_{\nu \neq \mu} (\xi_\mu - \xi_\nu)| = |\xi_\mu|^n \prod_{\nu \neq \mu} \left|1 - \frac{\xi_\nu}{\xi_\mu}\right|$$

the result (3.6) follows at once from (3.3). Equality in (3.6) holds precisely when equality holds in (3.4). Theorem 3.1 is proved.

Corollary 1. $\varkappa_1(\xi_\mu) \leqq 2 \prod_{\substack{\nu=1 \\ \nu \neq \mu}}^{n} \frac{1+\left|\frac{\xi_\nu}{\xi_\mu}\right|}{\left|1-\frac{\xi_\nu}{\xi_\mu}\right|}$.

Corollary 2. *Let n be even, and suppose that the zeros of f are pairwise symmetric with respect to the origin, say,*

$$\xi_{-\mu} = -\xi_\mu \neq 0, \qquad \mu = 1, 2, \ldots, \frac{n}{2}. \tag{3.7}$$

Then

$$\varkappa_1(\xi_{-\mu}) = \varkappa_1(\xi_\mu) \leqq \frac{2 \prod_{\substack{\nu=1 \\ \nu \neq \mu}}^{n/2} \left(1+\left|\frac{\xi_\nu}{\xi_\mu}\right|^2\right) - 1}{2 \prod_{\substack{\nu=1 \\ \nu \neq \mu}}^{n/2} \left|1-\frac{\xi_\nu^2}{\xi_\mu^2}\right|}, \qquad \mu = 1, 2, \ldots, \frac{n}{2}. \tag{3.8}$$

Equality holds in the second relation of (3.8) *if and only if* $\xi_\nu^2 = |\xi_\nu|^2 e^{i\phi}$, $\nu = 1, 2, \ldots, n/2$.

Proof. Letting

$$f(x) = \prod_{\nu=1}^{n/2} (x^2 - \xi_\nu^2) = \sum_{r=0}^{n/2} a_r x^{2r}, \qquad a_{n/2} = 1,$$

we have

$$\varkappa_1(\xi_\mu) = \sum_{k=0}^{n/2-1} \left|\frac{a_k}{\xi_\mu} \frac{\partial \xi_\mu}{\partial a_k}\right| = \frac{1}{2} \sum_{k=0}^{n/2-1} \left|\frac{a_k}{\xi_\mu^2} \frac{\partial \xi_\mu^2}{\partial a_k}\right|, \qquad \mu = 1, 2, \ldots, n/2.$$

Applying Theorem 3.1 (with n replaced by $n/2$ and the ξ's replaced by their squares) to the last summation, we get the inequality in (3.8), including the condition for equality. The first equality in (3.8) is obvious.

The invariance of $\varkappa_1$ with respect to scaling is reflected in the bounds of Theorem 3.1 and its corollaries. Interestingly enough, the bound in Corollary 1 is also invariant with respect to reciprocation.

We also remark that while ξ_μ is assumed to be a simple zero of f in Theorem 3.1 and its corollaries, some of the other zeros ξ_ν may well be multiple.

4. Examples

For specific configurations of zeros the results of the previous section permit us to work out the condition numbers of these zeros in closed form and to analyze their asymptotic behavior for large degrees, n. We begin with a well-known example due to Wilkinson [6, p. 41 ff.].

Example 4.1. $\xi_\nu = \nu$, $\nu = 1, 2, \ldots, n$.

Theorem 3.1, in this case, gives

$$\varkappa_1(\xi_\mu) = \frac{(\mu+n)! - \mu^n \mu!}{\mu!^2 (n-\mu)!}, \qquad \mu = 1, 2, \ldots, n. \tag{4.1}$$

Which of the roots is worst, which is best conditioned? We examine this question asymptotically, for large n, by letting

$$\mu = \tau n, \qquad 0 < \tau \leqq 1, \qquad n \to \infty.$$

We find from (4.1) that

$$\frac{\varkappa_1(\xi_{\tau n+1})}{\varkappa_1(\xi_{\tau n})} = \frac{(1-\tau)n}{\tau n+1} \frac{\dfrac{(\tau+1)n+1}{\tau n+1}\dfrac{[(\tau+1)n]!}{(\tau n)!} - (\tau n)^n\left(1+\dfrac{1}{\tau n}\right)^n}{\dfrac{[(\tau+1)n]!}{(\tau n)!} - (\tau n)^n}. \tag{4.2}$$

By Stirling's formula,

$$\frac{[(\tau+1)n]!}{(\tau n)!} \sim \sqrt{\frac{1+\tau}{\tau}}\left[\left(\frac{\tau+1}{\tau}\right)^\tau \frac{\tau+1}{e} n\right]^n, \qquad n\to\infty.$$

Since

$$\left(\frac{\tau+1}{\tau}\right)^\tau \frac{\tau+1}{e} > \tau \quad \text{on } 0<\tau\leqq 1,$$

the minued in the numerator and denominator of (4.2) dominates the subtrahend for large n. Consequently,

$$\frac{\varkappa_1(\xi_{\mu+1})}{\varkappa_1(\xi_\mu)} \sim \frac{1-\tau^2}{\tau^2} \quad \text{as } \mu=\tau n,\ n\to\infty. \tag{4.3}$$

If $1-\tau_0^2=\tau_0^2$, i.e., $\tau_0=1/\sqrt{2}$, we see that, asymptotically, the condition number $\varkappa_1(\xi_\mu)$ increases for $\mu<\tau_0 n$ and decreases for $\mu>\tau_0 n$, assuming a maximum at $\mu=\tau_0 n$. For $\varkappa_1(\xi_{\tau n})$ one finds

$$\varkappa_1(\xi_{\tau n}) \sim \frac{1}{2\pi(1-\tau)n}\left(\frac{1-\tau^2}{\tau^2}\right)^{\frac{1}{2}}\left[\frac{1+\tau}{1-\tau}\left(\frac{1-\tau^2}{\tau^2}\right)^\tau\right]^n, \qquad n\to\infty,$$

which, at the maximum, $\tau=\tau_0$, becomes

$$\begin{aligned} &\varkappa_1(\xi_{\tau n}) \sim \frac{1}{2\pi(1-\tau)n}\left(\frac{1+\tau}{1-\tau}\right)^n, \qquad n\to\infty, \\ &\tau=\tau_0=\frac{1}{\sqrt{2}}=0.70710\ldots, \qquad \frac{1+\tau_0}{1-\tau_0}=5.8284\ldots. \end{aligned} \tag{4.4}$$

The smallest condition occurs at $\mu=1$, and we find from (4.1) directly that

$$\varkappa_1(\xi_1) \sim n^2, \qquad n\to\infty. \tag{4.5}$$

Table 4.1 shows, for various values of n, the integer μ at which $\varkappa_1(\xi_\mu)$ attains its maximum, the asymptotic estimate $\mu\sim n/\sqrt{2}$, the value[1] of the maximum condition number, and, in the last column, its asymptotic estimate[1] from (4.4).

In the next example we determine how the condition of the roots is affected by a shift of the origin to the center of gravity of the roots.

Table 4.1. Maximum condition numbers for Example 4.1

n	μ	$n/\sqrt{2}$	$\varkappa_1(\xi_\mu)$	(4.4)
5	4	3.5355	5.8733 (2)	7.3097 (2)
10	7	7.0711	2.3244 (6)	2.4582 (6)
20	14	14.142	5.3952 (13)	5.5604 (13)
40	28	28.284	5.5698 (28)	5.6899 (28)
80	57	56.57	1.1806 (59)	1.1916 (59)

1 The integers in parentheses indicate powers of 10 by which the preceding numbers are to be multiplied.

Example 4.2. $\xi_\mu = \pm\mu$, $\mu = 1, 2, \ldots, n/2$, n even.

From Corollary 2 of Theorem 3.1 we now obtain

$$\varkappa_1(\xi_\mu) = \varkappa_1(-\xi_\mu) = \frac{\prod_{\nu=1}^{n/2}(\nu^2+\mu^2) - \mu^n}{\left(\frac{n}{2}+\mu\right)!\left(\frac{n}{2}-\mu\right)!}, \qquad \mu = 1, 2, \ldots, \frac{n}{2}.$$

Since

$$\prod_{\nu=1}^{n/2}(\nu^2+\mu^2) = \prod_{\nu=1}^{n/2}[(\nu+i\mu)(\nu-i\mu)] = \left|\frac{\Gamma\left(\frac{n}{2}+1+i\mu\right)}{\Gamma(1+i\mu)}\right|^2,$$

we can express the condition number in terms of the gamma function of a complex argument,

$$\varkappa_1(\xi_\mu) = \frac{\left|\Gamma\left(\frac{n}{2}+1+i\mu\right)\right|^2 - \mu^n\left|\Gamma(1+i\mu)\right|^2}{\left(\frac{n}{2}+\mu\right)!\left(\frac{n}{2}-\mu\right)!\left|\Gamma(1+i\mu)\right|^2}, \qquad \mu = 1, 2, \ldots, \frac{n}{2}. \tag{4.6}$$

We again determine the value of μ which maximizes $\varkappa_1(\xi_\mu)$. We set

$$\mu = \tau\frac{n}{2}, \qquad 0 < \tau \leqq 1,$$

and consider

$$\frac{\varkappa_1(\xi_{\mu+1})}{\varkappa_1(\xi_\mu)} = \frac{(1-\tau)\frac{n}{2}}{(1+\tau)\frac{n}{2}+1}\left|\frac{\Gamma\left(1+\tau\frac{n}{2}i\right)}{\Gamma\left(1+\left(\tau\frac{n}{2}+1\right)i\right)}\right|^2\gamma_n, \tag{4.7}$$

where

$$\gamma_n = \frac{\left|\Gamma\left(\frac{n}{2}+1+\left(\tau\frac{n}{2}+1\right)i\right)\right|^2 - \left(\tau\frac{n}{2}+1\right)^n\left|\Gamma\left(1+\left(\tau\frac{n}{2}+1\right)i\right)\right|^2}{\left|\Gamma\left(\frac{n}{2}+1+\tau\frac{n}{2}i\right)\right|^2 - \left(\tau\frac{n}{2}\right)^n\left|\Gamma\left(1+\tau\frac{n}{2}i\right)\right|^2}. \tag{4.8}$$

Using Stirling's formula, we obtain after some calculation,

$$\begin{aligned}\left|\Gamma\left(\frac{n}{2}+1+\tau\frac{n}{2}i\right)\right|^2 &\sim \sqrt{1+\tau^2}\,\pi n\left[\frac{1}{2e}\sqrt{1+\tau^2}\,e^{-\tau\tan^{-1}\tau}n\right]^n,\\ \left(\tau\frac{n}{2}\right)^n\left|\Gamma\left(1+\tau\frac{n}{2}i\right)\right|^2 &\sim \tau\pi n\left[\frac{\tau}{2}e^{-\tau\pi/2}n\right]^n, \qquad n\to\infty.\end{aligned} \tag{4.9}$$

We now show that the expression in the first line of (4.9) dominates the one in the second, asymptotically as $n\to\infty$, i.e.,

$$\frac{1}{2e}\sqrt{1+\tau^2}\,e^{-\tau\tan^{-1}\tau} > \frac{\tau}{2}e^{-\tau\pi/2}, \qquad 0 < \tau \leqq 1, \tag{4.10}$$

or, equivalently,

$$g(\tau) \overset{\text{def}}{=} \sqrt{1+\frac{1}{\tau^2}}\,e^{\frac{\pi}{2}\tau-\tau\tan^{-1}\tau-1} > 1 \qquad \text{for } 0 < \tau \leqq 1.$$

Since $g(1) = \sqrt{2}\exp\left(\frac{\pi}{4}-1\right) = 1.1410\ldots > 1$, it suffices to show that $g'(\tau) < 0$ on $0 < \tau \leqq 1$. But,

$$g'(\tau) = \frac{\sqrt{1+\tau^2}}{\tau^2}e^{\frac{\pi}{2}\tau-\tau\tan^{-1}\tau-1}\left\{\tau\left(\frac{\pi}{2}-\tan^{-1}\tau\right)-1\right\},$$

and for the function in curled brackets, say $h(\tau)$, we find

$$h(0)=-1<0, \qquad h(1)=\frac{\pi}{4}-1<0,$$

$$h'(\tau)=\frac{\pi}{2}-\tan^{-1}\tau-\frac{\tau}{1+\tau^2}>\frac{\pi}{2}-\frac{\pi}{4}-\frac{1}{2}=\frac{\pi}{4}-\frac{1}{2}>0 \qquad \text{on } 0<\tau\leqq 1,$$

so that $h(\tau)<0$ on $0<\tau\leqq 1$, hence also $g'(\tau)<0$.

Similarly to (4.9), one finds that

$$\begin{aligned}&\left|\Gamma\left(\frac{n}{2}+1+\left(\tau\frac{n}{2}+1\right)i\right)\right|^2\sim\sqrt{1+\tau^2}\,\pi n e^{-2\tan^{-1}\tau}\left[\frac{1}{2e}\sqrt{1+\tau^2}\,e^{-\tau\tan^{-1}\tau}n\right]^n,\\ &\left(\tau\frac{n}{2}+1\right)^n\left|\Gamma\left(1+\left(\tau\frac{n}{2}+1\right)i\right)\right|^2\sim\tau\pi n e^{\frac{2}{\tau}-\pi}\left[\frac{\tau}{2}e^{-\tau\pi/2}n\right]^n, \qquad n\to\infty,\end{aligned} \tag{4.11}$$

where again, by (4.10), the top expression is the asymptotically larger. Consequently, both in the numerator and denominator of (4.8), the second terms can be neglected. It follows that

$$\gamma_n\sim\frac{\left|\Gamma\left(\frac{n}{2}+1+\left(\tau\frac{n}{2}+1\right)i\right)\right|^2}{\left|\Gamma\left(\frac{n}{2}+1+\tau\frac{n}{2}i\right)\right|^2}\sim e^{-2\tan^{-1}\tau}, \qquad n\to\infty. \tag{4.12}$$

From (4.7), using again the second relations in (4.9), (4.11), we thus obtain

$$\frac{\varkappa_1(\xi_{\mu+1})}{\varkappa_1(\xi_\mu)}\sim\frac{1-\tau}{1+\tau}e^{\pi-2\tan^{-1}\tau}, \qquad \mu=\tau\frac{n}{2}, \qquad n\to\infty. \tag{4.13}$$

One readily verifies that the function on the right decreases monotonically from e^π to 0, as τ increases from 0 to 1. There exists, therefore, a unique value $\tau=\tau_0$ on the interval $(0,1)$ such that

$$\frac{1-\tau}{1+\tau}e^{\pi-2\tan^{-1}\tau}=1, \tag{4.14}$$

and we see that $\varkappa_1(\xi_\mu)$ increases for $\mu<\tau_0\frac{n}{2}$ and decreases for $\mu>\tau_0\frac{n}{2}$. In fact, $\tau_0=0.73409\ldots$.

Since

$$\left[(1+\tau)\frac{n}{2}\right]!\left[(1-\tau)\frac{n}{2}\right]!\sim\pi n\sqrt{1-\tau^2}\left[\frac{1}{2e}\sqrt{1-\tau^2}\left(\frac{1+\tau}{1-\tau}\right)^{\tau/2}n\right]^n, \qquad n\to\infty,$$

we obtain from (4.6), using once more (4.9) and (4.10),

$$\varkappa_1(\xi_{\tau n/2})\sim\frac{1}{\tau\pi n}\sqrt{\frac{1+\tau^2}{1-\tau^2}}\left\{\sqrt{\frac{1+\tau^2}{1-\tau^2}}\left[\frac{1-\tau}{1+\tau}e^{\pi-2\tan^{-1}\tau}\right]^{\tau/2}\right\}^n, \qquad n\to\infty.$$

At the maximum, $\tau=\tau_0$, in view of (4.14) this simplifies to

$$\begin{aligned}&\varkappa_1(\xi_\mu)\sim\frac{1}{\tau\pi n}\left(\sqrt{\frac{1+\tau^2}{1-\tau^2}}\right)^{n+1}, \qquad \mu=\tau\frac{n}{2}, \qquad n\to\infty,\\ &\tau=\tau_0=0.73409\ldots, \qquad \sqrt{\frac{1+\tau_0^2}{1-\tau_0^2}}=1.8268\ldots.\end{aligned} \tag{4.15}$$

The rate of growth of the maximum condition number is thus seen to be substantially smaller than in Example 4.1. Numerical values, arranged similarly as in Table 4.1, are shown in Table 4.2.

Table 4.2. Maximum condition numbers for Example 4.2

n	μ	$\tau_0 n/2$	$\varkappa_1(\xi_\mu)$	(4.15)
10	4	3.6705	2.7842 (1)	3.2799 (1)
20	7	7.3410	6.2788 (3)	6.7902 (3)
40	15	14.682	5.7290 (8)	5.8204 (8)
80	29	29.364	8.3726 (18)	8.5534 (18)

The smallest condition number occurs at $\mu=1$, and from (4.6) one finds by Stirling's formula that

$$\varkappa_1(\xi_1) \sim \frac{\sinh \pi}{\pi} = 3.6760\ldots, \qquad n\to\infty. \tag{4.16}$$

We next consider an example of a well-conditioned equation.

Example 4.3. (Wilkinson [6, p. 44ff.].) $\xi_\nu = 2^{-\nu}$, $\nu=1, 2, \ldots, n$.

From Corollary 1 of Theorem 3.1 we obtain after a short computation

$$\varkappa_1(\xi_\mu) \leqq 2\pi_\mu, \qquad \pi_\mu = \prod_{\nu=1}^{\mu-1} \frac{1+2^{-\nu}}{1-2^{-\nu}} \prod_{\lambda=1}^{n-\mu} \frac{1+2^{-\lambda}}{1-2^{-\lambda}}. \tag{4.17}$$

One readily verifies that π_μ is symmetric on $1\leqq\mu\leqq n$,

$$\pi_{n+1-\mu} = \pi_\mu, \qquad \mu=1, 2, \ldots, n,$$

and strictly increasing on the left half of this interval,

$$\pi_{\mu+1} > \pi_\mu \qquad \text{for } \mu < n/2.$$

The maximum of π_μ is therefore assumed at $\mu = \left[\frac{n+1}{2}\right]$,

$$\pi_\mu \leqq \pi_{n/2} = \frac{1-2^{-n/2}}{1+2^{-n/2}} \left(\prod_{\lambda=1}^{n/2} \frac{1+2^{-\lambda}}{1-2^{-\lambda}}\right)^2 < \left(\prod_{\lambda=1}^{n/2} \frac{1+2^{-\lambda}}{1-2^{-\lambda}}\right)^2 \qquad (n \text{ even}),$$

$$\pi_\mu \leqq \pi_{(n+1)/2} = \left(\prod_{\lambda=1}^{(n-1)/2} \frac{1+2^{-\lambda}}{1-2^{-\lambda}}\right)^2 \qquad (n \text{ odd}).$$

One computes $\prod_{\nu=1}^{\infty} [(1+2^{-\nu})/(1-2^{-\nu})] = 8.2559\ldots$, and therefore finds that

$$\varkappa_1(\xi_\mu) \leqq 137, \qquad \mu=1, 2, \ldots, n. \tag{4.18}$$

The condition is thus bounded by a relatively small number, uniformly in n. The minimum occurs at $\mu=1$, where

$$\varkappa_1(\xi_1) \leqq 2 \prod_{\lambda=1}^{n-1} \frac{1+2^{-\lambda}}{1-2^{-\lambda}} \sim 16.511\ldots.$$

As observed earlier, the bound in (4.17) is invariant under reciprocation. Hence, our analysis of Example 4.3 applies equally well to the case where $\xi_\nu = 2^\nu$, $\nu=1, 2, \ldots, n$.

The two examples which follow involve predominantly complex roots.

Example 4.4. (Roots of unity.) $\xi_\nu = e^{2\pi i\nu/n}$, $\nu = 1, 2, \ldots, n$.

Here, $f(x) = x^n - 1$, and from (3.2) we get immediately

$$\varkappa_1(\xi_\mu) = \frac{1}{n}, \quad \mu = 1, 2, \ldots, n. \tag{4.19}$$

Thus, all roots have the same (small) condition number $1/n$. It is an open question whether the root configuration of Example 4.4 is indeed optimal in some reasonable sense.

We note that inequality (3.6) does rather poorly on this example. It states, in fact, that

$$\varkappa_1(\xi_\mu) \leqq \frac{2 \cdot 2^{n-1} - 1}{\prod\limits_{\nu \neq \mu} |1 - e^{2\pi i(\nu-\mu)/n}|} = \frac{2^n - 1}{\prod\limits_{\nu \neq \mu} |e^{2\pi i\mu/n} - e^{2\pi i\nu/n}|}.$$

As the product in the denominator is simply $|f'(\xi_\mu)| = n$, we get

$$\varkappa_1(\xi_\mu) \leqq \frac{2^n - 1}{n},$$

a bound which is too large by a factor of $2^n - 1$.

Example 4.5. $e_n(\xi) = 0$ where $e_n(x) = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!}$.

We now have $f(x) = n!\, e_n(x)$, and 3.2 gives

$$\varkappa_1(\xi) = \frac{1}{|\xi e_n'(\xi)|} \sum_{k=0}^{n-1} \frac{1}{k!} |\xi|^k = \frac{e_{n-1}(|\xi|)}{|\xi e_{n-1}(\xi)|}.$$

The magnitude of $\varkappa_1(\xi)$ is seen to be related to the amount of cancellation which occurs when $e_{n-1}(\xi)$ is evaluated. A somewhat simpler expression is obtained by noting that $e_{n-1}(\xi) = -\xi^n/n!$. Thus,

$$\varkappa_1(\xi) = \frac{n!}{|\xi|^{n+1}} e_{n-1}(|\xi|). \tag{4.20}$$

From (2.12) we obtain the lower bound

$$\varkappa_1(\xi) \geqq \frac{1}{|\xi|}. \tag{4.21}$$

Given ξ, the expression in (4.20) is readily evaluated by recursion.

If n is even, all zeros of e_n are complex; if n is odd, all but one are complex. Numerical values and graphs for these zeros can be found in [4], [1]. We used the values published in [4] to compute $\varkappa_1(\xi_\mu)$ for each zero ξ_μ of e_n, $n = 5(5)20$. The results are shown in Table 4.3. The roots are listed in the order of decreasing moduli, and only those in the upper half-plane are shown. (The others are complex conjugates of those shown and have the same condition numbers.) It is seen that the condition worsens as one proceeds to roots with smaller moduli.

It is also found that the upper bound (3.6) for the condition number gradually weakens as n is increased. The bound overestimates the true value by a factor of 2–3, when $n = 5$, and by a factor of 200–400, when $n = 20$. The lower bound (4.21) is even worse.

Table 4.3. Condition numbers for Example 4.5

n	μ	ξ_μ	$\varkappa_1(\xi_\mu)$	n	μ	ξ_μ	$\varkappa_1(\xi_\mu)$
5	1	$0.2398+3.1283i$	2.2951	15	6	$-4.3272+3.0028i$	7.2239 (2)
	2	$-1.6495+1.6939i$	6.6384		7	$-4.8670+1.5176i$	1.0279 (3)
	3	-2.1806	9.1840		8	-5.0439	1.1553 (3)
10	1	$3.3749+5.6260i$	2.3077	20	1	$10.8046+9.2292i$	2.0619
	2	$0.0662+4.9677i$	1.1111 (1)		2	$5.7624+9.7555i$	1.4532 (1)
	3	$-1.8717+3.7702i$	3.2860 (1)		3	$2.3673+9.4134i$	7.4512 (1)
	4	$-3.0155+2.3352i$	6.5822 (1)		4	$-0.1684+8.6388i$	2.9584 (2)
	5	$-3.5539+0.7894i$	9.2502 (1)		5	$-2.1255+7.6041i$	9.3326 (2)
15	1	$6.9748+7.5746i$	2.1851		6	$-3.6406+6.3987i$	2.3799 (3)
	2	$2.7050+7.5509i$	1.3325 (1)		7	$-4.7903+5.0773i$	4.9676 (3)
	3	$-0.0586+6.8056i$	5.5484 (1)		8	$-5.6210+3.6771i$	8.5646 (3)
	4	$-2.0103+5.7117i$	1.7042 (2)		9	$-6.1611+2.2255i$	1.2274 (4)
	5	$-3.3949+4.4177i$	3.9842 (2)		10	$-6.4273+0.7450i$	1.4679 (4)

5. Equations in Orthogonal Polynomial Form

We assume now that $\{p_k\}$ is a system of polynomials orthogonal on (a, b) with respect to the nonnegative weight function w,

$$\int_a^b p_r(x)p_s(x)w(x)\,dx = \begin{cases} 0 & \text{if } r \neq s, \\ h_r & \text{if } r = s, \end{cases} \qquad r, s = 0, 1, 2, \ldots. \tag{5.1}$$

We represent the polynomial f in the form

$$f(x) = \sum_{k=0}^{n} a_k p_k(x), \qquad a_n = 1. \tag{5.2}$$

Given a simple zero ξ of f (normally located in the interval (a, b)), its condition is measured by

$$\varkappa_1(\xi) = \frac{1}{|\xi f'(\xi)|} \sum_{k=0}^{n-1} |a_k p_k(\xi)|. \tag{5.3}$$

Theorem 5.1. *The condition number $\varkappa_1(\xi)$ in (5.3) is invariant with respect to different normalizations of the orthogonal polynomials $\{p_k\}$ of (5.1). If $\{\pi_k\}$ denotes the system of orthonormal polynomials then*

$$\varkappa_1(\xi) \leqq \sqrt{n}\left(\int_a^b f^2(x)w(x)\,dx\right)^{\frac{1}{2}} \frac{\left(\sum_{k=0}^{n-1} |\pi_k(\xi)|^2\right)^{\frac{1}{2}}}{|\xi f'(\xi)|}. \tag{5.4}$$

Proof. The statement concerning normalization is a consequence of the remark (i) following Theorem 2.1. Thus, assuming $p_k = \pi_k$ in (5.2), (5.3), we have by orthogonality

$$\sum_{k=0}^{n-1} |a_k \pi_k(\xi)| = \sum_{k=0}^{n-1} \left|\int_a^b f(x)\pi_k(x)w(x)\,dx\right| |\pi_k(\xi)|, \tag{5.5}$$

and therefore, applying Schwarz's inequality twice, first for sums, and then for integrals,

$$\sum_{k=0}^{n-1} |a_k \pi_k(\xi)| \leqq \left(\sum_{k=0}^{n-1} \left| \int_a^b f(x)\pi_k(x)w(x)\,dx \right|^2\right)^{\frac{1}{2}} \left(\sum_{k=0}^{n-1} |\pi_k(\xi)|^2\right)^{\frac{1}{2}}$$
$$\leqq \left(\sum_{k=0}^{n-1} \int_a^b f^2(x)w(x)\,dx \int_a^b \pi_k^2(x)w(x)\,dx\right)^{\frac{1}{2}} \left(\sum_{k=0}^{n-1} |\pi_k(\xi)|^2\right)^{\frac{1}{2}}$$
$$= \left(n \int_a^b f^2(x)w(x)\,dx\right)^{\frac{1}{2}} \left(\sum_{k=0}^{n-1} |\pi_k(\xi)|^2\right)^{\frac{1}{2}}.$$

The inequality (5.4) now follows from (5.3). Theorem 5.1 is proved.

An alternative form for the sum in the numerator of (5.4) is given by

$$\sum_{k=0}^{n-1} |\pi_k(\xi)|^2 = \frac{k_{n-1}}{k_n} \{\pi_n'(\xi)\pi_{n-1}(\xi) - \pi_{n-1}'(\xi)\pi_n(\xi)\}, \tag{5.6}$$

where k_n is the leading coefficient of π_n. For many weight functions (on a finite interval $[a, b]$) one has [2, pp. 104–105]

$$\sum_{k=0}^{n-1} |\pi_k(\xi)|^2 = O(\sqrt{n}), \qquad \xi \in (a, b), \quad n \to \infty. \tag{5.7}$$

The estimate (5.4) given in Theorem 5.1 may be rather conservative. A first idea of the extent of overestimation can be gained by considering the special polynomial $f(x) = \pi_n(x)$, which has $a_k = 0$ for $k = 0, 1, \ldots, n-1$, and therefore $\varkappa_1(\xi) = 0$ for every zero ξ. From (5.4), (5.6) we get instead

$$\varkappa_1(\xi) \leqq \frac{1}{|\xi|} \left(\frac{n k_{n-1}}{k_n} \left| \frac{\pi_{n-1}(\xi)}{\pi_n'(\xi)} \right|\right)^{\frac{1}{2}}. \tag{5.8}$$

This can be quite large. For example, if $\pi_n(x) = \sqrt{\frac{2}{\pi}} T_n(x)$ is the Chebyshev polynomial, and $\xi = \xi_\mu = \cos\theta_\mu$ one of its zeros, then (5.8) gives

$$\varkappa_1(\xi_\mu) \leqq \frac{1}{\sqrt{2}} |\tan\theta_\mu|, \qquad \theta_\mu = \frac{2\mu - 1}{2n}\pi.$$

This is small for θ_μ near 0 or π, but may be quite large for θ_μ near $\pi/2$. Further observations on the quality of the estimate (5.4) are made in Section 7.

6. Computation of the Condition Number

We now indicate how the condition number (5.3), as well as its bound in (5.4), may be computed, given the set of zeros $\{\xi_\nu\}_{\nu=1}^n$ of f and the system of orthogonal polynomials. By Theorem 5.1 we may assume these polynomials to be normalized. Then

$$f(x) = k_n \prod_{\nu=1}^{n} (x - \xi_\nu) = \sum_{k=0}^{n} a_k \pi_k(x), \qquad a_n = 1,$$

where k_n is the leading coeffcient of $\pi_n(x)$. We let

$$\tilde{f}(x) = \prod_{\nu=1}^{n} (x - \xi_\nu).$$

By (5.3) and (5.5), we then have

$$\varkappa_1(\xi_\mu) = \frac{1}{|\xi_\mu \tilde{f}'(\xi_\mu)|} \sum_{k=0}^{n-1} \left| \int_a^b \tilde{f}(x)\, \pi_k(x) w(x)\, dx \right| |\pi_k(\xi_\mu)|, \qquad \mu = 1, 2, \ldots, n. \tag{6.1}$$

We first note that

$$|\xi_\mu \tilde{f}'(\xi_\mu)| = \left| \xi_\mu \prod_{\substack{\nu=1 \\ \nu \neq \mu}}^{n} (\xi_\mu - \xi_\nu) \right|. \tag{6.2}$$

Next, we expand $\tilde{f}$ in power form,

$$\tilde{f}(x) = \sum_{l=0}^{n} \sigma_{l,n} x^l, \tag{6.3}$$

where the coefficients are obtained by the following recursive scheme,

$$\begin{aligned} &\sigma_{-1,m} = 0, \qquad \sigma_{m,m} = 1 \qquad (m = 0, 1, \ldots, n), \\ &\sigma_{l,m} = \sigma_{l-1,m-1} - \xi_m \sigma_{l,m-1} \qquad (l = 0, 1, \ldots, m-1), \qquad m = 1, 2, \ldots, n. \end{aligned} \tag{6.4}$$

We then obtain

$$\int_a^b \tilde{f}(x)\, \pi_k(x) w(x)\, dx = \sum_{l=k}^{n} \sigma_{l,n} \int_a^b x^l \pi_k(x) w(x)\, dx,$$

where orthogonality was used to restrict the summation to $l \geqq k$. The constants

$$\mu_{k,l} = \int_a^b x^l \pi_k(x) w(x)\, dx$$

can in turn be computed resursively. From the three-term recurrence relation for orthonormal polynomials,

$$\beta_{k+1} \pi_{k+1}(x) = (x - \alpha_k) \pi_k(x) - \beta_k \pi_{k-1}(x), \qquad k = 0, 1, 2, \ldots, \qquad \pi_{-1}(x) = 0, \tag{6.5}$$

we find indeed that

$$\beta_{k+1} \mu_{k+1,l} = \mu_{k,l+1} - \alpha_k \mu_{k,l} - \beta_k \mu_{k-1,l}, \tag{6.6}$$

which permits us to generate the desired quantities $\mu_{k,l}$, $0 \leqq k \leqq n-1$, $k \leqq l \leqq n$, recursively from the starting values

$$\mu_{-1,l} = 0, \qquad \mu_{0,l} = \pi_0 \mu_l, \qquad 0 \leqq l \leqq 2n. \tag{6.7}$$

Here, μ_l is the l-th moment of weight function $w(x)$. In summary, then,

$$\varkappa_1(\xi_\mu) = \frac{1}{|\xi_\mu \prod_{\nu \neq \mu} (\xi_\mu - \xi_\nu)|} \sum_{k=0}^{n-1} \left| \sum_{l=k}^{n} \sigma_{l,n} \mu_{k,l} \right| |\pi_k(\xi_\mu)|, \tag{6.8}$$

where $\sigma_{l,n}$ is obtained from (6.4), $\mu_{k,l}$ from (6.6), (6.7), and $\pi_k(\xi_\mu)$ from (6.5).

Similarly, we can evaluate the upper bound in (5.4),

$$\varkappa_1(\xi_\mu) \leqq \frac{\sqrt{n}}{|\xi_\mu \tilde{f}'(\xi_\mu)|} \left(\int_a^b \tilde{f}^2(x) w(x)\, dx \right)^{\frac{1}{2}} \left(\sum_{k=0}^{n-1} |\pi_k(\xi_\mu)|^2 \right)^{\frac{1}{2}}, \tag{6.9}$$

where, by (6.3), defining $\sigma_{l,n}=0$ for $l>n$,

$$\int_a^b \tilde{f}^2(x)w(x)\,dx = \int_a^b \sum_{\nu=0}^{2n}\left(\sum_{l=0}^{\nu}\sigma_{l,n}\sigma_{\nu-l,n}\right)x^\nu w(x)\,dx = \sum_{\nu=0}^{2n}\mu_\nu \sum_{l=0}^{\nu}\sigma_{l,n}\sigma_{\nu-l,n}. \tag{6.10}$$

7. Examples

The formulas of Section 6 were used to calculate condition numbers for all examples considered in Section 4, and for a number of parametrizations involving classical orthogonal polynomials. The calculations were performed on the CDC 6500 computer in double precision arithmetic. Because of severe cancellation problems in some of the examples computed, we had to limit ourselves to degrees $n\leqq 20$.

Among the orthogonal polynomials which were tried are the Legendre polynomials, the Chebyshev polynomials of the first and second kind, the Laguerre polynomials, and the Hermite polynomials. We refer to them briefly by the symbols $P[a,b]$, $T[a,b]$, $U[a,b]$, $L[0,\infty]$, $H[-\infty,\infty]$, respectively, where the brackets enclose the respective intervals of orthogonality.

As is seen from (6.1), there are largely three factors influencing the magnitude of the condition number $\varkappa_1(\xi)$. They are

(1) the magnitude of the Fourier coefficients $a_k=\int_a^b \tilde{f}(x)\pi_k(x)w(x)\,dx$

(2) the magnitude of the orthogonal polynomials $\pi_k(\xi)$ evaluated at the zero ξ

(3) the magnitude of $\xi\tilde{f}'(\xi)$.

In the numerical examples below, we shall indicate their orders of magnitude (for $n=20$) in curled brackets. Thus,

$$\{-10,\ 0,\ -9/-13\}$$

at the bottom of Table 7.1 shall mean that

$$\max_{0\leqq k\leqq n-1}|a_k|\doteq 10^{-10},\qquad \max_{0\leqq k\leqq n-1}|\pi_k(\xi_\mu)|\doteq 10^{0}=1$$

$$\sum_{k=0}^{n-1}|a_k|\,|\pi_k(\xi_\mu)|\doteq 10^{-9},\qquad |\xi_\mu\tilde{f}'(\xi_\mu)|\doteq 10^{-13}\quad (n=20),$$

where μ is the index of the worst conditioned root ξ_μ. In this way we can learn what factors are most responsible for the widely differing magnitudes of condition numbers that will be encountered.

Example 7.1. $\xi_\nu=\nu/n$, $\nu=1,2,\ldots,n$.

We have scaled the roots of Example 4.1 so as to retain them within the interval $[0,1]$. For polynomials in power form, such a scaling does not affect the condition of the roots.

It appears natural to choose for p_k polynomials orthogonal on the interval $[0,1]$. In so doing, we find that the Chebyshev polynomials of the second kind perform best (in the sense of making $\max_\mu \varkappa_1(\xi_\mu)$ smallest), followed by the Legendre poly-

Table 7.1. Maximum condition numbers for Example 7.1 with $p_k \in U[0,1]$

n	$U[0,1]$
5	1.85
10	2.64 (1)
15	6.35 (2)
20	1.40 (4)
	$\{-10,0,-9/-13\}$

nomials and the Chebyshev polynomials of the first kind. In Table 7.1 we list maximum condition numbers in the most favorable case of Chebyshev polynomials of the second kind. The maximum of $\varkappa_1(\xi_\mu)$ is invariably assumed for μ near $n/2$.

Comparison with Table 4.1 shows a significant improvement in the condition of the roots. For $n=20$, e.g., we now have a maximum condition of 1.4×10^4, as compared to 5.40×10^{13} before. Nevertheless, the condition still grows unboundedly with n. We can see this from the Corollary to Theorem 2.1, where now

$$\tilde{p}_n(x) = 4^{-n} U_n(2x-1),$$

and

$$|\xi_\mu \tilde{l}'(\xi_\mu)| = n^{-n}\mu!\,(n-\mu)!.$$

Thus,

$$\varkappa_1(\xi_\mu) \geqq \left|\frac{\tilde{p}_n(\xi_\mu)}{\xi_\mu \tilde{l}'(\xi_\mu)}\right| = \left(\frac{n}{4}\right)^n \frac{|U_n(2\xi_\mu-1)|}{\mu!\,(n-\mu)!}.$$

Assuming n even, and specializing μ to $\mu = n/2$, so that $2\xi_\mu - 1 = 0$, we obtain

$$\max_\mu \varkappa_1(\xi_\mu) \geqq \varkappa_1(\xi_{n/2}) \geqq \left(\frac{n}{4}\right)^n \frac{1}{(n/2)!^2} \sim \frac{1}{\pi n}\left(\frac{e}{2}\right)^n, \quad n\to\infty. \tag{7.1}$$

It is instructive to examine the effect on the condition of choosing other intervals of orthogonality, both intervals containing $[0,1]$ in their interior as well as intervals being contained in $[0,1]$. A summary of results in this direction is shown in Table 7.2. It is seen that the condition grows catastrophically, even faster than in Example 4.1. The poor showing of the Laguerre polynomials is particularly worth noting. The rapid deterioration of the condition must be attributed in the first three cases to relatively large Fourier coefficients, and in the last case to large values of $|\pi_k(\xi_n)|$. As a matter of fact, in the last case, only the roots located in $(\frac{1}{2}, 1]$ are badly conditioned; those in $(0, \frac{1}{2}]$ are relatively well-conditioned.

Table 7.2. Maximum condition numbers for Example 7.1 in the case of unnatural intervals of orthogonality

n	$U[-1,1]$	$L[0,\infty]$	$H[-\infty,\infty]$	$P[0,\frac{1}{2}]$
5	8.81 (2)	1.94 (5)	4.33 (3)	5.35 (1)
10	9.62 (6)	7.94 (14)	2.36 (9)	1.27 (4)
15	1.82 (11)	3.22 (25)	6.04 (15)	3.59 (6)
20	3.82 (15)	5.69 (36)	3.80 (22)	1.12 (9)
	$\{2,0,2/-13\}$	$\{23,0,23/-13\}$	$\{9,0,9/-13\}$	$\{-9,12,-3/-12\}$

Computation of the upper bound (5.4) for $\varkappa_1(\xi_\mu)$, by means of (6.9), (6.10) revealed agreement with $\varkappa_1(\xi_\mu)$ to within one order of magnitude, in all cases except $P[0, \frac{1}{2}]$. The lower bound (2.12), in contrast, is quite weak, being too small by as many as four orders of magnitude, for the interval of orthogonality $[0, 1]$, and nine orders of magnitude for the interval $[-1, 1]$.

Example 7.2. $\xi_\mu = \pm 2\mu/n$, $\mu = 1, 2, \ldots, \frac{n}{2}$, n even.

The results in this case, on the whole, are very similar to the results in Example 7.1, except that condition numbers are uniformly lower. A few, corresponding to the natural interval of orthogonality $[-1, 1]$, and to Chebyshev polynomials of the second kind (which again are "best"), are shown in Table 7.3. As $n \to \infty$, the maximum condition number again tends to infinity, at least as fast as indicated in (7.1). Unnatural choices of the interval of orthogonality, like $[0, 1]$, $[0, \infty]$, or $[-\infty, \infty]$, give rise to comparatively much larger condition numbers, the Laguerre polynomials once again standing out as the worst offenders.

Table 7.3. Maximum condition numbers for Example 7.2 with $p_k \in U[-1, 1]$

n	$U[-1, 1]$
10	1.18
20	3.25 (2)
	$\{-5, 0, -5/-7\}$

Example 7.3. $\xi_\nu = 2^{-(\nu-1)}$, $\nu = 1, 2, \ldots, n$.

In Example 4.3, these roots are seen to be uniformly well-conditioned for equations in power form. Rather strikingly, it is found that they are extremely ill-conditioned under any orthogonal polynomial form of the equation, even choosing as interval of orthogonality the natural interval $[0, 1]$, and as orthogonal polynomials $p_k(x)$ the most favorable ones, the Chebyshev polynomials of the first kind. Table 7.4 gives some idea of the grimness of the situation. The factor most responsible for it is revealed in the schedule at the bottom of the table: extremely small values of the derivative of $\tilde{f}$ at some of the roots! As a matter of fact, one

Table 7.4. Maximum condition numbers for Example 7.3 with $p_k \in T[0, 1]$

n	$T[0, 1]$
5	5.03 (1)
10	1.44 (12)
15	1.19 (30)
20	3.58 (55)
	$\{-3, 0, -3/-58\}$

readily finds for the smallest root that

$$|\xi_n \tilde{f}'(\xi_n)| = \frac{1}{2^{n(n-1)/2}} \prod_{\lambda=1}^{n-1} (1-2^{-\lambda}) \sim \frac{0.28879}{2^{n(n-1)/2}}, \quad n \to \infty,$$

which indeed becomes small very rapidly. The Fourier coefficients of f, although reasonably small, are incapable of counteracting this kind of decay. In the power case, on the other hand, the small denominator $|\xi_n \tilde{f}'(\xi_n)|$ is neutralized by an equally small numerator,

$$\sum_{k=0}^{n-1} |a_k \xi_n^k| = \frac{1}{2^{(n+1)(n-2)/2}} \prod_{\lambda=1}^{n-1} (1+2^{-\lambda}) - 2^{-n(n-1)} \sim \frac{2.3842}{2^{(n+1)(n-2)/2}}, \quad n \to \infty.$$

We also note that the upper bound of Theorem 5.1 is found in all cases to overestimate the condition number by no more than one order of magnitude. The lower bound (2.12), as before, is considerably weaker.

Example 7.3a. $\xi_\nu = 2^{\nu-1}$, $\nu = 1, 2, \ldots, n$.

This example does not essentially differ from Example 7.3 if the equation is represented in power form (cf. Example 4.3). Using orthogonal polynomials, however, a substantial difference emerges. It seems natural, now, to resort to Laguerre polynomials, although other choices of orthogonal polynomials turn out to be almost equally good. We list some results for Laguerre and Chebyshev polynomials in Table 7.5. The conditions are comparable to those observed in Example 4.3.

Table 7.5. Maximum condition numbers for Example 7.3a with $p_k \in L[0, \infty]$ and $p_k \in U[0, 1]$

n	$L[0, \infty]$	$U[0, 1]$
5	2.24	2.50 (1)
10	3.20 (1)	9.51 (1)
15	9.65 (1)	1.27 (2)
20	1.26 (2)	1.34 (2)
	{56, 63, 89/87}	{56, 74, 78/76}

Example 7.4. (Roots of unity.) $\xi_\nu = e^{2\pi i \nu/n}$, $\nu = 1, 2, \ldots, n$.

Using the orthogonality interval $[-1, 1]$, Chebyshev polynomials of the second kind give the smallest condition numbers. A close second are the Legendre polynomials, followed by Chebyshev polynomials of the first kind. Some condition numbers are shown in Table 7.6. As one would expect, the condition is worst furthest away from the interval $[-1, 1]$, i.e., for ξ_μ near $\pm i$. The condition worsens some-

Table 7.6. Maximum condition numbers for Example 7.4 with $p_k \in U[-1, 1]$

n	$U[-1, 1]$
5	0.549
10	1.62
15	7.50
20	34.4

what if one chooses $[0, 1]$ as interval of orthogonality, and is particularly bad for Laguerre and Hermite polynomials.

A feature worthy of note is the relatively poor performance of the upper bound in Theorem 5.1 on this example. The bound overestimates the condition by as many as five orders of magnitude, when $[-1, 1]$ is used as interval of orthogonality, and by up to nine orders of magnitude, if the interval is $[0, 1]$.

Example 7.5. $e_n(\xi_\nu) = 0$ where $e_n(x) = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!}$.

Interestingly enough, all orthogonal polynomials, with the exception of the Laguerre polynomials, do about equally well on this example, the choice $[-1, 1]$ of the orthogonality interval faring slightly better than the choice $[0, 1]$. The results shown in Table 7.7, compared with those in Table 4.3, also show rather remarkably

Table 7.7. Maximum condition numbers for Example 7.5 with $p_k \in U[-1, 1]$

n	$U[-1, 1]$
5	9.28
10	9.30 (1)
15	1.16 (3)
20	1.47 (4)

that the condition of the roots is practically the same, no matter whether the equation is written in power form or in terms of orthogonal polynomials.

Another interesting feature in this example is the exceptionally poor quality of the upper bound in Theorem 5.1. For $n = 20$, e.g., the bound is too large by over 20 orders of magnitude! The breakdown can be traced to Schwarz's inequality, as applied to $\sum_{k=0}^{n-1} |a_k| \, |\pi_k(\xi)|$. The Fourier coefficients a_k happen to decrease at a rapid geometric rate, while the $|\pi_k(\xi)|$ increase, equally rapidly. This is a very unfavorable situation for Schwarz's inequality, as is exemplified by $a_k = 10^{-k}$, $b_k = 10^k$, in which case Schwarz's inequality gives $n = \sum_{k=0}^{n-1} a_k b_k \leqq 10^{n-1}$. The lower bound (2.12), this time, is substantially better, but still off by as many as five orders of magnitude.

It is difficult, of course, to draw general conclusions from a small sample of examples, particularly from examples as varied as those presented above. Nevertheless, some general trends are discernible. If the roots of the equation are predominantly real, and not accumulating near a finite point (as they do, e.g., in Example 7.3), then their condition seems indeed enhanced if the equation is represented in terms of orthogonal polynomials, provided the interval of orthogonality matches the interval spanned by the real roots as closely as possible. If the roots, on the other hand, are predominantly complex, then the choice of basis polynomials seems less critical. If the use of orthogonal polynomials is indicated, then Chebyshev polynomials of the second kind appear to be as good a choice as any, quite in contrast to Laguerre polynomials, or Hermite polynomials, which should be avoided.

References

1. Dejon, B., Nickel, K.: A never failing, fast convergent rootfinding algorithm, Constructive Aspects of the Fundamental Theorem of Algebra (Proc. Sympos., Zürich-Rüschlikon, 1967), 1–35. New York: Wiley-Interscience 1969
2. Freud, G.: Orthogonal polynomials. Oxford: Pergamon Press 1971
3. Gautschi, W.: On inverses of Vandermonde and confluent Vandermonde matrices. Numer. Math. **4**, 117–123 (1962)
4. Iverson, K. E.: The zeros of the partial sums of e^z. Math. Tables Aids Comput. **7**, 165–168 (1953)
5. Specht, W.: Die Lage der Nullstellen eines Polynoms. Math. Nachr. **15**, 353–374 (1956); ibid., **16**, 257–263 (1957); ibid., **16**, 369–389 (1957); ibid. **21**, 201–222 (1960)
6. Wilkinson, J. H.: Rounding errors in algebraic processes. Englewood Cliffs, N. J.: Prentice-Hall 1963

Prof. Dr. Walter Gautschi
Purdue University
Computer Sciences
Mathematical Sciences Building
Lafayette, Indiana 47907/USA

8.5. [51] “Norm Estimates for Inverses of Vandermonde Matrices”

---

[51] “Norm Estimates for Inverses of Vandermonde Matrices,” *Numer. Math.* **23**, 337–347 (1975).

Numer. Math. 23, 337—347 (1975)

Norm Estimates for Inverses of Vandermonde Matrices

Walter Gautschi

Received April 11, 1974

Summary. Formulas, or close two-sided estimates, are given for the norm of the inverse of a Vandermonde matrix when the constituent parameters are arranged in certain symmetric configurations in the complex plane. The effect of scaling the parameters is also investigated. Asymptotic estimates of the respective condition numbers are derived in special cases.

1. Introduction

In an earlier paper [2] we obtained norm inequalities for the inverse V_n^{-1} of a Vandermonde matrix $V(x_1, x_2, \ldots, x_n) \in \mathbb{C}^{n\times n}$, which become equalities if the complex parameters x_ν are all placed on a ray emanating from the origin. We now obtain equalities for $\|V_n^{-1}\|$ also in the case of parameters located symmetrically with respect to the origin on a straight line through the origin. For parameters which occur in conjugate complex pairs we give close upper and lower bounds for $\|V_n^{-1}\|$. We further examine how scaling of the parameters affects the magnitude of $\|V_n^{-1}\|$. Finally, as an application, we derive asymptotic estimates (for large n) for the condition number of Vandermonde matrices for special configurations of the parameters.

2. Preliminaries

We denote by σ_m the m-th elementary symmetric function in n complex variables,

$$\sigma_m = \sigma_m(x_1, x_2, \ldots, x_n) = \sum x_{\nu_1} x_{\nu_2} \ldots x_{\nu_m} \quad (1 \leqq m \leqq n), \qquad \sigma_0 = 1.$$

Lemma 2.1. *We have*

$$\sum_{m=0}^{n} |\sigma_m(x_1, x_2, \ldots, x_n)| \leqq \prod_{\nu=1}^{n} (1 + |x_\nu|), \tag{2.1}$$

where equality holds if and only if $x_\nu = |x_\nu| e^{i\phi}$, $\nu = 1, 2, \ldots, n$.

A proof of Lemma 2.1 is given in [2].

Lemma 2.2. *Let* $p_{2n}(x) = \prod_{\nu=1}^{n}(x^2 - x_\nu)$ *and*

$$(s + tx) p_{2n}(x) = \sum_{\mu=0}^{2n+1} c_\mu x^{2n-\mu+1}. \tag{2.2}$$

Then

$$\sum_{\mu=0}^{2n+1} |c_\mu| \leqq (|s|+|t|) \prod_{\nu=1}^{n} (1+|x_\nu|), \tag{2.3}$$

where equality holds if and only if $x_\nu = |x_\nu|\, e^{i\phi}$, $\nu = 1, 2, \ldots, n$.

Proof. Since

$$p_{2n}(x) = \sum_{\mu=0}^{n} (-1)^\mu \sigma_\mu(x_1, x_2, \ldots, x_n)\, x^{2n-2\mu},$$

we find for the coefficients c_μ in (2.2),

$$c_{2\mu} = (-1)^\mu t \sigma_\mu, \qquad c_{2\mu+1} = (-1)^\mu s \sigma_\mu, \qquad \mu = 0, 1, \ldots, n.$$

Consequently, using Lemma 2.1,

$$\sum_{\mu=0}^{2n+1} |c_\mu| = (|s|+|t|) \sum_{m=0}^{n} |\sigma_m(x_1, x_2, \ldots, x_n)| \leqq (|s|+|t|) \prod_{\nu=1}^{n} (1+|x_\nu|),$$

with equality as stated.

Lemma 2.3. *Given* $2n$ *real or complex numbers* $x_1, x_2, \ldots, x_{2n}$ *such that*

$$x_{n+\nu} = \bar{x}_\nu, \qquad \nu = 1, 2, \ldots, n, \tag{2.4}$$

and for all ν *either* $\operatorname{Re} x_\nu \geqq 0$, *or* $\operatorname{Re} x_\nu \leqq 0$, *let* $p_{2n}(x) = \prod_{\mu=1}^{2n} (x - x_\mu)$ *and*

$$(s+tx)\, p_{2n}(x) = \sum_{\mu=0}^{2n+1} c_\mu x^{2n-\mu+1}. \tag{2.5}$$

Then

$$\big||s|-|t|\big| \prod_{\nu=1}^{n} |1 \pm x_\nu|^2 \leqq \sum_{\mu=0}^{2n+1} |c_\mu| \leqq (|s|+|t|) \prod_{\nu=1}^{n} |1 \pm x_\nu|^2, \tag{2.6}$$

where the plus sign holds if all $\operatorname{Re} x_\nu \geqq 0$, *and the minus sign if all* $\operatorname{Re} x_\nu \leqq 0$.

Proof. We first observe that in

$$p_{2n}(x) = \sum_{\mu=0}^{2n} (-1)^\mu \sigma_\mu(x_1, x_2, \ldots, x_{2n})\, x^{2n-\mu} \tag{2.7}$$

we have

$$\sigma_\mu \geqq 0 \quad \text{if all } \operatorname{Re} x_\nu \geqq 0, \qquad (-1)^\mu \sigma_\mu \geqq 0 \quad \text{if all } \operatorname{Re} x_\nu \leqq 0.$$

In fact,

$$p_{2n}(x) = \prod_{\nu=1}^{n} [(x - x_\nu)(x - \bar{x}_\nu)] = \prod_{\nu=1}^{n} [x^2 - (2 \operatorname{Re} x_\nu)\, x + |x_\nu|^2],$$

and multiplying out the product on the right yields coefficients which alternate in sign, if all $\operatorname{Re} x_\nu \geqq 0$, and are nonnegative, if all $\operatorname{Re} x_\nu \leqq 0$. Consequently,

$$\sum_{\mu=0}^{2n} |\sigma_\mu(x_1, x_2, \ldots, x_{2n})| = \begin{cases} p_{2n}(-1) = \prod_{\nu=1}^{n} |1+x_\nu|^2 & \text{if all } \operatorname{Re} x_\nu \geqq 0, \\ p_{2n}(1) = \prod_{\nu=1}^{n} |1-x_\nu|^2 & \text{if all } \operatorname{Re} x_\nu \leqq 0. \end{cases} \tag{2.8}$$

For the coefficients c_μ in (2.5) we have

$$c_\mu = (-1)^\mu (t\sigma_\mu - s\sigma_{\mu-1}), \qquad \mu = 0, 1, \ldots, 2n+1,$$

where $\sigma_{-1} = \sigma_{2n+1} = 0$. Therefore,

$$\big||s| - |t|\big| \sum_{\mu=0}^{2n} |\sigma_\mu| \leqq \sum_{\mu=0}^{2n+1} |c_\mu| = \sum_{\mu=0}^{2n+1} |t\sigma_\mu - s\sigma_{\mu-1}| \leqq (|s| + |t|) \sum_{\mu=0}^{2n} |\sigma_\mu|,$$

from which (2.6) follows by virtue of (2.8).

3. Inversion of the Vandermonde Matrix

We denote the Vandermonde matrix of order n by

$$V_n = V(x_1, x_2, \ldots, x_n) = \begin{bmatrix} 1 & 1 \ldots 1 \\ x_1 & x_2 \ldots x_n \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ x_1^{n-1} & x_2^{n-1} \ldots x_n^{n-1} \end{bmatrix}, \tag{3.1}$$

where $x_1, x_2, \ldots, x_n$ are distinct complex numbers and $n > 1$. Its inverse can be obtained by solving the system of linear algebraic equations

$$\begin{aligned} u_1 + \quad & u_2 + \cdots + \quad u_n = v_1 \\ x_1 u_1 + x_2 \quad & u_2 + \cdots + x_n \quad u_n = v_2 \\ & \cdots\cdots\cdots\cdots\cdots\cdots \\ x_1^{n-1} u_1 + & x_2^{n-1} u_2 + \cdots + x_n^{n-1} u_n = v_n. \end{aligned} \tag{3.2}$$

Introducing the elementary Lagrange interpolation polynomials

$$l_\nu(x) = \prod_{\substack{\mu=1 \\ \mu \neq \nu}}^{n} \frac{x - x_\mu}{x_\nu - x_\mu} = a_{\nu n} x^{n-1} + a_{\nu, n-1} x^{n-2} + \cdots + a_{\nu 1}, \qquad \nu = 1, 2, \ldots, n, \tag{3.3}$$

which satisfy

$$l_\nu(x_\mu) = \begin{cases} 1 & \text{if } \nu = \mu \\ 0 & \text{if } \nu \neq \mu, \end{cases}$$

it is evident that by multiplying the μ-th Eq. (3.2) by $a_{\nu\mu}$, $\mu = 1, 2, \ldots, n$, and adding, we get

$$u_\nu = \sum_{\mu=1}^{n} a_{\nu\mu} v_\mu, \qquad \nu = 1, 2, \ldots, n.$$

Consequently,

$$V_n^{-1} = \begin{bmatrix} a_{11} & a_{12} \ldots a_{1n} \\ a_{21} & a_{22} \ldots a_{2n} \\ \ldots\ldots\ldots\ldots \\ a_{n1} & a_{n2} \ldots a_{nn} \end{bmatrix}. \tag{3.4}$$

4. Norm Inequalities for V_n^{-1}

We consider throughout the ∞-norm of V_n^{-1},

$$\|V_n^{-1}\|_\infty = \max_{1 \leqq \nu \leqq n} \sum_{\mu=1}^{n} |a_{\nu\mu}|.$$

Theorem 4.1. *$\|V^{-1}(x_1, x_2, \ldots, x_n)\|_\infty$ is a symmetric function in the variables $x_1, x_2, \ldots, x_n$.*

Proof. Interchanging two variables amounts to interchanging two columns of V_n, which in turn has the effect of interchanging two rows of V_n^{-1}. The value of $\|V_n^{-1}\|_\infty$ remains the same.

Theorem 4.2. *Let $\omega \neq 0$ be arbitrary complex, and*

$$V_n(\omega) = V(\omega x_1, \omega x_2, \ldots, \omega x_n).$$

Then $\|V_n^{-1}(\omega)\|_\infty$ depends only on $|\omega|$ and is strictly decreasing as a function of $|\omega|$.

Proof. Let $V_n = V_n(1)$, $V_n^{-1} = [a_{\nu\mu}]$. Since

$$V_n(\omega) = D(\omega) V_n, \qquad D(\omega) = \operatorname{diag}(1, \omega, \ldots, \omega^{n-1}),$$

we have $V_n^{-1}(\omega) = V_n^{-1} D^{-1}(\omega)$, i.e.,

$$V_n^{-1}(\omega) = \left[\frac{a_{\nu\mu}}{\omega^{\mu-1}}\right], \qquad \nu, \mu = 1, 2, \ldots, n.$$

It is clear, therefore, that the norm of $V_n^{-1}(\omega)$ depends only on $|\omega|$. Furthermore, if $|\omega_1| < |\omega_2|$, we have

$$\|V_n^{-1}(\omega_2)\|_\infty = \max_\nu \sum_{\mu=1}^n \frac{|a_{\nu\mu}|}{|\omega_2|^{\mu-1}} = \sum_{\mu=1}^n \frac{|a_{\nu_0\mu}|}{|\omega_2|^{\mu-1}}$$

$$< \sum_{\mu=1}^n \frac{|a_{\nu_0\mu}|}{|\omega_1|^{\mu-1}} \leqq \max_\nu \sum_{\mu=1}^n \frac{|a_{\nu\mu}|}{|\omega_1|^{\mu-1}} = \|V_n^{-1}(\omega_1)\|_\infty,$$

where strict inequality holds because of

$$a_{\nu_0 n} = \prod_{\mu \neq \nu_0} (x_{\nu_0} - x_\mu)^{-1} \neq 0.$$

This proves Theorem 4.2.

In [2] we have shown that

$$\|V_n^{-1}\|_\infty \leqq \max_{1 \leqq \nu \leqq n} \prod_{\substack{\mu=1 \\ \mu \neq \nu}}^n \frac{1 + |x_\mu|}{|x_\nu - x_\mu|}, \tag{4.1}$$

where equality holds if (but not only if) all x_ν are on the same ray through the origin,

$$x_\nu = |x_\nu| e^{i\phi}, \qquad \nu = 1, 2, \ldots, n. \tag{4.2}$$

In view of Theorem 4.2 we may assume $\phi = 0$ in (4.2), i.e., $x_\nu \geqq 0$, $\nu = 1, 2, \ldots, n$, in which case the equality in (4.1) can be given the alternative form

$$\|V_n^{-1}\|_\infty = \frac{|p_n(-1)|}{\min\limits_{1 \leqq \nu \leqq n} \{(1 + x_\nu)\,|p_n'(x_\nu)|\}} \qquad (x_\nu \geqq 0), \tag{4.1'}$$

where

$$p_n(x) = \prod_{\mu=1}^n (x - x_\mu). \tag{4.3}$$

We now wish to obtain a result analogous to (4.1') when the points x_ν are located symmetrically with respect to the origin on a straight line through the

origin. In view of Theorem 4.2 we may assume the straight line to coincide with the real axis.

Theorem 4.3. *Let x_ν be distinct real numbers such that*

$$x_\nu + x_{n+1-\nu} = 0, \qquad \nu = 1, 2, \ldots, n. \tag{4.4}$$

If $V_n = V(x_1, x_2, \ldots, x_n)$, we then have

$$\|V_n^{-1}\|_\infty = \begin{cases} \frac{1}{2} \max_\nu \left\{ \left(1 + \frac{1}{x_\nu}\right) \prod_{\mu \neq \nu} \frac{1 + x_\mu^2}{|x_\nu^2 - x_\mu^2|} \right\} & \text{if } n \text{ is even,} \\ \max_\nu \left\{ \varepsilon_\nu (1 + x_\nu) \prod_{\mu \neq \nu} \frac{1 + x_\mu^2}{|x_\nu^2 - x_\mu^2|} \right\} & \text{if } n \text{ is odd,} \end{cases} \tag{4.5}$$

where ν and μ vary over all integers for which $x_\nu \geqq 0$ and $x_\mu \geqq 0$, respectively, and where $\varepsilon_\nu = \frac{1}{2}$ when $x_\nu > 0$, and $\varepsilon_\nu = 1$ when $x_\nu = 0$. Alternatively,

$$\|V_n^{-1}\|_\infty = \frac{|p_n(i)|}{\min_\nu \left\{ \frac{1 + x_\nu^2}{1 + x_\nu} |p_n'(x_\nu)| \right\}}, \tag{4.5'}$$

where $p_n(x)$ is the polynomial in (4.3), *and the minimum is taken over all nonnegative abscissas.*

Proof. For the sake of definiteness we assume

$$x_\nu > 0 \quad \text{for } \nu = 1, 2, \ldots, [n/2], \qquad x_{(n+1)/2} = 0 \quad \text{if } n \text{ is odd.} \tag{4.6}$$

Let first n be even. The Lagrange polynomials (3.3) then are

$$l_\nu(x) = \frac{x + x_\nu}{2x_\nu} \prod_{\substack{\mu=1 \\ \mu \neq \nu}}^{n/2} \frac{x^2 - x_\mu^2}{x_\nu^2 - x_\mu^2}, \qquad l_{n+1-\nu}(x) = l_\nu(-x), \qquad \nu = 1, 2, \ldots, \frac{n}{2}.$$

It suffices in (3.4) to evaluate the sums $\sum_{\mu=1}^n |a_{\nu\mu}|$ for $1 \leqq \nu \leqq n/2$, the others (for $\nu > n/2$) having the same values. An application of (3.3), (3.4) and Lemma 2.2, in which n is to be replaced by $(n/2) - 1$, and s and t by

$$s = \frac{1}{2 \prod_{\substack{\mu=1 \\ \mu \neq \nu}}^{n/2} (x_\nu^2 - x_\mu^2)}, \qquad t = \frac{1}{2x_\nu \prod_{\substack{\mu=1 \\ \mu \neq \nu}}^{n/2} (x_\nu^2 - x_\mu^2)},$$

then gives the first result in (4.5). The second, for n odd, is obtained similarly, noting that

$$l_\nu(x) = \frac{x}{x_\nu} \frac{x + x_\nu}{2x_\nu} \prod_{\substack{\mu=1 \\ \mu \neq \nu}}^{(n-1)/2} \frac{x^2 - x_\mu^2}{x_\nu^2 - x_\mu^2}, \qquad l_{n+1-\nu}(x) = l_\nu(-x), \qquad \nu = 1, 2, \ldots, \frac{n-1}{2},$$

$$l_{(n+1)/2}(x) = \prod_{\mu=1}^{(n-1)/2} \frac{x^2 - x_\mu^2}{(-x_\mu^2)}.$$

The alternative form (4.5′) follows readily from (4.5) by observing that

$$p_n(x) = \prod_{\mu=1}^{n/2} (x^2 - x_\mu^2) \quad \text{if } n \text{ is even,}$$

and

$$p_n(x) = x \prod_{\mu=1}^{(n-1)/2} (x^2 - x_\mu^2) \quad \text{if } n \text{ is odd.}$$

Corollary. *If n is even and x_ν are symmetric points as in* (4.4), *then* (4.1) *holds with strict inequality.*

Proof. The bound in (4.1), again assuming (4.6), is

$$\frac{1}{2} \max_{1 \le \nu \le n/2} \left\{ \left(1 + \frac{1}{x_\nu}\right) \prod_{\substack{\mu=1 \\ \mu \ne \nu}}^{n/2} \frac{(1+x_\mu)^2}{|x_\nu^2 - x_\mu^2|} \right\},$$

which is larger than the top expression in (4.5) because of $(1+x_\mu)^2 > 1 + x_\mu^2$.

We next consider norm estimates for V_n^{-1} in the case of pairwise conjugate complex abscissas all located in the same half plane.

Theorem 4.4. *Let x_ν be distinct complex numbers such that*

$$x_{n+1-\nu} = \bar{x}_\nu \quad \text{for } \nu = 1, 2, \ldots, n \quad \text{and} \quad x_{(n+1)/2} = 0 \quad \text{if } n \text{ is odd,} \tag{4.7}$$

and such that for all ν either $\operatorname{Re} x_\nu \ge 0$ *or* $\operatorname{Re} x_\nu \le 0$. *If $V_n = V(x_1, x_2, \ldots, x_n)$, we then have for n even,*

$$\begin{aligned} &\max_{1 \le \nu \le n/2} \left\{ \frac{|1 - |x_\nu||}{|x_\nu - \bar{x}_\nu|} \prod_{\substack{\mu=1 \\ \mu \ne \nu}}^{n/2} \frac{|1 \pm x_\mu|^2}{|x_\nu - x_\mu|\,|x_\nu - \bar{x}_\mu|} \right\} \\ &\quad \le \|V_n^{-1}\|_\infty \le \max_{1 \le \nu \le n/2} \left\{ \frac{1 + |x_\nu|}{|x_\nu - \bar{x}_\nu|} \prod_{\substack{\mu=1 \\ \mu \ne \nu}}^{n/2} \frac{|1 \pm x_\mu|^2}{|x_\nu - x_\mu|\,|x_\nu - \bar{x}_\mu|} \right\}, \end{aligned} \tag{4.8}$$

and for n odd,

$$\begin{aligned} &\max_{1 \le \nu \le (n+1)/2} \left\{ \varepsilon_\nu |1 - |x_\nu|| \prod_{\substack{\mu=1 \\ \mu \ne \nu}}^{(n+1)/2} \frac{|1 \pm x_\mu|^2}{|x_\nu - x_\mu|\,|x_\nu - \bar{x}_\mu|} \right\} \\ &\quad \le \|V_n^{-1}\|_\infty \le \max_{1 \le \nu \le (n+1)/2} \left\{ \varepsilon_\nu (1 + |x_\nu|) \prod_{\substack{\mu=1 \\ \mu \ne \nu}}^{(n+1)/2} \frac{|1 \pm x_\mu|^2}{|x_\nu - x_\mu|\,|x_\nu - \bar{x}_\mu|} \right\}, \end{aligned} \tag{4.9}$$

where plus signs hold if $\operatorname{Re} x_\nu \ge 0$, *minus signs if* $\operatorname{Re} x_\nu \le 0$, *and where in* (4.9)

$$\varepsilon_{(n+1)/2} = 1, \qquad \varepsilon_\nu = \frac{|x_\nu|}{|x_\nu - \bar{x}_\nu|} \quad \text{for } 1 \le \nu \le (n-1)/2. \tag{4.10}$$

Alternatively,

$$\frac{|p_n(\mp 1)|}{\min_\nu \left\{ \frac{|1 \pm x_\nu|^2}{|1 - |x_\nu||} |p_n'(x_\nu)| \right\}} \le \|V_n^{-1}\|_\infty \le \frac{|p_n(\mp 1)|}{\min_\nu \left\{ \frac{|1 \pm x_\nu|^2}{1 + |x_\nu|} |p_n'(x_\nu)| \right\}}, \tag{4.11}$$

where $p_n(x)$ is the polynomial in (4.3), *and the minimum is taken over all ν with $1 \le \nu \le n/2$ when n is even, and over all ν with $1 \le \nu \le (n+1)/2$ when n is odd.*

We omit the proof of Theorem 4.4, since it is analogous to the proof of Theorem 4.3. Lemma 2.3 now plays the role of Lemma 2.2.

5. Scaling of the Abscissas

Let

$$V_n(\omega) = V(\omega x_1, \omega x_2, \ldots, \omega x_n), \quad \omega > 0.$$

How does the norm of $V_n^{-1}(\omega)$ compare with the norm of $V_n^{-1}(1)$? We shall answer this question first for positive abscissas x_ν, and then for symmetric real abscissas.

Theorem 5.1. *Let x_ν be distinct positive numbers. Then for $\omega > 0$,*

$$\frac{\omega}{\omega+1}\left|\frac{p_n\left(-\frac{1}{\omega}\right)}{p_n(-1)}\right| < \frac{\|V_n^{-1}(\omega)\|_\infty}{\|V_n^{-1}(1)\|_\infty} < (\omega+1)\left|\frac{p_n\left(-\frac{1}{\omega}\right)}{p_n(-1)}\right|, \tag{5.1}$$

where $p_n(x)$ is the polynomial in (4.3).

Proof. From (4.1′) we obtain

$$\|V_n^{-1}(\omega)\|_\infty = \frac{\left|p_n\left(-\frac{1}{\omega}\right)\right|}{\min\limits_{1\leqq\nu\leqq n}\left\{\left(\frac{1}{\omega}+x_\nu\right)|p_n'(x_\nu)|\right\}}$$

$$= \left|\frac{p_n\left(-\frac{1}{\omega}\right)}{p_n(-1)}\right| \cdot \frac{|p_n(-1)|}{\min\limits_{1\leqq\nu\leqq n}\{g_\omega(x_\nu)(1+x_\nu)|p_n'(x_\nu)|\}},$$

where

$$g_\omega(t) = \frac{\frac{1}{\omega}+t}{1+t}, \quad 0 \leqq t < \infty.$$

The theorem follows by observing (4.1′) and

$$\frac{1}{\omega+1} < g_\omega(t) < \frac{\omega+1}{\omega}, \quad 0 \leqq t < \infty.$$

Theorem 5.2. *Let n be even and x_ν be distinct real numbers such that*

$$x_\nu + x_{n+1-\nu} = 0 \quad \text{for } \nu = 1, 2, \ldots, n.$$

Then, for $\omega > 0$,

$$\frac{2(\sqrt{2}-1)\,\omega}{\omega+1}\left|\frac{p_n\left(\frac{i}{\omega}\right)}{p_n(i)}\right| < \frac{\|V_n^{-1}(\omega)\|_\infty}{\|V_n^{-1}(1)\|_\infty} < \frac{\omega+1}{2(\sqrt{2}-1)}\left|\frac{p_n\left(\frac{i}{\omega}\right)}{p_n(i)}\right|, \tag{5.2}$$

where $p_n(x)$ is the polynomial in (4.3).

Proof. From (4.5′) we obtain

$$\|V_n^{-1}(\omega)\|_\infty = \left|\frac{p_n\left(\frac{i}{\omega}\right)}{p_n(i)}\right| \cdot \frac{|p_n(i)|}{\min\limits_{\nu}\left\{g_\omega(x_\nu)\,\frac{1+x_\nu^2}{1+x_\nu}\,|p_n'(x_\nu)|\right\}},$$

where now

$$g_\omega(t) = \frac{\frac{1}{\omega^2}+t^2}{1+t^2}\,\frac{1+t}{\frac{1}{\omega}+t}, \quad 0 \leqq t < \infty.$$

We need to show that

$$\frac{2(\sqrt{2}-1)}{\omega+1} < g_\omega(t) < \frac{\omega+1}{2(\sqrt{2}-1)\,\omega}, \qquad 0 \leqq t < \infty. \tag{5.3}$$

We first note the identities

$$g_\omega(t) = \frac{1}{\omega} g_{1/\omega}\left(\frac{1}{t}\right), \qquad g_\omega(t) = \frac{1}{g_{1/\omega}(\omega t)}. \tag{5.4}$$

If $0 \leqq t \leqq 1$, the lower bound in (5.3) follows from

$$g_\omega(t) \geqq \frac{\frac{1}{\omega^2}+t^2}{\frac{1}{\omega}+t} = \frac{1}{\omega}\,\frac{1+\omega^2 t^2}{1+\omega t} > \frac{1}{\omega+1}\,\frac{1+\omega^2 t^2}{1+\omega t},$$

since $(1+y^2)/(1+y)$ for $y>0$ assumes the minimum value $2(\sqrt{2}-1)$ at $y=\sqrt{2}-1$. If $t>1$, we use the first identity in (5.4) to obtain again

$$g_\omega(t) > \frac{1}{\omega}\,\frac{2(\sqrt{2}-1)}{\frac{1}{\omega}+1} = \frac{2(\sqrt{2}-1)}{\omega+1}.$$

Combining the left inequality in (5.3) just established with the second identity in (5.4) gives the right inequality, and thus proves (5.3).

6. Examples

Norm estimates for V_n^{-1} imply estimates for the condition number of V_n. These in turn are of interest, e.g., in the study of the condition of polynomial interpolation [3]. In the examples which follow we derive asymptotic estimates for the condition number, assuming typical configurations of interpolation points.

Example 6.1 (equidistant points). $x_\nu = 1 - \frac{2(\nu-1)}{n-1}$, $\nu = 1, 2, \ldots, n$. We assume first n even. From (4.5) we find after some computation that

$$\|V_n^{-1}\|_\infty = \frac{\alpha_n}{\min\limits_{1 \leqq \nu \leqq n/2} \pi_\nu},$$

where

$$\alpha_n = \frac{1}{4^{n-2}} \left| \frac{\Gamma(n+i(n-1))}{\Gamma\left(\frac{n}{2}+i\,\frac{n-1}{2}\right)} \right|^2 \left| \frac{\Gamma\left(1+i\,\frac{n-1}{2}\right)}{\Gamma(1+i(n-1))} \right|^2,$$

$$\pi_\nu = [(n-1)^2 + (2\nu-1)^2] \left(\frac{n}{2}-\nu\right)! \left(\frac{n}{2}+\nu-2\right)!.$$

Since π_ν is increasing,

$$\min_{1 \leqq \nu \leqq n/2} \pi_\nu = \pi_1 = [(n-1)^2+1] \left(\frac{n}{2}-1\right)!^2,$$

and since $|\Gamma(1+iy)|^2 = |iy\Gamma(iy)|^2 = \pi y/\sinh(\pi y)$ for any real y, we obtain

$$\|V_n^{-1}\|_\infty = \frac{8}{4^n[(n-1)^2+1]\left(\frac{n}{2}-1\right)!^2}\,\frac{\sinh(\pi(n-1))}{\sinh\left(\pi\,\frac{n-1}{2}\right)} \left| \frac{\Gamma(n+i(n-1))}{\Gamma\left(\frac{n}{2}+i\,\frac{n-1}{2}\right)} \right|^2 \tag{6.1e}$$

(n even).

For n odd, we find similarly,

$$\|V_n^{-1}\|_\infty = \frac{\sinh\left(\pi\frac{n-1}{2}\right)}{\pi\frac{n-1}{2}\left(\frac{n-1}{2}\right)!^2}\left|\Gamma\left(\frac{n+1}{2}+i\frac{n-1}{2}\right)\right|^2 \quad (n \text{ odd}). \tag{6.1o}$$

Since

$$\operatorname{cond}_\infty V_n = \|V_n\|_\infty \|V_n^{-1}\|_\infty = n\|V_n^{-1}\|_\infty, \tag{6.2}$$

using Stirling's formula for the gamma function, and straightforward, but tedious, manipulations, we find from (6.1) that

$$\operatorname{cond}_\infty V_n \sim \frac{1}{\pi} e^{-\frac{\pi}{4}} e^{n\left(\frac{\pi}{4}+\frac{1}{2}\ln 2\right)}, \quad n \to \infty. \tag{6.3}$$

Some numerical values[1] are listed in Table 1.

Table 1. Condition of polynomial interpolation at equidistant points on $[-1, 1]$

n	$\operatorname{cond}_\infty V_n$	(6.3)
5	5.0000 (1)	4.1668 (1)
10	1.3625 (4)	1.1963 (4)
20	1.0535 (9)	9.8614 (8)
40	6.9269 (18)	6.7007 (18)
80	3.1456 (38)	3.0937 (38)

Example 6.2 (Chebyshev points). $x_\nu = \cos\theta_\nu$, $\theta_\nu = \frac{2\nu-1}{2n}\pi$, $\nu = 1, 2, \ldots, n$.

The abscissas x_ν are the zeros of the Chebyshev polynomial of the first kind, $T_n(x)$. Hence, by (4.5'), since $|T_n'(x_\nu)| = n/\sin\theta_\nu$, we find that

$$\|V_n^{-1}\|_\infty = \frac{|T_n(i)|}{n \cdot \min_\nu f(\theta_\nu)}, \tag{6.4}$$

where

$$f(\theta) = \frac{1+\cos^2\theta}{(1+\cos\theta)\sin\theta}, \quad 0 < \theta \leqq \pi/2.$$

An elementary calculation shows that $f(\theta)$ has a unique minimum on $[0, \pi/2]$, which is assumed at $\theta = \theta_0$, where

$$\cos\theta_0 = 2-\sqrt{3}, \quad f(\theta_0) = \frac{6-2\sqrt{3}}{3\sqrt{4\sqrt{3}-6}} = 2\cdot 3^{-3/4}.$$

Since the angles $\theta_\nu = \theta_{\nu,n}$ are equidistributed on the arc $0 \leqq \theta \leqq \pi/2$, there exists a sequence of integers ν_n with $0 < \nu_n < n$ such that $\theta_{\nu_n, n} \to \theta_0$ as $n \to \infty$. From

$$f(\theta_0) \leqq \min_\nu f(\theta_{\nu,n}) \leqq f(\theta_{\nu_n, n})$$

1 The integers in parentheses indicate powers of 10 by which the preceding numbers are to be multiplied.

it then follows that

$$\min_{\nu} f(\theta_{\nu,n}) \to f(\theta_0) \quad \text{as } n \to \infty.$$

Consequently, by (6.4),

$$\|V_n^{-1}\|_\infty \sim \frac{3^{3/4}}{2n} |T_n(i)|, \quad n \to \infty.$$

On the other hand [4, p. 194],

$$|T_n(i)| \sim \tfrac{1}{2}(1+\sqrt{2})^n, \quad n \to \infty,$$

so that, in view of (6.2),

$$\operatorname{cond}_\infty V_n \sim \frac{3^{3/4}}{4}(1+\sqrt{2})^n, \quad n \to \infty. \tag{6.5}$$

Some numerical values are listed in Table 2.

Table 2. Condition of polynomial interpolation at Chebyshev points

n	$\operatorname{cond}_\infty V_n$	(6.5)
5	4.1000 (1)	4.6737 (1)
10	3.7495 (3)	3.8330 (3)
20	2.5727 (7)	2.5781 (7)
40	1.1663 (15)	1.1663 (15)
80	2.3859 (30)	2.3869 (30)

Example 6.3. $x_\nu = 1 - e^{-i\omega_\nu h}$, $\nu = 1, 2, \ldots, n$ (even), $h > 0$,

$$0 < \omega_1 < \omega_2 < \cdots < \omega_{n/2}, \quad \omega_{\nu+n/2} = -\omega_\nu \quad \text{for } \nu = 1, 2, \ldots, n/2.$$

The interpolation problem corresponding to these complex abscissas arises in the construction of trigonometric multistep methods for ordinary differential equations with almost periodic solutions [1].

A short calculation, based on (4.11), gives the upper bound

$$\|V_n^{-1}\|_\infty \tag{6.6}$$
$$\leqq \frac{\prod_{\mu=1}^{n/2}\left[1+8\sin^2\left(\frac{1}{2}\omega_\mu h\right)\right]}{\min\limits_{1\leqq\nu\leqq n/2}\left\{\frac{1+8\sin^2(\frac{1}{2}\omega_\nu h)}{1+2\sin(\frac{1}{2}\omega_\nu h)}\cdot 2\sin(\omega_\nu h)\prod\limits_{\substack{\mu=1\\ \mu\neq\nu}}^{n/2}\left[4\sin\frac{1}{2}(\omega_\nu+\omega_\mu)h\sin\frac{1}{2}|\omega_\nu-\omega_\mu|h\right]\right\}},$$

and a similar lower bound in which $1+2\sin(\frac{1}{2}\omega_\nu h)$ in the denominator of (6.6) is replaced by $1-2\sin(\frac{1}{2}\omega_\nu h)$. For n fixed, and $h \to 0$, we find

$$\|V_n^{-1}\|_\infty \sim \frac{1}{2h^{n-1}\min\limits_{1\leqq\nu\leqq n/2}\left\{\omega_\nu\prod\limits_{\substack{\mu=1\\ \mu\neq\nu}}^{n/2}|\omega_\nu^2-\omega_\mu^2|\right\}} \quad (h\to 0). \tag{6.7}$$

The estimate (6.7) can also be obtained by using the approximations $x_\mu \doteq i\omega_\mu h$, rotating the abscissas through an angle of $\pi/2$, and then applying Theorem 4.3 with the simplifying approximation $(1+\omega_\nu h)\prod_{\mu\neq\nu}(1+\omega_\mu^2 h^2) \doteq 1$.

Example 6.4 (Roots of unity). $x_\nu = e^{2\pi i\nu/n}$, $\nu = 1, 2, \ldots, n$.

Although none of the previous estimates apply, we can obtain the inverse of the Vandermonde matrix directly by observing that the Lagrange interpolation polynomials are

$$l_\nu(x) = \frac{1}{n} \sum_{\mu=1}^{n} \left(\frac{x}{x_\nu}\right)^{\mu-1}, \qquad \nu = 1, 2, \ldots, n.$$

Consequently, by (3.3), (3.4),

$$\|V_n^{-1}\|_\infty = 1, \qquad \operatorname{cond}_\infty V_n = n.$$

Actually, the roots of unity are an optimal point configuration with regard to the spectral condition of Vandermonde matrices [3]. In fact, since $V_n^H V_n = n \cdot I_n$, we have $\operatorname{cond}_2 V_n = 1$.

References

1. Bettis, D. G.: Numerical integration of products of Fourier and ordinary polynomials. Numer. Math. **14**, 421–434 (1970)
2. Gautschi, W.: On inverses of Vandermonde and confluent Vandermonde matrices. Numer. Math. **4**, 117–123 (1962)
3. Singhal, K., Vlach, J.: Accuracy and speed of real and complex interpolation. Computing **11**, 147–158 (1973)
4. Szegö, G.: Orthogonal polynomials, 2nd rev. ed., Amer. Math. Soc. Colloq. Publ., Vol. **23**, Amer. Math. Soc., Providence, R.I., 1959

Prof. W. Gautschi
Department of Computer Sciences
Purdue University
Lafayette, Indiana 47907/U.S.A.

8.6. [62] "On Inverses of Vandermonde and Confluent Vandermonde Matrices III"

---

[62] "On Inverses of Vandermonde and Confluent Vandermonde Matrices III," *Numer. Math*. **29**, 445–450 (1978).

Numer. Math. 29, 445–450 (1978)

Numerische Mathematik

On Inverses of Vandermonde and Confluent Vandermonde Matrices III*

Walter Gautschi

Department of Computer Sciences, Purdue University Lafayette, IN 47907, USA

Summary. We derive lower bounds for the norm of the inverse Vandermonde matrix and the norm of certain inverse confluent Vandermonde matrices. They supplement upper bounds which were obtained in previous papers.

Subject Classifications. AMS(MOS): 15A12, 65F35; CR: 5.14.

1. Introduction

Norm estimates for the inverse of a Vandermonde matrix, or the inverse of confluent Vandermonde matrices, have been the subject of several previous papers [1, 2, 4]. The emphasis there was on upper bounds in the case of general complex nodes, or identities when the nodes are positive [1, 2] or real and symmetric with respect to the origin [4]. We now wish to supplement these results by providing lower bounds in the case of arbitrary complex nodes. We obtain these bounds by applying to appropriate polynomials Jensen's formula in the theory of analytic functions.

2. Jensen's Formula for Polynomials

Given a polynomial

$$p(z)=a_0+a_1 z+\cdots+a_n z^n, \quad a_n\neq 0, \tag{2.1}$$

with complex coefficients a_μ, let $\zeta_1, \zeta_2, \ldots, \zeta_n$ denote its zeros ordered such that

$$|\zeta_1|\leqq|\zeta_2|\leqq\cdots\leqq|\zeta_r|\leqq 1<|\zeta_{r+1}|\leqq|\zeta_{r+2}|\leqq\cdots\leqq|\zeta_n|.$$

Jensen's formula, applied to (2.1) on the unit circle, then gives [6]

* Sponsored in part by the United States Army under Contract No. DAAG29-75-C-0024 and the National Science Foundation under grant MCS 76-00842A01

$$|a_n \zeta_{r+1} \zeta_{r+2} \cdots \zeta_n| = \exp\left(\frac{1}{2\pi} \int_0^{2\pi} \ln |p(e^{i\theta})|\, d\theta\right),$$

hence, letting $M = \max_{0 \leqq \theta \leqq 2\pi} |p(e^{i\theta})|$,

$$|a_n \zeta_{r+1} \zeta_{r+2} \cdots \zeta_n| \leqq M \leqq \sum_{\mu=0}^{n} |a_\mu|. \tag{2.2}$$

Thus,

$$\sum_{\mu=0}^{n} |a_\mu| \geqq |a_n| \prod_{\nu=1}^{n} \max(1, |\zeta_\nu|). \tag{2.3}$$

Equality in (2.3) holds if and only if $a_0 = a_1 = \cdots = a_{n-1} = 0$, i.e., $p(z) = a_n z^n$. Indeed, if $p(z) = a_n z^n$, then (2.3) (with equality) is trivial. Conversely, if we have equality in (2.3), we must have equality in (2.2), hence, by Jensen's formula, $|p(e^{i\theta})| \equiv M$ for $0 \leqq \theta \leqq 2\pi$. Since

$$|p(e^{i\theta})|^2 = \sum_{k,l=0}^{n} a_k \bar{a}_l e^{i(k-l)\theta} = \sum_{\lambda=-n}^{n} c_\lambda e^{i\lambda\theta}$$

is a trigonometric polynomial, with coefficients

$$c_\lambda = \sum_{k=-\infty}^{\infty} a_k \bar{a}_{k-\lambda}, \qquad c_{-\lambda} = \bar{c}_\lambda$$

(the convention $a_\mu = 0$ if $\mu < 0$ or $\mu > n$ is used here), it can be constant equal to M^2 only if $c_n = c_{n-1} = \cdots = c_1 = 0$ and $c_0 = M^2$. The first condition, $c_n = 0$, implies $a_n \bar{a}_0 = 0$, hence $a_0 = 0$ (since $a_n \neq 0$). The second condition, $a_n \bar{a}_1 + a_{n-1} \bar{a}_0 = 0$, then gives $a_1 = 0$, and continuing in this manner, we find recursively $a_0 = a_1 = \cdots = a_{n-1} = 0$.

3. Inverse Vandermonde Matrix

We denote the Vandermonde matrix of order n by

$$V_n(z) = \begin{pmatrix} 1 & 1 & \dots & 1 \\ z_1 & z_2 & \dots & z_n \\ \vdots & \vdots & & \vdots \\ z_1^{n-1} & z_2^{n-1} & \dots & z_n^{n-1} \end{pmatrix}, \tag{3.1}$$

where $z^T = [z_1, z_2, \dots, z_n]$ is a vector of n complex numbers, called "nodes". If the nodes are mutually distinct, then $V_n(z)$ has an inverse, which we denote by

$$V_n^{-1}(z) = [u_{\lambda\mu}]_{\lambda,\mu=1}^{n}. \tag{3.2}$$

We are interested in the l_∞-norm of (3.2),

$$\|V_n^{-1}(z)\|_\infty = \max_{1 \leqq \lambda \leqq n} \sum_{\mu=1}^{n} |u_{\lambda\mu}|.$$

Theorem 3.1. *If $z_1, z_2, \ldots, z_n$ are mutually distinct complex numbers, and $n>1$, then*

$$\|V_n^{-1}(z)\|_\infty > \max_{1\leqq\lambda\leqq n} \prod_{\substack{\nu=1\\ \nu\neq\lambda}}^{n} \frac{\max(1,|z_\nu|)}{|z_\lambda - z_\nu|}. \tag{3.3}$$

Proof. We recall [4] that the elements $u_{\lambda\mu}$ in (3.2) are the coefficients of the fundamental Lagrange interpolation polynomials associated with the nodes z_ν,

$$\prod_{\substack{\nu=1\\ \nu\neq\lambda}}^{n} \frac{z-z_\nu}{z_\lambda - z_\nu} = u_{\lambda 1} + u_{\lambda 2} z + \cdots + u_{\lambda n} z^{n-1}. \tag{3.4}$$

Applying (2.3) and the remark following (2.3) to the polynomial of degree $n-1$ in (3.4), we find

$$\sum_{\mu=1}^{n} |u_{\lambda\mu}| > \left(\prod_{\nu\neq\lambda} \frac{1}{|z_\lambda - z_\nu|}\right) \prod_{\nu\neq\lambda} \max(1,|z_\nu|). \tag{3.5}$$

If λ_0 is the index λ for which the right-hand expression in (3.5) attains its maximum, then that maximum is less than $\sum_{\mu=1}^{n} |u_{\lambda_0\mu}|$, hence less or equal than $\max_{1\leqq\lambda\leqq n} \sum_{\mu=1}^{n} |u_{\lambda\mu}|$. This establishes (3.3) and proves Theorem 3.1.

The lower bound in (3.3) supplements the upper (attainable) bound in [1], which is of the same form as (3.3) except that the l_∞-norm of the 2-vectors $[1, z_\nu]$ in the numerator factors is replaced by the l_1-norm.

4. Inverse Confluent Vandermonde Matrices

The technique used in the proof of Theorem 3.1 can be adapted to confluent Vandermonde matrices. We illustrate this with the particular matrix

$$U_{2n}(z) = \begin{pmatrix} 1 & 1 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ z_1 & z_2 & \ldots & z_n & 1 & 1 & \ldots & 1 \\ z_1^2 & z_2^2 & \ldots & z_n^2 & 2z_1 & 2z_2 & \ldots & 2z_n \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ z_1^{2n-1} & z_2^{2n-1} & \ldots & z_n^{2n-1} & (2n-1)z_1^{2n-2} & (2n-1)z_2^{2n-2} & \ldots & (2n-1)z_n^{2n-2} \end{pmatrix} \tag{4.1}$$

considered previously in [1], [2].

Theorem 4.1. *If $z_1, z_2, \ldots, z_n$ are mutually distinct complex numbers, and $n>1$, then*

$$\|U_{2n}^{-1}(z)\|_\infty > \max_{1\leqq\lambda\leqq n} b_\lambda \prod_{\substack{\nu=1\\ \nu\neq\lambda}}^{n} \left(\frac{\max(1,|z_\nu|)}{|z_\lambda - z_\nu|}\right)^2, \tag{4.2}$$

where b_λ is the larger of the two quantities

$$b_\lambda^{(1)} = \max(1, |z_\lambda|),$$
$$b_\lambda^{(2)} = \max\left(2\left|\sum_{\nu\neq\lambda} 1/(z_\lambda - z_\nu)\right|, \left|1 + 2z_\lambda \sum_{\nu\neq\lambda} 1/(z_\lambda - z_\nu)\right|\right). \tag{4.3}$$

Proof. We have [2]

$$U_{2n}^{-1} = \begin{bmatrix} V \\ W \end{bmatrix}, \qquad V = [v_{\lambda\mu}], \qquad W = [w_{\lambda\mu}],$$

where

$$\begin{aligned} l_\lambda^2(z)[1 - 2l_\lambda'(z_\lambda)(z - z_\lambda)] &= \sum_{\mu=1}^{2n} v_{\lambda\mu} z^{\mu-1} \\ & \qquad 1 \leqq \lambda \leqq n, \\ l_\lambda^2(z)(z - z_\lambda) &= \sum_{\mu=1}^{2n} w_{\lambda\mu} z^{\mu-1} \end{aligned} \tag{4.4}$$

and $l_\lambda(z)$ denotes the fundamental Lagrange interpolation polynomial in (3.4). Applying (2.3) to the polynomials in (4.4), and taking note of the remark following (2.3), one finds

$$\sum_{\mu=1}^{2n} |v_{\lambda\mu}| > b_\lambda^{(2)} \prod_{\nu\neq\lambda} \left(\frac{\max(1, |z_\nu|)}{|z_\lambda - z_\nu|}\right)^2,$$
$$\sum_{\mu=1}^{2n} |w_{\lambda\mu}| > b_\lambda^{(1)} \prod_{\nu\neq\lambda} \left(\frac{\max(1, |z_\nu|)}{|z_\lambda - z_\nu|}\right)^2,$$

where $b_\lambda^{(1)}$, $b_\lambda^{(2)}$ are as defined in (4.3). Denoting the products $\prod_{\nu\neq\lambda}$ on the right by π_λ, and observing that $\|U_{2n}^{-1}\|_\infty = \max\left(\max_\lambda \sum_{\mu=1}^{2n} |v_{\lambda\mu}|, \max_\lambda \sum_{\mu=1}^{2n} |w_{\lambda\mu}|\right)$, an argument similar to the one after (3.5) will show that $b_\lambda^{(1)}\pi_\lambda < \|U_{2n}^{-1}\|_\infty$, $b_\lambda^{(2)}\pi_\lambda < \|U_{2n}^{-1}\|_\infty$ for all $\lambda = 1, 2, \ldots, n$, hence $\max(b_\lambda^{(1)}, b_\lambda^{(2)})\pi_\lambda < \|U_{2n}^{-1}\|_\infty$ for all $\lambda = 1, 2, \ldots, n$. This proves Theorem 4.1.

The lower bound in (4.2) supplements the (attainable) upper bound in [2], which is of the same form as (4.2) except that the l_∞-norm of the 2-vectors $[1, z_\nu]$ in the numerator factors, and the l_∞-norms defining $b_\lambda^{(1)}$ and $b_\lambda^{(2)}$, are all replaced by the respective l_1-norms. In the case of positive nodes z_ν another (usually sharper) lower bound can be found in [3, Theorem 2.1].

5. Examples

Example 5.1 (roots of unity). $z_\nu = e^{2\pi i(\nu-1)/n}$, $\nu = 1, 2, \ldots, n$. In view of

$$l_\lambda(z) = \frac{1}{n} \sum_{\mu=1}^{n} \left(\frac{z}{z_\lambda}\right)^{\mu-1}, \qquad \lambda = 1, 2, \ldots, n,$$

Table 1. Norm estimates for Example 5.2

		$\|V_n^{-1}\|_\infty$			$\|U_{2n}^{-1}\|_\infty$		
N	n	lower	true	upper	lower	true	upper
5	3	7.24(−1)	1.89	2.89	1.57	1.79(1)	4.19(1)
10	6	1.17	1.47(1)	3.75(1)	8.29	2.36(3)	1.56(4)
15	8	4.25	2.03(2)	5.45(2)	1.46(2)	6.18(5)	4.48(6)
20	11	1.17(1)	2.76(3)	1.20(4)	1.52(3)	1.59(8)	3.03(9)

Table 2. Norm estimates for Example 5.3

	$\|V_n^{-1}\|_\infty$			$\|U_{2n}^{-1}\|_\infty$		
n	lower	true	upper	lower	true	upper
5	1.12	2.08	3.93	4.76	2.21(1)	7.90(1)
10	2.44	5.22	1.45(1)	4.45(1)	2.52(2)	1.96(3)
15	7.45	1.69(1)	5.71(1)	6.08(2)	3.69(3)	4.22(4)
20	2.27(1)	5.36(1)	2.02(2)	7.54(3)	4.79(4)	6.84(5)

Table 3. Norm estimates for Example 5.4

		$\|V_n^{-1}\|_\infty$			$\|U_{2n}^{-1}\|_\infty$		
N	n	lower	true	upper	lower	true	upper
5	3	1.62	2.74	3.13	6.69	2.51(1)	3.29(1)
10	5	5.71	1.09(1)	1.27(1)	1.38(2)	6.15(2)	8.35(2)
15	8	3.07(1)	6.82(1)	8.63(1)	5.71(3)	3.24(4)	5.19(4)
20	10	1.60(2)	3.60(2)	4.49(2)	1.88(5)	1.07(6)	1.66(6)

we obtain from (3.4), and from (4.4) after a little computation,

$$\|V_n^{-1}(z)\|_\infty = 1, \qquad \|U_{2n}^{-1}(z)\|_\infty = 2 - \frac{1}{n}. \tag{5.1}$$

The lower bounds in (3.3) and (4.2) both evaluate to $1/n$, while the upper bounds in [1], [2] are $2^{n-1}/n$ and $(2n-1)\,4^{n-1}/n^2$, respectively.

Example 5.2 (roots of unity on half-circle). $z_\nu = e^{2\pi i(\nu-1)/N}$, $\nu = 1, 2, \ldots, n$, where $n = [N/2]+1$.

The true norms of V_n^{-1} and U_{2n}^{-1}, as well as the lower bounds of Theorems 3.1 and 4.1 and the upper bounds in [1], [2] are shown in Table 1 for $N = 5(5)20$[1]. It is interesting to note how deletion of the roots of unity on a half-circle results in substantially larger values of $\|V_n^{-1}\|_\infty$ and $\|U_{2n}^{-1}\|_\infty$.

[1] The integers in parentheses indicate exponents of 10

Example 5.3. $e_n(z_\nu)=0$, $\nu=1, 2, \ldots, n$, where $e_n(z)=1+z+\frac{z^2}{2!}+\cdots+\frac{z^n}{n!}$. Using the zeros of e_n, tabulated in [5], we obtain the results in Table 2.

Example 5.4. $e_N(z_\nu)=0$, $\operatorname{Im} z_\nu \geqq 0$, $\nu=1, 2, \ldots, n$, where $n=\left[\frac{N+1}{2}\right]$.

Similarly as in Example 5.2, deletion of the zeros in the lower half-plane has the effect of increasing the norms of V_n^{-1} and U_{2n}^{-1} (see Table 3).

References

1. Gautschi, W.: On inverses of Vandermonde and confluent Vandermonde matrices. Numer. Math. **4**, 117–123 (1962)
2. Gautschi, W.: On inverses of Vandermonde and confluent Vandermonde matrices II. Numer. Math. **5**, 425–430 (1963)
3. Gautschi, W.: Construction of Gauss-Christoffel quadrature formulas. Math. Comput. **22**, 251–270 (1968)
4. Gautschi, W.: Norm estimates for inverses of Vandermonde matrices. Numer. Math. **23**, 337–347 (1975)
5. Iverson, K.E.: The zeros of the partial sums of e^z. Math. Tables Aids Comput. **7**, 165–168 (1953)
6. Mahler, K.: An application of Jensen's formula to polynomials. Mathematika **7**, 98 – 100 (1960)

Received July 4, 1977

8.7. [64] "QUESTIONS OF NUMERICAL CONDITION RELATED TO POLYNOMIALS"

---

[64] "Questions of Numerical Condition Related to Polynomials," in *Symposium on recent advances in numerical analysis* (C. de Boor and G. H. Golub, eds.), 45–72, Academic Press, New York, 1978. [Revised and reprinted in *MAA Studies in Mathematics* **24**: *Studies in numerical analysis* (G. H. Golub, ed.), 140–177 (1984)].

QUESTIONS OF NUMERICAL CONDITION RELATED TO POLYNOMIALS†

*Walter Gautschi**

1. INTRODUCTION

Polynomials (in one variable) permeate much of classical numerical analysis, either in the role of approximators, or as gauge functions for a variety of numerical methods, or in the role of characteristic polynomials of one kind or another. It seems appropriate, therefore, to study some of their basic properties as they relate to computation. In the following we wish to consider one particular aspect of polynomials, namely, the extent to which they, or quantities related to them, are sensitive to small perturbations. In other words, we are interested in the numerical condition of polynomials. We shall examine from this angle three particular

†Revised and in part reprinted (with permission of the publisher) from *Recent Advances in Numerical Analysis* (C. de Boor and G. H. Golub, eds.), pp. 45–72, Academic Press, New York, 1978.

*Supported in part by the National Science Foundation under Grant MCS 7927158A01.

problem areas: (1) The representation of polynomials (polynomial bases); (2) Algebraic equations; (3) The problem of orthogonalization. Before embarking on these topics, however, we must briefly consider ways of measuring the condition of problems. We do this in the framework of maps from one normed space into another, for which we define appropriate condition numbers.

2. THE CONDITION OF MAPS

2.1. Nonlinear maps. Let X, Y be normed linear spaces, and let $y = f(x)$ define a map M: $\mathscr{D} \subset X \to Y$, with $\mathscr{D}$ an open domain. Let $\mathring{x} \in \mathscr{D}$ be fixed, and $\mathring{y} = f(\mathring{x})$, and assume that neither $\mathring{x}$ nor $\mathring{y}$ is the zero element in the respective space. The sensitivity of the map M at $\mathring{x}$, with respect to small relative changes in $\mathring{x}$, will be measured by the (*asymptotic*) *condition number* (see Rice [27])

$$\operatorname{cond}(M; \mathring{x}) = \lim_{\delta \to 0} \sup_{\|h\| = \delta} \left\{ \frac{\|f(\mathring{x}+h) - f(\mathring{x})\|}{\|f(\mathring{x})\|} \Big/ \frac{\|h\|}{\|\mathring{x}\|} \right\}, \tag{2.1}$$

provided the limit exists. The number in (2.1) measures the maximum amount by which a relative perturbation of $\mathring{x}$ (given by $\delta / \|\mathring{x}\|$) is magnified under the map M, in the limit of infinitesimal perturbations. Maps with large condition numbers are called *ill-conditioned.*

If M has a Fréchet derivative $[\partial f / \partial x]_0$ at $\mathring{x}$, then

$$\operatorname{cond}(M; \mathring{x}) = \frac{\|\mathring{x}\|}{\|\mathring{y}\|} \left\| \left[\frac{\partial f}{\partial x} \right]_0 \right\| \qquad (\mathring{y} = f(\mathring{x})). \tag{2.2}$$

In the important case of finite-dimensional spaces, $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$, the Fréchet derivative, as is well known, is the linear map defined by the Jacobian matrix of f. We may then use in (2.2) any family of vector norms and subordinate family of matrix norms (see Stewart [31], p. 177).

For composite maps $K \circ M$, the chain rule for Fréchet derivatives (see Ortega & Rheinboldt [25], p. 62) can be used to show that

$$\operatorname{cond}(K \circ M; \mathring{x}) \leqslant \operatorname{cond}(K; \mathring{y})\operatorname{cond}(M; \mathring{x}). \tag{2.3}$$

If the composite map is known to be ill-conditioned, the inequality (2.3) permits us to infer the ill-conditioning of (at least) one of the component maps.

2.2. Linear maps. If $M: y = f(x)$ is a linear (bounded) map, then

$$\sup_{\|h\| = \delta} \frac{\|f(\mathring{x}+h) - f(\mathring{x})\|}{\|h\|} = \sup_{\|h\| = \delta} \frac{\|f(h)\|}{\|h\|}$$

is independent of $\mathring{x}$ and δ and equal to the norm of M. Equation (2.1) then reduces to

$$\operatorname{cond}(M; \mathring{x}) = \frac{\|\mathring{x}\|}{\|\mathring{y}\|}\|M\| \qquad (M \text{ linear}, \ \mathring{y} = M\mathring{x}). \tag{2.4}$$

If in addition M is invertible, we can ask for the supremum of (2.4) as $\mathring{x}$ varies in X (or, equivalently, $\mathring{y}$ varies in MX), and we find, since $\mathring{x} = M^{-1}\mathring{y}$, that

$$\sup_{x \in X} \operatorname{cond}(M; x) = \|M^{-1}\|\|M\|. \tag{2.5}$$

The number on the right, usually referred to as the *condition number of* M, will be denoted by

$$\operatorname{cond} M = \|M^{-1}\|\|M\|. \tag{2.6}$$

We have, alternatively,

$$\operatorname{cond} M = \frac{\sup_{x \in X} (\|Mx\|/\|x\|)}{\inf_{x \in X} (\|Mx\|/\|x\|)}. \tag{2.7}$$

Condition numbers such as those proposed cannot be expected to do more than convey general guidelines as to the susceptibility of the respective maps to small changes in the elements of their domains. By their very definition they reflect "worst case" situations and therefore are inherently conservative measures.

3. THE CONDITION OF POLYNOMIAL BASES

Let $\mathbb{P}_{n-1}$ denote the class of (real) polynomials of degree $\leqslant n-1$, and let $p_1, p_2, \ldots, p_n$ be a basis in $\mathbb{P}_{n-1}$. For any $p \in \mathbb{P}_{n-1}$, we denote by $u_1, u_2, \ldots, u_n$ the coefficients of p with respect to this basis,

$$p(x) = \sum_{k=1}^{n} u_k p_k(x). \tag{3.1}$$

We wish to determine how strongly the values of p on some given finite interval $[a, b]$ react to small perturbations in the coefficients u_k and, vice versa, how the coefficients of p are affected by small changes in p.

The question may be formalized as one concerning the condition of the linear map $M_n : \mathbb{R}^n \to \mathbb{P}_{n-1}[a, b]$, which associates to each vector $u^T = [u_1, u_2, \ldots, u_n] \in \mathbb{R}^n$ the polynomial p in (3.1), restricted to $[a, b]$,

$$(M_n u)(x) = \sum_{k=1}^{n} u_k p_k(x), \qquad a \leqslant x \leqslant b. \tag{3.2}$$

We are thus interested in the condition number of M_n, see (2.6),

$$\operatorname{cond} M_n = \|M_n^{-1}\| \, \|M_n\|, \tag{3.3}$$

in particular, how fast it grows as $n \to \infty$, and how this growth depends on the particular interval chosen.

For definiteness, we consider only uniform norms, i.e., $\|u\| = \max_{1 \leqslant k \leqslant n} |u_k|$ in $\mathbb{R}^n$, and $\|p\| = \max_{a \leqslant x \leqslant b} |p(x)|$ in $\mathbb{P}_{n-1}[a, b]$, although there are circumstances in which other norms may be preferable (see, e.g., Geurts [19], Gautschi [16], Sections 3.1, 3.2).

We shall use the notation u_p to denote the coefficient vector of p,

$$u_p = M_n^{-1} p, \qquad p \in \mathbb{P}_{n-1}. \tag{3.4}$$

3.1. Power basis. For the power basis

$$p_k(x) = x^{k-1}, \qquad k = 1, 2, \ldots, n, \tag{3.5}$$

it is natural to assume an interval $[a, b]$ that contains the origin. We shall do so in the following, but other intervals could also be treated (in fact more easily). For definiteness, we assume further that $[a, b]$ is centered to the right of the origin, i.e., $0 \leqslant |a| \leqslant b$. It then follows immediately that

$$\|M_n\| = \sup_{\|u\|=1} \max_{a \leqslant x \leqslant b} \left| \sum_{k=1}^{n} u_k x^{k-1} \right| = \sum_{k=1}^{n} b^{k-1},$$

hence

$$\|M_n\| = \frac{b^n - 1}{b - 1}. \tag{3.6}$$

(It is understood, here and below, that the value of the function on the right equals n if $b = 1$.)

For the inverse map M_n^{-1} we have

$$\|M_n^{-1}\| = \sup_{\|p\|=1} \max_{1 \leqslant k \leqslant n} \frac{|p^{(k-1)}(0)|}{(k-1)!} = \max_{1 \leqslant k \leqslant n} \sup_{\|p\|=1} \frac{|p^{(k-1)}(0)|}{(k-1)!}.$$

Therefore, in terms of the linear functionals $\lambda_k : \mathbb{P}_{n-1}[a, b] \to \mathbb{R}$ defined by $\lambda_k p = p^{(k-1)}(0)/(k-1)!$,

$$\|M_n^{-1}\| = \max_{1 \leqslant k \leqslant n} \|\lambda_k\|. \tag{3.7}$$

Our problem thus reduces to determining the norm of λ_k. This is related to the problem of best uniform approximation of binomials $f_{n,\tau}(x) = (1 - |\tau|)x^n + \tau x^{n-1}$, where $-1 \leqslant \tau \leqslant 1$, by polynomials g of degree $\leqslant n-2$, which, in turn, gives rise to the Zolotarev polynomials

$$z_{n,\tau}(x) = \frac{1}{E_{n,\tau}} \left(f_{n,\tau}(x) - g^*_{n,\tau}(x) \right), \qquad -1 \leqslant \tau \leqslant 1,$$

where

$$E_{n,\tau} = \inf_{g \in \mathbb{P}_{n-2}} \|f_{n,\tau} - g\| = \|f_{n,\tau} - g^*_{n,\tau}\|.$$

The extremal for the functional λ_k, indeed, is a Zolotarev polynomial (of degree $n-1$, since we are working with $\mathbb{P}_{n-1}[a, b]$), that is, for $2 \leqslant k \leqslant n$,

$$\|\lambda_k\| = \sup_{\|p\|=1} |\lambda_k p| = |\lambda_k z_{n-1,\tau}| \text{ for some } \tau \in [-1, 1] \quad (3.8)$$

(see, e.g., Schönhage [30], Satz 6.11). Unfortunately, if the interval $[a, b]$ is arbitrary, the value of the parameter τ in (3.8) is not easily expressible, and may be different for different values of k. The exact determination of $\|M_n^{-1}\|$ in (3.7) indeed is cumbersome (see Voronovskaja [33], Ch. III and the appendix by V. A. Gusev). Upper bounds for $\|M_n^{-1}\|$ are obtained in Gautschi ([14], Theorem 4.1).

For the power basis (3.5), the most natural interval, however, is an interval symmetric about the origin, $[-\omega, \omega]$, $\omega > 0$. In this case, the Zolotarev polynomials reduce to Chebyshev polynomials of the first kind (Schönhage [30], p. 167). Making use of this, it then follows readily from (3.6) and (3.7) that

$$\operatorname{cond} M_n = \frac{\omega^n - 1}{\omega - 1} \max\left\{\|u_{T_{n-1}(x/\omega)}\|, \|u_{T_{n-2}(x/\omega)}\|\right\}, \quad (3.9)$$

where T_m denotes the Chebyshev polynomial of degree m, and $u_{T_m(x/\omega)}$ the coefficient vector of $T_m(x/\omega)$; see (3.4).

Using asymptotic estimates of $\|u_{T_m(x/\omega)}\|$ as $m \to \infty$, Gautschi ([14], Eq. (2.2)), it can be deduced from (3.9) that

$$(\operatorname{cond} M_n)^{1/n} \sim \begin{cases} 1 + \sqrt{1+\omega^2}, & \omega \geqslant 1, \\ \dfrac{1+\sqrt{1+\omega^2}}{\omega}, & \omega < 1, \end{cases} \quad \text{as } n \to \infty. \quad (3.10)$$

TABLE 3.1

	$(\text{cond } M_n)^{1/n}$				
n	$\omega = .1$	$\omega = .2$	$\omega = 1$	$\omega = 5$	$\omega = 10$
5	9.767	5.743	2.091	3.789	6.444
10	13.977	7.579	2.377	4.616	8.027
20	16.719	8.706	2.437	5.210	9.252
40	18.286	9.330	2.447	5.588	10.023
⋮	⋮	⋮	⋮	⋮	⋮
∞	20.050	10.099	2.414	6.099	11.050

The condition of M_n on $[-\omega, \omega]$.

The condition of M_n on $[-\omega, \omega]$ thus grows exponentially with n, the asymptotic growth rate being smallest (equal to $1+\sqrt{2}$) when $\omega = 1$. Some numerical values are shown in Table 3.1.†

Similar results hold for intervals $[0, \omega]$, $\omega > 0$, in which case (Gautschi [14])

$$(\text{cond } M_n)^{1/n} \sim \begin{cases} (1+\sqrt{1+\omega})^2, & \omega \geq 1, \\ \dfrac{(1+\sqrt{1+\omega})^2}{\omega}, & \omega < 1, \end{cases} \quad \text{as } n \to \infty. \tag{3.11}$$

Again, the minimum growth rate ($= (1+\sqrt{2})^2$) occurs when $\omega = 1$.

It is interesting to note that exponential growth of the condition is also observed for piecewise polynomial functions if represented

† The information in Table 3.1 might suggest that the asymptotic growth rate is approached monotonically. The reader, however, will have noticed that the limit rate in the case $\omega = 1$ is smaller than the seemingly increasing approach rates! In reality, the approach is indeed monotone, if $\omega < 1$, but changes from increasing to decreasing, if $\omega = 1$, and from decreasing to increasing, if $\omega > 1$. The changeover occurs near $n = 35$, if $\omega = 1$ (hence is not visible in Table 3.1), and near $(e/2\pi)\omega$, if $\omega \gg 1$. The latter would begin to be visible in Table 3.1 if $(e/2\pi)\omega \doteq 10$, i.e., $\omega \doteq 23$. The reason for this behavior can be found in the more precise relations $\text{cond } M_n \sim (\gamma^2 n)^{1/2}\rho^n$ if $\omega = 1$, and $\text{cond } M_n \sim (\gamma^2/n)^{1/2}\rho^n$, if $\omega \neq 1$, where ρ is the limit rate and $\gamma = \gamma(\omega)$ can be explicitly computed; Gautschi [14].

in terms of normalized B-splines. In fact, for splines of degree $k-1$, the condition of the B-spline basis is known to lie between $(1-1/k)2^{k-3/2}$ and $2k\cdot 9^k$; see de Boor [2], Lyche [23]. Empirical evidence seems to suggest that the conditon is indeed $O(2^k)$; de Boor [3].

3.2. Bases of orthogonal polynomials. We now consider the case of an orthogonal basis, i.e.,

$$p_k(x) = \pi_{k-1}(x), \qquad k = 1,2,\ldots,n, \tag{3.12}$$

where $\pi_0, \pi_1, \ldots, \pi_{n-1}$ are the first n of a sequence of polynomials orthogonal on the (finite) interval $[a, b]$ with respect to a nonnegative measure $d\sigma(x)$. We consider the condition of this basis on the interval of orthogonality, $[a, b]$. Since the coefficients u_k in (3.1) of any polynomial $p \in \mathbb{P}_{n-1}$ are now representable as Fourier coefficients of p, it is easy to estimate the condition of M_n with the aid of Schwarz' inequality. One finds (Gautschi [12])

$$\operatorname{cond} M_n \leqslant \max_{1\leqslant k\leqslant n}\left(\frac{\mu_0}{h_{k-1}}\right)^{1/2} \max_{a\leqslant x\leqslant b}\sum_{k=1}^{n}|\pi_{k-1}(x)|, \tag{3.13}$$

where

$$\mu_0 = \int_a^b d\sigma(x), \qquad h_k = \int_a^b \pi_k^2(x)\,d\sigma(x), \qquad k = 0,1,\ldots. \tag{3.14}$$

The first maximum in (3.13) is a bound for $\|M_n^{-1}\|$, the second an obvious bound for $\|M_n\|$. It should be noted that neither cond M_n nor the bound in (3.13) is invariant under different normalizations of the orthogonal polynomials $\{\pi_{k-1}\}$, and the bound indeed is minimized in the case of an orthonormal system.

It follows from (3.13) that the condition of an orthogonal basis, typically, exhibits only polynomial growth in n. For Chebyshev

polynomials $\pi_r = T_r$ on $[-1,1]$, for example, one finds

$$\text{cond } M_n \leqslant 2^{1/2}n \qquad (\pi_r = T_r),$$

while for Legendre polynomials $\pi_r = P_r$ on $[-1,1]$,

$$\text{cond } M_n \leqslant n(2n-1)^{1/2} \qquad (\pi_r = P_r).$$

The improvement over the power basis is substantial.

3.3. Lagrangian bases. All bases $\{p_k\}$ considered previously have the property that $\deg p_k = k-1$, $k = 1,2,3,\ldots$. We now consider an example of a basis in which each p_k is a polynomial of degree $n-1$, namely the familiar Lagrange polynomials

$$p_k(x) = l_k(x), \qquad l_k(x) = \prod_{\substack{\nu=1 \\ \nu \neq k}}^{n} \frac{x - s_\nu}{s_k - s_\nu}, \qquad k = 1,2,\ldots,n,$$

corresponding to a set of distinct nodes $s_1, s_2, \ldots, s_n$ in $[a,b]$. Lagrange's interpolation formula

$$p(x) = \sum_{k=1}^{n} p(s_k) l_k(x)$$

shows immediately that $u_k = p(s_k)$ in (3.1). By standard arguments in approximation theory, one finds $\|M_n\| = L_n$, $\|M_n^{-1}\| = 1$, where

$$L_n = \max_{a \leqslant x \leqslant b} \sum_{k=1}^{n} |l_k(x)|$$

is the Lebesgue constant for the nodes s_ν. Consequently (see also de Boor [5], p. 19),

$$\text{cond } M_n = L_n. \tag{3.15}$$

By a result of Faber and Bernstein, Natanson ([24], p. 24), one has

$$L_n > \frac{\ln n}{8\sqrt{\pi}}$$

for arbitrary (distinct) nodes s_ν, while for Chebyshev nodes, on the other hand,

$$L_n \sim \frac{2}{\pi} \ln n, \qquad n \to \infty$$

(see, e.g., Rivlin [28], p. 18). The basis consisting of the Lagrange polynomials $\{l_k\}$ for Chebyshev nodes, therefore, is optimally conditioned among all Lagrangian bases, and indeed among all polynomial bases (de Boor [4]), in the sense of attaining the optimal growth rate $O(\ln n)$.

4. THE CONDITION OF ALGEBRAIC EQUATIONS

We now turn our attention to roots of algebraic equations and their sensitivity to small changes in the coefficients. (We assume that the equation is expressed linearly in terms of basis polynomials.) An interesting, though largely unexplored, aspect of this question is the manner in which this sensitivity depends on the choice of polynomial basis. By far best understood is the case of equations expressed in the usual power form.

In order to give a formal statement of the problem, we assume, first of all, that the basis polynomials p_k have $\deg p_k = k-1$, $k = 1, 2, \ldots$, so that an algebraic equation of exact degree n can be written in normalized form

$$p(x) = 0, \qquad p(x) = p_{n+1}(x) + \sum_{k=1}^{n} u_k p_k(x), \qquad (4.1)$$

with leading coefficient 1. (To enhance clarity, we sometimes write $p(u; x)$ instead of $p(x)$, where $u = [u_1, u_2, \ldots, u_n]^T$.)

In general, one might be interested in just one, or in several, or collectively in all the roots of the equation, and again, there may be one single coefficient, or several, or all of them that are subject to perturbation. We treat all these cases in one, by considering q (simple) roots $\hat{\xi} = [\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_q]^T$, $1 \leqslant q \leqslant n$, of (4.1), corresponding to $u = \hat{u}$, and by introducing a multi-index $\mathbf{k} = (k_1, k_2, \ldots, k_p)$,

$1 \leqslant k_1 < k_2 < \cdots < k_p \leqslant n$, to indicate which of the coefficients in $\mathring{u}$ are to undergo changes. We write $\mathbf{k}^c$ for the multi-index complementary to $\mathbf{k}$, and denote by $\mathbf{u} \in \mathbb{R}^n$ the vector whose kth component is $\mathring{u}_k$, if $k \in \mathbf{k}^c$, and u_k, if $k \in \mathbf{k}$. There will be a neighborhood $\mathscr{D} = N(\mathring{u}_\mathbf{k}) \subset \mathbb{R}^p$ such that the equation (4.1) with $u = \mathbf{u}$, $u_\mathbf{k} \in \mathscr{D}$, continues to have q simple zeros $\xi = [\xi_1, \xi_2, \ldots, \xi_q]^T$ and $\xi \to \mathring{\xi}$ as $\mathbf{u} \to \mathring{u}$. We assume that neither $\mathring{\xi}$, nor $\mathring{u}_\mathbf{k}$, is the zero vector in the space $\mathbb{C}^q$ and $\mathbb{R}^p$, respectively. (Clearly, $\mathring{\xi} \neq \mathbf{0}$ if $q > 1$, since each $\mathring{\xi}_j$ is simple.) Our interest, then, is in the condition of the map $M_{\mathbf{k},q}: \mathscr{D} \subset \mathbb{R}^p \to \mathbb{C}^q$ defined by

$$M_{\mathbf{k},q}: \xi = \mathbf{f}(u_\mathbf{k}), \qquad u_\mathbf{k} \in \mathscr{D} \subset \mathbb{R}^p,$$

where $p(\mathbf{u}; f_j(u_\mathbf{k})) \equiv 0$ on $\mathscr{D}$, for each $j = 1, 2, \ldots, q$, and $\mathbf{f}(u_\mathbf{k}) \to \mathring{\xi}$ as $u_\mathbf{k} \to \mathring{u}_\mathbf{k}$.

It is now a straightforward matter to use (2.2) to calculate the condition number of the map $M_{\mathbf{k},q}$ at $\mathring{u}_\mathbf{k}$. If we denote by

$$V_{\mathbf{k},q}(\xi) = \begin{bmatrix} p_{k_1}(\xi_1) & p_{k_1}(\xi_2) & \cdots & p_{k_1}(\xi_q) \\ p_{k_2}(\xi_1) & p_{k_2}(\xi_2) & \cdots & p_{k_2}(\xi_q) \\ \cdots & \cdots & \cdots & \cdots \\ p_{k_p}(\xi_1) & p_{k_p}(\xi_2) & \cdots & p_{k_p}(\xi_q) \end{bmatrix} \in \mathbb{C}^{p \times q}$$

the "generalized Vandermonde matrix", and by $D(\mathbf{u}; \xi)$ the diagonal matrix

$$D(\mathbf{u}; \xi) = \operatorname{diag}\big[\, p'(\mathbf{u}; \xi_1), p'(\mathbf{u}; \xi_2), \ldots, p'(\mathbf{u}; \xi_q)\big] \in \mathbb{C}^{q \times q}$$

(where the prime in $p'(\mathbf{u}; x)$ indicates differentiation with respect to x), we find that

$$\operatorname{cond}\big(M_{\mathbf{k},q}; \mathring{u}_\mathbf{k}\big) = \frac{\|\mathring{u}_\mathbf{k}\|}{\|\mathring{\xi}\|} \big\| D^{-1}(\mathring{u}; \mathring{\xi}) V_{\mathbf{k},q}^T(\mathring{\xi}) \big\| \tag{4.2}$$

Specifically, if $\mathbf{k} = (1, 2, \ldots, n)$ and $q = n$, which is the extreme case of all roots being considered simultaneously and all coefficients undergoing changes, and if we choose the l_1-vector norm and

subordinate matrix norm, we get

$$\operatorname{cond}_1(M_{\mathbf{k},n};\mathring{u}) = \frac{\sum_{k=1}^{n}|\mathring{u}_k|}{\sum_{j=1}^{n}|\mathring{\xi}_j|}\cdot\max_{1\leqslant k\leqslant n}\sum_{j=1}^{n}\left|\frac{p_k(\mathring{\xi}_j)}{p'(\mathring{\xi}_j)}\right|, \tag{4.2a}$$

where $p'(\mathring{\xi}_j) = p'(\mathring{u};\mathring{\xi}_j)$. This provides the most overall description of the condition of the algebraic equation (4.1), assuming all roots are simple. The other extreme is $p = q = 1$, in which case we write $\mathbf{k} = k$, $\xi_1 = \xi$, and we find

$$\operatorname{cond}(M_{k,1};\mathring{u}) = \left|\frac{\mathring{u}_k p_k(\mathring{\xi})}{\mathring{\xi}p'(\mathring{\xi})}\right|, \qquad k = 1,2,\ldots,n. \tag{4.2b}$$

Each condition number in (4.2b) measures the sensitivity of the root $\mathring{\xi}$ to perturbations in one single (nonvanishing) coefficient, $\mathring{u}_k$, and provides the most detailed description of the condition of the root $\mathring{\xi}$. Note, in (4.2b), that only $\mathring{\xi}$ is assumed to be simple; some or all of the other roots may well be multiple.

A compromise between (4.2a) and (4.2b) for characterizing the condition of a single root, $\mathring{\xi}$, is $\operatorname{cond}\mathring{\xi} = \sum_{k=1}^{n}\operatorname{cond}(M_{k,1};\mathring{u})$, that is (we drop superscripts from now on),

$$\operatorname{cond}\xi = \frac{1}{|\xi p'(\xi)|}\sum_{k=1}^{n}|u_k p_k(\xi)|. \tag{4.3}$$

In the following, we adopt (4.3) as the condition number of the root ξ of (4.1). (Alternatively, we could use (4.2) with $q = 1$ and $\mathbf{k} = (1,2,\ldots,n)$.)

4.1. Equations in power form. Here, $p_k(x) = x^{k-1}$, and (4.3) assumes the form

$$\operatorname{cond}\xi = \frac{1}{|\xi p'(\xi)|}\sum_{k=1}^{n}|u_k\xi^{k-1}|. \tag{4.4}$$

The condition number (4.4) is easily seen to be invariant under scaling of the independent variable by an arbitrary complex number. Denoting the zeros of $p(x)$ by $\xi_1, \xi_2, \ldots, \xi_n$, additional insight may be provided by estimating the condition of one of these, ξ_μ, in terms of all of them. A result in this vein is the inequality (Gautschi [13])

$$\operatorname{cond} \xi_\mu \leqslant \frac{2 \prod_{\substack{\nu=1 \\ \nu \neq \mu}}^{n} \left(1 + \left|\frac{\xi_\nu}{\xi_\mu}\right|\right) - 1}{\prod_{\substack{\nu=1 \\ \nu \neq \mu}}^{n} \left|1 - \frac{\xi_\nu}{\xi_\mu}\right|}, \tag{4.5}$$

in which equality holds precisely when all zeros ξ_ν are located on a half-ray through the origin. A similar inequality, resp. equality, holds if the zeros are pairwise symmetric with respect to the origin. Note that the bound in (4.5), like the condition number itself, is invariant with respect to scaling.

We illustrate (4.5) by several examples, beginning with the well-known example due to Wilkinson of a severely ill-conditioned equation.

EXAMPLE 4.1 (Wilkinson [38], p. 41ff): $\xi_\nu = \nu$, $\nu = 1, 2, \ldots, n$.

This is a root configuration for which (4.5) holds with equality sign. There follows, by a simple computation,

$$\operatorname{cond} \xi_\mu = \frac{(\mu+n)! - \mu^n \mu!}{\mu!^2 (n-\mu)!}, \qquad \mu = 1, 2, \ldots, n.$$

An asymptotic analysis for large n will show (Gautschi [13]) that the worst conditioned root is the one near $n/\sqrt{2} = .7071\ldots n$. (For $n = 20$, the case considered by Wilkinson, the distinction goes to $\xi_{14} = 14$.) Its condition number grows exponentially,

$$\max_{1 \leqslant \mu \leqslant n} \operatorname{cond} \xi_\mu \sim \frac{1}{\pi(2-\sqrt{2})n} \left(\frac{\sqrt{2}+1}{\sqrt{2}-1}\right)^n, \qquad n \to \infty. \tag{4.6}$$

The best conditioned root is $\xi_1 = 1$, with a condition number that grows very slowly,

$$\operatorname{cond} \xi_1 \sim n^2, \qquad n \to \infty. \tag{4.7}$$

It is instructive to observe what happens if one of the coefficients u_k in

$$\prod_{\nu=1}^{n} (x-\nu) = x^n + u_n x^{n-1} + \cdots + u_1,$$

say the coefficient u_{k_0}, $k_0 = [(n+3)/2]$, is continuously perturbed,

$$u_{k_0}(t) = (1+t)u_{k_0}, \qquad 0 \leqslant t \leqslant \varepsilon,$$

all other coefficients being held constant. The resulting motion of the roots† is shown in Figure 4.1 for $n = 10$, $\varepsilon = 8 \times 10^{-7}$, and in Figure 4.2 for the ten "most active" roots in the case $n = 20$, $\varepsilon = 7 \times 10^{-14}$(!). Initially, of course, the zeros are all confined to move along the real axis. Not before long, however, a number of them will collide, each time branching off into pairs of conjugate complex roots. When $n = 10$, there are three collisions within $0 \leqslant t \leqslant \varepsilon$, occurring at $t_1 = 1.02567 \times 10^{-7}$, $t_2 = 1.53420 \times 10^{-7}$, and $t_3 = 7.21568 \times 10^{-7}$, and one further collision (not shown in Figure 4.1) at $t_4 = 1.17328 \times 10^{-4}$. These are the only collisions in $0 \leqslant t \leqslant 1$, as far as we could determine. For $n = 20$, there are five collisions within $0 \leqslant t \leqslant \varepsilon$, at approximately $t_1 = 4.0 \times 10^{-15}$, $t_2 = 5.4 \times 10^{-15}$, $t_3 = 8.9 \times 10^{-15}$, $t_4 = 2.75 \times 10^{-14}$, $t_5 = 6.25 \times 10^{-14}$, and several more later on (e.g., at $t_6 = 1.01 \times 10^{-12}$, $t_7 = 1.67 \times 10^{-12}$).

The behavior in Figures 4.1 and 4.2 may be viewed as an elementary example of a bifurcation phenomenon (catastrophe theory, if you will), the special feature here being the almost infinitesimal time scale on which the phenomenon takes place.

† The graphs were obtained by numerical integration of the differential equations satisfied by $\xi_\nu(t)$, $\nu = 1, 2, \ldots, n$. The exact instances of collision were determined by finding the t-zeros of the resultant of p and p'. The graphs after each collision represent the absolute values of the conjugate complex roots produced by the collision.

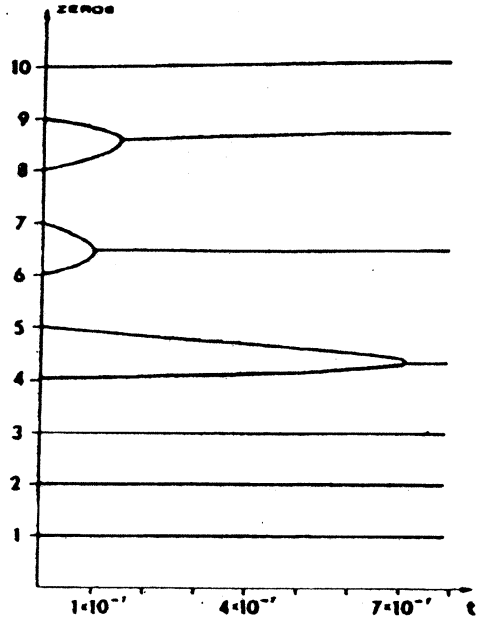


FIG. 4.1. Root paths for Example 4.1, $n = 10$.

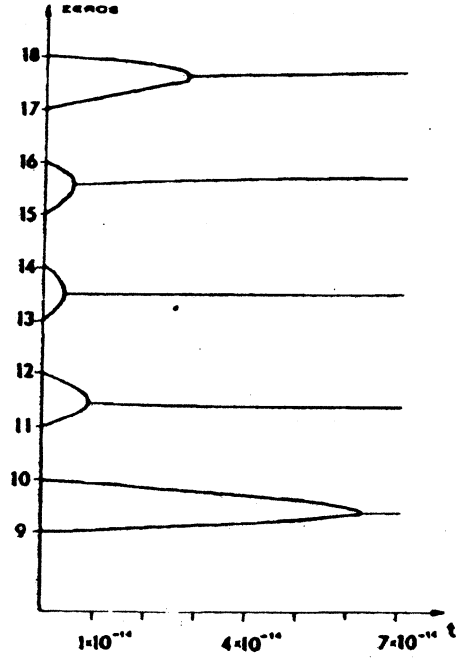


FIG. 4.2. Root paths for Example 4.1, $n = 20$.

EXAMPLE 4.2 (Wilkinson [38], p. 44ff): $\xi_\nu = 2^{-\nu}$, $\nu = 1, 2, \ldots, n$.

The roots ξ_ν accumulate rapidly near the origin, which at first might suggest that they become more and more ill-conditioned. In reality, however, they are all quite well-conditioned. This can be seen from the inequality

$$\operatorname{cond} \xi_\mu < 2 \prod_{\substack{\nu=1 \\ \nu \neq \mu}}^{n} \frac{1 + \left| \frac{\xi_\nu}{\xi_\mu} \right|}{\left| 1 - \frac{\xi_\nu}{\xi_\mu} \right|}, \tag{4.8}$$

which follows at once from (4.5), and which in the case at hand yields

$$\operatorname{cond} \xi_\mu < 2 \prod_{\nu=1}^{\infty} \left(\frac{1+2^{-\nu}}{1-2^{-\nu}} \right)^2 = 136.32\ldots. \tag{4.9}$$

The condition is thus bounded by a relatively small number (as condition numbers go), uniformly in n.

Since the bound in (4.8) is invariant with respect to reciprocation, the same result holds for the roots $\xi_\nu = 2^\nu$, $\nu = 1, 2, \ldots, n$.

EXAMPLE 4.3 (Roots of unity): $\xi_\nu = e^{2\pi i \nu / n}$, $\nu = 1, 2, \ldots, n$.

Since $p(x) = x^n - 1$, Equation (4.4) gives at once

$$\operatorname{cond} \xi_\mu = \frac{1}{n}, \qquad \mu = 1, 2, \ldots, n. \tag{4.10}$$

All roots are equally well-conditioned, the condition in fact getting better with increasing degree! The example, of course, is quite trivial, and (4.10) is just another way of saying that $(1+\varepsilon)^{1/n} - 1 \sim \varepsilon/n$ as $\varepsilon \to 0$.

4.2. Equations expressed in terms of orthogonal polynomials. We now assume that the equation is written in the form

$$p(x)=0, \qquad p(x)=\pi_n(x)+\sum_{k=1}^{n} u_k\pi_{k-1}(x), \tag{4.11}$$

where $\{\pi_r\}$ is a set of orthogonal polynomials. It is to be noted that the normalization in (4.11) is such that $p(x)$ and $\pi_n(x)$ have the same leading coefficient, if expressed in powers of x. The condition number of any (simple) root ξ of (4.11) is

$$\operatorname{cond}\xi=\frac{1}{|\xi p'(\xi)|}\sum_{k=1}^{n}|u_k\pi_{k-1}(\xi)|. \tag{4.12}$$

It is easily seen that $\operatorname{cond}\xi$ does not depend on the particular way the orthogonal polynomials π_r are normalized. Note, however, again, that changing the normalization of π_n also changes p, according to the normalization of the equation adopted in (4.11).

An easy lower bound can be had by noting that

$$\sum_{k=1}^{n}|u_k\pi_{k-1}(\xi)|\geq\left|\sum_{k=1}^{n}u_k\pi_{k-1}(\xi)\right|=|\pi_n(\xi)|.$$

Thus,

$$\operatorname{cond}\xi\geq\left|\frac{\pi_n(\xi)}{\xi p'(\xi)}\right| \tag{4.13}$$

For an upper bound we could apply Schwarz' inequality to the sum in (4.12), but the result is not particularly revealing. It appears difficult, indeed, to extract from (4.12) much detailed information concerning the qualitative behavior of the condition of ξ. We may note, however, that there are three factors which influence its magnitude: (i) the magnitude of the Fourier coefficients u_k of p; (ii) the magnitude of the orthogonal polynomials π_k evaluated at the root ξ; (iii) the magnitude of $\xi p'(\xi)$. Since orthogonal polynomials grow rapidly outside their interval of orthogonality, it

seems imperative, in view of (i) and (ii), that the interval of orthogonality be selected so as to contain ξ, if ξ is real.

It is quite possible that equations that are ill-conditioned in power form become well-conditioned when expanded in orthogonal polynomials, and vice versa. We can see this already by reexamining the examples discussed previously. We begin with Wilkinson's example, whose roots we now scale to be enclosed in the interval $[0,1]$. As we have noted earlier, such a scaling does not affect the condition of the roots, if the equation is in power form.

EXAMPLE 4.1′: $\xi_\nu = \nu/n$, $\nu = 1,2,\ldots,n$.

The condition number (4.12) can be computed for various (classical) polynomials $\{\pi_r\}$ orthogonal on $[0,1]$. It turns out that the Chebyshev polynomials of the second kind perform best (in the sense of making $\max_\mu \operatorname{cond} \xi_\mu$ smallest). Some numerical results are shown in the second column of Table 4.1. They are contrasted in the third column with the analogous condition numbers for the equation in power form (see Example 4.1). The improvement is clearly significant. We remark, nevertheless, that the condition still grows exponentially with n (though at a moderate rate), as can be deduced from the inequality (4.13); if n is even, e.g., one finds

$$\max_\mu \operatorname{cond} \xi_\mu \geqslant \frac{(n/4)^n}{(n/2)!^2} \sim \frac{1}{\pi n}(e/2)^n, \qquad n \to \infty.$$

TABLE 4.1

n	$\max_\mu \operatorname{cond} \xi_\mu$	
	Example 4.1′	Example 4.1
5	1.85	5.87×10^2
10	2.64×10^1	2.32×10^6
15	6.35×20^2	1.05×10^{10}
20	1.40×10^4	5.40×10^{13}

The condition of the roots in Examples 4.1′ and 4.1.

EXAMPLE 4.2′: $\xi_\nu = 2\cdot 2^{-\nu}$, $\nu = 1,2,\ldots,n$.

All orthogonal (on $[0,1]$) polynomials $\{\pi_r\}$ tried on this example led to condition numbers that grow extremely rapidly with n. For the "best" of these, the Chebyshev polynomials of the first kind, the results are shown in the second column of Table 4.2. The third column again contains the condition numbers of the same roots for the equation in power form. The contrast is striking!

It is not difficult to identify the culprit in Example 4.2′: it is the derivative of p at ξ_ν, which becomes extremely small as ν approaches n. Indeed, if p is normalized to have leading coefficient one, $p(x) = x^n + \cdots$, one finds for $\nu = n$ that

$$|\xi_n p'(\xi_n)| \sim \frac{.28879}{2^{n(n-1)/2}}, \qquad n \to \infty.$$

The Fourier coefficients u_k in (4.12), although reasonably small (of order 10^{-3}), are no match for this kind of decay! In the case of the power basis, the small denominator $|\xi_n p'(\xi_n)|$ in (4.4) is neutralized by an equally small numerator,

$$\sum_{k=1}^{n} |u_k \xi_n^{k-1}| \sim \frac{2.3842}{2^{(n+1)(n-2)/2}}, \qquad n \to \infty.$$

EXAMPLE 4.3′: $\xi_\nu = e^{2\pi i\nu/n}$, $\nu = 1,2,\ldots,n$.

TABLE 4.2

n	$\max_\mu \operatorname{cond} \xi_\mu$	
	Example 4.2′	Example 4.2
5	5.03×10^{1}	4.91×10^{1}
10	1.44×10^{12}	1.13×10^{2}
15	1.19×10^{30}	1.32×10^{2}
20	3.58×10^{55}	1.36×10^{2}

The condition of the roots in Examples 4.2′ and 4.2.

The roots of unity, extremely well-conditioned in the power basis (Example 4.3), continue to be quite well-conditioned in orthogonal bases, provided the interval of orthogonality is reasonably chosen. The most natural choice is $[-1,1]$, and all (classical) polynomials orthogonal on this interval do quite well, yielding condition numbers ranging from about .5 for $n=5$ to about 35 for $n=20$.

Just to show how an unreasonable choice of orthogonality interval may turn even the roots of unity into poorly conditioned roots, consider the case of Laguerre polynomials (orthogonal on $(0,\infty)$). Here,

$$\frac{(-1)^n}{n!}(x^n-1)=(-1)^n\left(1-\frac{1}{n!}\right)+\sum_{r=1}^{n}(-1)^{n-r}\binom{n}{r}L_r(x), \tag{4.14}$$

as can be derived from an integral formula for Laguerre polynomials (see Buchholz [7], p. 120, Eq. (4β)). Therefore, from (4.12), one gets

$$\operatorname{cond}\xi_\mu=(n-1)!\left(1-\frac{1}{n!}+\sum_{r=1}^{n-1}\binom{n}{r}\left|L_r(\xi_\mu)\right|\right).$$

Some numerical values are shown in Table 4.3; they speak for themselves!

TABLE 4.3

n	$\max_\mu \operatorname{cond}\xi_\mu$
5	3.17×10^{3}
10	5.45×10^{9}
20	9.71×10^{24}
40	1.92×10^{61}

The condition of the roots of unity in the Laguerre polynomial basis.

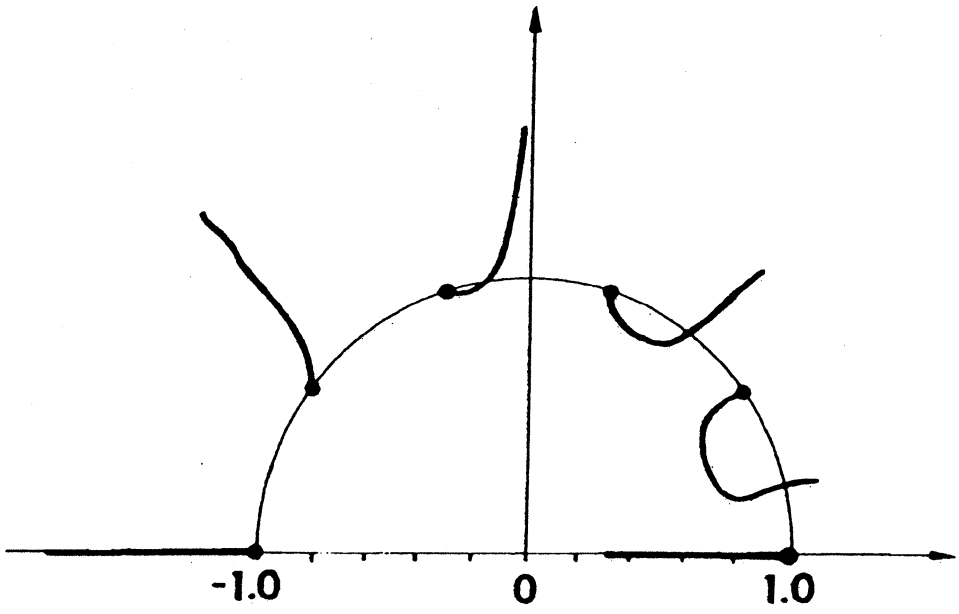


FIG. 4.3. Root paths for Example 4.3′, $n = 10$.

The case $n = 10$ is further illustrated in Figure 4.3, which shows the motion of the roots ξ_μ induced by a multiplication of the single coefficient

$$(-1)^{n-r_0}\binom{n}{r_0}, \qquad r_0 = \left[\frac{n+1}{2}\right],$$

in (4.14) by $1+t$ and variation of t from 0 to 10^{-8}.

5. GENERATION OF ORTHOGONAL POLYNOMIALS.

Generating orthogonal polynomials is fairly straightforward once the three-term recurrence relation, which they are known to satisfy, is explicitly available. Such is the case for all classical orthogonal polynomials. We are interested here in the more difficult task of generating the recurrence relation in cases where it is not explicitly known. A related problem is the construction of the Gaussian quadrature formulae; we use this connection to discuss the condition of the problem.

5.1. Statement of the problem. Let the desired (monic) polynomials $\{\pi_r\}$ be orthogonal with respect to a nonnegative measure $d\sigma(x)$ on the real line $\mathbb{R}$, where $\sigma(x)$ has at least $n+1$ points of increase. We assume that the first $2n$ moments of $d\sigma$ exist,

$$\mu_k = \int_{\mathbb{R}} x^k \, d\sigma(x), \qquad k = 0,1,2,\ldots,2n-1. \tag{5.1}$$

It is well known that the orthogonal polynomials π_r then exist for $r = 0,1,\ldots,n$, and satisfy a recurrence relation of the form

$$\pi_{-1}(x) = 0, \qquad \pi_0(x) = 1, \tag{5.2}$$

$$\pi_{k+1}(x) = (x-\alpha_k)\pi_k(x) - \beta_k \pi_{k-1}(x), \qquad k = 0,1,2,\ldots,n-1,$$

with real α_k and $\beta_k > 0$. (β_0 is arbitrary, but is conveniently defined as $\beta_0 = \int_{\mathbb{R}} d\sigma(x)$.) Associated with (5.2) is the symmetric tridiagonal matrix of order n,

$$J_n = \begin{bmatrix} \alpha_0 & \sqrt{\beta_1} & & & 0 \\ \sqrt{\beta_1} & \alpha_1 & \sqrt{\beta_2} & & \\ & \ddots & \ddots & \ddots & \\ & & & & \sqrt{\beta_{n-1}} \\ 0 & & & \sqrt{\beta_{n-1}} & \alpha_{n-1} \end{bmatrix}, \tag{5.3}$$

called *Jacobi matrix*, whose rth leading principal submatrix has the characteristic polynomial π_r, $r = 1,2,\ldots,n$. Let $\xi_1, \xi_2, \ldots, \xi_n$ be the zeros of π_n, hence the eigenvalues of J_n (they are all real and simple, as is well known). Let $v_j = [v_{1j}, v_{2j}, \ldots, v_{nj}]^T$ be the normalized eigenvector of J_n belonging to the eigenvalue ξ_j,

$$J_n v_j = \xi_j v_j, \qquad v_j^T v_j = 1.$$

Then (Wilf [37], Ch. 2, Exercise 9, Golub & Welsch [20])

$$\int_{\mathbb{R}} f(x)\, d\sigma(x) = \sum_{j=1}^{n} \lambda_j f(\xi_j) + R_n(f),$$
$$\lambda_j = \mu_0 v_{1j}^2, \qquad j = 1,2,\ldots,n, \tag{5.4}$$

is the *Gaussian quadrature formula* associated with the measure $d\sigma$, i.e., $R_n(f)=0$ for each $f\in\mathbb{P}_{2n-1}$.

There are classical procedures for generating the Jacobi matrix J_n, hence the Gaussian quadrature formula (5.4), from the given moments μ_k in (5.1). Unfortunately, as will be seen in the next subsection, the underlying map is likely to be severely ill-conditioned. The moments μ_k, indeed, are a poor way of "codifying" the measure $d\sigma$. An alternative way is through *modified moments*,

$$m_k=\int_{\mathbb{R}} p_k(x)\,d\sigma(x), \qquad k=0,1,2,\ldots,2n-1, \tag{5.5}$$

where $\{p_r\}$ is a suitably selected set of polynomials (see Sack & Donovan [29]). We shall assume that the p_r, like the π_r, satisfy a recurrence relation of the type

$$\begin{cases} p_{-1}(x)=0, \qquad p_0(x)=1, \\ p_{k+1}(x)=(x-a_k)p_k(x)-b_k p_{k-1}(x), \qquad k=0,1,2,\ldots,2n-2, \end{cases} \tag{5.6}$$

but now with coefficients a_k, b_k that are known. For example, $\{p_r\}$ may consist of a set of known (classical) orthogonal polynomials. If all a_k, b_k are zero, then $p_k(x)=x^k$, $k=0,1,2,\ldots$, and the modified moments reduce to ordinary moments, $m_k=\mu_k$, $k=0,1,2,\ldots$.

The problem we wish to consider is the following: Given the modified moments m_k in (5.5), determine the Jacobi matrix (5.3). In particular, this will also determine the orthogonal polynomials $\{\pi_r\}_{r=0}^n$, by virtue of (5.2), and the associated Gaussian quadrature formula, by virtue of (5.4).

The map in question thus is $K_n:\mathbb{R}^{2n}\to\mathbb{R}^{2n}$ which associates to the first $2n$ modified moments m_r the $2n$ recursion coefficients α_k,β_k, $k=0,1,\ldots,n-1$, for the respective orthogonal polynomials:

$$K_n: m\to\rho$$

$$m^T=[m_0,m_1,\ldots,m_{2n-1}], \qquad \rho^T=[\alpha_0,\ldots,\alpha_{n-1},\beta_0,\ldots,\beta_{n-1}]. \tag{5.7}$$

(Recall that $\beta_0 = \int_{\mathbb{R}} d\sigma(x) = m_0$.) For the purpose of analyzing the condition of the map K_n it is convenient to think of K_n as the composition of two maps,

$$K_n = H_n \circ G_n, \tag{5.8}$$

where $G_n : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ takes us from the modified moments m_r to the Gaussian quadrature rule,

$$G_n : m \to \gamma, \qquad \gamma^T = [\lambda_1, \ldots, \lambda_n, \xi_1, \ldots, \xi_n], \tag{5.9}$$

and $H_n : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ from the Gaussian quadrature rule to the recursion coefficients,

$$H_n : \gamma \to \rho.$$

The map H_n is usually (but not always) quite well-conditioned, as can be inferred from the discussion in Gautschi [16], Section 3.1. Here we look only at the map G_n which is by far the more critical one. We first demonstrate the ill-conditioned character of this map when $m = \mu$ are ordinary moments (5.1), and then see what can be gained from taking m to be modified moments with respect to a system of (classical) orthogonal polynomials.

5.2 Condition of the map G_n in the case of ordinary moments. Here $m = \mu$, or $a_k = b_k = 0$ in (5.6). For definiteness we assume $d\sigma(x)$ supported on the interval $[0,1]$ and normalized such that $\mu_0 = 1$. The Jacobian matrix of the map G_n in (5.9) is easily seen to be the inverse of a confluent Vandermonde matrix in the nodes ξ_j, multiplied (from the left) by the inverse of the diagonal matrix $\operatorname{diag}[1, \ldots, 1, \lambda_1, \ldots, \lambda_n]$. Using two-sided estimates of the (uniform) norm of inverses of confluent Vandermonde matrices, it is possible to prove (Gautschi [10]) that

$$\operatorname{cond}_\infty(G_n; \mu) > \frac{1}{2} \max_{1 \le j \le n} \left[\frac{\pi_n(-1)}{\pi_n'(\xi_j)} \right]^2, \tag{5.10}$$

where π_n is the (desired) orthogonal polynomial of degree n with

respect to the measure $d\sigma$. Since the point -1 is well outside the interval of orthogonality $[0,1]$, it is evident from (5.10) that the condition of G_n grows at least exponentially with n. An idea as to the numerical value of the growth rate can be had by considering the (representative) example $\pi_n = T_n^*$, the "shifted" Chebyshev polynomial of the first kind. In this case, (5.10) yields

$$\operatorname{cond}_\infty(G_n; \mu) > \frac{(3+\sqrt{8})^{2n}}{64n^2} \qquad (\pi_n = T_n^*). \tag{5.11}$$

The lower bound in (5.11) happens to grow at the same exponential rate as the (Turing-) condition number of the $n \times n$-Hilbert matrix!

5.3 Condition of the map G_n in the case of modified moments. We assume now that the vector m consists of modified moments

$$m^T = [m_0, m_1, \ldots, m_{2n-1}], \qquad m_k = \int_{\mathbf{R}} p_k(x)\sigma(x),$$

where $\{p_k\}$ is a system of (monic) polynomials orthogonal with respect to some measure $ds(x)$,

$$\int_{\mathbf{R}} p_k(x)\, p_l(x)\, ds(x) = 0, \qquad k \neq l.$$

The support of ds normally coincides with that of $d\sigma$, but need not do so necessarily. The analysis of the condition of the map $G_n: m \to \gamma$ in (5.9) is somewhat simplified if, instead of G_n, one considers the map

$$\tilde{G}_n: \tilde{m} \to \gamma, \qquad \gamma = [\lambda_1, \ldots, \lambda_n, \xi_1, \ldots, \xi_n], \tag{5.12}$$

where $\tilde{m}$ is the vector of *normalized* modified moments,

$$\tilde{m}_k = d_k^{-1/2} m_k, \qquad d_k = \int_{\mathbf{R}} p_k^2(x)\, ds(x),$$

$$k = 0, 1, \ldots, 2n-1. \tag{5.13}$$

The additional diagonal map, $D_n: m \to \tilde{m}$, of course, is quite harmless as far as the numerical condition is concerned, but makes the modified moments independent of the normalization of the polynomials p_k.

The condition of $\tilde{G}_n$ can be estimated rather realistically in terms of the fundamental Hermite interpolation polynomials associated with the Gaussian nodes $\xi_1, \ldots, \xi_n$. These are the polynomials of degree $2n-1$,

$$\begin{aligned} h_i(x) &= l_i^2(x)\left[1 - 2l_i'(\xi_i)(x-\xi_i)\right], \\ k_i(x) &= l_i^2(x)(x-\xi_i), \end{aligned} \qquad i = 1, 2, \ldots, n, \quad (5.14)$$

satisfying

$$\begin{aligned} h_i(\xi_j) &= \delta_{ij}, \qquad h_i'(\xi_j) = 0, \\ k_i(\xi_j) &= 0, \qquad k_i'(\xi_j) = \delta_{ij}, \end{aligned} \qquad i, j = 1, 2, \ldots, n. \quad (5.15)$$

Here, δ_{ij} is the Kronecker symbol and l_i in (5.14) are the fundamental Lagrange interpolation polynomials. Using the Euclidean norm $\|\cdot\|_2$, one can show (Gautschi [16], Section 3.3) that

$$\operatorname{cond}_2(\tilde{G}_n; \tilde{m}) \leqslant \frac{\|\tilde{m}\|_2}{\|\gamma\|_2} \left\{ \int_{\mathbf{R}} \sum_{i=1}^{n} \left(h_i^2(x) + \frac{1}{\lambda_i^2} k_i^2(x) \right) ds(x) \right\}^{1/2}. \quad (5.16)$$

It is interesting to note the interplay between the two measures $d\sigma$ and ds in this formula. Both the integrand and the vector γ depend solely on the target measure $d\sigma$. Integration, on the other hand, is with respect to the given measure ds, while the vector $\tilde{m}$ depends on both measures. To evaluate the integral in (5.16), once ξ_j and λ_j are known, one can use the $2n$-point Gaussian quadrature rule associated with ds, which produces the integral exactly (up to rounding errors). Since ds is usually one of the classical measures, the Gauss formula in question is readily available.

The critical quantity in (5.16) is the square root of the integral, which in fact represents the Frobenius norm of the Fréchet deriva-

tive of $\tilde{G}_n$; see Gautschi [16], Section 3.3. This quantity in turn depends critically on the behavior of the function

$$g_n(x) = \sum_{i=1}^{n} \left(h_i^2(x) + \frac{1}{\lambda_i^2} k_i^2(x) \right). \tag{5.17}$$

This is a nonnegative polynomial of degree $4n-2$ satisfying, in particular,

$$g_n(\xi_j) = 1, \qquad g_n'(\xi_j) = 0, \qquad j = 1,2,\ldots,n. \tag{5.18}$$

It is useful to classify the nodes ξ_j into *weak* nodes and *strong* nodes according as $g_n''(\xi_j) < 0$ or $g_n''(\xi_j) > 0$. Near a weak node, the function g_n is less than 1 and is likely (but not necessarily so; see Example 5.1) to remain below 1 between consecutive weak nodes. In contrast, g_n is larger than 1 near a strong node and must peak (possibly at very large values) on either side of it. We expect the map $\tilde{G}_n$ to be well-conditioned if the majority of the nodes are weak, and the strong nodes, if any, accumulate closely. This is often the case when the support of $d\sigma$ is a finite interval (see Examples 5.2 and 5.3). On the other hand, if there are nodes (strong or weak ones) which are separated by relatively large gaps, then there is a potential danger of g_n shooting up to considerable heights on the gaps, giving rise to ill-conditioning. This kind of predicament is likely to occur if the support of $d\sigma$ is an infinite interval or if it consists of separate intervals. In the former case, the gaps arise because of the relatively wide spacing of the absolutely larger nodes ξ_j, whereas in the latter case the gaps are the holes between the separate support intervals. Note that this all depends solely on the measure $d\sigma$. There are other factors depending on the measure ds which may also significantly influence the magnitude of $\operatorname{cond} \tilde{G}_n$; see, in particular, Examples 5.4 and 5.5.

A simple computation based on (5.14) and (5.17) shows that

$$\begin{aligned} \frac{1}{2} g_n''(\xi_j) &= 2l_j''(\xi_j) - 6\left[l_j'(\xi_j)\right]^2 + \lambda_j^{-2} \\ &= 2 \sum_{k \neq j} \sum_{l \neq j;\, l \neq k} \frac{1}{(\xi_j - \xi_k)(\xi_j - \xi_l)} \\ &\quad - 6\left(\sum_{k \neq j} \frac{1}{\xi_j - \xi_k} \right)^2 + \frac{1}{\lambda_j^2}. \end{aligned} \tag{5.19}$$

The node ξ_j is therefore strong or weak depending on whether the quantity on the right of (5.19) is positive or negative. Unfortunately, little is known concerning this matter, and a detailed analysis would seem to provide an interesting and rewarding area of research.

It is known, however, that all Chebyshev nodes $\xi_j = \cos \vartheta_j$, $\vartheta_j = (2j-1)\pi/(2n)$, $j = 1, 2, \ldots, n$ (corresponding to the Chebyshev measure $d\sigma(x) = (1-x^2)^{-\frac{1}{2}}\,dx$ on $[-1,1]$) are indeed weak; Gautschi [18]. For measures $d\sigma$ that behave similarly to the Chebyshev measure, one should expect, therefore, that most, if not all, nodes ξ_j are weak and hence that the map $\tilde{G}_n$ is quite well-conditioned; see Example 5.2.

5.4. An algorithm. A number of algorithms are known for carrying out the map K_n in (5.7) from the modified moments m_j to the recursion coefficients α_k, β_k. We describe a particularly simple one due, in the form given below, to Wheeler [34],† and in a different form, to Sack & Donovan [29]. The algorithm actually goes back to Chebyshev [8] who proposed it in the special case of ordinary moments ($a_k = b_k = 0$) and discrete measures $d\sigma$.

We introduce the "mixed moments"

$$\sigma_{kl} = \int_{\mathbf{R}} \pi_k(x)\, p_l(x)\, d\sigma(x), \qquad k, l \geqslant -1, \tag{5.20}$$

and note that by orthogonality, $\sigma_{kl} = 0$ for $k > l$, and

$$\int_{\mathbf{R}} \pi_k^2(x)\, d\sigma(x) = \int_{\mathbf{R}} \pi_k(x) x p_{k-1}(x)\, d\sigma(x) = \sigma_{kk}, \qquad k \geqslant 1.$$

The relation $\sigma_{k+1,k-1} = 0$, therefore, together with (5.2), yields immediately $\sigma_{kk} - \beta_k \sigma_{k-1,k-1} = 0$, hence

$$\beta_k = \frac{\sigma_{kk}}{\sigma_{k-1,k-1}}, \qquad k = 1, 2, 3, \ldots. \tag{5.21}$$

(Recall that β_0 is set equal to m_0.) Similarly, $\sigma_{k+1,k} = 0$ gives

$$\int_{\mathbf{R}} \pi_k(x) x p_k(x)\, d\sigma(x) - \alpha_k \sigma_{kk} - \beta_k \sigma_{k-1,k} = 0,$$

†Equation (3.4) in Wheeler [34] is misprinted; a_k, b_k should read a_l and b_l, respectively.

and using (5.6) in the form

$$xp_k(x) = p_{k+1}(x) + a_k p_k(x) + b_k p_{k-1}(x), \tag{5.22}$$

yields $\sigma_{k,k+1} + (a_k - \alpha_k)\sigma_{kk} - \beta_k \sigma_{k-1,k} = 0$, hence, together with (5.21),

$$\begin{cases} \alpha_0 = a_0 + \dfrac{\sigma_{01}}{\sigma_{00}}, \\ \alpha_k = a_k - \dfrac{\sigma_{k-1,k}}{\sigma_{k-1,k-1}} + \dfrac{\sigma_{k,k+1}}{\sigma_{kk}}, \qquad k = 1,2,3,\ldots. \end{cases} \tag{5.23}$$

The σ's in turn satisfy the recursion

$$\begin{aligned} \sigma_{kl} = \sigma_{k-1,l+1} &- (\alpha_{k-1} - a_l)\sigma_{k-1,l} \\ &- \beta_{k-1}\sigma_{k-2,l} + b_l \sigma_{k-1,l-1}, \end{aligned} \tag{5.24}$$

as follows from (5.2) and (5.22) (where k is replaced by l). To construct orthogonal polynomials π_r of degrees $r \leqslant n$, we thus have the following algorithm.

Initialization:

$$\begin{cases} \sigma_{-1,l} = 0, & l = 1,2,\ldots,2n-2, \\ \sigma_{0,l} = m_l, & l = 0,1,\ldots,2n-1, \\ \alpha_0 = a_0 + \dfrac{m_1}{m_0}, \\ \beta_0 = m_0. \end{cases}$$

Continuation: For $k = 1,2,\ldots,n-1$ (5.25)

$$\begin{cases} \sigma_{kl} = \sigma_{k-1,l+1} - (\alpha_{k-1} - a_l)\sigma_{k-1,l} - \beta_{k-1}\sigma_{k-2,l} \\ \qquad + b_l \sigma_{k-1,l-1}, \qquad l = k, k+1,\ldots,2n-k-1, \\ \alpha_k = a_k - \dfrac{\sigma_{k-1,k}}{\sigma_{k-1,k-1}} + \dfrac{\sigma_{k,k+1}}{\sigma_{kk}}, \\ \beta_k = \dfrac{\sigma_{kk}}{\sigma_{k-1,k-1}}. \end{cases}$$

The algorithm requires as input $\{m_l\}_{l=0}^{2n-1}$ and $\{a_k, b_k\}_{k=0}^{2n-2}$; it furnishes $\{\alpha_k, \beta_k\}_{k=0}^{n-1}$, hence the orthogonal polynomials $\{\pi_r\}_{r=0}^{n}$, and also, incidentally, the normalizing factors $\sigma_{kk} = \int_{\mathbf{R}} \pi_k^2(x)\, d\sigma(x)$, $k \leqslant n-1$. The number of multiplications and divisions required is $3n^2 - n - 1$, the number of additions, $4n^2 - 3n$; the algorithm thus involves $O(n^2)$ operations altogether.

The success of the algorithm (5.25), of course, depends on the ability to compute all required modified moments m_l accurately and reliably. Most frequently, these moments are obtained from recurrence relations, judiciously employed, as for example in the case of Chebyshev or Gegenbauer moments (Piessens & Branders [26], Branders [6], Luke [22], Lewanowicz [21]). Sometimes they can be computed directly in terms of special functions, or in integer form (Gautschi [11], Examples (ii), (iii), Wheeler & Blumstein [36], Blue [1], Gautschi [15], Gatteschi [9]). Still another possibility is to use a suitable discretization process (Gautschi [16], Section 2.5).

5.5 Examples. We begin with a measure of discrete type, already considered by Chebyshev [8].

EXAMPLE 5.1: $d\sigma(x) = (1/N)\sum_{k=0}^{N-1}\delta(x - k/N)$, where $\delta(\cdot)$ is the Dirac delta function.

The associated N orthogonal polynomials $\pi_0, \pi_1, \ldots, \pi_{N-1}$ are explicitly known. There is no need, therefore, to carry out maps such as G_n in (5.9). Nevertheless, we briefly consider the condition of this map in order to explain an interesting phenomenon observed previously in Gautschi [16], Example 4.1, namely the gradual worsening of the condition of $\tilde{G}_n$ as n approaches N. The underlying moments are those corresponding to the shifted (monic) Legendre polynomials $p_k(x) = (k!^2/(2k)!)P_k(2x-1)$, $0 \leqslant x \leqslant 1$.

Computation reveals that all zeros ξ_j of π_n, $n \leqslant N$, are weak nodes. Yet the condition of $\tilde{G}_n$ deteriorates significantly as n approaches N; see Table 5.1. The reason for this can be found in the following peculiar behavior of the function $g_n(x)$ of (5.17). In the central zone of the interval $[0,1]$, g_n wiggles rapidly, always

remaining $\leqslant 1$. In both end zones, however, g_n becomes $\geqslant 1$ and exhibits spikes of increasing magnitudes as one moves toward the endpoints. On the last interval $[\xi_n, 1]$, soon after g_n'' becomes positive, g_n increases rapidly to a global maximum at $x=1$. The behavior is illustrated in Table 5.1, where we show the approximate magnitude of the largest spike attained on $[\xi_{n-1}, \xi_n]$ (and of a similar spike on $[\xi_1, \xi_2]$), as well as $\max_{0 \leqslant x \leqslant 1} g_n(x) = g_n(1)$. The neighboring spikes to the left of $[\xi_{n-1}, \xi_n]$ (or to the right of $[\xi_1, \xi_2]$) are typically several orders of magnitude smaller. When n is relatively small, there may be no spikes at all, which is indicated in Table 5.1 by a dash. Also shown are the values of the estimate (5.16) of the condition of $\tilde{G}_n$. It can be seen that the main contribution to $\operatorname{cond} \tilde{G}_n$ comes from the two spikes of g_n on $[\xi_1, \xi_2]$ and $[\xi_{n-1}, \xi_n]$ and from the final upward surge of g_n on $[\xi_n, 1]$. The formation of these spikes and their increasing magnitudes as n approaches N is undoubtedly caused by the fact that the nodes ξ_j approach more and more a uniform distribution; they are exactly uniformly distributed when $n = N$. It is well known that interpolation polynomials (and those of Hermite are no exception!) are prone to violent oscillations in such cases.

EXAMPLE 5.2: $d\sigma(x) = [(1-k^2x^2)(1-x^2)]^{-1/2}\,dx$ on $[-1, 1]$, $0 < k < 1$.

This example also has been considered previously in Gautschi [16], Example 4.4, where the use of Chebyshev moments was found to work extremely well. Here we point out that all nodes ξ_j of the

TABLE 5.1

N	n	$\max_{[\xi_{n-1}, \xi_n]} g_n$	$g_n(1)$	$\operatorname{cond} \tilde{G}_n$	N	n	$\max_{[\xi_{n-1}, \xi_n]} g_n$	$g_n(1)$	$\operatorname{cond} \tilde{G}_n$
10	5	—	4×10^{2}	2.5×10^{0}	40	15	1.4×10^{0}	5×10^{5}	2.0×10^{1}
	10	8×10^{3}	3×10^{11}	6.3×10^{4}		25	9×10^{6}	3×10^{15}	1.3×10^{6}
20	5	—	3×10^{1}	7.9×10^{-1}		35	3×10^{22}	4×10^{32}	5.0×10^{14}
	10	—	1×10^{5}	1.9×10^{1}	80	10	—	5×10^{1}	4.9×10^{-1}
	15	6×10^{3}	3×10^{11}	3.0×10^{4}		20	1.2×10^{0}	2×10^{5}	6.5×10^{0}
	20	3×10^{14}	4×10^{23}	3.3×10^{10}		30	2×10^{3}	1×10^{11}	3.9×10^{3}
40	5	—	7×10^{0}	6.5×10^{-1}		40	1×10^{10}	2×10^{19}	4.8×10^{7}
	10	—	6×10^{2}	1.0×10^{0}		50	2×10^{20}	3×10^{30}	1.7×10^{13}

The behavior of $g_n(x)$ in Example 5.1 and the condition of $\tilde{G}_n$

n-point Gauss formula for $d\sigma$ appear to be weak nodes. This was verified numerically for various values of k^2 as close to 1 as $k^2 = .99$, and for values of n as large as $n = 80$. In all cases computed, moreover, the function g_n was found never to exceed 1 on $[-1, 1]$. No wonder, therefore, that the map $\tilde{G}_n$ is extremely well-conditioned!

EXAMPLE 5.3: $d\sigma(x) = x^\alpha \ln(1/x)\,dx$ on $[0,1]$, $\alpha > -1$.

Here, the modified moments with respect to the shifted Legendre polynomials $p_k(x) = (k!^2/(2k)!)P_k(2x-1)$ can be obtained explicitly. For example, if α is not an integer, then

$$\frac{(2l)!}{l!^2} m_l = \frac{1}{\alpha+1}\left\{\frac{1}{\alpha+1} + \sum_{k=1}^{l}\left(\frac{1}{\alpha+1+k} - \frac{1}{\alpha+1-k}\right)\right\} \times \prod_{k=1}^{l} \frac{\alpha+1-k}{\alpha+1+k}, \qquad l = 0,1,2,\ldots. \tag{5.26}$$

(Similar formulas hold for integral α; see Blue [1] for $\alpha = 0$, Gautschi [15] for $\alpha > 0$, and Gatteschi [9] for still more general cases.) The appropriate recursion coefficients for $\{p_r\}$ are

$$\begin{aligned} a_k &= \frac{1}{2}, && k = 0,1,2,\ldots, \\ b_k &= \frac{1}{4(4-k^{-2})}, && k = 1,2,3,\ldots. \end{aligned} \tag{5.27}$$

With the quantities in (5.26) and (5.27) as input, algorithm (5.25) now easily furnishes the recursion coefficients α_k, β_k, $0 \leqslant k \leqslant n-1$, for the orthogonal polynomials with respect to $d\sigma(x) = x^\alpha \ln(1/x)\,dx$. For $\alpha = -\frac{1}{2}$, and $n = 2,4,8,\ldots,80$, and single-precision computation on the CDC 6500 computer (approx. 14 decimal digit accuracy), the mean square errors $\varepsilon_n(\alpha,\beta) = (\sum_{k=0}^{n-1}[\varepsilon^2(\alpha_k) + \varepsilon^2(\beta_k)])^{1/2}$, where $\varepsilon(\alpha_k), \varepsilon(\beta_k)$ are the relative errors in the coefficients α_k, β_k, are shown in the left half of Table 5.2. The right half displays the analogous results for the power moments $\mu_l = (\alpha+1+l)^{-2}$, and $a_k = b_k = 0$, all k. In the first case, all coefficients are obtained close to machine precision, attesting not only to the extremely well-conditioned nature of the problem, but also to the

TABLE 5.2

	Legendre moments	power moments
n	$\varepsilon_n(\alpha,\beta)$	$\varepsilon_n(\alpha,\beta)$
2	9.22×10^{-14}	9.92×10^{-15}
4	2.42×10^{-13}	3.25×10^{-12}
8	6.53×10^{-13}	6.29×10^{-7}
12	1.09×10^{-12}	1.67×10^{-1}
20	1.29×10^{-12}	
40	1.98×10^{-12}	
80	5.03×10^{-12}	

Relative errors in the recursion coefficients α_k, β_k for Example 5.3.

stability of algorithm (5.25). In the second case, all accuracy is lost by the time n reaches 12, which confirms the severely ill-conditioned character of the problem of generating orthogonal polynomials from ordinary moments.

The underlying condition numbers, specifically the condition of the map H_n as computed in Gautschi [16], Equation (3.8)f and the estimates of $\operatorname{cond}\tilde{G}_n$ in (5.16) and of $\operatorname{cond}G_n$ in (5.10), are displayed in Table 5.3. Recall that Table 5.2 illustrates the accuracy of the map $K_n = H_n \circ G_n$; see (5.7). The algorithm (5.25) based on Legendre moments performs somewhat better than the condition numbers of H_n and $\tilde{G}_n$ in Table 5.3 would suggest. For power moments, the rapid loss of accuracy evidenced in Table 5.2 correlates very well (at least for $n \leqslant 12$) with the rapid growth of the condition number of G_n in Table 5.3.

n	$\operatorname{cond} H_n$	Legendre moments $\operatorname{cond}\tilde{G}_n \leqslant$	power moments $\operatorname{cond}G_n \geqslant$
2	2.37	5.58	5.42
4	5.31	1.89×10^{1}	4.27×10^{3}
8	1.52×10^{1}	6.55×10^{1}	1.69×10^{9}
12	2.43×10^{1}	1.29×10^{2}	1.06×10^{15}
20	4.56×10^{1}	2.79×10^{2}	7.25×10^{26}
40	1.09×10^{2}	7.03×10^{2}	7.92×10^{56}
80	2.54×10^{2}	1.66×10^{3}	3.55×10^{117}

TABLE 5.3
The condition of the maps H_n, $\tilde{G}_n$ and G_n for Example 5.3

The gradual (but slow) increase of $\operatorname{cond} \tilde{G}_n$ can be ascribed to a phenomenon similar to the one observed in Example 5.1, except that this time not all nodes ξ_j are weak, but only about the first two-thirds of them (when ordered increasingly). All remaining nodes are strong, giving rise to the development of spikes and final upward surges as in Example 5.1. The severity of these spikes, though, is much less here than shown in Table 5.1. The maximum peak is of the order of magnitude $1\times 10, 7\times 10^2, 2\times 10^4, 5\times 10^5$ for $n = 10, 20, 40, 80$, respectively, and the corresponding global maxima $g_n(1)$ have orders of magnitude $2\times 10^4, 1\times 10^6, 4\times 10^7$ and 1×10^9.

EXAMPLE 5.4: The half-range Hermite measure $d\sigma(x) = e^{-x^2}dx$ on $[0, \infty]$.

This example illustrates the potential ill-conditioning of the map $\tilde{G}_n$ in the case of measures supported on an infinite interval. Modified Hermite and Laguerre moments can be readily computed from the explicit power representations of the (monic) Hermite and Laguerre polynomials. One finds:

$$m_k = 2^{-k}\int_0^\infty e^{-x^2}H_k(x)\,dx$$

$$= \begin{cases} \frac{1}{2}\sqrt{\pi}, & k = 0, \\ \frac{1}{2}\sum_{r=0}^{\kappa}(-1)^r \prod_{i=r+1}^{\kappa} i \prod_{i=\kappa-r+1}^{\kappa}(i+\frac{1}{2}), & k = 2\kappa+1, \\ 0, & k = 2\kappa, \end{cases} \tag{5.28}$$

and

$$m_k = (-1)^k k!\int_0^\infty e^{-x^2}L_k(x)\,dx = \frac{(-1)^k}{2}\sum_{r=0}^{k}(-1)^r \frac{k!\Gamma\left(\frac{r+1}{2}\right)}{(k-r)!r!^2}. \tag{5.29}$$

As expected, the algorithm (5.25), in both cases, loses accuracy ather quickly, but perhaps unexpectedly, the loss is about twice as

TABLE 5.4

n	Hermite moments		Laguerre moments	
	cond $\tilde{G}_n$	$\varepsilon_n(\alpha,\beta)$	cond $\tilde{G}_n$	$\varepsilon_n(\alpha,\beta)$
2	1.29×10^{1}	4.06×10^{-14}	7.27×10^{1}	1.63×10^{-13}
4	2.11×10^{3}	3.28×10^{-12}	1.35×10^{6}	1.05×10^{-8}
6	5.66×10^{5}	7.75×10^{-10}	1.30×10^{11}	1.19×10^{-3}
8	1.86×10^{9}	3.39×10^{-6}	3.12×10^{16}	4.42×10^{2}
10	6.76×10^{10}	4.28×10^{-3}	1.42×10^{22}	—

The condition of $\tilde{G}_n$ in Example 5.4 for modified moments based on Hermite and Laguerre polynomials and mean square errors in the coefficients α_k, β_k, $k = 0,1,\ldots,n-1$.

large for Laguerre moments than for Hermite moments. An explanation is provided by the respective estimates (5.16) of cond $\tilde{G}_n$ shown in Table 5.4 together with the mean square (relative) error $\varepsilon_n(\alpha,\beta)$ of the α_k, β_k.†

The initial nodes ξ_j are again weak, as in the previous examples. All remaining nodes are strong and produce the now familiar peaking and surge phenomenon on $[0,\infty]$. On $[-\infty,0]$, g_n takes off to ∞. Even though the Hermite measure $ds(x) = e^{-x^2}dx$ has its support on $[-\infty,\infty]$ and, therefore, also contributes to cond $\tilde{G}_n$ through the values of $g_n(x)$ for $x<0$, the damping power of e^{-x^2} is much stronger for large $|x|$ than the damping power of e^{-x} for large $x>0$, which is the reason why the condition of $\tilde{G}_n$ turns out to be significantly smaller for Hermite moments than for Laguerre moments.

EXAMPLE 5.5:

$$d\sigma(x) = \begin{cases} \dfrac{1}{\pi}\dfrac{|x-\frac{1}{2}|}{\{x(1-x)(\frac{1}{3}-x)(\frac{2}{3}-x)\}^{1/2}}dx, & x\in(0,\tfrac{1}{3})\cup(\tfrac{2}{3},1), \\ 0 \quad \text{elsewhere.} \end{cases} \tag{5.30}$$

† The values of cond $\tilde{G}_n$ in the case of Hermite moments, as given in Gautschi [16], Table 4.8, are in error, being consistently somewhat too large. The error is due to an incorrect computation of $\|\tilde{m}\|_2$ using m_0 instead of $\tilde{m}_0 = \pi^{-1/4}m_0$.

This measure arises in the study of the diatomic linear chain (Wheeler [35]) and corresponds to the mass ratio $m/M = 1/2$, where m and M are the masses of the two kinds of particles alternating along the chain.

Wheeler [35] applies algorithm (5.25) to generate the associated orthogonal polynomials, using two choices of modified moments: Chebyshev moments, with $ds(x) = \pi^{-1}[x(1-x)]^{-1/2}dx$ on $[0,1]$, on the one hand, and modified moments with

$$ds(x) = \begin{cases} \dfrac{18}{\pi|x-1/2|}\left\{x(1-x)\left(x-\dfrac{1}{3}\right)\left(x-\dfrac{2}{3}\right)\right\}^{1/2}, & x \in (0,\tfrac{1}{3}) \cup (\tfrac{2}{3},1), \\ 0 \quad \text{elsewhere}, \end{cases} \tag{5.31}$$

on the other. He observes exponentially increasing instability (as n increases) in the first case, and perfect stability in the second. An explanation of this can be given on the basis of (5.16).

All zeros ξ_j of the orthogonal polynomial π_n associated with $d\sigma$, except possibly one, are known to congregate on the two support intervals $[0,1/3]$ and $[2/3,1]$ (Szegö [32], Theorem 3.41.2 and the sentence following it). In fact, by symmetry, the "hole" $[1/3,2/3]$ contains exactly one zero at $x = 1/2$, if n is odd, and none, if n is even. It was determined numerically that all zeros on the two support intervals are weak, and the one at $x = 1/2$ (if n is odd) is strong. The function g_n is wiggling on both support intervals, remaining $\leqslant 1$ there, and shoots up to a large single peak on the hole if n is even, and to a twin peak if n is odd. The peak values for $n = 5, 10, 20, 40$ are approximately $1.4, 6.5\times10^{2}, 2.4\times10^{8}, 1.1\times10^{20}$, respectively. In the case of Chebyshev moments, the integral $\int_0^1 g_n(x)\,ds(x)$ becomes large with increasing n, since the measure $ds(x)$ is supported on the entire interval $[0,1]$, including the hole $[1/3,2/3]$ where g_n is large. The condition of $\tilde{G}_n$ therefore gradually deteriorates; the condition numbers for $n = 5, 10, 20, 40$, in fact, are $7.3\times10^{-1}, 4.1, 1.5\times10^{3}, 6.2\times10^{8}$, respectively. In contrast, the measure $ds(x)$ in (5.31) is zero on the hole and thus gives rise to an

integral $\int_0^1 g_n(x)\,ds(x)$ which is bounded uniformly in n. Accordingly, the condition numbers remain quite small, namely 1.18, 1.07, 1.07, 1.07, respectively, for $n = 5, 10, 20, 40$.

The orthogonal polynomials relative to the measure $d\sigma$ in (5.30) have since been obtained explicitly (Gautschi [17]). Example 5.5, nevertheless, continues to be of interest, as it shows the importance of matching the supports of the two measures $d\sigma$ and ds in cases where $d\sigma$ is supported on separate intervals.

REFERENCES

1. J. L. Blue, "A Legendre polynomial integral," *Math. Comput.* 33 (1979), 739–741.
2. C. de Boor, "On calculating with B-splines," *J. Approximation Theory* 6 (1972), 50–62.
3. _____, "On local linear functionals which vanish at all B-splines but one," *Theory of Approximation with Applications* (Law, A. G. & Sahney, B. N., eds.), pp. 120–145, Academic Press, New York-San Francisco-London, 1976.
4. _____, personal communication, 1978.
5. _____, *A Practical Guide to Splines*, Springer-Verlag, New York-Heidelberg-Berlin, 1978.
6. M. Branders, "Application of Chebyshev polynomials in numerical integration" (Flemish), Thesis, Catholic University of Leuven, Belgium, 1976.
7. H. Buchholz, *Die konfluente hypergeometrische Funktion*, Springer-Verlag, Berlin-Göttingen-Heidelberg, 1953.
8. P. L. Chebyshev, "Sur l'interpolation par la méthode des moindres carrés," Mém. Acad. Impér. Sci. St. Pétersbourg (7) **1**, no. 15 (1859), 1–24. [Oeuvres I, 471–498]
9. L. Gatteschi, "On some orthogonal polynomial integrals," *Math. Comput.* **35** (1980), 1291–1298.
10. W. Gautschi, "Construction of Gauss-Christoffel quadrature formulas," *Math. Comput.* **22** (1968), 251–270.
11. _____, "On the construction of Gaussian quadrature rules from modified moments," *Math. Comput.* **24** (1970), 245–260.
12. _____, "The condition of orthogonal polynomials," *Math. Comput.*, **26** (1972), 923–924.
13. _____, "On the condition of algebraic equations," *Numer. Math.*, **21** (1973), 405–424.
14. _____, "The condition of polynomials in power form," *Math. Comput.*, **33** (1979), 343–352.
15. _____, "On the preceding paper, 'A Legendre polynomial integral' by J. L. Blue," *Math. Comput.*, **33** (1979), 742–743.
16. _____, "On generating orthogonal polynomials," *SIAM J. Sci. Statist. Comput.* **3** (1982), 289–317.

17. _____, "On some orthogonal polynomials of interest in theoretical chemistry," *BIT* **25** (1985), to appear.
18. _____, "On the sensitivity of orthogonal polynomials to perturbations in the moments," in preparation.
19. A. J. Geurts, "A contribution to the theory of condition," *Numer. Math.* **39** (1982), 85–96.
20. G. H. Golub and J. H. Welsch, "Calculation of Gauss quadrature rules," *Math. Comput.*, **23** (1969), 221–230.
21. S. Lewanowicz, "Construction of a recurrence relation for modified moments," *J. Comput. Appl. Math.*, **5** (1979), 193–206.
22. Y. L. Luke, *Algorithms for the Computation of Mathematical Functions*, Academic Press, New York-San Francisco-London, 1977.
23. T. Lyche, "A note on the condition numbers of the B-spline basis," *J. Approximation Theory*, **22** (1978), 202–205.
24. I. P. Natanson, *Constructive Function Theory*, Vol. III, Frederick Ungar Publ. Co., New York, 1965.
25. J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York-London, 1970.
26. R. Piessens and M. Branders, "The evaluation and application of some modified moments," *BIT*, **13** (1973), 443–450.
27. J. R. Rice, "A theory of condition," *SIAM J. Numer. Anal.*, **3** (1966), 287–310.
28. T. J. Rivlin, *The Chebyshev Polynomials*, John Wiley & Sons, London-Sydney-Toronto, 1974.
29. R. A. Sack, and A. F. Donovan, "An algorithm for Gaussian quadrature given modified moments," *Numer. Math.*, **18** (1971–72), 465–478.
30. A. Schönhage, *Approximationstheorie*, Walter de Gruyter & Co., Berlin-New York, 1971.
31. G. W. Stewart, *Introduction to Matrix Computations*, Academic Press, New York-London, 1973.
32. G. Szegö, *Orthogonal Polynomials*, AMS Colloquium Publications, Vol. 23, 4th ed., Providence, RI, 1975.
33. E. V. Voronovskaja, *The Functional Method and its Applications*, Translations of Mathematical Monographs, Vol. 28, American Mathematical Society, Providence, R.I., 1970.
34. J. C. Wheeler, "Modified moments and Gaussian quadratures," *Rocky Mountain J. Math.*, **4** (1974), 287–296.
35. _____, "Modified moments and continued fraction coefficients for the diatomic linear chain," *J. Chem. Phys.* **80** (1984), 472–476.
36. J. C. Wheeler and C. Blumstein, "Modified moments for harmonic solids," *Phys. Rev.*, B6 (1972), 4380–4382.
37. H. S. Wilf, *Mathematics for the Physical Sciences*, John Wiley & Sons, New York-London, 1962.
38. J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, Prentice-Hall, Englewood Cliffs, N.J., 1963.
39. _____, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

8.8. [66] “The Condition of Polynomials in Power Form”

---

[66] “The Condition of Polynomials in Power Form,” *Math. Comp.* **33**, 343–352 (1979).

MATHEMATICS OF COMPUTATION, VOLUME 33, NUMBER 145
JANUARY 1979, PAGES 343–352

The Condition of Polynomials in Power Form*

By Walter Gautschi

Abstract. A study is made of the numerical condition of the coordinate map M_n which associates to each polynomial of degree $\leqslant n-1$ on the compact interval $[a, b]$ the n-vector of its coefficients with respect to the power basis. It is shown that the condition number $\|M_n\|_\infty \|M_n^{-1}\|_\infty$ increases at an exponential rate if the interval $[a, b]$ is symmetric or on one side of the origin, the rate of growth being at least equal to $1+\sqrt{2}$. In the more difficult case of an asymmetric interval around the origin we obtain upper bounds for the condition number which also grow exponentially.

1. Introduction. Let $M_n: \mathbf{R}^n \longrightarrow \mathbf{P}_{n-1}$ be the linear map associating to each vector $u^T = [u_1, u_2, \ldots, u_n] \in \mathbf{R}^n$ the polynomial

$$p(x) = \sum_{k=1}^{n} u_k x^{k-1} \in \mathbf{P}_{n-1}, \qquad n \geqslant 2.$$

For any $p \in \mathbf{P}_{n-1}$ we shall write $u_p = M_n^{-1} p$, where M_n^{-1} is the inverse map of M_n. We define the *condition* of the map M_n, relative to the compact interval $[a, b]$, by

$$\operatorname{cond}_\infty M_n = \|M_n\|_\infty \|M_n^{-1}\|_\infty, \tag{1.1}$$

where the norms are $\|u\|_\infty = \max_{1\leqslant k\leqslant n} |u_k|$ (in $\mathbf{R}^n$) and $\|p\|_\infty = \max_{a\leqslant x\leqslant b} |p(x)|$ (in $\mathbf{P}_{n-1}[a, b]$). We are interested in the growth rate of $\operatorname{cond}_\infty M_n$ as $n \longrightarrow \infty$, and how this growth depends on the particular interval $[a, b]$ chosen.

The answer is relatively straightforward for symmetric intervals $[-\omega, \omega]$ and for intervals $[a, b]$ with $0 \leqslant a < b$, in which cases the condition number in (1.1) can be expressed explicitly in terms of $u_{T_{n-1}}$ (or $u_{T_{n-2}}$), where T_m denotes the Chebyshev polynomial of degree m on the appropriate interval (Theorems 3.1, 3.2). It will follow, in particular, that on $[-\omega, \omega]$ and $[0, \omega]$, $\omega > 0$, the condition grows exponentially with n, and that the minimum growth occurs precisely when $\omega = 1$, in which case $\operatorname{cond}_\infty M_n$ grows like $(1+\sqrt{2})^n$ on $[-1, 1]$ and like $(1+\sqrt{2})^{2n}$ on $[0, 1]$. This ought to be contrasted with the linear growth $\sqrt{2}\,n$ for the condition on $[-1, 1]$ of polynomials represented in terms of Chebyshev polynomials [1].

For asymmetric intervals $[a, b]$ with, say, $a < 0 < b$, $|a| < b$, the problem appears to be considerably more complex, and we are no longer able to ascertain the exact growth rate of (1.1). Instead, we obtain two upper bounds for $\operatorname{cond}_\infty M_n$, one being asymptotically sharp in the extreme case $|a| = b$, the other in the extreme case $a = 0$ (Theorem 4.1).

Received December 1, 1977; revised April 17, 1978.
AMS (MOS) subject classifications (1970). Primary 41A10; Secondary 65D99, 65G05.
Key words and phrases. Parametrization of polynomials, power basis, numerical condition.
*Sponsored in part by the National Science Foundation under grant MCS 76-00842A01.

2. **Preliminaries on the Coefficients of Chebyshev Polynomials.** In the following we need estimates for the largest coefficients in $T_n(x/\omega)$ and $T_n^*(x/\omega)$, where T_n is the Chebyshev polynomial of the first kind and T_n^* the "shifted" Chebyshev polynomial $T_n^*(x) = T_n(2x - 1)$.

It is well known that

$$T_n\left(\frac{x}{\omega}\right) = \sum_{k=0}^{[n/2]} c_k x^{n-2k}, \tag{2.1}$$

where

$$c_k = (-1)^k \frac{n}{2} \frac{(n-k-1)!}{k!(n-2k)!} \left(\frac{2}{\omega}\right)^{n-2k}, \qquad 0 \leqslant k \leqslant [n/2].$$

For fixed t, with $0 < t < \frac{1}{2}$, we put $k = tn$, and let $n \longrightarrow \infty$. Using Stirling's formula, we find

$$|c_{tn}| \sim \frac{n^{-\frac{1}{2}}}{2\sqrt{2\pi}} \frac{1}{\sqrt{t(1-t)(1-2t)}} \left(\frac{2}{\omega}\right)^n e^{ng(t)}, \qquad n \longrightarrow \infty,$$

where

$$g(t) = (1-t)\ln(1-t) - t\ln t - (1-2t)\ln(1-2t) - 2t\ln(2/\omega), \qquad 0 < t < \tfrac{1}{2}.$$

From $g(0) = 0$, $g(\frac{1}{2}) = -\ln(2/\omega)$, $g'(t) = \ln[(1-2t)^2\omega^2/4t(1-t)]$, it is seen that $g(t)$ has a unique maximum on $[0, \frac{1}{2}]$, assumed at

$$t = t_0 = \frac{1}{2}\left(1 - \frac{1}{\sqrt{1+\omega^2}}\right).$$

Since

$$g(t_0) = \ln\frac{1-t_0}{1-2t_0} = \ln[\tfrac{1}{2}(1+\sqrt{1+\omega^2})], \quad \sqrt{t_0(1-t_0)(1-2t_0)} = \tfrac{1}{2}\omega(1+\omega^2)^{-3/4},$$

we thus find for the maximum coefficient of $T_n(x/\omega)$ the asymptotic approximation

$$\|u_{T_n(x/\omega)}\|_\infty \sim \frac{1}{\sqrt{2\pi}} \frac{(1+\omega^2)^{3/4}}{\omega} n^{-\frac{1}{2}} \left(\frac{1+\sqrt{1+\omega^2}}{\omega}\right)^n, \qquad n \longrightarrow \infty. \tag{2.2}$$

For $\omega = 1$, this gives

$$\|u_{T_n}\|_\infty \sim \frac{2^{\frac{1}{4}}}{\sqrt{\pi}} n^{-\frac{1}{2}} (1+\sqrt{2})^n, \qquad n \longrightarrow \infty \quad (\omega = 1), \tag{2.2'}$$

which agrees with a result attributed to an (anonymous) referee in J. R. Rice [3, p. 304].

Since $T_n^*(x^2) = T_{2n}(x)$, the analogous result for $T_n^*(x/\omega)$ is readily obtained from (2.2) by replacing n by $2n$ and ω by $\sqrt{\omega}$,

$$\|u_{T_n^*(x/\omega)}\|_\infty \sim \frac{1}{2\sqrt{\pi}} \frac{(1+\omega)^{3/4}}{\sqrt{\omega}} n^{-\frac{1}{2}} \left(\frac{2+\omega+2\sqrt{1+\omega}}{\omega}\right)^n, \qquad n \longrightarrow \infty. \tag{2.3}$$

For $\omega = 1$, this gives

$$\|u_{T_n^*}\|_\infty \sim \frac{2^{-\frac{1}{4}}}{\sqrt{\pi}} n^{-\frac{1}{2}}(3+2\sqrt{2})^n, \quad n \to \infty \quad (\omega = 1). \tag{2.3'}$$

In Table 2.1 we compare the true values of $\|u_{T_n(x/\omega)}\|_\infty$ with their asymptotic approximations in (2.2) for selected values of n and ω.

ω	$n = 5$		$n = 10$		$n = 20$		$n = 40$	
	true	(2.2)	true	(2.2)	true	(2.2)	true	(2.2)
10	5.00(−1)	9.36(−1)	1.00	1.09	2.00	2.09	1.06(1)	1.09(1)
5	1.00	1.11	2.00	2.12	1.06(1)	1.09(1)	4.02(2)	4.11(2)
1	2.00(1)	2.46(1)	1.28(3)	1.43(3)	6.55(6)	6.79(6)	2.12(14)	2.17(14)
.2	5.00(4)	9.65(4)	5.00(9)	7.17(9)	5.00(19)	5.59(19)	5.00(39)	4.82(39)
.1	1.60(6)	5.82(6)	5.12(12)	1.33(13)	5.24(25)	9.91(25)	5.50(51)	7.72(51)

TABLE 2.1. The quality of the asymptotic formula (2.2)

We also note that

$$\|u_{T_n(x/\omega)}\|_\infty \geqslant \|u_{T_{n-1}(x/\omega)}\|_\infty, \quad n = 1, 2, 3, \ldots, \ \omega \leqslant 1, \tag{2.4}$$

where equality holds only for $n = 1$, $\omega = 1$. This follows easily from the three-term recurrence relation for Chebyshev polynomials and from the alternating character of the coefficients c_k in (2.1). The inequality in (2.4) holds for all $\omega \leqslant 2$, if n is restricted to $n \geqslant 2$, and it indeed holds for any fixed ω, if n is sufficiently large, as is seen from (2.2).

3. The Condition of M_n for Symmetric Intervals and for Intervals on One Side of the Origin. We shall always assume (without loss of generality) that our basic interval $[a, b]$ is centered to the right of the origin, so that $0 \leqslant |a| \leqslant b$. The Chebyshev polynomial T_m, adjusted to the interval $[a, b]$, will be denoted by $T_m[a, b]$,

$$T_m[a, b](x) = T_m\left(\frac{2x - a - b}{b - a}\right), \quad a \leqslant x \leqslant b.$$

Relative to any such interval $[a, b]$, the norm of the map M_n is easily seen to be

$$\|M_n\|_\infty = \sum_{k=1}^{n} b^{k-1} = \begin{cases} \dfrac{b^n - 1}{b - 1}, & b \neq 1, \\ n, & b = 1. \end{cases} \tag{3.1}$$

More delicate is the determination of $\|M_n^{-1}\|_\infty$, as this amounts to finding the norms of the linear functionals $\lambda_k: p \mapsto p^{(k-1)}(0)/(k-1)!$, $p \in \mathbf{P}_{n-1}[a, b]$, $k = 1, 2, \ldots, n$. Indeed,

$$\|M_n^{-1}\|_\infty = \max_{1 \leqslant k \leqslant n} \|\lambda_k\|_\infty. \tag{3.2}$$

While it is known [5, Satz 6.11] that, for $2 \leqslant k \leqslant n$, the extremal in $\mathbf{P}_{n-1}[a, b]$ for

the functional λ_k is a Zolotarev polynomial of degree $n-1$, it appears difficult, in the case of a general interval $[a, b]$, to pinpoint the parameter involved in the Zolotarev polynomial, and there may correspond different Zolotarev polynomials to different values of k. For these reasons the case of an arbitrary interval will be dealt with by other (less sophisticated and cruder) methods in Section 4.

For symmetric intervals $[-\omega, \omega]$, $\omega > 0$, on the other hand, the appropriate Zolotarev polynomials are known to be the Chebyshev polynomials T_{n-1} or T_{n-2}; indeed, $\|\lambda_k\|_\infty = |T_{n-1}^{(k-1)}[-\omega, \omega](0) + T_{n-2}^{(k-1)}[-\omega, \omega](0)|/(k-1)!$, $k = 1, 2, \ldots, n$, $n \geqslant 2$ [5, p. 167], and therefore,

$$\max_{1 \leqslant k \leqslant n} \|\lambda_k\|_\infty = \|u_{T_{n-1}[-\omega,\omega]+T_{n-2}[-\omega,\omega]}\|_\infty.$$

Since $T_n[-\omega, \omega](x) = T_n(x/\omega)$, and T_m is an even or odd polynomial, depending on the parity of m, we thus have, in view of (3.1), (3.2):

THEOREM 3.1. *The condition number* (1.1) *on* $[-\omega, \omega]$ *is given by*

$$\text{(3.3)} \qquad \operatorname{cond}_\infty M_n = \frac{\omega^n - 1}{\omega - 1} \max \{\|u_{T_{n-1}(x/\omega)}\|_\infty, \|u_{T_{n-2}(x/\omega)}\|_\infty\},$$

where $(\omega^n - 1)/(\omega - 1)$ (*here and in the sequel*) *is to be interpreted as having the value* n *if* $\omega = 1$.

It follows from (2.2) that for $\omega > 1$, $\omega = 1$, $0 < \omega < 1$, the condition of M_n for large n grows, respectively, like $(1 + \sqrt{1 + \omega^2})^n$, $(1 + \sqrt{2})^n$, $[(1 + \sqrt{1 + \omega^2})/\omega]^n$ (disregarding a factor $n^{\pm 1/2}$ and constant factors), so that the growth is smallest, asymptotically, when $\omega = 1$. Selected numerical values of cond M_n are shown in Table 3.1.

ω	$n = 5$	$n = 10$	$n = 20$	$n = 40$
10	1.11(4)	1.11(9)	2.11(19)	1.10(40)
5	7.81(2)	4.39(6)	2.17(14)	7.74(29)
1	4.00(1)	5.76(3)	5.45(7)	3.51(15)
.2	6.25(3)	6.25(8)	6.25(18)	6.25(38)
.1	8.89(4)	2.84(11)	2.91(24)	3.05(50)

TABLE 3.1. The condition of M_n on $[-\omega, \omega]$

Another special case which can be disposed of similarly is the case of an interval $[a, b]$ with $0 \leqslant a < b$. Here (see, e.g., [4, p. 93]) $\|\lambda_k\|_\infty = |T_{n-1}^{(k-1)}[a, b](0)|/(k-1)!$, and we can state

THEOREM 3.2. *The condition number* (1.1) *on* $[a, b]$, *where* $0 \leqslant a < b$, *is given by*

$$\text{(3.4)} \qquad \operatorname{cond}_\infty M_n = \frac{b^n - 1}{b - 1} \|u_{T_{n-1}[a,b]}\|_\infty.$$

We note that the expression on the right of (3.4), even for an arbitrary interval $[a, b]$, is always a lower bound for $\operatorname{cond}_\infty M_n$, since

$$\|M_n^{-1}\|_\infty = \sup_{p \in \mathbf{P}_{n-1}[a,b]} \frac{\|M_n^{-1} p\|_\infty}{\|p\|_\infty} \geqslant \|u_{T_{n-1}[a,b]}\|_\infty. \tag{3.5}$$

To illustrate Theorem 3.2, we consider the interval $[0, \omega]$, $\omega > 0$. Here, $T_{n-1}[0, \omega](x) = T^*_{n-1}(x/\omega)$, and depending on whether $\omega > 1$, $\omega = 1$, or $0 < \omega < 1$, Eq. (2.3) shows that the condition grows, respectively, like $(2 + \omega + 2\sqrt{1+\omega})^n$, $(3 + 2\sqrt{2})^n$ and $[(2 + \omega + 2\sqrt{1+\omega})/\omega]^n$, thus again slowest, asymptotically, when $\omega = 1$. Selected numerical values are shown in Table 3.2.

ω	$n = 5$	$n = 10$	$n = 20$	$n = 40$
10	3.56(4)	4.93(10)	1.80(23)	3.27(48)
5	5.00(3)	8.91(8)	3.67(19)	8.47(40)
1	1.28(3)	1.12(7)	7.34(14)	2.16(30)
.2	1.00(5)	3.20(11)	6.23(24)	3.02(51)
.1	1.42(6)	1.46(14)	1.53(30)	3.27(62)

TABLE 3.2. The condition of M_n on $[0, \omega]$

4. The Condition of M_n on an Arbitrary Interval. We now wish to make some progress towards the more difficult problem of estimating $\text{cond}_\infty M_n$ for an arbitrary right-centered interval $[a, b]$, $0 \leqslant |a| \leqslant b$. We content ourselves with establishing upper bounds for $\text{cond}_\infty M_n$. (A trivial, but not very useful, lower bound can be had from (3.1) and (3.5).)

Our main tool is the following simple observation.

LEMMA 4.1. *Let $s^T = [s_1, s_2, \ldots, s_n]$ be any vector of n distinct nodes in $[a, b]$ and $V_n(s)$ the corresponding Vandermonde matrix*

$$V_n(s) = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ s_1 & s_2 & \cdots & s_n \\ \cdots & \cdots & \cdots & \cdots \\ s_1^{n-1} & s_2^{n-1} & \cdots & s_n^{n-1} \end{bmatrix} \quad (a \leqslant s_\nu \leqslant b,\ \nu = 1, 2, \ldots, n). \tag{4.1}$$

Then

$$\|M_n^{-1}\|_\infty \leqslant n \|V_n^{-1}(s)\|_\infty. \tag{4.2}$$

Proof. Let

$$p(x) = \sum_{k=1}^{n} u_k x^{k-1}, \qquad a \leqslant x \leqslant b,$$

be an arbitrary polynomial of degree $\leqslant n - 1$. From

$$\sum_{k=1}^{n} s_\nu^{k-1} u_k = p(s_\nu), \qquad \nu = 1, 2, \ldots, n,$$

or, equivalently,

$$V_n^T(s)u = \pi, \quad u^T = [u_1, u_2, \ldots, u_n], \quad \pi^T = [p(s_1), p(s_2), \ldots, p(s_n)],$$

one gets immediately

$$\|u\|_\infty \leq \|u\|_1 \leq \|[V_n^{-1}(s)]^T\|_1 \|\pi\|_1 \leq n\|V_n^{-1}(s)\|_\infty \|\pi\|_\infty \leq n\|V_n^{-1}(s)\|_\infty \|p\|_\infty,$$

hence (4.2). □

It is tempting to optimize the bound in (4.2) by minimizing $\|V_n^{-1}(s)\|_\infty$ over all admissible node vectors s. Unfortunately, the corresponding optimal nodes are not known explicitly. We expect, however, the Chebyshev points on $[a, b]$ to provide a reasonably good alternative. In order to carry out the necessary computations, we need the following properties of Vandermonde matrices.

LEMMA 4.2 (SHIFT PROPERTY). *Let* $t = [t_1, t_2, \ldots, t_n]^T$ *and* $t - \mu = [t_1 - \mu, t_2 - \mu, \ldots, t_n - \mu]^T$. *Then*

$$V_n^{-1}(t - \mu) = V_n^{-1}(t)(D_n^{-1} P_n D_n)^T, \tag{4.3}$$

where $D_n = \operatorname{diag}(1, \mu, \mu^2, \ldots, \mu^{n-1})$ *and* P_n *is the initial* $(n \times n)$*-segment of the Pascal triangle, that is*

$$D_n^{-1} P_n D_n = \begin{bmatrix} 1 & \mu & \mu^2 & \mu^3 & \cdots \\ 0 & 1 & \binom{2}{1}\mu & \binom{3}{2}\mu^2 & \cdots \\ 0 & 0 & 1 & \binom{3}{1}\mu & \cdots \\ 0 & 0 & 0 & 1 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}_{(n \times n)}. \tag{4.4}$$

Proof. It is well known (see, e.g., [2]) that $V_n^{-1}(t) = [u_{\kappa\lambda}]$, where

$$\prod_{\substack{\nu=1 \\ \nu \neq \kappa}}^{n} \frac{x - t_\nu}{t_\kappa - t_\nu} \equiv \sum_{\lambda=1}^{n} u_{\kappa\lambda} x^{\lambda-1}.$$

The elements $u'_{\kappa\lambda}$ of $V_n^{-1}(t - \mu)$, therefore, are the coefficients of the polynomial

$$\prod_{\nu \neq \kappa} \frac{x + \mu - t_\nu}{t_\kappa - t_\nu} = \sum_{\rho=1}^{n} u_{\kappa\rho}(x + \mu)^{\rho-1} = \sum_{\rho=1}^{n} u_{\kappa\rho} \sum_{\lambda=1}^{\rho} \binom{\rho - 1}{\lambda - 1} x^{\lambda-1} \mu^{\rho-\lambda}$$

$$= \sum_{\lambda=1}^{n} x^{\lambda-1} \sum_{\rho=\lambda}^{n} u_{\kappa\rho} \binom{\rho - 1}{\lambda - 1} \mu^{\rho-\lambda},$$

that is,

$$u'_{\kappa\lambda} = \sum_{\rho=\lambda}^{n} u_{\kappa\rho} \binom{\rho - 1}{\lambda - 1} \mu^{\rho-\lambda}.$$

This, written in matrix form, is precisely (4.3). □

In the following two lemmas,

$$\cos\theta_\nu, \quad \theta_\nu = \frac{2\nu - 1}{2n}\pi, \qquad \nu = 1, 2, \ldots, n,$$

denote the Chebyshev points on $[-1, 1]$.

LEMMA 4.3. *If* $t_\nu = \tau\cos\theta_\nu$, $\nu = 1, 2, \ldots, n$, $\tau > 0$, *then*

$$n\|V_n^{-1}(t)\|_\infty \leqslant \frac{3^{3/4}}{4(\sqrt{2}-1)}(\tau + 1)\left|T_n\left(\frac{i}{\tau}\right)\right| \quad (i = \sqrt{-1}). \tag{4.5}$$

Proof. From [2, Theorem 5.2]** one has

$$n\|V_n^{-1}(t)\|_\infty \leqslant \frac{(\tau+1)n}{2(\sqrt{2}-1)}\left|\frac{T_n(i/\tau)}{T_n(i)}\right|\left\|V_n^{-1}\left(\frac{1}{\tau}t\right)\right\|_\infty,$$

and from [2, Example 6.2]

$$n\left\|V_n^{-1}\left(\frac{1}{\tau}t\right)\right\|_\infty \leqslant \frac{3^{3/4}}{2}|T_n(i)|.$$

LEMMA 4.4. *If* $t_\nu = \tau(1 + \cos\theta_\nu)$, $\nu = 1, 2, \ldots, n$, $\tau > 0$, *then*

$$n\|V_n^{-1}(t)\|_\infty \leqslant \frac{\tau}{\sqrt{1+2\tau}}\,T_n\left(\frac{1}{\tau} + 1\right). \tag{4.6}$$

Proof. From [2, Eq. (4.1′)] one obtains

$$n\|V_n^{-1}(t)\|_\infty \leqslant \frac{T_n(1/\tau + 1)}{\min\limits_{1\leqslant\nu\leqslant n}\left\{\dfrac{1/\tau + 1 + \cos\theta_\nu}{\sin\theta_\nu}\right\}}, \tag{4.7}$$

having used $|T_n'(\cos\theta_\nu)| = n/\sin\theta_\nu$. An elementary calculation will show that

$$f(\theta) = \frac{1/\tau + 1 + \cos\theta}{\sin\theta}$$

has a unique minimum on $0 < \theta < \pi$ at $\theta = \theta_0$, where $\cos\theta_0 = -\tau/(\tau+1)$. Thus

$$\min_{0<\theta<\pi} f(\theta) = \frac{1/\tau + 1 - \tau/(\tau+1)}{\sqrt{1 - \tau^2/(\tau+1)^2}} = \frac{1}{\tau}\sqrt{1+2\tau},$$

from which (4.6) follows by virtue of (4.7). □

Now the Chebyshev points on $[a, b]$ are given by

$$s_\nu = \frac{a+b}{2} + \frac{b-a}{2}\cos\theta_\nu = a + \frac{b-a}{2}(1 + \cos\theta_\nu), \qquad \nu = 1, 2, \ldots, n. \tag{4.8}$$

Each of these two representations suggests an application of the shift property in Lemma 4.2, the first with $t_\nu = \tau\cos\theta_\nu$, $\mu = -(a+b)/2$, the second with $t_\nu = \tau(1 + \cos\theta_\nu)$, $\mu = -a$, where $\tau = (b-a)/2$ in both. Observing also that

$$\|V_n^{-1}(t - \mu)\|_\infty \leqslant \|V_n^{-1}(t)\|_\infty\|D_n^{-1}P_nD_n\|_1 = (1 + |\mu|)^{n-1}\|V_n^{-1}(t)\|_\infty,$$

**Theorem 5.2 in [2] is stated for n even; the same theorem, however, also holds if n is odd.

and using Lemmas 4.3 and 4.4 to estimate $\|V_n^{-1}(t)\|_\infty$, we can easily estimate $\|V_n^{-1}(s)\|_\infty$ for the nodes in (4.8), hence $\|M_n^{-1}\|_\infty$ by Lemma 4.1, and finally $\operatorname{cond}_\infty M_n$, using (3.1). The result is stated as

THEOREM 4.1. *The condition number* (1.1) *on* $[a, b]$, *where* $0 \leq |a| \leq b$, *satisfies the inequality*

$$\operatorname{cond}_\infty M_n \leq \frac{3^{3/4}}{4(\sqrt{2}-1)} \frac{2+b-a}{2+b+a} \frac{b^n-1}{b-1} \left(1+\frac{b+a}{2}\right)^n \left|T_n\left(\frac{2i}{b-a}\right)\right|, \tag{4.9}$$

as well as the inequality

$$\operatorname{cond}_\infty M_n \leq \frac{b-a}{2(1+|a|)\sqrt{1+b-a}} \frac{b^n-1}{b-1} (1+|a|)^n T_n\left(\frac{2}{b-a}+1\right). \tag{4.10}$$

Theorem 4.1 holds for arbitrary intervals $[a, b]$, subject to $|a| \leq b$, but is of interest only in the case $a \leq 0 < b$ of an interval containing the origin. It will be useful to characterize such an interval by its "degree of asymmetry"

$$\alpha = (b+a)/(b-a), \qquad 0 \leq \alpha \leq 1,$$

and its half-width

$$\tau = (b-a)/2,$$

in terms of which $b = (1+\alpha)\tau$, $a = -(1-\alpha)\tau$.

We first examine the extreme cases $\alpha = 0$ (perfect symmetry) and $\alpha = 1$ (perfect asymmetry), typified by the intervals $[-\omega, \omega]$ and $[0, \omega]$, $\omega > 0$. In the first case, by virtue of

$$2\left|T_n\left(\frac{i}{\omega}\right)\right| = \left(\frac{1+\sqrt{1+\omega^2}}{\omega}\right)^n + \left(\frac{1-\sqrt{1+\omega^2}}{\omega}\right)^n \sim \left(\frac{1+\sqrt{1+\omega^2}}{\omega}\right)^n, \quad n \to \infty,$$

we find that the bound in (4.9) has the correct exponential growth rate as $n \to \infty$, which can be obtained from (3.3) and (2.2), while the bound in (4.10) grows at a larger exponential rate. (We say here that a sequence $\{c_n\}$ has exponential growth rate γ if $|c_{n+1}/c_n| \sim \gamma$ as $n \to \infty$.) The reverse is true in the second case, as can be seen from

$$2T_n\left(\frac{2}{\omega}+1\right) = \left(\frac{2+\omega+2\sqrt{1+\omega}}{\omega}\right)^n + \left(\frac{2+\omega-2\sqrt{1+\omega}}{\omega}\right)^n$$

$$\sim \left(\frac{2+\omega+2\sqrt{1+\omega}}{\omega}\right)^n, \quad n \to \infty,$$

and comparison with (3.4), (2.3). We, therefore, expect (4.9) to be sharper than (4.10) if the interval $[a, b]$ is more nearly symmetric (i.e., α small), and (4.10) better than (4.9) for more asymmetric intervals (α close to 1). That this is indeed the case can be seen by forming the ratio ρ of the exponential growth rates in (4.9) and (4.10), and expressing the result in terms of α and τ,

$$\rho = \frac{1+\alpha\tau}{1+(1-\alpha)\tau} \lambda(\tau), \quad \lambda(\tau) = \frac{1+\sqrt{1+\tau^2}}{1+\tau+\sqrt{1+2\tau}}.$$

One verifies that $\lambda(\tau) < 1$ for all τ, with $\lambda(0) = \lambda(\infty) = 1$, so that $\rho < 1$ certainly if $1 + \alpha\tau < 1 + (1 - \alpha)\tau$, i.e., $\alpha < \frac{1}{2}$. Thus, (4.9) *is asymptotically sharper than* (4.10) *whenever* $\alpha < \frac{1}{2}$. The condition on α is best possible for $\tau \to \infty$, but too stringent for specific finite values of τ. If $\tau = 1$, e.g., one finds (4.9) better than (4.10) whenever $\alpha < .8216 \ldots$, and as $\tau \to 0$, (4.9) is always better.

We illustrate Theorem 4.1 in Figure 4.1, where we plot the exponential growth rates of the bounds in (4.9) and (4.10) for intervals of fixed half-width $\tau = 1$, and asymmetries α varying from 0 to 1. (The growth rates are $(1 + \alpha)^2(1 + \sqrt{2})$ and $(1 + \alpha)(2 - \alpha)(2 + \sqrt{3})$, respectively.) The true asymptotic growth rate presumably interpolates somehow between the boundary values $1 + \sqrt{2}$ and $2(2 + \sqrt{3})$ (cf. the dashed line in Figure 4.1).

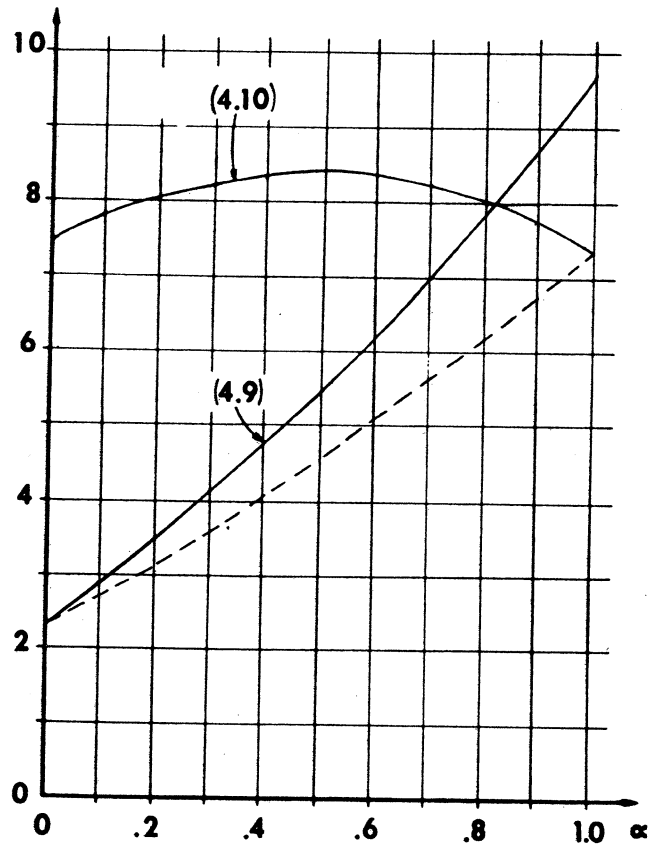


FIGURE 4.1. The asymptotic growth rates of the bounds in (4.9) and (4.10) for $a = -1 + \alpha$, $b = 1 + \alpha$, $0 \leqslant \alpha \leqslant 1$.

ACKNOWLEDGMENT. The author is indebted to Theodore J. Rivlin for pointing out the relevance of Zolotarev polynomials, which led to improved versions of Theorems 3.1 and 3.2.

Department of Computer Sciences
Purdue University
West Lafayette, Indiana 47907

1. W. GAUTSCHI, "The condition of orthogonal polynomials," *Math. Comp.*, v. 26, 1972, pp. 923–924.
2. W. GAUTSCHI, "Norm estimates for inverses of Vandermonde matrices," *Numer. Math.*, v. 23, 1975, pp. 337–347.
3. J. R. RICE, "A theory of condition," *SIAM J. Numer. Anal.*, v. 3, 1966, pp. 287–310.
4. T. J. RIVLIN, *The Chebyshev Polynomials*, Wiley, New York, 1974.
5. A. SCHÖNHAGE, *Approximationstheorie*, de Gruyter, Berlin and New York, 1971.

8.9. [83] "The Condition of Vandermonde-like Matrices Involving Orthogonal Polynomials"

---

[83] "The Condition of Vandermonde-like Matrices Involving Orthogonal Polynomials," *Linear Algebra Appl.* **52/53**, 293–300 (1983).

The Condition of Vandermonde-like Matrices Involving Orthogonal Polynomials*

Walter Gautschi
Department of Computer Sciences
Purdue University
West Lafayette, Indiana 47907

To my teacher, Alexander M. Ostrowski, in gratitude on his 90th birthday

Submitted by Richard A. Brualdi

ABSTRACT

The condition number (relative to the Frobenius norm) of the $n \times n$ matrix $P_n = [p_{i-1}(x_j)]_{i,j=1}^n$ is investigated, where $p_r(\cdot) = p_r(\cdot; d\lambda)$ are orthogonal polynomials with respect to some weight distribution $d\lambda$, and x_j are pairwise distinct real numbers. If the nodes x_j are the zeros of p_n, the condition number is either expressed, or estimated from below and above, in terms of the Christoffel numbers for $d\lambda$, depending on whether the p_r are normalized or not. For arbitrary real x_j and normalized p_r a lower bound of the condition number is obtained in terms of the Christoffel function evaluated at the nodes. Numerical results are given for minimizing the condition number as a function of the nodes for selected classical distributions $d\lambda$.

1. INTRODUCTION

Let $p_r(t) = p_r(t; d\lambda)$, $r = 0, 1, 2, \ldots$, denote a sequence of orthogonal polynomials relative to some positive measure $d\lambda(t)$ on the real line. If in the Vandermonde matrix the successive powers $1, t, t^2, \ldots$ are replaced by the successive orthogonal polynomials $p_0(t), p_1(t), p_2(t), \ldots$, there results the matrix

$$P_n = \begin{bmatrix} p_0(x_1) & p_0(x_2) & \cdots & p_0(x_n) \\ p_1(x_1) & p_1(x_2) & \cdots & p_1(x_n) \\ \cdots & \cdots & \cdots & \cdots \\ p_{n-1}(x_1) & p_{n-1}(x_2) & \cdots & p_{n-1}(x_n) \end{bmatrix}, \qquad p_r(t) = p_r(t; d\lambda), \tag{1.1}$$

*Work supported in part by the National Science Foundation under grant MCS-7927158.

which is nonsingular for pairwise distinct nodes $x_1, x_2, \ldots, x_n$. We shall assume here that all nodes are real. Our interest is in the condition of P_n. We find it convenient to consider the condition number

$$\operatorname{cond}_F(P_n) = \|P_n\|_F \|P_n^{-1}\|_F \tag{1.2}$$

with respect to the Frobenius norm $\|A\|_F = [\operatorname{tr}(A^T A)]^{1/2}$, or the closely related Turing condition number $\operatorname{cond}_T(P_n) = n^{-1} \operatorname{cond}_F(P_n)$.

In Section 2 we discuss the case of ortho*normal* polynomials $\{p_r(\cdot; d\lambda)\}$ and nodes at the zeros $\xi_\nu^{(n)}$ of p_n. Unnormalized polynomials are considered in Section 3, and arbitrary real nodes in Section 4. In Section 5 we comment on the problem of minimizing the condition number in (1.2).

2. ORTHONORMAL POLYNOMIALS—NODES AT ZEROS OF p_n

THEOREM 2.1. *Let $p_r(\cdot; d\lambda)$, $r = 0, 1, 2, \ldots$, be the orthonormal polynomials with respect to the (positive) measure $d\lambda$, and $x_\nu = \xi_\nu^{(n)}$, $\nu = 1, 2, \ldots, n$, the zeros of $p_n(\cdot; d\lambda)$. Let furthermore $\lambda_\nu = \lambda_\nu^{(n)}$, $\nu = 1, 2, \ldots, n$, denote the Christoffel numbers for $d\lambda$. Then*

$$\operatorname{cond}_F(P_n) = \left(\sum_{\nu=1}^{n} \lambda_\nu \sum_{\nu=1}^{n} \frac{1}{\lambda_\nu} \right)^{1/2}. \tag{2.1}$$

REMARK. If $m_A(\lambda)$, $m_H(\lambda)$ denote, respectively, the arithmetic and the harmonic mean of the (positive) numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$, the result (2.1) may be restated in terms of the Turing condition number as

$$\operatorname{cond}_T(P_n) = \left(\frac{m_A(\lambda)}{m_H(\lambda)} \right)^{1/2}. \tag{2.1'}$$

Letting $d\lambda$ vary, for any fixed positive integer n, over all positive measures which admit orthogonal polynomials of degree $\leqslant n$, it follows that $\operatorname{cond}_T(P_n)$, hence also $\operatorname{cond}_F(P_n)$, attains its minimum precisely when $\lambda_1 = \lambda_2 = \cdots = \lambda_n$. By a classical result [2] this is the case if and only if $\{p_r(\cdot; d\lambda)\}$ are the Chebyshev polynomials of the first kind.

TABLE 1

THE CONDITION OF P_n FOR SOME CLASSICAL ORTHOGONAL POLYNOMIALS

n	Legendre	Chebyshev 2nd kind	Laguerre	Hermite
5	5.362(0)	5.916(0)	2.076(2)	1.373(1)
10	1.155(1)	1.483(1)	1.005(6)	6.832(2)
20	2.494(1)	3.924(1)	7.770(13)	3.989(6)
40	5.367(1)	1.071(2)	1.924(30)	3.699(14)
80	1.148(2)	2.976(2)	6.607(63)	1.095(31)

Proof of Theorem 2.1. Let $P = P_n$ and $\Lambda = \operatorname{diag}(\lambda_1,\ldots,\lambda_n)$. From the discrete orthogonality property of orthonormal polynomials,

$$\sum_{\nu=1}^{n} \lambda_\nu p_r(\xi_\nu) p_s(\xi_\nu) = \left\{ \begin{matrix} 1 & \text{if} & r = s \\ 0 & \text{if} & r \neq s \end{matrix} \right\}, \qquad r, s = 0, 1, \ldots, n-1,$$

it follows that $P\Lambda^{1/2} = Q$ is an orthogonal matrix. Therefore,

$$P^T P = \Lambda^{-1/2} Q^T Q \Lambda^{-1/2} = \Lambda^{-1}, \qquad (P^{-1})^T P^{-1} = Q \Lambda Q^T,$$

so that

$$\|P\|_F^2 = \operatorname{tr}(P^T P) = \operatorname{tr}(\Lambda^{-1}),$$
$$\|P^{-1}\|_F^2 = \operatorname{tr}(Q \Lambda Q^T) = \operatorname{tr}(\Lambda). \qquad \blacksquare$$

The proof reveals that $1/\lambda_\nu$ are the squares of the singular values σ_ν of P, from which (2.1) follows also on account of

$$\operatorname{cond}_F(P_n) = \left(\sum_{\nu=1}^{n} \sigma_\nu^2 \sum_{\nu=1}^{n} \frac{1}{\sigma_\nu^2} \right)^{1/2}, \qquad \sigma_\nu = \sigma_\nu(P_n). \tag{2.2}$$

The numerical behavior of the condition number in (2.1) is illustrated in Table 1 for some classical orthogonal polynomials. (The numbers in parentheses indicate decimal exponents.)

3. UNNORMALIZED POLYNOMIALS

For unnormalized orthogonal polynomials there seems to be no result comparable in simplicity to (2.1). However, we can prove

THEOREM 3.1. *Let* $d_r = \int_{\mathbb{R}} p_{r-1}^2(t; d\lambda)\, d\lambda(t)$, $r = 1,2,\ldots$, *and* $\Delta_n^2 = \max d_r / \min d_r$, *where the maximum and minimum are taken over* $r = 1,2,\ldots,n$. *Then, in the notation of Theorem* 2.1, *if* $x_\nu = \xi_\nu^{(n)}$, $\nu = 1,2,\ldots,n$,

$$\frac{1}{\Delta_n} \leqslant \frac{\operatorname{cond}_F(P_n)}{\left(\sum_{\nu=1}^{n} \lambda_\nu \sum_{\nu=1}^{n} \frac{1}{\lambda_\nu}\right)^{1/2}} \leqslant \Delta_n. \tag{3.1}$$

Proof. Letting $P = P_n$ and $D = \operatorname{diag}(d_1, d_2, \ldots, d_n)$, we now find that $D^{-1/2} P \Lambda^{1/2} = Q$ is orthogonal. Therefore,

$$\operatorname{tr}(P^T P) = \operatorname{tr}(\Lambda^{-1/2} Q^T D Q \Lambda^{-1/2}),$$

$$\operatorname{tr}((P^{-1})^T P^{-1}) = \operatorname{tr}(D^{-1/2} Q \Lambda Q^T D^{-1/2}).$$

With

$$r_\nu = e_\nu^T Q^T D Q e_\nu, \qquad s_\nu = e_\nu^T Q \Lambda Q^T e_\nu, \tag{3.2}$$

where e_ν is the νth coordinate vector, we thus have

$$(\operatorname{cond}_F P)^2 = \sum_{\nu=1}^{n} \frac{r_\nu}{\lambda_\nu} \sum_{\nu=1}^{n} \frac{s_\nu}{d_\nu}. \tag{3.3}$$

Since $\|Qe_\nu\|_2 = \|Q^T e_\nu\|_2 = 1$, the quantities r_ν, s_ν in (3.2) are Rayleigh quotients of D and Λ, respectively; hence, in particular,

$$\min_r d_r \leqslant r_\nu \leqslant \max_r d_r. \tag{3.4}$$

Furthermore

$$\sum_{\nu=1}^{n} s_\nu = \operatorname{tr}(Q \Lambda Q^T) = \operatorname{tr}(\Lambda).$$

Therefore, (3.1) follows from (3.3) by replacing r_ν and d_ν by the bounds in (3.4). ∎

4. ARBITRARY REAL NODES

We now consider arbitrary real nodes x_ν, but assume *normalized* orthogonal polynomials $p_r(\cdot; d\lambda)$. We recall the definition of the Christoffel function (see, e.g. [1]):

$$\lambda_n(x_0) = \min_{\substack{p \in \mathbb{P}_{n-1} \\ p(x_0)=1}} \int_{\mathbb{R}} p^2(t)\, d\lambda(t), \qquad x_0 \in \mathbb{R}, \tag{4.1}$$

or, equivalently,

$$[\lambda_n(x)]^{-1} = \sum_{k=0}^{n-1} p_k^2(x), \qquad x \in \mathbb{R}. \tag{4.2}$$

THEOREM 4.1. *Let $x_1, x_2, \ldots, x_n$ be pairwise distinct real numbers and $\{p_r(\cdot; d\lambda)\}$ the orthonormal polynomials with respect to the (positive) measure $d\lambda$. Then*

$$\operatorname{cond}_F(P_n) \geqslant \left(\sum_{\nu=1}^{n} \lambda_n(x_\nu) \sum_{\nu=1}^{n} \frac{1}{\lambda_n(x_\nu)} \right)^{1/2}. \tag{4.3}$$

Proof. Let

$$l_\nu(t) = \prod_{\substack{\mu=1 \\ \mu \neq \nu}}^{n} \frac{t - x_\mu}{x_\nu - x_\mu}, \qquad \nu = 1, 2, \ldots, n,$$

be the fundamental Lagrange interpolation polynomials for the nodes $x_1, x_2, \ldots, x_n$, and let

$$l_\nu(t) = \sum_{\mu=1}^{n} a_{\nu\mu} p_{\mu-1}(t).$$

Then, as is easily seen,

$$P_n^{-1} = [a_{\nu\mu}].$$

Consequently,

$$\int_{\mathbb{R}} \sum_{\nu=1}^{n} l_\nu^2(t)\, d\lambda(t) = \int_{\mathbb{R}} \sum_{\nu} \sum_{\mu} a_{\nu\mu} p_{\mu-1}(t) \sum_{\kappa} a_{\nu\kappa} p_{\kappa-1}(t)\, d\lambda(t)$$

$$= \sum_{\nu} \sum_{\mu,\kappa} a_{\nu\mu} a_{\nu\kappa} \int_{\mathbb{R}} p_{\mu-1}(t) p_{\kappa-1}(t)\, d\lambda(t)$$

$$= \sum_{\nu} \sum_{\mu} a_{\nu\mu}^2,$$

and therefore

$$\|P_n^{-1}\|_F^2 = \int_{\mathbb{R}} \sum_{\nu=1}^{n} l_\nu^2(t)\, d\lambda(t). \tag{4.4}$$

Since $l_\nu \in \mathbb{P}_{n-1}$ and $l_\nu(x_\nu) = 1$, it follows from (4.1) that

$$\|P_n^{-1}\|_F^2 \geqslant \sum_{\nu=1}^{n} \lambda_n(x_\nu). \tag{4.5}$$

On the other hand, using (4.2),

$$\|P_n\|_F^2 = \sum_{\nu=1}^{n} \sum_{k=0}^{n-1} p_k^2(x_\nu) = \sum_{\nu=1}^{n} \frac{1}{\lambda_n(x_\nu)}. \tag{4.6}$$

The assertion (4.3) now follows immediately from (4.5), (4.6). ∎

We remark that (4.3) holds with equality if $x_\nu = \xi_\nu^{(n)}$, $\nu = 1,2,\ldots,n$, as follows from Theorem 2.1 and the fact that $\lambda_n(\xi_\nu^{(n)}) = \lambda_\nu^{(n)}$, $\nu = 1,2,\ldots,n$. We also remark that Theorem 4.1 remains valid, with essentially the same proof, if the nodes are complex and $\lambda_n(\cdot)$ is defined as in (4.1), with $p^2(t)$ replaced by $|p(t)|^2$.

5. MINIMIZING THE CONDITION NUMBER

An interesting problem is to determine the optimally conditioned matrix P_n for any fixed measure $d\lambda$, i.e. to find the nodes $x_1, x_2, \ldots, x_n$ which

minimize the condition number $\operatorname{cond}_F(P_n)$ over all pairwise distinct real nodes. We report here on attempts to solve this problem numerically.

Recall from (2.2) that

$$\operatorname{cond}_F(P_n) = n\left[\frac{m_A(\sigma^2)}{m_H(\sigma^2)}\right]^{1/2},$$

where $m_A(\sigma^2)$, $m_H(\sigma^2)$ are, respectively, the arithmetic and the harmonic mean of the squares of the singular values σ_ν of P_n. It follows that $\operatorname{cond}_F(P_n) \geqslant n$, so that the smallest possible condition number (attained for the Chebyshev measure and Chebyshev nodes; cf. Remark to Theorem 2.1) is equal to n.

Assuming normalized polynomials $p_r(\cdot; d\lambda)$, the condition number $\operatorname{cond}_F(P_n)$, or rather its square, can be written explicitly as the product of the two expressions in (4.4) and (4.6). Both expressions, including their gradients, can be computed fairly easily, the integral in (4.4) and similar integrals involved in the gradient being evaluated (exactly) by the n-point Gauss-Christoffel quadrature rule associated with the measure $d\lambda$. Using this computation in conjunction with a minimization algorithm, for which we selected the procedures in [3], we were able to obtain the results shown in Tables 2 and 3. Although only local extrema can be found in this manner, the closeness of the minimum to the absolute minimum n in some of the examples suggests that the results are indeed optimal to within the precision given.

In Table 2 we show the "optimal" nodes and the minimum condition number for Legendre polynomials ($d\lambda(t) = dt$ on $[-1, 1]$). Table 3 displays only the optimal condition number (without nodes) for some of the other

TABLE 2

OPTIMALLY CONDITIONED MATRIX P_n FOR LEGENDRE POLYNOMIALS

n	x_ν	$\operatorname{cond}_F(P_n)$	n	x_ν	$\operatorname{cond}_F(P_n)$
2	±.5773502692	2.0	20	±.9885188046	23.46822182
5	±.8780893894	5.229550605		±.9585058326	
	±.5336883454			±.9083037137	
	.0			±.8372548421	
10	±.9610897501	11.01832471		±.7462848447	
	±.8560330091			±.6372598211	
	±.6772857139			±.5126743680	
	±.4346101969			±.3755003472	
	±.1497603704			±.2290741509	
				±.0769922707	

TABLE 3

OPTIMAL CONDITION OF P_n FOR SOME CLASSICAL ORTHOGONAL POLYNOMIALS

n	Chebyshev 2nd kind	Laguerre	Hermite
2	2.0	2.0	2.0
5	5.544624008	2.106683223(1)	8.393706126
10	1.295377448(1)	1.340933671(3)	8.275431133(1)
20	3.240814863(1)	< 6.4040073(6)	7.476911820(3)

classical polynomials. In the case $n = 20$ of Laguerre polynomials the minimization algorithm could not be made to converge within a reasonable amount of time. Interestingly, some of the nodes in the Laguerre case turn out to be negative.

For $n = 2$ it can be shown by direct computation that the optimal condition always equals $\text{cond}_F(P_2) = 2$, and that the optimal nodes are the zeros ξ_1, ξ_2 of $p_2(\cdot; d\lambda)$, provided the measure $d\lambda$ is "symmetric" in the sense $\int_{\mathbf{R}} t\, d\lambda(t) = \int_{\mathbf{R}} t^3\, d\lambda(t) = 0$. In the Laguerre case, the optimal nodes are $x_1 = 0, x_2 = 2$.

REFERENCES

1 P. G. Nevai, Orthogonal polynomials, *Mem. Amer. Math. Soc.* 18, No. 213, Amer. Math. Soc., Providence, R.I., 1979.

2 C. Posse, Sur les quadratures, *Nouv. Ann. Math. (2)* 14:49–62 (1875).

3 D. F. Shanno and K. H. Phua, Remark on Algorithm 500, *ACM Trans. Math. Software* 6: 618–622 (1980).

Received 21 August 1981; revised 5 October 1981

8.10. [110] "Lower Bounds for the Condition Number of Vandermonde Matrices"

---

[110] (with G. Inglese) "Lower Bounds for the Condition Number of Vandermonde Matrices," *Numer. Math.* **52**, 241–250 (1988).

Numer. Math. 52, 241–250 (1988)

Numerische Mathematik

Lower Bounds for the Condition Number of Vandermonde Matrices*

Walter Gautschi[1,**] and Gabriele Inglese[2]

[1] Department of Computer Sciences, Purdue University, West Lafayette, IN 47907, USA
[2] CNR-Istituto Analisi Globale e Applicazioni, Via S. Marta 13/A, I-50139 Florence, Italy

Summary. We derive lower bounds for the ∞-condition number of the $n \times n$-Vandermonde matrix $V_n(x)$ in the cases where the node vector $x^T = [x_1, x_2, \ldots, x_n]$ has positive elements or real elements located symmetrically with respect to the origin. The bounds obtained grow exponentially in n, with $O(2^n)$ and $O(2^{n/2})$, respectively. We also compute the optimal spectral condition numbers of $V_n(x)$ for the two node configurations (including the optimal nodes) and compare them with the bounds obtained.

Subject Classifications: AMS (MOS): 15A12, 49D15, 65F35; CR: G1.3, G1.6.

1.1. Introduction

The condition of Vandermonde matrices

$$V_n(x) = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \\ \ldots & \ldots & \ldots & \ldots \\ x_1^{n-1} & x_2^{n-1} & \ldots & x_n^{n-1} \end{bmatrix}, \quad x^T = [x_1, x_2, \ldots, x_n], \quad n > 1, \tag{1.1}$$

where the nodes x_ν are real or complex numbers, and the related question of estimating the norm of $[V_n(x)]^{-1}$ have been studied in [2–5]. In [4] we considered the problem of minimizing the condition number

$$\kappa_{n,p}(x) = \operatorname{cond}_p V_n(x) = \|V_n(x)\|_p \, \|V_n^{-1}(x)\|_p, \tag{1.2}$$

where $p = \infty$, over all positive node vectors $x \in \mathbb{R}_+^n$, or all real symmetric node vectors $x \in \mathbb{R}^n$, $x_\nu + x_{n+1-\nu} = 0$ $(\nu = 1, 2, \ldots, n)$. We managed to obtain certain necessary conditions for optimality, computed optimal node configurations for

* Dedicated to the memory of James H. Wilkinson

** Supported, in part, by the National Science Foundation under grant CCR-8704404

$n=2$ and $n=3$ in the case of positive nodes, and for $2\leqq n\leqq 6$ in the case of symmetric nodes, but did not address the question of how fast

$$\kappa_{n,p}=\inf_{x}\kappa_{n,p}(x) \tag{1.3}$$

grows with n (when $p=\infty$). While the exact growth rate is still unknown, we now derive lower bounds for $\kappa_{n,\infty}$ which show that the growth of $\kappa_{n,\infty}$ is exponential, namely at least $O(2^n)$ and $O(2^{n/2})$ in the two respective cases. We also compute $\kappa_{n,2}$ for $2\leqq n\leqq 10$ in the former, and for $2\leqq n\leqq 16$ in the latter case, and depict the optimal nodes graphically.

We first recall from [4] some key formulas that will be needed. In the case of nonnegative nodes

$$x_1>x_2>\ldots>x_n\geqq 0, \tag{1.4}$$

we have

$$\kappa_{n,\infty}(x)=\max\{n, g_n(x)\}\cdot\max_{1\leqq\nu\leqq n} g_{n,\nu}(x), \tag{1.5}$$

where

$$g_n(x)=\sum_{\mu=1}^{n} x_\mu^{n-1}, \tag{1.6}$$

$$g_{n,\nu}(x)=\prod_{\substack{\mu=1\\ \mu\neq\nu}}^{n}\frac{1+x_\mu}{|x_\nu-x_\mu|}, \qquad \nu=1,2,\ldots,n. \tag{1.7}$$

For real symmetric nodes

$$\begin{gathered} x_\nu+x_{n+1-\nu}=0, \qquad \nu=1,2,\ldots,n,\\ x_1>x_2>\ldots>x_{[n/2]}>0 \end{gathered} \tag{1.8}$$

(note that $x_{(n+1)/2}=0$ if n is odd), we have

$$\kappa_{n,\infty}(x)=\max\left\{\frac{n}{2}, f_n(x)\right\}\cdot\max_{1\leqq\nu\leqq[(n+1)/2]} f_{n,\nu}(x), \tag{1.9}$$

where

$$f_n(x)=\sum_{\mu=1}^{[n/2]} x_\mu^{n-1}, \tag{1.10}$$

$$f_{n,\nu}(x)=\left(1+\frac{1}{x_\nu}\right)\prod_{\substack{\mu=1\\ \mu\neq\nu}}^{n/2}\frac{1+x_\mu^2}{|x_\nu^2-x_\mu^2|}, \qquad \nu=1,2,\ldots,n/2 \quad (n \text{ even}), \tag{1.11}$$

$$f_{n,\nu}(x)=\frac{1+x_\nu}{x_\nu^2}\prod_{\substack{\mu=1\\ \mu\neq\nu}}^{(n-1)/2}\frac{1+x_\mu^2}{|x_\nu^2-x_\mu^2|}, \qquad \nu=1,2,\ldots,(n-1)/2, \quad (n \text{ odd}), \tag{1.12}$$

$$f_{n,(n+1)/2}(x)=2\prod_{\mu=1}^{(n-1)/2}\left(1+\frac{1}{x_\mu^2}\right),$$

Empty products in (1.11), (1.12), when $n=2$ or $n=3$, are understood to have the value 1.

2. Positive Nodes

Although the following Theorem 2.1 will subsequently be sharpened, we state and prove it here because of its simplicity and elementary proof.

Theorem 2.1. *Let* $\kappa_{n,\infty}$ *be the infimum in* (1.3) (*for* $p=\infty$) *taken over all nonnegative nodes* (1.4). *Then, for* $n\geqq 2$,

$$\kappa_{n,\infty} > 2^{n-1}. \tag{2.1}$$

Proof. The optimal point is known to be finite (cf. the remarks preceding Theorem 3.1 of [4]). Letting

$$E_C = \{x\in\mathbb{R}^n : C = x_1 > x_2 > \ldots > x_n \geqq 0\},$$

it suffices therefore to show that

$$\kappa_{n,\infty}(x) > 2^{n-1}, \quad \text{all } x\in E_C, \text{ all } C>0. \tag{2.2}$$

At the heart of the proof is the elementary observation that

$$\inf_{0\leqq v<u\leqq C} \frac{1+u}{u-v} = 1+\frac{1}{C}, \tag{2.3}$$

where the infimum is attained for $u=C$, $v=0$.

Assume first $C>1$. Since, by (1.6), $g_n(x)\geqq C^{n-1}$ for $x\in E_C$, we have from (1.5), (1.7)

$$\begin{aligned}
\kappa_{n,\infty}(x) &\geqq C^{n-1} g_{n,n}(x) = C^{n-1} \prod_{\mu=1}^{n-1} \frac{1+x_\mu}{x_\mu - x_n} \\
&\geqq C^{n-1} \inf_{E_C} \prod_{\mu=1}^{n-1} \frac{1+x_\mu}{x_\mu - x_n} \geqq C^{n-1} \prod_{\mu=1}^{n-1} \inf_{0\leqq v<u\leqq C} \frac{1+u}{u-v} \\
&= C^{n-1}\left(1+\frac{1}{C}\right)^{n-1} = (1+C)^{n-1} > 2^{n-1},
\end{aligned}$$

where (2.3) has been used to evaluate the last infimum. Similarly, if $C\leqq 1$,

$$\kappa_{n,\infty}(x) \geqq n\cdot \prod_{\mu=1}^{n-1} \frac{1+x_\mu}{x_\mu - x_n} \geqq n\left(1+\frac{1}{C}\right)^{n-1} \geqq 2\cdot 2^{n-1} > 2^{n-1}$$

if $n\geqq 2$. □

We now improve upon Theorem 2.1 by establishing the following

Theorem 2.2. *Let* $\kappa_{n,\infty}$ *be as in Theorem* 2.1, *Then, for* $n\geqq 2$,

$$\kappa_{n,\infty} \geqq (n-1)\left\{1+\left(1-\frac{1}{n}\right)^{-1/(n-1)}\right\}^{n-1}. \tag{2.4}$$

In particular,

$$\kappa_{n,\infty} > (n-1)\cdot 2^{n-1}, \qquad n \geqq 2. \tag{2.5}$$

Proof. By Theorems 5.2 and 5.3 of [4], if $x=a$ is a minimum point of $\kappa_{n,\infty}(x)$, then

$$a_n = 0, \qquad g_n(a) = \sum_{\mu=1}^{n-1} a_\mu^{n-1} = n, \tag{2.6}$$

and, by (1.5),

$$\kappa_{n,\infty}(x) \geqq \kappa_{n,\infty}(a) = n \cdot \max_{1 \leqq \nu \leqq n} g_{n,\nu}(a).$$

In particular, therefore,

$$\kappa_{n,\infty}(x) \geqq n \cdot g_{n,n}(a) = n \prod_{\mu=1}^{n-1} \frac{1+a_\mu}{a_\mu}. \tag{2.7}$$

To get a lower bound, we minimize the product in (2.7) subject to the constraint in (2.6) (thereby changing the meaning of the variables a_μ). Using Lagrange multipliers, we obtain the necessary conditions

$$-\frac{1}{a_\nu^2} \prod_{\substack{\mu=1 \\ \mu \neq \nu}}^{n-1} \frac{1+a_\mu}{a_\mu} + \lambda(n-1)\, a_\nu^{n-2} = 0, \qquad \nu = 1, 2, \ldots, n-1,$$

or, equivalently,

$$\prod_{\mu=1}^{n-1} \frac{1+a_\mu}{a_\mu} = \lambda(n-1)\, a_\nu^{n-1}(1+a_\nu), \qquad \nu = 1, 2, \ldots, n-1.$$

This implies $a_1 = a_2 = \ldots = a_{n-1} = \alpha$, hence, by (2.6),

$$(n-1)\,\alpha^{n-1} = n, \qquad \alpha = \left(1 - \frac{1}{n}\right)^{-1/(n-1)}.$$

Substituting in (2.7) gives

$$\kappa_{n,\infty}(x) \geqq n \left(\frac{1+\alpha}{\alpha}\right)^{n-1},$$

which is (2.4). The corollary (2.5) is an immediate consequence of (2.4). □

Expanding the lower bound in (2.4) in powers of n^{-1}, we can also write

$$\kappa_{n,\infty} \geqq n \cdot 2^{n-1}(1 - \tfrac{1}{2}n^{-1} - \tfrac{1}{8}n^{-2} + \tfrac{1}{16}n^{-3} + \tfrac{19}{128}n^{-4} + \ldots), \qquad n \geqq 2. \tag{2.4'}$$

The five terms shown provide an accuracy of about 2 correct significant decimal digits when $n=2$, and 7 correct digits when $n=16$.

3. Symmetric Nodes

Since the infimum of $\kappa_{n,\infty}(x)$ over all $x\in\mathbb{R}^n$ is attained at a symmetric node configuration, if it is unique (see [4, Thm. 3.1]), the study of symmetric nodes is particularly appropriate. We have, in this case, results analogous to those in Theorems 2.1 and 2.2. Since the proofs are similar, we try to be brief.

Theorem 3.1. *Let $\kappa_{n,\infty}$ be the infimum in* (1.3) (*for $p=\infty$*) *taken over all nodes satisfying* (1.8). *Then, for $n\geqq 2$,*

$$\kappa_{n,\infty}\geqq 2^{n/2}. \tag{3.1}$$

If $n>2$, then (3.1) *holds with strict inequality.*

Proof. We now let

$$E_C=\{x\in\mathbb{R}^n\colon x_1=C>x_2>\ldots>x_{[n/2]}>0,\ x_\nu+x_{n+1-\nu}=0 \text{ all } \nu\}$$

and consider first the case n even. If $C>1$, then by (1.9)–(1.11),

$$\kappa_{n,\infty}(x)\geqq C^{n-1}\max_{1\leqq\nu\leqq n/2} f_{n,\nu}(x)\geqq C^{n-1}f_{n,n/2}(x)$$

$$=C^{n-1}\left(1+\frac{1}{x_{n/2}}\right)\prod_{\mu=1}^{(n/2)-1}\frac{1+x_\mu^2}{x_\mu^2-x_{n/2}^2},$$

and thus, by (2.3),

$$\kappa_{n,\infty}(x)\geqq C^{n-1}\left(1+\frac{1}{C}\right)\left(1+\frac{1}{C^2}\right)^{(n/2)-1}=(1+C)(1+C^2)^{(n/2)-1}>2^{n/2}.$$

Likewise, for $C\leqq 1$, if $n\geqq 4$,

$$\kappa_{n,\infty}(x)\geqq\frac{n}{2}f_{n,n/2}(x)\geqq 2\cdot 2\left(1+\frac{1}{C^2}\right)^{(n/2)-1}\geqq 2\cdot 2^{n/2}>2^{n/2}.$$

For $n=2$ one has $\kappa_{2,\infty}(x)\geqq 2$ (see [4, Eq. (4.1)]).

Consider now $n(\geqq 3)$ odd. Then, for $C>1$, by (1.9), (1.10), and (1.12),

$$\kappa_{n,\infty}(x)\geqq C^{n-1}\max_{1\leqq\nu\leqq(n+1)/2} f_{n,\nu}(x)\geqq C^{n-1}f_{n,(n+1)/2}(x)$$

$$=2C^{n-1}\prod_{\mu=1}^{(n-1)/2}\left(1+\frac{1}{x_\mu^2}\right)\geqq 2C^{n-1}\left(1+\frac{1}{C^2}\right)^{(n-1)/2}$$

$$=2(1+C^2)^{(n-1)/2}>2^{n/2},$$

and, for $C\leqq 1$,

$$\kappa_{n,\infty}(x)\geqq\frac{n}{2}f_{n,(n+1)/2}(x)=n\prod_{\mu=1}^{(n-1)/2}\left(1+\frac{1}{x_\mu^2}\right)\geqq n\cdot 2^{(n-1)/2}>2^{n/2}. \quad \square$$

Theorem 3.2. *Let $\kappa_{n,\infty}$ be as in Theorem 3.1. Then, for $n \geqq 4$,*

$$\kappa_{n,\infty} > \begin{cases} (n-2)\left\{1+\left(1-\frac{2}{n}\right)^{-2/(n-1)}\right\}^{(n-2)/2}, & n \text{ even}, \\ (n-3)\left\{1+\left(1-\frac{3}{n}\right)^{-2/(n-1)}\right\}^{(n-3)/2}, & n \text{ odd}. \end{cases} \tag{3.2}$$

In particular,

$$\kappa_{n,\infty} > \begin{cases} (n-2)\cdot 2^{(n-2)/2}, & n \text{ even}, \\ (n-3)\cdot 2^{(n-3)/2}, & n \text{ odd}. \end{cases} \tag{3.3}$$

Proof. By [4, Thm. 3.3], if $x=a$ is a minimum point, then

$$\kappa_{n,\infty}(x) \geqq \frac{n}{2} \max_{1 \leqq \nu \leqq [(n+1)/2]} f_{n,\nu}(a), \tag{3.4}$$

where

$$\sum_{\mu=1}^{[n/2]} a_\mu^{n-1} = \frac{n}{2} \tag{3.5}$$

and

$$a_1 > a_2 > \ldots > a_{[n/2]} > 0.$$

We assume first $n(\geqq 4)$ even. Then, by (3.4),

$$\kappa_{n,\infty}(x) \geqq \frac{n}{2} f_{n,n/2}(a) = \frac{n}{2}\left(1+\frac{1}{a_{n/2}}\right) \prod_{\mu=1}^{(n/2)-1} \frac{1+a_\mu^2}{a_\mu^2 - a_{n/2}^2} > \frac{n}{2}\cdot 2 \cdot \prod_{\mu=1}^{(n/2)-1} \frac{1+a_\mu^2}{a_\mu^2}. \tag{3.6}$$

We have used here $a_{n/2} \leqq 1$, which must certainly hold if (3.5) is to be true. We now minimize the last product in (3.6), subject to

$$\sum_{\mu=1}^{(n/2)-1} a_\mu^{n-1} = \frac{n}{2} - a_{n/2}^{n-1}. \tag{3.7}$$

(We may assume here that $a_{n/2} > 0$ is fixed.) Using Lagrange multipliers, we get

$$-\frac{2}{a_\nu^3} \prod_{\substack{\mu=1 \\ \mu \neq \nu}}^{(n/2)-1} \frac{1+a_\mu^2}{a_\mu^2} + \lambda(n-1)\, a_\nu^{n-2} = 0, \qquad \nu = 1, 2, \ldots, \frac{n}{2}-1,$$

or, equivalently,

$$\prod_{\mu=1}^{(n/2)-1} \frac{1+a_\mu^2}{a_\mu^2} = \frac{1}{2}\lambda(n-1)\, a_\nu^{n-1}(1+a_\nu^2), \qquad \nu = 1, 2, \ldots, \frac{n}{2}-1,$$

which implies $a_1 = a_2 = \ldots = a_{(n/2)-1} = \alpha$. By (3.7),

$$\left(\frac{n}{2}-1\right)\alpha^{n-1} = \frac{n}{2} - a_{n/2}^{n-1} < \frac{n}{2},$$

hence

$$\alpha<\left(1-\frac{2}{n}\right)^{-1/(n-1)}.$$

Therefore, by (3.6),

$$\kappa_{n,\infty}>n\cdot\left(\frac{1+\alpha^2}{\alpha^2}\right)^{(n-2)/2}>n\cdot\frac{\left\{1+\left(1-\frac{2}{n}\right)^{-2/(n-1)}\right\}^{(n-2)/2}}{\left(\frac{n}{n-2}\right)^{(n-2)/(n-1)}},$$

which, by increasing the denominator to $n/(n-2)$, yields the first inequality in (3.2).

Assuming now $n(\geqq 5)$ odd, we have

$$\begin{aligned}\kappa_{n,\infty}(x)&\geqq\frac{n}{2}\cdot f_{n,(n-1)/2}(a)=\frac{n}{2}\frac{1+a_{(n-1)/2}}{a^2_{(n-1)/2}}\prod_{\mu=1}^{(n-1)/2-1}\frac{1+a_\mu^2}{a_\mu^2-a^2_{(n-1)/2}}\\&>\frac{n}{2}\cdot 2\cdot\prod_{\mu=1}^{(n-1)/2-1}\frac{1+a_\mu^2}{a_\mu^2}.\end{aligned}\tag{3.8}$$

We are led to the same problem as before, namely to minimize the last product in (3.8) subject to

$$\sum_{\mu=1}^{(n-1)2-1}a_\mu^{n-1}=\frac{n}{2}-a^{n-1}_{(n-1)/2}.$$

We find $a_1=a_2=\ldots=a_{(n-3)/2}=\alpha$, with

$$\alpha<\left(1-\frac{3}{n}\right)^{-1/(n-1)},$$

hence, by (3.8),

$$\begin{aligned}\kappa_{n,\infty}(x)&>n\cdot\left(\frac{1+\alpha^2}{\alpha^2}\right)^{(n-3)/2}>n\cdot\frac{\left\{1+\left(1-\frac{3}{n}\right)^{-2/(n-1)}\right\}^{(n-3)/2}}{\left(\frac{n}{n-3}\right)^{(n-3)/(n-1)}}\\&>(n-3)\left\{1+\left(1-\frac{3}{n}\right)^{-2/(n-1)}\right\}^{(n-3)/2}.\quad\square\end{aligned}$$

For $n=2$ and $n=3$, we have trivially $\kappa_{2,\infty}=2$, $\kappa_{3,\infty}=5$ ([4, Eqs. (4.1), (4.2)]). We can write (3.2), in expanded form, as

$$\kappa_{n,\infty}>\begin{cases}n\cdot 2^{(n-2)/2}(1-n^{-1}-\frac{3}{2}n^{-2}-\frac{1}{2}n^{-3}-\frac{7}{24}n^{-4}+\ldots), & n\text{ (even)}\geqq 4,\\ n\cdot 2^{(n-3)/2}(1-\frac{3}{2}n^{-1}-\frac{33}{8}n^{-2}-\frac{39}{16}n^{-3}+\frac{75}{128}n^{-4}+\ldots), & n\text{ (odd)}\geqq 5.\end{cases}\tag{3.2'}$$

The accuracy provided by the five terms shown is about 3 correct significant decimal digits, when $n=4$, and increases to 6 correct digits for $n=15$.

4. Numerical Results

In order to assess the quality of the bounds obtained, it would be desirable to compute the optimum condition number $\kappa_{n,\infty}$ numerically. This would require the solution of nonlinearly constrained optimization problems [4, Eqs. (3.13), (5.10)] or nonlinear programming problems [4, Eqs. (3.15), (5.12)]. Since, at this time, there seems to be no easy access to reliable software in this area, we decided to minimize the spectral condition number,

$$\kappa_{n,2}(x)=\operatorname{cond}_2 V_n(x)=\frac{\sigma_1(V_n(x))}{\sigma_n(V_n(x))}, \tag{4.1}$$

where $\sigma_\nu=\sigma_\nu(V_n)$, $\sigma_1>\sigma_2>\ldots>\sigma_n$, are the singular values of V_n. This requires only unconstrained optimization and singular value decomposition, for which there exists standard software. Having found $\kappa_{n,2}=\inf \kappa_{n,2}(x)$, we can use the inequality

$$\kappa_{n,\infty}>\frac{1}{n}\kappa_{n,2} \tag{4.2}$$

to get a lower bound for $\kappa_{n,\infty}$. Indeed, if $\kappa_{n,\infty}=\operatorname{cond}_\infty V_n(a)$ and $\kappa_{n,2}=\operatorname{cond}_2 V_n(b)$, then, since $\|A\|_2\leqq\sqrt{n}\|A\|_\infty$ (see, e.g., [6, Eq. (2.2–14)]), we get $\kappa_{n,2}=\operatorname{cond}_2 V_n(b)<\operatorname{cond}_2 V_n(a)\leqq n \operatorname{cond}_\infty V_n(a)=n\kappa_{n,\infty}$. In the case of nonnegative nodes, we make the usual substitution

$$x_\nu=a_\nu+(b_\nu-a_\nu)\sin^2 t_\nu, \qquad \nu=1, 2, \ldots, n,$$

where a_ν, b_ν are lower and upper bounds for x_ν, in order to reduce the problem to an unconstrained problem in the variables t_ν. In our case, $a_\nu=0$, and we took for b_ν variously $b_\nu=2$, 2.5, and 3. Using the IMSL routine ZXMIN (cf. [7, pp. ZXMIN 1–4]) for minimization (with initial approximations $t_\nu=\nu\pi/(2n+2)$, $\nu=1, 2, \ldots, n$), and the EISPACK routine SVD (cf. [1, p. 265]) for singular value decomposition, to compute $\kappa_{n,2}$, we obtained for the lower bound in (4.2) the results in the second column of Table 1. (The computation was done in single precision on the CDC 6500. Integers in parentheses denote decimal exponents.) The routine ZXMIN, for n beyond 10, was unable to produce reliable answers. In the third column of Table 1 we show the lower bounds

Table 1. Lower bounds for $\kappa_{n,\infty}$ in (4.2) and Theorem 2.2 for $n=2(1)10$

n	(4.2)	Theorem 2.2	n	(4.2)	Theorem 2.2
			6	3.715 (2)	1.754 (2)
2	1.207	3.000	7	1.812 (3)	4.150 (2)
3	4.250	9.899	8	9.062 (3)	9.582 (2)
4	1.764 (1)	2.781 (1)	9	4.621 (4)	2.173 (3)
5	7.892 (1)	7.167 (1)	10	2.393 (5)	4.858 (3)

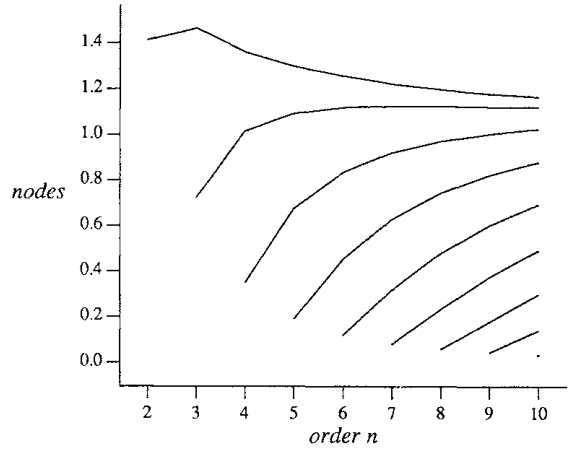


Fig. 1. Optimal positive nodes for $n=2(1)10$

computed from Theorem 2.2, Eq. (2.4). It can be seen that the bound from Theorem 2.2 is competitive with the one from (4.2) for about $n \leqq 5$, but then gradually weakens. Both bounds should be compared for $n=2$ and $n=3$ with the values $\kappa_{2,\infty}=3$, $\kappa_{3,\infty}=12.708$ computed in [4, Sect. 5].

The optimal nodes, as computed for the spectral norm, were found to have $x_n=0$. The positive nodes are depicted in Fig. 1, where the largest, second-largest, etc. are connected by straight lines for visual effect.

In the case of symmetric nodes, we used the same routines as above, with the Chebyshev nodes on $[-1, 1]$ as initial approximations. The results are shown in the second column of Table 2. We compare them in the third column with the lower bounds computed from Theorem 3.2, Eq. (3.2). In this case it was possible to go as far as $n=16$. Again, the bound in (3.2) is competitive with the one from (4.2) for about $n \leqq 10$, but then slowly deteriorates. Note also from [4, Sect. 4] that $\kappa_{2,\infty}=2$, $\kappa_{3,\infty}=5$, $\kappa_{4,\infty}=11.776$, $\kappa_{5,\infty}=21.456$, and $\kappa_{6,\infty}=51.330$.

The nonnegative optimal nodes in the symmetric case are shown graphically in Fig. 2.

Table 2. Lower bounds for $\kappa_{n,\infty}$ in (4.2) and Theorem 3.2 for $n=2(1)16$

n	(4.2)	Theorem 3.2	n	(4.2)	Theorem 3.2
			9	4.644 (1)	5.610 (1)
2	0.500		10	9.607 (1)	1.415 (2)
3	1.049		11	2.119 (2)	1.457 (2)
4	1.465	5.175	12	4.522 (2)	3.479 (2)
5	2.904	5.162	13	1.012 (3)	3.574 (2)
6	5.216	1.894 (1)	14	2.204 (3)	8.250 (2)
7	1.092 (1)	1.945 (1)	15	4.986 (3)	8.457 (2)
8	2.149 (1)	5.444 (1)	16	1.102 (4)	1.908 (3)

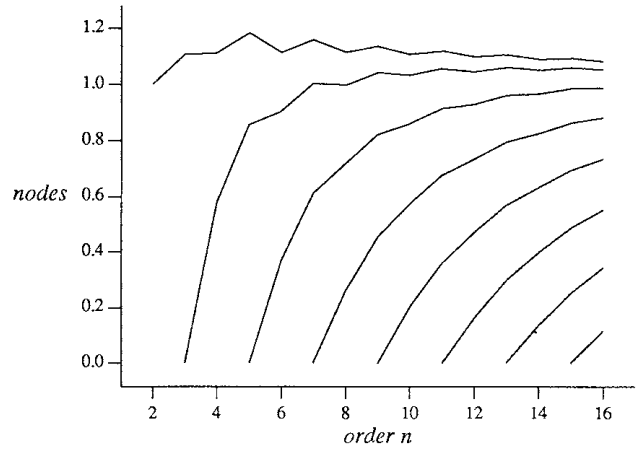


Fig. 2. Optimal symmetric nodes for $n=2(1)16$

Interestingly, the same results were obtained if the initial approximations were chosen to be nonsymmetric, for example the Chebyshev points on [0, 1]. (Since the routine takes considerably longer to converge in this case, we verified this only for $2 \leqq n \leqq 10$.) This seems to indicate that the optimally conditioned Vandermonde matrix (in the spectral norm) indeed has symmetric nodes.

References

1. Garbow, B.S., Boyle, J.M., Dongarra, J.J., Moler, C.B.: Matrix Eigensystem Routines – EISPACK Guide Extension. Lecture Notes in Computer Science, Vol. 51. Berlin, Heidelberg, New York: Springer 1977
2. Gautschi, W.: On Inverses of Vandermonde and Confluent Vandermonde Matrices. Numer. Math. **4**, 117–123 (1962)
3. Gautschi, W.: Norm Estimates for Inverses of Vandermonde Matrices. Numer. Math. **23**, 337–347 (1975)
4. Gautschi, W.: Optimally Conditioned Vandermonde Matrices. Numer. Math. **24**, 1–12 (1975)
5. Gautschi, W.: On Inverses of Vandermonde and Confluent Vandermonde Matrices III. Numer. Math. **29**, 445–450 (1978)
6. Golub, G.H., Van Loan, C.F.: Matrix Computations. Baltimore: Johns Hopkins University Press 1983
7. IMSL Library Reference Manual **4**, Edition 9, 1982

Received September 2, 1987 / October 28, 1987

8.11. [118] "How (Un)stable Are Vandermonde Systems?"

---

[118] "How (Un)stable Are Vandermonde systems?," in *Asymptotic and computational analysis* (R. Wong, ed.), 193–210, *Lecture Notes Pure Appl. Math.* **124** (1990).

How (Un)stable Are Vandermonde Systems?

Walter Gautschi Professor, Department of Computer Sciences, Purdue University, West Lafayette, Indiana

ABSTRACT. Results on the condition number of Vandermonde type matrices obtained during the last 25 years are reviewed. Equal emphasis is given to real and complex nodes. Recent work dealing with nodes placed sequentially on circular and elliptic contours in the complex plane receives special attention.

I. INTRODUCTION

Many problems in applied and numerical analysis eventually boil down to solving large systems of linear algebraic equations. Since the matrices and right-hand sides of such systems are typically the result of (sometimes extensive) computations, they are subject to an unavoidable level of noise caused by the rounding errors committed during their generation. It is then a matter of practical concern trying to estimate the effect of such uncertainties upon the solution of the system.

A common answer – and one which we shall adopt in the sequel – concerning any nonsingular system

$$Ax = b, \quad \det A \neq 0, \tag{1.1}$$

is to compute (or estimate) the *condition number*

$$\operatorname{cond} A = \|A\| \cdot \|A^{-1}\| \tag{1.2}$$

of the system, where $\|\cdot\|$ denotes a suitable matrix norm. Norms, for matrices

Work supported, in part, by the National Science Foundation under grant CCR-8704404.

$A = [a_{ij}]$, that will be used here are the ∞ - *norm*,

$$\|A\|_\infty = \max_i \sum_j |a_{ij}|, \tag{1.3}$$

the *Euclidean* (or spectral) *norm*,

$$\|A\|_2 = \sqrt{\rho(AA^H)}, \tag{1.4}$$

where $\rho(\cdot)$ denotes spectral radius, and the *Frobenius norm*,

$$\|A\|_F = \sqrt{\mathrm{tr}(AA^H)} = \sqrt{\sum_{i,j} |a_{ij}|^2}\ . \tag{1.5}$$

If $\varepsilon_x = \|\delta x\|/\|x\|$ is the relative error in the solution x of (1.1), caused by relative errors $\varepsilon_A = \|\delta A\|/\|A\|$, $\varepsilon_b = \|\delta b\|/\|b\|$ in the system, then the condition number in (1.2) indicates how much larger ε_x is compared to ε_A and ε_b, that is, roughly speaking,

$$\varepsilon_x \approx (\mathrm{cond}\, A)(\varepsilon_A + \varepsilon_b). \tag{1.6}$$

It always seemed important to us that the conditioning of matrices be investigated for many special classes of matrices. In this spirit, we began, 25 years ago, to take up the class of Vandermonde matrices. The original motivation came from unpleasant experiences with the computation of Gauss type quadrature rules from the moments of the underlying weight function. The sensitivity of the problem then indeed depends on the condition of certain (confluent) Vandermonde matrices with real nodes. Since then, we have intermittently looked at the conditioning of such matrices, considering not only real, but also complex nodes, and have enlarged the class of matrices by including Vandermonde-like matrices involving polynomial systems other than the system of powers. Here we present a brief survey of results obtained over the years, including also some original material (in Sections IV, V and VI).

To establish terminology and notation, we call a *Vandermonde matrix* a matrix of the form

$$V_n = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_n \\ \cdot & \cdot & \cdots & \cdot \\ z_1^{n-1} & z_2^{n-1} & \cdots & z_n^{n-1} \end{bmatrix}, \quad z_i \in \mathbb{C},\ n > 1, \tag{1.7}$$

where z_i are pairwise distinct real or complex numbers called the *nodes*. More generally, a *Vandermonde-like matrix*, with nodes z_i, is a matrix of the form

$$V_n = \begin{bmatrix} p_0(z_1) & p_0(z_2) & \cdots & p_0(z_n) \\ p_1(z_1) & p_1(z_2) & \cdots & p_1(z_n) \\ \cdot & \cdot & \cdots & \cdot \\ p_{n-1}(z_1) & p_{n-1}(z_2) & \cdots & p_{n-1}(z_n) \end{bmatrix}, \tag{1.8}$$

where $\{p_k\}$ is a system of linearly independent polynomials, often with $p_k \in \mathbf{P}_k$, the class of polynomials of degree k. Such matrices (or their transposed) are encountered, for example, when one deals with polynomial interpolation or with interpolatory approximation of linear functionals (see, e.g., [1]), the form (1.7) or (1.8) occurring depending on the choice of basis elements in polynomial spaces. Vandermonde systems with matrix (1.7) are also an important ingredient in Remes' algorithm for constructing best uniform polynomial approximations.

A brief outline of the paper is as follows. Sections II–IV are devoted to ordinary Vandermonde matrices V_n, (1.7). We begin in Section II with some basic inequalities for $\|V_n^{-1}\|_\infty$. These are then applied in Section III to obtain estimates for the condition number $\mathrm{cond}_\infty V_n$ for certain real node configurations. The bottom line here is that *V_n is ill-conditioned when all nodes are real.* Indeed, the condition number, in many cases (and perhaps always), grows exponentially with the order n of the matrix. The scenario changes drastically if one allows complex nodes. The roots of unity, for example, give rise to perfectly conditioned Vandermonde matrices. Other sequences of nodes on the unit circle are studied in Section IV. Some, like the Van der Corput sequence, perform nearly as well, and do so in a linear sequential, rather than triangular, fashion. Others, like "quasi-cyclic" sequences, do much worse. The remaining two sections discuss Vandermonde-like matrices with the polynomials p_k in (1.8) chosen to be orthogonal polynomials. In Section V we consider special real, as well as arbitrary complex nodes, in Section VI nodes placed sequentially on elliptic contours in the complex plane, and a Chebyshev system of polynomials p_k.

II. A BASIC INEQUALITY FOR INVERSES OF VANDERMONDE MATRICES

The inversion of a linear system with the Vandermonde matrix (1.7) as coefficient matrix can be easily described in terms of the elementary Lagrange interpolation polynomials

$$\ell_\lambda(z) = \prod_{\substack{\mu=1 \\ \mu \neq \lambda}}^{n} \frac{z - z_\mu}{z_\lambda - z_\mu}, \quad \lambda = 1,2,\ldots,n, \tag{2.1}$$

associated with the nodes $z_1, z_2, \ldots, z_n$. Indeed, if we expand ℓ_λ in powers of z,

$$\ell_\lambda(z) = \sum_{\mu=1}^{n} u_{\lambda\mu} z^{\mu-1}, \tag{2.2}$$

the inversion is accomplished by multiplying the μth equation by $u_{\lambda\mu}$, $\mu = 1,2,\ldots,n$, and adding up the results. In view of $\ell_\lambda(z_\nu) = \delta_{\lambda\nu}$ (the Kronecker delta), this will express the λth unknown linearly in terms of the right-hand members of the system. Hence,

$$V_n^{-1} = [u_{\lambda\mu}]_{\substack{1\le\lambda\le n\\ 1\le\mu\le n}}. \tag{2.3}$$

Combining (2.1) and (2.2) yields

$$u_{\lambda 1} + u_{\lambda 2}z + \cdots + u_{\lambda n}z^{n-1} = \pi_\lambda \prod_{\substack{\mu=1\\ \mu\neq\lambda}}^{n} (z - z_\mu), \tag{2.4}$$

where

$$\pi_\lambda = \prod_{\mu\neq\lambda} (z_\lambda - z_\mu)^{-1}. \tag{2.5}$$

Therefore, we have the alternative representation

$$u_{\lambda\mu} = (-1)^{\mu-1}\pi_\lambda \sigma_{n-\mu}^{\lambda}(z_1,\ldots,z_{\lambda-1},z_{\lambda+1},\ldots,z_n), \tag{2.6}$$

where σ_m^λ denotes the mth elementary symmetric function in the $n-1$ variables z_μ with z_λ removed.

Theorem 2.1. *For arbitrary $z_\nu \in \mathbb{C}$, with $z_\nu \neq z_\mu$ if $\nu \neq \mu$, there holds*

$$\max_\lambda \prod_{\mu\neq\lambda} \frac{\max(1,|z_\mu|)}{|z_\lambda - z_\mu|} < \|V_n^{-1}\|_\infty \le \max_\lambda \prod_{\mu\neq\lambda} \frac{1+|z_\mu|}{|z_\lambda - z_\mu|}, \tag{2.7}$$

where V_n is the matrix in (1.7). *The upper bound is attained if $z_\mu = |z_\mu| e^{i\theta}$, $\mu = 1,2,\ldots,n$, for some fixed $\theta \in \mathbb{R}$.*

Proof (Sketch). The upper bound in (2.7), and the statement about equality, follow from (2.6) and from a simple fact (see the Lemma in [7, p.118]) about elementary symmetric functions $\sigma_m = \sigma_m(x_1,\ldots,x_n)$ in n variables, namely that $\sum_{m=0}^{n} |\sigma_m| \le \prod_{\nu=1}^{n}(1+|x_\nu|)$, with equality precisely if all x_ν lie on the same ray through the origin.

For the lower bound we use the fact that

$$\sum_{\mu=0}^{n} |a_\mu| \ge |a_n| \prod_{\nu=1}^{n} \max(1,|\zeta_\nu|) \tag{2.8}$$

holds for any polynomial $p(z) = a_0 + a_1 z + \cdots + a_n z^n$, $a_n \neq 0$, having the zeros ζ_1,

$\zeta_2, \ldots, \zeta_n$, with equality if and only if $p(z) = a_n z^n$. This is a simple consequence of Jensen's formula (see [10, §2]). Applying (2.8) to the polynomial (of degree $n-1$) in (2.4) then easily yields the lower bound in (2.7); see again [10] for details. □

III. REAL NODES

An important case in which equality holds for the upper bound in (2.7) is when all nodes are nonnegative, $z_\nu = x_\nu \geq 0$. Then, letting

$$p_n(z) = \prod_{\nu=1}^{n} (z - x_\nu) \tag{3.1}$$

denote the node polynomial, we can write (2.7) in the form

$$\| V_n^{-1} \|_\infty = \frac{|p_n(-1)|}{\min_\nu \{(1 + x_\nu)|p_n'(x_\nu)|\}} \qquad (x_\nu \geq 0). \tag{3.2}$$

The techniques used in the first part of the proof of Theorem 2.1 can be adapted (cf. [8, Theorem 4.3]) to also deal with the case of real nodes located symmetrically with respect to the origin: $x_\nu \in \mathbb{R}$ with $x_\nu + x_{n+1-\nu} = 0$ for $\nu = 1, 2, \ldots, n$. In place of (3.2), one obtains

$$\| V_n^{-1} \|_\infty = \frac{|p_n(i)|}{\min\limits_{x_\nu \geq 0} \left\{ \dfrac{1 + x_\nu^2}{1 + x_\nu} |p_n'(x_\nu)| \right\}} \qquad (x_\nu + x_{n+1-\nu} = 0, \ x_\nu \in \mathbb{R}). \tag{3.3}$$

Since for given nodes x_ν the norm $\| V_n \|_\infty$ is easily calculated, the results (3.2), (3.3) allow us to evaluate the condition number $\mathrm{cond}_\infty V_n$ exactly in the respective cases. For example, if $|x_\nu| \leq 1$ for all ν, then

$$\mathrm{cond}_\infty V_n = n \| V_n^{-1} \|_\infty \qquad (|x_\nu| \leq 1). \tag{3.4}$$

We illustrate (3.2)–(3.4) with a number of examples, ordered in decreasing severity of ill-conditioning.

Example 3.1. Harmonic nodes $x_\nu = 1/\nu$, $\nu = 1, 2, \ldots, n$.

Here, an easy calculation gives $|p_n(-1)| = n+1$, and letting $\delta_\nu = (1 + x_\nu)|p_n'(x_\nu)|$, one finds

$$\delta_\nu = \frac{(\nu+1)!(n-\nu)!}{\nu^n n!}, \quad \nu = 1,2,\ldots,n.$$

There follows

$$\min_\nu \delta_\nu \le \delta_n = \frac{n+1}{n^n}.$$

(Actually, this holds with strict inequality, for $n > 2$, as can be shown by a more detailed analysis). Consequently, by (3.4) and (3.2),

$$\text{cond}_\infty V_n > n^{n+1} \quad (x_\nu = 1/\nu). \tag{3.5}$$

Note that the condition number in (3.5) grows more rapidly than $n!$, which is far worse than the condition of the notorious Hilbert matrix, which grows "only" exponentially!

Example 3.2. Equidistant nodes on [0,1]: $x_\nu = \frac{\nu-1}{n-1}$, $\nu = 1,2,\ldots,n$.

Defining δ_ν as in the previous example, an elementary computation gives

$$|p_n(-1)| = \frac{(2n-2)!}{(n-1)!(n-1)^{n-1}}, \quad \delta_\nu = \frac{(n+\nu-2)(\nu-1)!(n-\nu)!}{(n-1)^n}.$$

Putting $\nu = \kappa n$, $0 < \kappa < 1$, and studying $\delta_{\kappa n}$ for $n \to \infty$, reveals that, asymptotically, $\delta_{\kappa n}$ is a minimum when $\kappa = \frac{1}{2}$, and $\delta_{\frac{1}{2}n} \sim \pi e n (2e)^{-n}$ as $n \to \infty$. Combining this in Eq. (3.2) with the asymptotic expression for $|p_n(-1)|$, obtained by Stirling's formula, and noting that $\|V_n\|_\infty = n$, yields

$$\text{cond}_\infty V_n \sim \frac{\sqrt{2}}{4\pi} \cdot 8^n, \quad n \to \infty. \tag{3.6}$$

We are now down to exponential growth, but expect that the rate of growth can still be reduced by placing the nodes symmetrically with respect to the origin. This is confirmed in the next example.

Example 3.3. Equidistant nodes on [−1,1], $x_\nu = 1 - \frac{2(\nu-1)}{n-1}$, $\nu = 1,2,\ldots,n$.

Here we use (3.3). An asymptotic analysis similar to the one in the previous example, but more involved, shows that [8, Example 6.1]

$$\text{cond}_\infty V_n \sim \frac{1}{\pi} e^{-\frac{1}{4}\pi} e^{n(\frac{1}{4}\pi + \frac{1}{2}\ln 2)}, \quad n \to \infty. \tag{3.7}$$

Note that the exponential growth rate is now $\exp\left\{\frac{1}{4}\pi + \frac{1}{2}\ln 2\right\} = 3.1017\ldots$,

compared to 8 in the asymmetric case of Example 3.2.

Can we do better if we take the Chebyshev nodes in (–1,1)?

Example 3.4. Chebyshev nodes $x_\nu = \cos((2\nu - 1)\pi/2n)$, $\nu = 1,2,\ldots, n$.

Applying (3.3), one can prove [8, Example 6.2] that

$$\text{cond}_\infty V_n \sim \frac{3^{3/4}}{4}(1+\sqrt{2})^n, \quad n \to \infty. \tag{3.8}$$

Here the growth rate $1 + \sqrt{2} = 2.4142\ldots$ is indeed smaller than for equally spaced symmetric nodes, but not by a whole lot.

Seeing the condition of V_n continually improving through the series of examples above, one cannot help wondering whether there is an optimal set of real nodes $x^T = [x_1, x_2, \ldots, x_n]$ (say, with $x_1 > x_2 > \cdots > x_n$), and if so, what they are and what optimal growth rate of $\text{cond}_\infty V_n$ they produce as $n \to \infty$. As far as the existence of the optimum is concerned, the answer is easily seen to be affirmative (cf. [9]). We even conjecture (but have no proof as yet) that the optimal nodes are unique, subject to the above ordering. If this were true, it would follow [9, Theorem 3.1] that the optimal node configuration is symmetric with respect to the origin. Seen in this light, the recent result [14, Theorem 3.1]

$$\text{cond}_\infty V_n > 2^{n/2} \quad (n > 2,\ x_\nu \text{ symmetric}) \tag{3.9}$$

is of interest, since it shows that, accepting the above conjecture, the condition of Vandermonde matrices grows exponentially for *any* set of real nodes. Nevertheless, the growth rate indicated in (3.9) is not believed to be sharp, and the search for the optimal growth rate remains an interesting open problem. There is a result analogous to (3.9) for arbitrary positive nodes, namely [14, Theorem 2.1]

$$\text{cond}_\infty V_n > 2^{n-1} \quad (n > 1,\ x_\nu \geq 0). \tag{3.10}$$

Both bounds in (3.9) and (3.10) can be slightly sharpened (cf. Theorems 3.2 and 2.2, respectively, in [14]).

IV. COMPLEX NODES

The fact that real nodes lead to ill-conditioned Vandermonde matrices is not surprising if one considers that powers constitute, as is well known, a poor basis for polynomial approximation on the real line; see, e.g., the discussion of near linear dependence in [3, pp. 119–120], or of the conditioning of the power basis in [11]. In contrast, replacing the powers by Chebyshev polynomials, and considering the corresponding

Vandermonde-like matrices (1.8), can lead to perfectly conditioned matrices if one chooses the (real) nodes appropriately; cf. Section V below.

Now it so happens that the powers are indeed "Chebyshev polynomials" on any disc in the complex plane centered at the origin, in the sense of deviating least from zero (in the uniform norm on the disc, or on the circumference of the disc) among all monic polynomials of the same degree. Therefore, one expects better conditioning of ordinary Vandermonde matrices if one allows the nodes z_ν to be complex.

If we measure the condition in the Euclidean norm (1.4), and consider for simplicity the *unit* disc, then the nth roots of unity indeed minimize $\mathrm{cond}_2 V_n$; in fact,

$$\mathrm{cond}_2 V_n = 1 \text{ if } z_\nu = z_\nu^{(n)} = e^{i(\nu-1)2\pi/n}, \quad \nu = 1,2,\ldots,n. \tag{4.1}$$

This is an easy consequence of the orthogonality of trigonometric functions. The roots of unity, therefore, would seem to be an ideal choice for work on the unit disc, if it weren't for the fact that they form a *triangular array* of nodes, i.e., for each n, as indicated in (4.1), there are n distinct nodes $z_\nu^{(n)}$ which change as n is increased by 1. In applications to interpolation and quadrature, this would require that the function to be interpolated, or integrated, be obtained on a two-dimensional set of points. It is an interesting question to ask how well one can do with a *linear array* of nodes on the unit circle.

One naive answer to this is to first note that the set of kth roots of unity, for k even, contains as subset the $(k/2)$th roots of unity. We may therefore generate a linear sequence of nodes by adjoining to the 2^{k-1}th roots of unity every other 2^kth root of unity, going around the circle in the positive direction, and doing this for $k = 1,2,3,\ldots$. More precisely, if

$$2^{k-1} < \nu \le 2^k \tag{4.2}$$

for some $k \ge 1$, then

$$z_1 = 1, \quad z_\nu = e^{2\pi i(2(\nu-2^{k-1})-1)/2^k}. \tag{4.3}$$

We call this sequence, for lack of a better word, the *quasi-cyclic sequence* on the circle. (Such sequences have been used by Eiermann, Niethammer and Varga [4, p. 522] in the context of semiiterative methods for systems of linear algebraic equations.) With this choice of nodes, whenever n is a power of 2, the corresponding Vandermonde matrix V_n is perfectly conditioned, but otherwise, there is a chance, especially for large n, that the condition may deteriorate significantly. The bounds in Theorem 2.1, unfortunately, are too far apart to give much useful information. We therefore computed the condition number for V_n numerically, using, as seems natural on the circle, the Euclidean matrix norm (1.4). The results for $\mathrm{cond}_2 V_n$ are depicted on a logarithmic scale in Figure 4.1 for $3 \le n \le 64$. As expected, the condition number shoots up to considerable heights between two successive (large) powers of 2.

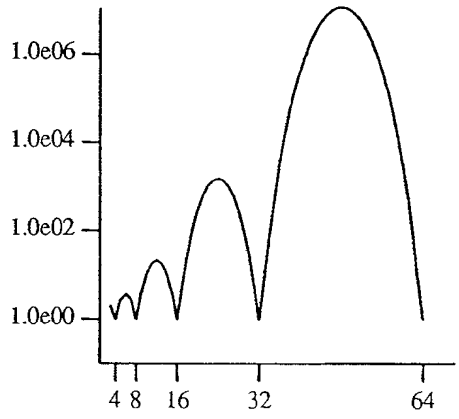


FIG. 4.1 *The condition of Vandermonde matrices* (1.7) *for* n=3(1)64 *with nodes taken from the quasi-cyclic sequence* (4.3).

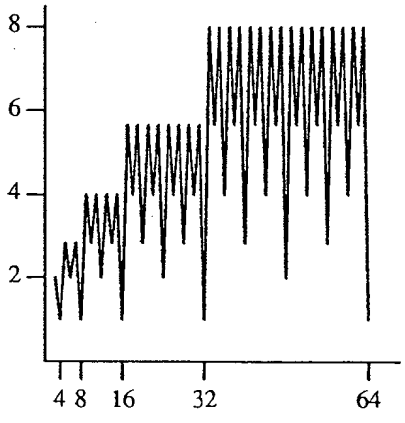


FIG. 4.2 *The condition of Vandermonde matrices* (1.7) *for* n=3(1)64 *with nodes taken from the Van der Corput sequence* (4.6).

How, then, can we avoid these large peaks? More specifically, for an integer n satisfying $2^{k-1} < n < 2^k$, in which order should we adjoin the set of 2^{k-1}th roots of unity by alternate 2^kth roots of unity such that $\max_{2^{k-1} < n < 2^k} \text{cond}_2 V_n$ is minimized? We don't know the solution to this problem, but a good candidate for an optimal (or nearly optimal) node sequence is obtained as follows.

For any integer $\nu \geq 0$, written in binary form

$$\nu = \sum_{j=0}^{\infty} \nu_j 2^j, \quad \nu_j \in \{0,1\}, \tag{4.4}$$

define the fraction $c_\nu \in [0,1)$ by

$$c_\nu = \sum_{j=0}^{\infty} \nu_j 2^{-j-1}. \tag{4.5}$$

The sequence $\{c_\nu\}_{\nu=0}^{\infty}$ is known as the *Van der Corput sequence*. We then take as nodes

$$z_\nu = e^{2\pi i c_{\nu-1}}, \quad \nu = 1,2,3, \ldots . \tag{4.6}$$

(Such nodes were used to good advantage by Fischer and Reichel [5, p. 228] in connection with the Richardson iteration method; see also Fischer and Reichel [6], Reichel and Opfer [17].) It is easily seen that for $n = 2^{k-1}$ the set $\{z_\nu: \nu = 1,2, \ldots, n\}$ consists of all the nth roots of unity, just like in the quasi-cyclic case. For values of n between 2^{k-1} and 2^k, however, the nodes (4.6) are picked in a zigzag manner from the 2^kth roots of unity, rather than cyclically around the circle as in (4.3). This achieves a more evenly distributed set of nodes, and one can hope that the condition number $\text{cond}_2 V_n$ remains correspondingly smaller. This is indeed confirmed by a computation for $3 \leq n \leq 64$, the results of which are summarized in Figure 4.2. (A similar picture, extended through $n = 148$, has previously been published by Reichel and Opfer [17].) Further computations [15] reveal a rather astonishing pattern for the eigenvalues and eigenspaces of the matrix $V_n V_n^H$ – a Hermitian Toeplitz matrix – which served as an inspiration for the work in [2]. There, it is proved, in particular, that all eigenvalues of $V_n V_n^H$ are powers of 2, the largest, $\lambda_{\max}$, always being equal to $\lambda_{\max} = 2^k$, and the smallest, $\lambda_{\min} = 2^\ell$, where $\ell = 0$ if n is odd, and $0 < \ell \leq k$ otherwise. There follows

$$\text{cond}_2 V_n \leq 2^{k/2} < \sqrt{2n} \quad (2^{k-1} < n \leq 2^k, \ z_\nu \text{ as in (4.6)}), \tag{4.7}$$

with equality on the left holding for every odd n. The various "levels" exhibited in Figure 4.2 thus have heights $2^{k/2}$, $k = 2,3,4, \ldots$. Comparison of Figures 4.2 and 4.1 clearly illustrates the significant improvement achieved by the Van der Corput sequence over the quasi-cyclic sequence. Similar phenomena on ellipses (and also on intervals) will be discussed in Section VI.

V. VANDERMONDE-LIKE MATRICES INVOLVING ORTHOGONAL POLYNOMIALS

As observed at the beginning of Section IV, the choice of orthogonal polynomials as bases in problems of approximation on the real line leads to Vandermonde-like matrices (1.8) which can be expected to have better condition than ordinary Vandermonde matrices. It is the purpose of this section to study the condition of such matrices in the case where

$$p_k(z) = p_k(z\,; d\sigma), \quad k = 0,1,2,\ldots, \tag{5.1}$$

are *orthonormal polynomials* with respect to some (positive) measure $d\sigma$ on the real line,

$$\int_{\mathbb{R}} p_r(x)p_s(x)d\sigma(x) = \delta_{rs} = \begin{cases} 1, & r = s, \\ 0, & r \neq s. \end{cases} \tag{5.2}$$

In most applications, the nodes z_ν are real and contained in the support of $d\sigma$,

$$z_\nu = x_\nu \in \mathbf{R}, \quad x_\nu \in \operatorname{supp} d\sigma. \tag{5.3}$$

A choice that appears particularly natural is that of the zeros of $p_n(\cdot\,; d\sigma)$,

$$x_\nu = x_\nu^{(n)}, \quad p_n(x_\nu\,; d\sigma) = 0, \quad \nu = 1,2,\ldots, n. \tag{5.4}$$

In this case the condition $\operatorname{cond}_F V_n$ of V_n (in the Frobenius norm (1.5)) can be expressed very simply in terms of *Christoffel numbers* $\gamma_\nu = \gamma_\nu^{(n)}(d\sigma)$ belonging to the measure $d\sigma$, i.e., in terms of the weights in the Gauss-Christoffel quadrature formula

$$\int_{\mathbb{R}} f(x)d\sigma(x) = \sum_{\nu=1}^{n} \gamma_\nu^{(n)} f(x_\nu^{(n)}) + R_n(f), \quad R_n(\mathbf{P}_{2n-1}) = 0. \tag{5.5}$$

(As is well known, $\gamma_\nu^{(n)} > 0$ for $\nu = 1,2,\ldots, n$.) Indeed, we have

Theorem 5.1. *The condition of V_n in* (1.8), *where p_k are the orthonormal polynomials* (5.1), (5.2) *and the nodes x_ν given by* (5.4), *equals*

$$\operatorname{cond}_F V_n = \left[\sum_{\nu=1}^{n} \gamma_\nu \sum_{\nu=1}^{n} \frac{1}{\gamma_\nu} \right]^{1/2}, \tag{5.6}$$

where $\gamma_\nu = \gamma_\nu^{(n)}(d\sigma)$ are the Christoffel numbers of $d\sigma$, and the norm used in (5.6) *is*

the Frobenius norm $\|\cdot\|_F$ *in* (1.5).

The proof rests on the fact that $1/\gamma_\nu^{(n)}$ are the squares of the singular values of V_n, which in turn is a consequence of the discrete orthogonality property of orthogonal polynomials (cf. [12]). Note also that $\text{cond}_F V_n \geq n$ for any set of positive numbers γ_ν.

Our first example is the analogue of the example involving the roots of unity, in the sense that it achieves optimality.

Example 5.1. The Chebyshev measure $d\sigma(x) = (1-x^2)^{-1/2}dx$ on $[-1,1]$.

Here, the Christoffel numbers $\gamma_\nu^{(n)}$ are all equal to π/n. Indeed, the Chebyshev measure is the only measure for which this is true for all n. (There are other measures, however, for which equality of Christoffel numbers holds for selected values of n; see, e.g., [13, §6], [16]). It then follows from (5.6) that $\text{cond}_F V_n = n$, i.e., V_n is optimally conditioned.

Nevertheless, optimality is achieved at a price: a *triangular* array of nodes (just like earlier with roots of unity). We will show in Section VI how one can find a *linear* array of nodes that also produces well-conditioned matrices V_n, (1.8).

Not much worse than the Chebyshev measure are those of Legendre and Chebyshev of the second kind.

Example 5.2. $d\sigma(x) = dx$ and $d\sigma(x) = (1-x^2)^{1/2}dx$ on $[-1,1]$.

Here one computes from (5.6) the following condition numbers for selected values of n:

TABLE 1 *The condition of Vandermonde-like matrices for Legendre and 2nd-kind Chebyshev polynomials*
(Numbers in parentheses indicate decimal exponents.)

n	Legendre	Chebyshev 2nd kind
5	5.362(0)	5.916(0)
10	1.155(1)	1.483(1)
20	2.494(1)	3.924(1)
40	5.367(1)	1.071(2)
80	1.148(2)	2.976(2)

In stark contrast, Laguerre and Hermite polynomials give rise to extremely ill-conditioned matrices V_n, for example, $\text{cond}_F V_{40} = 1.924(30)$ and $3.699(14)$ in the two respective cases. This is due to the presence of very small Christoffel numbers.

If z_ν are arbitrary complex nodes, one can prove a result similar to, but weaker than, (5.6); it involves the Christoffel function, rather than Christoffel numbers. We recall that the *Christoffel function* (for some measure $d\sigma$) is defined by

$$\gamma_n(z_0; d\sigma) = \min_{\substack{p \in \mathbb{P}_{n-1} \\ p(z_0)=1}} \int_{\mathbb{R}} |p(x)|^2 d\sigma(x), \quad z_0 \in \mathbb{C}, \tag{5.7}$$

where the minimum is over all complex polynomials of degree $\leq n-1$ taking on the value 1 at z_0. Alternatively,

$$[\gamma_n(z\,;d\sigma)]^{-1} = \sum_{k=0}^{n-1} |p_k(z\,;d\sigma)|^2. \tag{5.7'}$$

In place of (5.6) we then have [12]

$$\mathrm{cond}_F V_n \geq \left[\sum_{\nu=1}^{n} \gamma_n(z_\nu\,;d\sigma) \sum_{\nu=1}^{n} \frac{1}{\gamma_n(z_\nu\,;d\sigma)}\right]^{1/2}. \tag{5.8}$$

To prove (5.8), and at the same time give a version of (5.8) involving equality, one must first of all invert the matrix V_n in (1.8). This can be done similarly as in Section II for powers by expanding the fundamental Lagrange polynomial (2.1) not in powers, as in (2.2), but in the orthogonal polynomials $p_k(z) = p_k(z\,;d\sigma)$,

$$\ell_\nu(z) = \sum_{\mu=1}^{n} a_{\nu\mu} p_{\mu-1}(z), \quad \nu = 1,2,\ldots,n. \tag{5.9}$$

Then as before, one finds

$$V_n^{-1} = A, \quad A = [a_{\nu\mu}]. \tag{5.10}$$

Now

$$\begin{aligned}\int_{\mathbb{R}} \sum_{\nu=1}^{n} |\ell_\nu(x)|^2 d\sigma(x) &= \int_{\mathbb{R}} \sum_{\nu} \sum_{\mu} a_{\nu\mu} p_{\mu-1}(x) \sum_{\lambda} \bar{a}_{\nu\lambda} p_{\lambda-1}(x) d\sigma(x) \\ &= \sum_{\nu} \sum_{\mu,\lambda} a_{\nu\mu} \bar{a}_{\nu\lambda} \int_{\mathbb{R}} p_{\mu-1}(x) p_{\lambda-1}(x) d\sigma(x) \\ &= \sum_{\nu,\mu} |a_{\nu\mu}|^2\end{aligned}$$

on account of the orthonormality of the p_k. Consequently,

$$\|V_n^{-1}\|_F = \left[\int_{\mathbb{R}} \sum_{\nu=1}^{n} |\ell_\nu(x)|^2 d\sigma(x)\right]^{1/2}. \tag{5.11}$$

On the other hand,

$$\|V_n\|_F = \left[\sum_{\nu=1}^{n} \sum_{\mu=1}^{n} |p_{\mu-1}(z_\nu)|^2\right]^{1/2}, \tag{5.12}$$

which, on account of (5.7′), gives

$$\|V_n\|_F = \left[\sum_{\nu=1}^{n} \frac{1}{\gamma_n(z_\nu; d\sigma)}\right]^{1/2}. \tag{5.13}$$

The assertion (5.8) now follows by multiplying the two expressions in (5.11) and (5.13) and observing that

$$\int_{\mathbb{R}} |\ell_\nu(x)|^2 d\sigma(x) \geq \gamma_n(z_\nu; d\sigma)$$

by (5.7), since $\ell_\nu \in \mathbf{P}_{n-1}$ and $\ell_\nu(z_\nu) = 1$.

We note, however, that the product of (5.11) and (5.12) is exactly *equal* to $\text{cond}_F V_n$, and can easily be computed, at least for conventional measures $d\sigma$, the first factor by Gaussian quadrature and the other by recurrence.

Analogous results can be derived, in essentially the same way, for orthogonal polynomials $\{p_k(\cdot; d\sigma)\}$ that are *not* normalized (for example, for *monic* polynomials). Letting

$$d_k^2 = \int_{\mathbb{R}} p_k^2(x; d\sigma) d\sigma(x), \quad k = 0, 1, 2, \ldots, \tag{5.14}$$

and denoting $D = \text{diag}(d_0, d_1, \ldots, d_{n-1})$, the appropriate matrix norm to be used is then

$$\|A\|_{F,D} = \|D^{-1}AD\|_F \tag{5.15}$$

(which clearly satisfies all the axioms of a matrix norm, including submultiplicativity, $\|AB\|_{F,D} \leq \|A\|_{F,D}\|B\|_{F,D}$). In place of (5.11), one obtains

$$\|V_n^{-1}\|_{F,D} = \left[\int_{\mathbb{R}} \sum_{\nu=1}^{n} \frac{1}{d_{\nu-1}^2} |\ell_\nu(x)|^2 d\sigma(x)\right]^{1/2}, \tag{5.16}$$

and (5.12) must be modified to read

$$\|V_n\|_{F,D} = \left[\sum_{\nu=1}^{n} d_{\nu-1}^2 \sum_{\mu=1}^{n} \frac{1}{d_{\mu-1}^2} |p_{\mu-1}(z_\nu)|^2\right]^{1/2}. \tag{5.17}$$

The condition $\text{cond}_{F,D} V_n$ is then again computable as the product of (5.16) and (5.17), and can be estimated from below by

$$\mathrm{cond}_{F,D} V_n \geq \left[\sum_{\nu=1}^{n} \frac{\gamma_n(z_\nu;d\sigma)}{d_{\nu-1}^2} \sum_{\nu=1}^{n} \frac{d_{\nu-1}^2}{\gamma_n(z_\nu;d\sigma)} \right]^{1/2}. \tag{5.18}$$

For orthonormal polynomials, we have $D = I$, and the results (5.16)–(5.18) reduce to (5.11), (5.12) and (5.8).

VI. VANDERMONDE-LIKE MATRICES INVOLVING CHEBYSHEV POLYNOMIALS ON ELLIPSES

We have noted in Section IV that the powers are "Chebyshev polynomials" (i.e., monic polynomials of minimum uniform norm) on the disc. It is similarly known that the (monic) polynomials

$$p_0(z) = 1, \quad p_k(z) = 2\rho^{k/2} T_k\left[\frac{z}{2\sqrt{\rho}}\right], \quad k = 1,2,\ldots \quad (0 < \rho \leq 1), \tag{6.1}$$

where T_k denotes the Chebyshev polynomial of the first kind, are the "Chebyshev polynomials" on the ellipse $\mathcal{E}_\rho$ with boundary given by

$$\partial\mathcal{E}_\rho = \{z: z = e^{i\theta} + \rho e^{-i\theta},\ 0 \leq \theta \leq 2\pi\} \tag{6.2}$$

if $0 < \rho < 1$, and on the interval $[-2,2]$ (the limit of (6.2) as $\rho \to 1$), when $\rho = 1$; cf. [17]. (The ellipse $\mathcal{E}_\rho$ is scaled so as to have capacity 1.) This suggests the study of Vandermonde-like matrices (1.8) with polynomials p_k given by (6.1) and nodes located on the elliptic contour (6.2), either in quasi-cyclic order, or in the order determined by the Van der Corput sequence. Thus, in the former case, with ν given as in (4.2), and assuming $0 < \rho < 1$,

$$z_1 = 1 + \rho, \quad z_\nu = e^{i\theta_\nu} + \rho e^{-i\theta_\nu}, \quad \theta_\nu = 2\pi(2(\nu - 2^{k-1}) - 1)/2^k, \tag{6.3}$$

and in the latter case,

$$z_\nu = e^{2\pi i c_{\nu-1}} + \rho e^{-2\pi i c_{\nu-1}}, \tag{6.4}$$

where $\{c_\nu\}_{\nu=0}^{\infty}$ is the Van der Corput sequence (4.4), (4.5). In the limit case $\rho=1$, these formulae have to be slightly modified, since we do not want to run back and forth through the interval $[-2,2]$. We then assume, in the quasi-cyclic case,

$$1 + 2^{k-1} < \nu \leq 2^k + 1 \quad (k \geq 1), \tag{6.5}$$

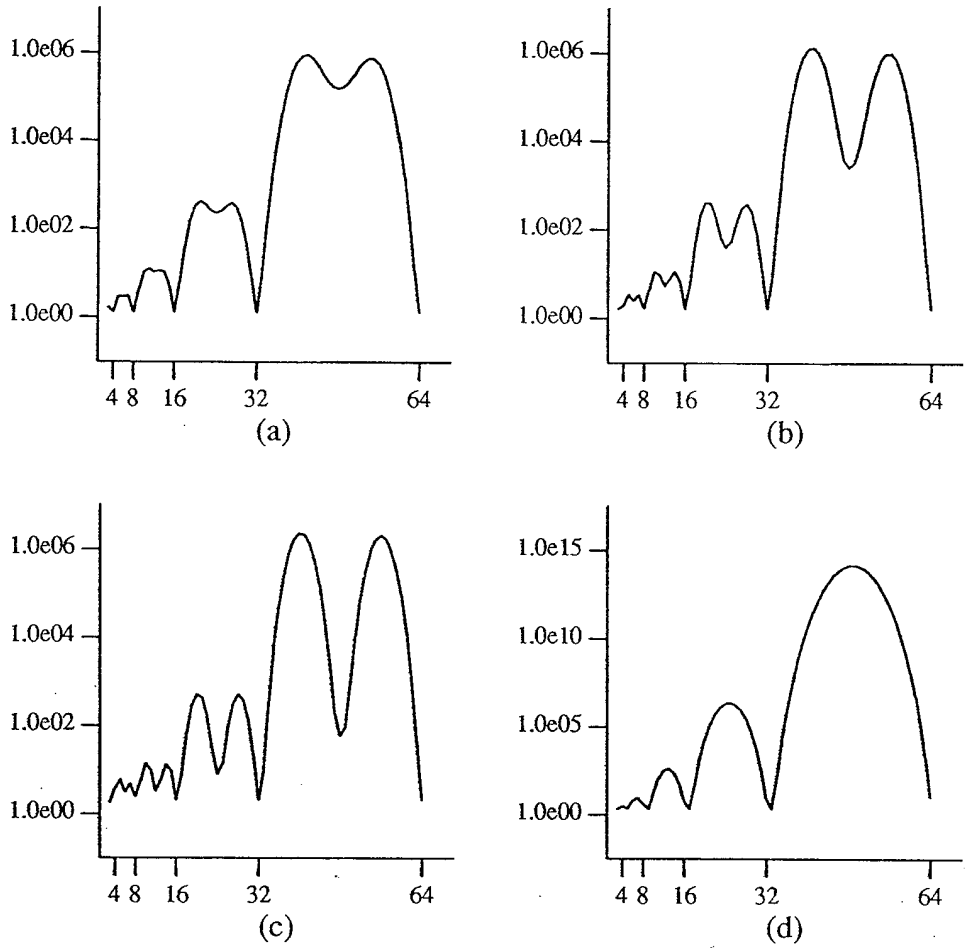


FIG. 6.1 *The condition of Vandermonde-like matrices* (1.8) *for* n=3(1)64 *involving "Chebyshev polynomials" on the ellipse* $\mathcal{E}_\rho$ *and nodes taken from the quasi-cyclic sequence* (6.3) *resp.* (6.3_1).
(a) $\rho = .25$ (b) $\rho = .5$ (c) $\rho = .75$ (d) $\rho = 1$.

and define

$$z_1 = -2, \quad z_2 = 2, \quad z_\nu = 2\cos \pi(2(\nu - 2^{k-1}) - 3)/2^k, \quad \nu = 3,4, \ldots ; \qquad (6.3_1)$$

in the case of the Van der Corput sequence, we let

$$z_1 = -2, \quad z_{\nu+1} = 2\cos(\pi c_{\nu-1}), \quad \nu = 1,2,3, \ldots . \qquad (6.4_1)$$

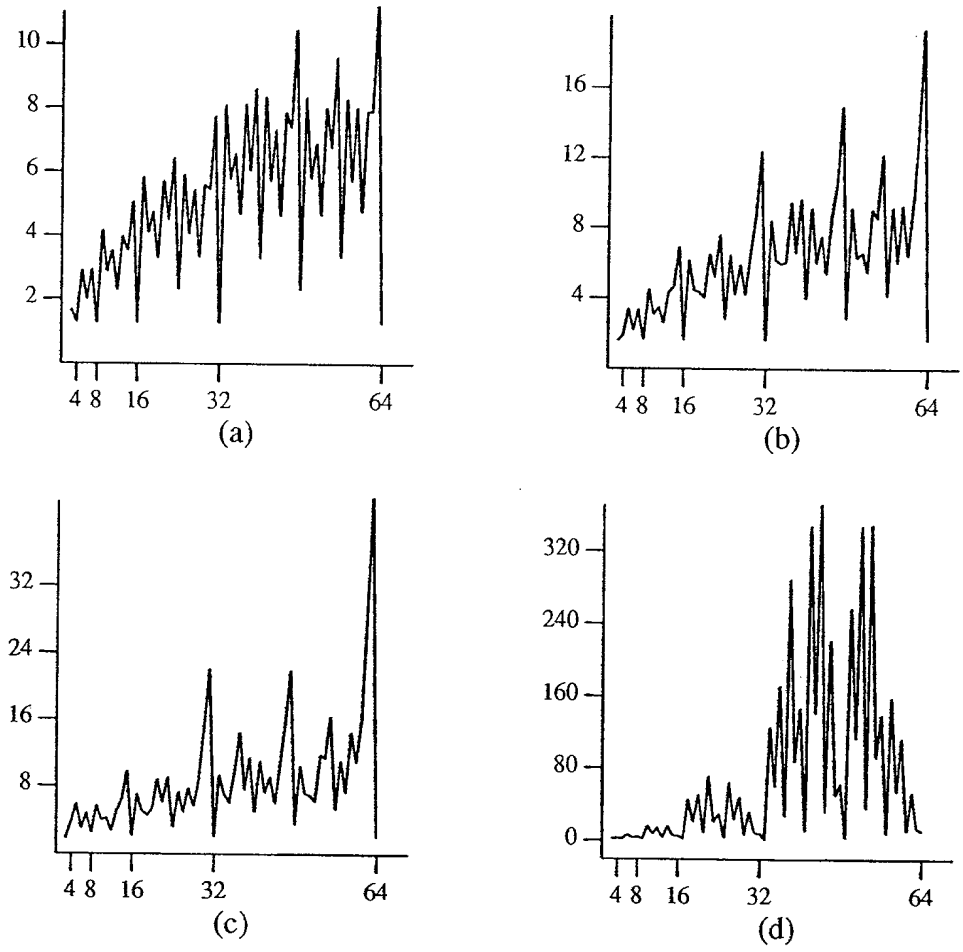


FIG. 6.2 *The condition of Vandermonde-like matrices* (1.8) *for* n=3(1)64 *involving "Chebyshev polynomials" on the ellipse* $\mathcal{E}_\rho$ *and nodes taken from the Van der Corput sequence* (6.4) *resp.* (6.4_1).
(a) $\rho = .25$ (b) $\rho = .5$ (c) $\rho = .75$ (d) $\rho = 1$.

We computed $\text{cond}_2 V_n$ for $3 \le n \le 64$ in the case of ellipses $\mathcal{E}_\rho$ with ρ = .25(.25).75 (for the case ρ=0, see Section IV), and for the segment [–2,2] (i.e., the case ρ=1). The results are shown graphically, on a logarithmic scale, in Figure 6.1 for nodes given by (6.3) [resp. (6.3_1)], and in Figure 6.2, on a linear scale, for the nodes (6.4) [resp. (6.4_1)]. As can be seen, Van der Corput sequences again perform significantly better than quasi-cyclic sequences. The condition, in fact, is known to grow at most polynomially in n, if $0 \le \rho < 1$, and at most like $n^{O(\log n)}$ if $\rho = 1$; cf. [17, Section 3]. In the case of quasi-cyclic sequences, when $0 < \rho < 1$, it is interesting to observe two large peaks between successive powers of 2, in contrast to the cases ρ=0 and ρ=1, which exhibit only one (a surprisingly large one when ρ=1).

ACKNOWLEDGMENT. The author is indebted to Professor Lothar Reichel for useful comments.

REFERENCES

1. Björck, A. and Elfving, T.: Algorithms for confluent Vandermonde systems, *Numer. Math.*, v. 21, 1973, pp. 130–137.

2. Córdova, A., Gautschi, W. and Ruscheweyh, S.: Vandermonde matrices on the circle: spectral properties and conditioning, submitted for publication.

3. Dahlquist, G. and Björck, Å.: *Numerical Methods*, Prentice-Hall, Englewood Cliffs, N.J., 1974.

4. Eiermann, M., Niethammer, W. and Varga, R.S.: A study of semiiterative methods for nonsymmetric systems of linear equations, *Numer. Math.*, v. 47, 1985, pp. 505–533.

5. Fischer, B. and Reichel, L.: A stable Richardson iteration method for complex linear systems, *Numer. Math.*, v. 54, 1988, pp. 225–242.

6. _____ and _____: Newton interpolation in Fejér and Chebyshev points, *Math. Comp.*, v. 53, 1989, to appear.

7. Gautschi, W.: On inverses of Vandermonde and confluent Vandermonde matrices, *Numer. Math.*, v. 4, 1962, pp. 117–123.

8. _____: Norm estimates for inverses of Vandermonde matrices, *Numer. Math.*, v. 23, 1975, pp. 337–347.

9. _____: Optimally conditioned Vandermonde matrices, *Numer. Math.*, v. 24, 1975, pp. 1–12.

10. _____: On inverses of Vandermonde and confluent Vandermonde matrices III, *Numer. Math.*, v. 29, 1978, pp. 445–450.

11. _____: The condition of polynomials in power form, *Math. Comp.*, v. 33, 1979, pp. 343–352.

12. _____: The condition of Vandermonde-like matrices involving orthogonal polynomials, *Linear Algebra Appl.*, v. 52/53, 1983, pp. 293–300.

13. _____: On some orthogonal polynomials of interest in theoretical chemistry, *BIT*, v. 24, 1984, pp. 473–483.

14. _____ and Inglese, G.: Lower bounds for the condition number of Vandermonde matrices, *Numer. Math.*, v. 52, 1988, pp. 241–250.

15. Opfer, G.: personal communication.

16. Peherstorfer, F.: On Gauss quadrature formulas with equal weights, *Numer. Math.*, v. 52, 1988, pp. 317–327.

17. Reichel, L. and Opfer G.: Chebyshev-Vandermonde systems, *Math. Comp.*, to appear.

8.12. [120] "Vandermonde Matrices on the Circle: Spectral Properties and Conditioning"

---

[120] (with A. Córdova and S. Ruscheweyh) "Vandermonde Matrices on the Circle: Spectral Properties and Conditioning," *Numer. Math.* **57**, 577–591 (1990).

Numer. Math. 57, 577–591 (1990)

Numerische Mathematik

Vandermonde Matrices on the Circle: Spectral Properties and Conditioning

Antonio Córdova[1,*], Walter Gautschi[2,], and Stephan Ruscheweyh[3,***]**

[1] Math. Institut, Universität Würzburg, D-8700 Würzburg, Federal Republic of Germany
[2] Department of Computer Sciences, Purdue University, West Lafayette, IN 47907, USA
[3] Departamento de Matemática, Universidad Técnica Federico Santa María, Valparaíso, Chile, and Math. Institut, Universität Würzburg, D-8700 Würzburg, Federal Republic of Germany

Received May 29, 1989

Dedicated to R. S. Varga on the occasion of his sixtieth birthday

Summary. We study Vandermonde matrices whose nodes are given by a Van der Corput sequence on the unit circle. Our primary interest is in the singular values of these matrices and the respective (spectral) condition numbers. Detailed information about multiplicities and eigenvectors, however, is also obtained. Two applications are given to the theory of polynomials.

Subject Classifications: AMS(MOS): 15A12, 15A18; CR: G1.3.

1 Introduction

Vandermonde matrices have a reputation of being ill-conditioned. This reputation is well deserved for Vandermonde matrices whose nodes are all real, in which case the condition number is expected to grow exponentially with the order of the matrix. (Exponential growth of the ∞-condition number has recently been proved [2] in the case of positive nodes, as well as for nodes located symmetrically with respect to the origin. The same is likely to hold for arbitrary real nodes, since the nodes of optimally conditioned Vandermonde matrices are conjectured to have the above symmetry property.) The situation changes drastically, however, if one allows complex nodes. For example, taking the nth roots of unity as nodes in an $(n \times n)$-Vandermonde matrix yields optimal (spectral) condition number 1 for each n [1,

* Research of A.C. supported by the Fundación Andes, Chile, and by the German Academic Exchange Service (DAAD), Federal Republic of Germany
** Research of W.G. supported, in part, by the National Science Foundation, USA, (Grant CCR-8704404)
*** Research of S.R. supported by the Fondo Nacional de Desarollo Cientßfico y Tecnológico (FONDECYT), Chile, (Grant 237/89), by the Universidad Técnica F. Santa Marßa, Valparaßso, Chile, (Grant 89.12.06), and by the German Academic Exchange Service (DAAD), Federal Republic of Germany

Example 6.4]. On the other hand, the roots of unity form a *triangular* array of points, which in applications – for example, polynomial interpolation or quadrature – may be inconvenient inasmuch as it requires that the function to be interpolated, resp. integrated, be evaluated on a two-dimensional array of points. Restricting oneself to *linear* sequences of points, and placing them all on the unit circle, gives rise to the following interesting question: To what extent is the well-conditioning of the respective Vandermonde matrices maintained? Intuitively, one expects well-conditioning if each segment of the sequence is as equally distributed on the circle as possible. A well-known sequence having such a property is the Van der Corput sequence; see e.g. [3, p. 127]. Our objective, then, is to study Vandermonde matrices whose nodes are the initial points of a Van der Corput sequence on the unit circle. The strikingly regular pattern of the singular values of such matrices, as exemplified by Fig. 4.1.1 in [4], has greatly stimulated our interest in this problem.

Given the integer $\nu \geqq 0$ in its binary representation

$$\nu = \sum_{j=0}^{\infty} \nu_j 2^j, \qquad \nu_j \in \{0, 1\},$$

define the fraction

$$c_\nu = \sum_{j=0}^{\infty} \nu_j 2^{-j-1}.$$

The sequence $\{c_\nu\}_{\nu=0}^{\infty}$ is called *Van der Corput sequence*. For $n \in \mathbb{N}$ define the $n \times n$ Vandermonde matrix

$$V_n = \begin{pmatrix} 1 & 1 & \dots & 1 \\ e^{2\pi i c_0} & e^{2\pi i c_1} & \dots & e^{2\pi i c_{n-1}} \\ \vdots & \vdots & \ddots & \vdots \\ e^{2\pi i (n-1) c_0} & e^{2\pi i (n-1) c_1} & \dots & e^{2\pi i (n-1) c_{n-1}} \end{pmatrix}.$$

We are interested in the (spectral) condition number of V_n,

$$\operatorname{cond} V_n = \left[\frac{\lambda_{\max}(S_n)}{\lambda_{\min}(S_n)} \right]^{1/2},$$

where $\lambda_{\max}(S_n)$ and $\lambda_{\min}(S_n)$ denote the largest and smallest eigenvalue of the matrix

$$S_n = V_n V_n^H.$$

In fact, our main result, Theorem 1, does much more: it describes completely all eigenvalues of S_n, and in the proof we even find explicitly all eigenspaces. The matrices S_n have a very interesting structure and may be worth further study. They turn out to be (Hermitian) Toeplitz matrices, namely

$$S_n = \operatorname{Toep}(s_0^{(n)}, \dots, s_{n-1}^{(n)}),$$

where

$$s_\mu^{(n)} = \sum_{\nu=0}^{n-1} e^{-2\pi i \mu c_\nu}, \qquad \mu = 0, 1, \dots. \tag{1.1}$$

To simplify the statement of Theorem 1, we introduce the following *alternating binary representation* of natural numbers n: let

$$2^{k-1} < n \leqq 2^k, \qquad k \in \{0, 1, \dots\}. \tag{1.2}$$

Then there exist a minimal integer $l(n) \geqq 0$ and integers

$$k = n_0 > n_1 > \ldots > n_{l(n)} \geqq 0,$$

all uniquely determined, such that

$$n = \sum_{j=0}^{l(n)} (-1)^j 2^{n_j}. \tag{1.3}$$

Theorem 1. *Let $n \in \mathbb{N}$ have the alternating binary representation* (1.3). *Then the eigenvalues of S_n are*

$$2^{n_j}, \quad j = 0, \ldots, l(n),$$

and the eigenvalue 2^{n_j} has the multiplicity

$$|2^{n_j} - 2(n \bmod 2^{n_j})|.$$

It is readily seen that the sum of the given multiplicities is indeed n. This follows even easier from the following obviously equivalent but less explicit corollary.

Corollary 2. *Let $n \in \mathbb{N}$ satisfy* (1.2). *If $n = 2^k$, then S_n has the only eigenvalue 2^k. If $n < 2^k$, set $n' = 2^k - n$. Then the following holds:*

i) *All eigenvalues of S_n are $\leqq 2^k$.*
ii) *2^k is eigenvalue of multiplicity $n - n'$ of S_n.*
iii) *All eigenvalues of $S_{n'}$ are also eigenvalues of S_n, with the same multiplicities.*

Another immediate consequence of Theorem 1 is

Corollary 3. *For each $n \in \mathbb{N}$ we have* cond $V_n < \sqrt{2n}$.

We are proving Theorem 1 not directly for S_n but for a similar real Toeplitz matrix T_n, using an explicit inductive construction of all eigenspaces. The vectors of a basis of these eigenspaces are given through the coefficients of certain polynomials with integer coefficients which come from an interesting unconventional three-term recurrence relation. These polynomials (of even degree) may deserve some independent interest because of the property described in Theorem 4, which is a consequence of our proof of Theorem 1. They are given as follows: let $e_1(z) \equiv 1$, $e_3(z) = 1 + z + z^2$, and for $k, l \in \mathbb{N}$, l odd, let

$$e_{2^k+l}(z) = \begin{cases} (1-2^k) e_l(z)(1+z^{2^k}) + z^l e_{2^k-l}(z), & 1 \leqq l < 2^{k-1}, \\ e_l(z)(1+z^{2^k}) + z^l e_{2^k-l}(z), & 2^{k-1} < l < 2^k. \end{cases} \tag{1.4}$$

Theorem 4. *The polynomials e_n (of degree $n-1$), n odd, have all their zeros on $|z| = 1$.*

Finally, we mention another application of the results in Theorem 1 to the theory of polynomials.

Theorem 5. *Let P be a polynomial of degree m, m even. Then we have*

$$\max_{|z| \leqq 1} \left| \sum_{j=0}^{m} P(e^{-2\pi i c_j} z) - P(0) \right| \leqq m \max_{|z| \leqq 1} |P(z)|.$$

This estimate is sharp for $P \equiv$ const.

2 A Generating Function for the $s_\mu^{(n)}$

We start with some elementary observations concerning the sequence c_ν.

Lemma 2.1. *The c_ν satisfy*

$$c_{2^{k-1}+\nu} = 2^{-k} + c_\nu, \quad 0 \leqq \nu < 2^{k-1}, \quad k \in \mathbb{N}, \tag{2.1}$$

and

$$\{2^k c_\nu, \nu = 0, \ldots, 2^{k-1}-1\} = \{2j : 0 \leqq j \leqq 2^{k-1}-1\},$$

$$\{2^k c_\nu, \nu = 2^{k-1}, \ldots, 2^k - 1\} = \{2j+1 : 0 \leqq j \leqq 2^{k-1}-1\}.$$

We omit the simple details of the proof. For $k, l \in \mathbb{N}$, $0 \leqq l < 2^k$, let

$$P_{2^k,l}(z) := \prod_{j=2^k-l}^{2^k-1} (1 - e^{-2\pi i c_j} z), \tag{2.2}$$

with the convention $P_{2^k,0} \equiv 1$.

Lemma 2.2. *Let k, l be as above and set*

$$l = \sum_{j=0}^{k-1} d_j 2^j, \quad d_j \in \{0, 1\}. \tag{2.3}$$

Then

$$P_{2^k,l}(e^{2\pi i c_{2^{k-1}-l}} z) = \prod_{j=0}^{k-1} (1 + d_j z^{2^j}). \tag{2.4}$$

Proof. We first prove the following recursion:

$$P_{2^k,l}(z) = P_{2^{k-1},l-d_{k-1}2^{k-1}}(z e^{-2\pi i(1-d_{k-1})2^{-k}})(1 + d_{k-1} z^{2^{k-1}}). \tag{2.5}$$

If $d_{k-1} = 0$, we use the index transformation $j \to 2^{k-1} + j$ in (2.2), and (2.1), to derive (2.5). If $d_{k-1} = 1$, then

$$\begin{aligned} P_{2^k,l}(z) &= \prod_{j=2^k-l}^{2^{k-1}-1} (1 - e^{-2\pi i c_j} z) \prod_{j=2^{k-1}}^{2^k-1} (1 - e^{-2\pi i c_j} z) \\ &= P_{2^{k-1},l-2^{k-1}}(z) \prod_{j=2^{k-1}}^{2^k-1} (1 - e^{-2\pi i c_j} z). \end{aligned}$$

To identify the second factor, we use the last formula of Lemma 2.1, and obtain

$$\begin{aligned} \prod_{j=2^{k-1}}^{2^k-1} (1 - e^{-2\pi i c_j} z) &= \prod_{j=1}^{2^{k-1}} \left(1 - e^{-2\pi i \frac{j}{2^{k-1}}} (e^{2i\pi 2^{-k}} z)\right) \\ &= 1 - (e^{2i\pi 2^{-k}} z)^{2^{k-1}} \\ &= 1 + d_{k-1} z^{2^{k-1}}. \end{aligned}$$

Repeated application of (2.5) yields

$$P_{2^k,l}(z) = \prod_{j=0}^{k-1} \left(1 + d_j \left(e^{-2\pi i \sum_{s=j+1}^{k-1} \frac{1-d_s}{2^{s+1}}} z\right)^{2^j}\right).$$

Hence, with

$$w = e^{2\pi i \sum_{s=0}^{k-1} \frac{1-d_s}{2^{s+1}}} = e^{2\pi i c_{2^k-1}},$$

we get

$$P_{2^k,1}(wz) = \prod_{j=0}^{k-1} \left(1 + d_j \left(e^{2\pi i \sum_{s=0}^{j} \frac{1-d_s}{2^{s+1}}} z\right)^{2^j}\right)$$

$$= \prod_{j=0}^{k-1} (1 + d_j e^{i\pi(1-d_j)} z^{2^j}),$$

the assertion. □

In the sequel we shall use the following notation: if f, g are analytic functions in $z=0$, we shall write

$$f \overset{n}{\sim} g$$

if $f-g$ has an n-fold zero in $z=0$.

Lemma 2.3. *Let $n \in \mathbb{N}$ satisfy $n \leqq 2^k$, $k \in \mathbb{N}$. If*

$$2^k - n = \sum_{j=0}^{k-1} d_j 2^j, \quad d_j \in \{0, 1\},$$

then

$$\sum_{j=1}^{n-1} \frac{s_j^{(n)}}{j} (e^{2\pi i c_{n-1}} z)^j \overset{n}{\sim} \sum_{j=0}^{k-1} \log(1 + d_j z^{2^j}). \tag{2.6}$$

Proof. We set

$$F_n(z) = \log P_{2^k, 2^k-n}(e^{2\pi i c_{n-1}} z),$$

so that

$$F_n'(z) = \sum_{j=n}^{2^k-1} \frac{-e^{-2\pi i (c_j - c_{n-1})}}{1 - e^{-2\pi i (c_j - c_{n-1})} z}$$

and

$$zF_n'(z) - 2^k + n = \sum_{j=n}^{2^k-1} \frac{-1}{1 - e^{-2\pi i (c_j - c_{n-1})} z}. \tag{2.7}$$

Now using the second part of Lemma 2.1, we see that

$$2^k \overset{2^k}{\sim} \sum_{j=0}^{2^k-1} \frac{1}{1 - e^{-2\pi i (c_j - c_{n-1})} z},$$

and thus

$$zF_n'(z) + n \overset{2^k}{\sim} \sum_{j=0}^{n-1} \frac{1}{1 - e^{-2\pi i (c_j - c_{n-1})} z}$$

$$\overset{n}{\sim} \sum_{\mu=0}^{n-1} \left(\sum_{j=0}^{n-1} e^{-2\pi i \mu c_j}\right) (e^{2\pi i c_{n-1}} z)^\mu.$$

This shows that

$$F_n'(z) \overset{n-1}{\sim} \sum_{j=1}^{n-1} s_j^{(n)} e^{2\pi i j c_{n-1}} z^{j-1}, \tag{2.8}$$

and by integration, using Lemma 2.1, we arrive at (2.6). □

We define the following sequences:

$$\sigma_j^{(n)} = s_j^{(n)} e^{2\pi i j c_{n-1}}, \quad j = 0, 1, \ldots, n-1, \quad n \in \mathbb{N},$$

and note from (2.6) that $\sigma_j^{(n)} \in \mathbb{Z}$ for all j, n. It is obvious that the matrices S_n and

$$S_n^* := \mathrm{Toep}\,(\sigma_0^{(n)}, \ldots, \sigma_{n-1}^{(n)})$$

are diagonally similar, and have the same eigenvalues with the same multiplicities (later we shall study a still slightly different set of matrices). For easier reference, we note down a consequence of (2.7), (2.8):

$$2^k - \sum_{j=0}^{n-1} \sigma_j^{(n)} z^j \overset{n}{\sim} \sum_{j=n}^{2^k-1} \frac{1}{1 - e^{-2\pi i (c_j - c_{n-1})} z}. \tag{2.9}$$

Corollary 2.4. *Let $n \in \mathbb{N}$ satisfy $n \leqq 2^k$, $k \in \mathbb{N}$. If $1 \leqq j \leqq n-1$ has the representation $j = 2^\mu (2m-1)$, where $\mu = 0, 1, \ldots$ and $m \in \mathbb{N}$, then $\sigma_0^{(n)} = n$ and*

$$\sigma_j^{(n)} = [(2^k - n) \bmod 2^{\mu+1}] - 2[(2^k - n) \bmod 2^\mu]. \tag{2.10}$$

Proof. From Lemma 2.3 we find that

$$\sigma_j^{(n)} = -\sum_{2^\nu | j} (-d_\nu)^{j 2^{-\nu}} 2^\nu,$$

which yields (2.10) after a simple computation. □

3 Reduction to the Case n Odd

The following is an immediate consequence of Corollary 2.4:

$$\left.\begin{aligned} \sigma_{2j}^{(2n)} &= 2\sigma_j^{(n)} \\ \sigma_{2j+1}^{(2n)} &= 0 \end{aligned}\right\}, \quad 0 \leqq j \leqq n-1.$$

Hence, if

$$(x_0, x_1, \ldots, x_{n-1})^t$$

is an eigenvector for S_n^* and the eigenvalue λ, then the two vectors

$$(x_0, 0, x_1, 0, \ldots, x_{n-1}, 0)^t, \quad (0, x_0, 0, x_1, \ldots, 0, x_{n-1})^t$$

are eigenvectors for S_{2n}^* and the eigenvalue 2λ. It is now easily seen that Theorem 1 holds for $2n$ if it holds for n. In view of this fact, we can restrict ourselves to the case n odd, which from now on will be always assumed.

4 The Largest Eigenvalue 2^k

The right-hand side of (2.9) can be written as

$$\sum_{j=n}^{2^k-1} \frac{1}{1 - e^{-2\pi i (c_j - c_{n-1})} z} = \int_0^{2\pi} \frac{d\mu(\phi)}{1 - e^{i\phi} z},$$

with a positive (discrete) measure μ. Hence,

$$2^k I_n - S_n^* = \left(\int_0^{2\pi} e^{i(l-m)\phi} d\mu(\phi) \right)_{0 \leqq l,m \leqq n-1},$$

which readily implies that $2^k I_n - S_n^*$ is positive semidefinite. This shows that all eigenvalues of S_n^* are $\leqq 2^k$ if

$$2^{k-1} \leqq n < 2^k, \quad k \in \mathbb{N}. \tag{4.1}$$

This proves assertion i) of Corollary 2. From now on we shall assume that n, k satisfy the relation (4.1). We shall use the following general result.

Lemma 4.1. *Let $m, n \in \mathbb{N}$, $m < n$. Let $\lambda_j \in \mathbb{R}$, $\varepsilon_j \in \mathbb{C}$ with $|\varepsilon_j| = 1$, $j = 1, \ldots, m$, and set*

$$\sum_{s=0}^{\infty} a_s z^s = \sum_{j=1}^{m} \frac{\lambda_j}{1 - \varepsilon_j z}.$$

Then $\lambda = 0$ is an eigenvalue of multiplicity at least $n - m$ for the (Hermitian) Toeplitz matrix

$$E = \mathrm{Toep}(a_0, a_1, \ldots, a_{n-1}).$$

Note that the following proof contains a description of $n - m$ linearly independent eigenvectors of E. We adopt the following terminology: a polynomial

$$P(z) = \sum_{j=0}^{n-1} b_j z^j$$

is said to *represent* the vector

$$\mathbf{e} = (b_0, b_1, \ldots, b_{n-1})^t \in \mathbb{C}^n.$$

Sometimes, if we say this, the actual degree of the polynomial may be less than $n-1$. In this case we assume that the missing high components of the vector are set to zero.

Proof of Lemma 4.1. We write

$$P(z) = \prod_{j=1}^{m} (\varepsilon_j - z),$$

and we want to show that the polynomials

$$P_s(z) := z^s P(z), \quad s = 0, \ldots, n-m-1,$$

represent (obviously linearly independent) eigenvectors for E and the eigenvalue zero. We can write

$$E = \sum_{j=1}^{m} \lambda_j \mathrm{Toep}(1, \varepsilon_j, \varepsilon_j^2, \ldots, \varepsilon_j^{n-1}).$$

Hence, it suffices to show that the vectors $\mathbf{e}_s$ represented by P_s are annihilated by every term in this sum, for instance by

$$\mathrm{Toep}(1, \varepsilon_1, \varepsilon_1^2, \ldots, \varepsilon_1^{n-1}).$$

We have

$$P_s(z) = (\varepsilon_1 - z) \sum_{j=0}^{m-1} r_j z^{j+s}.$$

Hence, the vectors $\mathbf{e}_s$ can be written as

$$\mathbf{e}_s = \sum_{j=0}^{m-1} r_j (0,\ldots,0,\underset{\text{row } j+s}{\underset{\uparrow}{\varepsilon_1}},-1,0,\ldots,0)^t,$$

from which the assertion becomes obvious. □

It is clear from (2.9) and Lemma 4.1 that S_n has 2^k as an eigenvalue of multiplicity *at least* $n-n'$ (in the notation of Corollary 2). It should also be noted that

$$z^s \prod_{j=0}^{k-1} (1+d_j z^{2^j}), \quad s=0,\ldots,n-n'-1, \quad n'=\sum_{j=0}^{k-1} d_j 2^j, \tag{4.2}$$

represent eigenvectors of S_n^* and the eigenvalue 2^k, as can be deduced from the proof of Lemma 4.1, formula (2.9), and Lemma 2.2.

To complete the proof of Corollary 2 (and Theorem 1), we only need to prove part iii) of the corollary for n odd. This will be done in the following two sections.

5 The Matrices T_n

Let $n \geqq 3$ be odd, $2^{k-1} < n < 2^k$, and write

$$2^k - n = \sum_{j=0}^{k-2} d_j 2^j, \quad d_j \in \{0,1\},$$

so that

$$n = 2^k + \sum_{j=0}^{k-2} (1-d_j)2^j - (2^{k-1}-1) = 1 + 2^{k-1} + \sum_{j=1}^{k-2} (1-d_j)2^j \quad (\text{since } d_0 = 1).$$

This implies

$$(n \bmod 2^s) + ((2^k - n) \bmod 2^s) = 2^s, \quad 1 \leqq s \leqq k-1. \tag{5.1}$$

Hence, for n odd and $\mu \geqq 1$, we obtain from (2.10) the relation

$$\sigma^{(n)}_{2^\mu(2m-1)} = 2(n \bmod 2^\mu) - (n \bmod 2^{\mu+1}), \quad 2 \leqq 2^\mu(2m-1) \leqq n-1,$$

while $\sigma_j^{(n)} = 1$ for j odd. For reasons of simplicity in the subsequent proofs, we now introduce the matrices

$$T_n = \text{Toep}\,(t_0^{(n)}, t_1^{(n)}, \ldots, t_{n-1}^{(n)}), \quad n \text{ odd},$$

where

$$t_0^{(n)} = n,$$

$$t^{(n)}_{2^\mu(2m-1)} = \begin{cases} (-1)^{\frac{n+1}{2}}, & \mu = 0, \\ 2(n \bmod 2^\mu) - (n \bmod 2^{\mu+1}), & \mu > 0. \end{cases}$$

It is here always understood that the index is in the proper range, and that $m \in \mathbb{N}$. It is immediately clear that the eigenvalues of T_n are identical with those of S_n and S_n^*. We shall now exhibit a number of important properties of the matrices T_n.

As before, we are relating $k=k(n)\in\mathbb{N}$ to the odd natural number n by

$$2^{k-1}<n<2^k,$$

and we are associating with n the two numbers

$$n':=2^k-n, \qquad n'':=n-2^{k-1}.$$

Lemma 5.1. *The following relations hold for* $n\geqq 5$:

$$t^{(n)}_{2^{k-1}+j}=t^{(n)}_j, \qquad j=1,\ldots,n''-1, \tag{5.2}$$

$$t^{(n)}_{2^{k-1}}=t^{(n)}_0-2^k, \tag{5.3}$$

$$t^{(n)}_{2^{k-2}+j}=t^{(n)}_j, \qquad j=1,\ldots,2^{k-2}-1, \tag{5.4}$$

$$t^{(n)}_j=t^{(n)}_{2^{k-1}-j}, \qquad j=1,\ldots,2^{k-1}-1, \tag{5.5}$$

$$t^{(n)}_j=t^{(n'')}_j, \qquad j=1,\ldots,n''-1, \tag{5.6}$$

$$t^{(n)}_j=-t^{(n')}_j, \qquad j=1,\ldots,n'-1. \tag{5.7}$$

Proof. (5.2): For j odd, this is clear. Now, for j even, we have

$$1<j=2^\mu(2m-1)\leqq n-1-2^{k-1}<2^{k-1}-1,$$

so that $1\leqq\mu\leqq k-2$. Hence,

$$2^{k-1}+j=2^\mu(2m-1+2^{k-1-\mu})=2^\mu(2m^*-1), \qquad m^*\in\mathbb{N},$$

and the result follows from the definition of $t^{(n)}_j$.

(5.3): We have $n=(n \bmod 2^k)=2^{k-1}+(n \bmod 2^{k-1})$. Hence,

$$\begin{aligned} t^{(n)}_{2^{k-1}} &= 2(n \bmod 2^{k-1})-(n \bmod 2^k) \\ &= -2^k+2(2^{k-1}+(n \bmod 2^{k-1}))-n \\ &= n-2^k \\ &= t^{(n)}_0-2^k. \end{aligned}$$

(5.4): As (5.2), but now using $1\leqq\mu\leqq k-3$, if $k>3$. If $k=3$, we only have $j=1$, an 'odd' case.

(5.5): The case j odd is again obvious. Let j be even with $1<j=2^\mu(2m-1)\leqq 2^{k-1}-2$, so that $1\leqq\mu\leqq k-2$. Hence,

$$\begin{aligned} 2^{k-1}-j &= 2^\mu(2(2^{k-2-\mu}-m+1)-1) \\ &= 2^\mu(2m^*-1). \end{aligned}$$

Since the number on the left is positive, we deduce $m^*\in\mathbb{N}$, and the result follows.

(5.6): We have

$$\frac{n+1}{2}=\frac{n''+1+2^{k-1}}{2}=\frac{n''+1}{2}+2^{k-2},$$

which gives the result for j odd. Next, for j even, write $1<j=2^{\mu}(2m-1)<n''<2^{k-1}$, so that $1\leqq\mu\leqq k-2$. Hence,

$$(n \bmod 2^{\mu})=(n-2^{k-1}) \bmod 2^{\mu},$$

$$(n \bmod 2^{\mu+1})=(n-2^{k-1}) \bmod 2^{\mu+1},$$

which proves this case.

(5.7): We have

$$\frac{n'+1}{2}=\frac{2^k-n+1}{2}=2^{k-1}+1-\frac{n+1}{2},$$

which shows

$$(-1)^{\frac{n'+1}{2}}=-(-1)^{\frac{n+1}{2}},$$

and hence the assertion for j odd. For j even, we use the definition of $t_j^{(n)}$ and (2.10), (5.1) to obtain ($j=2^{\mu}(2m-1), \mu>0$):

$$\begin{aligned} t_j^{(n)} &= \sigma_j^{(n)} \\ &\overset{(2.10)}{=} ((2^k-n) \bmod 2^{\mu+1})-2((2^k-n) \bmod 2^{\mu}) \\ &= -t_j^{(n')}. \quad \square \end{aligned}$$

To simplify the proof for the next lemma, we introduce the following periodic doubly infinite sequences:

$$a_{j+m2^{k-1}}^{(n)}=t_j^{(n)}, \quad j=1,\ldots,2^{k-1}, \quad m\in\mathbb{Z}, \quad 2^{k-1}<n<2^k,$$

which have the following properties ($n\geqq 5$):

(5.8) $$a_j^{(n)}=a_{-j}^{(n)}, \qquad j\in\mathbb{Z},$$

(5.9) $$a_j^{(n'')}=a_j^{(n)}, \qquad 1\leqq j\leqq 2^{k-1}-1, \quad n''>n',$$

(5.10) $$a_j^{(n')}=-a_j^{(n)}, \quad 1\leqq j\leqq 2^{k-1}-1, \quad n''<n'.$$

To prove (5.8), we assume without loss of generality that $1\leqq j\leqq 2^{k-1}-1$ (periodicity). But then we have by (5.5):

$$a_{-j}^{(n)}=a_{-j+2^{k-1}}^{(n)}=a_j^{(n)}.$$

Property (5.9) has to be proved only for $j=n'',\ldots,2^{k-1}-1$ because of (5.6). The assumption $n''>n'$ implies that $2^{k-2}<n''<2^{k-1}$, and hence that $a_j^{(n'')}$ has period 2^{k-2}. Then, using $1\leqq j-2^{k-2}<2^{k-2}$ in the following application of (5.4), we get

$$a_j^{(n'')}=a_{j-2^{k-2}}^{(n'')} \overset{(5.6)}{=} a_{j-2^{k-2}}^{(n)} \overset{(5.4)}{=} a_j^{(n)}.$$

A similar argument proves (5.10).

We note that by (5.3),

$$a_0^{(n)}=t_0^{(n)}-2^k,$$

and hence, using (5.8),

$$A_n := (a_{j-i}^{(n)})_{0\leqq i,j\leqq n-1}=T_n-2^kI_n.$$

It is convenient to associate with A_n the matrix

$$B_n := (a_{j-i+n''}^{(n)})_{\substack{0\leqq i\leqq n-1\\ 0\leqq j\leqq n'-1}}.$$

The following information about the structure of A_n is basic for the inductive argument in the next section.

Lemma 5.2. *Let $n \geqq 5$. Then we have*

$$A_n = \begin{pmatrix} T_{n''} - 2^{k-1} I_{n''} & M & T_{n''} - 2^{k-1} I_{n''} \\ M^t & -T_{n'} & M^t \\ T_{n''} - 2^{k-1} I_{n''} & M & T_{n''} - 2^{k-1} I_{n''} \end{pmatrix},$$

where

$$M = \begin{cases} B_{n''}, & n' < n'', \\ -B_{n'}^t, & n' > n''. \end{cases}$$

Proof. We write

$$A_n = \begin{pmatrix} M_{11} & M_{12} & M_{13} \\ M_{21} & M_{22} & M_{23} \\ M_{31} & M_{32} & M_{33} \end{pmatrix},$$

where $M_{11}, M_{13}, M_{31}, M_{33}$ and $n'' \times n''$ matrices, $M_{12}, M_{32}, M_{21}^t, M_{23}^t$ are $n'' \times n'$ matrices, and M_{22} is a $n' \times n'$ matrix. Since $n' + n'' = 2^{k-1}$, it follows immediately from the periodicity of $a_j^{(n)}$ and (5.8) that

$$M_{11} = M_{13} = M_{31} = M_{33},$$
$$M_{12} = M_{32} = M_{21}^t = M_{23}^t.$$

Hence, we can confine ourselves to the identification of M_{11}, M_{22}, M_{12}.

As for M_{11}, we have

$$\begin{aligned} M_{11} &= (a_{j-i}^{(n)})_{0 \leqq i,j \leqq n''-1} \\ &= (a_{|j-i|}^{(n)})_{0 \leqq i,j \leqq n''-1} \\ &= \mathrm{Toep}\,(t_0^{(n)} - 2^k, t_1^{(n)}, \ldots, t_{n''-1}^{(n)}) \\ &= \mathrm{Toep}\,(t_0^{(n'')} - 2^{k-1}, t_1^{(n'')}, \ldots, t_{n''-1}^{(n'')}) \qquad \text{(by (5.6))} \\ &= T_{n''} - 2^{k-1} I_{n''}. \end{aligned}$$

For M_{22} we have

$$\begin{aligned} M_{22} &= (a_{j-i}^{(n)})_{0 \leqq i,j \leqq n'-1} \\ &= \mathrm{Toep}\,(t_0^{(n)} - 2^k, t_1^{(n)}, \ldots, t_{n'-1}^{(n)}) \\ &= \mathrm{Toep}\,(-t_0^{(n')}, -t_1^{(n')}, \ldots, -t_{n'-1}^{(n')}) \qquad \text{(by (5.7))} \\ &= -T_{n'}. \end{aligned}$$

Now let $n' < n''$, which implies $2^{k-2} < n'' < 2^{k-1}$. We also observe that $n' = 2^{k-1} - n''$, which means $n' = (n'')'$; likewise, $(n'')'' = \frac{1}{2}(n'' - n')$. We then have

$$M_{12} = (a_{n''+j-i}^{(n)})_{\substack{0 \leqq i \leqq n''-1 \\ 0 \leqq j \leqq n'-1}}, \qquad B_{n''} = \left(a_{\frac{n''-n'}{2}+j-i}^{(n'')}\right)_{\substack{0 \leqq i \leqq n''-1 \\ 0 \leqq j \leqq n'-1}}.$$

If we set $s = n'' + j - i$, then $1 \leqq s \leqq n' + n'' - 1 = 2^{k-1} - 1$, and we have to show

$$a_s^{(n)} = a_{s-\frac{n'+n''}{2}}^{(n'')}, \qquad 1 \leqq s \leqq 2^{k-1} - 1.$$

But $a_j^{(n'')}$ has period $2^{k-2}=\dfrac{n'+n''}{2}$, and we are left with

$$a_s^{(n)}=a_s^{(n'')}, \qquad 1\leqq s\leqq 2^{k-1}-1,$$

which is (5.9). Hence, $M_{12}=B_{n''}$ in this case. A similar argument works for $n'>n''$, where $(n')''=\frac{1}{2}(n'-n'')$ and we have to use (5.10) instead of (5.9). □

6 Proof of Corollary 2

As we saw above, only part iii) of Corollary 2 for n odd remains to be proven. We shall prove the following, more precise statement.

Lemma 6.1. *Let $n\in\mathbb{N}$ be odd. Then the following hold:*

i) *$\lambda=1$ is the smallest eigenvalue of T_n.*

ii) *Every eigenvalue λ of T_n, except for 2^k, is also eigenvalue of $T_{n'}$, with the same multiplicity.*

iii) *To every eigenvalue λ of T_n there exists a polynomial $e_n(\lambda,z)$ of degree $\leqq n-m_n(\lambda)$, where $m_n(\lambda)$ is the multiplicity of λ, such that*

$$z^s e_n(\lambda,z), \qquad 0\leqq s\leqq m_n(\lambda)-1,$$

represents an eigenvector $\mathbf{e}_s^{(n)}(\lambda)$ of T_n for λ.

iv) *These polynomials obey the following recursion:*

$$e_1(1,z)\equiv 1,$$

$$e_3(1,z)=1+z+z^2,$$

$$e_n(\lambda,z)=\begin{cases}(1+z^{2^{k-1}})e_{n''}(\lambda,z)+z^{n''}e_{n'}(\lambda,z), & n''>n',\\ \left(1-\dfrac{2^{k-1}}{\lambda}\right)(1+z^{2^{k-1}})e_{n''}(\lambda,z)+z^{n''}e_{n'}(\lambda,z), & n'>n'',\end{cases} \quad n\geqq 5,\ \lambda<2^k,$$

$$e_n(2^k,z)=\prod_{j=0}^{k-1}\left(1+d_j((-1)^{\frac{n+1}{2}}z)^{2^j}\right), \qquad n'=\sum_{j=0}^{k-1}d_j 2^j.$$

In this formula, if λ does not happen to be an eigenvalue of $T_{n''}$, we set $e_{n''}(\lambda,z)\equiv 0$.

v) *The vectors $\mathbf{e}_s^{(n)}$ satisfy the following relations:*

$$B_n\cdot\mathbf{e}_s^{(n')}(\lambda)=-\lambda\mathbf{e}_s^{(n)}(\lambda), \qquad m_{n'}(\lambda)\neq 0,$$

$$B_n^t\cdot\mathbf{e}_s^{(n)}(\lambda)=(\lambda-2^k)\mathbf{e}_s^{(n')}(\lambda), \qquad m_n(\lambda)\neq 0,$$

for $s=0,1,\ldots,m_n(\lambda)-1$.

Proof. The truth of this lemma is readily checked for $n=1,3$, which permits us to start an induction. Note that i) is contained in ii). Assume we are done up to $n-2$, and let $2^{k-1}<n<2^k$, $k\geqq 3$. We now have to distinguish between two cases.

a) $n''>n'$. In this case we have $2^{k-2}<n''<2^{k-1}$, and hence

$$(n'')'=n'=2^{k-1}-n''.$$

Thus, by assumption ii), applied to n'', every eigenvalue λ of $T_{n'}$ is also eigenvalue of $T_{n''}$, with the same multiplicities (the largest one, 2^{k-1}, of $T_{n''}$ is not among them!).

Hence, for *all* eigenvalues λ of $T_{n'}$ we find from v) (applied to n''),

$$B_{n''}\cdot \mathbf{e}_s^{(n')}(\lambda)=-\lambda \mathbf{e}_s^{(n'')}(\lambda),$$
$$B_{n''}^t\cdot \mathbf{e}_s^{(n'')}(\lambda)=(\lambda-2^{k-1})\mathbf{e}_s^{(n')}(\lambda),$$

$0\leqq s\leqq m_{n'}(\lambda)-1$. Our recurrence iv) says that

$$\mathbf{e}_s^{(n)}(\lambda)=\begin{pmatrix}\mathbf{e}_s^{(n'')}(\lambda)\\ \mathbf{e}_s^{(n')}(\lambda)\\ \mathbf{e}_s^{(n'')}(\lambda)\end{pmatrix},\qquad s=0,\ldots,m_{n'}(\lambda)-1,$$

should be eigenvectors for T_n and λ. By Lemma 5.2 and the assumptions we find that

$$(T_n-2^kI_n)\cdot \mathbf{e}_s^{(n)}(\lambda)=2\begin{pmatrix}(T_{n''}-2^{k-1}I_{n''})\cdot \mathbf{e}_s^{(n'')}(\lambda)\\ B_{n''}^t\cdot \mathbf{e}_s^{(n'')}(\lambda)\\ (T_{n''}-2^{k-1}I_{n''})\cdot \mathbf{e}_s^{(n'')}(\lambda)\end{pmatrix}+\begin{pmatrix}B_{n''}\cdot \mathbf{e}_s^{(n')}(\lambda)\\ -T_{n'}\cdot \mathbf{e}_s^{(n')}(\lambda)\\ B_{n''}\cdot \mathbf{e}_s^{(n')}(\lambda)\end{pmatrix}$$
$$=(\lambda-2^k)\mathbf{e}_s^{(n)}(\lambda),$$

which proves this assertion. Since we do know already that T_n has the eigenvalue 2^k with multiplicity $n-n'$ and the mentioned polynomial $e_n(2^k,z)$ (see (4.2) and recall the sign changes made in the definition of $t_j^{(n)}$), the proof of iv) is completed for the present case a). We need to prove v). We have by Lemma 5.2

$$B_n\cdot \mathbf{e}_s^{(n')}(\lambda)=\begin{pmatrix}B_{n''}\cdot \mathbf{e}_s^{(n')}(\lambda)\\ -T_{n'}\cdot \mathbf{e}_s^{(n')}(\lambda)\\ B_{n''}\cdot \mathbf{e}_s^{(n')}(\lambda)\end{pmatrix}=-\lambda \mathbf{e}_s^{(n)}(\lambda)$$

for all eigenvalues of $T_{n'}$ by the assumptions, and similarly,

$$B_n^t\cdot \mathbf{e}_s^{(n)}(\lambda)=B_{n''}^t\cdot \mathbf{e}_s^{(n'')}(\lambda)-T_{n'}\cdot \mathbf{e}_s^{(n')}(\lambda)+B_{n''}^t\cdot \mathbf{e}_s^{(n'')}(\lambda)$$
$$=[(\lambda-2^{k-1})-\lambda+(\lambda-2^{k-1})]\cdot \mathbf{e}_s^{(n')}(\lambda)$$
$$=(\lambda-2^k)\cdot \mathbf{e}_s^{(n')}(\lambda)$$

for all eigenvalues of $T_{n'}$. For the remaining case $\lambda=2^k$ we note that $\mathbf{e}_s^{(n)}(2^k)$ is eigenvector to the eigenvalue 0 of the symmetric matrix $A_n=T_n-2^kI_n$. This implies by Lemma 5.2 that

$$B_n^t\cdot \mathbf{e}_s^{(n)}(2^k)=0,\qquad 0\leqq s\leqq m_n(2^k)-1,$$

which completes the proof of case a).

b) $n''<n'$. In this case we have $2^{k-2}<n'<2^{k-1}$, and hence

$$(n')'=n''=2^{k-1}-n'.$$

Thus, by assumption ii), applied to n', every eigenvalue λ of $T_{n''}$ is also eigenvalue of $T_{n'}$, with the same multiplicities. However, $T_{n'}$ has in addition the eigenvalue 2^{k-1} of multiplicity $n'-n''$. Here we find from the assumption that

$$B_{n'}\cdot \mathbf{e}_s^{(n'')}(\lambda)=-\lambda \mathbf{e}_s^{(n')}(\lambda),\qquad \lambda<2^{k-1},$$
$$B_{n'}^t\cdot \mathbf{e}_s^{(n')}(\lambda)=(\lambda-2^{k-1})\mathbf{e}_s^{(n'')}(\lambda),\qquad \lambda\leqq 2^{k-1},$$

for $0 \leqq s \leqq m_{n'}(\lambda) - 1$. The claimed eigenvectors for T_n and λ now have the form

$$\mathbf{e}_s^{(n)}(\lambda) = \begin{pmatrix} \left(1 - \frac{2^{k-1}}{\lambda}\right) \mathbf{e}_s^{(n'')}(\lambda) \\ \mathbf{e}_s^{(n')}(\lambda) \\ \left(1 - \frac{2^{k-1}}{\lambda}\right) \mathbf{e}_s^{(n'')}(\lambda) \end{pmatrix}, \quad s = 0, \ldots, m_{n'}(\lambda) - 1,$$

where $\mathbf{e}_s^{(n'')}(2^{k-1}) = 0$. Using Lemma 5.2 and the assumptions, we now find

$$\begin{aligned}(T_n - 2^k I_n) \cdot \mathbf{e}_s^{(n)}(\lambda) &= 2\left(1 - \frac{2^{k-1}}{\lambda}\right) \begin{pmatrix} (T_{n''} - 2^{k-1} I_{n''}) \cdot \mathbf{e}_s^{(n'')}(\lambda) \\ -B_{n'} \cdot \mathbf{e}_s^{(n'')}(\lambda) \\ (T_{n''} - 2^{k-1} I_{n''}) \cdot \mathbf{e}_s^{(n'')}(\lambda) \end{pmatrix} + \begin{pmatrix} -B_{n'}^t \cdot \mathbf{e}_s^{(n')}(\lambda) \\ -T_{n'} \cdot \mathbf{e}_s^{(n')}(\lambda) \\ -B_{n'}^t \cdot \mathbf{e}_s^{(n')}(\lambda) \end{pmatrix} \\ &= (\lambda - 2^k) \mathbf{e}_s^{(n)}(\lambda)\end{aligned}$$

(λ eigenvalue of $T_{n'}$), which proves iv). v) is verified in a similar fashion as in case a). □

It is clear that Lemma 6.1 completes the proof of Corollary 2. By a backward induction $n \to n' \to (n')' \to \ldots$ (check the alternating binary representations of n and n') we deduce easily that Corollary 2 implies Theorem 1, and we are done.

7 Proofs of Theorems 4 and 5

The polynomials $e_n(z)$ as described in the introduction are now immediately identified with the polynomials $e_n(1, z)$, which represent the eigenvector of the simple eigenvalue 1 of T_n, n odd. This eigenvalue is the smallest one of T_n, and thus $T_n - I_n$ is positive semidefinite (with eigenvalue zero), and therefore we must have

$$t_0^{(n)} - 1 + t_1^{(n)} z + \ldots + t_{n-1}^{(n)} z^{n-1} \overset{n}{\sim} \sum_{j=0}^{n-2} \frac{\lambda_j}{1 - \varepsilon_j z},$$

where

$$\lambda_j \geqq 0, \quad |\varepsilon_j| = 1, \quad j = 0, \ldots, n-2,$$

and

$$\sum_{j=0}^{n-2} \lambda_j = n - 1.$$

From Lemma 4.1 we can now deduce that we must have

$$\prod_{j=0}^{n-2} (\varepsilon_j - z) = C e_n(z),$$

since the product on the left represents, up to a factor, the only eigenvector to $T_n - I_n$, exactly as does the polynomial on the right. This proves Theorem 4. Theorem 5 is also a consequence of the fact that $S_{m+1} - I_{m+1}$ is positive semidefinite, using the representation (1.1) of the $s_\mu^{(m+1)}$. We omit the simple details.

Acknowledgements. The second author is indebted to Professor G. Opfer for bringing a preprint of [4] to his attention, for initial computations of the singular values of V_n, and for discussions of the problem at hand. The authors also like to thank Ruth Ruscheweyh, who did some of the computational work necessary to get better insight into the eigenspace structure, and also derived the recursion (1.4) from the numerical data obtained.

References

1. Gautschi, W.: Norm estimates for inverses of Vandermonde matrices. Numer. Math. **23**, 337–347 (1975)
2. Gautschi, W., Inglese, G.: Lower bounds for the condition number of Vandermonde matrices. Numer. Math. **52**, 241–250 (1988)
3. Kuipers, L., Niederreiter, H.: Uniform distribution of sequences. New York: Wiley 1974
4. Reichel, L., Opfer, G.: Chebyshev-Vandermonde systems. Math. Comput. (to appear)

8.13. [200] "Optimally scaled and optimally conditioned Vandermonde and Vandermonde-like matrices"

---

[200] "Optimally scaled and optimally conditioned Vandermonde and Vandermonde-like matrices," *BIT Numer. Math.* **51**, 103–125 (2011).

BIT Numer Math (2011) 51: 103–125
DOI 10.1007/s10543-010-0293-1

BIT

Optimally scaled and optimally conditioned Vandermonde and Vandermonde-like matrices

Walter Gautschi

Received: 1 July 2010 / Accepted: 16 October 2010 / Published online: 6 November 2010

Abstract Vandermonde matrices with real nodes are known to be severely ill-conditioned. We investigate numerically the extent to which the condition number of such matrices can be reduced, either by row-scaling or by optimal configurations of nodes. In the latter case we find empirically the condition of the optimally conditioned $n \times n$ Vandermonde matrix to grow exponentially at a rate slightly less than $(1+\sqrt{2})^n$. Much slower growth—essentially linear—is observed for optimally conditioned Vandermonde-Jacobi matrices. We also comment on the computational challenges involved in determining condition numbers of highly ill-conditioned matrices.

Keywords Singular value decomposition · Condition numbers · Vandermonde matrices · Optimal scaling · Optimal conditioning

Mathematics Subject Classification (2000) 15A18 · 15Bxx · 65F35

1 Introduction

The problem of optimally conditioned Vandermonde matrices,[1] that is, of determining a configuration of real nodes in a Vandermonde matrix that minimizes its condition number, has been addressed by us some time ago [3]. A related problem is that of

[1] As is well known, the attribution to Vandermonde of these matrices is incorrect. While it is true that Vandermonde made notable contributions to the theory of determinants, which he founded, and to the theory of equations and combinatorial analysis, there is no trace of "Vandermonde determinants", let alone "Vandermonde matrices", in any of his four mathematical papers. See Lebesgue [10, p. 207], who suggests that mistaking upper indices for powers may have been the source of this error.

Communicated by Lothar Reichel.

W. Gautschi (✉)
Department of Computer Sciences, Purdue University, West Lafayette, IN 47907-2066, USA
e-mail: wxg@cs.purdue.edu

optimal scaling of a Vandermonde matrix with given real nodes, which seems to have received less attention. There are indeed serious computational challenges brought on by the highly ill-conditioned matrices involved. In the case of condition numbers based on the Frobenius matrix norm, for example, the singular value decomposition of Vandermonde matrices can produce results that, at first sight, look reasonable, but in fact are wrong by many orders of magnitude. An example of this is exhibited in Sect. 2. To obtain reliable answers requires the use of variable-(high)precision computation. Further examples of optimal scaling are presented in Sect. 3, including one that is close to optimal conditioning. In Sect. 4 we first draw attention to, and correct, a small error in our example [3, (5.14)] of an optimally conditioned third-order Vandermonde matrix with real symmetric nodes. We then use high-precision calculations in combination with constrained and unconstrained optimization to compute optimally conditioned Vandermonde matrices of orders n up to 10, with nonnegative nodes in Sect. 4.1, with symmetric nodes in Sect. 4.2, and, for $n \leq 13$, with unconstrained real nodes in Sect. 4.3. In the last case we find numerically the optimal Frobenius condition number to grow exponentially at a rate somewhat less than $(1+\sqrt{2})^n$. In Sect. 5 we give an analogous discussion of the conditioning of Vandermonde-like matrices and show that optimally conditioned Vandermonde-Chebyshev matrices are perfectly conditioned in the Frobenius norm.

The implications of conditioning, based on condition numbers as here defined, to the sensitivity to errors in the data of Vandermonde systems, and its relevance to the numerical stability of specific solution algorithms, will not be considered in this paper. For an interesting discussion of this we refer to the work of N.J. Higham [8] for ordinary Vandermonde systems and [9] for Vandermonde-like systems.

2 Singular value decomposition: an instance of computational deception

In studying condition numbers of optimally scaled Vandermonde matrices, using the Frobenius norm, we had occasion to employ the singular value decomposition to compute

$$\operatorname{cond}_F(\boldsymbol{V}) = \sqrt{\sum_{\nu=1}^{n} \sigma_\nu^2 \sum_{\nu=1}^{n} \sigma_\nu^{-2}}, \tag{1}$$

where σ_ν are the singular values of the matrix $\boldsymbol{V} \in \mathbb{R}^{n\times n}$. In the course of these computations, we observed substantial discrepancies between computed results and those expected from theory, although the patterns of the former, at first sight, seemed rather reasonable. The trouble can be traced to an unreliable behavior in the case of highly ill-conditioned matrices of the singular value decomposition, unless executed in appropriately high precision.

Consider, for example, the Vandermonde matrix

$$\boldsymbol{V}(\boldsymbol{x}) = [v_{\nu,\mu}] \in \mathbb{R}^{n\times n}, \quad v_{\nu,\mu} = x_\mu^{\nu-1}, \tag{2}$$

where $n = 50$, and $x_\mu = 4(n+1-\mu)/n$, $\mu = 1, 2, \ldots, n$. Row-wise scaling, $\boldsymbol{D}\boldsymbol{V}(\boldsymbol{x})$, with $\boldsymbol{D} = \operatorname{diag}(1, R^{-1}, \ldots, R^{-(n-1)})$, when $R = x_1$, produces an equilibrated matrix $\boldsymbol{V}_{\mathrm{sc}}(R)$ (again a Vandermonde matrix), each row having the same ∞-norm

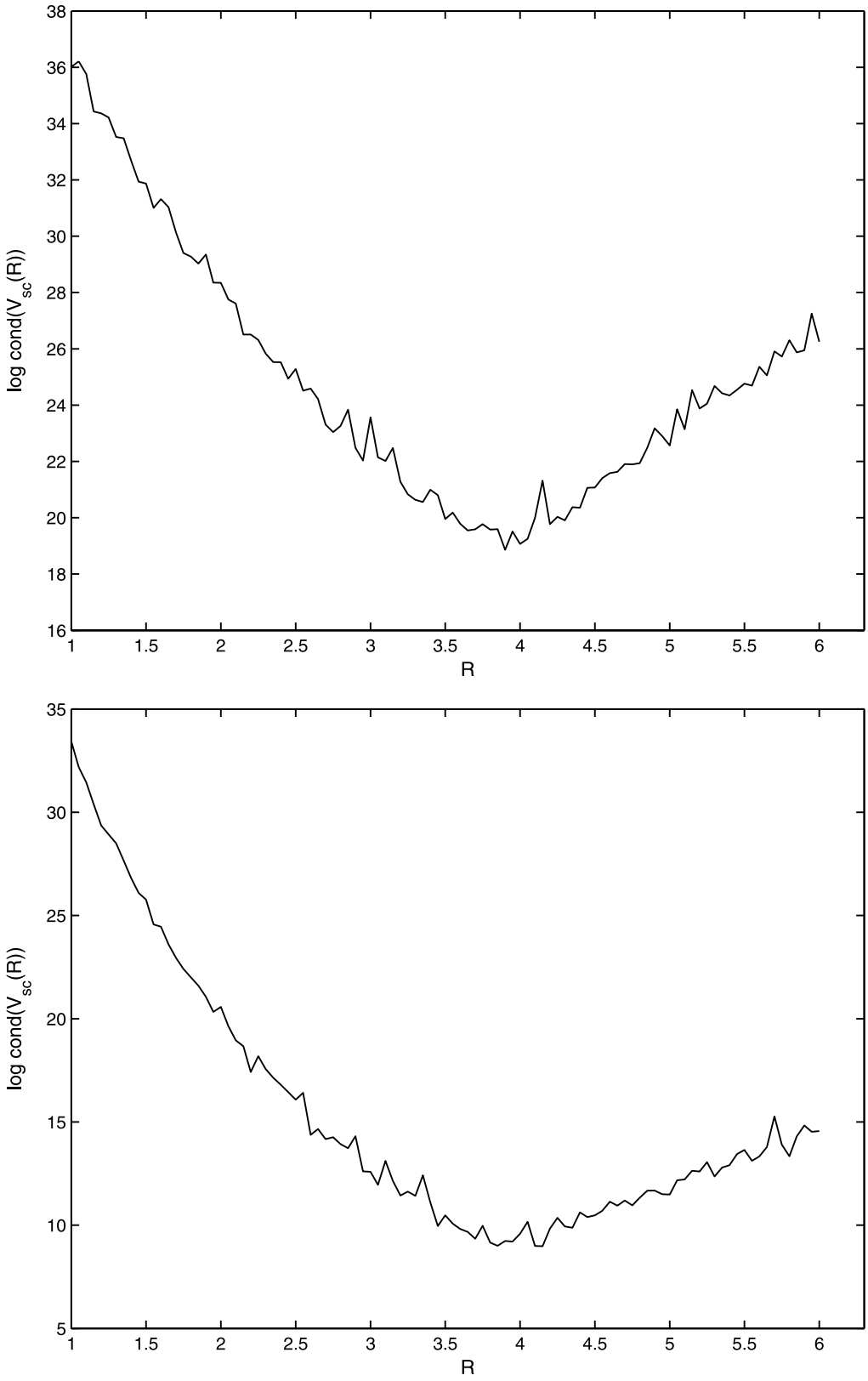


Fig. 1 Frobenius condition of scaled Vandermonde matrices (purported)

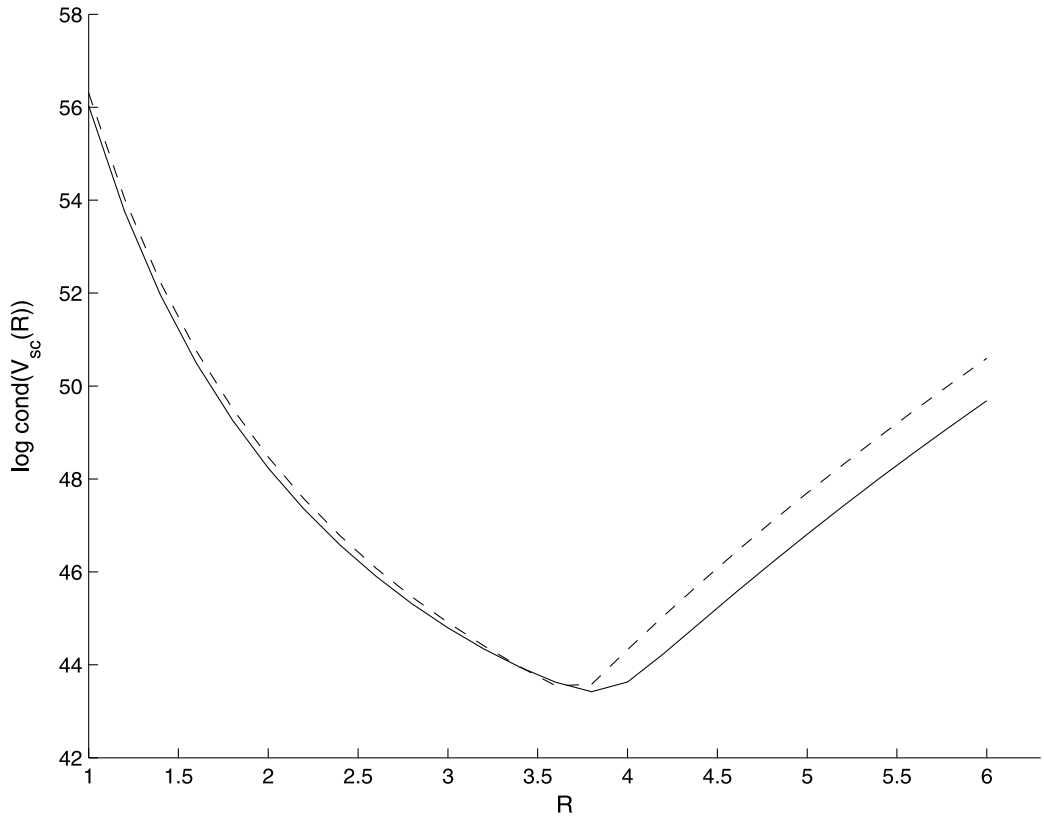


Fig. 2 Condition of scaled Vandermonde matrices (correct) *solid line*: using $\|\cdot\|_F$; *dashed line*: using $\|\cdot\|_\infty$

(= 1). This in turn minimizes the condition(-like) number $\max_{\nu,\mu}|a_{\nu,\mu}|\cdot\|\boldsymbol{A}^{-1}\|^*$, where $\|\cdot\|^*$ denotes any p-norm, or the Frobenius norm, of a matrix [13, Theorem 2.5(b)]. Wanting to illustrate this, we computed the Frobenius condition number of $\boldsymbol{V}_{\rm sc}(R)$, using the Matlab singular-value-decomposition routine `svd.m`, both in ordinary double-precision arithmetic and in symbolic/variable-precision arithmetic (with `digits` = 16), for $R = 1 : .05 : 6$. The results are depicted respectively on the top and bottom of Fig. 1. (The numerical data for the figure on the top are generated by the routine `Fig1a.m`[2], using `VdMsc.m`, and for the figure on the bottom by `sFig1b.m`, using `sVdMsc.m` with the test for complex singular values commented out. The curves themselves are produced respectively by the routines `plotFig1a.m` and `plotFig1b.m`.) Both figures look reasonable, showing a minimum near $R = 4$ $(= x_1)$, as expected, but the graph on the top is higher by 1–10 decimal orders than the one on the right, and both exhibit suspicious wiggles. Which one is more trustworthy?

The plain answer is: neither. The correct graph is shown in Fig. 2, another 20 decimal orders higher! Note that, according to this graph, the condition number of the matrix $\boldsymbol{V}$ itself ($R = 1$) is of the order 10^{56}.

[2]All Matlab routines (other than the standard ones) referred to in this paper can be found on the website http://www.cs.purdue.edu/archives/2002/wxg/codes/OCVdM.html.

How can these enormous discrepancies be explained? We discovered, in the case of the symbolic routine `svd.m`, that quite a few of the singular values computed are complex, in fact purely imaginary, some with rather large imaginary parts. For these imaginary σ_ν, the squares σ_ν^2 and σ_ν^{-2} are real again, but negative, and therefore in (1) have the effect of substantially lowering the Frobenius condition number or even making it complex. In the double-precision routine `svd.m`, all computed singular values appear to be positive, and it is not entirely clear what mechanism it is that caused the substantial underestimation of the condition number in that case.

Why do we think that the solid graph in Fig. 2 is indeed the correct one? First of all, we made sure that all computed singular values σ_ν are positive. This required us to carry out the computation in as much as 112-digit arithmetic. Secondly, the ∞-condition number

$$\mathrm{cond}_\infty(\boldsymbol{V}) = \|\boldsymbol{V}\|_\infty \|\boldsymbol{V}^{-1}\|_\infty \tag{3}$$

of a Vandermonde matrix with nonnegative node vector can be computed directly, without requiring matrix inversion (cf. [2, (4.1)]),

$$\mathrm{cond}_\infty(\boldsymbol{V}_n(\boldsymbol{x})) = \max_{1\le\nu\le n} \sum_{\mu=1}^{n} x_\mu^{\nu-1} \cdot \max_{1\le\nu\le n} \prod_{\substack{\mu=1\\ \mu\ne\nu}}^{n} \frac{1+x_\mu}{|x_\nu - x_\mu|}, \quad \boldsymbol{x} \ge \boldsymbol{0}. \tag{4}$$

(This is implemented in the Matlab routine `condVp.m`.) Using the ∞-norm instead of the Frobenius norm, we obtain the dashed curve in Fig. 2, which indeed is very much within the expected range from the curve for the Frobenius norm. (The routines that produced the numerical data for the two curves are `sFig2a.m` and `Fig2b.m`, the former using the routine `sVdMsc.m` and the latter using the routine `condVsc.m`. The curves themselves are produced by the routine `plotFig2.m`.)

The exact location of the minimum, together with the minimum value, can be determined with the help of the routine `fminbnd.m` of the Matlab Optimization Toolbox. Minimizing the functions defined by the routines `fFig2a.m` and `fFig2b.m`, one finds

$$\min_R \log_{10} \mathrm{cond}_F(\boldsymbol{V}_{\mathrm{sc}}(R)) = 43.416 \quad \text{attained at } R = 3.8340,$$

$$\min_R \log_{10} \mathrm{cond}_\infty(\boldsymbol{V}_{\mathrm{sc}}(R)) = 43.305 \quad \text{attained at } R = 3.7277.$$

Thus, while optimal scaling of the matrix $\boldsymbol{V}(\boldsymbol{x})$ reduces its condition number by more than ten decimal orders, the scaled matrix produced is still very much ill-conditioned. This is a pervasive phenomenon for Vandermonde matrices of this type, since any Vandermonde matrix of order n with nonnegative nodes cannot have an ∞-condition number less than [6, Theorem 2.2]

$$(n-1)\left\{1 + \left(1 - \frac{1}{n}\right)^{-1/(n-1)}\right\}^{n-1}, \tag{5}$$

which for $n = 50$ is $2.786\ldots \times 10^{16}$.

3 Equilibrated Vandermonde matrices

As already mentioned in Sect. 2, equilibration of a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with respect to the infinity vector norm, i.e., scaling it in such a way that all rows have the same infinity norm, has the effect of minimizing a condition(-like) number of the matrix. Similar facts hold for other vector norms. Thus, in the case of the ℓ_1-norm, equilibration minimizes the ∞-condition number $\mathrm{cond}_\infty(\boldsymbol{A}) = \|\boldsymbol{A}\|_\infty \|\boldsymbol{A}^{-1}\|_\infty$ [13, Theorem 2.5(a)], while in the ℓ_2-norm, it approximately minimizes $\mathrm{cond}(\boldsymbol{A}) = \|\boldsymbol{A}\|_2 \|\boldsymbol{A}^{-1}\|^*$ as well as $\mathrm{cond}(\boldsymbol{A}) = \|\boldsymbol{A}\|_F \|\boldsymbol{A}^{-1}\|^*$ within a factor $\sqrt{n}$, where $\|\cdot\|^*$ is any p-norm or the Frobenius norm (*ibid.*, Theorem 3.5(a)). In the case of Vandermonde matrices $\boldsymbol{A} = \boldsymbol{V}$ and the ℓ_p-norm, the scaled matrices, of course, are no longer Vandermonde matrices, unless $p = \infty$. To compute condition numbers, therefore, requires matrix inversion.

Since we are dealing with highly ill-conditioned matrices, we will use the singular value decomposition $\boldsymbol{V} = \boldsymbol{U}_\ell \boldsymbol{\Sigma} \boldsymbol{U}_r^T$ of the matrix $\boldsymbol{V}$, in combination with the computational precautions, in particular variable-(high)precision arithmetic, mentioned in Sect. 2, to obtain

$$\boldsymbol{V}^{-1} = \boldsymbol{U}_r \boldsymbol{\Sigma}^{-1} \boldsymbol{U}_\ell^T, \quad \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n), \quad \boldsymbol{U}_r \text{ and } \boldsymbol{U}_\ell \text{ orthogonal}, \tag{6}$$

where

$$\sigma_1 > \sigma_2 > \cdots > \sigma_n > 0 \tag{7}$$

are the singular values of $\boldsymbol{V}$. In the examples to be presented, we shall display condition numbers consistently in the Frobenius norm, computed in variable-precision arithmetic from (1), and implemented in the Matlab routine `sVdMsc.m`.

We begin with the example discussed in Sect. 2.

Example 3.1 The Vandermonde matrix (2) with

$$x_\mu = 4(n+1-\mu)/n, \quad \mu = 1, 2, \ldots, n. \tag{8}$$

We compute the condition numbers of the equilibrated matrices relative to the ℓ_p-norms for $p = 1, 2, \infty$, display them as a function of n for $n \le 50$ and compare them with the condition number of the original matrix. The results obtained, with the routine `sVdMsc1.m`, in 96-digit arithmetic, are depicted in Fig. 3; cf. Matlab routine `plotFig3_1.m`. It can be seen that equilibration in this case has a notable effect of reducing not only the magnitude of the condition number, but also its rate of growth. Yet, the improved condition is still too high for most applications in practice. It is also evident that the choice of vector norm in equilibration plays a relatively minor role, as the three respective graphs are almost indistinguishable.

Example 3.2 The Vandermonde matrix (2) with

$$x_\mu = \frac{n}{2}\left(-1 + 2\frac{\mu-1}{n-1}\right), \quad \mu = 1, 2, \ldots, n. \tag{9}$$

This is an example of symmetric, and increasingly spread out, nodes. Graphs analogous to those in Example 3.1, but for $n \le 20$, and computed with `sVdMsc2.m` in

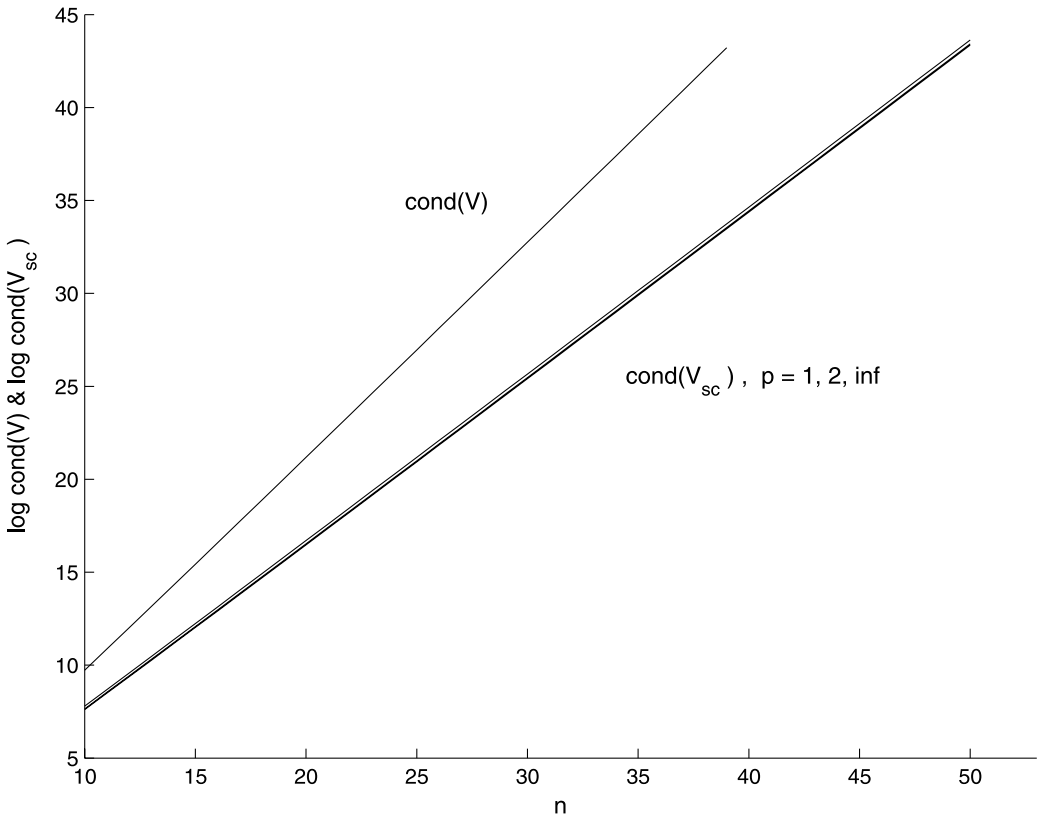


Fig. 3 Frobenius condition numbers of equilibrated and original matrix in Example 3.1

48-digit arithmetic, are shown in Fig. 4; cf. `plotFig3_2.m`. We have a behavior of the condition number of equilibrated matrices vs that of the original matrix which is similar to the one in Fig. 2, but showing markedly superior improvement of conditioning.

Example 3.3 Vandermonde matrix with Chebyshev nodes,

$$x_\mu = \cos\frac{2\mu - 1}{2n}\pi, \quad \mu = 1, 2, \ldots, n. \tag{10}$$

Here the graphs in Fig. 5, produced with `sVdMsc3.m` and `plotFig3_3.m` in 48-digit arithmetic, differ from the preceding graphs in a startling way: the graph for cond($\boldsymbol{V}$) seems to have disappeared! In fact, however, it merged with the other graphs for the equilibrated matrices, being practically identical with them. Equilibration, which, as we know, optimizes the condition of the matrix in one sense or another, has little effect in this case, which means that the matrix is already close to optimally conditioned. We will say more about this in Sect. 4.3.

We recall from [2, (6.5)] that the ∞-condition number of the $n \times n$ Vandermonde matrix with Chebyshev nodes has the asymptotic behavior

$$\mathrm{cond}_\infty(\boldsymbol{V}) \sim \frac{3^{3/4}}{4}(1 + \sqrt{2})^n, \quad n \to \infty, \tag{11}$$

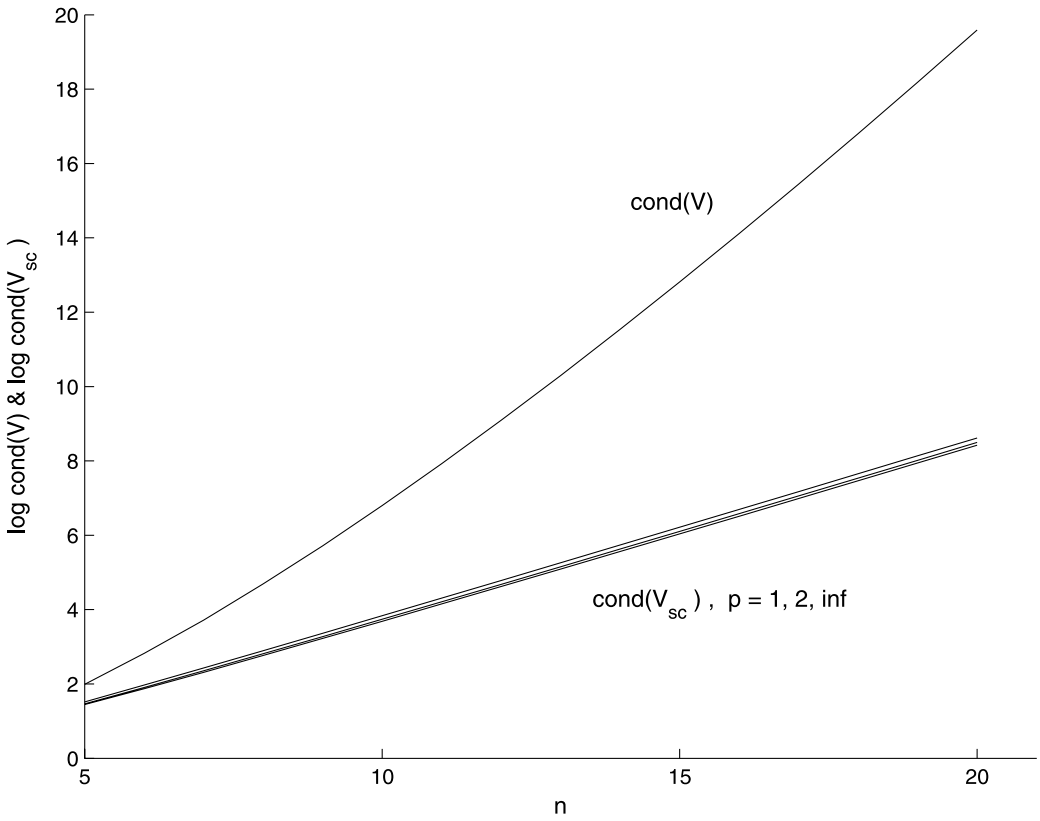


Fig. 4 Frobenius condition numbers of equilibrated and original matrix in Example 3.2

i.e., has exponential growth with rate $(1+\sqrt{2})^n$, which is confirmed in Fig. 5.

It may be worth noting that stretching or shrinking the nodes in (10) (by multiplying them by a constant $a > 1$ resp. $a < 1$) worsens the condition of $\boldsymbol{V}$, but leaves the condition numbers of the scaled matrices unchanged.

Minimality properties of condition numbers similar to those mentioned in Sect. 2 and at the beginning of this section hold also for column equilibration. Thus, the condition(-like) number defined in Sect. 2 is minimized by column equilibration in the ℓ_∞-norm [13, Theorem 2.5(c)], the condition number $\|\boldsymbol{A}\|_1\|\boldsymbol{A}^{-1}\|^*$ (for $\|\cdot\|^*$, see the first paragraph of this section) by column equilibration in the ℓ_1-norm (*ibid.*, Theorem 2.5(d)), and the condition number $\|\boldsymbol{A}\|_2\|\boldsymbol{A}^{-1}\|^*$ or $\|\boldsymbol{A}\|_F\|\boldsymbol{A}^{-1}\|^*$ is minimized approximately by column equilibration in the ℓ_2-norm (*ibid.*, Theorem 3.5(b)).

One could be tempted to equilibrate a matrix twice in succession, first by rows and then by columns, in an attempt to further optimize the condition of the matrix. However, since the first equilibration already minimizes a condition number of sorts, one cannot expect the second one to do more than reduce the condition number by an additional small amount, if at all. This has been confirmed for the Frobenius condition number in the Examples 3.1 and 3.2, where the improvement of the condition by the additional column equilibration is never more than about one decimal order and often much less.

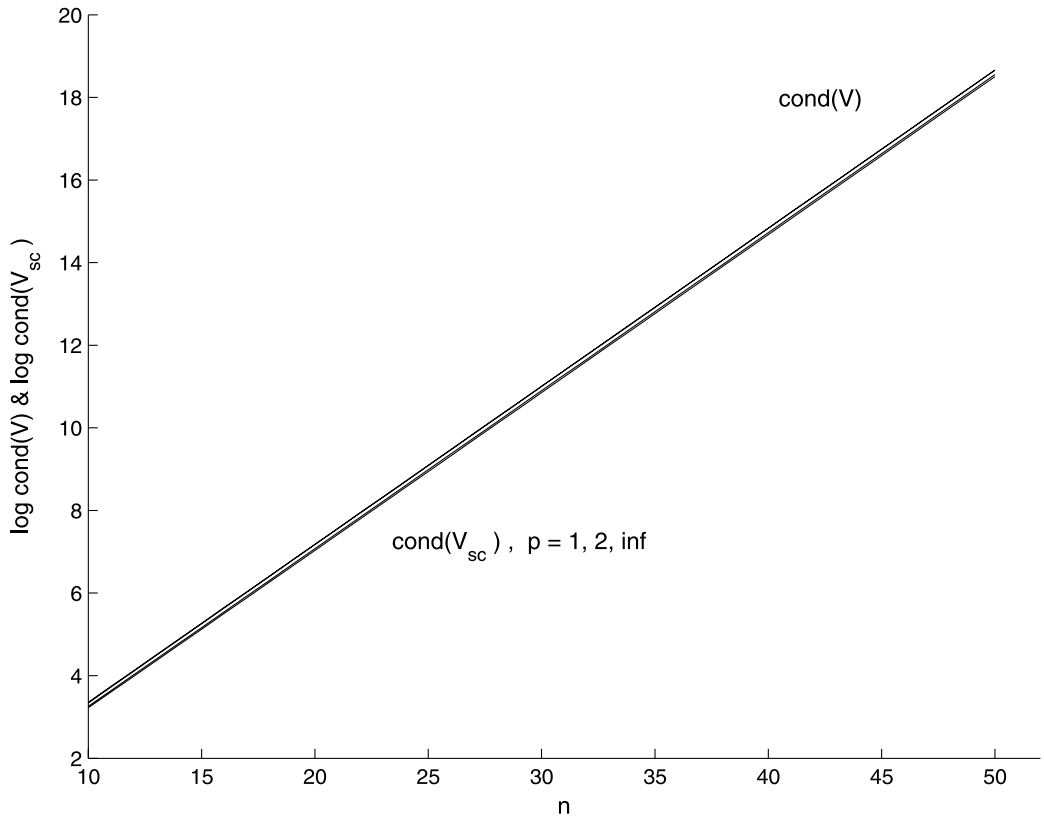


Fig. 5 Frobenius condition numbers of equilibrated and original matrix in Example 3.3

4 Optimally conditioned Vandermonde matrices

The word "optimal", in what follows, is to be understood as meaning "locally optimal". There is no easy way of establishing global optimality, although the possibility of there existing only one optimal point cannot be dismissed entirely.

4.1 Vandermonde matrices with nonnegative nodes

The problem of minimizing the infinity condition number (4) over all nonnegative node configurations

$$x_1 > x_2 > x_3 > \cdots > x_n \geq 0 \tag{12}$$

has been considered in [3, Sect. 5] and solved analytically for $n \leq 3$. For arbitrary n, it was shown (*ibid.*, Theorem 5.3) that the optimal node vector $\boldsymbol{x}^{\text{opt}}$ satisfies $x_n^{\text{opt}} = 0$, and moreover (*ibid.*, Theorem 5.2),

$$\sum_{\nu=1}^{n} (x_\nu^{\text{opt}})^{n-1} = n. \tag{13}$$

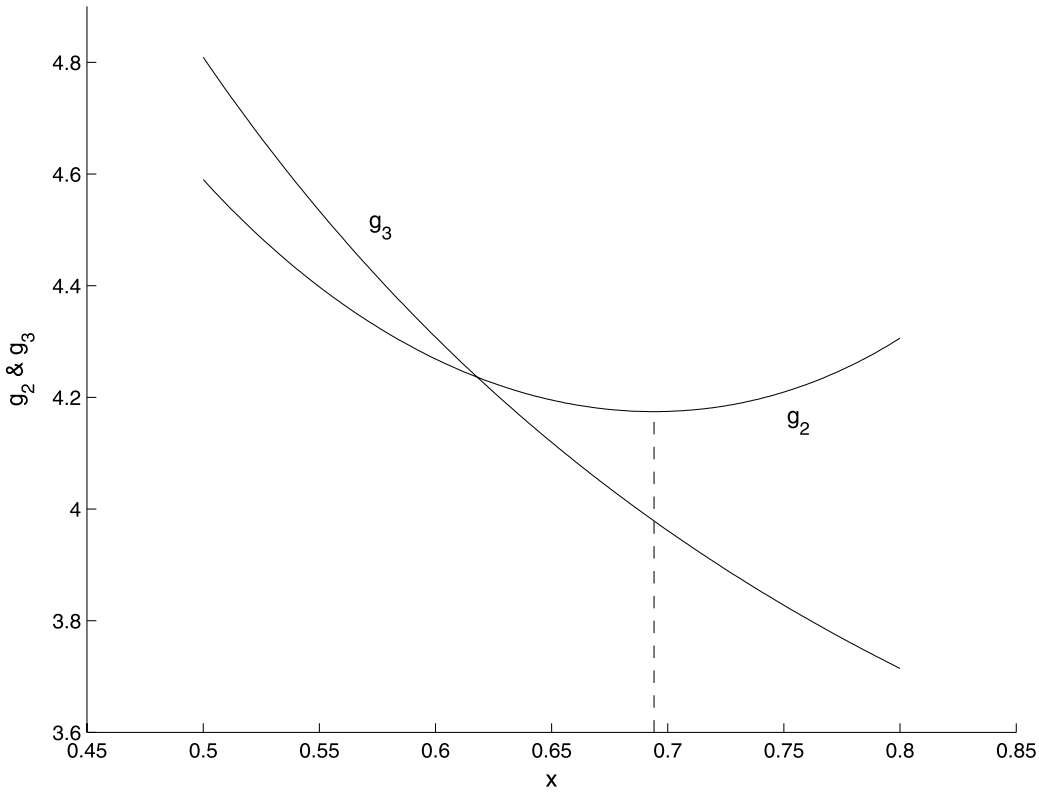


Fig. 6 The graphs of g_2 and g_3

We wish, however, to point out that the analytic solution for $n = 3$ given in the cited reference is slightly in error. The problem, in this case, boils down to solving

$$\max_{x_1>x_2>0}[g_2(x_1, x_2), g_3(x_1, x_2)] = \min$$

subject to $x_1^2 + x_2^2 = 3$, where [3, top of p. 12]

$$g_2(x_1, x_2) = \frac{1 + x_1}{x_2(x_1 - x_2)}, \qquad g_3(x_1, x_2) = \frac{(1 + x_1)(1 + x_2)}{x_1 x_2}.$$

Taking $x = x_2$ as the independent variable, one must solve the one-dimensional minimax problem

$$\max[g_2(\sqrt{3 - x^2}, x), g_3(\sqrt{3 - x^2}, x)] = \min \tag{14}$$

subject to $0 < x < \sqrt{3/2}$. The two graphs for g_2 and g_3 are shown in Fig. 6 in the critical portion of the interval $[0, \sqrt{3/2}]$. In [3], we erroneously assumed that the point of intersection of the two curves yields the minimum point of (14), whereas Fig. 6 shows that the minimum point occurs on the graph for g_2 somewhat to the right of the intersection. By elementary, but tedious, calculations one finds that the correct minimum point is located at the unique positive root of the equation

$$2x^6 - 13x^4 + 6x^3 + 39x^2 - 18 = 0,$$

Table 1 Optimal ∞-condition number and lower bound for nonnegative nodes

n	$\min(\mathrm{cond}_\infty(\boldsymbol{V}p))$	$\mathrm{lb}(\mathrm{cond}_\infty(\boldsymbol{V}p))$
2	3.0000(0)	3.0000(0)
3	1.2524(1)	9.8990(0)
4	7.1725(1)	2.7809(1)
5	3.6102(2)	7.1666(1)
6	1.8844(3)	1.7542(2)
7	9.9422(3)	4.1497(2)
8	5.3511(4)	9.5817(2)
9	2.9009(5)	2.1727(3)
10	1.5997(6)	4.8580(3)

which is x_2^{opt}, and for $\boldsymbol{x}^{\mathrm{opt}}$ yields

$$\begin{aligned} x_1^{\mathrm{opt}} &= 1.586908974551119, \\ x_2^{\mathrm{opt}} &= 0.6940604487284352, \\ x_3^{\mathrm{opt}} &= 0, \end{aligned} \tag{15}$$

giving

$$\mathrm{cond}_\infty(\boldsymbol{V}(\boldsymbol{x}^{\mathrm{opt}})) = 12.52354612417062 \quad (n = 3). \tag{16}$$

To go beyond order 3, we use the routine `fmincon.m` of the Matlab Optimization Toolbox to determine the optimal condition number (4) subject to (12) for $n \le 10$. This is implemented in the routine `optcondVp.m` using `condVp.m`. The required initial approximation $\boldsymbol{x}0$ to $\boldsymbol{x}^{\mathrm{opt}}$ for each n is found by extrapolating the $\boldsymbol{x}^{\mathrm{opt}}$ for the preceding values of n. The results are then checked against the identity (13). If there is insufficient agreement, the initial approximation is improved on the basis of the approximation of $\boldsymbol{x}^{\mathrm{opt}}$ currently at hand, and the routine is run again. This is repeated until sufficient agreement is achieved. A summary of the results so obtained, to five significant digits, is shown in Table 1. The last column, for comparison, lists the lower bound of the condition number according to (5). The respective optimal nodes x_ν^{opt} can be found on the website cited in footnote 2 in the file `xoptVp`.

4.2 Vandermonde matrices with symmetric nodes

If there is a unique (up to permutation of the nodes) optimally conditioned Vandermonde matrix in the ∞-matrix norm, its nodes must be distributed symmetrically with respect to the origin [3, Theorem 3.1],

$$x_1 > x_2 > \cdots > x_n, \qquad x_\nu + x_{n+1-\nu} = 0, \quad \nu = 1, 2, \ldots, n. \tag{17}$$

We therefore turn now our attention to this case of symmetry. It suffices here to consider the reduced vector

$$\boldsymbol{x}s = [x_1, x_2, \ldots, x_{\lfloor (n+1)/2 \rfloor}]^T \ge \boldsymbol{0}, \tag{18}$$

Table 2 Optimal ∞-condition numbers, those for Chebyshev nodes, and lower bounds for symmetric nodes

n	$\min(\mathrm{cond}_\infty(\boldsymbol{V}\boldsymbol{s}))$	$\mathrm{cond}_\infty(\boldsymbol{V}\boldsymbol{C}\boldsymbol{h})$	$\mathrm{lb}(\mathrm{cond}_\infty(\boldsymbol{V}\boldsymbol{s}))$
2	2.0000(0)	2.4142(0)	2.0000(0)
3	5.0000(0)	7.0000(0)	2.8284(0)
4	1.1776(1)	1.8942(1)	4.0000(0)
5	2.1456(1)	4.1000(1)	5.6569(0)
6	5.1330(1)	1.1282(2)	8.0000(0)
7	1.1060(2)	2.5984(2)	1.1314(1)
8	2.4222(2)	6.5152(2)	1.6000(1)
9	5.4541(2)	1.5727(3)	2.2627(1)
10	1.2282(3)	3.7495(3)	3.2000(1)

where $x_{\lfloor (n+1)/2 \rfloor} = 0$ whenever n is odd. In this case the ∞-condition number (3) of $\boldsymbol{V}$ is given explicitly by [2, Theorem 4.3]

$$\mathrm{cond}_\infty(\boldsymbol{V}_n(\boldsymbol{x})) = \max_\nu \sum_\mu x_\mu^{\nu-1} \cdot \begin{cases} \max_\nu [(1+\frac{1}{x_\nu}) \prod_{\mu\neq\nu} \frac{1+x_\mu^2}{|x_\nu^2 - x_\mu^2|}], & n \text{ even}, \\ 2\max_\nu [\varepsilon_\nu (1+x_\nu) \prod_{\mu\neq\nu} \frac{1+x_\mu^2}{|x_\nu^2 - x_\mu^2|}], & n \text{ odd}, \end{cases} \tag{19}$$

where ν and μ vary over all integers for which $x_\nu \geq 0$ and $x_\mu \geq 0$, respectively, and where $\varepsilon_\nu = \frac{1}{2}$ when $x_\nu > 0$, and $\varepsilon_\nu = 1$ when $x_\nu = 0$. (This is implemented in the Matlab routine `condVs.m`.)

The problem of finding the optimal $\boldsymbol{x}s$ has been solved analytically in [3, Sect. 4] for $n \leq 6$. Here we use `fmincon.m`, in `optcondVs.m`, to successfully reproduce these results and to extend them to $n = 10$. The procedure is similar to the one adopted in Sect. 4.1 for nonnegative nodes, the check being provided by the identity [3, Theorem 3.3]

$$\sum_{\nu=1}^{\lfloor (n+1)/2 \rfloor} (x_\nu^{\mathrm{opt}})^{n-1} = \frac{n}{2}. \tag{20}$$

The results are summarized in the second column of Table 2. The third column shows condition numbers for Vandermonde matrices with the Chebyshev nodes (10), and the last column the lower bound $2^{n/2}$ from [6, Theorem 3.1]. Optimal conditioning is now distinctly better than in the case of nonnegative nodes.

The optimal nodes x_ν^{opt} can be found on the website cited in footnote 2 in the file `xoptVs`.

4.3 More on optimal node configurations

We have seen in Example 3.3, and confirmed in Table 2, that Chebyshev nodes are close to optimal. We therefore use them as initial approximations in the unconstrained optimization routine `fminsearch.m` to compute optimally conditioned Vandermonde matrices relative to the Frobenius norm. (There are no longer explicit formulae for the ∞-condition number.) The objective function to be supplied to `fmin-`

Table 3 Optimal Frobenius condition numbers

n	$\min(\text{cond}_F(\boldsymbol{V}))$	n	$\min(\text{cond}_F(\boldsymbol{V}))$
2	2.0000(0)	8	2.9074(2)
3	4.5109(0)	9	6.8528(2)
4	1.0156(1)	10	1.6213(3)
5	2.3101(1)	11	3.8473(3)
6	5.3238(1)	12	9.1502(3)
7	1.2397(2)	13	2.1804(4)

`search.m` is generated in the routine `condV.m`, which uses symbolic/variable-precision tools to evaluate (in the routine `sVdMsc.m`) the Frobenius condition number of a Vandermonde matrix for arbitrary real nodes. The driver routine is `optcondV.m`. We succeeded, with only 16-digit computation, to determine optimally conditioned Vandermonde matrices up to order $n = 13$. (The machine time on our SUN workstation varied from a few seconds for the first few values of n to 55 minutes for $n = 13$. The computation for $n = 14$ failed after about an hour's worth of computing, producing the cryptic error message "integer too large in context". It appears that the message was generated in the symbolic `svd.m` routine.) The resulting condition numbers are shown in Table 3; the optimal nodes are listed in the file `xoptV` on the website cited in footnote 2. They are indeed symmetric with respect to the origin, at least to within the accuracy provided by the routine `fminsearch.m`, and are not much different from those in the file `xoptVs`.

In order to probe into possible alternative extrema, we repeated the computation with the initial approximations deliberately transformed to the interval $[0, 1]$. We found that the routine `fminsearch.m` converged to exactly the same symmetric solutions as obtained previously, and didn't even take any longer. This seems to suggest that for optimal conditioning of Vandermonde matrices we have both uniqueness and symmetry.

It is evident from our computations that the condition numbers of optimally scaled and optimally conditioned Vandermonde matrices grow at an exponential rate with respect to the order n. In the present case of Frobenius-optimal conditioning it appears that the exponential law for the optimal condition number holds not only for large n, but already for $n \geq 2$. Assuming this is the case, we find numerically, since for $n = 2$ the optimal condition number is 2, that

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \text{cond}_F(\boldsymbol{V}_n(\boldsymbol{x})) \approx 2 \times (2.32)^{n-2}. \tag{21}$$

A log-plot of (21) (dashed line) together with actually computed logarithms of the optimal condition numbers (indicated by stars) is shown in Fig. 7 and produced by `plotoptV.m`. The asymptotic law (11) of the ∞-condition number in the case of Chebyshev nodes is shown as a dashdot line. The rate of growth for the optimal condition numbers is seen to be just slightly less than the one, $(1 + \sqrt{2})^n$, for the ∞-condition numbers involving Chebyshev nodes.

For a related result involving the condition number in the Euclidean 2-norm, see also [1, Theorem 4.1].

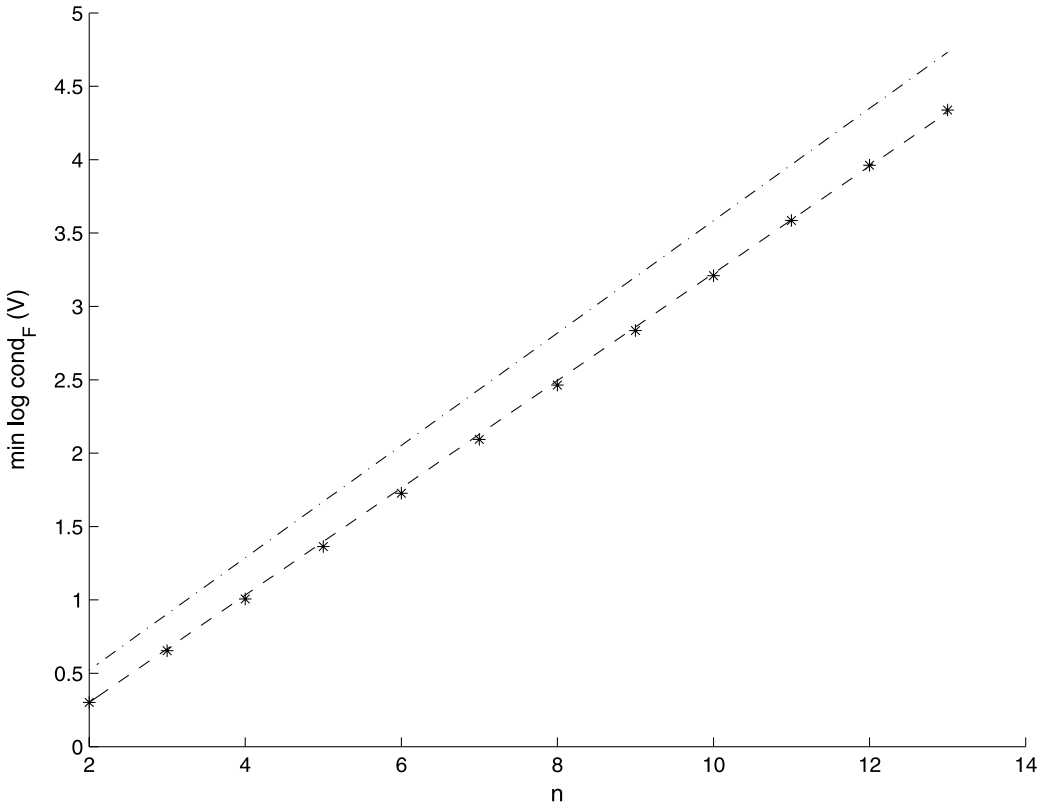


Fig. 7 Optimal Frobenius conditioning of Vandermonde matrices

5 Optimally and perfectly conditioned Vandermonde-like matrices

Vandermonde-like matrices, first considered in [4], are matrices of the form

$$\boldsymbol{W}_n = \boldsymbol{W}_n(\boldsymbol{x}) = \begin{bmatrix} p_0(x_1) & p_0(x_2) & \cdots & p_0(x_n) \\ p_1(x_1) & p_1(x_2) & \cdots & p_1(x_n) \\ p_2(x_1) & p_2(x_2) & \cdots & p_2(x_n) \\ \cdots & \cdots & \cdots & \cdots \\ p_{n-1}(x_1) & p_{n-1}(x_2) & \cdots & p_{n-1}(x_n) \end{bmatrix}, \tag{22}$$

where p_ν, $\nu = 0, 1, \ldots, n-1$, are polynomials of exact degree ν. Thus, the monomials x^ν in a Vandermonde matrix are now replaced by polynomials $p_\nu(x)$. An especially interesting example are polynomials

$$p_\nu(x) = p_\nu(x; \mathrm{d}\lambda), \quad \nu = 0, 1, 2, \ldots, \tag{23}$$

orthogonal with respect to a positive measure $\mathrm{d}\lambda$. Here, row-scaling amounts to renormalizing the orthogonal polynomials. A natural way to do this is to let p_ν in (23) be the ortho*normal* polynomials. We shall henceforth assume this to be the case unless stated otherwise.

If $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T$ is the vector of the zeros of $p_n(\cdot; \mathrm{d}\lambda)$—the "Gauss vector" for $\mathrm{d}\lambda$—then (cf. [4, Theorem 2.1])

$$\mathrm{cond}_F(\boldsymbol{W}_n(\boldsymbol{x})) = \sqrt{\sum_{\nu=1}^{n} \lambda_\nu \sum_{\nu=1}^{n} \frac{1}{\lambda_\nu}}, \tag{24}$$

where $\lambda_\nu = \lambda_\nu^{(n)}$, $\nu = 1, 2, \ldots, n$, are the Christoffel numbers for the measure $\mathrm{d}\lambda$.

5.1 General theory

We begin by noting that (cf. (1))

$$\mathrm{cond}_F(\boldsymbol{W}_n) = \sqrt{\sum_{\nu=1}^{n} \sigma_\nu^2 \sum_{\nu=1}^{n} \sigma_\nu^{-2}} = n\sqrt{\frac{m_A(\sigma^2)}{m_H(\sigma^2)}} \geq n, \tag{25}$$

where $m_A(\sigma^2)$ is the arithmetic mean of the quantities σ_ν^2 and $m_H(\sigma^2)$ their harmonic mean. As is well known, the former is larger than, or equal to, the latter and equal if and only if all singular values σ_ν are equal. Any matrix $\boldsymbol{W}_n$ for which equality holds in (25) is called *perfectly conditioned* with respect to the Frobenius norm.

In the case of Gauss vectors $\boldsymbol{x}$ we have similarly, from (24), that

$$\mathrm{cond}_F(\boldsymbol{W}_n(\boldsymbol{x})) = \sqrt{\sum_{\nu=1}^{n} \lambda_\nu \sum_{\nu=1}^{n} \lambda_\nu^{-1}} = n\sqrt{\frac{m_A(\lambda)}{m_H(\lambda)}} \geq n. \tag{26}$$

If equality holds in (26) for all n, then $\lambda_1^{(n)} = \lambda_2^{(n)} = \cdots = \lambda_n^{(n)}$ for all n, which in turn implies, by a classical result of Posse (cf. [5, Example 1.49]), that $\mathrm{d}\lambda$ must be the Chebyshev measure $\mathrm{d}\lambda(x) = \sqrt{1 - x^2}$ on $[-1, 1]$, and there is no other measure having the same property. Thus, *Vandermonde–Chebyshev matrices $\boldsymbol{W}_n(\boldsymbol{x})$ with Gauss–Chebyshev vectors $\boldsymbol{x}$ are the only Vandermonde-like matrices that for all n are perfectly conditioned with respect to the Frobenius norm.*

Orthonormalization is a major step toward optimizing the conditioning of $\boldsymbol{W}_n$. Also, it allows us to express the Frobenius condition number, for arbitrary vector $\boldsymbol{x}$, explicitly in the particularly simple form [4, (4.4) and (4.6)]

$$c := \mathrm{cond}_F(\boldsymbol{W}_n(\boldsymbol{x})) = \sqrt{\sum_{\nu=1}^{n} \sum_{k=0}^{n-1} p_k^2(x_\nu) \cdot \int_{\mathbb{R}} \sum_{\nu=1}^{n} \ell_\nu^2(t; \boldsymbol{x}) \mathrm{d}\lambda(t)}, \tag{27}$$

where ℓ_ν are the elementary Lagrange interpolation polynomials for the nodes $x_1, x_2, \ldots, x_n$,

$$\ell_\nu(t; \boldsymbol{x}) = \prod_{\substack{\mu=1 \\ \mu \neq \nu}}^{n} \frac{t - x_\mu}{x_\nu - x_\mu}, \quad \nu = 1, 2, \ldots, n.$$

Moreover, the gradient

$$\boldsymbol{g} := \operatorname{grad} c = [g_1, g_2, \ldots, g_n]^T, \quad g_i = \frac{\partial c}{\partial x_i},$$

computes to

$$g_i = \rho \sum_{k=1}^{n-1} p_k(x_i) p_k'(x_i) + \frac{1}{\rho} \int_{\mathbb{R}} \sum_{\nu=1}^{n} \ell_\nu(t; \boldsymbol{x}) \frac{\partial \ell_\nu}{\partial x_i}(t; \boldsymbol{x}) \mathrm{d}\lambda(t), \tag{28}$$

where

$$\frac{\partial \ell_\nu}{\partial x_i}(t; \boldsymbol{x}) = \begin{cases} \frac{1}{x_\nu - x_i} \frac{t - x_\nu}{t - x_i} \ell_\nu(t; \boldsymbol{x}) & \text{if } i \neq \nu, \\ -\ell_\nu(t; \boldsymbol{x}) \sum_{k \neq \nu} \frac{1}{x_\nu - x_k} & \text{if } i = \nu, \end{cases} \tag{29}$$

and

$$\rho = \sqrt{\frac{\int_{\mathbb{R}} \sum_{\nu=1}^{n} \ell_\nu^2(t; \boldsymbol{x}) \mathrm{d}\lambda(t)}{\sum_{\nu=1}^{n} \sum_{k=0}^{n-1} p_k^2(x_\nu)}}. \tag{30}$$

Both c and $\operatorname{grad} c$ can be computed in a straightforward way, the polynomials p_k and their derivatives by the three-term recurrence relation satisfied by orthonormal polynomials, and the integrals in (27) and (28) exactly by n-point Gauss quadrature relative to the measure $\mathrm{d}\lambda$. This is implemented in the routine `condVl.m`. Using this routine as input to the Matlab optimization routine `fminunc.m` allows us to compute optimally conditioned Vandermonde-like matrices; see the routine `opt-condVl.m`.

If $\boldsymbol{x}$ is the Gauss vector for $\mathrm{d}\lambda$, the formula for g_i can be simplified by noting that the integrand in (28) is a polynomial of degree $2n - 1$, so that n-point Gauss quadrature (for the measure $\mathrm{d}\lambda$) gives exactly

$$\int_{\mathbb{R}} \sum_{\nu=1}^{n} \ell_\nu(t; \boldsymbol{x}) \frac{\partial \ell_\nu}{\partial x_i}(t; \boldsymbol{x}) \mathrm{d}\lambda(t) = \sum_{\nu=1}^{n} \lambda_\nu \frac{\partial \ell_\nu}{\partial x_i}(x_\nu; \boldsymbol{x}) = -\lambda_i \sum_{\substack{k=1 \\ k \neq i}}^{n} \frac{1}{x_i - x_k},$$

where the last equality follows from (29). Likewise, using the well-known formula (see, e.g., [12, (3.4.8)])

$$\sum_{k=0}^{n-1} p_k^2(x_\nu) = \lambda_\nu^{-1},$$

one finds

$$\rho = \sqrt{\frac{\sum_{\nu=1}^{n} \lambda_\nu}{\sum_{\nu=1}^{n} \lambda_\nu^{-1}}}. \tag{31}$$

Table 4 Frobenius condition numbers of Vandermonde–Legendre matrices with monic and normalized polynomials, and optimal condition numbers

n	Monic	Normalized	Optimal	$\|g\|_\infty$	M'time
5	2.0433(01)	5.3624(00)	5.2296(00)	7.5011(–8)	.4
10	6.7459(02)	1.1550(01)	1.1018(01)	2.6139(–4)	5
20	7.0071(05)	2.4941(01)	2.3468(01)	1.5052(–3)	930
35	2.3097(10)	4.6321(01)	4.3348(01)	2.9835(–3)	1,061
50	7.5861(14)	6.8600(01)	6.4128(01)	–	–

Therefore,

$$g_i = \rho \sum_{k=1}^{n-1} p_k(x_i)\, p_k'(x_i) - \frac{\lambda_i}{\rho} \sum_{\substack{k=1 \\ k\neq i}}^{n} \frac{1}{x_i - x_k} \quad (\boldsymbol{x} = \text{Gauss vector for d}\lambda) \tag{32}$$

with ρ as given by (31).

5.2 Examples

Example 5.1 (Vandermonde–Legendre matrices) These are Vandermonde-like matrices with p_ν the Legendre polynomials. We first compare Frobenius condition numbers when the Legendre polynomials are monic with those for normalized polynomials. In both cases we take $\boldsymbol{x}$ to be the Gauss–Legendre node vector. For selected values of n, the two condition numbers are shown respectively in the second and third column of Table 4. They were computed, with identical results, in two ways, by actual scaling and by the explicit formula (24); see the routine `runVdMlsc.m`. It can be seen that normalization of the polynomials indeed reduces the conditioning dramatically, which is typical for other Jacobi polynomials as well. In the next column we show the optimal condition numbers computed (by `optcondVl.m`) as indicated above. Here the routine `fminunc` was used with, and without, gradient information. For the former, which is considerably slower, we show in the remaining columns the ∞-norm of the gradient[3] upon exiting the optimization routine, and the machine time in seconds (on our SUN workstation) expended to compute the optimal condition number. (When $n = 50$, convergence of the gradient-based routine could not be achieved within a reasonable amount of time.) Figure 8 has a graph of the optimal condition numbers together with the one (dashed line) for normalized Legendre polynomials. It is seen that the condition of Vandermonde–Legendre matrices with normalized polynomials and Gauss–Legendre vector $\boldsymbol{x}$ is nearly optimal.

[3] Another possible indicator of the accuracy is the measure of symmetry $\max_\nu |x_\nu + x_{n+1-\nu}|$ for the computed $\boldsymbol{x}^{\text{opt}}$.

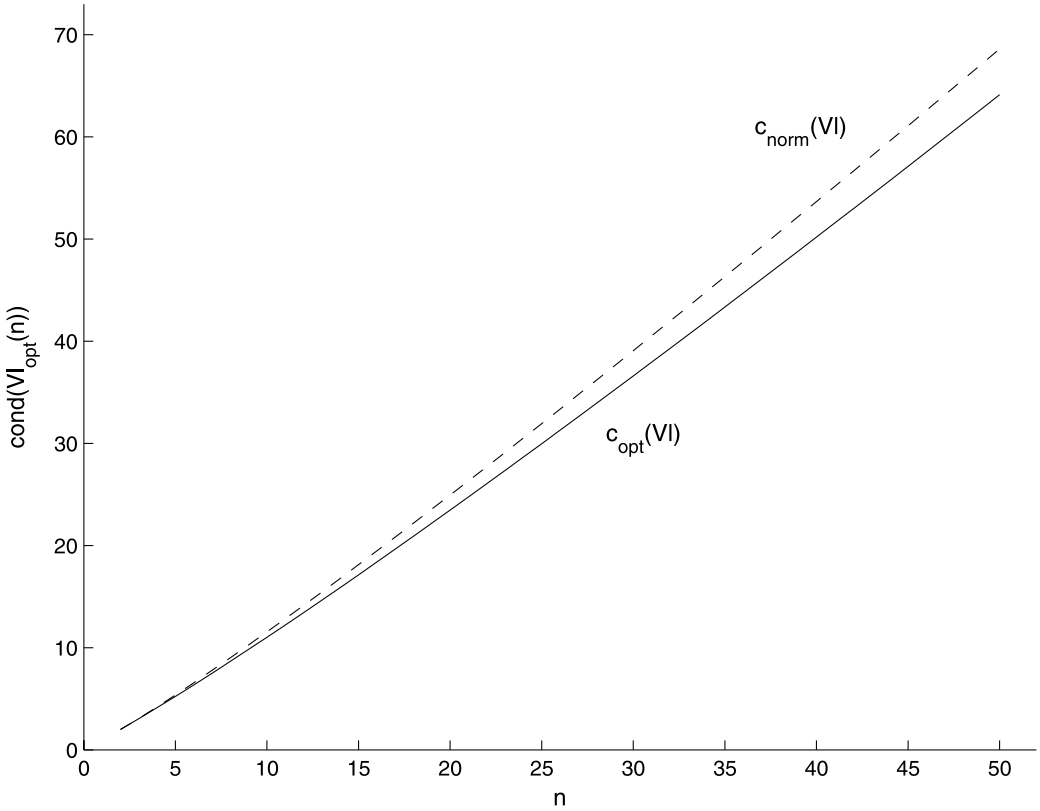


Fig. 8 Frobenius condition of optimal and normalized Vandermonde–Legendre matrices

Example 5.2 (Vandermonde–Chebyshev matrices) Here, p_ν are the Chebyshev polynomials of the first and second kind. For the former, when $\boldsymbol{x}$ is the respective Gauss–Chebyshev vector, the Vandermonde–Chebyshev matrix, as already mentioned, is perfectly conditioned,

$$\operatorname{cond}_F(\boldsymbol{W}_n(\boldsymbol{x})) = n \quad (\boldsymbol{x} = \text{first-kind Gauss–Chebyshev node vector}). \tag{33}$$

For the latter, $\lambda_\nu = \pi \sin^2(\nu\pi/(n+1))/(n+1)$, so that by (24)

$$\operatorname{cond}_F(\boldsymbol{W}_n(\boldsymbol{x})) = \sqrt{\sum_{\nu=1}^{n} \sin^2 \frac{\nu\pi}{n+1} \sum_{\nu=1}^{n} \frac{1}{\sin^2 \frac{\nu\pi}{n+1}}}$$
$$(\boldsymbol{x} = \text{second-kind Gauss–Chebyshev node vector}). \tag{34}$$

Frobenius condition numbers for monic and normalized Chebyshev polynomials, and Gauss–Chebyshev vector $\boldsymbol{x}$, as well as optimal condition numbers for normalized Chebyshev polynomials, are shown in Table 5.

Perfect conditioning (33) in the case of Chebyshev polynomials of the first kind implies zero gradient of $\boldsymbol{W}_n(\boldsymbol{x})$ at the Gauss–Chebyshev point $\boldsymbol{x}$. This can be nicely confirmed analytically. To begin with, $\lambda_\nu = \pi/n$, so that by (31), $\rho = \pi/n$. Furthermore, since $x_i = \cos\theta_i$, $\theta_i = (2i-1)\pi/2n$, one finds for $p_k(x) = \sqrt{2/\pi}\,T_k(x)$, $k \geq 1$,

Table 5 Frobenius condition numbers of Vandermonde–Chebyshev matrices with monic and normalized polynomials, and optimal condition numbers

n	Monic		Normalized		Optimal	
	Cheb1	Cheb2	Cheb1	Cheb2	Cheb1	Cheb2
5	1.6869(01)	2.4558(01)	5	5.9161	5	5.5446
10	5.3970(02)	8.5144(02)	10	14.832	10	12.954
20	5.5265(05)	9.0818(05)	20	39.243	20	32.408
35	1.8109(10)	3.0289(10)	35	88.148	35	70.299
50	5.9340(14)	9.9957(14)	50	148.66	50	–

that

$$\sum_{k=1}^{n-1} p_k(x_i)\,p_k'(x_i) = \frac{1}{\pi \sin\theta_i} \sum_{k=1}^{n-1} k \sin(2k\theta_i).$$

From the identity [7, (1.352.(1))]

$$\sum_{k=1}^{n-1} k \sin(2k\theta) = \frac{1}{4}\frac{\sin(2n\theta)}{\sin^2\theta} - \frac{n\cos((2n-1)\theta)}{2\sin\theta},$$

which, for $\theta = \theta_i$, yields $nx_i/(2\sqrt{1-x_i^2})$, one gets

$$\sum_{k=1}^{n-1} p_k(x_i)\,p_k'(x_i) = \frac{n}{2\pi}\frac{x_i}{1-x_i^2},$$

hence, by (32),

$$g_i = \frac{x_i}{2(1-x_i^2)} - \sum_{\substack{k=1\\k\neq i}}^{n} \frac{1}{x_i - x_k}. \tag{35}$$

Now,

$$\sum_{\substack{k=1\\k\neq i}}^{n} \frac{1}{x - x_k} = \frac{T_n'(x)}{T_n(x)} - \frac{1}{x - x_i} = \frac{(x-x_i)T_n'(x) - T_n(x)}{(x-x_i)T_n(x)},$$

which, for $x \to x_i$, by applying the rule of Bernoulli-L'Hôpital twice, yields

$$\sum_{\substack{k=1\\k\neq i}}^{n} \frac{1}{x_i - x_k} = \frac{T_n''(x_i)}{2T_n'(x_i)}.$$

On the other hand,

$$(1-x^2)T_n''(x) = xT_n'(x) - n^2 T_n(x),$$

Table 6 Frobenius condition numbers of Vandermonde–Laguerre matrices with monic and normalized polynomials, and optimal condition numbers

n	Monic	Normalized	Optimal
5	3.0362(03)	2.0757(02)	2.1067(01)
10	1.3860(11)	1.0047(06)	1.3409(03)
15	1.9857(20)	7.9047(09)	1.1749(05)
20	2.2351(30)	7.7699(13)	1.8408(08)
25	1.0948(41)	8.7177(17)	1.4895(12)

Table 7 Frobenius condition numbers of Vandermonde–Hermite matrices with monic and normalized polynomials, and optimal condition numbers

n	Monic	Normalized	Optimal
5	1.4185(01)	1.3731(1)	8.3937(0)
10	7.4990(03)	6.8318(2)	8.2754(1)
15	3.1256(07)	4.8288(4)	7.8092(2)
20	4.1560(11)	3.9886(6)	7.4769(3)
25	1.2370(16)	3.6159(8)	7.2162(4)

so that

$$\sum_{\substack{k=1\\k\neq i}}^{n} \frac{1}{x_i - x_k} = \frac{x_i T_n'(x_i)}{(1-x_i^2)\cdot 2T_n'(x_i)} = \frac{x_i}{2(1-x_i^2)},$$

giving $g_i = 0$ by (35).

Example 5.3 (Vandermonde–Laguerre matrices) Monic and normalized Laguerre polynomials and the Gauss–Laguerre node vector $\boldsymbol{x}$ give rise to Frobenius condition numbers shown in Table 6. Because of severe ill-conditioning, the second column (headed "monic") had to be computed in 48-digit arithmetic. Also, in the routine `fminunc.m` for computing optimal condition numbers the default value of 'MaxFunEvals' was increased to 5000. While normalized polynomials do lead to substantially smaller condition numbers than do monic polynomials, the matrices involved are still quite ill-conditioned. Optimal conditioning helps somewhat. A few of the optimal nodes are consistently negative.

Example 5.4 (Vandermonde–Hermite matrices) The condition of Vandermonde–Hermite matrices, with $\boldsymbol{x}$ the Gauss–Hermite vector, is roughly halfway between that for Vandermonde–Legendre and Vandermonde–Laguerre matrices, as is shown in Table 7. The optimal nodes are symmetric with respect to the origin, as in the case of Vandermonde matrices.

As a matter of curiosity, we remark that for $n = 2$, the optimally conditioned Vandermonde-like matrix $\boldsymbol{W}_2(\boldsymbol{x}; \mathrm{d}\lambda)$ is perfectly conditioned with respect to the Frobenius norm, for any (positive) measure $\mathrm{d}\lambda$. The proof, and a formula for $\boldsymbol{x}^{\mathrm{opt}} \in \mathbb{R}^2$, is given in the Appendix.

Acknowledgement The present work was motivated by recent attempts of C.-S. Liu, D.-L. Young, and C.-M. Fan [11] to alleviate ill-conditioning of Vandermonde matrices by suitable scaling.

Appendix

For $n = 2$, the Vandermonde-like matrix for the orthonormal polynomials p_0, p_1 has the form

$$\boldsymbol{W}(\boldsymbol{x}) = \begin{bmatrix} p_0 & p_0 \\ p_1(x_1) & p_1(x_2) \end{bmatrix}, \quad p_0 = \frac{1}{\sqrt{\beta_0}}, \quad \boldsymbol{x} = [x_1, x_2]^T \in \mathbb{R}^2, \tag{A.1}$$

where $p_1(x) = p_1(x; \mathrm{d}\lambda) = c_0 + c_1 x$, $x_1 > x_2$, and $\beta_0 = \int_{\mathbb{R}} \mathrm{d}\lambda(x)$. Its inverse is

$$\boldsymbol{W}^{-1}(\boldsymbol{x}) = \frac{1}{c_1(x_1 - x_2)} \begin{bmatrix} -p_1(x_2)/p_0 & 1 \\ p_1(x_1)/p_0 & -1 \end{bmatrix},$$

and hence

$$F(\boldsymbol{x}) := \mathrm{cond}_F(\boldsymbol{W}(\boldsymbol{x})) = \frac{1}{c_1 p_0 (x_1 - x_2)} [2p_0^2 + p_1^2(x_1) + p_1^2(x_2)]. \tag{A.2}$$

We have

$$\begin{aligned} \frac{\partial F}{\partial x_1} &= \frac{1}{c_1 p_0} \frac{2(x_1 - x_2) p_1(x_1) c_1 - [2p_0^2 + p_1^2(x_1) + p_1^2(x_2)]}{(x_1 - x_2)^2}, \\ \frac{\partial F}{\partial x_2} &= \frac{1}{c_1 p_0} \frac{2(x_1 - x_2) p_1(x_2) c_1 + [2p_0^2 + p_1^2(x_1) + p_1^2(x_2)]}{(x_1 - x_2)^2}. \end{aligned}$$

The extremal point $\boldsymbol{x}^{\mathrm{opt}} = [x_1, x_2]^T$ is determined by the equations

$$\begin{aligned} 2(x_1 - x_2) p_1(x_1) c_1 &= 2p_0^2 + p_1^2(x_1) + p_1^2(x_2), \\ 2(x_1 - x_2) p_1(x_2) c_1 &= -[2p_0^2 + p_1^2(x_1) + p_1^2(x_2)], \end{aligned} \tag{A.3}$$

which, subtracting and adding, are equivalent to

$$\begin{aligned} (x_1 - x_2)^2 c_1^2 &= 2p_0^2 + p_1^2(x_1) + p_1^2(x_2), \\ p_1(x_1) + p_1(x_2) &= 0. \end{aligned}$$

The latter equation yields immediately

$$x_1 + x_2 = -2\frac{c_0}{c_1}, \tag{A.4}$$

which, inserted in the former equation, gives, after a little computation,

$$(c_0 + c_1 x_1)^2 = p_0^2,$$

hence

$$x_1 = \frac{\pm p_0 - c_0}{c_1}. \tag{A.5}$$

At the extremal point $\boldsymbol{x}^{\text{opt}}$, by (A.2) and the first equation in (A.3), we get

$$F(\boldsymbol{x}^{\text{opt}}) = \frac{2(x_1 - x_2)p_1(x_1)c_1}{c_1 p_0(x_1 - x_2)} = 2\frac{p_1(x_1)}{p_0}.$$

Since, by (A.5), $p_1(x_1) = c_0 + c_1(\pm p_0 - c_0)/c_1 = \pm p_0$, we must take the plus sign, giving

$$F(\boldsymbol{x}^{\text{opt}}) = 2,$$

as was to be shown. Moreover, from (A.4) and (A.5),

$$x_1^{\text{opt}} = \frac{p_0 - c_0}{c_1}, \qquad x_2^{\text{opt}} = -\frac{p_0 + c_0}{c_1}. \tag{A.6}$$

We may express this more conveniently in terms of the moments μ_k of the measure $\mathrm{d}\lambda$,

$$\mu_k = \int_{\mathbb{R}} x^k \mathrm{d}\lambda(x), \quad k = 0, 1, 2, \ldots.$$

By orthogonality of p_1, we find

$$c_1 = \frac{-\mu_0}{\mu_1} c_0, \tag{A.7}$$

and by orthonormality, after some manipulation,

$$c_0 = \frac{-\mu_1}{\sqrt{\mu_0(\mu_0\mu_2 - \mu_1^2)}}, \tag{A.8}$$

where the radicand is positive by Schwarz's inequality. Insertion into (A.7) yields

$$c_1 = \sqrt{\frac{\mu_0}{\mu_0\mu_2 - \mu_1^2}}. \tag{A.9}$$

Substituting (A.8) and (A.9) in (A.6) and noting that $p_0 = 1/\sqrt{\mu_0}$ finally gives

$$x_1^{\text{opt}} = \frac{\mu_1 + \sqrt{\mu_0\mu_2 - \mu_1^2}}{\mu_0}, \qquad x_2^{\text{opt}} = \frac{\mu_1 - \sqrt{\mu_0\mu_2 - \mu_1^2}}{\mu_0}. \tag{A.10}$$

As expected, for symmetric measures ($\mu_1 = 0$), the optimal point $\boldsymbol{x}^{\text{opt}}$ is also symmetric with respect to the origin.

References

1. Beckermann, B.: The condition number of real Vandermonde, Krylov and positive definite Hankel matrices. Numer. Math. **85**(4), 553–577 (2000)
2. Gautschi, W.: Norm estimates for inverses of Vandermonde matrices. Numer. Math. **23**(4), 337–347 (1975)
3. Gautschi, W.: Optimally conditioned Vandermonde matrices. Numer. Math. **24**(1), 1–12 (1975)
4. Gautschi, W.: The condition of Vandermonde-like matrices involving orthogonal polynomials. Linear Algebra Appl. **52/53**, 293–300 (1983)
5. Gautschi, W.: Orthogonal Polynomials: Computation and Approximation. Oxford University Press, Oxford (2004)
6. Gautschi, W., Gabriele, I.: Lower bounds for the condition number of Vandermonde matrices. Numer. Math. **52**(3), 241–250 (1988)
7. Gradshteyn, I.S., Ryzhik, I.M.: Tables of Integrals, Series, and Products, 7th edn. Elsevier/Academic Press, Amsterdam (2007)
8. Higham, N.J.: Error analysis of the Björck–Pereyra algorithms. Numer. Math. **50**(5), 613–632 (1987)
9. Higham, N.J.: Fast solution of Vandermonde-like systems involving orthogonal polynomials. IMA J. Numer. Anal. **8**(4), 473–486 (1988)
10. Lebesgue, H.: L'œuvre mathématique de Vandermonde. Enseign. Math. (2) **1**, 203–223 (1956) [Originally published in 1940.]
11. Liu, C.-S., Young, D.-L., Fan, C.-M.: A highly accurate multi-scale polynomial interpolation by reducing the condition numbers of Vandermonde matrices. Manuscript
12. Szegö, G.: Orthogonal Polynomials. Colloquium Publ., vol. 23. Am. Math. Soc., Providence (1975)
13. van der Sluis, A.: Condition numbers and equilibration of matrices. Numer. Math. **14**(1), 14–23 (1969)

9

Papers on Special Functions

---

9 Some elementary inequalities relating to the gamma and incomplete gamma function, J. Math. and Phys. 38, 77–81 (1959)

10 Exponential integral $\int_1^\infty e^{-xt}t^{-n}dt$ for large values of n, J. Res. Nat. Bur. Standards 62, 123–125 (1959)

13 Recursive computation of the repeated integrals of the error function, Math. Comp. 15, 227–232 (1961)

39 Efficient computation of the complex error function, SIAM J. Numer. Anal. 7, 187–198 (1970)

47 A harmonic mean inequality for the gamma function, SIAM J. Math. Anal. 5, 278–281 (1974)

48 Some mean value inequalities for the gamma function, SIAM J. Math. Anal. 5, 282–292 (1974)

49 Computational methods in special functions — a survey, in *Theory and applications of special functions* (R. A. Askey, ed.), 1–98, Math. Res. Center, Univ. Wisconsin Publ. 35, Academic Press, New York, 1975

61 Anomalous convergence of a continued fraction for ratios of Kummer functions, Math. Comp. 31, 994–999 (1977)

68 A computational procedure for incomplete gamma functions, ACM Trans. Math. Software 5, 466–481 (1979)

72 (with F. Costabile) Lower bounds for the largest zeros of orthogonal polynomials, Boll. Un. Mat. Ital. (5) 17A, 516–522 (1980) (translated from Italian)

155 The incomplete gamma functions since Tricomi, in *Tricomi's ideas and contemporary applied mathematics*, 203–237, Atti Convegni Lincei 147, Accademia Nazionale dei Lincei, Roma, (1998)

168 Gauss quadrature approximations to hypergeometric and confluent hypergeometric functions, J. Comput. Appl. Math. 139, 173–187 (2002)

169 Computation of Bessel and Airy functions and of related Gaussian quadrature formulae, BIT 42, 110–118 (2002)

178 Numerical quadrature computation of the Macdonald function for complex orders, BIT Numer. Math. 45, 593–603 (2005)

182 (with P. Leopardi) Conjectured inequalities for Jacobi polynomials and their largest zeros, Numer. Algorithms 45, 217–230 (2007)

190 On a conjectured inequality for the largest zero of Jacobi polynomials, Numer. Algorithms 49, 195–198 (2008)

191 On conjectured inequalities for zeros of Jacobi polynomials, Numer. Algorithms 50, 93–96 (2009)

192 New conjectured inequalities for zeros of Jacobi polynomials, Numer. Algorithms 50, 293–296 (2009)

193 How sharp is Bernstein's inequality for Jacobi polynomials?, Electr. Trans. Numer. Anal. 36, 1–8 (2009)

199 The Lambert W-functions and some of their integrals: a case study of high-precision computation, Numer. Algorithms 57, 27–34 (2011)

203 Remark on "New conjectured inequalities for zeros of Jacobi polynomials by Walter Gautschi, Numer. Algorithms 50: 293–296 (2009)," Numer. Algorithms 57, 511 (2011)

9.1. [9] "SOME ELEMENTARY INEQUALITIES RELATING TO THE GAMMA AND INCOMPLETE GAMMA FUNCTION"

---

[9] "Some Elementary Inequalities Relating to the Gamma and Incomplete Gamma Function," *J. Math. and Phys.* **38**, 77–81 (1959).

Reprinted from JOURNAL OF MATHEMATICS AND PHYSICS
Vol. 38, No. 1, April, 1959
Printed in U.S.A.

SOME ELEMENTARY INEQUALITIES RELATING TO THE GAMMA AND INCOMPLETE GAMMA FUNCTION*

BY WALTER GAUTSCHI

1. In a recent note Y. Komatu [3] has proved the inequality

$$(1)\qquad \frac{1}{x+\sqrt{x^2+2}} < e^{x^2}\int_x^\infty e^{-t^2}\,dt < \frac{1}{x+\sqrt{x^2+1}}\quad (0 \leqq x < \infty).$$

The deviation of the bounds from the estimated function decreases monotonically to zero as x varies from zero to infinity. H. O. Pollak [5] has improved the upper bound by showing that

$$(2)\qquad e^{x^2}\int_x^\infty e^{-t^2}\,dt \leqq \frac{1}{x+\sqrt{x^2+4/\pi}}$$

with a deviation increasing from zero to a maximum value and decreasing, from there on, monotonically to zero.

We shall prove in section 3 the more general inequality

$$(3)\qquad \begin{aligned} &\tfrac{1}{2}((x^p+2)^{1/p}-x) < e^{x^p}\int_x^\infty e^{-t^p}\,dt \\ &\qquad \leqq c_p\left(\left(x^p+\frac{1}{c_p}\right)^{1/p}-x\right),\ c_p=\left\{\Gamma\left(1+\frac{1}{p}\right)\right\}^{p/(p-1)}\quad (0 \leqq x < \infty) \end{aligned}$$

where p is any real number >1.[1] For $p = 2$ the right-hand inequality in (3) reduces to (2) while the left-hand inequality reduces to the corresponding inequality in (1). The deviations of the bounds in the general p-case behave the same way as in the special case $p = 2$. Also, (3) remains valid if we replace c_p by 1. The quality of the bounds is indicated in Fig. 1.

By an easy transformation we can write (3) in terms of the complementary gamma function $\Gamma(a, x) = \int_x^\infty e^{-t}t^{a-1}\,dt$ as follows:

$$(4)\qquad \begin{aligned} &\tfrac{1}{2}p((x+2)^{1/p}-x^{1/p}) < e^x\Gamma(p^{-1},x) \\ &\qquad \leqq pc_p((x+c_p^{-1})^{1/p}-x^{1/p})\quad (0 \leqq x < \infty). \end{aligned}$$

In particular, if $p \to \infty$, we obtain an inequality for the exponential integral $E_1(x) = \Gamma(0, x)$:

$$(5)\qquad \tfrac{1}{2}\ln(1+2x^{-1}) \leqq e^xE_1(x) \leqq \ln(1+x^{-1})\quad (0 < x < \infty).$$

This improves an inequality due to E. Hopf [1]; the bounds in (5) exhibit the logarithmic singularity of $E_1(x)$ at $x = 0$.

* This paper was prepared under a National Bureau of Standards contract with American University.

[1] The integral in (3) for $p = 3$ occurs in heat transfer problems [2], for $p = 4$ in the study of electrical discharge through gases [6]. An application of (3) for general p is given in [4].

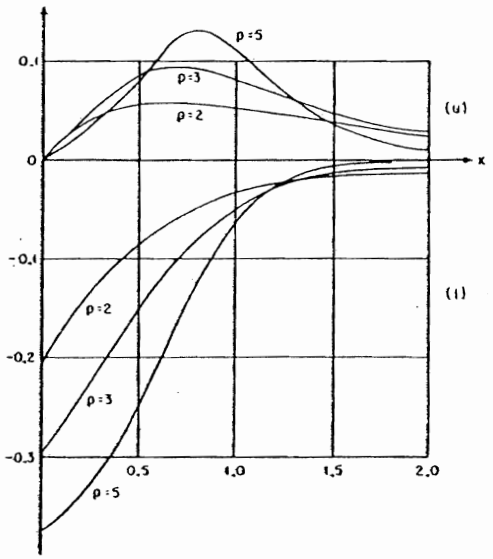


FIG. 1. Relative error of upper (u) and lower (l) bounds in (3)

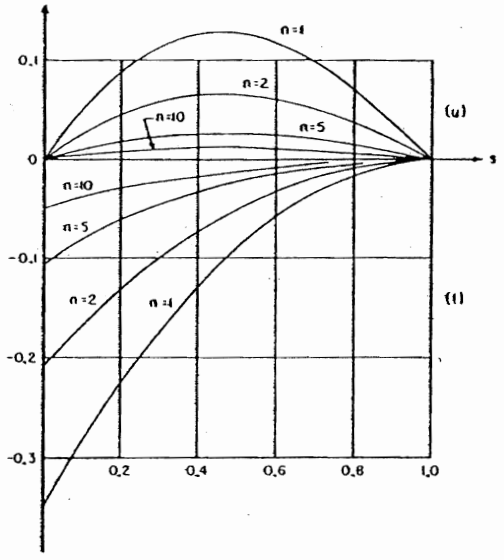


FIG. 2. Relative error of upper (u) and lower (l) bounds in (6)

2. From (4) we can deduce a simple inequality for the gamma function if we set $p = 1/s$, $x = 0$ and replace c_p by 1:

$$2^{s-1} \leqq \Gamma(1+s) \leqq 1 \qquad (0 \leqq s \leqq 1).$$

We shall prove in section 4 the sharper and more general inequality

$$(6) \qquad e^{(s-1)\psi(n+1)} \leqq \frac{\Gamma(n+s)}{\Gamma(n+1)} \leqq n^{s-1} \qquad (0 \leqq s \leqq 1,\ n = 1, 2, 3, \cdots)$$

where $\psi(n+1) = \sum_{k=1}^{n} 1/k - \gamma$ and $\gamma = 0.57721\ldots$ is the Euler-Mascheroni constant. We have equality in (6) only for $s = 1$ on the left-hand side and for $s = 0$ and $s = 1$ on the right-hand side. Fig. 2 illustrates the quality of the bounds. Since $\psi(n) < \ln n$ we have also

$$\left(\frac{1}{n+1}\right)^{1-s} \leqq \frac{\Gamma(n+s)}{\Gamma(n+1)} \leqq \left(\frac{1}{n}\right)^{1-s} \qquad (0 \leqq s \leqq 1). \tag{7}$$

It may be of interest to note that by letting $n \to \infty$ in (7) we obtain a simple proof of Euler's product formula in the segment $0 \leqq s \leqq 1$. In fact, (7) is equivalent to

$$\frac{n!(n+1)^{s-1}}{(s+1)(s+2)\cdots(s+n-1)} \leqq \Gamma(1+s) \leqq \frac{(n-1)!n^s}{(s+1)(s+2)\cdots(s+n-1)}. \tag{8}$$

Setting

$$\gamma_n(s) = \frac{(n-1)!n^s}{(s+1)(s+2)\cdots(s+n-1)}$$

we can write (8) in the form

$$\left(\frac{1}{1+1/n}\right)^{1-s} \gamma_n(s) \leqq \Gamma(1+s) \leqq \gamma_n(s).$$

Therefore

$$0 \leqq \gamma_n(s) - \Gamma(1+s) \leqq \Gamma(1+s)\{(1+1/n)^{1-s} - 1\} = 0(1/n) \quad (n \to \infty).$$

3. *Proof of* (3). Let

$$\Delta_p(a, x) = ae^{-x^p}((x^p + a^{-1})^{1/p} - x) - \int_x^\infty e^{-t^p}\,dt \quad (a > 0). \tag{9}$$

We have to prove that

$$\Delta_p(c_p, x) \geqq 0, \qquad \Delta_p(\tfrac{1}{2}, x) < 0 \qquad (0 \leqq x < \infty). \tag{10}$$

Differentiating (9) with respect to x we find

$$u^{p-1}(u^p - 1)e^{x^p}\frac{\partial}{\partial x}\Delta_p(a, x) = (1-a)u^{2p-1} - (p-a)u^p + (p+a-1)u^{p-1} - a \tag{11}$$

where

$$u = [1 + (1/ax^p)]^{1/p}, \qquad u \geqq 1. \tag{12}$$

Denoting the polynomial on the right-hand side of (11) by $g_p(u)$ we have

$$g_p(1) = g_p'(1) = 0, \qquad g_p''(1) = p(p-1)(1-2a). \tag{13}$$

Consider now the case $a = c_p$. We first note that

$$\Delta_p(a, \infty) = 0, \qquad \Delta_p(c_p, 0) = 0. \tag{14}$$

Next we notice that the coefficients of $g_p(u)$ alternate in sign. Since there are three sign changes we conclude from Descartes' rule that $g_p(u)$ has either three positive zeros or one. (13) shows that two zeros are located at $u = 1$; thus $g_p(u)$ has exactly three positive zeros. Furthermore, since $g_p''(1) < 0$ and $g_p(\infty) = \infty$, the third zero must be larger than 1. Therefore, as u increases from 1 to ∞ the polynomial $g_p(u)$ decreases from zero to a minimum value and from there on increases monotonically to ∞. On account of (11), (12) and (14) this means that $\Delta_p(c_p, x)$ increases from zero to a maximum value and from there on decreases monotonically to zero as x varies from zero to ∞. This proves the first relation in (10).

Consider next the case $a = \frac{1}{2}$. Again, Descartes' rule applies and from (13) it follows that all three positive zeros of $g_p(u)$ coincide at $u = 1$. Therefore $g_p(u) > 0$ for $u > 1$, from which follows $(\partial/\partial x)\Delta_p(\frac{1}{2}, x) > 0$ for $x > 0$. This proves the second relation in (10) since $\Delta_p(\frac{1}{2}, \infty) = 0$.

4. *Proof of* (6). Consider

$$f(s) = \frac{1}{1-s} \ln \left\{ \frac{\Gamma(n+s)}{\Gamma(n+1)} \right\} \qquad (0 \leqq s < 1).$$

We have $f(0) = \ln(1/n)$ and by using the rule of Bernoulli-L'Hospital

$$\lim_{s \to 1} f(s) = -\lim_{s \to 1} \psi(n+s) = -\psi(n+1)$$

where $\psi(x) = (d/dx)[\ln \Gamma(x)]$. We show that $f(s)$ is monotonically decreasing. We have

$$(1-s)f'(s) = f(s) + \psi(n+s).$$

Letting

$$\varphi(s) = (1-s)[f(s) + \psi(n+s)]$$

we have

$$\varphi(0) = \psi(n) - \ln n < 0, \qquad \varphi(1) = 0, \qquad \varphi'(s) = (1-s)\psi'(n+s).$$

Since $\psi'(n+s) > 0$ it follows that $\varphi(s) < 0$ and therefore $f'(s) < 0$ for $0 < s < 1$. Thus

$$-\psi(n+1) \leqq f(s) \leqq \ln(1/n)$$

which is equivalent to (6).

REFERENCES

[1] E. Hopf, Mathematical Problems of Radiative Equilibrium, Cambridge Tracts in Mathematics and Mathematical Physics, No. 31, Cambridge University Press 1934, p. 26.

[2] K. Yamagata, A Contribution to the Theory of Non-isothermal Laminar Flow of Fluids inside a Straight Tube of Circular Cross Section, Mem. Fac. Engr., Kyushu Imp. Univ. 8 (1940), pp. 365–449.

[3] Y. KOMATU, Elementary inequalities for Mills' ratio, Rep. Statist. Appl. Res. Un. Jap. Sci. Engrs. 4 (1955), pp. 69–70.

[4] G. FRANKLIN MONTGOMERY, On the Transmission Error Function for Meteor-Burst Communication, Proc. IRE 46 (1958), pp. 1423–1424.

[5] H. O. POLLAK, A Remark on "Elementary Inequalities for Mills' Ratio" by Yûsaku Komatu, Rep. Statist. Appl. Res. Un. Jap. Sci. Engrs. 4 (1956), p. 40.

[6] W. O. SCHUMANN, Elektrische Durchbruchfeldstärke von Gasen, Springer Berlin 1923.

AMERICAN UNIVERSITY
WASHINGTON, D. C.

(Received March 27, 1958)

9.2. [10] "Exponential Integral $\int_1^\infty e^{-xt}\, t^{-n} dt$ for Large Values of n"

---

[10] "Exponential Integral $\int_1^\infty e^{-xt}\, t^{-n} dt$ for Large Values of n," *J. Res. Nat. Bur. Standards* **62**, 123–125 (1959).

Journal of Research of the National Bureau of Standards Vol. 62, No. 3, March 1959 Research Paper 2941

Exponential Integral $\int_1^\infty e^{-xt}t^{-n}dt$ for Large Values of n [1]

Walter Gautschi

An asymptotic expansion is given which is well suited for numerical computation when n is large and x arbitrary positive.

1. Let

$$E_n(x)=\int_1^\infty e^{-xt}\,t^{-n}dt; \qquad x>0; \qquad n=1, 2, 3, \ldots . \tag{1}$$

By means of four integrations by parts, G. Blanch[2] has found the approximation

$$E_n(x)\approx\frac{e^{-x}}{x+n}\left[1+\frac{n}{(x+n)^2}+\frac{n(n-2x)}{(x+n)^4}+\frac{n(6x^2-8nx+n^2)}{(x+n)^6}\right]. \tag{2}$$

She also gives an integral representation for the error. Formula (2) has proved very efficient for computing $E_n(x)$ for large values of n in the whole range $x>0$. In what follows, the complete expansion is given, as well as error estimates.

Denote by $h_k(u)$ the polynomial (of degree $k-1$ if $k>0$), defined recursively by

$$h_{k+1}(u)=(1-2ku)h_k(u)+u(1+u)h_k'(u) \quad (k=0, 1, 2, \ldots), \quad h_0(u)=1. \tag{3}$$

Let

$$H_k(u)=\frac{h_k(u)}{(1+u)^{2k}}, \tag{4}$$

and let α_k, β_k be lower and upper bounds, respectively, for $H_k(u)$ in the interval $u\geq 0$:

$$\alpha_k\leq H_k(u)\leq\beta_k \qquad (u\geq 0). \tag{5}$$

Then it will be proved that

$$E_n(x)=\frac{e^{-x}}{x+n}\left[\sum_{\kappa=0}^{k-1}H_\kappa\left(\frac{x}{n}\right)n^{-\kappa}+R_k(x,n)\right], \tag{6}$$

$$\alpha_k n^{-k}\leq R_k(x,n)\leq\beta_k\left(1+\frac{1}{x+n-1}\right)n^{-k}. \tag{7}$$

---

[1] This paper was prepared under a National Bureau of Standards contract with The American University.

[2] G. Blanch, An asymptotic expansion for $E_n(x)=\int_1^\infty (e^{-xu}/u^n)du$, NBS Applied Math. Series 37, 61 (1954).

For reference, the first eight polynomials $h_k(u)$ and corresponding values of α_k, β_k are listed :[3]

$$h_0(u)=h_1(u)=1$$

$$h_2(u)=1-2u$$

$$h_3(u)=1-8u+6u^2$$

$$h_4(u)=1-22u+58u^2-24u^3$$

$$h_5(u)=1-52u+328u^2-444u^3+120u^4$$

$$h_6(u)=1-114u+1452u^2-4400u^3+3708u^4-720u^5$$

$$h_7(u)=1-240u+5610u^2-32120u^3+58140u^4-33984u^5+5040u^6$$

$$\alpha_1=0, \quad \alpha_2=-0.07, \quad \alpha_3=-0.18, \quad \alpha_4=-0.36, \quad \alpha_5=-0.60, \quad \alpha_6=-0.94, \quad \alpha_7=-1.4$$

$$\beta_1=\beta_2=\beta_3=\beta_4=\beta_5=\beta_6=1, \quad \beta_7=1.8.$$

2. Consider, more generally, the integral

$$I=\int_a^b e^{f(t)}dt, \tag{8}$$

where $f(t)$ is a real function defined on the finite or infinite interval (a, b). It is assumed that $f(t)$ has derivatives of any order in (a, b) and that $f'(t)\neq 0$. Following van der Corput and Franklin,[4] we define the sequence $g_k(t)$ by

$$g_0(t)=\frac{1}{f'(t)}, \qquad g_{k+1}(t)=\frac{g_k'(t)}{f'(t)} \qquad (k=0,1,2, \ldots). \tag{9}$$

Setting

$$I_k=\int_a^b g_k(t)f'(t)e^{f(t)}dt,$$

clearly $I_0=I$, and integration by parts yields

$$I_k=[g_k(t)e^{f(t)}]_a^b-\int_a^b g_k'(t)e^{f(t)}dt=v_k-I_{k+1}, \tag{10}$$

where

$$v_k=g_k(b)e^{f(b)}-g_k(a)e^{f(a)}. \tag{11}$$

Hence,

$$I=v_0-v_1+v_2-v_3+ \ \ldots \ +(-1)^{k-1}v_{k-1}+(-1)^k I_k, \tag{12}$$

$$I_k=\int_a^b g_{k-1}'(t)e^{f(t)}dt. \tag{13}$$

---

[3] The author is indebted to Mrs. L. K. Cherwinski and Mrs. B. H. Walter for the calculation of the α_k and β_k.

[4] J. G. van der Corput and Joel Franklin, Approximation of integrals by integration by parts, Nederl. Akad. Wetensch. Proc. Ser. [A] 54 213-219 (1951).

In case of an infinite interval (a, b) it has to be assumed that the values (11) exist for all k. Equation (10) then shows that the existence of the integral I_k implies the existence of I_{k+1}.

3. The integral in (8) is equal to $E_n(x)$ if

$$f(t) = -(xt + n \ln t), \qquad a = 1, \qquad b = \infty.$$

A short computation shows that with this definition of $f(t)$, the sequence $g_k(t)$ in (9) is equal to

$$g_k(t) = \frac{(-1)^{k+1} t}{xt+n} \frac{h_k(u)}{(1+u)^{2k}} n^{-k} = \frac{(-1)^{k+1} t}{xt+n} H_k(u) n^{-k}, \quad u = \frac{xt}{n},$$

where the $h_k(u)$ are the polynomials defined in (3) and $H_k(u)$ the rational functions (4). For the quantities v_k in (11), one obtains

$$v_k = \frac{(-1)^k e^{-x}}{x+n} H_k\left(\frac{x}{n}\right) n^{-k}.$$

Furthermore,

$$g'_{k-1}(t) = (-1)^k H_k(u) n^{-k}, \qquad u = \frac{xt}{n}.$$

Hence, from (12) and (13),

$$E_n(x) = \frac{e^{-x}}{x+n}\left[\sum_{\kappa=0}^{k-1} H_\kappa\left(\frac{x}{n}\right) n^{-\kappa}\right] + n^{-k} \int_1^\infty H_k\left(\frac{xt}{n}\right) e^{-xt} t^{-n} dt. \tag{14}$$

By (5)

$$\alpha_k E_n(x) \leq \int_1^\infty H_k\left(\frac{xt}{n}\right) e^{-xt} t^{-n} dt \leq \beta_k E_n(x),$$

and using the well-known inequality,[5]

$$\frac{1}{x+n} \leq e^x E_n(x) \leq \frac{1}{x+n-1} \qquad (x \geq 0),$$

one gets

$$\alpha_k \frac{e^{-x}}{x+n} \leq \int_1^\infty H_k\left(\frac{xt}{n}\right) e^{-xt} t^{-n} dt \leq \beta_k \frac{e^{-x}}{x+n-1}.$$

From this and (14) the formulas (6), (7) follow immediately.

It may be observed that the result (6), (7) holds also for nonintegral values of n with $n > 1$.

[5] E. Hopf, Mathematical problems of radiative equilibrium, Cambridge Tracts in Mathematics and Mathematical Physics, No. 31, p. 26 (Cambridge University Press, 1934).

WASHINGTON, October 14, 1958.

9.3. [13] "Recursive Computation of the Repeated Integrals of the Error Function"

---

[13] "Recursive Computation of the Repeated Integrals of the Error Function," *Math. Comp.* **15**, 227–232 (1961).

Recursive Computation of the Repeated Integrals of the Error Function

By Walter Gautschi

1. This paper is concerned with a special technique, originated by J. C. P. Miller [1, p. xvii], of computing a solution f_n of a second-order difference equation

$$y_{n+1} + a_n y_n + b_n y_{n-1} = 0 \qquad (n = 1, 2, 3, \cdots) \tag{1.1}$$

for $n = 0(1)N$, N large, in cases where (1.1) has a second solution, g_n, which ultimately grows much faster than f_n. Straightforward use of (1.1) is then not adequate, since rounding errors will "activate" the second solution g_n, which in turn will eventually overshadow the desired solution f_n. Miller's device consists of applying (1.1) in backward direction,

$$y_{n-1} = -b_n^{-1}(a_n y_n + y_{n+1}) \qquad (n = \nu - 1, \nu - 2, \cdots, 1; \nu > N), \tag{1.2}$$

starting with the initial values

$$y_{\nu-1} = \alpha, \qquad y_\nu = 0, \tag{1.3}$$

where α is any real number $\neq 0$. If ν is taken sufficiently large the values so obtained turn out to be approximately proportional to f_n in the range $0 \leqq n \leqq N$. The factor of proportionality may then be determined, e.g., by comparing y_0 with f_0.

This technique was originally devised [1] for the computation of Bessel functions $I_n(x)$, and has since then been applied to various other Bessel functions [2], [5], [9], to Legendre functions [8], and to the repeated integrals of the error function* [6],

$$\begin{aligned} i^n \operatorname{erfc} x &= \frac{2}{\sqrt{\pi}} \int_x^\infty \frac{(t-x)^n}{n!} e^{-t^2}\, dt \qquad (n = 0, 1, 2, \cdots), \\ i^{-1} \operatorname{erfc} x &= \frac{2}{\sqrt{\pi}} e^{-x^2}. \end{aligned} \tag{1.4}$$

An analogous technique for first-order difference equations is described in [4, p. 25].

We shall present in Section 2 a detailed description of Miller's procedure, paying special attention to the error term. In Section 3 we study the procedure as applied to the computation of the functions (1.4) and show that the process converges for any positive x, as $\nu \to \infty$. In Sections 4–5 estimates will be developed of how large ν must be taken to ensure any prescribed accuracy.

Received August 22, 1960.

* In this notation $i^n (n \geqq 0)$ should be interpreted as the nth power of the integral operator $i = \int_x^\infty$, so that

$$i^0 \operatorname{erfc} x = \operatorname{erfc} x, \qquad i^n \operatorname{erfc} x = \int_x^\infty i^{n-1} \operatorname{erfc} t\, dt \qquad (n = 1, 2, \cdots).$$

This notation for the repeated integrals of the error function, even though not entirely satisfactory, has become standard.

2. Consider the homogeneous second-order difference equation

$$y_{n+1} + a_n y_n + b_n y_{n-1} = 0 \qquad (n = 1, 2, 3, \cdots) \tag{2.1}$$

and assume that

$$b_n \neq 0 \quad \text{for all} \quad n \geqq 1. \tag{2.2}$$

Let f_n be the (nontrivial) solution of (2.1) to be computed for $n = 0(1)N$. We assume

$$f_0 \neq 0. \tag{2.3}$$

Let there be another solution g_n of (2.1), for which

$$g_n \neq 0 \quad \text{for all} \quad n \geqq 0, \tag{2.4}$$

and

$$\lim_{n\to\infty} \left| \frac{f_n}{g_n} \right| = 0. \tag{2.5}$$

It follows readily that f_n, g_n are linearly independent.

Now let $y_n^{(\nu)} (n = 0, 1, \cdots, \nu - 2; \nu > N)$ be the result of applying (2.1) in backward direction, starting with

$$y_{\nu-1}^{(\nu)} = \alpha, \qquad y_\nu^{(\nu)} = 0 \qquad (\alpha \neq 0). \tag{2.6}$$

These values, by (2.2), are well defined, and, as will presently be shown, $y_0^{(\nu)} \neq 0$ for ν sufficiently large. Let us then define

$$f_n^{(\nu)} = \frac{f_0}{y_0^{(\nu)}} y_n^{(\nu)}. \tag{2.7}$$

We show that for any fixed n,

$$\lim_{\nu\to\infty} f_n^{(\nu)} = f_n. \tag{2.8}$$

Moreover,

$$f_n^{(\nu)} = f_n \frac{1 - \dfrac{f_\nu}{g_\nu}\dfrac{g_n}{f_n}}{1 - \dfrac{f_\nu}{g_\nu}\dfrac{g_0}{f_0}}. \tag{2.9}$$

It is sufficient to prove (2.9), since (2.8) then follows from (2.5). Let $y_n^{(\nu)}$ be extended to all $n > \nu$ by means of (2.1). Then for every fixed ν the sequence $\{y_n^{(\nu)}\} (n = 0, 1, 2, \cdots)$ is a solution of (2.1), and therefore representable in the form

$$y_n^{(\nu)} = A^{(\nu)} f_n + B^{(\nu)} g_n \qquad (n \geqq 0).$$

By (2.6),

$$\begin{aligned} A^{(\nu)} f_{\nu-1} + B^{(\nu)} g_{\nu-1} &= \alpha, \\ A^{(\nu)} f_\nu + B^{(\nu)} g_\nu &= 0. \end{aligned} \tag{2.10}$$

Certainly, $A^{(\nu)} \neq 0$, since otherwise, by (2.4), $A^{(\nu)} = B^{(\nu)} = 0$, which contradicts the first equation in (2.10). From the second equation, $B^{(\nu)}/A^{(\nu)} = -f_\nu/g_\nu$. Therefore

$$y_n^{(\nu)} = A^{(\nu)}\left(f_n + \frac{B^{(\nu)}}{A^{(\nu)}} g_n\right) = A^{(\nu)}\left(f_n - \frac{f_\nu}{g_\nu} g_n\right).$$

If ν is sufficiently large it follows because of (2.3) and (2.5) that $y_0^{(\nu)} \neq 0$. By (2.7),

$$f_n^{(\nu)} = \frac{f_0 A^{(\nu)}\left(f_n - \frac{f_\nu}{g_\nu} g_n\right)}{A^{(\nu)}\left(f_0 - \frac{f_\nu}{g_\nu} g_0\right)} = f_n \frac{1 - \frac{f_\nu}{g_\nu}\frac{g_n}{f_n}}{1 - \frac{f_\nu}{g_\nu}\frac{g_0}{f_0}},$$

which proves (2.9).

It is convenient to define

$$\rho_n = \frac{f_n g_0}{g_n f_0} \qquad (n = 0, 1, 2, \cdots), \tag{2.11}$$

so that $\rho_n \to 0$ as $n \to \infty$, and

$$f_n^{(\nu)} = f_n \frac{1 - (\rho_\nu/\rho_n)}{1 - \rho_\nu}.$$

The relative error of $f_n^{(\nu)}$ is given by

$$\frac{f_n^{(\nu)} - f_n}{f_n} = \frac{\rho_\nu}{1 - \rho_\nu}\left(1 - \frac{1}{\rho_n}\right). \tag{2.12}$$

The approximations $f_n^{(\nu)}$ obviously do not depend on α, so that α can be chosen at will. If a high-speed computer is employed it is advisable to choose a small value for α to guard against "overflow" in the values of $y_n^{(\nu)}$.

3. Now let

$$f_n = i^{n-1}\operatorname{erfc} x, \times 70 \qquad (n = 0, 1, 2, \cdots). \tag{3.1}$$

Then f_n is a solution of

$$y_{n+1} + \frac{x}{n} y_n - \frac{1}{2n} y_{n-1} = 0 \qquad (n = 1, 2, 3, \cdots) \tag{3.2}$$

as is readily verified by writing

$$i^n \operatorname{erfc} x = \frac{2}{\sqrt{\pi}}\left(\frac{1}{n}\int_x^\infty \frac{(t-x)^{n-1}}{(n-1)!} te^{-t^2}\,dt - \frac{x}{n}\int_x^\infty \frac{(t-x)^{n-1}}{(n-1)!} e^{-t^2}\,dt\right)$$

and evaluating the first integral by parts. A second solution of (3.2) is given by

$$g_n = (-1)^n i^{n-1}\operatorname{erfc}(-x) \qquad (n = 0, 1, 2, \cdots). \tag{3.3}$$

It is clear that the assumptions (2.2)–(2.4) are satisfied in this case. We shall now verify (2.5), i.e.

$$\lim_{n\to\infty} \frac{i^n \operatorname{erfc} x}{i^n \operatorname{erfc}(-x)} = 0 \qquad (x > 0). \tag{3.4}$$

This then will prove the convergence of the procedure in Section 2, as applied to (3.2).

We recall that the repeated integrals of the error function are related to the parabolic cylinder functions $D_\nu(x)$ by [7, p. 76]

$$i^n \operatorname{erfc} x = \frac{e^{-\frac{1}{4}x^2}}{2^{(n-1)/2}\sqrt{\pi}} D_{-n-1}(x\sqrt{2}).$$

It is furthermore known [3, p. 123] that

$$D_{-n-1}(z) = \frac{\sqrt{\pi}e^{-\sqrt{n}z}}{2^{(n+1)/2}\Gamma\left(\frac{n}{2}+1\right)}\left[1 + O\left(\frac{1}{\sqrt{n}}\right)\right] \qquad (n \to \infty,\ z \text{ bounded}).$$

Therefore we obtain immediately for any fixed x, real or complex,

$$i^n \operatorname{erfc} x = \frac{e^{-\frac{1}{4}x^2}}{2^n\Gamma\left(\frac{n}{2}+1\right)} e^{-\sqrt{2n}x}\left[1 + O\left(\frac{1}{\sqrt{n}}\right)\right] \qquad (n \to \infty). \tag{3.5}$$

Hence,

$$\frac{i^n \operatorname{erfc} x}{i^n \operatorname{erfc}(-x)} \sim e^{-2\sqrt{2n}x} \qquad (n \to \infty),$$

which proves (3.4).

4. With f_n, g_n defined by (3.1) and (3.3) we have for the quantities ρ_n in (2.11)

$$\rho_n = (-1)^n \frac{i^{n-1} \operatorname{erfc} x}{i^{n-1} \operatorname{erfc}(-x)} \qquad (n = 0, 1, 2, \cdots). \tag{4.1}$$

It is shown in this section that for any fixed $x > 0$ the sequence $\{|\rho_n|\}$ is monotonically decreasing, i.e.,

$$\left|\frac{\rho_{n+1}}{\rho_n}\right| = \frac{i^n \operatorname{erfc} x}{i^{n-1} \operatorname{erfc} x} \cdot \frac{i^{n-1} \operatorname{erfc}(-x)}{i^n \operatorname{erfc}(-x)} < 1 \qquad (n \geqq 0). \tag{4.2}$$

Inequality (4.2) is obvious if $n = 0$ and, by (1.4), equivalent to

$$\int_x^\infty (t-x)^{n-1}e^{-t^2}\,dt \int_{-x}^\infty (s+x)^n e^{-s^2}\,ds - \int_x^\infty (t-x)^n e^{-t^2}\,dt \int_{-x}^\infty (s+x)^{n-1} e^{-s^2}\,ds > 0$$

if $n > 0$. By introducing new variables of integration, $t = u + x$, $s = v - x$, and writing the left-hand side as a double integral, one obtains

$$\iint_Q u^{n-1}v^{n-1}(v-u)e^{-(u+x)^2-(v-x)^2}\,du\,dv > 0, \tag{4.3}$$

where Q denotes the first quadrant $u \geqq 0$, $v \geqq 0$. Let Q_1, Q_2 denote the regions

$$Q_1: \quad 0 < u < v, \qquad Q_2: \quad 0 < v < u.$$

Interchanging variables of integration gives

$$\iint_{Q_2} u^{n-1}v^{n-1}(v-u)e^{-(u+x)^2-(v-x)^2}\,du\,dv = -\iint_{Q_1} u^{n-1}v^{n-1}(v-u)e^{-(v+x)^2-(u-x)^2}\,du\,dv.$$

Therefore (4.3) is equivalent to

$$\iint_{Q_1} u^{n-1}v^{n-1}(v-u)[e^{-(u+x)^2-(v-x)^2} - e^{-(v+x)^2-(u-x)^2}]\,du\,dv > 0.$$

Now, $u^{n-1}v^{n-1}(v-u) > 0$ in Q_1, and the same is true for the expression in brackets, since

$$-(u+x)^2 - (v-x)^2 > -(v+x)^2 - (u-x)^2 \quad \text{for} \quad u < v.$$

This proves (4.3), and thus (4.2).

5. We are now in a position to estimate ν such that for any given integer p,

$$|\,(f_n^{(\nu)} - f_n)/f_n\,| \leqq 10^{-p} \quad \text{for} \quad n = 0, 1, \cdots, N+1.$$

Here, $f_n^{(\nu)}$ denotes the approximations to $f_n = i^{n-1}\operatorname{erfc} x$ obtained by the procedure of Section 2.

Since, by (2.12),

$$|\,(f_n^{(\nu)} - f_n)/f_n\,| \leqq |\,\rho_\nu\,|\,(1 + |\,\rho_n\,|^{-1}) + 0(\rho_\nu^2),$$

and since $|\,\rho_n\,|^{-1}$ increases with n, by (4.2), it is sufficient to choose ν such that

$$|\,\rho_\nu\,|\,(1 + |\,\rho_{N+1}\,|^{-1}) \leqq 10^{-p}. \tag{5.1}$$

From (3.5) and (4.1) we have

$$|\,\rho_{n+1}\,| = e^{-2\sqrt{2}nx}\left[1 + O\left(\frac{1}{\sqrt{n}}\right)\right]. \tag{5.2}$$

Assuming N large enough to neglect the O-term in (5.2) for $n \geqq N$, the requirement (5.1) may be simplified to

$$e^{-2\sqrt{2}\nu x} \leqq \frac{10^{-p}}{1 + e^{2\sqrt{2N}x}},$$

or even to

$$e^{-2\sqrt{2}\nu x} \leqq \frac{10^{-p}}{2e^{2\sqrt{2N}x}}, \tag{5.3}$$

having made the right-hand bound smaller. Inequality (5.3) yields

$$\nu \geqq \left(\frac{2\sqrt{2N}x + p\ln 10 + \ln 2}{2\sqrt{2}x}\right)^2,$$

which gives us the desired estimate of ν.

Oak Ridge National Laboratory
Oak Ridge, Tennessee

1. British Association for the Advancement of Science, *Mathematical Tables, Volume X, Bessel Functions, Part II, Functions of Positive Integer Order*, Cambridge University Press, 1952.
2. F. J. Corbató & J. L. Uretsky, "Generation of spherical Bessel functions in digital computers," *J. Assoc. Comp. Mach.*, v. 6, 1959, p. 366–375.
3. A. Erdélyi, et al., *Higher Transcendental Functions*, Vol. II, McGraw-Hill Book Co., Inc., New York, 1953.
4. W. Gautschi, "Recursive Computation of certain integrals," *J. Assoc. Comput. Mach.*, v. 8, 1961, p. 21–40.
5. M. Goldstein & R. M. Thaler, "Recurrence techniques for the calculation of Bessel functions," *MTAC*, v. 13, 1959, p. 102–108.
6. J. Kaye, "A table of the first eleven repeated integrals of the error function," *J. Math. Phys.*, v. 34, 1955, p. 119–125. See also **RMT 58**, *MTAC*, v. 10, 1956, p. 176.
7. National Physical Laboratory, *Tables of Weber Parabolic Cylinder Functions*, J. C. P. Miller, Editor, H. M. Stationery Office, London, 1955.
8. A. Rotenberg, "The calculation of toroidal harmonics," *Math. Comp.*, v. 14, 1960, p. 274–276.
9. I. A. Stegun & M. Abramowitz, "Generation of Bessel functions on high-speed computers," *MTAC*, v. 11, 1957, p. 255–257.

9.4. [39] "EFFICIENT COMPUTATION OF THE COMPLEX ERROR FUNCTION"

---

[39] "Efficient Computation of the Complex Error Function," *SIAM J. Numer. Anal.* **7**, 187–198 (1970).

SIAM J. NUMER. ANAL.
Vol. 7, No. 1, March 1970

EFFICIENT COMPUTATION OF THE COMPLEX ERROR FUNCTION*

WALTER GAUTSCHI†

Abstract. The paper is concerned with the computation of $w(z) = \exp(-z^2)\operatorname{erfc}(-iz)$ for complex $z = x + iy$ in the first quadrant $Q_1: x \geqq 0, y \geqq 0$. Using Stieltjes' theory of continued fractions it is first observed that the Laplace continued fraction for $w(z)$, although divergent on the real line, represents $w(z)$ asymptotically for $z \to \infty$ in the sector $S: -\pi/4 < \arg z < 5\pi/4$. Specifically, the nth convergent approximates $w(z)$ to within an error of $O(z^{-2n-1})$ as $z \to \infty$ in S. A recursive procedure is then developed which permits evaluating $w(z)$ to a prescribed accuracy for any $z \in Q_1$. The procedure has the property that as $|z|$ becomes sufficiently large, it automatically reduces to the evaluation of the Laplace continued fraction, or, equivalently, to Gauss–Hermite quadrature of $(i/\pi)\int_{-\infty}^{\infty} \exp(-t^2)\,dt/(z-t)$.

1. Introduction. The error function of a complex variable, in more or less disguised forms, occurs in many branches of science and technology. Properties of this function, and computational methods, have been studied extensively. A useful survey, as of 1966, may be found in [1], and more recent work in [2], [11], [14]. In many applications the function must be evaluated a large number of times. It is therefore important to search for methods which are as efficient as possible. Current practice attempts to achieve the desired economy by adopting different methods in different regions of the complex plane. In this paper, instead, we propose a single algorithm which is uniformly effective for all complex arguments. A corresponding ALGOL procedure appears in [8].

In § 2 we review some relevant mathematical properties of the complex error function. Although much of this material is known, a few remarks are made which do not seem to be common knowledge. Among these is the observation that a certain continued fraction, known as the Laplace continued fraction, while divergent on the real line, approximates the error function asymptotically in the sense of Poincaré. The computational algorithm is developed in § 3. Basically, it consists of evaluating an approximation to a truncated Taylor expansion. The increment, h, as well as the number of terms, N, are made to depend on the argument z at which the function is evaluated. As $|z|$ increases, h and N decrease, until eventually both become zero, at which time the algorithm reduces to that of evaluating the Laplace continued fraction. Some performance characteristics, and data on testing the algorithm, are included in § 4.

2. Mathematical preliminaries. The function

$$w(z) = e^{-z^2}\operatorname{erfc}(-iz), \tag{2.1}$$

where $\operatorname{erfc}\zeta = (2/\sqrt{\pi})\int_{\zeta}^{\infty} e^{-t^2}\,dt$ denotes the complementary error function, was first introduced (and tabulated) by Faddeeva and Terent'ev [4]. As a function of the complex variable z, $w(z)$ represents an entire function, and has the property that both its real and imaginary parts are between zero and one for z in the first

* Received by the editors June 5, 1969, and in revised form August 6, 1969.

† Department of Computer Sciences, Purdue University, Lafayette, Indiana 47907. This work was supported, in part, by the National Aeronautics and Space Administration under grant NGR 15-005-039 and, in part, by Argonne National Laboratory.

quadrant of the complex plane. This property may well have been one of the motivations for considering (2.1) as the basic form of the error function for complex argument.

Closely related to (2.1) is the integral

$$f(z) = \int_{-\infty}^{\infty} \frac{e^{-t^2}}{z - t}\, dt. \tag{2.2}$$

We have in fact

$$w(z) = \begin{cases} \dfrac{i}{\pi} f(z), & \operatorname{Im} z > 0, \\ \dfrac{i}{\pi} f(z) + 2e^{-z^2}, & \operatorname{Im} z < 0. \end{cases} \tag{2.3}$$

While $w(z)$ is an entire function, $f(z)$ is analytic for all z not on the real line, and represents two analytic functions, one in the upper, another in the lower half-plane, neither of which is the analytic continuation of the other. For real z, the integral in (2.2) is meaningful only in the sense of a Cauchy principal value integral.

We note that (2.2) is a special case of a Stieltjes transform $\int_{-\infty}^{\infty} d\alpha(t)/(z - t)$. Most of the properties to be described below follow from Stieltjes' classical theory [12], [10] concerning integrals of this type, and are therefore applicable in other situations as well (e.g., in the computation of the complex exponential integral).

Expanding the integrand in (2.2) in descending powers of z, and integrating term by term, one obtains the asymptotic expansion

$$f(z) \sim \sum_{s=0}^{\infty} \frac{\mu_s}{z^{s+1}}, \qquad z \to \infty \text{ in } |\operatorname{Im} z| \geqq a, \quad a > 0 \text{ arbitrary}, \tag{2.4}$$

where

$$\mu_s = \int_{-\infty}^{\infty} t^s e^{-t^2}\, dt = \begin{cases} 0, & s \text{ odd}, \\ \Gamma((s+1)/2), & s \text{ even}, \end{cases} \tag{2.5}$$

are the moments of e^{-t^2}. Since e^{-z^2} has the zero asymptotic expansion in $-\pi/4 < \arg z < \pi/4$ and $3\pi/4 < \arg z < 5\pi/4$, it is not surprising, in view of (2.3), that

$$w(z) \sim \frac{i}{\pi} \sum_{s=0}^{\infty} \frac{\mu_s}{z^{s+1}}, \qquad z \to \infty \text{ in } -\pi/4 < \arg z < 5\pi/4. \tag{2.6}$$

With the (formal) expansion (2.4) is associated the continued fraction

$$\frac{\mu_0}{z-}\,\frac{1/2}{z-}\,\frac{1}{z-}\,\frac{3/2}{z-}\cdots, \tag{2.7}$$

known as the Laplace continued fraction. More precisely, (2.7) is associated with (2.4) in the following sense. Let

$$\frac{\mu_0}{z-}\,\frac{1/2}{z-}\,\frac{1}{z-}\,\frac{3/2}{z-}\cdots\frac{(n-1)/2}{z} = \frac{q_n(z)}{p_n(z)} \tag{2.8}$$

denote the nth convergent of the continued fraction (2.7). It is easily verified that $q_n(z)$ is a polynomial of degree $n-1$, while $p_n(z)$ is a monic polynomial of degree n. Then the quotient in (2.8), if expanded in descending powers of z,

$$\frac{q_n(z)}{p_n(z)} = \sum_{s=0}^{\infty} \frac{v_s^{(n)}}{z^{s+1}}, \tag{2.9}$$

yields a power series which agrees with that in (2.4) up to and including the term with z^{-2n}, i.e.,

$$v_s^{(n)} = \mu_s \qquad \text{for } s = 0, 1, 2, \cdots, 2n-1. \tag{2.10}$$

It is also known [13] that the continued fraction (2.7) in fact converges to $f(z)$ for every nonreal z.

Another remarkable connection of the continued fraction (2.7) with the integral in (2.2) is obtained if the rational function (2.8) is decomposed into partial fractions,

$$\frac{q_n(z)}{p_n(z)} = \sum_{k=1}^{n} \frac{\lambda_k^{(n)}}{z - t_k^{(n)}}. \tag{2.11}$$

(All zeros of $p_n(z)$ are known to be simple.) Expanding both sides of this equation in descending powers of z, and comparing coefficients of like powers, one finds that

$$v_s^{(n)} = \sum_{k=1}^{n} \lambda_k^{(n)} [t_k^{(n)}]^s, \qquad s = 0, 1, 2, \cdots.$$

In view of (2.10) it follows that

$$\mu_s = \sum_{k=1}^{n} \lambda_k^{(n)} [t_k^{(n)}]^s \qquad \text{for } s = 0, 1, 2, \cdots, 2n-1,$$

which, on account of (2.5), implies that $\lambda_k^{(n)}$ and $t_k^{(n)}$ are the weights and nodes, respectively, of n-point Gauss-Hermite quadrature. Consequently,

$$\lim_{n\to\infty} \sum_{k=1}^{n} \frac{\lambda_k^{(n)}}{z - t_k^{(n)}} = \int_{-\infty}^{\infty} \frac{e^{-t^2}}{z-t}\,dt, \qquad \operatorname{Im} z \neq 0, \tag{2.12}$$

i.e., Gauss-Hermite quadrature, applied to the integral in (2.2), converges for every z not on the real line.

Using the well-known remainder term of Gauss-Hermite quadrature it also follows that

$$\mu_{2n} - v_{2n}^{(n)} = \int_{-\infty}^{\infty} t^{2n} e^{-t^2}\,dt - \sum_{k=1}^{n} \lambda_k^{(n)} [t_k^{(n)}]^{2n} = \frac{\sqrt{\pi}\, n!}{2^n}. \tag{2.13}$$

It is interesting to observe that, although (2.12) does not converge if $z = x$ is real, the Gauss-Hermite quadrature sum (2.11) for fixed n and $z = x \to \infty$ nevertheless approximates $-i\pi w(x)$ to within an error of $O(x^{-2n-1})$. In fact, this is true as $z \to \infty$ in the sector $-\pi/4 < \arg z < 5\pi/4$. In other words, the quadrature sums (2.11), and thus *the convergents of the Laplace continued fraction* (2.7) *approximate* $-i\pi w(z)$ *asymptotically as* $z \to \infty$ *in* $-\pi/4 < \arg z < 5\pi/4$. This follows by

combining (2.6) and (2.9)–(2.11),

$$w(z) - \frac{i}{\pi}\sum_{k=1}^{n}\frac{\lambda_k^{(n)}}{z - t_k^{(n)}} = \frac{i}{\pi}\sum_{s=0}^{2n}\frac{\mu_s}{z^{s+1}} + O\left(\frac{1}{z^{2n+3}}\right) - \frac{i}{\pi}\sum_{s=0}^{\infty}\frac{\nu_s^{(n)}}{z^{s+1}}$$

$$= \frac{i}{\pi}\frac{\mu_{2n} - \nu_{2n}^{(n)}}{z^{2n+1}} + O\left(\frac{1}{z^{2n+3}}\right), \qquad z \to \infty \text{ in } -\frac{\pi}{4} < \arg z < \frac{5\pi}{4}.$$

We have used here the symmetry of the Gauss-Hermite weights and nodes, which implies that (2.11) is an odd function of z, and therefore $\nu_s^{(n)} = 0$ for s an odd integer. Also, the series in (2.9), in view of (2.11), obviously converges for $|z| > \max_k t_k^{(n)}$. By virtue of (2.13) we thus have

$$w(z) - \frac{i}{\pi}\sum_{k=1}^{n}\frac{\lambda_k^{(n)}}{z - t_k^{(n)}} = \frac{i}{\sqrt{\pi}}\frac{n!}{2^n}\frac{1}{z^{2n+1}} + O\left(\frac{1}{z^{2n+3}}\right), \tag{2.14}$$

$$z \to \infty \text{ in } -\frac{\pi}{4} < \arg z < \frac{5\pi}{4}.$$

If this is compared with the asymptotic expansion (2.6), i.e.,

$$w(z) - \frac{i}{\pi}\sum_{s=0}^{2n-2}\frac{\mu_s}{z^{s+1}} = \frac{i}{\pi}\frac{\mu_{2n}}{z^{2n+1}} + O\left(\frac{1}{z^{2n+3}}\right), \tag{2.15}$$

one notes that the leading term on the right in (2.14) is smaller than the corresponding term in (2.15) by a factor of

$$\frac{\sqrt{\pi}n!}{2^n\Gamma(n+\frac{1}{2})} \sim \sqrt{\pi}n^{\frac{1}{2}}2^{-n}, \qquad n \to \infty.$$

This is why Gauss-Hermite quadrature is so much more effective for computation than straightforward asymptotic expansion.

There is yet another approach to the continued fraction in (2.7), which involves the repeated integrals of the complementary error function. Consider (see [6] for notations)

$$w_n(z) = e^{-z^2}\mathrm{i}^n\operatorname{erfc}(-iz), \qquad n = 0, 1, 2, \cdots, \quad w_{-1}(z) = \frac{2}{\sqrt{\pi}}, \tag{2.16}$$

so that in particular

$$w_0(z) = w(z). \tag{2.17}$$

If $\operatorname{Im} z > 0$, the sequence $\{w_n(z)\}_{n=-1}^{\infty}$ is known to be a "minimal" solution of the linear second order difference equation

$$y_{n+1} - \frac{iz}{n+1}y_n - \frac{1}{2(n+1)}y_{n-1} = 0, \qquad n = 0, 1, 2, \cdots. \tag{2.18}$$

(For terminology, and subsequent development, see [7].) For any integer $N \geqq 0$, and $\nu > N$, define

$$r_\nu = 0, \qquad r_{n-1} = \frac{1/2}{-iz + (n+1)r_n}, \qquad n = \nu, \nu-1, \cdots, 0,$$

(2.19)

$$v_{-1} = \frac{2}{\sqrt{\pi}}, \quad v_n = r_{n-1}v_{n-1}, \qquad n = 0, 1, 2, \cdots, N.$$

We shall write $r_{n-1}^{[\nu]}(z), v_n^{[\nu]}(z)$ for r_{n-1}, v_n, if we wish to indicate their dependency on ν and z. It can then be shown that

$$\lim_{\nu\to\infty} v_n^{[\nu]}(z) = w_n(z), \qquad \operatorname{Im} z > 0, \quad n = -1, 0, 1, 2, \cdots, \tag{2.20}$$

and consequently,

$$\lim_{\nu\to\infty} r_{n-1}^{[\nu]}(z) = w_n(z)/w_{n-1}(z), \qquad \operatorname{Im} z > 0, \quad n = 0, 1, 2, \cdots.$$

In particular, by (2.17),

$$w(z) = \lim_{\nu\to\infty} v_0^{[\nu]}(z) = \frac{2}{\sqrt{\pi}} \lim_{\nu\to\infty} r_{-1}^{[\nu]}(z), \qquad \operatorname{Im} z > 0. \tag{2.21}$$

To see the connection with the Laplace continued fraction, let $n = 0$ in the second line of (2.19), and then in turn $n = 0, 1, 2, \cdots, \nu$ in the first line of (2.19). One obtains

$$\begin{aligned} v_0^{[\nu]}(z) &= \frac{1}{\sqrt{\pi}} \frac{1}{(-iz)+} \frac{1/2}{(-iz)+} \frac{1}{(-iz)+} \frac{3/2}{(-iz)+} \cdots \frac{\nu/2}{(-iz)} \\ &= \frac{i}{\pi} \frac{\mu_0}{z-} \frac{1/2}{z-} \frac{1}{z-} \frac{3/2}{z-} \cdots \frac{\nu/2}{z}, \end{aligned}$$

where the second expression follows from the first by an obvious equivalence transformation. Comparison with (2.8) shows that

$$v_0^{[\nu]}(z) = \frac{i}{\pi} \frac{q_{\nu+1}(z)}{p_{\nu+1}(z)}. \tag{2.22}$$

Curiously, the function $w_n(z)$ defined in (2.16) is related to the nth derivative of $w(z)$ by

$$w^{(n)}(z) = (2i)^n n! w_n(z), \qquad n = 0, 1, 2, \cdots, \tag{2.23}$$

a result apparently first observed in [5, p. 223].

3. Computational procedure. Our objective is to devise an efficient procedure for computing the function $w(z)$ defined in (2.1) to a given number d of correct decimal digits after the decimal point, i.e., to within an (absolute) error of $\frac{1}{2}10^{-d}$. We shall assume z to lie in the first quadrant Q_1 of the complex plane. This is no restriction of generality, since

$$w(-z) = 2e^{-z^2} - w(z), \qquad w(\bar{z}) = \overline{w(-z)} \tag{3.1}$$

can be used to continue w into the remaining quadrants.

As shown in (2.12), (2.14), Gauss–Hermite quadrature, or equivalently, the Laplace continued fraction (2.7), provides an effective means of computing $w(z)$ for $z \in Q_1$ and $|z|$ large. To obtain a more concrete idea as to the errors involved, we construct the altitude map of the meromorphic function

$$e_n(z) = w(z) - \frac{i}{\pi} \sum_{k=1}^{n} \frac{\lambda_k^{(n)}}{z - t_k^{(n)}},$$

i.e., the curves of constant modulus $|e_n(z)| = r$. These may be obtained by numerical integration of the differential equations

$$\frac{dx}{d\phi} = -\operatorname{Im}\left[\frac{e_n(z)}{e_n'(z)}\right], \qquad \frac{dy}{d\phi} = \operatorname{Re}\left[\frac{e_n(z)}{e_n'(z)}\right], \qquad z = x + iy,$$

subject to the initial conditions

$$x(0) = 0, \qquad y(0) = \eta,$$

where η is the root of $|e_n(iy)| = r$. Selected results are shown in Fig. 1, where $n = 9$ and $r = \frac{1}{2}10^{-d}$, $d = 2(2)10$.

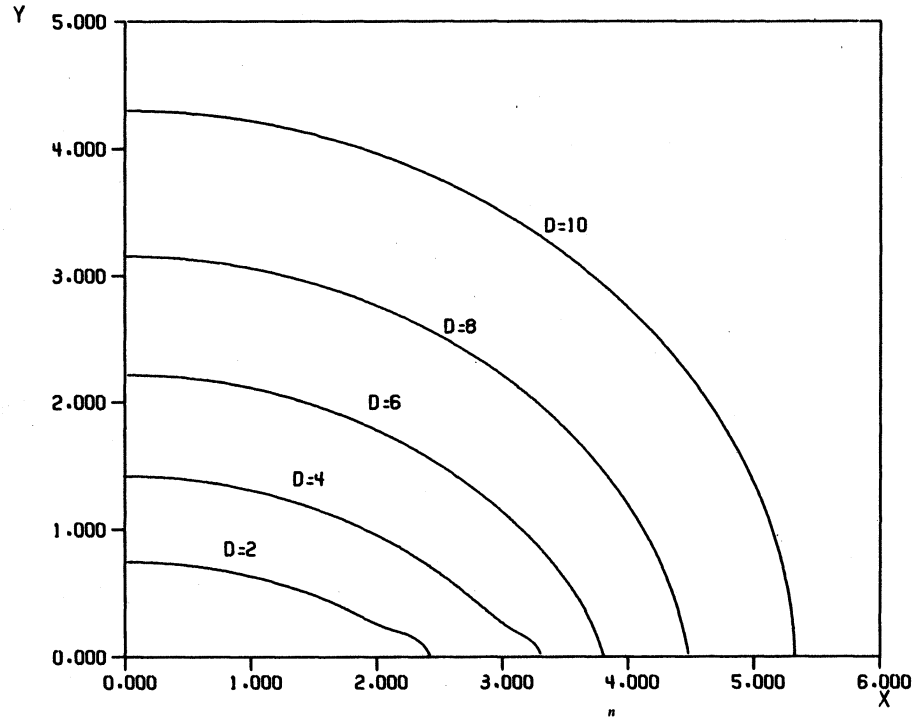


FIG. 1. *Altitude map of the function* $w(z) - (i/\pi) \sum_{k=1}^{n} \lambda_k^{(n)}/(z - t_k^{(n)})$, $n = 9$

Given any d, it is obviously possible to construct a rectangular region

$$R: \quad 0 \leqq x \leqq x_0, \qquad 0 \leqq y \leqq y_0 \tag{3.2}$$

outside of which 9-point Gauss–Hermite quadrature yields an accuracy of d decimal places. For $d = 10$, e.g., Fig. 1 suggests the choice $x_0 = 5.33$, $y_0 = 4.29$. Larger values of n would result in a smaller rectangle R, whereas smaller values of n would require a larger rectangle R for the same accuracy. It does not seem possible, in general, to arrive at an optimal choice of n, as such a choice would depend on the

relative frequency with which the procedure is used in various regions of the complex plane. The choice $n = 9$ appears to be a reasonable compromise, and we shall fix this value for what follows.

To compute $w(z)$ outside of R, we thus apply (2.19) with $\nu = 8$ and $N = 0$. In view of (2.22), (2.11), this is indeed equivalent to evaluating $w(z)$ from the integral representation (2.2), (2.3) by a 9-point Gauss–Hermite quadrature rule.

It remains to consider the case where $z \in R$. If $y = \operatorname{Im} z$ is relatively small, a common procedure consists of computing $w(z)$ from a Taylor expansion about $z_0 = \operatorname{Re} z$, or alternatively, to write

$$w(z) = e^{-z^2} + \frac{2i}{\sqrt{\pi}} F(z), \qquad F(z) = e^{-z^2} \int_0^z e^{t^2}\, dt,$$

and to expand $F(z)$ about $z_0 = \operatorname{Re} z$. There are two disadvantages to this approach:

(i) it requires the computation of Dawson's integral, $F(x)$, for $x = z_0$. Although good rational approximations are available for $F(x)$ (see, e.g., [3]), the necessity of computing $F(x)$ is apt to increase both the length of the program, and the total machine time, for computing $w(z)$;

(ii) the recursive computation of the expansion coefficients is subject to considerable loss of accuracy, particularly for large $x > 0$.

Interestingly enough, both these defects are removed if one expands "downward" rather than "upward", i.e., if one computes $w(z)$ from the Taylor expansion

$$w(z) = \sum_{n=0}^{\infty} \frac{w^{(n)}(z + ih)}{n!}(-ih)^n, \tag{3.3}$$

where $h > 0$ is suitably chosen. This approach has the further advantage of being related to the Laplace continued fraction approach, and in fact gives rise to an algorithm which generalizes algorithm (2.19) (used outside of R).

We observe from (2.23) that (3.3) can be written in the form

$$w(z) = \sum_{n=0}^{\infty} (2h)^n w_n(z + ih). \tag{3.4}$$

Approximating w_n by $v_n^{[\nu]}$ [cf. (2.20)], and truncating the infinite series, we are led to define

$$\sigma_N^{[\nu]}(z, h) = \sum_{n=0}^{N} (2h)^n v_n^{[\nu]}(z + ih). \tag{3.5}$$

Letting $t_n = (2h)^n v_n^{[\nu]}(z + ih)$, one obtains from (2.19) the following algorithm to compute (3.5),

$$\begin{aligned} &r_\nu = 0, \quad r_{n-1} = \frac{1/2}{h - iz + (n+1)r_n}, && n = \nu, \nu - 1, \cdots, 0, \\ &t_0 = \sigma_0 = \frac{2}{\sqrt{\pi}} r_{-1}, \\ &t_n = 2hr_{n-1}t_{n-1}, \quad \sigma_n = \sigma_{n-1} + t_n, && n = 1, 2, \cdots, N. \end{aligned} \tag{3.6}$$

We note that for $h = 0$, $N = 0$, algorithm (3.6) reduces to algorithm (2.19).

Given $\varepsilon = \frac{1}{2}10^{-d}$, it is clearly possible to determine N and ν (both depending on z, h, and ε) such that

$$|\sigma_N^{[\nu]}(z,h) - w(z)| \leqq \varepsilon. \tag{3.7}$$

Since the series in (3.4) converges for every z and h, we can indeed find N such that $|\sigma_N^{[\infty]}(z,h) - w(z)| \leqq \varepsilon/2$, and with N so determined, find $\nu > N$ such that $|\sigma_N^{[\nu]}(z,h) - \sigma_N^{[\infty]}(z,h)| \leqq \varepsilon/2$. The triangular inequality then yields the desired result (3.7).

Efficiency being one of our major concerns, we propose to

(i) let h, N and ν depend on z in such a way that $h = N = 0$, $\nu = 8$ for z outside of R, the rectangle introduced in (3.2);

(ii) empirically determine the smallest integers N and ν, subject to (i) and compatible with (3.7), for each $z \in R$.

The motivation behind the first of these objectives is to arrive at a *single* algorithm for computing $w(z)$ in all of Q_1, viz. algorithm (3.6), which, as was already observed, automatically reduces to the Laplace continued fraction algorithm (2.19) when $h = N = 0$. The objective can be attained in many different ways. Exploratory computations led us to set up h, N, ν tentatively in the form[1]

$$\begin{aligned} &h = h_0 s(z), \quad N = \{N_0 + N_1 s(z)\}, \quad \nu = \{\nu_0 + \nu_1 s(z)\} && \text{if } z \in R, \\ &h = N = 0, \qquad\qquad\qquad\quad \nu = 8 && \text{if } z \in Q_1 \backslash R, \end{aligned} \tag{3.8}$$

where

$$s(z) = \left(1 - \frac{y}{y_0}\right)\sqrt{1 - \left(\frac{x}{x_0}\right)^2} \qquad z = x + iy,$$

and $h_0, N_0, N_1, \nu_0, \nu_1$ are parameters which remain to be determined. Our second objective (ii) will serve to determine the last four of these parameters, while the first, h_0, will be chosen so as to minimize machine time.

A basic aid in this parameter study is a gauging routine Γ, which does the following: given $z = x + iy$, h, and ε, it returns nearly optimal values $N = N_\Gamma$, $\nu = \nu_\Gamma$ compatible with (3.7). The detailed steps involved in Γ are as follows:

(a) Select $N = N_{\max}$ sufficiently large (say, $N_{\max} = 40$).

(b) Determine $\nu_{\min}$, the smallest integer $\nu > N_{\max}$ such that $\max_{0 \leqq N \leqq N_{\max}} |\sigma_N^{[\nu+10]}(z,h) - \sigma_N^{[\nu]}(z,h)| \leqq \varepsilon/100$. The quantities $\sigma_N^{[\nu_{\min}]}(z,h)$, $N = 0, 1, \cdots, N_{\max}$, are considered sufficiently accurate to represent true partial sums of the Taylor series (3.4).

(c) Find N_Γ as the smallest integer $N \leqq N_{\max} - 3$ satisfying $|\sigma_{N+3}^{[\nu_{\min}]}(z,h) - \sigma_N^{[\nu_{\min}]}(z,h)| \leqq \varepsilon$. If there is no such integer N, increase $N_{\max}$ by 10, and repeat steps (a)–(c).

(d) Determine ν_Γ as the smallest integer ν satisfying $|\sigma_{N_\Gamma}^{[\nu+1]}(z,h) - \sigma_{N_\Gamma}^{[\nu]}(z,h)| \leqq \varepsilon$.

A first application of the gauging routine Γ is made in determining the parameter h_0. The choice of h_0 affects both the convergence of $\sigma_N^{[\nu]}(z,h)$, as $\nu \to \infty$, and the convergence of the Taylor expansion (3.4). In fact, large values of h_0 give rise to fast convergence of $\sigma_N^{[\nu]}(z,h)$, but slow convergence in (3.4), while

[1] $\{u\}$ denotes the integer closest to u, i.e., the largest integer contained in $u + 1/2$.

small values of h_0 yield slow convergence of $\sigma_N^{[\nu]}(z, h)$ (particularly if $y = \operatorname{Im} z$ is small), but fast convergence in (3.4). A good choice of h_0 is therefore one which strikes a balance between these two opposing effects. In order to search for such a value, we let $y = 0$ (where these effects are most pronounced) and apply Γ with input parameters $z = x$, $h = h_0\, s(x)$, ε, for selected values of x and h_0. After each application of Γ we measure the machine time required to compute $\sigma_N^{[\nu]}(x, h)$ by algorithm (3.6), where $\nu = \nu_\Gamma$, $N = N_\Gamma$ are the integers returned by Γ. With $x_0 = 5.33$, $\varepsilon = \frac{1}{2}10^{-10}$, the results are shown in Fig. 2, where machine time (in milliseconds) is plotted versus h_0 for $x = 0(1)5$. It is seen from these graphs that a good choice of h_0 is $h_0 = 1.6$.

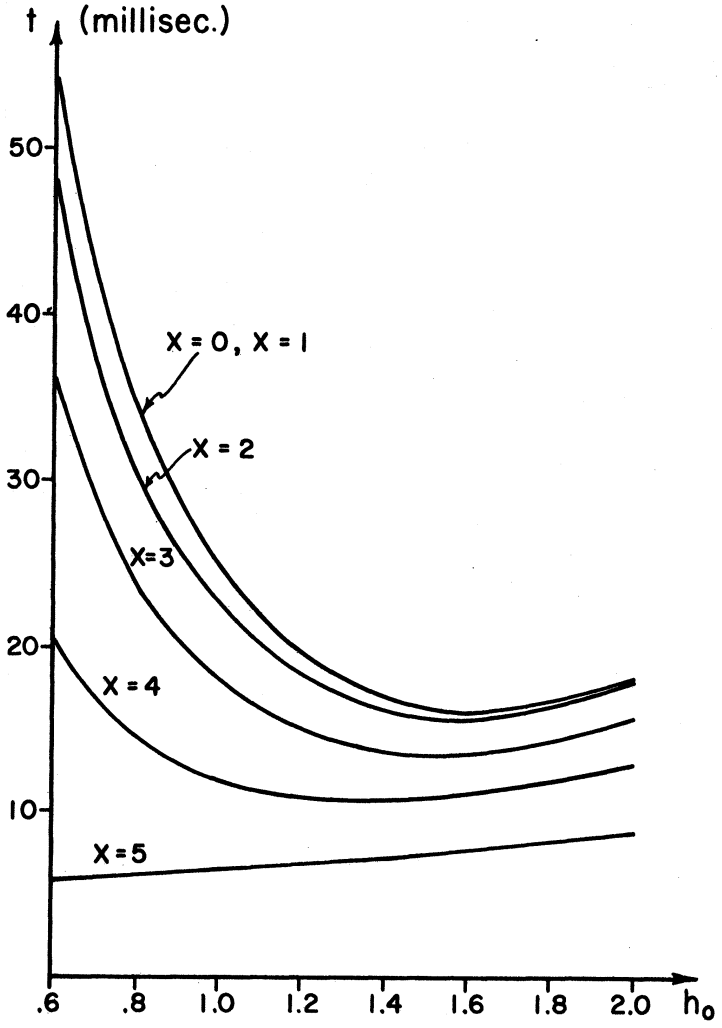


FIG. 2. *Machine time for algorithm* (3.6) *as a function of* h_0 *and* $z = x$

We next apply Γ to determine the parameters N_0, N_1, ν_0, ν_1 in (3.8). A first pair of constraints is obtained by letting $z = 0$ and requiring that

$$N_0 + N_1 = N_\Gamma^0, \qquad \nu_0 + \nu_1 = \nu_\Gamma^0, \tag{3.9}$$

where N_Γ^0, ν_Γ^0 are the results returned by Γ with input parameters $z = 0, h = h_0, \varepsilon$.

A narrow band near the separation line $y = y_0$ is then examined more carefully, since our preliminary computations indicated that N_Γ and v_Γ approach limits larger than 0 and 8, respectively, as $y \uparrow y_0$. With N_Γ^* the largest N_Γ, and v_Γ^* the largest v_Γ returned by Γ for input parameters z (near the line $y = y_0$), $h = h_0 s(z)$, ε, we let

$$N_0 = N_\Gamma^*, \qquad v_0 = v_\Gamma^*, \tag{3.10}$$

which, together with (3.9), determines the desired parameters uniquely. In our case of interest ($x_0 = 5.33$, $y_0 = 4.29$, $h_0 = 1.6$, $\varepsilon = \frac{1}{2}10^{-10}$), the results are $N_\Gamma^0 = 29$, $v_\Gamma^0 = 30$, $N_\Gamma^* = 6$, $v_\Gamma^* = 9$, giving

$$h = 1.6s(z), \quad N = \{6 + 23s(z)\}, \quad v = \{9 + 21s(z)\} \qquad \text{if } z \in R. \tag{3.11}$$

Note that $v > N$ for all z, as required in algorithm (3.6).

It remains to examine whether this choice of parameters is indeed compatible with (3.7). This is done by applying Γ with input parameters $z = x + iy$, $h = h_0 s(z)$, ε over a grid of points $z \in R$, and by checking the inequalities

$$\{N_0 + N_1 s(z)\} \geqq N_\Gamma(z), \qquad \{v_0 + v_1 s(z)\} \geqq v_\Gamma(z),$$

where $N_\Gamma(z)$ and $v_\Gamma(z)$ are the output values of Γ at the point z. Using the grid $x = 0(.5)5(.05)5.4$, $y = 0(.2)4(.05)4.4$, and $\varepsilon = \frac{1}{2}10^{-10}$, it is found that both inequalities are consistently satisfied.

With the values of h, N, v, defined in (3.8), (3.11), the desired function $w(z)$ is thus approximated by $\sigma_N^{[v]}(z, h)$ in (3.5), which in turn can be calculated by algorithm (3.6). The result is essentially the algorithm in [8], except that in this procedure the sum for $\sigma_N^{[v]}(z, h)$ is evaluated somewhat differently. Letting $s_n = [v_N^{[v]}(z + ih)]^{-1} \sum_{k=n+1}^{N} (2h)^k v_k^{[v]}(z + ih)$, $s_N = 0$, one can write indeed

$$\left.\begin{aligned} r_v &= 0, \qquad s_N = 0, \\ r_{n-1} &= \frac{1/2}{h - iz + (n+1)r_n}, \\ s_{n-1} &= r_{n-1}[(2h)^n + s_n], \qquad \text{if } n \leqq N \end{aligned}\right\} \quad n = v, v-1, \cdots, 0, \tag{3.12}$$

and then has

$$\sigma_N^{[v]}(z, h) = \begin{cases} \dfrac{2}{\sqrt{\pi}} s_{-1}, & h > 0, \\ \dfrac{2}{\sqrt{\pi}} r_{-1}, & h = 0. \end{cases} \tag{3.13}$$

The advantage of this algorithm over algorithm (3.6) is its increased speed on a digital computer [cf. § 4] together with the fact that no array of storage must be provided to hold the quantities r_{n-1}, $n = 1, 2, \cdots, N$.

4. Performance characteristics and tests. We begin by comparing the computing times of the algorithm in [8] (referred to below as "Algorithm 363") with those of a similar algorithm, in which (3.12), (3.13) is replaced by (3.6). Both algorithms are compiled and executed on the CDC 6500 computer.

We select five layers S_n, $n = 0, 1, \cdots, 4$, in R, defined by

$$S_n: \quad 0 \leqq x \leqq x_0, \qquad ny_0/5 \leqq y \leqq (n+1)y_0/5, \qquad x_0 = 5.33, \quad y_0 = 4.29,$$

and time the algorithms on each S_n. The average time on S_n is obtained by measuring the computing time of evaluating $w(z)$ for 1000 values of z, distributed uniformly in S_n, and by dividing the measured time by 1000. Both algorithms are timed similarly in the region outside of R (where computing time is independent of z). The results[2] are shown in Table 1. It is seen that the second algorithm is slower than the first by a factor of 1.6 to 2.2.

TABLE 1
Timing of Algorithm 363 and a related algorithm

z in	Computing time (in millisec.)	
	Algorithm 363	(3.6) replacing (3.12), (3.13)
S_0	6.7	14.5
S_1	6.0	12.6
S_2	5.2	10.7
S_3	4.4	8.7
S_4	3.6	6.8
$Q_1 \backslash R$	2.2	3.6

For comparison we also timed the library subroutine for the exponential function e^x for selected values of x in the interval $1 \leqq x \leqq 20$. The time observed was rather consistently .315 milliseconds. The computation of $w(z)$ (both real and imaginary part) by Algorithm 363 thus takes about as long as 7 to 21 exponentiations, depending on the location of the argument z.

In order to gain further confidence in the accuracy claimed, Algorithm 363 is run for $x = 0(.02)5.32(.005)5.35,\ 5.4(.2)6$, $y = 0(.02)4.28(.005)4.31,\ 4.4(.2)5$, the results being compared with those obtained by the same algorithm, where h is replaced by 1.6, N by 33, and ν by 36. (This combination of parameter values yields 14 correct decimal digits for z near the origin.) The largest absolute deviation is found to be 5.18×10^{-11} in the real part, and 4.91×10^{-11} in the imaginary part, suggesting that 10 decimal accuracy has indeed been attained.

Although we limited ourselves to an absolute error criterion, the relative error is not substantially larger for $z \in Q_1$, since $w(0) = 1$ and $|w(z)|$ decreases slowly (like $\pi^{-1/2}|z|^{-1}$) as $|z|$ increases. In fact, as $z \to \infty$ in Q_1, the relative error $\rho(z)$ is asymptotically given by $\rho(z) \sim 9!2^{-9}z^{-18}$, as can be seen from (2.14), with $n = 9$, and (2.6). Thus, e.g., $|\rho(z)| \doteq 1.03 \times 10^{-8}$ if $|z| = 4$, and $|\rho(z)| \doteq 1.86 \times 10^{-10}$ if $|z| = 5$. These estimates, it must be noted, do not necessarily apply to the relative errors in the real and imaginary parts individually. For example, $\operatorname{Im} w(iy) = 0$ for real y, which implies a large relative error in $\operatorname{Im} w(z)$ when z is very close to the imaginary axis. Similarly, $\operatorname{Re} w(x) = e^{-x^2}$, x real, implying a large relative error in $\operatorname{Re} w(z)$ for z very close to the real axis and $\operatorname{Re} z$ large.

[2] Such timings are subject to slight variations, even on the same computer, due to such incidental factors as compiler, executive system, clock reading routine, etc.

To illustrate the last remark, we use our algorithm to compute $\pi^{-1/2}$ Re $w(x+iy)$ for $x = 0(1)10$ and $y = s \cdot 10^{-r}$, $s = 1, 2, 3, 5, 7$, $r = 4, 3, 2$, and compare the results with those in Hummer's table [9]. Although some of the answers (for large x and small y) have order of magnitude 10^{-7}, there is still agreement to 8 significant digits (the precision in [9]), excepting occasional end figure errors of 1 to 7 units.

As a final note, we observe that loss of significant accuracy may also occur in the use of (3.1), since $w(-z)$ has zeros in the first quadrant Q_1, as may be inferred from the altitude chart in [6, p. 298].

Acknowledgments. The author is indebted to Professor Henry C. Thacher, Jr., for stimulating conversations, and to Mr. Thomas J. Aird for writing the program to produce the altitude map of Fig. 1.

REFERENCES

[1] B. H. ARMSTRONG, *Spectrum line profiles: the Voigt function*, J. Quant. Spectrosc. Radiat. Transfer, 7 (1967), pp. 61–88.

[2] C. CHIARELLA AND A. REICHEL, *On the evaluation of integrals related to the error function*, Math. Comp., 22 (1968), pp. 137–143.

[3] W. J. CODY, K. A. PACIOREK AND H. C. THACHER, JR., *Chebyshev approximations for Dawson's integral*, Ibid., to appear.

[4] V. N. FADDEEVA AND N. N. TERENT'EV, *Tables of values of the function* $w(z) = e^{-z^2}\left(1 + \frac{2i}{\sqrt{\pi}}\int_0^z e^{t^2}\,dt\right)$ *for complex argument*, Gosud. Izdat. Teh.-Teor. Lit., Moscow, 1954; English transl., Pergamon Press, New York, 1961.

[5] A. FLETCHER, J. C. P. MILLER AND L. ROSENHEAD, *An Index of Mathematical Tables*, Scientific Computing Service Limited, London, 1946.

[6] W. GAUTSCHI, *Error function and Fresnel integrals*, Handbook of Mathematical Functions, M. Abramowitz and I. A. Stegun, eds., Nat. Bur. Standards Appl. Math. Ser., vol. 55, 1964, pp. 295–329.

[7] ———, *Computational aspects of three-term recurrence relations*, SIAM Rev., 9 (1967), pp. 24–82.

[8] ———, *Algorithm 363—Complex error function*, Comm. ACM., 12 (1969), p. 635.

[9] D. G. HUMMER, *The Voigt function: an eight-significant-figure table and generating procedure*, Mem. Roy. Astronom. Soc., 70 (1965), pp. 1–31.

[10] O. PERRON, *Die Lehre von den Kettenbrüchen*, vol. II, 3rd ed., Teubner, Stuttgart, 1957.

[11] A. REICHEL, *Error estimates in simple quadrature with Voigt functions*, Math. Comp., 21 (1967), pp. 647–651.

[12] T. J. STIELTJES, *Reserches sur les fractions continues*, Ann. Fac. Sci. Univ. Toulouse, 8 (1894), pp. 1–122; 9 (1895), pp. 1–47; also in Oevres complètes de Thomas Jan Stieltjes, vol. II, P. Noordhoff, Groningen, 1918, pp. 402–566.

[13] O. SZÁSZ, *Bemerkungen zu Herrn Perrons Erweiterung eines Markoffschen Satzes über die Konvergenz gewisser Kettenbrüche*, Math. Ann., 76 (1915), pp. 301–314.

[14] H. C. THACHER, JR., *Computation of the complex error function by continued fractions*, Blanch Anniversary Volume, Aerospace Research Laboratory, U.S. Air Force, Washington, D.C., 1967, pp. 315–337.

9.5. [47] "A HARMONIC MEAN INEQUALITY FOR THE GAMMA FUNCTION"

---

[47] "A Harmonic Mean Inequality for the Gamma Function," *SIAM J. Math. Anal.* **5**, 278–281 (1974).

SIAM J. MATH. ANAL.
Vol. 5, No. 2, April 1974

A HARMONIC MEAN INEQUALITY FOR THE GAMMA FUNCTION*

WALTER GAUTSCHI†

Abstract. We prove that the harmonic mean of $\Gamma(x)$ and $\Gamma(1/x)$ is greater than or equal to $\Gamma(1) = 1$ for arbitrary $x > 0$.

1. Introduction. V. R. Rao Uppuluri [2] brought the following conjectured inequality to the author's attention:

$$\frac{2}{1/(\Gamma(x)) + 1/(\Gamma(1/x))} \geqq 1 \quad \text{on } 0 < x < \infty. \tag{1.1}$$

It states that the harmonic mean of $\Gamma(x)$ and $\Gamma(1/x)$ is always larger than or equal to $\Gamma(1) = 1$, equality being assumed for $x = 1$. Because of the well-known inequalities between the harmonic, geometric and arithmetic means, the conjecture implies these other inequalities, $\Gamma(x)\Gamma(1/x) \geqq 1$ and $\Gamma(x) + \Gamma(1/x) \geqq 2$.

The proof of (1.1), given below in §§ 2–5, is "computational" in the sense that it relies on certain isolated numerical values of the psi function

$$\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$$

and its derivative. This deficiency, however, is removed in § 6, where numerical values of only standard constants, such as π, ln 2, and Euler's constant γ, are required.

It suffices to prove (1.1) for $1 < x \leqq x_0$, where $x_0 = 1.4616\cdots$ is the positive minimum point of $\Gamma(x)$. In fact, the left-hand expression in (1.1) is clearly increasing on the interval (x_0, ∞). If we prove the inequality for $1 < x \leqq x_0$, it will hold for all $x > 1$, hence also for all positive $x < 1$, on account of its invariance under the substitution $x \to 1/x$.

2. Reformulation of the inequality. Letting

$$\phi(t) = \frac{1}{\Gamma(e^t)}, \qquad -\infty < t < \infty, \tag{2.1}$$

we may rewrite (1.1) in the form

$$\tfrac{1}{2}[\phi(t) + \phi(-t)] \leqq \phi(0), \tag{2.2}$$

which expresses a "symmetric concavity" property for ϕ. We must prove (2.2) for $0 < t \leqq \ln x_0$.

Using Taylor's theorem, we have for $t > 0$,

$$\tfrac{1}{2}[\phi(t) + \phi(-t)] - \phi(0) = \frac{t^2}{4}[\ddot{\phi}(\tau_1) + \ddot{\phi}(\tau_2)],$$

* Received by the editors May 24, 1972.

† Department of Computer Sciences, Purdue University, Lafayette, Indiana 47907.

where dots denote derivatives with respect to t, and

$$0 < \tau_1 < t, \quad -t < \tau_2 < 0. \tag{2.3}$$

We will show that

$$\ddot{\phi}(\tau_1) + \ddot{\phi}(\tau_2) < 0 \quad \text{on } 0 < t \leqq \ln x_0. \tag{2.4}$$

3. The second derivative of ϕ. Differentiating (2.1) we obtain

$$\dot{\phi}(t) = -e^t \frac{\Gamma'(e^t)}{[\Gamma(e^t)]^2} = -xy(x),$$

where

$$x = e^t, \quad y(x) = \frac{\psi(x)}{\Gamma(x)}.$$

Another differentiation gives

$$\ddot{\phi}(t) = \left(\frac{d}{dx}\dot{\phi}\right)\frac{dx}{dt} = x(-xy)' = -x(y + xy'),$$

where primes indicate differentiation with respect to x. Noting that

$$y\Gamma = \psi,$$
$$y'\Gamma = \psi' - y\Gamma' = \psi' - y\Gamma\psi = \psi' - \psi^2,$$

we may express the second derivative of ϕ in terms of ψ and ψ',

$$\ddot{\phi}(t) = -\frac{1}{\Gamma(x)}(x\psi + x^2\psi' - x^2\psi^2), \qquad x = e^t. \tag{3.1}$$

4. Some monotonicity properties. We now observe that both functions $x\psi(x)$ and $x^2\psi'(x)$ are monotonically increasing on the interval $1/x_0 < x < x_0$. To see this for the first, we use the known expansion [1]

$$x\psi(x) = -1 + (1 - \gamma)x + \sum_{m=1}^{\infty} \frac{x(x-1)}{(m+1)(m+x)}, \tag{4.1}$$

where $\gamma = .5772\cdots$ is Euler's constant. One checks that for $m > 0$,

$$\frac{x(x-1)}{m+x}$$

is monotonically increasing for $x > (1 + \sqrt{1 + 1/m})^{-1}$, hence in particular for $x > 1/x_0$. Monotonicity of $x\psi$ thus follows from (4.1). Since $\psi(x_0) = 0$, we also see that

$$x\psi(x) < 0 \quad \text{on } 1/x_0 < x < x_0.$$

For $x^2\psi'$, our assertion follows directly from

$$x^2\psi'(x) = \sum_{m=0}^{\infty} \left(\frac{x}{m+x}\right)^2,$$

$x/(m+x)$, for each $m \geqq 1$, being monotonically increasing for $x > 0$. We also note that

$$x^2\psi'(x) > 0 \quad \text{on } 1/x_0 < x < x_0.$$

5. Conclusion of the proof. We are now in a position to estimate the second derivative of ϕ in (3.1), first on the interval $1/x_0 < x < x_0$, then on $1 < x < x_0$.

On the first interval we have by the monotonicity properties of § 4,

$$x\psi + x^2\psi' - x^2\psi^2 \geqq x_0^{-1}\psi(1/x_0) + x_0^{-2}\psi'(1/x_0) - x_0^{-2}\psi^2(1/x_0). \tag{5.1}$$

Using linear interpolation in [1, Table 6.1] we find $\psi(1/x_0) = \psi(1 + 1/x_0) - x_0 = -1.2657\cdots$, $\psi'(1/x_0) = \psi'(1 + 1/x_0) + x_0^2 = 2.9392\cdots$, so that the lower bound in (5.1) is $-.2400\cdots$. From (3.1), since $\Gamma(x_0) = .8856\cdots$, we thus obtain

$$\ddot{\phi}(t) \leqq \frac{.2400\cdots}{\Gamma(x_0)} < .272 \quad \text{on } -\ln x_0 < t < \ln x_0.$$

On the second interval, similarly,

$$x\psi + x^2\psi' - x^2\psi^2 \geqq \psi(1) + \psi'(1) - \psi^2(1) = .7345\cdots, \tag{5.2}$$

where we have used [1]

$$\psi(1) = -\gamma, \quad \psi'(1) = \zeta(2) = \pi^2/6.$$

Thus,

$$\ddot{\phi}(t) \leqq \frac{-.7345\cdots}{\Gamma(1)} < -.734 \quad \text{on } 0 < t < \ln x_0.$$

The proof is now completed by recalling from (2.3) that

$$0 < \tau_1 < \ln x_0, \quad -\ln x_0 < \tau_2 < 0,$$

and hence

$$\ddot{\phi}(\tau_1) + \ddot{\phi}(\tau_2) < -.734 + .272 = -.462 < 0,$$

as we set out to show in (2.4).

6. A less computational variant of the proof. Reference to numerical values of $\psi(1/x_0)$ and $\psi'(1/x_0)$ in (5.1) can be avoided by observing that $x_0^{-1} > \frac{1}{2}$ and that $x\psi$ and $x^2\psi'$ are monotonically increasing on $\frac{1}{2} < x < x_0$. Using [1]

$$\psi(\tfrac{1}{2}) = -\gamma - 2\ln 2, \quad \psi'(\tfrac{1}{2}) = 3\zeta(2) = \pi^2/2,$$

we can thus write in place of (5.1),

$$x\psi + x^2\psi' - x^2\psi^2 \geqq \tfrac{1}{2}\psi(\tfrac{1}{2}) + \tfrac{1}{4}\psi'(\tfrac{1}{2}) - \tfrac{1}{4}\psi^2(\tfrac{1}{2}) = -.7118\cdots.$$

Together with the companion inequality (5.2), and (3.1), this gives

$$\ddot{\phi}(t)\Gamma(e^t) \leqq .712 \quad \text{on } -\ln x_0 < t < \ln x_0,$$
$$\ddot{\phi}(t)\Gamma(e^t) \leqq -.734 \quad \text{on } 0 < t < \ln x_0.$$

It follows, in particular, that

$$(6.1)\quad \ddot{\phi}(\tau_1) < 0 \quad \text{and} \quad \ddot{\phi}(\tau_1)\Gamma(e^{\tau_1}) + \ddot{\phi}(\tau_2)\Gamma(e^{\tau_2}) \leqq -.734 + .712 = -.022 < 0.$$

Were $\ddot{\phi}(\tau_2)$ negative or zero, our assertion (2.4) would follow immediately from the first inequality in (6.1). If $\ddot{\phi}(\tau_2)$ were positive, then $\ddot{\phi}(\tau_2)\Gamma(e^{\tau_2}) > \ddot{\phi}(\tau_2)\Gamma(e^{\tau_1})$, and the second inequality in (6.1) would give

$$0 > \ddot{\phi}(\tau_1)\Gamma(e^{\tau_1}) + \ddot{\phi}(\tau_2)\Gamma(e^{\tau_2}) > \Gamma(e^{\tau_1})[\ddot{\phi}(\tau_1) + \ddot{\phi}(\tau_2)],$$

that is again (2.4). Thus (2.4) is true in either case, and the proof, once more, is completed.

REFERENCES

[1] P. J. DAVIS, *Gamma function and related functions*, Handbook of Mathematical Functions, M. Abramowitz and I. A. Stegun, eds., NBS Applied Math. Ser., 55 (1964), pp. 253–293.

[2] V. R. RAO UPPULURI, Personal communication, April, 1972.

9.6. [48] "SOME MEAN VALUE INEQUALITIES FOR THE GAMMA FUNCTION"

---

[48] "Some Mean Value Inequalities for the Gamma Function," *SIAM J. Math. Anal.* **5**, 282–292 (1974).

SIAM J. MATH. ANAL.
Vol. 5, No. 2, April 1974

SOME MEAN VALUE INEQUALITIES FOR THE GAMMA FUNCTION*

In Memory of George E. Forsythe

WALTER GAUTSCHI†

Abstract. We determine the infimum of the harmonic mean of $\Gamma(x_1), \Gamma(x_2), \cdots, \Gamma(x_n)$ under the constraints $\prod_{k=1}^{n} x_k = 1$, all $x_k > 0$. We present numerical evidence for this infimum to be equal to $\Gamma(1) = 1$ if $n \leqq 8$, and show it to be less than 1 when $n > 8$. We also prove that the geometric mean of $\Gamma(x_1), \Gamma(x_2), \cdots, \Gamma(x_n)$ is always $\geqq 1$ under the same constraints, and that the geometric mean is the power mean with the smallest exponent for which this is true.

1. Introduction. In a recent note [1] we proved that the harmonic mean of $\Gamma(x)$ and $\Gamma(1/x)$ for $x > 0$ is never smaller than $\Gamma(1) = 1$, that is,

$$\frac{2}{1/\Gamma(x) + 1/\Gamma(1/x)} \geqq 1 \quad \text{for } 0 < x < \infty. \tag{1.1}$$

Equality, of course, is assumed when $x = 1$. We report here on attempts at generalizing (1.1) to more variables. A natural generalization would be $n/\sum_{k=1}^{n} 1/\Gamma(x_k) \geqq \Gamma((x_1 x_2 \cdots x_n)^{1/n})$, which, however, is readily dismissed as false by considering the case $n = 2$, $x_1 = 1$, x_2 large. More promising is the conjecture

$$\frac{n}{\sum_{k=1}^{n} 1/\Gamma(x_k)} \geqq 1 \quad \text{for all } x_k > 0 \text{ with } x_1 x_2 \cdots x_n = 1. \tag{1.2}$$

We present evidence that this inequality is in fact valid for $n = 1, 2, \cdots, 8$, but prove it to be false for $n \geqq 9$. We also determine the infimum of the expression on the left of (1.2) under the constraints listed in (1.2). We next show that for all $n \geqq 1$ we have

$$\left[\prod_{k=1}^{n} \Gamma(x_k)\right]^{1/n} \geqq 1 \quad \text{for all } x_k > 0 \text{ with } x_1 x_2 \cdots x_n = 1. \tag{1.3}$$

In terms of the power means

$$M_n^{[r]}(a_i) = \left(\frac{a_1^r + a_2^r + \cdots + a_n^r}{n}\right)^{1/r}, \tag{1.4}$$

the last inequality may be restated as $M_n^{[0]}(\Gamma(x_i)) \geqq 1$, for all $n \geqq 1$, and all $x_i > 0$ with $x_1 x_2 \cdots x_n = 1$. Since $M_n^{[r]}$ increases monotonically with r, the same statement holds for any power mean with $r \geqq 0$. We show, on the other hand, that the statement is false for any power mean with $r < 0$.

2. Main results. We denote by R^n the space of real vectors $\mathbf{x}^T = [x_1, x_2, \cdots, x_n]$ and by R_+^n the positive orthant $R_+^n = \{\mathbf{x} \in R^n : x_k > 0, k = 1, 2, \cdots, n\}$.

* Received by the editors July 18, 1972, and in revised form October 22, 1972.

† Department of Computer Sciences, Purdue University, Lafayette, Indiana 47907. This research was performed at the U.S.A.F. Aerospace Research Laboratories under Contract F33615-71-C-1463 with Technology Incorporated.

The constraints in (1.2) can then be written as

$$\mathbf{x} \in S_n, \quad \text{where } S_n = \left\{ \mathbf{x} \in R_+^n : \prod_{k=1}^{n} x_k = 1 \right\}.$$

Our main results are as follows.

THEOREM 1. *For $n = 1, 2, 3, \cdots$ we have*

$$\inf_{\mathbf{x} \in S_n} \frac{n}{\sum_{k=1}^{n} 1/\Gamma(x_k)} = \frac{1}{\gamma_n}, \tag{2.1}$$

where

$$\gamma_n = \max_{1 \leqq v \leqq n-1} \{ \max_{0 \leqq x \leqq 1} g_{n,v}(x) \}, \qquad g_{n,v}(x) \overset{\text{def}}{=} \frac{1}{n} \left[\frac{v}{\Gamma(x)} + \frac{n-v}{\Gamma(x^{-v/(n-v)})} \right]. \tag{2.2}$$

Moreover,

$$\gamma_n \to \frac{1}{\Gamma(x_0)} = 1.1291 \cdots \quad \text{as } n \to \infty, \tag{2.3}$$

where $x_0 = 1.4616 \cdots$ is the unique point at which $\Gamma(x)$ attains its minimum on the positive x-axis.

Equation (2.1), for $n = 1$ with $\gamma_1 = 1$, is trivial, since the only point x_1 satisfying the constraints is $x_1 = 1$. For $n > 1$, the maxima in (2.2) can easily be computed with the aid of a digital computer. It turns out (cf. § 5) that $\gamma_n = 1$ for $1 \leqq n \leqq 8$, and it will be shown that $\gamma_n > 1$ for $n \geqq 9$. The conjecture (1.2) thus seems true for $n \leqq 8$, but is certainly false for all $n \geqq 9$.

We also note from (2.3) that in the (obvious) inequality

$$\frac{n}{\sum_{k=1}^{n} 1/\Gamma(x_k)} \geqq \Gamma(x_0) = .88560 \cdots, \qquad n = 1, 2, 3, \cdots, \tag{2.4}$$

the constant on the right is best possible under the constraints $\mathbf{x} \in S_n$.

THEOREM 2.[1] *For $n = 1, 2, 3, \cdots$, we have*

$$\left[\prod_{k=1}^{n} \Gamma(x_k) \right]^{1/n} \geqq 1 \quad \text{for all } \mathbf{x} \in S_n. \tag{2.5}$$

THEOREM 3. *For the power means $M_n^{[r]}$ defined in* (1.4) *we have*

$$M_n^{[r]}(\Gamma(x_i)) \geqq 1 \quad \text{on } S_n \quad \text{for all } n \geqq 1 \tag{2.6}$$

if and only if $r \geqq 0$.

3. Auxiliary propositions. We need a few elementary properties of the psi function $\psi(x) = \Gamma'(x)/\Gamma(x)$ and some related functions.

PROPOSITION 1. *The function $x\psi(x)$ is convex for $x > 0$.*

Proof. We have

$$(x\psi)'' = (1/x^2)(x^3\psi'' + 2x^2\psi').$$

[1] An examination of inequality (2.5) was suggested to the author by Professor R. A. Askey.

From the known expansion

$$\psi(x) = -\frac{1}{x} + 1 - \gamma + \sum_{m=1}^{\infty} \frac{x-1}{(m+1)(m+x)}, \tag{3.1}$$

where $\gamma = .57721\cdots$ is Euler's constant, we obtain by two differentiations,

$$x^3\psi'' + 2x^2\psi' = -2 - 2\sum_{m=1}^{\infty}\left(\frac{x}{m+x}\right)^3 + 2 + 2\sum_{m=1}^{\infty}\left(\frac{x}{m+x}\right)^2$$

$$= 2x^2 \sum_{m=1}^{\infty} \frac{m}{(m+x)^3} > 0,$$

i.e., $(x\psi)'' > 0$ for $x > 0$.

The next result concerns the function

$$f(x) = x[\Gamma(x)]^r\psi(x) = \frac{x}{r}\frac{d}{dx}\{[\Gamma(x)]^r\}, \qquad r < 0, \tag{3.2}$$

where r is a fixed (negative) parameter. By x_0 we denote, as before, the abscissa of the minimum of $\Gamma(x)$.

PROPOSITION 2. *The function f in* (3.2) *vanishes at $x = 0$ and $x = x_0$, and is negative and unimodal on $0 < x < x_0$, i.e., there exists a ξ with $0 < \xi < x_0$ such that f decreases on $0 < x < \xi$ and increases on $\xi < x < x_0$.*

Proof. From the known power series expansion of $1/\Gamma(x)$, letting $\rho = |r|$, we find

$$f(x) = -x^{\rho} - (\rho+1)\gamma x^{\rho+1} + \cdots,$$

showing that $f(0) = 0$. (It is also seen, incidentally, that f need not be convex on $0 < x < x_0$; for example, if $\rho = 1$, then $f''(0) = -4\gamma < 0$.) By definition of x_0, we also have $f(x_0) = 0$.

To prove unimodality, we look at the derivative f'. A simple computation gives

$$x[\Gamma(x)]^{-r}f'(x) = x\psi(x) + rx^2\psi^2(x) + x^2\psi'(x). \tag{3.3}$$

Let

$$u(x) = x\psi(x) + rx^2\psi^2(x) + x^2\psi'(x). \tag{3.4}$$

From the power series expansion of $\psi(x+1)$, we obtain

$$x\psi(x) = x\psi(x+1) - 1 = -1 - \gamma x + \zeta(2)x^2 + \cdots.$$

This shows that the function $x\psi(x)$ decreases for small positive x; since it is convex by Proposition 1, and vanishes at $x = x_0$, it must have a unique minimum at some point ξ^* with $0 < \xi^* < x_0$. (In fact, $\xi^* \doteq .2161$.) As the derivative of $x\psi$ vanishes at this point, we have

$$\xi^*\psi'(\xi^*) + \psi(\xi^*) = 0. \tag{3.5}$$

We consider first the interval $0 < x < \xi^*$. On this interval, we have from (3.4), since $r < 0$ and $x^2\psi^2(x) > 1$,

$$u(x) < U(x), \quad U(x) = x\psi(x) + r + x^2\psi'(x), \qquad 0 < x < \xi^*. \tag{3.6}$$

Note that

$$U(0) = -1 + r + 1 = r < 0, \qquad U(\xi^*) = r < 0, \tag{3.7}$$

the second relation being a consequence of (3.5). We now show that $U(x)$ is convex on $0 < x < \xi^*$. By Proposition 1, it suffices to show that the last term in $U(x)$ is convex, i.e.,

$$(x^2\psi')'' = 2\psi' + 4x\psi'' + x^2\psi''' > 0, \qquad 0 < x < \xi^*. \tag{3.8}$$

Repeated differentiation of (3.1), however, gives

$$(x^2\psi')'' = \sum_{m=1}^{\infty} \frac{2m(m - 2x)}{(m + x)^4},$$

which is certainly positive if $0 < x < \frac{1}{2}$, hence, in particular, if $0 < x < \xi^*$. From (3.6), (3.7), and the convexity of $U(x)$ just established, it now follows that $u(x) < r < 0$ on $0 < x < \xi^*$, i.e., by virtue of (3.3), (3.4),

$$f'(x) < 0 \quad \text{on } 0 < x < \xi^*. \tag{3.9}$$

On the remaining interval $\xi^* < x < x_0$, the function $x\psi(x)$, while still negative, increases monotonically. Since also $x^2\psi'(x)$ increases monotonically for $x > 0$ (cf. [1]), it follows from (3.4) that $u(x)$ is monotonically increasing on $\xi^* < x < x_0$. Moreover, $u(x_0) = x_0^2\psi'(x_0) > 0$. Hence there is a unique point ξ, with $\xi^* < \xi < x_0$, such that $u(\xi) = 0$, and thus $u(x) < 0$ for $0 < x < \xi$ and $u(x) > 0$ for $\xi < x < x_0$. In view of (3.3), (3.4), this implies unimodality of f.

4. Proof of Theorem 1. We assume $n \geqq 2$, since the case $n = 1$, as we pointed out, is trivial. For short, let

$$\gamma(x_1, x_2, \cdots, x_n) = \frac{1}{n} \sum_{k=1}^{n} \frac{1}{\Gamma(x_k)}.$$

Since obviously

$$\gamma(x_1, x_2, \cdots, x_n) \leqq \frac{1}{\Gamma(x_0)} \quad \text{for all } \mathbf{x} \in R_+^n,$$

the function γ is bounded from above in all of R_+^n, and hence, in particular, on S_n. We denote

$$\sigma_n = \sup_{\mathbf{x} \in S_n} \gamma(x_1, x_2, \cdots, x_n) < \infty. \tag{4.1}$$

We want to prove that

$$\sigma_n = \gamma_n.$$

We distinguish two major cases (not a priori mutually exclusive):

Case I. The supremum σ_n is "assumed at infinity", i.e., there exists a sequence of vectors $\mathbf{x}^{(r)} \in S_n$ such that

$$\|\mathbf{x}^{(r)}\| \to \infty, \quad \gamma(\mathbf{x}^{(r)}) \to \sigma_n \quad \text{as } r \to \infty. \tag{4.2}$$

By virtue of the first relation, and the fact that $\mathbf{x}^{(r)} \in S_n$, there must exist a subsequence of $\mathbf{x}^{(r)}$ for which at least one component tends to ∞ and another tends

to 0. Let us write again $\mathbf{x}^{(r)}$ for this subsequence, and for definiteness, assume that

$$x_1^{(r)} \to \infty, \quad x_2^{(r)} \to 0 \quad \text{as } r \to \infty. \tag{4.3}$$

Since

$$\begin{aligned} n\gamma(\mathbf{x}^{(r)}) &= \frac{1}{\Gamma(x_1^{(r)})} + \frac{1}{\Gamma(x_2^{(r)})} + \sum_{k=3}^{n} \frac{1}{\Gamma(x_k^{(r)})} \\ &\leqq \frac{1}{\Gamma(x_1^{(r)})} + \frac{1}{\Gamma(x_2^{(r)})} + \frac{n-2}{\Gamma(x_0)}, \end{aligned}$$

we obtain from (4.2), (4.3), by letting $r \to \infty$ in this inequality,

$$\sigma_n \leqq \frac{1 - 2/n}{\Gamma(x_0)}. \tag{4.4}$$

We show that equality holds in (4.4). Define $\mathbf{x}(t)$ by

$$x_1(t) = tc, \quad x_2(t) = c/t, \quad x_3 = \cdots = x_n = x_0, \qquad c = x_0^{-(n-2)/2}.$$

Clearly, $\mathbf{x}(t) \in S_n$ for all $t > 0$, and

$$n\gamma(\mathbf{x}(t)) = \frac{1}{\Gamma(tc)} + \frac{1}{\Gamma(c/t)} + \frac{n-2}{\Gamma(x_0)}.$$

Letting $t \to \infty$ gives $\gamma(\mathbf{x}(t)) \to (1 - 2/n)/\Gamma(x_0)$, and therefore strict inequality cannot hold in (4.4).

Thus, in Case I, we conclude that

$$\sigma_n = \frac{1 - 2/n}{\Gamma(x_0)}. \tag{4.5}$$

Case II. The supremum σ_n is assumed at a finite point $\mathbf{x} = \mathbf{s}$ of S_n,

$$\gamma(\mathbf{x}) \leqq \gamma(\mathbf{s}) \quad \text{for all } \mathbf{x} \in S_n. \tag{4.6}$$

The function γ thus has on S_n a global maximum at $\mathbf{s}$.

Using Lagrange multipliers, it follows that $\mathbf{s}^T = [s_1, s_2, \cdots, s_n]$ must be a solution of the system of equations

$$\frac{\partial}{\partial x_i}\left[\gamma(x_1, x_2, \cdots, x_n) + \lambda\left(\prod_{k=1}^{n} x_k - 1\right)\right] = 0, \quad i = 1, 2, \cdots, n,$$

$$\prod_{k=1}^{n} x_k - 1 = 0,$$

that is,

$$-\frac{\Gamma'(s_i)}{[\Gamma(s_i)]^2} + \lambda n \prod_{\substack{k=1 \\ k \neq i}}^{n} s_k = 0, \qquad i = 1, 2, \cdots, n, \tag{4.7}$$

$$\prod_{k=1}^{n} s_k = 1. \tag{4.8}$$

Multiplying the ith equation in (4.7) by s_i, and taking note of (4.8), we find

$$\lambda n = f(s_1) = f(s_2) = \cdots = f(s_n), \tag{4.9}$$

where the function f is as in (3.2), with $r = -1$. Since $f(x) < 0$ for $0 < x < x_0$ and $f(x) \geqq 0$ for $x \geqq x_0$, it follows from (4.9) that either all s_k are between 0 and x_0, or all s_k are $\geqq x_0$. The latter, however, is excluded by (4.8), since $x_0 > 1$. Consequently,

$$0 < s_k < x_0, \qquad k = 1, 2, \cdots, n. \tag{4.10}$$

Now, using Proposition 2 of § 3, according to which f is unimodal on $0 < x < x_0$, we conclude from (4.9) and (4.10) that only one of two situations can arise:

IIa. All s_k are the same. By (4.8), this implies $s_1 = s_2 = \cdots = s_n = 1$, and so $\sigma_n = \gamma(\mathbf{s}) = 1$ in this case.

IIb. There are exactly *two* distinct s_k, say,

$$s_1 = s_2 = \cdots = s_\nu < s_{\nu+1} = s_{\nu+2} = \cdots = s_n, \qquad 1 \leqq \nu < n,$$

$$s_1^\nu s_n^{n-\nu} = 1.$$

We then have

$$\sigma_n = \gamma(\mathbf{s}) = \frac{1}{n}\left[\frac{\nu}{\Gamma(s_1)} + \frac{n-\nu}{\Gamma(s_n)}\right], \qquad 0 < s_1 < s_n < x_0.$$

Since $s_n = s_1^{-\nu/(n-\nu)}$ and $s_1 < s_n$, it follows that $0 < s_1 < 1$. (In fact, $x_0^{-(n-\nu)/\nu} < s_1 < 1$, by virtue of $s_n < x_0$.) According to the definition of $g_{n,\nu}$ in (2.2), we thus have

$$\sigma_n = g_{n,\nu}(s_1), \qquad 0 < s_1 < 1. \tag{4.11}$$

Furthermore, by (4.9), s_1 is a solution of the equation

$$f(x) = f(x^{-\nu/(n-\nu)}). \tag{4.12}$$

One checks readily that the roots of (4.12) are precisely the stationary points of $g_{n,\nu}(x)$. Since

$$\gamma(\underbrace{x, x, \cdots, x}_{\nu\text{-times}}, \underbrace{y, y, \cdots, y}_{(n-\nu)\text{-times}}) = g_{n,\nu}(x), \qquad y = x^{-\nu/(n-\nu)},$$

where the argument of γ is a point on S_n for each $x > 0$, and since σ_n is the global maximum of γ on S_n, the stationary point (4.11) cannot be other than a local maximum. There are now two possibilities:

IIba. For no integer ν with $1 \leqq \nu \leqq n-1$ does $g_{n,\nu}(x)$ have a local maximum on $(0, 1)$.

IIbb. There is at least one integer ν, $1 \leqq \nu \leqq n-1$, for which $g_{n,\nu}(x)$ has a local maximum on $(0, 1)$.

Case IIba is incompatible with Case IIb, so that IIa necessarily applies, and $\sigma_n = 1$. In Case IIbb we must look for the largest local maximum (if there are several, corresponding to different values of ν), which is then equal to σ_n if larger than 1. Otherwise, $\sigma_n = 1$ from Case IIa.

Summarizing Case II, we can write

$$\sigma_n = \max_{1 \leqq \nu \leqq n-1} \{\max_{0 \leqq x \leqq 1} g_{n,\nu}(x)\} = \gamma_n,$$

where the inner maximum picks up a local maximum of $g_{n,\nu}$, if it is larger than 1, or the value $g_{n,\nu}(1) = 1$, if it is less than 1 or nonexistent. With Case I, equation (4.5),

taken into account, we thus have

$$\sigma_n = \max\left[\frac{1-2/n}{\Gamma(x_0)}, \gamma_n\right].$$

Observing, however, that

$$\gamma_n \geqq \max_{0\leqq x\leqq 1} g_{n,1}(x) \geqq g_{n,1}(x_0^{-(n-1)}) = \frac{1}{n}\left[\frac{1}{\Gamma(x_0^{-(n-1)})} + \frac{n-1}{\Gamma(x_0)}\right]$$

$$> \frac{1-1/n}{\Gamma(x_0)} > \frac{1-2/n}{\Gamma(x_0)},$$

we see that in fact $\sigma_n = \gamma_n$, proving (2.1).

Noting further that

$$g_{n,\nu}(x) \leqq \frac{1}{n}\left[\frac{\nu}{\Gamma(x_0)} + \frac{n-\nu}{\Gamma(x_0)}\right] = \frac{1}{\Gamma(x_0)} \quad \text{on } 0 \leqq x \leqq 1,$$

we have $\gamma_n \leqq 1/\Gamma(x_0)$, and thus

$$\frac{1-1/n}{\Gamma(x_0)} < \gamma_n \leqq \frac{1}{\Gamma(x_0)}, \qquad n = 1, 2, 3, \cdots,$$

showing that $\lim_{n\to\infty} \gamma_n = 1/\Gamma(x_0)$, as claimed in (2.3). Theorem 1 is now proved.

5. Numerical results and graphs. In this section we present some information concerning the functions $g_{n,\nu}(x)$ in (2.2) which was obtained by extensive numerical computation, using the CDC 6500 computer.

First of all, we observe that for large n many of the functions $g_{n,\nu}(x)$ do in fact have local maxima in $0 < x < 1$. This can be seen by noting that

$$g_{n,\nu}(x_0^{-(n-\nu)/\nu}) = \frac{1}{n}\left[\frac{\nu}{\Gamma(x_0^{-(n-\nu)/\nu})} + \frac{n-\nu}{\Gamma(x_0)}\right] > \frac{1-\nu/n}{\Gamma(x_0)},$$

so that $g_{n,\nu}(x_0^{-(n-\nu)/\nu}) > 1$ whenever $(1-\nu/n)/\Gamma(x_0) > 1$, i.e., whenever

$$\frac{\nu}{n} < 1 - \Gamma(x_0) = .1143\cdots. \tag{5.1}$$

Since $g_{n,\nu}(0) = 0$, $g_{n,\nu}(1) = 1$, the presence of a local maximum in the case of (5.1) is thus evident.

More detailed computations, covering the range $2 \leqq n \leqq 30$, $1 \leqq \nu \leqq n-1$, revealed that:

(i) $g_{n,\nu}(x)$ is monotonically increasing on $0 \leqq x \leqq 1$ for $n \leqq 6$, $1 \leqq \nu \leqq n-1$.

(ii) $g_{n,1}(x)$ for $n \geqq 7$ has a unique local maximum on $(0, 1)$ which is less than 1 for $n = 7$ and $n = 8$, but larger than 1 for $n \geqq 9$.

(iii) $g_{n,\nu}(x)$ for $\nu = 2, 3, 4$ has a local maximum only for $n \geqq 14, n \geqq 21, n \geqq 28$, respectively, each being smaller than the respective maximum of $g_{n,1}$. As ν increases from 2 to 4, the maxima in question decrease.

(iv) $g_{n,\nu}(x)$ for $7 \leqq n \leqq 30$, $5 \leqq \nu \leqq n-1$ is monotonically increasing on $0 \leqq x \leqq 1$.

The numerical results suggest the conjecture that the relative maxima of $g_{n,\nu}$ decrease as ν increases (with n held fixed), but we do not have a proof for this. Some critical portions of the "dominant" curves $y = g_{n,1}(x)$, $7 \leqq n \leqq 10$, are shown in Fig. 1.

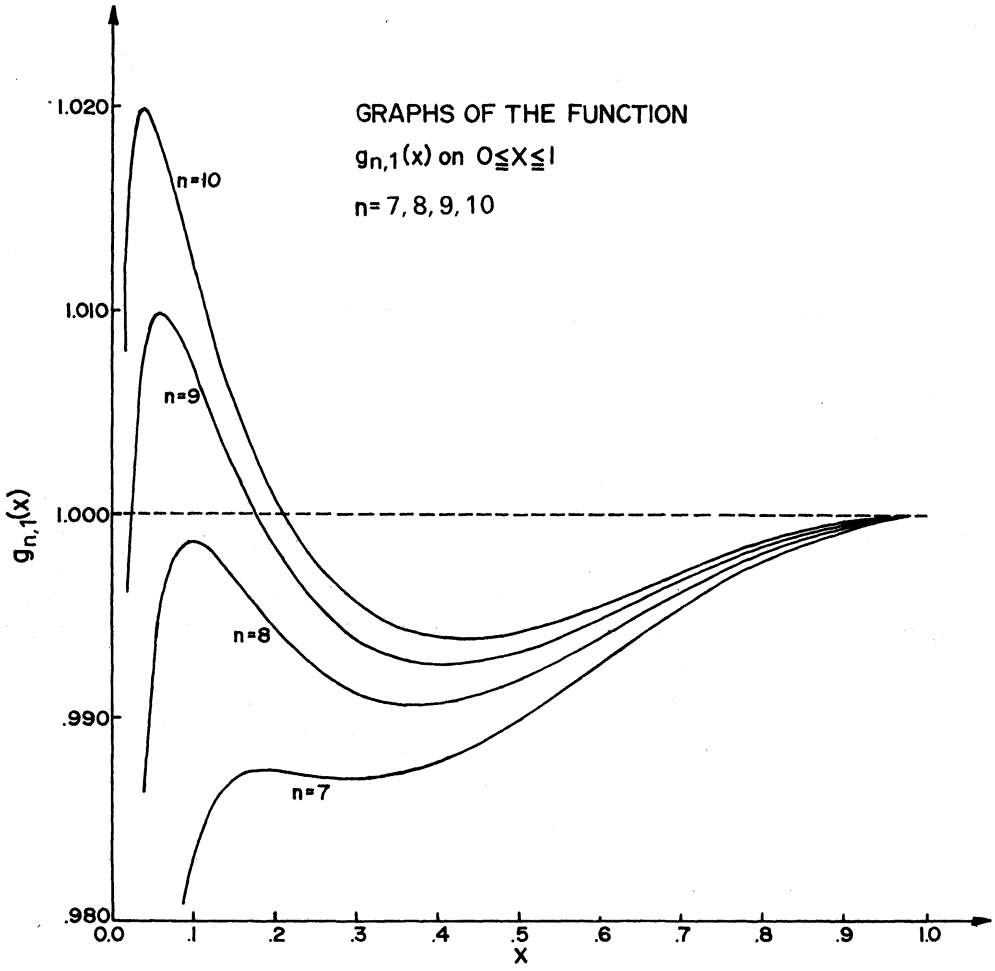


FIG. 1. *Graphs of* $y = g_{n,1}(x)$ *for* $7 \leqq n \leqq 10$

Based on the numerical evidence described above in (i) and (ii), we may infer with confidence that[2]

$$\gamma_n = 1 \quad \text{for } 1 \leqq n \leqq 8. \tag{5.2}$$

From (5.1) with $\nu = 1$, on the other hand, we see that $\gamma_n > 1$ whenever $n > 1/(1 - \Gamma(x_0)) = 8.741 \cdots$, i.e.,

$$\gamma_n > 1 \quad \text{for all } n \geqq 9. \tag{5.3}$$

[2] Equation (5.2) is trivial for $n = 1$, and established in [1] for $n = 2$.

The local maxima γ_n^* of $g_{n,1}$, $7 \leqq n \leqq 30$, were computed more accurately by applying Newton's method to the equation (4.12) with $\nu = 1$. A binary search method was used to obtain fairly accurate initial approximations. The results, believed to be accurate to all digits shown, are displayed in Table 1. (Observe that $\gamma_n^* = \gamma_n$ for $n \geqq 9$.)

TABLE 1[3]

Local maxima $\gamma_n^* = g_{n,1}(\xi_n^*)$ *of* $g_{n,1}(x)$ *for* $7 \leqq n \leqq 30$

n	ξ_n^*	γ_n^*	n	ξ_n^*	γ_n^*
7	1.900855126(−1)	.9874040859	19	1.087835945(−3)	1.069800731
8	9.819583769(−2)	.9986294355	20	7.425128883(−4)	1.072752220
9	6.005471800(−2)	1.009798259	21	5.071386946(−4)	1.075427804
10	3.840859500(−2)	1.019864207	22	3.465411640(−4)	1.077863530
11	2.512555627(−2)	1.028706548	23	2.368819775(−4)	1.080089658
12	1.666215274(−2)	1.036420644	24	1.619635354(−4)	1.082131717
13	1.114939565(−2)	1.043152112	25	1.107593956(−4)	1.084011358
14	7.506998631(−3)	1.049045683	26	7.575306970(−5)	1.085747033
15	5.076813299(−3)	1.054229841	27	5.181558830(−5)	1.087354549
16	3.444215694(−3)	1.058813803	28	3.544457248(−5)	1.088847512
17	2.341999266(−3)	1.062888730	29	2.424710624(−5)	1.090237691
18	1.595180690(−3)	1.066530189	30	1.658765573(−5)	1.091535309

[3] The integers in parentheses denote powers of 10 by which the preceding numbers are to be multiplied.

6. Proof of Theorem 2. The proof follows similar lines of reasoning as the proof of Theorem 1. We can therefore be brief. Letting

$$\gamma(x_1, x_2, \cdots, x_n) = \frac{1}{n} \sum_{k=1}^{n} \ln \Gamma(x_k),$$

we denote

$$\sigma_n = \inf_{\mathbf{x} \in S_n} \gamma(x_1, x_2, \cdots, x_n) > -\infty, \tag{6.1}$$

and propose to prove that

$$\sigma_n = 0.$$

The infimum in (6.1) cannot be assumed at infinity, since otherwise there would be a sequence of vectors $\mathbf{x}^{(r)} \in S_n$ satisfying (4.2), (4.3), hence

$$n\gamma(\mathbf{x}^{(r)}) = \ln \Gamma(x_1^{(r)}) + \ln \Gamma(x_2^{(r)}) + \sum_{k=3}^{n} \ln \Gamma(x_k^{(r)})$$

$$\geqq \ln \Gamma(x_1^{(r)}) + \ln \Gamma(x_2^{(r)}) + (n-2)\Gamma(x_0) \to \infty \quad \text{as } r \to \infty.$$

The function $\gamma(\mathbf{x})$ thus assumes a minimum on S_n at some finite point $\mathbf{x} = \mathbf{s} \in S_n$,

$$\gamma(\mathbf{x}) \geqq \gamma(\mathbf{s}) \quad \text{for all } \mathbf{x} \in S_n.$$

On using Lagrange multipliers, it follows that $\mathbf{s}^T = [s_1, s_2, \cdots, s_n]$ must satisfy

$$\phi(s_1) = \phi(s_2) = \cdots = \phi(s_n), \tag{6.2}$$

$$\prod_{k=1}^{n} s_k = 1, \tag{6.3}$$

where

$$\phi(x) \overset{\text{def}}{=} x\psi(x).$$

Since $\phi(x) < 0$ for $0 < x < x_0$, and $\phi(x) \geqq 0$ for $x \geqq x_0$, we conclude from (6.2), (6.3) that

$$0 < s_k < x_0, \qquad k = 1, 2, \cdots, n.$$

From Proposition 1 we know that $\phi(x)$ is convex for $x > 0$, and from the proof of Proposition 2, that

$$\phi(0) = -1, \quad \phi'(0) < 0, \quad \phi(x_0) = 0.$$

There are thus points ξ^*, ξ_0, with $0 < \xi^* < \xi_0$, such that $\phi(0) = \phi(\xi_0) = -1$ and $\phi(x)$ is monotonically decreasing on $0 \leqq x < \xi^*$ and monotonically increasing on $\xi^* < x \leqq x_0$. Since $\phi(1) = -\gamma > -1$, we have in fact $0 < \xi_0 < 1$.

From (6.2) we now conclude that only one of two situations can hold:

(a) All s_k are the same. Then $s_1 = s_2 = \cdots = s_n = 1$, giving $\sigma_n = 0$.

(b) There are exactly two distinct s_k, say,

$$0 < s_1 = s_2 = \cdots = s_\nu < s_{\nu+1} = s_{\nu+2} = \cdots = s_n, \qquad 1 \leqq \nu < n,$$

such that

$$0 < s_1 < \xi^* < s_n < \xi_0 < 1.$$

Since the last inequalities imply $s_1^\nu s_n^{n-\nu} < 1$, in contradiction to (6.3), case (b) is impossible, leaving us with case (a), i.e., $\sigma_n = 0$. Theorem 2 is proved.

7. Proof of Theorem 3. We have already observed in § 1 that (2.6) is true for all $r \geqq 0$. It suffices therefore to show that (2.6) is false for $r < 0$.

By an obvious adaptation of the proof of Theorem 1, one finds that

$$\inf_{\mathbf{x} \in S_n} M_n^{[r]}(\Gamma(x_i)) = \gamma_n^{1/r}, \qquad r < 0, \tag{7.1}$$

where

$$\gamma_n = \max_{1 \leqq \nu \leqq n-1} \{ \max_{0 \leqq x \leqq 1} g_{n,\nu}(x) \},$$

$$g_{n,\nu}(x) \overset{\text{def}}{=} \frac{1}{n}\{\nu[\Gamma(x)]^r + (n-\nu)[\Gamma(x^{-\nu/(n-\nu)})]^r\}.$$

Now arguing as in (5.3), we have

$$\gamma_n \geqq \max_{0 \leqq x \leqq 1} g_{n,1}(x) \geqq g_{n,1}(x_0^{-(n-1)})$$

$$= \frac{1}{n}\{[\Gamma(x_0^{-(n-1)})]^r + (n-1)[\Gamma(x_0)]^r\} > \left(1 - \frac{1}{n}\right)[\Gamma(x_0)]^r,$$

from which it follows that $\gamma_n > 1$ as soon as

$$n > \frac{1}{1 - [\Gamma(x_0)]^{-r}}.$$

For all these values of n, the infimum in (7.1) is <1, and thus the inequality (2.6) false. This proves Theorem 3.

Acknowledgments. The author gratefully acknowledges helpful discussions with Professors A. Ostrowski, H. Rubin, and Dr. O. Shisha. He is also indebted to James C. Caslin for writing the computer programs which produced the numerical table and the graphs of § 5.

REFERENCE

[1] W. GAUTSCHI, *A harmonic mean inequality for the gamma function*, this journal, 5 (1974), pp. 278–281.

9.7. [49] "Computational Methods in Special Functions — A Survey"

---

[49] "Computational Methods in Special Functions — A Survey," in *Theory and applications of special functions* (R. A. Askey, ed.), 1–98, *Math. Res. Center, Univ. Wisconsin Publ.* **35**, Academic Press, NewYork, 1975.

Reprinted from:
THEORY AND APPLICATION OF SPECIAL FUNCTIONS

ACADEMIC PRESS, INC.
New York San Francisco London

Computational Methods in Special Functions–A Survey

Walter Gautschi

Introduction

Scientific computing often requires special functions. In the past, the need for numerical values was partly satisfied by extensive mathematical tables. Today, with powerful digital computers available, such values are obtained almost invariably by direct computation. We wish to review here the principal methods used in computing special functions.

We may group these methods into two large classes, namely those based on direct approximation, and those based on functional equations. Among the former, we consider only rational approximation methods (§1). We thus leave aside a multitude of possible expansions in terms of other special functions. These expansions, indeed, while often helpful, still leave us with the problem of evaluating the special functions involved. Among the functional equations most useful for computation are linear and nonlinear recurrence relations. These are discussed in §§2 and 3. We omit references to other functional equations, such as differential and integral equations, since we consider them of secondary importance in our context. In §4 we give a brief account of the current state of computer software development for special functions.

Due to limitations in time and space, a number of important topics are omitted in this survey. Nothing is said, e.g., about elementary functions and special computational techniques related to them. Good accounts of this can be found in Lyusternik, Chervonenkis and Yanpol'skii [1965] and Fike [1968]. Other topics not covered include methods based on numerical quadrature and on Euler-Maclaurin and Poisson summation formulas, the computation of zeros of special functions and of inverse functions, and the computation of special constants to very high precision.

Few references are given to computer algorithms for special functions, as they can be retrieved from the indices in the journal "Communications of the ACM" and in "Collected Algorithms from CACM" (a looseleaf collection issued by ACM of all algorithms published in Comm. ACM since 1960). Another topic dealt with only superficially are asymptotic methods, as these are discussed more fully elsewhere in this volume.

There are not many general references on computational methods for special functions. The only book devoted entirely to this subject is Hart et al. [1968]. The two volumes of Luke [1969] also contain much relevant material, and informative survey articles have been written by Bulirsch and Stoer [1968] and Thacher [1969].

As to notations for special functions, we try to be consistent with Abramowitz and Stegun [1964]. With regard to bibliographic references, we give special emphasis to the literature of the past twenty years or so. Little attempt has been made to trace all results back to the original sources.

§1. Methods based on preliminary approximation

Our concern in this paragraph is with the approximation of a given function of a real or complex variable by means of "simpler" functions. Most attractive among these simpler functions are polynomials and rational functions, since they can be evaluated by a finite number of rational operations. Hence we restrict ourselves to polynomial and rational approximation. One should keep in mind, however, that other means of approximation, e.g., expansions in special functions like Bessel functions, can be equally effective if one takes advantage of appropriate recursive schemes of computation. (cf., e.g., 1.5.3, 2.2.2.)

The selection of a particular rational approximation depends on a number of circumstances. If the region of interest is an interval on the real line and our objective is to produce an approximation of high efficiency, and if we are prepared to expend the necessary effort, then we may seek to obtain a best rational approximation, i.e., one whose maximum

error on the interval in question is as small as possible. This is often the preferred choice in computer subroutines. If, on the other hand, we are dealing with functions of a complex variable, or functions of several variables, we are led to use analytic approximation methods, the constructive theory of best approximation in the multivariate case still being in its infancy. (See, e.g., Collatz [1968], Williams [1972], Harris [1973], Fletcher, Grant and Hebden [1974], Watson [1975].) Even if we decide to construct a best approximation, in the process of doing so we still need to be able to calculate the function to high accuracy. Here again, analytic methods can be useful.

With regard to polynomial vs. rational approximation, folklore has it that "in some overall sense, rational approximation is essentially no better than polynomial approximation" (Newman [1964]). Precise theorems to this effect (Walsh [1968b], Feinerman and Newman [1974, p. 71 ff]) add further support to this contention. Experience, nevertheless, seems to show that for the special functions encountered in everyday practice, rational approximations are in fact somewhat superior.

In designing a rational approximation, certain preliminary decisions need to be made regarding the best form in which to approximate the function, the choice of auxiliary variables, and the best type of segmentation of the independent variable. As there is little theory to go by, such decisions are usually made by trial and error. Taylor series, or asymptotic expansions, usually suggest appropriate forms. For the problem of segmentation, see Lawson [1964], Collatz [1965], Meinardus [1966], [1964, §11 of English translation], Hawkins [1972].

1.1. Best rational approximation

Many computer subroutines for special functions employ rational approximations in appropriate segments of the real line. If the subroutine operates in an environment in which every value of the independent variable is equally likely to occur, it is natural to design the approximation in such a way that the error on each segment is "uniformly

distributed", and about the same from segment to segment. In this way, no user is going to be punished if he happens to prefer one particular region over another. The logical conclusion of this philosophy is to employ the principle of best uniform approximation (Chebyshev approximation) on each segment and to arrange the maximum error to be about the same from segment to segment. The "uniform distribution" of the error is then guaranteed by the equi-oscillation property of the best approximation (cf. 1.1.1).

The theory of best uniform approximation is an important chapter of approximation theory, and is dealt with in a number of excellent books. We mention, e.g., Achieser [1956], Davis [1963], Meinardus [1964], Natanson [1964], Rice [1964b], [1969], Cheney [1966], Werner [1966], Rivlin [1969], Walsh [1969], Schönhage [1971], Feinerman and Newman [1974]. A treatise on numerical methods of Chebyshev approximation (not including, however, rational approximation) is Remez [1969]. Practical aspects of generating rational and polynomial approximations are reviewed by Cody [1970].

1.1.1. Best uniform rational approximation. We denote by $\mathbb{P}_n$ the class of polynomials of degree $\leq n$, and by $\mathbb{R}_{m,n}$ the family of rational functions

$$r(x) = \frac{p(x)}{q(x)}, \quad p \in \mathbb{P}_n, \quad q \in \mathbb{P}_m, \quad q \not\equiv 0 \,. \tag{1}$$

Given a real-valued continuous function f on the compact interval $[a,b]$, there exists a unique element $r^*_{m,n} \in \mathbb{R}_{m,n}$ such that

$$\|r^*_{m,n} - f\|_\infty \leq \|r - f\|_\infty \quad \text{for all } r \in \mathbb{R}_{m,n} \,. \tag{2}$$

Here the norm is $\|u\|_\infty = \max_{a\leq x\leq b} |u(x)|$ or, more generally, $\|u\|_\infty = \max_{a\leq x\leq b} w(x)|u(x)|$, where w is a positive weight function. One calls

$r^*_{m,n}$ the rational function of best uniform approximation to f from $\mathbb{R}_{m,n}$ (or briefly the rational Chebyshev approximation of f from $\mathbb{R}_{m,n}$). The associated error is denoted by

$$E_{m,n}(f) = \|r^*_{m,n} - f\|_\infty \quad . \tag{3}$$

In particular, there is a unique polynomial $p^*_n \in \mathbb{P}_n$ of best uniform approximation, with associated error $E_n(f) = E_{0,n}(f)$. The array of rational functions

$$\begin{array}{llll} r^*_{0,0} & r^*_{0,1} & r^*_{0,2} & \cdots \\ r^*_{1,0} & r^*_{1,1} & r^*_{1,2} & \cdots \\ r^*_{2,0} & r^*_{2,1} & r^*_{2,2} & \cdots \\ \cdots & \cdots & \cdots & \end{array}$$

is referred to as the L_∞ Walsh array of f on $[a,b]$.

The best approximation $r^*_{m,n}$ is characterized by the equi-oscillation property, which states (excepting certain degenerate cases) that the error curve $w(r^*_{m,n} - f)$ assumes its extreme value (3) at $m+n+2$ consecutive points of $[a,b]$ with alternating signs (Achieser [1956, p. 55]). Moreover (barring again degeneracies), if $r \in \mathbb{R}_{m,n}$ is any rational function bounded on $[a,b]$ which has the oscillation property, i.e., an error curve $e = w(r-f)$ assuming values of alternating sign on $m+n+2$ consecutive points $x_i \in [a,b]$, say,

$$e(x_i) = (-1)^i \lambda_i, \quad \lambda_i > 0, \qquad i = 1, 2, \ldots, m+n+2 \quad ,$$

then (Achieser [1956, p. 52])

$$\min_i \lambda_i \leq E_{m,n}(f) \leq \|e\|_\infty \quad . \tag{4}$$

Concerning the behavior of $E_{m,n}(f)$ as m and n both tend to infinity, little is known. If m, or n, remains fixed, there are asymptotic results for meromorphic functions, due to Walsh [1964b], [1965], [1968a], while in the polynomial case $m = 0$ one has the classical results of Jackson and Bernstein. The former states that $E_n(f) = o(n^{-p})$ if $f \in C^p[a, b]$, the latter that $\limsup [E_n(f)]^{1/n} < 1$ precisely if f is analytic on $[a, b]$, and $[E_n(f)]^{1/n} = o(1)$ precisely if f is entire (see, e.g., Natanson [1964, pp. 127, 183]).

1.1.2. A list of available Chebyshev approximations. Some entries of the Walsh array, often those along or near the diagonal $m = n$, have proven to yield remarkably efficient approximations for many of the special functions in current use. Table 1 lists those for which (numerically constructed) rational Chebyshev approximations are available. The first column shows the function being approximated, in the notation of Abramowitz and Stegun [1964]. The second column records the segmentation used, where $[a_0, a_1, \ldots, a_s]$ is written to indicate that the interval $[a_0, a_s]$ is broken up into segments $[a_{i-1}, a_i]$, $i = 1, 2, \ldots, s$. The exact form of the function which is being approximated, as well as the type (m, n) of rational function, usually changes from segment to segment in a manner not shown in the table. The third column tells whether the approximant is truly rational or polynomial. The fourth column indicates the approximate range of accuracy, where S is to be read as "significant decimal digits" and D as "decimal digits after the decimal point". The final column gives the source of the approximation. For an extensive bibliography of approximations see also Hart et al. [1968, pp. 161-179].

Table 1. Chebyshev approximations to special functions

f(x)	segmentation	type	accuracy	reference
$E_1(x)$	$[0, 1, 4, \infty]$	rat.	2-20S	Cody & Thacher [1968]
Ei(x)	$[0, 6, 12, 24, \infty]$	rat.	3-20S	Cody & Thacher [1969]

f(x)	segmentation	type	accuracy	reference
$\Gamma(x)$	[2, 3]	pol.	7-18S	Werner & Collinge [1961]
$\ell n\,\Gamma(x)$	[.5, 1.5, 4, 12]	rat.	2-17S	Cody & Hillstrom [1967]
$\Gamma(x)$	$[2, 3]^{(1)}$	pol.&rat.	1-24D	Hart et al. [1968]
$\ell n\,\Gamma(x)-(x-\frac{1}{2})\ell n x + x-\ell n\sqrt{2\pi}$	[8, 1000]	"	8-18D	"
"	[12, 1000]	"	9-23D	"
$\arg\Gamma(1+ix)$	$[0, 2, 4, \infty]$	rat.	4-20S	Cody & Hillstrom [1970]
$\psi(x)$	[1, 2]	pol.	6-8D	Moody [1967]
$\psi(x)$	$[.5, 3, \infty]$	rat.	2-20S	Cody, Strecok&Thacher [1973]
erfc x	[0, 10]	rat.	1-23D	Hart et al. [1968]
"	[0, 20]	"	4-6D	"
"	[0, 4]	"	1-9D	"
"	[0, 8]	"	1-16S	"
"	[8, 100]	"	3-17S	"
erf x	[0, .5]	rat.	5-19S	Cody [1969]
erfc x	$[.46875, 4, \infty]$	rat.	2-18S	"
$e^{-x^2}\int_0^x e^{t^2}dt$	$[0, 2.5, 3.5, 5, \infty]$	rat.	1-21S	Cody, Paciorek&Thacher [1970]
C(x), S(x)	$[0, 1.2, 1.6, 1.9, 2, 4, \infty]$	rat.	2-18S	Cody [1968]
$J_n(x), I_n(x), Y_n(x), K_n(x)$ $n=0, 1$	[0, 8]	pol.	2-7D	Werner [1958/59]
$J_\nu(x)$, $I_\nu(x)$ $\nu = -\frac{2}{3}, -\frac{1}{3}, \frac{1}{3}, \frac{2}{3}$	[0, 4]	pol.	10D	Bhagwandin [1962]
$K_\nu(x)$, $\nu = \frac{1}{3}, \frac{2}{3}$	$[4, \infty]$	pol.	10D	"
$H_\nu^{(1)}(x)$, $\nu = \frac{1}{3}, \frac{2}{3}$	$[4, \infty]$	rat.	10D	"
$I_0(x), I_1(x)$	[0, 8, 70]	rat.	8S	Gargantini [1966]
$K_0(x)$	[0, .1, 8]	rat.	8S	"

f(x)	segmentation	type	accuracy	reference
$K_1(x)$	$[0, 8]$	rat.	7S	Gargantini [1966]
$J_0(x), J_1(x), Y_0(x), Y_1(x)$	$[0, 8, \infty]$	rat.	3-25D	Hart et al. [1968]
$I_0(x), I_1(x)$	$[0, 1]$	rat.	2-23S	Russon & Blair [1969]
$K_0(x), K_1(x)$	$[0, 1, \infty]$	rat.	2-23S	"
$I_0(x), I_1(x)$	$[0, 15, \infty]$	rat.	8-23S	Blair [1974]
$I_0(x), I_1(x)$	$[0, 15, \infty]$	rat.	1-23S	Blair & Edwards [1974]
$Ki_r(x)$, r=1, 2, 3	$[0, \infty]$	rat.	2-7S	Gargantini & Pomentale [1964]
$\int_0^x I_0(t)dt$	$[0, 8, 30]$	rat.	8-9S	Gargantini [1966]
$\int_x^\infty K_0(t)dt$	$[0, .1, 8, 70]$	rat.	7S	"
$G_0(\eta, 2\eta), G_0'(\eta, 2\eta)$	$[1, 2, 3.5, 15]$	rat.	13-14S	Strecok & Gregory [1972]
$G_0(\eta, 1), G_0'(\eta, 1)$	$[0, 1]$	rat.	16S	"
$G_0(\eta, 30), G_0'(\eta, 30)$	$[15, 18.5, 22]$	rat.	13-14S	"
$\ell n(G_0(\eta, 30))$, $\ell n(-G_0'(\eta, 30))$	$[22, 30]$	rat.	13S	"
$\int_0^{\pi/2} (1-x^2 \sin^2 t)^{\pm \frac{1}{2}} dt$	$[0, 1]$	pol.	4-17D	Cody [1965]
$\zeta(x)$	$[.5, 5, 11, 25, 55]$	rat.	8-22S	Cody, Hillstrom & Thacher [1971]
$x^{-1} \int_0^x t^k (e^t-1)^{-1} dt$ $k = 1, 2, 3, 4$	$[0, 10]$	rat.	2-5D	Thacher [1960]
$\int_0^\infty t^{\frac{1}{2}} (e^{t-x}+1)^{-1} dt$	$[-\infty, 1, \infty]$	pol.	3S	Werner & Raymann [1963]
$\int_0^\infty t^\nu (e^{t-x}+1)^{-1} dt$ $\nu = -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}$	$[-\infty, 1, 4, \infty]$	rat.	2-10S	Cody & Thacher [1967]

(1) range incorrectly stated in Hart et al. [1968].

1.1.3. Computation of Chebyshev approximations. Most, if not all, of the approximations in Table 1 were generated by some version of Remez' second algorithm. This is a procedure, originally devised for polynomials (Remes [1934]) and later extended to rational functions, which attempts to achieve the equi-oscillation property in an iterative fashion. The object of the iteration, basically, is to move the two bounds in (4) ever closer together. There are many variants of the procedure, differing somewhat in the technical execution of each iteration step. Detailed descriptions of these can be found in some books on approximation theory, e.g., Meinardus [1964], Rice [1964b], Cheney [1966], Werner [1966], Remez [1969], Rivlin [1969], as well as in survey articles by Cheney and Southard [1963], Stiefel [1959], [1964], Fraser [1965], Ralston [1967], Krabs [1969], Cody [1970]. Computer algorithms are given in Stoer [1964], Werner [1966], Cody and Stoer [1966/67], Werner, Stoer and Bommas [1967], Cody, Fraser and Hart [1968], Huddleston [1972], Johnson and Blair [1973]. The construction of rational Chebyshev approximants, in spite of the many aids available, is still a tricky business due to the possibility of near-degeneracies. For a discussion of this, the reader is referred to Rice [1964a], Cody [1970], Huddleston [1972], Ralston [1973].

There are other methods of obtaining best rational approximations which rely more heavily on mathematical programming. Some of these are referenced in Lee and Roberts [1973] and compared there with Remez' algorithm. Others, more recently, are proposed by Har-El and Kaniel [1973] and Kaufman and Taylor [1974].

1.2. Truncated Chebyshev expansion

There is some effort involved in generating a best rational, or even polynomial, approximation to a given function f. Polynomials which approximate f "nearly best" can be obtained more easily by truncating the Chebyshev expansion of f.

Assuming that the interval of interest has been transformed to $[-1,1]$, we can formally expand f into a series of Chebyshev polynomials,

$$(1) \qquad f(x) = \frac{1}{2}a_0 + \sum_{k=1}^{\infty} a_k T_k(x), \qquad -1 \le x \le 1 ,$$

where

$$(2) \qquad a_k = \frac{2}{\pi} \int_0^{\pi} f(\cos\theta) \cos k\theta \, d\theta, \qquad k = 0, 1, 2, \ldots .$$

In effect, (1) is the Fourier cosine expansion of $f(\cos\theta)$. It converges uniformly and absolutely on $[-1, 1]$ if $f \in C[-1, 1]$ and $f' \in L_p[-1, 1]$, $p > 1$ (Zygmund [1959, p. 242]). The polynomials referred to above are the partial sums of (1),

$$(3) \qquad s_n(f;x) = \frac{1}{2} a_0 + \sum_{k=1}^{n} a_k T_k(x), \qquad n = 0, 1, 2, \ldots .$$

The classical source on Chebyshev polynomials and their applications is Lanczos' introduction in National Bureau of Standards [1952]. More recent accounts can be found in the books of Fox and Parker [1968] and Rivlin [1974].

1.2.1. Convergence. The functions f encountered in practice are usually quite smooth, typically real-valued analytic on $[-1, 1]$ and holomorphic in a domain of the complex plane enclosing the segment $[-1, 1]$. If ϵ is the eccentricity of the largest ellipse, with foci at ± 1, in which f is holomorphic, then (1) converges like a geometric series with ratio $\epsilon/(1+\sqrt{1-\epsilon^2})$ (see, e.g., Werner [1966, §20], Rivlin [1974, p. 143]). For entire functions one has $\epsilon = 0$, and the convergence is supergeometric.

Scraton [1970] observes that convergence can be enhanced if one uses a suitable bilinear, rather than linear, transformation of variables to obtain the canonical interval $[-1, 1]$. Experimental evidence of this has previously been presented by Thacher [1966].

Compared with expansions of f in other orthogonal polynomials, particularly ultraspherical polynomials $P_k^{(\alpha, \alpha)}$, Lanczos early recognized

(National Bureau of Standards [1952]) that convergence is most rapid when $\alpha = -1/2$, i.e., when the expansion is indeed in Chebyshev polynomials. Some firm results in this direction, for restricted classes of functions, are due to Rivlin and Wilson [1969] and Handscomb [1973].

Closely related to convergence is the asymptotic behavior of the expansion coefficients a_k as $k \to \infty$. This is studied in detail by Elliott [1964] for meromorphic functions, and also for functions with a branchpoint at an endpoint of the basic interval, and by Elliott and Szekeres [1965] for entire functions. The case of logarithmic and branch-point singularities on the real line, and combinations of such, is treated by Chawla [1966/67] and Piessens and Criegers [1974]. It is not uncommon to also find essential singularities at an endpoint or midpoint of $[-1,1]$. This occurs, e.g., if the original interval is infinite and f has an essential singularity at infinity. Mapping the interval onto $[-1,1]$ by a reciprocal transformation carries the singularity into a point of $[-1,1]$. The extent to which this slows down the convergence of (1) is studied by Miller [1966]. Asymptotic results for the expansion coefficients in the case of generalized hypergeometric functions are given by Németh [1974].

1.2.2. Relation to best uniform approximation. Letting

$$S_n(f) = \max_{-1 \le x \le 1} |s_n(f;x) - f(x)| , \tag{4}$$

we clearly have $E_n(f) \le S_n(f)$, where $E_n(f)$ is the error of best uniform approximation of f by polynomials of degree n. The difference between $S_n(f)$ and $E_n(f)$ can be remarkably small if f is smooth. This can be seen from de La Vallée Poussin's inequality [1919, p. 107]

$$\left| \sum_{r=0}^{\infty} a_{(2r+1)(n+1)} \right| \le E_n(f) \le S_n(f) \le \sum_{k=n+1}^{\infty} |a_k| , \tag{5}$$

and from other similar results (Hornecker [1958], [1960], Hewers and Zeller [1960/61], Blum and Curtis [1961], Cheney [1966, p. 131], Rivlin

[1974, p. 139ff]). If $a_{k+1} = o(a_k)$, for example, it follows from (5) that $S_n(f) \sim E_n(f)$ as $n \to \infty$. Even for larger classes of functions, e.g., the class $C^{n+1}_{M_n}$ of functions $f \in C^{n+1}[-1,1]$ with $\max_{-1\le x\le 1} |f^{(n+1)}(x)| \le M_n$, the spread is still infinitesimal in the sense (Remez and Gavriljuk [1963])

$$\sup_{f\in C^{n+1}_{M_n}} S_n(f) = [1 + O(\tfrac{1}{n})] \sup_{f\in C^{n+1}_{M_n}} E_n(f), \qquad n \to \infty . \tag{6}$$

Widening the class further to include all continuous functions $f \in C[-1,1]$ we have from the theory of orthogonal series (Alexits [1961, Theorem 4.5.1]) that

$$1 \le \frac{S_n(f)}{E_n(f)} \le 1 + \lambda_n , \tag{7}$$

where λ_n is the Lebesgue constant for Fourier series (Zygmund [1959, p. 67]). Although these constants eventually grow logarithmically with n (Fejér [1910]), they are fairly small in the domain of common interest. It is known that λ_n is monotonically increasing, in fact totally monotone (Szegö [1921]), and $\lambda_1 = 1.436$, $\lambda_{1000} = 4.07$ (Powell [1967]). The error of the truncated Chebyshev expansion, in the range $1 \le n \le 1000$, is therefore never worse than five times the error of the corresponding best uniform approximation.

When f is a polynomial of degree $n+1$, then in fact $S_n(f) = E_n(f)$. For polynomials of degree $> n+1$ the ratios in (7) are investigated by Clenshaw [1964], Lam and Elliott [1972] and Elliott and Lam [1973]. Some of this work, however, is based on conjectures. For related work, see also Riess and Johnson [1972].

It is possible to modify the truncated Chebyshev expansion so as to bring it closer to the best uniform approximation (Hornecker [1958], [1960], Korneĭčuk and Širikova [1968], Širikova [1970]). Other modifications can be made to fit interpolatory conditions at the end points (Cohen

[1971]). This may be useful in segmented approximation when continuity at the joints is desirable.

Using a method reminiscent of Lanczos' τ-method, Stolyarčuk [1974a, b] obtains explicit polynomial approximations to the sine integral, error function, and Bessel functions of integer order, which are valid on an arbitrary interval and are infinitesimally close to the best polynomial approximations on that interval as the degree tends to infinity.

1.2.3. Calculation of expansion coefficients. There are a number of methods available to calculate (or approximate) the expansion coefficients a_k. Some will now be considered.

(i) Fourier analysis. Since we are dealing with Fourier coefficients, we can enlist the techniques of harmonic analysis, and thus, for example, approximate a_k, $k \leq n$, by

$$\alpha_k^{(n)} = \frac{2}{n} \sum_{j=0}^{n}{}'' f(x_j)\, T_k(x_j), \qquad x_j = \cos(j\pi/n) \quad . \tag{8}$$

(The primes on the summation sign indicate that the first and last term is to be halved.) Since $T_k(x_j) = T_j(x_k)$, the sum in (8) can be evaluated effectively by Clenshaw's algorithm (cf. 1.5.1(ii)).

It is a relatively simple matter to increase the accuracy of (8), by doubling n, if one observes that about half of the terms in (8) can be reused, and only half of the $\alpha_k^{(n)}$ need to be computed, by virtue of

$$\alpha_k^{(n)} = \alpha_k^{(2n)} + \alpha_{2n-k}^{(2n)}$$

(Clenshaw [1964], Torii and Makinouchi [1968]).

(ii) Rearrangement of power series. The coefficients a_k of the Chebyshev expansion (1) are related to the coefficients c_k of the Maclaurin series, $f(x) = \sum_{k=0}^{\infty} c_k x^k$, by the linear transformation

$$(9)\qquad \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} u_{00} & u_{01} & u_{02} & \cdots \\ 0 & u_{11} & u_{12} & \cdots \\ 0 & 0 & u_{22} & \cdots \\ & & \cdots\cdots\cdots & \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \end{bmatrix},$$

where

$$u_{ij} = \begin{cases} 2^{1-j}\binom{j}{\frac{j-i}{2}} & \text{if } i+j \text{ is even}, \quad j \geq i \geq 0, \\ 0 & \text{otherwise} \end{cases}$$

(Minnick [1957], De Vogelaere [1959]). As some of the coefficients c_k may be quite large, and of different signs, the application of (9) is likely to require high-precision work. Another complication occurs if the power series converges very slowly (Clenshaw [1962]). The infinite series implied in (9) then also converge very slowly, although, sometimes, they respond well to nonlinear acceleration techniques (Thacher [1964]).

(iii) Recurrence relations. In many cases of practical interest it is possible to derive recurrence relations for the coefficients a_k, either directly from the integral representation (2), or indirectly via differential equations. In using these recursions, a certain amount of skill is required to maintain numerical stability (Clenshaw [1962], Luke and Wimp [1963], Németh [1965], [1974], Clenshaw and Picken [1966], Hangelbroek [1967], Wood [1967], Luke [1969 , Vol. II, §12.5], [1971b, c], [1972a]).

(iv) Numerical quadrature. The integral in (2) can be approximated directly by numerical quadrature. Eq. (8), in fact, is an example. For others, see Rivlin [1974, p. 153ff] and Bjalkova [1963].

(v) Explicit formulas. Explicit formulas for a_k in terms of easily computed functions are known for a number of important special functions, e.g., Bessel functions J_ν, I_ν, Y_ν, K_ν (Wimp [1962], Cylkowski [1966/68]),

Dawson's integral (Hummer [1964]), $\psi(a+x)$, $\ln \Gamma(a+x)$, Ci(x), Si(x) (Wimp [1961]). Luke and Wimp [1963] express the expansion coefficients for confluent hypergeometric functions in terms of Meijer's G-function.

1.2.4. Tables of Chebyshev expansions and computer programs. The most extensive tables are those of Clenshaw [1962], Clenshaw and Picken [1966], and Luke [1969, Vol. II, Ch. XVII]. References to additional tables are given in Luke [1969, Vol. II, pp. 287-291]. Among the more recent specialized tables are those of Németh [1967] for Stirling's series, Strecok [1968] for the inverse error function, Wood [1968] for Clausen's integral, Ng, Devine and Tooper [1969] for Bose-Einstein functions, Wimp and Luke [1969] for modified Bessel functions and their incomplete Laplace transform, Kölbig, Mignaco and Remiddi [1970] for generalized polylogarithms, Németh [1971] for Airy functions, Németh [1972] for zeros of Bessel functions J_ν (considered as functions of ν), Németh [1974] for the integrals $\int_0^\infty t^{-\frac{1}{2}}\exp(-t-t^2/x^2)dt$, $\int_0^\infty (x+t)^{-1}\exp(-t^2)dt$, and Sheorey [1974] for Coulomb wave functions.

An interesting and potentially useful idea, advanced by Clenshaw and Picken [1966] and pursued further by Luke [1971b, c], [1972a], is to provide "miniaturized" tables for functions of several variables. These are tables of coefficients in multiple Chebyshev series. The idea is carried out for Bessel functions of real argument and real order.

A set of ALGOL procedures facilitating the use of Chebyshev expansions is given in Clenshaw, Miller and Woodger [1962/63]. FORTRAN programs for generating Chebyshev expansion coefficients can be found in Håvie [1968] and Amos and Daniel [1972].

1.3. Taylor series and asymptotic expansion

A special function f is often naturally represented in the form

$$(1) \qquad f(z) = \alpha(z)\,g(z), \qquad g(z) \sim \sum_{k=0}^{\infty}{}' c_k (z - z_0)^k, \qquad c_0 = 1 ,$$

where the factor $\alpha(z)$ may vanish at z_0, be singular there, or represent

some other peculiar behavior. The expansion for g is a Taylor series if it converges to $g(z)$ at some $z \neq z_0$, hence in some circle $|z-z_0|<\rho$, $\rho > 0$. It is called an asymptotic expansion if it possibly diverges for every $z \neq z_0$, but for each n $(n = 1, 2, 3, \ldots)$ obeys the law

$$g(z) - \sum_{k=0}^{n-1} c_k (z-z_0)^k = O((z-z_0)^n) \quad \text{as} \quad z \to z_0 \,. \tag{2}$$

It is customary, then, to write (2) in terms of descending powers of ζ, where $\zeta = (z-z_0)^{-1}$.

We will not give here a systematic account of Taylor's series and of asymptotic expansions, but limit ourselves to a few remarks on the computational uses of these expansions, and to an example. We refer to Olver [1974] for a thorough treatment of asymptotic expansions and their application to special functions.

1.3.1. Computational uses. As a computational tool, Taylor series are most useful near the point of expansion, z_0, and then indeed may be quite effective. Further away from z_0 one runs into several problems, notably slow convergence, or absence of it, and severe cancellation of terms, with the attendant loss of significant digits. Asymptotic expansions, likewise, may be quite useful sufficiently close to z_0. The accuracy obtainable from a divergent asymptotic expansion, however, is limited at any fixed $z \neq z_0$, in contrast to convergent expansions. Also, error bounds are not always available, and the evaluation of higher order terms may be laborious.

Both expansions may serve purposes other than direct evaluation of functions. For one, they suggest an appropriate form in which to seek best rational approximations. For another, they may be used as input to some of the methods of 1.2, 1.4 for generating polynomial or rational approximations (cf., in particular, 1.2.3(ii), 1.4.1, 1.4.2, 1.4.5).

Nontrivial problems arise in the expansion of functions of several complex variables. Expanding in one variable leaves the coefficients to

be functions of the remaining variables. This creates challenging problems of effective computation, satisfactory rate of convergence, etc. An example in point is the Taylor expansion of the Bessel function $K_\nu(z)$ of complex order and complex argument, which is treated by Temme [1973]. Another example will be discussed below.

A further important problem is the computation of the Taylor expansion coefficients c_k, when z_0 is an arbitrary point in the complex plane. (In particular, this yields $g(z_0) = c_0$.) There are various approaches one can take: numerical quadrature on Cauchy's integral (Lyness and Sande [1971]), recursive computation of higher derivatives (as, e.g., in Gautschi [1966] and Gautschi and Klein [1970]), or more general backward recurrence techniques in cases where g satisfies a linear differential equation with polynomial coefficients (Thacher [1972], and work of Thacher in progress). The more obvious process of analytic continuation (Henrici [1966]), unfortunately, is inherently unstable.

1.3.2. An example (Van de Vel [1969]). Consider the incomplete elliptic integral of the first kind (cf. 3.1.1),

$$F(\varphi, k) = \int_0^{\varphi} (1 - k^2 \sin^2\theta)^{-\frac{1}{2}}\, d\theta, \quad 0 \le k \le 1, \quad 0 \le \varphi < \pi/2 \ , \tag{3}$$

where φ is the amplitude and k the modulus of F. The developments to be made for (3) apply similarly to the integral of the second kind. The complementary modulus k' is defined by

$$k' = \sqrt{1-k^2} \ , \tag{4}$$

and the complete integral by

$$\mathbb{K}(k) = F(\tfrac{\pi}{2}, k) = \int_0^{\pi/2} (1 - k^2 \sin^2\theta)^{-\frac{1}{2}}\, d\theta, \quad 0 \le k < 1 \ . \tag{5}$$

We are interested in Taylor's expansion of F with respect to the modulus k.

The most obvious attack is to expand the integrand in a binomial series and to integrate term by term. The result is

$$(6) \qquad F(\varphi, k) = \sum_{r=0}^{\infty} (-1)^r \binom{-\frac{1}{2}}{r} \sigma_r(\varphi) k^{2r}, \qquad \sigma_r(\varphi) = \int_0^{\varphi} \sin^{2r}\theta \, d\theta .$$

For the σ_r one can find a simple recurrence formula. The series (6) converges geometrically, with an asymptotic quotient $k^2 \sin^2\varphi$. We have rapid convergence, therefore, if k is small, but slow convergence, if k is near 1 and φ near $\pi/2$.

When k is near 1, then (4) suggests finding an expansion in k'. This can be achieved by writing (3) as

$$F(\varphi, k) = \int_0^{\varphi} \frac{d\theta}{\cos\theta[1 + k'^2 \tan^2\theta]^{\frac{1}{2}}} ,$$

and again using the binomial expansion,

$$(7) \qquad F(\varphi, k) = \sum_{r=0}^{\infty} \binom{-\frac{1}{2}}{r} \tau_r(\varphi) k'^{2r}, \qquad \tau_r(\varphi) = \int_0^{\varphi} \frac{\sin^{2r}\theta}{\cos^{2r+1}\theta} d\theta .$$

As before, the τ_r can be generated by a simple recursion. The asymptotic convergence quotient of the series (7) is now $k'^2 \tan^2\varphi$, and thus satisfactory if k' is small and φ not too close to $\pi/2$.

It remains to deal with the last contingency, viz., φ near $\pi/2$. Here we write

$$\mathbb{K}(k) - F(\varphi, k) = \int_0^{\frac{\pi}{2} - \varphi} \frac{d\theta}{\cos\theta[k'^2 + \tan^2\theta]^{\frac{1}{2}}} ,$$

and make the change of variables $\tan\theta = k' \tan\psi$. The result is

$$\mathbb{K}(k) - F(\varphi, k) = \int_0^u \frac{\cos\theta}{\cos\psi}\, d\psi = \int_0^u \frac{d\psi}{\cos\psi[1 + k'^2 \tan^2\psi]^{\frac{1}{2}}} ,$$

where

$$u = \cot^{-1}(k' \tan\varphi) .$$

Therefore, if $k'^2 \tan^2 u < 1$, i.e., $\varphi > \pi/4$, we can expand in a binomial series and find

$$\mathbb{K}(k) - F(\varphi, k) = \sum_{r=0}^{\infty} \binom{-\frac{1}{2}}{r} \tau_r(u) k'^{2r} . \tag{8}$$

We now have a series whose convergence quotient is $k'^2 \tan^2 u = \cot^2\varphi$, thus independent of k, and which converges more rapidly, the closer φ is to $\pi/2$. Note, however, that (8) requires the computation of the complete elliptic integral. (For this, see 3.4.)

It is easily verified that for any k and φ in the region $0 \leq k \leq 1$, $0 \leq \varphi < \pi/2$, at least one of the series (6), (7), (8) converges geometrically with an asymptotic quotient $\leq 1/2$.

Other methods of computation, based on Gauss and Landen transformations, will be considered in 3.3. These are sometimes (but not always) more efficient than the expansions considered here.

1.4. Padé and continued fraction approximations

Given a formal power series about some point z_0 in the complex plane, one can associate with it certain rational functions having highest order contact with the power series at z_0. The rational functions in turn can be interpreted as convergents of continued fractions. These often converge faster, or in larger domains, than the original series, and may even converge when the series diverges. It is this property which makes them useful as a tool of approximation. Without loss of generality we shall assume the point of contact at the origin, $z_0 = 0$.

The basic references are Wall [1948], Perron [1957] and Khovanskii [1963]. On Padé approximation there are survey articles by Gragg [1972] and Chisholm [1973b], as well as a forthcoming book by Baker [1975]. Informative surveys on the use and application of Padé approximants and continued fractions can be found in the collection of articles edited by Baker and Gammel [1970], and in recent conference proceedings, e.g., Graves-Morris [1973a, b] and Jones and Thron [1974b]. We single out the extensive survey of Wynn [1974], containing many references, both to original sources and to newer developments. A good introduction into the numerical evaluation of continued fractions is Blanch [1964]. For a collection of computer algorithms see Wynn [1966b].

1.4.1. Padé table. Let

$$f(z) \sim c_0 + c_1 z + c_2 z^2 + \ldots, \quad c_0 \neq 0 \ , \tag{1}$$

be a formal power series, and ν, μ two nonnegative integers. It is possible to determine polynomials $\hat{p}_{\nu,\mu} \in \mathbb{P}_\mu$, $\hat{q}_{\nu,\mu} \in \mathbb{P}_\nu$, with $\hat{q}_{\nu,\mu} \not\equiv 0$, such that

$$\hat{q}_{\nu,\mu}(z)\, f(z) - \hat{p}_{\nu,\mu}(z) = (z^{\nu+\mu+1}) \ , \tag{2}$$

where the symbol on the right stands for a formal power series beginning with a power z^k, $k \geq \nu + \mu + 1$. Although the polynomials $\hat{p}_{\nu,\mu}$ and $\hat{q}_{\nu,\mu}$ are not unique, they determine a unique rational function $\hat{p}_{\nu,\mu}(z)/\hat{q}_{\nu,\mu}(z)$, which may be expressed, in irreducible form, as

$$[\nu,\mu]_f(z) = \frac{p_{\nu,\mu}(z)}{q_{\nu,\mu}(z)}, \quad p_{\nu,\mu} \in \mathbb{P}_\mu, \quad q_{\nu,\mu} \in \mathbb{P}_\nu, \quad q_{\nu,\mu}(0) = 1 \ . \tag{3}$$

One calls $[\nu,\mu]_f$ the Padé approximant of order ν, μ generated by $f(z)$ (Wall [1948, p. 377ff], Perron [1957, p. 235ff]). We note from (2) and (3) that

$$[\nu, \mu]_f = \frac{1}{[\mu, \nu]_{\frac{1}{f}}}, \quad \nu \geq 0, \quad \mu \geq 0 \quad . \tag{4}$$

The array of rational functions

$$\begin{matrix} [0,0]_f & [0,1]_f & [0,2]_f & \dots \\ [1,0]_f & [1,1]_f & [1,2]_f & \dots \\ [2,0]_f & [2,1]_f & [2,2]_f & \dots \\ \dots & \dots & \dots & \dots \end{matrix} \tag{5}$$

is called the Padé table of f.

If $f(z) - [\nu, \mu]_f(z) = (z^{r+1})$, and (z^{r+1}) cannot be replaced by (z^{s+1}) with $s > r$, we say that $[\nu, \mu]_f$ has contact of order r with f. A Padé table in which each approximant $[\nu, \mu]_f$ has contact of order $\nu + \mu$,

$$f(z) - [\nu, \mu]_f(z) = (z^{\nu+\mu+1}) \quad , \tag{6}$$

is called normal. A necessary and sufficient condition for this is (Wall [1948, p. 398])

$$\Delta_{m,n} = \det \begin{bmatrix} c_{n-m} & c_{n-m+1} & \dots & c_n \\ c_{n-m+1} & c_{n-m+2} & \dots & c_{n+1} \\ \dots & \dots & \dots & \dots \\ c_n & c_{n+1} & \dots & c_{n+m} \end{bmatrix} \neq 0, \quad n, m = 0, 1, 2, \dots \; . \tag{7}$$

(The convention $c_k = 0$ for $k < 0$ is used here.) If this condition holds, $p_{\nu,\mu}$ and $q_{\nu,\mu}$ in (3) are of exact degrees μ and ν, respectively. In the abnormal case, identical approximants lie in square blocks of the Padé table of the form $[i+r, j+s]_f$ $(r, s = 0, 1, \dots, k)$, each approximant of this block having contact of order $i + j + k$.

The question of convergence, $[\nu,\mu]_f \to F$ as ν,μ, or both, tend to infinity, where F is a function associated in some way with f, is a difficult one, depending, as it does, on the behavior of the poles of $[\nu,\mu]_f$. We refer to Baker [1965], [1970], and Chisholm [1973c], for summaries of results and conjectures, and to Wynn [1972], Jones and Thron [1975], for more recent results.

1.4.2 Corresponding continued fractions. If the series in (1) is such that

$$\Delta_{m,m} \neq 0 \quad \text{for} \quad m = 0, 1, 2, \ldots , \tag{8}$$

we can associate with it an infinite J-fraction,

$$\sum_{k=0}^{\infty} c_k z^k \sim \frac{b_0}{1-a_0 z-} \frac{b_1 z^2}{1-a_1 z-} \frac{b_2 z^2}{1-a_2 z-} \cdots , \quad b_k \neq 0, \quad b_0 = c_0 . \tag{9}$$

If the series is such that

$$\Delta_{m,m} \neq 0, \quad \Delta_{m,m+1} \neq 0 \quad \text{for} \quad m = 0, 1, 2, \ldots , \tag{10}$$

we can also associate an infinite S-fraction,

$$\sum_{k=0}^{\infty} c_k z^k \sim \frac{s_0}{1-} \frac{s_1 z}{1-} \frac{s_2 z}{1-} \frac{s_3 z}{1-} \cdots , \quad s_k \neq 0, \quad s_0 = c_0 . \tag{11}$$

Both continued fractions are completely characterized by their contact properties: the p-th convergent of the J-fraction ($p = 1, 2, 3, \ldots$) has contact of order $2p$, that of the S-fraction contact of order p, with the series (1). The J-fraction, in fact, is a contraction of the S-fraction.

The correspondences (9) and (11) are often written for series in descending powers of z (usually asymptotic series), in which case they assume the form (Wall [1948, pp. 197, 202])

$$\sum_{k=0}^{\infty} \frac{c_k}{z^{k+1}} \sim \frac{b_0}{z-a_0-} \frac{b_1}{z-a_1-} \frac{b_2}{z-a_2-} \cdots , \tag{9'}$$

$$(11') \qquad \sum_{k=0}^{\infty} \frac{c_k}{z^{k+1}} \sim \frac{s_0}{z-} \frac{s_1}{1-} \frac{s_2}{z-} \frac{s_3}{1-} \cdots .$$

An important special case of (8), namely $\Delta_{m,m} > 0$, occurs precisely when $\{c_k\}$ is a moment sequence (Wall [1948, p. 325]),

$$(12) \qquad c_k = \int_{-\infty}^{\infty} t^k \, d\phi(t), \qquad k = 0, 1, 2, \ldots ,$$

with ϕ a bounded nondecreasing function having infinitely many points of increase. The series (1), called Stieltjes series, is then the formal expansion of a Stieltjes transform,

$$(13) \qquad \int_{-\infty}^{\infty} \frac{d\phi(t)}{1-tz} \sim c_0 + c_1 z + c_2 z^2 + \cdots .$$

The continued fraction (9) associated with (13) has all a_k real, and all $b_k > 0$ (Perron [1957, p. 193]). Its convergents, as well as the convergents of (9'), are expressible in terms of the orthogonal polynomials $\{\pi_k(t)\}$ belonging to $d\phi(t)$, or in terms of Gaussian quadrature. For example, in the case of (9'),

$$(14) \qquad \int_{-\infty}^{\infty} \frac{d\phi(t)}{z-t} \sim \frac{c_0}{z} + \frac{c_1}{z^2} + \cdots \sim \frac{b_0}{z-a_0-} \frac{b_1}{z-a_1-} \cdots ,$$

we have

$$(15) \qquad \frac{b_0}{z-a_0-} \frac{b_1}{z-a_1-} \cdots \frac{b_{p-1}}{z-a_{p-1}} = \frac{1}{\pi_p(z)} \int_{-\infty}^{\infty} \frac{\pi_p(z) - \pi_p(t)}{z-t} \, d\phi(t)$$

$$(16) \qquad = \sum_{k=1}^{p} \frac{\omega_k^{(p)}}{z-\tau_k^{(p)}} ,$$

where $\tau_k^{(p)}$ are the zeros of $\pi_p(t)$ and $\omega_k^{(p)}$ the associated Christoffel

numbers. The polynomials $\pi_k(z)$ are thus the denominators of the continued fraction in (14), the associated orthogonal polynomials

$$\sigma_k(z) = \int_{-\infty}^{\infty} \frac{\pi_k(z)-\pi_k(t)}{z-t}\, d\phi(t)$$

the numerators. Both satisfy the same recurrence formula,

$$y_{r+1} = (z-a_r)y_r - b_r y_{r-1}, \qquad r = 0, 1, 2, \ldots , \tag{17}$$

where $y_0 = 1$, $y_{-1} = 0$ for $\{\pi_k\}$, and $y_0 = 0$, $y_{-1} = -1$ for $\{\sigma_k\}$. This is meaningful not only for Stieltjes series, but for any series which has an associated J-fraction, provided orthogonality is defined algebraically (Wall [1948, p. 192]). We also note that in terms of the continued fraction (11), we have

$$\begin{aligned} a_0 &= s_1, \quad b_0 = s_0 , \\ a_r &= s_{2r} + s_{2r+1} \\ b_r &= s_{2r-1}\, s_{2r} \end{aligned} \Bigg\} \quad r = 1, 2, 3, \ldots . \tag{18}$$

A special case of (10), similarly, is $\Delta_{m,m} > 0$, $\Delta_{m,m+1} > 0$, and obtains precisely when (12) holds for some measure $d\phi(t)$ vanishing for $t < 0$ (Wall [1948, p. 327]). In this case, $s_k > 0$ for all $k \geq 0$ in (11), a source of useful inequalities when z is real and negative.

With regard to convergence of the continued fractions in (9) and (11), and their limits, we refer to Perron [1957, p. 145ff].

1.4.3. Relation between Padé table and continued fractions. Assume that the series (1) is normal. The conditions (8) and (10) are then valid not only for the given series, but also for all delayed series

$$f_m(z) \sim c_m + c_{m+1}z + c_{m+2}z^2 + \ldots, \qquad m = 0, 1, 2, \ldots . \tag{1_m}$$

Each of these, therefore, has an associated J-fraction

$$(9_m) \qquad f_m(z) \sim \frac{b_0^{(m)}}{1-a_0^{(m)}z-} \frac{b_1^{(m)}z^2}{1-a_1^{(m)}z-} \frac{b_2^{(m)}z^2}{1-a_2^{(m)}z-} \cdots, \quad b_k^{(m)} \neq 0, \; b_0^{(m)} = c_m \; ,$$

and an associated S-fraction,

$$(11_m) \qquad f_m(z) \sim \frac{s_0^{(m)}}{1-} \frac{s_1^{(m)}z}{1-} \frac{s_2^{(m)}z}{1-} \frac{s_3^{(m)}z}{1-} \cdots, \quad s_k^{(m)} \neq 0, \quad s_0^{(m)} = c_m \; .$$

It turns out (Wall [1948, p. 380]) that the entries of the Padé table for $f = f_0$ in the stairlike sequence

$$\begin{array}{llll} [0, m-1] & [0, m] & & \\ & [1, m] & [1, m+1] & \\ & & [2, m+1] & [2, m+2] \\ & & & \ddots \end{array}$$

are identical with the successive convergents of the continued fraction

$$(19) \qquad c_0 + c_1 z + \cdots + c_{m-1} z^{m-1} + \frac{s_0^{(m)} z^m}{1-} \frac{s_1^{(m)}z}{1-} \frac{s_2^{(m)}z}{1-} \cdots \; ,$$

while those along the para-diagonal

$$\begin{array}{lll} [0, m-1] & & \\ & [1, m] & \\ & & [2, m+1] \\ & & \ddots \end{array}$$

are the successive convergents of

$$(20) \qquad c_0 + c_1 z + \cdots + c_{m-1} z^{m-1} + \frac{b_0^{(m)} z^m}{1-a_0^{(m)}z-} \frac{b_1^{(m)}z^2}{1-a_1^{(m)}z-} \frac{b_2^{(m)}z^2}{1-a_2^{(m)}z-} \cdots \; .$$

As in (15), the latter are expressible in terms of the orthogonal polynomials $\{\pi_k^{(m)}\}$ belonging to the measure $t^m\, d\phi(t)$. (See, in this connection, Allen, Chui, Madych, Narcowich and Smith [1974]). Similar statements can be obtained for the entries in the lower half of the Padé table by using (4).

We remark that in the case of convergence, the continued fraction

$$\frac{1}{1-}\ \frac{s_1^{(m)}z}{1-}\ \frac{s_2^{(m)}z}{1-}\cdots$$

in (19), and the analogous continued fraction in (20), serve as "converging factor", being the factor by which the last term $c_m z^m$ is to be multiplied in order to obtain the correct limit of the series (1).

1.4.4. Algorithms. The entries of the Padé table may be generated either in explicit form, as ratios of polynomials, or in their continued fraction form (19). For the former, there are a number of recursive schemes for generating the polynomials in question (Wynn [1960], Baker [1970], [1973], Longman [1971], Watson [1973]). For the latter, one has the quotient-difference (qd-) algorithm (Rutishauser [1954a, b], [1957], Henrici [1958], [1963], [1967]), which consists in generating the qd-array

$$\begin{array}{ccccc}
e_0^{(0)} & & & & \\
 & q_1^{(0)} & & & \\
e_0^{(1)} & & e_1^{(0)} & & \\
 & q_1^{(1)} & & q_2^{(0)} & \\
e_0^{(2)} & & e_1^{(1)} & & e_2^{(0)} \\
 & q_1^{(2)} & & q_2^{(1)} & \\
e_0^{(3)} & & e_1^{(2)} & & e_2^{(1)} \\
 & q_1^{(3)} & & q_2^{(2)} & \\
\vdots & & e_1^{(3)} & & e_2^{(2)} \\
 & \vdots & & q_2^{(3)} & \\
 & & \vdots & & e_2^{(3)} \\
 & & & \vdots & \vdots
\end{array}$$

from left to right by means of

$$e_0^{(n)} = 0, \quad q_1^{(n)} = \frac{c_{n+1}}{c_n}, \qquad n = 0, 1, 2, \ldots ,$$

$$\left.\begin{aligned} e_k^{(n)} &= q_k^{(n+1)} - q_k^{(n)} + e_{k-1}^{(n+1)} \\ q_{k+1}^{(n)} &= \frac{e_k^{(n+1)}}{e_k^{(n)}}\, q_k^{(n+1)} \end{aligned}\right\} \quad k = 1, 2, 3, \ldots, \quad n = 0, 1, 2, \ldots .$$

The coefficients in the continued fraction (19) are then given by

$$s_0^{(m)} = c_m, \quad s_{2k-1}^{(m)} = q_k^{(m)}, \quad s_{2k}^{(m)} = e_k^{(m)}, \quad k = 1, 2, 3, \ldots, \quad m = 0, 1, 2, \ldots .$$

Unfortunately, the generation of the qd-array, as described, is unstable, and should be carried out in high precision, or with some other precautions (Gargantini and Henrici [1967]). Thacher [1971] notes, however, that inaccuracies in the higher order coefficients $s_k^{(m)}$ need not necessarily imply an inaccurate value of the continued fraction (19).

In some instances one has explicit expressions for the $e_k^{(n)}$, $q_k^{(n)}$, for example, in the case of the complex error function (Thacher [1967]), or for certain special hypergeometric and confluent hypergeometric functions (Wynn [1960], Henrici [1963]). For series (1), with $c_m = \prod_{\mu=0}^{m-1} \{(a - q^{\alpha+\mu})(b - q^{\beta+\mu})^{-1}\}$, Wynn [1967] gives closed expressions for the numbers $e_k^{(n)}$, $q_k^{(n)}$, and also for the numerator and denominator polynomials of the approximants in the upper half of the Padé table. Limiting forms of these results (obtained, e.g., when $a = b = 1$, $q \to 1$) yield all cases in which these numbers and polynomials are known in closed form.

There are other algorithms, notably the ϵ-algorithm and related methods due to Wynn [1956], [1961], [1966a], which operate directly on the entries of the Padé table. Their most important use, probably, is in the calculation of numerical values for a sequence of Padé approximants,

e.g., the values at $z = 1$ in an attempt to speed the convergence of $\sum_{k=0}^{\infty} c_k$.

1.4.5. Applications to special functions. The qd-algorithm, either applied to a Taylor series or to an asymptotic expansion, has been used by many authors to obtain the corresponding S-fraction explicitly or numerically. We mention the work of Gargantini and Henrici [1967] on the Bessel function $K_0(z)$ and more general confluent hypergeometric functions, the work of Thacher [1967] on the complex error function, of Cody and Thacher [1968] and Chipman [1972] on the exponential integral $E_1(z)$ and related integrals, of Strecok and Gregory [1972] on the irregular Coulomb wave function along the transition line, and the study of Shenton and Bowman [1971] on the polygamma functions $\psi^{(n)}(z)$. Jacobs and Lambert [1972] apply S-fractions to polylogarithms of a complex argument, while Barlow [1974] does the same to generalized polylogarithms.

Earlier, Fair [1964] uses Lanczos' τ-method for obtaining the J-fraction for functions defined by Riccati differential equations, and applies the technique to confluent hypergeometric functions and Bessel functions of the first and second kind. Fair and Luke [1967] further apply it to incomplete elliptic integrals (cf. also Luke [1969, Vol. II, p. 77ff]).

For large classes of functions, including Gauss hypergeometric functions and the incomplete gamma function, Luke [1969, Vol. II, Chs. XIII and XIV], [1970b], [1971a], [1975] gives explicit expressions for the Padé entries on the diagonal, and immediately above, as well as appraisals of the errors. Those for the incomplete gamma function also serve to approximate the gamma function in the complex plane. See, however, Ng [1975] for a comparison with other methods. Tables of Padé coefficients are given in Luke [1969, Vol. II, p. 402ff] for the exponential, sine, and cosine integrals and for the error function. Golden, McGuire and Nuttall [1973] give an experimental study of the diagonal Padé approximants in the case of Hankel functions of the first and second kind.

Gaussian quadrature, or the equivalent J-fraction in (15), have been used by Todd [1954] for evaluating the complex exponential integral, and by Gautschi [1970] for evaluating the complex error function. In the latter work, the continued fraction approach is combined with a Taylor series approach, there being a gradual transition from one to the other as the complex argument decreases in magnitude.

1.4.6. Error estimates. It is important to have reasonably good estimates of the error due to premature truncation of a continued fraction. One distinguishes between a priori estimates, which are expressed directly in terms of the elements of the continued fraction, and a posteriori estimates, which depend on the knowledge of a finite number (usually two or three) of convergents. Concerning the latter, we mention the elegant work of Henrici [1965] and Henrici and Pfluger [1966] on Stieltjes fractions, in which a sequence of nested lens-shaped regions is constructed the intersection of which contains the value of the continued fraction. For more recent extensions of this work, as well as for other types of estimates, we refer to the survey of Jones [1974].

For a large number of continued fraction expansions of special functions, Wynn [1962a, b] gives "efficiency profiles", i.e., tables from which the order of convergents can be determined as a function of the (real) argument and the accuracy desired.

1.4.7. Generalizations. In view of the contact properties of Padé and continued fraction approximations, one expects these approximations to be best near the point of contact, and to gradually worsen away from it. There is, in fact, a close relationship between the best uniform rational approximants on small discs $|z| \leq \epsilon$, or small intervals $0 \leq z \leq \epsilon$, and the Padé approximant, the former tending to the latter as $\epsilon \to 0$ (Walsh [1964a], [1974], Chui, Shisha and Smith [1974]). The reason for this behavior is largely due to the employment of powers in setting up the Padé table. To obtain a more balanced rational approximation on a given interval, it has been suggested to use systems of orthogonal

polynomials instead, and to proceed similarly as in 1.4.1, starting with the appropriate orthogonal expansion of f. It will be noted that the analogue of (2) is still a linear problem, but the analogue of (6) is not. The original work along this line is due to Maehly [1956], [1958] (see also Kogbetliantz [1960], Spielberg [1961b]), who uses Chebyshev polynomials, and is continued by Cheney [1966, p. 177ff], Holdeman [1969] and Fleischer [1972]. These authors use the linear approach. The nonlinear problem, which is closer in spirit to Padé approximation, has only recently been considered (Common [1969], Fleischer [1973a, b], Frankel and Gragg [1973], Clenshaw and Lord [1974], Gragg and Johnson [1974]). The use of Chebyshev polynomials often leads to nearly best rational approximations (Clenshaw [1974]).

In another direction, one might generalize Padé and continued fraction approximation by imposing contact conditions not only at one, but at several points (typically, at the origin and at infinity). See Baker, Rushbrooke and Gilbert [1964] and Baker [1970] for recent attempts in this direction, and McCabe [1974] for an interesting continued fraction approach. The potential of this approach remains largely to be explored.

Finally, we mention generalizations of Padé approximation to functions of two variables by Chisholm [1973a], Hughes Jones and Makinson [1974], Graves-Morris, Hughes Jones and Makinson [1974], Common and Graves-Morris [1974].

1.4.8. Other rational approximations. We already mentioned the τ-method (Lanczos [1956, pp. 464-507]) applied to linear and nonlinear differential equations as a source of rational approximations (Luke [1955], [1958], [1959/60], Guerra [1969], Verbeeck [1970]). Other sources are Maehly's economization of continued fractions and related techniques (Maehly [1960], Spielberg [1961a], Ralston [1963]), Hornecker's method of modifying the Chebyshev expansion (Hornecker [1959a, b], [1960]), the method of Luke and co-workers (Luke [1969, Vol. II, Ch. XI]) on generalized hypergeometric functions and functions representable as Laplace transforms, and the nonlinear sequence-to-sequence transformation of Levin applied to the partial sums of power series (Levin [1973], Longman

[1973]). Integrating Padé approximants for the square root, Luke [1968], [1970a] obtains rational approximations to the three normal forms of incomplete elliptic integrals, including asymptotic estimates of the error. We also mention the curious ad-hoc approximation to the gamma function $\Gamma(z)$ on $\operatorname{Re} z \geq 1$ due to Lanczos [1964].

1.5. Representation and evaluation of approximations

Once an approximation to a special function has been constructed, it is often possible to represent this approximation in different mathematically equivalent forms. Each form in turn suggests one or several algorithms of evaluation. Although mathematically equivalent, these forms may behave quite differently under evaluation in finite precision. It is important to select a representation, and a corresponding evaluation algorithm, which to the maximum extent possible is invulnerable to the vagaries of finite precision arithmetic.

With regard to representation, what one aims for is well-conditioning. This means that the value of the particular functional form be insensitive to small perturbations in the parameters (coefficients) involved. With regard to algorithms, one strives for economy and stability, i.e., few arithmetic operations and maximum resistance to rounding errors. It is a rare instance where all three of these requirements are in complete harmony with each other.

We discuss some possible representations and algorithms for polynomial and rational approximations, and then consider an algorithm for evaluating approximations in the form of orthogonal sums.

1.5.1. Polynomials

(i) Power form. A polynomial of degree n is most frequently represented in the form

$$p(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n \,, \tag{1}$$

which can be evaluated rather economically by a scheme ascribed to Horner [1819] (but already known to Newton (Ostrowski [1954])),

$$(2)\qquad \begin{cases} u_n = a_n \\ u_k = xu_{k+1} + a_k, \qquad k = n-1,\ n-2,\ \ldots,\ 0\ , \\ p(x) = u_0\ . \end{cases}$$

The scheme requires n multiplications and n additions. With regard to addition, this is optimal (Ostrowski [1954]). The conditioning of the form (1) (at the point x) depends on the relative magnitudes of the quantities $\max_k |a_k x^k|$ and $|p(x)|$. If the former is much larger than the latter, then (1) is ill-conditioned at x. Horner's scheme is generally stable, but can be moderately, and in some cases severely, unstable (Wilkinson [1963, p. 36], Reimer and Zeller [1967], Reimer [1968]). The Chebyshev polynomials, of all, are particularly vulnerable (Reimer [1971]).

(ii) Chebyshev polynomial form. Every polynomial of degree n can be represented in terms of Chebyshev polynomials as (cf. 1.2.3(ii))

$$(3)\qquad p(x) = \frac{1}{2}\, a_0 + \sum_{k=1}^{n} a_k\, T_k(x)\ .$$

One of the attractive features of this form is the possibility of obtaining a sequence of approximations of varying accuracy by merely truncating (3) at consecutive terms. For the evaluation of $p(x)$ one has an algorithm due to Clenshaw [1955],

$$(4)\qquad \begin{cases} u_n = a_n,\quad u_{n+1} = 0\ , \\ u_k = 2xu_{k+1} - u_{k+2} + a_k, \qquad k = n-1,\ n-2,\ \ldots,\ 0\ , \\ p(x) = \frac{1}{2}(u_0 - u_2)\ , \end{cases}$$

requiring $2n$ additions and n multiplications (cf. 1.5.3). Although more time-consuming than Horner's scheme, Clenshaw's algorithm is often preferred on account of its more favorable stability properties. See Newbery [1974] for a comparative study.

(iii) Root product form. This is the form obtained by factoring the polynomial into its linear and quadratic factors,

$$(5) \qquad p(x) = a_n \prod_{k=1}^{r} (x-x_k) \prod_{k=r+1}^{r+s} [(x-x_k)^2 + y_k], \quad y_k > 0, \quad r + 2s = n \ .$$

Like Horner's scheme, this form requires n additions and n multiplications. For maximum stability, however, the differences $x - x_k$ must be evaluated with care: Assuming x machine representable (in floating-point arithmetic), and denoting by x_k^* the machine representable part of x_k, and by r_k the remainder,

$$x_k = x_k^* + r_k \ ,$$

one should evaluate $x - x_k$ in two steps as $(x - x_k^*) - r_k$, thereby preserving as much significance as possible when x is close to x_k. Note that this doubles the number of additions. The construction of the form (5) requires some effort, namely the calculation of all zeros of p, but this effort may be rewarded by a well-conditioned representation.

(iv) Newton form. In a sense intermediate between (1) and (5) is Newton's form

$$(6) \qquad \begin{aligned} p(x) = a_0 &+ a_1(x-x_0) + a_2(x-x_0)(x-x_1) + \ldots \\ &+ a_n(x-x_0)(x-x_1) \ldots (x-x_{n-1}) \ , \end{aligned}$$

which reduces to (1) if all $x_k = 0$, and to (5) (with $s = 0$), if $a_k = 0$ for $k < n$. We have the Horner-type evaluation scheme

$$(7) \qquad \begin{cases} u_n = a_n \ , \\ u_k = (x-x_k)u_{k+1} + a_k, \qquad k = n-1, \ n-2, \ \ldots, \ 0 \ , \\ p(x) = u_0 \ , \end{cases}$$

which is quite stable if the differences $x-x_k$ are evaluated as above in (iii), and the parameters x_k, a_k are selected to make the two additive terms on the right of (7) of equal sign. This can always be done (Mesztenyi and Witzgall [1967]). A special form of (6) has proved useful, e.g., in approximating modified Bessel functions (Blair and Edwards [1974])

(v) Lagrange form. Given any $n+1$ distinct real numbers $x_0, x_1, \ldots, x_n$, we may represent a polynomial of degree n in its Lagrange form

$$(8) \qquad p(x) = \sum_{k=0}^{n} a_k \ell_k(x), \quad \ell_k(x) = \prod_{\substack{r=0 \\ r\neq k}}^{n} \frac{x-x_r}{x_k-x_r}, \qquad a_k = p(x_k) ,$$

familiar from interpolation theory. It is evaluated most conveniently in the barycentric form (see, e.g., Bulirsch and Rutishauser [1968])

$$(9) \qquad p(x) = \frac{\sum_{k=0}^{n} a_k \frac{\lambda_k}{x-x_k}}{\sum_{k=0}^{n} \frac{\lambda_k}{x-x_k}} \qquad (x \neq x_i, \quad i = 0, 1, \ldots, n) ,$$

where $\lambda_k = \prod_{r\neq k} (x_k - x_r)^{-1}$ are precomputed constants.

(vi) Ultraeconomic forms. There are a number of representations, due to Motzkin, Belaga, Pan, and others, which require only of the order $n/2$ multiplications and n additions. While these forms are highly interesting from the standpoint of complexity theory, their practical merits are not entirely clear. For one thing, they tend to be poorly conditioned (Rice [1965], Fike [1967]), although this matter deserves further analysis. For another, the time saving gained by fewer multiplications may well be lost on some computers by the need for more memory transactions (Cody [1967]).

1.5.2. Rational functions

(i) Polynomial ratio form. This is the collective name given to

all the forms that can be obtained by representing the polynomials p and q in

$$r(x) = \frac{p(x)}{q(x)} \tag{10}$$

in any one of the forms discussed in 1.5.1. Since division is a stable operation, the conditioning and stability properties of r depend entirely on those of p and q. Occasionally it is preferable (see, e.g., Cody and Hillstrom [1970, p. 676]) to write the two polynomials in descending powers of x.

(ii) Continued fraction forms. Intrinsically different are representations of r in terms of continued fractions. There are many different types of continued fractions that can be used in this connection. We mention only the J-fractions (cf. 1.4.2), which are of the form

$$r(x) = \frac{r_1}{x+s_1+} \; \frac{r_2}{x+s_2+} \; \cdots \; \frac{r_n}{x+s_n}, \quad r_k \neq 0 \text{ all } k, \tag{11}$$

and refer to Hart et al. [1968, p. 73ff] for others. The continued fraction (11) represents a rational function in $\mathbb{R}_{n,n-1}$. Conversely, a rational function in $\mathbb{R}_{n,n-1}$ can be represented in the form (11), unless certain determinants in the coefficients of p and q happen to vanish (Wall [1948, p. 165]). Conversion algorithms are given in Hart et al. [1968, pp. 155-160].

For the evaluation of (11) one proceeds most easily "from tail to head", according to

$$\begin{cases} u_{n+1} = 0, \\ u_k = \dfrac{r_k}{x+s_k+u_{k+1}}, \qquad k = n, n-1, \ldots, 1, \\ r(x) = u_1. \end{cases} \tag{12}$$

This requires $2n-1$ additions and n divisions, which, unless division is very slow, compares favorably with the $2n-1$ additions, $2n-1$ multiplications, and 1 division, required with Horner's scheme in (10), and even more favorably with the evaluation of the continued fraction by means of the fundamental three-term recurrence relation. The algorithm (12) is not only more economical than Horner's scheme, but also more stable, in general. There are, however, exceptions (Cody and Hillstrom [1967, p. 203]). The stability of evaluation schemes for continued fractions is discussed by Macon and Baskervill [1956], Blanch [1964] and Jones and Thron [1974a, c].

1.5.3. Orthogonal sums. The Chebyshev polynomials T_k in (3) are a special case of orthogonal polynomials, $\{\pi_k\}$, which are known to satisfy a recurrence relation of the form (cf. 1.4.2 (17))

$$\pi_{r+1} = \alpha_r(x)\,\pi_r + \beta_r(x)\,\pi_{r-1}, \qquad r = 1, 2, 3, \ldots \quad . \tag{13}$$

Other (nonpolynomial) systems of special functions also satisfy relations of this type. When expanding a given function in terms of π_k, it is useful to have an efficient algorithm for evaluating a partial sum,

$$s(x) = \sum_{k=0}^{n} a_k\,\pi_k(x) \quad . \tag{14}$$

One such algorithm is Clenshaw's algorithm (Clenshaw [1955]), a generalization of the algorithm in (4),

$$\begin{cases} u_n = a_n, \quad u_{n+1} = 0 \,, \\ u_k = \alpha_k(x)\,u_{k+1} + \beta_{k+1}(x)\,u_{k+2} + a_k \\ \qquad\qquad k = n-1,\ n-2,\ \ldots,\ 0 \,, \\ s(x) = u_0\pi_0(x) + u_1[\pi_1(x) - \alpha_0(x)\,\pi_0(x)] \quad . \end{cases} \tag{15}$$

The validity of (15) is best seen by writing (13) in matrix form (Deuflhard [1974]) as

$$\begin{bmatrix} 1 & & & & \\ -\alpha_0 & 1 & & O & \\ -\beta_1 & -\alpha_1 & 1 & & \\ & \ddots & \ddots & \ddots & \\ O & & -\beta_{n-1} & -\alpha_{n-1} & 1 \end{bmatrix} \begin{bmatrix} \pi_0 \\ \pi_1 \\ \pi_2 \\ \vdots \\ \pi_n \end{bmatrix} = \begin{bmatrix} \pi_0 \\ \pi_1 - \alpha_0 \pi_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} ,$$

or briefly as $L\pi = \rho$, and noting that the recurrence in (15) is simply $L^T u = a$. Thus,

$$s(x) = a^T \pi = a^T L^{-1} \rho = ((L^T)^{-1} a)^T \rho = u^T \rho ,$$

which is the last line of (15). The argument clearly extends to functions π_k satisfying recurrence relations of order larger than two. Other possible extensions are considered in Saffren and Ng [1971].

For orthogonal polynomials $\{\pi_k\}$ one has $\pi_1 - \alpha_0(x)\pi_0 = 0$, and the last line in (15) simplifies to $s(x) = u_0 \pi_0$. The derivative $s'(x)$ can be computed by a similar algorithm (Smith [1965], Cooper [1967]).

Applying (15) to the Chebyshev polynomials T_k, U_k of the first and second kind, and noting that $T_k(\cos\theta) = \cos k\theta$, $(\sin\theta)\, U_k(\cos\theta) = \sin(k+1)\theta$, one is led to an algorithm for evaluating trigonometric sums, known as Goertzel's algorithm (Goertzel [1958], [1960]), or Watt's algorithm (Watt [1958/59]).

Floating-point error analyses of Clenshaw's and Goertzel's algorithms are given by Gentleman [1969/70], Newbery [1973] and Cox [unpubl.]. Although usually quite stable, these algorithms are not without pitfalls. Clenshaw's algorithm (15), e.g., should be avoided if $\{\pi_k\}$ is a minimal solution of (13) (cf. 2.2.1). The computation of $s(x)$ in the final step of (15) then likely leads to large cancellation errors (Elliott [1968]). Goertzel's algorithm, in turn, suffers from substantial

accumulation of rounding errors if θ is small modulo π (Gentleman [1969/70]). It can be stabilized either by incorporating phase shifts (Newbery [1973]), or by reformulating the recurrence in a manner proposed by Reinsch (Stoer [1972, p. 64]).

For computational experiments with Clenshaw's algorithm see Ng [1968/69].

§2. Methods based on linear recurrence relations

It is often necessary to compute not just one particular function, but a whole sequence of special functions. The task is considerably simplified if the members of the sequence satisfy a recurrence relation. It is then possible to compute each member recursively in terms of those already computed. The process is not only fast, but also well adapted to modern computing machinery, and may be useful even if only one member of the sequence is desired.

Most recurrences of interest in special functions are linear difference equations. The particular solution desired is often rapidly decaying, but embedded in a family of growing solutions. The question of numerical stability then becomes a central issue. In order to keep the dominant solutions in check, special precautions need to be adopted. The nature of these precautions is the subject of this paragraph.

Computational aspects of recurrence relations have been reviewed by several writers, notably Fox [1965], Gautschi [1967], [1972], Wimp [1970], and Amos [1970].

2.1. First-order recurrence relations

The simplest linear recurrence is

$$y_{n+1} = a_n y_n, \qquad n = 0, 1, 2, \ldots , \tag{1}$$

where y_0 and $a_n \neq 0$ are given numbers. Multiplication being a stable operation, errors due to rounding will essentially accumulate linearly

with n, making (1) a stable computational process. A classic example is the recurrence relation for the gamma function.

As we proceed to inhomogeneous recurrences,

$$y_{n+1} = a_n y_n + b_n, \qquad n = 0, 1, 2, \ldots , \tag{2}$$

the stability characteristics may change significantly. The relation (2) indeed involves repeated additions, thus potentially unstable operations. It suffices that the two terms on the right be nearly equal in magnitude and opposite in sign to cause significant loss of accuracy, due to "cancellation". If this happens repeatedly, the computation may quickly deteriorate, giving rise to numerical instability.

2.1.1. A simple analysis of numerical stability. Suppose $f_n \neq 0$ is a solution of (2) that we wish to compute. It is instructive to examine how a relative error ϵ in f_n, committed at $n = s$ (s for "starting"), affects the value of f_n at $n = t$ (t for "terminal"), where $t \lesseqgtr s$, assuming that no further errors are being introduced. If we denote the perturbed solution by f_n^*, so that $f_s^* = (1 + \epsilon) f_s$, we find by a simple computation that

$$f_t^* = (1 + \frac{\rho_t}{\rho_s}\epsilon) f_t , \tag{3}$$

where

$$\rho_n = \frac{f_0 h_n}{f_n} , \tag{4}$$

and h_n is the solution of the homogeneous recurrence (1), with $h_0 = 1$. Going from s to t, the relative error is thus amplified if $|\rho_t| > |\rho_s|$, and damped if $|\rho_t| < |\rho_s|$. In an effort to maintain optimal numerical stability, the recurrence (2), therefore, should be applied in the direction of decreasing $|\rho_n|$, whenever practicable.

An important special case is

$$\lim_{n\to\infty} |\rho_n| = \infty \,, \tag{5}$$

where $|\rho_n|$ diverges monotonically. The recurrence (2) is then unstable in the forward direction, the ratio $|\rho_t/\rho_s|$ being unbounded for $t > s$, but stable in the backward direction, the same ratio now being bounded by 1. More than that, we can start the recursion arbitrarily with $f_\nu^* = 0$, for some ν sufficiently large, and recur downward to some fixed n, thereby obtaining f_n to arbitrarily high accuracy. This is because the initial error, $\epsilon = -1$, according to (3), will be damped by a factor of $|\rho_n/\rho_\nu|$, which can be made arbitrarily small by choosing ν large enough. All intermediate rounding errors, moreover, are being consistently damped.

We can interprete (5) by saying that the particular solution of (2) desired is dominated by the "complementary solution" of (2), i.e., the solution of the corresponding homogeneous recurrence (1). It should be clear on intuitive grounds that forward recurrence cannot be stable under these circumstances.

We remark that similar stability considerations apply to general systems of linear difference equations (Gautschi [1972]).

2.1.2. Applications to special functions. Although not many special functions obey relations of the type (2), there are some which do, e.g., certain integrals in the theory of molecular structure (Gautschi [1961]), the incomplete gamma function (Kohútová [1970], Amos and Burgmeier [1973]), in particular the exponential integrals $E_n(z)$ (Gautschi [1973]), and successive derivatives of $f(z)/z$ (Gautschi [1966], [1972], Gautschi and Klein [1970]). The techniques indicated above provide effective schemes of computation in all these cases.

2.2. Homogeneous second-order recurrence relations

We assume now, more importantly, that f_n satisfies a three-term recurrence relation

(1) $$y_{n+1} + a_n y_n + b_n y_{n-1} = 0, \qquad n = 1, 2, 3, \ldots ,$$

where, for simplicity,

(2) $$f_n \neq 0, \quad b_n \neq 0 \quad \text{for all} \quad n .$$

Given f_0 and f_1, we can use (1) in turn for $n = 1, 2, \ldots$ to successively calculate $f_2, f_3, \ldots$. This is quite effective if f_0 and f_1 are easily calculated and the recurrence (1) is numerically stable. We expect the latter to be the case if no solution of (1) grows faster than f_n . An important example of such a recursion is the one for orthogonal polynomials, $f_n = \pi_n$, where the second solution is the sequence of associated orthogonal polynomials, $g_n = \sigma_n$ (cf. 1.4.2), and where by a theorem of Markov the ratios σ_n/π_n converge to a finite limit, the corresponding Stieltjes integral, at least outside the interval of orthogonality (Perron [1957, p. 198ff]). The recurrence relation is reputed to be stable even on the interval of orthogonality, except possibly in the vicinities of the endpoints.

If there are solutions which grow much faster than f_n, then forward recursion on (1), as in 2.1(5), is bound to fail. Such is the case if the solution f_n is minimal.

2.2.1. Minimal solutions. We call a solution f_n of (1) minimal, if for every other, linearly independent, solution g_n we have

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

All solutions g_n, for which (3) holds, are called dominant. A minimal solution, if one exists, is unique apart from a constant factor. It can be specified by imposing a single condition, e.g.,

(4) $$f_0 = s ,$$

or more generally,

$$\sum_{m=0}^{\infty} \lambda_m f_m = s \,, \tag{5}$$

where s and λ_m are given numbers.

Defining

$$r_n = f_{n+1}/f_n \,, \qquad n = 0, 1, 2, \ldots \,, \tag{6}$$

we have by a result of Pincherle (see, e.g., Perron [1957, Satz 2.46C], Gautschi [1967]) that

$$r_{n-1} = \frac{f_n}{f_{n-1}} = \frac{-b_n}{a_n -} \frac{b_{n+1}}{a_{n+1} -} \frac{b_{n+2}}{a_{n+2} -} \cdots, \qquad n = 1, 2, 3, \ldots \,, \tag{7}$$

where the continued fractions converge precisely if (1) has a (nonvanishing) minimal solution, f_n. In principle, therefore, all ratios r_{n-1} are known, and (5) gives

$$f_0 = \frac{s}{\sum_{m=0}^{\infty} \lambda_m r_0 r_1 \cdots r_{m-1}} \,, \tag{8}$$

from which

$$f_n = r_{n-1} f_{n-1}, \qquad n = 1, 2, 3, \ldots \,, \tag{9}$$

by virtue of (6).

2.2.2. Algorithms for minimal solutions. Any implementation of the approach just described will involve, explicitly or implicitly, the truncated continued fractions

$$r_{n-1}^{(\nu)} = \frac{-b_n}{a_n -} \frac{b_{n+1}}{a_{n+1} -} \cdots \frac{b_\nu}{a_\nu} \,, \qquad n = 1, 2, \ldots, \nu \,. \tag{10}$$

We assume they all exist. (They do for ν sufficiently large.) For simplicity of exposition, we consider the case of prescribed f_0, see (4).

(i) Nonlinear backward recursion. Evaluating (10) recursively from behind, and then using an approximate version of (9), we get the algorithm (Gautschi [1967])

$$(11)\qquad \begin{cases} r_{\nu}^{(\nu)} = 0, \quad r_{n-1}^{(\nu)} = -\dfrac{b_n}{a_n + r_n^{(\nu)}}, & n = \nu, \nu-1, \ldots, 1, \\[2ex] f_0^{(\nu)} = f_0, \quad f_n^{(\nu)} = r_{n-1}^{(\nu)} f_{n-1}^{(\nu)}, & n = 1, 2, \ldots, \nu. \end{cases}$$

From Pincherle's result it follows that

$$(12)\qquad \lim_{\nu\to\infty} f_n^{(\nu)} = f_n$$

for any fixed n.

The major inconvenience with (11) is the fact that we do not always know an appropriate value of ν ahead of time, and may have to repeat (11) several times, with increasing ν, until the $f_n^{(\nu)}$ converge to the desired accuracy.

Replacing $r_{\nu}^{(\nu)} = 0$ in (11) by $r_{\nu}^{(\nu)} = \rho_{\nu}$, a suitable approximation of $r_{\nu} = f_{\nu+1}/f_{\nu}$, often leads to improved convergence (Gautschi [1967, pp. 38, 40], Scraton [1972]).

(ii) Linear algebraic system. The approximations $f_n^{(\nu)}$ in (11) can be identified with the solution of the tridiagonal system

$$(13)\qquad \begin{bmatrix} a_1 & 1 & & & \\ b_2 & a_2 & 1 & & O \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots \\ O & & & b_{\nu} & a_{\nu} \end{bmatrix} \begin{bmatrix} f_1^{(\nu)} \\ f_2^{(\nu)} \\ \vdots \\ \vdots \\ f_{\nu}^{(\nu)} \end{bmatrix} = \begin{bmatrix} -b_1 f_0 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix},$$

which is formally obtained from (1) by setting $y_{\nu+1} = 0$.

(iii) Miller's backward recurrence algorithm. We may start the recurrence (1) with

$$\eta_\nu = 1, \quad \eta_{\nu+1} = 0 \ , \tag{14}$$

and use it in the backward direction to obtain $\eta_n = \eta_n^{(\nu)}$, $n = \nu-1, \nu-2, \ldots, 1$. In effect, we produce a solution of the linear system (13), where f_0 on the right is replaced by $\eta_0^{(\nu)}$. Consequently,

$$f_n^{(\nu)} = \frac{f_0}{\eta_0^{(\nu)}} \eta_n^{(\nu)}, \qquad n = 0, 1, \ldots, \nu \ . \tag{15}$$

Generating $\eta_n^{(\nu)}$ as described, and then $f_n^{(\nu)}$ by (15), is known as Miller's algorithm (British Association for the Advancement of Science [1952, p. xvii]). It has the same disadvantage as noted in (i). In addition, the quantities $\eta_n^{(\nu)}$ may become large enough to cause overflow on a computer.

(iv) Olver's algorithm. Miller's algorithm can be thought of as solving the system (13) by a form of Gauss elimination, in which the elimination is performed backwards, from the last equation to the first, and the solution then obtained by forward substitution. The algorithm proposed by Olver [1967a] uses the more conventional forward elimination followed by back substitution. To describe it, let

$$\left.\begin{aligned} p_0 &= 0, \quad p_1 = 1, \quad e_0 = f_0 \ , \\ p_{n+1} &= -a_n p_n - b_n p_{n-1} \\ e_n &= b_n e_{n-1} \end{aligned}\right\} \qquad n = 1, 2, \ldots, \nu \ . \tag{16}$$

Then

$$f_{\nu+1}^{(\nu)} = 0, \quad p_{n+1} f_n^{(\nu)} - p_n f_{n+1}^{(\nu)} = e_n, \qquad n = \nu, \nu-1, \ldots, 1 \ , \tag{17}$$

which yields $f_\nu^{(\nu)}, f_{\nu-1}^{(\nu)}, \ldots, f_1^{(\nu)}$ in this order, provided none of the p_n vanishes.

We note from (17) that

$$\frac{f_n^{(\nu)}}{p_n} - \frac{f_{n+1}^{(\nu)}}{p_{n+1}} = \frac{e_n}{p_n p_{n+1}} \quad ,$$

so that

$$f_n^{(\nu)} = p_n \sum_{k=n}^{\nu} \frac{e_k}{p_k p_{k+1}}, \qquad n = 1, 2, \ldots, \nu \quad . \tag{18}$$

In particular, by (12),

$$f_n = p_n \sum_{k=n}^{\infty} \frac{e_k}{p_k p_{k+1}} \quad . \tag{19}$$

It follows that $f_n^{(\nu)}$ has relative error

$$\frac{f_n - f_n^{(\nu)}}{f_n} = \frac{\sum_{k=\nu+1}^{\infty} e_k / p_k p_{k+1}}{\sum_{k=n}^{\infty} e_k / p_k p_{k+1}} \doteq \frac{e_{\nu+1}}{p_{\nu+1} p_{\nu+2}} \Big/ \frac{e_n}{p_n p_{n+1}} \quad , \tag{20}$$

the approximation on the far right being valid if the series in (19) converges rapidly. (Using the techniques in Olver [1967b] one could estimate the series more carefully and thus obtain a rigorous error bound). If we wish to obtain f_n to within a relative error ϵ, we may thus iterate with (16) until a value of ν is reached for which

$$\left| \frac{e_{\nu+1}}{p_{\nu+1} p_{\nu+2}} \right| \leq \epsilon \min_n \left| \frac{e_n}{p_n p_{n+1}} \right| \quad , \tag{21}$$

the minimum being taken over all values n of interest. With ν so determined, the $f_n^{(\nu)}$ are then obtained as described in (17). It is this feature of automatically determining ν, which makes Olver's algorithm attractive.

(v) Olver's and Miller's algorithm combined. In some applications, the recursion in (16) for p_n is mildly unstable, initially, although ultimately it is always stable. Olver and Sookne [1972] therefore suggest applying the procedure (16), which serves mainly to determine the cutoff-index ν, only in a region $n \geq n_0$ of perfect stability for the p-recursion, starting with $p_{n_0} = 0$, $p_{n_0+1} = 1$ as before, but with $e_{n_0} = 1$. Once ν is determined, the desired approximations are then obtained by recurring backward, as in Miller's procedure, starting with $f^{(\nu)}_{\nu+1} = 0$, $f^{(\nu)}_{\nu} = e_\nu/p_{\nu+1}$, and by a final normalization, as in (15).

We remark that all algorithms described can be extended to accommodate the more general "normalization condition" (5). This is an important point, inasmuch as the algorithms so extended do not require the calculation of any particular value of f_n (such as f_0 above). For details, we refer to the cited references.

2.2.3. Applications to special functions. The algorithms of 2.2.2 have been applied to a large number of special functions. The first major applications involved Bessel functions and Coulomb wave functions, whose recurrence relations are similar in nature. Further applications soon followed, e.g., to Legendre functions, incomplete beta and gamma functions, repeated integrals of the error function, and others. Detailed references, up to about 1965, can be found in Gautschi [1967]. More recently, in connection with Bessel functions, Mechel [1968] and Cylkowski [1971] discuss appropriate choices of the starting index ν in Miller's algorithm, while Amos [1974] proposes accurate starting values from uniform asymptotic expansions. The latter approach, combined with Taylor expansion where appropriate, is carefully implemented in Amos and Daniel [1973] and Amos, Daniel and Weston [unpubl.]. Ratios of successive Bessel functions (and of other functions, e.g., the repeated integrals of the error function) can also be computed by an iterative algorithm based on certain inequalities satisfied by these ratios (Amos [1973], [1974]). For Bessel functions, this is implemented in Amos and Daniel [1973]. Still on Bessel functions, we mention the work of Luke

[1972b], which relates Miller's algorithm to certain rational approximations in the theory of hypergeometric functions, and the computer implementation and certification of Olver's algorithm by Sookne [1973a, b, c, d]. Sidonskiĭ [1967] has a related algorithm for Bessel functions of integer order and real argument, furnishing upper and lower bounds. Hitotumatu [1967/68] recommends a nonlinear normalization condition in place of the linear condition (5). On Coulomb wave functions we note a recent improvement by Gautschi [1969] on the recurrence algorithm (i), and refer to Wills [1971] for a procedure very similar to Olver's. Kölbig [1972] gives a survey of computational methods for Coulomb wave functions. Legendre functions are discussed by Fettis [1967] and more recently by Amos and Bulgren [1969] in connection with series expansions for the bivariate t-distribution in statistics. Bardo and Ruedenberg [1971] revisit the repeated integrals of the error function. Temme [1972] applies algorithm (i) to certain Laplace integrals connected with van Wijngaarden's transformation of formal series.

The stability of forward recurrence is analyzed by Wimp [1971/72], and in the case of orthogonal polynomials of the Laguerre and Hermite type, by Baburin and Lebedev [1967].

2.3. Inhomogeneous second-order and higher-order recurrence relations

Some of the more esoteric functions are solutions of inhomogeneous second-order recurrence relations,

$$y_{n+1} + a_n y_n + b_n y_{n-1} = c_n, \qquad n = 1, 2, 3, \ldots \quad . \tag{1}$$

Others satisfy recurrences of even higher order. The latter are also encountered in the computation of expansion coefficients, e.g., the coefficients in a Taylor series or a series in Chebyshev polynomials. Frequently, the solutions of interest are of the recessive type, in which case some of the algorithms described in 2.2.2, suitably extended, are again effective.

2.3.1. Subdominant solutions of inhomogeneous second-order recurrence relations. Assume that the homogeneous recurrence associated with (1) has a pair of linearly independent solutions g_n and h_n, of which g_n is minimal (with $g_0 \neq 0$), hence h_n dominant. We then call a solution f_n of (1) subdominant if

$$\lim_{n\to\infty} f_n/h_n = 0 \quad . \tag{2}$$

A subdominant solution may or may not dominate the minimal solution g_n. If it does, neither forward nor backward recurrence is entirely satisfactory.

In analogy to 2.2(13) we consider the linear algebraic system

$$\begin{bmatrix} a_1 & 1 & & & O \\ b_2 & a_2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ O & & & b_\nu & a_\nu \end{bmatrix} \begin{bmatrix} f_1^{(\nu)} \\ f_2^{(\nu)} \\ \vdots \\ f_\nu^{(\nu)} \end{bmatrix} = \begin{bmatrix} c_1 - b_1 f_0 \\ c_2 \\ \vdots \\ c_\nu \end{bmatrix} . \tag{3}$$

If this system has a solution for all ν sufficiently large, and if f_n is a subdominant solution of (1), then by a result of Olver [1967a],

$$\lim_{\nu\to\infty} f_n^{(\nu)} = f_n \quad . \tag{4}$$

The algorithms of Olver and Olver and Sookne (cf. 2.2.2(iv), (v)) thus extend readily to the case of subdominant solutions. So does in particular Olver's device for determining the appropriate ν and estimating the error (Olver [1967a, b]). Related algorithms are also discussed in Amos and Burgmeier [1973].

Olver applies his algorithm to Anger-Weber and Struve functions, while Amos and Burgmeier apply theirs to numerous other special functions, including incomplete Laplace transforms, and moments, of Bessel

and Struve functions, the incomplete gamma function and Lommel functions. Sadowski and Lozier [1972] give an interesting application of Olver's algorithm to certain definite integrals in plasma physics, involving Chebyshev polynomials. Similar integrals are also treated by Piessens and Branders [1973].

2.3.2. Higher-order recurrence relations. Miller's algorithm is applicable to recurrence relations of arbitrary order, but, unless substantially modified, is effective only for solutions which are "sufficiently minimal". For a penetrating study of this we refer to Wimp [1969]. There are applications to hypergeometric and confluent hypergeometric functions in Wimp [1969], as well as in Wimp [1974], and another application in Wimp and Luke [1969]. Thacher [1972] discusses Miller's algorithm in connection with the solution in power series of linear differential equations with polynomial coefficients and relates minimality of the expansion coefficients to the singularities of the differential equation.

Given enough information about the growth pattern of fundamental solutions, approaches via boundary value problems appear to be more widely applicable. By imposing the right boundary conditions, it is sometimes possible to filter out a desired solution which is neither minimal nor dominant. The principal references in this direction are Oliver [1966/67], [1968a, b].

§3. Nonlinear recurrence algorithms for elliptic integrals and elliptic functions

Some functions of several variables, notably elliptic integrals, have the remarkable property that their values remain unchanged as the variables undergo certain nonlinear transformations. Repeated application of these transformations, moreover, causes the variables to converge rapidly to certain limiting values, for which the functions can be evaluated by elementary means. These invariance properties thus give

rise to interesting and powerful recursive algorithms for computing the functions in question.

3.1. Elliptic integrals and Jacobian elliptic functions

3.1.1. Definitions and special values. The best known functions enjoying invariance properties of the type indicated are the elliptic integrals of the first, second, and third kind. In Legendre's canonical form, they are, respectively,

$$F(\varphi, k) = \int_0^{\varphi} \frac{d\theta}{\sqrt{1-k^2\sin^2\theta}} , \tag{1}$$

$$E(\varphi, k) = \int_0^{\varphi} \sqrt{1-k^2\sin^2\theta}\, d\theta , \tag{2}$$

$$\Pi(\varphi, n, k) = \int_0^{\varphi} \frac{d\theta}{(1+n\sin^2\theta)\sqrt{1-k^2\sin^2\theta}} . \tag{3}$$

The variable k is known as the modulus; we assume it in the interval $0 \le k \le 1$. The complementary modulus k' is defined by

$$k' = \sqrt{1-k^2} . \tag{4}$$

The variable φ is called the amplitude, and we assume that $0 \le \varphi \le \pi/2$. The variable n in (3) may take on arbitrary values, provided the integral is interpreted in the sense of a Cauchy principal value, should n be negative and $1 + n\sin^2\varphi < 0$.

The integrals (1)-(3) are called complete, or incomplete, depending on whether $\varphi = \pi/2$, or $\varphi < \pi/2$. The complete elliptic integrals are usually denoted by

$$\mathbb{K}(k) = F(\tfrac{\pi}{2}, k), \quad \mathbb{E}(k) = E(\tfrac{\pi}{2}, k), \quad \Pi(n, k) = \Pi(\tfrac{\pi}{2}, n, k) . \tag{5}$$

As $k \downarrow 0$, or $k \uparrow 1$, we have the limiting values

$$\lim_{k \downarrow 0} F(\varphi, k) = \lim_{k \downarrow 0} E(\varphi, k) = \varphi \quad , \tag{6}$$

$$\lim_{k \uparrow 1} F(\varphi, k) = \tanh^{-1}(\sin \varphi) \ (0 \leq \varphi < \pi/2), \quad \lim_{k \uparrow 1} E(\varphi, k) = \sin \varphi \quad . \tag{7}$$

Similar, but more complicated formulas hold for $\Pi(\varphi, n, k)$ (see, e.g., Byrd and Friedman [1971, p. 10]). We also note

$$F(\varphi, k) \sim \ell n \frac{4}{\cos \varphi + \sqrt{1-k^2 \sin^2 \varphi}}, \quad E(\varphi, k) \sim 1 \text{ as } k \uparrow 1, \ \varphi \uparrow \pi/2 \quad , \tag{8}$$

where the first relation is given by Carlson [1965, p. 39]; see also Nellis and Carlson [1966, p. 228].

Considering k fixed, the function $u = F(\varphi, k)$ is monotone in φ, and thus possesses an inverse function,

$$\varphi = \operatorname{am} u \quad , \tag{9}$$

the <u>amplitude function</u>. In terms of it one defines <u>Jacobian elliptic functions</u> by

$$\operatorname{sn} u = \sin \varphi, \quad \operatorname{cn} u = \cos \varphi, \quad \operatorname{dn} u = \sqrt{1-k^2 \sin^2 \varphi} \quad . \tag{10}$$

3.1.2. <u>Gauss transformations vs. Landen transformations</u>. One distinguishes between <u>Gauss transformations</u> and <u>Landen transformations</u>, and for each between <u>descending</u> and <u>ascending</u> transformations. (Terminology, however, varies). In a descending transformation, the modulus k always decreases; in an ascending transformation, it always increases.

In a Gauss transformation, the amplitude φ varies in parallel with k (i.e., φ and k both increase or both decrease), while in a Landen transformation they vary in opposite directions. Repeated application of a descending transformation causes k to converge down to zero, while φ converges down to some limiting value φ_∞ in a Gauss transformation and up to ∞ in a Landen transformation. The former, therefore, eventually invokes the equations in (6). Repeated application of an ascending transformation, instead, causes k to converge upward to 1, while φ converges upward to $\pi/2$ in a Gauss transformation and down to some limiting value φ_∞ in a Landen transformation. The former, therefore, eventually invokes the relations in (8), the latter those in (7).

In describing these transformations, we limit ourselves to elliptic integrals of the first kind, and must refer to the literature for the others. An early treatment of computational algorithms for elliptic functions and integrals is King [1924]. We follow more closely the work of Carlson [1965], who develops the algorithms in a unified way, at least for integrals of the first two kinds. Hofsommer and van de Riet [1963] have ALGOL programs for integrals of the first and second kind, using Landen transformations, as well as programs for elliptic functions, based on ascending Landen and descending Gauss transformations. See also Neuman [1969/70a, b] and Kami, Kiyoto and Arakawa [1971a, b]. Descending transformations for integrals of the third kind are discussed by Ward [1960] in the case of complete integrals, and by Fettis [1965] in the case of incomplete integrals. A thorough treatment of descending Gauss and Landen transformations for integrals of all three kinds, complete with ALGOL procedures, is given in Bulirsch [1965a, b], and more definitively, especially as regards integrals of the third kind, in Bulirsch [1969a, b]. In the latter work, more general transformations, ascribed to Bartky, and extensions thereof, are used effectively. A good introduction into these developments is Bulirsch and Stoer [1968]. For the theory of elliptic integrals and elliptic functions we refer to the books of Neville [1944], [1971] and Tricomi [1948], [1951].

We begin with Gauss' process of the arithmetic-geometric mean, which underlies all algorithms for elliptic functions.

3.2. Gauss' algorithm of the arithmetic-geometric mean

Starting with $a_0 > b_0 > 0$, Gauss' algorithm generates two sequences $\{a_n\}$, $\{b_n\}$ by compounding the arithmetic and the geometric mean in the following manner,

$$(1) \qquad \begin{cases} a_{n+1} = \frac{1}{2}(a_n + b_n), \\ b_{n+1} = \sqrt{a_n b_n}, \end{cases} \qquad n = 0, 1, 2, \ldots .$$

Since the iteration functions are homogeneous of degree 1, only the ratio b_0/a_0 matters.

The arithmetic mean being larger than the geometric mean, we have $a_n > b_n$ for all n, and therefore $b_0 < a_{n+1} < a_n$, $b_n < b_{n+1} < a_0$. It follows that $\{a_n\}$ and $\{b_n\}$ both converge monotonically to certain limits, which, by letting $n \to \infty$ in (1), are readily found to be equal. The common limit is denoted by $M = M(a_0, b_0)$, and called the arithmetic-geometric mean of a_0 and b_0. Clearly, $b_0 < M(a_0, b_0) < a_0$.

In order to discuss the rate of convergence, it is convenient to introduce

$$(2) \qquad \epsilon_n = \frac{a_n - b_n}{a_n + b_n} .$$

One finds by a simple computation that

$$(3) \qquad \epsilon_{n+1} = \left(\frac{\epsilon_n}{1 + \sqrt{1 - \epsilon_n^2}} \right)^2 , \qquad n = 0, 1, 2, \ldots .$$

The sequence $\{\epsilon_n\}$, therefore, converges monotonically and quadratically to zero. Since

$$0 < a_n - M < (a_0 + M)\epsilon_n, \quad 0 < M - b_n < 2M\epsilon_n , \tag{4}$$

we see that also $\{a_n\}$ and $\{b_n\}$ converge quadratically. We note from (2) that

$$\frac{b_n}{a_n} = 1 - 2\epsilon_n + O(\epsilon_n^2), \quad n \to \infty . \tag{5}$$

Quadratic convergence is a common feature of more general processes of compounding means (Lehmer [1971]). For variants of Gauss' algorithm (none of which quadratically convergent, however), and for many historical notes, see also Carlson [1971]. For complex variables, the algorithm is discussed by Fettis and Caslin [1969] and Morita and Horiguchi [1972/73].

In applications to elliptic integrals, the ratio b_0/a_0 will be identified with either the modulus k, or the complementary modulus k'. The algorithm (1) then generates a sequence of transformed moduli $k_n = b_n/a_n$, or $k'_n = b_n/a_n$, respectively, where in the former case

$$k_{n+1} = \frac{2\sqrt{k_n}}{1+k_n}, \qquad n = 0, 1, 2, \ldots, \ k_0 = k , \tag{6}$$

and in the latter,

$$k'_{n+1} = \frac{2\sqrt{k'_n}}{1+k'_n}, \qquad n = 0, 1, 2, \ldots, \ k'_0 = k' . \tag{7'}$$

An equivalent form of (7') is

$$k_{n+1} = \frac{1-k'_n}{1+k'_n}, \qquad n = 0, 1, 2, \ldots, \ k_0 = k . \tag{7}$$

Since the modulus increases in (6), and decreases in (7), we call (6) an ascending and (7) a descending transformation. The convergence is to 1 and 0, respectively, and quadratic in both cases.

The choice of the transformation is dictated by the speed of convergence, which depends on the magnitude of $\epsilon_0 = (1-b_0/a_0)/(1+b_0/a_0)$. Since we want ϵ_0 small, we choose an ascending transformation if $k^2 > \frac{1}{2}$, and a descending transformation otherwise, so that in either case $1 > (b_0/a_0)^2 \geq \frac{1}{2}$, and thus

$$\epsilon_0 \leq \frac{1-2^{-\frac{1}{2}}}{1+2^{-\frac{1}{2}}} < .172 \ .$$

From (3) we then find that

$$(8) \qquad \epsilon_{n+1} = \left(\frac{\epsilon_n}{1+\sqrt{1-\epsilon_n^2}} \right)^2 < \frac{\epsilon_n^2}{(1+\sqrt{1-\epsilon_0^2})^2} < \frac{\epsilon_n^2}{3.94} \ ,$$

and so,

$$(9) \quad \epsilon_1 < .00751, \ \epsilon_2 < 1.44 \times 10^{-5}, \ \epsilon_3 < 5.27 \times 10^{-11}, \ \epsilon_4 < 7.05 \times 10^{-22}, \ldots \ ,$$

illustrating the quadratic nature of convergence.

3.3. Computational algorithms based on Gauss and Landen transformations

3.3.1. Descending Gauss transformation. We define

$$(1) \qquad \begin{cases} a_0 = 1, \ b_0 = k', \ t_0 = \csc \varphi \ , \\ a_{n+1} = \frac{1}{2}(a_n + b_n) \ , \\ b_{n+1} = \sqrt{a_n b_n} \ , \qquad n = 0, 1, 2, \ldots \ . \\ t_{n+1} = \frac{1}{2}(t_n + \sqrt{t_n^2 - a_n^2 + b_n^2}) \ , \end{cases}$$

One verifies without difficulty that t_n and a_n/t_n both decrease. Hence, $a_n/t_n \leq 1$, and t_n must converge,

$$t_n \downarrow T, \quad n \to \infty ,$$

where $T \geq M$. The speed of convergence is comparable to that of ϵ_n, in the sense

$$t_n - T \sim \frac{M^2}{T}\epsilon_n, \quad n \to \infty . \tag{2}$$

To see this, observe from the last relation in (1), and from 3.2(5), that

$$t_{n+1} = \frac{1}{2}t_n\left\{1 + \sqrt{1-\left(\frac{a_n}{t_n}\right)^2\left[1-\left(\frac{b_n}{a_n}\right)^2\right]}\right\} = \frac{1}{2}t_n\left\{1 + \sqrt{1-\left(\frac{a_n}{t_n}\right)^2\left[4\epsilon_n + O(\epsilon_n^2)\right]}\right\}$$

$$= \frac{1}{2}t_n\left\{1+1-2\left(\frac{M}{T}\right)^2\epsilon_n + o(\epsilon_n)\right\} = t_n\left\{1-\left(\frac{M}{T}\right)^2\epsilon_n + o(\epsilon_n)\right\} ,$$

from which

$$t_n - t_{n+1} = \frac{M^2}{T}\epsilon_n + o(\epsilon_n) .$$

Since ϵ_n converges quadratically, in particular $\epsilon_{n+1} < \epsilon_n^2$, we easily obtain, for any $p \geq 0$,

$$t_n - t_{n+p+1} = \frac{M^2}{T}\epsilon_n + o(\epsilon_n) ,$$

from which (2) follows by letting $p \to \infty$.

We now set

$$\frac{a_n}{t_n} = \sin\varphi_n \ (0 < \varphi_n < \tfrac{\pi}{2}), \quad \frac{b_n}{a_n} = k'_n, \qquad n = 0, 1, 2, \ldots , \tag{3}$$

which for $n = 0$ is consistent with the first relations in (1) (if $\varphi_0 = \varphi$, $k'_0 = k'$). The last relation in (1) can then be written in trigonometric form as

$$\sin \varphi_{n+1} = \frac{(1+k_n') \sin \varphi_n}{1+\sqrt{1 - k_n^2 \sin^2 \varphi_n}} , \qquad n = 0, 1, 2, \ldots .$$

If in the integral $F(\varphi_n, k_n) = \int_0^{\varphi_n} (1 - k_n^2 \sin^2 \theta)^{-\frac{1}{2}} d\theta$ we make the change of variables

$$\sin \lambda = \frac{(1+k_n') \sin \theta}{1+\sqrt{1 - k_n^2 \sin^2 \theta}} , \quad 0 < \theta \leq \varphi_n ,$$

we find, after a little computation, that

$$(4) \qquad \frac{1}{a_n} F(\varphi_n, k_n) = \frac{1}{a_{n+1}} F(\varphi_{n+1}, k_{n+1}), \qquad n = 0, 1, 2, \ldots .$$

This is the descending Gauss transformation for elliptic integrals $F(\varphi, k)$. Since $k_n \downarrow 0$, and recalling 3.1(6), we conclude

$$(5) \qquad F(\varphi, k) = \lim_{n \to \infty} \frac{1}{a_n} F(\varphi_n, k_n) = \frac{1}{M} \sin^{-1} \frac{M}{T} .$$

Thus, $F(\varphi, k)$ may be approximated by evaluating $a_n^{-1} \sin^{-1}(a_n/t_n)$ for some n sufficiently large. Observing that

$$0 < \frac{a_n}{t_n} - \frac{M}{T} = \frac{(a_n - M)T + M(T - t_n)}{t_n T} < \frac{a_n - M}{T} ,$$

and using Taylor's theorem, and 3.2(4), we find

$$\left| \frac{1}{M} \sin^{-1} \frac{M}{T} - \frac{1}{a_n} \sin^{-1} \frac{a_n}{t_n} \right| \leq \frac{M+1}{M^2} \left(\frac{\pi}{2} + \sec \varphi\right) \epsilon_n .$$

For $a_n^{-1} \sin^{-1}(a_n/t_n)$ to be an acceptable approximation to $F(\varphi, k)$, it

suffices, therefore, that ϵ_n be sufficiently small [which for most purposes will be the case when $n = 3$ or $n = 4$; cf. 3.2(9)].

3.3.2. Ascending Landen transformation. We define

$$(6)\qquad \begin{cases} a_0 = 1, \quad b_0 = k, \quad s_0 = \cot\varphi \ , \\ a_{n+1} = \frac{1}{2}(a_n + b_n) \ , \\ b_{n+1} = \sqrt{a_n b_n} \ , \qquad n = 0, 1, 2, \ldots \ . \\ s_{n+1} = \frac{1}{2}\left(s_n + \sqrt{s_n^2 + a_n^2 - b_n^2}\right) , \end{cases}$$

Clearly, s_n increases, while a_n/s_n decreases. An argument similar to the one surrounding (2) shows that $s_n \uparrow S$, where $S < \infty$, and in fact,

$$(7)\qquad S - s_n \sim \frac{M^2}{S}\epsilon_n, \qquad n \to \infty \ .$$

Letting

$$(8)\qquad \frac{a_n}{s_n} = \tan\varphi_n \ (0 < \varphi_n < \tfrac{\pi}{2}), \quad \frac{b_n}{a_n} = k_n, \qquad n = 0, 1, 2, \ldots \ ,$$

we can recast the last relation in (6) in the trigonometric form

$$\tan\varphi_{n+1} = \frac{(1+k_n)\tan\varphi_n}{1 + \sqrt{1 + k_n'^2 \tan^2\varphi_n}}, \qquad n = 0, 1, 2, \ldots \ .$$

Similarly as in 3.3.1, it follows that

$$(9)\qquad \frac{1}{a_n} F(\varphi_n, k_n) = \frac{1}{a_{n+1}} F(\varphi_{n+1}, k_{n+1}), \qquad n = 0, 1, 2, \ldots \ ,$$

which is known as the ascending Landen transformation. Making use of 3.1(7), we now obtain

$$F(\varphi, k) = \lim_{n\to\infty} \frac{1}{a_n} F(\varphi_n, k_n) = \frac{1}{M} \sinh^{-1} \frac{M}{S} \quad . \tag{10}$$

When S is small, there is some loss of significant figures in computing $s_n^2 + a_n^2 - b_n^2$. We can avoid this by introducing

$$d_n = \sqrt{a_n^2 - b_n^2} \quad , \tag{11}$$

and computing $s_n^2 + a_n^2 - b_n^2 = s_n^2 + d_n^2$, where the d_n are generated recursively by

$$d_{n+1} = \frac{d_n^2}{4a_{n+1}} , \qquad n = 0, 1, 2, \ldots \quad . \tag{12}$$

3.3.3. Ascending Gauss transformation. We define

$$\begin{cases} a_0 = 1, \quad b_0 = k, \quad q_0 = \csc\varphi \quad , \\ a_{n+1} = \frac{1}{2}(a_n + b_n) \quad , \\ b_{n+1} = \sqrt{a_n b_n} \quad , \qquad n = 0, 1, 2, \ldots \quad . \\ q_{n+1} = \frac{1}{2}\left(q_n + \frac{a_n b_n}{q_n}\right) \quad , \end{cases} \tag{13}$$

One verifies without difficulty that $q_n \geq a_n$ for all n. Consequently, $q_{n+1} < q_n$, and the sequence $\{q_n\}$, being monotone decreasing and bounded from below by M, converges to some limit. It is easily seen that the limit is M,

$$q_n \downarrow M, \quad n \to \infty \quad .$$

We set

$$(14)\qquad \frac{a_n}{q_n} = \sin\varphi_n \quad (0<\varphi_n<\tfrac{\pi}{2}), \quad \frac{b_n}{a_n} = k_n, \qquad n = 0, 1, 2, \ldots ,$$

and rewrite the last relation in (13) as

$$\sin\varphi_{n+1} = \frac{(1+k_n)\sin\varphi_n}{1+k_n\sin^2\varphi_n}, \qquad n = 0, 1, 2, \ldots ,$$

which shows that φ_n indeed increases. The ascending Gauss transformation takes the form

$$(15)\qquad \frac{1}{2^n a_n} F(\varphi_n, k_n) = \frac{1}{2^{n+1} a_{n+1}} F(\varphi_{n+1}, k_{n+1}), \qquad n = 0, 1, 2, \ldots .$$

In contrast to the previous transformations, $F(\varphi_n, k_n)$ no longer remains bounded as $n\to\infty$. Indeed, simultaneously $\varphi_n \uparrow \pi/2$ and $k_n \uparrow 1$, and so, by 3.1(8),

$$F(\varphi_n, k_n) \sim \ln\frac{4}{\cos\varphi_n + \sqrt{1-k_n^2\sin^2\varphi_n}}, \quad n\to\infty .$$

From (15) we obtain

$$(16)\qquad F(\varphi, k) = \frac{1}{M}\lim_{n\to\infty}\left\{2^{-n}\ln\frac{2q_n+2a_n}{\sqrt{q_n^2-a_n^2}+\sqrt{q_n^2-b_n^2}}\right\} .$$

It suffices to evaluate the expression on the right for n large enough so that ϵ_n is negligible compared to 1.

The denominator

$$e_n = \sqrt{q_n^2-a_n^2}+\sqrt{q_n^2-b_n^2}$$

can be computed without loss of significance by means of

$$4q_n e_{n+1} = e_n^2 + (a_n - b_n)^2 \quad ,$$

where the second term on the right is substantially smaller than the first, $(a_n - b_n)^2 \leq \epsilon_n e_n^2$. The cancellation error incurred in computing $a_n - b_n$, therefore, is of no consequence.

3.3.4. Descending Landen transformation. We define

$$(17) \qquad \begin{cases} a_0 = 1, \quad b_0 = k_0', \quad p_0 = \cot\varphi \quad , \\ a_{n+1} = \frac{1}{2}(a_n + b_n) \quad , \\ b_{n+1} = \sqrt{a_n b_n} \quad , \qquad\qquad n = 0, 1, 2, \ldots \quad . \\ p_{n+1} = \frac{1}{2}\left(p_n - \frac{a_n b_n}{p_n} \right) \quad , \end{cases}$$

This time, p_n cannot possibly tend to a finite limit P, as this would imply $P^2 = -M^2$, which is absurd. Neither need p_n preserve its sign.

Letting

$$(18) \qquad \frac{a_n}{p_n} = \tan\varphi_n, \quad \frac{b_n}{a_n} = k_n', \qquad n = 0, 1, 2, \ldots \quad ,$$

and writing the last relation of (17) in trigonometric form,

$$\tan\varphi_{n+1} = \frac{(1+k_n')\tan\varphi_n}{1 - k_n' \tan^2\varphi_n}, \qquad n = 0, 1, 2, \ldots \quad ,$$

we find however that φ_n increases, if we take (Carlson [1965])

$$(19) \qquad i_n \frac{\pi}{2} < \varphi_n \leq (i_n + 1)\frac{\pi}{2} \quad ,$$

where

$$(20)\qquad i_0 = 0, \quad i_n = \begin{cases} 2i_{n-1} & \text{if } p_n \geq 0 \ , \\ 2i_{n-1}+1 & \text{if } p_n < 0 \text{ or } p_n = \infty \ . \end{cases}$$

The descending Landen transformation states that

$$(21)\qquad \frac{1}{2^n a_n} F(\varphi_n, k_n) = \frac{1}{2^{n+1} a_{n+1}} F(\varphi_{n+1}, k_{n+1}), \qquad n = 0, 1, 2, \ldots \ ,$$

and consequently, since $k_n \downarrow 0$, that

$$(22)\qquad F(\varphi, k) = \frac{1}{M} \lim_{n\to\infty} 2^{-n} \varphi_n, \qquad \varphi_n = \tan^{-1} \frac{a_n}{p_n} \ .$$

The branch of the inverse tangent is to be taken in conformity with (19) and (20).

3.4. Complete elliptic integrals

All four transformations discussed in 3.3 apply equally for complete integrals. Some of them, however, simplify.

Thus, in the descending Gauss transformation, we find that $t_n = a_n$ for all n, which reduces the algorithm 3.3(1), and 3.3(5), to

$$(1)\qquad \begin{cases} a_0 = 1, \quad b_0 = k' \ , \\ a_{n+1} = \frac{1}{2}(a_n + b_n) \ , \\ b_{n+1} = \sqrt{a_n b_n} \ , \\ \mathbb{K}(k) = \frac{\pi}{2M} \ . \end{cases} \qquad n = 0, 1, 2, \ldots \ ,$$

The arithmetic-geometric mean $M = M(1, k')$ is thus seen to be related to the complete elliptic integral of the first kind, $\mathbb{K}(k)$.

Similarly, in the descending Landen transformation, we have $p_0 = 0$,

and thus $p_n = \infty$ for all $n \geq 1$, which, by 3.3(20) has the consequence that $i_n = 2^n - 1$. By 3.3(19), therefore, $\varphi_n = 2^{n-1}\pi$, and 3.3(22) then reestablishes (1). The descending Gauss and Landen transformations thus become identical.

Not so for the ascending transformations. In the ascending Gauss transformation, we find $q_n = a_n$ for all n, and 3.3(13), together with 3.3(16), where n is conveniently replaced by $n+1$, simplify to

$$(2)\qquad \begin{cases} a_0 = 1, \quad b_0 = k, \\ a_{n+1} = \frac{1}{2}(a_n + b_n), \\ b_{n+1} = \sqrt{a_n b_n}, \qquad n = 0, 1, 2, \ldots, \\ \mathbb{K}(k) = \frac{1}{2M} \lim_{n\to\infty} \left(2^{-n} \ln \frac{4}{\epsilon_n}\right). \end{cases}$$

The ascending Landen transformation, finally, neither simplifies, nor does it preserve the completeness of the integral.

3.5. Jacobian elliptic functions

All four algorithms of 3.3, suitably reversed, yield algorithms for computing Jacobian elliptic functions. We recall that, by definition,

$$(1)\qquad \operatorname{sn} u = \sin\varphi, \quad \text{where } u = F(\varphi, k), \quad 0 \leq u \leq \mathbb{K}(k).$$

In the case of the descending Gauss transformation, e.g., we need to compute $\operatorname{sn} u = 1/t_0$ in 3.3(1), knowing that $T = \lim_{n\to\infty} t_n = M/\sin(Mu)$ by virtue of 3.3(5). We accomplish this by using the Gauss arithmetic-geometric mean process to compute M, hence T, and then reversing the recursion for t_n in 3.3(1) to compute t_0. Thus (Salzer [1962], Hofsommer and van de Riet [1963], Carlson [1965]),

$$
(2)\qquad \begin{cases} a_0 = 1, \quad b_0 = k' \ , \\ \left.\begin{aligned} a_{n+1} &= \tfrac{1}{2}(a_n + b_n) \\ b_{n+1} &= \sqrt{a_n b_n} \end{aligned}\right\} \quad n = 0, 1, \ldots, \nu-1 \ , \\ t_\nu^{(\nu)} = a_\nu / \sin(a_\nu u) \ , \\ t_{n-1}^{(\nu)} = t_n^{(\nu)} + \dfrac{a_{n-1}^2 - b_{n-1}^2}{4 t_n^{(\nu)}}, \qquad n = \nu, \nu-1, \ldots, 1 \ , \\ \mathrm{sn}\, u = \dfrac{1}{t_0^{(\nu)}}, \quad \mathrm{cn}\, u = \sqrt{[t_0^{(\nu)}]^2 - 1}\ \mathrm{sn}\, u, \quad \mathrm{dn}\, u = (2t_1^{(\nu)} - t_0^{(\nu)})\, \mathrm{sn}\, u \ , \end{cases}
$$

where ν is chosen large enough for $a_\nu - M$ to be negligible. (By 3.2(4), this will be the case if ϵ_ν is negligible compared to $1/2$). A simpler form of the t-recursion results from using 3.3.2(11) and (12),

$$
(2')\qquad t_{n-1}^{(\nu)} = t_n^{(\nu)} + \frac{a_n d_n}{t_n^{(\nu)}}, \qquad n = \nu, \nu-1, \ldots, 1 \ .
$$

From the ascending Landen transformation we obtain (Southard [1963], Hofsommer and van de Riet [1963], Carlson [1965]), similarly,

$$
(3)\qquad \begin{cases} a_0 = 1, \quad b_0 = k \ , \\ \left.\begin{aligned} a_{n+1} &= \tfrac{1}{2}(a_n + b_n) \\ b_{n+1} &= \sqrt{a_n b_n} \end{aligned}\right\} \quad n = 0, 1, \ldots, \nu-1 \ , \\ s_\nu^{(\nu)} = a_\nu / \sinh(a_\nu u) \ , \\ s_{n-1}^{(\nu)} = s_n^{(\nu)} - \dfrac{a_n d_n}{s_n^{(\nu)}} \ , \qquad n = \nu, \nu-1, \ldots, 1 \ , \\ \mathrm{sn}\, u = \dfrac{1}{\sqrt{1 + [s_0^{(\nu)}]^2}}, \quad \mathrm{cn}\, u = s_0^{(\nu)}\, \mathrm{sn}\, u, \quad \mathrm{dn}\, u = (2 s_1^{(\nu)} - s_0^{(\nu)})\, \mathrm{sn}\, u \ . \end{cases}
$$

According to the discussion at the end of 3.2, the ascending algorithm (3) is faster than the descending algorithm (2) when $k^2 > \frac{1}{2}$.

§4. Computer software for special functions

Good numerical methods need to be made easily accessible to the interested user. One way of doing this is by providing computer programs written in one of the higher-level languages such as FORTRAN or ALGOL. For special functions, as well as for many other mathematical problem areas, a great number of such programs are in fact available, and have been so for some time. There are published algorithms in specialized journals (e.g., Comm. ACM, Numer. Math., BIT, Computer Physics Comm., Applied Statistics, and ACM Trans. Mathematical Software), and many others in user's group libraries, commercial libraries, local subroutine libraries, etc. Unfortunately, the quality of these algorithms and programs varies enormously. It has been felt, therefore, in recent years, that libraries should be established by selecting a few algorithms, known for their outstanding quality, implementing them carefully into reliable and thoroughly tested pieces of computer software, integrating the pieces into larger, well-streamlined, and easy-to-use collections of subroutines, and finally releasing these collections to the computing public, with provisions for updating them at regular intervals.

This is not the place to enter into a discussion of the many design objectives and desirable attributes of such packages, nor of explaining the considerable difficulties in trying to attain them; we refer for this to Rice [1971] and Cody [1974]. We would like to draw attention, however, to two current efforts in this direction, one in the United States known as the NATS project (National Activity to Test Software), the other in England, known as the NAG project (Numerical Algorithms Group, formerly Nottingham Algorithms Group). The former's original objective is to produce high-quality software for two restricted problem areas, namely matrix eigensystem problems, and special functions, for which initial packages have been released in 1972 and 1973 under the names EISPACK

and FUNPACK, respectively. The latter's objectives are similar, but embrace a wider problem area – essentially all the major numerical analysis problems. The most recent version ("mark 4") was completed in 1973. For a general description of the NATS project we refer to Boyle et al. [1972] and Smith, Boyle and Cody [1974], and for a discussion of the NAG project to Ford and Hague [1974] and Ford and Sayers [1974]. We briefly compare the two efforts, as far as they concern special functions.

4.1. NATS software for special functions

The special function package of the NATS project – FUNPACK – is developed and maintained under the direction of Cody at Argonne National Laboratory (Cody [1975]). His principal decisions in designing FUNPACK are, first of all, to adopt FORTRAN as the exclusive language of the package, and, secondly, to limit the programs to three different lines of computers, namely the IBM 360-370 series, the CDC6000-7000 series, and Univac 1108. Accordingly, only three accuracy requirements have to be dealt with, roughly 14 significant decimal digits on CDC equipment, and 16, respectively 18, decimal digits for the hardware double-precision arithmetic on IBM and Univac equipment. The package, therefore, is designed to perform well on these particular machines, and is not expected, nor intended, to be easily transportable to other machines.

The limitation to three different precisions has a major influence in the selection of approximation methods. Most attractive, under the circumstances, are rational Chebyshev approximations, both by virtue of their efficiency and uniform accuracy. This, in fact, is the choice made in FUNPACK. The current version I includes subroutines for six special functions – the exponential integral, the complete elliptic integrals of the first and second kind, Dawson's integral, and the Bessel functions K_0 and K_1. All of them are computed from appropriate best rational approximations. Plans are underway to extend the package to include sequences of functions and multivariate functions.

All the programming of the package, as well as the initial testing, was done at Argonne National Laboratory, which has IBM equipment. Similar tests were run on CDC equipment at the University of Texas, and on Univac equipment at the University of Wisconsin. After this initial testing and "tuning" of the routines, they were subjected to additional field tests on the same type of computers, running, however, with different FORTRAN compilers and under a variety of operating systems, some in batch mode, others in time sharing mode. Only after successful completion of all field tests, in September of 1973, was the first version of the package released.

4.2. NAG software for special functions

The special function chapter of the NAG library is being developed by Schonfelder at the University of Birmingham (Schonfelder [1974a, b]). While the basic objectives, and methods of testing, are similar to those of the NATS project, there are some significant differences. For one, all programs in the NAG library are written separately in two languages, FORTRAN and ALGOL. For another, the library is designed to be highly portable, i.e., to run, with a minimal amount of changes, on a wide variety of different machines. Finally, coverage is hoped to eventually include all the major functions in Abramowitz and Stegun [1964] - roughly fifty separate functions. At the moment (Schonfelder [1975]), the list of functions for the forthcoming edition ("mark 5") is to include the exponential , sine , and cosine integral, the gamma function, the error function and Fresnel integrals, and the Bessel functions J_0, J_1, Y_0, Y_1, I_0, I_1, K_0, K_1 . Plans exist to cover functions of several variables, and of complex variables, but implementation appears to be several years in the future (Schonfelder [1975]).

The choices made for the methods of computation reflect the multi-machine character of the NAG library. Preference, in fact, is given to expansions in Chebyshev polynomials, which can be truncated easily to fit various machine precisions, although they may be somewhat inferior in efficiency compared to rational approximations.

4.3. Other software for special functions

Good subroutines for special functions can be found in other mathematical subroutine libraries, e.g., the Boeing library and handbook (Newbery [1971]), containing programs in FORTRAN, and the NUMAL library (Numerical procedures in ALGOL 60) developed at the Mathematical Centre in Amsterdam (den Heijer et al. [1974]). The latter has appeared in seven volumes, volume 6 being devoted to special functions. In addition, there are a number of commercial subroutine packages. IBM offers SSP (Scientific Subroutine Package), currently in its 5th edition, and SLMATH (Subroutine Library Mathematics) and its PL/1 version, PLMATH, while IMSL (International Mathematical Statistical Libraries) regularly issues revised editions of its library.

We listed only those library projects, relevant to special functions, which are most familiar to us, realizing that there are undoubtedly many others.

Acknowledgments. The author is indebted to Profs. F. W. J. Olver and H. Thacher, Jr., who read the entire manuscript and suggested several improvements. He also gratefully acknowledges helpful comments by Prof. P. Wynn on section 1.4 of the manuscript.

REFERENCES

Abramowitz, M., and Stegun, I. A. (1964): Handbook of mathematical functions, Nat. Bur. Standards Appl. Math. Ser. 55.

Achieser, N. I. (1956): Theory of Approximation, Frederick Ungar Publ. Co., New York.

Alexits, G. (1961): Convergence problems of orthogonal series, Pergamon Press, New York-Oxford-Paris.

Allen, G. D., Chui, C. K., Madych, W. R., Narcowich, F. J., and Smith, P. W. (1974): Padé approximation and orthogonal polynomials, Bull. Austral. Math. Soc. 10, 263-270.

Amos, D. E. (1970): Significant digit computation of certain distribution functions, Proc. Sympos. on Empirical Bayes Estimation and Computing in Statistics, pp. 165-180. Texas Tech. Press, Lubbock, Tex.

_____ (1973): Bounds on iterated coerror functions and their ratios, Math. Comp. 27, 413-427.

_____ (1974): Computation of modified Bessel functions and their ratios, Math. Comp. 28, 239-251.

_____ and Bulgren, W. G. (1969): On the computation of a bivariate t-distribution, Math. Comp. 23, 319-333.

_____ and Burgmeier, J. W. (1973): Computation with three-term, linear, nonhomogeneous recursion relations, SIAM Rev. 15, 335-351.

_____ and Daniel, S. L. (1972): CDC 6600 utility routines for Chebyshev approximation and function inversion, Report SC-DR-720917, Sandia Laboratories, Albuquerque, New Mexico.

_____ and _____ (1973): CDC 6600 codes for Bessel functions $I_\nu(x)$, $\mathrm{ber}_\nu(x)$, $\mathrm{bei}_\nu(x)$ and ratios $I_{\nu+1}(x)/I_\nu(x)$, Report SLA-73-0072, Sandia Laboratories, Albuquerque, New Mexico.

_____, _____, and Weston, M, K. (unpubl.): CDC 6600 subroutines IBESS and JBESS for Bessel functions $I_\nu(x)$ and $J_\nu(x)$, $x \geq 0$, $\nu \geq 0$.

Baburin, O. V., and Lebedev, V. I. (1967): The computation of tables of roots and weights of Hermite and Laguerre polynomials for n = 1(1)101 (Russian), Ž. Vyčisl. Mat. i Mat. Fiz. 7, 1021-1030.

Baker, G. A., Jr. (1965): The theory and application of the Padé approximant method, in "Advances in Theoretical Physics" (K. A. Brueckner, ed.), Vol. 1, pp. 1-58. Academic Press, New York-London.

_____ (1970): The Padé approximant method and some related generalizations, in "The Padé Approximant in Theoretical Physics" (G. A. Baker, Jr. and J. L. Gammel, eds.), pp. 1-39. Academic Press, New York-London.

Baker, G. A., Jr. (1973): Recursive calculation of Padé approximants, in "Padé Approximants and their Applications" (P. R. Graves-Morris, ed.), pp. 83-91. Academic Press, London-New York.

_____ (1975): Essentials of Padé Approximants, Academic Press, New York-London.

_____ and Gammel, J. L., eds. (1970): The Padé Approximant in Theoretical Physics, Academic Press, New York-London.

_____, Rushbrooke, G. S., and Gilbert, H. E. (1964): High-temperature series expansions for the spin-$\frac{1}{2}$ Heisenberg model by the method of irreducible representations of the symmetric group, Phys. Rev. 135, A1272-A1277.

Bardo, R. D., and Ruedenberg, K. (1971): Numerical analysis and evaluation of normalized repeated integrals of the error function and related functions, J. Computational Phys. 8, 167-174.

Barlow, R. H. (1974): Convergent continued fraction approximants to generalised polylogarithms, BIT 14, 112-116.

Bhagwandin, K. (1962): L'approximation uniforme des fonctions d'Airy-Stokes et fonctions de Bessel d'indices fractionnaires, 2^{e} Congr. Assoc. Française Calcul Traitement Information, Paris, 1961, pp. 137-145. Gauthier-Villars, Paris.

Bjalkova, A. I. (1963): Computation of Fourier-Chebyshev coefficients (Russian), Metody Vyčisl. 1, 27-29.

Blair, J. M. (1974): Rational Chebyshev approximations for the modified Bessel functions $I_0(x)$ and $I_1(x)$, Math. Comp. 28, 581-583.

_____ and Edwards, C. A. (1974): Stable rational minimax approximations to the modified Bessel functions $I_0(x)$ and $I_1(x)$, Report AECL - 4928, Atomic Energy of Canada Limited, Chalk River Nuclear Laboratories, Chalk River, Ontario.

Blanch, G. (1964): Numerical evaluation of continued fractions, SIAM Rev. 6, 383-421.

Blum, E. K., and Curtis, P. C., Jr. (1961): Asymptotic behavior of the best polynomial approximation, J. Assoc. Comput. Mach. 8, 645-647.

Boyle, J. M., Cody, W. J., Cowell, W. R., Garbow, B. S., Ikebe, Y., Moler, C. B., and Smith, B. T. (1972): NATS - a collaborative effort to certify and disseminate mathematical software, Proc. ACM Annual Conference, 1972, vol. II, pp. 630-635. Assoc. Comput. Mach., New York.

British Association for the Advancement of Science (1952): Mathematical Tables, vol. X, Bessel functions, Part II, Functions of Positive Integer Order, Cambridge University Press.

Bulirsch, R. (1965a): Numerical calculation of elliptic integrals and elliptic functions, Numer. Math. 7, 78-90.

_____ (1965b): Numerical calculation of elliptic integrals and elliptic functions II, Numer. Math. 7, 353-354.

_____ (1969a): An extension of the Bartky-transformation to incomplete elliptic integrals of the third kind, Numer. Math. 13, 266-284.

_____ (1969b): Numerical calculation of elliptic integrals and elliptic functions III, Numer. Math. 13, 305-315.

_____ and Rutishauser, H. (1968): Interpolation und genäherte Quadratur, in "Mathematische Hilfsmittel des Ingenieurs" (R. Sauer and I. Szabó, eds.), Teil III, pp. 232-319. Springer-Verlag, Berlin-Heidelberg-New York.

_____ and Stoer, J. (1968): Darstellung von Funktionen in Rechenautomaten, in "Mathematische Hilfsmittel des Ingenieurs" (R. Sauer and I. Szabó, eds.), Teil III, pp. 352-446. Springer-Verlag, Berlin-Heidelberg-New York.

Byrd, P. F., and Friedman, M. D. (1971): Handbook of elliptic integrals for engineers and scientists, 2nd ed., Springer-Verlag, New York-Heidelberg-Berlin.

Carlson, B. C. (1965): On computing elliptic integrals and functions, J. Math. and Phys. 44, 36-51.

_____ (1971): Algorithms involving arithmetic and geometric means, Amer. Math. Monthly 78, 496-505.

Chawla, M. M. (1966/67): A note on the estimation of the coefficients in the Chebyshev series expansion of a function having a logarithmic singularity, Comput. J. 9, 413.

Cheney, E. W. (1966): Introduction to Approximation Theory, McGraw-Hill, New York-Toronto-London.

_____ and Southard, T. H. (1963): A survey of methods for rational approximation, with particular reference to a new method based on a formula of Darboux, SIAM Rev. 5, 219-231.

Chipman, D. M. (1972): The numerical computation of two transcendental functions related to the exponential integral, Math. Comp. 26, 241-249.

Chisholm, J. S. R. (1973a): Rational approximants defined from double power series, Math. Comp. 27, 841-848.

_____ (1973b): Mathematical theory of Padé approximants, in "Padé Approximants" (P. R. Graves-Morris, ed.), Lectures delivered at a summer school held at the University of Kent, July 1972, pp. 1-18. The Institute of Physics, London-Bristol.

_____ (1973c): Convergence properties of Padé approximants, in "Padé Approximants and their Applications" (P. R. Graves-Morris, ed.), pp. 11-21. Academic Press, London-New York.

Chui, C.K., Shisha, O. and Smith, P. W. (1974): Padé approximants as limits of best rational approximants, J. Approximation Theory 12, 201-204.

Clenshaw, C. W. (1955): A note on the summation of Chebyshev series, Math. Tables Aids Comput. 9, 118-120.

_____ (1962): Chebyshev series for mathematical functions, National Physical Laboratory Mathematical Tables, Vol. 5. Her Majesty's Stationery Office, London.

_____ (1964): A comparison of "best" polynomial approximations with truncated Chebyshev series expansions, J. Soc. Indust. Appl. Math. Ser. B Numer. Anal. 1, 26-37.

Clenshaw, C. W. (1974): Rational approximations for special functions, in "Software for Numerical Mathematics" (D. J. Evans, ed.), pp. 275-284. Academic Press, London-New York.

_____ and Picken, S. M. (1966): Chebyshev series for Bessel functions of fractional order, National Physical Laboratory Mathematical Tables, Vol. 8. Her Majesty's Stationery Office, London.

_____ and Lord, K. (1974): Rational approximations from Chebyshev series, in "Studies in Numerical Analysis" (B. K. P. Scaife, ed.), pp. 95-113. Academic Press, London-New York.

_____, Miller, G. F., and Woodger, M. (1962/63): Algorithms for special functions I, Numer. Math. 4, 403-419.

Cody, W. J. (1965): Chebyshev approximations for the complete elliptic integrals K and E, Math. Comp. 19, 105-112. {Corrigenda: ibid. 20 (1966), 207}.

_____ (1967): Another aspect of economical polynomials, Letters to the Editor, Comm. ACM 10, 531.

_____ (1968): Chebyshev approximations for the Fresnel integrals, Math. Comp. 22, 450-453.

_____ (1969): Rational Chebyshev approximations for the error function, Math. Comp. 23, 631-637.

_____ (1970): A survey of practical rational and polynomial approximation of functions, SIAM Rev. 12, 400-423. {Reprinted in: SIAM Studies in Appl. Math. 6 (1970), 86-109}.

_____ (1974): The construction of numerical subroutine libraries, SIAM Rev. 16, 36-46.

_____ (1975): The FUNPACK package of special function subroutines, ACM Trans. Mathematical Software 1, 13-25.

_____ and Hillstrom, K. E. (1967): Chebyshev approximations for the natural logarithm of the gamma function, Math. Comp. 21, 198-203.

_____ and _____ (1970): Chebyshev approximations for the Coulomb phase shift, Math. Comp. 24, 671-677. {Corrigendum: ibid. 26 (1972), 1031}.

Cody, W.J., and Stoer, J. (1966/67): Rational Chebyshev approximation using interpolation, Numer. Math. 9, 177-188.

_____ and Thacher, H.C., Jr. (1967): Rational Chebyshev approximations for Fermi-Dirac integrals of orders - 1/2, 1/2 and 3/2, Math. Comp. 21, 30-40.

_____ and _____ (1968): Rational Chebyshev approximations for the exponential integral $E_1(x)$, Math. Comp. 22, 641-649.

_____ and _____ (1969): Chebyshev approximations for the exponential integral Ei(x), Math. Comp. 23, 289-303.

_____, Fraser, W., and Hart, J.F. (1968): Rational Chebyshev approximation using linear equations, Numer. Math. 12, 242-251.

_____, Hillstrom, K.E., and Thacher, H.C., Jr. (1971): Chebyshev approximations for the Riemann zeta function, Math. Comp. 25, 537-547.

_____, Paciorek, K.A., and Thacher, H.C., Jr. (1970): Chebyshev approximations for Dawson's integral, Math. Comp. 24, 171-178.

_____, Strecok, A.J., and Thacher, H.C., Jr. (1973): Chebyshev approximations for the psi function, Math. Comp. 27, 123-127.

Cohen, E.A., Jr. (1971): Note on a truncated Chebyshev series modified to match function values at interval endpoints, SIAM J. Numer. Anal. 8, 754-756.

Collatz, L. (1965): Einschliessungssatz für die Minimalabweichung bei der Segmentapproximation, Simpos. Internaz. Appl. Analisi Fis. Mat., Cagliari - Sassari 1964, pp. 11-21. Edizioni Cremonese, Roma.

_____ (1968): Zur numerischen Behandlung der rationalen Tschebyscheff-Approximation bei mehreren unabhängigen Veränderlichen, Apl. Mat. 13, 137-146.

Common, A.K. (1969): Properties of Legendre expansions related to series of Stieltjes and applications to $\pi - \pi$ scattering, Nuovo Cimento 63A, 863-891.

Feinerman, R. P., and Newman, D. J. (1974): Polynomial approximation, The Williams and Wilkins Co., Baltimore.

Fejér, L. (1910): Lebesguesche Konstanten und divergente Fourierreihen, J. Reine Angew. Math. 138, 22-53.

Fettis, H. E. (1965): Calculation of elliptic integrals of the third kind by means of Gauss' transformation, Math. Comp. 19, 97-104.

_____ (1967): Calculation of toroidal harmonics without recourse to elliptic integrals, in: "Blanch Anniversary Volume" (B. Mond, ed.), pp. 21-34. Aerospace Research Lab., U. S. Air Force, Washington, D. C

_____ and Caslin, J. C. (1969): A table of the complete elliptic integral of the first kind for complex values of the modulus I, II, Reports ARL 69-0172 and 69-0173, Wright-Patterson Air Force Base, Ohio.

Fike, C. T. (1967): Methods of evaluating polynomial approximations in function evaluation routines, Comm. ACM 10, 175-178.

_____ (1968): Computer Evaluation of Mathematical Functions, Prentice-Hall, Englewood Cliffs, N. J.

Fleischer, J. (1972): Analytic continuation of scattering amplitudes and Padé approximants, Nucl. Phys. B37, 59-76. {Erratum: ibid. B44 (1972), 641}.

_____ (1973a): Nonlinear Padé approximants for Legendre series, J. Mathematical Phys. 14, 246-248.

_____ (1973b): Generalizations of Padé approximants, in: "Padé Approximants" (P. R. Graves-Morris, ed.), Lectures delivered at a summer school held at the University of Kent, July 1972, pp. 126-131. The Institute of Physics, London-Bristol.

Fletcher, R., Grant, J. A., and Hebden, M. D. (1974): Linear minimax approximation as the limit of best L_p-approximation, SIAM J. Numer. Anal. 11, 123-136.

Ford, B., and Hague, S. J. (1974): The organisation of numerical algorithms libraries, in: "Software for Numerical Mathematics" (D. J. Evans, Ed.), pp. 357-372. Academic Press, London-New York.

Common, A. K., and Graves-Morris, P. R. (1974): Some properties of Chisholm approximants, J. Inst. Math. Appl. 13, 229-232.

Cooper, G. J. (1967): The evaluation of the coefficients in a Chebyshev expansion, Comput. J. 10, 94-100.

Cox, M. G. (unpubl.): Numerical computations associated with Chebyshev polynomials.

Cylkowski, Z. (1966/68): Chebyshev series expansions of the functions $J_\nu(kx)/(kx)^\nu$ and $I_\nu(kx)/(kx)^\nu$, Zastos. Mat. 9, 413-415.

_____ (1971): Remarks on the evaluation of the Bessel functions from the recurrent formula, Zastos. Mat. 12, 217-220.

Davis, P. J. (1963): Interpolation and Approximation, Blaisdell Publ. Co., New York-Toronto-London.

Deuflhard, P. (1974): On algorithms for the summation of certain special functions, Bericht Nr. 7407, Techn. Univ. München, Abteilung Mathematik.

De Vogelaere, R. (1959): Remarks on the paper "Tchebysheff approximations for power series", J. Assoc. Comput. Mach. 6, 111-114.

Elliott, D. (1964): The evaluation and estimation of the coefficients in the Chebyshev series expansion of a function, Math. Comp. 18, 274-284.

_____ (1968): Error analysis of an algorithm for summing certain finite series, J. Austral. Math. Soc. 8, 213-221.

_____ and Szekeres, G. (1965): Some estimates of the coefficients in the Chebyshev series expansion of a function, Math. Comp. 19, 25-32.

_____ and Lam, B. (1973): An estimate of $E_n(f)$ for large n, SIAM J. Numer. Anal. 10, 1091-1102.

Fair, W. (1964): Padé approximation to the solution of the Ricatti equation, Math. Comp. 18, 627-634.

_____ and Luke, Y. L. (1967): Rational approximations to the incomplete elliptic integrals of the first and second kinds, Math. Comp. 21, 418-422.

Ford, B., and Sayers, D. (1974): Developing a single numerical algorithms library for different machine ranges, in: "Mathematical Software II", pp. 234-237, Informal Proceedings of a Conference, Purdue Univ., May 29-31, 1974.

Fox, L. (1965): The proper use of recurrence relations, Math. Gaz. 49, 371-387.

_____, and Parker, I. B. (1968): Chebyshev Polynomials in Numerical Analysis, Oxford University Press, London-New York-Toronto.

Frankel, A. P., and Gragg, W. B. (1973): Algorithmic almost best uniform rational approximation with error bounds (abstract), SIAM Rev. 15, 418-419.

Fraser, W. (1965): A survey of methods of computing minimax and near-minimax polynomial approximations for functions of a single independent variable, J. Assoc. Comput. Mach. 12, 295-314.

Gargantini, I. (1966): On the application of the process of equalization of maxima to obtain rational approximation to certain modified Bessel functions, Comm. ACM 9, 859-863.

_____ and Henrici, P. (1967): A continued fraction algorithm for the computation of higher transcendental functions in the complex plane, Math. Comp. 21, 18-29.

_____ and Pomentale, T. (1964): Rational Chebyshev approximations to the Bessel function integrals $Ki_s(x)$, Comm. ACM 7, 727-730.

Gautschi, W. (1961): Recursive computation of certain integrals, J. Assoc. Comput. Mach. 8, 21-40.

_____ (1966): Computation of successive derivatives of $f(z)/z$, Math. Comp. 20, 209-214.

_____ (1967): Computational aspects of three-term recurrence relations, SIAM Rev. 9, 24-82.

_____ (1969): An application of minimal solutions of three-term recurrences to Coulomb wave functions, Aequationes Math. 2, 171-176.

Gautschi, W. (1970): Efficient computation of the complex error function, SIAM J. Numer. Anal. 7, 187-198.

_____ (1972): Zur Numerik rekurrenter Relationen, Computing 9, 107-126. {English translation in: Report ARL 73-0005, Aerospace Research Laboratories, Wright-Patterson Air Force Base, Ohio, 1973}.

_____ (1973): Algorithm 471-Exponential integrals, Comm. ACM 16, 761-763.

_____ and Klein, B.J. (1970): Recursive computation of certain derivatives - A study of error propagation, Comm. ACM 13, 7-9.

Gentleman, W. M. (1969/70): An error analysis of Goertzel's (Watt's) method for computing Fourier coefficients, Comput. J. 12, 160-165.

Goertzel, G. (1958): An algorithm for the evaluation of finite trigonometric series, Amer. Math. Monthly 65, 34-35.

_____ (1960): Fourier analysis, in "Mathematical Methods for Digital Computers" (A. Ralston and H.S. Wilf, eds.), pp. 258-262. Wiley, New York-London.

Golden, J.E., McGuire, J.H., and Nuttall, J. (1973): Calculating Bessel functions with Padé approximants, J. Math. Anal. Appl. 43, 754-767.

Gragg, W.B. (1972): The Padé table and its relation to certain algorithms of numerical analysis, SIAM Rev. 14, 1-62.

_____ and Johnson, G.D. (1974): The Laurent-Padé table, Proc. IFIP Congress 74, pp. 632-637. North-Holland Publ. Co.

Graves-Morris, P.R., ed. (1973a): Proceedings of the Canterbury Summer School on Padé Approximants and their Applications, Institute of Physics.

_____ (1973b): Proceedings of the 1972 Canterbury International Conference on Padé Approximants and their Applications. Academic Press, London-New York.

_____, Hughes Jones, R., and Makinson, G.J. (1974): The calculation of some rational approximants in two variables, J. Inst. Math. Appl. 13, 311-320.

Guerra, S. (1969): Sul calcolo approssimato di particolari funzioni ipergeometriche confluenti con la tecnica del "τ-method", Calcolo 6, 213-223.

Handscomb, D.C. (1973): The relative sizes of the terms in Chebyshev and other ultraspherical expansions, J. Inst. Math. Appl. 11, 241-246.

Hangelbroek, R.J. (1967): Numerical approximation of Fresnel integrals by means of Chebyshev polynomials, J. Engrg. Math. 1, 37-50.

Har-El, J., and Kaniel, S. (1973): Linear programming method for rational approximation, Israel J. Math. 16, 343-349.

Harris, R.M. (1973): Uniform approximation of functions through partitioning, J. Approximation Theory 7, 239-255.

Hart, J.F., Cheney, E.W., Lawson, C.L., Maehly, H.J., Mesztenyi, C.K., Rice, J.R., Thacher, H.C., Jr., and Witzgall, C. (1968): Computer Approximations, Wiley, New York-London-Sydney.

Håvie, T. (1968): CHECOF - double precision calculation of the coefficients in a Chebyshev expansion, CERN 6600 Series Program Library C320.

Hawkins, D.M. (1972): On the choice of segments in piecewise approximation, J. Inst. Math. Appl. 9, 250-256.

den Heijer, C., Hemker, P.W., van der Houwen, P.J., Temme, N.M., and Winter, D.T., eds. (1974): NUMAL - A library of numerical procedures in ALGOL 60, vols. 0-7. Mathematisch Centrum, Amsterdam.

Henrici, P. (1958): The quotient-difference algorithm, Nat. Bur. Standards Appl. Math. Ser. 49, 23-46.

_____ (1963): Some applications of the quotient-difference algorithm, in: "High Speed Computing and Experimental Arithmetic", pp. 159-183. Proc. Sympos. Appl. Math. 15, Amer. Math. Soc., Providence, R. I.

Henrici, P. (1965): Error bounds for computations with continued fractions, in: "Error in Digital Computation", Vol. 2, pp. 39-53. Proc. Sympos. Math. Res. Center, U.S. Army, Univ. Wisconsin, Madison, Wis., 1965. Wiley, New York.

_____ (1966): An algorithm for analytic continuation, SIAM J. Numer. Anal. 3, 67-78.

_____ (1967): Quotient-difference algorithms, in: "Mathematical Methods for Digital Computers", Vol. II (A. Ralston and H.S. Wilf, eds.), pp. 37-62. Wiley, New York-London-Sydney.

_____ and Pfluger, P. (1966): Truncation error estimates for Stieltjes fractions, Numer. Math. 9, 120-138.

Hewers, W., and Zeller, K. (1960/61): Tschebyscheff-Approximation und Tschebyscheff-Entwicklung, Ann. Univ. Sci. Budapest. Eötvös. Sect. Math. 3-4, 91-93.

Hitotumatu, S. (1967/68): On the numerical computation of Bessel functions through continued fraction, Comment. Math. Univ. St. Paul. 16, 89-113.

Hofsommer, D.J., and van de Riet, R.P. (1963): On the numerical calculation of elliptic integrals of the first and second kind and the elliptic functions of Jacobi, Numer. Math. 5, 291-302.

Holdeman, J.T., Jr. (1969): A method for the approximation of functions defined by formal series expansions in orthogonal polynomials, Math. Comp. 23, 275-287.

Hornecker, G. (1958): Evaluation approchée de la meilleure approximation polynomiale d'ordre n de f(x) sur un segment fini [a, b], Chiffres 1, 157-169.

_____ (1959a): Approximations rationnelles voisines de la meilleure approximation au sens de Tchebycheff, C.R. Acad. Sci. Paris 249, 939-941.

_____ (1959b): Détermination des meilleures approximations rationnelles (au sens de Tchebichef) de fonctions réelles d'une variable sur un segment fini et des bornes d'erreur correspondantes, C.R. Acad. Sci. Paris 249, 2265-2267.

Hornecker, G. (1960): Méthodes pratiques pour la détermination approchée de la meilleure approximation polynômiale ou rationnelle, Chiffres 3, 193-228.

Horner, W. G. (1819): A new method of solving numerical equations of all orders by continuous approximation, Philos. Trans. Roy. Soc. London, part I, 308-335.

Huddleston, R. E. (1972): REHRAT - A program for best min-max rational approximation, Report SCL-DR-720370, Sandia Laboratories, Livermore, California.

Hughes Jones, R., and Makinson, G. J. (1974): The generation of Chisholm rational polynomial approximants to power series in two variables, J. Inst. Math. Appl. 13, 299-310.

Hummer, D. G. (1964): Expansion of Dawson's function in a series of Chebyshev polynomials, Math. Comp. 18, 317-319.

Jacobs, D., and Lambert, F. (1972): On the numerical calculation of polylogarithms, BIT 12, 581-585.

Johnson, J. H., and Blair, J. M. (1973): REMES2: A Fortran program to calculate rational minimax approximations to a given function, Report AECL-4210, Atomic Energy of Canada Limited, Chalk River Nuclear Laboratories, Chalk River, Ontario.

Jones, W. B. (1974): Analysis of truncation error of approximations based on the Padé table and continued fractions, Rocky Mountain J. Math. 4, 241-250.

_____ and Thron, W. J. (1974a): Numerical stability in evaluating continued fractions, Math. Comp. 28, 795-810.

_____ and _____, eds. (1974b): Proceedings of the international conference on Padé approximants, continued fractions and related topics, Rocky Mountain J. Math. 4, 135-397.

_____ and _____ (1974c): Rounding error in evaluating continued fraction expansions, Proceedings ACM Annual Conference, November 1974, San Diego, pp. 11-18. Association for Computing Machinery, New York.

Jones, W. B. and Thron, W.J. (1975): On convergence of Padé approximants, SIAM J. Math. Anal. 6, 9-16.

Kami, Y., Kiyoto, S., and Arakawa, T. (1971a): Method for numerical calculation of the standard elliptic integrals of the first and second kind (Japanese), Rep. Univ. Electro-Commun. 22, 99-108.

_____, _____, and _____ (1971b): Programming method on accurate values of the elliptic integral of the third kind (Japanese), Rep. Univ. Electro-Commun. 22, 109-118.

Kaufman, E. H., Jr., and Taylor, G. D. (1974): An application of linear programming to rational approximation, Rocky Mountain J. Math. 4, 371-373.

King, L. V. (1924): On the Direct Numerical Calculation of Elliptic Functions and Integrals, Cambridge Univ. Press, London.

Khovanskii, A. N. (1963): The Application of Continued Fractions and their Generalizations to Problems in Approximation Theory, translated from Russian by Peter Wynn. P. Noordhoff N. V., Groningen.

Kogbetliantz, E. G. (1960): Generation of elementary functions, in: "Mathematical Methods for Digital Computers" (A. Ralston and H.S. Wilf, eds.), pp. 7-35. Wiley, New York-London.

Kohútová, E. (1970): Stabilitätsbedingungen von rekurrenten Relationen und deren Anwendung, Apl. Mat. 15, 207-212.

Kölbig, K. S. (1972): Remarks on the computation of Coulomb wavefunctions, Computer Physics Comm. 4, 214-220.

_____, Mignaco, J. A., and Remiddi, E. (1970): On Nielsen's generalized polylogarithms and their numerical calculation, BIT 10, 38-73.

Korneĭčuk, A. A., and Širikova, N. Ju. (1968): An iterational method of determining the polynomial of best approximation (Russian), Ž. Vyčisl. Mat. i Mat. Fiz. 8, 670-674.

Krabs, W. (1969): Gleichmässige Approximation von Funktionen, B. I - Hochschultaschenbücher 247/247a, Überblicke Math. 3, 39-69.

Lam, B., and Elliott, D. (1972): On a conjecture of C. W. Clenshaw, SIAM J. Numer. Anal. 9, 44-52.

Lanczos, C. (1956): Applied Analysis, Prentice Hall, Englewood Cliffs, N. J.

_____ (1964): A precision approximation to the gamma function, J. Soc. Indust. Appl. Math. Ser. B Numer. Anal. 1, 86-96.

de La Vallée Poussin, C. (1919): Leçons sur l'approximation des fonctions d'une variable réelle, Gauthier-Villars, Paris.

Lawson, C. L. (1964): Characteristic properties of the segmented rational minimax approximation problem, Numer. Math. 6, 293-301.

Lee, C. M., and Roberts, F. D. K. (1973): A comparison of algorithms for rational ℓ_∞ approximation, Math. Comp. 27, 111-121.

Lehmer, D. H. (1971): On the compounding of certain means, J. Math. Anal. Appl. 36, 183-200.

Levin, D. (1973): Development of non-linear transformations for improving convergence of sequences, Intern. J. Comput. Math. 3, 371-388.

Longman, I. M. (1971): Computation of the Padé table, Intern. J. Comput. Math. 3, 53-64.

_____ (1973): On the generation of rational function approximations for Laplace transform inversion with an application to viscoelasticity, SIAM J. Appl. Math. 24, 429-440.

Luke, Y. L. (1955): Remarks on the τ-method for the solution of linear differential equations with rational coefficients, J. Soc. Indust. Appl. Math. 3, 179-191.

_____ (1958): The Padé table and the τ-method, J. Math. and Phys. 37, 110-127.

_____ (1959/60): On economic representations of transcendental functions, J. Math. and Phys. 38, 279-294.

_____ (1968): Approximations for elliptic integrals, Math. Comp. 22, 627-634.

Luke, Y. L. (1969): The Special Functions and their Approximations, Vols. I, II, Academic Press, New York-London.

_____ (1970a): Further approximations for elliptic integrals, Math. Comp. 24, 191-198.

_____ (1970b): Evaluation of the gamma function by means of Padé approximations, SIAM J. Math. Anal. 1, 266-281.

_____ (1971a): Rational approximations for the logarithmic derivative of the gamma function, Applicable Anal. 1, 65-73.

_____ (1971b): Miniaturized tables of Bessel functions, Math. Comp. 25, 323-330.

_____ (1971c): Miniaturized tables of Bessel functions II, Math. Comp. 25, 789-795.

_____ (1972a): Miniaturized tables of Bessel functions III, Math. Comp. 26, 237-240. {Corrigendum: ibid 26 (1972), no.120, loose microfiche supplement A1-A7}.

_____ (1972b): On generating Bessel functions by use of the backward recurrence formula, Report ARL 72-0030, Aerospace Research Laboratories, Wright-Patterson Air Force Base, Ohio.

_____ (1975): On the error in the Padé approximants for a form of the incomplete gamma function including the exponential function, to appear in: SIAM J. Math. Anal.

_____ and Wimp, J. (1963): Jacobi polynomial expansions of a generalized hypergeometric function over a semi-infinite ray, Math. Comp. 17, 395-404.

Lyness, J. N., and Sande, G. (1971): Algorithm 413-ENTCAF and ENTCRE: Evaluation of normalized Taylor coefficients of an analytic function, Comm. ACM 14, 669-675.

Lyusternik, L. A., Chervonenkis, O. A., and Yanpol'skii, A. R. (1965): Handbook for Computing Elementary Functions, Pergamon Press, New York.

Macon, N., and Baskervill, M. (1956): On the generation of errors in the digital evaluation of continued fractions, J. Assoc. Comput. Mach. 3, 199-202.

Maehly, H.J. (1956): Monthly Progress Report, Oct. 1956, Institute for Advanced Study.

_____ (1958): First Interim Progress Report on Rational Approximations, June 23, 1958, Project NR 044-196, Princeton University.

_____ (1960): Methods for fitting rational approximations I. Telescoping procedures for continued fractions, J. Assoc. Comput. Mach. 7, 150-162.

Mechel, Fr. (1968): Improvement in recurrence techniques for the computation of Bessel functions of integral order, Math. Comp. 22, 202-205.

Meinardus, G. (1964): Approximation von Funktionen und ihre numerische Behandlung, Springer, Berlin-New York. {Expanded English translation in: Springer Tracts in Natural Philosophy, Vol. 13, 1967}.

_____ (1966): Zur Segmentapproximation mit Polynomen, Z. Angew. Math. Mech. 46, 239-246.

Mesztenyi, C., and Witzgall, C. (1967): Stable evaluation of polynomials, J. Res. Nat. Bur. Standards 71B, 11-17.

Miller, G.F. (1966): On the convergence of the Chebyshev series for functions possessing a singularity in the range of representation, SIAM J. Numer. Anal. 3, 390-409.

Minnick, R.C. (1957): Tshebysheff approximations for power series, J. Assoc. Comput. Mach. 4, 487-504.

Moody, W.T. (1967): Approximations for the psi (digamma) function, Math. Comp. 21, 112.

Morita, T., and Horiguchi, T. (1972/73): Convergence of the arithmetic-geometric mean procedure for the complex variables and the calculation of the complete elliptic integrals with complex modulus, Numer. Math. 20, 425-430.

McCabe, J.H. (1974): A continued fraction expansion, with a truncation error estimate, for Dawson's integral, Math. Comp. 28, 811-816.

Natanson, I.P. (1964): Constructive Function Theory, vol. I, Uniform Approximation. Frederick Ungar Publ. Co., New York.

National Bureau of Standards (1952): Tables of Chebyshev polynomials $S_n(x)$ and $C_n(x)$, Appl. Math. Ser. 9 .

Nellis, W.J., and Carlson, B.C. (1966): Reduction and evaluation of elliptic integrals, Math. Comp. 20, 223-231.

Németh, G. (1965): Chebyshev expansions for Fresnel integrals, Numer. Math. 7, 310-312.

_____ (1967): Chebyshev expansion of the Stirling series (Hungarian), Mat. Lapok 18, 329-333.

_____ (1971): Chebyshev polynomial expansions of Airy functions, their zeros, derivatives, first and second integrals (Hungarian), Magyar Tud. Akad. Mat. Fiz. Oszt. Közl. 20, 13-33.

_____ (1972): Tables of the expansions of the first 10 zeros of Bessel functions (Russian), Comm. Joint Inst. Nuclear Res., Dubna, Report 5-6336, March 1972.

_____ (1974): Expansion of generalized hypergeometric functions in Chebyshev polynomials, Collection of Scientific Papers in Collaboration of Joint Institute for Nuclear Research, Dubna, USSR and Central Research Institute for Physics. Algorithms and Programs for Solution of Some Problems in Physics, pp. 57-91. Central Research Institute, Budapest.

Neuman, E. (1969/70a): On the calculation of elliptic integrals of the second and third kinds, Zastos. Mat. 11, 91-94.

_____ (1969/70b): Elliptic integrals of the second and third kinds, Zastos. Mat. 11, 99-102.

Neville, E.H. (1944): Jacobian Elliptic Functions, Oxford University Press, London.

_____ (1971): Elliptic Functions: a Primer, prepared for publication by W.J. Langford. Pergamon Press, Oxford-New York-Toronto-Sydney-Braunschweig.

Newbery, A.C.R. (1971): The Boeing library and handbook of mathematical routines, in "Mathematical Software" (J.R. Rice, ed.), pp. 153-169. Academic Press, New York-London.

Newbery, A. C. R. (1973): Error analysis for Fourier series evaluation, Math. Comp. 27, 639-644.

_____ (1974): Error analysis for polynomial evaluation, Math. Comp. 28, 789-793.

Newman, D. J. (1964): Rational approximation to $|x|$, Michigan Math. J. 11, 11-14.

Ng, E. W. (1968/69): On the direct summation of series involving higher transcendental functions, J. Computational Phys. 3, 334-338.

_____ (1975): A comparison of computational methods and algorithms for the complex gamma function, ACM Trans. Mathematical Software 1, 56-70.

_____, Devine, C. J., and Tooper, R. F. (1969): Chebyshev polynomial expansion of Bose-Einstein functions of orders 1 to 10, Math. Comp. 23, 639-643.

Oliver, J. (1966/67): Relative error propagation in the recursive solution of linear recurrence relations, Numer. Math. 9, 323-340.

_____ (1968a): The numerical solution of linear recurrence relations, Numer. Math. 11, 349-360.

_____ (1968b): An extension of Olver's error estimation technique for linear recurrence relations, Numer. Math. 12, 459-467.

Olver, F. W. J. (1967a): Numerical solution of second-order linear difference equations, J. Res. Nat. Bur. Standards 71B, 111-129.

_____ (1967b): Bounds for the solutions of second-order linear difference equations, J. Res. Nat. Bur. Standards 71B, 161-166.

_____ (1974): Asymptotics and Special Functions, Academic Press, New York-London.

_____ and Sookne, D. J. (1972): Note on backward recurrence algorithms, Math. Comp. 26, 941-947.

Ostrowski, A. (1954): On two problems in abstract algebra connected with Horner's rule, in: "Studies in Mathematics and Mechanics Presented to Richard von Mises", pp. 40-48. Academic Press, New York.

Perron, O. (1957): Die Lehre von den Kettenbrüchen, Vol. II, 3rd ed., Teubner Verlagsges., Stuttgart.

Piessens, R., and Branders, M. (1973): The evaluation and application of some modified moments, BIT 13, 443-450.

_____ and Criegers, R. (1974): Estimation asymptotique des coefficients du développement en série de polynômes de Chebyshev d'une fonction ayant certaines singularités, C.R. Acad. Sci. Paris A278, 405-407.

Powell, M.J.D. (1967): On the maximum errors of polynomial approximations defined by interpolation and by least squares criteria, Comput. J. 9, 404-407.

Ralston, A. (1963): On economization of rational functions, J. Assoc. Comput. Mach. 10, 278-282.

_____ (1967): Rational Chebyshev approximation, in "Mathematical Methods for Digital Computers", Vol. II (A. Ralston and H.S. Wilf, eds.), pp. 264-284. Wiley, New York-London-Sydney.

_____ (1973): Some aspects of degeneracy in rational approximations, J. Inst. Math. Appl. 11, 157-170.

Reimer, M. (1968): Bounds for the Horner sums, SIAM J. Numer. Anal. 5, 461-469.

_____ (1971): Numerische Stabilität beim Horner-Schema, Z. Angew. Math. Mech. 51, T71-T72.

_____ and Zeller, K. (1967): Abschätzung der Teilsummen reeller Polynome, Math. Z. 99, 101-104.

Remes, E. [Remez, E. Ja.] (1934): Sur le calcul effectif des polynomes d'approximation de Tchebichef, C.R. Acad. Sci. Paris 199, 337-340.

_____ (1969): Fundamentals of Numerical Methods of Chebyshev Approximation (Russian), "Naukova Dumka", Kiev.

_____ and Gavriljuk, V.T. (1963): Some remarks on polynomial Chebyshev approximations of functions compared to the intervals of expansions in Chebyshev polynomials (Russian), Ukrain. Mat. Ž. 15, 46-57.

Rice, J.R. (1964a): On the L_∞ Walsh arrays for $\Gamma(x)$ and Erfc(x), Math. Comp. 18, 617-626.

_____ (1964b): The Approximation of Functions, Vol. I: Linear Theory. Addison-Wesley Publ. Co., Reading, Mass.-London.

_____ (1965): On the conditioning of polynomial and rational forms, Numer. Math. 7, 426-435.

_____ (1969): The Approximation of Functions, Vol. II: Nonlinear and Multivariate Theory. Addison-Wesley Publ. Co., Reading, Mass.-London-Don Mills, Ont.

_____ (1971): The challenge for mathematical software, in: "Mathematical Software" (J.R. Rice, ed.), pp. 27-41. Academic Press, New York-London.

Riess, R.D., and Johnson, L.W. (1972): Estimates for $E_n(x^{n+2m})$, Aequationes Math. 8, 258-262.

Rivlin, T.J. (1969): An Introduction to the Approximation of Functions, Blaisdell Publ. Co., Waltham, Mass.-Toronto-London.

_____ (1974): The Chebyshev Polynomials, Wiley, New York-London-Sydney-Toronto.

_____ and Wilson, M.W. (1969): An optimal property of Chebyshev expansions, J. Approximation Theory 2, 312-317.

Russon, A.E., and Blair, J.M. (1969): Rational function minimax approximations for the Bessel functions $K_0(x)$ and $K_1(x)$, Report AECL-3461, Atomic Energy of Canada Limited, Chalk River, Ontario.

Rutishauser, H. (1954a): Der Quotienten-Differenzen-Algorithmus, Z. Angew. Math. Phys. 5, 233-251.

_____ (1954b): Anwendungen des Quotienten-Differenzen-Algorithmus, Z. Angew. Math. Phys. 5, 496-508.

_____ (1957): Der Quotienten-Differenzen-Algorithmus, Birkhäuser Verlag, Basel.

Sadowski, W.L., and Lozier, D.W. (1972): Use of Olver's algorithm to evaluate certain definite integrals of plasma physics involving Chebyshev polynomials, J. Computational Phys. 10, 607-613.

Saffren, M. M., and Ng, E. W. (1971): Recursive algorithms for the summation of certain series, SIAM J. Math. Anal. 2, 31-36.

Salzer, H. E. (1962): Quick calculation of Jacobian elliptic functions, Comm. ACM 5, 399.

Schonfelder, J. L. (1974a): The NAG library and its special function chapter, in: "International Computing Symposium 1973" (A. Günther et al., eds.), pp. 109-116. North-Holland Publ. Co.

_____ (1974b): Special functions in the NAG library, in: "Software for Numerical Mathematics" (D. J. Evans, ed.), pp. 285-300. Academic Press, London-New York.

_____ (1975): private communication.

Schönhage, A. (1971): Approximationstheorie, Walter de Gruyter, Berlin-New York.

Scraton, R. E. (1970): A method for improving the convergence of Chebyshev series, Comput. J. 13, 202-203.

_____ (1972): A modification of Miller's recurrence algorithm, BIT 12, 242-251.

Shenton, L. R., and Bowman, K. O. (1971): Continued fractions for the psi function and its derivatives, SIAM J. Appl. Math. 20, 547-554.

Sheorey, V. B. (1974): Chebyshev expansions for wave functions, Computer Physics Comm. 7, 1-12.

Sidonskiĭ, O. B. (1967): Computation of Bessel functions from the recurrence relation by the double-sweep method (Russian), Izv. Sibirsk. Otdel. Akad. Nauk SSSR 1967, 3-7.

Širikova, N. Ju. (1970): A formula for the polynomial of best approximation, obtained by means of a computer (Russian), Ž. Vyčisl. Mat. i Mat. Fiz. 10, 181-183.

Smith, F. J. (1965): An algorithm for summing orthogonal polynomial series and their derivatives with applications to curve-fitting and interpolation, Math. Comp. 19, 33-36.

Smith, B. T., Boyle, J. M., and Cody, W. J. (1974): The NATS approach to quality software, in "Software for Numerical Mathematics" (D. J. Evans, ed.), pp. 393-405. Academic Press, London-New York.

Sookne, D. J. (1973a): Bessel functions I and J of complex argument and integer order, J. Res. Nat. Bur. Standards 77B, 111-114.

_____ (1973b): Certification of an algorithm for Bessel functions of real argument, J. Res. Nat. Bur. Standards 77B, 115-124.

_____ (1973c): Bessel functions of real argument and integer order, J. Res. Nat. Bur. Standards 77B, 125-132.

_____ (1973d): Certification of an algorithm for Bessel functions of complex argument, J. Res. Nat. Bur. Standards 77B, 133-136.

Southard, T. H. (1963): On the evaluation of the Jacobian elliptic and related functions, Mathematical Note No. 329, Boeing Scientific Research Laboratory.

Spielberg, K. (1961a): Representation of power series in terms of polynomials, rational approximations and continued fractions, J. Assoc. Comput. Mach. 8, 613-627.

_____ (1961b): Efficient continued fraction approximations to elementary functions, Math. Comp. 15, 409-417.

Stiefel, E. L. (1959): Numerical methods of Tchebycheff approximation, in: "On Numerical Approximation" (R. E. Langer, ed.), pp. 217-232. University of Wisconsin Press, Madison.

_____ (1964): Methods - old and new - for solving the Tchebycheff approximation problem, J. Soc. Indust. Appl. Math. Ser. B Numer. Anal. 1, 164-176.

Stoer, J. (1964): A direct method for Chebyshev approximation by rational functions, J. Assoc. Comput. Mach. 11, 59-69.

_____ (1972): Einführung in die numerische Mathematik I, Springer Verlag, Berlin-Heidelberg-New York.

Stoljarčuk, V. K. (1974a): The construction, for functions si x and $\phi(x) = 2\pi^{-\frac{1}{2}} \int_0^x e^{-t^2} dt$, of polynomials that realize a close-to-best approximation (Russian), Ukrain. Mat. Ž. 26, 216-226.

Stoljarčuk, V. K. (1974b): Uniform approximation, by polynomials on a segment, of Bessel functions with integer index (Russian), Ukrain. Mat. Ž. 26, 683-686.

Strecok, A. J. (1968): On the calculation of the inverse of the error function, Math. Comp. 22, 144-158.

_____ and Gregory, J. A. (1972): High precision evaluation of the irregular Coulomb wave functions, Math. Comp. 26, 955-961.

Szegö, G. (1921): Über die Lebesgueschen Konstanten bei den Fourierschen Reihen, Math. Z. 9, 163-166.

Temme, N. M. (1972): Numerical evaluation of functions arising from transformations of formal series, Report TW 134/72, Mathematisch Centrum Amsterdam.

_____ (1973): On the numerical evaluation of the modified Bessel function of the third kind, Report TN 72/73, Mathematisch Centrum Amsterdam.

Thacher, H. C., Jr. (1960): Rational approximations for the Debye functions, J. Chem. Phys. 32, 638.

_____ (1964): Conversion of a power to a series of Chebyshev polynomials, Comm. ACM 7, 181-182.

_____ (1966): Independent variable transformations in approximation, Proc. IFIP Congress 65, Vol. 2, pp. 576-577. Spartan Books, Washington, D. C.

_____ (1967): Computation of the complex error function by continued fractions, in: "Blanch Anniversary Volume" (B. Mond, ed.), pp. 315-337. Aerospace Research Lab., U. S. Air Force, Washington, D. C.

_____ (1969): Computational methods for mathematical functions, Report 32-1324, Jet Propulsion Laboratory, Pasadena, California.

_____ (1971): Making special arithmetics available, in: "Mathematical Software" (J. R. Rice, ed.), pp. 113-119. Academic Press, New York-London.

Thacher, H. C., Jr. (1972): Series solutions to differential equations by backward recurrence, Proc. IFIP Congress 71, Vol. 2, pp. 1287-1291. North-Holland Publ. Co., Amsterdam-London.

Todd, J. (1954): Evaluation of the exponential integral for large complex arguments, J. Res. Nat. Bur. Standards 52, 313-317.

Torii, T., and Makinouchi, S. (1968): An efficient algorithm for Chebyshev expansion, Information Processing in Japan 8, 89-92.

Tricomi, F. (1948): Elliptische Funktionen, translated and edited by Maximilian Krafft. Akad. Verlagsges., Leipzig.

_____(1951): Funzioni ellittiche, 2nd ed., Nicola Zanichelli Editore, Bologna.

Van de Vel, H. (1969): On the series expansion method for computing incomplete elliptic integrals of the first and second kinds, Math. Comp. 23, 61-69.

Verbeeck, P. (1970): Rational approximations for exponential integrals $E_n(x)$, Acad. Roy. Belg. Bull. Cl. Sci. (5) 56, 1064-1072.

Wall, H. S. (1948): Analytic Theory of Continued Fractions, D. Van Nostrand Co., New York.

Walsh, J. L. (1964a): Padé approximants as limits of rational functions of best approximation, J. Math. Mech. 13, 305-312.

_____(1964b): The convergence of sequences of rational functions of best approximation, Math. Ann. 155, 252-264.

_____(1965): The convergence of sequences of rational functions of best approximation II, Trans. Amer. Math. Soc. 116, 227-237.

_____(1968a): The convergence of sequences of rational functions of best approximation III, Trans. Amer. Math. Soc. 130, 167-183.

_____(1968b): Degree of approximation by rational functions and polynomials, Michigan Math. J. 15, 109-110.

_____(1969): Interpolation and Approximation by Rational Functions in the Complex Domain, Amer. Math. Soc. Colloq. Publ., Vol. 20, 5th ed., Amer. Math. Soc., Providence, R. I.

Walsh, J. L. (1974): Padé approximants as limits of rational functions of best approximation, real domain, J. Approximation Theory 11, 225-230.

Ward, M. (1960): The calculation of the complete elliptic integral of the third kind, Amer. Math. Monthly 67, 205-213.

Watson, G. A. (1975): A multiple exchange algorithm for multivariate Chebyshev approximation, SIAM J. Numer. Anal. 12, 46-52.

Watson, P. J. S. (1973): Algorithms for differentiation and integration, in: "Padé Approximants and their Applications" (P. R. Graves-Morris, ed.), pp. 93-97. Academic Press, London-New York.

Watt, J. M. (1958/59): A note on the evaluation of trigonometric series, Comput. J. 1, 162.

Werner, H. (1958/59): Tschebyscheffsche Approximationen für Bessel-Funktionen, Nukleonik 1, 60-63.

_____ (1966): Vorlesung über Approximationstheorie, Lecture Notes in Mathematics 14, Springer-Verlag, Berlin-New York.

_____ and Collinge, R. (1961): Chebyshev approximations to the gamma function, Math. Comp. 15, 195-197.

_____ and Raymann, G. (1963): An approximation to the Fermi integral $F_{\frac{1}{2}}(x)$, Math. Comp. 17, 193-194.

_____, Stoer, J., and Bommas, W. (1967): Rational Chebyshev approximation, Numer. Math. 10, 289-306.

Wilkinson, J. H. (1963): Rounding Errors in Algebraic Processes, Prentice-Hall, Englewood Cliffs, N. J.

Williams, J. (1972): Numerical Chebyshev approximation in the complex plane, SIAM J. Numer. Anal. 9, 638-649.

Wills, J. G. (1971): On the use of recursion relations in the numerical evaluation of spherical Bessel functions and Coulomb functions, J. Computational Phys. 8, 162-166.

Wimp, J. (1961): Polynomial approximations to integral transforms, Math. Comp. 15, 174-178.

Wimp, J. (1962): Polynomial expansions of Bessel functions and some associated functions, Math. Comp. 16, 446-458.

_____ (1969): On recursive computation, Report ARL 69-0186, Wright-Patterson Air Force Base, Ohio.

_____ (1970): Recent developments in recursive computation, SIAM Studies in Appl. Math. 6, 110-123.

_____ (1971/72): Forward computation in second order difference equations, Applicable Anal. 1, 325-329.

_____ (1974): On the computation of Tricomi's ψ function, Computing 13, 195-203.

_____ and Luke, Y. L. (1969): An algorithm for generating sequences defined by nonhomogeneous difference equations, Rend. Circ. Mat. Palermo (2) 18, 251-275.

Wood, V. E. (1967): Chebyshev expansions for integrals of the error function, Math. Comp. 21, 494-496.

_____ (1968): Efficient calculation of Clausen's integral, Math. Comp. 22, 883-884.

Wynn, P. (1956): On a device for computing the $e_m(S_n)$ transformation, Math. Tables Aids Comput. 10, 91-96.

_____ (1960): The rational approximation of functions which are formally defined by a power series expansion, Math. Comp. 14, 147-186.

_____ (1961): L'ϵ-algoritmo e la tavola di Padé, Rend. Mat. e Appl. (5) 20, 403-408.

_____ (1962a): Numerical efficiency profile functions, Nederl. Akad. Wetensch. Proc. Ser. A 65 = Indag. Math. 24, 118-126.

_____ (1962b): The numerical efficiency of certain continued fraction expansions IA, Nederl. Akad. Wetensch. Proc. Ser. A65 = Indag. Math. 24, 127-137; IB, ibid., 138-148.

_____ (1966a): Upon systems of recursions which obtain among the quotients of the Padé table, Numer. Math. 8, 264-269.

Wynn, P. (1966b): An arsenal of Algol procedures for the evaluation of continued fractions and for effecting the epsilon algorithm, Rev. Française Traitement Information Chiffres 9, 327-362.

_____ (1967): A general system of orthogonal polynomials, Quart. J. Math. Oxford Ser. (2) 18, 81-96.

_____ (1972): Upon a convergence result in the theory of the Padé table, Trans. Amer. Math. Soc. 165, 239-249.

_____ (1974): Some recent developments in the theories of continued fractions and the Padé table, Rocky Mountain J. Math. 4, 297-323.

Zygmund, A. (1959): Trigonometric Series, Vol. I, Cambridge University Press.

9.8. [61] "Anomalous Convergence of a Continued Fraction for Ratios of Kummer Functions"

---

[61] "Anomalous Convergence of a Continued Fraction for Ratios of Kummer Functions," *Math. Comp.* **31**, 994–999 (1977).

MATHEMATICS OF COMPUTATION, VOLUME 31, NUMBER 140
OCTOBER 1977, PAGES 994–999

Anomalous Convergence of a Continued Fraction for Ratios of Kummer Functions*

By Walter Gautschi

Abstract. We exhibit a phenomenon of apparent convergence to the wrong limit in connection with a continued fraction of Perron for ratios of Kummer functions. The phenomenon is further illustrated in the special cases of Bessel functions and incomplete gamma functions.

1. Introduction. From the differential equation satisfied by Kummer's function

$$M(a, b; z) = 1 + \frac{a}{b}\frac{z}{1!} + \frac{a(a+1)}{b(b+1)}\frac{z^2}{2!} + \dots,$$

Perron [4, p. 278] develops the following continued fraction,

$$\frac{zM'(a, b; z)}{M(a, b; z)} = \frac{az}{b - z+}\ \frac{(a+1)z}{b+1-z+}\ \frac{(a+2)z}{b+2-z+}\cdots, \qquad b \neq 0, -1, -2, \dots, \tag{1.1}$$

where $M'(a, b; z) = (d/dz)M(a, b; z) = (a/b)M(a+1, b+1; z)$. While the continued fraction converges for any complex z not a zero of $M(a, b; z)$, the convergence behavior can be extremely deceptive, when $|z| \gg \max(|a|, |b|)$, particularly if $\operatorname{Re} z > 0$. The point is illustrated by concrete examples involving Bessel and incomplete gamma functions.

2. The Phenomenon of Apparent Convergence to the Wrong Limit. We assume, for simplicity, that $a \neq 0$ and $b - z \neq 0, -1, -2, \dots$. Equation (1.1) can then be written in the form

$$\frac{b-z}{a}\frac{M'(a, b; z)}{M(a, b; z)} = \frac{1}{1+}\frac{a_1}{1+}\frac{a_2}{1+}\cdots, \tag{2.1}$$

where

$$a_k = \frac{(a+k)z}{(b-z+k-1)(b-z+k)}, \qquad k = 1, 2, 3, \dots. \tag{2.2}$$

Alternatively (cf., e.g., Wall [6, p. 17ff]),

$$\frac{b-z}{a}\frac{M'(a, b; z)}{M(a, b; z)} = \sum_{k=0}^{\infty} p_k, \tag{2.3}$$

where

$$p_0 = 1, \qquad p_k = \rho_1\rho_2\cdots\rho_k, \qquad k = 1, 2, 3, \dots,$$

Received January 20, 1977.
AMS (MOS) subject classifications (1970). Primary 33A30, 33A40, 40A15, 65D20.
Key words and phrases. Continued fractions, apparent convergence to the wrong limit, Bessel functions, incomplete gamma functions.
*Sponsored by the United States Army under Contract No. DAAG29-75-C-0024 and The National Science Foundation under grant MCS 76-00842.

and

$$\rho_0 = 0, \qquad \rho_k = \frac{-a_k(1 + \rho_{k-1})}{1 + a_k(1 + \rho_{k-1})}, \qquad k = 1, 2, 3, \ldots . \tag{2.4}$$

The infinite series in (2.3) represents the continued fraction in (2.1) in the sense that the nth partial sum of the former is equal to the nth convergent of the latter, $n = 1, 2, 3, \ldots$.

Evidently, the terms p_k in the series of (2.3) decrease or increase in absolute value according as $|\rho_k| < 1$ or $|\rho_k| > 1$, respectively. It is useful, therefore, to examine the behavior of $|\rho_k|$ as a function of k.

Assuming $|\rho_{k-1}| < 1$, then $|\rho_k| < 1$ certainly if $|a_k| \leqslant \frac{1}{4}$. On the other hand, by (2.2), if $|z| > |b| + k$, then $|a_k| < (|a| + k)|z|(|z| - |b| - k)^{-2}$, and an elementary calculation shows the upper bound for $|a_k|$ to be $\leqslant \frac{1}{4}$ if

$$|z| \geqslant 2(|b| + 2|a| + 3k). \tag{2.5}$$

It follows that (2.5), together with $|\rho_{k-1}| < 1$, implies $|\rho_k| < 1$. Since, initially, $\rho_0 = 0$, we obtain by induction that $|\rho_k| < 1$ for all k satisfying (2.5).

If $|z|$ is large, we see that the terms p_k in (2.3) must decrease initially, the rate of decrease being greater the larger $|z|$. The continued fraction in (2.1) then gives the appearance of converging rapidly to a value of the order of magnitude 1, yielding for $M'(a, b; z)/M(a, b; z)$ a value approximately equal to $-a/z$. This is obviously the wrong answer, if Re $z > 0$. Indeed, from known asymptotic formulas [5, Eq. 13.1.4],

$$\frac{M'(a, b; z)}{M(a, b; z)} \sim 1 \quad \text{as } |z| \longrightarrow \infty \text{ in Re } z > 0. \tag{2.6}$$

What is likely to happen, then, is that the terms p_k, after the initial descent, begin to increase again, and converge to zero only after reaching some peak values which are sufficiently large so as to contribute to a limit consistent with (2.6). It is only during the "final descent" of the terms p_k that the correct limit will be attained (assuming no rounding errors).

The phenomenon of apparent convergence, while prevalent for Re $z > 0$ and $|z|$ large, need not occur if Re $z < 0$, since in this case [5, Eq. 13.1.5]

$$\frac{M'(a, b; z)}{M(a, b; z)} \sim -\frac{a}{z} \quad \text{as } |z| \longrightarrow \infty \text{ in Re } z < 0. \tag{2.7}$$

Nevertheless, we will see in examples that the phenomenon persists if $\pi/2 \leqslant |\arg z| < \pi$, albeit in a weakened form.

3. The Case of Real z. It is instructive to examine in more detail the case of real arguments z and real parameters a, b satisfying $0 < a + 1$ if $z > 0$, and $0 < a + 1 \leqslant b$ if $z < 0$.

It will be convenient to introduce the quantities

$$\sigma_k = 1 + \rho_k, \qquad k = 0, 1, 2, \ldots , \tag{3.1}$$

for which the recursion (2.4) gives

$$\sigma_0 = 1, \quad \sigma_k = \frac{1}{1 + a_k \sigma_{k-1}}, \quad k = 1, 2, 3, \ldots . \tag{3.2}$$

3.1. *The Case* $z = x > 0$, $a + 1 > 0$. We consider two subcases, (i) $b - x > 0$, (ii) $b - x < 0$.

In case (i), it follows from (2.2) that $a_k > 0$ for all $k \geqslant 1$, hence from (3.2) that $0 < \sigma_k < 1$ for all $k \geqslant 1$, and therefore from (3.1) that $-1 < \rho_k < 0$. We see that *the terms* p_k *in* (2.3) *alternate in sign and decrease monotonically in modulus.* In fact, since $a_k \longrightarrow 0$, hence $a_k \sigma_{k-1} \longrightarrow 0$ as $k \longrightarrow \infty$, we have $\sigma_k \longrightarrow 1$, and so $\rho_k \longrightarrow 0$, meaning that *the series in* (2.3) *converges faster than any geometric series.* Indeed,

$$\rho_k \sim -\frac{x}{k} \quad \text{as } k \longrightarrow \infty, \tag{3.3}$$

as is easily verified.

In case (ii), there exists a unique integer $k_0 \geqslant 1$ such that $x - b < k_0 < x - b + 1$. Therefore, $a_{k_0} < 0$, while $a_k > 0$ for all $k \neq k_0$. It follows as before that $-1 < \rho_k < 0$ for $k < k_0$. If $\sigma_{k_0} > 0$ (even though $a_{k_0} < 0$), then $-1 < \rho_k < 0$ also for all $k \geqslant k_0$, and we have the same alternating and supergeometric convergence behavior as in case (i). If, however, $\sigma_{k_0} < 0$ (which will be the case if x is large), then $|\rho_{k_0}| > 1$. Since $a_k > 0$ for $k > k_0$, the inequality $|\rho_k| > 1$ will persist as long as σ_k remains negative. Eventually, however, σ_k has to turn positive (the series in (2.3) being convergent), and from this point on, all subsequent σ's remain positive, and the corresponding ρ's less than one in modulus. Therefore, if x is large, *the sequence* $\{|p_k|\}$ *initially decreases, then increases, and finally decreases to zero* at a supergeometric rate given by (3.3). The "dip-and-peak" effect is more pronounced, the larger x, and is what gives rise to the phenomenon of apparent convergence.

3.2. *The Case* $z = -x < 0$, $0 < a + 1 \leqslant b$. This time, $a_k < 0$ for all $k \geqslant 1$. Noting that the function $x(b + x + k - 1)^{-1}(b + x + k)^{-1}$ on $0 \leqslant x < \infty$ assumes a unique maximum at $(b + k - 1)^{1/2}(b + k)^{1/2}$, we find that

$$|a_k| \leqslant \frac{a + k}{(\sqrt{b + k - 1} + \sqrt{b + k})^2}, \quad k \geqslant 1,$$

and thus, in particular,

$$|a_k| \leqslant \frac{a + k}{2b + 2k - 1 + 2\sqrt{b + k - 1}\sqrt{b + k}} < \frac{a + k}{2b + 2k - 1 + 2(b + k - 1)}$$
$$< \frac{1}{4}\frac{a + k}{b + k - 1}.$$

Since $a + 1 \leqslant b$, it follows that

$$a_k < 0, \quad |a_k| < \tfrac{1}{4} \quad \text{for all } k \geqslant 1.$$

From this, and (3.2), we deduce inductively

$$1 < \sigma_k < \frac{2(k + 1)}{k + 2} \quad \text{for all } k \geqslant 1,$$

hence, in particular, $0 < \rho_k < 1$ for all $k \geqslant 1$. *The series in* (2.3) *is now a series of*

positive monotonically decreasing terms, and convergence thus monotone and, as before, supergeometric.

4. Examples.

4.1. *Bessel Functions.* We specialize (1.1) to $a = \nu + \frac{1}{2}$, $b = 2\nu + 1$, where $\nu \geqslant \frac{1}{2}$, and use $M(\nu + \frac{1}{2}, 2\nu + 1; z) = \Gamma(1 + \nu)\exp(\frac{1}{2}z)(\frac{1}{4}z)^{-\nu}I_\nu(\frac{1}{2}z)$, together with the differential-difference relation $I'_\nu(z) = I_{\nu-1}(z) - \nu I_\nu(z)/z$, to obtain

$$\frac{1}{2}z\left\{\frac{I_{\nu-1}(\frac{1}{2}z)}{I_\nu(\frac{1}{2}z)} - \frac{4\nu}{z} + 1\right\}$$

$$= \frac{\left(\nu + \frac{1}{2}\right)z}{2\nu + 1 - z +} \frac{\left(\nu + \frac{3}{2}\right)z}{2\nu + 2 - z +} \frac{\left(\nu + \frac{5}{2}\right)z}{2\nu + 3 - z +} \cdots = \frac{\left(\nu + \frac{1}{2}\right)z}{2\nu + 1 - z} \sum_{k=0}^{\infty} p_k. \tag{4.1}$$

In Figure 4.1, the moduli of the terms, $|p_k|$, are plotted in function of k, for $\nu = 1$, and $z = re^{i\varphi}$, $r = 10, 20, 40$, $\varphi = 0, \pi/8, 2\pi/8, \ldots, \pi$. The behavior of $|p_k|$, when r is fixed and φ varies between 0 and $\pi/2$, is almost identical and is represented by *one* curve in Figure 4.1. The dependence on φ is shown only in the case $r = 40$, but is analogous for the other values of r.

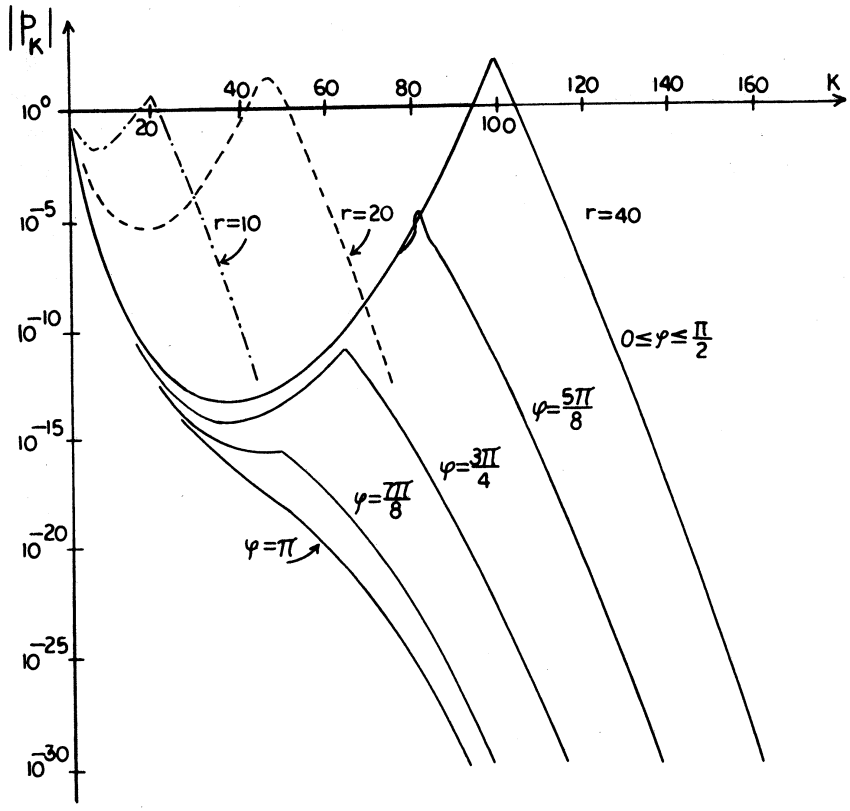


FIGURE 4.1

Anomalous convergence of (4.1) *for* $\nu = 1$ *and* $z = re^{i\varphi}$

It is seen, typically, that the terms $|p_k|$ decrease rapidly at the beginning, then bottom out and rise to a sharp peak, before finally converging to zero. The dip of the curve, and the upswing that follows, are quite substantial if $0 \leqslant \varphi \leqslant \pi/2$ and r is large, for reasons explained in subsection 3.1, case (ii). As φ increases from $\pi/2$ to π, the peaking of the curves gradually weakens and finally disappears when $\varphi = \pi$. A similar attenuation takes place upon increasing the value of ν, as is to be expected from the discussion in subsection 3.1, case (i).

The seriousness of the convergence anomaly can be seen, e.g., in the case $\nu = 1$, $r = 40$, $0 \leqslant \varphi \leqslant \pi/2$. If we require ten decimal digit accuracy, we will attain it at about $k = 15$ and retain it through about $k = 60$, the partial sums in (4.1) all having the same value to ten decimal digits in the range $15 \leqslant k \leqslant 60$. This "apparent limit", of course, is totally incorrect, as the main contribution to the series comes from the few terms around $k = 100$. The situation is aggravated by the fact that the numerical process of generating the terms p_k is accompanied by a substantial loss of accuracy during the upswing of the curve, amounting to a loss of about 16 digits when $r = 40$. A further complication is the apparent lack of warning signals: Known a posteriori error estimates (see, e.g., [3]) either do not apply, or seem to apply only in the region of "final descent".

On the other hand, the convergence behavior of the continued fraction in (4.1) is quite acceptable when $\varphi = \pi$, i.e., $z = -x$, $x > 0$, in which case (4.1) can be given the form

$$\frac{1}{2}x\left\{\frac{I_{\nu-1}(\tfrac{1}{2}x)}{I_\nu(\tfrac{1}{2}x)} - \frac{4\nu}{x} - 1\right\} = \frac{-\left(\nu+\frac{1}{2}\right)x}{2\nu+1+x-}\ \frac{\left(\nu+\frac{3}{2}\right)x}{2\nu+2+x-}\ \frac{\left(\nu+\frac{5}{2}\right)x}{2\nu+3+x-}\cdots. \tag{4.2}$$

Convergence is more rapid the larger x and/or ν. The use of this continued fraction, in combination with Gauss' continued fraction, is further discussed in [2].

4.2. *Incomplete Gamma Function.* We have $M(a, a+1; z) = a(-z)^{-a}\gamma(a, -z)$, where $\gamma(a, \cdot)$ denotes the incomplete gamma function. Noting that

$$M'(a, a+1; z) = \frac{a}{a+1}M(a+1, a+2; z),$$

Eq. (1.1) now takes the form

$$\frac{\gamma(a+1, -z)}{\gamma(a, -z)} = \frac{-az}{a+1-z+}\ \frac{(a+1)z}{a+2-z+}\ \frac{(a+2)z}{a+3-z+}\cdots. \tag{4.3}$$

The convergence behavior of (4.3) for $a > 0$ appears to be quite analogous to that of (4.1), exhibiting the phenomenon of apparent convergence for z large in the complex plane cut along the negative real axis. Along the negative real axis, we have monotone convergence, if $a > -1$, according to the results of subsection 3.2. In this case,

$$\frac{\gamma(a+1, x)}{\gamma(a, x)} = \frac{ax}{a+1+x-}\ \frac{(a+1)x}{a+2+x-}\ \frac{(a+2)x}{a+3+x-}\cdots, \qquad x > 0. \tag{4.4}$$

We may combine this with $\gamma(a+1, x) = a\gamma(a, x) - x^a e^{-x}$ to obtain

$$x^{-a}e^x\gamma(a, x) = \frac{1}{a-}\ \frac{ax}{a+1+x-}\ \frac{(a+1)x}{a+2+x-}\ \frac{(a+2)x}{a+3+x-}\cdots, \qquad x > 0. \tag{4.5}$$

The use of this continued fraction, in combination with other methods, to evaluate incomplete gamma functions is discussed in [1].

Department of Computer Sciences
Purdue University
Lafayette, Indiana 47907

1. W. GAUTSCHI, "An evaluation procedure for incomplete gamma functions," *ACM Trans. Mathematical Software.* (To appear.)

2. W. GAUTSCHI & J. SLAVIK, "On the computation of modified Bessel function ratios," *Math. Comp.* (To appear.)

3. W. B. JONES, "Analysis of truncation error of approximations based on the Padé table and continued fractions," *Rocky Mountain J. Math.*, v. 4, 1974, pp. 241–250. MR 49 #6551.

4. O. PERRON, *Die Lehre von den Kettenbrüchen,* Vol. II, 3rd ed., Teubner, Stuttgart, 1957. MR 19, 25.

5. L. J. SLATER, "Confluent hypergeometric functions," *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (M. Abramowitz & I. A. Stegun, Editors), Nat. Bur. Standards, Appl. Math. Ser., no. 55, Superintendent of Documents, U. S. Government Printing Office, Washington, D. C., 1964, pp. 503–535. MR 29 #4914.

6. H. S. WALL, *Analytic Theory of Continued Fractions,* Van Nostrand, New York, 1948; reprint, Chelsea, New York, 1967. MR 10, 32.

9.9. [68] "A Computational Procedure for Incomplete Gamma Functions"

---

[68] "A Computational Procedure for Incomplete Gamma Functions," *ACM Trans. Math. Software* **5**, 466–481 (1979).

A Computational Procedure for Incomplete Gamma Functions

WALTER GAUTSCHI
Purdue University

We develop a computational procedure, based on Taylor's series and continued fractions, for evaluating Tricomi's incomplete gamma function $\gamma^*(a, x) = (x^{-a}/\Gamma(a))\int_0^x e^{-t}t^{a-1}\,dt$ and the complementary incomplete gamma function $\Gamma(a, x) = \int_x^\infty e^{-t}t^{a-1}\,dt$, suitably normalized, in the region $x \geq 0$, $-\infty < a < \infty$.

Key Words and Phrases: computation of incomplete gamma functions, Taylor's series, continued fractions
CR Categories: 5.12
The Algorithm: Incomplete Gamma Functions. *ACM Trans. Math. Software 5,* 4(Dec. 1979), 482-489.

1. INTRODUCTION

The incomplete gamma function and its complementary function are usually defined by

$$\gamma(a, x) = \int_0^x e^{-t}t^{a-1}\,dt, \qquad \Gamma(a, x) = \int_x^\infty e^{-t}t^{a-1}\,dt. \tag{1.1}$$

By Euler's integral for the gamma function,

$$\gamma(a, x) + \Gamma(a, x) = \Gamma(a). \tag{1.2}$$

We are interested in computing both functions for arbitrary x, a in the half-plane

$$\mathcal{H} = \{(x, a) : x \geq 0, -\infty < a < \infty\}.$$

The function $\Gamma(a, x)$ is meaningful everywhere in $\mathcal{H}$, except along the negative a-axis, where it becomes infinite. The definition of $\gamma(a, x)$ is less satisfactory, inasmuch as it requires $a > 0$. The difficulty, however, is easily resolved by adopting Tricomi's version [14] of the incomplete gamma function,

$$\gamma^*(a, x) = \frac{x^{-a}}{\Gamma(a)}\gamma(a, x), \tag{1.3}$$

This work was supported in part by the U.S. Army under Contract DAAG29-75-C-0024 and in part by the National Science Foundation under Grant MCS 76-00842.
Author's address. Department of Computer Sciences, Purdue University, Mathematical Sciences Building, Room 442, West Lafayette, IN 47907.

which can be continued analytically into the entire (x, a)-plane, resulting in an entire function both in a and x,

$$\gamma^*(a, x) = \frac{e^{-x}M(1, a+1; x)}{\Gamma(a+1)} = \frac{M(a, a+1; -x)}{\Gamma(a+1)}. \tag{1.4}$$

Here,

$$M(a, b; z) = 1 + \frac{a}{b}\frac{z}{1!} + \frac{a(a+1)}{b(b+1)}\frac{z^2}{2!} + \cdots$$

is Kummer's function. Moreover, $\gamma^*(a, x)$ is real-valued for a and x both real, in contrast to $\Gamma(a, x)$, which becomes complex for negative x.

Our objective, then, is to compute the functions $\gamma^*(a, x)$ and $\Gamma(a, x)$, suitably normalized, to any prescribed accuracy for arbitrary x, a in $\mathscr{H}$. We do not attempt here to compute $\gamma^*(a, x)$ for negative x, which may well be a more difficult (but, fortunately, less important) task. We accomplish our task by selecting one of the two functions as *primary function*, to be computed first, and computing the other in terms of the primary function by means of

$$\Gamma(a, x) = \Gamma(a)\{1 - x^a\gamma^*(a, x)\} \tag{1.5}$$

or

$$\gamma^*(a, x) = x^{-a}\left\{1 - \frac{\Gamma(a, x)}{\Gamma(a)}\right\}. \tag{1.6}$$

If $\gamma^*(a, x)$ is the primary function, we evaluate it by Taylor's series. For $\Gamma(a, x)$ we use a combination of methods, including direct evaluation based partly on power series, recursive computation, and the classical continued fraction of Legendre. Although our procedure is valid throughout the region $\mathscr{H}$, excessively large values of a and x will strain it, particularly when $a \doteq x \gg 1$ (cf. Section 5). In such cases it may be preferable to use asymptotic methods, e.g. the uniform asymptotic expansions of Temme [12]. We shall not consider these here, however, nor do we implement them in our algorithm.

An evaluation procedure of the generality attempted here is likely to be of interest in many diverse areas of application. Widely used special cases of $\gamma^*(a, x)$ or $\Gamma(a, x)$ include Pearson's form of the incomplete gamma function [10],

$$I(u, p) = (u\sqrt{p+1})^{p+1}\gamma^*(p+1, u\sqrt{p+1}), \qquad u \geq 0, \quad p > -1, \tag{1.7}$$

the χ^2-probability distribution functions

$$P(\chi^2 \mid \nu) = \left(\frac{1}{2}\chi^2\right)^{\nu/2}\gamma^*\left(\frac{\nu}{2}, \frac{1}{2}\chi^2\right), \qquad Q(\chi^2 \mid \nu) = \frac{1}{\Gamma(\nu/2)}\Gamma\left(\frac{\nu}{2}, \frac{1}{2}\chi^2\right), \tag{1.8}$$

the exponential integrals

$$E_\nu(x) = x^{\nu-1}\Gamma(-\nu+1, x) \tag{1.9}$$

(which, for $\nu = -n$, a negative integer, yield the molecular integrals $A_n(x)$ [7]), and the error functions

$$\operatorname{erf} x = x\gamma^*(\tfrac{1}{2}, x^2), \qquad \operatorname{erfc} x = (1/\sqrt{\pi})\Gamma(\tfrac{1}{2}, x^2). \tag{1.10}$$

When a is integer-valued, $\gamma^*(a, x)$ becomes an elementary function,

$$\gamma^*(-n, x) = x^n, \qquad \gamma^*(n+1, x) = x^{-(n+1)}[1 - e^{-x}e_n(x)], \quad n = 0, 1, 2, \ldots, \tag{1.11}$$

where $e_n(x) = \sum_{k=0}^{n} x^k/k!$.

2. NORMALIZATION AND ASYMPTOTIC BEHAVIOR

The purpose of normalizing functions is twofold: In the first place, one wants to scale the function in such a way that underflow or overflow on a computer is avoided in as large a region as possible. In the second place, one wants to bring the function into a form in which it is used most naturally and conveniently in applications. There is little doubt as to what the proper normalization ought to be for $\Gamma(a, x)$ and $\gamma^*(a, x)$, when a is a positive number. The formulas (1.7), (1.8), (1.10), and (1.11), indeed, all point toward the normalization

$$G(a, x) = \frac{\Gamma(a, x)}{\Gamma(a)}, \qquad g^*(a, x) = x^a\gamma^*(a, x), \quad 0 \le x < \infty, \ a > 0. \tag{2.1}$$

We then have, by (1.5),

$$G(a, x) + g^*(a, x) = 1, \quad 0 \le x < \infty, \ a > 0.$$

It is equally clear that division by $\Gamma(a)$ to normalize $\Gamma(a, x)$, when a is negative or zero, is undesirable, as this would generate functions identically zero for $x > 0$, when a is integer-valued, and would cause complications in evaluating exponential and molecular integrals (cf. (1.9)). Growth considerations, on the other hand, suggest a multiplicative factor $e^x x^{-a}$. The function $\gamma^*(a, x)$ behaves rather capriciously for $a < 0$ and is not easily normalized. We decided (somewhat reluctantly) to adopt the same normalization as in (2.1), primarily for reasons of uniformity and good behavior for large a and x. We are doing this, however, at the expense of introducing a singularity along the line $x = 0$. For nonpositive a, we thus define

$$G(a, x) = e^x x^{-a}\Gamma(a, x), \qquad g^*(a, x) = x^a\gamma^*(a, x), \quad 0 \le x < \infty, \ a \le 0. \tag{2.2}$$

Our efforts will be directed towards computing $G(a, x)$ and $g^*(a, x)$ in the region $\mathscr{H}$.

It is useful to briefly indicate the behavior of $G(a, x)$ and $g^*(a, x)$ in the various parts of the region $\mathscr{H}$. The limit values, as x approaches zero for fixed a, are readily found to be

$$\begin{aligned} &G(a, 0) = 1, && g^*(a, 0) = 0 \quad \text{if } a > 0, \\ &G(a, 0) = \infty, && g^*(a, 0) = 1 \quad \text{if } a = 0, \\ &G(a, 0) = 1/|a|, && g^*(a, 0) = \infty \quad \text{if } a < 0. \end{aligned} \tag{2.3}$$

(It should be noted that $g^*(a, x)$, considered as a function of two independent variables, is indeterminate at $a = 0$, $x = 0$.) If $|a|$ is bounded and x large, we

deduce from well-known asymptotic formulas [13, p. 174],

$$\begin{aligned} G(a, x) &\sim \frac{e^{-x}x^{a-1}}{\Gamma(a)}, \quad a > 0 \text{ bounded}, \\ G(a, x) &\sim \frac{1}{x}, \qquad a \leq 0 \text{ bounded}, \quad x \to \infty. \\ g^*(a, x) &\sim 1, \qquad |a| \text{ bounded}, \end{aligned} \tag{2.4}$$

Equally simple is the case $|x|$ bounded and $a \to \infty$ (over positive values of a), in which case [13, p. 175]

$$\begin{aligned} G(a, x) &\sim 1, \qquad |x| \text{ bounded}, \\ & \qquad\qquad\qquad a(>0) \to \infty. \\ g^*(a, x) &\sim \frac{e^{-x}x^{a}}{\Gamma(a+1)}, \quad |x| \text{ bounded}, \end{aligned} \tag{2.5}$$

An indication of the behavior of these functions, when both variables are large, can be gained by setting $x = \rho a$, $\rho > 0$ fixed, and letting $a \to \infty$. Laplace's method, applied to the integrals in (1.1), then gives

$$G(a, \rho a) \sim \begin{cases} 1, & 0 < \rho < 1, \\ \frac{1}{2}, & \rho = 1, \qquad a \to \infty, \\ \dfrac{\rho^a e^{-(\rho-1)a}}{\sqrt{2\pi a}\,(\rho - 1)}, & \rho > 1, \end{cases}$$

$$g^*(a, \rho a) \sim \begin{cases} \dfrac{\rho^a e^{(1-\rho)a}}{\sqrt{2\pi a}\,(1 - \rho)}, & 0 < \rho < 1, \\ \frac{1}{2}, & \rho = 1, \qquad a \to \infty. \\ 1, & \rho > 1, \end{cases} \tag{2.6}$$

Similarly,

$$G(-a, \rho a) \sim \frac{1}{(\rho + 1)a}, \quad 0 < \rho < \infty, \; a \to \infty,$$

$$g^*(-a, \rho a) \sim \begin{cases} \dfrac{2 \sin \pi a}{\sqrt{2\pi a}\,(\rho + 1)} \rho^{-a} e^{-a(\rho+1)} & \text{if } \rho e^{\rho+1} < 1, \\ a \not\equiv 0 \pmod 1, \qquad a \to \infty. \\ 1 \quad \text{if } \rho e^{\rho+1} \geq 1, \end{cases} \tag{2.7}$$

3. CHOICE OF PRIMARY FUNCTION

Either of the two functions $\Gamma(a, x)$, $\gamma^*(a, x)$ can be expressed in terms of the other by means of the relations

$$\Gamma(a, x) = \Gamma(a)\{1 - x^a\gamma^*(a, x)\}, \qquad \gamma^*(a, x) = x^{-a}\left\{1 - \frac{\Gamma(a, x)}{\Gamma(a)}\right\}. \tag{3.1}$$

In our choice of primary function, we are guided primarily by considerations of numerical stability. We must be careful not to lose excessively in accuracy when

we perform the subtractions indicated in braces in (3.1). No such loss occurs if the absolute value of the respective difference is larger than, or equal to, $\frac{1}{2}$. This criterion is easily expressed in terms of the ratio

$$r(a, x) = \frac{\Gamma(a, x)}{\Gamma(a)}. \tag{3.2}$$

Indeed, the first relation in (3.1) is stable exactly if $|r(a, x)| \geq \frac{1}{2}$, while the second is stable in either of the two cases $r(a, x) \geq \frac{3}{2}$ and $r(a, x) \leq \frac{1}{2}$. As a consequence, *an ideal choice of the primary function is* $\gamma^*(a, x)$ *if* $\frac{1}{2} \leq r(a, x) \leq \frac{3}{2}$, *and* $\Gamma(a, x)$ *if* $|r(a, x)| \leq \frac{1}{2}$; *in all remaining cases either choice is satisfactory.*

For the practical implementation of this criterion, consider first $a > 0$, $x > 0$. In this case, $0 < r(a, x) < 1$, and $r(a, x)$ increases monotonically in the variable a ([14, p. 276]). Since $\lim_{a\to 0} r(a, x) = 0$ and, by (2.5), $\lim_{a\to\infty} r(a, x) = 1$, there is a unique curve $a = \alpha(x)$ in the first quadrant $x > 0$, $a > 0$, along which $r(a, x) = \frac{1}{2}$, and $r(a, x) \gtrless \frac{1}{2}$ depending on whether $a \gtrless \alpha(x)$. Since, by (2.6), $r(x, x) \sim \frac{1}{2}$ as $x \to \infty$, we have $\alpha(x) \doteq x$ for x large. By numerical computation it is found that in fact $\alpha(x) \doteq x$ for all (except very small) positive x, the value of $\alpha(x)$ consistently being slightly larger than x. As $x \to 0$ one finds $\alpha(x) \sim \ln \frac{1}{2}/\ln x$, which suggests the approximation $\alpha(x) \doteq \alpha^*(x)$, where

$$\alpha^*(x) = \begin{cases} x + \frac{1}{4}, & \frac{1}{4} \leq x < \infty, \\ \ln \frac{1}{2}/\ln x, & 0 < x \leq \frac{1}{4}. \end{cases} \tag{3.3}$$

The proper choice of primary function thus is $\Gamma(a, x)$ (resp. $G(a, x)$) if $0 < a \leq \alpha(x)$, and $\gamma^*(a, x)$ (resp. $g^*(a, x)$) if $a > \alpha(x)$, where $\alpha(x)$ may be approximated by $\alpha^*(x)$ in (3.3).

In the case $a \leq 0$, $x > 0$, the second relation in (3.1) is stable if $\Gamma(a) < 0$, i.e. if

$$-m - 1 < a < -m, \tag{3.4}$$

where $m \geq 0$ is an even integer. If $m \geq 1$ is an odd integer and a as in (3.4), then for x not too large there is a possibility that $\gamma^*(a, x)$ will vanish. The second relation in (3.1) is then subject to cancellation errors. A similar problem of cancellation, however, would occur if $\gamma^*(a, x)$ were calculated directly (e.g. from its Taylor expansion in the variable x). Furthermore, if $\gamma^*(a, x)$ were the primary function, the first relation in (3.1) would create serious (though not unsurmountable) computational difficulties for values of a near (or equal!) to a nonpositive integer (cf. the relevant discussion in [3]). All these considerations lead us to adopt $\Gamma(a, x)$ (resp. $G(a, x)$) as the primary function, whenever $a \leq 0$.

In summary, then, *our choice of primary function is* $\Gamma(a, x)$ (*resp.* $G(a, x)$), *if* $-\infty < a \leq \alpha(x)$, *and* $\gamma^*(a, x)$ (*resp.* $g^*(a, x)$), *if* $a > \alpha(x)$. Here, $\alpha(x)$ is adequately approximated by $\alpha^*(x)$ in (3.3).

4. THE COMPUTATION OF $G(a, x)$

As discussed in Section 3, it suffices to consider the region $-\infty < a \leq \alpha^*(x)$, $x \geq 0$. We shall break up this region into the following three subregions:

Region I: $0 \leq x \leq x_0$, $-\frac{1}{2} \leq a \leq \alpha^*(x)$.
Region II: $0 \leq x \leq x_0$, $-\infty < a < -\frac{1}{2}$.
Region III: $x > x_0$, $-\infty < a \leq \alpha^*(x)$.

The breakpoint x_0 will be chosen to have the value $x_0 = 1.5$. (A motivation for this choice is given in subsection 4.1.) We use a different method of computation in each of these three subregions. In Region I we first compute $\Gamma(a, x)$ directly from (1.4) and (1.5), and then use (2.1) or (2.2), depending on whether $a > 0$ or $a \leq 0$, to obtain $G(a, x)$. In Region II we employ a recurrence relation in the variable a, the starting value being computed by the method appropriate for Region I (except, possibly, when $x < \frac{1}{4}$). In Region III we use a continued fraction due to Legendre. We now proceed to describe and justify these various methods in more detail.

4.1 Direct Computation of $\Gamma(a, x)$ and $G(a, x)$ for $0 < x \leq x_0$, $-\frac{1}{2} \leq a \leq \alpha^*(x)$

Using (1.5), we can write

$$\Gamma(a, x) = \Gamma(a) - \frac{x^a}{a} + \frac{x^a}{a}[1 - \Gamma(a + 1)\gamma^*(a, x)]. \tag{4.1}$$

We let

$$u = \Gamma(a) - \frac{x^a}{a},$$
$$v = \frac{x^a}{a}[1 - \Gamma(a + 1)\gamma^*(a, x)], \tag{4.2}$$

and propose to use

$$\Gamma(a, x) = u + v \tag{4.3}$$

as a basis of computation in Region I. The breakpoint x_0 will be determined, among other things, from the requirement that the relative error generated in (4.3) (due to respective errors in u and v) be within acceptable limits.

Before analyzing these errors, we observe that both quantities u and v have finite limits as $a \to 0$, when $x > 0$. Indeed,

$$\lim_{a \to 0} u = -\gamma - \ln x, \qquad \lim_{a \to 0} v = E_1(x) + \gamma + \ln x, \tag{4.4}$$

where $\gamma = .57721...$ is Euler's constant and $E_1(x)$ is the exponential integral. The first relation follows at once from

$$u = \frac{\Gamma(1 + a) - 1}{a} - \frac{x^a - 1}{a}, \tag{4.5}$$

the second from (4.3) by letting $a \to 0$ and noting that $\Gamma(0, x) = E_1(x)$. Furthermore, from (1.1) and (1.3), we have

$$v = \int_0^x t^{a-1}(1 - e^{-t})dt, \tag{4.6}$$

valid not only for $a > 0$, but even for $a > -1$. In particular, therefore,

$$v > 0 \quad \text{if } a > -1, x > 0. \tag{4.7}$$

Using Taylor's expansion in (4.6) it is possible to compute v very accurately, essentially to machine precision. The same can be said for u, except that near the

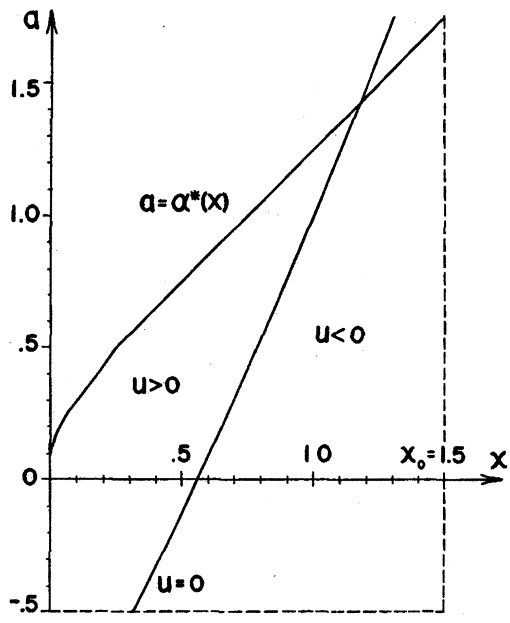


Fig. 1. The subregions $u \gtreqless 0$ in Region I

line where $u = 0$ (see Figure 1) the precision will be attained only in terms of the absolute error, not the relative error. If the absolute and relative error of u is e_u and ϵ_u, respectively, and ϵ_v is the relative error of v, then the relative error ϵ_Γ of $\Gamma(a, x)$, computed by (4.3), will be

$$\epsilon_\Gamma = \frac{e_u + v\epsilon_v}{u + v} = \frac{u\epsilon_u + v\epsilon_v}{u + v}.$$

Therefore, if $\epsilon = \max(|\epsilon_u|, |\epsilon_v|)$, we have, in view of (4.7),

$$|\epsilon_\Gamma| \leq \epsilon \quad \text{if} \quad u > 0, \qquad |\epsilon_\Gamma| \leq \left(1 + \frac{2|u|}{u + v}\right)\epsilon \quad \text{if} \quad u < 0. \tag{4.8}$$

Similarly, if $e = \max(|e_u|, |\epsilon_v|)$, then

$$|\epsilon_\Gamma| \leq \frac{1 + v}{u + v}\, e. \tag{4.9}$$

It is seen from the first relation in (4.8) that (4.3) is perfectly stable if $u > 0$, except possibly when u is very close to zero, in which case the absolute (not relative) error of u is what matters. Even then, however, one finds that the error magnification in (4.3) is negligible, since along the line $u = 0$ in Region I the factor $1 + 1/v$ multiplying e in (4.9) is always less than 3.8. In the subregion $u < 0$ of Region I it has been determined by computation[1] that the magnification factor

$$\mu(a, x) = 1 + \frac{2|u|}{u + v}$$

[1] A preliminary version of an algorithm for computing $\Gamma(a, x)$ and $\gamma^*(a, x)$ (see [3]) was used for this purpose.

Table I Maximum Error Magnification in Formula (4.3)

x_0	0 5	1.0	1.5	2.0	2.5	3.0
$\mu(-\frac{1}{2}, x_0)$	3.426	18.34	56.25	142.6	327.3	706.5

in (4.8) decreases monotonically as a function of a. It is easily verified, moreover, that in the same subregion the quantity $\mu(a, x)$ increases monotonically as a function of x. Therefore, the maximum error magnification occurs at the corner $(-\frac{1}{2}, x_0)$ of Region I. Table I shows the value of μ at this corner point in dependence on x_0. A similar behavior is exhibited by the magnification factor in (4.9). Its values at $(-\frac{1}{2}, x_0)$, however, are generally smaller than those in Table I.

Since the continued fraction used in Region III converges rather more slowly when x gets small, we have an interest in choosing x_0 as large as possible. Unfortunately, this runs counter the increased instability of (4.3). By way of a compromise, we will adopt the value $x_0 = 1.5$, thus accepting a possible loss of between 1 and 2 decimal digits. This choice of x_0 also strikes a reasonable balance in the computational work on either side of the boundary line separating Region III from Region I.

For the actual computation of u, we use (4.5) when $|a| < \frac{1}{2}$ and the first of (4.2), otherwise. The term in (4.5) involving the gamma function will be written in the form

$$\frac{\Gamma(1+a)-1}{a} = -\frac{1}{a}\cdot\Gamma(1+a)\cdot\left\{\frac{1}{\Gamma(1+a)}-1\right\}, \quad |a|<\frac{1}{2},$$

and evaluated using the Taylor expansions of $[\Gamma(1+a)]^{-1}$ and $[\Gamma(1+a)]^{-1}-1$, respectively. High-precision values of the necessary coefficients are available in [16, table 5]. Similarly, for the remaining term we write

$$\frac{x^a-1}{a} = \frac{e^{a\ln x}-1}{a\ln x}\cdot\ln x,$$

and evaluate the first factor on the right by Taylor expansion whenever $|a \ln x| < 1$.

The computation of v is most easily accomplished by series expansion. From (4.6) we find immediately

$$v = -x^a \sum_{n=1}^{\infty} \frac{(-x)^n}{(a+n)n!}, \quad a > -1, \tag{4.10}$$

or, equivalently,

$$v = \frac{x^{a+1}}{a+1}\sum_{k=0}^{\infty} t_k, \qquad t_k = \frac{(a+1)(-x)^k}{(a+k+1)(k+1)!}, \quad k = 0, 1, 2, \ldots.$$

The terms t_k can be obtained recursively by

$$t_0 = 1, \qquad t_k = -\frac{(a+k)x}{(a+k+1)(k+1)}\,t_{k-1}, \quad k = 1, 2, 3, \ldots.$$

In an effort to reduce the number of arithmetic operations, we define $p_k = (a + k)x$, $q_k = (a + k + 1)(k + 1)$, $r_k = a + 2k + 3$, and generate $\{t_k\}$ by means of

$$p_0 = ax, \quad q_0 = a + 1, \quad r_0 = a + 3, \quad t_0 = 1,$$
$$\left.\begin{aligned} p_k &= p_{k-1} + x, \\ q_k &= q_{k-1} + r_{k-1}, \\ r_k &= r_{k-1} + 2, \\ t_k &= -p_k \cdot t_{k-1}/q_k, \end{aligned}\right\} \quad k = 1, 2, 3, \ldots.$$

This requires only three additions, one multiplication, and one division per iteration step.

It is worth noting that overflow poses no serious threat in computing $\Gamma(a, x)$ as described. Indeed, $\Gamma(a, x)$ decreases in x, hence is largest along the left boundary of Region I. The respective boundary values are finite, equal to $\frac{1}{2}\Gamma(a) = (1/2a)\Gamma(a + 1)$, if $a > 0$, and infinite, if $a \leq 0$. As $x \to 0$ for fixed $a \leq 0$, $\Gamma(a, x)$ behaves like $E_1(x) \sim -\gamma - \ln x$, if $a = 0$, and like $-x^a/a$, if $a < 0$. In all cases ($a > 0$, and $-\frac{1}{2} < a \leq 0$) the values of $\Gamma(a, x)$ are machine representable if a, $1/a$, and x are.

Having computed $\Gamma(a, x)$, one obtains $G(a, x)$ from the first relations in (2.1) and (2.2), according as $a > 0$ or $a \leq 0$, respectively. The secondary function $g^*(a, x)$ then follows from $G(a, x)$ by

$$g^*(a, x) = \begin{cases} 1 - \dfrac{e^{-x}x^a G(a, x)}{\Gamma(a)}, & a < 0, \\ 1, \quad a = 0, & \\ 1 - G(a, x), & a > 0. \end{cases} \tag{4.11}$$

4.2 Recursive Computation of $G(a, x)$ for $0 < x \leq x_0$, $-\infty < a < -\frac{1}{2}$

We let[2] $m = [\frac{1}{2} - a]$, so that

$$a = -m + \epsilon, \qquad -\tfrac{1}{2} < \epsilon \leq \tfrac{1}{2},$$
$$G(a, x) = G(-m + \epsilon, x),$$

where m is an integer greater than or equal to 1. Defining $G_n = G(-n + \epsilon, x)$, $n = 0, 1, 2, \ldots$, the well-known recurrence relation in the variable a, satisfied by $\Gamma(a, x)$, yields[3]

$$G_0 = G(\epsilon, x),$$
$$G_n = \frac{1}{n - \epsilon}(1 - x\,G_{n-1}), \quad n = 1, 2, \ldots, m. \tag{4.12}$$

[2] The symbol $[r]$ denotes the largest integer less than or equal to r.

[3] The normalization (2.2) for $G(\epsilon, X)$ must be adopted here, even if $\epsilon > 0$.

The error propagation pattern in (4.12) is very similar for all ϵ in $-\frac{1}{2} < \epsilon \leq \frac{1}{2}$. When x is small ($x \lesssim 0.2$), the error is consistently damped for all n. When x is larger, there is an initial interval $1 \leq n \leq n_0$ in which the error is amplified, and a subsequent interval $n > n_0$ of rapid error damping. As x increases, both n_0 and the maximum error amplification increases. The latter, however, is well within acceptable limits, if $x \leq x_0 = 1.5$, the error never being amplified by more than a factor of 5.7. The case $\epsilon = 0$, which is typical, is analyzed in [5, example 5.4 and fig. 3]. (Note, in this connection, that $G(-m, x) = e^x E_{m+1}(x)$, where $E_{m+1}(x)$ is the exponential integral of order $m + 1$.) The recurrence relation (4.12), therefore, is extremely stable in the region in which it is being used.

The initial value $G_0 = G(\epsilon, x)$ can be computed by the method appropriate for Region I (see Section 4.1), except when $x < \frac{1}{4}$ and $\epsilon > \alpha^*(x)$, in which case $g^*(\epsilon, x)$ is computed first (see Section 5), whereupon $G(\epsilon, x)$ is obtained in a stable manner from $g^*(\epsilon, x)$, using $G(\epsilon, x) = \Gamma(\epsilon)e^x x^{-\epsilon}(1 - g^*(\epsilon, x))$ (cf. footnote 3).

4.3 Computation of G(a, x) for $x > x_0$, $-\infty < a \leq \alpha^*(x)$ by Legendre's Continued Fraction

The following continued fraction, due to Legendre, is well known ([11, p. 103; 1, eq. 6.5.31]),

$$x^{-a}e^x\Gamma(a, x) = \frac{1}{x+}\,\frac{1-a}{1+}\,\frac{1}{x+}\,\frac{2-a}{1+}\,\frac{2}{x+}\cdots. \tag{4.13}$$

It converges for any $x > 0$ and for arbitrary real a. We can write (4.13) in contracted form as

$$x^{-a}e^x\Gamma(a, x) = \frac{\beta_0}{x + \alpha_0 +}\,\frac{\beta_1}{x + \alpha_1 +}\,\frac{\beta_2}{x + \alpha_2 +}\cdots,$$

$$\alpha_k = 2k + 1 - a, \qquad k = 0, 1, 2, \ldots,$$

$$\beta_0 = 1, \qquad \beta_k = k(a - k), \qquad k = 1, 2, 3, \ldots,$$

or, alternatively, in the form

$$(x + 1 - a)x^{-a}e^x\Gamma(a, x) = \frac{1}{1+}\,\frac{a_1}{1+}\,\frac{a_2}{1+}\,\frac{a_3}{1+}\cdots, \tag{4.14}$$

where

$$a_k = \frac{k(a - k)}{(x + 2k - 1 - a)(x + 2k + 1 - a)}, \qquad k = 1, 2, 3, \ldots. \tag{4.15}$$

We investigate the convergence character of the continued fraction in (4.14) for $x > x_0 = 1.5$, $-\infty < a \leq \alpha^*(x)$, which is Region III, in which (4.14) is going to be used.

It is well known (cf., e.g. [15, p. 17ff]) that any continued fraction of the form (4.14) can be evaluated as an infinite series,

$$\frac{1}{1+}\,\frac{a_1}{1+}\,\frac{a_2}{1+}\,\frac{a_3}{1+}\cdots = \sum_{k=0}^{\infty} t_k, \tag{4.16}$$

where

$$t_0 = 1, \qquad t_k = \rho_1\rho_2\cdots\rho_k, \qquad k = 1, 2, 3, \ldots, \tag{4.17}$$

$$\rho_0 = 0, \qquad \rho_k = \frac{-a_k(1+\rho_{k-1})}{1+a_k(1+\rho_{k-1})}, \qquad k = 1, 2, 3, \ldots. \tag{4.18}$$

The nth partial sum in (4.16), in fact, is equal to the nth convergent of the continued fraction, $n = 1, 2, 3, \ldots$. If we let $\sigma_k = 1 + \rho_k$, then the recursion for ρ_k in (4.18) translates into the following recursion for σ_k:

$$\sigma_0 = 1, \quad \sigma_k = \frac{1}{1+a_k\sigma_{k-1}}, \quad k = 1, 2, 3, \ldots. \tag{4.19}$$

Consider now the case of a_k as given in (4.15). If $k < a$ (thus $a > 1$), then $a_k > 0$ (since $a \le x + \frac{1}{4}$), and it follows inductively from (4.19) that $0 < \sigma_k < 1$; hence $-1 < \rho_k < 0$. In view of (4.17), this means that (4.16) initially behaves like an alternating series with terms decreasing monotonically in absolute value.

If $k > a$, then $a_k < 0$, and σ_k may become larger than 1. However, if $0 < \sigma_{k-1} \le 2$, we claim that $1 < \sigma_k \le 2$ whenever $x \ge \frac{1}{4}$. Indeed, for the upper bound we must show that $1 + a_k\sigma_{k-1} \ge \frac{1}{2}$, i.e. $a_k\sigma_{k-1} \ge -\frac{1}{2}$, or, equivalently, $|a_k|\sigma_{k-1} \le \frac{1}{2}$. Since $\sigma_{k-1} \le 2$, it suffices to show $|a_k| \le \frac{1}{4}$, which is equivalent to $1 \le (x-a)^2 + 4kx$. Since $k \ge 1$ and $x > 0$, the latter is certainly true if $x \ge \frac{1}{4}$, which proves the assertion $\sigma_k \le 2$. The other inequality, $1 < \sigma_k$, is an easy consequence of $1 + a_k\sigma_{k-1} > 0$, established in the course of the argument just given, and the negativity of a_k. Since for the first k with $k > a$ we have $0 < \sigma_{k-1} \le 1$ (by virtue of the discussion in the preceding paragraph, or by virtue of $\sigma_0 = 1$), it follows inductively that $1 < \sigma_k \le 2$ for all $k > a$, hence $0 < \rho_k \le 1$. In the case $k = a$, we have $a_k = 0$ and $\sigma_k = 1$, thus $\rho_k = 0$, and the argument again applies.

We have shown that $|\rho_k| \le 1$ for all $k \ge 1$, that is, the terms in the series of (4.16) are nonincreasing in modulus, whenever $-\infty < a \le \alpha^*(x)$, $x \ge \frac{1}{4}$, in particular, therefore, when (x, a) is in Region III under consideration. Moreover, the series changes from an alternating series (if $a > 1$), initially, to a monotone series, ultimately.

In the region $a > \alpha^*(x)$, convergence of Legendre's continued fraction may deteriorate considerably in speed, which, together with the appropriate choice of primary function, is the reason we prefer a different method for $a > \alpha^*(x)$ (cf. Section 5).

Computationally, the summation in (4.16), with the a_k given in (4.15), can be simplified similarly as in (4.10). We now define $p_k = -k(a-k)$, $q_k = (x + 2k - 1 - a)(x + 2k + 1 - a)$, $r_k = 4(x + 2k + 1 - a)$, $s_k = 2k - a + 1$, and generate the terms t_k in (4.16) by means of

$$p_0 = 0, \quad q_0 = (x - 1 - a)(x + 1 - a), \quad r_0 = 4(x + 1 - a),$$
$$s_0 = -a + 1, \quad \rho_0 = 0, \quad t_0 = 1,$$

$$\left.\begin{aligned} p_k &= p_{k-1} + s_{k-1} \\ q_k &= q_{k-1} + r_{k-1} \\ r_k &= r_{k-1} + 8 \\ s_k &= s_{k-1} + 2 \\ \tau_k &= p_k(1 + \rho_{k-1}) \\ \rho_k &= \frac{\tau_k}{q_k - \tau_k} \\ t_k &= \rho_k t_{k-1} \end{aligned}\right\} \quad k = 1, 2, 3, \ldots. \tag{4.20}$$

This requires six additions, two multiplications, and one division per term.

5. THE COMPUTATION OF $g^*(a, x) = x^a\gamma^*(a, x)$

We need to consider only the region $a > \alpha^*(x)$, $x \geq 0$, in which $g^*(a, x)$ is the primary function (cf. Section 3). Among the tools available for computing $\gamma^*(a, x)$ are the two power series

$$\Gamma(a + 1)e^x\gamma^*(a, x) = e^x \sum_{n=0}^{\infty} \frac{a(-x)^n}{(a + n)n!} = \Gamma(a + 1) \sum_{n=0}^{\infty} \frac{x^n}{\Gamma(a + n + 1)}, \tag{5.1}$$

which follow immediately from (1.4), and the continued fraction

$$\Gamma(a + 1)e^x\gamma^*(a, x) = \frac{1}{1-} \quad \frac{x}{a + 1 + x -} \quad \frac{(a + 1)x}{a + 2 + x -} \quad \frac{(a + 2)x}{a + 3 + x -} \cdots, \tag{5.2}$$

which can be derived from Perron's continued fraction for ratios of Kummer functions [4]. In our preliminary work [3] we used the first series in (5.1), if $x \leq 1.5$, and the continued fraction (5.2), if $x > 1.5$. Our preference for the alternating series in (5.1) was motivated by the fact that $\gamma^*(a, x)$ in [3] served as primary function in the whole strip $0 \leq x \leq 1.5$, $-\infty < a < \infty$. In this case the first series in (5.1) has the advantage of terminating after the first term, if $a = 0$, and of presenting similar simplifications if a is a negative integer. These advantages had to be reconciled with problems of internal cancellation, which increase as x gets larger. In the present setup, these considerations become irrelevant, and indeed for $a > \alpha^*(x)$ the second series in (5.1) is clearly more attractive, all terms being positive (hence no cancellation), and convergence being quite rapid, even for x relatively large (in which case $a > x + \frac{1}{4}$).

How does this series compare with the continued fraction (5.2)? Rather surprisingly, the answer is: They are identical! In other words, the successive convergents of the continued fraction are identical with the successive partial sums of the series. To see this, let A_n, B_n be the numerators and denominators of the continued fraction in (5.2), so that, in particular,

$$B_1 = 1, \qquad B_2 = a + 1,$$

$$B_n = (a + n - 1 + x)B_{n-1} - (a + n - 2)xB_{n-2}, \quad n = 3, 4, \ldots.$$

One easily verifies by induction that

$$B_1 = 1, \qquad B_n = (a + 1)(a + 2) \cdots (a + n - 1), \quad n = 2, 3, \ldots. \tag{5.3}$$

From the theory of continued fractions it is known that

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

where $a_1 = 1$, $a_2 = -x$, $a_n = -(a + n - 2)x$ $(n > 2)$ are the partial numerators in (5.2). It follows, by virtue of (5.3), that

$$\frac{A_n}{B_n} - \frac{A_{n-1}}{B_{n-1}} = \frac{x^{n-1}}{(a + 1)(a + 2) \cdots (a + n - 1)}, \quad n \geq 2;$$

hence

$$\frac{A_n}{B_n} = 1 + \sum_{k=2}^{n} \left(\frac{A_k}{B_k} - \frac{A_{k-1}}{B_{k-1}} \right) = 1 + \sum_{k=2}^{n} \frac{x^{k-1}}{(a + 1)(a + 2) \cdots (a + k - 1)},$$

which is the nth partial sum of the series on the far right of (5.1).

Since series are easier to compute than continued fractions, we propose to compute $g^*(a, x)$ by

$$g^*(a, x) = x^a e^{-x} \sum_{n=0}^{\infty} \frac{x^n}{\Gamma(a + n + 1)} \tag{5.4}$$

everywhere in the region $a > \alpha^*(x)$.

The use of (5.4) in the region $a > \alpha^*(x)$ is comparable, with regard to computational effort, to the use of Legendre's continued fraction in the neighboring region $a < \alpha^*(x)$, $x > 1.5$, except when x is very large and $a \doteq \alpha^*(x)$, in which case Legendre's continued fraction is more efficient. Some pertinent data are shown in Table II. We determined the number of iterations required for 8 decimal digit accuracy in Legendre's continued fraction (4.14), when $a = \alpha^*(x)(1 - h)$, and in the power series (5.4), when $a = \alpha^*(x)(1 + h)$, where h was given the values 0.001, 0.25, 0.5, 0.75, 1.0, 1.5, 2.0, 2.5, 3.0, and $x = 10, 20, 40, 80, \ldots, 10240$. Table II shows for each x the minimum and maximum number of iterations

Table II. Number of Iterations in Legendre's Continued Fraction (4.14) and in Taylor's Expansion (5.4) for 8-Digit Accuracy

	$x = 10$		$x = 20$		$x = 40$		$x = 80$		$x = 160$		$x = 320$	
	min	max	min	max	min	max	min	max	min	max	min	max
Legendre	7	10	6	15	5	19	4	25	4	32	3	41
Taylor	13	24	13	31	14	42	14	57	14	77	14	106
	$x = 640$		$x = 1280$		$x = 2560$		$x = 5120$		$x = 10240$			
Legendre	3	52	3	65	3	82	2	101	2	124		
Taylor	14	146	14	202	14	279	14	387	14	536		

as h varies over the values specified. The number of iterations consistently decreases with increasing $|h|$, so that the maximum occurs on the boundary line $a = \alpha^*(x)$. In order to properly evaluate the data in Table II, one must keep in mind that each iteration in Legendre's continued fraction, using the algorithm in (4.20), requires seven additions, two multiplications, and one division, whereas each iteration in Taylor's series requires only two additions, one multiplication, and one division. Thus, Legendre's continued fraction is $2 - 2\frac{1}{2}$ times as expensive, per iteration, as Taylor's series.

6. TESTING

The algorithm in [2] and a double precision version of it were tested extensively on the CDC 6500 computer at Purdue University against a double precision version of the procedure in [3]. The double precision algorithms were used to provide reference values for checking the single precision algorithm, and on a few occasions, to check against high precision tables (notably the 14S tables in [8]). Other reference values were taken from various mathematical tables in the literature.

The tests include:

(i) the error functions (1.10), checked against tables 7.1 and 7.3 in [1];
(ii) the case (1.11) of integer values $a = n$, $-20 \leq n \leq 20$;
(iii) the exponential integral $E_\nu(x)$ in (1.9) for integer values $\nu = n$, $0 \leq n \leq 20$, and fractional values of ν in $0 \leq \nu \leq 1$, checked against tables I, II, III in [9];
(iv) Pearson's incomplete gamma function (1.7), checked against tables I and II in [10];
(v) the incomplete gamma function $P(a, x) = (x/2)^a\gamma^*(a, x/2)$, checked against the tables in [6];
(vi) the χ^2 distribution (1.8), checked against table 26.7 in [1];
(vii) the molecular integral $A_n(x)$, checked against table 1 in [7] and the more accurate tables in [8].

An important feature of our algorithm is the automatic monitoring of overflow and underflow conditions. This is accomplished by first computing the logarithm of the desired quantities and by making the tests for overflow and underflow on the logarithms. As a result, minor inaccuracies are introduced in the final exponentiation, which become particularly noticeable if the result is near the overflow or underflow limit.

7. SEQUENCES OF INCOMPLETE GAMMA FUNCTIONS

Expansions in terms of incomplete gamma functions require the generation of sequences $G_n = G(\alpha + n, x)$ or $g_n^* = g^*(\alpha + n, x)$ for fixed α and $n = 0, 1, 2, \ldots$, or of suitably scaled sequences $\{\lambda_n G_n\}$, $\{\lambda_n^* g_n^*\}$, where $\lambda_n \neq 0$, $\lambda_n^* \neq 0$ are scale factors. (For the purpose of the following discussion, the choice of these factors is immaterial; we shall assume, therefore, $\lambda_n = \lambda_n^* = 1$.) It would be wasteful to compute the G_n and g_n^* individually, for each n, by some evaluation procedure (such as the one developed in Sections 3–5). More efficient is the use of recurrence

relations satisfied by G_n and g_n^*. We discuss this in the case $\alpha > 0$, $x > 0$, which is a case of practical importance.

7.1 Generation of $G_n = G(\alpha + n, x)$

From the difference equation $G(a + 1, x) = G(a, x) + x^a e^{-x}/\Gamma(a + 1)$, letting $a = \alpha + n$, one finds immediately the recurrence relation

$$G_{n+1} = G_n + \frac{x^{\alpha+n}e^{-x}}{\Gamma(\alpha + n + 1)}, \quad n = 0, 1, 2, \ldots. \tag{7.1}$$

The numerical stability of (7.1) is determined by the solution $h_n = 1$ of the associated homogeneous recurrence relation, through the "amplification factors" [5]

$$\rho_n = \left|\frac{G_0 h_n}{G_n}\right| = \frac{\Gamma(\alpha, x)}{\Gamma(\alpha)}\frac{\Gamma(\alpha + n)}{\Gamma(\alpha + n, x)}. \tag{7.2}$$

Indeed, if s and t are arbitrary nonnegative integers, a small (relative) error ϵ injected into (7.1) at $n = s$ will propagate into a (relative) error $\epsilon\rho_t/\rho_s$ at $n = t$, causing the error to be damped if $\rho_t < \rho_s$ and magnified if $\rho_t > \rho_s$. To achieve consistent error damping, hence perfect numerical stability, the recurrence relation (7.1) ought to be applied *in the direction of decreasing* ρ_n.

Since $\Gamma(a, x)/\Gamma(a)$ increases from 0 to 1 on the interval $0 < a < \infty$ [14, p. 276], we see from (7.2) that ρ_n decreases monotonically from 1 to $\Gamma(\alpha, x)/\Gamma(\alpha)$, as n increases from 0 to ∞. It follows that *the recurrence relation* (7.1) *is perfectly stable in the forward direction.* The proper way to compute the sequence $\{G_n\}$, therefore, consists in first evaluating $G_0 = G(\alpha, x)$ (using our evaluation procedure, for example), and then applying (7.1) for $n = 0, 1, 2, \ldots$ to successively generate as many of the G_n as desired.

7.2 Generation of $g_n^* = g_n^*(\alpha + n, x)$

From the difference equation $g^*(a + 1, x) = g^*(a, x) - x^a e^{-x}/\Gamma(a + 1)$, we now find the recurrence relation

$$g_{n+1}^* = g_n^* - \frac{x^{\alpha+n}e^{-x}}{\Gamma(\alpha + n + 1)}, \quad n = 0, 1, 2, \ldots, \tag{7.3}$$

which has associated the amplification factors

$$\rho_n^* = \left|\frac{g_0^* h_n^*}{g_n^*}\right| = \frac{\gamma(\alpha, x)}{\Gamma(\alpha)}\frac{\Gamma(\alpha + n)}{\gamma(\alpha + n, x)}, \tag{7.4}$$

since $h_n^* = 1$ and $g_n^* = \gamma(\alpha + n, x)/\Gamma(\alpha + n)$. Noting that $\Gamma(a, x)/\Gamma(a) = 1 - \gamma(a, x)/\Gamma(a)$, and that the ratio on the left increases monotonically from 0 to 1 as a function of a, it follows that $\gamma(a, x)/\Gamma(a)$ decreases monotonically from 1 to 0, hence that ρ_n^* increases monotonically from 1 to ∞ as n increases from 0 to ∞. Therefore, *the recurrence relation* (7.3) *is perfectly stable in the backward direction.* Wishing to compute g_n^* for $n = 0, 1, 2, \ldots, N$, say, we should therefore use our evaluation procedure on $g_N^* = g^*(\alpha + N, x)$, and then employ (7.3) in the

form

$$g_n^* = g_{n+1}^* + \frac{x^{\alpha+n}e^{-x}}{\Gamma(\alpha+n+1)}, \quad n = N-1, N-2, \ldots, 0,$$

to generate all remaining values of g_n^*.

ACKNOWLEDGMENT

The author is indebted to N. M. Temme for suggesting the approach in Sections 4.1 and 4.2.

REFERENCES

1. ABRAMOWITZ, M., AND STEGUN, I.A., EDS. *Handbook of Mathematical Functions.* Nat. Bur. Standards Appl. Math. Series 55, U.S. Govt. Printing Office, 1964.
2. GAUTSCHI, W. Algorithm 542. Incomplete gamma functions. *ACM Trans. Math. Software 5,* 4 (Dec. 1979), 482–489.
3. GAUTSCHI, W. An evaluation procedure for incomplete gamma functions. MRC Tech. Summary Rep. 1717, Mathematics Res. Ctr., U. of Wisconsin, Madison, Feb. 1977.
4. GAUTSCHI, W Anomalous convergence of a continued fraction for ratios of Kummer functions. *Math. Comp. 31* (1977), 994–999.
5. GAUTSCHI, W Zur Numerik rekurrenter Relationen. *Comptg. 9* (1972), 107–126. (English translation in Aerospace Res. Lab. Rep. ARL 73-0005, Wright Patterson AFB, Feb. 1973.)
6. KHAMIS, S H. *Tables of the Incomplete Gamma Function Ratio.* Justus von Liebig Verlag, Darmstadt, 1965.
7. KOTANI, M., AMENIYA, A., ISHIGURO, E., AND KIMURA, T. *Table of Molecular Integrals.* Maruzen Co , Ltd., Tokyo, 1955
8. MILLER, J., GERHAUSEN, J.M., AND MATSEN, F.A. *Quantum Chemistry Integrals and Tables* U. of Texas Press, Austin, 1959.
9. PAGUROVA, V I. *Tables of the Exponential Integral* $E_r(x) = \int_1^\infty e^{-xu}u^{-r}\,du$. (Translated from the Russian by D. G. Fry), Pergamon Press, New York, 1961.
10. PEARSON, K., Ed., *Tables of the Incomplete* Γ*-Function* Biometrika Office, U. College, Cambridge U. Press, Cambridge, 1934.
11. PERRON, O. *Die Lehre von den Kettenbruchen, Vol. II,* 3rd ed., B.G. Teubner, Stuttgart, 1957.
12. TEMME, N.M. The asymptotic expansion of the incomplete gamma functions. Rep. TW 165/77, Stichting Mathematisch Centrum, Amsterdam, 1977.
13. TRICOMI, F.G. Funzioni ipergeometriche confluenti. Edizioni Cremonese, Rome, 1954.
14 TRICOMI, F.G. Sulla funzione gamma incompleta. *Ann. Mat. Pura Appl. (4) 31* (1950), 263–279.
15. WALL, H.S. *Analytic Theory of Continued Fractions* Van Nostrand, New York, 1948. (Reprinted in 1967 by Chelsea Publ Co., Bronx, N.Y)
16. WRENCH, J.W., JR. Concerning two series for the gamma function. *Math. Comp. 22* (1968), 617–626.

Received April 1977, revised August 1978

9.10. [72] "Lower bounds for the largest zeros of orthogonal polynomials"

---

[72] (with F. Costabile) "Lower bounds for the largest zeros of orthogonal polynomials," *Boll. Un. Mat. Ital.* (5) **17A**, 516–522 (1980) (translated from Italian).

Lower bounds for the largest zeros of orthogonal polynomials*

F. COSTABILE (Cosenza) and W. GAUTSCHI (Lafayette)

Abstract. For the largest zero of orthogonal polynomials we derive lower bounds which are somewhat sharper than those classically known.

1. Basic theorems

Let $P_n(x)$ be a polynomial of degree n having distinct positive zeros, denoted by

$$r_1 > r_2 > \cdots > r_n > 0. \tag{1.1}$$

We define

$$u_k = \sum_{i=1}^{n} c_i r_i^k, \quad k = 0, 1, 2, \ldots, \tag{1.2}$$

where c_i are constants.

Theorem 1.1. *If the c_i in* (1.2) *are positive, and $n > 1$, then, putting*

$$\sigma_k = \frac{u_{k+1}}{u_k}, \tag{1.3}$$

one has

$$\lim_{k\to\infty} \sigma_k = r_1, \quad \sigma_k < \sigma_{k+1} < r_1, \quad k = 0, 1, 2, \ldots . \tag{1.4}$$

Proof. Since

* English translation by Walter Gautschi of "Stime per difetto per gli zeri più grandi dei polinomi ortogonali", *Bolletino U. M. I.* (5) **17**-A (1980), 516–522.

$$u_k = c_1 r_1^k \left[1 + \sum_{i=2}^{n} \frac{c_i}{c_1} \left(\frac{r_i}{r_1}\right)^k\right], \quad c_1 > 0,$$

from (1.1) and (1.3) there follows

$$\lim_{k\to\infty} \sigma_k = r_1.$$

Applying Cauchy's inequality, we find

$$u_{k+1}^2 = \left(\sum_{i=1}^{n} c_i r_i^{k+1}\right)^2 = \left(\sum_{i=1}^{n} \sqrt{c_i} r_i^{k/2} \sqrt{c_i} r_i^{k/2+1}\right)^2$$
$$< \sum_{i=1}^{n} c_i r_i^k \sum_{i=1}^{n} c_i r_i^{k+2} = u_k u_{k+2},$$

from which, dividing by $u_k u_{k+1}$, the second relations in (1.4) follow.

Theorem 1.2. *Let $\{Q_n(x)\}$ be a sequence of polynomials orthogonal with respect to the (nonnegative and integrable) weight function $p(x)$ on the finite or semi-infinite interval $[a, b]$, $a \geq 0$*[1]*. Let*

$$r_1^{(n)} > r_2^{(n)} > \cdots > r_n^{(n)} > 0$$

be the zeros of $Q_n(x)$, and

$$m_k = \int_a^b x^k p(x)dx < \infty, \quad k = 0, 1, 2, \ldots \tag{1.5}$$

the kth moments of the weight function $p(x)$. Putting

$$\tau_k = \frac{m_{k+1}}{m_k}, \quad k = 0, 1, 2, \ldots, \tag{1.6}$$

there holds

(a) $\{\tau_k\}$ is monotonically increasing

(b) $r_1^{(n)} > \tau_{2n-1} - \dfrac{1}{m_{2n-1}} \displaystyle\int_a^b \omega_n^2(x)p(x)dx > \tau_{2n-2}, \quad n > 1,$

with

$$\omega_n(x) = \prod_{i=1}^{n} (x - r_i^{(n)}).$$

[1] The original paper has $a > 0$. (Translator's note)

Proof. For the constants c_i in (1.2) we choose the Christoffel numbers, so that

$$\int_a^b f(x)p(x)dx = \sum_{i=1}^{n} c_i f(r_i^{(n)}) + \frac{1}{(2n)!} f^{(2n)}(\xi) \int_a^b \omega_n^2(x)p(x)dx, \quad \xi \in (a,b), \tag{1.7}$$

holds for any function f having a continuous $(2n)$th derivative.

Letting $f(x) = x^k$ in (1.7) gives

$$m_k = u_k, \quad k = 0, 1, \ldots, 2n-1,$$
$$m_{2n} = u_{2n} + \int_a^b \omega_n^2(x)p(x)dx,$$

from which

$$\sigma_k = \tau_k, \quad k = 0, 1, \ldots, 2n-2,$$
$$\sigma_{2n-1} = \tau_{2n-1} - \frac{1}{m_{2n-1}} \int_a^b \omega_n^2(x)p(x)dx.$$

From Theorem 1.1 there follows

$$r_1^{(n)} > \sigma_{2n-1} = \tau_{2n-1} - \frac{1}{m_{2n-1}} \int_a^b \omega_n^2(x)p(x)dx > \sigma_{2n-2} = \tau_{2n-2},$$

that is, (b).

The assertion (a) is obtained by applying Schwarz's inequality in place of Cauchy's, in a manner similar to the proof of Theorem 1.1.

2. Applications to classical orthogonal polynomials

2.1. Jacobi polynomials

Let $P_n^{(\alpha,\beta)}(x)$ be the Jacobi polynomial of degree n, that is, the polynomial orthogonal with respect to the weight function $(1-x)^\alpha(1+x)^\beta$ with $\alpha, \beta > -1$ on the interval $[-1, 1]$. With

$$1 > \xi_1^{(n)} > \xi_2^{(n)} > \cdots > \xi_n^{(n)} > -1$$

denoting the zeros of $P_n^{(\alpha,\beta)}(x)$, we have the following theorem.

Theorem 2.1. *For $n > 1$, there holds*

$$\xi_1^{(n)} > H_n, \tag{2.1}$$

where

$$H_n = \frac{2n+\beta-\alpha-1}{2n+\beta+\alpha+1} - \frac{2\Gamma(n+\alpha+1)\Gamma(n+\beta+1)\Gamma(n+\alpha+\beta+1)\Gamma(n+1)}{\Gamma(2n+\alpha+\beta+2)\Gamma(2n+\beta)\Gamma(\alpha+1)}. \tag{2.2}$$

Proof. We observe, first of all, that the polynomial $P_n^{(\alpha,\beta)}(x)$, by means of the linear transformation

$$x = 2y - 1,$$

becomes the polynomial $\overline{P}_n^{(\alpha,\beta)}(y)$, orthogonal on $[0,1]$ with respect to the weight function $(1-y)^\alpha y^\beta$. We can therefore apply Theorem 1.2 to the polynomial $\overline{P}_n^{(\alpha,\beta)}(y)$. For this purpose, we note that

$$m_k = \int_0^1 (1-y)^\alpha y^{\beta+k} dy = B(k+\beta+1, \alpha+1) = \frac{\Gamma(\alpha+1)\Gamma(k+\beta+1)}{\Gamma(k+\alpha+\beta+2)},$$

and therefore

$$\tau_k = \frac{m_{k+1}}{m_k} = \frac{k+\beta+1}{k+\alpha+\beta+2}.$$

If $\overline{\xi}_1^{(n)}$ indicates the largest zero of $\overline{P}_n^{(\alpha,\beta)}(y)$, from Theorem 1.2 one gets

$$\overline{\xi}_1^{(n)} > \overline{H}_n > \tau_{2n-2},$$

with

$$\overline{H}_n = \tau_{2n-1} - \frac{1}{m_{2n-1}} \int_0^1 \omega_n^2(y)(1-y)^\alpha y^\beta dy = \frac{2n+\beta}{2n+\alpha+\beta+1} - \frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)\Gamma(n+\alpha+\beta+1)\Gamma(n+1)}{\Gamma(2n+\alpha+\beta+2)\Gamma(2n+\beta)\Gamma(\alpha+1)}.$$

Taking into account that $\xi_1^{(n)} = 2\overline{\xi}_1^{(n)} - 1$, $H_n = 2\overline{H}_n - 1$, one obtains (2.1), (2.2).

Observation 2.1. The lower bound for $\xi_1^{(n)}$ just obtained is sharper than the one that can be obtained from Laguerre's theorem. The latter, in fact, is [1, p. 119]

$$\xi_1^{(n)} > \frac{2n+\beta-\alpha-2}{2n+\beta+\alpha},$$

whereas

$$H_n = 2\sigma_{2n-1} - 1 > 2\sigma_{2n-2} - 1 = 2\tau_{2n-2} - 1 = \frac{2n+\beta-\alpha-2}{2n+\beta+\alpha}.$$

Observation 2.2. The subtrahend in (2.2) is an infinitesimal quantity as $n \to \infty$, tending to zero like

$$[\pi/2^{\alpha+2\beta-1}\Gamma(\alpha+1)]n^{\alpha+1}2^{-4n}.$$

2.2. Ultraspherical polynomials

Let $P_n^{(\alpha,\alpha)}(x)$ be the ultraspherical polynomial of degree n, that is, the polynomial orthogonal on $[-1,1]$ with respect to the weight function $(1-x^2)^\alpha$, and $\xi_i^{(n)}$, $i = 1,2,\ldots,n$, the respective zeros in decreasing order. We then have the following theorem.

Theorem 2.2. *For $n > 2$, there holds*

$$\xi_1^{(n)} > \left(\frac{2n-3}{2n+2\alpha-1}\right)^{\frac{1}{2}}. \tag{2.3}$$

Proof. Because of the symmetry of the interval and weight function, putting

$$P_n^{(\alpha,\alpha)}(x) = \begin{cases} p_{n/2}^{(\alpha,\alpha)}(x^2) & \text{if } n \text{ is even,} \\ x q_{\lfloor n/2 \rfloor}^{(\alpha,\alpha)}(x^2) & \text{if } n \text{ is odd,} \end{cases}$$

the polynomials $p_{n/2}^{(\alpha,\alpha)}(t)$ and $q_{\lfloor n/2 \rfloor}^{(\alpha,\alpha)}(t)$ form an orthogonal system on the interval $[0,1]$ with respect to the weigth functions $(1-t)^\alpha t^{-\frac{1}{2}}$ and $(1-t)^\alpha t^{\frac{1}{2}}$. Furthermore, if by $\bar{\xi}_i^{(n)}$, $i = 1,2,\ldots,\lfloor n/2 \rfloor$,[2] we denote the zeros of $p_{n/2}^{(\alpha,\alpha)}(t)$ $\left(q_{\lfloor n/2 \rfloor}^{(\alpha,\alpha)}(t)\right)$, one obviously has

$$\xi_1^{(n)} = \sqrt{\bar{\xi}_1^{(n)}}. \tag{2.4}$$

We now apply Theorem 1.2 to the polynomial $p_{n/2}^{(\alpha,\alpha)}(t)$.

Since

$$m_k = \int_0^1 (1-t)^\alpha t^{k-\frac{1}{2}} dt = \frac{\Gamma(k+\frac{1}{2})\Gamma(\alpha+1)}{\Gamma(k+\alpha+3/2)},$$

and therefore

$$\tau_k = \frac{m_{k+1}}{m_k} = \frac{k+1/2}{k+\alpha+3/2},$$

it follows that

$$\bar{\xi}_1^{(n)} > \frac{n-3/2}{n+\alpha-1/2},$$

[2] The original has $i = 1,2,\ldots,2\lfloor n/2 \rfloor$. (Translator's note)

that is, (2.3), by virtue of (2.4).

In a similar manner, one deals with the case $n > 1$ odd.

Observation 2.3. From Laguerre's theorem [1, p. 119] one obtains the bound

$$\xi_1^{(n)} > \left(\frac{n-1}{n+2\alpha+1}\right)^{\frac{1}{2}} \tag{2.5}$$

which is worse, if $n > 2$, than the one just derived.

Observation 2.4. One checks, with a simple calculation, that

$$\left(\frac{2n-3}{2n+2\alpha-1}\right)^{\frac{1}{2}} > \frac{2n-1}{2n+2\alpha+1} \quad \text{for } n > 3,$$

so that the bound (2.3) is sharper than the one in (2.1), (2.2) for $\alpha = \beta$.

In Table 2.1 we indicate the quality of our bound (2.3) in comparison with Laguerre's bound, (2.5), for various values of α and n.

2.3. Laguerre and Hermite polynomials

Let $L_n^{(\alpha)}$ be the polynomial of degree n orthogonal on the interval $(0, \infty)$ with respect to the weight function $e^{-x}x^{\alpha}$, $\alpha > -1$, and

$$\xi_1^{(n)} > \xi_2^{(n)} > \cdots > \xi_n^{(n)} > 0$$

the respective zeros. From Theorem 1.2, taking into account that $m_k = \Gamma(\alpha+k+1)$, one has immediately the following theorem.

Theorem 2.3. *For $n > 1$, there holds*

$$\xi_1^{(n)} > H_n, \tag{2.6}$$

where

$$H_n = 2n + \alpha - \frac{n!\Gamma(n+\alpha+1)}{\Gamma(2n+\alpha)} > 2n + \alpha - 1. \tag{2.7}$$

For the Hermite polynomial $H_n(x)$ of degree n and its zeros $\xi_1^{(n)} > \xi_2^{(n)} > \cdots > \xi_n^{(n)}$ one obtains by a reasoning similar to the one in the proof of Theorem 2.2 the following theorem.

Theorem 2.4. *For $n > 2$, there holds*

$$\xi_1^{(n)} > \sqrt{n - 3/2}. \tag{2.8}$$

Table 1. Comparison of (2.3) with (2.5).

α	n	$\xi_1^{(n)}$	(2.3)	(2.5)
−0.80	5	0.98008	0.97260	0.95346
	10	0.99533	0.98844	0.97849
	20	0.99887	0.99464	0.98964
	40	0.99972	0.99741	0.99491
	80	0.99993	0.99873	0.99748
−0.40	5	0.94171	0.92394	0.87706
	10	0.98501	0.96647	0.93934
	20	0.99621	0.98417	0.96984
	40	0.99905	0.99230	0.98496
	80	0.99976	0.99620	0.99249
0.00	5	0.90618	0.88192	0.81650
	10	0.97391	0.94591	0.90453
	20	0.99313	0.97402	0.95119
	40	0.99824	0.98726	0.97530
	80	0.99955	0.99369	0.98758
0.75	5	0.84761	0.81650	0.73030
	10	0.95220	0.91064	0.84853
	20	0.98653	0.95581	0.91894
	40	0.99642	0.97802	0.95794
	80	0.99908	0.98904	0.97856
1.50	5	0.79821	0.76376	0.66667
	10	0.93039	0.87905	0.80178
	20	0.97919	0.93859	0.88976
	40	0.99428	0.96903	0.94147
	80	0.99850	0.98445	0.96978
3.00	5	0.71988	0.68313	0.57735
	10	0.88860	0.82462	0.72761
	20	0.96318	0.90676	0.83887
	40	0.98925	0.95178	0.91093
	80	0.99708	0.97546	0.95291

Observation 2.5. From Laguerre's theorem [1, p. 119] one has for the largest zero of the nth-degree Laguerre and Hermite polynomial respectively the bounds

$$\xi_1^{(n)} > 2n + \alpha - 1,$$

$$\xi_1^{(n)} > \sqrt{(n-1)/2}, \quad n > 2,$$

which turn out to be worse than the analogous bounds furnished by Theorems 2.3 and 2.4.

References

[1] SZEGÖ, G.: *Orthogonal polynomials*, Colloquium Publications XXIII, 4th ed., American Mathematical Society, Providence, R. I., 1975.

Received at the Office of the U. M. I.
on November 28, 1979

9.11. [155] "THE INCOMPLETE GAMMA FUNCTIONS SINCE TRICOMI"

---

[155] "The Incomplete Gamma Functions Since Tricomi," in *Tricomi's ideas and contemporary applied mathematics*, 203–237, *Atti Convegni Lincei* **147** (1998).

WALTER GAUTSCHI(*)

THE INCOMPLETE GAMMA FUNCTIONS SINCE TRICOMI

1. THE INCOMPLETE GAMMA FUNCTIONS UP TO 1950

Tricomi considered his work on the asymptotic behavior of Laguerre polynomials and their zeros among his «chief contributions to the theory of special functions» ([153, p. 56]). Nevertheless, the incomplete gamma function held a special fascination for him, as he was fond of calling it affectionately the Cinderella of special functions. I feel especially privileged to talk about this topic here, since the only time I met Tricomi in person was shortly before his death when he honored me by his presence in a colloquium lecture I gave in Turin. It was precisely the incomplete gamma functions and methods for computing them that I was talking about, a subject in which Tricomi still expressed a vivid interest.

The incomplete gamma functions arise from Euler's integral for the gamma function,

$$\Gamma(a) = \int_0^\infty e^{-t} t^{a-1} dt\,,$$

by decomposing it into an integral from 0 to x, and another from x to ∞,

$$\begin{aligned} \gamma(a,x) &= \int_0^x e^{-t} t^{a-1} dt, \quad \operatorname{Re} a > 0; \\ \Gamma(a,x) &= \int_x^\infty e^{-t} t^{a-1} dt, \quad |\arg x| < \pi\,. \end{aligned} \tag{1.1}$$

Historically, this decomposition was first studied in 1877 for $x = 1$ by Prym[118], apparently in an attempt to collect the poles at $a = 0, -1, -2, \ldots$ of the gamma function in the first (more manageable) integral, $\gamma(a, 1)$, leaving the second integral, $\Gamma(a, 1)$, an entire function. The functions (1.1), therefore, are sometimes referred to as Prym's functions. For general $x > 0$ (even for $x < 0$), however, the second integral in (1.1) already appears in Legendre's *Exercises* [85, pp. 399-343] and in some of his later works.

(*) Department of Computer Sciences - Purdue University - WEST LAFAYETTE, IN 47907-1398 (U.S.A.)

Noteworthy special cases of (1.1) are obtained when $a = 1 \pm n$ is an integer. Specifically, for $n \geq 0$,

$$\gamma(1+n,x) = n![1 - e^{-x}e_n(x)], \tag{1.2}$$

$$\Gamma(1+n,x) = n!e^{-x}e_n(x), \tag{1.3}$$

$$\Gamma(1-n,x) = x^{1-n}E_n(x), \tag{1.4}$$

where $e_n(x) = 1 + x + x^2/2! + \cdots + x^n/n!$, $n = 0, 1, 2, \ldots$, are the partial sums of the exponential series, and

$$E_n(x) = \int_1^\infty e^{-xt}t^{-n}dt, \quad n = 0, 1, 2, \ldots, \tag{1.5}$$

the exponential integrals. The latter occur prominently in astrophysics and nuclear physics and include (for $n = 1$) such functions as the logarithmic, sine, and cosine integrals. The function $\gamma(a,x)$ has a pole when a is a negative integer or zero; see, however, (2.1) and (2.2). When $a = \frac{1}{2}$ one obtains the error functions

$$\operatorname{erf} x = \frac{1}{\sqrt{\pi}}\gamma\left(\frac{1}{2}, x^2\right), \quad \operatorname{erfc} x = 1 - \operatorname{erf} x = \frac{1}{\sqrt{\pi}}\Gamma\left(\frac{1}{2}, x^2\right) \tag{1.6}$$

and their close relatives such as the Fresnel integrals.

The older theory of the incomplete gamma function, including series expansions of various kinds, asymptotic expansions, differentiation and recurrence relations, continued fractions, etc., can be found in Nielsen [103, Kap. II, XV, XXI], and further material, especially integral representations, in Böhmer [15, Kap. V]. The basic theory, however, remained rather stable, until in the late 1940s, as a result of his involvement in the Bateman project, Tricomi fully recognized the importance of these functions and revitalized their theory by adding important contributions of his own (see § 2) and by summarizing the knowledge as of 1950 in the second volume of the Bateman project [40, Ch. IX, pp. 133-151]. He gave a more detailed exposition, in the context of the theory of confluent hypergeometric functions, in his monograph [151, §§ 4.1–4.6].

One aspect of incomplete gamma functions, namely their real and complex zeros, does not receive an entirely adequate coverage in these works, in part, perhaps, because Tricomi's interest was in the x-zeros for fixed a, while work done in the early 1900s was exclusively concerned with a-zeros for fixed x. The earliest investigations dealt with the real negative zeros of $\gamma(a,x)$ for Prym's choice $x = 1$. Increasingly sharper localizations of these zeros were obtained in work of Haskins [59], Gronwall [56], and Walther [158]. Rasch [120] was the first to consider the case of arbitrary fixed $x > 0$, and Hille and Rasch [60] the case of $x < 0$. Complex zeros were already studied by Gronwall [56], who showed in the case $x = 1$ that there are exactly two conjugate complex pairs of them. They were subsequently computed to seven decimals by Franklin [43].

Nielsen [103] proved that all zeros of $\Gamma(a,x)$, for $x > 0$, lie in the half-plane $\operatorname{Re} a > x$. Rasch [120] gave an asymptotic formula for the number $M(x)$ of pairs of conjugate complex a-zeros of $\gamma(a,x)$ as $x \to \infty$. Hille and Rasch [60] already in 1929, and Mahler [96] in 1930, investigated the behavior of the zeros when x is a fixed complex number; they also identified zero-free regions in the complex a-plane.

Other texts on confluent hypergeometric functions are the one by Buchholz [17] published shortly before Tricomi's monograph, and one published later by Slater [125]. The former is based on Whittaker's definition [160, Ch. 16] of the confluent hypergeometric functions, a definition not favored by Tricomi; the latter also contains numerical tables. A detailed treatment of the probability integral and some of its generalizations, notably $\Phi(z,a) = \pi^{-1/2}\gamma\left(\frac{1}{2}a, z^2\right)$, can be found in a monograph by Hadži [58].

2. Tricomi's contributions

2.1. *Normalization*

The integral $\gamma(a,x)$ has the inconvenience of not only having poles at the nonpositive integers $a = 0, -1, -2, \ldots$, but also representing a multivalued function of the complex variable x, owing to the fractional power in the integrand. Both these inconveniences can be avoided by introducing, as Tricomi does in [146] and Böhmer before him in [15, pp. 124–125], the function

$$\gamma^*(a,x) = \frac{x^{-a}}{\Gamma(a)}\gamma(a,x)\,, \tag{2.1}$$

which is an entire function in a as well as in x and real-valued for real a and real x (also for $x < 0$). In particular,

$$\gamma^*(-n,x) = x^n, \quad n = 0,1,2,\ldots\,. \tag{2.2}$$

In terms of the function (2.1), both incomplete gamma functions in (1.1) can be represented as

$$\gamma(a,x) = \Gamma(a)x^a\gamma^*(a,x), \quad \Gamma(a,x) = \Gamma(a)[1 - x^a\gamma^*(a,x)]\,, \tag{2.3}$$

where fractional powers of x, as always in this theory, are to be understood as having their principal values. Tricomi finds it useful to introduce yet another form of the incomplete gamma function, namely

$$\gamma_1(a,x) = \Gamma(a)x^a\gamma^*(a,-x)\,, \tag{2.4}$$

for which, as he notes (cf. [151, p. 161]), one has

$$\gamma_1(a,x) = \int_0^x e^t t^{a-1}dt, \quad \operatorname{Re} a > 0\,. \tag{2.5}$$

This function allows the values of $\gamma(a,x)$ above and below the branch cut along the negative real axis to be expressed as

$$\gamma(a,-x\pm i0)=e^{\pm a\pi i}\gamma_1(a,x)\,,\quad x>0\,. \tag{2.6}$$

2.2. *Series expansions*

To the classical power series expansions Tricomi in [147] adds expansions in Bessel functions, which he obtains as special cases of similar expansions he derived for the confluent hypergeometric functions. Characteristically, Tricomi adopts a form of the Bessel functions which makes them entire functions of both the variable and the order, namely

$$J_\nu^*(x)=x^{-\nu/2}J_\nu(2\sqrt{x})\,. \tag{2.7}$$

(Tricomi's notation for them is $E_\nu(x)$.) In terms of these functions, he derives the expansion

$$\gamma^*(a,x)=e^{-x}\sum_{n=0}^{\infty}e_n(-1)x^nJ_{a+n}^*(-x)\,, \tag{2.8}$$

where $e_n(\cdot)$ is the $(n+1)$st partial sum of the exponential series (cf. § 1). For real arguments a and x, one can write (2.8) as an expansion in (ordinary) Bessel functions $J_{a+n}(2\sqrt{|x|})$ if x is negative (see [147, 2d equation (39)], where, however, the factor x^n should read $x^{n/2}$), and as a similar expansion in modified Bessel functions $I_{a+n}(2\sqrt{x})$ if x is positive. Both converge rather well, when $0\le a<1$ (for other values of a the recurrence relation (6.4) can be used), but the former suffers increasingly from internal cancellations as $|x|$ becomes large.

For good measure, Tricomi obtains yet another expansion,

$$\gamma^*(a,x)=e^{-x/2}\sum_{n=0}^{\infty}c_n\left(\frac{x}{2}\right)^nJ_{a+n}^*\left(\frac{a-1}{2}x\right)\,, \tag{2.9}$$

where the coefficients c_n can be obtained recursively from[1]

$$\begin{aligned}&c_0=1,\ c_1=0\,,\\ &c_n=c_{n-2}+L_n^{(1-a-n)}(1-a)\,,\end{aligned} \tag{2.10}$$

and $L_n^{(\alpha)}$ are the Laguerre polynomials. The peculiar form $\lambda_n(y)=L_n^{(y-n)}(y)$ of the Laguerre polynomials appearing in (2.10) is studied by Tricomi in [148], where he derived the recursion

$$\begin{aligned}&\lambda_0(y)=1,\ \lambda_1(y)=0\,,\\ &\lambda_{n+1}(y)=-\frac{1}{n+1}[n\lambda_n(y)+y\lambda_{n-1}(y)],\ n=1,2,\dots\,.\end{aligned} \tag{2.11}$$

[1]There is a misprint in [147, line after equation (41)] in that A_1^* (our c_1) is erroneously defined to be 1 instead of 0.

(The same polynomials are also used by Temme [134] in uniform asymptotic expansions of Laplace transforms.) The series (2.9) seems to converge (for $0 \leq a < 1$) somewhat faster than (2.8) and, for $x < 0$, also suffers less from internal cancellations. We used it (in IEEE Standard double precision) to produce the plots in § 2.4 as well as the graphs in figure 5.

Although Tricomi refers to the polynomials $\lambda_n(y)$ as being nonorthogonal, Carlitz [19] showed that in fact $(-1)^{n+1}(n+2)x^{n+2}\lambda_{n+2}(x^{-2})$, $n = 0, 1, 2, \ldots$, is a set of (monic) orthogonal polynomials relative to a measure that is discretely supported on the points $x_j = \pm j^{-1/2}$ with jumps $\frac{1}{2}j^{j-1}e^{-j}/j!$, $j = 1, 2, 3 \ldots$. These polynomials occur also as «random walk polynomials» in the work of Karlin and McGregor [70, Appendix B] on birth and death processes. Their asymptotics and zero distribution are studied in [52] and [53].

Other series expansions which may be original with Tricomi are an expansion [146, equation (44)] of $\Gamma(a, x)$ in Laguerre polynomials $L_n^{(a)}(x)$, and an expansion (ibid., equation (45)) of $\gamma(a, \lambda x)$ in $\{\gamma(a + n, x)\}$.

2.3. *Asymptotics*

The asymptotic behavior of the incomplete gamma functions is elementary when only one of the two parameters a and x tends to infinity. More interesting (and also more difficult) is the behavior when $|a|$ and $|x|$ become large simultaneously. Here Tricomi shows in [146] (see also [151, § 4.3]) that the matter depends on whether a and x are not near each other, or x is near $a - 1$, as $|a|$ and $|x|$ both tend to infinity. In the first case, he proves from the integral representation of $\Gamma(a + 1, x)$ that

$$\Gamma(a+1,x) = \frac{e^{-x}x^{a+1}}{x-a}\left[1 - \frac{a}{(x-a)^2} + \frac{2a}{(x-a)^3} + O\left(\frac{a^2}{(x-a)^4}\right)\right] \tag{2.12}$$

as the modulus of $\sqrt{a}/(x-a)$ tends to zero and its argument ultimately remains between $-3\pi/4$ and $3\pi/4$. He in fact has the complete asymptotic expansion in explicit (though complicated) form. In the second case there are two subcases depending on whether Re a is positive or negative. Equivalently, Tricomi considers the functions $\gamma(1 + a, x)$ and $\gamma_1(1 - a, x)$ separately, both under the assumption Re $a > 0$. In the first subcase, again from the integral representation (1.1), he finds, when a and y are both real and y bounded, that

$$\gamma(a+1, a+\sqrt{2a}y) = \frac{1}{2}\Gamma(a+1)\left[1 + \operatorname{erf} y - \frac{2}{3}\sqrt{\frac{2}{a\pi}}(1+y^2)e^{-y^2} + O\left(\frac{1}{a}\right)\right], \tag{2.13}$$

$$a \to \infty.$$

(For a simplified derivation of (2.13), see also [152].) A similar result holds for complex a, y (with Re $a > 0$), and again, Tricomi is able to write down the complete asymptotic expansion.

As a by-product of (2.13) and (1.3), one obtains a nice asymptotic estimate for $e_n(x)$ near $x = n$, namely

$$e_n(n + \sqrt{2n}y) = \frac{1}{2}e^{n+\sqrt{2n}y}\left[\operatorname{erfc} y - \frac{2}{3}\sqrt{\frac{2}{n\pi}}(1+y^2)e^{-y^2} + O\left(\frac{1}{n}\right)\right], \tag{2.14}$$
$$n \to \infty.$$

In the second subcase, Tricomi finds, for real $a > 0$ and $y \in \mathbb{R}$ bounded, that

$$\gamma_1(1-a, a+\sqrt{2a}y) = \frac{1}{\Gamma(a)}\left[-\pi\cot a\pi + 2\sqrt{\pi}\int_0^y e^{t^2}dt\right.$$
$$\left. + \frac{2}{3}\sqrt{\frac{2\pi}{a}}(1-y^2)e^{y^2} + O\left(\frac{1}{a}\right)\right], \quad a \to \infty. \tag{2.15}$$

2.4. *Zeros*

In [147] (see also [151, § 4.4]) Tricomi studies the zeros of $\gamma^*(a,x)$, $a \in \mathbb{R}$, $x \in \mathbb{R}$, considered as a function of x for fixed a. Except possibly for $x = 0$, these zeros coincide with the zeros of $\gamma(a,x)$ or $\gamma_1(a,-x)$. Tricomi gives a complete description of these zeros and, more generally, a remarkable contour map of the function $\gamma^*(a,x)$, i.e., of the lines $\gamma^*(a,x) = \text{const}$. In figure 1 we reproduce this map, and also provide the associated surface plot; they were generated by the MATLAB commands contour and surf, respectively. The function itself was computed with the help of the series expansion (2.9) for $0 \le a < 1$ and the recurrence relation (6.4) for other values of a.

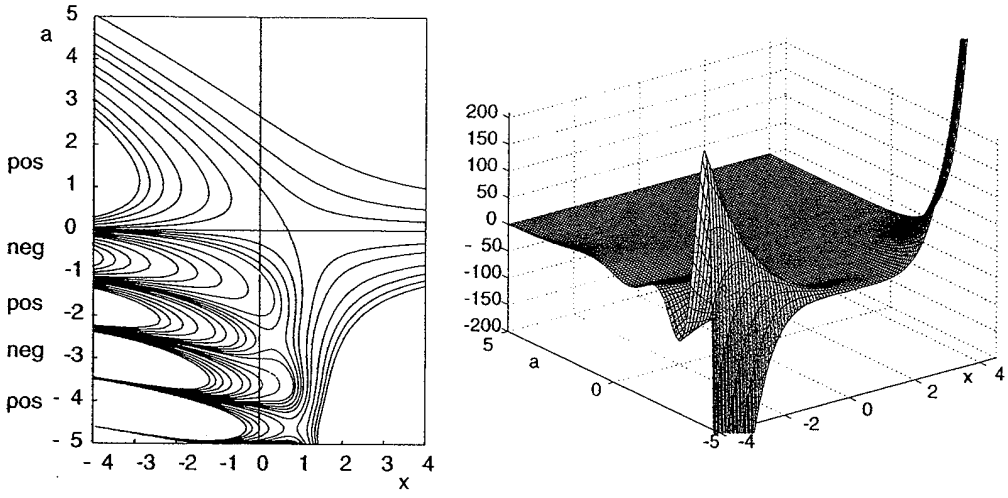


Fig. 1. — Contour map and surface plot of $\gamma^*(a,x)$. The lines in the contour map correspond to altitudes $-6(1) - 2(.5)0(.25)1(.5)2(1)6$, the zero line being red.

From his asymptotic results in [146], Tricomi derives the following asymptotic approximations (as corrected by Kölbig [78]) for the real zeros: for the positive zeros $x_+(a)$ of $\gamma^*(a,x)$,

$$x_+(a) = \tau|a| - \frac{\tau}{1+\tau}\log\left[\frac{(1+\tau)\sqrt{(1+|a|)\pi/2}}{\sin a\pi}\right] + O\left(\left(\frac{\log|a|}{a}\right)^2\right), \tag{2.16}$$
$$a<0,\ \sin a\pi>0,\ a\to-\infty,$$

where $\tau = .27846454\ldots$ is the unique positive root of the equation $1+x+\log x = 0$; for the negative zeros,

$$x_-(a) = -(1+|a|) - \sqrt{2(1+|a|)}y_0 + O(|a|^{-1/2}), \quad a\to-\infty, \tag{2.17}$$

where $y_0 = y_0(a)$ is the unique root of $\int_0^y e^{t^2}dt = \frac{\sqrt{\pi}}{2}\cot(|a|\pi)$, provided $|a|$ is not too close to a positive integer.

2.5. *Inequalities and monotonicity*

Obviously,

$$g(a,x) := \frac{\gamma(a,x)}{\Gamma(a)}, \quad a>0, \quad x>0, \tag{2.18}$$

is a probability distribution on $[0,\infty]$; thus, in particular, $0\le g(a,x)\le 1$ and g is monotonically increasing in x. In [147] Tricomi proves that g is monotone also in a, namely decreasing. Interestingly enough, Tricomi uncovers similar monotonicity properties also in the regions $a<0$, $x>0$ and $a>0, x<0$. In the former region,

$$G(a,x) := -ae^x x^{-a}\Gamma(a,x), \tag{2.19}$$

and in the latter,

$$g_1(a,|x|) := ae^{-|x|}|x|^{-a}\gamma_1(a,|x|), \tag{2.20}$$

are both between 0 and 1 and are monotone in x as well as in a. More difficult (not surprisingly in view of figure 1) is the region $a<0, x<0$. Here Tricomi manages to prove that $|g^*(a,x)|\le 1$, where

$$g^*(a,x) := \frac{e^x\gamma^*(a,x)}{\Gamma(|a|+1)}. \tag{2.21}$$

Moreover, as a function of x, with a held fixed, g^* has one, or at most two, maxima or minima.

2.6. *Applications*

2.6.1 *Number theory*

It is known from a well-known theorem of Lagrange that each positive integer can be decomposed into a sum of (at most) four perfect squares, whereas

only some integers are decomposable into a sum of two squares, and even fewer into a sum of two cubes, or two fourth powers, or, more generally, two kth powers, $k = 2, 3, 4, \ldots$. The problem of determining the distribution $N_k(x)$ of all positive integers $\leq x$ that are the sum of exactly two kth powers seems to have led Tricomi in 1938 to his first encounter with the incomplete gamma function. By probabilistic heuristics, and at times — as he says [151, p. 286], «acrobatic» — arguments, Tricomi in [143] indeed arrives at the approximation

$$N_k(x) \approx x - \frac{k}{k-2} A_k^{\frac{k}{k-2}} \Gamma\left(-\frac{k}{k-2}, A_k x^{\frac{2-k}{k}}\right), \quad k \geq 3, \tag{2.22}$$

where

$$A_k = \frac{[\Gamma(1/k)]^2}{2k^2\Gamma(2/k)}.$$

The nature of the approximation in (2.22), given its roundabout derivation, is of course unclear, but definitely is not asymptotic for $x \to \infty$, since a similarly derived approximation for $k = 2$ gives $N_2(x) \approx (1 - e^{-\pi/8})x$, whereas, by a result of Landau, $N_2(x) \sim bx/\sqrt{\log x}$ (cf. [86]), with a well-determined constant b approximately equal to .764 (cf. [151, p. 289]). Nevertheless, for x not too large, the formula (2.22) seems to give excellent results, as Tricomi demonstrates for $k = 3$ and $x \leq 2000$.

Precise asymptotic results have been obtained only more recently, for example in [61] for $k = 3$, and in [62] for k (odd) ≥ 5.

2.6.2. *Random walks*

A problem of interest in physics, biology, and other areas of science, is the following. Given randomly n unit vectors in Euclidean space $\mathbb{R}^d$, what is the probability $P_n(r) = \Pr(\|s\| < r)$ that their sum s has length $< r$, where $0 \leq r \leq n$? The problem has been solved in 1906 for $d = 2$ by J.C. Kluyver (even for vectors of arbitrary fixed lengths), with full details, for $d = 3$, supplied later in 1919 by Lord Rayleigh. In the case of general d, the result is derived in Watson [159, p. 421], where $P_n(r)$ is given in the form of an integral involving Bessel functions (here written in terms of Tricomi's Bessel functions),

$$P_n(r) = [\Gamma(d/2)]^{n-1}(r^2/n)^{d/2} \int_0^\infty t^{(d/2)-1} J^*_{d/2}(r^2 t/n)[J^*_{(d/2)-1}(t/n)]^n dt. \tag{2.23}$$

What is of particular interest in applications is the behavior of $P_n(r)$ as $n \to \infty$. Watson already studied this informally by applying the method of steepest descent and arriving at an asymptotic approximation involving ${}_1F_1\left(\frac{d}{2}; \frac{d}{2}+1; -\frac{r^2 d}{2n}\right)$, hence the incomplete gamma function. In [149] Tricomi, by a more rigorous approach using power series and Laplace transform techniques, improves upon

Watson's result, showing that

$$P_n(r) = x^{\frac{d}{2}} \left[\gamma^* \left(\frac{d}{2}, x \right) - \frac{1}{\Gamma(d/2)} \frac{e^{-x}}{n} \left(\frac{1}{2} - \frac{x}{d+2} \right) \right] + O(n^{-2}), \tag{2.24}$$

where

$$x = \frac{r^2 d}{2n}.$$

The behavior of the leading term in (2.24) is illustrated in figure 2 for $n = 10$ and $d = 2, 3, 5, 10$.

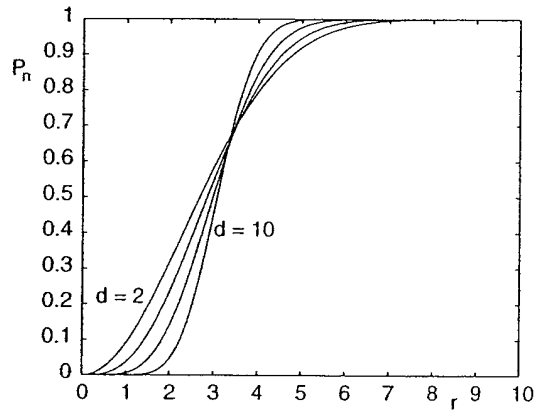


Fig. 2. — The probability $P_n(r)$ for $n = 10$ and $d = 2, 3, 5, 10$.

From (2.24) it takes a quick calculation for Tricomi to determine the mean value $\bar{r}$ of r, namely

$$\bar{r} = \int_0^\infty r \frac{dP_n}{dr} dr = \frac{\Gamma((d+1)/2)}{\Gamma(d/2)} \sqrt{\frac{2n}{d}} \left[1 + \frac{1}{4(d+2)n} + O(n^{-2}) \right]. \tag{2.25}$$

2.6.3. *Laguerre's equation*

The Laguerre polynomial $L_n^{(\alpha)}(x)$, as is well known, is a solution of the linear second-order differential equation

$$xy'' + (\alpha + 1 - x)y' + ny = 0, \tag{2.26}$$

a special case of the confluent hypergeometric equation. The second solution, $y_2(x)$, therefore, must be a confluent hypergeometric function, which Tricomi in [150], when α is not an integer, identifies explicitly in terms of the incomplete gamma function and products of Laguerre polynomials. Specifically,

$$y_2(x) = L_n^{(\alpha)}(x)\gamma_1(-\alpha, x) + e^x x^{-\alpha} \sum_{k=1}^{n} \frac{1}{k} L_{n-k}^{(\alpha+k)}(x) L_{k-1}^{(\alpha-k)}(-x). \tag{2.27}$$

(For γ_1, see (2.4).) There is an analogous formula involving $\Gamma(0, -x)$ when α is an integer.

2.7. *Miscellanea*

Without in any way wanting to disparage the results contained in this subsection, it seems fair to say that they lie at the fringes of the general theory of incomplete gamma functions and are therefore mentioned only in passing.

One concerns the gamma function itself, more precisely the ratio of two gamma functions, $\Gamma(z+a)/\Gamma(z+b)$, for which in [145] and [154] the complete asymptotic expansion in descending powers of z is derived, with an explicit characterization of the coefficients and the precise conditions of validity. Error bounds for this and similar expansions have later been obtained by Frenzen [44], [45].

In order to derive (2.15), Tricomi made use of the following (apparently new) integral representation of $\gamma^*(a,x)$ for real a and x,

$$\gamma^*(a,x) = \frac{e^{-x}}{\Gamma(a)\sin a\pi}\operatorname{Re}\left\{e^{-a\pi i}\int_0^\infty e^{-ixt}(1+it)^{a-1}dt\right\}. \tag{2.28}$$

Another curious integral representation is the one for the «norm» of the incomplete gamma function, $\Gamma(a,ix)\Gamma(a,-ix)$, which in [144] (or [40, § 9.3, equation (6)]) is expressed in terms of the Laplace transform of a sum of two conjugate complex hypergeometric functions.

3. Asymptotics

3.1. *An improved approximation (2.13)*

Formula (2.13), after division by $\Gamma(a+1)$, can be interpreted as an asymptotic approximation of the gamma distribution in terms of the normal distribution $\Phi(x) = (2\pi)^{-1/2}\int_{-\infty}^{x} e^{-t^2/2}dt$, the leading term in (2.13) indeed being $\Phi(\sqrt{2}y)$. A more accurate approximation has been derived by Pagurova [111] by statistical arguments; it involves derivatives of the normal distribution, hence Hermite polynomials $He_n(x)$,

$$\begin{aligned}\frac{\gamma(a,a+x\sqrt{a})}{\Gamma(a)} = \Phi(x) - \frac{e^{-x^2/2}}{\sqrt{2\pi}}\Bigg\{&\frac{1}{3\sqrt{a}}He_2(x) + \frac{1}{2a}\left[\frac{1}{2}He_3(x) + \frac{1}{9}He_5(x)\right]\\ &+ \frac{1}{a\sqrt{a}}\left[\frac{1}{5}He_4(x) + \frac{1}{12}He_6(x) + \frac{1}{162}He_8(x)\right]\\ &+ \frac{1}{6a^2}\left[He_5(x) + \frac{47}{80}He_7(x) + \frac{1}{12}He_9(x) + \frac{1}{324}He_{11}(x)\right]\\ &+ \frac{1}{a^2\sqrt{a}}\left[\frac{1}{7}He_6(x) + \frac{19}{180}He_8(x) + \frac{31}{1440}He_{10}(x)\right.\\ &\left.+ \frac{1}{648}He_{12}(x) + \frac{1}{29160}He_{14}(x)\right] + O(a^{-3})\Bigg\}.\end{aligned} \tag{3.1}$$

(The $a^{-1/2}$ term in (3.1), with $He_2(x) = x^2 - 1$, is consistent with the cor-

responding term in (2.13). This can be seen by applying the recurrence relation (6.4) to the left-hand side of (2.13), letting $x = \sqrt{2}y$, and applying elementary asymptotics to the additive term coming from the recurrence relation.) A similar, even more accurate, approximation (without the $a^{-1/2}$ term) is also derived, but it involves on the right-hand side a more complicated variable.

3.2. *Uniform asymptotics*

In deriving asymptotic results for large $|a|$, Tricomi found it necessary (cf. (2.12) and (2.13)) to distinguish cases according to the magnitude of $|x|$ relative to $|a|$. One of the major advances since Tricomi's work in this area is the development of asymptotic expansions for large a that hold uniformly for, say, all $x \geq 0$. There is a price to be paid, however, for uniformity: For one, the expansion involves not only elementary but also transcendental functions, specifically the error function erfc in our case; for another, the calculation of the expansion coefficients is much more intricate.

Uniform asymptotic expansions for the incomplete gamma functions were first derived by Temme [131], [132] (see also [140, § 11.2.4]). His point of departure is the integral representation

$$\frac{\Gamma(a,z)}{\Gamma(a)} = \frac{e^{-a\phi(\lambda)}}{2\pi i}\int_{c-i\infty}^{c+i\infty} e^{a\phi(t)}\frac{dt}{\lambda - t}, \quad 0 < c < \lambda, \tag{3.2}$$

where

$$\phi(t) = t - 1 - \ln t, \quad \lambda = \frac{z}{a}. \tag{3.3}$$

The integrand in (3.2) has a saddle point at $t = 1$. Changing the contour of integration into a path of steepest descent, and separating out the pole close to the saddle point (when $\lambda \approx 1$), Temme arrives at asymptotic representations of the type

$$\begin{aligned} \frac{\Gamma(a,z)}{\Gamma(a)} &= \frac{1}{2}\operatorname{erfc}\left(\eta\sqrt{a/2}\right) + R_a(\eta), \\ \frac{\gamma(a,z)}{\Gamma(a)} &= \frac{1}{2}\operatorname{erfc}\left(-\eta\sqrt{a/2}\right) - R_a(\eta), \\ R_a(\eta) &\sim \frac{e^{-\frac{1}{2}a\eta^2}}{\sqrt{2\pi a}}\sum_{n=0}^{\infty}\frac{c_n(\eta)}{a^n}, \quad a \to \infty, \end{aligned} \tag{3.4}$$

where

$$\eta = (\lambda - 1)\sqrt{2\frac{\lambda - 1 - \ln\lambda}{(\lambda - 1)^2}}. \tag{3.5}$$

(When $\lambda > 0$, then $\eta = \pm\sqrt{2(\lambda - 1 - \ln\lambda)}$ with the plus sign for $\lambda > 1$ and the minus sign for $\lambda < 1$.) The asymptotic expansion of $R_a(\eta)$ is valid for a going to infinity over positive values, and is uniform for all $\lambda \geq 0$, i.e., for all $z \geq 0$.

Its validity, indeed, can be established for complex a and z as well; that is, (3.4) is valid as $a \to \infty$ uniformly in $|\arg a| \leq \pi - \varepsilon_1$ and $|\arg z/a| \leq 2\pi - \varepsilon_2$, where ε_1, ε_2 are positive numbers with $0 < \varepsilon_1 < \pi$, $0 < \varepsilon_2 < 2\pi$.

As to the coefficients $c_n(\eta)$ in (3.4), they are holomorphic functions of η and can be computed for small $|\eta|$ by a power series expansion, and for other values of $|\eta|$ by recurrence. For details, as well as for estimates of the remainder terms when the expansion in (3.4) is truncated at some finite $n = N - 1$, we must refer to the original paper [132]; also see [114]. An extensive set of Taylor coefficients for $c_n(\eta)$ is given in [30, Appendix F]. The growth of $c_n(\eta)$ as $n \to \infty$ is studied in [34].

For a rearranged version of the expansion (3.4), in the context of the Riemann zeta function, see also [113, Appendix A].

It is interesting to note the role played by the error function in (3.4). If $z = \lambda a$, with a and λ positive, then $\Gamma(a, \lambda a)$ as a function of λ exhibits a sharp decrease near the transition point $\lambda = 1$, the decrease being sharper the larger a. Elementary functions would have a hard time describing this kind of behavior, but the error function does a nice job of it; this is shown in figure 3.

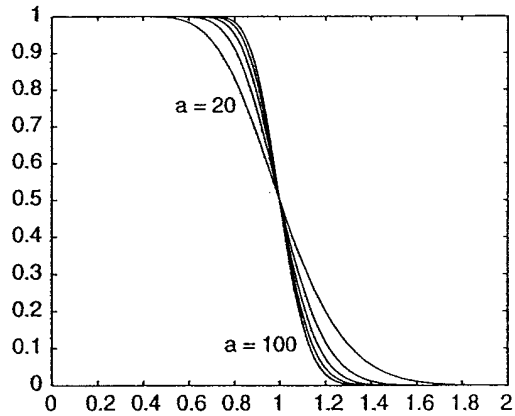


Fig. 3. — The leading term $\frac{1}{2}\operatorname{erfc}(\eta\sqrt{a/2})$ in (3.4) as a function of λ in $0 < \lambda \leq 2$, for $a = 20(20)100$.

At the transition point $\lambda = 1$ one has $\eta = 0$, and from the first of (3.4) one gets

$$\frac{\Gamma(a,a)}{\Gamma(a)} \sim \frac{1}{2} + \frac{1}{\sqrt{2\pi a}} \sum_{n=0}^{\infty} \frac{c_n(0)}{a^n}, \quad a \to \infty.$$

A similar expansion in which the factor multiplying the series is replaced by the asymptotically equivalent factor $\Gamma(a+1)(e/a)^a$ is given in [89], together with an asymptotic representation of the coefficients for large indices and the first eleven coefficients expressed exactly in rational form.

The expansions (3.4) do not cover negative values of a, but there are similar uniform expansions for $\gamma(-a,-z)$ and $\Gamma(-a,-z)$, involving the same coefficients $c_n(\eta)$, that are valid in $|\arg a| \leq \pi - \delta_1, |\arg \lambda| \leq 2\pi - \delta_2$, with δ_1, δ_2 arbitrarily small positive constants (cf. [139]). Of special interest is the expansion for $\gamma^*(-a,-x)$, where a and x are positive and γ^* real but oscillatory [139, equation (3.11)].

An alternative derivation of (3.4) and, with similar methods, of Tricomi's expansions in § 2.3, along with numerical comparisons, is given by Schell in [121].

Applying differential equations rather than integral representations, specifically the asymptotic theory of linear second-order differential equations with almost coalescent turning points, Dunster in [31] derives an alternative asymptotic approximation (not expansion) for $\Gamma(a,z)$ that also involves the complementary error function, but an auxiliary variable ζ rather more complicated than the η in Temme's expansion (3.4). The approximation holds, for example, when $a \to \infty$, uniformly for z in a domain containing the positive real axis, but there are other possible interpretations of its asymptotic character. For details, we refer to the original paper [31, Remarks on p. 1346].

3.3. *The generalized exponential integral*

If we take $n = p$ in (1.4) to be an arbitrary complex number, we are led to consider the generalized exponential integral

$$E_p(z) = z^{p-1}\Gamma(1-p,z) = z^{p-1}\int_z^\infty \frac{e^{-t}}{t^p}dt\,. \tag{3.6}$$

Even though closely related to the incomplete gamma function, it arises in this form in many applications and has attracted a considerable amount of interest in recent years.

3.3.1. *Asymptotic expansion for $p \to \infty$*

If p goes to ∞ over positive values $p > 1$, and x is an arbitrary nonnegative number, it was shown in [46] by elementary means, involving integration by parts, that

$$E_p(x) = \frac{e^{-x}}{x+p}\left[\sum_{k=0}^{n-1} H_k\left(\frac{x}{p}\right)p^{-k} + \Theta_n(x,p)p^{-n}\right], \tag{3.7}$$

where

$$\alpha_n \leq \Theta_n(x,p) \leq \beta_n\left(1 + \frac{1}{x+p-1}\right). \tag{3.8}$$

Here,

$$H_k(u) = \frac{h_k(u)}{(1+u)^{2k}}, \qquad k = 0,1,2,\ldots,$$

where $h_k(u)$ is a polynomial of degree $k-1$ (if $k>0$) defined recursively by

$$h_0(u)=1,\ h_{k+1}(u)=(1-2ku)h_k(u)+u(1+u)h_k'(u),\quad k=0,1,2,\dots,$$

and α_n,β_n are lower and upper bounds, respectively, of $H_n(u)$ on the interval $u\geq 0$. The first eight polynomials $h_k(u)$ and respective constants α_k,β_k are given explicitly in [46]. For improvements, both in the error bounds and the approximations, see also [6, § 3]. More recently, Dunster [32, Thm. 2.1] showed that the same[(2)] expansion (3.7) with a different bound on the error term (and different notations), holds uniformly for complex x in the domain $|\arg x|\leq \pi-\delta$, provided $p>(1-\cos\delta)^{-1}$.

An alternative asymptotic approximation for $p\ (>0)\to\infty$, valid for complex x in a domain containing the negative real axis, is given in [32, Thm. 3.1]. Similarly to the asymptotic approximation for the incomplete gamma function derived by the same author in [31], it also involves the complementary error function and a rather complicated auxiliary variable ζ. For an asymptotic expansion, including error bounds, see also [33, Thms. 5.1 and 5.2].

3.3.2. *Stokes's phenomenon and uniform exponential improvement*

For fixed p, as $z\to\infty$ in $|\arg z|\leq\frac{3}{2}\pi-\delta$, where δ is an arbitrarily small positive number, one has the classical asymptotic expansion

$$E_p(z)\sim\frac{e^{-z}}{z}\sum_{k=0}^{\infty}(-1)^k\frac{(p)_k}{z^k}, \tag{3.9}$$

where $(p)_k$ denotes the ascending factorial $p(p+1)\cdots(p+k-1)$. In the sector $\frac{1}{2}\pi+\delta\leq\arg z\leq\frac{7}{2}\pi-\delta$, which partly overlaps with the preceding sector, one has an asymptotic expansion just like (3.9) but with an additional term

$$\frac{2\pi i e^{-p\pi i}}{\Gamma(p)}z^{p-1}. \tag{3.10}$$

In the common sector $\frac{1}{2}\pi+\delta\leq\arg z\leq\frac{3}{2}\pi-\delta$ this term is exponentially small compared to the main term in (3.9). Nevertheless, as z crosses the line $\arg z=\pi$, there occurs a rapid, though smooth, change in the form of the asymptotic expansion. This is known as the Stokes phenomenon. It has been analyzed in a formal (but insightful) manner by Berry [11] and more rigorously by Olver [107], who writes the remainder term in (3.9), if truncated after the nth term, as follows,

$$E_p(z)=\frac{e^{-z}}{z}\sum_{k=0}^{n-1}(-1)^k\frac{(p)_k}{z^k}+\frac{2\pi i e^{-p\pi i}}{\Gamma(p)}z^{p-1}T_{n+p}(z), \tag{3.11}$$

[(2)]There is a sign error in equations (2.11) and (2.13) of [32], where the second term on the right should be subtracted instead of added.

where

$$T_q(z) = \frac{e^{q\pi i}\Gamma(q)}{2\pi i}\frac{E_q(z)}{z^{q-1}}. \tag{3.12}$$

The interest lies in the sector $\frac{1}{2}\pi < \arg z < \frac{3}{2}\pi$ (containing the «Stokes line» $\arg z = \pi$), where the factor T_{n+p} in (3.11) acts as a «Stokes multiplier» (cf. (3.10)). If n is chosen optimally, i.e., the series (3.9) is truncated just before its numerically smallest term, and if $z = \rho e^{i(\pi+\theta)}$, $-\frac{1}{2}\pi \le \theta \le \frac{1}{2}\pi$, then $n = \rho - p + \alpha$, where α is complex and bounded in absolute value as $\rho \to \infty$. By a delicate analysis, Olver then finds an asymptotic representation of the Stokes multiplier in the form

$$T_{n+p}(z) \sim \frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\eta\sqrt{\rho/2}\right) + \frac{e^{-\frac{1}{2}\rho\eta^2}}{\sqrt{2\pi\rho}}\sum_{k=0}^{\infty} c_k(\theta,\alpha)\rho^{-k}, \quad \rho \to \infty, \tag{3.13}$$

which holds uniformly for $\theta \in \left[-\frac{1}{2}\pi, \frac{1}{2}\pi\right]$ and for $|\alpha|$ bounded. Here η is a complex-valued function of θ defined by

$$\frac{1}{2}\eta^2 = 1 + i\theta - e^{i\theta}.$$

(The branch with $\operatorname{Re}\eta > 0$ is taken for $\theta > 0$ and the one with $\operatorname{Re}\eta < 0$ for $\theta < 0$.) The (complex) error function in (3.13) plays a similar role as the error function in (3.4), the transition point now being at $\theta = 0$. Plots of the real and imaginary parts of the leading term in (3.13) are shown in figure 4 as functions of λ, where $\theta = \lambda\pi/2$.

An alternative discussion of Stokes's phenomenon is given more recently by Dunster in [32, §§ 4 and 5].

In [108] it is shown that choosing n optimally as described, and expanding the remainder term in (3.11) in descending powers of ρ, provides in the domain $|\arg z| \le \pi - \delta$ a «uniformly exponentially improved» approximation in the

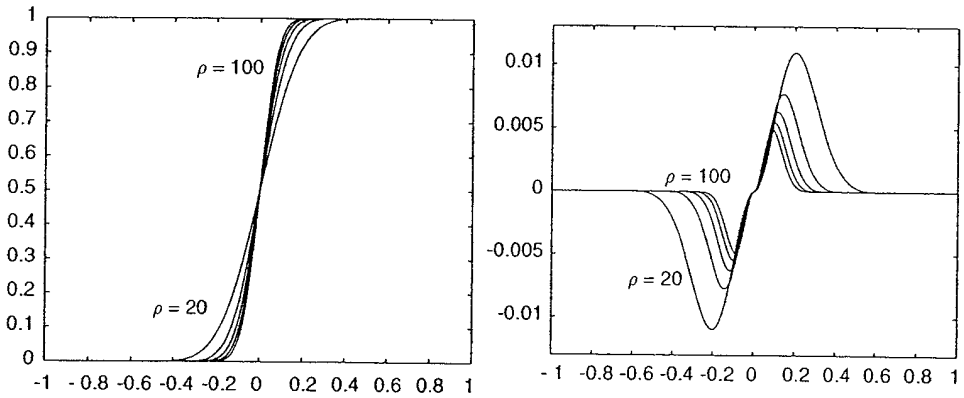


Fig. 4. — The real and imaginary parts of the leading term in (3.13) as functions of λ, $\theta = \lambda\pi/2$, $-1 \le \lambda \le 1$, for $\rho = 20(20)100$.

sense that truncating the remainder expansion at any fixed term yields, in combination with the (optimally) truncated expansion (3.9), an approximation of $E_p(z)$ whose relative error is exponentially small, uniformly in the domain indicated. In the same work the asymptotic nature of the expansion (3.13) is further analyzed, its validity extended to the sector $-2\pi + \delta \le \theta \le 2\pi - \delta$, and new, more convenient, formulae are given for the coefficients. (Note, however, that the notation in [108] differs somewhat from the notation in [107].)

Much of what is discussed in this § 3.3, and more, is nicely summarized by Olver in [110]. See also § 7.2 for repeated re-expansion of remainders.

4. Inverse functions and zeros

4.1. *Inverse functions*

For any $a > 0$, the functions

$$P(a,x) = \frac{\gamma(a,x)}{\Gamma(a)}, \quad Q(a,x) = \frac{\Gamma(a,x)}{\Gamma(a)}, \tag{4.1}$$

satisfying $P(a,x) + Q(a,x) = 1$, are cumulative probability functions on the interval $x \ge 0$. For example, $a = \frac{\nu}{2}$, $x = \frac{1}{2}\chi^2$, where $\nu \ge 1$ is an integer, yields the chi-square probability functions with ν degrees of freedom. For this reason, their inverse functions are important in statistical applications, where, given any p with $0 < p < 1$, or $q = 1 - p$, one is interested in determining x such that

$$P(a,x) = p, \quad \text{or (equivalently)}, \quad Q(a,x) = q\,. \tag{4.2}$$

In principle, this amounts to solving a nonlinear equation, for which many iterative methods are available such as Newton's method. In practice, however, these would require good initial approximations as well as repeated evaluations of the incomplete gamma function, both of which can render the inversion costly. It is desirable, therefore, to be able to solve the equations (4.2) more directly and economically.

The case $a = \frac{1}{2}$ of the error function (cf. (1.6)) is particularly simple, as the inverse error function is a function of a single real variable on [0,1] and hence accessible to approximation-theoretic methods. Thus, Strecok in [129] uses Chebyshev expansions in appropriate variables and ranges to obtain accuracies in the region $[0,1 - 10^{-300}]$ of at least 18 significant decimal digits, whereas Blair *et al.* [13] use rational approximations to obtain even higher accuracies of up to 23 digits on a larger domain, $[0,1 - 10^{-10000}]$. They also provide an asymptotic series approximation acurate to at least 25 digits for the remaining interval $[1 - 10^{-10000}, 1]$.

For general a, Temme [136] employs his uniform asymptotic expansion (3.4) to do the inversion. Thus, the second equation in (4.2), for example, in

combination with the first in (3.4), takes the form

$$\frac{1}{2}\operatorname{erfc}\left(\eta\sqrt{a/2}\right) + R_a(\eta) = q\,. \tag{4.3}$$

This is first solved for η, whereupon (3.5) is used to solve for λ, which by (3.3) finally yields $x = a\lambda$. To solve (4.3) for η, one can take as initial approximation η_0 the solution of

$$\frac{1}{2}\operatorname{erfc}\left(\eta_0\sqrt{a/2}\right) = q\,, \tag{4.4}$$

which is computable in terms of the inverse error function previously discussed. Then for η one seeks an asymptotic expansion of the form

$$\eta \sim \eta_0 + \frac{\varepsilon_1}{a} + \frac{\varepsilon_2}{a^2} + \frac{\varepsilon_3}{a^3} + \cdots\,, \quad a \to \infty\,. \tag{4.5}$$

The determination of the coefficients ε_i is laborious. It is shown in [136] that they are analytic functions of η_0 for $|\eta_0| < 2\sqrt{\pi}$ and therefore can be expanded in powers of η_0. For the first four coefficients ε_i, the power series are given in exact rational form up to 17, 11, 9, and 7 terms, respectively. For larger values of η_0, the same four coefficients are expressed algebraically in terms of η_0, $\varepsilon_1 = \eta_0^{-1}\ln f(\eta_0)$ and $f(\eta_0) = \frac{\eta_0}{\lambda_0 - 1}$, where λ_0 is the solution of (3.5) for $\eta = \eta_0$. The procedure is particularly effective for large values of a, but yields typically 3 to 4 correct decimal digits already for $a = 1$ or $a = 2$.

For the chi-square distribution an alternative asymptotic inversion is described in [41], which is valid for small q and a fixed.

4.2. *Real zeros*

The unique positive zero of $\operatorname{Ei}(x) = \frac{1}{2}[E_1(-x + i0) + E_1(-x - i0)]$ is given to 30 decimal places in [29, p. 300]. Asymptotic approximations to the positive zeros of the sine and cosine integrals can be found in [37] and [127], respectively.

Tricomi's interest, as noted in § 1, was in the zeros of $\gamma(a, x)$ considered as a function of x for fixed $a < 0$. Little (to the author's knowledge) has been done on this problem beyond Tricomi's work. Curiously, though, the negative zero $x_-(a)$ of $\gamma(a, x)$ for $-1 < a < 0$ has received some scrutiny in connection with a probability density (encountered by Mandelbrot) whose Fourier transform is $[\Gamma(1 + a)\gamma^*(a, -is)]^{-1}$. Lew [87] indeed shows that $x_-(a)$ decreases monotonically in $[-1, 0]$ (which can also be read off from the contour map of figure 1) and satisfies the inequalities

$$1 - \frac{1}{|a|} < x_-(a) < \ln|a|, \quad -1 < a < 0\,. \tag{4.6}$$

4.3. *Complex zeros*

The study of complex zeros becomes interesting already for some of the special cases (1.2)–(1.6) of the incomplete gamma function. Thus, e.g., the complex zeros of $e_n(nz)$ (cf. (1.3)) and their asymptotics as $n \to \infty$ have received a great

deal of attention; see, e.g., Varga's monograph [156, Ch. 4]. Asymptotic approximations to the zeros of the complex error function $w(z) = e^{-z^2}\operatorname{erfc}(-iz)$ and to those of $\operatorname{erf} z$ are given in [42], including tables[3] to 11 significant digits of the first 100 of them. For the complex zeros of the Fresnel integrals, see [80].

For an asymptotic analysis of the complex zeros of $\Gamma(a,z)$, Temme in [138] uses the same method as described previously in § 4.1, except that in (4.4) he puts $q = 0$ and takes for $\eta_0\sqrt{a/2}$ one of the complex zeros of the complementary error function. In particular, curves in the complex λ-plane are identified which are approached by the λ-zeros of $\Gamma(a,\lambda a)$ as $a \to \infty$ over positive values. A branch of this curve is the Szegö curve known from [156], which has been studied in connection with integer values of a.

In the work of Kölbig [76], [77], [78] the focus is on the complex zeros of $\gamma(a,x)$ considered as a function of a for fixed real x. From the contour map in figure 1 it is evident that the line $x = \text{const}$, for x suitably chosen, has two intersections with the zero level curve of γ^* in each of the intervals $-2m < a < 1-2m, m = 1,2,3,\ldots$ (visible in figure 1 explicitly for $m = 1$ and $m = 2$). These intersections move toward each other as x is increased and eventually coalesce. If the point of coalescence is denoted by (a_m^*, x_m^*), the double zero of γ^* (or γ) at this point will split into a pair of conjugate complex zeros upon further increase of x beyond x_m^*. Thus, for each $m = 1,2,3,\ldots$ there is a pair of conjugate complex trajectories in the complex a-plane emanating from a_m^* along which γ vanishes. Using Tricomi's result (2.16), Kölbig in [78] gives the approximations $a_m^* \sim 1 - 2m - .623021$ and $x_m^* \sim .556929m - .108906\ln m - .299840$ and in [76] he provides graphs and tables of the first five (eight in [77]) trajectories $a = a_m(x)$, $x \geq x_m^*$, in the upper half-plane. In [78] the concern is with the trajectories $a = a_m(x)/x$, i.e., the zero curves of $\gamma(xa,x)$ in the complex a-plane, and plots of the first eight of them are shown. As $m \to \infty$, according to a result of Mahler [96], they approach a limiting curve, which is also shown.

5. Inequalities and monotonicity

5.1. *Inequalities*

An early inequality of some generality for the incomplete gamma function is the author's inequality [47]

$$(5.1)\quad \frac{1}{2a}[(x+2)^a - x^a] < e^x\Gamma(a,x) \leq \frac{c_a}{a}[(x+c_a^{-1})^a - x^a],\ 0 \leq x < \infty,\ 0 < a < 1,$$

[3]The heading of Table 2 in [42] is incorrect; it should be «Zeros of $w(x)$» or «Zeros of $\operatorname{Erfc}(-iz)$».

where

$$c_a = [\Gamma(1+a)]^{\frac{1}{1-a}}.$$

For $a = \frac{1}{2}$, the second inequality reduces to one of Pollak [117], the first to one of Komatu [79], for «Mills' ratio» $e^{x^2}\int_x^\infty e^{-t^2}dt$. For sharper bounds regarding this ratio, see also [16], and for related inequalities, [82]. As $a \uparrow 1$, both bounds tend to 1, which is the value of $e^x\Gamma(1,x)$, since $\Gamma(1,x) = e^{-x}$. As $a \downarrow 0$, one obtains an inequality for the exponential integral,

$$\frac{1}{2}\ln\left(1+\frac{2}{x}\right) \le e^x E_1(x) \le \ln\left(1+\frac{1}{x}\right), \quad 0 < x < \infty, \tag{5.2}$$

which sharpens an inequality due to E. Hopf [63, p. 26]. Another special case of (5.1) obtains by setting $x = 0$ and using $\Gamma(1+a) \le 1$ on [0,1],

$$2^{a-1} \le \Gamma(1+a) \le 1, \quad 0 \le a \le 1.$$

This has been sharpened and generalized in [47] to[4] (ψ is the logarithmic derivative of the gamma function)

$$x^{1-a} \le \frac{\Gamma(x+1)}{\Gamma(x+a)} \le \exp[(1-a)\psi(x+1)], \quad 0 \le a \le 1, \quad x > 0, \tag{5.3}$$

which in turn has been the subject of numerous improvements and extensions; see, e.g., [39], [155], [124], [64], [84], [74], [91], [81], [82], [142], [98, §§ 2,3], [116, § 3], [112], [51]. Further inequalities for the gamma function can be found in [101, § 3.6], [72], [20], [71], [94], [73], [126], [123], [69], [119], [36], [83], [93], [38, § 3], [54], [55].

An alternative inequality for the incomplete gamma function was recently obtained by Alzer [4], who proved

$$(1-e^{-s_a x})^a < \frac{\gamma(a,x)}{\Gamma(a)} < (1-e^{-r_a x})^a, \quad 0 \le x < \infty, \quad a > 0, \quad a \ne 1, \tag{5.4}$$

where

$$r_a = \begin{cases} [\Gamma(1+a)]^{-1/a} & \text{if } 0 < a < 1, \\ 1 & \text{if } a > 1, \end{cases} \qquad s_a = \begin{cases} 1 & \text{if } 0 < a < 1, \\ [\Gamma(1+a)]^{-1/a} & \text{if } a > 1. \end{cases}$$

For $a = \frac{1}{2}$, this reduces to inequalities of Chu [28] for the error function $\operatorname{erf} x$. As $a \to 1$, both bounds tend to $1 - e^{-x}$, which is the value of $\gamma(1,x)$. Rewriting (5.4) in terms of $\frac{\Gamma(a,x)}{\Gamma(a)} = 1 - \frac{\gamma(a,x)}{\Gamma(a)}$, and letting $a \downarrow 0$, yields a new inequality for the exponential integral,

$$-\ln(1-e^{-cx}) \le E_1(x) \le -\ln(1-e^{-x}), \quad 0 < x < \infty, \tag{5.5}$$

[4] Actually, (5.3) was proved in [47] only for x an integer $n = 1, 2, 3, \ldots$, but the proof given is valid for arbitrary $x > 0$ (cf. *Math. Reviews 21*, Review 2067).

where c is expressible in terms of Euler's constant γ as $c = e^{\gamma} = 1.7810724\ldots$. In the domain $0 < a < 1$ the inequalities for $\Gamma(a,x)$ derivable from (5.4) are sharper than those in (5.1) if x is small, but weaker if x is large. The right inequality in (5.5) is always weaker than the corresponding inequality in (5.2), whereas the left inequality is sharper for small x and weaker for large x.

Upper bounds for $\Gamma(a,x)$ in the domain $x \geq 0$, $a \geq 1$ are also derived in [14]. The rather special, but pretty, sequence of inequalities,

$$\frac{\Gamma(n,n)}{\Gamma(n)} < \frac{1}{2} < \frac{\Gamma(n,n-1)}{\Gamma(n)}, \quad n = 1,2,3,\ldots, \tag{5.6}$$

is proved in [157] and attributed to G. Lochs.

5.2. *Monotonicity*

Monotonicity, convexity, and higher monotonicity results abound for the gamma function, but seem to be scarce for incomplete gamma functions. Absolute monotonicity, i.e., positivity of all derivatives, has been shown in [35, § 4b] for the sum of squares of «Hermite functions», which are expressible in terms of confluent hypergeometric functions. Lorch [90] has monotonicity results for ratios of Whittaker functions. Convexity and logarithmic convexity of $\Gamma(a+1,a)/\Gamma(a+1)$ on $[0,\infty)$ are shown by Temme [133]. The fact that this function decreases monotonically from 1 to $\frac{1}{2}$ has been shown previously by Van De Lune [95]. According to the well-known Bohr-Mollerup-Artin theorem, logarithmic convexity, on the other hand, lies at the heart of the gamma function, as, together with the difference equation and normalization, it characterizes the gamma function uniquely (cf. Artin [9]).

A function f is said to be completely monotonic on an interval I if it has derivatives of all orders in I and $(-1)^n f^{(n)}(x) \geq 0$ on I for $n = 0,1,2,\ldots$. It is called strictly completely monotonic if strict inequality holds for each n. Remarkably, many functions involving the gamma and/or the psi function are completely monotonic. Bustoz and Ismail [18], for example, prove this for the functions

$$\left(1-\frac{1}{2x}\right)^{-1/2} \frac{\Gamma^2\left(x+\frac{1}{2}\right)}{\Gamma(x)\Gamma(x+1)} \quad \text{and} \quad \left(1+\frac{1}{2x}\right)^{-1/2} \frac{\Gamma(x)\Gamma(x+1)}{\Gamma^2\left(x+\frac{1}{2}\right)} \tag{5.7}$$

on the interval $(\frac{1}{2},\infty)$ and $(0,\infty)$, respectively. Likewise, they show that

$$\frac{\Gamma(x+s)}{\Gamma(x+1)} \exp\left[(1-s)\psi\left(x+\frac{1}{2}(s+1)\right)\right] \quad \text{and} \quad \frac{\Gamma(x+1)\left(x+\frac{1}{2}s\right)^{s-1}}{\Gamma(x+s)}, \quad 0 \leq s \leq 1, \tag{5.8}$$

are completely monotonic on $(0,\infty)$, and strictly so if $0 < s < 1$. Furthermore,

$$\frac{\Gamma(x+1)}{\Gamma(x+s)}\exp[(s-1)\psi(x+\sqrt{s})]$$

$$\text{and}\quad \frac{\Gamma(x+s)\left[x-\frac{1}{2}+\sqrt{s+\frac{1}{4}}\right]^{1-s}}{\Gamma(x+1)}, \quad 0<s<1, \tag{5.9}$$

are strictly decreasing on $(0,\infty)$, which, together with (5.8), generalizes inequalities of Kershaw [74]. Other examples are given in [66] and [2]. Far-reaching results are proved by Alzer in [3]. Thus, for example, $x[\ln x - \psi(x)]$ is shown to be strictly completely monotonic on $(0,\infty)$, which extends a monotonicity and convexity result of Anderson *et al.* [8, § 3]. The convexity on $(0,\infty)$ of $x\psi(x)$, proved by the author [48], is generalized to

$$0 < (-1)^n x^{n-1}[x\psi(x)]^{(n)} < (n-2)!, \quad x>0, \quad n \geq 2 \tag{5.10}$$

([3, Thm. 4]). All remainders in the asymptotic expansion of $\ln\Gamma(x)$ for $x \to \infty$ are completely monotonic. More precisely, if

$$R_n(x) = \ln\Gamma(x) - \left(x-\frac{1}{2}\right)\ln x + x - \frac{1}{2}\ln(2\pi) - \sum_{k=1}^{n}\frac{B_{2k}}{2k(2k-1)x^{2k-1}}, \tag{5.11}$$
$$n = 0,1,2\ldots,$$

where B_{2k} are the Bernoulli numbers, then $(-1)^n R_n(x)$ is completely monotonic on $(0,\infty)$ ([3, Thm. 8]). This was proved earlier by Muldoon [102] for $n = 0$, whereas convexity and concavity for general n were shown by Merkle [98]. Another remarkable result is the complete monotonicity on $(0,\infty)$ of

$$\prod_{\nu=1}^{n}\frac{\Gamma(x+a_\nu)}{\Gamma(x+b_\nu)} \tag{5.12}$$

([3, Thm. 10]), provided

$$0 \leq a_1 \leq a_2 \leq \cdots \leq a_n, \quad 0 \leq b_1 \leq b_2 \leq \cdots \leq b_n,$$
$$\sum_{\nu=1}^{\mu} a_\nu \leq \sum_{\nu=1}^{\mu} b_\nu \quad \text{for } \mu = 1,2,\ldots,n.$$

This generalizes a result of Bustoz and Ismail [18, Thm. 6] for $n = 2$ and monotonicity results of Stolarsky [128, § 8] and Maligranda *et al.* [97, Thm. 2].

A monotonicity result of Kershaw and Laforgia [75], according to which on $(0,\infty)$ the function $\left[\Gamma\left(1+\frac{1}{x}\right)\right]^x$ decreases, and $x\left[\Gamma\left(1+\frac{1}{x}\right)\right]^x$ increases, extends an earlier inequality of Minc and Sathre [100]. See also [116, § 5] for additional monotonicity results of this kind. Logarithmic convexity on $\mathbb{R}_+$ of $\Gamma(2x)/x\Gamma^2(x)$ and logarithmic concavity of $\Gamma(2x)/\Gamma^2(x)$ are proved by Merkle [99].

The q-analogue of the gamma function also enjoys inequalities and higher monotonicity properties, many of which extend those in this § 5.2. For a good account of this, the reader is referred to Ismail and Muldoon [65].

6. Numerical methods

6.1. *General procedures*

As with other special functions, numerical methods for computing incomplete gamma functions rely on a variety of standard tools. Thus, asymptotics is used by Takenaga [130] to evaluate $\Gamma(a,x)$ for large a. In a series of papers, Chiccoli *et al.* [23], [24], [25], [26], [27] use asymptotic approximations, Taylor and other series expansions (including Tricomi's series (2.8)), and recurrence relations, to evaluate the generalized exponential integral $E_p(x)$ for arbitrary positive p and x. A combination of forward and backward recurrence is the principal tool in the work of Amos [5], [7] to compute the exponential integrals for integer values $p = n$ of p and positive, resp. complex x. Difficulties near the negative real axis are overcome by an analytic continuation scheme. Allasia and Besenghi [1] propose quadrature methods, in particular the composite trapezoidal rule, to evaluate $\Gamma(a,x)$ for $a < -1, x > 0$ and provide detailed error analyses. The use of (unstable) forward recurrence to compute the «molecular integrals» $\{\gamma(n+1,x)\}$ is analyzed in [88]. A fairly comprehensive procedure for evaluating incomplete gamma functions in the domain $-\infty < a < \infty, x \geq 0$ is described in [49]. If there is a weakness in this procedure, it is the fact of becoming computer-intensive when a and x are both very large and almost equal. This, however, has been corrected by DiDonato and Morris [30], who use, among other things, Temme's uniform asymptotic expansion (cf. § 3.2) to compute $\gamma(a,x)/\Gamma(a)$ and $\Gamma(a,x)/\Gamma(a)$ for $a > 0$, $x \geq 0$, and by Temme himself [135], who uses (3.4) in the critical region, with the asymptotic series (in a) for $R_a(\eta)$ replaced by a more manageable Taylor series (in η). DiDonato and Morris also describe an inversion procedure which uses a third-order iterative method along with an elaborate scheme of computing a good initial value. Expansion in Chebyshev polynomials is used by Barakat [10] to compute $\gamma(a,z)$ for real a and purely imaginary z. Techniques based on continued fractions are employed by Jones and Thron [68] and Jacobsen *et al.* [67], and still other, especially asymptotic, techniques by Temme [137], to compute $\gamma(a,z)$ and $\Gamma(a,z)$ for complex a and z. There is an extensive literature dealing with the computation of special univariate cases of the incomplete gamma function, such as the exponential integral $E_1(x)$ and the error function and their close relatives, both for real and complex arguments. For this, as well as for relevant software, including software for incomplete gamma functions, we refer to the comprehensive documentation in [92]. Here we concentrate on real parameters and the stable use of recurrence relations.

6.2. *Recurrence relations*

The recurrence relations satisfied by incomplete gamma functions are linear, inhomogeneous, first-order difference equations of the form

$$y_n = a_n y_{n-1} + b_n, \quad n = 1, 2, 3, \dots, \quad a_n \neq 0, \tag{6.1}$$

where the coefficients a_n, b_n depend on x and/or a. Given y_0, the recurrence (6.1) defines uniquely the sequence $\{y_n\}_{n=0}^{\infty}$. It is important, however, to know how robust the recurrence is to small perturbations such as rounding errors. An informative answer to this is provided by the «amplification factors» $\omega_{s\to t}$, which determine the effect of a small relative error ε at $n = s$ («s» for starting) upon the value at $n = t$ («t» for «terminal»), assuming exact arithmetic. Thus, if the desired solution of (6.1) is $\{f_n\}$, and if $y_s = f_s(1+\varepsilon)$, then $y_t = f_t(1 + \omega_{s\to t} \cdot \varepsilon)$ in exact arithmetic. Here s may be less than t, which is the case in forward recursion, or $s > t$, in which case (6.1) is applied in reverse order (computing y_{n-1} in terms of y_n). An easy calculation (cf. [50]) will show that

$$\omega_{s\to t} = \frac{\rho_t}{\rho_s}, \tag{6.2}$$

where

$$\rho_n = \frac{f_0 h_n}{f_n}, \quad h_n = a_n a_{n-1} \cdots a_0 \quad (a_0 = 1) \tag{6.3}$$

(assuming $f_0 \neq 0$). Here, h_n is a solution of the homogeneous equation (6.1) (with all $b_n = 0$). Knowledge of the quantities $\{\rho_n\}$ is thus sufficient to determine all amplification factors in (6.2). Note that $\rho_n = \omega_{0\to n}$.

A first example is $\gamma^*(a,x)$, which satisfies the recurrence relation

$$\gamma^*(a+1,x) = \frac{1}{x}\left[\gamma^*(a,x) - \frac{e^{-x}}{\Gamma(a+1)}\right]. \tag{6.4}$$

Once we know $\gamma^*(a,x)$ for $0 \leq a < 1$, repeated application of this relation allows us to obtain $\gamma^*(a,x)$ for any $a \geq 1$, and also for any $a < 0$ if we apply (6.4) in the reverse order. Consider first the case of positive parameters,

$$\gamma_n^* = \gamma^*(a+n,x), \quad n = 0, 1, 2, \dots, \quad 0 \leq a < 1. \tag{6.5}$$

Then (6.4) yields

$$\gamma_n^* = \frac{1}{x}\left[\gamma_{n-1}^* - \frac{e^{-x}}{\Gamma(a+n)}\right], \quad n = 1, 2, \dots; \quad \gamma_0^* = \gamma^*(a,x), \tag{6.6}$$

a relation of the form (6.1). Since here $h_n = x^{-n}$, we get

$$\rho_n = \frac{\gamma^*(a,x)}{\gamma^*(a+n,x)x^n}, \quad n = 0, 1, 2, \dots. \tag{6.7}$$

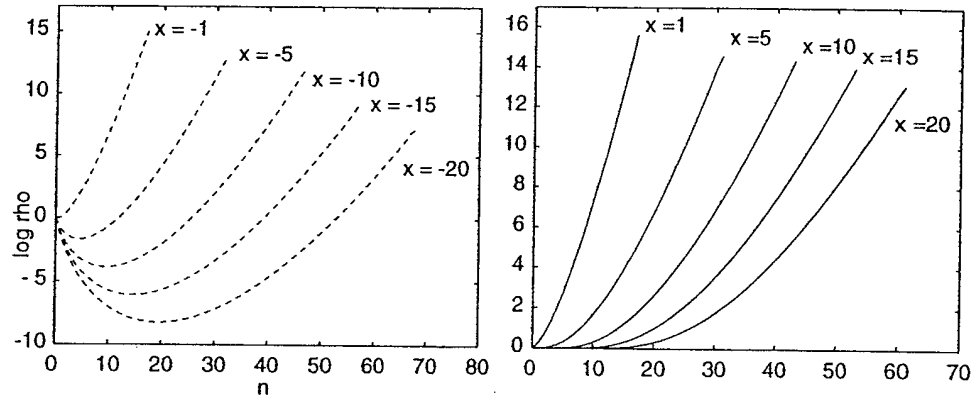


Fig. 5. — Amplification factors $|\omega_{0\to n}|$ for the incomplete gamma function $\{\gamma^*(\frac{1}{2}+n,x)\}$.

The behavior of the corresponding amplification factors $|\omega_{0\to n}| = |\rho_n|$ is similar for all values of a in $[0,1]$; figure 5 shows them (on a logarithmic scale) for $a = \frac{1}{2}$ and for selected negative values of x on the left, and positive values of x on the right. (The case $x = 0$ is uninteresting, since $\gamma^*(a,0) = 1/\Gamma(a+1)$.) In either case, $|\rho_n| \to \infty$ as $n \to \infty$, but in the first case there is a significant downward dip to a minimum at about $n = |x|$ before $|\rho_n|$ grows monotonically to ∞, whereas in the second case $|\rho_n|$ increases monotonically from the start. This has important computational implications. The fact that $|\rho_n| \to \infty$ by (6.2) indeed implies that $|\omega_{s\to t}|$ for t fixed becomes arbitrarily small as $s \to \infty$. In effect, this means that backward recurrence in (6.6) from some large $n = \nu$ down to any fixed n produces arbitrarily accurate results if ν is chosen large enough, regardless of the choice of starting value. The latter, therefore, may conveniently be taken to be zero. This procedure is particularly effective for positive x, because of the monotonicity of $|\rho_n|$. When $x < 0$, backward recurrence should not proceed below $n \approx |x|$, since otherwise, by the nature of the dashed curves in figure 5, one would run into a regime of significant error amplification. The values of γ_n^* for n smaller than $|x|$ therefore must be generated by forward recurrence.

The case of negative parameters can be expected to be more complicated since we are getting into regions containing zero curves (cf. figure 1 and § 2.4). Here the recursion for $\gamma_n^* = \gamma^*(a-n,x)$, $0 \le a < 1$, is

$$\gamma_n^* = x\gamma_{n-1}^* + \frac{e^{-x}}{\Gamma(a-n+1)}, \quad n = 1,2\ldots; \quad \gamma_0^* = \gamma^*(a,x), \tag{6.8}$$

where the second term on the right is to be replaced by zero if $a = 0$. Limited exploration suggests that the amplification factors $|\rho_n| = |\omega_{0\to n}|$, when $x > 0$, are of the order of magnitude 1 for a while before decreasing rapidly, while

for $x < 0$ they also eventually decrease at a similar speed, but may assume relatively large values (especially if $|x|$ is large) before they do so. Nevertheless, the recurrence (6.8), overall, is reasonably stable in forward direction.

What has been said about $\gamma^*(a,x)$ holds also for $\gamma(a,x)$, since the quantities ρ_n in (6.3) are invariant with respect to any scaling transformation $y_n \mapsto c_n y_n, c_n \neq 0$.

A second example is the complementary incomplete gamma function $\Gamma(a,x)$, for which the recurrence relation reads

$$\Gamma(a+1,x) = a\Gamma(a,x) + x^a e^{-x}. \tag{6.9}$$

Its use for generating $\Gamma_n^+ = \Gamma(a+n,x)$ and $\Gamma_n^- = \Gamma(a-n,x), n = 0,1,2,\ldots,$ $0 \leq a < 1$, can be discussed along lines similar to the preceding, except that x is restricted to positive values. One finds that the respective amplification factors $|\rho_n^+|$ and $|\rho_n^-|$ behave much like those in figure 5, but upside-down. That is, $|\rho_n^+| = \rho_n^+$ decreases monotonically whereas $|\rho_n^-|$ initially increases to a maximum near $n = |x|$ before decreasing to zero, the maximum being larger the larger $|x|$. The monotonicity of $\rho_n^+ = \frac{\Gamma(a,x)}{\Gamma(a)} / \frac{\Gamma(a+n,x)}{\Gamma(a+n)}$ follows from the monotonicity of $\Gamma(a,x)/\Gamma(a)$ as a function of a, proved by Tricomi (cf. § 2.5). This means that the recurrence for Γ_n^+ is perfectly stable in forward direction, whereas the one for Γ_n^- should be started at a value of n near $|x|$, with backward [forward] recurrence being applied for the smaller [larger] values of n. The starting value can be computed by a continued fraction, for example. Note that Γ_n^- is related to the generalized exponential integral by $\Gamma_n^- = x^{n-a} E_{1-a+n}(x)$ (cf. equation (3.6)).

7. Applications

Many of the special cases of the incomplete gamma function are widely used in the applied sciences. Thus, the exponential integrals $E_p(x)$ for $p > 0$ play a significant role in transport theory and fluid flow, and for negative integer values of p furnish basic auxiliary functions in molecular physics. The error functions are frequently encountered in heat conduction, and the Fresnel integrals in Fresnel diffraction, problems. The complex error function $e^{-z^2}\,\mathrm{erfc}(-iz)$ is important in plasma wave problems, where it is known as «plasma dispersion function», in astrophysics and Lorentz/Doppler line broadening, where the real and imaginary parts go under the name of «Voigt functions», and in the design of particle accelerators. Finally, the incomplete gamma function ratios and their special cases are used extensively in statistical applications. Rather than reviewing these «external» applications (a nearly impossible task), we limit ourselves to a few recent «internal» applications that we happen to be familiar with, i.e., applications within the theory of special functions.

7.1. *Expansions in incomplete gamma functions*

7.1.1. *The Riemann zeta function on the critical line*

Efforts to verify the Riemann Hypothesis, according to which all zeros of $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$ in Re $s > 0$ lie on the critical line Re $s = \frac{1}{2}$, require high-precision calculation of the zeta function for $s = \frac{1}{2} + it$ and t very large. Presently, the most efficient methods are based on the Riemann-Siegel formula and some of its recent improvements; see, e.g., Odlyzko [104] and Berry [12]. A promising alternative method has been developed by Paris [113] and Paris and Cang [115], who use an expansion of the zeta function in incomplete gamma functions in combination with (essentially) Temme's uniform asymptotic expansion (cf. § 3.2). Once a reliable estimate of the truncation error becomes available, the expansion could provide a useful tool for the rigorous verification of the Riemann Hypothesis.

For an expansion in incomplete gamma functions of more general Dirichlet series, see also [57, p. 106].

7.1.2. *A generalization of the incomplete gamma function*

The following generalization of the incomplete gamma function,

$$\Gamma(a,x;b) = \int_x^{\infty} t^{a-1}e^{-t-b/t}dt, \quad x>0, \quad a>0, \quad b\geq 0\,, \tag{7.1}$$

has been studied in [21]. By expanding $e^{-b/t}$ into a Taylor series in t^{-1} one obtains an expansion in incomplete gamma functions [21], [22],

$$\Gamma(a,x;b) = \sum_{n=0}^{\infty} \frac{(-b)^n}{n!}\Gamma(a-n,x)\,. \tag{7.2}$$

It is just the Maclaurin expansion of $\Gamma(\cdot,\cdot;b)$, an entire function of b. When $a>0$ and $b\geq 0$ are restricted to bounded intervals, then (7.2) can also be viewed as an asymptotic expansion for $x\to\infty$. The incomplete gamma functions $\Gamma(a-n,x)$ required in (7.2) can be generated recursively as discussed in § 6.2. There is also an asymptotic expansion of $\Gamma(a,x;b)$ for large a analogous to (3.4), involving the complementary error function [22, equation (5.2)].

7.1.3. *Fermi-Dirac integrals*

The Fermi-Dirac integral

$$F_{p-1}(x) = \frac{1}{\Gamma(p)}\int_0^{\infty} \frac{t^{p-1}}{1+e^{t-x}}dt, \quad p>0, \quad x\in\mathbb{R}\,, \tag{7.3}$$

for negative values of x is easily evaluated by the series

$$F_{p-1}(x) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}e^{xn}}{n^p}, \quad x<0\,.$$

More difficult is the case of (large) positive x and p. Writing $p = ax$ and assuming $N - 1 < a < N$ for some integer $N \geq 1$, Temme and Olde Daalhuis [141], improving on previous work of Schell [122], obtain the representation

$$F_{p-1} = \sum_{n=1}^{N-1} \frac{(-1)^{n-1} e^{xn}}{n^p} + G_{p-1}(x) + H_{p-1}(x)\,, \tag{7.4}$$

where the terms in the sum on the right (if $N \geq 3$) decrease monotonically. By contour integration in the complex plane, the function $G_{p-1}(x)$ is expressible as the term with $n = N$ in the summation of (7.4) multiplied by an incomplete gamma function ratio,

$$G_{p-1}(x) = \frac{(-1)^{N-1} e^{xN}}{N^p} \cdot \frac{\Gamma(p, Nx)}{\Gamma(p)}\,. \tag{7.5}$$

Here Temme's uniform asymptotic expansion for $p \to \infty$ (cf. § 3.2) is applicable, also when $a = N$. The last term, $H_{p-1}(x)$, in (7.4) can be formally expanded in descending powers of x. Both terms G_{p-1}, H_{p-1} are negligible when N is large.

7.2. *Hyperasymptotics*

A process of successive re-expansion of remainder terms in asymptotic expansions, called hyperasymptotics, is developed in [109] for solutions of the confluent hypergeometric equation, and in [105], [106] for solutions of more general linear homogeneous second-order differential equations having an irregular singularity of rank one at infinity. By truncating the classical Poincaré expansion after a judiciously selected number of terms, one re-expands the corresponding remainder term to obtain a «first-level» expansion, the Poincaré expansion being at level zero. This first-level expansion is a series in generalized exponential integrals (cf. § 3.3). If that series in turn is judiciously truncated, its remainder term is re-expanded to obtain a «second-level» expansion; it proceeds in functions called «hyperterminants», which are repeated infinite integrals of the generalized exponential integral. The process can be repeated indefinitely. An important feature of this sequence of re-expansions is that at each step the error is reduced by an exponentially small factor, of which the «exponential improvement» of the first-level expansion, mentioned at the end of § 3.3.2, is just one example.

ACKNOWLEDGMENTS – The author gratefully acknowledges comments on a preliminary version of this paper received from Richard Askey, Mourad Ismail, Kurt Kölbig, Andrew Odlyzko, Frank Olver, and Nico Temme.

References

[1] G. Allasia - R. Besenghi, *Numerical calculation of incomplete gamma functions by the trapezoidal rule*, Numer. Math., 50, 1987, 419-428.

[2] H. Alzer, *Some gamma function inequalities*, Math. Comp., 60, 1993, 337-346.

[3] H. Alzer, *On some inequalities for the gamma and psi functions*, Math. Comp., 66, 1997, 373-389.

[4] H. Alzer, *On some inequalities for the incomplete gamma function*, Math. Comp., 66, 1997, 771-778.

[5] D.E. Amos, *Computation of exponential integrals*, ACM Trans. Math. Software, 6, 1980, 365-377.

[6] D.E. Amos, *Uniform asymptotic expansions for exponential integrals* $E_n(x)$ *and Bickley functions* $\mathrm{Ki}_n(x)$, ACM Trans. Math. Software, 9, 1983, 467-479.

[7] D.E. Amos, *Computation of exponential integrals of a complex argument*, ACM Trans. Math. Software, 16, 1990, 169-177.

[8] G.D. Anderson - R.W. Barnard - K.C. Richards - M.K. Vamanamurthy - M. Vuorinen, *Inequalities for zero-balanced hypergeometric functions*, Trans. Amer. Math. Soc., 347, 1995, 1713-1723.

[9] E. Artin, *Einführung in die Theorie der Gammafunktion*, Hamburger Mathematische Einzelschriften, 11. Heft, Teubner, Leipzig 1931. (English transl. by M. Butler, Athena Series: Selected Topics in Mathematics, Holt, Rinehart, and Winston, New York 1964).

[10] R. Barakat, *Evaluation of the incomplete gamma function of imaginary argument by Chebyshev polynomials*, Math. Comp., 15, 1961, 7-11.

[11] M.V. Berry, *Uniform asymptotic smoothing of Stokes's discontinuities*, Proc. Roy. Soc. London, Ser. A, 422, 1989, 7-21.

[12] M.V. Berry, *The Riemann-Siegel expansion for the zeta function: high orders and remainders*, Proc. Roy. Soc. London, Ser. A, 450, 1995, 439-462.

[13] J.M. Blair - C.A. Edwards - J.H. Johnson, *Rational Chebyshev approximations for the inverse of the error function*, Math. Comp., 30, 1976, 827-830.

[14] G. Boese, *Eine majorante asymptotische Abschätzung für die unvollständige Gammafunktion*, Z. Angew. Math. Mech., 52, 1972, 552-553.

[15] P.E. Böhmer, *Differenzengleichungen und bestimmte Integrale*, Koehler, Leipzig 1939.

[16] A.V. Boyd, *Inequalities for Mills' ratio*, Rep. Statist. Appl. Res. Un. Jap. Sci. Engrs., 6, 1959, 44-46.

[17] H. Buchholz, *Die konfluente hypergeometrische Funktion mit besonderer Berücksichtigung ihrer Anwendungen*, Ergebnisse der angewandten Mathematik, Bd. 2, Springer, Berlin 1953. (English transl. by H. Lichtblau and K. Wetzel, Springer Tracts in Natural Philosophy, 15, Springer, New York 1959).

[18] J. Bustoz - M.E.H. Ismail, *On gamma function inequalities*, Math. Comp., 47, 1986, 659-667.

[19] L. Carlitz, *On some polynomials of Tricomi*, Boll. Un. Mat. Ital., (3), 13, 1958, 58-64.

[20] I.V. Čebaevskaja, *Extension of the limits of applicability of certain inequalities for gamma functions* (Russian), Moskov. Gos. Ped. Inst. Učen. Zap., 460, 1972, 38-42.

[21] M.A. Chaudhry - S.M. Zubair, *Generalized incomplete gamma functions with applications*, J. Comput. Appl. Math., 55, 1994, 99-124.

[22] M.A. Chaudhry - N.M. Temme - E.J.M. Veling, *Asymptotics and closed form of a generalized incomplete gamma function*, J. Comput. Appl. Math., 67, 1996, 371-379.

[23] C. CHICCOLI - S. LORENZUTTA - G. MAINO, *A numerical method for generalized exponential integrals*, Comput. Math. Appl., 14, 1987, 261-268.

[24] C. CHICCOLI - S. LORENZUTTA - G. MAINO, *On the evaluation of generalized exponential integrals $E_\nu(x)$*, J. Comput. Phys., 78, 1988, 278-287.

[25] C. CHICCOLI - S. LORENZUTTA - G. MAINO, *Calculation of exponential integrals of real order*, Internat. J. Comput. Math., 31, 1989, 125-135.

[26] C. CHICCOLI - S. LORENZUTTA - G. MAINO, *A note on a Tricomi expansion for the generalized exponential integral and related functions*, Nuovo Cimento, B (11), 103, 1989, 563-568.

[27] C. CHICCOLI - S. LORENZUTTA - G. MAINO, *On a Tricomi series representation for the generalized exponential integral*, Internat. J. Comput. Math., 31, 1990, 257-262.

[28] J.T. CHU, *On bounds for the normal integral*, Biometrika, 42, 1955, 263-265.

[29] W.J. CODY - H.C. THACHER JR, *Chebyshev approximations for the exponential integral* $\mathrm{Ei}(x)$, Math. Comp., 23, 1969, 289-303.

[30] A.R. DI DONATO - A.H. MORRIS JR, *Computation of the incomplete gamma function ratios and their inverse*, ACM Trans. Math. Software, 12, 1986, 377-393.

[31] T.M. DUNSTER, *Asymptotic solutions of second-order linear differential equations having almost coalescent turning points, with an application to the incomplete gamma function*, Proc. Roy. Soc. London, Ser. A, 452, 1996, 1331-1349.

[32] T.M. DUNSTER, *Asymptotics of the generalized exponential integral, and error bounds in the uniform asymptotic smoothing of its Stokes discontinuities*, Proc. Roy. Soc. London, Ser. A, 452, 1996, 1351-1367.

[33] T.M. DUNSTER, *Error analysis in a uniform asymptotic expansion for the generalised exponential integral*, J. Comput. Appl. Math., 80, 1997, 127-161.

[34] T.M. DUNSTER - R.B. PARIS - S. CANG, *On the high-order coefficients in the uniform asymptotic expansion for the incomplete gamma function*, Submitted for publication.

[35] L. DURAND, *Nicholson-type integrals for products of Gegenbauer functions and related topics*, in: *Theory and application of special functions* (R.A. Askey, ed.), Academic Press, New York 1975, 353-374.

[36] J. DUTKA, *On some gamma function inequalities*, SIAM J. Math. Anal., 16, 1985, 180-185.

[37] V.K. DZJADYK - A.I. STEPANEC, *The sequence of zeros of the integral sine* (Russian), in: *Metric questions of the theory of functions and mappings*, No. 2 (Russian), Izdat. «Naukova Dumka», Kiev 1971, 64-73.

[38] Á. ELBERT - A. LAFORGIA, *An inequality for Legendre polynomials*, J. Math. Phys., 35, 1994, 1348-1360.

[39] T. ERBER, *The gamma function inequalities of Gurland and Gautschi*, Skand. Aktuarietidskr., 43, 1960, 27-28.

[40] A. ERDÉLYI, ed., *Higher transcendental functions*, Vol. II. Based, in part, on notes left by H. Bateman and compiled by the staff of the Bateman manuscript project, McGraw-Hill, New York 1953.

[41] H.E. FETTIS, *An asymptotic expansion for the upper percentage points of the χ^2-distribution*, Math. Comp., 33, 1979, 1059-1064.

[42] H.E. FETTIS - J.C. CASLIN - K.R. CRAMER, *Complex zeros of the error function and of the complementary error function*, Math. Comp., 27, 1973, 401-407.

[43] P. FRANKLIN, *Calculation of the complex zeros of the function $P(z)$ complementary to the incomplete gamma function*, Ann. of Math., (2), 21, 1919, 61-63.

[44] C.L. FRENZEN, *Error bounds for asymptotic expansions of the ratio of two gamma functions*, SIAM J. Math. Anal., 18, 1987, 890-896.

[45] C.L. FRENZEN, *Error bounds for the asymptotic expansion of the ratio of two gamma functions with complex argument*, SIAM J. Math. Anal., 23, 1992, 505-511.

[46] W. GAUTSCHI, *Exponential integral* $\int_1^\infty e^{-xt}t^{-n}dt$ *for large values of* n, J. Res. Nat. Bur. Standards, 62, 1959, 123-125.

[47] W. GAUTSCHI, *Some elementary inequalities relating to the gamma and incomplete gamma function*, J. Math. and Phys., 38, 1959, 77-81.

[48] W. GAUTSCHI, *Some mean value inequalities for the gamma function*, SIAM J. Math. Anal., 5, 1974, 282-292.

[49] W. GAUTSCHI, *A computational procedure for incomplete gamma functions*, ACM Trans. Math. Software, 5, 1979, 466-481.

[50] W. GAUTSCHI, *The computation of special functions by linear difference equations*, in: *Advances in difference equations* (S. Elaydi, I. Györi, and G. Ladas, eds.), Gordon and Breach, Amsterdam 1997, 213-243.

[51] C. GIORDANO - A. LAFORGIA - J. PEČARIĆ, *Unified treatment of Gautschi-Kershaw type inequalities for the gamma function*, Proc. VIII Sympos. Orthogonal Polynomials, to appear.

[52] W.M.Y. GOH - J. WIMP, *On the asymptotics of the Tricomi-Carlitz polynomials and their zero distribution I*, SIAM J. Math. Anal., 25, 1994, 420-428.

[53] W.M.Y. GOH - J. WIMP, *The zero distribution of the Tricomi-Carlitz polynomials*, Comput. Math. Appl., 33, 1997, 119-127.

[54] L. GORDON, *The stochastic approach to the gamma function*, Amer. Math. Monthly, 101, 1994, 858-865.

[55] P.J. GRABNER - R.F. TICHY - U.T. ZIMMERMANN, *Inequalities for the gamma function with applications to permanents*, Discrete Math., 154, 1996, 53-62.

[56] T.H. GRONWALL, *Sur les zéros des fonctions* $P(z)$ *et* $Q(z)$ *associées à la fonction gamma*, Ann. École Norm. Sup., (3), 33, 1916, 381-393.

[57] A. GUTHMANN, *Asymptotische Entwicklungen für unvollständige Gammafunktionen*, Forum Math., 3, 1991, 105-141.

[58] P.I. HADŽI, *The probability function: Integrals, series, and some generalizations* (Russian), Akad. Nauk Moldavsk. SSR, Kišinev 1971.

[59] C.N. HASKINS, *On the zeros of the function,* $P(z)$*, complementary to the incomplete gamma function*, Trans. Amer. Math. Soc., 16, 1915, 405-412.

[60] E. HILLE - G. RASCH, *Über die Nullstellen der unvollständigen Gammafunktion* $P(z,\rho)$*. II. Geometrisches über die Nullstellen*, Math. Z., 29, 1929, 319-334.

[61] C. HOOLEY, *On the representations of a number as the sum of two cubes*, Math. Z., 82, 1963, 259-266.

[62] C. HOOLEY, *On another sieve method and the numbers that are a sum of the hth powers*, Proc. London Math. Soc., (3), 43, 1981, 73-109.

[63] E. HOPF, *Mathematical problems of radiative equilibrium*, Cambridge Tracts in Math. and Math. Phys., no. 31, Cambridge University Press, 1934.

[64] C.O. IMORU, *A note on an inequality for the gamma function*, Internat. J. Math. Math. Sci., 1, 1978, 227-233.

[65] M.E.H. ISMAIL - M.E. MULDOON, *Inequalities and monotonicity properties for gamma and q-gamma functions*, in: *Approximation and computation: A festschrift in honor of Walter Gautschi* (R.V.M. Zahar, ed.), Internat. Ser. Numer. Math., 119, Birkhäuser, Boston, MA, 1994, 309-323.

[66] M.E.H. ISMAIL - L. LORCH - M.E. MULDOON, *Completely monotonic functions associated with the gamma function and its q-analogues*, J. Math. Anal. Appl., 116, 1986, 1-9.

[67] L. JACOBSEN - W.B. JONES - H. WAADELAND, *Further results on the computation of incomplete gamma functions*, in: *Analytic theory of continued fractions II* (W.J. Thron, ed.), Lecture Notes in Math., 1199, Springer, Berlin 1986, 67-89.

[68] W.B. JONES - W.J. THRON, *On the computation of incomplete gamma functions in the complex domain*, J. Comput. Appl. Math., 12-13, 1985, 401-417.

[69] D.G. KABE, *On some inequalities satisfied by beta and gamma functions*, South African Statist. J., 12, 1978, 25-31.

[70] S. KARLIN - J. MCGREGOR, *Many server queueing processes with Poisson input and exponential service times*, Pacific J. Math., 8, 1958, 87-118.

[71] J.D. KEČKIĆ - M.S. STANKOVIĆ, *Some inequalities for special functions*, Publ. Inst. Math. (Beograd), N.S., 13 (27), 1972, 51-54.

[72] J.D. KEČKIĆ - P.M. VASIĆ, *Some inequalities for the gamma function*, Publ. Inst. Math. (Beograd), N.S., 11 (25), 1971, 107-114.

[73] A.W. KEMP, *On gamma function inequalities*, Skand. Aktuarietidskr., 56, 1973, 65-69.

[74] D. KERSHAW, *Some extensions of W. Gautschi's inequalities for the gamma function*, Math. Comp., 41, 1983, 607-611.

[75] D. KERSHAW - A. LAFORGIA, *Monotonicity results for the gamma function*, Atti Accad. Sci. Torino Cl. Sci. Fis. Mat. Natur., 119, 1985, 127-133.

[76] K.S. KÖLBIG, *Complex zeros of an incomplete Riemann zeta function and of the incomplete gamma function*, Math. Comp., 24, 1970, 679-696.

[77] K.S. KÖLBIG, *Complex zeros of two incomplete Riemann zeta functions*, Math. Comp., 26, 1972, 551-565.

[78] K.S. KÖLBIG, *On the zeros of the incomplete gamma function*, Math. Comp., 26, 1972, 751-755.

[79] Y. KOMATU, *Elementary inequalities for Mills' ratio*, Rep. Statist. Appl. Res. Un. Jap. Sci. Engrs., 4, 1955, 69-70.

[80] E. KREYSZIG, *On the zeros of the Fresnel integrals*, Canadian J. Math., 9, 1957, 118-131.

[81] A. LAFORGIA, *Further inequalities for the gamma function*, Math. Comp., 42, 1984, 597-600.

[82] A. LAFORGIA - S. SISMONDI, *Monotonicity results and inequalities for the gamma and error functions*, J. Comput. Appl. Math., 23, 1988, 25-33.

[83] A. LAFORGIA - S. SISMONDI, *A geometric mean inequality for the gamma function*, Boll. Un. Mat. Ital., A (7), 3, 1989, 339-342.

[84] I.B. LAZAREVIĆ - A. LUPAŞ, *Functional equations for Wallis and gamma functions*, Univ. Beograd. Publ. Elektrotehn. Fak. Ser. Mat. Fiz., no. 461-497, 1979, 245-251.

[85] A.M. LEGENDRE, *Exercises de calcul intégral sur divers ordres de transcendantes et sur les quadratures*, Vol. 1, Courcier, Paris 1811.

[86] P. LÉVY, *Observations sur le mémoire de M. F. Tricomi: «Sulla frequenza dei numeri interi decomponibili nella somma di due potenze k-esime»*, Atti Accad. Sci. Torino, 75, 1939, 177-183.

[87] J.S. LEW, *On the Darling-Mandelbrot probability density and the zeros of some incomplete gamma functions*, Constructive Approx., 10, 1994, 15-30.

[88] Y. LI - X. DONG - S. PAN, *Computation of auxiliary functions in STO molecular integrals up to arbitrary accuracy. I. Evaluation of incomplete gamma function $E_n(x)$ by forward recursion*, Internat. J. Quant. Chem., 45, 1993, 3-14.

[89] G. LOCHS, *Über die Koeffizienten der asymptotischen Reihen für den Korrekturfaktor der Stirlingschen Formel und einen speziellen Wert der unvollständigen Gammafunktion*, Österreich. Akad. Wiss. Math.-Natur. Kl. Sitzungsber. II, 196, 1987, 27-37 and Supplement.

[90] L. LORCH, *Inequalities for some Whittaker functions*, Arch. Math. (Brno), 3, 1967, 1-9.

[91] L. LORCH, *Inequalities for ultraspherical polynomials and the gamma function*, J. Approx. Theory, 40, 1984, 115-120.

[92] D.W. LOZIER - F.W.J. OLVER, *Numerical evaluation of special functions*, in: *Mathematics of Computation 1943-1993: A half-century of computational mathematics* (W. Gautschi, ed.), Proc. Sympos. Appl. Math., 48, Amer. Math. Soc., Providence, RI, 1994, 79-125.

[93] L.G. LUCHT, *Mittelwertungleichungen für Lösungen gewisser Differenzengleichungen*, Aequationes Math., 39, 1990, 204-209.

[94] Y.L. LUKE, *Inequalities for the gamma function and its logarithmic derivative*, Math. Balkanica, 2, 1972, 118-123.

[95] J. VAN DE LUNE, *A note on Euler's (incomplete) gamma function*, Report ZN 61, Mathematisch Centrum, Amsterdam 1975.

[96] K. MAHLER, *Über die Nullstellen der unvollstaendigen Gammafunktionen*, Rend. Circ. Mat. Palermo, 54, 1930, 1-41.

[97] L. MALIGRANDA - J.E. PEČARIĆ - L.E. PERSSON, *Stolarsky's inequality with general weights*, Proc. Amer. Math. Soc., 123, 1995, 2113-2118.

[98] M. MERKLE, *Logarithmic convexity and inequalities for the gamma function*, J. Math. Anal. Appl., 203, 1996, 369-380.

[99] M. MERKLE, *On log-convexity of a ratio of gamma functions*, Univ. Beograd. Publ. Elektrotehn. Fak. Ser. Mat., 8, 1997, 114-119.

[100] H. MINC - L. SATHRE, *Some inequalities involving* $(r!)^{1/r}$, Proc. Edinburgh Math. Soc., (2), 14, 1964, 41-46.

[101] D.S. MITRINOVIĆ, *Analytic inequalities* (in cooperation with P.M. Vasić), Die Grundlehren der mathematischen Wissenschaften, 165, Springer, New York 1970.

[102] M.E. MULDOON, *Some monotonicity properties and characterizations of the gamma function*, Aequationes Math., 18, 1978, 54-63.

[103] N. NIELSEN, *Handbuch der Theorie der Gammafunktion*, Teubner, Leipzig 1906.

[104] A.M. ODLYZKO, *Analytic computations in number theory*, in: *Mathematics of Computation 1943-1993: A half-century of computational mathematics* (W. Gautschi, ed.), Proc. Sympos. Appl. Math., 48, Amer. Math. Soc., Providence, RI, 1994, 451-463.

[105] A.B. OLDE DAALHUIS - F.W.J. OLVER, *Hyperasymptotic solutions of second-order linear differential equations. I*, Methods Appl. Anal., 2, 1995, 173-197.

[106] A.B. OLDE DAALHUIS, *Hyperasymptotic solutions of second-order linear differential equations. II*, Methods Appl. Anal., 2, 1995, 198-211.

[107] F.W.J. OLVER, *On Stokes' phenomenon and converging factors*, in: *Asymptotic and computational analysis* (R. Wong, ed.), Lecture Notes in Pure and Appl. Math., 124, Dekker, New York 1990, 329-355.

[108] F.W.J. OLVER, *Uniform, exponentially improved, asymptotic expansions for the generalized exponential integral*, SIAM J. Math. Anal., 22, 1991, 1460-1474.

[109] F.W.J. OLVER, *Exponentially-improved asymptotic solutions of ordinary differential equations. I. The confluent hypergeometric function*, SIAM J. Math. Anal., 24, 1993, 756-767.

[110] F.W.J. OLVER, *The generalized exponential integral*, in: *Approximation and computation: A festschrift in honor of Walter Gautschi* (R.V.M. Zahar, ed.), Internat. Ser. Numer. Math., 119, Birkhäuser, Boston, MA, 1994, 497-510.

[111] V.I. PAGUROVA, *An asymptotic formula for the incomplete gamma function* (Russian), Zh. Vychisl. Mat. i Mat. Fiz., 5, 1965, 118-121. (Engl. transl. in U.S.S.R. Comput. Math. and Math. Phys., 5, 162-166).

[112] B. PALUMBO, *A generalization of some inequalities for the gamma function*, J. Comput. Appl. Math., 88, 1998, 255-268.

[113] R.B. PARIS, *An asymptotic representation for the Riemann zeta function on the critical line*, Proc. Roy. Soc. London, Ser. A, 446, 1994, 565-587.

[114] R.B. PARIS, *Error bounds for the uniform asymptotic expansion of the incomplete gamma function*, Tech. Rep. MACS(97:01), Division of Mathematical Sciences, University of Abertay Dundee, Dundee DD1 1HG, 1997.

[115] R.B. PARIS - S. CANG, *An asymptotic representation for* $\zeta(\frac{1}{2}+it)$, Methods Appl. Anal., 4, 1997, 449-470.

[116] J. PEČARIĆ - G. ALLASIA - C. GIORDANO, *Convexity and the gamma function*, Res. Rep. 41/1997, Department of Mathematics, University of Turin, 1-13. Submitted for publication.

[117] H.O. POLLAK, *A remark on «Elementary inequalities for Mills' ratio» by Yûsaku Komatu*, Rep. Statist. Appl. Res. Un. Jap. Sci. Engrs., 4, 1956, 110.

[118] F.E. PRYM, *Zur Theorie der Gammafunction*, J. Reine Angew. Math., 82, 1877, 165-172.

[119] B. RAJA RAO, *On an inequality satisfied by the gamma function and its logarithmic derivatives*, Metron, 39, 1981, 125-131.

[120] G. RASCH, *Über die Nullstellen der unvollständigen Gammafunktion* $P(z,\rho)$. *I. Die reellen Nullstellen von* $P(z,\rho)$ *bei positivem reellem* ρ, Math. Z., 29, 1929, 300-318.

[121] H.-J. SCHELL, *Asymptotische Entwicklungen für die unvollständige Gammafunktion*, Wiss. Z. Tech. Hochsch. Karl-Marx-Stadt, 22, 1980, 477-485.

[122] H.-J. SCHELL, *Über das asymptotische Verhalten des Fermi-Dirac-Integrals*, Z. Anal. Anwendungen, 6, 1987, 421-438.

[123] J.B. SELLIAH, *An inequality satisfied by the gamma function*, Canad. Math. Bull., 19, 1976, 85-87.

[124] D.N. SHANBHAG, *On some inequalities satisfied by the gamma function*, Skand. Aktuarietidskr., 50, 1967, 45-49.

[125] L.J. SLATER, *Confluent hypergeometric functions*, Cambridge University Press, New York 1960.

[126] D.V. SLAVIĆ, *On inequalities for* $\frac{\Gamma(x+1)}{\Gamma(x+\frac{1}{2})}$, Univ. Beograd. Publ. Elektrotehn. Fak. Ser. Mat. Fiz., 498-541, 1975, 17-20.

[127] A.I. STEPANEC, *Asymptotic formulas for the roots of the cosine integral* (Russian), in: *Questions of the theory of the approximation of functions* (Russian), Akad. Nauk Ukrain. SSR, Inst. Mat., Kiev 1980, 154-159.

[128] K.B. STOLARSKY, *From Wythoff's Nim to Chebyshev's inequality*, Amer. Math. Monthly, 98, 1991, 889-900.

[129] A.J. STRECOK, *On the calculation of the inverse of the error function*, Math. Comp., 22, 1968, 144-158.

[130] R. TAKENAGA, *On the evaluation of the incomplete gamma function*, Math. Comp., 20, 1966, 606-610.

[131] N.M. TEMME, *Uniform asymptotic expansions of the incomplete gamma functions and the incomplete beta functions*, Math. Comp., 29, 1975, 1109-1114.

[132] N.M. TEMME, *The asymptotic expansion of the incomplete gamma functions*, SIAM J. Math. Anal., 10, 1979, 757-766.

[133] N.M. TEMME, *Some problems in connection with the incomplete gamma functions*, Report TW 205/80, Stichting Mathematisch Centrum, Amsterdam 1980.

[134] N.M. TEMME, *Traces to Tricomi in recent work on special functions and asymptotics of integrals*, in: *Mathematical analysis* (J.M. Rassias, ed.), Teubner-Texte Math., 79, Teubner, Leipzig 1985, 236-249.

[135] N.M. TEMME, *On the computation of the incomplete gamma functions for large values of the parameters*, in: *Algorithms for approximation* (J. C. Mason and M.G. Cox, eds.), Inst. Math. Appl. Conf. Ser. New Ser., 10, Oxford Univ. Press, New York 1987, 479-489.

[136] N.M. TEMME, *Asymptotic inversion of incomplete gamma functions*, Math. Comp., 58, 1992, 755-764.

[137] N.M. TEMME, *Computational aspects of incomplete gamma functions with large complex parameters*, in: *Approximation and computation: A festschrift in honor of Walter Gautschi* (R.V.M. Zahar, ed.), Internat. Ser. Numer. Math., 119, Birkhäuser, Boston, MA, 1994, 551-562.

[138] N.M. TEMME, *Asymptotics of zeros of incomplete gamma functions*, Ann. Numer. Math., 2, 1995, 415-423.

[139] N.M. TEMME, *Uniform asymptotics for the incomplete gamma functions starting from negative values of the parameters*, Methods Appl. Anal., 3, 1996, 335-344.

[140] N.M. TEMME, *Special functions: An introduction to the classical functions of mathematical physics*, Wiley, New York 1996.

[141] N.M. TEMME - A.B. OLDE DAALHUIS, *Uniform asymptotic approximation of Fermi-Dirac integrals*, J. Comput. Appl. Math., 31, 1990, 383-387.

[142] A. TORTORICI MACALUSO, *Generalizzazione di alcune disuguaglianze per la funzione gamma*, Riv. Mat. Univ. Parma, (4), 13, 1987, 373-378.

[143] F. TRICOMI, *Sulla frequenza dei numeri interi decomponibili nella somma di due potenze k-esime*, Atti Accad. Sci. Torino, Cl. I, 74, 1938-39, 369-380.

[144] F.G. TRICOMI, *Una formula sulla norma della funzione gamma incompleta*, Boll. Unione Mat. Ital., (3), 4, 1949, 341-344.

[145] F. TRICOMI, *Sviluppo in serie asintotica del rapporto* $\Gamma(z+\alpha):\Gamma(z+\beta)$, Rend. Sem. Mat. Torino, 9, 1949-50, 343-351.

[146] F.G. TRICOMI, *Asymptotische Eigenschaften der unvollständigen Gammafunktion*, Math. Z., 53, 1950, 136-148.

[147] F.G. TRICOMI, *Sulla funzione gamma incompleta*, Ann. Mat. Pura Appl., (4), 31, 1950, 263-279.

[148] F.G. TRICOMI, *A class of non-orthogonal polynomials related to those of Laguerre*, J. Analyse Math., 1, 1951, 209-231.

[149] F.G. TRICOMI, *Applicazione della funzione gamma incompleta allo studio della somma di vettori casuali*, Boll. Un. Mat. Ital., (3), 6, 1951, 189-194.

[150] F.G. TRICOMI, *La seconda soluzione dell'equazione di Laguerre*, Boll. Un. Mat. Ital., (3), 7, 1952, 1-4.

[151] F.G. TRICOMI, *Funzioni ipergeometriche confluenti*, Edizioni Cremonese, Roma 1954.

[152] F. TRICOMI, *Sul comportamento asintotico della funzione gamma incompleta* $\Gamma(\alpha,x)$ *al simultaneo divergere di* α *e* x, Univ. Nac. Tucumán Rev., Ser. A, 14, 1962, 333-339.

[153] F.G. TRICOMI, *La mia vita di matematico attraverso la cronistoria dei miei lavori (Bibliografia commentata 1916-1967)*, Casa Editrice Dott. Antonio Milani, Padova 1967.

[154] F.G TRICOMI - A. ERDÉLYI, *The asymptotic expansion of a ratio of gamma functions*, Pacific J. Math., 1, 1951, 133-142.

[155] V.R.R. UPPULURI, *A stronger version of Gautschi's inequality satisfied by the gamma function*, Skand. Aktuarietidskr., 47, 1964, 51-52.

[156] R.S. VARGA, *Scientific computation on mathematical problems and conjectures*, CBMS-NSF Regional Conf. Ser. in Appl. Math., 60, SIAM, Philadelphia 1990.

[157] L. VIETORIS, *Dritter Beweis der die unvollständige Gammafunktion betreffenden Lochsschen Ungleichungen*, Österreich. Akad. Wiss. Math. - Natur. Kl. Sitzungsber., II, 192, 1983, 83-91.

[158] A. WALTHER, *Über die reellen Nullstellen der unvollständigen Gammafunktion* $P(z)$, Math. Z., 23, 1925, 238-245.

[159] G.N. WATSON, *A treatise on the theory of Bessel functions*, 2nd ed., Cambridge University Press, 1952.

[160] E.T. WHITTAKER - G.N. WATSON, *A course of modern analysis: An introduction to the general theory of infinite processes and of analytic functions; with an account of the principal transcendental functions*, 4th ed., Cambridge University Press, 1940. (Reprinted 1962).

9.12. [168] "Gauss quadrature approximations to hypergeometric and confluent hypergeometric functions"

---

[168] "Gauss quadrature approximations to hypergeometric and confluent hypergeometric functions," *J. Comput. Appl. Math.* **139**, 173–187 (2002).

ELSEVIER

Journal of Computational and Applied Mathematics 139 (2002) 173–187

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.com/locate/cam

Gauss quadrature approximations to hypergeometric and confluent hypergeometric functions

Walter Gautschi *

Department of Computer Sciences, Purdue University, West Lafayette, IN 47907-1398, USA

Received 6 November 2000

Abstract

Integral representations of hypergeometric and confluent hypergeometric functions with real parameters and complex arguments are used to approximate these functions by Gaussian quadrature. An analysis is given of the errors involved and of estimates of the number of Gauss points required to achieve any given accuracy. Numerical examples illustrate the theory.

0. Introduction

Hypergeometric and confluent hypergeometric functions admit integral representations for parameter values restricted by certain inequalities. Applying Gaussian quadrature to these integrals produces approximations valid in the whole complex plane, for the confluent hypergeometric function, and in the whole complex plane cut along the segment $[1,\infty]$ of the positive real axis, for the hypergeometric function. The former are studied in Section 1. Numerical evidence, graphically displayed, demonstrates the effectiveness of these quadrature approximations. The error is analyzed both in terms of derivatives of the integrand (in the case of the confluent hypergeometric function) and in terms of derivative-free contour integral representations of the remainder term. Both approaches lead to a priori estimates of the number of Gauss points needed for given accuracy. Analogous analyses are given in Section 2 for the hypergeometric function. The paper ends with Section 3 containing some concluding remarks.

* Tel.: +1-765-494-1995; fax: +1-765-494-0739.
E-mail address: wxg@cs.purdue.edu (W. Gautschi).

1. Confluent hypergeometric functions

1.1. Quadrature approximation

Our interest is in the confluent hypergeometric function $M(a,b;z)$, also known as Kummer's function, when a and b are positive parameters satisfying $b > a > 0$ and z is real or complex. One then has the well-known integral representation (see, e.g. [1, Eq. (13.2.1)])

$$M(a,b;z) = \frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)} \int_0^1 \mathrm{e}^{zt} w_{a,b}(t)\,\mathrm{d}t, \tag{1.1}$$

where

$$w_{a,b}(t) = (1-t)^{b-a-1} t^{a-1} \tag{1.2}$$

is a Jacobi weight function on the interval $[0,1]$ with Jacobi parameters $\alpha = b-a-1$, $\beta = a-1$. We propose to approximate the integral in (1.1) by an n-point Gauss–Jacobi quadrature rule

$$\int_0^1 \mathrm{e}^{zt} w_{a,b}(t)\,\mathrm{d}t = \sum_{k=1}^{n} w_k^{\mathrm{J}} \mathrm{e}^{z t_k^{\mathrm{J}}} + E_n(a,b;z), \tag{1.3}$$

where t_k^{J} and w_k^{J} are the Gauss–Jacobi nodes and weights for the interval $[0,1]$ (rather than the standard interval $[-1,1]$). These quadrature rules, after the change of variables $t \mapsto \frac{1}{2}(1+t)$, are readily generated by the software provided in [6].

1.2. Numerical data

Approximation (1.3) clearly converges as $n \to \infty$ for any complex z, since the integrand is continuous on $[0,1]$, in fact an entire function of t. With regard to the quality of convergence, we first describe some numerical tests.

We experimentally determined the smallest value of n that, for given a, b, and z, yields a prescribed relative accuracy ε (absolute accuracy ε if the result is less than 1 in absolute value). The left frame of Fig. 1 shows the maximum of this smallest n as a function of real $z = x$, $-100 \leqslant x \leqslant 100$, where the maximum is taken over $a = 0.25(0.25)5$, $b = (a+0.25)(0.25)5$ and $\varepsilon = \frac{1}{2} \cdot 10^{-10}$. The right frame displays the analogous n as estimated below in (1.9), (1.24) (and modified as described in the last paragraph of Section 1.4). It is seen that for real $z = x$ in $[-100, 100]$ and 10-digit accuracy, a Gauss rule of not more than 30 points will do, the case $x < 0$ being slightly more favorable than the case $x > 0$. Note also that for real $z = x$, the quadrature sum in (1.3) consists of positive terms only, so that its evaluation is perfectly stable.

Analogous values of n—experimental and estimated via (1.9) below—for complex $z = r\mathrm{e}^{\mathrm{i}\varphi}$, $0 < r \leqslant 100$, $0 \leqslant \varphi \leqslant \pi$, are depicted in Fig. 2. Here, a 50-point Gauss rule will suffice.

1.3. Error estimates in terms of derivatives

If $u(t)$ and $v(t)$ denote the real and imaginary parts of e^{zt}, one has from the well-known error formula for Gaussian quadrature [2, Theorem, p. 98], applied separately to the real and imaginary

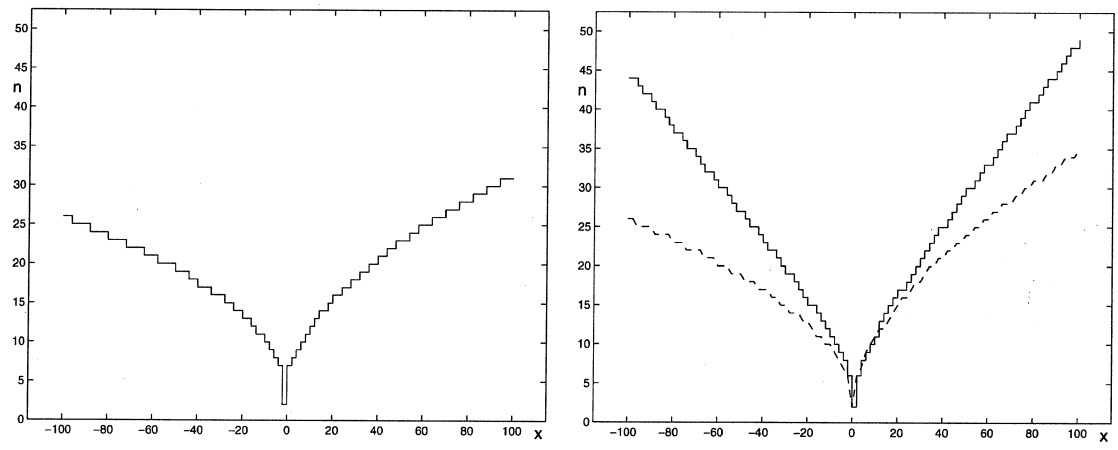


Fig. 1. Values of n of n-point quadrature rules (1.3) yielding 10-digit accuracy for real arguments z; left: experimentally determined, right: estimated.

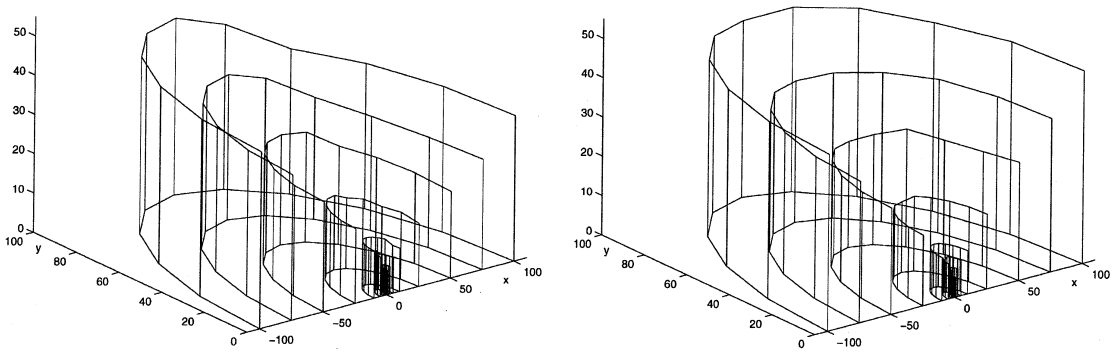


Fig. 2. Values of n of n-point quadrature rules yielding 10-digit accuracy for complex arguments z; left: experimentally determined, right: estimated.

part of (1.3),

$$E_n(a,b;z) = \frac{1}{(2n)!}[u^{(2n)}(\tau_u) + \mathrm{i}v^{(2n)}(\tau_v)]\gamma_n(a,b), \tag{1.4}$$

where

$$\gamma_n(a,b) = \int_0^1 [\pi_n(t;w_{a,b})]^2 w_{a,b}(t)\,\mathrm{d}t.$$

Here, $\pi_n(\,\cdot\,;w_{a,b})$ is the monic Jacobi polynomial relative to the interval $[0,1]$ (the polynomial $G_n(b-1,a,t)$ in the notation of [1, Table 22.2]), and τ_u, τ_v are certain numbers between 0 and 1. From [1, Table 22.2] one has

$$\gamma_n(a,b) = \frac{n!\Gamma(n+a)\Gamma(n+b-1)\Gamma(n+b-a)}{(2n+b-1)\Gamma^2(2n+b-1)}.$$

For $n = 0$, the quantity γ_0 is just the reciprocal of the numerical factor multiplying the integral in (1.1); thus,

$$M(a,b;z) = \frac{1}{\gamma_0(a,b)} \int_0^1 \mathrm{e}^{zt} w_{a,b}(t)\,\mathrm{d}t. \tag{1.1'}$$

Using Stirling's formula for the gamma function, one finds, after straightforward computation,

$$\gamma_n(a,b) \sim \frac{\pi}{2^{2b-3}} 2^{-4n}, \quad n \to \infty. \tag{1.5}$$

On the other hand, letting $z = x + \mathrm{i}y = r\mathrm{e}^{\mathrm{i}\varphi}$, one has

$$\frac{\mathrm{d}^{2n}}{\mathrm{d}t^{2n}} \mathrm{e}^{zt} = z^{2n}\mathrm{e}^{zt} = r^{2n}\mathrm{e}^{xt+\mathrm{i}(yt+2n\varphi)},$$

so that

$$|u^{(2n)}(t)| \leqslant r^{2n}\mathrm{e}^{xt}, \quad |v^{(2n)}(t)| \leqslant r^{2n}\mathrm{e}^{xt}.$$

There follows from (1.4) that

$$|E_n(a,b;z)| \leqslant \frac{\delta(y)\gamma_n(a,b)}{(2n)!}|z|^{2n} e_+^x, \tag{1.6}$$

where

$$\delta(y) = \begin{cases} 1 & \text{if } y = 0, \\ 2 & \text{otherwise,} \end{cases} \qquad e_+^x = \begin{cases} 1 & \text{if } x \leqslant 0, \\ \mathrm{e}^x & \text{otherwise,} \end{cases}$$

or, asymptotically for $n \to \infty$, with the help of (1.5) and Stirling's formula for $(2n)!$,

$$|E_n(a,b;z)| \lesssim \sqrt{\frac{\pi}{n}} \frac{1}{2^{2b-3}} \left(\frac{\mathrm{e}|z|}{8n}\right)^{2n} e_+^x. \tag{1.7}$$

Actually, if $y = 0$, the bound on the right could be halved, but this is of little consequence. It is interesting to note that the error bound in (1.7) does not depend on a, and could be made independent of b as well by letting $b = 0$.

1.4. Estimate of n for prescribed accuracy

In order to estimate the number n of Gauss points needed to achieve an error tolerance $\varepsilon > 0$, we disregard the factor $\sqrt{\pi/n}$ in (1.7) and note that $|E_n| \leqslant \varepsilon$ if

$$\left(\frac{\mathrm{e}|z|}{8n}\right)^{2n} \leqslant \frac{2^{2b-3}}{e_+^x}\varepsilon$$

or, equivalently,

$$\frac{8n}{\mathrm{e}|z|} \ln \frac{8n}{\mathrm{e}|z|} \geqslant \frac{4}{\mathrm{e}|z|} \left\{ x_+ + (3-2b)\ln 2 + \ln \frac{1}{\varepsilon} \right\}, \tag{1.8}$$

where $x_+ = x$ if $x \geqslant 0$ and $x_+ = 0$ if $x < 0$. Denoting by $t(s)$ the inverse function of $s = t \ln t$, one has $t \ln t \geqslant c$, $c \geqslant 0$, if and only if $t \geqslant t(c)$. Hence, (1.8) can be given the form

$$n \geqslant \frac{\mathrm{e}|z|}{8} t\left(\frac{4}{\mathrm{e}|z|}\left[x_+ + (3-2b)\ln 2 + \ln\frac{1}{\varepsilon}\right]\right). \tag{1.9}$$

Low-accuracy approximations of the function $t(s)$ are given in [5, pp. 51–52].

The analysis given here is aimed towards attaining a given *absolute* error in (1.3), not relative error. In practice, however, it is usually the relative error that one wants to control, i.e., the relative error

$$\frac{M(a,b;z) - (1/\gamma_0(a,b))\sum_{k=1}^{n} w_k^{\mathrm{J}} \mathrm{e}^{zt_k^{\mathrm{J}}}}{M(a,b;z)} = \frac{E_n(a,b;z)}{\int_0^1 \mathrm{e}^{zt} w_{a,b}(t)\,\mathrm{d}t}.$$

This is particularly so if $|M(a,b;z)| > 1$. (Otherwise, control of the absolute error is often adequate.) Thus, we want

$$|E_n(a,b;z)| \leqslant \varepsilon \left|\int_0^1 \mathrm{e}^{zt} w_{a,b}(t)\,\mathrm{d}t\right| \quad \text{if} \quad \left|\int_0^1 \mathrm{e}^{zt} w_{a,b}(t)\,\mathrm{d}t\right| > \gamma_0$$

and

$$|E_n(a,b;z)| \leqslant \varepsilon\gamma_0 \quad \text{if} \quad \left|\int_0^1 \mathrm{e}^{zt} w_{a,b}(t)\,\mathrm{d}t\right| \leqslant \gamma_0.$$

While this would seem to require knowledge of the integral, a rough estimate of it suffices. Once in possession of such an estimate, one can simply apply the formula for n in (1.9), but replace ε, respectively, by ε multiplied by the absolute value of that estimate, or by $\varepsilon\gamma_0$. The estimate in question can be obtained, e.g., by applying a low-order Gaussian quadrature rule to the integral in (1.1). Extensive tests have shown that a 10-point rule should be adequate for this purpose. The solid graph in the right frame of Fig. 1 displays the estimate of n so obtained.

1.5. Derivative-free error estimates

The change of variables $t \mapsto \frac{1}{2}(1+t)$ transforms the integral in (1.1) into one over $[-1,1]$,

$$M(a,b;z) = \frac{1}{\gamma_0^{\mathrm{J}}(a,b)} \int_{-1}^{1} \mathrm{e}^{(1/2)z(1+t)} w_{a,b}^{\mathrm{J}}(t)\,\mathrm{d}t. \tag{1.10}$$

Here, γ_0^{J}, and more generally, γ_n^{J}, are the normalization factors of the standard (monic) Jacobi polynomials,

$$\gamma_n^{\mathrm{J}}(a,b) = \int_{-1}^{1} [\pi_n^{\mathrm{J}}(t; w_{a,b}^{\mathrm{J}})]^2 w_{a,b}^{\mathrm{J}}(t)\,\mathrm{d}t, \tag{1.11}$$

where

$$w_{a,b}^{\mathrm{J}}(t) = (1-t)^{b-a-1}(1+t)^{a-1}.$$

The error term of the n-point Gauss–Jacobi quadrature rule applied to the integral in (1.10) will be denoted by $E_n^{\mathrm{J}}(a,b;z)$, and the integrand by

$$f^{\mathrm{J}}(t;z) = \mathrm{e}^{(1/2)z(1+t)}. \tag{1.12}$$

A derivative-free estimate of E_n^{J} is given by (see, e.g., [7])

$$E_n^{\mathrm{J}}(a,b;z)=\frac{1}{2\pi\mathrm{i}}\oint_{\Gamma}K_n(\zeta)f^{\mathrm{J}}(\zeta;z)\,\mathrm{d}\zeta, \tag{1.13}$$

where for Γ one may take a circle $\Gamma=C_R$ of radius $R>1$ centered at the origin. The "kernel" K_n of the quadrature error can be represented, e.g., by

$$K_n(\zeta)=\frac{\rho_n(\zeta)}{\pi_n(\zeta)}, \tag{1.14}$$

where $\pi_n(\zeta)=\pi_n(\zeta;w_{a,b}^{\mathrm{J}})$ is the monic Jacobi polynomial belonging to the weight function $w_{a,b}^{\mathrm{J}}$, and

$$\rho_n(\zeta)=\rho_n(\zeta;w_{a,b}^{\mathrm{J}})=\int_{-1}^{1}\frac{\pi_n(t)}{\zeta-t}w_{a,b}^{\mathrm{J}}(t)\,\mathrm{d}t,\quad \zeta\in\mathbb{C}\setminus[-1,1].$$

The kernel tends to 0 as $n\to\infty$ for any complex $\zeta\notin[-1,1]$, the faster the further away ζ is from the interval $[-1,1]$. It is easily computed by recursion (see [7, Section 4]), provided ζ is not too close to $[-1,1]$. From [7, Section 3] it follows, moreover, that

$$\max_{\zeta\in C_R}|K_n(\zeta)|=\begin{cases}K_n(R) & \text{if } b\leqslant 2a,\\ K_n(-R) & \text{if } b>2a.\end{cases} \tag{1.15}$$

Consequently, by (1.13),

$$|E_n^{\mathrm{J}}(a,b;z)|\leqslant|RK_n(\pm R)|\max_{\zeta\in C_R}|f^{\mathrm{J}}(\zeta;z)|, \tag{1.16}$$

where the plus or minus sign holds according as $b\leqslant 2a$ or $b>2a$. The maximum on the right is easily found to be

$$\max_{\zeta\in C_R}|f^{\mathrm{J}}(\zeta;z)|=\mathrm{e}^{(1/2)r(R+\cos\varphi)}, \tag{1.17}$$

where, as before, $z=r\mathrm{e}^{\mathrm{i}\varphi}$. Thus,

$$|E_n^{\mathrm{J}}(a,b;z)|\leqslant|RK_n(\pm R)|\mathrm{e}^{(1/2)r(R+\cos\varphi)}. \tag{1.18}$$

This holds for any $R>1$. Optimizing the bound then yields

$$|E_n^{\mathrm{J}}(a,b;z)|\leqslant\min_{R>1}\{|RK_n(\pm R)|\mathrm{e}^{(1/2)r(R+\cos\varphi)}\},\quad z=r\mathrm{e}^{\mathrm{i}\varphi}. \tag{1.19}$$

In the numerical work described below, the kernel K_n has been computed by means of the double-precision routine `dkern` of [6]. (It calls for the additional routines `d1mach`, `drecur`, `nu0jac`, and `dknum`.)

Table 1 illustrates this bound for $a=0.5$ and $b=2.5$. For each n and $r=|z|$, there are four entries. The two upper ones are bounds for the relative error when $\varphi=0$ and absolute error when $\varphi=\pi$. The two lower ones are the optimal values of R that yield the minimum in (1.19) for these two values of φ. According to the discussion at the end of Section 1.4, bounds for the relative (resp. absolute) error are obtained by dividing the bound in (1.19), respectively, by $\gamma_0^{\mathrm{J}}M(a,b;r)$ and by γ_0^{J}, since for $\varphi=0$ and π one has, respectively, $M(a,b,r)>1$ and $M(a,b;-r)<1$. It is seen that the errors for $\varphi=\pi$ are consistently smaller than those for $\varphi=0$, confirming what was experimentally observed in the left frame of Fig. 1. Evidently, the reason for this is the presence of the term $\cos\varphi$ in the exponential on the right of (1.19).

Table 1
Optimized error bounds for $a = 0.5$, $b = 2.5$

n	5	5	10	10	20	20	50	50
r	$\varphi = 0$	$\varphi = \pi$	$\varphi = 0$	$\varphi = \pi$	$\varphi = 0$	$\varphi = \pi$	$\varphi = 0$	$\varphi = \pi$
1	0.19e − 11	0.87e − 12	0.35e − 29	0.16e − 29	0.13e − 70	0.60e − 71	0.13e − 216	0.61e − 217
	20.2	20.2	40.1	40.1	80.1	80.0	200.0	200.0
5	0.43e − 4	0.18e − 5	0.64e − 15	0.26e − 16	0.20e − 42	0.82e − 44	0.17e − 146	0.68e − 148
	4.26	4.26	8.14	8.14	16.1	16.1	40.0	40.0
10	0.38e − 1	0.33e − 3	0.40e − 9	0.34e − 11	0.11e − 30	0.92e − 33	0.90e − 117	0.77e − 119
	2.37	2.37	4.20	4.20	8.10	8.10	20.0	20.0
20	0.11e + 2	0.21e − 1	0.43e − 4	0.85e − 7	0.66e − 20	0.12e − 22	0.43e − 88	0.85e − 91
	1.51	1.51	2.30	2.30	4.16	4.16	10.1	10.1
50	0.29e + 4	0.88e + 0	0.46e + 1	0.14e − 2	0.41e − 8	0.13e − 11	0.24e − 53	0.75e − 57
	1.12	1.12	1.33	1.33	1.92	1.92	4.14	4.14
100	0.65e + 5	0.50e + 1	0.17e + 4	0.13e + 0	0.92e − 2	0.70e − 6	0.23e − 31	0.17e − 35
	1.04	1.04	1.10	1.10	1.30	1.30	2.25	2.25

Comparing the error bounds in Table 1 with error bounds derivable from (1.7), one finds general agreement within 1–2 orders of magnitude when $r = 1, 5$, or 10. For larger values of r, however, the bounds in Table 1 become progressively better, by as much as 17 orders of magnitude (when $r = 100$ and $n = 50$).

1.6. *Asymptotics for* $K_n(\pm R)$

If $K_n^{(\alpha,\beta)}(\zeta) = \rho_n^{(\alpha,\beta)}(\zeta)/\pi_n^{(\alpha,\beta)}(\zeta)$ denotes kernel (1.14) for the Jacobi weight function $(1-t)^\alpha(1+t)^\beta$, it is known (see [4], [3, Appendix A.1]) that, for ζ away from $[-1, 1]$,

$$K_n^{(\alpha,\beta)}(\zeta) \sim c_n(\alpha,\beta)\frac{(\zeta-1)^\alpha(\zeta+1)^\beta}{(\zeta+\sqrt{\zeta^2-1})^{2n+\alpha+\beta+1}}, \quad n \to \infty, \tag{1.20}$$

where $-\pi < \arg(\zeta \pm 1) < \pi$ and

$$c_n(\alpha,\beta) = 2^{4n+2(\alpha+\beta+1)}\frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)\Gamma(n+1)\Gamma(n+\alpha+\beta+1)}{\Gamma(2n+\alpha+\beta+2)\Gamma(2n+\alpha+\beta+1)}.$$

Applying Stirling's formula to the gamma functions above yields

$$c_n(\alpha,\beta) \sim 2\pi. \tag{1.21}$$

We need (1.20) for real $\zeta = R$ and $\zeta = -R$, $R > 1$. From the well-known property of Jacobi polynomials $\pi_n^{(\alpha,\beta)}(-\zeta) = (-1)^n\pi_n^{(\beta,\alpha)}(\zeta)$, which implies $\rho_n^{(\alpha,\beta)}(-\zeta) = -\rho_n^{(\beta,\alpha)}(\zeta)$, one obtains from (1.20), (1.21), as $n \to \infty$,

$$|K_n^{(\alpha,\beta)}(\pm R)| \sim 2\pi\frac{|\pm R-1|^\alpha\,|\pm R+1|^\beta}{(R+\sqrt{R^2-1})^{2n+\alpha+\beta+1}}, \quad R > 1. \tag{1.22}$$

Inserting this into (1.19), with $\alpha = b - a - 1$, $\beta = a - 1$, one gets

$$|E_n^{\mathrm{J}}(a,b;z)| \lesssim \min_{R>1}\left\{2\pi\frac{R|\pm R-1|^{b-a-1}|\pm R+1|^{a-1}}{(R+\sqrt{R^2-1})^{2n+b-1}}\mathrm{e}^{(1/2)r(R+\cos\varphi)}\right\}, \quad z = r\mathrm{e}^{\mathrm{i}\varphi}. \tag{1.23}$$

Using this bound in place of the one in (1.19), we reproduced Table 1 and found almost perfect agreement. (The largest discrepancy observed was 2 units in the second decimal digit of the mantissas.)

It is also possible to estimate n such that $|E_n^{\mathrm{J}}(a,b;z)| \leqslant \varepsilon$ by combining (1.22) with (1.18). One finds

$$n \geqslant \frac{1}{2}\left(1 - b + \min_{R>1}\frac{(1/2)r(R+\cos\varphi) + \ln(2\pi R|\pm R-1|^{b-a-1}|\pm R+1|^{a-1}) + \ln(1/\varepsilon)}{\ln(R+\sqrt{R^2-1})}\right). \tag{1.24}$$

The minimum on the right is best obtained numerically, by evaluating the objective function at the zero of its derivative (and making sure that it is indeed a minimum).

The dashed graph in the right frame of Fig. 1 shows the estimate of n as obtained from (1.24) with $z = r\mathrm{e}^{\mathrm{i}\varphi}$, $\varphi = 0$ and π, when ε is modified as discussed at the end of Section 1.4.

2. Hypergeometric functions

2.1. Quadrature approximation

We now consider the hypergeometric function $F(a,b;c;z)$, for real parameters a and $c > b > 0$, and real or complex z in the complex plane cut along the line from 1 to ∞. We will assume throughout that $z \neq 0$, since otherwise one trivially has $F(a,b;c;0) = 1$. The integral representation then is (cf. [1, Eq. 15.3.1])

$$F(a,b;c;z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1 (1-zt)^{-a} w_{b,c}(t)\,\mathrm{d}t, \tag{2.1}$$

where $w_{b,c}$ is the Jacobi weight (1.2) with a replaced by b, and b by c. The Gauss quadrature approximation of the integral in (2.1) is

$$\int_0^1 (1-zt)^{-a} w_{b,c}(t)\,\mathrm{d}t = \sum_{k=1}^{n} w_k^{\mathrm{J}}(1-zt_k^{\mathrm{J}})^{-a} + E_n(a,b;c;z), \tag{2.2}$$

the Jacobi nodes t_k^{J} and weights w_k^{J} now referring to the Jacobi weight function $w_{b,c}$. (The use of (2.2) was already suggested in 1955 by Karmazina [8], who provided tables of the Gauss nodes t_k^{J} and weights w_k^{J} for selected values of the parameters b and c, but gave no analysis of the error.) In analogy to (1.1′), one has

$$F(a,b;c;z) = \frac{1}{\gamma_0(b,c)}\int_0^1 (1-tz)^{-a} w_{b,c}(t)\,\mathrm{d}t. \tag{2.1′}$$

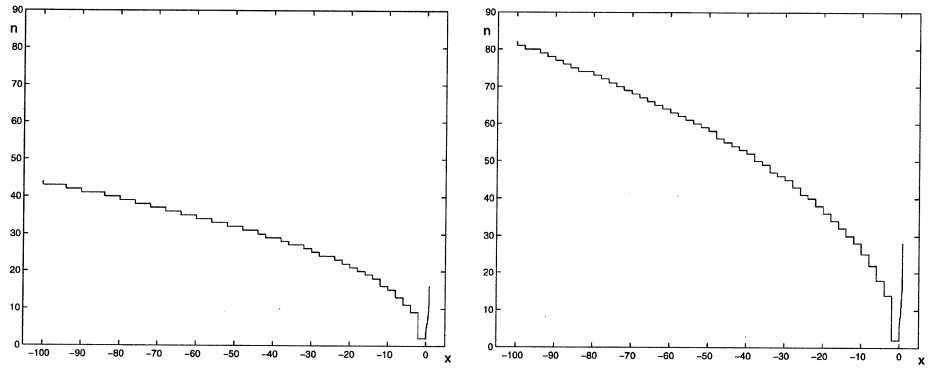


Fig. 3. Values of n of n-point quadrature rules (2.2) yielding 10-digit accuracy for real arguments z; left: $a < 0$, right: $a > 0$.

2.2. Numerical data

The Gauss quadrature approximation in (2.2) converges for any z in the complex plane cut along $[1,\infty]$, since the integrand then is a continuous function of t. However, the error analysis given for the confluent hypergeometric function in Sections 1.3 and 1.4 does not carry over. The reason is that the $2n$th derivative of the integrand now is

$$\frac{d^{2n}}{dt^{2n}}(1-zt)^{-a} = (a)_{2n}(1-zt)^{-a-2n}z^{2n}, \tag{2.3}$$

with $(a)_{2n} = a(a+1)\cdots(a+2n-1)$ the Pochhammer symbol. The latter behaves like $(2n)!$ for large n, which cancels the $(2n)!$ in the denominator of the error formula analogous to (1.4). What remains, in spite of the rapid decrease of $\gamma_n(b,c)$, will no longer necessarily tend to zero as $n \to \infty$. The derivative-free error estimation of Section 1.5, on the other hand, applies also in the present context if appropriately modified. Before discussing this, it may be useful to make some preliminary remarks and describe numerical tests.

A first observation that can be made is that the quadrature error $E_n(a,b;c;z)$ will behave differently depending on whether a is negative or positive. In the former case, the integrand

$$f(t;z) = (1-zt)^{-a} \tag{2.4}$$

of (2.1) is a "polynomial-like" function of t, and in fact an outright polynomial if a is a negative integer. Moreover, the Gauss formula (2.2) has zero error, $E_n(a,b;c;z) = 0$, if a is a nonpositive integer $a \geqslant -2n+1$. If a is negative but not an integer, and z is on the cut $[1,\infty]$, then $f(t;z)$ has a branch-point singularity for some $t \in (0,1]$, but remains bounded. None of this is true if a is positive. One therefore expects convergence of the Gauss formula (2.2) to be faster for negative values of a, and slower for positive values. This is borne out in the graphs of Fig. 3 for real $z=x$, $-100 \leqslant x < 1$, and in the graphs of Fig. 4 for complex $z = 1 + re^{i\varphi}$, $0 < r \leqslant 100$, $\pi \geqslant \varphi \geqslant \frac{1}{4}\pi$ (note the slightly different notation compared to the one in Section 1). In both the figures the maximum values of n are shown that are required for 10-digit accuracy, the maximum being taken over $a = -5(0.5)0$ (resp. $a = 0.5(0.5)5$). In either case, one clearly needs considerably larger values of n to achieve a

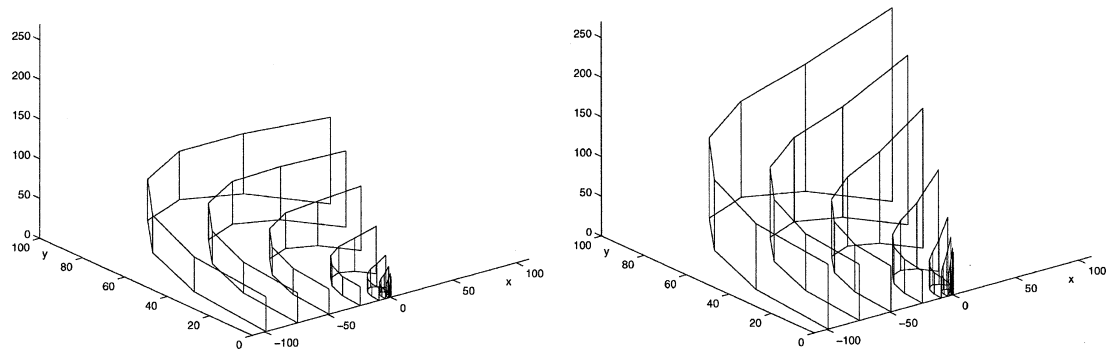


Fig. 4. Values of n of n-point quadrature rules (2.2) yielding 10-digit accuracy for complex arguments z; left: $a < 0$, right: $a > 0$.

prescribed accuracy than is the case for confluent hypergeometric functions, and the situation gets worse as z approaches the cut $[1,\infty]$. (This is why we omitted φ-values between 0 and $\frac{1}{4}\pi$.)

2.3. Derivative-free error estimates

We make the same change of variables as at the beginning of Section 1.5, thus writing

$$F(a,b;c;z) = \frac{1}{\gamma_0^{\mathrm{J}}(b,c)} \int_{-1}^{1} [1 - \tfrac{1}{2}z(1+t)]^{-a} w_{b,c}^{\mathrm{J}}(t)\,\mathrm{d}t. \tag{2.1''}$$

The integrand

$$f^{\mathrm{J}}(t;z) = [1 - \frac{1}{2}z(1+t)]^{-a}, \quad z \in \mathbb{C} \setminus [1,\infty] \tag{2.5}$$

is holomorphic in the variable t in the whole complex plane cut along the line $\lambda/z - 1$, $\lambda \geqslant 2$. We shall denote this cut plane by $\mathbb{C}_z^*$. The cut lies entirely outside the unit circle if and only if $|\lambda/z - 1|^2 > 1$ for all $\lambda \geqslant 2$, which is easily seen to be equivalent to $\mathrm{Re}\, z < 1$. In this case, the method used in Section 1.5 to estimate the quadrature error still applies provided one takes circular contours about the origin with radii R satisfying $1 < R < |2/z - 1|$. Otherwise, a family of confocal ellipses (with foci at ± 1) can be used that are sufficiently slim to leave the cut outside. These, of course, can also be used in the former case. We discuss the two cases separately.

2.3.1. The case $\mathrm{Re}\, z < 1$

With $E_n^{\mathrm{J}}(a,b;c;z)$ denoting the error term of Gauss–Jacobi quadrature applied to the integral in (2.1''), we have

$$E_n^{\mathrm{J}}(a,b;c;z) = \frac{1}{2\pi \mathrm{i}} \oint_{\Gamma} K_n(\zeta) f^{\mathrm{J}}(\zeta;z)\,\mathrm{d}\zeta. \tag{2.6}$$

Here, K_n is the same kernel as in (1.14), but with the parameters a,b replaced by b,c, respectively, and Γ is any contour encircling $[-1,1]$ and lying inside the cut plane $\mathbb{C}_z^*$. If $\Gamma = C_R$ with

$1 < R < |2/z - 1|$, then, as in (1.16),

$$|E_n^{\mathrm{J}}(a,b;c;z)| \leqslant |RK_n(\pm R)| \max_{\zeta \in C_R} |f^{\mathrm{J}}(\zeta;z)|, \tag{2.7}$$

with the plus or minus sign holding according as $c \leqslant 2b$ or $c > 2b$. It remains to determine the maximum on the right.

Writing $z = x + \mathrm{i}y$, $x < 1$ and $\zeta = R\mathrm{e}^{\mathrm{i}\theta}$, we have

$$|1 - \tfrac{1}{2}z(1+\zeta)| = h(\theta),$$

$$h^2(\theta) = 1 - x + \tfrac{1}{4}|z|^2(R^2+1) - R[(x - \tfrac{1}{2}|z|^2)\cos\theta - y\sin\theta], \quad 0 \leqslant \theta < 2\pi.$$

If we let

$$\begin{aligned} \mu_R^+(z) &= \max_{\zeta \in C_R} |1 - \tfrac{1}{2}z(1+\zeta)|, \\ \mu_R^-(z) &= \min_{\zeta \in C_R} |1 - \tfrac{1}{2}z(1+\zeta)|, \end{aligned} \tag{2.8}$$

we have

$$[\mu_R^{\pm}(z)]^2 = 1 - x + \tfrac{1}{4}|z|^2(R^2+1) \pm R\sqrt{(x - \tfrac{1}{2}|z|^2)^2 + y^2},$$

which simplifies to

$$\mu_R^{\pm}(z) = \tfrac{1}{2}\big|\,|z-2| \pm R|z|\,\big|. \tag{2.9}$$

It follows from (2.5) that

$$\max_{\zeta \in C_R} |f^{\mathrm{J}}(\zeta;z)| = \begin{cases} [\mu_R^+(z)]^{-a} & \text{if } a < 0, \\ [\mu_R^-(z)]^{-a} & \text{if } a > 0. \end{cases} \tag{2.10}$$

Combining (2.7) and (2.10), and optimizing the error bound at the same time, we get

$$|E_n^{\mathrm{J}}(a,b;c;z)| \leqslant \begin{cases} \min\limits_R |RK_n(\pm R)|\,[\mu_R^+(z)]^{-a} & \text{if } a < 0, \\ \min\limits_R |RK_n(\pm R)|\,[\mu_R^-(z)]^{-a} & \text{if } a > 0, \end{cases} \tag{2.11}$$

where the minima are taken over all R with $1 < R < |2/z - 1|$.

We illustrate (2.11) for real $z = x$ and $a = 1.5$, $b = 0.5$, $c = 2.5$. In this case, the second line of (2.11) applies, and the kernel K_n is to be evaluated at $-R$, since $c > 2b$. Moreover, $\mu_R^-(x) = 1/2(2 - x - R|x|)$, and the admissible values of R are $1 < R < 1 + 2/|x|$ when $x < 0$, and $1 < R < 2/x - 1$ when $0 < x < 1$. Table 2 shows bounds (2.11) divided by $\gamma_0^{\mathrm{J}}(b,c)|F(a,b;c;x)|$ or by $\gamma_0^{\mathrm{J}}(b,c)$ (if $|F(a,b;c;x)|$ is larger (resp. smaller) than 1) for selected values of x and n.

As z approaches the imaginary axis, the radius R of the circle C_R approaches 1 and the computation of $K_n(\pm R)$ becomes more difficult. The use of elliptic contours would alleviate this problem somewhat; cf. the discussion in the next subsection.

Table 2
Optimized error bounds for $a = 1.5$, $b = 0.5$, $c = 2.5$

x	$n = 5$	$n = 10$	$n = 20$	$n = 50$	$n = 100$
0.9	0.22e + 1	0.70e − 2	0.36e − 7	0.11e − 23	0.11e–51
0.6	0.63e − 4	0.52e − 10	0.15e − 22	0.83e − 61	0.41e − 125
0.3	0.27e − 8	0.21e − 18	0.54e − 39	0.18e − 101	0.41e − 206
−1	0.13e − 5	0.72e − 13	0.94e − 28	0.42e − 73	0.32e − 149
−5	0.12e − 1	0.50e − 5	0.37e − 12	0.36e − 34	0.22e − 71
−10	0.16e + 0	0.72e − 3	0.70e − 8	0.15e − 23	0.40e − 50
−20	0.11e + 1	0.28e − 1	0.93e − 5	0.91e − 16	0.14e − 34
−50	0.63e + 1	0.78e + 0	0.62e − 2	0.94e − 9	0.14e − 20
−100	0.15e + 2	0.42e + 1	0.17e + 0	0.34e − 5	0.19e − 13

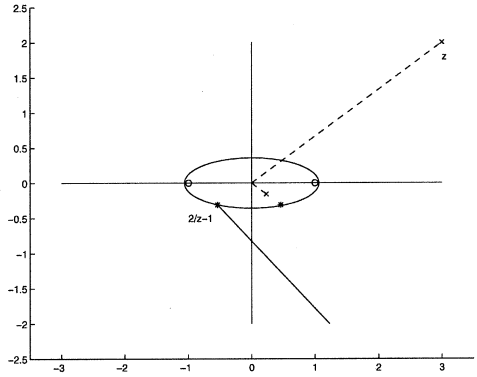


Fig. 5. The limiting ellipse $\mathscr{E}_{\rho^*}$.

2.3.2. The case $\operatorname{Re} z \geqslant 1$

Here, the cut of $\mathbb{C}_z^*$ intrudes into the unit disc, and we can no longer use circular contours Γ in (2.6). We use instead elliptic contours $\Gamma = \mathscr{E}_\rho$, $\rho > 1$, i.e., ellipses with foci at ± 1 and sum of the semiaxes equal to ρ (cf. [7], Section 5). The parameter ρ must be selected sufficiently small for $\mathscr{E}_\rho$ to avoid the cut. The limiting ellipse $\mathscr{E}_{\rho^*}$ is the one that passes through the end point $z^* = 2/z - 1$ of the cut (see Fig. 5).

To determine ρ^*, let $z^* = x^* + \mathrm{i}y^*$. The ellipse $\mathscr{E}_\rho$ in parametric form is

$$\mathscr{E}_\rho = \{\zeta \in \mathbb{C} : \zeta = \tfrac{1}{2}(\rho \mathrm{e}^{\mathrm{i}\theta} + \tfrac{1}{\rho}\mathrm{e}^{-\mathrm{i}\theta}),\ 0 \leqslant \theta \leqslant 2\pi\}.$$

In order for z^* to be on $\mathscr{E}_\rho$, we must have

$$\left(\rho + \frac{1}{\rho}\right)\cos\theta = 2x^*,$$

$$\left(\rho - \frac{1}{\rho}\right)\sin\theta = 2y^*$$

Table 3
Values of ρ^* for selected values of r and φ

r	$\varphi = \pi/2$	$\varphi = 3\pi/8$	$\varphi = \pi/4$
1	2.4142136	1.8708684	1.4966058
5	1.7968705	1.5535565	1.3409415
10	1.5395351	1.3923775	1.2506770
20	1.3645232	1.2727733	1.1787546
50	1.2197560	1.1680016	1.1123979
100	1.1513638	1.1167928	1.0788811

for some θ, which implies

$$\frac{4x^{*2}}{(\rho+1/\rho)^2}+\frac{4y^{*2}}{(\rho-1/\rho)^2}=1.$$

This amounts to an algebraic equation of degree 4 in ρ^2, namely

$$\rho^8-4(x^{*2}+y^{*2})\rho^6+2(4x^{*2}-4y^{*2}-1)\rho^4-4(x^{*2}+y^{*2})\rho^2+1=0. \tag{2.12}$$

On geometric grounds, there must be a unique real root larger than 1. This root is the desired critical value ρ^* of the parameter ρ. The admissible elliptic contours Γ therefore are the confocal ellipses $\mathscr{E}_\rho$ with $1<\rho<\rho^*$. If, as above, we write $z=1+re^{i\varphi}$, then $z^*=(1-re^{i\varphi})/(1+re^{i\varphi})$, i.e.,

$$x^*=-\frac{r^2-1}{r^2+2\cos\varphi+1},\quad y^*=-\frac{2\sin\varphi}{r^2+2\cos\varphi+1}.$$

The real root $\rho^*>1$ of (2.12) for these values of x^* and y^* is shown in Table 3 for selected values of r and φ.

Evidently, as φ decreases to 0, that is, z approaches the cut $[1,\infty]$, the value of ρ^* tends to 1, i.e., the ellipses $\mathscr{E}_{\rho^*}$ become progressively slimmer and degenerate to the interval $[-1,1]$ in the limit. As a consequence, convergence of the quadrature rule slows down, as was observed experimentally in Fig. 4.

We now have from (2.6) that

$$E_n^{\mathrm{J}}(a,b;c;z)=\frac{1}{2\pi i}\oint_{\mathscr{E}_\rho}K_n(\zeta)f^{\mathrm{J}}(\zeta;z)\,d\zeta,\quad 1<\rho<\rho^*. \tag{2.13}$$

To avoid the determination of $\max_{\zeta\in\mathscr{E}_\rho}|f^{\mathrm{J}}(\zeta;z)|$, which is rather cumbersome, we estimate E_n^{J} as follows:

$$|E_n^{\mathrm{J}}(a,b;c;z)|\leqslant\frac{1}{2\pi}\max_{\zeta\in\mathscr{E}_\rho}|K_n(\zeta)|\oint_{\mathscr{E}_\rho}|f^{\mathrm{J}}(\zeta;z)|\,|d\zeta|,\quad 1<\rho<\rho^*. \tag{2.14}$$

Here, the integral on the right can be approximated by applying the composite trapezoidal rule to

$$\oint_{\mathscr{E}_\rho}|f^{\mathrm{J}}(\zeta;z)|\,|d\zeta|=\frac{1}{2}\int_0^{2\pi}|f^{\mathrm{J}}(\zeta;z)|\sqrt{\rho^2-2\cos 2\theta+\rho^{-2}}\,d\theta \tag{2.15}$$

and the maximum of the kernel can be computed numerically.

Table 4
Relative error bounds for the hypergeometric function $F(a,b;c;z)$, $a=1.5$, $b=0.5$, $c=2.5$ and $z=re^{i\varphi}$, $\varphi=\pi/2,\ 3\pi/8,\ \pi/4$

r	$n=25$	$n=50$	$n=100$	$n=200$
1	0.24e − 15	0.89e − 32	0.12e − 64	0.22e − 130
	0.41e − 10	0.14e − 21	0.15e − 44	0.18e − 90
	0.16e − 5	0.85e − 13	0.25e − 27	0.22e − 56
5	0.53e − 10	0.10e − 20	0.39e − 42	0.58e − 85
	0.55e − 7	0.61e − 15	0.73e − 31	0.10e − 62
	0.66e − 4	0.38e − 9	0.13e − 19	0.14e − 40
10	0.29e − 7	0.47e − 15	0.12e − 30	0.84e − 62
	0.37e − 5	0.44e − 11	0.61e − 23	0.12e − 46
	0.78e − 3	0.84e − 7	0.96e − 15	0.13e − 30
20	0.31e − 5	0.86e − 11	0.66e − 22	0.39e − 44
	0.97e − 4	0.50e − 8	0.13e − 16	0.96e − 34
	0.48e − 2	0.60e − 5	0.93e − 11	0.23e − 22
50	0.16e − 3	0.50e − 7	0.46e − 14	0.39e − 28
	0.15e − 2	0.28e − 5	0.94e − 11	0.11e − 21
	0.24e − 1	0.33e − 3	0.59e − 7	0.20e − 14
100	0.95e − 3	0.31e − 5	0.34e − 10	0.40e − 20
	0.48e − 2	0.56e − 4	0.74e − 8	0.13e − 15
	0.48e − 1	0.24e − 2	0.52e − 5	0.25e − 10

As an illustration, we implemented this for $a = 1.5$, $b = 0.5$, $c = 2.5$, and for $z = 1 + re^{i\varphi}$ with the same values of r and φ as in Table 3. It is found, in this case, that the maximum of $K_n(\zeta)$ in (2.14) is consistently attained at the left extreme point $-(\rho + 1/\rho)$ of the ellipse $\mathscr{E}_\rho$. Moreover, using a composite trapezoidal rule on (2.15) with 100 subintervals is found to yield at least 2–3 correct decimal digits for all ρ not too close to ρ^*, specifically for $\rho = 1 + \lambda(\rho^* - 1)$, $\lambda = 0.1(0.1)0.8$. Invariably, the bound in (2.14) is found to decrease as λ increases through these values. In Table 4, we show the error bound in (2.14) for the ρ-value corresponding to $\lambda = 0.8$. The three vertical entries for each r and n correspond to the three values $\pi/2$, $3\pi/8$, and $\pi/4$ of φ. A true optimization of the bound over all $1 < \rho < \rho^*$, similarly as was done in (1.19), is not feasible in this case, since the optimum seems to occur at a value of ρ very close to ρ^*, for which the numerical evaluation of integral (2.15) becomes unreliable. Table 4 actually shows the error bound for $\rho = 1 + 0.8(\rho^* - 1)$ divided by $\gamma_0^J(b,c)|F(a,b;c;z)|$ if $|F(a,b;c;z)| > 1$, or divided by $\gamma_0^J(b,c)$ otherwise. This represents a bound on the relative (resp. absolute) error of the quadrature approximation to $F(a,b;c;z)$; cf. the discussion at the end of Section 1.4.

3. Concluding remarks

It has been shown that Gaussian quadrature applied to the integral representation of confluent hypergeometric and hypergeometric functions is a powerful tool to evaluate these functions in large domains of the complex plane. An inherent limitation of this approach is the restriction of the parameters a, b, and c, if they are real, to satisfying the inequalities $b > a > 0$ (resp. $c > b > 0$). The evaluation of these functions for other real values of the parameters can be accomplished, in principle,

by using appropriate recurrence relations (see, e.g., [1, Eqs. 13.4.1–13.4.7 and 15.2.10–15.2.27]). Complex values of the parameters are also accessible to our approach, but would require complex Gauss–Jacobi quadrature rules; see, e.g., Nuttal and Wherry [9], who use such rules in scattering theory, or Theocaris and Ioakimidis [10], who use them in elasticity theory. Further investigation of this would be interesting, but is beyond the scope of the present paper.

References

[1] M. Abramowitz, I.A. Stegun (Eds.), Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. NBS Applied Mathematics Series, Vol. 55, U.S. Government Printing Office, Washington, DC, 1964.
[2] P.J. Davis, P. Rabinowitz, Methods of Numerical Integration, 2nd Edition, Academic Press, Orlando, FL, 1984.
[3] J.D. Donaldson, D. Elliott, A unified approach to quadrature rules with asymptotic estimates of their remainders, SIAM J. Numer. Anal. 9 (1972) 573–602.
[4] D. Elliott, Uniform asymptotic expansions of the Jacobi polynomials and an associated function, Math. Comp. 25 (1971) 309–315.
[5] W. Gautschi, Computational aspects of three-term recurrence relations, SIAM Rev. 9 (1967) 24–82.
[6] W. Gautschi, Algorithm 726: ORTHPOL—a package of routines for generating orthogonal polynomials and Gauss-type quadrature rules, ACM Trans. Math. Software 20 (1994) 21–62.
[7] W. Gautschi, R.S. Varga, Error bounds for Gaussian quadrature of analytic functions, SIAM J. Numer. Anal. 20 (1983) 1170–1186.
[8] L.N. Karmazina, On a method of computation of the hypergeometric function, Vyčisl. Mat. Vyčisl. Tehn. 2 (1955) 111–115 (in Russian).
[9] J. Nuttal, C.J. Wherry, Gaussian integration for complex weight functions, J. Inst. Math. Appl. 21 (1978) 165–170.
[10] P.S. Theocaris, N.I. Ioakimidis, On the numerical solution of Cauchy type singular integral equations and the determination of stress intensity factors in case of complex singularities, Z. Angew. Math. Phys. 28 (1977) 1085–1098.

9.13. [169] "COMPUTATION OF BESSEL AND AIRY FUNCTIONS AND OF RELATED GAUSSIAN QUADRATURE FORMULAE"

---

[169] "Computation of Bessel and Airy Functions and of Related Gaussian Quadrature Formulae," *BIT* **42**, 110–118 (2002).

BIT
2002, Vol. 42, No. 1, pp. 110–118
0006-3835/02/4201-0110 $16.00

COMPUTATION OF BESSEL AND AIRY FUNCTIONS AND OF RELATED GAUSSIAN QUADRATURE FORMULAE *

WALTER GAUTSCHI

Department of Computer Sciences, Purdue University, West Lafayette, IN 47907, USA. email: wxg@cs.purdue.edu

Abstract.

Procedures are described for the high-precision calculation of the modified Bessel function $K_\nu(x)$, $0 < \nu < 1$, and the Airy function $\mathrm{Ai}(x)$, for positive arguments x, as prerequisites for generating Gaussian quadrature rules having these functions as weight function.

AMS subject classification: 33C10, 33C45, 65D20, 65D32.

Key words: Modified Bessel function, Airy function, Gaussian quadrature.

1 Introduction.

Integrals involving modified Bessel functions K_ν for $\nu = \frac{1}{3}$ and the Airy function Ai occur naturally in some physics applications (see, e.g., Gordon [5, 6]); the weight function K_0 has also found use in the asymptotic estimation of oscillatory integral transforms (see Wong [12], Gautschi [4, Example 6.1, p. 94]). Efficient evaluation of such integrals calls for Gaussian quadrature rules having these functions as weight function. Such rules (involving up to 10 points) were already given in [6] for the modified Bessel function $K_{\frac{1}{3}}$ and in [5, 9] for the Airy function. It appears, however, that the latter are in error because of an incorrect calculation of the relevant moments. Here we develop Gaussian quadrature formulae for both weight functions with up to $n = 40$ points. The main task is to find the first n recursion coefficients in the three-term recurrence relation for the relevant orthogonal polynomials. These can be obtained to arbitrary precision from the known moments by symbolic computation, or else, to standard machine precisions, by general procedures developed earlier in [3]. To apply these procedures, it is important to have routines that calculate modified Bessel functions and the Airy function to high accuracy. Such routines are described in Section 2 for Bessel functions, and in Section 3 for the Airy function. Section 4 discusses the computation of the respective Gaussian quadrature rules. An appendix contains the first 40 of the requisite recursion coefficients to 28 decimals.

*Received December 2000. Communicated by Kaj Madsen.

2 Modified Bessel functions.

Our interest is in generating values to high accuracy of the modified Bessel function $K_\nu(x)$, where $0 < \nu < 1$ and $x > 0$, as a preparation for generating the Gaussian quadrature rules in Sections 4.1 and 4.2. A case of particular interest is $\nu = \frac{1}{3}$, but the procedure we develop is applicable also for other values of ν except those close to 0 and 1. It is not our intention, here, to develop a general-purpose production code. There are a number of such codes available in the literature (see, e.g., [8, Sections 4.1, 5.1]), that—like ours—are suitable also for high-accuracy work. What we find worth observing is the apparently novel use of integral representations and related generalized Gauss–Laguerre quadratures to compute modified Bessel functions for moderately large, and large, real arguments. The approach, in fact, is potentially useful also for complex arguments.

When x is relatively small, say $0 < x \le 2$, we use, as others have done before, the representation (cf. [1, Eq. 9.6.2])

$$K_\nu(x) = \tfrac{1}{2}\pi \, \frac{I_{-\nu}(x) - I_\nu(x)}{\sin \pi\nu} \,, \tag{2.1}$$

and evaluate $I_{\pm\nu}(x)$ by Taylor expansion ([1, Eq. 9.6.10])

$$I_{\pm\nu}(x) = (\tfrac{1}{2}x)^{\pm\nu} \sum_{k=0}^{\infty} \frac{(\frac{1}{4}x^2)^k}{k!\Gamma(k+1\pm\nu)} \,. \tag{2.2}$$

If ν is close to 0 or 1, considerable cancellation occurs in the numerator of (2.1). This could be dealt with, if deemed necessary, by special additional procedures (cf., e.g., [10, Section II]). Here we simply assume $0.05 \le \nu \le 0.95$, which limits the loss of accuracy to at most two (or three, if x is near 2) decimal digits.

For $x > 2$, we use the integral representation ([1, Eq. 9.6.23])

$$K_\nu(x) = \frac{\sqrt{\pi}}{2^\nu \Gamma(\nu+\frac{1}{2})} \, \frac{e^{-x}}{\sqrt{x}} \int_0^\infty \left(2 + \frac{t}{x}\right)^{\nu-\frac{1}{2}} \cdot t^{\nu-\frac{1}{2}} e^{-t} dt, \tag{2.3}$$

where the integral is conveniently evaluated by generalized Gauss–Laguerre quadrature with parameter $\alpha = \nu - \frac{1}{2}$,

$$\int_0^\infty \left(2 + \frac{t}{x}\right)^{\nu-\frac{1}{2}} \cdot t^{\nu-\frac{1}{2}} e^{-t} dt \simeq \sum_{k=1}^{n} w_k^L \left(2 + \frac{t_k^L}{x}\right)^{\nu-\frac{1}{2}}, \quad x > 2. \tag{2.4}$$

Here, t_k^L, w_k^L are the nodes and weights of the generalized Gauss–Laguerre quadrature rule. (The dependence on n is suppressed in the notation.) These, for $\nu = \frac{1}{3}$, were generated by double-precision resp. quadruple-precision analogues of the procedures `recur` and `gauss` of [3]. While it would be unreasonable to expect convergence to full machine precision as $n \to \infty$, it was found that in double and quadruple precision, "numerical convergence" occurs to an accuracy of $10\times\varepsilon_{\text{dble}}$ resp. $1000\times\varepsilon_{\text{quad}}$, where $\varepsilon_{\text{dble}} \simeq .111\times10^{-15}$, $\varepsilon_{\text{quad}} \simeq .963\times10^{-34}$ are respectively the IEEE double- and quadruple-precision machine precisions. In other words, the approximants stabilized to these accuracies at certain values of n, which are shown in Table 2.1.

Table 2.1: Number of Gauss points in (2.4), with $\nu = \frac{1}{3}$, required for double- and quadruple-precision accuracy.

ν	0	1/6	1/3	1/2	2/3	5/6	1
n double	23	23	22	1	21	22	22
n quadruple	87	85	83	1	82	82	82

(The data is for $x = 2$; as x increases, n decreases.)

In many applications (including the one in Section 4.1; cf. (4.7)), it is better to compute $e^x K_\nu(x)$. This is also the function tabulated (for $\nu = \frac{1}{3}$) in [11, Table III].

We remark that our evaluation procedure is also applicable for x in the complex plane cut along the negative real axis, but we will not pursue this here any further.

3 The Airy function.

The Airy function is related to the modified Bessel function $K_{\frac{1}{3}}$ as follows (cf. [1, Eq. 10.4.14]):

$$\mathrm{Ai}(x) = \frac{1}{\pi}\sqrt{\frac{x}{3}}\, K_{\frac{1}{3}}(\zeta), \quad \zeta = \tfrac{2}{3}x^{\frac{3}{2}}. \tag{3.1}$$

Using (2.1) with $\nu = \frac{1}{3}$, one gets

$$\mathrm{Ai}(x) = \frac{1}{3}\sqrt{x}\,[I_{-\frac{1}{3}}(\zeta) - I_{\frac{1}{3}}(\zeta)]. \tag{3.2}$$

Here, for $0 < \zeta \le 2$, i.e., $0 < x \le 3^{2/3} = 2.08008\ldots$, both Bessel functions can be evaluated by Taylor expansion as in (2.2) with $\nu = \frac{1}{3}$.

For $\zeta > 2$, we use the integral representation (2.3) for $K_{\frac{1}{3}}$ in conjunction with (3.1) to obtain

$$\mathrm{Ai}(x) = \frac{1}{\sqrt{\pi}}\,\frac{\zeta^{-\frac{1}{6}}e^{-\zeta}}{(48)^{\frac{1}{6}}\Gamma(\frac{5}{6})}\int_0^\infty \left(2 + \frac{t}{\zeta}\right)^{-\frac{1}{6}} \cdot t^{-\frac{1}{6}}e^{-t}dt, \tag{3.3}$$

with ζ as defined in (3.1). Now generalized Gauss–Laguerre quadrature is appropriate with Laguerre parameter $\alpha = -\frac{1}{6}$. According to Table 2.1 ($\nu = \frac{1}{3}$), a 22-point formula yields double-precision accuracy and a 83-point formula quadruple-precision accuracy.

This procedure can be used also for complex x, at least in the sector $|\arg x| < \frac{2}{3}\pi$.

4 Gauss quadratures.

4.1 Gauss quadrature with Bessel weight function.

We define the weight function

$$w(x) = \frac{2}{\pi}\cos(\tfrac{1}{2}\nu\pi)K_\nu(x), \quad 0 < x < \infty. \tag{4.1}$$

Its moments can be calculated explicitly by (cf. [7, 6.561.16])

$$\mu_k = \int_0^\infty x^k w(x)dx = \frac{1}{\pi}\cos(\tfrac{1}{2}\nu\pi)\cdot 2^k\Gamma(\tfrac{1}{2}(k+1+\nu))\Gamma(\tfrac{1}{2}(k+1-\nu)). \tag{4.2}$$

In particular, for $k = 0$,

$$\mu_0 = \frac{1}{\pi}\cos(\tfrac{1}{2}\nu\pi)\cdot\Gamma(\tfrac{1}{2}(1+\nu))\Gamma(\tfrac{1}{2}(1-\nu)) = 1 \tag{4.3}$$

by virtue of the reflection formula ([1, 6.1.17]) for $\Gamma(z)\Gamma(1-z)$, $z = \frac{1}{2}(1+\nu)$. Thus, (4.1) is a normalized weight function.

In principle, the first $2n$ moments μ_k of w can be used to generate the n-point Gauss formula for w. It is well known, however, that this becomes quickly unstable as n increases. A way around this problem is to employ a symbolic computation package combined with extended-precision arithmetic. A Maple script, named `cheb.mws`[1], has been developed for this purpose and has been used to produce the required recursion coefficients α_k, β_k (cf. (4.4) below) for $0 \le k \le 39$ to as many as 100 decimal digits[2]. It can be accessed at
`http://www.cs.purdue.edu/archives/2001/wxg/codes/cheb.mws`.
(A text file `cheb.txt` can be found at the same URL.)

Here, we describe a stable numerical procedure—a four-pronged discretization procedure (cf. [4, Section 6])—that discretizes the inner product for the weight function w and generates the corresponding discrete orthogonal polynomials. If the discretization is chosen judiciously, the discrete orthogonal polynomials converge to the desired ones as the discretization is made increasingly finer. The first n recursion coefficients α_k, β_k in the three-term recurrence relation

$$\begin{gathered}\pi_{k+1}(x) = (x-\alpha_k)\pi_k(x) - \beta_k\pi_{k-1}(x), \quad k = 0, 1, \ldots, n-1,\\ \pi_0(x) = 1, \quad \pi_{-1}(x) = 0,\end{gathered} \tag{4.4}$$

for the monic orthogonal polynomials (where $\beta_0 = \int_0^\infty w(x)dx$) are computed by the Stieltjes procedure ([4, Section 6.3]).

The discretization we choose makes use of a composition of the positive real axis into three subintervals, $\mathbb{R}_+ = (0, x_0] \cup [x_0, x_1] \cup [x_1, \infty)$, with x_0, x_1 still to be selected such that $0 < x_0 \le 1$, $1 < x_1 < \infty$. In the first subinterval, the behavior of $K_\nu(x)$ for small x must be properly accounted for. One has [1, Eqs. 9.6.2 and 9.6.10]

$$K_\nu(x) = \frac{\pi}{2\sin\nu\pi}\left\{\frac{(\frac{1}{2}x)^{-\nu}}{\Gamma(1-\nu)}S_{-\nu}(x) - \frac{(\frac{1}{2}x)^{\nu}}{\Gamma(1+\nu)}S_\nu(x)\right\}, \tag{4.5}$$

where

$$S_{\pm\nu}(x) = \sum_{k=0}^\infty \frac{(\frac{1}{4}x^2)^k\Gamma(1\pm\nu)}{k!\Gamma(k+1\pm\nu)}.$$

[1] This is written for Maple.Release 5.

[2] The author is indebted to Oscar Chinellato at the Institute for Scientific Computing of the ETH Zurich, Switzerland, for translating a slightly edited version of our ORTHPOL routine `cheb` (cf. [3]) into a Maple script.

The two distinct behaviors $x^{-\nu}$ and x^{ν} as $x \to 0$ need to be treated separately for purposes of integration. Indeed, the following composition for integrals over $(0, x_0]$ is suggested:

$$\int_0^{x_0} p(x)K_\nu(x)dx = \frac{2^{\nu-1}\pi}{\sin\nu\pi \cdot \Gamma(1-\nu)} \int_0^{x_0} p(x)S_{-\nu}(x) \cdot x^{-\nu}dx - \frac{\pi}{2^{\nu+1}\sin\nu\pi \cdot \Gamma(1+\nu)} \int_0^{x_0} p(x)S_\nu(x) \cdot x^{\nu}dx. \tag{4.6}$$

The first integral on the right is approximated by an N-point Gauss–Jacobi quadrature rule relative to the interval $(0, x_0]$ with Jacobi parameters $\alpha = 0$, $\beta = -\nu$, the second by a similar N-point Gauss–Jacobi rule with parameters $\alpha = 0$, $\beta = \nu$. In the second interval, we apply the ordinary N-point Gauss–Legendre rule transformed to the interval $[x_0, x_1]$. For the last interval, we write

$$\int_{x_1}^{\infty} p(x)K_\nu(x)dx = e^{-x_1} \int_0^{\infty} p(x_1+t)[e^{x_1+t}K_\nu(x_1+t)] \cdot e^{-t}dt \tag{4.7}$$

and approximate the integral on the right by an N-point Gauss–Laguerre quadrature rule, the function $e^x K_\nu(x)$, $x \geq x_1$, being computed from the integral representation (2.3) as discussed in the text following (2.3).

We call this a "four-pronged" discretization procedure since four different quadrature rules are employed to discretize the integral $\int_0^\infty p(x)K_\nu(x)dx$ (and with it the inner product relative to the weight function (4.1)): two Gauss–Jacobi rules for approximating the two integrals on the right of (4.6), the Gauss rule over $[x_0, x_1]$, and the Gauss–Laguerre rule to deal with the integral on the right of (4.7).

The parameters x_0, x_1 are chosen in an attempt to reduce the number N of quadrature terms required to achieve a given accuracy. Some limited experimentation suggested the choice $x_0 = 1$ and $x_1 = 10$. When $\nu = \frac{1}{3}$ and $n = 10$, then $N = 41$ will yield a relative accuracy of 1000 times the double machine precision (about 12 decimal-digit accuracy), and $N = 71$ a relative accuracy of 10^5 times the quadruple machine precision (about 29 decimal digits). For $n = 40$ the respective numbers are both $N = 81$.

The values of the recursion coefficients α_k, β_k (for $\nu = \frac{1}{3}$) to 28 decimal digits are given in the appendix, Table A1, for $k = 0, 1, \ldots, 39$. These allow us to generate the respective Gauss and Gauss–Radau quadrature rules for up to 40 points by well-known eigenvalue/eigenvector techniques [4, Section 4]. Double- and quadruple-precision fortran programs producing these coefficients are accessible at `http://www.cs.purdue.edu/archives/2001/wxg/codes/` in the files `dOPbess.f` and `qOPbess.f`. The coefficients themselves to 28 decimals can be found in the file `coeffbess` at the same URL. The file `ORTHPOLq` contains the quadruple-precision routines of the package ORTHPOL in [3].

4.2 Gauss quadrature with Airy weight function.

We define a weight function proportional to the one in Eq. (1.4) of [9],

$$w(x) = \frac{2^{\frac{2}{3}}\pi}{3^{\frac{5}{6}}\Gamma(\frac{2}{3})} x^{-\frac{2}{3}} e^{-x} \mathrm{Ai}((\tfrac{3}{2}x)^{\frac{2}{3}}), \quad 0 < x < \infty. \tag{4.8}$$

By (3.1), one has

$$w(x) = \frac{2^{\frac{1}{3}}}{3\Gamma(\frac{2}{3})}\, x^{-\frac{1}{3}} e^{-x} K_{\frac{1}{3}}(x). \tag{4.9}$$

As in [9], we use [7, 6.621.3] to obtain for the moments (correctly)

$$\mu_k = \int_0^\infty x^k w(x)dx = \frac{2^{\frac{2}{3}}\sqrt{\pi}}{\Gamma(\frac{2}{3})}\, \frac{k!\Gamma(k+\frac{1}{3})}{(6k+1)2^k\Gamma(k+\frac{1}{6})}\,. \tag{4.10}$$

In particular,

$$\mu_0 = \frac{2^{\frac{2}{3}}\sqrt{\pi}}{\Gamma(\frac{2}{3})}\, \frac{\Gamma(\frac{1}{3})}{\Gamma(\frac{1}{6})} = 1 \tag{4.11}$$

by virtue of the duplication formula [1, 6.1.19] for $\Gamma(2z)$, $z = \frac{1}{6}$. Thus, the weight function (4.8) is normalized.

The three-term recurrence relation (4.4) for this weight function can again be obtained by symbolic computation using the Maple script `cheb.mws` (cf. §4.1). Numerically, on the other hand, we may use, similarly as in Section 4.1, a three-pronged discretization method. First, integration against the weight function is "regularized" by means of the change of variable $x \mapsto x^3/3$; thus,

$$\int_0^\infty p(x)w(x)dx = \frac{2^{\frac{2}{3}}\pi}{3^{\frac{1}{6}}\Gamma(\frac{2}{3})} \int_0^\infty p(\tfrac{1}{3}x^3)\mathrm{Ai}(2^{-\frac{2}{3}}x^2)\cdot e^{-x^3/3}dx. \tag{4.12}$$

The discretization is effected by using the decomposition $\mathbb{R}_+ = (0,2]\cup[2,6]\cup[6,\infty)$ and N-point Gauss–Legendre quadrature on the first two intervals, and N-point Gauss quadrature relative to the weight function $e^{-x^3/3}$, $0 < x < \infty$, on the last interval (after transforming the integral over $[6,\infty)$ to one over $[0,\infty)$). The latter quadrature rules have been generated by a "general-purpose" discretization method [4, p. 95]; see also [2]. In this way, when $n = 10$, then $N = 51$ (in double precision) yields about 12-digit accuracy, $N = 91$ (in quadruple precision) an accuracy of about 29 digits. For $n = 40$, the respective numbers are $N = 81$ and $N = 161$.

Values of the recursion coefficients α_k, β_k to 28 decimal digits are given in the appendix, Table A2, for $k = 0, 1, \ldots, 39$. These again permit the generation of Gauss and Gauss–Radau quadrature rules with up to 40 points. Double- and quadruple-precision fortran programs producing these coefficients are accessible at

`http://www.cs.purdue.edu/archives/2001/wxg/codes/` in the files`dOPairy.f` and `qOPairy.f`. The coefficients themselves to 28 decimals can be found in the file `coeffairy` at the same URL.

Acknowledgement.

The author gratefully acknowledges helpful comments from two anonymous referees.

A Appendix.

Recursion coefficients α_k and β_k for the orthogonal polynomials relative to the weight function $\frac{\sqrt{3}}{\pi}K_{\frac{1}{3}}(x)$, $0 < x < \infty$, are given in Table A.1.

Table A.1: Recursion coefficients for orthogonal polynomials with Bessel weight function.

k	dalpha(k)	dbeta(k)
0	0.57735026918962576450914878O5D+00	0.10000000000000000000000000000D+01
1	0.25403411844343533636840254634D+01	0.55555555555555555555555555556D+00
2	0.45303250992775569729406193 92D+01	0.30800000000000000000000000000D+01
3	0.65252632053670247545844211 46D+01	0.75973088960101947114934127 92D+01
4	0.85220916815723164954105718 28D+01	0.14111288950643590484767646 85D+02
5	0.10519870690325796817093273 97D+02	0.22623283756519972141184890 43D+02
6	0.12518205342004118426986613 10D+02	0.33133934495118279090152051 75D+02
7	0.14516897450708475899825403 47D+02	0.45643601984576447974202422 03D+02
8	0.16515835338909712945017052 03D+02	0.60152511693871654445741830 86D+02
9	0.18514950703700216501467869 56D+02	0.76660815122790272370295821 99D+02
10	0.20514199134596376599961869 34D+02	0.95168619641531512764202519 70D+02
11	0.22513550350966397331553300 55D+02	0.11567600451295364022414377 43D+03
12	0.24512982906925447589739452 21D+02	0.13818303016614297368305199 92D+03
13	0.26512481134031380745844017 42D+02	0.16268974388883212711028968 29D+03
14	0.28512033284217889886196819 52D+02	0.18919618349092681270044297 81D+03
15	0.30511630353379984494986073 62D+02	0.21770237975590841525221055 42D+03
16	0.32511265309289715164513256 88D+02	0.24820835813554385367738681 70D+03
17	0.34510932569315403938872506 2D+02	0.28071413995440991178675067 45D+03
18	0.36510627637760244734914891 60D+02	0.31521974328668330503422661 85D+03
19	0.38510346848218176853857670 70D+02	0.35172518360774645238197113 69D+03
20	0.40510087176814417926038440 69D+02	0.39023047428733980675972347 09D+03
21	0.42509846104386257094909113 93D+02	0.43073562696885280867041753 67D+03
22	0.44509621513140441276089535 58D+02	0.47324065186525903076257111 71D+03
23	0.46509411608040199994936432 28D+02	0.51774555799300470520410315 95D+03
24	0.48509214856221347986475885 26D+02	0.56425035335901673936349506 13D+03
25	0.50509029939747541853289263 80D+02	0.61275504511180750405912503 02D+03
26	0.52508855718368037952718831 28D+02	0.66325963966474388749188552 82D+03
27	0.54508691199868481800078655 62D+02	0.71576414279749246320929380 47D+03
28	0.56508535516250920027796913 10D+02	0.77026855974017792568324480 17D+03
29	0.58508387904435593094235233 07D+02	0.82677289524371904890660617 45D+03
30	0.60508247690504090767465010 85D+02	0.88527715363901570667787427 62D+03
31	0.62508114276740777255813280 07D+02	0.94578133888707076553849392 25D+03
32	0.64507987130903653941908285 50D+02	0.10082854546216859666327285 10D+04
33	0.66507865777285184445942878 97D+02	0.10727895041860321418401763 89D+04
34	0.68507749789220609720584340 23D+02	0.11392934906641335695677578 33D+04
35	0.70507638782774714709547449 45D+02	0.12077974169081041187988146 82D+04
36	0.72507532411394094660919210 10D+02	0.12783012855618146698424410 05D+04
37	0.74507430361355164228014596 83D+02	0.13508050990815465982217713 77D+04
38	0.76507332347871680219924933 44D+02	0.14253088597540870680261682 99D+04
39	0.78507238111751765159290348 05D+02	0.15018125697126426815746413 94D+04

Recursion coefficients α_k and β_k for the orthogonal polynomials relative to the weight function $\frac{2^{2/3}\pi}{3^{5/6}\Gamma(2/3)}\, x^{-\frac{2}{3}} e^{-x} \mathrm{Ai}\big((\frac{3}{2}x)^{\frac{2}{3}}\big)$, $0 < x < \infty$, are given in Table A.2.

Table A.2: Recursion coefficients for orthogonal polynomials with Airy weight function.

k	dalpha(k)	dbeta(k)
0	0.14285714285714285714285714 29D+00	0.10000000000000000000000000 00D+01
1	0.11105088302150725651337646 44D+01	0.67503924646781789638932496 08D-01
2	0.21039232623033664697190466 50D+01	0.61764355206314549464773663 22D+00
3	0.31006129236194680602040417 49D+01	0.16652223695797287639935981 04D+01
4	0.40985229900533503002596773 56D+01	0.32116654379338133014426731 08D+01
5	0.50970480179122379522111913 66D+01	0.52574397914282669985698940 03D+01
6	0.60959347721427352894050676 51D+01	0.78027628718161412450789162 30D+01
7	0.70950557426151749696957611 15D+01	0.10847755827663314637317852 13D+02
8	0.80943387049797903045725228 53D+01	0.14392494051279503008919829 57D+02
9	0.90937392474629931366999661 23D+01	0.18437028129221289287730985 12D+02
10	0.10093228346099565290956749 25D+02	0.22981393910509113133201030 04D+02
11	0.11092786118994277963895130 21D+02	0.28025617876564395301895131 39D+02
12	0.12092398425735241831496679 65D+02	0.33569720237015983129850790 15D+02
13	0.13092054896137796054305526 64D+02	0.39613716825662470924767630 52D+02
14	0.14091747728196667776798864 68D+02	0.46157620319174718546589178 80D+02
15	0.15091470923783667057958572 63D+02	0.53201441052135684780032004 69D+02
16	0.16091219785281719839537880 68D+02	0.60745187580452629573767385 63D+02
17	0.17090990573969440137935584 92D+02	0.68788867081900320364309351 16D+02
18	0.18090780272089838774453518 34D+02	0.77332485647813606942498593 41D+02
19	0.19090586413340360720405440 31D+02	0.86376048499994747266102355 45D+02
20	0.20090406959676781196253599 98D+02	0.95919560154988984853375918 88D+02
21	0.21090240210178251886147338 37D+02	0.10596302455052778144792737 89D+03
22	0.22090084732554680300801113 79D+02	0.11650644514426334671872770 61D+03
23	0.23089939310933302448587962 06D+02	0.12754982499186694087724614 53D+03
24	0.24089802905539742800085714 54D+02	0.13909316680952576066464148 52D+03
25	0.25089674621197700228155754 63D+02	0.15113647302448387041451470 93D+03
26	0.26089553682454295865206351 28D+02	0.16367974581630744705402971 92D+03
27	0.27089439413744318231582362 92D+02	0.17672298715087261457351113 46D+03
28	0.28089331223429595213827617 89D+02	0.19026619880858475806527647 75D+03
29	0.29089228590849281318522073 32D+02	0.20430938240798204060362816 48D+03
30	0.30089131055731901297556390 56D+02	0.21885253942561325212625418 85D+03
31	0.31089038209476330674573859 95D+02	0.23389567121288418866806529 98D+03
32	0.32088949687923881056051201 54D+02	0.24943877901041894670919058 99D+03
33	0.33088865165329150952378707 07D+02	0.26548186396036985718116802 51D+03
34	0.34088784349301510252239270 20D+02	0.28202492711702309565791942 38D+03
35	0.35088706976537763785392615 56D+02	0.29906796945597970768703581 57D+03
36	0.36088632809203769272943704 07D+02	0.31661099188213910906157393 19D+03
37	0.37088561631851495611747249 58D+02	0.33465399523667056339923916 24D+03
38	0.38088493248780320814079768 38D+02	0.35319698030312511600256047 49D+03
39	0.39088427481768837660306608 97D+02	0.37223994781281404077676699 42D+03

REFERENCES

1. M. Abramowitz and I. A. Stegun (eds.), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, NBS Applied Mathematics Series, Vol. 55, U.S. Government Printing Office, Washington, DC, 1964.
2. W. Gautschi, *How and how not to check Gaussian quadrature formulae*, BIT, 23 (1983), pp. 209–216.
3. W. Gautschi, *Algorithm 726: ORTHPOL—A package of routines for generating orthogonal polynomials and Gauss-type quadrature rules*, ACM Trans. Math. Software, 20 (1994), pp. 21–62.
4. W. Gautschi, *Orthogonal polynomials: applications and computation*, Acta Numerica, 5 (1996), pp. 45–119.
5. R. G. Gordon, *New method for constructing wavefunctions for bound states and scattering*, J. Chem. Phys., 51 (1969), pp. 14–25.
6. R. G. Gordon, *Constructing wavefunctions for nonlocal potentials*, J. Chem. Phys., 52 (1970), pp. 6211–6217.
7. I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series, and Products*, Academic Press, San Diego, CA, 2000.
8. D. W. Lozier and F. W. J. Olver, *Numerical evaluation of special functions*, in Mathematics of Computation 1943–1993: A half-century of computational mathematics, Vancouver, BC, 1993, W. Gautschi, ed., Proc. Sympos. Appl. Math., Vol. 48, Amer. Math. Soc., Providence, RI, 1994, pp. 79–125.
9. Z. Schulten, D. G. M. Anderson, and R. G. Gordon, *An algorithm for the evaluation of the complex Airy functions*, J. Comput. Phys., 31 (1979), pp. 60–75.
10. N. M. Temme, *On the numerical evaluation of the modified Bessel function of the third kind*, J. Comput. Phys., 19 (1975), pp. 324–337.
11. G. N. Watson, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge University Press, Cambridge, 1958.
12. R. Wong, *Quadrature formulas for oscillatory integral transforms*, Numer. Math., 39 (1982), pp. 351–360.

9.14. [178] "NUMERICAL QUADRATURE COMPUTATION OF THE MACDONALD FUNCTION FOR COMPLEX ORDERS"

---

[178] "Numerical Quadrature Computation of the Macdonald Function for Complex Orders," *BIT Numer. Math.* **45**, 593–603 (2005).

BIT Numerical Mathematics (2005) 45: 593–603
DOI: 10.1007/s10543-005-0020-5

NUMERICAL QUADRATURE COMPUTATION OF THE MACDONALD FUNCTION FOR COMPLEX ORDERS*

WALTER GAUTSCHI[1]

[1] *Department of Computer Sciences, Purdue University, West Lafayette, IN 47907, USA. email: wxg@cs.purdue.edu*

Abstract.

The use of Gaussian quadrature formulae is explored for the computation of the Macdonald function (modified Bessel function) of complex orders and positive arguments. It is shown that for arguments larger than one, Gaussian quadrature applied to the integral representation of this function is a viable approach, provided the (nonclassical) weight function is suitably chosen. In combination with Gauss–Legendre quadrature the approach works also for arguments smaller than one. For very small arguments, power series can be used. A Matlab routine is provided that implements this approach.

AMS subject classification (2000): 33-04, 33C10, 65D15, 65D32.

Key words: Macdonald function, modified Bessel function, complex order, Gauss quadrature approximation, Matlab software.

1 Introduction.

The Macdonald function (or modified Bessel function) $K_\nu(x)$ with complex order $\nu = \alpha + \mathrm{i}\beta$ and positive argument $x > 0$ is of some importance in a number of applied areas. Little attention, nevertheless, seems to have been paid to developing computational procedures and software for this function, except when the order $\nu = \mathrm{i}\beta$ is purely imaginary. In this case a variety of methods have been studied in [3] for extended domains of the (x, β)-plane; related software is available in [6]. Generally, among the most prominent methods used to compute special functions are power series expansions, asymptotic expansions, and continued fractions. Here we promote the use of numerical quadrature, a technique often neglected in the literature; see, however, [9, Ch. IV, §6] and [5].

The point of departure is the pair of integral representations [1, Eq. 9.6.24]

$$\begin{aligned} \operatorname{Re} K_{\alpha+\mathrm{i}\beta}(x) &= \int_0^\infty e^{-x\cosh t} \cosh \alpha t \cos \beta t \, \mathrm{d}t, \\ \operatorname{Im} K_{\alpha+\mathrm{i}\beta}(x) &= \int_0^\infty e^{-x\cosh t} \sinh \alpha t \sin \beta t \, \mathrm{d}t. \end{aligned} \tag{1.1}$$

* Received December 2004. Accepted April 2005. Communicated by Tom Lyche.

Clearly, it suffices to consider $\alpha \geq 0$ and, in view of the recurrence relation for K_ν (which is stable in forward direction of ν), $0 \leq \alpha < 2$. Large values of β give rise to rapid oscillations and hence to numerical difficulties. A proper computational treatment of this case, either by asymptotics or deformation of the path of integration, is beyond the scope of this paper. Instead we shall assume $0 \leq |\beta| \leq 10$. With α and β thus restricted, a numerical quadrature procedure will be developed which is effective for essentially all $x > 0$ of practical interest. A Matlab implementation of the procedure yields at least 9, and usually more, correct significant digits (except near zeros), and a quadruple-precision Fortran routine about 26 digits or more. To make our routine applicable also when $|\beta| > 10$, we rely, in this case, on the symbolic Matlab `mfun`-routine `BesselK`.

In trying to use numerical integration in (1.1), one has to be cognizant of the extremely rapid decay of the first factor in the integrands as $t \to \infty$. We shall deal with this by employing Gaussian quadrature relative to the weight function $w(t) = \exp(-e^t)$ on $[0, \infty]$. This turns out to be effective when x is not small, say $x \geq 1$. Naturally, these quadrature formulae are not classical and must be generated numerically. Suitably decomposing the interval of integration into two parts, $[0, \infty] = [0, c] \cup [c, \infty]$, and using Gauss–Legendre quadrature over the first, and the newly generated Gaussian quadratures over the second interval, allows us to deal also with considerably smaller values of x, say $x \geq .01$. Power series expansions can be used for values of x still smaller.

Parameter values of particular interest are $\alpha = 0$ and $\alpha = \frac{1}{2}$, which yield the kernels in the ordinary and modified Kontorovich–Lebedev integral transforms, respectively [2, Ch. 3, §3], [7]. These in turn have been used for the solution of Dirichlet and other boundary value problems in wedge-shaped domains.

In §2 we discuss how the new Gaussian quadrature formulae can be generated. In §3 we develop the computational procedure for $x \geq 1$, in §5 for $.01 < x < 1$, and in §7 for $0 < x \leq .01$. Numerical results will be presented in §§4 and 6, plots of $K_{\mathrm{i}\beta}$ and $K_{1/2+\mathrm{i}\beta}$ in §8.

All pieces of software referenced in this paper, including the Matlab package `OPQ`, can be downloaded from the web site

`http://www.cs.purdue.edu/archives/2002/wxg/codes/`

2 Gauss quadrature with weight function $w(t) = \exp(-e^t)$ on $[0, \infty]$.

We need to construct the orthogonal polynomials with respect to the weight function w, that is, the coefficients α_k, β_k in the three-term recurrence relation

$$\begin{gathered} \pi_{k+1}(t) = (t - \alpha_k)\pi_k(t) - \beta_k \pi_{k-1}(t), \quad k = 0, 1, 2, \ldots, \\ \pi_{-1}(t) = 0, \quad \pi_0(t) = 1, \end{gathered} \tag{2.1}$$

satisfied by the orthogonal polynomials $\pi_k(\,\cdot\,) = \pi_k(\,\cdot\,; w)$. Once these coefficients are known, the n-point Gauss formula for w,

$$\int_0^\infty f(t)w(t)\mathrm{d}t = \sum_{\nu=1}^{n} \lambda_\nu^G f(\tau_\nu^G) + R_n^G(f), \tag{2.2}$$

is readily obtained in terms of eigenvalues and eigenvectors of the $n \times n$ Jacobi matrix

$$J_n(w) = \begin{bmatrix} \alpha_0 & \sqrt{\beta_1} & & \mathbf{0} \\ \sqrt{\beta_1} & \alpha_1 & \ddots & \\ & \ddots & \ddots & \sqrt{\beta_{n-1}} \\ \mathbf{0} & & \sqrt{\beta_{n-1}} & \alpha_{n-1} \end{bmatrix}. \tag{2.3}$$

Indeed (cf. [4, §3.1.1.1]), τ_ν^G are the eigenvalues of J_n and $\lambda_\nu^G = \beta_0 v_{\nu,1}^2$, where $\beta_0 = \int_0^\infty w(t)\mathrm{d}t$ and $v_{\nu,1}$ is the first component of the normalized eigenvector v_ν corresponding to the eigenvalue τ_ν^G.

To compute $J_n(w)$ and β_0, we apply a multiple-component discretization method ([4, §2.2.4]) based on the decomposition $[0,\infty] = [0,.75] \cup [.75,1.5] \cup [1.5,3] \cup [3,\infty]$ and the use of Fejér quadrature rules as a general-purpose means of discretization. This is implemented in the following Matlab routine using the package `OPQ`; the weight function w has to be properly identified at the end of the file `quadgp.m`.

```
global mc mp iq idelta irout DM uv AB
N=100; eps0=.5e-12;
mc=4; mp=0; iq=0; idelta=2; Mmax=900;
AB=[[0 .75];[.75 1.5];[1.5 3];[3 Inf]];
[ab,Mcap,kount]=mcdis(N,eps0,@quadgp,Mmax);
```

The choices made of `N` and `eps0` led to `Mcap` = 801 and `kount` = 6 iterations. The results are stored in the `N`×2 array `abmacdonald`. The routine

```
load -ascii abmacdonald;
xw=gauss(n,abmacdonald);
```

then produces the `n`-point Gauss formula (`n` ≤ `N`) with the nodes and weights stored in the first resp. second column of the `n`×2 array `xw`.

More accurate values of the first 100 recurrence coefficients, produced by a quadruple-precision Fortran program, are provided to 30 decimal digits in the file `qabmacdonald`.

3 The function $K_\nu(x)$ for $x \geq 1$.

The rapidly decaying factor $\exp(-x\cosh t)$ in (1.1), as already mentioned, ought to be treated as a weight function. To avoid its dependence on x, we write

$$e^{-x\cosh t} = e^{-\frac{1}{2}xe^t} \cdot e^{-\frac{1}{2}xe^{-t}} = e^{-\frac{1}{2}x} \cdot e^{-\frac{1}{2}x(e^t-1)} \cdot e^{-\frac{1}{2}xe^{-t}} \tag{3.1}$$

and define a new variable u by

$$\tfrac{1}{2}x(e^t-1)=e^u-1, \tag{3.2}$$

so that $0\le u\le\infty$ when $0\le t\le\infty$. Letting

$$h(u)=1+\left(\tfrac{1}{2}x-1\right)e^{-u}, \tag{3.3}$$

one computes

$$\begin{aligned}
&e^{-x\cosh t}=e^{1-\frac{1}{2}x}e^{-e^u}e^{-\frac{1}{4}x^2e^{-u}/h},\\
&\cosh\alpha t=\tfrac{1}{2}(2h/x)^\alpha e^{\alpha u}\left(1+(2h/x)^{-2\alpha}e^{-2\alpha u}\right),\\
&\cos\beta t=\cos\beta(u+\ln(2h/x)),\\
&\mathrm{d}t=\mathrm{d}u/h,
\end{aligned} \tag{3.4}$$

with a similar formula for $\sinh\alpha t$ having a minus sign in place of the plus sign. Consequently, from (1.1),

$$\begin{aligned}
\operatorname{Re}K_{\alpha+\mathrm{i}\beta}(x)&=\tfrac{1}{2}(2/x)^\alpha e^{1-\frac{1}{2}x}\int_0^\infty e^{-e^u}f(u)\mathrm{d}u,\\
\operatorname{Im}K_{\alpha+\mathrm{i}\beta}(x)&=\tfrac{1}{2}(2/x)^\alpha e^{1-\frac{1}{2}x}\int_0^\infty e^{-e^u}g(u)\mathrm{d}u,
\end{aligned} \tag{3.5}$$

where

$$\begin{aligned}
f(u)&=e^{\alpha u-\frac{1}{4}x^2e^{-u}/h}h^{-(\alpha+1)}\left(h^{2\alpha}+(xe^{-u}/2)^{2\alpha}\right)\cos\beta(u+\ln(2h/x)),\\
g(u)&=e^{\alpha u-\frac{1}{4}x^2e^{-u}/h}h^{-(\alpha+1)}\left(h^{2\alpha}-(xe^{-u}/2)^{2\alpha}\right)\sin\beta(u+\ln(2h/x)),
\end{aligned} \tag{3.6}$$

with $h=h(u)$ defined in (3.3). Thus, both integrals in (3.5) can be evaluated by the Gauss formulae developed in §2.

4 Numerical results for $x\ge 1$.

Given $x\ge 1$ and (without restriction of generality) $\beta\ge 0$, we need to determine the smallest order n of the Gauss quadrature rule that yields results to a prescribed relative accuracy ε_0 when applied to (3.5). For $\varepsilon_0=\frac{1}{2}\times 10^{-9}$, numerical experiments were conducted with $\alpha=0:\frac{1}{2}:2$, $\beta=0:10$, and selected values of x between 1 and 100. They revealed that relatively low-order Gauss formulae suffice to achieve the desired accuracy. The number n of Gauss points required has consistently been ≤ 29 and usually much smaller. This is illustrated in Table 4.1 for a typical value of β and the two values $\alpha=0$, $\alpha=\frac{1}{2}$. The integers n_r and n_i refer to real and imaginary part, respectively.

It thus appears that in the range $x\ge 1$, $0\le\alpha<2$, $0\le\beta\le 10$, a 30-point Gauss formula will be adequate throughout. This formula is stored in the 30×2 array `xwmacdonald`.

Table 4.1: Gauss quadrature for the integrals in (3.5).

β	α	x	n_r	n_i	$\operatorname{Re} K_{\alpha+\mathrm{i}\beta}(x)$	$\operatorname{Im} K_{\alpha+\mathrm{i}\beta}(x)$
5.00	0.00	1.00	19	2	3.80461828e–04	0.00000000e+00
		5.00	14	2	3.18591025e–04	0.00000000e+00
		10.00	14	2	5.27812177e–06	0.00000000e+00
		20.00	14	2	3.11005908e–10	0.00000000e+00
		50.00	16	2	2.66182488e–23	0.00000000e+00
		100.00	21	2	4.11189777e–45	0.00000000e+00
	0.50	1.00	19	18	6.75850406e–04	2.64552074e–04
		5.00	14	14	2.85418288e–04	1.66486655e–04
		10.00	13	14	5.18618578e–06	1.30924941e–06
		20.00	14	15	3.10593229e–10	3.84530876e–11
		50.00	16	18	2.66517386e–23	1.32270773e–24
		100.00	21	22	4.11574681e–45	1.02447087e–46

The same cannot be said for smaller values of x. Indeed, the required order n of Gauss quadrature increases rapidly with decreasing x. The difficulty is caused, in part, by the behavior of the function $1/h$ in (3.6), which for small x exhibits a steep boundary layer in the vicinity of $u = 0$.

In the next section we show how the difficulty can be resolved.

5 The function $K_\nu(x)$ for $x < 1$.

The idea is to split the integral into two parts and use Gauss–Legendre quadrature on one, and our special Gauss formula on the other. Thus, with c a fixed constant (which will be determined presently), we write

$$\operatorname{Re} K_{\alpha+\mathrm{i}\beta}(x) = \left(\int_0^c + \int_c^\infty\right) e^{-x\cosh t}\cosh\alpha t\cos\beta t\,\mathrm{d}t, \tag{5.1}$$

and similarly for the imaginary part. In the second integral, we change variables, $t \mapsto \tau + c$, so that

$$\int_c^\infty e^{-x\cosh t}\cosh\alpha t\cos\beta t\,\mathrm{d}t = \int_0^\infty e^{-x\cosh(\tau+c)}\cosh\alpha(\tau+c)\cos\beta(\tau+c)\mathrm{d}\tau.$$

Similarly as in §3, we write the first factor on the right in the form

$$\begin{aligned} e^{-x\cosh(\tau+c)} &= e^{-\frac{1}{2}xe^{\tau}(\cosh c+\sinh c)-\frac{1}{2}xe^{-\tau}(\cosh c-\sinh c)} \\ &= e^{-\frac{1}{2}xe^{c}e^{\tau}}\cdot e^{-\frac{1}{2}xe^{-c}e^{-\tau}}, \end{aligned}$$

or, with

$$\xi = xe^c, \tag{5.2}$$

in the form

$$e^{-\frac{1}{2}\xi}\cdot e^{-\frac{1}{2}\xi(e^{\tau}-1)}\cdot e^{-\frac{1}{2}\xi e^{-2c}e^{-\tau}}.$$

This looks exactly like (3.1) and suggests the new variable u defined by

$$\tfrac{1}{2}\,\xi(e^{\tau}-1)=e^{u}-1. \tag{5.3}$$

Letting $\eta(u)=1+(\frac{1}{2}\,\xi-1)e^{-u}$ and carrying out a computation analogous to that in (3.4) yields

$$\int_c^\infty e^{-x\cosh t}\cosh\alpha t\cos\beta t\,\mathrm{d}t=\tfrac{1}{2}(2/\xi)^{\alpha}e^{1-\frac{1}{2}\,\xi+\alpha c}\int_0^\infty e^{-e^{u}}\varphi(u)\mathrm{d}u,$$

where

$$\varphi(u)=\\ e^{\alpha u-\frac{1}{4}\,\xi^2e^{-2c}e^{-u}/\eta}\,\eta^{-(\alpha+1)}\big(\eta^{2\alpha}+(\xi e^{-(u+c)}/2)^{2\alpha}\big)\cos\beta(u+\ln(2\eta/\xi)+c)\mathrm{d}u.$$

Likewise,

$$\int_c^\infty e^{-x\cosh t}\sinh\alpha t\sin\beta t\,\mathrm{d}t=\tfrac{1}{2}(2/\xi)^{\alpha}e^{1-\frac{1}{2}\,\xi+\alpha c}\int_0^\infty e^{-e^{u}}\gamma(u)\mathrm{d}u,$$

where

$$\gamma(u)=\\ e^{\alpha u-\frac{1}{4}\,\xi^2e^{-2c}e^{-u}/\eta}\,\eta^{-(\alpha+1)}\big(\eta^{2\alpha}-(\xi e^{-(u+c)}/2)^{2\alpha}\big)\sin\beta(u+\ln(2\eta/\xi)+c)\mathrm{d}u.$$

This is very much like (3.5) and (3.6). Since $x=1$ was an admissible value before, we expect $\xi=1$ to be admissible now. Therefore, we define c from (5.2) to be $c=\ln(1/x)$ and obtain

$$\int_c^\infty e^{-x\cosh t}\cosh\alpha t\cos\beta t\,\mathrm{d}t=\tfrac{1}{2}(2/x)^{\alpha}e^{\frac{1}{2}}\int_0^\infty e^{-e^{u}}\varphi_1(u)\mathrm{d}u,$$
$$\int_c^\infty e^{-x\cosh t}\sinh\alpha t\sin\beta t\,\mathrm{d}t=\tfrac{1}{2}(2/x)^{\alpha}e^{\frac{1}{2}}\int_0^\infty e^{-e^{u}}\gamma_1(u)\mathrm{d}u,$$

with

$$\begin{aligned}\varphi_1(u)&=e^{\alpha u-\frac{1}{4}\,x^2e^{-u}/\eta_1}\,\eta_1^{-(\alpha+1)}\,\big(\eta_1^{2\alpha}+(xe^{-u}/2)^{2\alpha}\big)\cos\beta(u+\ln(2\eta_1/x)),\\ \gamma_1(u)&=e^{\alpha u-\frac{1}{4}\,x^2e^{-u}/\eta_1}\,\eta_1^{-(\alpha+1)}\,\big(\eta_1^{2\alpha}-(xe^{-u}/2)^{2\alpha}\big)\sin\beta(u+\ln(2\eta_1/x)),\end{aligned} \tag{5.4}$$

and

$$c=\ln(1/x),\quad \eta_1=1-\tfrac{1}{2}e^{-u}. \tag{5.5}$$

Eqs. (5.4) are the same as Eqs. (3.6), if in the latter h is replaced by η_1.

The first integral in (5.1) is written as

$$\int_0^c e^{-x\cosh t}\cosh\alpha t\cos\beta t\,\mathrm{d}t=c\int_0^1 e^{-x\cosh(\tau c)}\cosh(\alpha\tau c)\cos(\beta\tau c)\mathrm{d}\tau,$$

and similarly for the imaginary part. Therefore, altogether,

$$
\begin{aligned}
\operatorname{Re} K_{\alpha+\mathrm{i}\beta}(x) &= c\int_0^1 e^{-x\cosh(\tau c)}\cosh(\alpha\tau c)\cos(\beta\tau c)\mathrm{d}\tau \\
&\qquad +\tfrac{1}{2}(2/x)^{\alpha}e^{\frac{1}{2}}\int_0^\infty e^{-e^u}\varphi_1(u)\mathrm{d}u, \\
\operatorname{Im} K_{\alpha+\mathrm{i}\beta}(x) &= c\int_0^1 e^{-x\cosh(\tau c)}\sinh(\alpha\tau c)\sin(\beta\tau c)\mathrm{d}\tau \\
&\qquad +\tfrac{1}{2}(2/x)^{\alpha}e^{\frac{1}{2}}\int_0^\infty e^{-e^u}\gamma_1(u)\mathrm{d}u,
\end{aligned}
\tag{5.6}
$$

with φ_1, γ_1 defined in (5.4) and c in (5.5). We now apply Gauss–Legendre quadrature on $[0,1]$ to the first integrals in (5.6), and our Gauss formula of §2 to the second.

6 Numerical results for $.01 < x < 1$.

Numerical experimentation for values of α and β as in §4 and selected values of x between .01 and 1 revealed that 30-point Gauss rules of both types yield the same, if not better, accuracy achieved earlier. It turns out, however, that the two parts on the right of (5.6) have a tendency to cancel each other, more so the larger β. We illustrate this in the typical case of $\beta = 5$ and $\alpha = 0$, $\alpha = \frac{1}{2}$. The degree *can* of cancellation between two numbers is measured by the logarithm (to base 10) of the ratio of the absolutely larger of the two numbers divided by the absolute value of their (algebraic) sum. This measure roughly indicates the number of decimal digits lost, owing to cancellation. Results are shown in Table 6.1.

Table 6.1: Gauss quadrature for the integrals in (5.6).

β	α	x	$\operatorname{Re} K_{\alpha+\mathrm{i}\beta}(x)$	$\operatorname{Im} K_{\alpha+\mathrm{i}\beta}(x)$	*canr*	*cani*
5.00	0.00	1.0000	3.80461828e–04	0.00000000e+00	0.000	0.000
		0.5000	–4.24117148e–04	0.00000000e+00	1.804	0.000
		0.1000	–2.37141870e–05	0.00000000e+00	3.674	0.000
		0.0500	–1.15770402e–04	0.00000000e+00	2.895	0.000
		0.0100	–3.89483091e–04	0.00000000e+00	2.406	0.000
	0.50	1.0000	6.75850406e–04	2.64552074e–04	0.000	0.000
		0.5000	–8.39993536e–04	–5.72771511e–04	1.587	1.793
		0.1000	1.47550860e–03	–1.57009337e–03	2.106	1.764
		0.0500	–2.70500618e–03	1.43843540e–03	1.862	2.141
		0.0100	–2.02652762e–03	–6.58012217e–03	2.424	1.682

Apart from cancellation, the Gauss rules with 30 points yield 12–14 correct digits uniformly for $.01 \le x \le 1$ and $0 \le \beta \le 10$. This higher accuracy of the Gauss quadratures more than compensates for the loss of accuracy caused by cancellation.

7 The function $K_\nu(x)$ for $0 < x \le .01$.

To compute $K_\nu(x)$ for small values of x, we combine

$$K_\nu(x) = \frac{\pi/2}{\sin \nu\pi}\,(I_{-\nu}(x) - I_\nu(x)), \quad \nu \notin \mathbb{N}, \tag{7.1}$$

with the first three terms of the well-known power series expansion of $I_{\pm\nu}$. Making use of the reflection formula for the gamma function,

$$\Gamma(\nu)\Gamma(1-\nu) = \frac{\pi}{\sin \nu\pi}\,,$$

one obtains

$$\begin{aligned} K_\nu(x) = \frac{1}{2}\left(\frac{x}{2}\right)^{-\nu} \Bigg\{ \Gamma(\nu)\left[1 + \frac{x^2}{4(1-\nu)} + \frac{x^4}{32(1-\nu)(2-\nu)}\right] + O(x^6) \\ + \left(\frac{x}{2}\right)^{2\nu} \Gamma(-\nu)\left[1 + \frac{x^2}{4(1+\nu)} + \frac{x^4}{32(1+\nu)(2+\nu)} + O(x^6)\right]\Bigg\}. \end{aligned} \tag{7.2}$$

Since we assume $\operatorname{Re}\nu < 2$, the second expression in curled brackets is significant. When $\nu = 1$, we compute $K_1(x)$ directly from [1, Eq. 9.6.11], using only the first three terms in the power series involved,

$$\begin{aligned} K_1(x) = \frac{1}{x} - \frac{x}{2}\ln(2/x)\left[1 + \tfrac{1}{8}x^2 + \tfrac{1}{192}x^4\right] \\ - \frac{x}{4}\left[1 - 2\gamma + \tfrac{1}{8}\left(\tfrac{5}{2} - 2\gamma\right)x^2 + \tfrac{1}{192}\left(\tfrac{10}{3} - 2\gamma\right)x^4\right] + O(x^6\ln(1/x)), \end{aligned} \tag{7.3}$$

where $\gamma = .57721\ldots$ is Euler's constant. When $\nu = 0$, the analogous result follows from [1, Eqs. 9.6.12 and 9.6.13],

$$K_0(x) = (\ln(2/x) - \gamma)\left(1 + \tfrac{1}{4}x^2 + \tfrac{1}{64}x^4\right) + \tfrac{1}{4}x^2 + \tfrac{3}{128}x^4 + O(x^6\ln(1/x)). \tag{7.4}$$

Naturally, when ν is very close to 0, 1, or 2, and x not very small, some cancellation must be expected to occur in (7.2).

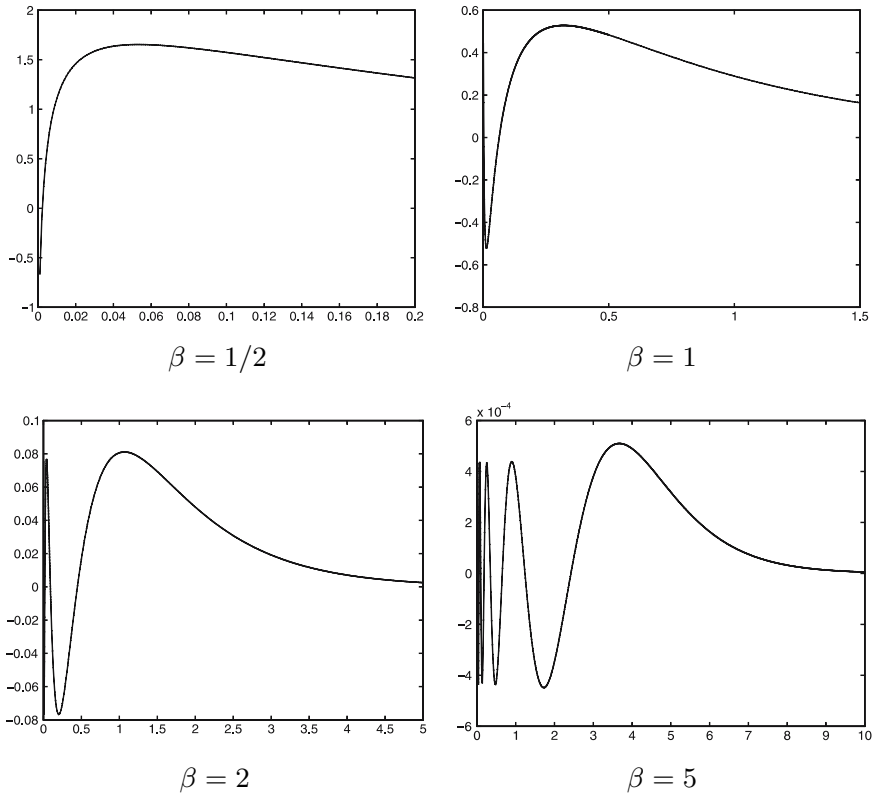
A Matlab routine named `macdonald.m` for computing $K_{\alpha+\mathrm{i}\beta}(x)$ is provided on the web site mentioned at the end of §1. It has been tested against selected 7-digit values of the tables in [8] and [10], as well as against the Matlab (symbolic) `mfun`-routine `BesselK`. The required 30-point Gauss–Legendre rule is stored in the 30×2 array `xwleg01`. We also prepared a quadruple-precision Fortran routine `qmacdonald.f` which, for $x \ge .0001$, provides answers to about 26 correct decimal digits. It requires 55- resp. 60-point Gauss rules, which are provided in the arrays `xg`, `wg` resp. `xgl`, `wgl` in the common statement of the program. They must be read in from the file `qxwmcdleg` by the calling program. For an example, see the program `qmcd.f` and its calling sequence `qmcd`.

8 Plots of $K_{\mathrm{i}\beta}$ and $K_{1/2+\mathrm{i}\beta}$.

While for large x the modified Bessel function $K_\nu(x)$, for fixed (real or complex) ν exhibits exponential decay [1, Eq. 9.7.2],

$$K_\nu(x) \sim \sqrt{\frac{\pi}{2x}}\, e^{-x}, \quad x \to \infty, \tag{8.1}$$

the behavior near $x = 0$ is more complicated, as follows from (7.2).



Closeups near $x = 0$ in the cases $\beta = 2$ and $\beta = 5$:

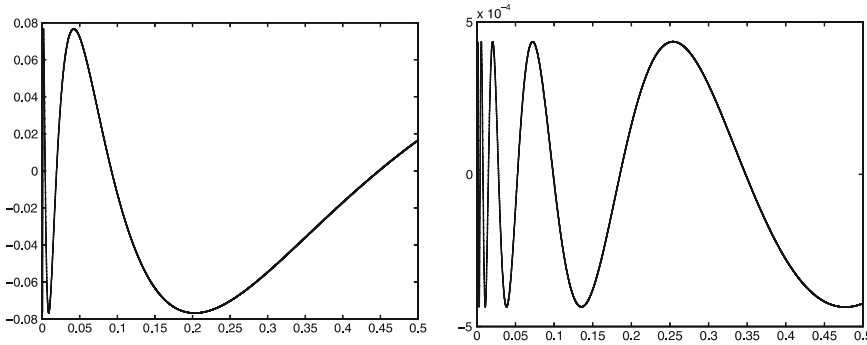
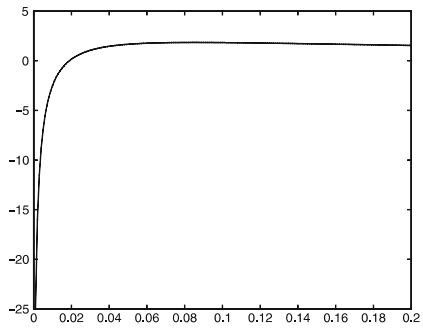
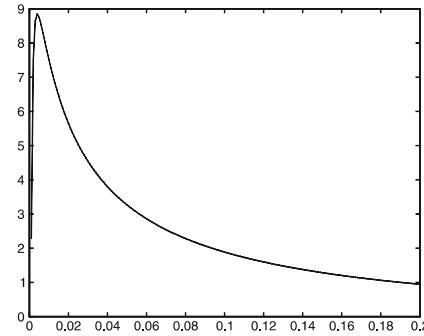


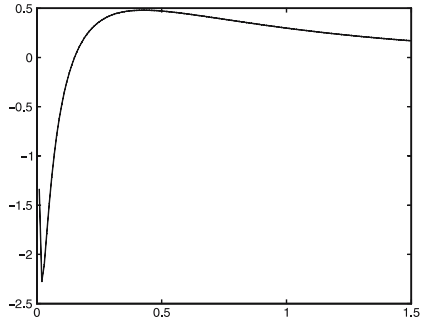
Figure 8.1: The function $K_{\mathrm{i}\beta}(x)$.

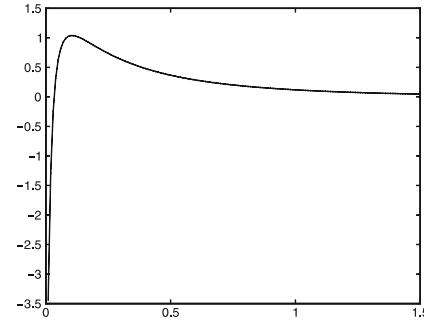
The dense oscillations near $x = 0$, typical in all cases, pose interesting questions as to good methods of calculating the Kontorovich–Lebedev integral transforms. We hope to address this problem in future work.



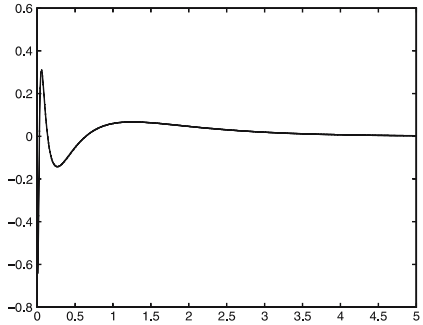


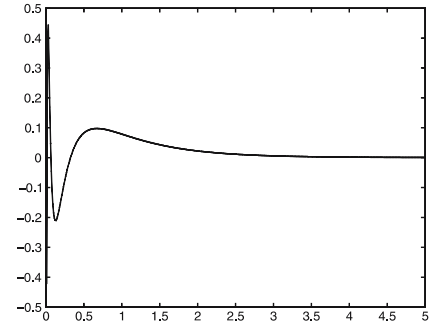
$\operatorname{Re} K_{1/2+\mathrm{i}\beta}$ $\beta = 1/2$ $\operatorname{Im} K_{1/2+\mathrm{i}\beta}$



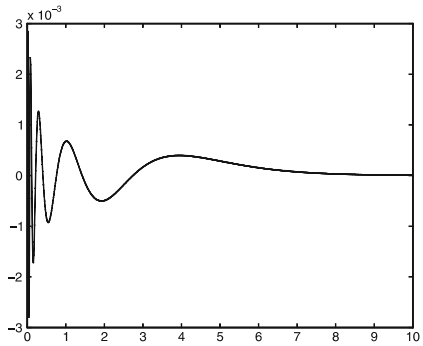


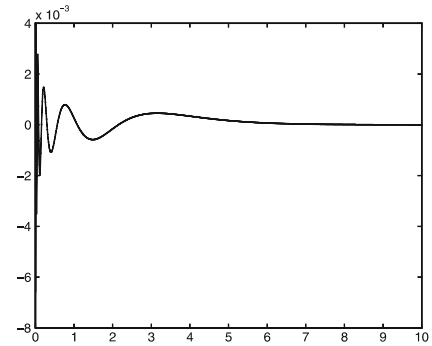
$\operatorname{Re} K_{1/2+\mathrm{i}\beta}$ $\beta = 1$ $\operatorname{Im} K_{1/2+\mathrm{i}\beta}$





$\operatorname{Re} K_{1/2+\mathrm{i}\beta}$ $\beta = 2$ $\operatorname{Im} K_{1/2+\mathrm{i}\beta}$





$\operatorname{Re} K_{1/2+\mathrm{i}\beta}$ $\beta = 5$ $\operatorname{Im} K_{1/2+\mathrm{i}\beta}$

Figure 8.2: The function $K_{1/2+\mathrm{i}\beta}(x)$.

Acknowledgment.

The author is indebted to Dr. Juri Rappoport for having drawn the author's attention to the problem at hand, and for providing relevant references.

REFERENCES

1. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, National Bureau of Standards, Applied Mathematics Series 55 (1964), U.S. Government Printing Office, Washington, D.C.
2. V. A. Ditkin and A. P. Prudnikov, *Integral Transforms and Operational Calculus* (Russian), 2nd edn., Izdat. "Nauka", Moscow, 1974 [English translation of the 1st edn., Pergamon Press, Oxford, 1965].
3. B. R. Fabijonas, D. W. Lozier, and J. M. Rappoport, *Algorithms and codes for the Macdonald function: recent progress and comparisons*, J. Comput. Appl. Math. 161 (2003), pp. 179–192.
4. W. Gautschi, *Orthogonal Polynomials: Computation and Approximation*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 2004.
5. A. Gil, J. Segura, and N. M. Temme, *Computing special functions by using quadrature rules*, Numer. Algorithms 33 (2003), pp. 265–275.
6. A. Gil, J. Segura, and N. M. Temme, *Algorithm 831: modified Bessel functions of imaginary order and positive argument*, ACM Trans. Math. Softw. 30 (2004), pp. 159–164.
7. N. N. Lebedev and I. P. Skal'skaya, *Some Integral Transforms Related to the Kontorovich–Lebedev Transform* (Russian), in Problems of Mathematical Physics (Russian), Nauka, Leningrad, pp. 68–79, 1976.
8. J. M. Rappoport, *Tables of Modified Bessel Functions* $K_{1/2+\mathrm{i}\beta}(x)$ (Russian), Nauka, Moscow, 1979.
9. R. Wong, *Asymptotic Approximations of Integrals*, Computer Science and Scientific Computing, Academic Press, Boston, 1989.
10. M. I. Žurina and L. N. Karmazina, *Tables of Modified Bessel Functions with Imaginary Index* $K_{\mathrm{i}\tau}(x)$ (Russian), Vyčisl. Centr Akad. Nauk SSSR, Moscow, 1967.

9.15. [182] "Conjectured inequalities for Jacobi polynomials and their largest zeros"

---

[182] (with P. Leopardi) "Conjectured inequalities for Jacobi polynomials and their largest zeros," *Numer. Algorithms* **45**, 217–230 (2007).

Numer Algor (2007) 45:217–230
DOI 10.1007/s11075-007-9067-5

ORIGINAL PAPER

Conjectured inequalities for Jacobi polynomials and their largest zeros

Walter Gautschi · Paul Leopardi

Received: 17 November 2006 / Accepted: 22 January 2007 /
Published online: 27 March 2007

Abstract Inequalities are conjectured for the Jacobi polynomials $P_n^{(\alpha,\beta)}$ and their largest zeros. Special attention is given to the cases $\beta = \alpha - 1$ and $\beta = \alpha$.

Keywords Jacobi polynomials · Zeros · Inequalities

Mathematics Subject Classification (2000) 33C45

1 Introduction

Special Jacobi polynomials $P_n^{(\alpha,\beta)}(x)$ with parameters $\beta = \alpha - 1$ or $\beta = \alpha$ are frequently encountered in multivariate polynomial approximation on spherical surfaces, in which case α is related to the space dimension; see, e.g., [3, 4, Section 14.1]. Technical properties, especially inequalities, for these polynomials can be a valuable aid in simplifying various estimates in the theory of spherical approximation. In this paper inequalities are studied related to the largest zeros of Jacobi polynomials and also inequalities involving the Jacobi polynomials themselves, more precisely, the scaled polynomials having the value 1 at $x = 1$. All inequalities are only conjectured to hold, but compelling evidence is provided, both numerical and analytic, in support of their validity.

W. Gautschi (✉)
Department of Computer Sciences, Purdue University,
West Lafayette, IN 47907-2066, USA
e-mail: wxg@cs.purdue.edu

P. Leopardi
School of Mathematics and Statistics, University of New South Wales,
Sydney, NSW 2052, Australia
e-mail: leopardi@maths.unsw.edu.au

The special Jacobi polynomials with $\beta = \alpha - 1$ are considered in Section 2, with inequalities for the largest zeros being discussed in Section 2.1, and inequalities for the scaled polynomials in Section 2.2. In Section 3, the analogous problems, and a variation thereof, for general Jacobi polynomials are taken up. Some special cases that can be proved rigorously are mentioned in Section 4.

2 Special Jacobi polynomials

2.1 Largest zeros

Let $x_n^{(\alpha)} = \cos \Theta_n^{(\alpha)}$, $0 < \Theta_n^{(\alpha)} < \pi$, be the largest zero of the Jacobi polynomial $P_n^{(\alpha,\alpha-1)}(x)$, $\alpha > 0$. Our conjecture relates to the inequality

$$n\Theta_n^{(\alpha)} < (n+1)\Theta_{n+1}^{(\alpha)}. \tag{2.1}$$

From the interlacing property of the zeros of orthogonal polynomials it is known that the sequence $\{\Theta_n^{(\alpha)}\}$ is monotonically decreasing. Inequality (2.1), if true, places a limit on the relative decrement, $(\Theta_n^{(\alpha)} - \Theta_{n+1}^{(\alpha)})/\Theta_{n+1}^{(\alpha)} < 1/n$.

Conjecture 1 *Given $\alpha > 0$, there are two alternatives: either* (2.1) *holds for all $n = 1, 2, 3, \ldots$, or* (2.1) *is false for $n = 1$. In other words, the validity of* (2.1) *for $n = 1$ implies the validity of* (2.1) *for all $n \geq 1$.*

Numerical evidence for Conjecture 1 was obtained with the help of the Matlab package OPQ available on the web site http://www.cs.purdue.edu/archives/2002/wxg/codes. The following routine is at the core of the verification effort:

```
ab=r_jacobi(n+1,a,a-1);
for k=1:n
 xw=gauss(k,ab); xw1=gauss(k+1,ab);
 theta=acos(xw(k,1)); theta1=acos(xw1(k+1,1));
 if k*theta >= (k+1)*theta1
  [k*theta,(k+1)*theta1], a, k, error('conjecture 1 false')
 end
end
```

The first command generates the recursion coefficients for the special Jacobi polynomials, which are used in the routine `gauss` to compute the nodes and weights of the respective Gaussian quadrature rules. Only the nodes, stored (in increasing order) in the first column of the array `xw` resp. `xw1` are of interest here.

When the verification routine is run with `n` = 100 and `a` = [0.5 : 0.01 : 1, 1.1 : 0.1 : 10, 10.5 : 0.5 : 20], the error statement is never invoked. On the other hand, when `a` = 0.5 : −0.01 : 0.01, the error message appears with `a` = 0.13, `n` = 1, and likewise, when `a` = 0.14 : −0.0001 : 0.13, it appears with

a = 0.1350, n = 1. It thus appears that Conjecture 1 is true, and that inequality (2.1) holds for all $n \geq 1$ and for all $\alpha > \alpha_0$, where $0.1350 < \alpha_0 < 0.1351$. In order to determine α_0 more precisely, we examine the case $n = 1$.

From the recurrence relation for Jacobi polynomials (see, e.g., [6, eqn (4.5.1)]) one finds

$$P_1^{(\alpha,\alpha-1)}(x) = \frac{1}{2}\left((2\alpha+1)x+1\right),$$
$$4P_2^{(\alpha,\alpha-1)}(x) = (\alpha+1)\left((2\alpha+3)x^2+2x-1\right). \tag{2.2}$$

Therefore,

$$x_1^{(\alpha)} = -\frac{1}{2\alpha+1}, \quad x_2^{(\alpha)} = \frac{1}{1+\sqrt{2\alpha+4}}, \tag{2.3}$$

and (2.1) for $n = 1$ is equivalent to

$$\arccos\left(-\frac{1}{2\alpha+1}\right) < 2\arccos\frac{1}{1+\sqrt{2\alpha+4}},$$

or, using $\arccos(-t) = \pi - \arccos(t)$, equivalent to

$$2\arccos\frac{1}{1+\sqrt{2\alpha+4}} + \arccos\frac{1}{2\alpha+1} - \pi > 0. \tag{2.4}$$

The left-hand side is a strictly increasing function of α, negative for $\alpha = 0$ and tending to $\frac{1}{2}\pi$ as $\alpha \to \infty$. Therefore, if α_0 is the unique root of

$$2\arccos\frac{1}{1+\sqrt{2\alpha+4}} + \arccos\frac{1}{2\alpha+1} - \pi = 0, \tag{2.5}$$

then (2.4), and hence (2.1) for $n = 1$, holds exactly if $\alpha > \alpha_0$. Using the Matlab routine `fzero`, one finds

$$\alpha_0 = 0.13507978085964. \tag{2.6}$$

Thus, *if Conjecture* 1 *is true, then* (2.1) *holds for all* $n \geq 1$ *precisely if* $\alpha > \alpha_0$.

2.2 Scaled polynomials

For the remainder of this paper, we use the abbreviated notation

$$\widetilde{P}_n^{(\alpha,\beta)}(x) := \frac{P_n^{(\alpha,\beta)}(x)}{P_n^{(\alpha,\beta)}(1)}. \tag{2.7}$$

The conjecture for the Jacobi polynomials themselves involves the inequality

$$\widetilde{P}_n^{(\alpha,\alpha-1)}\left(\cos\frac{\theta}{n}\right) < \widetilde{P}_{n+1}^{(\alpha,\alpha-1)}\left(\cos\frac{\theta}{n+1}\right). \tag{2.8}$$

With notation as in Section 2.1 we consider two intervals for θ,

$$0 < \theta < \Theta_1^{(\alpha)}, \quad \text{and} \quad 0 < \theta < \pi, \tag{2.9}$$

where

$$\cos \Theta_1^{(\alpha)} = x_1^{(\alpha)} = -\frac{1}{2\alpha + 1}. \tag{2.10}$$

Conjecture 2 *Given $\alpha > 0$, there are two alternatives for each of the two intervals* (2.9)*: either* (2.8) *holds for all $n = 1, 2, 3, \ldots$ and all θ in the respective interval, or* (2.8) *is false for $n = 1$ and some θ in the respective interval. In other words, the validity of* (2.8) *for $n = 1$ implies the validity of* (2.8) *for all $n \geq 1$.*

The verification routine for Conjecture 2 is a bit more intricate than the one for Conjecture 1. Its core is shown below.

```
ab=r_jacobi(n+1,a,a-1);
th1=acos(-1/(2*a+1));
% th1=pi;
for nu=1:N
   th=nu*th1/(N+1);
   for k=1:n
     x0=1; x=cos(th/k); y=cos(th/(k+1));
     p0=0; p01=1; px=0; px1=1; py=0; py1=1;
     for r=1:k+1
       p0m1=p0; p0=p01; pxm1=px; px=px1; pym1=py; py=py1;
       p01=(x0-ab(r,1))*p0-ab(r,2)*p0m1;
       px1=(x-ab(r,1))*px-ab(r,2)*pxm1;
       py1=(y-ab(r,1))*py-ab(r,2)*pym1;
     end
     if px/p0 >= py1/p01
      [px/p0,py1/p01],a,k,nu,error('conjecture 2 false')
     end
   end
end
```

Run with n = 100, N = 1000, and a as in Section 2.1, the routine for the first interval of (2.9) produces the same results as in Section 2.1, provided N is increased to N = 5000 for the last set of a-values. Conjecture 2 thus appears to be true, and inequality (2.8) valid for $0 < \theta < \Theta_1^{(\alpha)}$ precisely if $\alpha > \alpha_0$. In the case of the second interval $0 < \theta < \pi$, the first set of a-values, when N = 1000, again produces no error message, the second set, with N = 5000, an error message with a = 0.28, n = 1, and a = 0.29 : −0.001 : 0.28 an error message with a = 0.280, n = 1. Inequality (2.8) for the second interval thus seems to hold if $\alpha > \alpha_1$, where $0.280 < \alpha_1 < 0.290$.

To get more precise information, we analyze the case $n = 1$, i.e.,

$$\widetilde{P}_1^{(\alpha,\alpha-1)}(\cos\theta) < \widetilde{P}_2^{(\alpha,\alpha-1)}\left(\cos\frac{\theta}{2}\right). \tag{2.11}$$

From (2.2), we have

$$\widetilde{P}_1^{(\alpha,\alpha-1)}(\cos\theta) = \frac{(2\alpha+1)\cos\theta + 1}{2(\alpha+1)},$$
$$\widetilde{P}_2^{(\alpha,\alpha-1)}\left(\cos\frac{\theta}{2}\right) = \frac{(2\alpha+3)\cos^2\frac{\theta}{2} + 2\cos\frac{\theta}{2} - 1}{2(\alpha+2)}, \tag{2.12}$$

so that (2.11), using $\cos\theta = 2\cos^2\frac{\theta}{2} - 1$, becomes

$$(1 + 5\alpha + 2\alpha^2)\cos^2\frac{\theta}{2} - 2(1+\alpha)\cos\frac{\theta}{2} + (1 - 3\alpha - 2\alpha^2) < 0,$$

or, simplifying,

$$(u-1)[(1+5\alpha+2\alpha^2)u - (1-3\alpha-2\alpha^2)] < 0, \quad u := \cos\frac{\theta}{2}.$$

Since $u - 1 < 0$ on either interval (2.9), this is the same as

$$(1 + 5\alpha + 2\alpha^2)u - (1 - 3\alpha - 2\alpha^2) > 0,$$

or, since $1 + 5\alpha + 2\alpha^2 > 0$,

$$u > \frac{1 - 3\alpha - 2\alpha^2}{1 + 5\alpha + 2\alpha^2}, \quad u := \cos\frac{\theta}{2}. \tag{2.13}$$

Consider first the interval $0 < \theta < \Theta_1^{(\alpha)}$. Then (2.13) holds precisely if

$$\cos\frac{\Theta_1^{(\alpha)}}{2} = \sqrt{\frac{1 + \cos\Theta_1^{(\alpha)}}{2}} = \sqrt{\frac{\alpha}{2\alpha+1}} > \frac{1 - 3\alpha - 2\alpha^2}{1 + 5\alpha + 2\alpha^2}. \tag{2.14}$$

Using the Matlab routine `fzero`, one finds

$$\alpha > \alpha_0, \tag{2.15}$$

where, interestingly, α_0 is exactly the same as in (2.6).

On the second interval $0 < \theta < \pi$, we have (2.13) precisely if

$$\cos\frac{\pi}{2} = 0 > \frac{1 - 3\alpha - 2\alpha^2}{1 + 5\alpha + 2\alpha^2},$$

i.e., if

$$\alpha > \alpha_1 = \frac{1}{4}(\sqrt{17} - 3) = .28077640640442. \tag{2.16}$$

In summary, *if Conjecture* 2 *is true, then the inequality* (2.8) *holds for all* $n \geq 1$ *on the first interval* (2.9) *precisely if* $\alpha > \alpha_0$, *and on the second interval precisely if* $\alpha > \alpha_1$, *where* α_0, α_1 *are given by* (2.6) *and* (2.16), *respectively.*

We remark that by squaring (2.14) and removing the root $\alpha = -1$, one finds that α_0 is the smallest positive root of the quartic equation

$$4\alpha^4 + 4\alpha^3 - 11\alpha^2 - 6\alpha + 1 = 0. \tag{2.17}$$

The same equation can be obtained from (2.5), written in the form

$$2\arccos\frac{1}{1+\sqrt{2\alpha+4}} = \arccos\left(-\frac{1}{2\alpha+1}\right). \tag{2.18}$$

Indeed, observing that $2\arccos t = \arccos(2t^2-1)$, (2.18) implies

$$\frac{2}{(1+\sqrt{2\alpha+4})^2} - 1 = -\frac{1}{2\alpha+1},$$

or

$$\alpha(1+\sqrt{2\alpha+4})^2 = 2\alpha+1.$$

By an elementary calculation, this yields (2.17).

3 General Jacobi polynomials

3.1 Largest zeros

We now denote by $x_n^{(\alpha,\beta)} = \cos\Theta_n^{(\alpha,\beta)}$, $0 < \Theta_n^{(\alpha,\beta)} < \pi$, the largest zero of the Jacobi polynomial $P_n^{(\alpha,\beta)}(x)$, $\alpha > -1$, $\beta > -1$. We consider the inequality analogous to (2.1),

$$n\Theta_n^{(\alpha,\beta)} < (n+1)\Theta_{n+1}^{(\alpha,\beta)}. \tag{3.1}$$

The case $\alpha = \beta = -1/2$ of Chebyshev polynomials is exceptional here, since $\Theta_n = \pi/2n$, and both sides of (3.1) are identically equal to $\pi/2$.

Using an obvious extension of the Matlab routine in Section 2.1, we are led to conjecture:

Conjecture 3 *Given $\alpha > -1$, $\beta > -1$, there are two alternatives: either* (3.1) *holds for all $n = 1, 2, 3, \ldots$, or* (3.1) *is false for $n = 1$. In other words, the validity of* (3.1) *for $n = 1$ implies the validity of* (3.1) *for all $n \geq 1$.*

It is known that

$$\lim_{n\to\infty} n\Theta_n^{(\alpha,\beta)} = j_1^{(\alpha)}, \tag{3.2}$$

where $j_1^{(\alpha)}$ is the first positive zero of the Bessel function J_α (cf. [6, Theorem 8.1.2]). Conjecture 3, if true, then states that the validity of (3.1) for $n = 1$ implies that convergence in (3.2) is monotone increasing.

The following is our evidence for Conjecture 3. Running the (extended) verification routine of Section 2.1 with n up to 100, and for each $\alpha = 1.01 : 0.01 : 1.2, 1.3 : 0.1 : 5.0, 6 : 1 : 20$ for $\beta = -0.99 : 0.01 : -0.80, -0.7 : 0.1 : 0.0, 1 : 1 : 20$, no error message was encountered, suggesting that the inequality (3.1) holds for all $n \geq 1$ in the infinite domain $\alpha > 1$, $\beta > -1$. When one takes $\alpha = 0.9 : -0.1 : -0.9$, however, and for each of these α goes through

$\beta = 20 : -1 : 0, -0.1 : -0.1 : -0.9, -0.89 : -0.01 : -0.99, -0.999$, then an error message appears, always with n = 1, for the following pairs of values (α, β):

α	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
β	−0.999	−0.999	−0.99	−0.98	−0.97	−0.95	−0.92	−0.90	−0.9	−0.9

α	−0.1	−0.2	−0.3	−0.4	−0.5	−0.6	−0.7	−0.8	−0.9
β	−0.8	−0.8	−0.7	−0.6	−0.5	−0.5	−0.4	−0.3	−0.2

The results suggest that in the strip $-1 < \alpha < 1, \beta > -1$, there exists a curve, monotonically decreasing from 0 to −1, above which (3.1) holds for all $n \geq 1$, and below which inequality (3.1) fails for $n = 1$. We will compute this curve more accurately when, as we now begin to do, the case $n = 1$ is examined.

In analogy to (2.2), we find

$$\begin{aligned} P_1^{(\alpha,\beta)}(x) &= \frac{1}{2}\left((\alpha+\beta+2)x+\alpha-\beta\right), \\ 8P_2^{(\alpha,\beta)}(x) &= (\alpha+\beta+3)(\alpha+\beta+4)x^2 + 2(\alpha+\beta+3)(\alpha-\beta)x \\ &\quad + (\alpha-\beta)^2 - (\alpha+\beta+4), \end{aligned} \tag{3.3}$$

from which

$$\begin{aligned} x_1^{(\alpha,\beta)} &= -\frac{\alpha-\beta}{\alpha+\beta+2}, \\ x_2^{(\alpha,\beta)} &= \frac{1}{\alpha+\beta+4}\left[-(\alpha-\beta)+2\sqrt{2+\frac{\alpha\beta-2}{\alpha+\beta+3}}\right]. \end{aligned} \tag{3.4}$$

Inequality (3.1), therefore, analogously to (2.5), can be given the form

$$\begin{aligned} &2\arccos\left(\frac{1}{\alpha+\beta+4}\left[-(\alpha-\beta)+2\sqrt{2+\frac{\alpha\beta-2}{\alpha+\beta+3}}\right]\right) \\ &\quad + \arccos\frac{\alpha-\beta}{\alpha+\beta+2} - \pi > 0. \end{aligned} \tag{3.5}$$

When $\alpha = \beta = -\frac{1}{2}$, this gives $2\frac{\pi}{4} + \frac{\pi}{2} - \pi = 0$, i.e., equality in (3.1), as was already noted above. The same is true for $\alpha = 1$ and $\beta \to -1$, and for $\alpha > 1$ and $\beta \to \infty$. When $\alpha > 1$ is fixed, and β increases from −1 to ∞, the graph of (3.5) sharply increases from a positive value to a maximum and then decreases monotonically to zero, so that (3.5) holds for all $\alpha > 1, \beta > -1$, in agreement with what was found numerically above.

When $-1 < \alpha < 1$ is fixed, the equation in β resulting from replacing inequality in (3.5) by equality, can be solved numerically by the Matlab routine `fzero`. This produces the curve shown in Fig 1. Inequality (3.1) for $n = 1$ thus holds in the region above this curve, and, together with $(-1 < \alpha < 1, \beta > 0) \cup (1 < \alpha < \infty, \beta > -1)$, this is the region of validity of the inequality for all $n \geq 1$ if Conjecture 3 is true.

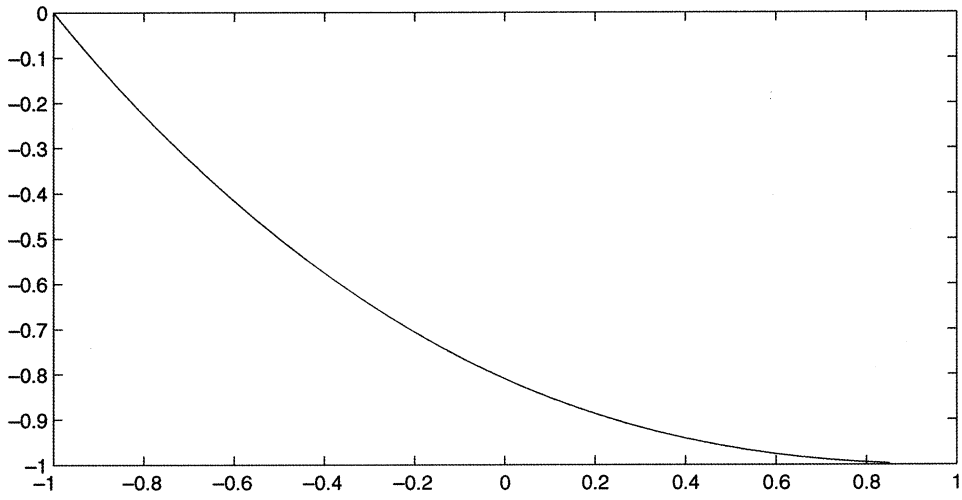


Fig. 1 The boundary curve of the domain of validity for (3.1)

The graph also sheds new light on the result found in Section 2.1: If the inequality is to be true for $(\alpha, \beta = \alpha - 1)$, then the point of intersection of the line $\beta = \alpha - 1$ with the boundary curve of Fig. 1 determines α_0. Setting $\beta = \alpha - 1$ in (3.5) and replacing the inequality sign with the equality sign indeed yields (2.5). Similarly, if $\beta = \alpha$, the point of intersection of the line $\beta = \alpha$ with the curve yields $\alpha = -\frac{1}{2}$, as is easily verified. Inequality (3.5) thus holds for $\beta = \alpha > -\frac{1}{2}$, and therefore, if Conjecture 3 is true, inequality (3.1) in the ultraspherical case $\beta = \alpha$ holds for all $\alpha > -\frac{1}{2}$. It is actually known to hold for $-\frac{1}{2} < \alpha < \frac{1}{2}$ in the sharper form $(n + \alpha + \frac{1}{2})\Theta_n < (n + \alpha + \frac{3}{2})\Theta_{n+1}$; cf. [6, Section 6.3(5), p. 127]. However, this sharper inequality ceases to hold when $\alpha \geq \frac{1}{2}$.

3.2 Scaled polynomials

The inequality to be studied here is

$$\widetilde{P}_n^{(\alpha,\beta)}\left(\cos\frac{\theta}{n}\right) < \widetilde{P}_{n+1}^{(\alpha,\beta)}\left(\cos\frac{\theta}{n+1}\right), \tag{3.6}$$

with $\widetilde{P}$ defined by (2.7), on either of the two intervals

$$0 < \theta < \Theta_1^{(\alpha,\beta)}, \quad 0 < \theta < \pi, \tag{3.7}$$

where

$$\cos\Theta_1^{(\alpha,\beta)} = x_1^{(\alpha,\beta)} = -\frac{\alpha - \beta}{\alpha + \beta + 2}. \tag{3.8}$$

We note again the exceptional case $\alpha = \beta = -1/2$, in which both sides of (3.6) are identically equal to $\cos\theta$.

Conjecture 4 *Given $\alpha > -1$, $\beta > -1$, there are two alternatives for each of the two intervals* (3.7)*: either* (3.6) *holds for all $n = 1, 2, 3, \ldots$ and all θ in the respective interval, or* (3.6) *is false for $n = 1$ and some θ in the respective interval. In other words, the validity of* (3.6) *for $n = 1$ implies the validity of* (3.6) *for all $n \geq 1$.*

Since

$$P_n^{(\alpha,\beta)}(1) = \binom{n+\alpha}{n} \sim \frac{n^\alpha}{\Gamma(\alpha+1)} \quad \text{as } n \to \infty,$$

the result in [6, Theorem 8.1.1] can be rephrased in the form

$$\lim_{n\to\infty} \widetilde{P}_n^{(\alpha,\beta)}\left(\cos\frac{\theta}{n}\right) = \Gamma(\alpha+1)\left(\frac{\theta}{2}\right)^{-\alpha} J_\alpha(\theta), \tag{3.9}$$

where J_α is the Bessel function of order α. Therefore, Conjecture 4, if true, states that the validity of (3.6) for $n = 1$ implies that convergence in (3.9) is monotone increasing.

The Matlab script of Section 2.2 is easily adapted to deal with the conjecture (3.6) for general Jacobi polynomials. When run with the same data as used to verify Conjecture 3, with $n = 100$ and $N = 1000$, similar results were obtained as in Section 3.1, i.e., a strong indication that (3.6) holds on either interval (3.7) for all $n \geq 1$ whenever $\alpha > 1$ and $\beta > -1$, while for α in the interval $(-1, 1)$ the same is true for β above a certain curve that extends from the point $(\alpha, \beta) = (-1, 0)$ down to the point $(\alpha, \beta) = (1, -1)$. For β below that curve, the conjecture fails consistently when $n = 1$. As will be seen, the initial part of this curve, for $-1 < \alpha < -\frac{1}{2}$, is the straight line $\beta = -\alpha - 1$.

This all will become more clear by analyzing (3.6) in the case $n = 1$,

$$\widetilde{P}_1^{(\alpha,\beta)}(\cos\theta) < \widetilde{P}_2^{(\alpha,\beta)}\left(\cos\frac{\theta}{2}\right). \tag{3.10}$$

From (3.3) we first note that

$$\widetilde{P}_1^{(\alpha,\beta)}(\cos\theta) = \frac{(\alpha+\beta+2)\cos\theta + \alpha - \beta}{2(\alpha+1)}$$

and

$$\widetilde{P}_2^{(\alpha,\beta)}\left(\cos\frac{\theta}{2}\right) = \frac{N(\alpha,\theta)}{4(\alpha+1)(\alpha+2)},$$

where

$$\begin{aligned} N(\alpha,\theta) &= (\alpha+\beta+3)(\alpha+\beta+4)\cos^2\frac{\theta}{2} \\ &\quad + 2(\alpha+\beta+3)(\alpha-\beta)\cos\frac{\theta}{2} + (\alpha-\beta)^2 - (\alpha+\beta+4). \end{aligned}$$

The inequality (3.10) then becomes, after simplification,

$$(u-1)[(3\alpha^2 + 2\alpha\beta + 9\alpha - \beta^2 + \beta + 4)u + \alpha^2 + 2\alpha\beta + \beta^2 + 3\alpha + 7\beta + 4] < 0,$$

with u as in (2.13). Again, since $u - 1 < 0$ on either of the two intervals (3.7), the inequality to be studied is

$$(3\alpha^2 + 2\alpha\beta + 9\alpha - \beta^2 + \beta + 4)u + \alpha^2 + 2\alpha\beta + \beta^2 + 3\alpha + 7\beta + 4 > 0,$$
$$u := \cos\frac{\theta}{2}. \tag{3.11}$$

Lemma 3.1 *Let a, b be real numbers, and consider the inequality*

$$au + b > 0 \quad \text{on } u_0 < u < 1,\ u_0 \geq 0. \tag{3.12}$$

If $a + b \geq 0$, then (3.12) is always true except when $a = b = 0$ or $a > 0$, $b < 0$, and $u_0 < -b/a$. If $a + b < 0$, then (3.12) is never true.

Proof Immediate on geometric grounds. □

We now apply Lemma 3.1 to (3.11), i.e., to

$$a = 3\alpha^2 + 2\alpha\beta + 9\alpha - \beta^2 + \beta + 4,$$
$$b = \alpha^2 + 2\alpha\beta + \beta^2 + 3\alpha + 7\beta + 4. \tag{3.13}$$

Here, one computes

$$a + b = 4(\alpha + 2)(\alpha + \beta + 1).$$

Since $\alpha + 2 > 0$, inequality (3.11) is false on either of the two intervals (3.7) if $\alpha + \beta + 1 < 0$. In the case $\alpha + \beta + 1 \geq 0$ it is false if $a = b = 0$, which implies $\alpha = \beta = -\frac{1}{2}$, or if

$$a > 0,\ b < 0, \text{ and } u_0 < -b/a, \tag{3.14}$$

with a, b as defined in (3.13). The curve $b = 0$ is given by

$$\beta = -\alpha - \frac{7}{2} + \frac{1}{2}\sqrt{16\alpha + 33}, \quad -1 < \alpha < 1.$$

By plotting the respective curves in the (α, β)-plane, one finds that $b < 0$ combined with $\alpha + \beta + 1 \geq 0$ and $\beta \geq -1$, cuts out the domain D shown in Fig. 2. Inequality (3.11) thus holds for all (α, β) located above the upper boundary curve of D and to the right of the line $\alpha + \beta + 1 = 0$, and for those (α, β) in the interior of D precisely if $u_0 > -b/a$ in (3.14). On the first interval (3.7), this will be true precisely if

$$u_0 = \cos\frac{\Theta_1^{(\alpha,\beta)}}{2} = \sqrt{\frac{1 + \cos\Theta_1^{(\alpha,\beta)}}{2}} = \sqrt{\frac{\beta + 1}{\alpha + \beta + 2}}$$
$$> -\frac{\alpha^2 + 2\alpha\beta + \beta^2 + 3\alpha + 7\beta + 4}{3\alpha^2 + 2\alpha\beta + 9\alpha - \beta^2 + \beta + 4}. \tag{3.15}$$

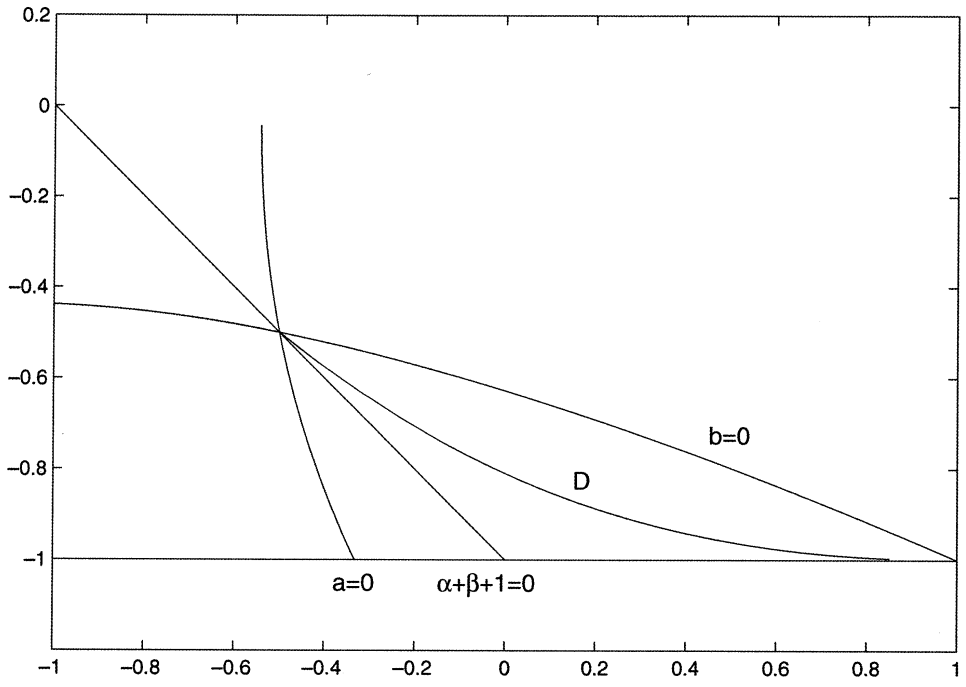


Fig. 2 The boundary curves for the domain of validity of (3.10)

This is the curve plotted inside the domain D of Fig. 2, above which inequality (3.11) is true, and below which it is false. This, together with the discussion above, completely delineates the domain of validity of (3.10) on the first interval (3.7). On the second interval we have $u_0 = \cos\frac{\pi}{2} = 0$, and the third inequality in (3.14) is a consequence of the other two. Thus, (3.10) is false in all of D, and the domain of validity of (3.10) is the region above the upper boundary of D, to the right of the line $\alpha + \beta + 1 = 0$, and of course bounded by the lines $\alpha = -1$ and $\beta = -1$. If Conjecture 4 is true, the same domains of validity hold for the inequality (3.6).

We remark that the special case $\beta = \alpha - 1$, $\alpha > 0$, turns (3.15) into (2.14), and the inequality $b > 0$ into (2.16). Likewise, the line $\beta = \alpha$, $\alpha > -1$, passes through the point $(-\frac{1}{2}, -\frac{1}{2})$ where all the curves in Fig. 2 intersect. Consequently, (3.10), and if Conjecture 4 is valid, (3.6), is true for all $\beta = \alpha > -\frac{1}{2}$. Koumandos (2005, personal communication), in fact, has shown that (3.6) is true whenever $|\alpha| = |\beta| = \frac{1}{2}$ except for $\alpha = \beta = -\frac{1}{2}$.

In order to lend still more credence to the validity of Conjecture 4, we ran the (extended) Matlab routine of Section 2.2 with (α, β) slightly above and below (at a distance of .01 from) the boundary curves of the domain of validity for (3.10). As expected, no error message appeared when (α, β) is above the boundary curve, and error messages consistently with n = 1 otherwise. (Only in the case of $u_0 = 0$, the maximum value of n had to be lowered to $n = 50$ to obtain sufficient numerical resolution along the straight part of the boundary curve).

3.3 An alternative conjecture

Examination of the graphs of $\widetilde{P}_n^{(\alpha,\beta)}\left(\cos\frac{\theta}{n}\right)$ for numerous values of α, β and n suggests that Conjectures 3 and 4 can be combined into the following conjecture.

Conjecture 5 *Given* $\alpha > -1$, $\beta > -1$, *if* (3.6) *holds for* $n = 1$ *and* $0 < \theta \leqslant \Theta_1^{(\alpha,\beta)}$, *then* (3.6) *holds for* $0 < \theta \leqslant n\Theta_n^{(\alpha,\beta)}$ *for all* $n = 1, 2, 3, \ldots$.

If (3.6) holds for $\theta = n\Theta_n^{(\alpha,\beta)}$ then we have

$$\widetilde{P}_{n+1}^{(\alpha,\beta)}\left(\cos\frac{n\Theta_n}{n+1}\right) > 0 = \widetilde{P}_{n+1}^{(\alpha,\beta)}(\cos\Theta_{n+1})$$

and therefore $n\Theta_n^{(\alpha,\beta)} < (n+1)\Theta_{n+1}^{(\alpha,\beta)}$. In other words, if the premise of Conjecture 5 is true, then the conjecture implies (3.1).

To gain confidence in Conjecture 5, the verification routine of Section 2.1 was further modified. Following is the core of the Matlab routine used to verify Conjecture 5.

```
ab=r_jacobi(n+1,a,b);
th1=n*pi; Nn=N*n;
negpx=zeros(1,n);
p1=zeros(1,n+1); p0=0; p01=1; x0=1;
for r=1:n+1
  p0m1=p0; p0=p01;
  p01=(x0-ab(r,1))*p0-ab(r,2)*p0m1;
  p1(r)=p01;
end
for nu=1:Nn
   th=nu*th1/(Nn+1);
   for k=1:n
     if negpx(k) == 0
       x=cos(th/k); y=cos(th/(k+1));
       px=0; px1=1; py=0; py1=1;
       for r=1:k+1
         pxm1=px; px=px1; pym1=py; py=py1;
         px1=(x-ab(r,1))*px-ab(r,2)*pxm1;
         py1=(y-ab(r,1))*py-ab(r,2)*pym1;
       end
       if px < 0
         negpx(k) = nu;
```

```
          else
            if px/p1(k) >= py1/p1(k+1)
              [px/p1(k),py1/p1(k+1)], a, b, k, th, ...
              error('conjecture 5 is false')
            end
          end
        end
      end
    end
```

(To avoid overflow when $\alpha + \beta + 2 > 128$, the statement defining `mu` in the routine `r_jacobi.m` was modified by evaluating the expression involving the gamma function by first taking its logarithm and then exponentiating the result).

This routine was run with `N` $= 15$, `n` $= 128$ and the following values of `a` and `b`:

1. `a` $= 2^{\mu} - 1$, `b` $= 2^{\nu} - 1$, with $\mu, \nu \in \{-1, -0.9, \ldots, 6\}$,
2. `a` $\in \{-0.95, -0.9, \ldots, -0.55\}$, `b` $= 2^{\nu} - 1$, with $\nu \in \{-1, -0.9, \ldots, 6\}$, subject to `a` $+$ `b` $+ 1 > 0$,
3. `b` $\in \{-0.95, -0.9, \ldots, -0.55\}$, `a` $= 2^{\mu} - 1$, with $\mu \in \{1, 1.1, \ldots, 6\}$,
4. `b` $\in \{-0.95, -0.9, \ldots, -0.55\}$, `a` $= 2^{\mu} - 1$, with $\mu \in \{-1, -0.9, \ldots, 0.9\}$, subject to `a` $= \alpha$, `b` $= \beta$, such that $\alpha + \beta + 1 > 0$, $\beta < -\alpha - \frac{7}{2} + \frac{1}{2}\sqrt{16\alpha + 33}$ and (3.15) holds.

In all cases, the error message was not seen.

4 Partial results

Apart from the result of Szegö [6, Section 6.3(5), p. 127] and Koumandos (2005, personal communication), referred to above, the inequalities (3.1) and (3.6) are so far known to hold in only a few cases.

For the case where (α, β) lies in the square $\left(-\frac{1}{2}, \frac{1}{2}\right)^2$, the inequality (3.1) can be proven for $n \geqslant 2$ either as a result of the inequalities of Gatteschi [1, Theorem 1.5, p. 1550], or directly using a version of the Sturm comparison theorem as formulated by Szegö [5, p. 3].

The paper [2] uses a different formulation of the Sturm comparison theorem to show that (3.6) holds for $n \geqslant 1, \alpha \geqslant \beta > -\frac{1}{2}, 0 < \theta \leqslant \frac{\pi}{2}$.

References

1. Gatteschi, L.: New inequalities for the zeros of Jacobi polynomials. SIAM J. Math. Anal. **18**, 1549–1562 (1987)
2. Leopardi, P.: Positive weight quadrature on the sphere and monotonicities of Jacobi polynomials. Numer. Algor. doi:10.1007/s11075-007-9073-7 (2007)

3. Reimer, M.: Hyperinterpolation on the sphere at the minimal projection order. J. Approx. Theory **104**, 272–286 (2000)
4. Reimer, M.: Multivariate polynomial approximation. Internat. Ser. Numer. Math. **144**, (2003) (Birkhäuser, Basel)
5. Szegö, G.: Inequalities for the zeros of Legendre polynomials and related functions. Trans. Amer. Math. Soc. **39**, 1–17 (1936)
6. Szegö, G.: Orthogonal Polynomials, vol. 23, 4th edn. Colloquium Publications, Amer. Math. Soc., Providence, RI (1975)

9.16. [190] "On a conjectured inequality for the largest zero of Jacobi polynomials"

---

[190] "On a conjectured inequality for the largest zero of Jacobi polynomials," *Numer. Algorithms* **49**, 195–198 (2008).

Numer Algor (2008) 49:195–198
DOI 10.1007/s11075-008-9207-6

ORIGINAL PAPER

On a conjectured inequality for the largest zero of Jacobi polynomials

Walter Gautschi

Received: 18 March 2008 / Accepted: 9 April 2008 /
Published online: 22 May 2008

Abstract P. Leopardi and the author recently investigated, among other things, the validity of the inequality $n\theta_n^{(\alpha,\beta)} < (n+1)\theta_{n+1}^{(\alpha,\beta)}$ between the largest zero $x_n = \cos\theta_n^{(\alpha,\beta)}$ and $x_{n+1} = \cos\theta_{n+1}^{(\alpha,\beta)}$ of the Jacobi polynomial $P_n^{(\alpha,\beta)}(x)$ resp. $P_{n+1}^{(\alpha,\beta)}(x)$, $\alpha > -1$, $\beta > -1$. The domain in the parameter space (α, β) in which the inequality holds for all $n \geq 1$, conjectured by us, is shown here to require a small adjustment—the deletion of a very narrow lens-shaped region in the square $\{-1 < \alpha < -1/2,\ -1/2 < \beta < 0\}$.

Keywords Jacobi polynomials · Zeros · Inequalities

Mathematics Subject Classification (2000) 33C45

1 Introduction

Let $x_n^{(\alpha,\beta)} = \cos\theta_n^{(\alpha,\beta)}$ be the largest zero of the Jacobi polynomial $P_n^{(\alpha,\beta)}(x)$, where $\alpha > -1$, $\beta > -1$. In [3, §3], the following inequality was considered,

$$n\theta_n^{(\alpha,\beta)} < (n+1)\theta_{n+1}^{(\alpha,\beta)}, \tag{1}$$

where the case $\alpha = \beta = -1/2$ is to be excluded, since both sides of (1) are then equal to $\pi/2$. It was conjectured that the validity of (1) for $n = 1$ implies the validity of (1) for all $n \geq 1$. Using an asymptotic result of L. Gatteschi, we show here that the conjecture is false in a small subregion of the square $-1 < \alpha < -1/2$, $-1/2 < \beta < 0$,

In memoriam Luigi Gatteschi.

W. Gautschi (✉)
Department of Computer Sciences, Purdue University,
West Lafayette, IN 47907-2066, USA
e-mail: wxg@cs.purdue.edu

but believe it to remain valid in the rest of the parameter plane. A revised conjecture is formulated.

2 The conjecture and its disproof

Based on a fair amount of numerical testing, the following conjecture was formulated in [3, §3].

Conjecture *Given $\alpha > -1$, $\beta > -1$, there are two alternatives: either* (1) *holds for all $n = 1, 2, 3, \ldots$, or* (1) *is false for $n = 1$. In other words, the validity of* (1) *for $n = 1$ implies the validity of* (1) *for all $n \geq 1$.*

In order to delineate the conjectured domain of validity of (1), we defined a curve

$$\mathcal{B}: \quad \beta = \beta(\alpha), \quad -1 < \alpha < 1, \tag{2}$$

monotonically descending from the point $(-1, 0)$ to the point $(1, -1)$, above which (1) is true for $n = 1$ (and hence for all $n \geq 1$ if the conjecture is true), and on and below which (1) is false for $n = 1$. Specifically, $\beta = \beta(\alpha)$ is the solution of the equation

$$2\arccos\left(\frac{1}{\alpha+\beta+4}\left[-(\alpha-\beta)+2\sqrt{2+\frac{\alpha\beta-2}{\alpha+\beta+3}}\right]\right) + \arccos\frac{\alpha-\beta}{\alpha+\beta+2} - \pi = 0. \tag{3}$$

The asymptotic behavior for large degree n of the zeros of Jacobi polynomials (ordered decreasingly) has been studied extensively by L. Gatteschi (see also the paper [2] in this issue). A particular result that is relevant to us holds for the kth zero, where k is fixed, and specialized to the case $k = 1$ reads as follows [1, Theorem 4.1]: If $\alpha > -1$ and $\beta > -1$ (actually, β could be arbitrary real), then

$$\theta_n^{(\alpha,\beta)} = \frac{j_{\alpha,1}}{\nu} + O(n^{-5}), \quad n \to \infty, \tag{4}$$

where $j_{\alpha,1}$ is the first positive zero of the Bessel function J_α and

$$\nu = \nu(n) = \left[\left(n + \frac{\alpha+\beta+1}{2}\right)^2 + \frac{1-\alpha^2-3\beta^2}{12}\right]^{1/2}. \tag{5}$$

From this, it follows that

$$\frac{\theta_n^{(\alpha,\beta)}}{\theta_{n+1}^{(\alpha,\beta)}} = \frac{\frac{j_{\alpha,1}}{\nu(n)} + O(n^{-5})}{\frac{j_{\alpha,1}}{\nu(n+1)} + O(n^{-5})} = \frac{\nu(n+1)}{\nu(n)} + O(n^{-4}).$$

Expanding the function

$$\frac{\nu(n+1)}{\nu(n)} = \left[\frac{\left(1+\frac{\alpha+\beta+3}{2n}\right)^2 + \frac{1-\alpha^2-3\beta^2}{12n^2}}{\left(1+\frac{\alpha+\beta+1}{2n}\right)^2 + \frac{1-\alpha^2-3\beta^2}{12n^2}}\right]^{1/2}$$

in descending powers of n, using Maple V, one finds

$$\frac{\theta_n^{(\alpha,\beta)}}{\theta_{n+1}^{(\alpha,\beta)}} = 1 + n^{-1} - \frac{1}{2}(\alpha+\beta+1)n^{-2} + \frac{1}{6}(2\alpha^2+3\alpha\beta+3\beta^2+3\alpha+3\beta+1)n^{-3} + O(n^{-4}). \tag{6}$$

Theorem *If $\alpha+\beta+1>0$, the inequality* (1) *is valid for n sufficiently large. If $\alpha+\beta+1=0$, the same is true when (α,β) is located on the open line segment from $(-1,0)$ to $(-1/2,-1/2)$, but for large enough n is false on the half-open line segment from $(-1/2,-1/2)$ (inclusive) to $(0,-1)$. If $\alpha+\beta+1<0$, the inequality* (1) *is false for n sufficiently large.*

Proof Inequality (1) can be written in the form

$$\frac{\theta_n^{(\alpha,\beta)}}{\theta_{n+1}^{(\alpha,\beta)}} < 1 + n^{-1}. \tag{7}$$

From (6) it can be seen that the ratio on the left of (7), if $\alpha+\beta+1>0$, is less than $1+n^{-1}$ if n is sufficiently large, so that (7), and thus also (1), is true for such values of n. This proves the first part of the theorem. To prove the second part, we have to examine the coefficient of n^{-3} on the right of (6) in the case that $\beta=-\alpha-1$, i.e. the expression $\frac{1}{6}(2\alpha+1)(\alpha+1)$. The latter, under the assumption that $\alpha>-1$, is negative precisely if $\alpha<-1/2$. This proves the second part of the theorem, while the last part follows by the same argument used in the first part. □

In order to disprove the conjecture, all we need to show is that the domains $\alpha+\beta+1<0$ and $\beta(\alpha)>0$ have a nonempty intersection $\mathcal{S}$ in the square $\{-1<\alpha<0, -1<\beta<0\}$. But this becomes evident if we look at the graph of $\beta=-\alpha-1-\beta(\alpha)$ shown in Fig. 1. Indeed, if we pick a point (indicated by a circle in Fig. 1) at the center of the positive bulge of this graph, which corresponds to the point $\alpha=-.75$,

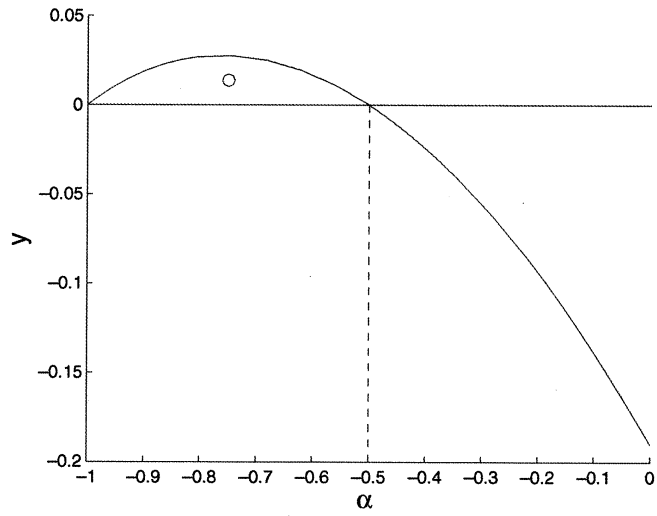


Fig. 1 The graph of $\beta=-\alpha-1-\beta(\alpha)$

$\beta = \frac{1}{2}(\beta(\alpha) - \alpha - 1) \approx -.2638$ in $\mathcal{S}$, we find that the inequality (1) is true for $n = 1$ and $n = 2$, but false for $3 \leq n \leq 100$. In our earlier testing we somehow missed the intersection $\mathcal{S}$, which is so slim as to be barely visible by the naked eye; see [3, Fig. 1].

3 A revised conjecture

We conclude with formulating our new conjecture.

Revised Conjecture *With the exception of the point $\alpha = \beta = -1/2$, the domain of validity in the (α, β)-plane of the inequality* (1) *for all $n \geq 1$ is the subdomain $\mathcal{D}$ of all admissible $\{(\alpha, \beta) : \alpha > -1, \beta > -1\}$ bounded below by the line segment $\mathcal{C}_1$ from the point $(-1, 0)$ to $(-1/2, -1/2)$, the part $\mathcal{C}_2 = \{(\alpha, \beta) : \beta = \beta(\alpha), -1/2 \leq \alpha < 1\}$ of the curve $\mathcal{B}$, and the line $\mathcal{C}_3 = \{(\alpha, \beta) : 1 \leq \alpha < \infty, \beta = -1\}$.*

In effect, the only difference between the original and the revised conjecture is the replacement of the curved segment $\{(\alpha, \beta) : \beta = \beta(\alpha), -1 < \alpha < -1/2\}$ in the original boundary of the domain of validity by the line segment from $(-1, 0)$ to $(-1/2, -1/2)$. In particular, Conjecture 1 of [3] relating to the Jacobi polynomials $P_n^{(\alpha,\alpha-1)}$ remains unaffected by this change, and so does the conjecture for ultraspherical polynomials $P_n^{(\alpha,\alpha)}$.

It may be of interest to compare some of the statements in our theorem with actual computational results. On the line segment $\mathcal{C}_1$ (which forms part of the boundary of $\mathcal{D}$), the inequality (1) according to the theorem holds for n sufficiently large. Computation suggests that it holds for all $n \geq 1$. Likewise, on the continuation of the line segment, from $(-1/2, -1/2)$ to $(0, -1)$ (which is not part of the boundary of $\mathcal{D}$), inequality (1) by our theorem is false for n sufficiently large. Computation suggests that it is false for all $n \geq 1$.

References

1. Gatteschi, L.: On the zeros of Jacobi polynomials and Bessel functions. In: International conference on special functions: theory and computation (Turin, 1984). Rend. Sem. Mat. Univ. Politec. Torino (Special Issue), pp. 149–177 (1985)
2. Gautschi, W., Giordano, C.: Luigi Gatteschi's work on asymptotics of special functions and their zeros. Numer. Algorithms. doi:10.1007/s11075-008-9208-5
3. Gautschi, W., Leopardi, P.: Conjectured inequalities for Jacobi polynomials and their largest zeros. Numer. Algorithms **45**(1–4), 217–230 (2007)

9.17. [191] "On conjectured inequalities for zeros of Jacobi polynomials"

---

[191] "On conjectured inequalities for zeros of Jacobi polynomials," *Numer. Algorithms* **50**, 93–96 (2009).

Numer Algor (2009) 50:93–96
DOI 10.1007/s11075-008-9217-4

ORIGINAL PAPER

On conjectured inequalities for zeros of Jacobi polynomials

Walter Gautschi

Received: 15 May 2008 / Accepted: 26 May 2008 /
Published online: 27 June 2008

Abstract Inequalities for the largest zero of Jacobi polynomials, conjectured recently by us and in joint work with P. Leopardi, are here extended to all zeros of Jacobi polynomials, and new relevant conjectures are formulated based on extensive computation.

Keywords Jacobi polynomials · Zeros · Inequalities

Mathematics Subject Classification (2000) 33C45

1 Introduction

Let $x_{n,r}^{(\alpha,\beta)} = \cos\theta_{n,r}^{(\alpha,\beta)}$, $r = 1, 2, \ldots, n$, be the zeros in descending order of the Jacobi polynomial $P_n^{(\alpha,\beta)}$,

$$0 < \theta_{n,1}^{(\alpha,\beta)} < \theta_{n,2}^{(\alpha,\beta)} < \cdots < \theta_{n,n}^{(\alpha,\beta)} < \pi. \tag{1}$$

The object here is to determine (numerically) the domain of validity in the (α, β)-plane of the inequalities

$$n\,\theta_{n,r}^{(\alpha,\beta)} < (n+1)\,\theta_{n+1,r}^{(\alpha,\beta)}, \quad r = 1, 2, \ldots, n. \tag{2}$$

For $r = 1$, this was done in [2] and slightly revised in [1]. Since in this case, i.e., for r fixed, the inequality (2) can be proven to hold for all sufficiently large values of n in the conjectured domain of validity, this domain does not in any

W. Gautschi (✉)
Department of Computer Sciences, Purdue University,
West Lafayette, 47907-2066 IN, USA
e-mail: wxg@cs.purdue.edu

way depend on n. If we include all values of r, as in (2), this will no longer be true. On the contrary, the set of inequalities (2), especially the one for $r = n$, appears to be false for all n sufficiently large in extended portions of the (α, β)-plane. Thus, if we assume $1 \le n \le N$, the domain of validity of (2) will depend on N. We will consider the cases $N = 50$, $N = 100$, and $N = 200$, and on the basis of the results obtained, conjecture that as $N \to \infty$ there is a limit domain in the (α, β)-plane in which the inequalities (2) hold unrestrictedly for all n.

We begin in Section 2 with focusing on the smallest zero, $r = n$, in the case of ultraspherical polynomials, before dealing with the general case and general Jacobi polynomials in Section 3.

2 Ultraspherical polynomials

We let $\alpha = \beta = \lambda - 1/2$ and denote $\theta_{n,r}^{(\alpha,\beta)} = \theta_{n,r}^{(\lambda)}$. We may assume $\lambda > 0$, since we know from [2] that this is necessary for (2) to hold when $r = 1$. Suppose now that $r = n$. By symmetry,

$$\theta_{n,n}^{(\lambda)} = \pi - \theta_{n,1}^{(\lambda)} \quad \text{and} \quad \theta_{n+1,n}^{(\lambda)} = \pi - \theta_{n+1,2}^{(\lambda)},$$

so that (2) for $r = n$ takes the form

$$(n+1)\,\theta_{n+1,2}^{(\lambda)} - n\,\theta_{n,1}^{(\lambda)} < \pi. \tag{3}$$

In the special case $\lambda = 1$ one has $\theta_{n,1} = \pi/(n+1)$, $\theta_{n+1,2} = 2\pi/(n+2)$, and the left-hand side of (3) becomes

$$\frac{n^2 + 2n + 2}{(n+1)(n+2)}\,\pi < \pi.$$

Here, (3) is valid for all $n \ge 1$. Numerical computation suggests that the same is true for $0 \le \lambda \le 1$.

For each of the three cases $N (= \max n) = 50, 100, 200$, we now let λ increase from 1 in steps of $\Delta\lambda = .1$ until a $\lambda = \lambda^*$ is found for which (3) is false for some $n \le N$. Then a method of bisection is applied to the interval $[\lambda^* - \Delta\lambda, \lambda^*]$ in order to zero-in to a more accurate value $\lambda = \lambda_0$ and interval $(0, \lambda_0]$ in which (3) holds for all $n \le N$. This is implemented in the Matlab routine `uspconj.m`, which uses `uspineq.m`. (These routines, as well as the others referenced in this note, can be downloaded from the web site http://www.cs.purdue.edu/archives/2002/wxg/codes by clicking on `CIZJP`.) For the values $\alpha_0 = \lambda_0 - 1/2$ it is found that

$$\alpha_0 = 2.2009763331 \quad \text{if} \quad N = 50,$$
$$\alpha_0 = 1.0605211988 \quad \text{if} \quad N = 100,$$
$$\alpha_0 = .7412587491 \quad \text{if} \quad N = 200.$$

A set of 1000 points randomly selected in $(0, \lambda_0]$ was then successfully tested to add further credence for the desired property (3) to indeed hold in $(0, \lambda_0]$.

3 General Jacobi polynomials

We adopt the following strategy to explore the general case of Jacobi polynomials $P_n^{(\alpha,\beta)}$, $\alpha > -1$, $\beta > -1$, with $1 \le n \le N$ and N as in Section 1. We first determine, in α-steps $\Delta\alpha = .02$, that the inequalities (2) (always for $n \le N$) are valid on the horizontal line segment $\mathcal{H} = \{(\alpha, \beta) : -.5 < \alpha \le 5, \beta = -.5\}$ as well as on the downward diagonal segment $\mathcal{K} = \{(\alpha, \beta) : -1 < \alpha < -.5, \beta = -\alpha - 1\}$. Next, consider the vertical lines $\mathcal{L}_\alpha$ with abscissae $\alpha = -.5 : \Delta\alpha : 5$ and denote by $\mathcal{L}_\alpha^+$ [$\mathcal{L}_\alpha^-$] the part of $\mathcal{L}_\alpha$ above [below] $\mathcal{K}$ resp. $\mathcal{H}$. On each $\mathcal{L}_\alpha^+$ we move upwards from $\mathcal{K}$ resp. $\mathcal{H}$ in steps of $\Delta\beta = .1$ until a point (α, β^*) is encountered for which one of the inequalities is false. (Invariably, it turned out to be the one for $r = n, n \le N$.) Thereafter, similarly as in Section 2, we apply a method of bisection on the β-interval $[\beta^* - \Delta\beta, \beta^*]$ to determine a more precise value β_0 so that on the vertical line segment $\mathcal{L}_\alpha^+$ bounded above by the point (α, β_0) all inequalities (2) are valid for $n \le N$. This gives rise to the slightly concave curves $\mathcal{B}_N$ shown in Fig. 1 for $N = 50, 100, 200$. (When $N = 200$, to avoid excessive computing times, we checked only the last inequality in (2) for $r = n$.)

We know from [1] that vertically below $\mathcal{K}$ the inequality (2) for $r = 1$ is false for all n sufficiently large. Therefore, it remains to examine the line segments $\mathcal{L}_\alpha^-$ for $\alpha > -.5$. Here we do the same as for $\mathcal{L}_\alpha^+$, but moving downwards in steps of $\Delta\beta = .02$. This gives rise to the curves $\mathcal{C}_N$, also shown in Fig. 1.

All of this is implemented in the routine `testconj.m`, which uses `jacconj.m` and `ineq.m`. Some of the control commands in the second of these routines must be adapted, as indicated by commented-out statements, to whether the curves $\mathcal{B}_N$ or $\mathcal{C}_N$ are to be computed, and in the former case also to whether or not $-1 < \alpha \le -.5$.

The domain of validity of (2) for $1 \le n \le N$ is thus *the domain D_N bounded above by $\mathcal{B}_N$, on the left by the vertical segment at $\alpha = -1$ between $\mathcal{B}_N$ and $\mathcal{K}$, and below by $\mathcal{K}$ followed by $\mathcal{C}_N$*. As seen in Fig. 1, the curves $\mathcal{B}_N$ and $\mathcal{C}_N$

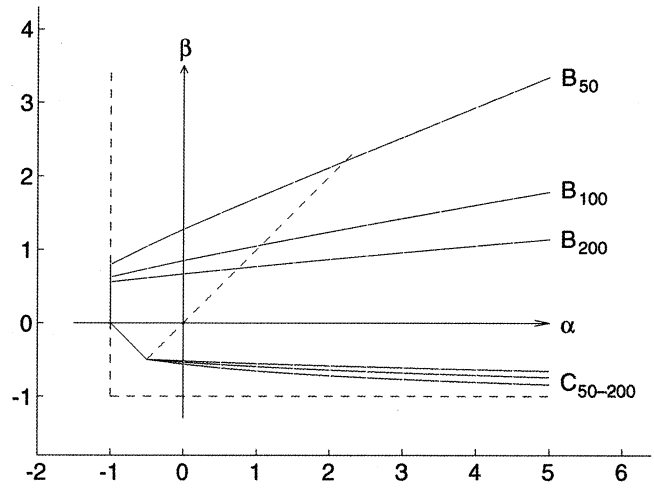


Fig. 1 Domains of validity of (2)

turn downward resp. upward with increasing N. Very likely they tend toward horizontal lines $\beta = \pm.5$ as $N \to \infty$. If so, the *inequalities* (2) *are expected to be valid for all n in the horizontal strip* $S = \{(\alpha, \beta) : \alpha > -1, |\beta| \leq .5\}$ *cut off on the left by the line segment* $\mathcal{K}$, and, as always, with the point $(-.5, -.5)$ removed. This is consistent with the findings in Section 2 for ultraspherical polynomials, and in fact was partially (checking only the last inequality in (2)) reinforced for $N = 500$ on 100 points selected randomly in the strip $\{(\alpha, \beta) : -.5 \leq \alpha \leq 20, |\beta| \leq .5\}$.

References

1. Gautschi, W.: On a conjectured inequality for the largest zero of Jacobi polynomials. Numer. Algorithms (2008). doi:10.1007/s11075-008-9207-6
2. Gautschi, W., Leopardi, P.: Conjectured inequalities for Jacobi polynomials and their largest zeros. Numer. Algorithms **45**(1–4), 217–230 (2007)

9.18. [192] "New conjectured inequalities for zeros of Jacobi polynomials"

---

[192] "New conjectured inequalities for zeros of Jacobi polynomials," *Numer. Algorithms* **50**, 293–296 (2009).

Numer Algor (2009) 50:293–296
DOI 10.1007/s11075-008-9230-7

ORIGINAL PAPER

New conjectured inequalities for zeros of Jacobi polynomials

Walter Gautschi

Received: 8 July 2008 / Accepted: 28 July 2008 /
Published online: 16 August 2008

Abstract Inequalities recently conjectured for all zeros of Jacobi polynomials $P_n^{(\alpha,\beta)}$ of all degrees n are modified and conjectured to hold (in reverse direction) in considerably larger domains of the (α, β)-plane.

Keywords Jacobi polynomials · Zeros · Inequalities

Mathematics Subject Classification (2000) 33C45

1 Introduction

Inequalities for the largest zero of Jacobi polynomials, recently conjectured by Leopardi and the author [5], and slightly revised in [2], have been extended by us in [3] to all zeros of Jacobi polynomials. They state that for each $r = 1, 2, \ldots, n$ the sequence $\{n\theta_{n,r}^{(\alpha,\beta)}\}$ in an appropriate domain $\mathcal{D}$ of the (α, β)-plane is monotonically increasing, where $\cos\theta_{n,r}^{(\alpha,\beta)} = x_{n,r}^{(\alpha,\beta)}$ are the zeros in descending order of the Jacobi polynomial $P_n^{(\alpha,\beta)}(x)$. When $r \geq 1$ is fixed, this sequence tends to the rth positive zero of the Bessel function J_α, and convergence is monotone increasing in $\mathcal{D}$ if the conjecture is true. In the theory of Jacobi polynomials and their zeros, it is often observed that the factor $n + (\alpha + \beta + 1)/2$ is more natural than n, and therefore Askey (Email of May 13, 2008) asked the author to examine computationally whether similar inequalities hold with the factor so modified, and in what domains $\mathcal{D}$. Here we report on the results of these investigations.

W. Gautschi (✉)
Department of Computer Sciences, Purdue University, West Lafayette, IN 47907-2066, USA
e-mail: wxg@cs.purdue.edu

2 A preliminary asymptotic result

Explorative computations suggested that the inequalities to be studied should be

$$(n+(\alpha+\beta+1)/2)\theta_{n,r}^{(\alpha,\beta)} > (n+(\alpha+\beta+3)/2)\theta_{n+1,r}^{(\alpha,\beta)}, \quad r=1,2,\ldots,n. \tag{1}$$

We first examine what to expect if r is fixed and $n \to \infty$.

Using a result of Gatteschi [1, Thm. 4.1] (cf. also [4, Section 5.6]), we found in [2, eq (6)] that for r fixed, there holds, as $n \to \infty$,

$$\frac{\theta_{n,r}^{(\alpha,\beta)}}{\theta_{n+1,r}^{(\alpha,\beta)}} = 1 + n^{-1} - \frac{1}{2}(\alpha+\beta+1)n^{-2} + \frac{1}{6}\left(2\alpha^2+3\alpha\beta+3\beta^2+3\alpha+3\beta+1\right)n^{-3} + O\left(n^{-4}\right), \tag{2}$$

where $\alpha > -1$ and β can be arbitrary real. The inequality in (1) can be written as

$$\frac{\theta_{n,r}^{(\alpha,\beta)}}{\theta_{n+1,r}^{(\alpha,\beta)}} > \frac{1+(\alpha+\beta+3)/(2n)}{1+(\alpha+\beta+1)/(2n)} = 1 + n^{-1} - \frac{1}{2}(\alpha+\beta+1)n^{-2} + \frac{1}{4}(\alpha+\beta+1)^2 n^{-3} + O\left(n^{-4}\right). \tag{3}$$

We have agreement of this expansion with the one in (2) up to, and including, the n^{-2} term, whereas the inequalities in [2] give agreement only up to the n^{-1} term. Comparing (3) with (2), we see that the inequality (1) with r fixed holds for all sufficiently large n if

$$\frac{1}{4}(\alpha+\beta+1)^2 < \frac{1}{6}\left(2\alpha^2+3\alpha\beta+3\beta^2+3\alpha+3\beta+1\right),$$

that is, if $\alpha^2+3\beta^2 > 1$, and is false for all n sufficiently large if $\alpha^2+3\beta^2 < 1$. We conclude that, if (1) is to hold for all n and all r, then (α, β) must be outside the ellipse

$$\mathcal{E}: \quad \alpha^2+3\beta^2 = 1, \tag{4}$$

or possibly on the ellipse. Note, however, that the four points with $|\alpha| = |\beta| = 1/2$, which are all located on the circumference of $\mathcal{E}$, are exceptional in that equality holds in (1) for all n and r.

3 First conjectures

When testing inequalities computationally, it is of course difficult to decide whether two real numbers x, y are almost equal, or exactly equal, if we only know their computed (in Matlab arithmetic) values x^*, y^*, respectively. In formulating our conjectures, we therefore adopted the following "working definition": For a small positive ε (close to machine precision), we say that

x is computationally equal to y (in formula, $x \stackrel{*}{=} y$) if $|x^* - y^*| < \varepsilon$ regardless of whether $x^* < y^*$ or $x^* \geq y^*$, that x is computationally less than y ($x \stackrel{*}{<} y$) if $x^* < y^*$ and $|x^* - y^*| \geq \varepsilon$, and that x is computationally larger than y ($x \stackrel{*}{>} y$) if $x^* > y^*$ and $|x^* - y^*| \geq \varepsilon$. As for ordinary order relations, also their computational analogues form a trichotomy. For our present computations in Matlab, we take $\varepsilon = 2 \times 10^{-10}$.

Our first conjectures relate to values of α and β for which the point (α, β) is located in the interior of the ellipse $\mathcal{E}$ of (4),

$$(\alpha, \beta) \in \text{int}(\mathcal{E}). \tag{5}$$

For such points we know from Section 1 that the inequality (1) is false for n sufficiently large. We conjecture it to be false for all $n \geq 1$, at least in a large portion of $\mathcal{E}$.

Conjecture 1 *For* $(\alpha, \beta) \in \mathcal{R}^0_{1/2}$, *where* $\mathcal{R}^0_{1/2} = \{(\alpha, \beta) : |\alpha| \leq 1/2, |\beta| \leq 1/2, \alpha^2 + \beta^2 \neq 1/2\} \subset \text{int}(\mathcal{E})$, *the inequality* (1) *holds with* $>$ *replaced by* $<$ *for all* $n = 1, 2, 3, \ldots$.

Conjecture 1 was tested, using the same software as in [2, 3, 5], for all n with $1 \leq n \leq 100$ and was verified for all grid points $\alpha = ih$, $\beta = jh$, $i, j \in \mathbb{N}$, $h = .01$ contained in $\mathcal{R}^0_{1/2}$. We also note that Conjecture 1 is consistent with the "remarkable property" ([6, Section 6.3(5)]) that the inequality (1) with $<$ instead of $>$ holds for $\alpha = \beta = \lambda - 1/2$, $0 < \lambda < 1$, i.e., on the open line segment from the point $(-1/2, -1/2)$ to the point $(1/2, 1/2)$ lying entirely in $\mathcal{R}^0_{1/2}$. The proof of this property, given in [6], is by a Sturm comparison theorem.

Conjecture 2 *In the parts of the ellipse* $\mathcal{E}$ *adjoining* $\mathcal{R}^0_{1/2}$ *on the left and on the right, including the boundaries, the inequality* (1) *holds for all* $n = 1, 2, 3, \ldots$ *with* $>$ *replaced by* $\leq$.

This conjecture was tested similarly as Conjecture 1. Occasionally, points (α, β) and values of n and r were encountered for which the two sides of the inequality are computationally equal (in the sense defined above). We therefore cannot exclude the possible occurrence of equality, for which reason equality sign is included in the conjecture.

In the remaining parts of $\mathcal{E}$ (on top and bottom of $\mathcal{R}^0_{1/2}$) either inequality sign in (1) was observed, and therefore no general statement can be made for these regions, except for the two boundaries of $\mathcal{E}$, where (1) holds with $\geq$ sign.

4 Main conjecture

Our main conjecture, however, concerns the inequalities as stated in (1).

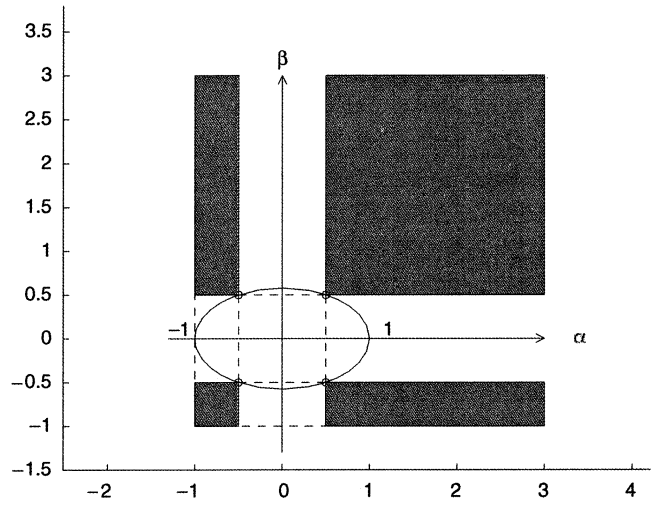


Fig. 1 Domains of validity of (1)

Main Conjecture *The inequalities* (1) *hold in the four rectangular regions (shaded in Fig.* 1 *and extending to infinity in both the* α*- and* β*-directions) having a vertex in one of the four points with* $|\alpha| = |\beta| = 1/2$*, including the boundaries lying in* $\alpha > -1$*,* $\beta > -1$*, but excluding (as already mentioned in Section* 2*) the four vertices on the ellipse* $\mathcal{E}$*.*

The conjecture was verified computationally for $1 \leq n \leq 100$ and for α- and β-values in $[-.99 : .01 : -.50], [.50 : .01 : 1.00 : .05 : 2.00 : .1 : 5.0 : .5 : 10 : 30]$. The remaining regions were similarly investigated. It was found that the inequalities (1) also hold in the strips on top and at the bottom of the ellipse $\mathcal{E}$ (with base interval $-1/2 < \alpha < 1/2$), possibly with equality holding near the ellipse. On the strip to the right of $\mathcal{E}$, as well as in the small remaining pieces to the left of $\mathcal{E}$, either inequality sign can occur in (1).

References

1. Gatteschi, L.: On the zeros of Jacobi polynomials and Bessel functions. In: International Conference on Special Functions: Theory and Computation (Turin, 1984). Rend. Semin. Mat. Univ. Politec. Torino (special issue) **1985**, 149–177 (1985)
2. Gautschi, W.: On a conjectured inequality for the largest zero of Jacobi polynomials. Numer. Algorithms (2008, in press)
3. Gautschi, W.: On conjectured inequalities for zeros of Jacobi polynomials. Numer. Algorithms (2008, in press)
4. Gautschi, W., Giordano, C.: Luigi Gatteschi's work on aymptotics of special functions and their zeros. Numer. Algorithms (2008, in press)
5. Gautschi, W., Leopardi, P.: Conjectured inequalities for Jacobi polynomials and their largest zeros. Numer. Algorithms **45**(1–4), 217–230 (2007)
6. Szegö, G.: Orthogonal Polynomials, 4th edn., vol. 23. American Mathematical Society, Colloquium Publications. American Mathematical Society, Providence (1975)

9.19. [193] "HOW SHARP IS BERNSTEIN'S INEQUALITY FOR JACOBI POLYNOMIALS?"

---

[193] "How Sharp is Bernstein's Inequality for Jacobi Polynomials?" *Electr.Trans. Numer. Anal.* **36**, 1–8 (2009).

Electronic Transactions on Numerical Analysis.
Volume 36, pp. 1-8, 2009.

ISSN 1068-9613.

ETNA
Kent State University
http://etna.math.kent.edu

HOW SHARP IS BERNSTEIN'S INEQUALITY FOR JACOBI POLYNOMIALS?*

WALTER GAUTSCHI†

Dedicated to Richard S. Varga on his 80th birthday

Abstract. Bernstein's inequality for Jacobi polynomials $P_n^{(\alpha,\beta)}$, established in 1987 by P. Baratella for the region $\mathcal{R}_{1/2} = \{|\alpha| \le 1/2, |\beta| \le 1/2\}$, and subsequently supplied with an improved constant by Y. Chow, L. Gatteschi, and R. Wong, is analyzed here analytically and, above all, computationally with regard to validity and sharpness, not only in the original region $\mathcal{R}_{1/2}$, but also in larger regions $\mathcal{R}_s = \{-1/2 \le \alpha \le s, -1/2 \le \beta \le s\}$, $s > 1/2$. Computation suggests that the inequality holds with new, somewhat larger, constants in any region $\mathcal{R}_s$. Best constants are provided for $s = 1:.5:4$ and $s = 5:1:10$. Our work also sheds new light on the so-called Erdélyi–Magnus–Nevai conjecture for orthonormal Jacobi polynomials, adding further support for its validity and suggesting .66198126... as the best constant implied in the conjecture.

Key words. Bernstein's inequality, Jacobi polynomials, sharpness, Erdélyi–Magnus–Nevai conjecture

AMS subject classifications. 33C45, 41A17

1. Introduction. Bernstein's inequality for Legendre polynomials P_n, slightly sharpened by Antonov and Holševnikov [1] and Lorch [5], states that for $n = 1, 2, 3, \ldots$,

$$(\sin\theta)^{1/2}|P_n(\cos\theta)| < \left(\frac{2}{\pi}\right)^{1/2}\left(n+\tfrac{1}{2}\right)^{-1/2}, \quad 0 \le \theta \le \pi. \tag{1.1}$$

According to Bernstein, the constant $(2/\pi)^{1/2}$ is best possible. An extension of (1.1) to ultraspherical polynomials $P_n^{(\lambda)} = P_n^{(\lambda-1/2,\lambda-1/2)}$, $0 < \lambda < 1$, is due to Lorch [6], and a further extension to Jacobi polynomials $P_n^{(\alpha,\beta)}$ with $|\alpha| \le 1/2$, $|\beta| \le 1/2$ to Baratella [2]. Chow, Gatteschi, and Wong [3], by sharpening her constant, improved Baratella's result to read

$$\begin{aligned}(\sin\tfrac{1}{2}\theta)^{\alpha+1/2}(\cos\tfrac{1}{2}\theta)^{\beta+1/2}|P_n^{(\alpha,\beta)}(\cos\theta)| &\le \frac{\Gamma(q+1)}{\Gamma(1/2)}\binom{n+q}{n}N^{-q-1/2},\\ N = n+(\alpha+\beta+1)/2,\ \ 0 \le \theta \le \pi,\end{aligned} \tag{1.2}$$

where $q = \max(\alpha,\beta)$ and $|\alpha| \le 1/2$, $|\beta| \le 1/2$. Equality sign is included in (1.2), since in the case $\alpha = \beta = \mp 1/2$ the inequality reduces to $|\cos(n\theta)| \le 1$ resp. $|\sin((n+1)\theta)| \le 1$, and in the case $\alpha = \pm 1/2$, $\beta = \mp 1/2$ to $|\sin(n+1/2)\theta| \le 1$ resp. $|\cos(n+1/2)\theta| \le 1$. It appears, though, that strict inequality holds in all other cases.

Squaring both sides of the inequality (1.2) and writing the result in terms of $x = \cos\theta$ and the orthonormal Jacobi polynomial $\hat{P}_n^{(\alpha,\beta)}$ yields (if $\beta \ge \alpha$; cf. (4.1))

$$\begin{aligned}&(1-x)^{\alpha+1/2}(1+x)^{\beta+1/2}[\hat{P}_n^{(\alpha,\beta)}(x)]^2\\ &\quad\le \frac{2\,\Gamma(n+\alpha+\beta+1)\Gamma(n+\beta+1)}{\pi\,\Gamma(n+\alpha+1)n!(n+(\alpha+\beta+1)/2)^{2\beta}}, \quad |\alpha| \le 1/2,\ |\beta| \le 1/2.\end{aligned} \tag{1.3}$$

*Received April 25, 2008. Accepted for publication November 22, 2008. Published online June 10, 2009. Recommended by V. Andriyevskyy.

†Department of Computer Sciences, Purdue University, West Lafayette, Indiana 47907-2066, USA (wxg@cs.purdue.edu).

ETNA
Kent State University
http://etna.math.kent.edu

Since as $n \to \infty$ the right-hand side is $\sim 2/\pi$, it follows that the left-hand side is $O(1)$ for $|x| \leq 1$, which proves the Erdélyi–Magnus–Nevai conjecture

$$(1.4) \qquad (1-x)^{\alpha+1/2}(1+x)^{\beta+1/2}[\hat{P}_n^{(\alpha,\beta)}(x)]^2 = O\left(\max[1,(\alpha^2+\beta^2)^{1/4}]\right)$$

[7, p. 604] (see also [4]) on the domain $|\alpha| \leq 1/2$, $|\beta| \leq 1/2$. The constant on the right of (1.3) takes on the value $2/\pi$ not only at $n = \infty$, but also at $n = 1$ when $\beta = 0$ or $|\alpha| = |\beta| = 1/2$. It is probably for $n = 1$ and $\beta = 1/2$ that the maximum is attained, near $\alpha = -.0691$, its value being .64297807.

Incidentally, if we denote the ratio of the left-hand side of (1.2) and the right-hand side (as in (3.2), (3.3)) by $c_n F_n(x)$, we have

$$(1.5) \qquad (1-x)^{\alpha+1/2}(1+x)^{\beta+1/2}[\hat{P}_n^{(\alpha,\beta)}(x)]^2 = \gamma_n c_n^2 F_n^2(x),$$

where

$$(1.6) \qquad \gamma_n = \frac{2\,\Gamma(n+\alpha+\beta+1)\Gamma(n+\beta+1)}{\pi\,\Gamma(n+\alpha+1)n!(n+(\alpha+\beta+1)/2)2^{2\beta}}.$$

While the constant $\Gamma(q+1)/\Gamma(1/2)$ in (1.2), when $\alpha = \beta = 0$, is best possible, it does not follow necessarily that the same is true in the general case, although asymptotic arguments will suggest that it is. In this note, the sharpness of the inequality is determined computationally, at least for $n \leq 100$, in the square $|\alpha| \leq 1/2$, $|\beta| \leq 1/2$. Outside thereof, it is examined to what extent the inequality is an underestimation. We will also experiment with different choices of the parameter q, which, asymptotically, is irrelevant.

All of this will be done by computing the infinity norm $\rho_n = \rho_n(\alpha,\beta,q)$ (on the interval $0 \leq \theta \leq \pi$) of the ratio of the left-hand side of (1.2) divided by the right-hand side. This is an important quantity inasmuch as it allows us to assess the quality of the inequality (1.2) on a domain $\mathcal{D}$ of the parameter space (n,α,β,q). In fact, let $\rho_{\mathcal{D}}^+ = \max_{\mathcal{D}} \rho_n(\alpha,\beta,q)$ and $\rho_{\mathcal{D}}^- = \min_{\mathcal{D}} \rho_n(\alpha,\beta,q)$. Then, if $\rho_{\mathcal{D}}^+ \leq 1$, i.e., the inequality holds on $\mathcal{D}$, on a scale from 0 to 1, the best degree of sharpness of (1.2) on $\mathcal{D}$ is $\rho_{\mathcal{D}}^+$, and the worst degree of sharpness on $\mathcal{D}$ is $\rho_{\mathcal{D}}^-$. If $\rho_{\mathcal{D}}^+ > 1$, then the inequality on the domain $\mathcal{D}$ should be modified by multiplying the right-hand side by $\rho_{\mathcal{D}}^+$, to make it valid on $\mathcal{D}$. The best and worst degrees of sharpness, $\hat{\rho}_{\mathcal{D}}^+$, $\hat{\rho}_{\mathcal{D}}^-$ of the modified inequality are then $\hat{\rho}_{\mathcal{D}}^+ = 1$, $\hat{\rho}_{\mathcal{D}}^- = \rho_{\mathcal{D}}^-/\rho_{\mathcal{D}}^+$.

2. The constant in (1.2) is sharp. An elementary computation, using Stirling's formula, will show that the right-hand side of (1.2), as $n \to \infty$, is asymptotically equivalent to $(\pi n)^{-1/2}$, regardless of the values of the parameters α, β, and q. The inequality (1.2) thus says that the function on the left, multiplied by $(\pi n)^{1/2}$, is less than, or equal to, a constant that tends to 1 as $n \to \infty$. But Darboux's formula [8, Theorem 8.21.8] tells us that this same expression, at least on a compact subinterval of $0 < \theta < \pi$, but for arbitrary real α and β, is $\leq 1 + O(1/n)$, where the constant 1 is best possible (bounding, as it does, a cosine function). This not only shows that the constant $\Gamma(q+1)/\Gamma(1/2)$ in (1.2) is indeed best possible, but also that the inequality, with the constant somewhat enlarged, may well hold in larger domains of the (α,β)-plane. The purpose of this note is to explore this computationally in some detail.

3. Bernstein's inequality for monic Jacobi polynomials. In what follows, we prefer to use the monic Jacobi polynomial $\pi_n^{(\alpha,\beta)}$, i. e.,

$$P_n^{(\alpha,\beta)}(x) = k_n \pi_n^{(\alpha,\beta)}(x), \quad k_n = 2^{-n}\binom{2n+\alpha+\beta}{n},$$

ETNA
Kent State University
http://etna.math.kent.edu

and we shall write it as

$$\pi_n^{(\alpha,\beta)}(x) = \prod_{r=1}^{n}(x - x_r) \tag{3.1}$$

in terms of the zeros $x_r = x_{n,r}^{(\alpha,\beta)}$ (in ascending order) of the Jacobi polynomial $P_n^{(\alpha,\beta)}$. If we divide both sides of (1.2) by the expression on its right-hand side, and let $x = \cos\theta$, Bernstein's inequality takes the form

$$c_n|F_n(x)| \le 1, \quad -1 \le x \le 1, \tag{3.2}$$

where

$$\begin{aligned} c_n = c_n(\alpha,\beta,q) &= \frac{\sqrt{\pi}(n+(\alpha+\beta+1)/2)^{q+1/2}\binom{2n+\alpha+\beta}{n}}{\Gamma(q+1)2^{n+(\alpha+\beta+1)/2}\binom{n+q}{n}}, \\ F_n(x) = F_n^{(\alpha,\beta)}(x) &= (1-x)^{(2\alpha+1)/4}(1+x)^{(2\beta+1)/4}\pi_n^{(\alpha,\beta)}(x), \end{aligned} \tag{3.3}$$

where $q = \max(\alpha,\beta)$. Since we later consider q to be an independent parameter, we include it in the constant c_n as one of three parameters. Notice that

$$c_n(\alpha,\beta,q) = c_n(\beta,\alpha,q),$$

regardless of how $q = q(\alpha,\beta)$ is defined so long as $q(\alpha,\beta) = q(\beta,\alpha)$.

4. The infinity norm $\|F_n\|_\infty$ of F_n. We now wish to compute $\|F_n\|_\infty = \max_{-1\le x\le 1}|F_n(x)|$. Since by the reflection formula for Jacobi polynomials,

$$\|F_n^{(\alpha,\beta)}\|_\infty = \|F_n^{(\beta,\alpha)}\|_\infty,$$

it suffices to consider $\beta \ge \alpha$, and since $\|F_n^{(\alpha,\beta)}\|_\infty = \infty$ if $2\alpha + 1 < 0$, we may assume

$$\beta \ge \alpha \ge -1/2. \tag{4.1}$$

Computing $\|F_n\|_\infty$ amounts to computing the local extrema of F_n in the interior of the interval $[-1,1]$ along with $|F_n(\pm 1)|$. With regard to the former, we have

$$F_n'(x) = \tfrac{1}{2}(1-x)^{(2\alpha-3)/4}(1+x)^{(2\beta-3)/4}\{[\beta-\alpha-(\alpha+\beta+1)x]\pi_n^{(\alpha,\beta)}(x)+2(1-x^2)\pi_n^{(\alpha,\beta)\prime}(x)\},$$

so that the local extrema occur at those roots of the equation $[\beta-\alpha-(\alpha+\beta+1)x]\pi_n^{(\alpha,\beta)}(x)+2(1-x^2)\pi_n^{(\alpha,\beta)\prime}(x) = 0$ that are inside $(-1,1)$, that is, dividing by $\pi_n^{(\alpha,\beta)}$ and noting (3.1), at the respective roots of

$$f(x) = 0, \quad f(x) = \beta - \alpha - (\alpha+\beta+1)x + 2(1-x^2)\sum_{r=1}^{n}\frac{1}{x-x_r}. \tag{4.2}$$

There can be at most $n+1$ real roots. To discuss their location, we first observe that

$$f(-1) = 2\beta + 1, \quad f(1) = -(2\alpha+1).$$

It is clear from from (4.2) that

$$f(x_r + 0) = +\infty, \quad f(x_r - 0) = -\infty, \quad r = 1,2,\ldots,n,$$

ETNA
Kent State University
http://etna.math.kent.edu

and on each interval (x_r, x_{r+1}), $r = 1, 2, \ldots n-1$, the function f descends monotonically (cf. Section 5) from $+\infty$ to $-\infty$. It therefore crosses the real line exactly once, accounting for $n-1$ internal extrema. We distinguish three cases with regard to the parameter α. If, first, $2\alpha + 1 > 0$, and hence by (4.1) also $2\beta + 1 > 0$, then $f(-1) > 0$ and $f(1) < 0$, so that there are two more roots, one each in $(-1, x_1)$ and $(x_n, 1)$, accounting for two more internal extrema, and thus for a complete set of $n+1$ extrema. If, secondly, $2\alpha + 1 = 0$, there are two subcases: $2\beta + 1 > 0$ and $2\beta + 1 = 0$. In the former, there is still a local extremum in $(-1, x_1)$, but none in $(x_n, 1)$; in the latter, both these lateral intervals are devoid of local extrema (in fact, this is one of the trivial cases noted in Section 1, in which $c_n \|F_n\|_\infty = 1$.) Finally, in the third case, $2\alpha + 1 < 0$, as was already mentioned, $\|F_n\|_\infty = \infty$.

5. Computing $\|F_n\|_\infty$ in terms of local extrema. To compute a local extremum of F_n, say in the interval (a, b), $-1 \le a < b \le 1$, we use Newton's method applied to the equation (4.2), with the midpoint of the interval (a, b) as the initial approximation,

$$x^{(i+1)} = x^{(i)} - \frac{f(x^{(i)})}{f'(x^{(i)})}, \quad i = 0, 1, 2, \ldots, \quad x^{(0)} = (a+b)/2. \tag{5.1}$$

Since the interval (a, b) in our application is small and f rapidly descending from $+\infty$ to $-\infty$ (i.e., f' is large negative), Newton's iteration (5.1) converges very quickly. The derivative of f is easily computed from (4.2),

$$f'(x) = -(\alpha + \beta + 1) - 2\sum_{r=1}^{n} \frac{x^2 - 2x_r x + 1}{(x - x_r)^2}.$$

Since $\alpha + \beta + 1 \ge 0$ by (4.1), and the discriminants of the quadratics in the numerator on the right are $-4(1 - x_r^2) < 0$, each term of the sum is positive and $f'(x) < 0$ on (a, b), as already noted in the previous section. Thus we arrive at the following

Computational procedure.
If $\alpha > -1/2$, apply (5.1) to the intervals $(a, b) = (x_r, x_{r+1})$, $r = 0, 1, 2, \ldots, n$ (where $x_0 = -1$, $x_{n+1} = 1$), giving $\xi_r = x^{(\infty)}$. Then, since $F_n(\pm 1) = 0$,

$$\|F_n\|_\infty = \max_{0 \le r \le n} |F_n(\xi_r)|, \quad 2\beta + 1 \ge 2\alpha + 1 > 0. \tag{5.2}$$

If $\alpha = -1/2$ and $\beta > -1/2$, do the same, but in (5.2) let r run only up to $n-1$, and compute

$$\|F_n\|_\infty = \max\left\{F_n(1), \max_{0 \le r \le n-1} |F_n(\xi_r)|\right\}, \quad 2\beta + 1 > 0 = 2\alpha + 1. \tag{5.3}$$

If $\alpha = \beta = -1/2$, put $c_n \|F_n\|_\infty = 1$.

The Matlab script `bernstein.m` listed in the Appendix implements this procedure and for any given n, α, β, q outputs $\rho_n(\alpha, \beta, q) = c_n(\alpha, \beta, q)\|F_n^{(\alpha,\beta)}\|_\infty$.

6. Numerical results. In this section we present numerical results for the square $|\alpha| \le 1/2$, $|\beta| \le 1/2$. We determine $\rho_{\mathcal{D}}^+$ and $\rho_{\mathcal{D}}^-$ (cf. Section 1) on the domain $\mathcal{D} = \{n = [5\ 10\ 20\ 50\ 100], \alpha = -.5 : .01 : .5, \beta = \alpha : .01 : .5, q\}$, where in turn $q = q^+ = \max(\alpha, \beta) = \beta$, $q = q^- = \min(\alpha, \beta) = \alpha$, $q = -.75 : .25 : 1$. The results are shown in Table 6.1.

It was observed that the sequence $\{\rho_n(\alpha, \beta, q)\}$ is monotone, either increasing or decreasing. Therefore, if $n_0 \le n \le n_1$, it would suffice to compute ρ_n for $n = n_0$ and

TABLE 6.1
Sharpness of (1.2) on the square $|\alpha| \le 1/2$, $|\beta| \le 1/2$, with selected values of q.

$q \to$	q^+	q^-	0	.25	.5	.75	1
$\rho_{\mathcal{D}}^+$	1.0000	1.0000	1.0230	1.0169	1.0000	.9997	.9988
$\rho_{\mathcal{D}}^-$	.9978	.9978	.9754	.9468	.9091	.8639	.8128

$q \to$	$-.25$	$-.5$	$-.75$
$\rho_{\mathcal{D}}^+$	1.0174	1.0000	1.0000
$\rho_{\mathcal{D}}^-$	.9532	.9167	.8707

TABLE 6.2
Sharpness of (the modified) Bernstein's inequality (1.2) with the right-hand side multiplied by $\rho_{\mathcal{D}_s}^+$ on the square $-1/2 \le \alpha \le s$, $-1/2 \le \beta \le s$.

s	$\rho_{\mathcal{D}_s}^+$	$\hat{\rho}_{\mathcal{D}_s}^-$
1.0	1.038670463288	.960631920975
1.5	1.077936370739	.925639053930
2.0	1.119905216638	.890950401502
2.5	1.166112996124	.855646070084
3.0	1.217697600829	.819398840672
3.5	1.275581233437	.782215962616
4.0	1.340588974513	.744284804200
5.0	1.495211643984	.667316902208
6.0	1.688484850743	.590932161440
7.0	1.928648600010	.517346707121
8.0	2.225950341336	.448248994544
9.0	2.593289070919	.384754639811
10.0	3.046949165887	.327468542495

$n = n_1$, since $\max_{n_0 \le n \le n_1} \rho_n = \max(\rho_{n_0}, \rho_{n_1})$ and $\min_{n_0 \le n \le n_1} \rho_n = \min(\rho_{n_0}, \rho_{n_1})$. Consequently, $\rho_{\mathcal{D}}^+ = \max(\max_{\mathcal{D}} \rho_{n_0}, \max_{\mathcal{D}} \rho_{n_1})$ and $\rho_{\mathcal{D}}^- = \min(\min_{\mathcal{D}} \rho_{n_0}, \min_{\mathcal{D}} \rho_{n_1})$. In other words, if monotonicity in fact holds true, $\mathcal{D} = \{n_0 \le n \le n_1, \dots\}$ may be replaced by $\mathcal{D} = \{n = \{n_0, n_1\}, \dots\}$. In all our experiments we have verified that indeed the results for $\rho_{\mathcal{D}}^+$ and $\rho_{\mathcal{D}}^-$ are the same whether we restrict n to the smallest and largest value, or include intermediate values as well.

It can be seen from Table 6.1 that the choices $q = q^+$ and $q = q^-$ yield by far the best degrees of sharpness, both choices being essentially identical in quality. Naturally, if we lower $n_0 = 5$ to $n_0 = 1$, the sharpness deteriorates (to $\rho_{\mathcal{D}}^- = .9406$ for both choices of q), while increasing n_0 to, say, $n_0 = 10$ improves sharpness (to $\rho_{\mathcal{D}}^- = .9994$ for both choices of q).

7. Bernstein's inequality on larger domains. We now explore the sharpness resp. validity of (1.2) in the larger regions $\mathcal{R}_s = \{-1/2 \le \alpha \le s, -1/2 \le \beta \le s\}$, where, to begin with, $s = 1, 2, 5$, and 10. We define $\mathcal{D} = \mathcal{D}_s = \{n = [5\ 10\ 20\ 50\ 100], (\alpha, \beta) \in \mathcal{R}_s\}$, $s \ge 1/2$. We found that $\rho_{\mathcal{D}_s}^- = \rho_{\mathcal{D}_{1/2}}^-$ for all $s > 1/2$, and computations based on successively finer screenings near the minimum point $(\alpha^-, \beta^-) \in \mathcal{R}_{1/2}$ for $\rho_{\mathcal{D}_{1/2}}^-$ yielded

$$\begin{aligned} \rho_{\mathcal{D}_{1/2}}^- &= .997780002408 \quad (\text{where } q = q^+), \\ \rho_{\mathcal{D}_{1/2}}^- &= .997804307519 \quad (\text{where } q = q^-). \end{aligned} \tag{7.1}$$

For $\rho^+_{\mathcal{D}_s}$ we found, when $s = 1, 2, 5, 10$, regardless of whether $q = q^+$ or $q = q^-$, that the maximum $\rho^+_{\mathcal{D}_s} = \max_{\mathcal{D}_s} \rho_n$ is always attained at the upper right-hand corner $(\alpha, \beta) = (s, s)$ of the square $\mathcal{R}_s$. Assuming this to be true in general, we computed Table 6.2 for $\rho^+_{\mathcal{D}_s}$ and (cf. Section 1) $\hat{\rho}^-_{\mathcal{D}_s} = \rho^-_{\mathcal{D}_s}/\rho^+_{\mathcal{D}_s}$, where we used the first of the two values for $\rho^-_{\mathcal{D}_s} = \rho^-_{\mathcal{D}_{1/2}}$ in (7.1). (The other value, of course, gives very similar results.)

It can be seen that the sharpness of the inequality, even for $s = 10$, is still well within one order of magnitude. What is remarkable is also that the results are exacly the same if we let n go up to 200, so that the results are likely to be valid for all $n \geq 5$.

As a final experiment, we recomputed the second column of Table 6.2 with $\mathcal{D}_s = \{n = [5\ 24\ 43\ \ 62\ 81\ 100],\ \{(\alpha, \beta)\} \subset \mathcal{R}_s\}$, where $\{(\alpha, \beta)\}$ is a set of 1,000 randomly generated pairs (α, β) in $\mathcal{R}_s$. We verified that the results are all strictly smaller than those in Table 6.2, the smallest and largest deviations being 3.0770×10^{-5} (for $s = 3.5$) resp. 6.2961×10^{-3} (for $s = 9$).

We remark that the property of the maximum $\rho^+_{\mathcal{D}_s}$ being attained at $\alpha = \beta = s$ has been verified also if n =[1:10, 20, 25, 50, 75, 100] in the definition of $\mathcal{D}_s$ and also for $\max_n \sqrt{\gamma_n} c_n \times \|F_n\|_\infty$ in (1.5). The property, therefore, is likely to hold for any $n \geq 1$ and any $s \geq 1/2$; if so, it would allow to extend the upper bound for

$$\max_{-1 \leq x \leq 1} (1 - x)^{\alpha + 1/2} (1 + x)^{\beta + 1/2} [\hat{P}_n^{(\alpha, \beta)}(x)]^2,$$

proved for $\alpha = \beta \geq (1 + \sqrt{2})/4$ in [4, Equation (4)] to arbitrary $\alpha > -1/2$, $\beta > -1/2$, lending added support for the validity of the Erdélyi–Magnus–Nevai conjecture. Indeed, further calculations along the lines reported on in Table 6.2, but for $n \geq 1$, in particular the computation for $\alpha = \beta = s$ of the quantity

$$\max_n \gamma_n c_n^2 \|F_n\|_\infty^2 / \max\left(1, (2s^2)^{1/4}\right)$$

for s = [.5 : .01 : 1 2 : 10 20 50] and s = .706 : .0001 : .708 reveals that it attains a global maximum .66198126... at $s = 1/\sqrt{2}$. This suggests that the best constant implied in the Erdélyi–Magnus–Nevai conjecture (1.4) is .66198126....

Appendix. In the following Matlab script, the routines `r_jacobi` and `gauss` are part of a software package `OPQ`, which can be downloaded, along with the routine below, auxiliary routines, and a driver, from

`http://www.cs.purdue.edu/archives/2002/wxg/codes/BIJ.html`

ETNA
Kent State University
http://etna.math.kent.edu


```
% BERNSTEIN Sharpness of Bernstein's inequality for
%       Jacobi polynomials P_n(a,b;·) with b>=a>=-1/2.
%       The output is c_n || F_n ||.
%
function rho=bernstein(n,a,b,q)
if a<-1/2 | b<-1/2 | b<a
   disp('parameters a and/or b not in range')
   return
end tol=1e2*eps;
pnum=1; pden=1; p2=1;
for nu=1:n
   pnum=(1+(n+a+b)/nu)*pnum;
   pden=(1+q/nu)*pden;
   p2=(1-1/(2*nu))*p2;
end
c0=pnum/2^(n+(a+b+1)/2);
c1=(n+(a+b+1)/2)^(q+1/2)/(gamma(1+q)*pden);
c2=sqrt(pi)*c1*p2;
c=sqrt(pi)*c0*c1;
extr=zeros(n+1,1);
%
% When applying this routine for the same values
% of a and b, but many different values of n, the
% following command, for better efficiency, should
% be called outside the n-loop with n set equal to
% the largest n-value in the loop and the array ab
% included among the input parameters of this routine.
%
ab=r_jacobi(n,a,b);
xw=gauss(n,ab);
x=xw(:,1);
x1=[-1 x' 1]';
k0=1; k1=n+1;
if a==-1/2
   if b>-1/2
     k1=n;
   else
     rho=1;
     return
   end
end
for k=k0:k1
   t0=0; t1=(x1(k)+x1(k+1))/2;
   while abs(t1-t0)>tol
     t0=t1;
     t1=t0-fbern(t0,a,b,x)/f1bern(t0,a,b,x);
   end
   p=prod(t1-x);
   extr(k)=(1-t1)^(a/2+1/4)*(1+t1)^(b/2+1/4)*abs(p);
end
rho=c*max(extr);
if a==-1/2
   if c2>rho
     rho=c2;
   end
end
```

```
% FBERN A function f needed in Bernstein's inequality
%      for Jacobi polynomials
%
function y=fbern(t,a,b,x)
y=b-a-(a+b+1)*t+2*(1-t^2)*sum(1./(t-x));


% F1BERN The function f' needed in Bernstein's inequality
%      for Jacobi polynomials
%
function y=f1bern(t,a,b,x)
y=-(a+b+1)-2*sum((t^2-2*t*x+1)./(t-x).^2);
```

REFERENCES

[1] V. A. ANTONOV AND K. V. HOLŠHEVNIKOV, *Estimation of a remainder of a Legendre polynomial generating function expansion (generalization and refinement of the Bernšteĭn inequality*, Vestnik Leningrad. Univ. Mat. Mekh. Astronom., vyp. 3 (1980), pp. 5–7, 128 (in Russian).

[2] P. BARATELLA, *Bounds for the error term in Hilb formula for Jacobi polynomials*, Atti Accad. Sci. Torino Cl. Sci. Fis. Mat. Natur., 120 (1986), pp. 207–223.

[3] Y. CHOW, L. GATTESCHI, AND R. WONG, *A Bernstein-type inequality for the Jacobi polynomial*, Proc. Amer. Math. Soc., 121 (1994), pp. 703–709.

[4] I. KRASIKOV, *On the Erdélyi–Magnus–Nevai conjecture for Jacobi polynomials*, Constr. Approx., 28 (2008), pp. 113–125.

[5] L. LORCH, *Alternative proof of a sharpened form of Bernstein's inequality for Legendre polynomials*, Applicable Anal., 14 (1982/83), pp. 237–240. [Corrigendum, *ibid.* 50 (1993), p. 47.]

[6] L. LORCH, *Inequalities for ultraspherical polynomials and the gamma function*, J. Approx. Theory, 40 (1984), pp. 115–120.

[7] P. NEVAI, T. ERDÉLYI, AND A. P. MAGNUS, *Generalized Jacobi weights, Christoffel functions, and Jacobi polynomials*, SIAM J. Math. Anal., 25 (1994), pp. 602–614. [Erratum, *ibid.* 25 (1994), p. 1461.]

[8] G. SZEGŐ, *Orthogonal Polynomials*, 4th ed., American Mathematical Society, Providence, RI, 1975.

9.20. [199] "The Lambert W-functions and some of their integrals: a case study of high-precision computation"

---

[199] "The Lambert W-functions and some of their integrals: a case study of high-precision computation," *Numer. Algorithms* **57**, 27–34 (2011).

Numer Algor (2011) 57:27–34
DOI 10.1007/s11075-010-9409-6

ORIGINAL PAPER

The Lambert W-functions and some of their integrals: a case study of high-precision computation

Walter Gautschi

Received: 18 February 2010 / Accepted: 27 June 2010 /
Published online: 16 July 2010

Abstract The real-valued Lambert W-functions considered here are $w_0(y)$ and $w_{-1}(y)$, solutions of $we^w = y$, $-1/e < y < 0$, with values respectively in $(-1, 0)$ and $(-\infty, -1)$. A study is made of the numerical evaluation to high precision of these functions and of the integrals $\int_1^\infty [-w_0(-xe^{-x})]^\alpha x^{-\beta} \mathrm{d}x$, $\alpha > 0$, $\beta \in \mathbb{R}$, and $\int_0^1 [-w_{-1}(-xe^{-x})]^\alpha x^{-\beta} \mathrm{d}x$, $\alpha > -1$, $\beta < 1$. For the latter we use known integral representations and their evaluation by nonstandard Gaussian quadrature, if $\alpha \neq \beta$, and explicit formulae involving the trigamma function, if $\alpha = \beta$.

Keywords Lambert W-functions · Integrals of Lambert W-functions · Nonstandard Gaussian quadrature · Variable-precision computation

Mathematics Subject Classifications (2010) 33B99 · 33F05 · 65D20 · 65D30

1 Introduction

The (real-valued) Lambert W-functions are solutions of the nonlinear equation

$$we^w = y, \quad y \in \mathbb{R}. \tag{1.1}$$

If $y > 0$, there is a unique real solution, $w(y)$, satisfying $0 < w(y) < \infty$. If $-1/e \leq y < 0$, there are exactly two real solutions, $w_0(y)$ and $w_{-1}(y)$, satisfying respectively $-1 \leq w_0(y) < 0$ and $-\infty < w_{-1}(y) \leq -1$. Clearly, $w(0+) = 0$, $w_0(0-) = 0$, and $w_{-1}(0-) = -\infty$, while $w_0(-1/e) = w_{-1}(-1/e) = -1$. For

W. Gautschi ( )
Department of Computer Sciences, Purdue University, West Lafayette, IN 47907-1398, USA
e-mail: wxg@cs.purdue.edu

$y < -1/e$, there are no real solutions of (1.1). For a discussion of the various branches of the Lambert W-functions, also in the complex plane, see [3].

We note that

$$-w_0(-xe^{-x}) \equiv x \quad \text{for } 0 \le x \le 1, \tag{1.2}$$

and likewise

$$-w_{-1}(-xe^{-x}) \equiv x \quad \text{for } 1 \le x < \infty. \tag{1.3}$$

Indeed, by definition,

$$-xe^{-x} \equiv w_0(-xe^{-x})e^{w_0(-xe^{-x})}, \quad 0 \le x \le 1.$$

This identity remains valid if $w_0(-xe^{-x})$ at both occurrences is replaced by $-x$, which is in $[-1, 0]$. By uniqueness of w_0, there follows (1.2). The identity (1.3) is proved similarly.

Our interest here is in the computation (to high precision) of the three Lambert functions and of the integrals

$$I_{0,[1,\infty]}(\alpha, \beta) = \int_1^\infty f_0(x; \alpha, \beta)\mathrm{d}x \quad \text{and} \quad I_{1,[0,1]}(\alpha, \beta) = \int_0^1 f_1(x; \alpha, \beta)\mathrm{d}x, \tag{1.4}$$

where

$$\begin{aligned} f_0(x; \alpha, \beta) &= [-w_0(-xe^{-x})]^\alpha x^{-\beta}, \ \alpha > 0; \\ f_1(x; \alpha, \beta) &= [-w_{-1}(-xe^{-x})]^\alpha x^{-\beta}, \ \alpha > -1, \beta < 1. \end{aligned} \tag{1.5}$$

Both integrals present numerical difficulties because of singularities of the integrands at the upper resp. lower end point of integration.

For $\alpha = \beta$, the integrals are explicitly known [8],

$$\begin{aligned} I_{0,[1,\infty]}(\alpha, \alpha) &= \alpha\psi_1(\alpha) - 1, \ \alpha > 0; \\ I_{1,[0,1]}(\alpha, \alpha) &= \alpha\psi_1(1-\alpha) + 1, \ |\alpha| < 1, \end{aligned} \tag{1.6}$$

where ψ_1 is the trigamma function. Their sum equals

$$\alpha(\psi_1(\alpha) + \psi_1(1-\alpha)) = \alpha\left[\frac{\pi}{\sin(\alpha\pi)}\right]^2, \tag{1.7}$$

by the reflection formula for ψ_1 (cf. [1, Eq. 6.4.7]). The Matlab routines[1] `sI01infaa.m`[2] and `sI101aa.m`[2] implement (1.6) in variable-precision arithmetic.

[1] All Matlab routines referred to in this paper can be downloaded from the web site http://www.cs.purdue.edu/archives/2002/wxg/codes/LAMBERTW.html. They make use of additional routines in the packages `OPQ`, `SOPQ` found on the same web site.

[2] This routine requires Matlab Version 7.8 (R2009a) or later.

If $\alpha \neq \beta$, then [8]

$$I_{0,[1,\infty]}(\alpha,\beta) = \frac{1}{\alpha-\beta+1}\left[-1+\alpha\int_0^1 u^{\alpha-1}\left(\frac{\ln(1/u)}{1-u}\right)^{\alpha-\beta+1}\mathrm{d}u\right], \quad \alpha > 0, \tag{1.8}$$

which for $\alpha = \beta$ reduces to (1.6) in view of [7, Eq. 4.251.4]. Similarly [8],

$$I_{1,[0,1]}(\alpha,\beta) = \frac{1}{\alpha-\beta+1}\left[1+\alpha\int_1^\infty u^{\alpha-1}\left(\frac{\ln u}{u-1}\right)^{\alpha-\beta+1}\mathrm{d}u\right], \ \alpha > -1, \ \beta < 1. \tag{1.9}$$

Both formulae lend themselves to numerical evaluation by appropriate (non-standard) Gaussian quadrature; see Sections 3, 4 for details.

Although the Lambert functions will not be used explicitly in what follows, it may be of interest to briefly consider computational methods for their evaluation. This is done in Section 2.

2 Computing the Lambert W-functions

There are of course many possible ways of solving the equation (1.1).[3] A simple and generally reliable method is Newton's method which, by choosing the initial approximations judiciously, allows us to compute all three Lambert functions defined in Section 1 (only the last two of them being of interest here). For y near and above $-1/e$, Newton's method, however, suffers from loss of accuracy and consequent slow, or even lack of, convergence. In this case a power series expansion method is proposed. For y near and below zero, Newton's method for w_{-1}, while numerically stable, may take many iterations (some 30 in Matlab double precision, when $y = -10^{-10}$) to converge.

2.1 Newton's method

The Newton iteration for (1.1) is

$$w^{[\nu+1]} = \frac{(w^{[\nu]})^2 + ye^{-w^{[\nu]}}}{1+w^{[\nu]}}, \quad \nu = 0, 1, 2, \ldots, \tag{2.1}$$

where as initial value $w^{[0]}$ we take

$$w^{[0]} = \begin{cases} y & \text{if } 0 < y < e, \\ \ln(y/\ln y) & \text{if } y \geq e \end{cases} \tag{2.2}$$

for $w(y)$, and

$$w^{[0]} = \begin{cases} 0 & \text{for } w_0(y), \\ -2 & \text{for } w_{-1}(y). \end{cases} \tag{2.3}$$

[3] For a recent discussion of numerical methods, see [2].

The choice in (2.2), when $y \geq e$, is motivated by the asymptotic behavior of $w(y)$ as $y \to \infty$, and the choice in (2.3) by our desire to have monotone convergence. The latter, in theory, is guaranteed (cf. [5, Example 6.4, p. 233]) by the convexity/concavity properties of the function we^w and the fact that $w = -2$ is an inflection point of the curve $y = we^w$.

Near the point $(w, y) = (-1, -1/e)$, where the graph of $y = we^w$ has a horizontal tangent, Newton's method converges only linearly and may even fail to converge because of cancellation errors in the denominator of (2.1). Indeed, $w^{[\nu]}$ comes arbitrarily close to -1 when y approaches $-1/e$ from above. To avoid this difficulty, we use an appropriate power series expansion (cf. Section 2.2).

2.2 Power series solution

Let

$$y = -\frac{1}{e} + x^2, \quad x > 0. \tag{2.4}$$

Then the solution w of (1.1) admits an expansion in powers of x,

$$w = -1 + c_1 x + c_2 x^2 + c_3 x^3 + \cdots. \tag{2.5}$$

Matching the power series of we^w against the (finite) series (2.4), we find

$$c_1 = \pm\sqrt{2e}, \tag{2.6}$$

and, with the help of Maple, that

$$\begin{aligned}
c_2 &= -\frac{1}{3}\,c_1^2,\\
c_3 &= -\left(c_2c_1^2 + \frac{1}{8}\,c_1^4 + \frac{1}{2}\,c_2^2\right)/c_1,\\
c_4 &= -\left(c_3c_2 + c_3c_1^2 + c_1c_2^2 + \frac{1}{2}\,c_2c_1^3 + \frac{1}{30}\,c_1^5\right)/c_1,\\
c_5 &= -\left(c_4c_2 + c_4c_1^2 + \frac{1}{2}\,c_3c_1^3 + \frac{3}{4}\,c_2^2c_1^2 + \frac{1}{6}\,c_2c_1^4\right.\\
&\qquad \left. + \frac{1}{2}\,c_3^2 + \frac{1}{3}\,c_2^3 + \frac{1}{144}\,c_1^6 + 2c_1c_2c_3\right)/c_1,\\
c_6 &= -\left(2c_4c_1c_2 + \frac{3}{2}\,c_3c_2c_1^2 + \frac{1}{840}\,c_1^7 + c_5c_2 + c_4c_3\right.\\
&\qquad + c_5c_1^2 + \frac{1}{2}\,c_4c_1^3 + c_1c_3^2 + c_3c_2^2 + \frac{1}{6}\,c_3c_1^4\\
&\qquad \left. + \frac{1}{2}\,c_1c_2^3 + \frac{1}{3}\,c_2^2c_1^3 + \frac{1}{24}\,c_2c_1^5\right)/c_1,\\
&\cdots \qquad\qquad \cdots
\end{aligned} \tag{2.7}$$

Clearly, for $w = w_{-1}$ we must select the minus sign in (2.6), and for $w = w_0$ the plus sign. Substituting (2.6) in (2.7), we then get, again with the help of Maple,

$$w_{-1}(x) = -1 - \sqrt{2e}\,x - \frac{2}{3}\,e\,x^2 - \frac{11}{36}\sqrt{2e^3}\,x^3 - \frac{43}{135}\,e^2\,x^4 - \frac{769}{4320}\sqrt{2e^5}\,x^5 - \frac{1768}{8505}\,e^3\,x^6 + \cdots \tag{2.8}$$

and

$$w_0(x) = -1 + \sqrt{2e}\,x - \frac{2}{3}\,e\,x^2 + \frac{11}{36}\sqrt{2e^3}\,x^3 - \frac{43}{135}\,e^2\,x^4 + \frac{769}{4320}\sqrt{2e^5}\,x^5 - \frac{1768}{8505}\,e^3\,x^6 + \cdots . \tag{2.9}$$

A series expansion closely related to (2.8) is the power series expansion of $-w_{-1}(-\exp(-1 + z^2/2))$ in [4, Eq. (48)]; other related series can be found in [4, Section 3.2]. Both expansions (2.8) and (2.9) are appropriate for, say, $x^2 \le .5 \times 10^{-4}$, while Newton's method is adequate in all other cases. In symbolic routines, using variable-precision artithmetic, only Newton's method needs to be used, together with appropriate precautions near the branch point $(w, y) = (-1, -1/e)$ (cf. Section 2.3).

2.3 Matlab implementation

The procedures described in Sections 2.1 and 2.2 are implemented in the Matlab routines `wofy.m`, `wofy0.m`, `wofy1.m`. The respective symbolic analogues are `swofy.m`, `swofy0.m`, `swofy1.m`. These use only Newton's method; to counteract the loss of accuracy in `swofy0.m` and `swofy1.m` when y is near and above $-1/e$, the working precision in these two routines must be selected sufficiently larger than the target precision. For `wofy0.m` one needs 4, 6, 10 more digits than in the target precision when the distance of y from $-1/e$ is respectively 10^{-5}, 10^{-10}, and 10^{-20}; for `swofy1.m` the numbers are 4, 7, and 12 digits, respectively.

3 The integrals $I_{0,[1,\infty]}(\alpha, \beta)$ and $I_{0,[0,1]}(\alpha, \beta)$

For the evaluation of $I_{0,[1,\infty]}(\alpha, \beta)$, $\alpha \neq \beta$, we use the integral representation (1.8). In view of the logarithmic/algebraic singularity at the lower limit of the integral in (1.8), and its regularity at the upper limit, we decompose the integral into two parts: the first extended from 0 to $1/e$, the second from $1/e$ to 1. The former is written as

$$I_{[0,1/e]} = \int_0^{1/e} (1-x)^{-(\alpha-\beta+1)} \cdot x^{\alpha-1}[\ln(1/x)]^{\alpha-\beta+1}\mathrm{d}x, \tag{3.1}$$

the second factor being treated as a weight function,

$$v(x; \alpha, \beta) = x^{\alpha-1}[\ln(1/x)]^{\alpha-\beta+1}, \quad 0 < x \le 1/e, \tag{3.2}$$

with the intention of applying Gauss quadrature relative to this weight function. The second part,

$$I_{[1/e,1]} = \int_{1/e}^{1} u^{\alpha-1} \left(\frac{\ln(1/u)}{1-u} \right)^{\alpha-\beta+1} \mathrm{d}u,$$

after the change of variable $u = 1 - x(1 - 1/e)$, becomes

$$I_{[1/e,1]} = (1-1/e) \int_0^1 [1-(1-1/e)x]^{\alpha-1} \left(\frac{-\ln(1-(1-1/e)x)}{(1-1/e)x} \right)^{\alpha-\beta+1} \mathrm{d}x \tag{3.3}$$

and is amenable to Gauss–Legendre quadrature on [0, 1].

With regard to Gauss quadrature for the (nonstandard) weight function (3.2), we generate the relevant orthogonal polynomials (see [6] for details) by the variable-precision Chebyshev algorithm from the moments

$$\begin{aligned} \mu_k(v;\alpha,\beta) &= \int_0^{1/e} x^{k+\alpha-1} [\ln(1/x)]^{\alpha-\beta+1} \mathrm{d}x = \int_1^{\infty} t^{\alpha-\beta+1} e^{-(k+\alpha)t} \mathrm{d}t \\ &= \frac{1}{(k+\alpha)^{\alpha-\beta+2}} \Gamma(\alpha-\beta+2, k+\alpha), \quad k = 0, 1, 2, \ldots . \end{aligned}$$

These are generated (in variable-precision arithmetic) by the Matlab routine `smomvab.m`[2]. The Matlab routine `sr_vab.m` then generates the first N recurrence coefficients of the required orthogonal polynomials and stores them in the $N \times 2$ array `abv`. These in turn allow us to generate the desired N-point Gaussian quadrature rule using the `SOPQ` routine `sgauss.m`. The evaluation of $I_{0,[1,\infty]}(\alpha, \beta)$ from (1.8), using (3.1) and (3.3), is implemented in the Matlab routine `sI01infab.m`.

We ran this procedure in 32-digit arithmetic for $\alpha = 2, 1, \frac{1}{2}$ ($\alpha = 0$ is trivial), and for each of these values for $\beta = 2, 1, \frac{1}{2}, 0, -\frac{1}{2}, -1, -2$ ($\alpha = 1$, $\beta = 2$ being trivial). The most time-consuming part of these calculations is the generation of the N recurrence coefficients by the routine `sr_vab`, which, when $N = 50$, took about 12 minutes or less on the Sun Ultra 5 workstation, assuming a good estimate of `dig0`—the initial number of digits used in the routine `sr_vab`. Tables of these coefficients can be found on the web site http://www.cs.purdue.edu/archives/2001/wxg/tables in files whose names start with "`abv`".

A sample of results is shown in Table 1, where n denotes the number of quadrature points needed for 32-digit accuracy and `dig` the number of digits required in the routine `sr_vab` to obtain the recurrence coefficients accurate to 32 digits. The number n is seen to be less than 30, which is remarkably small for this type of accuracy. When $\alpha = \beta$, there is agreement with the 32-digit results obtained by the routine `sI01infaa.m` except for an endfigure discrepancy of one unit in the case $\alpha = \beta = 2$.

The integral $I_{0,[0,1]}(\alpha, \beta)$, by (1.2), is equal to $1/(\alpha - \beta + 1)$ if $\alpha - \beta + 1 > 0$.

Table 1 The integral $I_{0,[1,\infty]}(\alpha, \beta)$ for selected values of α, β

α	β	n	dig	$I_{0,[1,\infty]}(\alpha, \beta)$
2	2	27	115	0.28986813369645287294483033329204
	0	26	115	0.55242099404393096881202067693593
	−2	25	110	1.58573825833907392618631364042 74
1	1	29	115	0.64493406684822643647241516664603
	0	26	115	1.14493406684822643647241516664 6
	−1	28	110	2.51365763667448738853881999482 42
$\frac{1}{2}$	2	27	110	0.56150165251555182684424279164016
	0	28	110	2.33383591550890144676101870325 41
	−2	27	110	38.72763356997972400830840316563 4

4 The integrals $I_{1,[0,1]}(\alpha, \beta)$ and $I_{1,[1,\infty]}(\alpha, \beta)$

In order to evaluate $I_{1,[0,1]}(\alpha, \beta)$ from the integral representation (1.9), we first make the change of variable $u = 1/x$ in the integral of (1.9) to write it as

$$\int_0^1 x^{-\beta} \left(\frac{\ln(1/x)}{1-x} \right)^{\alpha-\beta+1} \mathrm{d}x.$$

This has the same form as the integral in (1.8). Hence, we deal with it in the same way as was done in Section 3, i.e., decompose it into two integrals analogous to (3.1) and (3.3). The first is calculated by Gauss quadrature with respect to the weight function

$$u(x; \alpha, \beta) = x^{-\beta}[\ln(1/x)]^{\alpha-\beta+1}, \quad 0 < x \le 1/e, \tag{4.1}$$

the second by Gauss–Legendre quadrature of

$$(1-1/e) \int_0^1 [1-(1-1/e)x]^{-\beta} \left(\frac{-\ln(1-(1-1/e)x)}{(1-1/e)x} \right)^{\alpha-\beta+1} \mathrm{d}x.$$

The moments of the weight function (4.1) are given by

$$\mu_k(u; \alpha, \beta) = \frac{1}{(k-\beta+1)^{\alpha-\beta+2}} \Gamma(\alpha-\beta+2, k-\beta+1), \quad k = 0, 1, 2, \ldots,$$

and are evaluated by the routine `smomuab.m`[2]. The Matlab routine `sr_uab.m` then generates the recurrence coefficients for the required orthogonal polynomials and stores them in the array `abu`. The integral $I_{1,[0,1]}(\alpha, \beta)$ itself is computed by the routine `sI101ab.m`.

The procedure was run in 32-digit arithmetic for $\alpha = 2, 1, \frac{1}{2}, -\frac{1}{2}$ and for each of these values for $\beta = \frac{1}{2}, 0, -\frac{1}{2}, -1, -2$ (the case $\alpha = -\frac{1}{2}$, $\beta = \frac{1}{2}$ being trivial). The first 50 recurrence coefficients required in these computations, generated by the routine `sr_uab.m`, are retrievable on the web site mentioned in Section 3 from files whose names start with "`abu`".

Selected results are shown in Table 2 in the same format as used in Table 1.

Table 2 The integral $I_{1,[0,1]}(\alpha, \beta)$ for selected values of α, β

α	β	n	dig	$I_{1,[0,1]}(\alpha, \beta)$
2	$\frac{1}{2}$	27	110	33.343927985540712124255844272507
	0	27	110	6.027315273348974777077639989648 3
	-2	25	110	0.664705074209279053639332239992 06
$\frac{1}{2}$	$\frac{1}{2}$	26	110	3.467401100272339654708622749969
	0	26	115	1.420019688788567361130468925952 8
	-1	25	110	0.614174874181793783900301161758 49
$-\frac{1}{2}$	0	26	115	0.742915501211264886889271571242 86
	-1	25	115	0.417045121211600390849790576851 56
	-2	24	115	0.292165139638023691849423622411 38

When $\alpha = \beta$, the results are in complete agreement with those furnished by the routine `sI101aa.m` with `dig` $= 32$.

The integral $I_{1,[1,\infty]}(\alpha, \beta)$, by virtue of (1.3), exists only if $\beta - \alpha - 1 > 0$, and then equals $1/(\beta - \alpha - 1)$.

Acknowledgements The computational problem in the case $\alpha = \beta$ was brought to the author's attention by Tony Tam. The author is indebted to Robert M. Corless for the References [2–4], the first of which, but not the others, having been familiar to the author at the time of writing this paper.

References

1. Abramowitz, M., Stegun, I.A.: Handbook of mathematical functions. National Bureau of Standards, Applied Mathematics Series 55, U.S. Government Printing Office, Washington, DC (1964)
2. Barry, D.A., Li, L., Jeng, D.-S.: Comments on numerical evaluation of the Lambert W-functions and application to generation of generalized Gaussian noise with exponent 1/2. IEEE Trans. Signal Process. **52**, 1456–1458 (2004)
3. Corless, R.M., Gonnet, G.H., Hare, D.E.G., Jeffrey, D.J., Knuth, D.E.: On the Lambert W-functions. Adv. Comput. Math. **5**, 329–359 (1996)
4. Corless, R.M., Jeffrey, D.J., Knuth, D.E.: A sequence of series for the Lambert W-functions. In: Proceedings of the 1997 International Symposium on Symbolic and Algebraic Computation (Kihei, HI), pp. 197–204. ACM, New York (1997, electronic)
5. Gautschi, W.: Numerical Analysis: An Introduction. Birkhäuser, Boston (1997)
6. Gautschi, W.: Variable-precision recurrence coefficients for non-stand-ard orthogonal polynomials. Numer. Algorithms **52**, 409–418 (2009)
7. Gradshteyn, I.S., Ryzhik, I.M.: Table of integrals, series, and products, 7th edn. Elsevier/Academic Press, Amsterdam (2007)
8. Yu, Y.: Personal communication, October (2009)

9.21. [203] "Remark on 'New conjectured inequalities for zeros of Jacobi polynomials' by Walter Gautschi, Numer. Algorithms 50: 293–296 (2009)"

---

[203] "Remark on 'New conjectured inequalities for zeros of Jacobi polynomials' by Walter Gautschi, Numer. Algorithms 50: 293–296 (2009)," *Numer. Algorithms* **57**, 511 (2011).

Numer Algor (2011) 57:511
DOI 10.1007/s11075-010-9442-5

Remark on "New conjectured inequalities for zeros of Jacobi polynomials" by Walter Gautschi, Numer. Algorithms 50:293–296 (2009)

Walter Gautschi

Received: 16 December 2010 / Accepted: 16 December 2010 /
Published online: 27 January 2011

Keywords Jacobi polynomials · Zeros · Inequalities

Mathematics Subject Classification (2010) 33C45

Martin Muldoon kindly brought to the author's attention that Conjecture 1 of the paper cited in the title has been proved by Sturm comparison methods in Theorem 3.1(ii) of the paper "On the spacing of the zeros of some classical orthogonal polynomials" by S. Ahmed, A. Laforgia, and M. E. Muldoon, J. Lond. Math. Soc. 25:246–252 (1982).

W. Gautschi (✉)
Department of Computer Sciences, Purdue University, West Lafayette, IN 47907-2066, USA
e-mail: wxg@cs.purdue.edu

10

Papers on Interpolation and Approximation

---

41 Attenuation factors in practical Fourier analysis, Numer. Math. 18, 373–400 (1972)

86 On Padé approximants associated with Hamburger series, Calcolo 20, 111–127 (1983)

87 On the convergence behavior of continued fractions with real elements, Math. Comp. 40, 337–342 (1983)

89 Discrete approximations to spherically symmetric distributions, Numer. Math. 44, 53–60 (1984)

100 (with G. V. Milovanović) Spline approximations to spherically symmetric distributions, Numer. Math. 49, 111–121 (1986)

102 (with M. Frontini and G. V. Milovanović) Moment-preserving spline approximation on finite intervals, Numer. Math. 50, 503–518 (1987)

132 On mean convergence of extended Lagrange interpolation, J. Comput. Appl. Math. 43, 19–35 (1992)

147 (with S. Li) On quadrature convergence of extended Lagrange interpolation, Math. Comp. 65, 1249–1256 (1996)

165 Remark: "Barycentric formulae for cardinal (SINC-) interpolants by Jean-Paul Berrut," Numer. Math. 87, 791–792 (2001)

202 Experimental mathematics involving orthogonal polynomials, in *Approximation and computation — in honor of Gradimir V. Milovanović* (W. Gautschi, G. Mastroianni, and Th. M. Rassias, eds.), 117–134, Springer Optim. Appl. 42 (2011)

10.1. [41] “Attenuation Factors in Practical Fourier Analysis”

---

[41] “Attenuation Factors in Practical Fourier Analysis,” *Numer. Math.* **18**, 373–400 (1972).

Numer. Math. 18, 373–400 (1972)

Attenuation Factors in Practical Fourier Analysis*

WALTER GAUTSCHI

Department of Computer Sciences, Purdue University, Lafayette, Indiana 47907

Received May 26, 1971

Summary. Given a 2π-periodic function f, it is desired to approximate its n-th Fourier coefficient $c_n(f)$ in terms of function values f_μ at N equidistant abscisses

$$x_\mu = \mu 2\pi/N, \quad \mu = 0, 1, \ldots, N-1.$$

A time-honored procedure consists in interpolating f at these points by some 2π-periodic function φ and approximating $c_n(f)$ by $c_n(\varphi)$. In a number of cases, where φ is piecewise polynomial, it has been known that $c_n(\varphi) = \tau_n \hat{c}_n(f)$, where $\hat{c}_n(f)$ is the trapezoidal rule approximation of $c_n(f)$ and τ_n is independent of f. Our interest is in the factors τ_n, called attenuation factors. We first clarify the conditions on the approximation process P: $f \to \varphi$ under which such attenuation factors arise. It turns out that a necessary and sufficient condition is linearity and translation invariance of P. The latter means that shifting the periodic data $f = \{f_\mu\}$ one place to the right has the effect of shifting $\varphi = Pf$ by the same amount. An explicit formula for τ_n is obtained for any process P which is linear and translation invariant. For interpolation processes it suffices to obtain a factorization $c_n(\varphi) = \omega(n)\psi_f(n)$, where ω does not depend on f and $\psi_f(n)$ has period N. This also implies existence of attenuation factors τ_n, which are expressible in terms of ω. The results can be extended in two directions: First, the process P may also approximate successive derivative values $f_\mu^{(\varkappa)}$, $\varkappa = 0, 1, \ldots, k-1$, of the function f, in which case formulas of the type $c_n(\varphi) = \sum_{\varkappa=0}^{k-1} \tau_{n,\varkappa} \hat{c}_n(f^{(\varkappa)})$ emerge. Secondly, P may be translation invariant over r subintervals, $r > 1$, in which case $c_n(\varphi) = \sum_{\varrho=0}^{r-1} \tau_{n,\varrho} \hat{c}_{n+\varrho N/r}(f)$. All results are illustrated by a number of examples, in which φ are polynomial and nonpolynomial spline interpolants, including deficient splines, as well as other piecewise polynomial interpolants. These include approximants of low and medium continuity classes permitting arbitrarily high degree of approximation.

1. Introduction

The problem of practical Fourier analysis often presents itself in the following form. It is desired to calculate the Fourier coefficients

$$c_n = c_n(f) \overset{\text{def}}{=} \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-inx} dx, \quad n = 0, \pm 1, \pm 2, \ldots \tag{1.1}$$

of a 2π-periodic real-valued function f which is known, or can be calculated, on a set of discrete points

$$x_\mu = \mu \frac{2\pi}{N}, \quad \mu = 0, 1, \ldots, N-1. \tag{1.2}$$

* This work was carried out while the author was Visiting Professor at the Mathematical Institute of the Technical University of Munich, Germany. The work was supported in part by a Fulbright research grant.

Derivative values, in addition to function values, may also be available at these points.

If no information is known about f, other than the values

$$f_\mu = f(x_\mu), \quad \mu = 0, 1, \ldots, N-1, \tag{1.3}$$

a reasonable approximation to $c_n(f)$ is given by

$$\hat{c}_n = \hat{c}_n(f) \overset{\text{def}}{=} \frac{1}{N} \sum_{\mu=0}^{N-1} f_\mu e^{-inx_\mu}. \tag{1.4}$$

This may be justified by noting that for each integer $s \geqq 0$, with $2s+1 \leqq N$, the trigonometric polynomial of order s,

$$\hat{t}_s(x) = \sum_{n=-s}^{s} \hat{c}_n(f) e^{inx}, \tag{1.5}$$

formed with the coefficients (1.4), approximates f best among all polynomials of the same type in the sense of the discrete L_2-norm $\|g\|_2 = \left(\sum_{\mu=0}^{N-1} |g(x_\mu)|^2 \right)^{\frac{1}{2}}$. If $2s+1 = N$, then (1.5) in fact represents the unique interpolation polynomial belonging to the data (1.3). A more practical consideration in support of $\hat{c}_n$ is the fact that sums such as those in (1.4) can be calculated on binary digital computers very efficiently by the algorithm of Cooley and Tukey [4], now commonly known as the "fast Fourier transform" [5, 9]. It may also be noted that the approximations $\hat{c}_n$ share with the exact Fourier coefficients c_n the symmetry property

$$\hat{c}_{-n} = \overline{\hat{c}_n}, \quad \text{all } n. \tag{1.6}$$

In the presence of additional information about the function f, however, the choice (1.4) may fall short in reflecting essential properties of Fourier coefficients. If it is known, e.g., that f has an $(r-1)$-st derivative which is absolutely continuous on the real line, then $c_n = o(n^{-r})$ as $n \to \infty$. The approximations $\hat{c}_n$, on the other hand, have period N,

$$\hat{c}_{n+N} = \hat{c}_n, \quad \text{all } n, \tag{1.7}$$

and are thus unable to simulate the asymptotic behavior of c_n.

For this reason one often attempts to approximate f by some φ which shares with f some of its smoothness properties, and then takes $c_n(\varphi)$ to approximate $c_n(f)$. In many cases where φ interpolates f at the points x_μ, and is made up of polynomial pieces over the subintervals $(x_\mu, x_{\mu+1})$, it has been found that

$$c_n(\varphi) = \tau_n \hat{c}_n(f), \quad \text{all } n, \tag{1.8}$$

where τ_n are certain universal factors not depending on f. These are referred to as *attenuation factors*. By choosing φ judiciously one can arrange these factors to go to zero at a rate comparable to that of the $c_n(f)$. The fact that a relation of the type (1.8) still permits use of the fast Fourier transform is another attractive feature of (1.8).

An instance of (1.8), as pointed out by Yuškov [19], was already discovered in 1898 by Oumoff [12]. He considers the broken line interpolant φ and states (without proof) that (1.8) is valid with

$$\tau_n = \left[\frac{\sin(\pi n/N)}{\pi n/N}\right]^2. \tag{1.9}$$

The same result was rediscovered by Dällenbach [6] in 1921, based on calculations of H. Weyl, and again much later by Chao [3]. Dällenbach also considers a cubic interpolation scheme and determines the associated attenuation factors. Further instances of (1.8) are discussed by Eagle [7].

In the terminology of spline functions the broken line interpolant is a polynomial spline interpolation function of degree 1. It is interesting to note that Runge [14, p. 193ff.] already in 1904 constructed periodic spline interpolants of degree 2 (without, of course, naming them as such) in connection with Fourier analysis, although he did not obtain the respective attenuation factors. Spline interpolants of higher degrees were used systematically by Eagle [7] in 1928, and ten years later, independently, by Quade and Collatz [13]. Eagle gives a remarkably elegant derivation of the attenuation factors in the case of splines (which he calls "lath functions"). A more lengthy derivation is given by Quade and Collatz, whose major concern, however, are the approximation-theoretic aspects of the problem. These are also discussed more recently by Ehlich [8] and Golomb [10]. Further short derivations of attenuation factors can be found in Bauer and Stetter [2], who also consider the generalization to Fourier transforms.

In choosing φ it does not suffice, of course, to simulate the smoothness properties of f. One also wants local accuracy, i.e., φ should approximate f as closely as possible on each subinterval $(x_\mu, x_{\mu+1})$. In this respect, the splines have the (perhaps undesirable) feature of correlating accuracy and smoothness: increasing the degree of the spline also increases its degree of smoothness. It seems worthwhile to make available a repertoire of other approximants φ which are capable of fitting f to high accuracy and yet have low, or only moderate, degrees of smoothness. We will partially meet this need in Section 5 where examples are provided of piecewise polynomial approximants φ whose polynomial degrees are $2r-1$, and whose degrees of smoothness are 1 (Example 5.2) and r (Example 5.3). We shall also give an example of a nonpolynomial approximant φ, viz., a generalized spline function belonging to a linear differential operator with constant coefficients (Example 5.4). In all cases, the respective attenuation factors will be specified explicitly.

In view of the multitude of possible attenuation factors the question naturally arises as to the precise conditions on the approximation process $P: f \to \varphi$ in order to yield a formula of the type (1.8). We shall give a complete answer to this question in Section 3, Theorems 3.1 and 3.2. We show, in essence, that for (1.8) to hold, it is necessary and sufficient that P be linear and translation invariant over an interval of length $h = 2\pi/N$. The latter means that by shifting the periodic data $f = \{f_\mu\}$ one place to the right, the (periodic) function $\varphi = Pf$ is shifted by the same amount. An explicit formula for the attenuation factor τ_n is also obtained for any linear process P which is translation invariant in this sense.

These results extend readily to Hermite-type approximation processes (involving any number of successive derivative values), where formulas of the type

$$c_n(\varphi) = \sum_s \tau_{n,s}\hat{c}_n(f^{(s)}) \tag{1.10}$$

emerge. This is again illustrated in Section 5 (Example 5.6), where φ is taken to be a periodic spline interpolant of degree $2r-1$ and deficiency k (in a terminology of Ahlberg *et al.* [1]). As special cases this includes ordinary splines ($k=1$) and Hermite interpolation polynomials ($k=r$).

For interpolation processes it will also be shown (Theorem 3.3) that for (1.8) to hold, it suffices to find a factorization

$$c_n(\varphi) = \omega(n)\psi_f(n), \tag{1.11}$$

where $\psi_f(n)$ is an *arbitrary* N-periodic function, and $\omega(n)$ is independent of f. The attenuation factor τ_n can then be expressed in terms of ω. This result is often more convenient for deriving attenuation factors than the explicit formula provided in Theorem 3.1.

Besides (1.8), formulas of a somewhat different character exist in the literature. For example, if N is even, and f is interpolated by quadratic polynomials over panels of two consecutive subintervals, then Yuškov [19] has shown that

$$c_n(\varphi) = \tau_{n,0}\hat{c}_n(f) + \tau_{n,1}\hat{c}_{n+N/2}(f), \quad \text{all } n. \tag{1.12}$$

There are now two attenuation factors associated with this interpolation process. Using cubic interpolation over three consecutive intervals gives rise to three attenuation factors (Yuškov [20]). We prove in section 4 the considerably more general result that a formula of the type

$$c_n(\varphi) = \sum_{\varrho=0}^{r-1} \tau_{n,\varrho}\hat{c}_{n+\varrho N/r}(f) \tag{1.13}$$

is valid precisely when the process $P\colon f\to\varphi$ is linear and translation invariant over r subintervals. This is once more illustrated in Section 5 (Example 5.5), where the formulas of Yuškov (and more general formulas) are rederived in a particularly transparent manner.

Section 2 contains some auxiliary results concerning the functions $\sigma_k(z) = \sum_{\nu=-\infty}^{\infty} [z/(\nu+z)]^{k+1}$, which will find use in the subsequent sections.

2. Mathematical Preliminaries

In the following sections some properties of the functions

$$\sigma_k(z) = \sum_{\nu=-\infty}^{\infty} \left(\frac{z}{\nu+z}\right)^{k+1}; \quad k=0, 1, 2, \ldots; z\neq 0 \pmod 1 \tag{2.1}$$

will be needed. (If $k=0$, the summation is to be understood in the sense of a principal value.) The functions (2.1) have been introduced and studied by Ehlich [8], who also gives explicit expressions for $1\leqq k\leqq 11$ and numerical tables. For $k=0$, as is well known,

$$\sigma_0(z) = \pi z \cot \pi z. \tag{2.2}$$

Proposition 2.1.

$$\sigma_{k+1}(z) = \sigma_k(z) - \frac{z}{k+1}\sigma_k'(z), \qquad k = 0, 1, 2, \ldots .$$

Proof. From

$$z^{-(k+1)}\,\sigma_k(z) = \sum_{\nu=-\infty}^{\infty} (\nu+z)^{-(k+1)}$$

one gets

$$\begin{aligned}\frac{d}{dz}\left[z^{-(k+1)}\,\sigma_k(z)\right] &= -(k+1)\sum_{\nu=-\infty}^{\infty}(\nu+z)^{-(k+2)} \\ &= -(k+1)z^{-(k+2)}\,\sigma_{k+1}(z).\end{aligned}$$

Solving for σ_{k+1}, and carrying out the differentiation, gives the desired result.

Proposition 2.2.

$$\sigma_k(z) = \left(\frac{\pi z}{\sin \pi z}\right)^{k+1} q_{k-1}(\cos \pi z), \qquad k = 1, 2, \ldots,$$

where $q_{k-1}(t)$ is a polynomial of degree $k-1$. In fact,

$$q_0(t) = 1, \qquad q_k(t) = t q_{k-1}(t) + \frac{1-t^2}{k+1}\,q_{k-1}'(t), \qquad k = 1, 2, 3, \ldots,$$

so that $q_k(t)$ is even [odd] if k is even [odd].

Proof. By Proposition 2.1, and (2.2),

$$\sigma_1(z) = \pi z \cot \pi z - z\left(\pi \cot \pi z - \frac{\pi^2 z}{\sin^2 \pi z}\right) = \left(\frac{\pi z}{\sin \pi z}\right)^2.$$

The assertion is thus true for $k = 1$, with $q_0(t) = 1$. Proceeding by induction, assume that the proposition holds for some k. Then, by Proposition 2.1, after some elementary computation,

$$\begin{aligned}\sigma_{k+1}(z) &= \sigma_k(z) - \frac{z}{k+1}\,\sigma_k'(z) \\ &= \left(\frac{\pi z}{\sin \pi z}\right)^{k+2}\left\{\cos \pi z \cdot q_{k-1}(\cos \pi z) + \frac{1-\cos^2 \pi z}{k+1}\,q_{k-1}'(\cos \pi z)\right\},\end{aligned}$$

from which the assertion follows.

Proposition 2.3. *For z real and not an integer the matrix*

$$H_{s,p}(z) = \begin{bmatrix} \sigma_s(z) & \sigma_{s+1}(z) & \ldots & \sigma_{s+p-1}(z) \\ \sigma_{s+1}(z) & \sigma_{s+2}(z) & \ldots & \sigma_{s+p}(z) \\ \ldots & \ldots & \ldots & \ldots \\ \sigma_{s+p-1}(z) & \sigma_{s+p}(z) & \ldots & \sigma_{s+2p-2}(z) \end{bmatrix} \tag{2.3}$$

is positive definite if s is an odd integer $\geqq 1$ and p is any integer $\geqq 1$. In particular,

$$\det H_{s,p}(z) > 0, \qquad z \neq 0 \pmod 1, \qquad s\,(\text{odd}) \geqq 1, \qquad p \geqq 1. \tag{2.4}$$

Proof[1]. Let $x^T=[\xi_1, \xi_2, \ldots, \xi_p]$. Then

$$\begin{aligned} x^H H_{s,p}(z)\,x &= \sum_{i,j=1}^{p} \sigma_{s-2+i+j}(z)\bar{\xi}_i \xi_j \\ &= \sum_{i,j=1}^{p} \sum_{\nu=-\infty}^{\infty} \left(\frac{z}{\nu+z}\right)^{s-1+i+j} \bar{\xi}_i \xi_j \\ &= \sum_{\nu=-\infty}^{\infty} \left(\frac{z}{\nu+z}\right)^{s+1} \sum_{i=1}^{p} \bar{\xi}_i \left(\frac{z}{\nu+z}\right)^{i-1} \sum_{j=1}^{p} \xi_j \left(\frac{z}{\nu+z}\right)^{j-1} \\ &= \sum_{\nu=-\infty}^{\infty} \left(\frac{z}{\nu+z}\right)^{s+1} \left| \sum_{i=1}^{p} \xi_i \left(\frac{z}{\nu+z}\right)^{i-1} \right|^2. \end{aligned}$$

The last expression is nonnegative, if s is odd, and can only vanish if the polynomial $\psi(t)=\sum_{i=1}^{p} \xi_i t^{i-1}$ vanishes at all points $z/(\nu+z)$, $\nu=0, \pm 1, \pm 2, \ldots$, i.e., if $\psi(t)\equiv 0$.

Although not needed in the sequel, the following result is offered because of possible independent interest.

Proposition 2.4. *The sequence* $\{\sigma_{k+1}(z)\}_{k=0}^{\infty}$ *has the generating function*

$$\sum_{k=0}^{\infty} \sigma_{k+1}(z)\, y^k = \frac{\pi z}{y} \frac{\sin \pi z y}{\sin(\pi z(1-y)) \sin \pi z}. \tag{2.4}$$

Proof. By (2.1),

$$\begin{aligned} \sum_{k=0}^{\infty} \sigma_{k+1}(z)\, y^k &= \sum_{k=0}^{\infty} y^k \sum_{\nu=-\infty}^{\infty} \left(\frac{z}{\nu+z}\right)^{k+2} \\ &= \sum_{\nu=-\infty}^{\infty} \left(\frac{z}{\nu+z}\right)^{2} \sum_{k=0}^{\infty} \left(\frac{zy}{\nu+z}\right)^{k} = \sum_{\nu=-\infty}^{\infty} \left(\frac{z}{\nu+z}\right)^{2} \frac{1}{1-\frac{zy}{\nu+z}} \\ &= \sum_{\nu=-\infty}^{\infty} \left(\frac{z}{\nu+z}\right)^{2} \frac{\nu+z}{\nu+z-zy} = \frac{z}{y} \sum_{\nu=-\infty}^{\infty} \left(-\frac{1}{\nu+z} + \frac{1}{\nu+z-zy}\right). \end{aligned}$$

Therefore, using (2.2),

$$\sum_{k=0}^{\infty} \sigma_{k+1}(z)\, y^k = \frac{\pi z}{y} [\cot(\pi z(1-y)) - \cot \pi z],$$

which is the same as (2.4).

3. Single Attenuation Factors

We denote by $\boldsymbol{N}$ the set of integers, $\boldsymbol{N}=\{0, \pm 1, \pm 2, \ldots\}$, and by $\boldsymbol{R}$ the set of reals. The set of periodic data will be denoted by

$$\boldsymbol{F}=[\{f_m\}_{m\in\boldsymbol{N}}:\ f_m\in\boldsymbol{R},\ f_{m+N}=f_m,\ \text{all } m\in\boldsymbol{N}].$$

Each member of $\boldsymbol{F}$ may be thought of as the sequence of function values $f_m=f(x_m)$, $x_m=m2\pi/N$, $m=0, \pm 1, \pm 2, \ldots$, for some 2π-periodic function $f(x)$. Evidently,

1 The author is indebted to Professor W. B. Gragg for suggesting the idea for the proof.

$\boldsymbol{F}$ is an N-dimensional linear vector space under the usual definitions of addition and scalar multiplication of sequences. As basis we take $e_0, e_1, \ldots, e_{N-1}$, where the m-th component of e_μ is given by

$$(e_\mu)_m = \begin{cases} 1 & \text{if } m = \mu \pmod N, \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

Each $f \in \boldsymbol{F}$ then has the representation

$$f = \sum_{\mu=0}^{N-1} f_\mu e_\mu, \quad f = \{f_m\} \in \boldsymbol{F}. \tag{3.2}$$

Besides $\boldsymbol{F}$, we consider the linear space of 2π-periodic continuous real-valued functions, which we denote by $\mathscr{F}$. Our approximation process is then thought of as an operator $P: \boldsymbol{F} \to \mathscr{F}$ from $\boldsymbol{F}$ into $\mathscr{F}$. If $\varphi = Pf$ satisfies $\varphi(x_\mu) = f_\mu$, $\mu = 0, 1, \ldots, N-1$, we call P an *interpolation operator*.

We define the shift operator E in $\boldsymbol{F}$ as usual by $(Ef)_m = f_{m+1}$, all $m \in \boldsymbol{N}$, and in $\mathscr{F}$ by $(E\varphi)(x) = \varphi(x+h)$, all $x \in \boldsymbol{R}$, where $h = 2\pi/N$.

Definition. An operator $P: \boldsymbol{F} \to \mathscr{F}$ is called *translation invariant* if

$$P(Ef) = E(Pf), \quad \text{all} \quad f \in \boldsymbol{F}. \tag{3.3}$$

Clearly, (3.3) implies the more general identity

$$P(E^\lambda f) = E^\lambda (Pf), \quad \text{all} \quad \lambda \in \boldsymbol{N}. \tag{3.4}$$

Also note from (3.1) that

$$e_\mu = E^{-\mu} e_0, \quad \mu = 0, 1, 2, \ldots, N-1. \tag{3.5}$$

Theorem 3.1. *Let $P: \boldsymbol{F} \to \mathscr{F}$ be a linear operator from $\boldsymbol{F}$ into $\mathscr{F}$, and $\varphi = Pf$. If P is translation invariant, then*

$$c_n(\varphi) = \tau_n \hat{c}_n(f), \quad \textit{all} \quad n \in \boldsymbol{N}, \quad \textit{all} \quad f \in \boldsymbol{F}, \tag{3.6}$$

where

$$\tau_n = N c_n(\eta_0), \quad \eta_0 = P e_0. \tag{3.7}$$

Proof. Let $f = \{f_m\}$ be an arbitrary element of $\boldsymbol{F}$. From (3.2) we get by the linearity of P, and using (3.4), (3.5),

$$\varphi = Pf = \sum_{\mu=0}^{N-1} f_\mu P e_\mu = \sum_{\mu=0}^{N-1} f_\mu P(E^{-\mu} e_0)$$
$$= \sum_{\mu=0}^{N-1} f_\mu E^{-\mu}(P e_0),$$

that is, the following representation of φ as a discrete convolution,

$$\varphi(x) = \sum_{\mu=0}^{N-1} f_\mu \eta_0(x - x_\mu).$$

Therefore,

$$\begin{aligned} c_n(\varphi) &= \frac{1}{2\pi}\int_0^{2\pi} \varphi(x)e^{-inx}dx \\ &= \frac{1}{2\pi}\sum_{\mu=0}^{N-1} f_\mu \int_0^{2\pi} \eta_0(x-x_\mu)e^{-inx}dx \\ &= \frac{1}{N}\sum_{\mu=0}^{N-1} f_\mu e^{-inx_\mu}\,\frac{N}{2\pi}\int_0^{2\pi} \eta_0(x-x_\mu)e^{-in(x-x_\mu)}\,dx \\ &= \hat{c}_n(f)\cdot N c_n(\eta_0), \end{aligned}$$

since η_0, as an alement of $\mathscr{F}$, is 2π-periodic. This proves (3.6), (3.7).

Theorem 3.1 admits the following converse.

Theorem 3.2. *Suppose each element of $\mathscr{F}_0 \subset \mathscr{F}$ has an uniformly convergent Fourier series. Let $P\colon \boldsymbol{F} \to \mathscr{F}_0$ be an operator such that*

$$c_n(Pf) = \tau_n \hat{c}_n(f), \quad \textit{all} \quad n \in \boldsymbol{N}, \quad \textit{all} \quad f \in \boldsymbol{F}. \tag{3.8}$$

Then the operator P is necessarily linear and translation invariant.

Proof. By assumption,

$$\varphi(x) = (Pf)(x) = \sum_{n=-\infty}^{\infty} c_n(\varphi)e^{inx} = \sum_{n=-\infty}^{\infty} \tau_n \hat{c}_n(f)e^{inx},$$

from which the linearity of P is evident. Moreover,

$$\begin{aligned} P(Ef)(x) &= \sum_{n=-\infty}^{\infty} \tau_n \hat{c}_n(Ef)e^{inx} \\ &= \sum_{n=-\infty}^{\infty} \tau_n e^{inh}\hat{c}_n(f)e^{inx} = \sum_{n=-\infty}^{\infty} \tau_n \hat{c}_n(f)e^{in(x+h)} \\ &= [E(Pf)](x), \end{aligned}$$

i.e., P is translation invariant.

Both theorems can readily be extended to include approximation processes which involve not only function values, but also any number of successive derivative values.

The data space $\boldsymbol{F}$ then consists of elements f which are k-tuples of N-periodic sequences,

$$f = \begin{bmatrix} f_m \\ f'_m \\ \vdots \\ f_m^{(k-1)} \end{bmatrix}_{m\in\boldsymbol{N}}. \tag{3.9}$$

$\boldsymbol{F}$ has thus dimension kN. Defining $\mathscr{F}$ to be the linear space of 2π-periodic $(k-1)$-times continuously differentiable functions, an interpolation process P is an operator from $\boldsymbol{F}$ into $\mathscr{F}$ such that $\varphi = Pf$ satisfies $\varphi^{(\varkappa)}(x_\mu) = f_\mu^{(\varkappa)}$, $\varkappa = 0, 1, \ldots, k-1$, $\mu = 0, 1, \ldots, N-1$.

The definition of translation invariance of an operator P: $\boldsymbol{F}\to\mathscr{F}$ remains as before, if the shift operator E in $\boldsymbol{F}$ is understood to act simultaneously on all sequences $\{f_m^{(s)}\}$ in (3.9).

Theorem 3.1 now extends as follows. *If the operator* P: $\boldsymbol{F}\to\mathscr{F}$ *is linear and translation invariant, and* $\varphi = Pf$, *then*

$$c_n(\varphi) = \sum_{s=0}^{k-1} \tau_{n,s}\, \hat{c}_n(f^{(s)}), \quad \textit{all} \quad n\in\boldsymbol{N}, \tag{3.10}$$

where

$$\tau_{n,s} = N c_n(\eta_{0,s}),$$

$$\eta_{0,0} = P\begin{bmatrix} e_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \eta_{0,1} = P\begin{bmatrix} 0 \\ e_0 \\ \vdots \\ 0 \end{bmatrix}, \ldots, \eta_{0,k-1} = P\begin{bmatrix} 0 \\ 0 \\ \vdots \\ e_0 \end{bmatrix}. \tag{3.11}$$

The proof is virtually the same as before. Also the analogue of Theorem 3.2 is immediate.

Eq. (3.7), in principle, can be used to calculate the attenuation factors for any linear process P which is translation invariant. In simple cases this approach has already been taken by Eagle [7]. In more complicated situations the theorem which follows may be more convenient.

For the purpose of this theorem we define $\mathscr{F}$ to be the linear space of 2π-periodic continuous functions φ whose Fourier series converge at each x_μ,

$$\varphi(x_\mu) = \sum_{m=-\infty}^{\infty} c_m(\varphi)\, e^{imx_\mu}, \quad \mu = 0, 1, \ldots, N-1. \tag{3.12}$$

(The summation in (3.12) and in similar formulas in the sequel is always to be understood in the sense of a principal value.)

If $\varphi\in\mathscr{F}$, then by (1.4), (3.12), for every $n\in\boldsymbol{N}$,

$$\begin{aligned} \hat{c}_n(\varphi) &= \frac{1}{N}\sum_{\mu=0}^{N-1}\left(\sum_{m=-\infty}^{\infty} c_m(\varphi)\, e^{imx_\mu}\right) e^{-inx_\mu} \\ &= \sum_{m=-\infty}^{\infty} c_m(\varphi)\left(\frac{1}{N}\sum_{\mu=0}^{N-1} e^{i(m-n)x_\mu}\right). \end{aligned}$$

Since for any $p\in\boldsymbol{N}$,

$$\frac{1}{N}\sum_{\mu=0}^{N-1} e^{ipx_\mu} = \begin{cases} 1 & \text{if} \quad p = 0 \pmod N \\ 0 & \text{if} \quad p \neq 0 \pmod N, \end{cases} \tag{3.13}$$

it follows, as is well known, that

$$\hat{c}_n(\varphi) = \sum_{\nu=-\infty}^{\infty} c_{\nu N+n}(\varphi), \quad \text{all} \quad n\in\boldsymbol{N}, \quad \text{all} \quad \varphi\in\mathscr{F}. \tag{3.14}$$

Theorem 3.3. *Let* P: $\boldsymbol{F}\to\mathscr{F}$ *be an arbitrary interpolation operator, and* $\varphi = Pf$. *Then for*

$$c_n(\varphi) = \tau_n \hat{c}_n(f), \quad \textit{all} \quad n\in\boldsymbol{N}, \quad \textit{all} \quad f\in\boldsymbol{F} \tag{3.15}$$

to hold with some τ_n *not depending on* f *it is necessary and sufficient that two functions* $\omega(n)$, $\psi_f(n)$ *exist having the following properties:*

(i) $\omega(n)$ *is defined for all* $n \in \boldsymbol{N} \setminus \boldsymbol{N}_0$ *and is independent of* f, *where* $\boldsymbol{N}_0$ *is either the empty set, or a subset of* $\boldsymbol{N}$ *with the property that* $n \in \boldsymbol{N}_0$ *implies* $n + \nu N \notin \boldsymbol{N}_0$, *all* $\nu \in \boldsymbol{N} \setminus \{0\}$;

(ii) $\psi_f(n+N) = \psi_f(n)$, *all* $n \in \boldsymbol{N}$, *all* $f \in \boldsymbol{F}$;

(ii$_0$) $\psi_f(n) = 0$, *all* $n \in \boldsymbol{N}_0$, *all* $f \in \boldsymbol{F}$;

(iii) $c_n(\varphi) = \omega(n)\psi_f(n)$, *all* $n \in \boldsymbol{N} \setminus \boldsymbol{N}_0$, *all* $f \in \boldsymbol{F}$.

If $\omega(n)$, $\psi_f(n)$ *with the stated properties exist, then in fact*

$$\tau_n = \frac{\omega(n)}{\sum_{\nu=-\infty}^{\infty} \omega(\nu N + n)}, \quad \text{all } n \notin \boldsymbol{N}_0 \ (\mathrm{mod}\, N)^2, \tag{3.16}$$

where the series in the denominator converges and has sum $\neq 0$. *Moreover,*

$$\begin{aligned} &\tau_n = 1, \quad \text{all } n \in \boldsymbol{N}_0, \\ &\tau_{n+\nu N} = 0, \quad \text{all } n \in \boldsymbol{N}_0, \quad \nu \in \boldsymbol{N} \setminus \{0\}. \end{aligned} \tag{3.16$_0$}$$

Remarks. 1. The function $\omega(n)$ in (iii) is not uniquely determined, but only up to an arbitrary N-periodic (nonvanishing) factor. Such a factor, of course, does not affect the value of τ_n in (3.16).

2. If (3.15) holds for an interpolation process P which carries $f \equiv 1$ into $\varphi \equiv 1$, then (3.16$_0$) necessarily holds with $\boldsymbol{N}_0 = \{0\}$. To see this, simply apply (3.15) with $\varphi \equiv 1$, $f \equiv 1$, and observe that for this choice

$$c_n = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{if } n \neq 0, \end{cases} \qquad \hat{c}_n = \begin{cases} 1 & \text{if } n = 0 \ (\mathrm{mod}\, N) \\ 0 & \text{if } n \neq 0 \ (\mathrm{mod}\, N). \end{cases}$$

3. If (3.15) holds for an interpolation process P which preserves symmetry about π, i.e., for which

$$f_m = f_{N-m} \quad \text{implies} \quad \varphi(x) = \varphi(2\pi - x), \tag{3.17}$$

then τ_n is necessarily real-valued. This follows from the fact that both $\hat{c}_n(f)$ and $c_n(\varphi)$ are real for any f, $\varphi = Pf$ satisfying (3.17). All examples considered later will meet the condition in (3.17). It is not difficult, however, to construct interpolation processes which violate (3.17). For example, if the restriction of φ to $[x_\mu, x_{\mu+1}]$, $\mu = 0, 1, \ldots, N-1$, is taken to be the quadratic polynomial interpolating f at x_μ, $x_{\mu+1}$, $x_{\mu+2}$, then symmetry is clearly destroyed. Accordingly, one finds in this case that

$$\omega(n) = \frac{1}{n^2}\left(\frac{1}{2} + \frac{i}{n}\right), \qquad \psi_f(n) = \sum_{\mu=0}^{N-1} \Delta^3 f_{\mu-1} e^{-inx_\mu},$$

and $\omega(n)$ is indeed complex-valued.

Proof of Theorem 3.3. (a) Sufficiency. Assume that $\omega(n)$, $\psi_f(n)$ with the properties (i)–(iii) exist. According to (iii) we have that

$$c_n(\varphi) = \omega(n)\psi_f(n), \qquad \varphi = Pf,$$

2 If $\boldsymbol{N}_0$ is empty, the quantifier is to be interpreted as "all $n \in \boldsymbol{N}$", otherwise "all $n \neq n_0 \ (\mathrm{mod}\, N)$" where n_0 is an arbitrary integer in $\boldsymbol{N}_0$.

holds for all $f \in \boldsymbol{F}$ and all $n \in \boldsymbol{N} \setminus \boldsymbol{N}_0$. Consequently, by (3.14), noting that $\hat{c}_n(\varphi) = \hat{c}_n(f)$ and using (ii), we get

$$\begin{aligned} \hat{c}_n(f) &= \sum_{\nu=-\infty}^{\infty} c_{\nu N+n}(\varphi) = \sum_{\nu=-\infty}^{\infty} \omega(\nu N+n)\psi_f(\nu N+n) \\ &= \sum_{\nu=-\infty}^{\infty} [\omega(\nu N+n)\psi_f(n)], \quad \text{all} \quad n \notin \boldsymbol{N}_0 \ (\mathrm{mod}\ N), \quad \text{all} \quad f \in \boldsymbol{F}, \end{aligned} \tag{3.18}$$

where the last series converges (possibly trivially, if $\psi_f(n) = 0$).

We show that in fact

$$\sum_{\nu=-\infty}^{\infty} \omega(\nu N+n) \tag{3.19}$$

converges and has a nonzero sum. For this it suffices to exhibit an $f \in \boldsymbol{F}$ for which $\hat{c}_n(f) \neq 0$. Because then, $\psi_f(n) \neq 0$ for this particular f, by virtue of (3.18), and the assertion follows, again from (3.18). The choice $f \in \boldsymbol{F}$, however, such that

$$f_0 = 1, \quad f_\mu = 0 \quad \text{for} \quad \mu = 1, 2, \ldots, N-1,$$

will do, since then $\hat{c}_n(f) = 1/N \neq 0$.

Convergence of (3.19) being assured, we now conclude from (3.18) that

$$\hat{c}_n(f) = \left[\sum_{\nu=-\infty}^{\infty} \omega(\nu N+n)\right] \psi_f(n), \quad \text{all} \quad n \notin \boldsymbol{N}_0 \ (\mathrm{mod}\ N), \quad \text{all} \quad f \in \boldsymbol{F}.$$

Multiplying through by $\omega(n)$, we get

$$\begin{aligned} \hat{c}_n(f)\omega(n) &= \left[\sum_{\nu=-\infty}^{\infty} \omega(\nu N+n)\right] \omega(n)\psi_f(n) \\ &= \left[\sum_{\nu=-\infty}^{\infty} \omega(\nu N+n)\right] c_n(\varphi), \end{aligned}$$

proving (3.15) for $n \notin \boldsymbol{N}_0 \ (\mathrm{mod}\ N)$, with τ_n as defined in (3.16).

Assuming now $n \in \boldsymbol{N}_0$, we have by (ii_0) that $\psi_f(n) = 0$, all $f \in \boldsymbol{F}$, and by (ii) that $\psi_f(n + \nu N) = 0$, all $f \in \boldsymbol{F}$, $\nu \in \boldsymbol{N}$. Since by assumption $n + \nu N \notin \boldsymbol{N}_0$ for $\nu \neq 0$, it follows from (iii) that $c_{n+\nu N}(\varphi) = 0$, all $f \in \boldsymbol{F}$, $n \in \boldsymbol{N}_0$, $\nu \in \boldsymbol{N} \setminus \{0\}$. This proves (3.15) for integers of the form $n + \nu N$, $n \in \boldsymbol{N}_0$, $\nu \neq 0$, with $\tau_{n+\nu N} = 0$. From this, and (3.14), noting again that $\hat{c}_n(\varphi) = \hat{c}_n(f)$, it follows further that $\hat{c}_n(f) = c_n(\varphi)$, $n \in \boldsymbol{N}_0$, proving (3.15) for $n \in \boldsymbol{N}_0$, with $\tau_n = 1$.

(b) The necessity of (i)–(iii) is trivial, since we may take $\omega(n) = \tau_n$, $\psi_f(n) = \hat{c}_n(f)$, and $\boldsymbol{N}_0$ the empty set. Theorem 3.3 is proved.

4. Several Attenuation Factors

The concept of translation invariance introduced in the previous section will be generalized as follows.

Definition. An operator P: $\boldsymbol{F} \to \mathscr{F}$ is called *r-translation invariant* (r an integer), if

$$P(E^r f) = E^r(Pf), \quad \text{all} \quad f \in \boldsymbol{F}. \tag{4.1}$$

Translation invariance in the previous sense thus coincides with 1-translation invariance. It implies r-translation invariance for each integer r. On the other hand, P may be r-translation invariant, for some fixed $r>1$, without being 1-translation invariant. Note that (4.1) implies

$$P(E^{\lambda r}f)=E^{\lambda r}(Pf), \quad \text{all} \quad \lambda\in\boldsymbol{N}. \tag{4.2}$$

Theorem 4.1. *Let $r\geqq 1$ be an integer and N be divisible by r,*

$$N=rq. \tag{4.3}$$

Let P: $\boldsymbol{F}\to\mathscr{F}$ be a linear operator from $\boldsymbol{F}$ into $\mathscr{F}$, and $\varphi=Pf$. If P is r-translation invariant, then

$$c_n(\varphi)=\sum_{\varrho=0}^{r-1}\tau_{n,\varrho}\hat{c}_{n+\varrho q}(f), \quad \textit{all} \quad n\in\boldsymbol{N}, \quad \textit{all} \quad f\in\boldsymbol{F}, \tag{4.4}$$

where

$$\tau_{n,\varrho}=\frac{N}{r}\sum_{\sigma=0}^{r-1}e^{i(n+\varrho q)x_\sigma}c_n(\eta_\sigma), \qquad \eta_\sigma=Pe_\sigma. \tag{4.5}$$

Remarks. 1. If P is 1-translation invariant, and thus also r-translation invariant for any $r\geqq 1$, then (4.4) reduces to (3.6). In fact, observing that

$$\eta_\sigma=Pe_\sigma=E^{-\sigma}Pe_0, \qquad \eta_\sigma(x)=\eta_0(x-x_\sigma),$$

and thus

$$c_n(\eta_\sigma)=\frac{1}{2\pi}\int_0^{2\pi}\eta_0(x-x_\sigma)e^{-inx}dx=e^{-inx_\sigma}c_n(\eta_0),$$

we find from (4.5),

$$\tau_{n,\varrho}=\frac{N}{r}c_n(\eta_0)\sum_{\sigma=0}^{r-1}e^{i\varrho q x_\sigma}=\begin{cases}Nc_n(\eta_0) & \text{if} \quad \varrho=0\\ 0 & \text{if} \quad 0<\varrho\leqq r-1.\end{cases}$$

2. The result (4.4) may be given a slightly different form by breaking up the sum on the right of (4.4) into two pieces, according to $\sum_{\varrho=0}^{r-1}=\sum_{\varrho=0}^{[r/2]}+\sum_{\varrho=[r/2]+1}^{r-1}$, and introducing the new variable of summation $\varrho'=\varrho-r$ in the second piece. Using (1.7), one gets

$$c_n(\varphi)=\sum_{\varrho=[-r/2]+1}^{[r/2]}\hat{\tau}_{n,\varrho}\hat{c}_{n+\varrho q}(f), \tag{4.4'}$$

where

$$\hat{\tau}_{n,\varrho}=\begin{cases}\tau_{n,\varrho} & \text{if} \quad \varrho\geqq 0\\ \tau_{n,\varrho+r} & \text{if} \quad \varrho<0.\end{cases} \tag{4.5'}$$

Proof of Theorem 4.1. Using the representation (3.2) we have for each $f\in\boldsymbol{F}$,

$$f=\sum_{\mu=0}^{N-1}f_\mu e_\mu=\sum_{\varrho=0}^{r-1}\sum_{\lambda=0}^{q-1}f_{\varrho+\lambda r}e_{\varrho+\lambda r}$$
$$=\sum_{\varrho=0}^{r-1}\sum_{\lambda=0}^{q-1}f_{\varrho+\lambda r}E^{-\lambda r}e_\varrho.$$

By linearity of P, and (4.2),

$$\varphi = Pf = \sum_{\varrho=0}^{r-1} \sum_{\lambda=0}^{q-1} f_{\varrho+\lambda r} P(E^{-\lambda r} e_\varrho)$$
$$= \sum_{\varrho=0}^{r-1} \sum_{\lambda=0}^{q-1} f_{\varrho+\lambda r} E^{-\lambda r}(P e_\varrho),$$

i.e.,

$$\varphi(x) = \sum_{\varrho=0}^{r-1} \sum_{\lambda=0}^{q-1} f_{\varrho+\lambda r} \eta_\varrho(x - x_{\lambda r}).$$

Therefore,

$$c_n(\varphi) = \frac{1}{2\pi} \sum_{\varrho=0}^{r-1} \sum_{\lambda=0}^{q-1} f_{\varrho+\lambda r} \int_0^{2\pi} \eta_\varrho(x - x_{\lambda r}) e^{-inx} dx$$
$$= \sum_{\varrho=0}^{r-1} \sum_{\lambda=0}^{q-1} f_{\varrho+\lambda r} e^{-inx_{\lambda r}} c_n(\eta_\varrho)$$
$$= \sum_{\varrho=0}^{r-1} c_n(\eta_\varrho) e^{inx_\varrho} \sum_{\lambda=0}^{q-1} f_{\varrho+\lambda r} e^{-inx_{\varrho+\lambda r}}.$$

Letting

$$d_\varrho = d_{\varrho,n} = \sum_{\lambda=0}^{q-1} f_{\varrho+\lambda r} e^{-inx_{\varrho+\lambda r}}, \qquad \varrho = 0, 1, \ldots, r-1,$$

we thus have

$$c_n(\varphi) = \sum_{\varrho=0}^{r-1} c_n(\eta_\varrho) e^{inx_\varrho} d_\varrho. \tag{4.6}$$

Now observe that

$$\hat{c}_n(f) = \frac{1}{N} \sum_{\varrho=0}^{r-1} d_\varrho.$$

More generally,

$$\hat{c}_{n+\sigma q}(f) = \frac{1}{N} \sum_{\varrho=0}^{r-1} \sum_{\lambda=0}^{q-1} f_{\varrho+\lambda r} e^{-i(n+\sigma q)x_{\varrho+\lambda r}}, \qquad \sigma = 0, 1, \ldots, r-1. \tag{4.7}$$

Since

$$(n+\sigma q)x_{\varrho+\lambda r} = n x_{\varrho+\lambda r} + \sigma q(\varrho+\lambda r)\frac{2\pi}{rq}$$
$$= n x_{\varrho+\lambda r} + \sigma q x_\varrho + \sigma\lambda \cdot 2\pi,$$

Eq. (4.7) simplifies to

$$\hat{c}_{n+\sigma q}(f) = \frac{1}{N} \sum_{\varrho=0}^{r-1} e^{-i\sigma q x_\varrho} \sum_{\lambda=0}^{q-1} f_{\varrho+\lambda r} e^{-inx_{\varrho+\lambda r}},$$

i.e., to

$$N\hat{c}_{n+\sigma q}(f) = \sum_{\varrho=0}^{r-1} e^{-i\sigma\varrho 2\pi/r} d_\varrho, \qquad \sigma = 0, 1, \ldots, r-1. \tag{4.8}$$

We have obtained a system of r linear equations in r "unknowns" d_ϱ. The coefficient matrix is the Vandermonde matrix

$$V(1, e^{-2\pi i/r}, e^{-2\cdot 2\pi i/r}, \ldots, e^{-(r-1)2\pi i/r})$$

in the r-th roots of unity. Since the inverse of V is readily found to be

$$V^{-1}=[v_{\sigma\varrho}], \qquad v_{\sigma\varrho}=\frac{1}{r}e^{\sigma\varrho 2\pi i/r},$$

one obtains from (4.8),

$$d_\sigma=\frac{N}{r}\sum_{\varrho=0}^{r-1}e^{\sigma\varrho 2\pi i/r}\,\hat{c}_{n+\varrho q}(f),$$

and thus from (4.6),

$$\begin{aligned}c_n(\varphi)&=\sum_{\sigma=0}^{r-1}c_n(\eta_\sigma)\,e^{inx_\sigma}d_\sigma\\&=\frac{N}{r}\sum_{\sigma=0}^{r-1}c_n(\eta_\sigma)e^{inx_\sigma}\sum_{\varrho=0}^{r-1}e^{\sigma\varrho 2\pi i/r}\,\hat{c}_{n+\varrho q}(f)\\&=\sum_{\varrho=0}^{r-1}\left(\frac{N}{r}\sum_{\sigma=0}^{r-1}c_n(\eta_\sigma)\,e^{inx_\sigma+\sigma\varrho 2\pi i/r}\right)\hat{c}_{n+\varrho q}(f).\end{aligned}$$

This proves (4.4) and (4.5).

The following converse of Theorem 4.1 is proved in the same manner as Theorem 3.2.

Theorem 4.2. *Suppose each element of $\mathscr{F}_0\subset\mathscr{F}$ has an uniformly convergent Fourier series. Let $P\colon \boldsymbol{F}\to\mathscr{F}_0$ be an operator such that*

$$c_n(Pf)=\sum_{\varrho=0}^{r-1}\tau_{n,\varrho}\,\hat{c}_{n+\varrho q}(f), \quad \textit{all} \quad n\in\boldsymbol{N}, \quad \textit{all} \quad f\in\boldsymbol{F}. \tag{4.9}$$

Then the operator P is necessarily linear and r-translation invariant.

It is clear that in analogy to the discussion in Section 3 we could also treat Hermite approximation processes which are r-translation invariant. However, we shall not pursue this here any further.

5. Examples

We begin by recalling a well-known representation for Fourier coefficients. Suppose g is a 2π-periodic function which, together with all derivatives of orders up to and including the $(k+1)$-st, is piecewise continuous. Let $\xi_\sigma^{(s)}$ denote the points of discontinuity of $g^{(s)}$ in the half-open interval $[0,2\pi)$, and $\delta g_\sigma^{(s)}$ the respective jumps

$$\delta g_\sigma^{(s)}\overset{\text{def}}{=}\lim_{x\downarrow 0}\left[g^{(s)}(\xi_\sigma^{(s)}+x)-g^{(s)}(\xi_\sigma^{(s)}-x)\right]. \tag{5.1}$$

Then, for $n\neq 0$,

$$\begin{aligned}2\pi c_n(g)=&\frac{1}{in}\sum_\sigma\delta g_\sigma e^{-in\xi_\sigma}+\frac{1}{(in)^2}\sum_\sigma\delta g_\sigma' e^{-in\xi_\sigma'}+\cdots\\&+\frac{1}{(in)^{k+1}}\sum_\sigma\delta g_\sigma^{(k)}e^{-in\xi_\sigma^{(k)}}+\frac{1}{(in)^{k+1}}\int_0^{2\pi}g^{(k+1)}(x)\,e^{-inx}\,dx.\end{aligned} \tag{5.2}$$

The proof follows directly from a repeated application of integration by parts.

Eq. (5.2), applied to $g=\varphi$, will be useful in deriving the factorization in (iii) of Theorem 3.3.

We note that the validity of (5.2) is not restricted to integer values of n. The result holds for arbitrary real n, provided the "jumps" at $\xi_0=0$ are defined by

$$\delta g_0^{(s)} \stackrel{\text{def}}{=} \lim_{x \downarrow 0} [g^{(s)}(x) - e^{-2\pi i n} g^{(s)}(-x)], \tag{5.3}$$

regardless of whether $g^{(s)}$ is continuous at 0 or not.

5.1. *Single Attenuation Factors*

For completeness we include as first example the well-known case where $\varphi = Pf$ is a polynomial spline function (Eagle [7], Quade and Collatz [13], Bauer and Stetter [2], Ehlich [8], Golomb [10]).

Example 5.1. Given $f \in \boldsymbol{F}$, let $\varphi = Pf$ denote the periodic spline interpolant of degree $2r-1$, i.e.,

$$\begin{aligned} &\varphi \in C^{2r-2}(-\infty, \infty), \\ &\varphi(x+2\pi) = \varphi(x), \quad \text{all} \quad x \in \boldsymbol{R}, \\ &\varphi(x_\mu) = f_\mu, \quad \mu = 0, 1, \ldots, N-1, \\ &\varphi^{(2r)}(x) = 0, \quad x \neq x_\mu. \end{aligned} \tag{5.4}$$

It is known that φ exists uniquely.

We first illustrate the use of Theorem 3.3. Applying (5.2) to φ, with $k=2r-2$, we get

$$\begin{aligned} 2\pi c_n(\varphi) &= \frac{1}{(in)^{2r-1}} \int_0^{2\pi} \varphi^{(2r-1)}(x) e^{-inx} dx \\ &= \frac{1}{(in)^{2r-1}} \sum_{\mu=0}^{N-1} \varphi^{(2r-1)}\left(x_\mu + \frac{1}{2}h\right) \int_{x_\mu}^{x_{\mu+1}} e^{-inx} dx, \quad n \neq 0, \end{aligned}$$

$\varphi^{(2r-1)}$ being piecewise constant. Thus,

$$c_n(\varphi) = \omega(n) \psi_f(n), \quad n \neq 0,$$

with

$$\omega(n) = \frac{1}{n^{2r}}, \qquad \psi_f(n) = \frac{(-1)^r}{2\pi} in \sum_{\mu=0}^{N-1} \varphi^{(2r-1)}\left(x_\mu + \frac{1}{2}h\right) \int_{x_\mu}^{x_{\mu+1}} e^{-inx} dx.$$

Clearly, $\psi_f(0)=0$, all $f \in \boldsymbol{F}$, and the relation

$$in \int_{x_\mu}^{x_{\mu+1}} e^{-inx} dx = e^{-inx_\mu} - e^{-inx_{\mu+1}}$$

shows that $\psi_f(n)$ has period N. All conditions (i)–(iii) of Theorem 3.3, with $\boldsymbol{N}_0 = \{0\}$, being verified, we conclude that

$$c_n(\varphi) = \tau_n \hat{c}_n(f), \quad \text{all} \quad n \in \boldsymbol{N}, \quad \text{all} \quad f \in \boldsymbol{F},$$

with $\tau_0 = 1$, $\tau_{\nu N} = 0$ for $\nu \neq 0$, and

$$\tau_n = \frac{1}{n^{2r} \sum_{\nu=-\infty}^{\infty} (\nu N + n)^{-2r}} \quad \text{for} \quad n \neq 0 \pmod N.$$

In terms of the functions $\sigma_k(z)$ defined in (2.1), we can write more briefly

$$\tau_n = \frac{1}{\sigma_{2r-1}(z)}, \quad z = \frac{n}{N}, \quad n \neq 0 \pmod N. \tag{5.5}$$

The same result can also be derived from Theorem 3.1[3]. Let, in fact,

$$\psi(u) = \left[\frac{\sin(u/2)}{u/2}\right]^{2r}, \qquad \Phi(u) = \sum_{\nu=-\infty}^{\infty} \psi(u + 2\pi\nu).$$

Schoenberg [16] has shown that

$$L_r(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\psi(u)}{\Phi(u)} e^{iut} du \tag{5.6}$$

is a cardinal spline interpolant of degree $2r-1$, i.e., of continuity class $C^{2r-2}(-\infty, \infty)$, a polynomial of degree $2r-1$ on each open interval $(m, m+1)$, $m \in \boldsymbol{N}$, and satisfying the unit interpolation conditions

$$L_r(m) = \begin{cases} 1 & \text{if} \quad m = 0 \\ 0 & \text{if} \quad m \neq 0, \end{cases} \qquad m \in \boldsymbol{N}.$$

It follows from this that

$$\eta_0(x) = \sum_{\varkappa=-\infty}^{\infty} L_r\left(\frac{x}{h} + \varkappa N\right),$$

so that (3.7) gives

$$\begin{aligned} \tau_n &= \frac{N}{2\pi} \int_0^{2\pi} \eta_0(x) e^{-inx} dx \\ &= \frac{1}{h} \int_0^{2\pi} \sum_{\varkappa=-\infty}^{\infty} L_r\left(\frac{x}{h} + \varkappa N\right) e^{-inx} dx \\ &= \frac{1}{h} \sum_{\varkappa=-\infty}^{\infty} \int_0^{2\pi} L_r\left(\frac{x}{h} + \varkappa N\right) e^{-inx} dx \\ &= \sum_{\varkappa=-\infty}^{\infty} \int_{\varkappa N}^{(\varkappa+1)N} L_r(t) e^{-inht} dt = \int_{-\infty}^{\infty} L_r(t) e^{-inht} dt. \end{aligned}$$

Inverting the Fourier transform in (5.6), on the other hand, shows that

$$\frac{\psi(u)}{\Phi(u)} = \int_{-\infty}^{\infty} L_r(t) e^{-iut} dt,$$

so that

$$\tau_n = \frac{\psi(nh)}{\Phi(nh)} = \frac{\psi(2\pi z)}{\sum_{\nu=-\infty}^{\infty} \psi(2\pi(z+\nu))} = \frac{1}{\sum_{\nu=-\infty}^{\infty} \left(\frac{z}{z+\nu}\right)^{2r}} = \frac{1}{\sigma_{2r-1}(z)},$$

in agreement with (5.5).

3 The derivation which follows is due to Dr. Christian H. Reinsch.

Example 5.2. Let $f \in \boldsymbol{F}$, and $p(x) = p_\mu(x)$ be the polynomial of degree $2r-1$ interpolating f at the $2r$ points $x_{\mu+\lambda}$, $\lambda = -r+1, -r+2, \ldots, r-1, r$. Define $\varphi(x) = p_\mu(x)$ on $[x_\mu, x_{\mu+1}]$, $\mu = 0, \pm 1, \pm 2, \ldots$.

Clearly, $\varphi \in C(-\infty, \infty)$. All derivatives $\varphi^{(s)}$, $s = 1, 2, \ldots, 2r-1$, are continuous on each open interval $(x_\mu, x_{\mu+1})$, but will have jumps at the points x_μ. Since $\varphi^{(2r)}(x) = 0$, a.e., we have from (5.2), applied to φ,

$$2\pi c_n(\varphi) = \sum_{s=1}^{2r-1} \frac{1}{(in)^{s+1}} \sum_{\mu=0}^{N-1} \delta\varphi_\mu^{(s)} e^{-inx_\mu}, \qquad n \neq 0. \tag{5.7}$$

We proceed to calculate the jumps $\delta\varphi_\mu^{(s)}$. Let

$$P(t) = p_\mu(x_\mu + th), \qquad h = \frac{2\pi}{N}.$$

By Newton's interpolation formula we have

$$P(t) = \sum_{k=0}^{2r-1} \binom{t+r-1}{k} \Delta^k f_{\mu-r+1}.$$

Therefore,

$$P^{(s)}(0) = \sum_{k=s}^{2r-1} \binom{t+r-1}{k}^{(s)}_{t=0} \Delta^k f_{\mu-r+1}, \qquad s = 1, 2, \ldots, 2r-1. \tag{5.8$_0$}$$

$$P^{(s)}(1) = \sum_{k=s}^{2r-1} \binom{t+r-1}{k}^{(s)}_{t=1} \Delta^k f_{\mu-r+1}, \tag{5.8$_1$}$$

We obtain $h^s \delta\varphi_\mu^{(s)}$ by subtracting from (5.8$_0$) the relation (5.8$_1$) with μ replaced by $\mu-1$,

$$h^s \delta\varphi_\mu^{(s)} = \sum_{k=s}^{2r-1} \left\{ \binom{t+r-1}{k}^{(s)}_{t=0} \Delta^k f_{\mu-r+1} - \binom{t+r-1}{k}^{(s)}_{t=1} \Delta^k f_{\mu-r} \right\}.$$

The sum on the right, however, "collapses", since

$$\begin{aligned}
&\binom{t+r-1}{k}^{(s)}_{t=0} \Delta^k f_{\mu-r+1} - \binom{t+r-1}{k}^{(s)}_{t=1} \Delta^k f_{\mu-r} \\
&= \binom{t+r-1}{k}^{(s)}_{t=0} [\Delta^k f_{\mu-r+1} - \Delta^k f_{\mu-r}] - \left[\binom{t+r-1}{k}^{(s)}_{t=1} - \binom{t+r-1}{k}^{(s)}_{t=0} \right] \Delta^k f_{\mu-r} \\
&= \binom{t+r-1}{k}^{(s)}_{t=0} \Delta^{k+1} f_{\mu-r} - \binom{t+r-1}{k-1}^{(s)}_{t=0} \Delta^k f_{\mu-r}.
\end{aligned}$$

As a result,

$$h^s \delta\varphi_\mu^{(s)} = \binom{t+r-1}{2r-1}^{(s)}_{t=0} \Delta^{2r} f_{\mu-r}, \qquad s = 1, 2, \ldots, 2r-1. \tag{5.9}$$

Since

$$\binom{t+r-1}{2r-1} = \frac{1}{(2r-1)!} [t^2 - (r-1)^2] \ldots [t^2 - 1] t$$

is an odd function, all derivatives of even order at $t = 0$ vanish in (5.9). Letting

$$\gamma_{rs} = (-1)^{r+s} \binom{t+r-1}{2r-1}^{(2s-1)}_{t=0}, \qquad s = 1, 2, \ldots, r,$$

we thus obtain from (5.7)

$$2\pi c_n(\varphi) = (-1)^r h \sum_{s=1}^{r} \frac{\gamma_{rs}}{(nh)^{2s}} \sum_{\mu=0}^{N-1} (\Delta^{2r} f_{\mu-r}) e^{-inx_\mu}, \qquad n \neq 0.$$

This is the desired factorization $c_n(\varphi) = \omega(n)\psi_f(n)$, with

$$\omega(n) = \sum_{s=1}^{r} \frac{\gamma_{rs}}{(hn)^{2s}}, \qquad n \neq 0,$$

$$\psi_f(n) = \frac{(-1)^r h}{2\pi} \sum_{\mu=0}^{N-1} (\Delta^{2r} f_{\mu-r}) e^{-inx_\mu}.$$

It is indeed evident that $\psi_f(n)$ has period N. Moreover, since with f also all differences $\Delta^s f$ have period N, one has with $g_\mu = \Delta^{2r-1} f_{\mu-r}$ that

$$\sum_{\mu=0}^{N-1} \Delta^{2r} f_{\mu-r} = \sum_{\mu=0}^{N-1} \Delta g_\mu = g_N - g_0 = 0,$$

i.e., $\psi_f(0) = 0$. Theorem 3.3 is thus applicable with $\boldsymbol{N}_0 = \{0\}$, giving

$$c_n(\varphi) = \tau_n \hat{c}_n(f), \quad \text{all} \quad n \in \boldsymbol{N}, \quad \text{all} \quad f \in \boldsymbol{F},$$

with

$$\tau_0 = 1, \qquad \tau_{\nu N} = 0 \quad \text{for} \quad \nu \neq 0,$$

and

$$\tau_n = \frac{\sum_{s=1}^{r} \gamma_{rs} (2\pi z)^{2r-2s}}{\sum_{s=1}^{r} \gamma_{rs} (2\pi z)^{2r-2s} \sigma_{2s-1}(z)}, \qquad z = \frac{n}{N}, \qquad n \neq 0 \pmod{N}. \tag{5.10}$$

A short table of the coefficients γ_{rs} follows.

Table 5.1. *The coefficients γ_{rs} in* (5.10)

r \ s	1	2	3	4
1	1			
2	1/6	1		
3	1/30	1/4	1	
4	1/140	7/120	1/3	1

Using Proposition 2.2 on obtains from (5.10), after some computation,

$$\tau_n = \left(\frac{\sin \pi z}{\pi z}\right)^2 \qquad (r = 1), \tag{5.10$_1$}$$

$$\tau_n = \left(\frac{\sin \pi z}{\pi z}\right)^4 \left[1 + \frac{2}{3}(\pi z)^2\right] \qquad (r = 2), \tag{5.10$_2$}$$

$$\tau_n = \left(\frac{\sin \pi z}{\pi z}\right)^6 \left[1 + (\pi z)^2 + \frac{8}{15}(\pi z)^4\right] \qquad (r = 3), \tag{5.10$_3$}$$

$$\tau_n = \left(\frac{\sin \pi z}{\pi z}\right)^8 \left[1 + \frac{4}{3}(\pi z)^2 + \frac{14}{15}(\pi z)^4 + \frac{16}{35}(\pi z)^6\right] \qquad (r = 4). \tag{5.10$_4$}$$

The first of these, (5.10_1), of course, agrees with the attenuation factor for spline interpolants of degree one; the second, for $r=2$, is due to Dällenbach [6].

Comparing (5.10_1)–(5.10_4) with the corresponding attenuation factors for splines,

$$\tau_n=\left(\frac{\sin \pi z}{\pi z}\right)^{2r}\frac{1}{q_{2r-2}(\cos \pi z)}$$

[cf. (5.5) and Proposition 2.2], one notes that the polynomial factors in the brackets of (5.10_1)–(5.10_4) are precisely the beginning terms of the power series expansion of $1/q_{2r-2}(\cos \pi z)$. Thus, for small $z=n/N$, the interpolation polynomials of Example 5.2 give only slightly different Fourier coefficients compared to the spline interpolants of Example 5.1. For large z, however, the attenuation factors show different behavior, as they should, the two approximants belonging to different continuity classes.

Example 5.3. Let δ denote the central difference operator, $\delta y_\mu=y_{\mu+\frac{1}{2}}-y_{\mu-\frac{1}{2}}$, and $\bar{\delta}^{2k-1}y_\mu=\frac{1}{2}(\delta^{2k-1}y_{\mu-\frac{1}{2}}+\delta^{2k-1}y_{\mu+\frac{1}{2}})$ the mean odd differences. There are unique finite difference expressions

$$(5.11)\qquad (L_{2s}y)_\mu=\sum_{k=s}^{r-1}a_{sk}\delta^{2k}y_\mu,\qquad (L_{2s-1}y)_\mu=\sum_{k=s}^{r-1}b_{sk}\bar{\delta}^{2k-1}y_\mu$$

such that

$$(5.12)\qquad h^\varrho y^{(\varrho)}(x_\mu)=(L_\varrho y)_\mu,\qquad \varrho=0,1,\ldots,2r-2$$

is valid for any polynomial y of degree $\leqq 2r-2$. The coefficients a_{sk} and b_{sk} in (5.11) indeed [11, p. 136] are the coefficients in the power series expansions

$$\left[2\sinh^{-1}\frac{z}{2}\right]^{2s}=\sum_{k=s}^{\infty}a_{sk}z^{2k},\qquad \frac{\left[2\sinh^{-1}\frac{z}{2}\right]^{2s-1}}{\sqrt{1+z^2/4}}=\sum_{k=s}^{\infty}b_{sk}z^{2k-1}.$$

Given $f\in\boldsymbol{F}$, let now $p(x)=p_\mu(x)$ be the unique polynomial of degree $2r-1$ satisfying

$$h^s p^{(s)}(x_\mu)=(L_s f)_\mu,\qquad h^s p^{(s)}(x_{\mu+1})=(L_s f)_{\mu+1},\qquad s=0,1,\ldots,r-1,$$

where $h=2\pi/N$, and let $\varphi(x)=p_\mu(x)$ on $[x_\mu, x_{\mu+1}]$, $\mu=0, \pm1, \pm2, \ldots$. Clearly, $\varphi\in C^{r-1}(-\infty, \infty)$, so that φ has now a degree of smoothness which is about midway between those in Examples 5.1 and 5.2.

Letting $P(t)=p_\mu(x_\mu+th)$, the well-known formula for Hermite interpolation gives

$$(5.13)\qquad P(t)=\sum_{s=0}^{r-1}h_s(t)(L_s f)_\mu+\sum_{s=0}^{r-1}(-1)^s h_s(1-t)(L_s f)_{\mu+1},$$

where

$$h_s(t)=\frac{1}{s!}t^s(1-t)^r\sum_{\sigma=0}^{r-s-1}\binom{r+\sigma-1}{\sigma}t^\sigma,\qquad s=0,1,\ldots,r-1.$$

Similarly as in the previous example, we may now calculate the jumps $\delta\varphi_\mu^{(\varrho)}$ for $r\leqq\varrho\leqq 2r-1$, and then use (5.2) to find a factorization for $c_n(\varphi)$. We omit the somewhat lengthy calculations and content ourselves in stating the final result.

One finds that

$$c_n(\varphi) = \omega(n)\psi_f(n), \qquad n \neq 0,$$

where

$$\psi_f(n) = \sum_{\mu=0}^{N-1} f_\mu e^{-in x_\mu} \tag{5.14}$$

and

$$\omega(n) = \sum_{\varrho=\left[\frac{r+1}{2}\right]}^{r-1} \frac{A_\varrho(z)}{(2\pi z)^{2\varrho+1}} + \sum_{\varrho=\left[\frac{r}{2}\right]}^{r-1} \frac{B_\varrho(z)}{(2\pi z)^{2\varrho+2}}, \qquad z = \frac{n}{N}. \tag{5.15}$$

Here,

$$\begin{aligned} A_\varrho(z) = 2(-1)^\varrho \Big\{ &\sin 2\pi z \sum_{s=0}^{[(r-1)/2]} h_{2s}^{(2\varrho)}(1)\alpha_s(\sin \pi z) \\ &+ \cos \pi z \sum_{s=1}^{[r/2]} [h_{2s-1}^{(2\varrho)}(1) \cos 2\pi z - h_{2s-1}^{(2\varrho)}(0)]\, \beta_s(\sin \pi z) \Big\}, \end{aligned} \tag{5.16}$$

$$\begin{aligned} B_\varrho(z) = 2(-1)^{\varrho+1} \Big\{ &\sum_{s=0}^{[(r-1)/2]} [h_{2s}^{(2\varrho+1)}(0) - h_{2s}^{(2\varrho+1)}(1) \cos 2\pi z]\alpha_s(\sin \pi z) \\ &+ \cos \pi z \sin 2\pi z \sum_{s=1}^{[r/2]} h_{2s-1}^{(2\varrho+1)}(1)\beta_s(\sin \pi z) \Big\}, \end{aligned} \tag{5.17}$$

where

$$\alpha_s(z) = \sum_{k=s}^{r-1} (-1)^k a_{sk}(2z)^{2k}, \qquad \beta_s(z) = \sum_{k=s}^{r-1} (-1)^k b_{sk}(2z)^{2k-1}. \tag{5.18}$$

It is evident that $\psi_f(n) = N\hat{c}_n(f)$ in (5.14) has period N. Also, $A_\varrho(z)$ and $B_\varrho(z)$ both have a factor $\sin^2 \pi z$. This follows for A_ϱ from the identity

$$h_0(t) = h_1(t) - h_1(1-t) + 1 - t,$$ [4]

which implies

$$h_0^{(2\varrho)}(1) = h_1^{(2\varrho)}(1) - h_1^{(2\varrho)}(0), \qquad \text{all} \quad \varrho \geqq 0,$$

and for B_ϱ from

$$h_0(t) = 1 - h_0(1-t),$$ [4]

which implies

$$h_0^{(2\varrho+1)}(0) = h_0^{(2\varrho+1)}(1), \quad \text{all} \quad \varrho \geqq 0.$$

The common factor $\sin^2(\pi n/N)$ (which is N-periodic and vanishes at $n=0$) can be transferred from $\omega(n)$ to $\psi_f(n)$, with the result that Theorem 3.3 becomes applicable with $\boldsymbol{N}_0 = \{0\}$. The first few attenuation factors, which follow from (3.16) and (5.15)–(5.18) are listed below:

$$\tau_n = \left(\frac{\sin \pi z}{\pi z}\right)^2 \qquad (r=1), \tag{5.19$_1$}$$

$$\tau_n = \left(\frac{\sin \pi z}{\pi z}\right)^4 [3 - 2\pi z \cot \pi z] \qquad (r=2), \tag{5.19$_2$}$$

$$\tau_n = \left(\frac{\sin \pi z}{\pi z}\right)^6 [25 - 7(\pi z)^2 - 24\pi z \cot \pi z] \qquad (r=3), \tag{5.19$_3$}$$

$$\tau_n = \left(\frac{\sin \pi z}{\pi z}\right)^8 \left[623 - \frac{734}{3}(\pi z)^2 - \left(622\pi z - \frac{116}{3}(\pi z)^3\right)\cot \pi z\right] \qquad (r=4). \tag{5.19$_4$}$$

4 This is most quickly seen by checking that the function on the right-hand side has the same interpolatory properties as $h_0(t)$, viz., $h_0^{(s)}(0) = \delta_{s0}$, $h_0^{(s)}(1) = 0$, $s = 0, 1, \ldots, r-1$.

The first of these is again the attenuation factor for broken line approximants; the second is due to Eagle [7].

Interestingly enough, the expressions in brackets in (5.19_{1-4}) have power series expansions whose initial terms are precisely those given in the brackets of (5.10_{1-4}).

Example 5.4. Given $f \in \boldsymbol{F}$, we now take for φ a generalized periodic spline interpolant corresponding to the linear differential operator

$$L = D^r + a_1 D^{r-1} + \cdots + a_r, \quad D = d/dx, \tag{5.20}$$

where a_ϱ are real constants. This means that

$$\begin{aligned} &\varphi \in C^{2r-2}(-\infty, \infty), \\ &\varphi(x + 2\pi) = \varphi(x), \quad \text{all} \quad x \in \boldsymbol{R}, \\ &\varphi(x_\mu) = f_\mu, \quad \mu = 0, 1, \ldots, N-1, \\ &(L^* L\varphi)(x) = 0, \quad x \neq x_\mu, \end{aligned} \tag{5.21}$$

where L^* denotes the formal adjoint of L,

$$L^* = (-1)^r D^r + (-1)^{r-1} a_1 D^{r-1} + \cdots + a_r. \tag{5.20*}$$

The existence of such a spline interpolant is assured if L has the property that

$$Ly = 0, \quad y(x_\mu) = 0 \quad (\mu = 0, 1, \ldots, N) \quad \text{implies} \quad y \equiv 0$$

(Ahlberg *et al.* [1], p. 199). This is the case, e.g., if $N \geqq r$ and if we assume that the characteristic polynomial of L,

$$\alpha(t) = t^r + a_1 t^{r-1} + \cdots + a_r,$$

has distinct zeros t_ϱ, $\varrho = 1, 2, \ldots, r$, such that

$$t_\varrho - t_\sigma \neq 0 \pmod{iN} \quad \text{for} \quad \varrho \neq \sigma. \tag{5.22}$$

The fundamental solution set $y_\varrho = \exp(t_\varrho x)$, $\varrho = 1, 2, \ldots, r$, of $Ly = 0$ is then indeed unisolvent on any interval. We also assume that none of the nonvanishing zeros t_ϱ is an integer multiple of i. This implies that

$$\lambda(t) = (-1)^r \alpha(t)\alpha(-t) = t^{2r} + l_1 t^{2r-2} + \cdots + l_r, \tag{5.23}$$

which (apart from the sign) is the characteristic polynomial of L^*L, does not vanish at an integer multiple of i, except possibly at zero.

Applying now (5.2) to $\varphi^{(2\varrho)}$, $\varrho = 0, 1, \ldots, r-1$, we get for $n \neq 0$,

$$2\pi c_n(\varphi^{(2\varrho)}) = \frac{1}{(in)^{2r-2\varrho}} \sum_{\mu=0}^{N-1} \delta\varphi_\mu^{(2r-1)} e^{-inx_\mu} + \frac{1}{(in)^{2r-2\varrho}} \int_0^{2\pi} \varphi^{(2r)}(x) e^{-inx} dx,$$
$$\varrho = 0, 1, \ldots, r-1.$$

Using the last relation in (5.21) and the fact that $\lambda(t)$ in (5.23) is the characteristic polynomial of L^*L, we can write

$$c_n(\varphi^{(2\varrho)}) = \frac{1}{(in)^{2r-2\varrho}} (\psi_f(n) - \gamma_n), \quad \varrho = 0, 1, \ldots, r-1, \tag{5.24}$$

where

$$\psi_f(n)=\frac{1}{2\pi}\sum_{\mu=0}^{N-1}\delta\varphi_\mu^{(2r-1)}e^{-inx_\mu}, \tag{5.25}$$

$$\gamma_n=\frac{1}{2\pi}\int_0^{2\pi}[l_1\varphi^{(2r-2)}(x)+\cdots+l_r\varphi(x)]e^{-inx}dx=\sum_{\varrho=0}^{r-1}l_{r-\varrho}c_n(\varphi^{(2\varrho)}). \tag{5.26}$$

Multiplying (5.24) by $l_{r-\varrho}$, $\varrho=0, 1, \ldots, r-1$, and adding up the results, we get

$$\gamma_n=\left[\sum_{\varrho=0}^{r-1}\frac{l_{r-\varrho}}{(in)^{2r-2\varrho}}\right](\psi_f(n)-\gamma_n)=[(in)^{-2r}\lambda(in)-1](\psi_f(n)-\gamma_n),$$

i.e., since by assumption $\lambda(in)\neq 0$,

$$\gamma_n=\left[1-\frac{(in)^{2r}}{\lambda(in)}\right]\psi_f(n).$$

Now (5.24), with $\varrho=0$, gives the desired relation

$$c_n(\varphi)=\omega(n)\psi_f(n), \quad n\neq 0, \tag{5.27}$$

with

$$\omega(n)=\frac{1}{\lambda(in)}. \tag{5.28}$$

Clearly, $\psi_f(n)$ in (5.25) has period N. If $\lambda(0)\neq 0$, Theorem 3.3 applies with the empty set for $\boldsymbol{N}_0$. (Eq. (5.27) then also holds for $n=0$, as can be concluded from (5.27′) below by letting $n\to 0$.) Whether or not this case has any practical merits is questionable, since the interpolation process P in this case does not reproduce the function $f\equiv 1$, since $L^*L\varphi=0$ has no nontrivial constants among its solutions.

If, on the other hand, $\lambda(0)=0$, then $\psi_f(0)=0$, all $f\in\boldsymbol{F}$, as we now proceed to show. We have noted earlier that the result (5.2) holds for arbitrary real n, if one observes the definition (5.3). The preceding derivation, therefore, can be carried through under this more general assumption on n, giving in place of (5.27)

$$c_n(\varphi)=\frac{1}{\lambda(in)}\left\{\frac{1-e^{-2\pi in}}{2\pi in}\left[\lambda(in)F_0-\sum_{\varrho=0}^{r-1}l_{r-\varrho}F_\varrho\right]+\psi_f(n)\right\}, \quad n\,(\text{real})\neq 0, \tag{5.27′}$$

where

$$F_\varrho=\varphi^{(2\varrho)}(0)+\frac{1}{in}\varphi^{(2\varrho+1)}(0)+\cdots+\frac{1}{(in)^{2r-2\varrho-2}}\varphi^{(2r-2)}(0), \quad \varrho=0, 1, \ldots, r-1,$$

and the definition (5.25) of $\psi_f(n)$ is to be adapted in accordance with (5.3). Since for $\varrho=0, 1, \ldots, r-1$,

$$l_{r-\varrho}F_\varrho=l_{r-\varrho}(in)^{2\varrho}F_0+O(n), \quad \text{as} \quad n\to 0,$$

we see that

$$\begin{aligned}\lambda(in)F_0-&\sum_{\varrho=0}^{r-1}l_{r-\varrho}F_\varrho\\ &=\lambda(in)F_0-[\lambda(in)-(in)^{2r}]F_0+O(n)\\ &=(in)^{2r}F_0+O(n)=O(n), \quad \text{as} \quad n\to 0,\end{aligned}$$

proving indeed that $\psi_f(n)\to 0$ as $n\to 0$, all $f\in\boldsymbol{F}$. Thus, Theorem 3.3 applies with $\boldsymbol{N}_0=\{0\}$.

In summary, then,

$$c_n(\varphi) = \tau_n \hat{c}_n(f), \quad \text{all} \quad n \in \boldsymbol{N}, \quad \text{all} \quad f \in \boldsymbol{F}, \tag{5.29}$$

with

$$\tau_n = \frac{1}{\sum_{\nu=-\infty}^{\infty} \frac{\lambda(in)}{\lambda(i(\nu N+n))}} \tag{5.30}$$

for all $n \in \boldsymbol{N}$, if $\lambda(0) \neq 0$, and for all $n \not\equiv 0 \pmod{N}$, if $\lambda(0) = 0$. In the latter case, $\tau_0 = 1$, and $\tau_{\nu N} = 0$, $\nu \neq 0$.

For the "splines in tension", considered by Schweikert [17], we have $L = D(D-\sigma)$, and thus $L^*L = D^4 - \sigma^2 D^2$, i.e., $\varphi(x) = c_1 e^{\sigma x} + c_2 e^{-\sigma x} + c_3 + c_4 x$ on each subinterval $(x_\mu, x_{\mu+1})$. In this case, $\lambda(in) = n^4 + \sigma^2 n^2$.

5.2. Several Attenuation Factors

The following example generalizes constructions and results due to Yuškov [20].

Example 5.5. Let $r \geqq 2$ be an integer, and assume N divisible by r, say, $N = rq$. Letting $p(x) = p_\mu(x)$ be the unique polynomial of degree $\leqq r$ satisfying

$$p(x_{\mu r+s}) = f_{\mu r+s}, \qquad s = 0, 1, \ldots, r, \tag{5.31}$$

we define $\varphi(x) = p_\mu(x)$ on $[x_{\mu r}, x_{(\mu+1)r}]$, $\mu = 0, 1, \ldots, q-1$.

The interpolation process $\varphi = Pf$ of Example 5.5 is clearly linear and r-translation invariant. Hence, by Theorem 4.1,

$$c_n(\varphi) = \sum_{\varrho=0}^{r-1} \tau_{n,\varrho} \hat{c}_{n+\varrho q}(f), \quad \text{all} \quad n \in \boldsymbol{N}, \quad \text{all} \quad f \in \boldsymbol{F}. \tag{5.32}$$

In order to calculate the attenuation factors $\tau_{n,\varrho}$, we denote by

$$l_{r,\sigma}(t) = \prod_{\substack{\varrho=0 \\ \varrho \neq \sigma}}^{r} \left(\frac{t-\varrho}{\sigma-\varrho}\right), \qquad \sigma = 0, 1, \ldots, r$$

the fundamental Lagrange interpolation polynomials belonging to the set of abscissas $\{0, 1, \ldots, r\}$. Then, with $h = 2\pi/N$,

$$\eta_0(x) = \begin{cases} l_{r,0}(t), & x = th, & \text{for} \quad 0 \leqq t \leqq r, \\ l_{r,0}(-t), & x = 2\pi + th, & \text{for} \quad -r \leqq t \leqq 0, \\ 0 & & \text{for} \quad x_r \leqq x \leqq x_{(q-1)r}, \end{cases}$$

so that

$$\begin{aligned} c_n(\eta_0) &= \frac{1}{2\pi} \int_0^{2\pi} \eta_0(x) e^{-inx} dx \\ &= \frac{h}{2\pi} \left\{ \int_0^r l_{r,0}(t) e^{-inht} dt + \int_{-r}^0 l_{r,0}(-t) e^{-in(2\pi+th)} dt \right\} \\ &= \frac{h}{2\pi} \int_0^r l_{r,0}(t) [e^{-inht} + e^{inht}] dt, \end{aligned}$$

i.e.,

$$c_n(\eta_0) = \frac{h}{\pi}\int_0^r l_{r,0}(t)\cos nht\,dt.$$

Similarly, for $0<\sigma\leqq r-1$,

$$\eta_\sigma(x) = \begin{cases} l_{r,\sigma}(t), & x=th, \quad \text{for} \quad 0\leqq t\leqq r, \\ 0 & \text{for} \quad x_r\leqq x\leqq 2\pi, \end{cases}$$

giving

$$c_n(\eta_\sigma) = \frac{h}{2\pi}\int_0^r l_{r,\sigma}(t)e^{-inht}dt, \qquad 0<\sigma\leqq r-1.$$

Therefore, applying (4.5),

$$\tau_{n,\varrho} = \frac{1}{r}\left\{2\int_0^r l_{r,0}(t)\cos nht + \sum_{\sigma=1}^{r-1} e^{i\varrho\sigma 2\pi/r}\int_0^r l_{r,\sigma}(t)e^{-inh(t-\sigma)}dt\right\}. \tag{5.33}$$

Observing that

$$l_{r,\sigma}(t) = l_{r,r-\sigma}(r-t), \qquad \sigma=0,1,\ldots,r, \tag{5.34}$$

we can transform the sum in (5.33) as follows,

$$\begin{aligned} &\sum_{\sigma=1}^{r-1} e^{i\varrho\sigma 2\pi/r}\int_0^r l_{r,\sigma}(t)e^{-inh(t-\sigma)}dt \\ &= \sum_{\sigma=1}^{r-1} e^{i\varrho(r-\sigma)2\pi/r}\int_0^r l_{r,r-\sigma}(t)e^{-inh(t-r+\sigma)}dt \\ &= \sum_{\sigma=1}^{r-1} e^{-i\varrho\sigma 2\pi/r}\int_0^r l_{r,r-\sigma}(r-\tau)e^{-inh(-\tau+\sigma)}d\tau \\ &= \sum_{\sigma=1}^{r-1} e^{-i\varrho\sigma 2\pi/r}\int_0^r l_{r,\sigma}(t)e^{inh(t-\sigma)}dt. \end{aligned}$$

This shows that the sum in (5.33) is real, so that (5.33) simplifies to

$$\tau_{n,\varrho} = \frac{1}{r}\left\{2\int_0^r l_{r,0}(t)\cos nht + \sum_{\sigma=1}^{r-1}\int_0^r l_{r,\sigma}(t)\cos[nh(t-\sigma)-\varrho\sigma 2\pi/r]dt\right\}.$$

Again using (5.34), we can further write

$$\tau_{n,\varrho} = \frac{1}{r}\sum_{\sigma=0}^{r}\int_0^r l_{r,\sigma}(t)\cos[nh(t-\sigma)-\varrho\sigma 2\pi/r]dt,$$

or, since symmetric terms (with indices σ and $r-\sigma$) are equal,

$$\tau_{n,\varrho} = \frac{2}{r}\sum_{\sigma=0}^{(r-1)/2}\int_0^r l_{r,\sigma}(t)\cos[nh(t-\sigma)-\varrho\sigma 2\pi/r]dt \quad \text{if } r \text{ is odd}, \tag{5.35}$$

$$\begin{aligned} \tau_{n,\varrho} = \frac{1}{r}\Bigg\{&(-1)^\varrho\int_0^r l_{r,r/2}(t)\cos[nh(t-r/2)]dt \\ &+2\sum_{\sigma=0}^{(r/2)-1}\int_0^r l_{r,\sigma}(t)\cos[nh(t-\sigma)-\varrho\sigma 2\pi/r]dt\Bigg\} \quad \text{if } r \text{ is even}. \end{aligned} \tag{5.35'}$$

For $r=2$, e.g., one obtains from (5.35′) by an elementary computation,

$$(5.35_2') \qquad \tau_{n,\varrho}=\left(\frac{\sin \pi z}{\pi z}\right)^3 [\cos(\pi z+\varrho\pi/2)+\pi z \sin(\pi z+\varrho\pi/2)] \quad (r=2;\ \varrho=0,1),$$

where, as before, $z=n/N$. Similarly, (5.35) for $r=3$ gives

$$\begin{aligned}(5.35_3) \qquad \tau_{n,\varrho} &= -\frac{1}{36(\pi z)^2}\left[2\cos 6\pi z-9\cos(4\pi z-\varrho 2\pi/3)\right.\\ &\quad \left.+18\cos(2\pi z+\varrho 2\pi/3)-11\right]\\ &\quad +\frac{1}{12(\pi z)^3}\left[\sin 6\pi z-4\sin(4\pi z-\varrho 2\pi/3)+5\sin(2\pi z+\varrho 2\pi/3)\right]\\ &\quad +\frac{1}{24(\pi z)^4}\left[\cos 6\pi z-3\cos(4\pi z-\varrho 2\pi/3)\right.\\ &\quad \left.+3\cos(2\pi z+\varrho 2\pi/3)-1\right] \quad (r=3;\ \varrho=0,1,2).\end{aligned}$$

Both results $(5.35_2')$ and (5.35_3), in a somewhat different form, were already obtained by Yuškov [20].

The case $r=N=8$ is considered by Salzer [15] who has numerical tables for $0\leqq n\leqq 24$.

5.3. *Attenuation Factors Associated with Derivatives*

Example 5.6. Let k and r be integers with $1\leqq k\leqq r$. Define φ as follows:

$$\begin{aligned}&\varphi\in C^{2r-1-k}(-\infty,\infty),\\ &\varphi(x+2\pi)=\varphi(x), \quad \text{all} \quad x\in\boldsymbol{R},\\ &\varphi^{(\varkappa)}(x_\mu)=f_\mu^{(\varkappa)}, \quad \varkappa=0,1,\ldots,k-1;\ \ \mu=0,1,\ldots,N-1,\\ &\varphi^{(2r)}(x)=0, \quad x\neq x_\mu.\end{aligned} \tag{5.36}$$

In the terminology of spline functions, φ is a periodic spline interpolant of degree $2r-1$ and deficiency k. It exists and is uniquely determined (Ahlberg *et al.* [1, pp. 167–168]). The special case $k=1$ gives ordinary splines, considered in Example 5.1. If $k=r$, we are dealing with Hermite interpolation of order $r-1$ on each subinterval $[x_\mu, x_{\mu+1}]$.

Applying (5.2) to $\varphi^{(s)}$, $s=0,1,\ldots,k-1$, we get for $n\neq 0$,

$$(5.37) \quad c_n(\varphi^{(s)})=\frac{1}{2\pi}\sum_{\varkappa=0}^{k-1}\frac{1}{(in)^{2r+1+\varkappa-s-k}}\sum_{\mu=0}^{N-1}\delta\varphi_\mu^{(2r+\varkappa-k)}e^{-in x_\mu}, \quad s=0,1,\ldots,k-1.$$

For $n\neq 0 \pmod N$ one computes

$$\sum_{\nu=-\infty}^{\infty}\frac{1}{[i(\nu N+n)]^{k+1}}=\frac{1}{(in)^{k+1}}\sigma_k(z), \quad z=\frac{n}{N}, \quad k=0,1,2,\ldots.$$

Moreover, by construction,

$$\hat{c}_n(\varphi^{(s)})=\hat{c}_n(f^{(s)}), \quad s=0,1,\ldots,k-1.$$

Applying (3.14) to $\varphi^{(s)}$, we thus obtain

$$\hat{c}_n(f^{(s)}) = \sum_{\nu=-\infty}^{\infty} c_{\nu N+n}(\varphi^{(s)})$$

$$= \frac{1}{2\pi} \sum_{\varkappa=0}^{k-1} \frac{\sigma_{2r+\varkappa-s-k}(z)}{(in)^{2r+1+\varkappa-s-k}} \sum_{\mu=0}^{N-1} \delta\varphi_\mu^{(2r+\varkappa-k)} e^{-inx_\mu}, \quad s=0, 1, \ldots, k-1. \tag{5.38}$$

Defining

$$d_\varkappa = \frac{1}{2\pi} \frac{1}{(in)^{2r+1+\varkappa-k}} \sum_{\mu=0}^{N-1} \delta\varphi_\mu^{(2r+\varkappa-k)} e^{-inx_\mu}, \quad \varkappa=0, 1, \ldots, k-1,$$

we can write

$$c_n(\varphi) = \sum_{\varkappa=0}^{k-1} d_\varkappa, \tag{5.39}$$

by virtue of (5.37) with $s=0$. On the other hand, (5.38) represents a system of k linear equations in the k unknowns $d_\varkappa$,

$$\sum_{\varkappa=0}^{k-1} (in)^s \sigma_{2r+\varkappa-s-k}(z)\, d_\varkappa = \hat{c}_n(f^{(s)}), \quad s=0, 1, \ldots, k-1. \tag{5.40}$$

Inverting this system, and substituting the result in (5.39) gives $c_n(\varphi)$ as a linear combination of the $\hat{c}_n(f^{(s)})$,

$$c_n(\varphi) = \sum_{s=0}^{k-1} \frac{\tau_{n,s}}{(in)^s} \hat{c}_n(f^{(s)}), \quad n \not\equiv 0 \pmod{N}. \tag{5.41}$$

The coefficient matrix of the system (5.40) is given by

$$A = \begin{bmatrix} 1 & & 0 \\ & in & \\ & \ddots & \\ 0 & & (in)^{k-1} \end{bmatrix} \begin{bmatrix} \sigma_{2r-k} & \sigma_{2r-k+1} & \cdots & \sigma_{2r-1} \\ \sigma_{2r-k-1} & \sigma_{2r-k} & \cdots & \sigma_{2r-2} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{2r-2k+1} & \sigma_{2r-2k+2} & \cdots & \sigma_{2r-k} \end{bmatrix}.$$

Except for the ordering of the rows, the second factor of A is identical with the matrix $H_{s,p}$ of Proposition 2.3, if we define $s=2r-2k+1$ and $p=k$. Since $s \geqq 1$ is odd, and $p \geqq 1$, it follows from (2.4) that $\det A \neq 0$ for $n \not\equiv 0 \pmod{N}$. One checks easily that the attenuation factors $\tau_{n,s}$ in (5.41), $s=0, 1, \ldots, k-1$, are just the column sums of $H_{2r-2k+1,k}^{-1}(z)$, taken in the order from right to left.

A slightly more explicit expression for $\tau_{n,s}$ may be obtained from Proposition 2.2. In fact, if we let

$$Q(z) = \begin{bmatrix} q_{2r-k-1} & q_{2r-k} & \cdots & q_{2r-2} \\ q_{2r-k-2} & q_{2r-k-1} & \cdots & q_{2r-3} \\ \cdots & \cdots & \cdots & \cdots \\ q_{2r-2k} & q_{2r-2k+1} & \cdots & q_{2r-k-1} \end{bmatrix}, \quad Q^{-1}(z) = [\omega_{\varkappa s}]_{\varkappa,s=0}^{k-1},$$

where $q_s = q_s(\cos \pi z)$, then an elementary calculation yields

$$\tau_{n,s} = \sum_{\varkappa=0}^{k-1} \left(\frac{\sin \pi z}{\pi z}\right)^{2r+\varkappa-k-s+1} \omega_{\varkappa s}(\cos \pi z), \quad s=0, 1, \ldots, k-1. \tag{5.42}$$

For $r=k=2$, e.g., one finds

$$\begin{aligned}\tau_{n,0} &= -\frac{3\sin\pi z}{(\pi z)^3}\left(\cos\pi z - \frac{\sin\pi z}{\pi z}\right),\\ \tau_{n,1} &= \frac{3}{(\pi z)^2}\left(\frac{2}{3}\cos^2\pi z + \frac{1}{3} - \frac{\sin\pi z}{\pi z}\cos\pi z\right),\end{aligned} \qquad (r=k=2). \tag{5.43}$$

This corresponds to cubic Hermite interpolation on each subinterval. The procedure has already been discussed by Serebrennikov [18], who expresses the result in the form of an additive correction term to Dällenbach's result, but does not note the connection with attenuation factors.

The following limiting relations are worth noting,

$$\begin{aligned}&\tau_{n,0}\to 1, \quad (in)^{-1}\tau_{n,1}\to 0 \quad \text{as} \quad n\to 0,\\ &\tau_{n,0}\to 0, \quad (in)^{-1}\tau_{n,1}\to \frac{3}{(\pi\nu)^2}(i\nu N)^{-1} \quad \text{as} \quad n\to \nu N, \quad \nu\neq 0.\end{aligned} \tag{5.44}$$

It can be verified that with these limiting values Eq. (5.41) (for $r=k=2$) also holds true when $n=0 \pmod N$.

For $r=3$, $k=2$, Eq. (5.42) gives

$$\begin{aligned}\tau_{n,0} &= -\frac{45\sin\pi z}{(\pi z)^5(1+\sin^2\pi z)}\left[q_3(\cos\pi z) - \frac{\sin\pi z}{\pi z}q_2(\cos\pi z)\right],\\ \tau_{n,1} &= \frac{45}{(\pi z)^4(1+\sin^2\pi z)}\left[q_4(\cos\pi z) - \frac{\sin\pi z}{\pi z}q_3(\cos\pi z)\right],\end{aligned} \qquad (r=3, k=2), \tag{5.45}$$

where again limit relations similar to those in (5.44) are valid.

For $r=k=3$, finally,

$$\begin{aligned}\tau_{n,0} &= -\frac{15}{(\pi z)^4}\left[\sin^2\pi z + 3\frac{\sin\pi z}{\pi z}\cos\pi z - 3\left(\frac{\sin\pi z}{\pi z}\right)^2\right],\\ \tau_{n,1} &= \frac{3}{(\pi z)^3}\left[4\sin\pi z\cos\pi z + \frac{1}{\pi z}(1+14\cos^2\pi z) - \frac{15}{(\pi z)^2}\sin\pi z\cos\pi z\right],\\ \tau_{n,2} &= \frac{3}{(\pi z)^2}\left[1+\sin^2\pi z + 4\frac{\sin\pi z}{\pi z}\cos\pi z - 5\left(\frac{\sin\pi z}{\pi z}\right)^2\right].\end{aligned} \tag{5.46}$$

One checks that

$$\tau_{n,0}\to 1, \quad (in)^{-1}\tau_{n,1}\to 0, \quad (in)^{-2}\tau_{n,2}\to \frac{1}{15}\left(\frac{\pi}{N}\right)^2 \quad \text{as} \quad n\to 0,$$

which is in agreement with the relation

$$c_0(\varphi) = \hat{c}_0(f) + \frac{1}{60}h^2\hat{c}_0(f'')$$

obtained by 5-th degree Hermite interpolation.

Acknowledgment. The author is greatly indebted to Dr. Christian H. Reinsch for many valuable discussions which helped clarify and simplify the exposition at several places. In particular, the present version of Theorem 3.1, as well as Theorem 3.2, are due to Dr. Reinsch. He clearly recognized the role of translation invariance for the existence of attenuation factors, which the author had expressed only implicitly in a preliminary version of Theorem 3.1.

References

1. Ahlberg, J. H., Nilson, E. N., Walsh, J. L.: The theory of splines and their application. New York-London: Academic Press 1967.
2. Bauer, F. L., Stetter, H. J.: Zur numerischen Fourier-Transformation. Numer. Math. **1**, 208–220 (1959).
3. Chao, F. H.: A new method of practical harmonic analysis [Chinese]. Acta Math. Sinica **6**, 433–451 (1956).
4. Cooley, J. W., Tukey, J. W.: An algorithm for the machine calculation of complex Fourier series. Math. Comp. **19**, 297–301 (1965).
5. — Lewis, P. A. W., Welch, P. D.: The fast Fourier transform and its applications. IEEE Trans. Education E-**12**, 27–34 (1969).
6. Dällenbach, W.: Verschärftes rechnerisches Verfahren der harmonischen Analyse. Arch. Elektrotechnik **10**, 277–282 (1921).
7. Eagle, A.: On the relations between the Fourier constants of a periodic function and the coefficients determined by harmonic analysis. Philos. Mag. **5** (**7**), 113–132 (1928).
8. Ehlich, H.: Untersuchungen zur numerischen Fourieranalyse. Math. Z. **91**, 380–420 (1966).
9. Gentleman, W. M., Sande, G.: Fast Fourier transforms—for fun and profit, 1966 Fall Joint Computer Conference, AFIPS Proc., vol. 29. Washington D. C.: Spartan 1966.
10. Golomb, M.: Approximation by periodic spline interpolants on uniform meshes. J. Approximation Theory **1**, 26–65 (1968).
11. Hildebrand, F. B.: Introduction to numerical analysis. New York: McGraw-Hill 1956.
12. Oumoff, N.: Sur l'application de la méthode de Mr. Ludimar Hermann à l'analyse des courbes périodiques. Le Physiologiste Russe **1**, 52–64 (1898/99).
13. Quade, W., Collatz, L.: Zur Interpolationstheorie der reellen periodischen Funktionen. Sitzungsber. Preuss. Akad. Wiss. **30**, 383–429 (1938).
14. Runge, C.: Theorie und Praxis der Reihen. Leipzig: G. J. Göschen'sche Verlagshandlung 1904.
15. Salzer, H. E.: Formulas for calculating Fourier coefficients. J. Math. Phys. **36**, 96–98 (1957).
16. Schoenberg, I. J.: On spline interpolation at all integer points of the real axis. Colloquium on the Theory of Approximation of Functions (Cluj, 1967). Mathematica (Cluj) **10** (**33**), 151–170 (1968).
17. Schweikert, D. G.: An interpolation curve using a spline in tension. J. Math. and Phys. **45**, 312–317 (1966).
18. Serebrennikov, M. G.: A more exact method of harmonic analysis of empirical periodic curves [Russian]. Akad. Nauk SSSR. Prikl. Mat. Meh. **12**, 227–232 (1948).
19. Yuškov, P. P.: The practical harmonic analysis of empirical functions when the given curve is replaced by another approximating the given one by tracing [Russian]. Akad. Nauk SSSR. Inž. Sbornik **6**, 197–210 (1950).
20. — On the correction of the coefficients obtained in the usual practical harmonic analysis [Russian]. Akad. Nauk SSSR. Inž. Sbornik **10**, 213–222 (1951).

Prof. Dr. Walter Gautschi
Dept. of Computer Sciences
Purdue University
Lafayette, Indiana 47907/USA

10.2. [86] "ON PADÉ APPROXIMANTS ASSOCIATED WITH HAMBURGER SERIES"

---

[86] "On Padé Approximants Associated with Hamburger Series," *Calcolo* **20**, 111–127 (1983).

ON PADÉ APPROXIMANTS ASSOCIATED WITH HAMBURGER SERIES (1)

Dedicated to Professor Aldo Ghizzetti on his 75th birthday

W. GAUTSCHI (2)

ABSTRACT - We discuss three (only loosely connected) aspects of Padé approximants associated with Hamburger series: (i) Normality criteria, expressed in terms of orthogonal polynomials (ii) Inequalities for expansion coefficients (iii) Computational methods.

1. Introduction.

A formal power series

$$f(z)=\mu_0+\mu_1 z+\mu_2 z^2+ \ldots \tag{1.1}$$

is called a *Hamburger series* if its coefficients are moments

$$\mu_k=\int_{\mathbf{R}} t^k \, d\lambda(t) \tag{1.2}$$

of a bounded nondecreasing function $\lambda(t)$ having infinitely many points of increase; (1.1) is called a *Stieltjes series* if (1.2) holds with a distribution $d\lambda$ supported on the nonnegative real axis $\mathbf{R}_+$. There is a well-known connection between Padé approximants associated with a Stieltjes or Hamburger series and polynomials orthogonal with respect to the distribution $d\lambda$. (For a recent exposition, emphasizing this point of view, see [1]). The Padé approximants on the first subdiagonal of the Padé table (3), indeed, are

---

(1) Sponsored in part by the National Science Foundation under grant MCS-7927158A1.

(2) Department of Computer Sciences, Purdue University, West Lafayette, U.S.A..

(3) We arrange the Padé table so that fractions with constant denominator degree appear in the same row.

$$f[n-1,n](z)=\sum_{\nu=1}^{n}\frac{\lambda_\nu^{(n)}}{1-\tau_\nu^{(n)}z},\quad n=1,2,3,\dots, \tag{1.3$_0$}$$

where $\tau_\nu^{(n)}$ are the zeros of the n-th degree orthogonal polynomial $\pi_n(t;d\lambda)$ and $\lambda_\nu^{(n)}$ the corresponding Christoffel numbers. More generally, given any integer $j\geq 0$, we have for Stieltjes series, and also for Hamburger series if j is even,

$$f[n-1+j,n](z)=\mu_0+\mu_1 z+\dots+\mu_{j-1}z^{j-1}+z^j\sum_{\nu=1}^{n}\frac{\lambda_{\nu,j}^{(n)}}{1-\tau_{\nu,j}^{(n)}z}, \tag{1.3$_j$}$$

where $\tau_{\nu,j}^{(n)}$ are the zeros of $\pi_{n,j}(\cdot)=\pi_n(\cdot;d\lambda_j)$ and $\lambda_{\nu,j}^{(n)}$ the associated Christoffel numbers, $d\lambda_j$ now being the measure

$$d\lambda_j(t)=t^j\,d\lambda(t). \tag{1.4}$$

The integer n in (1.3$_j$) may be any nonnegative integer, if $j>0$, assuming the usual convention that empty sums are zero. Since for Stieltjes series the support of $d\lambda$ is in $\mathbf{R}_+$, the measure $d\lambda_j$ is positive definite, hence defines a unique set of (monic) orthogonal polynomials $\pi_{n,j}$, $n=0,1,2,\dots$(cf., e. g., [7, § 2.2]). The same is true for Hamburger series, if j is even.

Padé tables associated with Stieltjes series are known to be always normal. This, of course, is no longer true for Hamburger series, an extreme example being furnished by a symmetric distribution $d\lambda$ on a symmetric interval, in which case the formal power series (1.1) proceeds in even powers of z and no entry of the Padé table is normal. (The Padé table, however, is seminormal in a sense defined by Gragg; see [9, p. 16 and Theorem 7.3]). In Section 2 we formulate and prove a condition, in terms of orthogonal polynomials, for all entries $f[n-1+j,n]$, $n\geq 1$, $j\geq 0$, of the Padé table associated with a Hamburger series to be normal (Corollary to Theorem 2.1).

Expanding the rational part of (1.3$_j$) in powers of z, one gets

$$f[n-1+j,n](z)=\mu_0+\mu_1 z+\dots+\mu_{2n-1+j}z^{2n-1+j}+\sum_{k=2n}^{\infty}\mu_{k,j}^{(n)}z^{k+j}, \tag{1.5}$$

where

$$\mu_{k,j}^{(n)}=\sum_{\nu=1}^{n}\lambda_{\nu,j}^{(n)}[\tau_{\nu,j}^{(n)}]^k,\quad k=0,1,2,\dots. \tag{1.6}$$

We have used the fact that

$$\mu_{k,j}^{(n)}=\mu_{k+j}\ \text{if}\ 0\leq k<2n,\ j\geq 0. \tag{1.7}$$

The expansion coefficients $\mu_{k,j}^{(n)}$, when (1.1) is a Stieltjes series, satisfy the following interesting inequalities,

$$\mu_{k,j}^{(n+1)}-\mu_{k,j}^{(n)}>0 \text{ for all } k\geq 2n, \quad n=1,2,3,\ldots, \quad j=0,1,2,\ldots; \tag{1.8}$$

see the proof of Theorem 5.2.7 in Baker & Graves-Morris [2]. This means that the coefficients of some fixed power z^{k+j} in the expansion of the Padé approximants down a diagonal $[n-1+j, n]$, $n=1,2,\ldots, j\geq 0$, are monotonically increasing until they become, and stay, equal to an exact moment.

We show in Section 3 (see Corollary to Theorem 3.2) that (1.8) remains true for Padé approximants associated with a large class of Hamburger series, provided j is restricted to an even integer. Our tool, which produces the assertion rather quickly, is a recent result of Hunter [10] on orthogonal polynomials, for which we also give a slightly simplified proof.

Finally, in Section 4, we discuss several methods, all based on the representation (1.3_{2j}), of computing Padé approximants $f[n-1+2j, n]$. Numerical stability being of particular concern to us, we avoid methods that depart from the moments. We assume, instead, that we are given the measure $d\lambda(t)$, or equivalently, the recursion coefficients of the associated orthogonal polynomials, and use these as input to our procedures for generating the moments μ_k and Gauss-Christoffel data $\tau_{\nu,2j}^{(n)}$, $\lambda_{\nu,2j}^{(n)}$ required in (1.3_{2j}).

2. Normality criteria.

An entry of the Padé table is said to be *normal* if the same entry does not occur in any other location of the Padé table. The Padé table is normal if each of its entries is normal. Every Padé table associated with a Stieltjes series is known to be normal [13, p. 390]. We now formulate necessary and sufficient conditions for normality in the case of a Padé table associated with a Hamburger series.

The measure $d\lambda_j(t)$ of (1.4), when $j\geq 0$ is even, gives rise to (monic) orthogonal polynomials $\pi_{n,j}$ and polynomials of the second kind,

$$\sigma_{n,j}(z)=\int_{\mathbf{R}}\frac{\pi_{n,j}(z)-\pi_{n,j}(t)}{z-t}\,d\lambda_j(t), \tag{2.1}$$

where $\pi_{n,j}$ and $\sigma_{n,j}$ are of exact degree n and $n-1$, respectively. We also need the function

$$\rho_{n,j}(z)=\int_{\mathbf{R}}\frac{\pi_{n,j}(t)}{z-t}\,d\lambda_j(t), \tag{2.2}$$

where z is assumed outside the support of $d\lambda_j$. We write σ_n and ρ_n for $\sigma_{n,0}$ and $\rho_{n,0}$. Clearly,

$$\pi_{n,j}(z)\int_{\mathrm{R}}\frac{d\lambda_j(t)}{z-t}=\sigma_{n,j}(z)+\rho_{n,j}(z). \tag{2.3}$$

Also, as is well-known (see, e. g., [4, § 1.4]),

$$\frac{\sigma_{n,j}(z)}{\pi_{n,j}(z)}=\sum_{\nu=1}^{n}\frac{\lambda_{\nu,j}^{(n)}}{z-\tau_{\nu,j}^{(n)}}, \tag{2.4}$$

where $\tau_{\nu,j}^{(n)}$ are the zeros of $\pi_{n,j}$ and $\lambda_{\nu,j}^{(n)}$ the corresponding Christoffel numbers. Note that the limit $\rho_{n,j}(0)=\lim_{z\to 0}\rho_{n,j}(z)$, if $j>0$, formally exists, since by (2.3) and (2.4),

$$\begin{aligned}-\rho_{n,j}(0)&=\pi_{n,j}(0)\int_{\mathrm{R}}\frac{d\lambda_j(t)}{t}+\sigma_{n,j}(0)\\ &=\pi_{n,j}(0)\left\{\mu_{j-1}-\sum_{\nu=1}^{n}\frac{\lambda_{\nu,j}^{(n)}}{\tau_{\nu,j}^{(n)}}\right\},\end{aligned} \tag{2.5}$$

if $\pi_{n,j}(0)\neq 0$, and similarly, with an expression involving $\pi'_{n,j}(0)$, if $\pi_{n,j}(0)=0$.

THEOREM 2.1. *The entry* $f[n-1+j,n]$, $n\geq 1$, j (*even*)≥ 0, *in the Padé table associated with the Hamburger series* (1.1), (1.2) *is normal if and only if*

$$\pi_n(0)\,\sigma_n(0)\neq 0 \text{ in the case } j=0, \tag{2.6}$$

and

$$\pi_{n,j}(0)\,\rho_{n,j}(0)\neq 0 \text{ in the case } j>0. \tag{2.7}$$

Proof. Assume first $j>0$. Then, by (1.3_j) and (2.4),

$$f[n-1+j,n](z)=\frac{\pi_{n,j}^*(z)[\mu_0+\mu_1 z+\ldots+\mu_{j-1}z^{j-1}]+z^j\,\sigma_{n,j}^*(z)}{\pi_{n,j}^*(z)}, \tag{2.8_j}$$

where $\pi_{n,j}^*(z)=z^n\,\pi_{n,j}(1/z)$ and $\sigma_{n,j}^*(z)=z^{n-1}\,\sigma_{n,j}(1/z)$. From (2.4) it can be seen that the zeros of $\sigma_{n,j}$ are real and alternate with the zeros of $\pi_{n,j}$ (if $n>1$). It follows that $\pi_{n,j}$ and $\sigma_{n,j}$ have no common zeros, hence neither do $\pi_{n,j}^*$ and $\sigma_{n,j}^*$. Consequently, the fraction in (2.8_j) is irreducible, if $n>1$. The same is trivially true (since $\sigma_{1,j}^*(z)=\mu_j>0$) if $n=1$. A well-known theorem [12, Satz 5.3] then tells us that the entry (2.8_j) is normal if and only if its numerator and

denominator polynomials are of exact degree $n-1+j$ and n, respectively, and $\mu_{2n+j}-\mu_{2n,j}^{(n)} \neq 0$ [cf. (1.5)].

The last condition is always satisfied, since Markov's formula for the remainder term $R_n(\cdot)$ in Gaussian integration (with measure $d\lambda_j$) yields

$$\mu_{2n+j}-\mu_{2n,j}^{(n)}=R_n(t^{2n})=\int_{\mathbf{R}} \pi^2_{n,j}(t)\, d\lambda_j(t)>0. \tag{2.9}$$

Now the coefficient of the power z^n in $\pi_{n,j}^*$ is $\pi_{n,j}(0)$, while the coefficient of the power z^{n-1+j} in the numerator polynomial of (2.8_j) is $\pi_{n,j}(0)\,\mu_{j-1}+\sigma_{n,j}(0)$, or, by (2.3),

$$\pi_{n,j}(0)\,\mu_{j-1}-\pi_{n,j}(0)\int_{\mathbf{R}} \frac{d\lambda_j(t)}{t}-\rho_{n,j}(0)=-\rho_{n,j}(0).$$

Therefore, the degrees of the numerator and denominator polynomials in (2.8_j) are exactly $n-1+j$ and n, respectively, if and only if $\pi_{n,j}(0)\,\rho_{n,j}(0)\neq 0$. This proves (2.7).

The proof in the case $j=0$ follows similarly from

$$f[n-1,n]=\frac{\sigma_n^*(z)}{\pi_n^*(z)}, \tag{2.8_0}$$

the degree condition (for the numerator polynomial) now reading $\sigma_n(0)\neq 0$ in place of $\rho_{n,j}(0)\neq 0$. □

REMARKS.

1. Theorem 2.1 holds also for $n=0$, $j>0$, if the condition in (2.7) is replaced by $\mu_{j-1}\neq 0$.

2. By virtue of (2.4), (2.5), the conditions (2.6) and (2.7) can also be written in the form

$$\pi_n(0)\sum_{\nu=1}^{n}\frac{\lambda_\nu^{(n)}}{\tau_\nu^{(n)}}\neq 0, \quad j=0, \tag{2.6'}$$

$$\pi_{n,j}(0)\left\{\mu_{j-1}-\sum_{\nu=1}^{n}\frac{\lambda_{\nu,j}^{(n)}}{\tau_{\nu,j}^{(n)}}\right\}\neq 0, \quad j\,(\text{even})>0. \tag{2.7'}$$

3. The condition $\pi_{n,j}(0)\neq 0$, $j\,(\text{even})\geq 0$, implies the existence of $\pi_{n,j+1}$; indeed,

$$\pi_{n,j+1}(t)=\frac{1}{t}\left[\pi_{n+1,j}(t)-\frac{\pi_{n+1,j}(0)}{\pi_{n,j}(0)}\pi_{n,j}(t)\right].$$

The representation (1.3_{j+1}) therefore holds for $f[n+j,n]$, j (even) ≥ 0, provided $\pi_{n,j+1}$ has n distinct zeros.

4. The usual criterion for normality of $f[n-1+j,n]$, j (even) ≥ 0, in terms of determinants, is [12, Satz 5.4]

$$\Delta_{n-1,j}\,\Delta_{n,j}\,\Delta_{n-1,j+1}\,\Delta_{n,j-1}\neq 0, \tag{2.10}$$

where

$$\Delta_{n,k}=\det\begin{bmatrix}\mu_k & \mu_{k+1} & \cdots & \mu_{k+n}\\ \mu_{k+1} & \mu_{k+2} & \cdots & \mu_{k+n+1}\\ \cdots & \cdots & \cdots & \cdots\\ \mu_{k+n} & \mu_{k+n+1} & \cdots & \mu_{k+2n}\end{bmatrix},\quad k\geq -1 \tag{2.11}$$

(with the understanding that $\mu_{-1}=0$). The conditions implied by (2.10) can easily be recovered from (2.6), (2.7). It suffices, first of all, to show $\Delta_{n-1,j+1}\,\Delta_{n,j-1}\neq 0$, since from the theory of the Hamburger moment problem, $\Delta_{n-1,j}>0$, $\Delta_{n,j}>0$ if $j\geq 0$ is even (cf., e. g., [13, p. 325]). The representation of orthogonal polynomials in determinant form,

$$\pi_{n,j}(z)=\frac{1}{\Delta_{n-1,j}}\det\begin{bmatrix}\mu_j & \cdots & \mu_{j+n-1} & 1\\ \mu_{j+1} & \cdots & \mu_{j+n} & z\\ \cdots & \cdots & \cdots & \cdots\\ \mu_{j+n} & \cdots & \mu_{j+2n-1} & z^n\end{bmatrix} \tag{2.12}$$

(see, e. g., [7, Eq. (2.2.7)]), yields

$$(-1)^n\,\pi_{n,j}(0)=\Delta_{n-1,j+1}/\Delta_{n-1,j},\quad j\geq 0,$$

and, for $j>0$,

$$-\rho_{n,j}(0)=\int_{\mathbf{R}}\pi_{n,j}(t)\,t^{j-1}\,d\lambda(t)$$

$$=\frac{1}{\Delta_{n-1,j}}\int_{\mathbf{R}}\det\begin{bmatrix}\mu_j & \cdots & \mu_{j+n-1} & t^{j-1}\\ \mu_{j+1} & \cdots & \mu_{j+n} & t^j\\ \cdots & \cdots & \cdots & \cdots\\ \mu_{j+n} & \cdots & \mu_{j+2n-1} & t^{j+n-1}\end{bmatrix}d\lambda(t)=(-1)^n\,\Delta_{n,j-1}/\Delta_{n-1,j}.$$

The condition $\pi_{n,j}(0)\neq 0$ thus is equivalent to $\Delta_{n-1,j+1}\neq 0$, while $\rho_{n,j}(0)\neq 0$ (for $j>0$) is equivalent to $\Delta_{n,j-1}\neq 0$. Finally, if $j=0$,

$$\sigma_n(0)=\int_{\mathbf{R}} \frac{\pi_n(t)-\pi_n(0)}{t}\, d\lambda(t)=$$

$$=\frac{1}{\Delta_{n-1,0}}\int_{\mathbf{R}} \frac{1}{t}\left\{\det\begin{bmatrix}\mu_0 & \cdots & \mu_{n-1} & 1\\ \mu_1 & \cdots & \mu_n & t\\ \cdots & \cdots & \cdots & \cdots\\ \mu_n & \cdots & \mu_{2n-1} & t^n\end{bmatrix}-\det\begin{bmatrix}\mu_0 & \cdots & \mu_{n-1} & 1\\ \mu_1 & \cdots & \mu_n & 0\\ \cdots & \cdots & \cdots & \cdots\\ \mu_n & \cdots & \mu_{2n-1} & 0\end{bmatrix}\right\} d\lambda(t)$$

$$=\frac{1}{\Delta_{n-1,0}}\int_{\mathbf{R}} \det\begin{bmatrix}\mu_0 & \cdots & \mu_{n-1} & 0\\ \mu_1 & \cdots & \mu_n & 1\\ \cdots & \cdots & \cdots & \cdots\\ \mu_n & \cdots & \mu_{2n-1} & t^{n-1}\end{bmatrix} d\lambda(t)$$

$$=(-1)^n \Delta_{n,-1}/\Delta_{n-1,0},$$

so that $\sigma_n(0)\neq 0$ is equivalent to $\Delta_{n,-1}\neq 0$.

As a consequence of Remark 4, note that $\sigma_n(0)\neq 0$, all $n\geq 1$, is equivalent to $\Delta_{n,-1}\neq 0$, all $n\geq 1$, while $\pi_{n,j}(0)\neq 0$, all $n\geq 1$, all j (even)≥ 0, is equivalent to $\Delta_{n,k}\neq 0$, all $n\geq 0$, all k (odd)≥ 1. Since $\Delta_{n,j}>0$ whenever j is even, it follows that the conditions

(2.13) $$\sigma_n(0)\neq 0, \quad \text{all } n\geq 1,$$

and

(2.14) $$\pi_{n,j}(0)\neq 0, \quad \text{all } n\geq 1, \quad \text{all } j \text{ (even)}\geq 0,$$

together are sufficient, and also necessary, for (2.10) to hold for all $n\geq 1$ and all $j\geq 0$. This establishes:

COROLLARY TO THEOREM 2.1. *Every entry $f[n-1+j, n]$, $n\geq 1$, in the Padé table associated with the Hamburger series* (1.1), (1.2), *regardless of whether $j\geq 0$ is even or odd, is normal if and only if* (2.13) *and* (2.14) *hold.*

3. Inequalities for Padé expansion coefficients.

We assume in this section that

(3.1) $$d\lambda(t)=w(t)\,dt,$$

where $w(t)$ is a nonnegative weight function on a symmetric interval I: $-\infty \leq -a \leq t \leq a \leq \infty$, $a>0$, continuous on the open interval $(-a, a)$, and such that all moments μ_k in (1.2) exist and $\mu_0>0$. We denote the associated (monic) orthogonal polynomial of degree n by $\pi_n(\cdot)=\pi_n(\cdot; d\lambda)$, its zeros by $\tau_\nu^{(n)}$, and the Christoffel numbers by $\lambda_\nu^{(n)}$. Note that $z^n \pi_n(1/z)$ is a polynomial of degree $\leq n$, equal to 1 at $z=0$. Let

$$\frac{1}{z^n \pi_n(1/z)} = c_0^{(n)} + c_1^{(n)} z + c_2^{(n)} z^2 + \dots, \quad c_0^{(n)} = 1, \tag{3.2}$$

be the Maclaurin expansion of its reciprocal.

THEOREM 3.1. (*Hunter* [10]) (*a*) *If* $w(t)/w(-t)$ *is strictly increasing on* I, *then*

$$c_k^{(n)} > 0, \quad k=0, 1, 2, \dots; \quad n=1, 2, 3\dots. \tag{3.3}$$

(*b*) *If* $w(t)=w(-t)$ *on* I, *and* $n \geq 2$, *then*

$$c_k^{(n)} > 0 \text{ if } k \text{ is even}, \quad c_k^{(n)} = 0 \text{ if } k \text{ is odd}. \tag{3.4}$$

(*c*) *If* $w(t)/w(-t)$ *is strictly decreasing on* I, *then*

$$(-1)^k c_k^{(n)} > 0, \quad k=0, 1, 2, \dots; \quad n=1, 2, 3, \dots, \tag{3.5}$$

PROOF (cf. Hunter [10]). Let

$$w(t, \sigma) = \sigma w(t) + (1-\sigma) w(-t), \quad 0 \leq \sigma \leq 1,$$

and let $\tau_1(\sigma) > \tau_2(\sigma) > \dots > \tau_n(\sigma)$ be the zeros of $\pi_n(\cdot; w(t, \sigma)\, dt)$. Consider

$$\frac{1}{z^n \pi_n\left[\frac{1}{z}; w(t, \sigma)\, dt\right]} = \prod_{\nu=1}^{n} \frac{1}{1-\tau_\nu(\sigma) z}$$

$$= q(z) \prod_{\nu=1}^{[n/2]} \left\{ \frac{1}{[1-\tau_\nu(\sigma) z]\,[1-\tau_{n+1-\nu}(\sigma) z]} \right\}, \tag{3.6}$$

where

$$q(z) = \begin{cases} 1, & n \text{ even}, \\ \dfrac{1}{1-\tau_{(n+1)/2}(\sigma) z}, & n \text{ odd}. \end{cases}$$

(*a*) We prove that $q(z)$ (if n is odd) and each product in curled brackets in (3.6) (if $n \geq 2$) has a Maclaurin expansion with all coefficients positive if $\frac{1}{2} < \sigma \leq 1$. Since $w(t, 1) = w(t)$, hence $\tau_\nu^{(n)} = \tau_\nu(1)$, $\nu = 1, 2, \ldots, n$, the assertion (3.3) then follows immediately from (3.6) with $\sigma = 1$.

Write, for short, $\tau_\nu = \tau_\nu(\sigma)$, $\tau_\nu^* = \tau_{n+1-\nu}(\sigma)$. Since $w(t, \frac{1}{2})$ is an even function, the zeros $\tau_\nu(\sigma)$ for $\sigma = \frac{1}{2}$ are symmetric with respect to the origin. Furthermore, each zero $\tau_\nu(\sigma)$, under the assumption of (*a*), increases monotonically on $\frac{1}{2} \leq \sigma \leq 1$ (see, e. g., [10]). If all $\tau_\nu(\sigma)$ are positive, the assertion is obvious in view of

$$\frac{1}{1-\tau_\nu z} = 1 + \tau_\nu z + \tau_\nu^2 z^2 + \ldots, \quad \tau_\nu > 0.$$

It suffices, therefore, to consider pairs of zeros τ_ν, τ_ν^* (if $n \geq 2$) such that

$$-a < \tau_\nu^* \leq 0 < \tau_\nu < a, \quad |\tau_\nu^*| < \tau_\nu.$$

Write $\tau_\nu^* = -\gamma_\nu \tau_\nu$, $0 \leq \gamma_\nu < 1$. Then

$$\begin{aligned} \frac{1}{(1-\tau_\nu z)(1-\tau_\nu^* z)} &= \frac{1}{(1-\tau_\nu z)(1+\gamma_\nu \tau_\nu z)} \\ &= \frac{1}{1-[\tau_\nu(1-\gamma_\nu) z + \gamma_\nu \tau_\nu^2 z^2]} = 1 + [\ldots] + [\ldots]^2 + \ldots, \end{aligned} \tag{3.7}$$

where the content of the brackets on the right is the same as in the denominator immediately to the left. Since $\tau_\nu > 0$ and $0 \leq \gamma_\nu < 1$, the coefficients of z and z^2 in these brackets are positive and nonnegative, respectively, hence (3.7), when fully expanded in powers of z, can produce only positive coefficients. The same is true for $q(z)$ if n is odd, since by the monotonicity of the zeros, $\tau_{(n+1)/2}(\sigma) > 0$ for $\frac{1}{2} < \sigma \leq 1$.

(*b*) In this case of symmetry, $w(t, \sigma) = w(t)$ is even and $\gamma_\nu = 1$ in (3.7); the expansion of (3.6) contains only even powers of z, each, as before, with a positive coefficient, and $q(z) \equiv 1$.

(*c*) Applying part (*a*) to the weight function $w(-t)$ and its associated (monic) orthogonal polynomials $(-1)^n \pi_n(-z)$ yields positive coefficients in the expansion of $[(-z)^n \pi_n(-1/z)]^{-1}$, hence alternating coefficients in the expansion of $[z^n \pi_n(1/z)]^{-1}$. □

We are now in a position to prove the following theorem for the expansion coefficients $\mu_k^{(n)} = \mu_{k,0}^{(n)}$ in (1.5).

THEOREM 3.2. *Let*

$$\mu_k^{(n)} = \sum_{\nu=1}^{n} \lambda_\nu^{(n)} [\tau_\nu^{(n)}]^k, \quad k=1,2,3,\ldots; \quad n=1,2,3,\ldots. \tag{3.8}$$

(a) If $w(t)/w(-t)$ is strictly increasing on I, then

$$0<\mu_k^{(1)}<\mu_k^{(2)}<\ldots<\mu_k^{([k/2]+1)}=\mu_k^{([k/2]+2)}=\ldots=\mu_k. \tag{3.9}$$

(b) If $w(t)=w(-t)$ on I, then

$$0=\mu_k^{(1)}<\mu_k^{(2)}<\ldots<\mu_k^{([k/2]+1)}=\mu_k^{([k/2]+2)}=\ldots=\mu_k \text{ if } k \text{ is even} \tag{3.10}$$

(while, trivially, $0=\mu_k^{(1)}=\mu_k^{(2)}=\ldots=\mu_k$ if k is odd).

(c) If $w(t)/w(-t)$ is strictly decreasing on I, then

$$0<\mu_k^{(1)}<\mu_k^{(2)}<\ldots<\mu_k^{([k/2]+1)}=\mu_k^{([k/2]+2)}=\ldots=\mu_k \quad \text{if } k \text{ is even} \tag{3.11}$$

and

$$\mu_k=\ldots=\mu_k^{([k/2]+2)}=\mu_k^{([k/2]+1)}<\mu_k^{([k/2])}<\ldots<\mu_k^{(2)}<\mu_k^{(1)}<0 \quad \text{if } k \text{ is odd.} \tag{3.12}$$

PROOF. By (1.5), (1.6), with $j=0$, we have

$$f[n-1,n](z)=\mu_0+\mu_1 z+\ldots+\mu_{2n-1} z^{2n-1}+\sum_{k=2n}^{\infty} \mu_k^{(n)} z^k. \tag{3.13}$$

With $h_n=\int_I \pi_n^2(t)\,d\lambda(t)$ denoting the normalization factor for the orthogonal polynomial π_n, it is known (see, e. g., [11], where $-z$ is used in place of our z), and easily verified, that

$$f[n,n+1](z)-f[n-1,n](z)=\frac{h_n}{z\pi_n(1/z)\,\pi_{n+1}(1/z)}.$$

By (3.13), this can be rewritten in the form

$$\sum_{l=0}^{\infty} (\mu_{2n+l}^{(n+1)}-\mu_{2n+l}^{(n)}) z^l=\frac{h_n}{z^n\,\pi_n(1/z)\,z^{n+1}\,\pi_{n+1}(1/z)}. \tag{3.14}$$

Now in the case *(a)* we apply Theorem 3.1 *(a)* to the expansion of both $[z^n\,\pi_n(1/z)]^{-1}$ and $[z^{n+1}\,\pi_{n+1}(1/z)]^{-1}$ on the right of (3.14) and conclude that the product expansion has all coefficients positive, hence $\mu_{2n+l}^{(n+1)}-\mu_{2n+l}^{(n)}>0$ for all $l\geq 0$. This proves all «inner» inequalities in (3.9). The outer inequality on the left,

$$\mu_k^{(1)}=\lambda_1^{(1)}\,[\tau_1^{(1)}]^k>0,$$

follows from the positivity of the Christoffel number $\lambda_1^{(1)}$ and from $\tau_1^{(1)}>0$, which in turn is a consequence of the monotonicity property for the root $\tau_1(\sigma)$ (used in the proof of Theorem 3.1). The equalities on the right of (3.9) follow from (1.7). Parts (*b*) and (*c*) of Theorem 3.2 follow similarly from Theorem 3.1 (*b*) and (*c*). □

COROLLARY TO THEOREM 3.2. *The inequalities* (3.9) - (3.12) *of Theorem* 3.2, *under the appropriate assumption on w, hold also for the quantities* $\mu_{k,j}^{(n)}$ *defined in* (1.6), *provided j is even.*

PROOF. If w satisfies one of the assumptions (*a*), (*b*), (*c*) of Theorem 3.2, then $w_j(t)=t^j w(t)$, j even, satisfies the same assumption. □

EXAMPLE 3.1. Jacobi distribution $w(t)=(1-t)^\alpha(1+t)^\beta$ on $[-1,1]$. Here,

$$\frac{w(t)}{w(-t)}=\left(\frac{1+t}{1-t}\right)^{\beta-\alpha},$$

which is strictly increasing if $\alpha<\beta$, equal to 1 if $\alpha=\beta$, and strictly decreasing if $\alpha>\beta$. Accordingly, we have (3.9) if $\alpha<\beta$, (3.10) if $\alpha=\beta$, and (3.11), (3.12) if $\alpha>\beta$.

EXAMPLE 3.2. The special cases $\alpha=\pm\frac{1}{2}$, $\beta=\pm\frac{1}{2}$ of Example 3.1 yield interesting trigonometric inequalities, the simplest of which (for $\alpha=\beta=-\frac{1}{2}$) are

$$\frac{1}{n}\sum_{\nu=1}^{[n/2]}\left[\cos\left(\frac{2\nu-1}{2n}\pi\right)\right]^{2k}<\frac{1}{n+1}\sum_{\nu=1}^{[(n+1)/2]}\left[\cos\left(\frac{2\nu-1}{2n+2}\pi\right)\right]^{2k}, \tag{3.15}$$

$$k=2,3,4,\ldots;\quad n=2,3,\ldots,k.$$

EXAMPLE 3.3. A translate of the logistic distribution:

$$w(t)=\frac{1}{\vartheta}\,\frac{e^{-(t-\mu)/\vartheta}}{[1+e^{-(t-\mu)/\vartheta}]^2},\quad -\infty<t<\infty,\quad \vartheta>0.$$

An elementary computation will show that we are in the case (*a*), (*b*), (*c*) of Theorem 3.2 depending on whether $\mu>0$, $\mu=0$, or $\mu<0$, respectively.

4. Computational methods.

Instead of considering as input data the moments μ_k in (1.2), which, as is

well-known, give usually rise to ill-conditioned problems, we assume here that we are given the measure $d\lambda(t)$, and that it be required to generate from it the desired Padé approximant. We concentrate on the approximant $f[n-1+2j, n]$, $n \geq 1$, $j \geq 0$, referring to Remark 3 of Section 2 for the case of $f[n+2j, n]$. The measure $d\lambda$, being positive definite, generates a set of (monic) orthogonal polynomials $\pi_k(\cdot) = \pi_k(\cdot; d\lambda)$ satisfying

$$\begin{aligned} &\pi_{k+1}(z) = (z-\alpha_k)\,\pi_k(z) - \beta_k\,\pi_{k-1}(z), \quad k=0, 1, 2, \dots, \\ &\pi_{-1}(z) = 0, \ \pi_0(z) = 1, \end{aligned} \tag{4.1}$$

where α_k, β_k are real numbers and $\beta_k > 0$. Although β_0 is arbitrary, we find it convenient to define $\beta_0 = \int_{\mathbf{R}} d\lambda(t)$.

Given the measure $d\lambda$, a number of (usually stable) methods are known for generating the coefficients α_k, β_k, $k=0, 1, 2, \dots$; see Gautschi [6]. We assume, therefore, that we are given the first $n+j$ of these coefficients: α_k, β_k, $k=0, 1, 2, \dots, n+j-1$. (For «classical» measures they are known explicitly). From these, it will be possible to compute not only the moments $\mu_0, \mu_1, \dots, \mu_{2j-1}$, but also the Gauss-Christoffel data $\lambda_{\nu,2j}^{(n)}$, $\tau_{\nu,2j}^{(n)}$, all in a stable manner. Together, they determine the desired Padé approximant $f[n-1+2j, n]$ according to (1.3_{2j}). Our motivation to proceed in this manner derives mainly from two considerations: First, we wish to maintain a high degree of numerical stability in generating the approximant in question, and also obtain it in a form conducive to stable evaluation. Secondly, it is desirable to employ mathematical software which, by now, ought to be part of the standard computing repertoire.

Basically, the problem is to compute the recursion coefficients $\alpha_{k,2j}$, $\beta_{k,2j}$ associated with the measure $d\lambda_{2j}(t) = t^{2j}\,d\lambda(t)$, given the recursion coefficients $\alpha_k = \alpha_{k,0}$, $\beta_k = \beta_{k,0}$. There are several methods of accomplishing this; we discuss three of them based, respectively, on Christoffel's theorem, Chebyshev's algorithm, and the QR algorithm. It suffices to consider $j=1$, since the general case can be treated by repeated application of the special one.

4.1. *An algorithm derived from Christoffel's theorem.* Our first algorithm obtains from multiplying the measure $d\lambda(t)$ twice by t, each time employing Galant's algorithmic version [3] of Christoffel's theorem to generate the new recursion coefficients. The result is (cf. also Eq. (4.1) in Gautschi [5], where $z=0$):

$$(4.2)\qquad \left.\begin{aligned} &\hat{e}_{-1}=0,\ q_0=\alpha_0\\ &e_k=\beta_{k+1}/q_k\\ &\hat{q}_k=q_k+e_k-\hat{e}_{k-1}\\ &\hat{\beta}_k=\hat{q}_k\,\hat{e}_{k-1}\\ &q_{k+1}=\alpha_{k+1}-e_k\\ &\hat{e}_k=q_{k+1}\,e_k/\hat{q}_k\\ &\hat{\alpha}_k=\hat{q}_k+\hat{e}_k \end{aligned}\right\} \quad k=0,1,2,\dots,n-1,$$

where $\hat{\alpha}_k, \hat{\beta}_k$ denote the recursion coefficients for the measure $d\lambda_2(t)=t^2 d\lambda(t)$. In (4.2), $\hat{\beta}_0$ is set equal to zero. If we wish to adhere to our convention $\hat{\beta}_0=\int_{\mathbf{R}} d\lambda_2(t)$, we must redefine $\hat{\beta}_0$ as

$$(4.2_0)\qquad \hat{\beta}_0=\beta_0(\beta_1+\alpha_0^2).$$

This is obtained by representing t^2 in terms of the orthogonal polynomials π_k,

$$(4.3)\qquad t^2=c_0\pi_0(t)+c_1\pi_1(t)+c_2\pi_2(t),$$

comparing coefficients of equal powers on the right and left, and making use of $\pi_1(t)=t-\alpha_0$, $\pi_2(t)=(t-\alpha_1)(t-\alpha_0)-\beta_1$. One finds

$$(4.4)\qquad c_0=\beta_1+\alpha_0^2,\quad c_1=\alpha_0+\alpha_1,\quad c_2=1,$$

from which, by orthogonality, $\hat{\beta}_0=\int_{\mathbf{R}} t^2\,d\lambda(t)=c_0\int_{\mathbf{R}} d\lambda(t)=(\beta_1+\alpha_0^2)\,\beta_0$, as claimed in (4.2_0).

The algorithm (4.2), (4.2_0) produces the first n of the desired recursion coefficients in terms of the first $n+1$ given coefficients. In the general case of $d\lambda_{2j}$, one needs the first $n+j$ recursion coefficients α_k, β_k, $k=0,1,\dots,n+j-1$, in order to produce $\alpha_{k,2j}, \beta_{k,2j}$, $k=0,1,\dots,n-1$, by j-fold repetition of (4.2), (4.2_0).

A serious deficiency of this algorithm is the fact that it breaks down whenever $\alpha_0=0$ (division by zero in $e_0=\beta_1/q_0$), which happens, for example, if $d\lambda$ is a symmetric measure. Also, considerable loss of accuracy is observed in cases where $d\lambda$ is «nearly symmetric», i. e., the coefficients α_k are relatively small. The two algorithms that follow are more robust in this respect.

4.2. *Modified Chebyshev algorithm.* Another implementation of Christoffel's theorem, using modified moments, has been described in Gautschi [4, p. 123]. The «modified moments» of $d\lambda_2(t)=t^2\,d\lambda(t)$ with respect to the polynomials $\pi_k(\cdot)=\pi_k(\cdot;d\lambda)$, in view of (4.3), are

$$\nu_k=\int_{\mathbf{R}}\pi_k(t)\,d\lambda_2(t)=\int_{\mathbf{R}}\pi_k(t)\,t^2\,d\lambda(t)$$

$$=\begin{cases}c_k\displaystyle\int_{\mathbf{R}}\pi_k^2(t)\,d\lambda(t) & \text{if } k\leq 2,\\ 0 & \text{if } k>2.\end{cases}$$

Since $\displaystyle\int_{\mathbf{R}}\pi_k^2\,d\lambda(t)=\beta_0\beta_1\ldots\beta_k$, and taking note of (4.4), one finds

$$\nu_0=\beta_0(\beta_1+\alpha_0^2),\quad \nu_1=\beta_0\beta_1(\alpha_0+\alpha_1),\quad \nu_2=\beta_0\beta_1\beta_2,$$
$$\nu_k=0 \text{ for } k>2. \tag{4.5}$$

Given these modified moments of $d\lambda_2$, the modified Chebyshev algorithm produces the desired recursion coefficients for $d\lambda_2$ in terms of certain quantities $\sigma_{k,l}$ generated recursively from the ν_k (cf., e. g., Gautschi [6, § 2.4]). The algorithm, in fact, simplifies considerably, since $\nu_k=0$, $k>2$, which implies $\sigma_{k,k+3}=0$, all $k\geq 0$. Using the notation $u_k=\sigma_{k,k}$, $v_k=\sigma_{k,k+1}$, $w_k=\sigma_{k,k+2}$, the algorithm can be written in the form

Initialization:

$$\left.\begin{aligned}&u_0=\beta_0(\beta_1+\alpha_0^2)\\ &v_0=\beta_0\beta_1(\alpha_0+\alpha_1)\\ &w_0=\beta_0\beta_1\beta_2\\ &w_{-1}=0\end{aligned}\right\}\ \text{if } n>1$$
$$\hat{\alpha}_0=\alpha_0+\frac{v_0}{u_0}$$
$$\hat{\beta}_0=u_0 \tag{4.6$_0$}$$

Continuation: for $k=1,2,\ldots,n-1$:

$$
\begin{aligned}
u_k &= w_{k-1} - (\hat{\hat{\alpha}}_{k-1} - \alpha_k)\, v_{k-1} - \hat{\hat{\beta}}_{k-1}\, w_{k-2} + \beta_k\, u_{k-1} \\
v_k &= -(\hat{\hat{\alpha}}_{k-1} - \alpha_{k+1})\, w_{k-1} + \beta_{k+1}\, v_{k-1} \\
w_k &= \beta_{k+2}\, w_{k-1} \quad (\text{if } k < n-1) \\
\hat{\hat{\alpha}}_k &= \alpha_k + \frac{v_k}{u_k} - \frac{v_{k-1}}{u_{k-1}} \\
\hat{\hat{\beta}}_k &= \frac{u_k}{u_{k-1}} .
\end{aligned}
\tag{4.6}
$$

Given the recursion coefficients α_k, β_k, $0 \le k \le n$, for $d\lambda$, this determines uniquely and unequivocally (since $u_k = \sigma_{k,k} > 0$) the recursion coefficients $\hat{\hat{\alpha}}_k$, $\hat{\hat{\beta}}_k$, $0 \le k \le n-1$, for $d\lambda_2$. As before, the coefficients $\alpha_{k,2j}, \beta_{k,2j}$ can be obtained by repeating the process j-times.

4.3 *QR algorithm.* Golub &Kautsky [8, Corollary 1 to Lemma 4] recently observed that the recursion coefficients for $d\lambda_2(t) = t^2\, d\lambda(t)$ can be obtained from those for $d\lambda(t)$ by applying one step of the QR algorithm (with zero shift) to a symmetric tridiagonal matrix. More precisely, if the first n recursion coefficients $\hat{\hat{\alpha}}_k, \hat{\hat{\beta}}_k$, $k = 0, 1, \ldots, n-1$, of $d\lambda_2$ are desired, one applies the QR step to the (symmetric, tridiagonal) Jacobi matrix of order $n+2$ belonging to $d\lambda$ [with diagonal elements α_k, $k = 0, 1, \ldots, n+1$, and first side-diagonal elements $\sqrt{\beta_k}$, $k = 1, 2, \ldots, n+1$] and discards the last two rows and columns in the result. The matrix of order n so obtained is the Jacobi matrix for $d\lambda_2$. Using the square root free implementation of the QR algorithm, described in Wilkinson [14, p. 567], one is led to the following algorithm (in the notations of [14], except that a_j is replaced by α_{j-1} and b_j by $\sqrt{\beta_{j-1}}$):

$$
\begin{aligned}
& u_0 = 0,\ c_0 = 1,\ \hat{\hat{\beta}}_0 = \beta_0\,(\beta_1 + \alpha_0^2) \\
& \left.\begin{aligned}
& \gamma_k = \alpha_{k-1} - u_{k-1} \\
& p_k^2 = \begin{cases} \gamma_k^2 / c_{k-1}^2 & \text{if } c_{k-1} \neq 0 \\ c_{k-2}^2\, \beta_{k-1} & \text{if } c_{k-1} = 0 \end{cases} \\
& \hat{\hat{\beta}}_{k-1} = s_{k-1}^2\,(p_k^2 + \beta_k) \quad \text{if } k > 1 \\
& s_k^2 = \beta_k / (p_k^2 + \beta_k) \\
& c_k^2 = p_k^2 / (p_k^2 + \beta_k) \\
& u_k = s_k^2\,(\gamma_k + \alpha_k) \\
& \hat{\hat{\alpha}}_{k-1} = \gamma_k + u_k
\end{aligned}\right\} \quad k = 1, 2, \ldots, n.
\end{aligned}
\tag{4.7}
$$

As before, j-fold repetition of this algorithm, with n suitably increased, yields the recursion coefficients for $d\lambda_{2j}(t)=t^{2j}\,d\lambda(t)$.

Numerically, Algorithm (4.7) appears to produce slightly more accurate results than the algorithm in (4.6), but otherwise they are comparable. Algorithm (4.2), as already observed, looses accuracy in nearly symmetric situations. This is illustrated in Table 4.1, where the maximum relative errors in $\hat{\alpha}_k$, $\hat{\beta}_k$, $k=0, 1, \ldots, n-1$, are shown in the case $d\lambda(t)=(1+t)^\varepsilon\,dt$ on $[-1, 1]$, with $\varepsilon=.1, .01, \ldots, .00001$, $n=20$, using Algorithms (4.7), (4.6) and (4.2) based, respectively, on the QR algorithm, Chebyshev's algorithm and Christoffel's theorem. (Numbers in parentheses indicate decimal exponents. The computations were performed on the CDC 6500 computer, which has a machine precision of approx. 3.55×10^{-15} in single precision).

	QR		Chebyshev		Christoffel	
ε	err α	err β	err α	err β	err α	err β
.1	7.49(-14)	2.48(-14)	1.05(-12)	9.38(-14)	1.28(-11)	1.24(-12)
.01	8.18(-14)	1.67(-14)	7.00(-13)	7.03(-14)	1.89(-9)	1.88(-10)
.001	6.82(-14)	2.25(-14)	4.47(-13)	6.60(-14)	2.97(-7)	2.89(-8)
.0001	8.79(-14)	1.74(-14)	9.07(-13)	6.92(-14)	1.05(-5)	1.02(-6)
.00001	9.30(-14)	1.57(-14)	7.80(-13)	6.97(-14)	3.39(-4)	4.90(-5)

TABLE 4.1. *Numerical performance of the algorithms* (4.7), (4.6) *and* (4.2) *in the case* $d\lambda(t)=(1+t)^\varepsilon\,dt$ *on* $[-1, 1]$, $n=20$.

Once the recursion coefficients $\alpha_{k,2j}$, $\beta_{k,2j}$ have been obtained, it is a simple matter to produce from the corresponding Jacobi matrix the Christoffel numbers $\lambda_{\nu,2j}^{(n)}$ and the nodes $\tau_{\nu,2j}^{(n)}$ required in (1.3_{2j}). For appropriate methods see, e. g., Gautschi [4, § 5.1]. Likewise, the moments μ_k, $k=0, 1, \ldots, 2j-1$, can be computed by applying the j-point Gauss-Christoffel quadrature rule (associated with the measure $d\lambda$) to the integrals in (1.2).

For Stieltjes series the appropriate algorithm is the Cholesky LR algorithm, Golub & Kautsky [8, Theorem 3], or the closely related algorithms in Galant [3] and Gautschi [5].

Acknowledgment.

The author is indebted to Richard A. Askey for reminding him of the results in reference [10] and to William B. Gragg Jr. for useful discussions.

REFERENCES

[1] G. D. ALLEN, C. K. CHUI, W. R. MADYCH, F.J. NACOWICH, P. W. SMITH, *Padé approximation of Stieltjes series,* J. Approx. Theory **14** (1975), 302-316.

[2] G. A. BAKER, JR., P. GRAVES-MORRIS, *Padé approximants, part I: basic theory* 1981. Addison-Wesley, Reading, Mass.

[3] D. GALANT, *An implementation of Christoffel's theorem in the theory of orthogonal polynomials,* Math. Comp. **25** (1971), 111-113.

[4] W. GAUTSCHI, *A survey of Gauss-Christoffel quadrature formulae,* in «E. B. Christoffel, The Influence of his Work on Mathematics and the Physical Sciences» 1981. P. L. Butzer, F. Fehér, eds. Birkhäuser, Basel, 72-147.

[5] W. GAUTSCHI, *An algorithmic implementation of the generalized Christoffel theorem,* in: «Numerische Integration» 1982. G. Hämmerlin, ed., 89-106. Internat. Ser. Numer. Math. **57**, Birkhäuser, Basel.

[6] W. GAUTSCHI, *On generating orthogonal polynomials,* SIAM J. Sci. Statist. Comput.. **3** (1982), 289-317.

[7] A. GHIZZETTI, A. OSSICINI, *Polinomi ortogonali e problema dei momenti,* Pubbl. Istit. Mat. Appl. Fac. Ing. Univ. Stud. Roma **231**, Rome, 1981.

[8] G. H. GOLUB, J. KAUTSKY, *Calculation of Gauss quadratures with multiple free and fixed knots,* Numer. Math. **41** (1983), 147-163.

[9] W. B. GRAGG, *The Padé table and its relation to certain algorithms of numerical analysis,* SIAM Rev. **14** (1972), 1-62.

[10] D. B. HUNTER, *Some properties of orthogonal polynomials,* Math. Comp. **29** (1975), 559-565.

[11] J. KARLSSON, B. VON SYDOW, *The convergence of Padé approximants to series of Stieltjes,* Ark. Mat. **14** (1976), 43-53.

[12] O. PERRON, *Die Lehre von den Kettenbrüchen,* 1957, **II**, 3rd ed., B. G. Teubner, Stuttgart.

[13] H. S. WALL, *Analytic theory of continued fractions,* Chelsea, Bronx, N. Y., 1948.

[14] J. H. WILKINSON, *The algebraic eigenvalue problem,* Clarendon Press, Oxford, .1965.

10.3. [87] "On the Convergence Behavior of Continued Fractions with Real Elements"

---

[87] "On the Convergence Behavior of Continued Fractions with Real Elements," *Math. Comp.* **40**, 337–342 (1983).

MATHEMATICS OF COMPUTATION
VOLUME 40, NUMBER 161
JANUARY 1983, PAGES 337–342

On the Convergence Behavior of Continued Fractions with Real Elements*

By Walter Gautschi

Abstract. We define the notion of transient (geometric) convergence rate for infinite series and continued fractions. For a class of continued fractions with real elements we prove a monotonicity property for such convergence rates which helps explain the effectiveness of certain continued fractions known to converge "only" sublinearly. This is illustrated in the case of Legendre's continued fraction for the incomplete gamma function.

1. Introduction. Continued fractions, as is well known, can be viewed in terms of infinite series. To describe the convergence behavior of a series it is useful to consider the notion of *transient* (geometric) *convergence rate*. Given a convergent series $\sum_{n=0}^{\infty} t_n$, the nth transient convergence rate is the quantity $|\rho_n|$, $n = 1, 2, \ldots$, where $t_n = \rho_n t_{n-1}$ (assuming $t_{n-1} \neq 0$). If $\lim_{n\to\infty} |\rho_n| = r$, $0 \leq r \leq 1$, convergence is *linear* (geometric) with convergence rate r, if $0 < r < 1$, *superlinear*, if $r = 0$, and *sublinear* if $r = 1$. It is important to note, however, that these concepts are asymptotic in nature, hence not necessarily relevant for numerical (finite!) computation. Thus, a series need not be dismissed as useless, simply because it converges only sublinearly. The approach of $|\rho_n|$ to the limit 1 indeed may be so slow that the series has "converged to machine precision" long before $|\rho_n|$ reaches the neighborhood of 1. For this reason, convergence of a series ought to be judged on the basis of the complete sequence $\{\rho_n\}$ of convergence rates, and not just on the basis of asymptotic properties of ρ_n. In this connection, properties of monotone behavior significantly add to the understanding of the quality of convergence.

The purpose of this note is to prove a criterion for the sequence $\{|\rho_n|\}$ to be (ultimately) monotonically increasing, in the case where the partial sums of the series are convergents of a continued fraction with real elements. We illustrate the result with Legendre's continued fraction for the incomplete gamma function, which, though sublinearly convergent, provides an effective tool of numerical computation.

2. Continued Fractions and Infinite Series. We consider continued fractions of the form

$$c = \frac{1}{1+}\,\frac{a_1}{1+}\,\frac{a_2}{1+}\cdots, \tag{2.1}$$

where, for some integer $k_0 \geq 1$,

$$\begin{aligned} &a_k > 0 \quad \text{for } 1 \leq k \leq k_0 - 1, \\ &a_k < 0 \quad \text{and} \quad |a_k| \leq \tfrac{1}{4} \quad \text{for } k \geq k_0. \end{aligned} \tag{2.2}$$

Received February 16, 1982.
1980 *Mathematics Subject Classification.* Primary 40A15; Secondary 33A70.
Key words and phrases. Convergence of real continued fractions, Legendre's continued fraction for the incomplete gamma function.
*Sponsored in part by the National Science Foundation under Grant MCS-7927158.

0025-5718/82/0000-0715/$02.25

It can be seen from Worpitzky's theorem (Henrici [3, p. 506]) that the tail of the continued fraction (2.1) beginning with the element a_{k_0}, hence also the complete continued fraction, converges. The infinite series

$$s = \sum_{k=0}^{\infty} t_k \qquad (t_0 = 1) \tag{2.3}$$

is *equivalent* to the continued fraction (2.1) if its nth partial sum

$$s_n = 1 + \sum_{k=1}^{n-1} t_k \tag{2.4}$$

is equal to the nth convergent of c, for each $n = 1, 2, 3, \ldots$. According to Euler,

$$s_1 = 1, \quad s_{k+1} = s_k + t_k, \qquad k = 1, 2, 3, \ldots, \tag{2.5}$$

where

$$\left.\begin{aligned} &\rho_0 = 0, \qquad t_0 = 1, \\ &\rho_k = \frac{-a_k(1 + \rho_{k-1})}{1 + a_k(1 + \rho_{k-1})} \\ &t_k = \rho_k t_{k-1} \end{aligned}\right\} \qquad k = 1, 2, 3, \ldots. \tag{2.6}$$

This represents a convenient algorithm for evaluating the continued fraction c, and is also useful for analyzing qualitative properties of convergence. Note indeed that the quantities ρ_n in (2.6) yield the transient convergence rates $|\rho_n|$ of the series (2.3).

Slightly more convenient for analytical purposes are the quantities $\sigma_k = 1 + \rho_k$, which satisfy

$$\sigma_0 = 1, \quad \sigma_k = \frac{1}{1 + a_k \sigma_{k-1}}, \qquad k = 1, 2, 3, \ldots. \tag{2.7}$$

3. Convergence Behavior. Some first insights into the convergence behavior of the continued fraction (2.1) can be gained from the following lemma.

LEMMA 3.1. *If the partial numerators a_k in* (2.1) *satisfy* (2.2), *then the quantities σ_k in* (2.7) *satisfy*

$$0 < \sigma_k < 1 \quad \textit{for } 1 \leq k \leq k_0 - 1, \tag{3.1}$$

and

$$1 < \sigma_k \leq \frac{2(k - k_0 + 2)}{k - k_0 + 3} \quad \textit{for } k \geq k_0. \tag{3.2}$$

Proof. The inequalities (3.1) follow immediately from the positivity of a_k and (2.7). To prove (3.2), we use induction. Since $-\frac{1}{4} \leq a_{k_0} < 0$ and $0 < \sigma_{k_0 - 1} \leq 1$, we have $1 < \sigma_{k_0} \leq 4/3$, so that (3.2) is true for $k = k_0$. Assuming its truth for some $k \geq k_0$, we obtain

$$1 < \sigma_{k+1} = \frac{1}{1 + a_{k+1}\sigma_k} \leq \frac{1}{1 - \frac{1}{4}\frac{2(k - k_0 + 2)}{k - k_0 + 3}} = \frac{2(k - k_0 + 3)}{k - k_0 + 4},$$

which is (3.2) with k replaced by $k + 1$. □

Lemma 3.1, in particular, implies $0 < \sigma_k < 2$, hence $-1 < \rho_k < 1$, for all $k \geq 1$. The series (2.3), therefore, has terms that are strictly decreasing in absolute value. Furthermore, by (3.1) and (3.2),

$$(3.3) \quad -1 < \rho_k < 0 \quad \text{for } 1 \leq k \leq k_0 - 1, \quad \text{and} \quad 0 < \rho_k < 1 \quad \text{for } k \geq k_0,$$

so that the series initially (if $k_0 > 1$) behaves like an alternating series and subsequently turns into a monotone series.

A more detailed description of convergence is provided by the following theorem.

THEOREM 3.1. *If the partial numerators a_k in* (2.1) *satisfy* (2.2), *and in addition* $-\frac{1}{4} \leq a_{k+1} \leq a_k < 0$ *for* $k \geq k_0$, *then*

$$(3.4) \quad -1 < \rho_k < 0 \quad \textit{for } 1 \leq k \leq k_0 - 1 \quad \textit{and} \quad \rho_{k+1} > \rho_k > 0 \quad \textit{for } k \geq k_0.$$

In particular,

$$(3.5) \qquad \lim_{k \to \infty} \rho_k = \rho, \qquad \rho = \frac{1 - \sqrt{1 + 4a}}{1 + \sqrt{1 + 4a}},$$

where $a = \lim_{k \to \infty} a_k$; *the continued fraction* (2.1) *converges linearly, with convergence rate* ρ, *if* $a > -\frac{1}{4}$, *and sublinearly if* $a = -\frac{1}{4}$.

Proof. The first inequalities in (3.4) have already been noted in (3.3). The others are equivalent to $\sigma_{k+1} > \sigma_k > 1$ for $k \geq k_0$. Since $\sigma_k > 1$, by (3.2), it suffices to prove

$$(3.6) \qquad \sigma_{k+1} > \sigma_k \quad \text{for } k \geq k_0.$$

We first show

$$(3.7) \qquad \sigma_{k-1} < \frac{2}{1 + \sqrt{1 - 4|a_k|}} \quad \text{for } k \geq k_0.$$

This is true for $k = k_0$, since by (3.1) (and (2.7), if $k_0 = 1$) $\sigma_{k_0 - 1} \leq 1$, while the expression on the right of (3.7) is greater than 1. Using induction, assume that (3.7) holds for some $k \geq k_0$. Then

$$(3.8) \qquad \begin{aligned} \sigma_k = \frac{1}{1 + a_k \sigma_{k-1}} &< \frac{1}{1 - |a_k| \dfrac{2}{1 + \sqrt{1 - 4|a_k|}}} \\ &\leq \frac{1}{1 - \dfrac{2|a_{k+1}|}{1 + \sqrt{1 - 4|a_{k+1}|}}}, \end{aligned}$$

where in the last inequality we have used $|a_{k+1}| \geq |a_k|$. Now observe that, for any $\alpha \leq \frac{1}{4}$,

$$\frac{1}{1 - \dfrac{2\alpha}{1 + \sqrt{1 - 4\alpha}}} = \frac{1 + \sqrt{1 - 4\alpha}}{1 + \sqrt{1 - 4\alpha} - 2\alpha}$$

$$= \frac{1 - (1 - 4\alpha)}{1 - (1 - 4\alpha) - 2\alpha\left(1 - \sqrt{1 - 4\alpha}\right)} = \frac{2}{1 + \sqrt{1 - 4\alpha}}.$$

Using this in (3.8), with $\alpha = |a_{k+1}|$, yields (3.7) with k replaced by $k+1$, and thus establishes (3.7) for all $k \geq k_0$.

Now (3.6), in view of (2.7), is equivalent to

$$\frac{1}{1 + a_{k+1}\sigma_k} > \sigma_k \quad \text{for } k \geq k_0,$$

which in turn, since $1 + a_{k+1}\sigma_k > 0$ and $a_{k+1} < 0$ for $k \geq k_0$, is equivalent to

$$|a_{k+1}|\sigma_k^2 - \sigma_k + 1 > 0.$$

The quadratic function $|a_{k+1}|t^2 - t + 1$ is convex and has two real zeros $t_{1,k+1} < t_{2,k+1}$, the smaller of which is

$$t_{1,k+1} = \frac{2}{1 + \sqrt{1 - 4|a_{k+1}|}}.$$

By (3.7), $\sigma_k < t_{1,k+1}$, hence $|a_{k+1}|\sigma_k^2 - \sigma_k + 1 > 0$, which implies $\sigma_{k+1} > \sigma_k$. This proves (3.6).

Since the sequence $\{a_k\}$ is monotonically decreasing for $k \geq k_0$, and bounded below by $-\frac{1}{4}$, the limit $\lim_{k\to\infty} a_k = a$ exists, and $-\frac{1}{4} \leq a < 0$, since $a_{k_0} < 0$. Similarly, $\lim_{k\to\infty} \rho_k = \rho$, $0 < \rho \leq 1$, and $\lim_{k\to\infty} \sigma_k = \sigma$ with $\sigma = 1 + \rho$. Going to the limit $k \to \infty$ in (2.7) then gives

$$\sigma = \frac{1}{1 + a\sigma}, \qquad \sigma = \frac{2}{1 \pm \sqrt{1 + 4a}}.$$

Since $\sigma \leq 2$ and $-\frac{1}{4} \leq a < 0$, the minus sign in the last equation for σ cannot hold (unless $a = -\frac{1}{4}$), and we conclude that

$$\sigma = \frac{2}{1 + \sqrt{1 + 4a}}, \qquad \rho = \sigma - 1 = \frac{1 - \sqrt{1 + 4a}}{1 + \sqrt{1 + 4a}},$$

which is (3.5). The last statement of the theorem is an immediate consequence of (3.5). This completes the proof of Theorem 3.1.

4. Truncation. In practice, the continued fraction (2.1) is evaluated by carrying out (2.5) and (2.6) for $k = 1, 2, \ldots, n$ and taking s_{n+1} to approximate the value of s (or c) of the continued fraction. It is important, then, to be able to choose n in such a way that s_{n+1} approximates s to any prescribed accuracy.

Assuming first $k_0 = 1$, hence $0 < \rho_k < 1$ by (3.3) and $0 < t_k < 1$, it follows from a result of Merkes [4, Eq. (12)] that

$$|s - s_{n+1}| \leq \frac{1 + \rho_n}{1 - \rho_n} t_n. \tag{4.1}$$

This suggests the following *stopping rule*: Given a prescribed (relative) accuracy ε, stop the recursion (2.6) at the first integer $k = n$ for which

$$(1 + \rho_n)t_n \leq (1 - \rho_n)s_{n+1}\varepsilon. \tag{4.2}$$

By (4.1), this implies $|s - s_{n+1}| \leq s_{n+1}\varepsilon$, hence

$$\frac{|s - s_{n+1}|}{s + |s_{n+1} - s|} \leq \frac{|s - s_{n+1}|}{s_{n+1}} \leq \varepsilon,$$

from which $|s - s_{n+1}| \leqslant s\varepsilon + |s_{n+1} - s|\varepsilon$, that is,

$$\left|\frac{s - s_{n+1}}{s}\right| \leqslant \frac{\varepsilon}{1-\varepsilon}. \tag{4.3}$$

Our stopping rule therefore achieves the desired accuracy, at least asymptotically for $\varepsilon \to 0$.

To avail oneself of this simple stopping rule, when $k_0 > 1$, one ought to first evaluate the "tail"

$$c_{k_0} = \frac{1}{1+}\,\frac{a_{k_0}}{1+}\,\frac{a_{k_0+1}}{1+}\cdots \tag{4.4}$$

of the continued fraction (2.1), to which Merkes' result applies, and then compute

$$c_k = \frac{1}{1 + a_k c_{k+1}} \quad \text{for } k = k_0 - 1, k_0 - 2, \ldots, 1, \tag{4.5}$$

to get the complete continued fraction $c = c_1$. Since $c_{k_0} > 0$ and $a_k > 0$ for $k < k_0$, the computation in (4.5) involves the addition of positive numbers and division, hence only numerically stable operations.

5. An Example. Theorem 3.1 is applicable to Legendre's continued fraction for the incomplete gamma function,

$$\begin{aligned}(x - \alpha + 1)x^{-\alpha}e^{x}\Gamma(\alpha, x) &= \frac{1}{1+}\,\frac{a_1}{1+}\,\frac{a_2}{1+}\cdots,\\ a_k &= \frac{k(\alpha - k)}{(x - \alpha + 2k - 1)(x - \alpha + 2k + 1)}, \qquad k = 1, 2, 3, \ldots,\end{aligned} \tag{5.1}$$

which is used in [1, p. 475], [2] to compute the incomplete gamma function in the domain D: $x \geqslant 1.5$, $-\infty < \alpha \leqslant x + \frac{1}{4}$. Assuming α not a positive integer (otherwise, the continued fraction (5.1) would terminate and our assumption (2.2) would be violated), we have for $(x, \alpha) \in D$

$$k_0 = \begin{cases} 1 & \text{if } \alpha < 1, \\ 1 + [\alpha] & \text{if } \alpha > 1. \end{cases} \tag{5.2}$$

If $k \geqslant k_0$, the condition $|a_k| \leqslant \frac{1}{4}$ is equivalent to $(x - \alpha)^2 + 4kx \geqslant 1$, hence satisfied if $x \geqslant \frac{1}{4}$ (since $k \geqslant 1$). An elementary calculation furthermore shows that $|a_{k+1}| \geqslant |a_k|$ for $k \geqslant k_0$ whenever $x \geqslant \frac{1}{2}$. It follows, in particular, that all assumptions of Theorem 3.1 are satisfied when $(x, \alpha) \in D$. Since clearly $a = \lim_{k\to\infty} a_k = -\frac{1}{4}$, we are in a case of sublinear convergence. (This is also noted by Henrici [3, p. 629] by way of a different analysis.) Nevertheless, the continued fraction is known to be quite useful as a computational tool, at least in a domain such as D. The reason for this is readily understood on the basis of Theorem 3.1: Although the transient convergence rates ρ_k eventually increase monotonically to 1, the limit is approached quite slowly. We can see this from Table 5.1 which, in the case of the continued fraction (4.4), and for selected x and α, displays the values of

$$n_\nu = \max\left(k: |\rho_k| \leqslant \frac{\nu}{4}\right) \quad \text{and} \quad \varepsilon_\nu = \frac{4+\nu}{4-\nu} t_{n_\nu}, \qquad \nu = 1, 2, 3.$$

TABLE 5.1

Convergence behavior of the continued fraction (5.1)

(*Numbers in parentheses indicate decimal exponents.*)

x	α	n_1	ε_1	n_2	ε_2	n_3	ε_3
1.5	1.75	3	1.9(−3)	13	2.9(−7)	75	6.7(−18)
	.875	3	9.2(−4)	13	1.2(−7)	75	2.7(−18)
	0.0	3	6.8(−3)	13	1.4(−6)	75	3.7(−17)
	−3.5	3	7.3(−3)	13	1.4(−6)	75	3.5(−17)
	−7.0	4	8.9(−4)	15	5.8(−8)	79	5.5(−19)
5.0	5.25	9	1.9(−8)	41	5.4(−21)	243	2.9(−56)
	2.625	10	1.0(−9)	41	7.3(−22)	243	3.0(−57)
	0.0	10	9.8(−10)	42	2.9(−22)	244	1.6(−57)
	−10.5	9	1.9(−8)	42	1.6(−21)	244	8.5(−57)
	−21.0	12	2.7(−11)	48	1.2(−25)	254	8.8(−62)
10.0	10.25	18	1.1(−16)	81	1.2(−41)	483	3.0(−112)
	5.125	18	8.6(−17)	81	7.1(−42)	483	1.5(−112)
	0.0	20	8.0(−20)	83	8.1(−45)	485	1.9(−115)
	−20.5	19	1.7(−17)	83	6.9(−43)	485	1.6(−113)
	−41.0	24	1.5(−22)	95	5.0(−51)	504	4.2(−123)
20.0	20.25	36	3.0(−33)	161	4.8(−83)	963	2.9(−224)
	10.125	36	1.2(−33)	162	5.2(−84)	964	3.1(−225)
	0.0	40	2.9(−40)	165	3.4(−90)	967	1.6(−231)
	−40.5	38	4.3(−35)	165	8.8(−86)	968	3.4(−227)
	−81.0	48	3.1(−45)	188	1.2(−101)	1004	6.8(−246)

Note that by virtue of (4.1), and the fact that $1 < s = c_{k_0} \leqslant 2$ (cf. [3, Theorem 12.3c]),

$$\left|\frac{s - s_{n_\nu+1}}{s}\right| \leqslant |s - s_{n_\nu+1}| \leqslant \varepsilon_\nu, \qquad \nu = 1, 2, 3. \tag{5.3}$$

Thus, for example, if $x = 5$, $\alpha = 0$, by the time the transient convergence rate has risen to $\frac{1}{2}$, the continued fraction has already converged to within a (relative) error of about 3×10^{-22}.

Department of Computer Sciences
Purdue University
West Lafayette, Indiana 47907

1. W. GAUTSCHI, "A computational procedure for incomplete gamma functions," *ACM Trans. Math. Software*, v. 5, 1979, pp. 466–481.

2. W. GAUTSCHI, "Algorithm 542—Incomplete gamma function," *ACM Trans. Math. Software*, v. 5, 1979, pp. 482–489.

3. P. HENRICI, *Applied and Computational Complex Analysis*, Vol. 2, Wiley, New York, 1977.

4. E. P. MERKES, "On truncation errors for continued fraction computations," *SIAM J. Numer. Anal.*, v. 3, 1966, pp. 486–496.

10.4. [89] "Discrete Approximations to Spherically Symmetric Distributions"

---

[89] "Discrete Approximations to Spherically Symmetric Distributions," *Numer. Math.* **44**, 53–60 (1984).

Numer. Math. 44, 53–60 (1984)

Numerische Mathematik

Discrete Approximations to Spherically Symmetric Distributions*

Dedicated to Fritz Bauer on the occasion of his 60th birthday

Walter Gautschi

Purdue University, Department of Computer Sciences, Mathematical Sciences Building, Room 442, West Lafayette, Indiana 47907, USA

Summary. We consider the approximation of spherically symmetric distributions in $\mathbb{R}^d$ by linear combinations of Heaviside step functions or Dirac delta functions. The approximations are required to faithfully reproduce as many moments as possible. We discuss stable methods of computing such approximations, taking advantage of the close connection with Gauss-Christoffel quadrature. Numerical results are presented for the distributions of Maxwell, Bose-Einstein, and Fermi-Dirac.

Subject Classifications: AMS(MOS): 65D15 CR: 5.13.

1. Introduction

There is some interest among physicists in approximating the distribution functions of statistical mechanics by discrete functions – either linear combinations of Dirac delta functions or linear combinations of Heaviside step functions. For the Maxwell velocity distribution, Laframboise and Stauffer [9] and Calder, Laframboise and Stauffer [1] construct such approximations which are optimal in the sense of matching as many moments as possible. The resulting equations are solved in [9] by what amounts to Prony's method and in [1] by a reduction to an eigenvalue problem involving Hankel matrices. Both methods are classical; for the former, see, e.g. Hildebrand [8, §9.4], for the latter, Szegö [10, Eq. (2.2.9)]. They are subject to severe ill-conditioning, however, as is well-known. Here we point out that both approximation problems can be formulated in terms of Gauss-Christoffel quadrature, and can therefore be brought into the realm of stable modern methods of constructing orthogonal polynomials; see Gautschi [5]. In particular, algorithmic implementations of Christoffel's theorem (Galant [3], Gautschi [6]) find application here. We use these methods to generate numerical data for the distributions of Maxwell, Bose-Einstein, and Fermi-Dirac.

* Work supported in part by the National Science Foundation under Grant MCS-7927158A1

2. Approximation by Step Functions

We consider a function f which is spherically symmetric in $\mathbb{R}^d$, hence a function only of the radial distance, $f=f(r)$, $0\leqq r<\infty$. We impose the following conditions on f:

(i) $f\in C^1[0,\infty]$ and $f'(r)\leqq 0$ on $[0,\infty]$.

(ii) The integrals $\int_0^\infty f(r)r^m\,dr$, $\int_0^\infty f'(r)r^m\,dr$, $m=0,1,2,\ldots$, all exist and are finite.

In particular, f has to be nonnegative on $[0,\infty]$. This will not be required in this section.

Noting that

$$\int_0^\infty f(r)r^m\,dr=\frac{1}{m+1}f(r)r^{m+1}\Big|_0^\infty-\frac{1}{m+1}\int_0^\infty f'(r)r^{m+1}\,dr,\qquad m=0,1,2,\ldots,$$

it follows from (ii) that $r^{m+1}f(r)$ has a limit as $r\to\infty$, which of course must be equal to zero. Likewise, $f(r)\to 0$ as $r\to\infty$, as follows from (i) and (ii). Thus,

$$\lim_{r\to\infty} r^m f(r)=0,\qquad m=0,1,2,\ldots. \tag{2.1}$$

We wish to approximate f by a linear combination of Heaviside step functions,

$$f(r)\approx\tilde{f}(r),\qquad \tilde{f}(r)=\sum_{\nu=1}^{n} a_\nu H(r_\nu-r), \tag{2.2}$$

where $H(t)=0$ if $t\leqq 0$, $H(t)=1$ if $t>0$. Defining the moments of a function $g(r)$, as in [1], [9], by $\int_0^\infty g(r)r^j\,dV$, $j=0,1,2,\ldots$, where $dV=[2\pi^{d/2}/\Gamma(d/2)]r^{d-1}\,dr$ is the volume element of the spherical shell in $\mathbb{R}^d$, $d>1$, and $dV=dr$ if $d=1$, we require that f and $\tilde{f}$ have the same moments of orders up to $2n-1$, i.e.,

$$\int_0^\infty \sum_{\nu=1}^{n} a_\nu H(r_\nu-r)r^{j+d-1}\,dr=\int_0^\infty f(r)r^{j+d-1}\,dr,\qquad j=0,1,\ldots,2n-1,$$

or, which is the same,

$$\sum_{\nu=1}^{n} a_\nu\int_0^{r_\nu} r^{j+d-1}\,dr=\int_0^\infty f(r)r^{j+d-1}\,dr,\qquad j=0,1,\ldots,2n-1.$$

Carrying out the integration on the left, and integrating by parts on the right, we get, upon using (2.1),

$$\sum_{\nu=1}^{n}(a_\nu r_\nu^d)r_\nu^j=\int_0^\infty[-r^d f'(r)]r^j\,dr,\qquad j=0,1,\ldots,2n-1. \tag{2.3}$$

These are precisely the equations for n-point Gauss-Christoffel quadrature relative to the (nonnegative) integration measure

$$d\lambda(r)=-r^d f'(r)\,dr\qquad \text{on } [0,\infty]. \tag{2.4}$$

Hence, r_ν *in* (2.2) *are the Gaussian abscissas relative to the measure* $d\lambda(r)$ *in* (2.4) *and* $\lambda_\nu = a_\nu r_\nu^d$ *are the corresponding Christoffel numbers.* Once the nodes r_ν and weights λ_ν have been computed, the coefficients a_ν in (2.2) are simply obtained by $a_\nu = \lambda_\nu r_\nu^{-d}$, $\nu = 1, 2, \ldots, n$.

If $d=1$, the functions $f(r)$ and $\tilde{f}(r)$ are to be extended to negative values of r symmetrically with respect to the origin.

To compute the n-point Gauss-Christoffel formula in question, it suffices to generate the coefficients α_k, β_k, $k=0, 1, \ldots, n-1$, in the recursion formula

$$\begin{aligned} \pi_{k+1}(t) &= (t-\alpha_k)\pi_k(t) - \beta_k \pi_{k-1}(t), \quad k=0, 1, \ldots, n-1, \\ \pi_{-1}(t) &= 0, \quad \pi_0(t) = 1, \end{aligned} \tag{2.5}$$

for the respective (monic) orthogonal polynomials $\pi_k(\cdot) = \pi_k(\cdot\,; d\lambda)$. The desired nodes r_ν are then the eigenvalues of the Jacobi matrix

$$J_n = \begin{bmatrix} \alpha_0 & \sqrt{\beta_1} & & 0 \\ \sqrt{\beta_1} & \alpha_1 & \sqrt{\beta_2} & \\ & \ddots & \ddots & \ddots \sqrt{\beta_{n-1}} \\ 0 & & \sqrt{\beta_{n-1}} & \alpha_{n-1} \end{bmatrix},$$

and the weights λ_ν are expressible as $\lambda_\nu = \beta_0 v_{\nu,1}^2$, $\beta_0 = \int_0^\infty d\lambda(r)$, in terms of the first components $v_{\nu,1}$ of the corresponding normalized eigenvectors. They all can be readily computed using standard software of linear algebra; see, e.g., Gautschi [4, § 5.1].

3. Examples

Example 3.1. The Maxwell distribution (cf. [9, 1]) $f(r) = \pi^{-d/2} e^{-r^2}$ on $[0, \infty]$.

In this case, (2.4) yields

$$d\lambda(r) = \frac{2}{\pi^{d/2}} r^{d+1} e^{-r^2} dr \quad \text{on } [0, \infty]. \tag{3.1}$$

The recursion coefficients α_k, β_k in (2.5) that correspond to this measure can be computed in two different ways. We can start with the recursion coefficients α_k^0, β_k^0 for the measure $d\lambda^0(r) = \frac{2}{\sqrt{\pi}} e^{-r^2} dr$ on $[0, \infty]$, which are available in Galant [2] for $0 \leqq k \leqq 19$ to 20 significant decimal digits, and then keep multiplying the measure by r ($d+1$ times), each time generating the corresponding recursion coefficients by the algorithms described in Galant [3] or Gautschi [6]. The coefficient β_0, at the end, is then adjusted to conform to the normalization adopted in (3.1). Alternatively, we may compute the α_k, β_k directly from (3.1), using the discretized Stieltjes procedure as described in Gautschi [5,

Example 4.6]. We have used both these approaches, at the same time extending Galant's table up to $k=49$ and recomputing it to 25 decimal places. Using double precision on the CDC 6500 (ca. 29 significant decimal digits) we observed agreement to 25 decimal places. The results are listed in Table 1 of the Appendix. They can be used, as described at the end of Sect. 2, to produce Gauss-Christoffel formulae with as many as 50 terms. For reasons of space, we refrain from tabulating any of them.

Example 3.2. The Bose-Einstein distribution $f(r)=(\beta e^r-1)^{-1}$ on $[0,\infty]$, $\beta\geqq 1, d\geqq 2$.

The measure of interest now is

$$d\lambda(r)=\beta r^{d-2}\left(\frac{r}{\beta-e^{-r}}\right)^2 e^{-r}dr \quad \text{on } [0,\infty]. \tag{3.2}$$

Since for large r the distribution behaves like $d\lambda(r)\sim\beta^{-1}r^d e^{-r}dr$, we generated the recursion coefficients by the discretized Stieltjes procedure, using the Gauss-Laguerre quadrature rule to carry out the discretization (see [7] for a similar application). As a check, we also used a discretization based on the Fejér quadrature rule applied to each subinterval of the decomposition $[0,\infty]=[0,10]\cup[10,100]\cup[100,500]\cup[500,\infty]$ (cf. [5, §2.2]). In the case $\beta=1$, $d=3$ that we computed, the largest discrepancy observed was 1 unit in the 25th decimal place. The results are shown in Table 2 of the Appendix.

Example 3.3. The Fermi-Dirac distribution $f(r)=(\beta e^{r^2}+1)^{-1}$ on $[0,\infty]$, $\beta>0$.

Here,

$$d\lambda(r)=2\beta\frac{r^{d+1}e^{-r^2}}{(\beta+e^{-r^2})^2}dr \quad \text{on } [0,\infty].$$

For large r, this measure behaves similarly as the one in Example 3.1. Therefore, we used the same method as in Example 3.1 (the second one described there) to generate the recursion coefficients α_k, β_k. The results for $\beta=1$, $d=3$, are shown in Table 3 of the Appendix.

4. Approximation by Dirac Delta Functions

We now assume that $f\in C[0,\infty]$, $f(r)\geqq 0$ on $[0,\infty]$, and that the integrals $\int_0^\infty f(r)r^m dr$, $m=0,1,2,\ldots$, all exist and are finite, with $\int_0^\infty f(r)dr>0$.

Approximating f by a linear combination of Dirac delta functions,

$$f(r)\approx\tilde{f}(r), \quad \tilde{f}(r)=\sum_{\nu=1}^{n}a_\nu\delta(r-r_\nu), \tag{4.1}$$

and using the same moment-matching procedure as in Sect. 2, one is led immediately to the equations

$$\sum_{\nu=1}^{n}(a_\nu r_\nu^{d-1})r_\nu^j=\int_0^\infty[f(r)r^{d-1}]r^j dr, \quad j=0,1,\ldots,2n-1. \tag{4.2}$$

Thus, r_ν *are the Gaussian abscissas relative to the (nonnegative) measure* $d\mu(r) = r^{d-1} f(r)\,dr$, *and* $a_\nu = \mu_\nu r_\nu^{1-d}$, $\nu = 1, 2, \ldots, n$, *where* μ_ν *are the corresponding Christoffel numbers.*

For the Maxwell distribution of Example 3.1 this yields $d\mu(r) = \pi^{-d/2} r^{d-1} e^{-r^2} dr$, so that, when $d=3$, we can use the results of Example 3.1 relative to the case $d=1$, with an obvious modification of β_0. For the Bose-Einstein distribution, one has $d\mu(r) = r^{d-1}(\beta e^r - 1)^{-1} dr$, to which the numerical results of [7] apply, if $d=2$ and $\beta=1$. In other cases, one can easily adapt the methods used in Example 3.2. The same holds for the Fermi-Dirac distribution of Example 3.3.

Appendix (Tables 1–3)
see pages 58–60

References

1. Calder, A.C., Laframboise, J.G., Stauffer, A.D.: Optimum step-function approximation of the Maxwell distribution (unpublished)
2. Galant, D.: Gauss quadrature rules for the evaluation of $2\pi^{-\frac{1}{2}} \int_0^\infty \exp(-x^2) f(x)\,dx$. Math. Comput. **23** (1969), Review **42**, 676–677. Loose microfiche suppl. E
3. Galant, D.: An implementation of Christoffel's theorem in the theory of orthogonal polynomials. Math. Comput. **25**, 111–113 (1971)
4. Gautschi, W.: A survey of Gauss-Christoffel quadrature formulae. In: E.B. Christoffel, The Influence of his Work on Mathematics and the Physical Sciences (P.L. Butzer, F. Fehér, eds.), pp. 72–147. Basel: Birkhäuser 1981
5. Gautschi, W.: On generating orthogonal polynomials. SIAM J. Sci. Statist. Comput. **3**, 289–317 (1982)
6. Gautschi, W.: An algorithmic implementation of the generalized Christoffel theorem. In: Numerische Integration (G. Hämmerlin, ed.), pp. 89–106. Intern. Ser. Numer. Math. 57. Basel: Birkhäuser 1982
7. Gautschi, W., Milovanović, G.V.: Gaussian quadrature involving Einstein and Fermi functions with an application to convergence acceleration of series, submitted for publication.
8. Hildebrand, F.B.: Introduction to Numerical Analysis, 2nd ed., New York: McGraw-Hill 1974
9. Laframboise, J.G., Stauffer, A.D.: Optimum discrete approximation of the Maxwell distribution. AIAA Journal **7**, 520–523 (1969)
10. Szegö, G.: Orthogonal Polynomials, AMS Colloq. Publications, Vol. **23**, 4th ed. 2nd printing. Providence, R.I.: Amer. Math. Soc. 1978

Received July 5, 1983 / September 7, 1983

Appendix

Table 1. Recursion coefficients α_k, β_k, $0 \leqq k \leqq 49$, for the orthogonal polynomials relative to $d\lambda(r) = (2\pi^{-d/2}) r^{d+1} e^{-r^2} dr$ on $[0, \infty]$, $d = 1, 2, 3$

k	alpha(k)	d=1	beta(k)
0	1.12837916709551257389615 9d+00		5.00000000000000000000000d-01
1	1.35966303977353728816253 5d+00		2.26760455264837313848929 9d-01
2	1.57050685966162520985904 0d+00		4.24555963008947044546972 2d-01
3	1.76189656622377423648909 6d+00		6.08952240746833213673450 7d-01
4	1.93706156967307123825138 3d+00		7.86768249182040312617653 5d-01
5	2.09900068122281124538828 6d+00		9.61024226113657068297327 0d-01
6	2.25007811295547393114984 0d+00		1.13317191411251725904408 9d+00
7	2.39210114905658864499363 59d+00		1.30397657148621623790927 9d+00
8	2.52645775810874315262656 5d+00		1.47387552119593201202127 2d+00
9	2.65422974071936777065207 3d+00		1.64313567529621146080137 8d+00
10	2.77627463550050414354420 6d+00		1.81192880818458627695666 9d+00
11	2.89328343285258558440353 6d+00		1.98037034009195658092446 8d+00
12	3.00582135373188585573078 3d+00		2.14854063708880130270200 7d+00
13	3.11435703492043657663243 3d+00		2.31649735236061182153380 7d+00
14	3.21928375471469894797115 9d+00		2.48428290668097476804679 6d+00
15	3.32093512761387066463479 2d+00		2.65192919994905529103352 5d+00
16	3.41959689634575018339919 1d+00		2.81946067823118927234860 6d+00
17	3.51551592558701389709076 4d+00		2.98689638827132340364523 9d+00
18	3.60890715736024041842682 7d+00		3.15425138867275756417239 2d+00
19	3.69995905926075497434556 6d+00		3.32153774088127133809533 4d+00
20	3.78883794248459683039701 8d+00		3.48876521891299772181844 6d+00
21	3.87569142118627799439524 8d+00		3.65594182667608919878439 9d+00
22	3.96065121150184497837629 0d+00		3.82307418106699951058340 8d+00
23	4.04383541702560846095621 2d+00		3.99016779976196812548410 2d+00
24	4.12535041072162113969715 5d+00		4.15722732024735518273605 5d+00
25	4.20529239662322020021523 8d+00		4.32425666851159658771455 7d+00
26	4.28374871517211987043580 4d+00		4.49125919039133625789956 3d+00
27	4.36079894159977699291855 0d+00		4.65823775486987596946013 0d+00
28	4.43651581593218797403427 0d+00		4.82519483607238879806938 4d+00
29	4.51096603501158132052737 3d+00		4.99213257891163597349946 0d+00
30	4.58421093067425063914619 4d+00		5.15905285206072929057097 4d+00
31	4.65630705340349502994313 4d+00		5.32595729102599150990747 9d+00
32	4.72730667703013949816988 0d+00		5.49284733339887017338038 9d+00
33	4.79725823711805096474670 2d+00		5.65972424790739019573889 6d+00
34	4.86620671335548395503701 8d+00		5.82658915849561970921334 1d+00
35	4.93419396443174735081487 7d+00		5.99344306439839910493043 3d+00
36	5.00125902240532599575691 2d+00		6.16028685696691725415565 0d+00
37	5.06743835238324439066411 3d+00		6.32712133384240567541928 2d+00
38	5.13276608237051159978248 1d+00		6.49394721095298377355344 8d+00
39	5.19727420736570768595260 9d+00		6.66076513271388675760712 0d+00
40	5.26099277113770343977200 5d+00		6.82757568073726814625283 4d+00
41	5.32395002859081302649659 3d+00		6.99437938129955981036350 1d+00
42	5.38617259118920347290587 7d+00		7.16117671176832134736885 4d+00
43	5.44768555754868890977376 0d+00		7.32796810615386025182007 1d+00
44	5.50851263100132141720411 5d+00		7.49475395992157020075768 7d+00
45	5.56867622568447241592240 2d+00		7.66153463417732955152293 9d+00
46	5.62819756249259341356439 6d+00		7.82831045931920928457772 0d+00
47	5.68709675604948877267995 1d+00		7.99508173823322077261880 2d+00
48	5.74539289370601904634668 3d+00		8.16184874909816081424561 3d+00
49	5.80310410743804922396673 7d+00		8.32861174785421611385018 5d+00

k	alpha(k)	d=2	beta(k)
0	1.32934038817913702047362 6d+00		3.18309886183790671537767 5d-01
1	1.52510839211617326737044 3d+00		2.32854132355741303364763 1d-01
2	1.70983244800631686955396 3d+00		4.41190259941079382636294 9d-01
3	1.88249026306903723847805 5d+00		6.35282739803646336297367 1d-01
4	2.04393022006948993469547 1d+00		8.20947669646620636259171 5d-01
5	2.19547874008179108080001 0d+00		1.00140158584006580871620 1d+00
6	2.33843672005313984762910 0d+00		1.17847151600880543324101 9d+00
7	2.47393805844232930215422 1d+00		1.35324219029830782477397 3d+00
8	2.60293369962146918948672 6d+00		1.52638829269229022149455 6d+00
9	2.72621273950877376913347 3d+00		1.69834786268260098470878 9d+00
10	2.84443045519362768751616 3d+00		1.86941623625746583267092 3d+00
11	2.95813401105555555914049 6d+00		2.03979914930950615968719 2d+00
12	3.06778370638205944395128 4d+00		2.20964402303105075930294 7d+00
13	3.17376988598977502052571 3d+00		2.37905910782570330796747 4d+00
14	3.27642622868559197312429 9d+00		2.54812560295874713708406 7d+00
15	3.37604016946224805095393 3d+00		2.71690556502236183354444 8d+00
16	3.47286110320074992878732 1d+00		2.88544720915495362911700 8d+00
17	3.56710688135160661679222 1d+00		3.05378854872031652361168 0d+00
18	3.65896899447959346893565 3d+00		3.22195994829369855093537 0d+00
19	3.74861673868409550802061 9d+00		3.38998594914329414771370 8d+00
20	3.83620059141663857699321 8d+00		3.55788659731412146430489 4d+00
21	3.92185496775842097966069 9d+00		3.72567842510051094393195 2d+00
22	4.00570048755078211670762 4d+00		3.89337518676808379417 4208d+00
23	4.08784585339669837815327 9d+00		4.06098841726734265026918 8d+00
24	4.16838941678831245840663 2d+00		4.22852786160007610052203 1d+00
25	4.24742049246696004715549 4d+00		4.39600180840711622661024 3d+00
26	4.32502046812218451420208 5d+00		4.56341735176461036270464 0d+00
27	4.40126374661418665196336 7d+00		4.73078059855954956368711 9d+00
28	4.47621855027817216928501 2d+00		4.89809683418010303080868 0d+00
29	4.54994761096653045693791 1d+00		5.06537065596547562061932 2d+00
30	4.62250876488445129981723 7d+00		5.23260608149448639337278 8d+00
31	4.69395546766483361267055 5d+00		5.39980663707193814037222 1d+00
32	4.76433724227721810849840 9d+00		5.56697543050747502850515 0d+00
33	4.83370007009924051763133 0d+00		5.73411521134283729046647 1d+00
34	4.90208673366675583429656 9d+00		5.90122842097976001548088 4d+00
35	4.96953711816091327637362 3d+00		6.06831723462863683616344 0d+00
36	5.03608847751113453337120 6d+00		6.23538359659228824037524 0d+00
37	5.10177567003380158687507 0d+00		6.40242925008729476294894 9d+00
38	5.16663136774235945531146 9d+00		6.56945576256386010999710 8d+00
39	5.23068624282036893763839 6d+00		6.73646454729685590564925 2d+00
40	5.29396913421732792563810 3d+00		6.90345688187287649873911 9d+00
41	5.35650719688619646827802 1d+00		7.07043392408135888464492 1d+00
42	5.41832603581440608318862 5d+00		7.23739672562502317976772 4d+00
43	5.47944982669310583104413 1d+00		7.40434624399072022736241 6d+00
44	5.53990142481162541198367 8d+00		7.57128335276217224552848 5d+00
45	5.59970246354688872491897 1d+00		7.73820885060795042660684 7d+00
46	5.65887344363374620453662 1d+00		7.90512346913895492426268 4d+00
47	5.71743381424620233369654 0d+00		8.07202787979779169342701 8d+00
48	5.77540204678664970936920 7d+00		8.23892269991633061498300 8d+00
49	5.83279570216668541469303 1d+00		8.40580849805624658644573 4d+00

Table 1 (continued)

k	alpha(k) d=3	beta(k)
0	1.50450555612735009852821 2d+00	2.38732414637843003653325 6d-01
1	1.67676456897896715036815 0d+00	2.36463031581933002398097 6d-01
2	1.84235863723599013827811 4d+00	4.51997548634845510697040 5d-01
3	1.99999820030432538041693 1d+00	6.53717103167099781240251 5d-01
4	2.14965863667216479564362 5d+00	8.46290095612359792570781 6d-01
5	2.29184622102607721511519 5d+00	1.03267765016840999543642 6d+00
6	2.42723456684855126541260 1d+00	1.21476324900007922973096 7d+00
7	2.55650567481530430817210 4d+00	1.39376910849544648868398 5d+00
8	2.68028873516834793936808 6d+00	1.57050962614169905489403 8d+00
9	2.79914213927399350339191 8d+00	1.74554266998795854918013 6d+00
10	2.91355335782303494334117 4d+00	1.91926061415265260743167 4d+00
11	3.02394543942357798380364 4d+00	2.09194621696546549869097 6d+00
12	3.13068531064097966137895 1d+00	2.26380776242387835576191 3d+00
13	3.23409191879086260394300 6d+00	2.43500172331291432405151 9d+00
14	3.33444351109594745278233 5d+00	2.60564773749872224000199 7d+00
15	3.43198387833784804101422 1d+00	2.77583873381140791633472 4d+00
16	3.52692760669936065061009 9d+00	2.94564792501769513960175 5d+00
17	3.61946445677608238084546 3d+00	3.11513373208422485353538 0d+00
18	3.70976300401976348358933 6d+00	3.28434331405039402506629 6d+00
19	3.79797366618769660295194 8d+00	3.45331513995586627843355 2d+00
20	3.88423122668150204824377 6d+00	3.62208089098892065511290 2d+00
21	3.96865694480032556407427 9d+00	3.79066688668339319567978 5d+00
22	4.05136032759829420679492 0d+00	3.95909516780475245303148 8d+00
23	4.13244062406608277894417 3d+00	4.12738432815788979104481 4d+00
24	4.21198809080460413845242 9d+00	4.29555016041043451243939 7d+00
25	4.29008506897444502503547 3d+00	4.46360616250975129026053 1d+00
26	4.36680690475049425267499 4d+00	4.63156393845277998180678 9d+00
27	4.44222273945551205331749 6d+00	4.79943351817062783364300 3d+00
28	4.51639619069722342281046 6d+00	4.96722361489373804287838 8d+00
29	4.58938594194722229590293 1d+00	5.13494183376187143591559 4d+00
30	4.66124625487884819120852 5d+00	5.30259484209533570654987 0d+00
31	4.73202741626733116295213 2d+00	5.47018850928257609846693 0d+00
32	4.80177612922383711931466 6d+00	5.63772802241175006098030 3d+00
33	4.87053585688696292906955 4d+00	5.80521798240439382535101 5d+00
34	4.93834712535307129451765 8d+00	5.97266248437398393906288 0d+00
35	5.00524779152944857596217 6d+00	6.14006518514307321056644 4d+00
36	5.07127328069335754141081 6d+00	6.30742936024650447758768 3d+00
37	5.13645679779748261850570 8d+00	6.47475795227912597544524 5d+00
38	5.20082951594776685455172 1d+00	6.64205361208090478229277 6d+00
39	5.26442074496914184131628 4d+00	6.80931873396561233082635 5d+00
40	5.32725808254893281257048 3d+00	6.97655548597293330562680 9d+00
41	5.38936755009137905496263 7d+00	7.14376583594413444558967 3d+00
42	5.45077371511737550243223 9d+00	7.31095157407791242727137 4d+00
43	5.51149801791035039006748 4d+00	7.47811433250780796868345 9d+00
44	5.57156779094169974896316 7d+00	7.64525560234957350955136 9d+00
45	5.63099851076784302025095 3d+00	7.81237674859145441121329 9d+00
46	5.68981171925630360979673 9d+00	7.97947902313887968230308 4d+00
47	5.74802617921753454348915 8d+00	8.14656357627474679134424 9d+00
48	5.80565972672519096872461 2d+00	8.31363146675512466598343 6d+00
49	5.86272933365127288224537 5d+00	8.48068367072605625592384 9d+00

Table 2. Recursion coefficients α_k, β_k, $0 \leqq k \leqq 49$, for the orthogonal polynomials relative to $d\lambda(r) = r^{d-2}(r/(1-e^{-r}))^2 e^{-r} dr$ on $[0, \infty]$, $d=3$

k	alpha(k) d=3	beta(k)
0	3.60157071055875186151296 8d+00	7.21234141895756571239842 9d+00
1	5.60694617736899675209298 9d+00	4.28124530359196823321648 6d+00
2	7.64323449224082718523886 9d+00	1.02657366976386124979183 0d+01
3	9.67593540725853188776572 2d+00	1.81687822483505327340439 6d+01
4	1.17022549841574204972093 5d+01	2.80513592410012357458671 3d+01
5	1.37234388716497913725403 8d+01	3.99309423897161463971866 4d+01
6	1.57408106904824925767973 4d+01	5.38126987590820409022417 6d+01
7	1.77553406594379456423817 3d+01	6.96981425811362525353174 8d+01
8	1.97677074400745443629049 2d+01	8.75876047542754849489854 6d+01
9	2.17783902245112492477665 4d+01	1.07481004752111054883026 6d+02
10	2.37877347852261801529480 3d+01	1.29378119501057669183113 4d+02
11	2.57959964598976138156937 2d+01	1.53278685528809034268362 5d+02
12	2.78033679268517043141269 8d+01	1.79182439494876718922232 4d+02
13	2.98099974398470861275013 7d+01	2.07089133955679181301421 6d+02
14	3.18160010637674093044568 7d+01	2.36998542640757392571969 4d+02
15	3.38214710718735894892652 6d+01	2.68910461204901723963690 8d+02
16	3.58264818336580903202359 4d+01	3.02824706080362108923734 0d+02
17	3.78310940258521784217920 3d+01	3.38741112618175887611017 d+02
18	3.98353576992189940819874 8d+01	3.76659533060533687602602 6d+02
19	4.18393145491948024994489 5d+01	4.16579834584603334558224 4d+02
20	4.38429996225003932655754 5d+01	4.58501897515252324436693 0d+02
21	4.58464426175056333595561 9d+01	5.02425613736286842419229 6d+02
22	4.78496688875169458892538 6d+01	5.48350885298445545751233 9d+02
23	4.98527002237712429217698 5d+01	5.96277623208878666223038 3d+02
24	5.18555554729661350152664 1d+01	6.46205746382079786053326 3d+02
25	5.38582510290319347211336 9d+01	6.98135180731658352676376 3d+02
26	5.58608012282738183679340 0d+01	7.52065858383649585785246 5d+02
27	5.78632186695113622623626 3d+01	8.07997716994102984517688 7d+02
28	5.98655144754534127047954 9d+01	8.65930699155904595568894 8d+02
29	6.18676985076271638336970 9d+01	9.25864751881906034425933 6d+02
30	6.38697795442980001584206 5d+01	9.87799826153343658455914 3d+02
31	6.58717654286743733130651 9d+01	1.05173587652420059844184 8d+03
32	6.78736631930839354992921 8d+01	1.11767286077359642226775 d+03
33	6.98754791635889011043441 0d+01	1.18561073959950451214400 3d+03
34	7.18772190485776141295072 9d+01	1.25554947634812215772978 3d+03
35	7.38788801415199259199677 d+01	1.32748903677407982979302 2d+03
36	7.58804907485735852491547 9d+01	1.40142938882739914247911 6d+03
37	7.78820315175954045767769 3d+01	1.47737050246371578756420 d+03
38	7.98835142121637000966165 1d+01	1.55531234947479332074095 0d+03
39	8.18849238970197392504723 d+01	1.63525490233678224860463 9d+03
40	8.38863193099727317685391 8d+01	1.71719813907403994101956 d+03
41	8.58876479663385461741814 1d+01	1.80114203313663174990033 0d+03
42	8.78889311131037228486736 1d+01	1.88708656328989169605975 d+03
43	8.98901712895041032618637 4d+01	1.97503170851463993456163 2d+03
44	9.18913708408198260768638 8d+01	2.06497744891684032833554 4d+03
45	9.38925319370099491624105 5d+01	2.15692376564564015146495 d+03
46	9.58936565892053773132660 3d+01	2.25087064081886959958607 d+03
47	9.78947466643449354506882 d+01	2.34681805745519504483152 8d+03
48	9.98958038981966089948328 0d+01	2.44476599941221987389246 5d+03
49	1.01896829906970303773303 5d+02	2.54471445132991279380069 5d+03

Table 3. Recursion coefficients α_k, β_k, $0 \leqq k \leqq 49$, for the orthogonal polynomials relative to $d\lambda(r) = 2r^{d+1}(1+e^{-r^2})^{-2}e^{-r^2}dr$ on $[0, \infty]$, $d=3$

k	alpha(k) d=3	beta(k)
0	1.61721366181088227757278 8d+00	1.01714084272965151096846 3d+00
1	1.75713994950795538913306 7d+00	2.18061980915625611100199 6d-01
2	1.89840876440205654763324 6d+00	4.39919994187263242640967 1d-01
3	2.04113437133384689917768 3d+00	6.49461248920188517688662 4d-01
4	2.18175126296503180630016 4d+00	8.47496417278577834742490 4d-01
5	2.31809160387917701666790 1d+00	1.03738944052946265239465 3d+00
6	2.44945112928291214556372 9d+00	1.22176280892065268078985 1d+00
7	2.57580201354362564799287 8d+00	1.40229890744035784378252 6d+00
8	2.69738051595278069451968 3d+00	1.58008496105517196863829 1d+00
9	2.81451407181926027345343 0d+00	1.75584417001046877618805 1d+00
10	2.92754716744444284669611 3d+00	1.93007183052557567559172 0d+00
11	3.03681027583154982091619 7d+00	2.10311674873776337814751 4d+00
12	3.14260805223135855005467 8d+00	2.27523082755672140682977 8d+00
13	3.24521616255568105449755 2d+00	2.44660010126794373345753 8d+00
14	3.34488190906809980756985 7d+00	2.61736470884990548241611 4d+00
15	3.44182640764664740423273 5d+00	2.78763208930883983057032 1d+00
16	3.53624726853869396795382 5d+00	2.95748591173815234200027 5d+00
17	3.62832130023064764567747 2d+00	3.12699225283954579645715 9d+00
18	3.71820702903760119690712 9d+00	3.29620395621945199339944 5d+00
19	3.80604695855596913979875 2d+00	3.46516376466159620098222 4d+00
20	3.89196955545987367742636 0d+00	3.63390660797969122306476 6d+00
21	3.97609097623433622970060 5d+00	3.80246129925739794547311 4d+00
22	4.05851656062019174786688 5d+00	3.97085180974693215910867 4d+00
23	4.13934212047971963032791 9d+00	4.13909823914669687621579 6d+00
24	4.21865505187263587340273 3d+00	4.30721756257898993665256 8d+00
25	4.29653529558115719541433 7d+00	4.47522421178081185934695 3d+00
26	4.37305616824808649721052 3d+00	4.64313053175031659073077 2d+00
27	4.44828508322623740310737 1d+00	4.81094714280589844867247 0d+00
28	4.52228417741790847356419 7d+00	4.97868323007834203630550 3d+00
29	4.59511085789587069317145 6d+00	5.14634677680299214784863 4d+00
30	4.66681827995453006994470 0d+00	5.31394475370328609313649 7d+00
31	4.73745576641934109649373 0d+00	5.48148327378584147126483 3d+00
32	4.80706917650829189720152 4d+00	5.64896771967853364800607 0d+00
33	4.87570123125222526371275 1d+00	5.81640284901471483112214 2d+00
34	4.94339180140348592855837 4d+00	5.98379288214419228110354 9d+00
35	5.01017816286130358169260 7d+00	6.15114157552570207424933 7d+00
36	5.07609522388831190021149 5d+00	6.31845228344865415146658 1d+00
37	5.14117572776093738286645 6d+00	6.48572801018794554363964 6d+00
38	5.20545043396633523704951 0d+00	6.65297145427400794005895 7d+00
39	5.26894828061285632695071 7d+00	6.82018504623119423602155 8d+00
40	5.33169653034541960197453 2d+00	6.98737098087910863490820 5d+00
41	5.39372090173984549967095 5d+00	7.15453124508715184434328 3d+00
42	5.45504568788144870713468 4d+00	7.32166764171008139597889 8d+00
43	5.51569386360496768531293 8d+00	7.48878181030247417956045 8d+00
44	5.57568718267859817676280 8d+00	7.65587524510554264301948 3d+00
45	5.63504626604902242586355 7d+00	7.82294931071537239030818 3d+00
46	5.69379068212235145667298 8d+00	7.99000525577313449656318 8d+00
47	5.75139901993406402863091 6d+00	8.15704422496193705283014 0d+00
48	5.80950895595620708713294 2d+00	8.32406726954918738535816 8d+00
49	5.86651731519971464443358 3d+00	8.49107535667565521448380 0d+00

10.5. [100] "Spline Approximations to Spherically Symmetric Distributions"

---

[100] (with G. V. Milovanović) "Spline Approximations to Spherically Symmetric Distributions," *Numer. Math.* **49**, 111–121 (1986).

Numer. Math. 49, 111-121 (1986)

Numerische Mathematik

Spline Approximations to Spherically Symmetric Distributions*

Walter Gautschi[1] and Gradimir V. Milovanović

[1] Department of Computer Sciences, Computer Science Building, Room 164C, Purdue University, West Lafayette, Indiana 47907, USA

[2] Faculty of Electronic Engineering, Department of Mathematics, University of Niš, P.O. Box 73,18000 Niš, Yugoslavia

Summary. We discuss the problem of approximating a function f of the radial distance r in $\mathbb{R}^d$ on $0 \leqq r < \infty$ by a spline function of degree m with n (variable) knots. The spline is to be constructed so as to match the first $2n$ moments of f. We show that if a solution exists, it can be obtained from an n-point Gauss-Christoffel quadrature formula relative to an appropriate moment functional or, if f is suitably restricted, relative to a measure, both depending on f. The moment functional and the measure may or may not be positive definite. Pointwise convergence is discussed as $n \to \infty$. Examples are given including distributions from statistical mechanics.

Subject Classifications: AMS (MOS): 41A15, 65D32; 33A65; CR: G1.2.

1. Introduction

Following earlier work of Laframboise and Stauffer [10] and Calder, Laframboise and Stauffer [1], one of us in [8] considered the problem of approximating a function $f(r)$ of the radial distance $r = \|x\|$, $0 \leqq r < \infty$, in $\mathbb{R}^d$, $d \geqq 1$, by a piecewise constant function of r (and also by a linear combination of Dirac delta functions). The approximation was to preserve as many moments of f as possible. It was found that the problem can be solved by means of Gauss-Christoffel quadrature. Here we extend this work to spline approximation of arbitrary degree. Under suitable assumptions on f it will be shown that the problem has a unique solution if and only if certain Gaussian quadrature rules exist corresponding to a (possibly nonpositive) moment functional or weight distribution depending on f. Existence and uniqueness is assured if f is completely monotonic on $[0, \infty)$. Pointwise convergence of our approximation process depends on a convergence property of the Gauss-Christoffel quadrature rule. A number of examples are presented illustrating the quality of approximation.

* The work of the first author was supported in part by the National Science Foundation under grant DCR-8320561

2. Moment-Preserving Approximation by Spline Functions

A spline function of degree $m \geqq 0$ on the interval $0 \leqq r < \infty$, vanishing at $r = \infty$, with $n \geqq 1$ positive knots r_ν, $\nu = 1, 2, \ldots, n$, can be written in the form

$$s_n(r) = \sum_{\nu=1}^{n} a_\nu (r_\nu - r)_+^m, \qquad 0 \leqq r < \infty, \tag{2.1}$$

where a_ν are real numbers and the plus sign on the right is the cutoff symbol, $t_+ = t$ if $t > 0$ and $t_+ = 0$ if $t \leqq 0$. Given a function $f(r)$ on $0 \leqq r < \infty$, we wish to determine $s_n(r)$ such that

$$\int_0^\infty r^j s_n(r)\, dV = \int_0^\infty r^j f(r)\, dV, \qquad j = 0, 1, \ldots, 2n-1, \tag{2.2}$$

where $dV = [2\pi^{d/2}/\Gamma(d/2)]\, r^{d-1}\, dr$ is the volume element of the spherical shell in $\mathbb{R}^d$ if $d > 1$, and $dV = dr$ if $d = 1$. In other words, we want s_n to faithfully reproduce the first $2n$ spherical moments of f.

A first approach to this problem can be based on the moment functional

$$\mathscr{L} t^j = \mu_j, \qquad \mu_j = \frac{(j+d+m)!}{m!(j+d-1)!} \int_0^\infty r^{j+d-1} f(r)\, dr, \qquad j = 0, 1, 2, \ldots. \tag{2.3}$$

The functional $\mathscr{L}$, by virtue of (2.3), and being linear, is well defined for any polynomial, and therefore gives rise to the concept of orthogonality with respect to the functional $\mathscr{L}$: Two polynomials p and q are orthogonal with respect to $\mathscr{L}$ if $\mathscr{L}(p \cdot q) = 0$ (cf. [2, Chapter 1, Sect. 2]).

Theorem 2.1. *Given f with $\int_0^\infty r^{j+d-1} f(r)\, dr$, $j = 0, 1, \ldots, 2n-1$, finite, there exists a unique spline function s_n of the form (2.1) with distinct positive knots r_ν and satisfying (2.2) if and only if there exists a unique (monic) polynomial $\pi_n(\cdot\,; \mathscr{L})$ of degree n orthogonal with respect to $\mathscr{L}$ to all lower-degree polynomials and having zeros $r_\nu^{(n)}$, $\nu = 1, 2, \ldots, n$, that are all simple and positive. In that event, the knots r_ν and weights a_ν in (2.1) are given by*

$$r_\nu = r_\nu^{(n)}, \quad a_\nu = r_\nu^{-(m+d)} w_\nu, \qquad \nu = 1, 2, \ldots, n, \tag{2.4}$$

where $\{w_\nu\}$ is the (unique) solution of the Vandermonde system

$$\sum_{\nu=1}^{n} w_\nu r_\nu^j = \mu_j, \qquad j = 0, 1, \ldots, n-1. \tag{2.5}$$

Proof. Substituting (2.1) in (2.2) yields, since $r_\nu > 0$,

$$\sum_{\nu=1}^{n} a_\nu \int_0^{r_\nu} r^{j+d-1} (r_\nu - r)^m\, dr = \int_0^\infty r^{j+d-1} f(r)\, dr, \qquad j = 0, 1, \ldots, 2n-1. \tag{2.6}$$

Introducing on the left the new variable of integration t through $r=tr_\nu$ gives

$$\sum_{\nu=1}^{n} a_\nu r_\nu^{j+d+m} \int_0^1 t^{j+d-1}(1-t)^m\,dt = \int_0^\infty r^{j+d-1} f(r)\,dr.$$

The integral on the left is the well-known beta integral which can be expressed in terms of factorials. There results

$$\sum_{\nu=1}^{n} w_\nu r_\nu^j = \mu_j, \qquad j=0,1,\ldots,2n-1, \tag{2.7}$$

where

$$w_\nu = a_\nu r_\nu^{d+m}, \qquad \nu=1,2,\ldots,n, \tag{2.8}$$

and μ_j is given by (2.3). By virtue of the first relation in (2.3), the system of nonlinear equations (2.7) can be written in the form

$$\sum_{\nu=1}^{n} w_\nu p(r_\nu) = \mathscr{L} p, \qquad \text{all } p\in\mathbb{P}_{2n-1}, \tag{2.9}$$

which identifies r_ν and w_ν as the nodes and weights of the "Gaussian quadrature formula" for the functional $\mathscr{L}$. It is well known (see, e.g., [6, § 1.3]) that (2.9) is equivalent to the following two conditions:

(i) The formula (2.9) is interpolatory, i.e., valid for every $p\in\mathbb{P}_{n-1}$;

(ii) The node polynomial $\omega(r)=\prod_{\nu=1}^{n}(r-r_\nu)$ is orthogonal with respect to $\mathscr{L}$ to all polynomials of degree $<n$.

The second condition identifies ω as $\omega(\cdot)=\pi_n(\cdot;\mathscr{L})$ and the knots r_ν as the zeros of $\pi_n(\cdot;\mathscr{L})$. The first condition is equivalent to (2.5). □

It is well known that $\pi_n(\cdot;\mathscr{L})$ exists uniquely if and only if

$$\det\begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_n \\ \mu_1 & \mu_2 & \cdots & \mu_{n+1} \\ \cdots & \cdots & \cdots & \cdots \\ \mu_n & \mu_{n+1} & \cdots & \mu_{2n} \end{bmatrix} \tag{2.10}$$

While Theorem 2.1 is of some theoretical interest, it does not lend itself to constructive purposes because of the well-known ill-conditioning associated with power moments.

By further restricting the class of functions f, it is possible, however, to relate our problem to Gauss-Christoffel quadrature relative to an absolutely continuous measure supported on $[0,\infty]$ (and depending of f). Therefore, recently developed stable methods of constructing orthogonal polynomials (see, e.g., [7]) can be brought to bear upon the problem.

Theorem 2.2. *Let f be such that the integrals $\int_0^\infty r^{j+d-1} f(r)\,dr$, $j=0,1,\ldots,2n-1$, converge and, in addition, that*

$$f\in C^{m+1}[0,\infty], \qquad \lim_{r\to\infty} r^{2n-1+d+\mu} f^{(\mu)}(r)=0, \qquad \mu=0,1,\ldots,m. \tag{2.11}$$

Then a spline function s_n of the form (2.1) *with positive knots r_ν, that satisfies* (2.2), *exists and is unique if and only if the measure*

$$d\lambda(r)=\frac{(-1)^{m+1}}{m!}r^{m+d}f^{(m+1)}(r)\,dr \quad on \quad [0,\infty) \tag{2.12}$$

admits an n-point Gauss-Christoffel quadrature formula

$$\int_0^\infty p(r)\,d\lambda(r)=\sum_{\nu=1}^{n}\lambda_\nu^{(n)}p(r_\nu^{(n)}), \qquad p\in\mathbb{P}_{2n-1}, \tag{2.13}$$

with distinct positive nodes $r_\nu^{(n)}$. In that event, the knots r_ν and weights a_ν in (2.1) *are given by*

$$r_\nu=r_\nu^{(n)},\quad a_\nu=r_\nu^{-(m+d)}\lambda_\nu^{(n)}, \qquad \nu=1,2,\ldots,n. \tag{2.14}$$

Remark. The case $m=0$ of Theorem 2.2 has been obtained in [8].

Proof of Theorem 2.2. The left-hand side in (2.6), through m integrations by parts, can be seen to be equal to

$$\begin{aligned} &m![(j+d)(j+d+1)\cdots(j+d+m-1)]^{-1}\sum_{\nu=1}^{n}a_\nu\int_0^{r_\nu}r^{j+d+m-1}\,dr \\ &=m![(j+d)(j+d+1)\cdots(j+d+m)]^{-1}\sum_{\nu=1}^{n}a_\nu r_\nu^{j+d+m}. \end{aligned} \tag{2.15}$$

The integral on the right of (2.6) is transformed similarly by $m+1$ integrations by parts. We carry out the first of these in detail to exhibit the reasonings involved. We have, for any $b>0$,

$$\int_0^b r^{j+d-1}f(r)\,dr=(j+d)^{-1}r^{j+d}f(r)\Big|_0^b-(j+d)^{-1}\int_0^b r^{j+d}f'(r)\,dr.$$

The integrated term clearly vanishes at $r=0$ and tends to zero as $r=b\to\infty$ by the second assumption in (2.11) with $\mu=0$. Since $j\leqq 2n-1$ and the integral on the left converges by assumption, we conclude the convergence of the integral on the right as $b\to\infty$. Therefore,

$$\int_0^\infty r^{j+d-1}f(r)\,dr=-(j+d)^{-1}\int_0^\infty r^{j+d}f'(r)\,dr.$$

Continuing in this manner, using the second assumption in (2.11) to show convergence to zero of the integrated term at the upper limit (its value at $r=0$ always being zero) and the existence of $\int_0^\infty r^{j+d-1+\mu}f^{(\mu)}(r)\,dr$ already established to infer the existence of $\int_0^\infty r^{j+d+\mu}f^{(\mu+1)}(r)\,dr$, $\mu=1,2,\ldots,m$, we arrive at

$$\int_0^\infty r^{j+d-1}f(r)\,dr=(-1)^{m+1}[(j+d)(j+d+1)\cdots(j+d+m)]^{-1}\int_0^\infty r^{j+d+m}f^{(m+1)}(r)\,dr. \tag{2.16}$$

Comparing (2.16) with (2.15), we see that Eqs. (2.6), and hence Eqs. (2.2), are equivalent to

$$\sum_{\nu=1}^{n} (a_\nu r_\nu^{m+d}) r_\nu^j = \int_0^\infty \left[\frac{(-1)^{m+1}}{m!} r^{m+d} f^{(m+1)}(r) \right] r^j dr, \quad j=0,1,\ldots,2n-1.$$

These are precisely the conditions for r_ν to be the nodes of the Gauss-Christoffel formula (2.12), (2.13) and $a_\nu r_\nu^{m+d}$ their weights.

The nodes $r_\nu^{(n)}$, being the zeros of the orthogonal polynomial $\pi_n(\cdot; d\lambda)$ (if it exists), are uniquely determined, hence also the weights $\lambda_\nu^{(n)}$. □

If f is completely monotonic on $[0, \infty)$ (see, e.g., Widder [12, p. 145 ff.]) then $d\lambda(r)$ in (2.12) is a positive measure for every m. Moreover, the first $2n$ moments exist by virtue of the assumptions made on f in Theorem 2.2. The Gauss-Christoffel quadrature rule (2.13) therefore exists uniquely, all nodes $r_\nu^{(n)}$ being distinct and positive and all weights $\lambda_\nu^{(n)}$ positive. The latter implies $a_\nu > 0$, $\nu = 1, 2, \ldots, n$, in (2.1).

Theorem 2.3. *Given f as in Theorem 2.2, assume that the measure $d\lambda$ in (2.12) admits an n-point Gauss-Christoffel quadrature formula (2.13) with distinct positive nodes $r_\nu = r_\nu^{(n)}$. Define*

$$\sigma_r(t) = t^{-(m+d)}(t-r)_+^m. \tag{2.17}$$

Then, for any $r > 0$, we have for the error of the approximation (2.1), (2.2),

$$f(r) - s_n(r) = R_n(\sigma_r; d\lambda), \tag{2.18}$$

where $R_n(g; d\lambda)$ is the remainder term in the Gauss-Christoffel quadrature formula (2.12), (2.13),

$$\int_0^\infty g(t)\, d\lambda(t) = \sum_{\nu=1}^{n} \lambda_\nu^{(n)} g(r_\nu^{(n)}) + R_n(g; d\lambda). \tag{2.19}$$

Proof. By Taylor's formula, one has for any $b > 0$,

$$f(r) = f(b) + f'(b)(r-b) + \cdots + \frac{f^{(m)}(b)}{m!}(r-b)^m + \frac{1}{m!}\int_b^r (r-t)^m f^{(m+1)}(t)\, dt. \tag{2.20}$$

Since by (2.11), $\lim_{t\to\infty} t^\mu f^{(\mu)}(t) = 0$ for $\mu = 0, 1, \ldots, m$, we obtain from (2.20), letting $b \to \infty$,

$$f(r) = \frac{(-1)^{m+1}}{m!} \int_r^\infty (t-r)^m f^{(m+1)}(t)\, dt = \frac{(-1)^{m+1}}{m!} \int_0^\infty (t-r)_+^m f^{(m+1)}(t)\, dt,$$

hence, by (2.12) and (2.17),

$$f(r) = \int_0^\infty \sigma_r(t)\, d\lambda(t). \tag{2.21}$$

On the other hand, by (2.1) and (2.14),

$$s_n(r)=\sum_{\nu=1}^{n}\lambda_\nu r_\nu^{-(m+d)}(r_\nu-r)_+^m=\sum_{\nu=1}^{n}\lambda_\nu\sigma_r(r_\nu). \tag{2.22}$$

Subtracting (2.22) from (2.21) yields (2.18). □

To discuss convergence as $n\to\infty$ (for fixed m), we assume f to satisfy the assumptions of Theorem 2.2 for all $n=1,2,3,\ldots$. Then, by Theorem 2.3, our approximation process converges pointwise (at r), as $n\to\infty$, if and only if the Gauss-Christoffel quadrature formula (2.19) converges when applied to the special function $g(t)=\sigma_r(t)$ in (2.17). Since σ_r is uniformly bounded on $\mathbb{R}$, this is true, for example, if $d\lambda$ is a positive measure and the moment problem for $d\lambda$ on $[-\infty,\infty]$ (with $d\lambda(t)=0$ for $t<0$) is determined (cf. [4, Chapter 3, Theorem 1.1]).

3. Examples

We begin with, perhaps, the simplest example - the exponential distribution in $\mathbb{R}^d$. All computations reported in this section were done on the CDC 6500 computer in single precision (machine precision $\approx 3.55\times 10^{-15}$), except for Table 2, which was computed in double precision.

Example 3.1. $f(r)=c_d e^{-r}$ on $[0,\infty)$, where $c_1=1$, $c_d=\Gamma(d/2)/(2\Gamma(d)\pi^{d/2})$ if $d>1$.

For this distribution the measure (2.12) becomes the generalized Laguerre measure

$$d\lambda(r)=\frac{c_d}{m!}r^{m+d}e^{-r}dr,\qquad 0\leqq r<\infty. \tag{3.1}$$

The knots r_ν, therefore, are the zeros of the generalized Laguerre polynomial $L_n^{(\alpha)}$ with parameter $\alpha=m+d$, and the weights a_ν follow readily from (2.14) in terms of the corresponding Christoffel numbers $\lambda_\nu^{(n)}$. It is a straightforward matter to calculate the desired spline (2.1) for any value of m, d and n.

Table 1 shows approximate values of the resulting maximum absolute errors $\max_{0\leqq r\leqq r_n}|s_n(r)-f(r)|$, for $m=1,2,3$; $d=1,2,3$; and $n=5,10,20,40$. (Numbers in parentheses indicate decimal exponents.) Clearly, $|s_n(r)-f(r)|=f(r)$ for $r\geqq r_n$. Since the moment problem for the measure $d\lambda$ in (3.1) is determined (see, e.g., [4, Chapter 2, Theorem 5.2]), it follows from the remark at the end of Sect. 2 that $s_n(r)\to f(r)$ as $n\to\infty$, for any fixed $r>0$.

It is likely that convergence also takes place if n is fixed and $m\to\infty$. When $n=1$, for example,

$$s_1(r)=c_d\frac{(m+1)\cdots(m+d)}{(m+d+1)^d}\left(1-\frac{r}{m+d+1}\right)_+^m, \tag{3.2}$$

which implies $s_1(r)=c_d e^{-r}+O(m^{-1})$ as $m\to\infty$. For other values of n, and selected values of r (with $d=1$), the convergence behavior as $m\to\infty$ is illustrated in Table 2, which shows the respective absolute errors.

Our next example is the Bose-Einstein distribution; for simplicity we do not normalize it to have unit integral over space.

Example 3.2. $f(r)=(\alpha e^r-1)^{-1}$, $\alpha>1$ if $d=1$ and $\alpha\geqq 1$ if $d\geqq 2$.

Table 1. Accuracy of the spline approximation for Example 3.1

n	$d=1$			$d=2$			$d=3$		
	$m=1$	$m=2$	$m=3$	$m=1$	$m=2$	$m=3$	$m=1$	$m=2$	$m=3$
5	5.9 (-2)	1.8 (-2)	7.9 (-3)	2.4 (-2)	1.1 (-2)	5.9 (-3)	1.2 (-2)	6.5 (-3)	3.9 (-3)
10	1.8 (-2)	3.5 (-3)	1.0 (-3)	8.9 (-3)	2.7 (-3)	9.4 (-4)	5.0 (-3)	1.9 (-3)	7.6 (-4)
20	1.5 (-2)	1.2 (-3)	1.9 (-4)	2.8 (-3)	4.9 (-4)	1.0 (-4)	1.7 (-3)	3.9 (-4)	9.8 (-5)
40	7.5 (-3)	4.2 (-4)	4.7 (-5)	1.2 (-3)	7.6 (-5)	8.8 (-6)	5.1 (-4)	6.5 (-5)	9.2 (-6)

Table 2. Convergence behavior as $m\to\infty$ of the spline approximation for Example 3.1

m	$r=5$			$r=1.0$			$r=5.0$		
	$n=5$	$n=10$	$n=20$	$n=5$	$n=10$	$n=20$	$n=5$	$n=10$	$n=20$
5	7.3 (-4)	3.2 (-5)	3.3 (-7)	5.1 (-4)	1.6 (-5)	3.5 (-6)	1.2 (-4)	2.5 (-5)	1.3 (-6)
10	6.7 (-5)	6.3 (-7)	7.3 (-10)	4.4 (-5)	2.1 (-7)	3.3 (-9)	8.3 (-7)	1.8 (-8)	1.3 (-9)
20	4.0 (-6)	4.6 (-9)	2.4 (-13)	2.6 (-6)	1.5 (-9)	6.3 (-13)	2.1 (-7)	7.0 (-10)	4.6 (-13)
40	1.8 (-7)	1.5 (-11)	1.1 (-17)	1.2 (-7)	4.9 (-12)	2.2 (-17)	9.9 (-9)	1.8 (-12)	6.1 (-18)
80	7.0 (-9)	3.0 (-14)	1.0 (-22)	4.6 (-9)	9.5 (-15)	1.8 (-22)	3.7 (-10)	2.8 (-15)	2.8 (-23)

It can be shown by induction that

$$f^{(m+1)}(r)=(-1)^{m+1}f(r)\sum_{k=0}^{m+1}q_{m+1,k}[f(r)]^k,$$

where

$$\left.\begin{aligned} q_{1,0}&=q_{1,1}=1,\\ q_{\mu+1,0}&=q_{\mu,0},\\ q_{\mu+1,\kappa}&=\kappa q_{\mu,\kappa-1}+(\kappa+1)q_{\mu,\kappa}, \qquad \kappa=1,\ldots,\mu\\ q_{\mu+1,\mu+1}&=(\mu+1)q_{\mu,\mu} \end{aligned}\right\}\ \mu=1,\ldots,m.$$

The measure (2.12) thus becomes

$$d\lambda(r)=\frac{r^{m+d}}{m!}f(r)\sum_{k=0}^{m+1}q_{m+1,k}[f(r)]^k dr, \qquad 0\leqq r<\infty, \tag{3.3}$$

and is clearly positive. All moments of $d\lambda$ exist, if $\alpha>1$, for arbitrary $m\geqq 0$ and $d\geqq 1$. The same is true for $\alpha=1$, if $d\geqq 2$, since $d\lambda(r)\sim(m+1)r^{d-2}[r/(e^r-1)]^{m+2}$ as $r\to 0$. In these cases, the moment problem for $d\lambda$ is determined, since $d\lambda(r)\sim(\alpha m!)^{-1}r^{m+d}e^{-r}dr$ as $r\to\infty$ (cf. [4, Chapter 2, Theorem 5.2]), and therefore $s_n(r)\to f(r)$ as $n\to\infty$.

The function $f(r)$, however, is unbounded near the origin, when $\alpha=1$, which renders approximation by low-degree splines difficult. In the range where f is significant, and not too close to $r=0$, the accuracy attainable is typically about 1–10 percent.

Table 3. Relative accuracy of the spline approximation for Example 3.2

n	$m=1$			$m=2$			$m=3$		
	ν	r_ν	rel. err.	ν	r_ν	rel. err.	ν	r_ν	rel. err.
5	1	1.272		1	1.468		1	1.646	
	2	3.771	5.5(-1)	2	4.333	2.7(-1)	2	4.885	1.4(-1)
	3	7.152	7.5(-1)	3	7.992	4.8(-1)	3	8.821	2.9(-1)
10	1	0.597		1	0.664		1	0.724	
	2	1.910	3.6(-1)	2	2.142	1.7(-1)	2	2.350	6.6(-2)
	3	3.757	3.2(-1)	3	4.199	1.0(-1)	3	4.625	3.9(-2)
	4	6.057	4.3(-1)	4	6.677	1.6(-1)	4	7.286	6.4(-2)
20	1	0.271		1	0.293		1	0.313	
	2	0.895	3.1(-1)	2	0.971	1.4(-1)	2	1.037	5.5(-2)
	3	1.837	1.7(-1)	3	2.002	4.1(-2)	3	2.146	1.4(-2)
	4	3.049	1.6(-1)	4	3.328	3.2(-2)	4	3.584	9.1(-3)
	5	4.500	1.9(-1)	5	4.890	4.1(-2)	5	5.264	1.1(-2)
	6	6.187	2.4(-1)	6	6.675	5.8(-2)	6	7.151	1.6(-2)
40	1	0.123		1	0.131		1	0.137	
	2	0.412	2.9(-1)	2	0.436	1.3(-1)	2	0.458	5.1(-2)
	3	0.859	1.3(-1)	3	0.912	3.2(-2)	3	0.958	9.6(-3)
	4	1.458	8.8(-2)	4	1.551	1.2(-2)	4	1.631	4.4(-3)
	5	2.197	7.3(-2)	5	2.343	1.0(-2)	5	2.469	2.7(-3)
	6	3.065	8.0(-2)	6	3.271	1.0(-2)	6	3.456	2.2(-3)
	7	4.055	9.3(-2)	7	4.321	1.2(-2)	7	4.569	2.3(-3)
	8	5.164	1.1(-1)	8	5.487	1.6(-2)	8	5.796	3.0(-3)
	9	6.393	1.3(-1)	9	6.770	2.0(-2)	9	7.134	4.1(-3)

Maximum relative errors in some of the early intervals $[r_\nu, r_{\nu+1}]$, $\nu=1,2,3,\ldots$, are shown in Table 3 for $\alpha=1$, $d=3$, $1\leqq m\leqq 3$, and $n=5$, 10, 20, 40. The Gauss-Christoffel quadrature formula for the measure (3.3) was obtained by first computing the recursion coefficients of the respective orthogonal polynomials by a discretized Stieltjes procedure, similarly as in [8, Example 3.2], and then using well-known methods to compute the Gauss-Christoffel formula in terms of the eigensystem of the associated Jacobi matrix; see, e.g., [5, 9].

Our last example deals with the Maxwell velocity distribution treated previously in [8] for $m=0$.

Example 3.3. $f(r)=\pi^{-d/2}e^{-r^2}$ on $[0,\infty]$.

The measure (2.12) here becomes

$$d\lambda(r)=\frac{\pi^{-d/2}}{m!}r^{m+d}H_{m+1}(r)e^{-r^2}dr, \qquad 0\leqq r<\infty, \tag{3.4}$$

where H_{m+1} is the Hermite polynomial of degree $m+1$. If $m>0$, as we assume, H_{m+1} changes sign at least once on $(0,\infty)$, so that $d\lambda$ is no longer a positive measure. The existence of the Gauss-Christoffel quadrature formula (2.13) is therefore in doubt, and even if it exits, we cannot be sure that its nodes are all simple and positive as in the previous examples. The matter depends on

whether the nth degree orthogonal polynomial $\pi_n(\cdot; d\lambda)$ relative to $d\lambda$ exists, and in addition whether its zeros - the nodes $r_\nu^{(n)}$ in (2.13) - are distinct and positive. If so, the solution of our approximation problem is given by (2.14), where the $\lambda_\nu^{(n)}$ are uniquely determined by the nodes $r_\nu^{(n)}$; if not, the problem has no solution.

To resolve these issues computationally, we try to generate the recurrence relation (i.e., the coefficients α_k, β_k) for the (monic) orthogonal polynomials $\pi_k(\cdot) = \pi_k(\cdot; d\lambda)$,

$$\begin{aligned} &\pi_{k+1}(r) = (r-\alpha_k)\,\pi_k(r) - \beta_k\,\pi_{k-1}(r), \qquad k = 0, 1, \ldots, n-1, \\ &\pi_{-1}(r) = 0, \qquad \pi_0(r) = 1, \end{aligned} \tag{3.5}$$

by a discretized Stieltjes procedure; cf. [7, Example 4.6]. If the procedure does not break down, that is, $\beta_k \neq 0$ for $k = 0, 1, \ldots, n-1$, then $\pi_n(\cdot; d\lambda)$ exists uniquely. Its zeros $r_\nu^{(n)}$ are the eigenvalues of the (nonsymmetric) Jacobi matrix

$$J_n(d\lambda) = \begin{bmatrix} \alpha_0 & 1 & & & 0 \\ \beta_1 & \alpha_1 & 1 & & \\ & \beta_2 & \alpha_2 & \ddots & \\ & & \ddots & \ddots & 1 \\ 0 & & & \beta_{n-1} & \alpha_{n-1} \end{bmatrix}. \tag{3.6}$$

Since some of the β's are expected to be negative, we are not attempting to symmetrize the matrix J_n, as is customary, and possible, in the classical case of positive measures. From (3.5) it follows easily that the columns of the matrix

$$P_n(d\lambda) = [\pi_{\mu-1}(r_\nu^{(n)}; d\lambda)]_{\mu,\nu=1}^{n} \tag{3.7}$$

are the eigenvectors of $J_n(d\lambda)$, normalized to have the first component equal to 1. Putting in turn $p(r) = \pi_{\mu-1}(r; d\lambda)$, $\mu = 1, 2, \ldots, n$, in (2.13), and observing that $\int_0^\infty \pi_{\mu-1}(r)\,d\lambda(r) = \mu_0\,\delta_{\mu,1}$, where $\mu_0 = \int_0^\infty d\lambda(r)$ and $\delta_{\mu,1}$ is the Kronecker delta, one obtains for the vector $\lambda^T = [\lambda_1^{(n)}, \lambda_2^{(n)}, \ldots, \lambda_n^{(n)}]$ the system of linear algebraic equations

$$P_n(d\lambda)\,\lambda = \mu_0\,e_1, \qquad e_1^T = [1, 0, \ldots, 0]. \tag{3.8}$$

We have carried out the computation for the cases $m = 1, 2, 3$; $d = 1, 2, 3$; and $n = 1(1)20$. All coefficients β_k were found to be different from zero, but quite a few of them negative; see Table 4. Interestingly, the negative β's seem to occur in pairs of two.

With the α's and β's at hand, we used the EISPACK routine HQR2 [11, p. 248] to compute the eigenvalues and eigenvectors of $J_n(d\lambda)$ and, if all eigenvalues are positive, the LINPACK routines SGECO, SGESL [3, Chapter 1] to solve the system (3.8). A summary of the results is presented in Table 5. A dash indicates the presence of a negative eigenvalue and an asterisk the presence of a pair of conjugate complex eigenvalues. In all cases computed, there were never more than one negative eigenvalue or more than one pair of complex

Table 4. The sign of the coefficients β_k in (3.5) for Example 3.3

d	m	$\beta_k<0$ for $k=$
1	1	2-3, 6-7, 10-11, 15-16
	2	1-2, 4-5, 7-8, 11-12, 14-15, 18-19
	3	1-2, 4-5, 9-10, 16-17
2	1	3-4, 7-8, 12-13, 17-18
	2	2-3, 5-6, 8-9, 12-13, 15-16, 19
	3	1-2, 4-5, 10-11, 16-17
3	1	4-5, 8-9, 13-14, 18-19
	2	2-3, 6-7, 9-10, 13-14, 17-18
	3	2-3, 5-6, 10-11, 17-18

Table 5. Existence and accuracy of the spline approximation for Example 3.3

n	$d=1$			$d=2$			$d=3$		
	$m=1$	$m=2$	$m=3$	$m=1$	$m=2$	$m=3$	$m=1$	$m=2$	$m=3$
1	3.9(-2)	1.0(-1)	1.4(-1)	3.8(-2)	7.1(-2)	1.3(-1)	4.4(-2)	2.8(-2)	8.2(-2)
2	4.6(-2)	–	1.3(-1)	3.5(-2)	8.2(-2)	–	1.4(-2)	5.7(-2)	8.8(-2)
3	–	6.2(-3)	1.4(-3)	3.8(-2)	1.3(-1)	5.9(-3)	2.5(-2)	–	4.0(-2)
4	2.1(-2)	3.8(-3)	1.2(-3)	–	4.5(-3)	9.6(-4)	2.5(-2)	8.3(-3)	1.4(-3)
5	1.5(-2)	–	8.8(-4)	8.9(-3)	5.9(-3)	–	–	5.9(-3)	1.8(-3)
6	1.4(-2)	1.8(-3)	*	6.4(-3)	–	8.6(-4)	7.3(-3)	6.0(-3)	3.2(-3)
7	–	1.3(-3)	*	6.9(-3)	7.2(-4)	6.8(-4)	5.3(-3)	4.9(-2)	6.5(-4)
8	9.1(-3)	–	2.0(-4)	–	8.1(-4)	*	6.1(-3)	7.7(-4)	5.7(-4)
9	7.2(-3)	9.5(-4)	1.5(-4)	3.9(-2)	–	8.1(-5)	–	1.0(-3)	*
10	6.8(-3)	6.4(-4)	–	3.7(-3)	4.2(-4)	1.4(-4)	–	–	2.3(-1)
11	–	6.3(-4)	6.4(-5)	3.4(-3)	2.9(-4)	*	1.9(-3)	1.9(-4)	–
12	–	–	*	3.4(-3)	2.9(-4)	5.0(-5)	2.0(-3)	2.3(-4)	*
13	5.4(-3)	3.8(-4)	*	–	–	4.3(-5)	2.1(-3)	2.3(-4)	5.9(-5)
14	4.8(-3)	3.5(-4)	4.9(-5)	2.8(-3)	1.9(-4)	*	–	–	5.0(-5)
15	4.8(-3)	–	4.5(-5)	2.5(-3)	1.7(-4)	2.5(-5)	–	9.4(-5)	*
16	–	3.0(-4)	4.5(-5)	2.3(-3)	–	2.5(-5)	1.3(-3)	8.2(-5)	–
17	–	2.2(-4)	2.4(-5)	2.3(-3)	–	–	1.1(-3)	8.2(-5)	–
18	3.6(-3)	2.2(-4)	*	–	1.1(-4)	1.5(-5)	1.1(-3)	–	*
19	3.3(-3)	–	*	1.9(-3)	1.1(-4)	1.2(-5)	–	5.8(-5)	1.2(-5)
20	3.2(-3)	1.8(-4)	*	1.7(-3)	–	*	–	5.4(-5)	8.5(-6)

eigenvalues. The numbers shown in Table 5 represent (approximately) the maximum absolute errors, $\max_{0\leq r\leq r_n} |s_n(r)-f(r)|$; they are usually (but not always) attained at one of the early knots r_ν of the spline.

Unlike in the previous examples, the weights α_ν in (2.1) are no longer necessarily positive, the solution of (3.8) having components of either sign, in general.

Acknowledgment. The authors are indebted to the referee for pointing out that the approximation problem (2.1), (2.2) is equivalent to the system of nonlinear equations (2.7).

References

1. Calder, A.C., Laframboise, J.G., Stauffer, A.D.: Optimum step-function approximation of the Maxwell distribution (Unpublished)
2. Chihara, T.S.: An Introduction to Orthogonal Polynomials. New York: Gordon and Breach 1978
3. Dongarra, J.J., Bunch, J.R., Moler, C.B., Stewart, G.W.: LINPACK Users' Guide. Philadelphia: SIAM 1979
4. Freud, G.: Orthogonal Polynomials. New York: Pergamon Press 1981
5. Gautschi, W.: On generating Gaussian quadrature rules. In: Numerische Integration (G. Hämmerlin, ed.), pp. 147–154, ISNM Vol. 45. Basel: Birkhäuser 1979
6. Gautschi, W.: A survey of Gauss-Christoffel quadrature formulae. In: E.B. Christoffel (P.L. Butzer, F. Fehér, eds.), pp. 72–147. Basel: Birkhäuser 1981
7. Gautschi, W.: On generating orthogonal polynomials. SIAM J. Sci. Stat. Comput. **3**, 289–317 (1982)
8. Gautschi, W.: Discrete approximations to spherically symmetric distributions. Numer. Math. **44**, 53–60 (1984)
9. Golub, G.H., Welsch, J.H.: Calculation of Gauss quadrature rules. Math. Comput. **23**, 221–230 (1969)
10. Laframboise, J.G., Stauffer, A.D.: Optimum discrete approximation of the Maxwell distribution. AIAA J. **7**, 520–523 (1969)
11. Smith, B.T., Boyle, J.M., Garbow, B.S., Ikebe, Y., Klema, V.C., Moler, C.B.: Matrix Eigensystem Routines – EISPACK Guide. Lec. Notes Comput. Sci. Vol. 6. Berlin, Heidelberg, New York: Springer 1974
12. Widder, D.V.: The Laplace Transform. Princeton University Press 1941

Received August 21, 1985 / January 31, 1986

10.6. [102] "Moment-Preserving Spline Approximation on Finite Intervals"

---

[102] (with M. Frontini and G. V. Milovanović) "Moment-Preserving Spline Approximation on Finite Intervals," *Numer. Math.* **50**, 503–518 (1987).

Numer. Math. 50, 503–518 (1987)

Numerische Mathematik

Moment-Preserving Spline Approximation on Finite Intervals*

Marco Frontini[1], Walter Gautschi[2], and Gradimir V. Milovanović[3]

[1] Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci, 32, I-20133 Milano, Italy
[2] Department of Computer Sciences, Purdue University, West Lafayette, IN 47907, USA
[3] Faculty of Electronic Engineering, Department of Mathematics, P.O. Box 73, Y-18000 Niš, Yugoslavia

Summary. Continuing previous wotk, we discuss the problem of approximating a function f on the interval $[0,1]$ by a spline function of degree m, with n (variable) knots, matching as many of the initial moments of f as possible. Additional constraints on the derivatives of the approximation at one endpoint of $[0,1]$ may also be imposed. We show that, if the approximations exist, they can be represented in terms of generalized Gauss-Lobatto and Gauss-Radau quadrature rules relative to appropriate moment functionals or measures (depending on f). Pointwise convergence as $n\to\infty$, for fixed $m>0$, is shown for functions f that are completely monotonic on $[0,1]$, among others. Numerical examples conclude the paper.

Subject Classifications: AMS: Primary 41A15, 65D32; Secondary 33A65; CR: G1.2, G1.4.

1. Introduction

In previous papers [4, 6] two of us dealt with the problem of approximating a given function f on $[0,\infty]$ by a spline function of fixed degree (with variable knots) in such a way as to reproduce as many moments of f as possible. Having had in mind applications to physics, our functions $f=f(r)$ were considered functions of the radial distance $r=\|x\|$ of a vector $x\in\mathbb{R}^d$, and accordingly the moments were "spherical moments". We now wish to consider the analogous problem on an arbitrary finite interval. In this case, the interpretation of the independent variable as a radial distance is no longer meaningful, and our functions $f=f(t)$, therefore, are now simply functions of a real variable t on some given interval $[a,b]$. The case of a semi-infinite interval having been treated in our previous work, we restrict attention here to the case of a finite

* The work of the first author was supported by the *Ministero della Pubblica Istruzione* and by the *Consiglio Nazionale delle Ricerche*. The work of the second author was supported, in part, by the National Science Foundation under grant DCR-8320561

interval, which can be standardized to $[a, b]=[0, 1]$. The case of the whole real line, $[a, b]=\mathbb{R}$, is also of interest, as is the case of periodic splines. Both, however, appear to be less amenable to the type of analysis we are going to give, and will not be considered here.

2. Spline Approximation on [0, 1]

A spline function of degree $m \geqq 0$, with n (distinct) knots $\tau_1, \tau_2, \ldots, \tau_n$ in the interior of $[0, 1]$, can be written in terms of truncated powers in the form

$$s_{n,m}(t)=p_m(t)+\sum_{\nu=1}^{n} a_\nu(\tau_\nu-t)_+^m, \qquad 0 \leqq t \leqq 1, \tag{2.1}$$

where a_ν are real numbers and p_m is a polynomial of degree $\leqq m$. (Our choice of truncated powers distinguishes the right endpoint of $[0, 1]$ in the sense that $s_{n,m}(t) \equiv p_m(t)$, $t \geqq 1$.) We consider two related problems:

Problem I. Determine $s_{n,m}$ in (2.1) such that

$$\int_0^1 t^j s_{n,m}(t)\,dt=\int_0^1 t^j f(t)\,dt, \qquad j=0, 1, \ldots, 2n+m. \tag{2.2}$$

Problem I.* Determine $s_{n,m}$ in (2.1) such that

$$s_{n,m}^{(k)}(1)=f^{(k)}(1), \qquad k=0, 1, \ldots, m, \tag{2.3}$$

and such that (2.2) holds for $j=0, 1, \ldots, 2n-1$. Here we must assume that f has m derivatives at $t=1$, all being known.

Both problems will be solved in two ways: first in terms of moment functionals, then in terms of Gauss-Christoffel quadrature. The former approach requires only the existence and knowledge of the moments of f involved; the latter requires additional regularity of f, but lends itself better to stable implementations.

2.1. Solution of Problems I and I* by Moment Functionals

We first consider Problem I. Let

$$\mu_j=\frac{(m+j+1)!}{m!\,j!}\int_0^1 t^j f(t)\,dt, \qquad j=0, 1, \ldots, 2n+m, \tag{2.4}$$

where the moments of f on the right are assumed to exist. (They do, of course, if f is integrable on $[0, 1]$.) We define a linear functional $\mathscr{L}$ on the set of polynomials of the form $t^{m+1}p(t)$, $p \in \mathbb{P}_{2n+m}$, by

$$\mathscr{L}(t^{m+1} \cdot t^j)=\mu_j, \qquad j=0, 1, \ldots, 2n+m. \tag{2.5}$$

Then the inner product

$$(p, q) = \mathscr{L}(t^{m+1}(1-t)^{m+1} p \cdot q) \tag{2.6}$$

is well defined for any polynomials p, q for which $p \cdot q \in \mathbb{P}_{2n-1}$. In particular, we can define (if it exists) the monic polynomial $\pi_n(\cdot) = \pi_n(\cdot; \mathscr{L})$ of degree n orthogonal with respect to the inner product (2.6) to all polynomials of lower degree,

$$\begin{aligned} &\deg \pi_n = n, \qquad \pi_n(t) = t^n + \ldots, \\ &(\pi_n, q) = 0, \qquad \text{all } q \in \mathbb{P}_{n-1}. \end{aligned} \tag{2.7}$$

Theorem 2.1. *There exists a unique spline function on* $[0, 1]$,

$$s_{n,m}(t) = p_m(t) + \sum_{\nu=1}^{n} a_\nu (\tau_\nu - t)_+^m, \qquad 0 < \tau_\nu < 1, \ \tau_\nu \neq \tau_\mu \text{ for } \nu \neq \mu, \tag{2.8}$$

satisfying the $2n+m+1$ *moment equations* (2.2) *of Problem I if and only if the orthogonal polynomial* $\pi_n(\cdot) = \pi_n(\cdot; \mathscr{L})$ *in* (2.7) *exists uniquely and has n distinct real zeros* $\tau_\nu^{(n)}$, $\nu = 1, 2, \ldots, n$, *all contained in the open interval* $(0, 1)$. *The knots* τ_ν *in* (2.8) *are then precisely these zeros,*

$$\tau_\nu = \tau_\nu^{(n)}, \qquad \nu = 1, 2, \ldots, n, \tag{2.9}$$

while the coefficients a_ν *and the quantities*

$$b_k = \frac{(-1)^k}{m!} p_m^{(k)}(1), \qquad k = 0, 1, \ldots, m, \tag{2.10}$$

(which uniquely determine p_m *in* (2.8)) *are obtained uniquely from the linear system*

$$\mathscr{L}_0(t^{m+1} p) = \mathscr{L}(t^{m+1} p) \qquad \text{all } p \in \mathbb{P}_{n+m}, \tag{2.11}$$

where

$$\mathscr{L}_0(g) = \sum_{k=0}^{m} b_k g^{(m-k)}(1) + \sum_{\nu=1}^{n} a_\nu g(\tau_\nu), \qquad \tau_\nu = \tau_\nu^{(n)}. \tag{2.12}$$

Proof. Substituting (2.1) in (2.2), and observing that $0 < \tau_\nu < 1$, gives

$$\begin{aligned} &\int_0^1 t^j p_m(t)\, dt + \sum_{\nu=1}^{n} a_\nu \int_0^{\tau_\nu} t^j (\tau_\nu - t)^m\, dt = \int_0^1 t^j f(t)\, dt, \\ &\qquad j = 0, 1, \ldots, 2n+m. \end{aligned} \tag{2.13}$$

Changing variables, $t = \tau\, \tau_\nu$, in the ν-th integral of the summation, one obtains

$$\begin{aligned} \int_0^{\tau_\nu} t^j (\tau_\nu - t)^m\, dt &= \tau_\nu^{m+j+1} \int_0^1 \tau^j (1-\tau)^m\, d\tau \\ &= \frac{j!\, m!}{(m+j+1)!} \tau_\nu^{m+j+1}. \end{aligned} \tag{2.14}$$

Using m integrations by parts in the first integral of (2.13) yields

$$\int_0^1 t^j p_m(t)\,dt = \frac{j!\,m!}{(m+j+1)!} \sum_{k=0}^{m} b_k \left[\frac{d^{m-k}}{dt^{m-k}} t^{m+1+j}\right]_{t=1}, \tag{2.15}$$

where b_k is defined in (2.10). Inserting (2.14) and (2.15) in (2.13) and dividing through by $j!\,m!/(m+j+1)!$ gives

$$\mathscr{L}_0(t^{m+1}\cdot t^j) = \mu_j, \quad j=0,1,\ldots,2n+m,$$

where μ_j is defined by (2.4) and $\mathscr{L}_0$ by (2.12). Therefore, using (2.5) and the linearity of $\mathscr{L}_0$ and $\mathscr{L}$,

$$\mathscr{L}_0(t^{m+1}p) = \mathscr{L}(t^{m+1}p), \quad \text{all } p\in\mathbb{P}_{2n+m}. \tag{2.16}$$

Thus, the moment equations (2.2) and Eqs. (2.16) are equivalent.

Let now π_n denote the "knot polynomial"

$$\pi_n(t) = \prod_{\nu=1}^{n} (t-\tau_\nu) \tag{2.17}$$

having the knots τ_ν of the spline (2.8) as zeros. Then, by the definition of the inner product (2.6) we have, for any $q\in\mathbb{P}_{n-1}$,

$$(\pi_n, q) = \mathscr{L}(t^{m+1}(1-t)^{m+1}\pi_n\cdot q) = \mathscr{L}_0(t^{m+1}(1-t)^{m+1}\pi_n\cdot q), \tag{2.18}$$

by (2.16), since $(1-t)^{m+1}\pi_n\cdot q\in\mathbb{P}_{2n+m}$. Therefore, $(\pi_n, q)=0$ by the definition (2.12) of $\mathscr{L}_0$ and the fact that $\pi_n(\tau_\nu)=0$, $\nu=1,2,\ldots,n$. It follows that the knots τ_ν must be the zeros of the orthogonal polynomial $\pi_n(\cdot;\mathscr{L})$ of (2.7). This proves the necessity of the condition asserted in Theorem 2.1. Furthermore, the system (2.11) is a trivial consequence of (2.16); with $\tau_\nu=\tau_\nu^{(n)}$ determined, (2.11) is essentially a confluent Vandermonde system, hence nonsingular.

To prove the sufficiency of the condition, together with (2.11), we must show that they imply (2.16). Thus, let $p\in\mathbb{P}_{2n+m}$ be an arbitrary polynomial of degree $\leqq 2n+m$. Let q and r be the quotient and remainder of p upon division by $(1-t)^{m+1}\pi_n(t)$, where $\pi_n(\cdot)=\pi_n(\cdot;\mathscr{L})$,

$$p(t) = (1-t)^{m+1}\pi_n(t)\,q(t) + r(t), \quad q\in\mathbb{P}_{n-1},\ r\in\mathbb{P}_{n+m}. \tag{2.19}$$

Then,

$$\begin{aligned}
\mathscr{L}(t^{m+1}p) &= \mathscr{L}(t^{m+1}(1-t)^{m+1}\pi_n\cdot q) + \mathscr{L}(t^{m+1}r)\\
&= \mathscr{L}(t^{m+1}r) \qquad [\text{by (2.7)}]\\
&= \mathscr{L}_0(t^{m+1}r) \qquad [\text{by (2.11)}]\\
&= \mathscr{L}_0(t^{m+1}p) - \mathscr{L}_0(t^{m+1}(1-t)^{m+1}\pi_n\cdot q) \qquad [\text{by (2.19)}]\\
&= \mathscr{L}_0(t^{m+1}p) \qquad [\text{since } \pi_n(\tau_\nu)=0].
\end{aligned}$$

This proves (2.16). □

The solution of Problem I* can be effected similarly, if one observes, in view of $0<\tau_\nu<1$, that

$$s_{n,m}^{(k)}(1) = p_m^{(k)}(1), \quad k=0,1,\ldots,m. \tag{2.20}$$

By (2.3), therefore, $p_m^{(k)}(1)=f^{(k)}(1)$, $k=0, 1, \ldots, m$, so that the moment equations in question can now be written as

$$\sum_{\nu=1}^{n} a_\nu \int_0^{\tau_\nu} t^j(\tau_\nu - t)^m\, dt = \int_0^1 t^j \left[f(t) - \sum_{k=0}^{m} \frac{f^{(k)}(1)}{k!}(t-1)^k \right] dt, \quad j=0, 1, \ldots, 2n-1. \tag{2.21}$$

In analogy to (2.4) we define

$$\mu_j^* = \frac{(m+j+1)!}{m!\, j!} \int_0^1 t^j \left[f(t) - \sum_{k=0}^{m} \frac{f^{(k)}(1)}{k!}(t-1)^k \right] dt, \quad j=0, 1, \ldots, 2n-1, \tag{2.22}$$

which gives rise to the linear functional $\mathscr{L}^*$ on polynomials of the form $t^{m+1}p(t)$, $p\in\mathbb{P}_{2n-1}$, defined by

$$\mathscr{L}^*(t^{m+1}\cdot t^j)=\mu_j^*, \quad j=0, 1, \ldots, 2n-1, \tag{2.23}$$

and the inner product

$$(p, q)^* = \mathscr{L}^*(t^{m+1}\, p\cdot q), \quad p\cdot q\in\mathbb{P}_{2n-1}. \tag{2.24}$$

The orthogonal polynomial $\pi_n^*(\cdot)=\pi_n(\cdot; \mathscr{L}^*)$ is now defined by

$$\begin{aligned} &\deg \pi_n^* = n, \quad && \pi_n^*(t)=t^n+\ldots, \\ &(\pi_n^*, q)^* = 0, \quad && \text{all } q\in\mathbb{P}_{n-1}. \end{aligned} \tag{2.25}$$

Then the result for Problem I*, analogous to Theorem 2.1, is given by the following

Theorem 2.2. *There exists a unique spline function on* $[0, 1]$,

$$s_{n,m}^*(t) = p_m^*(t) + \sum_{\nu=1}^{n} a_\nu^*(\tau_\nu^* - t)_+^m, \quad 0<\tau_\nu^*<1,\ \tau_\nu^*\neq\tau_\mu^* \text{ for } \nu\neq\mu, \tag{2.26}$$

satisfying (2.3) *and the* $2n$ *moment equations of Problem I* if and only if the orthogonal polynomial* $\pi_n^*(\cdot)=\pi_n(\cdot; \mathscr{L}^*)$ *in* (2.25) *exists uniquely and has* n *distinct real zeros* $\tau_\nu^{(n)*}$, $\nu=1, 2, \ldots, n$, *all contained in the open interval* $(0, 1)$. *The knots* τ_ν^* *in* (2.26) *are then precisely these zeros,*

$$\tau_\nu^* = \tau_\nu^{(n)*}, \quad \nu=1, 2, \ldots, n, \tag{2.27}$$

the polynomial p_m^* *is given by*

$$p_m^*(t) = \sum_{k=0}^{m} \frac{f^{(k)}(1)}{k!}(t-1)^k, \tag{2.28}$$

and the coefficients a_ν^* *are obtained uniquely from the linear system*

$$\mathscr{L}_0^*(t^{m+1}\, p) = \mathscr{L}^*(t^{m+1}\, p), \quad \text{all } p\in\mathbb{P}_{n-1}, \tag{2.29}$$

where

$$\mathscr{L}_0^*(g)=\sum_{\nu=1}^{n} a_\nu^* g(\tau_\nu^*), \qquad \tau_\nu^*=\tau_\nu^{(n)*}. \tag{2.30}$$

The proof is entirely analogous to the proof of Theorem 2.1 and is omitted.

The functions $s_{n,m}$ and $s_{n,m}^*$ of Theorems 2.1 and 2.2 may be thought of as solutions of finite moment problems in terms of spline functions.

2.2. *Solution of Problems I and I* by Gauss-Christoffel Quadrature*

While the solution of Problems I, I* given in the previous subsection has some intrinsic mathematical interest, it is suspect, computationally, because of its reliance on the "moments" (2.4) and (2.22), which are likely to create ill-conditioning. For constructive purposes, it is better to reduce these problems to Gauss-Christoffel quadrature with respect to an absolutely continuous measure, as was similarly done in [4, 6]. This requires more regularity of f; we shall assume, in fact, that $f\in C^{m+1}[0,1]$. (This hypothesis could be slightly weakened.) We also assume that $f^{(k)}(1)$, $k=0,1,\ldots,m$, are known, and that $f\notin\mathbb{P}_m$ (otherwise, trivially, $s_{n,m}\equiv f$).

Again, we first consider Problem I. Applying (2.14), (2.15) and $m+1$ integrations by parts to the last integral in the moment equations (2.13) now results in

$$\sum_{k=0}^{m} b_k\left[\frac{d^{m-k}}{dt^{m-k}}t^{m+1+j}\right]_{t=1}+\sum_{\nu=1}^{n} a_\nu \tau_\nu^{m+1+j}$$
$$=\sum_{k=0}^{m}\phi_k\left[\frac{d^{m-k}}{dt^{m-k}}t^{m+1+j}\right]_{t=1}+\frac{(-1)^{m+1}}{m!}\int_0^1 f^{(m+1)}(t)\,t^{m+1+j}\,dt, \tag{2.31}$$
$$j=0,1,\ldots,2n+m,$$

where

$$b_k=\frac{(-1)^k p_m^{(k)}(1)}{m!}, \qquad \phi_k=\frac{(-1)^k f^{(k)}(1)}{m!}, \qquad k=0,1,\ldots,m. \tag{2.32}$$

Defining the measure

$$d\lambda_m(t)=\frac{(-1)^{m+1}}{m!}f^{(m+1)}(t)\,dt \quad \text{on } [0,1], \tag{2.33}$$

we can rewrite (2.31), similarly as in (2.16), in the form

$$\mathscr{L}_0(t^{m+1}p)=\mathscr{L}(t^{m+1}p), \qquad \text{all } p\in\mathbb{P}_{2n+m}, \tag{2.34}$$

where $\mathscr{L}_0$ is defined in (2.12), but $\mathscr{L}$ is now defined by

$$\mathscr{L}(g)=\sum_{k=0}^{m}\phi_k g^{(m-k)}(1)+\int_0^1 g(t)\,d\lambda_m(t). \tag{2.35}$$

The resolution of (2.34) is now verbatim the same as in the proof of Theorem 2.1, the inner product again being defined as in (2.6), but now with $\mathscr{L}$ given in (2.35). This yields

Theorem 2.3. *Assume that* $f\in C^{m+1}[0,1]$. *There exists a unique spline function* (2.8) *on* $[0,1]$ *satisfying the* $2n+m+1$ *moment equations* (2.2) *of Problem I if and only if the orthogonal polynomial* $\pi_n(\cdot)=\pi_n(\cdot;\mathscr{L})$ *in* (2.7) *relative to the inner product* (2.6), (2.35) *exists uniquely and has* n *distinct real zeros* $\tau_\nu^{(n)}$, $\nu=1,2,\ldots,n$, *all contained in the open interval* $(0,1)$. *The knots* τ_ν *in* (2.8) *are then precisely these zeros,*

$$\tau_\nu=\tau_\nu^{(n)}, \qquad \nu=1,2,\ldots,n, \tag{2.36}$$

while the coefficients a_ν, *and the quantities* b_k *in* (2.32) (*which uniquely determine* p_m *in* (2.8)), *are obtained uniquely from the linear system*

$$\mathscr{L}_0(t^{m+1}p)=\mathscr{L}(t^{m+1}p), \qquad \textit{all } p\in\mathbb{P}_{n+m}, \tag{2.37}$$

where $\mathscr{L}_0$, $\mathscr{L}$ *are defined, respectively, by* (2.12) *and* (2.35).

The result of Theorem 2.3 has been announced without proof in [5, § 3.3]. It can also be interpreted in terms of the generalized Gauss-Lobatto quadrature formula (relative to the measure $d\lambda_m$ in (2.33)),

$$\begin{aligned}\int_0^1 g(t)\,d\lambda_m(t)=&\sum_{k=0}^{m}[A_k g^{(k)}(0)+B_k g^{(k)}(1)]\\&+\sum_{\nu=1}^{n}\lambda_\nu^{(n)}g(\tau_\nu^{(n)})+R_{n,m}(g;d\lambda_m),\end{aligned} \tag{2.38}$$

where

$$R_{n,m}(g;d\lambda_m)=0, \qquad \text{all } g\in\mathbb{P}_{2n+2m+1}. \tag{2.39}$$

This quadrature formula, in turn, is known to be related to the Gauss-Christoffel quadrature formula

$$\int_0^1 g(t)\,d\sigma_m(t)=\sum_{\nu=1}^{n}\sigma_\nu^{(n)}g(\tau_\nu^{(n)})+R_n(g;d\sigma_m), \qquad R_n(\mathbb{P}_{2n-1};d\sigma_m)=0, \tag{2.40}$$

with respect to the measure

$$d\sigma_m(t)=t^{m+1}(1-t)^{m+1}\,d\lambda_m(t) \qquad \text{on } [0,1]. \tag{2.41}$$

Indeed, the nodes $\tau_\nu^{(n)}$ in (2.38) and (2.40) are the same (equal to the zeros of $\pi_n(\cdot;d\sigma_m)$), while the weights $\lambda_\nu^{(n)}$ in (2.38) are expressible in terms of those in (2.40) by

$$\lambda_\nu^{(n)}=[\tau_\nu^{(n)}(1-\tau_\nu^{(n)})]^{-(m+1)}\sigma_\nu^{(n)}, \qquad \nu=1,2,\ldots,n. \tag{2.42}$$

Furthermore, the coefficients A_k, B_k in (2.38) can be obtained from the linear system

$$R_{n,m}(p;d\lambda_m)=0, \qquad \text{all } p\in\mathbb{P}_{2m+1}. \tag{2.43}$$

Now we note that the inner product (2.6), in view of (2.35), can be written in the form

$$(p,q)=\int_0^1 t^{m+1}(1-t)^{m+1}p(t)\,q(t)\,d\lambda_m(t)=\int_0^1 p(t)\,q(t)\,d\sigma_m(t). \tag{2.6'}$$

Therefore, the knots τ_ν in (2.36) are precisely the nodes in (2.40), hence those in (2.38). Putting $g(t)=t^{m+1}p(t)$, $p\in\mathbb{P}_{2n+m}$, in (2.38) and noting (2.39) yields

$$\sum_{k=0}^{m} B_k \frac{d^k}{dt^k}[t^{m+1}p(t)]_{t=1} + \sum_{\nu=1}^{n} \lambda_\nu^{(n)}[\tau_\nu^{(n)}]^{m+1}p(\tau_\nu^{(n)}) = \int_0^1 t^{m+1}p(t)\,d\lambda_m(t), \quad \text{all } p\in\mathbb{P}_{2n+m},$$

which is identical to (2.34), if we identify

$$b_k-\phi_k=B_{m-k}, \quad k=0,1,\ldots,m; \quad a_\nu=\lambda_\nu^{(n)}, \quad \nu=1,2,\ldots,n.$$

Since under the assumptions of Theorem 2.3 the solution of (2.34) is unique, we have shown the following

Corollary 1 to Theorem 2.3. *If the conditions of Theorem 2.3 are satisfied, then the spline function* (2.8) *solving Problem I is given by*

$$\tau_\nu=\tau_\nu^{(n)}, \quad a_\nu=\lambda_\nu^{(n)}, \quad \nu=1,2,\ldots,n, \tag{2.44}$$

where $\tau_\nu^{(n)}$ *are the interior nodes of the generalized Gauss-Lobatto quadrature formula* (2.38) [*or the nodes of the Gauss-Christoffel formula* (2.40)] *and* $\lambda_\nu^{(n)}$ *the corresponding weights in* (2.38) [*or* (2.42)], *while*

$$p_m^{(k)}(1)=f^{(k)}(1)+(-1)^k m!\,B_{m-k}, \quad k=0,1,\ldots,m, \tag{2.45}$$

where B_{m-k} *is the coefficient of* $g^{(m-k)}(1)$ *in the Gauss-Lobatto formula* (2.38).

We remark that the conditions of Theorem 2.3 are satisfied for each $m=0,1,2,\ldots$ if f is completely monotonic on $[0,1]$ (cf. [8, p. 145ff.]), since $d\lambda_m$, and hence also $d\sigma_m$, is then a positive measure. We have, moreover, the following

Corollary 2 to Theorem 2.3. *If* f *is completely monotonic on* $[0,1]$ *and for some* $m\geqq 0$,

$$m!\,B_{m-\mu}+(-1)^\mu f^{(\mu)}(1)>0, \quad \mu=0,1,\ldots,m, \tag{2.46}$$

then so is $s_{n,m}$ *for each* $n\geqq 1$; *more precisely,*

$$(-1)^k s_{n,m}^{(k)}(t)\begin{cases}>0 & \text{if } k=0,1,\ldots,m,\\ =0 & \text{if } k>m,\end{cases} \tag{2.47}$$

for each $t\in[0,1]$ *for which* $s_{n,m}^{(k)}(t)$ *is defined.*

Proof. The assumption (2.46) implies $(-1)^\mu p_m^{(\mu)}(1)>0$, $\mu=0,1,\ldots,m$, hence the positivity on $[0,1]$ of $(-1)^k p_m^{(k)}(t)=(-1)^k\left[\sum_{\mu=0}^{m}(-1)^\mu \mu!^{-1}p_m^{(\mu)}(1)(1-t)^\mu\right]^{(k)}$ for $k=0,1,\ldots,m$. Since $a_\nu>0$, by (2.44), and $(-1)^k[(\tau_\nu-t)_+^m]^{(k)}\geqq 0$, $k=0,1,\ldots,m$, whenever the derivative exists, the assertion (2.47) follows. □

We note that (2.46) restricts only those $B_0, B_1, \ldots, B_m$ that are negative. In the case of the infinite interval $[0, \infty]$, considered in [6], the property (2.47) (with $\geqq$ in place of $>$) follows directly from (2.8), since $p_m(t) \equiv 0$.

Turning now to Problem I*, we note that (2.20) again implies $p_m^{(k)}(1) = f^{(k)}(1)$, hence $b_k = \phi_k$, $k = 0, 1, \ldots, m$. The moment equations in question thus simplify to

$$\mathscr{L}_0^*(t^{m+1} p) = \mathscr{L}^*(t^{m+1} p), \quad \text{all } p \in \mathbb{P}_{2n-1}, \tag{2.48}$$

where $\mathscr{L}_0^*$ is given by (2.30) and $\mathscr{L}^*$ by

$$\mathscr{L}^*(g) = \int_0^1 g(t)\, d\lambda_m(t). \tag{2.49}$$

The analogue of Theorem 2.2, therefore, is as follows.

Theorem 2.4. *Assume that $f \in C^{m+1}[0, 1]$. There exists a unique spline function* (2.26) *on* $[0, 1]$ *satisfying* (2.3) *and the $2n$ moment equations of Problem I* if and only if the orthogonal polynomial $\pi_n^*(\cdot) = \pi_n(\cdot; \mathscr{L}^*)$ in* (2.25) *relative to the inner product* (2.24), (2.49) *exists uniquely and has n distinct real zeros $\tau_\nu^{(n)*}$, $\nu = 1, 2, \ldots, n$, all contained in the open interval* $(0, 1)$. *The knots τ_ν^* in* (2.26) *are then precisely these zeros,*

$$\tau_\nu^* = \tau_\nu^{(n)*}, \quad \nu = 1, 2, \ldots, n, \tag{2.50}$$

the polynomial p_m^* is given by

$$p_m^*(t) = \sum_{k=0}^{m} \frac{f^{(k)}(1)}{k!} (t-1)^k, \tag{2.51}$$

and the coefficients a_ν^* are obtained uniquely from the linear system

$$\mathscr{L}_0^*(t^{m+1} p) = \mathscr{L}^*(t^{m+1} p), \quad \text{all } p \in \mathbb{P}_{n-1}, \tag{2.52}$$

where $\mathscr{L}_0^*$, $\mathscr{L}^*$ are defined, respectively, by (2.30) *and* (2.49).

Underlying Theorem 2.4 is now the generalized Gauss-Radau quadrature formula,

$$\begin{aligned} \int_0^1 g(t)\, d\lambda_m(t) &= \sum_{k=0}^{m} A_k^* g^{(k)}(0) + \sum_{\nu=1}^{n} \lambda_\nu^{(n)*} g(\tau_\nu^{(n)*}) + R_{n,m}^*(g; d\lambda_m), \\ R_{n,m}^*(g; d\lambda_m) &= 0, \quad \text{all } g \in \mathbb{P}_{2n+m}, \end{aligned} \tag{2.53}$$

or the related Gauss-Christoffel formula

$$\int_0^1 g(t)\, d\sigma_m^*(t) = \sum_{\nu=1}^{n} \sigma_\nu^{(n)*} g(\tau_\nu^{(n)*}) + R_n^*(g; d\sigma_m^*), \quad R_n^*(\mathbb{P}_{2n-1}; d\sigma_m^*) = 0 \tag{2.54}$$

for the measure

$$d\sigma_m^*(t) = t^{m+1}\, d\lambda_m(t) \quad \text{on } [0, 1]. \tag{2.55}$$

Again, the nodes $\tau_\nu^{(n)*}$ in (2.53) and (2.54) are identical, whereas

$$\lambda_\nu^{(n)*} = [\tau_\nu^{(n)*}]^{-(m+1)} \sigma_\nu^{(n)*}, \quad \nu = 1, 2, \ldots, n. \tag{2.56}$$

One has, in fact,

Corollary 1 to Theorem 2.4. *If the conditions of Theorem 2.4 are satisfied, then the spline function* (2.26) *solving Problem I* is given by* p_m^* *as in* (2.51) *and by*

$$\tau_\nu^* = \tau_\nu^{(n)*}, \qquad a_\nu^* = \lambda_\nu^{(n)*}, \qquad \nu = 1, 2, \ldots, n, \tag{2.57}$$

where $\tau_\nu^{(n)*}$ *are the interior nodes of the generalized Gauss-Radau formula* (2.53) [*or the nodes of the Gauss-Christoffel formula* (2.54)] *and* $\lambda_\nu^{(n)*}$ *the corresponding weights in* (2.53) [*or* (2.56)].

Corollary 2 to Theorem 2.4. *If* f *is completely monotonic on* $[0, 1]$ *then so is* $s_{n,m}^*$ *for each* $n \geqq 1$, $m \geqq 0$; *more precisely,*

$$(-1)^k s_{n,m}^{*(k)}(t) \begin{cases} > 0 & \text{if } k = 0, 1, \ldots, m, \\ = 0 & \text{if } k > m, \end{cases} \tag{2.58}$$

for each $t \in [0, 1]$ *for which* $s_{n,m}^{*(k)}(t)$ *is defined.*

The proofs are analogous to the proofs of Corollaries 1 and 2 to Theorem 2.3 and are omitted.

To obtain the Gauss-Christoffel formulae in question, one must be able to generate the orthogonal polynomials relative to the measures $d\sigma_m$ and $d\sigma_m^*$ in (2.41) and (2.55), respectively. For this, the methods discussed in [2] and [3] (see also [1, § 5]) are often helpful.

3. Error and Convergence of Approximation

Similarly as in [6], the error of the spline approximants $s_{n,m}$ and $s_{n,m}^*$ constructed in Sect. 2 can be expressed in terms of the quadrature error of the generalized Gauss-Lobatto and Gauss-Radau formulae (2.38) and (2.53), respectively, when applied to a special function. This is the content of the next two theorems.

Theorem 3.1. *Assume the conditions of Theorem 2.3 are satisfied. Then, for any* x *with* $0 < x < 1$, *the spline function* $s_{n,m}$ *in* (2.8), *solving Problem I, approximates* f *with an error given by*

$$f(x) - s_{n,m}(x) = R_{n,m}(\rho_x; d\lambda_m), \tag{3.1}$$

where $R_{n,m}(\cdot; d\lambda_m)$ *is the remainder term in the generalized Gauss-Lobatto quadrature formula* (2.38) (*relative to the measure* $d\lambda_m$ *in* (2.33)) *and* ρ_x *is given by*

$$\rho_x(t) = (t - x)_+^m, \qquad 0 \leqq t \leqq 1. \tag{3.2}$$

Alternatively, we have

$$f(x) - s_{n,m}(x) = R_n(\sigma_x; d\sigma_m), \tag{3.3}$$

where $R_n(\cdot; d\sigma_m)$ *is the remainder term in the Gauss-Christoffel quadrature formula* (2.40) (*relative to the measure* $d\sigma_m$ *in* (2.41)) *and* σ_x *is given by*

$$\sigma_x(t) = t^{-(m+1)}(1 - t)^{-(m+1)}[\rho_x(t) - q_{2m+1}(\rho_x; t)], \tag{3.4}$$

$q_{2m+1}(\rho_x;\cdot)$ *being the polynomial of degree* $\leqq 2m+1$ *interpolating to* ρ_x *and its first* m *derivatives* $\rho_x^{(k)}$, $k=1,2,\ldots,m$, *at* $t=0$ *and* $t=1$.

Proof. By Taylor's theorem,

$$\begin{aligned} f(x) &= \sum_{k=0}^{m} \frac{1}{k!} f^{(k)}(1)(x-1)^k + \frac{1}{m!}\int_1^x (x-t)^m f^{(m+1)}(t)\,dt \\ &= \sum_{k=0}^{m} \frac{1}{k!} f^{(k)}(1)(x-1)^k + \frac{(-1)^{m+1}}{m!}\int_x^1 (t-x)^m f^{(m+1)}(t)\,dt \\ &= \sum_{k=0}^{m} \frac{1}{k!} f^{(k)}(1)(x-1)^k + \int_0^1 \rho_x(t)\,d\lambda_m(t). \end{aligned}$$

By (2.44),

$$s_{n,m}(x) = \sum_{k=0}^{m} \frac{1}{k!} p_m^{(k)}(1)(x-1)^k + \sum_{\nu=1}^{n} \lambda_\nu^{(n)}(\tau_\nu^{(n)}-x)_+^m.$$

Subtracting this from the preceding equation gives

$$\begin{aligned} f(x)-s_{n,m}(x) = \int_0^1 \rho_x(t)\,d\lambda_m(t) &+ \sum_{k=0}^{m} \frac{1}{k!}[f^{(k)}(1)-p_m^{(k)}(1)](x-1)^k \\ &- \sum_{\nu=1}^{n} \lambda_\nu^{(n)}(\tau_\nu^{(n)}-x)_+^m, \end{aligned}$$

which, by virtue of (2.45) and (3.2), yields

$$f(x)-s_{n,m}(x) = \int_0^1 \rho_x(t)\,d\lambda_m(t) - \sum_{k=0}^{m} \frac{m!}{k!} B_{m-k}(1-x)^k - \sum_{\nu=1}^{n} \lambda_\nu^{(n)} \rho_x(\tau_\nu^{(n)}).$$

But

$$\rho_x^{(k)}(0)=0, \qquad \rho_x^{(k)}(1) = \frac{m!}{(m-k)!}(1-x)^{m-k}, \qquad k=0,1,\ldots,m,$$

so that

$$\begin{aligned} f(x)-s_{n,m}(x) &= \int_0^1 \rho_x(t)\,d\lambda_m(t) - \sum_{k=0}^{m} B_{m-k}\rho_x^{(m-k)}(1) - \sum_{\nu=1}^{n} \lambda_\nu^{(n)}\rho_x(\tau_\nu^{(n)}) \\ &= R_{n,m}(\rho_x; d\lambda_m), \end{aligned}$$

as claimed in (3.1).

To prove (3.3), it suffices to observe that for any function h that has zeros of multiplicity $m+1$ at $t=0$ and $t=1$ one obtains from (2.38), (2.40) and (2.42), by putting $g(t)=t^{-(m+1)}(1-t)^{-(m+1)}h(t)$ in (2.40), that

$$R_n(t^{-(m+1)}(1-t)^{-(m+1)}h; d\sigma_m) = R_{n,m}(h; d\lambda_m). \tag{3.5}$$

In particular, for $h(t)=\rho_x(t)-q_{2m+1}(\rho_x;t)$, since $R_{n,m}(q_{2m+1}; d\lambda_m)=0$, one gets $R_{n,m}(\rho_x; d\lambda_m)=R_{n,m}(\rho_x-q_{2m+1}; d\lambda_m)=R_n(\sigma_x; d\sigma_m)$, with σ_x given by (3.4). □

Theorem 3.2. *Assume the conditions of Theorem 2.4 are satisfied. Then, for any* x *with* $0<x<1$, *the spline function* $s_{n,m}^*$ *in* (2.26), *solving Problem I*, approximates* f *with an error given by*

$$f(x)-s^*_{n,m}(x)=R^*_{n,m}(\rho_x; d\lambda_m), \tag{3.6}$$

where $R^*_{n,m}(\cdot; d\lambda_m)$ *is the remainder term in the generalized Gauss-Radau quadrature formula* (2.53) (*relative to the measure* $d\lambda_m$ *in* (2.33)) *and* ρ_x *is given by* (3.2). *Alternatively, we have*

$$f(x)-s^*_{n,m}(x)=R^*_n(\sigma^*_x; d\sigma^*_m), \tag{3.7}$$

where $R^*_n(\cdot; d\sigma^*_m)$ *is the remainder term in the Gauss-Christoffel quadrature formula* (2.54) (*relative to the measure* $d\sigma^*_m$ *in* (2.55)) *and* σ^*_x *is given by*

$$\sigma^*_x(t)=t^{-(m+1)}\rho_x(t)=t^{-(m+1)}(t-x)^m_+, \qquad 0\leqq t\leqq 1. \tag{3.8}$$

Proof. Equation (3.6) is proved similarly as Eq. (3.1) in Theorem 3.1. The alternative formula (3.7) follows readily from $R^*_{n,m}(\rho_x; d\lambda_m)=R^*_n(\sigma^*_x; d\sigma^*_m)$. □

If $f\in C^{m+1}[0,1]$ is such that $d\lambda_m$ in (2.33) is a positive measure (for example, if f is completely monotonic on $[0,1]$), then the approximations $s_{n,m}$ and $s^*_{n,m}$ exist uniquely by Theorems 2.3 and 2.4, respectively. Moreover, for fixed $m>0$ and x, with $0<x<1$, we have

$$R_n(\sigma_x; d\sigma_m)\to 0 \qquad \text{as } n\to\infty$$

since σ_x is continuous on $[0,1]$ and $d\sigma_m$ is also a positive measure. Therefore, by (3.3), we have pointwise convergence $s_{n,m}\to f$ as $n\to\infty$. The analogous fact for $s^*_{n,m}$ follows likewise from (3.7) and the continuity of σ^*_x on $[0,1]$. Thus, we have

Theorem 3.3. *If* $f\in C^{m+1}[0,1]$ *and* $d\lambda_m$ *in* (2.33) *is a positive measure, then the approximations* $s_{n,m}$ *and* $s^*_{n,m}$ *constructed in Sect. 2 converge pointwise to* f *in* $(0,1)$, *as* $n\to\infty$ *for fixed* $m>0$.

We finally note that the formulas (3.1) and (3.3), resp. (3.6) and (3.7), by differentiating them repeatedly with respect to x, yield representations for the errors $f^{(k)}-s^{(k)}_{n,m}$ and $f^{(k)}-s^{*(k)}_{n,m}$ in the derivatives, respectively.

4. Examples

We illustrate the spline approximations of Theorems 2.3 and 2.4 (or their corollaries) in the case of exponential and trigonometric functions. All computations reported on were carried out on the CDC 6500 computer in single precision (machine precision $\approx 3.55\times 10^{-15}$).

Example 4.1. $f(t)=e^{-ct}$, $0\leqq t\leqq 1$, $c>0$.

This is an example of a completely monotonic function, for which the associated measure (2.33) is thus positive; indeed,

$$d\lambda_m(t)=\frac{c^{m+1}}{m!}e^{-ct}dt \qquad \text{on } [0,1]. \tag{4.1}$$

Problems I, I* therefore have unique solutions by Theorems 2.3 and 2.4. In terms of the generalized Gauss-Lobatto formula

$$\int_0^1 g(t)\,e^{-ct}\,dt=\sum_{k=0}^{m}[\bar{A}_k\,g^{(k)}(0)+\bar{B}_k\,g^{(k)}(1)]+\sum_{\nu=1}^{n}\bar{\lambda}_\nu^{(n)}\,g(\tau_\nu^{(n)})+\bar{R}_{n,m}(g;e^{-ct}\,dt) \tag{4.2}$$

and the generalized Gauss-Radau formula

$$\int_0^1 g(t)\,e^{-ct}\,dt=\sum_{k=0}^{m}\bar{A}_k^*\,g^{(k)}(0)+\sum_{\nu=1}^{n}\bar{\lambda}_\nu^{(n)*}\,g(\tau_\nu^{(n)*})+\bar{R}_{n,m}^*(g;e^{-ct}\,dt), \tag{4.3}$$

we have from Corollary 1 to Theorem 2.3,

$$\begin{gathered}\tau_\nu=\tau_\nu^{(n)},\qquad a_\nu=\frac{c^{m+1}}{m!}\,\bar{\lambda}_\nu^{(n)},\qquad \nu=1,2,\ldots,n,\\ p_m(t)=\frac{c^{m+1}}{m!}\sum_{k=0}^{m}\frac{m!}{k!}\,[c^{k-m-1}\,e^{-c}+\bar{B}_{m-k}]\,(1-t)^k\end{gathered} \tag{4.4}$$

for the spline $s_{n,m}$ in (2.8), solving Problem I, and from Corollary 1 to Theorem 2.4,

$$\begin{gathered}\tau_\nu^*=\tau_\nu^{(n)*},\qquad a_\nu^*=\frac{c^{m+1}}{m!}\,\bar{\lambda}_\nu^{(n)*},\qquad \nu=1,2,\ldots,n,\\ p_m^*(t)=\frac{c^{m+1}}{m!}\sum_{k=0}^{m}\frac{m!}{k!}\,c^{k-m-1}\,e^{-c}(1-t)^k\end{gathered} \tag{4.5}$$

for the spline $s_{n,m}^*$ in (2.26), solving Problem I*.

The Gaussian nodes and weights in (4.2) and (4.3) were obtained in the usual way (see, e.g., [2, p. 290]) in terms of the eigensystems of the Jacobi matrices $J_n(t^{m+1}(1-t)^{m+1}e^{-ct}\,dt)$ and $J_n(t^{m+1}e^{-ct}\,dt)$, respectively. The latter were generated from the Jacobi matrix $J_{n+2m+2}(e^{-ct}\,dt)$, resp. $J_{n+m+1}(e^{-ct}\,dt)$, by repeated application of the algorithms in [3, §4.1] corresponding to multiplication of a measure by $t(1-t)$ and t, respectively. (Alternatively, algorithms based on the QR algorithm, as in [7], could be used for the same purpose.) Finally, $J_{n+2m+2}(e^{-ct}\,dt)$ was computed by the discretized Stieltjes algorithm (see [2, §2.2]), the Fejér quadrature rule having been used as the modus of discretization.

As to the coefficients $\bar{A}_k$, $\bar{B}_k$ in the boundary terms of (4.2), they were computed from the linear system of equations

$$\bar{R}_{n,m}(p;e^{-ct}\,dt)=0,\qquad \text{all } p\in\mathbb{P}_{2m+1}, \tag{4.6}$$

where the first $2m+2$ orthogonal polynomials $\{\pi_k(\cdot;e^{-ct}\,dt)\}_{k\geq 0}$ (whose Jacobi matrix J_{n+2m+2} has already been generated!) were used as basis in the polynomial space $\mathbb{P}_{2m+1}$ of (4.6). The coefficients $\bar{A}_k^*$ in (4.3) are not needed.

The accuracy of the spline approximations $s_{n,m}$ and $s_{n,m}^*$ thus obtained is shown in Table 4.1 for $n=5$, 10, 20, 40; $m=0(1)3$; and $c=1$, 2, 4. Displayed are (two-digit approximations to) the respective maximum absolute errors on $[0,1]$.

Table 4.1. Accuracy of the spline approximations $s_{n,m}$ and $s^*_{n,m}$ for Example 4.1. (Numbers in parentheses denote decimal exponents).

c	n	$\max_{0\leq t\leq 1} \lvert s_{n,m}(t)-e^{-ct}\rvert$				$\max_{0\leq t\leq 1} \lvert s^*_{n,m}(t)-e^{-ct}\rvert$			
		$m=0$	$m=1$	$m=2$	$m=3$	$m=0$	$m=1$	$m=2$	$m=3$
1	5	8.0 (−2)	2.4 (−3)	4.0 (−5)	9.7 (−7)	8.8 (−2)	3.3 (−3)	6.8 (−5)	2.4 (−6)
	10	4.6 (−2)	8.6 (−4)	8.6 (−6)	1.4 (−7)	4.8 (−2)	1.0 (−3)	1.2 (−5)	2.5 (−7)
	20	2.5 (−2)	2.6 (−4)	1.5 (−6)	1.5 (−8)	2.5 (−2)	2.9 (−4)	1.9 (−6)	2.1 (−8)
	40	1.3 (−2)	7.3 (−5)	2.4 (−7)	1.4 (−9)	1.3 (−2)	7.7 (−5)	2.7 (−7)	1.6 (−9)
2	5	1.3 (−1)	7.0 (−3)	2.1 (−4)	9.8 (−6)	1.3 (−1)	9.1 (−3)	3.8 (−4)	2.4 (−5)
	10	7.1 (−2)	2.4 (−3)	4.6 (−5)	1.5 (−6)	7.5 (−2)	2.8 (−3)	6.5 (−5)	2.6 (−6)
	20	3.9 (−2)	7.4 (−4)	8.4 (−6)	1.6 (−7)	4.0 (−2)	8.1 (−4)	1.0 (−5)	2.3 (−7)
	40	2.0 (−2)	2.1 (−4)	1.3 (−6)	1.4 (−8)	2.0 (−2)	2.2 (−4)	1.4 (−6)	1.7 (−8)
4	5	1.7 (−1)	1.6 (−2)	8.7 (−4)	7.8 (−5)	1.7 (−1)	1.9 (−2)	1.5 (−3)	2.5 (−4)
	10	1.0 (−1)	5.7 (−3)	2.0 (−4)	1.1 (−5)	1.1 (−1)	6.7 (−3)	2.7 (−4)	2.0 (−5)
	20	5.7 (−2)	1.8 (−3)	3.6 (−5)	1.3 (−6)	5.8 (−2)	2.0 (−3)	4.3 (−5)	1.8 (−6)
	40	3.0 (−2)	5.1 (−4)	5.7 (−6)	1.2 (−7)	3.0 (−2)	5.3 (−4)	6.2 (−6)	1.4 (−7)

For $m=0$, 1, and 3, the maxima are almost always attained at a knot of the spline, about half-way (or somewhat less) through the interval. The only exception observed was for $s^*_{n,m}$, $n=5$, $m=3$, $c=4$, where the maximum occurs at $t=0$. When $m=2$, the maxima are usually attained between two such knots. The linear system (4.6) (in the orthogonal basis mentioned) was found to be relatively well-conditioned, the worst condition number (occurring for $m=3$) being approx. 2.5×10^3.

It is seen that the approximation error is more easily reduced by increasing m rather than n. Also, the spline $s_{n,m}$ is only marginally more accurate than the spline $s^*_{n,m}$. The additional effort required in computing $s_{n,m}$, therefore, seems hardly justified, if uniform approximation is indeed the main objective. If moment-matching is more important, however, the spline $s_{n,m}$ would be preferable, as it matches $m+1$ additional moments.

The coefficients of p_m, i.e., the expressions in the brackets of (4.4), turned out to be positive for all values of m, n and c tried, so that the computed splines $s_{n,m}$ are completely monotonic in the sense of (2.47). The analogous property for $s^*_{n,m}$ follows from Corollary 2 to Theorem 2.4.

Example 4.2. $f(t)=\sin\dfrac{\pi}{2}t$, $0\leqq t\leqq 1$.

Here, the function f, though not completely monotonic, still has derivatives that are all of constant sign on [0, 1]. Therefore, the measure $d\lambda_m$ in (2.33), i.e.,

$$d\lambda_m(t)=\frac{(-1)^{[m/2]+1}}{m!}\left(\frac{\pi}{2}\right)^{m+1}\begin{Bmatrix}\cos\dfrac{\pi}{2}t\\ \\ \sin\dfrac{\pi}{2}t\end{Bmatrix}dt \qquad \text{on } [0,1], \tag{4.7}$$

where the cosine or sine is taken according as m is even or odd, admits a unique system of (monic) orthogonal polynomials, and Problems I and I* both have unique solutions for each m and n. Observing that the substitution $t \to 1-t$ carries the cosine into the sine, and vice versa, it suffices to generate the orthogonal polynomials for one of the trigonometric measures only, say $\cos((\pi/2)t)\,dt$. If α_k^c, β_k^c, are the coefficients in the corresponding recurrence relation

$$\begin{aligned} \pi_{k+1}(t) &= (t-\alpha_k)\,\pi_k(t) - \beta_k\,\pi_{k-1}(t), \qquad k=0,1,2,\ldots, \\ \pi_{-1}(t) &= 0, \qquad \pi_0(t)=1, \end{aligned} \tag{4.8}$$

then the coefficients α_k^s, β_k^s for the sine-measure are

$$\alpha_k^s = 1-\alpha_k^c, \qquad \beta_k^s = \beta_k^c, \qquad k=0,1,2,\ldots. \tag{4.9}$$

A similar remark applies to the generalized Lobatto measures (2.41) [but not to the generalized Radau measures (2.55)]. The constants multiplying the trigonometric measures in (4.7), of course, simply give rise to analogous multiplicative constants in the quadrature rules (2.38) and (2.53).

Techniques similar to those in Example 4.1 were used to compute the desired spline approximants in the present example.

Table 4.2. Accuracy of the spline approximations $s_{n,m}$ and $s_{n,m}^*$ for Example 4.2.

n	$\max_{0\le t\le 1}\left\lvert s_{n,m}(t)-\sin\frac{\pi}{2}t\right\rvert$				$\max_{0\le t\le 1}\left\lvert s_{n,m}^*(t)-\sin\frac{\pi}{2}t\right\rvert$			
	$m=0$	$m=1$	$m=2$	$m=3$	$m=0$	$m=1$	$m=2$	$m=3$
5	1.4 (−1)	6.5 (−3)	1.7 (−4)	6.2 (−6)	1.5 (−1)	8.8 (−3)	2.7 (−4)	1.5 (−5)
10	8.4 (−2)	2.4 (−3)	3.7 (−5)	9.4 (−7)	8.8 (−2)	2.8 (−3)	5.0 (−5)	1.6 (−6)
20	4.6 (−2)	7.6 (−4)	6.8 (−6)	1.1 (−7)	4.7 (−2)	8.2 (−4)	8.2 (−6)	1.4 (−7)
40	2.4 (−2)	2.1 (−4)	1.1 (−6)	9.6 (−9)	2.4 (−2)	2.2 (−4)	1.2 (−6)	1.1 (−8)

Their accuracy is shown in Table 4.2; the error behaves rather similarly as the error in Example 4.1 for $c=2$.

References

1. Gautschi, W.: A survey of Gauss-Christoffel quadrature formulae. In: E.B. Christoffel, Butzer, P.L., Fehér, F. (eds.), pp. 72–147. Basel: Birkhäuser 1981
2. Gautschi, W.: On generating orthogonal polynomials. SIAM J. Sci. Stat. Comput. **3**, 289–317 (1982)
3. Gautschi, W.: An algorithmic implementation of the generalized Christoffel theorem. In: Numerische Integration. Hämmerlin, G. (ed.). Internat. Ser. Numer. Math., vol. 57, pp. 89–106. Basel: Birkhäuser 1982

4. Gautschi, W.: Discrete approximations to spherically symmetric distributions. Numer. Math. **44**, 53–60 (1984)
5. Gautschi, W.: Some new applications of orthogonal polynomials. In: Polynômes orthogonaux et applications. Brezinski, C., Draux, A., Magnus, A.P., Maroni, P., Ronveaux, A. (eds.). Lecture Notes Math., vol. 1171, pp. 63–73. Berlin-Heidelberg-New York-Tokyo: Springer 1985
6. Gautschi, W., Milovanović, G.V.: Spline approximations to spherically symmetric distributions. Numer. Math. **49**, 111–121 (1986)
7. Golub, G.H., Kautsky, J.: Calculation of Gauss quadratures with multiple free and fixed knots. Numer. Math. **41**, 147–163 (1983)
8. Widder, D.V.: The Laplace Transform. Princeton: University Press 1941

Received July 21, 1986/November 17, 1986

10.7. [132] “On mean convergence of extended Lagrange interpolation”

---

[132] “On mean convergence of extended Lagrange interpolation,” *J. Comput. Appl. Math.* **43**, 19–35 (1992).

Journal of Computational and Applied Mathematics 43 (1992) 19–35
North-Holland

CAM 1240

On mean convergence of extended Lagrange interpolation

Walter Gautschi

Department of Computer Sciences, Purdue University, West Lafayette, IN 47907, United States

Received 22 February 1991
Revised 11 November 1991

Abstract

Gautschi, W., On mean convergence of extended Lagrange interpolation, Journal of Computational and Applied Mathematics 43 (1992) 19–35.

Lagrange interpolation to any continuous function on $[-1, 1]$ at the zeros of orthogonal polynomials is known to converge in the mean. Here, following Bellen, we study mean convergence of Lagrange interpolation on an extended set of nodes that includes, in addition to the n zeros of the orthogonal (relative to some positive weight function w) polynomial π_n of degree n, other $n+1$ nodes, which in turn are zeros of an orthogonal polynomial $\hat{\pi}_{n+1}$ of degree $n+1$ corresponding to the weight function $\hat{w}_n = \pi_n^2 w$. A sufficient criterion of Bellen for mean convergence (as $n \to \infty$) of such extended Lagrange interpolation, for arbitrary continuous functions, is shown to fail for Chebyshev weight functions of the first, third and fourth kind. (It holds trivially for Chebyshev weights of the second kind.) Based on extensive computations, it is conjectured, on the other hand, that the criterion is satisfied for certain Jacobi weights with parameters α and β suitably restricted. Necessary conditions for mean convergence, due to Erdős and Turán, are shown to be violated for the three kinds of Chebyshev weights mentioned above. For smooth functions, a comparison is made of the speed of convergence of simple vs. extended Lagrange interpolation.

Keywords: Extended Lagrange interpolation; convergence in the mean; orthogonal polynomials.

1. Introduction

Let $\pi_n(\cdot; w)$, $n \geqslant 1$, denote the nth-degree orthogonal polynomial on $(-1, 1)$ with respect to a positive weight function w. It is well known [6] that the Lagrange polynomial $(L_n f)(\cdot)$ of degree $\leqslant n-1$ interpolating f at the n zeros $\tau_i = \tau_i^{(n)}$ of π_n converges to f in the mean whenever f is a continuous function on $[-1, 1]$,

$$\lim_{n \to \infty} \| f - L_n f \|_w = 0, \quad \text{all } f \in C[-1, 1]. \tag{1.1}$$

Correspondence to: Prof. W. Gautschi, Department of Computer Sciences, Purdue University, West Lafayette, IN 47907, United States. e-mail: wxg@cs.purdue.edu.

Here the norm is the weighted L_2-norm,

$$\|u\|_w = \left\{\int_{-1}^{1} u^2(t)w(t)\,dt\right\}^{1/2}.$$

Suppose we adjoin to the n zeros τ_i an additional $n+1$ nodes $\hat{\tau}_j = \hat{\tau}_j^{(n)}$, $j = 1, 2, \ldots, n+1$, distinct among themselves and from the τ_i, and form the Lagrange polynomial $(\hat{L}_{2n+1}f)(\cdot)$ of degree $\leqslant 2n$ interpolating f on the union of the nodes, $\{\tau_i\} \cup \{\hat{\tau}_j\}$. Is it still true that

$$\lim_{n\to\infty} \|f - \hat{L}_{2n+1}f\|_w = 0, \quad \text{all } f \in C[-1, 1]? \tag{1.2}$$

The answer can no longer be expected to be an unqualified "yes", since the behavior of $\hat{L}_{2n+1}f$ will strongly depend on the kind of additional nodes introduced. A natural choice for these nodes would be the zeros of $\pi_{n+1}(\cdot; w)$. Unfortunately, no criteria are known that would be applicable to prove mean convergence for all $C[-1, 1]$ in this case. An interesting choice, however, has recently been discussed by Bellen [2], who takes the nodes $\hat{\tau}_j$ to be the zeros of the polynomial $\hat{\pi}_{n+1}(\cdot) = \pi_{n+1}(\cdot; \pi_n^2 w)$ of degree $n+1$ orthogonal to all lower-degree polynomials with respect to the (positive) weight function $\hat{w}_n = \pi_n^2 w$. He proves, in this case, that (1.2) indeed is true provided that the ratio

$$M(n; w) := \frac{\|\pi_n\|_w^2}{\min\limits_{1\leqslant j\leqslant n+1} \pi_n^2(\hat{\tau}_j)} \tag{1.3}$$

remains uniformly bounded,

$$M(n; w) = O(1), \quad \text{as } n \to \infty. \tag{1.4}$$

Note that (1.4) precludes any of the $\hat{\tau}_j$ from becoming equal, or too close, to any of the τ_i, and like the τ_i, they, too, are distinct from one another. (Clearly, it is irrelevant how one normalizes the polynomial π_n in (1.3).) Other types of extended Lagrange interpolation are studied in [1] for Lipschitz-continuous functions $f \in \text{Lip}\,\gamma$, $\gamma > \frac{1}{2}$, and in [3–5] with a view toward uniform convergence.

We say that the nodes $\hat{\tau}_j$ interlace with the nodes τ_i, if they satisfy, when ordered decreasingly,

$$\hat{\tau}_1 > \tau_1 > \hat{\tau}_2 > \tau_2 > \cdots > \hat{\tau}_n > \tau_n > \hat{\tau}_{n+1}. \tag{1.5}$$

If the leading coefficient of $\hat{\pi}_{n+1}$ is positive, then (1.5) is equivalent to

$$\operatorname{sgn} \hat{\pi}_{n+1}(\tau_i) = (-1)^i, \quad i = 1, 2, \ldots, n. \tag{1.5'}$$

The only weight function w for which (1.4) (and also (1.5)) is known to be true is

$$w(t) = (1-t^2)^{1/2}, \quad -1 < t < 1. \tag{1.6}$$

In this case, $\pi_n = U_n$ is the Chebyshev polynomial of the second kind, and $\hat{\pi}_{n+1} = T_{n+1}$ the $(n+1)$st-degree Chebyshev polynomial of the first (cf. [2]). In Section 2 we show that (1.4) is false for any of the other three Chebyshev weights, even though the respective nodes τ_i and $\hat{\tau}_j$ interlace. In Section 5 we discuss the validity of (1.4) numerically, based on methods described in Section 4, when w is a Jacobi weight function,

$$w^{(\alpha,\beta)}(t) = (1-t)^{\alpha}(1+t)^{\beta}, \quad -1 < t < 1, \tag{1.7}$$

in particular, a Gegenbauer weight $w^{(\alpha,\alpha)}$, with the parameters α, β suitably restricted. It suffices, in (1.7), to consider $\beta \geqslant \alpha$, since

$$M(n; w^{(\alpha,\beta)}) = M(n; w^{(\beta,\alpha)}). \tag{1.8}$$

This follows easily from the identity $P_n^{(\alpha,\beta)}(t) = (-1)^n P_n^{(\beta,\alpha)}(-t)$ for Jacobi polynomials. We conjecture that (1.4) is valid for Jacobi weights $w^{(\alpha,\beta)}$ with $0 \leqslant \alpha \leqslant 1.6$, $\alpha \leqslant \beta < \beta_0$, where $1.55 < \beta_0 < 1.65$, and for Jacobi–Gegenbauer weights $w^{(\alpha,\alpha)}$ in the sharper form $\lim_{n\to\infty} M(n, w^{(\alpha,\alpha)}) = \frac{1}{2}\pi$, provided $0 \leqslant \alpha < \alpha^0$, where $1.6 < \alpha^0 < 1.7$. The case of negative α seems more subtle, and we dare not conjecture (1.4) except for Gegenbauer weights $w^{(\alpha,\alpha)}$ with $-\alpha_0 < \alpha \leqslant 0$ for some α_0 near and slightly larger than 0.31.

It should be borne in mind, however, that (1.4) is merely a sufficient condition for convergence in the mean (cf. (1.2)), and its failure to be satisfied does not necessarily invalidate (1.2). A condition that *does* invalidate (1.2) has been given by Erdős and Turán [6, Theorem III] in terms of the function

$$L(n; w) := \int_{-1}^{1} \left(\sum_{i=1}^{n} l_i^2(t) + \sum_{j=1}^{n+1} \hat{l}_j^2(t) \right) w(t)\, dt, \tag{1.9}$$

where l_i and $\hat{l}_j$ are the elementary Lagrange interpolation polynomials for the point set $\{\tau_i\} \cup \{\hat{\tau}_j\}$,

$$\begin{aligned} l_i(t) &= \frac{\pi_n(t)\hat{\pi}_{n+1}(t)}{(t-\tau_i)\pi_n'(\tau_i)\hat{\pi}_{n+1}(\tau_i)}, \quad i = 1, 2, \ldots, n, \\ \hat{l}_j(t) &= \frac{\hat{\pi}_{n+1}(t)\pi_n(t)}{(t-\hat{\tau}_j)\hat{\pi}_{n+1}'(\hat{\tau}_j)\pi_n(\hat{\tau}_j)}, \quad j = 1, 2, \ldots, n+1. \end{aligned} \tag{1.10}$$

Indeed, if

$$\overline{\lim_{n\to\infty}}\, L(n; w) = +\infty, \tag{1.11}$$

it was shown [1] in [6] that there exists an $f \in C[-1, 1]$ such that

$$\overline{\lim_{n\to\infty}} \left\| f - \hat{L}_{2n+1} f \right\|_w = +\infty. \tag{1.12}$$

In Section 3 it will be shown that (1.11) is true for all Chebyshev weight functions other than the one of second kind, which establishes that mean convergence (1.2) does no longer hold in these cases. It should be stressed, nevertheless, that this negative result is not so much a critique of the special choice of interpolation nodes, as it is a reflection of the very large class of functions considered. Adding only a slight amount of regularity, for example, Lipschitz continuity with parameter larger than $\frac{1}{2}$, would already restore convergence. Indeed, for Chebyshev weights w, the referee has kindly pointed out that $\| f - \hat{L}_{2n+1} f \|_w \leqslant \text{const.} \cdot n^{1/2} E_n(f)$, where $E_n(f)$ is the error of best uniform approximation of f by polynomials of degree $\leqslant n$. For still more regularity, in particular analyticity, see also Section 6.

[1] [6, Theorem III], valid for an arbitrary triangular matrix of interpolation nodes, assumes $w(t) = 1$, but the proof goes through for an arbitrary weight function w.

2. The ratio $M(n; w)$ for Chebyshev weights

2.1. First-kind Chebyshev weight function

In this subsection we take the weight function w to be

$$w = w_1, \qquad w_1(t) = (1-t^2)^{-1/2}, \quad -1 < t < 1. \tag{2.1}$$

The corresponding orthogonal polynomial is the Chebyshev polynomial of the first kind,

$$\pi_n(t) = T_n(t), \qquad T_n(\cos\theta) = \cos n\theta. \tag{2.2}$$

As in Section 1, we assume $n \geqslant 1$. We claim that

$$\hat{\pi}_{n+1}(t) = T_{n+1}(t) - \tfrac{1}{2}T_{n-1}(t). \tag{2.3}$$

Indeed, using repeatedly the well-known identity

$$T_nT_m = \tfrac{1}{2}(T_{n+m} + T_{|n-m|}), \tag{2.4}$$

we have for any $p \in \mathbb{P}_n$,

$$\begin{aligned}\int_{-1}^{1}(T_{n+1} - \tfrac{1}{2}T_{n-1})pT_n^2w_1\,\mathrm{d}t &= \frac{1}{2}\int_{-1}^{1}\left[(T_{2n+1}+T_1) - \tfrac{1}{2}(T_{2n-1}+T_1)\right]pT_nw_1\,\mathrm{d}t \\ &= \frac{1}{2}\int_{-1}^{1}(T_{2n+1} - \tfrac{1}{2}T_{2n-1} + \tfrac{1}{2}T_1)pT_nw_1\,\mathrm{d}t \\ &= \frac{1}{4}\int_{-1}^{1}(T_{3n+1} - \tfrac{1}{2}T_{3n-1} + \tfrac{3}{2}T_{n+1})pw_1\,\mathrm{d}t = 0,\end{aligned}$$

the last equality, since $p \in \mathbb{P}_n$, on account of the orthogonality of the Chebyshev polynomials. This proves (2.3).

With

$$\tau_i = \cos\theta_i, \quad \theta_i = \cos\left(\frac{2i-1}{2n}\pi\right), \quad i = 1, 2, \ldots, n, \tag{2.5}$$

denoting the zeros of T_n, it follows easily from (2.3) and (2.2) that

$$\hat{\pi}_{n+1}(\tau_i) = \tfrac{3}{2}(-1)^i \sin\theta_i, \quad i = 1, 2, \ldots, n,$$

so that (1.5′), and hence the interlacing property (1.5), holds for the zeros $\hat{\tau}_j$ of $\hat{\pi}_{n+1}$.

Letting $t = \cos\theta$ in (2.3), we can write the equation $\hat{\pi}_{n+1}(t) = 0$ in trigonometric form $\cos(n+1)\theta - \frac{1}{2}\cos(n-1)\theta = 0$, or, with the help of the addition theorem for the cosine, in the form

$$\tan n\theta \tan\theta = \tfrac{1}{3}, \quad 0 < \theta < \pi. \tag{2.6}$$

Since with π_n also $\hat{\pi}_{n+1}$ is an even polynomial, its zeros $\hat{\tau}_j$ are symmetric with respect to the origin; it suffices therefore to consider (2.6) in $0 < \theta < \frac{1}{2}\pi$. Using

$$\tan^2\theta = \frac{1}{\cos^2\theta} - 1 \quad \text{and} \quad t = \cos\theta,$$

we can write (2.6), when squared, in the form

$$\left(\frac{1}{T_n^2(t)} - 1\right)\left(\frac{1}{t^2} - 1\right) = \frac{1}{9},$$

or, equivalently, in the form

$$T_n^2(t) = \frac{9(1-t^2)}{9-8t^2}. \tag{2.7}$$

The rational function on the right decreases monotonically on (0, 1); therefore, by the symmetry of the $\hat{\tau}_j$,

$$\min_{1\leq j\leq n+1} T_n^2(\hat{\tau}_j) = \frac{9(1-\hat{\tau}_1^2)}{9-8\hat{\tau}_1^2}.$$

Since $\| T_n \|_{w_1}^2 = \frac{1}{2}\pi$ for $n \geq 1$, we get from the definition in (1.3) that

$$M(n; w_1) = \tfrac{1}{18}\pi \frac{9-8\hat{\tau}_1^2}{1-\hat{\tau}_1^2}. \tag{2.8}$$

Writing (2.6) in the form

$$\tan n\theta = \frac{1}{3\tan\theta}, \tag{2.6'}$$

and examining the graphs of the two functions on the left and right, immediately yields

$$0 < \hat{\theta}_1 < \frac{\pi}{2n}$$

for the smallest positive root, $\theta = \hat{\theta}_1$, of (2.6'). Consequently,

$$\cos\frac{\pi}{2n} < \hat{\tau}_1 = \cos\hat{\theta}_1 < 1,$$

and by (2.8),

$$M(n; w_1) > \tfrac{1}{18}\pi \frac{1}{\sin^2(\pi/2n)}. \tag{2.9}$$

This shows that $M(n; w_1)$ grows to ∞ as $n \to \infty$, at least like $O(n^2)$.

2.2. *Second-kind Chebyshev weight function*

For completeness, we include here the Chebyshev weight function of the second kind,

$$w = w_2, \qquad w_2(t) = (1-t^2)^{1/2}, \quad -1 < t < 1, \tag{2.10}$$

for which

$$\pi_n(t) = U_n(t), \qquad U_n(\cos\theta) = \frac{\sin(n+1)\theta}{\sin\theta}. \tag{2.11}$$

Since $T_{n+1}U_n = \frac{1}{2}U_{2n+1}$, we have, for any $p \in \mathbb{P}_n$,

$$\int_{-1}^{1} T_{n+1} p U_n^2 w_2 \, dt = \frac{1}{2}\int_{-1}^{1} U_{2n+1} p U_n w_2 \, dt = 0,$$

by orthogonality, since $pU_n \in \mathbb{P}_{2n}$. Therefore (as already observed in [2]),

$$\hat{\pi}_{n+1}(t) = T_{n+1}(t). \tag{2.12}$$

The zeros $\hat{\tau}_j$ of T_{n+1} clearly interlace with those of $\pi_n = U_n$. Furthermore, with

$$\hat{\tau}_j = \cos\hat{\theta}_j, \quad \hat{\theta}_j = \frac{2j-1}{2n+2}\pi,$$

we have

$$U_n(\hat{\tau}_j) = \frac{\sin(n+1)\hat{\theta}_j}{\sin\hat{\theta}_j} = \frac{(-1)^{j-1}}{\sin((2j-1)\pi/(2n+2))},$$

from which

$$\min_{1\leqslant j\leqslant n+1} U_n^2(\hat{\tau}_j) = \begin{cases} 1, & n \text{ even}, \\ \dfrac{1}{\sin^2(\frac{1}{2}\pi n/(n+1))} = \dfrac{1}{\cos^2(\pi/(2n+2))}, & n \text{ odd}. \end{cases}$$

Therefore, since $\|U_n\|_{w_2}^2 = \frac{1}{2}\pi$,

$$M(n; w_2) = \begin{cases} \frac{1}{2}\pi, & n \text{ even}, \\ \frac{1}{2}\pi\cos^2\dfrac{\pi}{2n+2}, & n \text{ odd}. \end{cases} \tag{2.13}$$

We see that $M(n; w_2)$ now indeed satisfies (1.4); specifically, $M(n; w_2) \leqslant \frac{1}{2}\pi$ for all $n \geqslant 1$, and

$$\lim_{n\to\infty} M(n; w_2) = \tfrac{1}{2}\pi. \tag{2.14}$$

2.3. *Third- and fourth-kind Chebyshev weight functions*

These are the Jacobi weights (1.7) with $\alpha = -\frac{1}{2}$, $\beta = \frac{1}{2}$ and $\alpha = \frac{1}{2}$, $\beta = -\frac{1}{2}$, respectively. As remarked in (1.8), it suffices to consider the first of these,

$$w = w_3, \qquad w_3(t) = (1-t)^{-1/2}(1+t)^{1/2}, \quad -1 < t < 1. \tag{2.15}$$

The corresponding orthogonal polynomial is

$$\pi_n(t) = V_n(t), \qquad V_n(\cos\theta) = \frac{\cos(n+\frac{1}{2})\theta}{\cos\frac{1}{2}\theta}. \tag{2.16}$$

Here we have

$$\hat{\pi}_{n+1}(t) = T_{n+1}(t) - \tfrac{1}{2}T_n(t). \tag{2.17}$$

Indeed, noting that

$$T_nV_n = \tfrac{1}{2}(V_{2n} + 1), \qquad T_{n+1}V_n = \tfrac{1}{2}(V_{2n+1} + 1),$$

we compute, for any $p \in \mathbb{P}_n$,

$$\begin{aligned}\int_{-1}^{1} (T_{n+1} - \tfrac{1}{2}T_n)pV_n^2 w_3 \, dt &= \frac{1}{2}\int_{-1}^{1} \left[(V_{2n+1} + 1) - \tfrac{1}{2}(V_{2n} + 1)\right] pV_n w_3 \, dt \\ &= \frac{1}{2}\int_{-1}^{1} \left[V_{2n+1} - \tfrac{1}{2}(V_{2n} - 1)\right] pV_n w_3 \, dt \\ &= \frac{1}{4}\int_{-1}^{1} (1 - V_{2n}) pV_n w_3 \, dt. \end{aligned} \tag{2.18}$$

An elementary calculation shows that

$$\int_{-1}^{1} V_n^2(t) w_3(t) \, dt = \pi, \quad \text{for } n \geqslant 1. \tag{2.19}$$

Therefore, letting $p = V_n$ in (2.18) gives

$$\begin{aligned}\int_{-1}^{1} (1 - V_{2n}) V_n^2 w_3 \, dt &= \int_{-1}^{1} V_n^2 w_3 \, dt - \int_{-1}^{1} V_{2n} V_n^2 w_3 \, dt \\ &= \int_{-1}^{1} V_n^2 w_3 \, dt - \int_{-1}^{1} V_{2n}^2 w_3 \, dt = 0,\end{aligned}$$

since V_n^2 differs from V_{2n} by a polynomial of degree $< 2n$ and the last two integrals are the same by (2.19). Letting $p = V_m$, $m < n$, in (2.18) gives

$$\int_{-1}^{1} (1 - V_{2n}) V_m V_n w_3 \, dt = \int_{-1}^{1} V_m V_n w_3 \, dt - \int_{-1}^{1} V_{2n} V_m V_n w_3 \, dt = 0,$$

by orthogonality. Thus, the integral on the far right of (2.18), hence the one on the left, vanishes for every $p \in \mathbb{P}_n$, which proves (2.17).

It is again a simple matter to compute

$$\hat{\pi}_{n+1}(\tau_i) = \tfrac{3}{2}(-1)^i \sin\left(\frac{2i-1}{2n+1}\tfrac{1}{2}\pi\right), \quad i = 1, 2, \ldots, n, \tag{2.20}$$

for the zeros τ_i of V_n, and hence to verify the interlacing property (1.5).

The equation $\hat{\pi}_{n+1}(t) = 0$ in trigonometric form ($t = \cos\theta$) is $\cos(n+1)\theta - \frac{1}{2}\cos n\theta = 0$, which, by writing $(n+1)\theta = (n+\frac{1}{2})\theta + \frac{1}{2}\theta$ and $n\theta = (n+\frac{1}{2})\theta - \frac{1}{2}\theta$ and using the addition theorem for the cosine, becomes

$$\tan(n + \tfrac{1}{2})\theta \tan \tfrac{1}{2}\theta = \tfrac{1}{3}, \quad 0 < \theta < \pi. \tag{2.21}$$

By manipulations similar to those in Section 2.1, this can be written as

$$V_n^2(t) = \frac{9(1-t)}{(1+t)(5-4t)}, \quad t = \cos\theta. \tag{2.22}$$

The rational function on the right decreases monotonically on $(-1, 1)$, implying that

$$\min_{1\leqslant j\leqslant n+1} V_n^2(\hat{\tau}_j) = \frac{9(1-\hat{\tau}_1)}{(1+\hat{\tau}_1)(5-4\hat{\tau}_1)}.$$

Combining this with $\| V_n \|_{w_3}^2 = \pi$ (cf. (2.19)), we get

$$M(n; w_3) = \tfrac{1}{9}\pi \frac{(1+\hat{\tau}_1)(5-4\hat{\tau}_1)}{1-\hat{\tau}_1}. \tag{2.23}$$

Similarly as in Section 2.1, we find that $0 < \hat{\theta}_1 < \pi/(2n+1)$, hence $\hat{\tau}_1 = \cos\hat{\theta}_1 > \cos(\pi/(2n+1))$, and thus, again by the monotonicity of the rational function in (2.22),

$$M(n; w_3) > \tfrac{1}{9}\pi \frac{[1+\cos(\pi/(2n+1))][5-4\cos(\pi/(2n+1))]}{1-\cos(\pi/(2n+1))}.$$

Thus, finally,

$$M(n; w_3) > \tfrac{1}{9}\pi \cot^2 \frac{\pi}{2(2n+1)}. \tag{2.24}$$

Again, $M(n; w_3)$ grows to ∞ as $n \to \infty$, at least like $O(n^2)$.

3. The function $L(n; w)$ for Chebyshev weights

3.1. First-kind Chebyshev weight function

We begin with the case $w_1(t) = (1-t^2)^{-1/2}$. Since

$$L(n; w_1) > \int_{-1}^{1} \sum_{i=1}^{n} l_i^2(t) w_1(t)\, dt, \tag{3.1}$$

it suffices, for showing (1.11), that the right-hand side of (3.1) is unbounded as $n \to \infty$. We may choose in (1.10)

$$\pi_n(t) = T_n(t), \qquad \hat{\pi}_{n+1}(t) = T_{n+1}(t) - \tfrac{1}{2}T_{n-1}(t) \tag{3.2}$$

(cf. (2.3)). Letting, as in (2.5), $\tau_i = \cos\theta_i$, one easily computes

$$\pi_n'(\tau_i) = \frac{(-1)^{i-1}n}{\sin\theta_i}, \qquad \hat{\pi}_{n+1}(\tau_i) = \tfrac{3}{2}(-1)^i \sin\theta_i, \tag{3.3}$$

so that

$$\pi_n'(\tau_i)\hat{\pi}_{n+1}(\tau_i) = -\tfrac{3}{2}n. \tag{3.4}$$

Furthermore, using (2.4),

$$\begin{aligned}\hat{\pi}_{n+1}^2(t) &= T_{n+1}^2 - T_{n+1}T_{n-1} + \tfrac{1}{4}T_{n-1}^2 \\ &= \tfrac{1}{2}\left[(T_{2n+2}+1) - (T_{2n}+T_2) + \tfrac{1}{4}(T_{2n-2}+1)\right] \\ &= \tfrac{1}{2}(T_{2n+2} - T_{2n} + \tfrac{1}{4}T_{2n-2} - T_2 + \tfrac{5}{4}),\end{aligned}$$

hence, by orthogonality,

$$\int_{-1}^{1}\left(\frac{\pi_n(t)}{t-\tau_i}\right)^2 \hat{\pi}_{n+1}^2(t)w_1(t)\,dt = \frac{1}{8}\int_{-1}^{1}\left(\frac{T_n}{t-\tau_i}\right)^2 [T_{2n-2}-4T_2+5]w_1\,dt. \tag{3.5}$$

Now for the first term on the right we use the fact that $(T_n/(t-\tau_i))^2 = 2T_{2n-2}$ modulo $\mathbb{P}_{2n-3}$, if $n>1$, so that, again by orthogonality,

$$\int_{-1}^{1}\left(\frac{T_n}{t-\tau_i}\right)^2 T_{2n-2}w_1\,dt = 2\int_{-1}^{1} T_{2n-2}^2 w_1\,dt = \pi, \quad n>1. \tag{3.6}$$

The remaining part of the integral on the right of (3.5) can be evaluated by the n-point Gauss–Chebyshev quadrature rule with remainder term. Since

$$\frac{1}{(2n)!}\left\{\left(\frac{T_n}{t-\tau_i}\right)^2(-4T_2+5)\right\}^{(2n)} = -2^{2n}$$

and $T_n(\tau_k)=0$, $k=1,2,\dots,n$, we get, upon using (3.3),

$$\begin{aligned}
&\int_{-1}^{1}\left(\frac{T_n}{t-\tau_i}\right)^2(-4T_2+5)w_1\,dt \\
&= \frac{\pi}{n}[T_n'(\tau_i)]^2[-4T_2(\tau_i)+5] - 2^{2n}\int_{-1}^{1}\left[\frac{1}{2^{n-1}}T_n(t)\right]^2 w_1(t)\,dt \\
&= \frac{\pi}{n}\frac{n^2}{\sin^2\theta_i}[5-4\cos 2\theta_i] - 4\cdot\tfrac{1}{2}\pi > \frac{\pi n}{\sin^2\theta_i} - 2\pi.
\end{aligned} \tag{3.7}$$

Inserting (3.6) and (3.7) into (3.5) gives

$$\int_{-1}^{1}\left(\frac{\pi_n}{t-\tau_i}\right)^2 \hat{\pi}_{n+1}^2 w_1\,dt > \tfrac{1}{8}\pi\left(\frac{n}{\sin^2\theta_i}-1\right).$$

Therefore, by (1.10) and (3.4),

$$\begin{aligned}
\int_{-1}^{1}\sum_{i=1}^{n} l_i^2(t)w_1\,dt &> \frac{\pi}{8\cdot\frac{9}{4}n^2}\left(n\sum_{i=1}^{n}\frac{1}{\sin^2\theta_i}-n\right) = \frac{\pi}{18n}\left(\sum_{i=1}^{n}\frac{1}{\sin^2\theta_i}-1\right) \\
&> \frac{\pi}{18n}\left(\frac{1}{\sin^2\theta_1}-1\right) > \tfrac{1}{18}\pi\left(\frac{4}{\pi^2}n-\frac{1}{n}\right),
\end{aligned}$$

since $\sin\theta_1 = \sin(\pi/2n) < \pi/2n$. This shows that the right-hand side of (3.1), hence also the left-hand side, is unbounded as $n\to\infty$.

3.2. *Second-kind Chebyshev weight function*

Since here we know that $M(n; w_2)$ satisfies (1.4) (cf. (2.14)), we must necessarily have uniform boundedness of $L(n; w_2)$. We show, in fact, that

$$L(n; w_2) = \tfrac{1}{2}\pi, \quad \text{for all } n \geqslant 1. \tag{3.8}$$

Indeed, since $\pi_n = U_n$, $\hat{\pi}_{n+1} = T_{n+1}$ (cf. (2.11), (2.12)), and the set $\{\tau_i\} \cup \{\hat{\tau}_j\}$ consists precisely of the zeros of U_{2n+1}, we have

$$L(n; w_2) = \sum_{i=1}^{2n+1} \gamma_i^{(2n+1)}(w_2) = \int_{-1}^{1} w_2(t)\, dt,$$

where $\gamma_i^{(2n+1)}(w_2)$ are the weights in the $(2n+1)$-point Gauss formula for w_2. From this, (3.8) follows immediately.

3.3. *Third- and fourth-kind Chebyshev weight functions*

As in (1.8), one easily shows that

$$L(n; w^{(\alpha,\beta)}) = L(n; w^{(\beta,\alpha)}). \tag{3.9}$$

It suffices therefore to assume $\alpha = -\frac{1}{2}$, $\beta = \frac{1}{2}$, that is,

$$w(t) = w_3(t), \qquad w_3(t) = (1-t)^{-1/2}(1+t)^{1/2}.$$

We will show that $\int_{-1}^{1} \sum_{i=1}^{n} l_i^2(t) w_3(t)\, dt$ is unbounded as $n \to \infty$, which, as in (3.1), will prove the same for $L(n; w_3)$. Since $\pi_n = V_n$ and $\hat{\pi}_{n+1} = T_{n+1} - \frac{1}{2}T_n$ (cf. (2.16), (2.17)), we have

$$\begin{aligned}
\int_{-1}^{1} \left(\frac{\pi_n}{t-\tau_i}\right)^2 \hat{\pi}_{n+1}^2 w_3\, dt &= \int_{-1}^{1} (1+t)\left(\frac{V_n}{t-\tau_i}\right)^2 \left(T_{n+1} - \tfrac{1}{2}T_n\right)^2 \frac{w_3}{1+t}\, dt \\
&= \int_{-1}^{1} (1+t)\left(\frac{V_n}{t-\tau_i}\right)^2 \left(T_{n+1}^2 - T_{n+1}T_n + \tfrac{1}{4}T_n^2\right) w_1\, dt \\
&= \frac{1}{2}\int_{-1}^{1} (1+t)\left(\frac{V_n}{t-\tau_i}\right)^2 \left[(T_{2n+2}+1) - (T_{2n+1} + T_1)\right. \\
&\qquad \left. + \tfrac{1}{4}(T_{2n}+1)\right] w_1\, dt.
\end{aligned}$$

By orthogonality, this simplifies to

$$\frac{1}{8}\int_{-1}^{1} (1+t)\left(\frac{V_n}{t-\tau_i}\right)^2 [-4T_1 + 5] w_1\, dt = \frac{1}{8}\int_{-1}^{1} \left(\frac{V_n}{t-\tau_i}\right)^2 [-4T_1 + 5] w_3\, dt.$$

Now, with $\gamma_i = \gamma_i^{(n)}(w_3)$ denoting the Christoffel numbers for the n-point Gauss formula for $w = w_3$, we get, noting that the integrand in the last integral is a polynomial of degree $2n-1$, and $V_n(\tau_k) = 0$, that

$$\int_{-1}^{1} \left(\frac{\pi_n}{t-\tau_i}\right)^2 \hat{\pi}_{n+1}^2 w_3\, dt = \tfrac{1}{8}\gamma_i [\pi_n'(\tau_i)]^2 [5 - 4\tau_i].$$

Since furthermore, by (2.20),

$$\hat{\pi}_{n+1}(\tau_i) = \tfrac{3}{2}(-1)^i \sin \tfrac{1}{2}\theta_i, \quad \tau_i = \cos\theta_i, \ \theta_i = \frac{2i-1}{2n+1}\pi,$$

we obtain from (1.10)

$$\int_{-1}^{1} \sum_{i=1}^{n} l_i^2(t) w_3(t)\, dt = \frac{1}{8} \sum_{i=1}^{n} \gamma_i \frac{5-4\tau_i}{\frac{9}{4}\sin^2 \frac{1}{2}\theta_i}$$
$$= \frac{1}{9} \sum_{i=1}^{n} \gamma_i \frac{5-4\tau_i}{1-\tau_i} > \frac{1}{9} \frac{\gamma_1}{1-\tau_1}. \tag{3.10}$$

Noting that $1-\tau_1 = 2\sin^2 \frac{1}{2}\theta_1$ and, as is well known,

$$\gamma_1 = \frac{4\pi}{2n+1} \cos^2 \tfrac{1}{2}\theta_1,$$

we see that the lower bound in (3.10) is

$$\frac{2\pi}{9(2n+1)\tan^2 \frac{1}{2}\theta_1} = \frac{2\pi}{9(2n+1)\tan^2(\pi/(2(2n+1)))} = O(n), \quad n \to \infty,$$

which proves the assertion.

4. Computational methods

For Chebyshev weight functions w, the polynomials $\hat{\pi}_{n+1}$ (cf. (2.3), (2.12) and (2.17)) are easily computed, as they are all orthogonal with respect to either the Chebyshev weight function of the first kind, as in the case of (2.12), or with respect to this same weight function divided by a quadratic or linear polynomial, as in the other two cases (cf., e.g., [8, §5]). In particular, the three-term recurrence relation for these polynomials is explicitly known. This is no longer true for other weight functions, where computational methods have to be employed to generate the desired polynomials.

The key constituents of any computation involving orthogonal polynomials relative to a weight function $\omega \geqslant 0$ are the recursion coefficients $\alpha_k = \alpha_k(\omega)$, $\beta_k = \beta_k(\omega)$ in the basic three-term recurrence relation satisfied by the (monic) polynomials $\pi_k(\cdot) = \pi_k(\cdot\,; \omega)$,

$$\begin{aligned} &\pi_{k+1}(t) = (t-\alpha_k)\pi_k(t) - \beta_k \pi_{k-1}(t), \quad k = 0, 1, 2, \ldots, \\ &\pi_{-1}(t) = 0, \qquad \pi_0(t) = 1 \end{aligned} \tag{4.1}$$

(cf. [7]). The Jacobi matrix $J(\omega)$ of the weight function ω is the infinite symmetric tridiagonal matrix

$$J(\omega) = \operatorname{tri}\left(\alpha_0, \alpha_1, \alpha_2, \ldots; \sqrt{\beta_1}, \sqrt{\beta_2}, \sqrt{\beta_3}, \ldots\right) \tag{4.2}$$

having the α's on the main diagonal, and the square roots of the β's on the two side diagonals. Its leading principal minor matrix of order m will be denoted by

$$J_m = J_m(\omega) = J(\omega)_{m \times m}. \tag{4.3}$$

In terms of the matrix J_m, the zeros $\tau_\mu = \tau_\mu^{(m)}$ of $\pi_m(\cdot\,; \omega)$ are best computed as eigenvalues of J_m,

$$\det(J_m - \tau_\mu I_m) = 0, \quad \mu = 1, 2, \ldots, m, \tag{4.4}$$

using the QR algorithm with judiciously selected shifts.

The following observation, due to [9], will be the basis for our computational method. Given the Jacobi matrix $J_{m+1}(\omega)$ of order $m+1$ for the weight function $\omega(t)$, one can obtain the Jacobi matrix of order m, $J_m((\cdot-\tau)^2\omega)$, for the weight function $(t-\tau)^2\omega(t)$, $\tau\in\mathbb{R}$, by one step of the QR algorithm with shift τ: from the QR decomposition

$$J_{m+1}(\omega)-\tau I_{m+1}=QR,\quad Q \text{ orthogonal, } R \text{ upper triangular, diag } R\geqslant 0, \tag{4.5}$$

one obtains

$$J_m\big((\cdot-\tau)^2\omega\big)=(RQ+\tau I_{m+1})_{m\times m}, \tag{4.6}$$

discarding, as indicated by the subscript, the last row and last column of the transformed matrix. An efficient and stable algorithm for carrying out this transformation can be found, e.g., in [10, p.567].

If we denote $\beta_0(\omega)=\int\omega(t)\,dt$, where the integral extends over the support of ω, then we can also get $\beta_0((\cdot-\tau)^2\omega)$ by the simple formula

$$\beta_0\big((\cdot-\tau)^2\omega\big)=\beta_0(\omega)\big[\beta_1(\omega)+(\tau-\alpha_0(\omega))^2\big]. \tag{4.7}$$

Indeed, if α_0, β_0, β_1 are the quantities appearing on the right of (4.7), we have, as is well known,

$$\int t\omega(t)\,dt=\alpha_0\beta_0,\qquad \int(t-\alpha_0)^2\omega(t)\,dt=\beta_0\beta_1.$$

Expanding the square in the second formula, and using the first, we find

$$\int t^2\omega(t)\,dt=2\alpha_0\cdot\alpha_0\beta_0-\alpha_0^2\beta_0+\beta_0\beta_1=\beta_0(\beta_1+\alpha_0^2),$$

and hence

$$\int(t-\tau)^2\omega(t)\,dt=\beta_0(\beta_1+\alpha_0^2)-2\tau\alpha_0\beta_0+\tau^2\beta_0=\beta_0(\beta_1+\alpha_0^2-2\tau\alpha_0+\tau^2),$$

which is (4.7).

Now let $\omega=w$, and write $\hat{w}_n$ in the form

$$\hat{w}_n(t)=w(t)\pi_n^2(t;w)=w(t)\prod_{i=1}^{n}(t-\tau_i)^2, \tag{4.8}$$

where $\tau_i=\tau_i^{(n)}$ are the zeros of $\pi_n(\cdot;w)$. Define

$$\hat{w}(t;k)=w(t)\prod_{i=1}^{k}(t-\tau_i)^2, \tag{4.9}$$

so that

$$\hat{w}(t;0)=w(t),\qquad \hat{w}(t;n)=\hat{w}_n(t).$$

Let $J_{m,k}$ denote the Jacobi matrix of order m of the "intermediate" weight function $w(\cdot;k)$,

$$J_{m,k}=J_m(\hat{w}(\cdot;k)),\quad k=0,1,\ldots,n. \tag{4.10}$$

Since

$$\hat{w}(t;k)=(t-\tau_k)^2\hat{w}(t;k-1),\quad k=1,2,\ldots,n, \tag{4.11}$$

it is clear that $J_{n+1,n}$ can be obtained from $J_{2n+1,0}$ by n successive transformations of the type (4.5), (4.6) with shifts τ_k:

$$J_{2n-k+2,k-1} - \tau_k I_{2n-k+2} = QR, \qquad J_{2n-k+1,k} = (RQ + \tau_k I_{2n-k+2})_{2n-k+1\times 2n-k+1}. \tag{4.12}$$

Upon completion of (4.12) for $k = 1, 2, \ldots, n$, we will have the desired Jacobi matrix

$$J_{n+1,n} = J_{n+1}(\hat{w}_n). \tag{4.13}$$

The zeros τ_k of π_n required in (4.12), as well as those of $\hat{\pi}_{n+1}$, are computed as eigenvalues of $J_n(w)$ and $J_{n+1}(\hat{w}_n)$, respectively (cf. (4.4)). Also, since in our implementation we successively update β_0 by means of (4.7), the numerator in (1.3) will simply be

$$\|\pi_n\|_w^2 = \int_{-1}^{1} \pi_n^2(t) w(t)\, dt = \int_{-1}^{1} \hat{w}_n(t)\, dt = \hat{\beta}_0,$$

the final β_0-coefficient.

Our experience has indicated that this procedure is quite stable, even for values of n as large as $n = 320$. The results reported in the next section are all obtained in this manner.

5. Numerical study for Jacobi weight functions

The reason why in Sections 2.1–2.3 the polynomials $\hat{\pi}_{n+1}$ for the four Chebyshev weights could be obtained analytically is the fact that in all these cases, $\hat{\pi}_{k,n}(\cdot) = \pi_k(\cdot; \pi_n^2 w)$ is one of the Chebyshev polynomials for each $k \leqslant n$. This is no longer true for general Jacobi weights, not even in the simple case $\alpha = \frac{1}{2}$, $\beta = \frac{3}{2}$. Here, the Jacobi matrix $J_{n+1}(\hat{w}_n^{(1/2,3/2)})$, computed by the methods of Section 4, turns out to be a nontrivial perturbation of the Jacobi matrix $J_{n+1}(\hat{w}_n^{(1/2,1/2)})$ $[= J_{n+1}(w^{(-1/2,-1/2)})]$, showing that the $\hat{\pi}_{k,n}$ are no longer pure Chebyshev polynomials in the range $k \leqslant n$. This also suggests expanding $\hat{\pi}_{n+1}$ in Chebyshev polynomials of the first kind. In doing so, one finds by computation that all expansion coefficients are different from zero (and alternating in sign). It appears unlikely, therefore, that analytic methods will be successful in this case, let alone in the case of general Jacobi weight functions. We therefore undertake to explore the problem computationally.

It is, of course, intrinsically impossible to demonstrate the validity of (1.4) by (a finite number of) computations. Nevertheless, extensive and well-targeted computations can be carried out to come up with certain conjectures, which we now formulate. Each conjecture will be followed by numerical (and other) evidence supporting it.

Conjecture 5.1. *For the Jacobi–Gegenbauer weight function $w = w^{(\alpha,\alpha)}$, there holds*

$$\lim_{n\to\infty} M(n; w^{(\alpha,\alpha)}) = \tfrac{1}{2}\pi, \quad \text{if } 0 \leqslant \alpha < \alpha^0, \tag{5.1}$$

where α^0 is some number between 1.6 and 1.7. (For $\alpha = 0$, this was conjectured in [2].)

Numerical evidence. Let the minimum in the denominator of (1.3) be attained for $j = j_n$,

$$\min_{1\leqslant j\leqslant n+1} \pi_n^2(\hat{\tau}_j) = \pi_n^2(\hat{\tau}_{j_n}), \quad 1 \leqslant j_n \leqslant n+1. \tag{5.2}$$

We say that the minimum is attained "in the middle", if

$$j_n = \begin{cases} \frac{1}{2}(n+2), & n \text{ even}, \\ \frac{1}{2}(n+1) \text{ or } \frac{1}{2}(n+3), & n \text{ odd}. \end{cases} \tag{5.3}$$

This terminology is justified by virtue of the $n+1$ nodes $\hat{\tau}_j$ being symmetric with respect to the origin. Note, in particular, that (5.3), for n even, implies $\hat{\tau}_{j_n} = 0$.

Now for $w = w^{(\alpha,\alpha)}$, $\alpha > -1$, one easily computes

$$\frac{\|\pi_n\|_w^2}{\pi_n^2(0)} = \sqrt{\pi}\,\frac{2^n(\frac{1}{2}n)!^2\Gamma(\alpha+1+\frac{1}{2}n)}{n!(n+\alpha+\frac{1}{2})\Gamma(\alpha+\frac{1}{2}(n+1))} \sim \tfrac{1}{2}\pi \quad \text{as } n\to\infty,\ w = w^{(\alpha,\alpha)}. \tag{5.4}$$

Therefore, if (5.3) holds, then the limit relation in Conjecture 5.1 is valid when restricted to even n (and very likely for unrestricted n as well). By computation we have found that for $\alpha = 0$, 0.5, 1.0, 1.5, 1.6, the minimum (5.2) is consistently attained in the middle, whenever $n = 1,\ldots,80$ and $n = 159$, 160, 239, 240, 319, 320. In each case computed, the interlacing property (1.5) also holds. Moreover, $M(240; w)$ and $M(320; w)$ agree with $\frac{1}{2}\pi$ to at least 5 decimal digits. When $\alpha = 1.7$, this pattern changes significantly: the minimum is still attained in the middle for $n = 1,\ldots,117$, but no longer so for $n = 118,\ldots,130$, in which cases it is attained "near the ends" ($j_n = 7$ or 8). The interlacing property, while true for $1 \leqslant n \leqslant 130$, breaks down at $n = 131$. As n is further increased, the ratio $M(n; w)$ takes on larger and larger values, for example, $M(160; w) = 45.964$ and $M(320; w) = 223.78$.

Figure 5.1 shows the behavior of $M(n; w^{(\alpha,\alpha)})$ for $1 \leqslant n \leqslant 160$; Fig. 5.1(a) is for $\alpha = 1.6$, Fig. 5.1(b) for $\alpha = 1.7$ (in a logarithmic scale!).

Thus, it appears that the proven behavior of $M(n; w^{(\alpha,\alpha)})$ in the case $\alpha = \frac{1}{2}$ (cf. (2.14)) is typical for all $0 \leqslant \alpha \leqslant 1.6$, but certainly not for $\alpha = 1.7$.

Conjecture 5.2. *The limit relation* (5.1) *holds for* $-\alpha_0 < \alpha \leqslant 0$, *where* α_0 *is some number between* 0.31 *and* 0.3125.

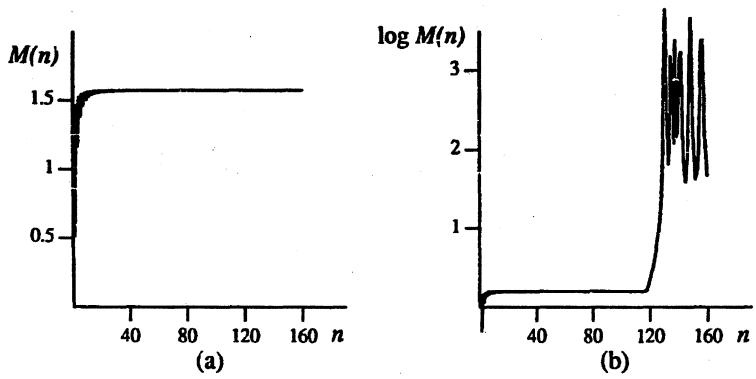


Fig. 5.1. $M(n; w^{(\alpha,\alpha)})$ for $1 \leqslant n \leqslant 160$; (a) $\alpha = 1.6$; (b) $\alpha = 1.7$.

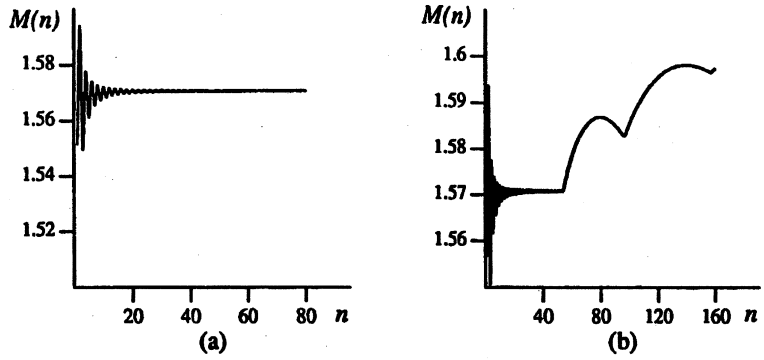


Fig. 5.2. $M(n; w^{(\alpha,\alpha)})$ for (a) $\alpha = -0.31$, $1 \leqslant n \leqslant 80$; (b) $\alpha = -0.3125$, $1 \leqslant n \leqslant 160$.

Numerical evidence. Verification of Conjecture 5.2 is made more difficult by the apparent fact that in the range under consideration, the interlacing property holds regardless of whether (1.4) is valid or not. We have found, however, that for $\alpha = -0.1, -0.2, -0.3, -0.31$, the minimum in (5.2) is consistently attained in the middle for the same values of n ($\leqslant 320$) as used in the discussion of Conjecture 5.1. This pattern changes when $\alpha = -0.3125$, in which case the minimum is attained in the middle for $1 \leqslant n \leqslant 53$, but near the ends ($j_n = 1$ or 2) for $54 \leqslant n \leqslant 80$. Along with this abrupt change in pattern goes a change in the behavior of the function $M(n; w)$; see Fig. 5.2. While the magnitude of $M(n; w)$ remains relatively small, even for n as larger as 320, the sudden development of distinct "vaults" raises some legitimate doubts as to uniform boundedness. (For still smaller values of $\alpha > -\frac{1}{2}$, the vaults seem to spread out over larger n-domains, making a determination of boundedness even more problematic.)

Conjecture 5.3. *For the Jacobi weight* $w = w^{(\alpha,\beta)}$, *the relation* (1.4) *is valid if* $0 \leqslant \alpha \leqslant 1.6$, $\alpha \leqslant \beta < \beta_0$, *where* β_0 *is some number* (*depending on* α) *between* 1.55 *and* 1.65.

Numerical evidence. For each $\alpha = 0(0.5)1.5$ we computed $M(n; w^{(\alpha,\beta)})$ for $\beta = 0(0.1)1.5$ and $n = 20, 41, 60, 79, 80$ and found the results to approach a limit to within 3–4 decimal digits. For the same values of α, we further scrutinized the case $\beta = 1.5$ by computing $M(n; w^{(\alpha,\beta)})$ for $n = 40, 81, 160, 241, 320$. The last two values (for $n = 241$ and $n = 320$) consistently agreed to 2–4 decimal digits. The same was observed for $\beta = 1.55$. In all cases, the interlacing property was found to hold. When $\beta = 1.6$, however, $M(n; w^{(\alpha,\beta)})$ takes on values of the order 10^3–10^5 when $n = 320$, and the interlacing property consistently breaks down for some $n \leqslant 320$, for each of the above α's, except $\alpha = 1.5$. In the cases $\alpha = 1.5(0.05)1.6$, $\beta = 1.6$, we still seem to have convergence as $n \to \infty$, but no longer so if $\beta = 1.65$ for the same three values of α.

For $\alpha < 0$ and $\beta > \alpha$, the numerical results seem inconclusive, as they do not permit a distinction between (slow) divergence and convergence. We are not prepared to make any conjecture in this range.

6. Simple vs. extended interpolation for smooth functions

Extended interpolation can be used, in practice, to check the accuracy of (simple) interpolation at the zeros of orthogonal polynomials. Thus, one would compare the (simple) interpolant $L_n f$ with the extended interpolant $\hat{L}_{2n+1} f$ (as defined in Section 1) to see how well they agree. This can be done at a cost of $2n+1$ function values. If we were to do the same with simple interpolation alone, and insisted on equal cost, we would have to compare $L_n f$ with $L_{n+1} f$, since all nodes change going from n to $n+1$. This has the serious disadvantage that the reference interpolant we are comparing with, i.e., $L_{n+1} f$, is only modestly more accurate than the interpolant to be checked, $L_n f$. In extended interpolation, on the other hand, the reference interpolant can be expected to be substantially more accurate, at least when f is sufficiently smooth.

To analyze the matter further, assume that $f \in C^{2n+1}[-1, 1]$, and let the scaled kth derivative of f be bounded on $[-1, 1]$ by M_k,

$$\frac{1}{k!} \| f^{(k)} \|_\infty \leq M_k, \quad k = 0, 1, \ldots, 2n+1. \tag{6.1}$$

Then it follows from interpolation theory that

$$\| f - L_{n+1} f \|_w^2 \leq M_{n+1}^2 \int_{-1}^{1} \pi_{n+1}^2(t; w) w(t)\, dt, \tag{6.2}$$

while

$$\| f - \hat{L}_{2n+1} f \|_w^2 \leq M_{2n+1}^2 \int_{-1}^{1} \hat{\pi}_{n+1}^2(t) \pi_n^2(t; w) w(t)\, dt, \tag{6.3}$$

the polynomials π_n, π_{n+1} and $\hat{\pi}_{n+1}$ all being assumed monic. Since

$$\int_{-1}^{1} \pi_{n+1}^2 w\, dt = \beta_0 \beta_1 \cdots \beta_n \beta_{n+1},$$

where the β's are the recursion coefficients $\beta_k(w)$ for the orthogonal polynomials $\{\pi_k(\cdot; w)\}$ (cf. (4.1)), and similarly

$$\int_{-1}^{1} \hat{\pi}_{n+1}^2 \pi_n^2 w\, dt = \hat{\beta}_0 \hat{\beta}_1 \cdots \hat{\beta}_n \hat{\beta}_{n+1},$$

where $\hat{\beta}_k = \beta_k(\hat{w}_n)$ (cf. (4.13)), the ratio ρ_n of the upper bounds in (6.3) and (6.2) is

$$\rho_n = \frac{\hat{\beta}_0 \hat{\beta}_1 \cdots \hat{\beta}_n \hat{\beta}_{n+1}}{\beta_0 \beta_1 \cdots \beta_n \beta_{n+1}} \left(\frac{M_{2n+1}}{M_{n+1}} \right)^2.$$

Since $\hat{\beta}_0 = \int_{-1}^{1} \pi_n^2 w\, dt = \beta_0 \beta_1 \cdots \beta_n$, this simplifies to

$$\rho_n = \omega_n \left(\frac{M_{2n+1}}{M_{n+1}} \right)^2, \quad \omega_n = \frac{\hat{\beta}_1 \hat{\beta}_2 \cdots \hat{\beta}_{n+1}}{\beta_{n+1}}. \tag{6.4}$$

Although conceivably M_{2n+1} is considerably larger than M_{n+1}, the quantity ω_n goes to zero rather quickly, so that the L_2-error of $\hat{L}_{2n+1} f$ is typically much smaller than that of $L_{n+1} f$.

Table 6.1
The quantities ω_n, $n = 5(5)40$, for Gegenbauer weights $w^{(\alpha,\alpha)}$, $\alpha = 0, 1, 2, 5, 10$

n	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$	$\alpha = 5$	$\alpha = 10$
5	$1.092 \cdot 10^{-3}$	$9.923 \cdot 10^{-4}$	$1.075 \cdot 10^{-3}$	$7.987 \cdot 10^{-4}$	$2.930 \cdot 10^{-4}$
10	$1.082 \cdot 10^{-6}$	$1.024 \cdot 10^{-6}$	$1.373 \cdot 10^{-6}$	$9.282 \cdot 10^{-7}$	$2.686 \cdot 10^{-7}$
15	$1.063 \cdot 10^{-9}$	$1.033 \cdot 10^{-9}$	$1.548 \cdot 10^{-9}$	$9.113 \cdot 10^{-10}$	$3.922 \cdot 10^{-10}$
20	$1.041 \cdot 10^{-12}$	$1.030 \cdot 10^{-12}$	$1.642 \cdot 10^{-12}$	$9.295 \cdot 10^{-13}$	$4.524 \cdot 10^{-13}$
25	$1.018 \cdot 10^{-15}$	$1.021 \cdot 10^{-15}$	$1.688 \cdot 10^{-15}$	$9.805 \cdot 10^{-16}$	$4.548 \cdot 10^{-16}$
30	$9.952 \cdot 10^{-19}$	$1.009 \cdot 10^{-18}$	$1.705 \cdot 10^{-18}$	$1.046 \cdot 10^{-18}$	$4.779 \cdot 10^{-19}$
35	$9.726 \cdot 10^{-22}$	$9.939 \cdot 10^{-22}$	$1.705 \cdot 10^{-21}$	$1.111 \cdot 10^{-21}$	$5.361 \cdot 10^{-22}$
40	$9.505 \cdot 10^{-25}$	$9.778 \cdot 10^{-25}$	$1.693 \cdot 10^{-24}$	$1.168 \cdot 10^{-24}$	$6.136 \cdot 10^{-25}$

Some numerical values of ω_n, for Gegenbauer weights $w^{(\alpha,\alpha)}$, $\alpha = 0, 1, 2, 5, 10$, are shown in Table 6.1.

For $w = w_1$ (Chebyshev weight of the first kind) we know from [8] that $\hat{\beta}_1 = \beta_1 = \frac{1}{2}$, $\hat{\beta}_k = \beta_k = \frac{1}{4}$ for $2 \leqslant k \leqslant n-1$, $\hat{\beta}_n = \frac{3}{8}$ and $\hat{\beta}_{n+1} = \frac{1}{8}$, so that $\omega_n = 3 \cdot 2^{-(2n+1)}$. Similarly, for $w = w_2$, we get $\omega_n = 2^{-2n}$. If, then, f is analytic in the disk $|z| \leqslant r$ with $r > 1$, one finds $\rho_n = O(2^{-2n}(r-1)^{-2n})$ as $n \to \infty$, hence $\rho_n = o(1)$ if $r > \frac{3}{2}$.

Acknowledgements

The author is indebted to Professor Giuseppe Mastroianni for a simplification in the proof of (3.8) and to the referee for useful comments.

References

[1] A. Bellen, A note on mean convergence of Lagrange interpolation, *J. Approx. Theory* **33** (1981) 85–95.
[2] A. Bellen, Alcuni problemi aperti sulla convergenza in media dell' interpolazione Lagrangiana estesa, *Rend. Istit. Mat. Univ. Trieste* **20** (1988) Fasc. suppl., 1–9.
[3] G. Criscuolo, G. Mastroianni and D. Occorsio, Convergence of extended Lagrange interpolation, *Math. Comp.* **55** (1990) 197–212.
[4] G. Criscuolo, G. Mastroianni and D. Occorsio, Uniform convergence of derivatives of extended Lagrange interpolation, *Numer. Math.* **60** (1991) 195–218.
[5] G. Criscuolo, G. Mastroianni and P. Vértesi, Pointwise simultaneous convergence of extended Lagrange interpolation with additional knots, *Math. Comp.*, to appear.
[6] P. Erdős and P. Turán, On interpolation. I, *Ann. of Math.* **38** (1937) 142–155.
[7] W. Gautschi, Computational aspects of orthogonal polynomials, in: P. Nevai, Ed., *Orthogonal Polynomials: Theory and Practice*, NATO Adv. Sci. Inst. Ser. C: Math. Phys. Sci. **294** (Kluwer Academic Publishers, Dordrecht, 1990) 181–216.
[8] W. Gautschi and S. Li, A set of orthogonal polynomials induced by a given orthogonal polynomial, submitted for publication.
[9] J. Kautsky and G.H. Golub, On the calculation of Jacobi matrices, *Linear Algebra Appl.* **52/53** (1983) 439–455.
[10] J.H. Wilkinson, *The Algebraic Eigenvalue Problem* (Clarendon Press, Oxford, 1965).

10.8. [147] "ON QUADRATURE CONVERGENCE OF EXTENDED LAGRANGE INTERPOLATION"

---

[147] (with S. Li) "On Quadrature Convergence of Extended Lagrange Interpolation," *Math. Comp.* **65**, 1249–1256 (1996).

MATHEMATICS OF COMPUTATION
Volume 65, Number 215
July 1996, Pages 1249–1256

ON QUADRATURE CONVERGENCE OF EXTENDED LAGRANGE INTERPOLATION

WALTER GAUTSCHI AND SHIKANG LI

ABSTRACT. Quadrature convergence of the extended Lagrange interpolant $L_{2n+1}f$ for any continuous function f is studied, where the interpolation nodes are the n zeros τ_i of an orthogonal polynomial of degree n and the $n+1$ zeros $\hat{\tau}_j$ of the corresponding "induced" orthogonal polynomial of degree $n+1$. It is found that, unlike convergence in the mean, quadrature convergence *does* hold for all four Chebyshev weight functions. This is shown by establishing the positivity of the underlying quadrature rule, whose weights are obtained explicitly. Necessary and sufficient conditions for positivity are also obtained in cases where the nodes τ_i and $\hat{\tau}_j$ interlace, and the conditions are checked numerically for the Jacobi weight function with parameters α and β. It is conjectured, in this case, that quadrature convergence holds for $|\alpha| \leq \frac{1}{2}$, $|\beta| \leq \frac{1}{2}$.

1. INTRODUCTION

If $\pi_n(\cdot\,;w)$, $n \geq 1$, denotes the nth-degree orthogonal polynomial on $[-1,1]$ with respect to a positive weight function w, and $(L_nf)(\cdot)$ the Lagrange interpolation polynomial of degree $< n$ interpolating f at the zeros $\{\tau_i\}$ of π_n, it is a well-known result of Erdös and Turán [2] that L_nf converges in the mean to f for any continuous function f. That is,

$$\lim_{n\to\infty} \| f - L_nf \|_w = 0, \text{ all } f \in C[-1,1], \tag{1.1}$$

where $\| g \|_w = \left(\int_{-1}^{1} g^2(t)w(t)dt\right)^{1/2}$. Attempts have been made in the past to obtain an analogous result for the extended Lagrange interpolant $(\hat{L}_{2n+1}f)(\cdot)$ interpolating f at $2n+1$ points — the n points $\{\tau_i\}$ and $n+1$ additional points $\{\hat{\tau}_j\}$ suitably chosen. A particularly interesting choice of the $\hat{\tau}_j$, first suggested by Bellen [1], is given by the zeros of $\hat{\pi}_{n+1}$, the polynomial $\hat{\pi}_{n+1}(\cdot) = \pi_{n+1}(\cdot\,;\pi_n^2 w)$ of degree $n+1$ "induced by π_n", i.e., orthogonal relative to the weight function $\pi_n^2 w$ (cf. [5]). Concrete results have only been obtained in the case of Chebyshev weight functions. The one of the second kind, $w(t) = (1-t^2)^{1/2}$, is particularly easy, since in this case $\{\tau_i\} \cup \{\hat{\tau}_j\}$ are precisely the zeros of $U_nT_{n+1} = U_{2n+1}$ (cf. [1]), and one is led back to the Erdös-Turán result. For all other three Chebyshev weight functions, however, one of us [3] has shown that mean convergence cannot hold for all continuous functions.

Received by the editor April 20,1995.
1991 *Mathematics Subject Classification.* Primary 41A05, 65D32; Secondary 33C45.

It may be interesting to ask the same question for what Erdös and Turán called *quadrature convergence*. In their scenario, that would mean

$$\lim_{n\to\infty}\int_{-1}^{1}[f(t)-(L_nf)(t)]w(t)dt=0,\ \text{all } f\in C[-1,1], \tag{1.2}$$

which is obviously true, since the integral over L_nf is just the n-point Gauss quadrature sum relative to the weight function w. Is it true that the same holds for extended interpolation,

$$\lim_{n\to\infty}\int_{-1}^{1}[f(t)-(\hat{L}_{2n+1}f)(t)]w(t)dt=0,\ \text{all } f\in C[-1,1]? \tag{1.3}$$

The answer is yes, if the underlying quadrature rule has all weights positive, as follows from a classical result of Pólya [6]. We will show in §2 of this note that this is indeed the case, for all four Chebyshev weight functions, and in the process also determine explicitly the weights of the quadrature rules involved. Moreover, it will be shown in §3 that positivity also holds if the nodes τ_i and $\hat{\tau}_j$ interlace, provided the Gauss weights for the weight function w satisfy certain inequalities. The latter are checked numerically for the Jacobi weight function $w^{(\alpha,\beta)}(t)=(1-t)^\alpha(1+t)^\beta$, and evidence is produced suggesting that the quadrature weights in question are indeed positive if $|\alpha|\le\frac{1}{2}$, $|\beta|\le\frac{1}{2}$.

One could be tempted to take the zeros of π_{n+1} as the additional nodes $\hat{\tau}_j$ since interlacing is then guaranteed. However, the quadrature rule implied by (1.3) is then simply the $(n+1)$-point Gaussian rule for w (all nodes τ_i receive weight zero), and we are back again to the Erdös-Turán result!

2. Chebyshev weight functions

The weights of the interpolatory quadrature rule implied by (1.3) are given by

$$\lambda_i=\int_{-1}^{1}\frac{\pi_n(t)\hat{\pi}_{n+1}(t)}{(t-\tau_i)\pi_n'(\tau_i)\hat{\pi}_{n+1}(\tau_i)}\,w(t)dt,\qquad i=1,2,\dots,n; \tag{2.1}$$

$$\mu_j=\int_{-1}^{1}\frac{\pi_n(t)\hat{\pi}_{n+1}(t)}{(t-\hat{\tau}_j)\pi_n(\hat{\tau}_j)\hat{\pi}_{n+1}'(\hat{\tau}_j)}\,w(t)dt,\qquad j=1,2,\dots,n+1, \tag{2.2}$$

where $\pi_n(\cdot)=\pi_n(\,\cdot\,;w)$ and $\hat{\pi}_{n+1}(\cdot)=\pi_{n+1}(\,\cdot\,;\pi_n^2w)$. The rule has degree of exactness equal to $2n$. For reasons indicated in the Introduction, it suffices to look at Chebyshev weights of the first, third, and fourth kind.

2.1. Chebyshev weight of the first kind.

Here the weight function is $w_1(t)=(1-t^2)^{-1/2}$, and π_n is the Chebyshev polynomial of the first kind,

$$\pi_n(t)=T_n(t),\qquad T_n(\cos\theta)=\cos n\theta, \tag{2.3}$$

whereas $\hat{\pi}_{n+1}$ is given by [3]

$$\hat{\pi}_{n+1}(t)=\ T_{n+1}(t)-\tfrac{1}{2}\,T_{n-1}(t),\quad n\ge 1. \tag{2.4}$$

Theorem 2.1. *For $w_1(t)=(1-t^2)^{-1/2}$, the quadrature weights λ_i and μ_j in (2.1), (2.2) are given by*

$$\lambda_i=\frac{\pi}{3n},\qquad i=1,2,\dots,n; \tag{2.5}$$

$$\mu_j = \frac{2\pi}{3} \frac{1}{n + \dfrac{3}{9 - 8\hat{\tau}_j^2}}, \qquad j = 1, 2, \ldots, n+1, \tag{2.6}$$

where $\hat{\tau}_j$ are the zeros of $\hat{\pi}_{n+1}$. All weights are positive.

Proof. It follows easily from (2.3) and (2.4) that $\pi_n'(\tau_i) = n(-1)^{i-1}/\sin\theta_i$ and $\hat{\pi}_{n+1}(\tau_i) = \frac{3}{2}\,(-1)^i \sin\theta_i$, where $\theta_i = (2i-1)\pi/2n$, so that

$$\pi_n'(\tau_i)\hat{\pi}_{n+1}(\tau_i) = -\ \tfrac{3}{2}\ n. \tag{2.7}$$

It remains, for λ_i, to evaluate the integral

$$\int_{-1}^{1} \frac{T_n(t)[T_{n+1}(t) - \frac{1}{2}\,T_{n-1}(t)]}{t - \tau_i}\, w_1(t)dt.$$

Since τ_i is a zero of T_n, the integral, by orthogonality of the T_m, reduces to

$$-\ \frac{1}{2} \int_{-1}^{1} \frac{T_n(t)}{t - \tau_i}\, T_{n-1}(t) w_1(t)dt,$$

which in turn is equal to $-\ \frac{\pi}{2}$. This follows by observing, if $n > 1$, that

$$\frac{T_n(t)}{t - \tau_i} = 2T_{n-1}(t) + \text{ lower-degree terms},$$

by orthogonality, and by using

$$\int_{-1}^{1} T_m^2(t) w_1(t)dt = \frac{\pi}{2}, \qquad m \geq 1.$$

For $n = 1$, the reasoning is the same except for the factor and divisor 2 in the last two formulae, which must be replaced by 1. The result (2.5) now follows immediately.

To evaluate the constant in the denominator of (2.2), we let

$$\hat{\tau}_j = \cos\hat{\theta}_j$$

and obtain from

$$\hat{\pi}_{n+1}(\cos\theta) = \ \cos(n+1)\theta - \tfrac{1}{2}\cos(n-1)\theta$$

by differentiation and the addition formula for the sine

$$\hat{\pi}_{n+1}'(\hat{\tau}_j) = \frac{1}{2\sin\hat{\theta}_j}\{(n+3)\sin n\hat{\theta}_j \cos\hat{\theta}_j + (3n+1)\cos n\hat{\theta}_j \sin\hat{\theta}_j\}. \tag{2.8}$$

Since

$$\cos(n+1)\hat{\theta}_j - \tfrac{1}{2}\cos(n-1)\hat{\theta}_j = 0,$$

and using here the addition formula for the cosine, we find

$$\sin n\hat{\theta}_j = \frac{1}{3}\,\frac{\cos n\hat{\theta}_j \cos\hat{\theta}_j}{\sin\hat{\theta}_j}.$$

Together with (2.8), this yields after a simple computation

$$\pi_n(\hat{\tau}_j)\hat{\pi}_{n+1}'(\hat{\tau}_j) = \frac{1}{2}\,T_n^2(\hat{\tau}_j)\left\{\frac{n+3}{3}\,\frac{\hat{\tau}_j^2}{1-\hat{\tau}_j^2} + 3n + 1\right\}.$$

It is known from [3, Eq. (2.7)] that $T_n^2(t) = 9(1-t^2)/(9-8t^2)$ for $t = \hat{\tau}_j$. Therefore,

$$\pi_n(\hat{\tau}_j)\hat{\pi}'_{n+1}(\hat{\tau}_j) = \frac{3(9n+3-8n\hat{\tau}_j^2)}{2(9-8\hat{\tau}_j^2)} = \frac{3}{2}\left(n + \frac{3}{9-8\hat{\tau}_j^2}\right). \tag{2.9}$$

For the integral in (2.2), we proceed as follows:

$$\int_{-1}^{1} \frac{T_n(t)[T_{n+1}(t) - \frac{1}{2}\,T_{n-1}(t)]}{t-\hat{\tau}_j}\, w_1(t)dt = 2\int_{-1}^{1} T_n^2(t)w_1(t)dt = \pi. \tag{2.10}$$

The first equality is a result of the orthogonality of the T_m and the fact that

$$\frac{T_{n+1}(t) - \frac{1}{2}\,T_{n-1}(t)}{t-\hat{\tau}_j} = 2T_n(t) + \text{ lower-degree terms.}$$

Combining (2.10) and (2.9) yields (2.6).

The positivity of the quadrature weights is an immediate consequence of $-1 < \hat{\tau}_j < 1$ for the μ_j, and trivial for the λ_i. □

Since $\sum_{i=1}^{n}\lambda_i + \sum_{j=1}^{n+1}\mu_j = \pi$, it follows from Theorem 2.1 that the nodes $\hat{\tau}_j$ must satisfy

$$\sum_{j=1}^{n+1} \frac{1}{n + \dfrac{3}{9-8\hat{\tau}_j^2}} = 1. \tag{2.11}$$

2.2. **Chebyshev weights of the third and fourth kind.** Because of the remark at the beginning of §3.2 below, it suffices to examine the Chebyshev weight function of the third kind, $w_3(t) = (1-t)^{-1/2}(1+t)^{1/2}$, for which the relevant polynomials are

$$\pi_n(t) = V_n(t), \qquad V_n(\cos\theta) = \frac{\cos\left(n+\frac{1}{2}\right)\theta}{\cos\frac{1}{2}\theta} \tag{2.12}$$

and [3, Eq. (2.17)]

$$\hat{\pi}_{n+1}(t) = T_{n+1}(t) - \tfrac{1}{2}\,T_n(t), \qquad n \geq 1. \tag{2.13}$$

Theorem 2.2. *For $w_3(t) = (1-t)^{-1/2}(1+t)^{1/2}$, the quadrature weights λ_i and μ_j in (2.1), (2.2) are given by*

$$\lambda_i = \frac{2\pi}{3}\,\frac{1+\tau_i}{2n+1}, \qquad i = 1,2,\dots,n; \tag{2.14}$$

$$\mu_j = \frac{2\pi}{3}\,\frac{1+\hat{\tau}_j}{n + \dfrac{4-2\hat{\tau}_j}{5-4\hat{\tau}_j}}, \qquad j = 1,2,\dots,n+1, \tag{2.15}$$

where τ_i and $\hat{\tau}_j$ are the zeros of π_n and $\hat{\pi}_{n+1}$, respectively. All weights are positive.

Proof. From (2.12) and (2.13), one obtains by an elementary computation that $\pi'_n(\tau_i) = \left(n+\frac{1}{2}\right)(-1)^{i-1}/\left(\cos\frac{1}{2}\theta_i \sin\theta_i\right)$ and $\hat{\pi}_{n+1}(\tau_i) = \frac{3}{2}\,(-1)^i \sin\frac{1}{2}\theta_i$, where $\theta_i = (2i-1)\pi/(2n+1)$, so that the constant in the integral of (2.1) is

$$\pi'_n(\tau_i)\hat{\pi}_{n+1}(\tau_i) = -\frac{3}{4}\,\frac{2n+1}{1+\tau_i}. \tag{2.16}$$

The integral itself is

$$I_n = \int_{-1}^{1} \frac{V_n(t)}{t-\tau_i}\,[T_{n+1}(t) - \frac{1}{2}\,T_n(t)]w_3(t)dt$$
$$= \int_{-1}^{1} \frac{V_n(t)}{t-\tau_i}\,(1+t)[T_{n+1}(t) - \frac{1}{2}\,T_n(t)]w_1(t)dt.$$

Since, by the recurrence relation for the T_m, we have

$$(1+t)[T_{n+1}(t) - \tfrac{1}{2}\,T_n(t)] = \tfrac{1}{2}\,T_{n+2}(t) + \tfrac{3}{4}\,T_{n+1}(t) - \tfrac{1}{4}\,T_{n-1}(t),$$

we can use the orthogonality of the T_m with respect to w_1 to simplify:

$$I_n = -\frac{1}{4}\int_{-1}^{1} \frac{V_n(t)}{t-\tau_i}\,T_{n-1}(t)w_1(t)dt.$$

Now $V_n(t)$ has leading coefficient 2^n, if $n \geq 2$, so that

$$I_n = -\int_{-1}^{1} T_{n-1}^2(t)w_1(t)dt = -\frac{\pi}{2}\,, \qquad n \geq 2.$$

The same result holds also for $n = 1$. Combining it with (2.16) yields (2.14).

Letting as before $\hat{\tau}_j = \cos\hat{\theta}_j$, putting $t = \cos\theta$ in (2.13), and differentiating with respect to θ gives

$$\hat{\pi}'_{n+1}(\hat{\tau}_j) = \frac{n+2}{4}\,\frac{\sin\left(n+\frac{1}{2}\right)\hat{\theta}_j}{\sin\frac{1}{2}\hat{\theta}_j} + \frac{3n+2}{4}\,\frac{\cos\left(n+\frac{1}{2}\right)\hat{\theta}_j}{\cos\frac{1}{2}\hat{\theta}_j}\,.$$

Since

$$\cos(n+1)\hat{\theta}_j - \tfrac{1}{2}\cos n\hat{\theta}_j = 0,$$

this simplifies to

$$\hat{\pi}'_{n+1}(\hat{\tau}_j) = \frac{V_n(\hat{\tau}_j)}{4}\left(\frac{n+2}{3}\,\frac{1+\hat{\tau}_j}{1-\hat{\tau}_j} + 3n+2\right).$$

From [3, Eq. (2.22)] it is known that

$$V_n^2(t) = \frac{9(1-t)}{(1+t)(5-4t)} \text{ for } t = \hat{\tau}_j,$$

so that the constant in the integral of (2.2) becomes

$$\pi_n(\hat{\tau}_j)\hat{\tau}'_{n+1}(\hat{\tau}_j) = \frac{3}{2}\,\frac{n + \dfrac{4-2\hat{\tau}_j}{5-4\hat{\tau}_j}}{1+\hat{\tau}_j}\,. \tag{2.17}$$

The integral, on the other hand, is

$$\int_{-1}^{1} \frac{T_{n+1}(t) - \frac{1}{2}\,T_n(t)}{t-\hat{\tau}_j}\,V_n(t)w_3(t)dt,$$

which, since

$$\frac{T_{n+1}(t) - \frac{1}{2}\,T_n(t)}{t-\hat{\tau}_j} = V_n(t) + \text{ lower-degree terms},$$

reduces to

$$\int_{-1}^{1} V_n^2(t)w_3(t)dt = \pi, \qquad n \geq 1.$$

Together with (2.17), this yields (2.15).

The positivity of the weights is evident from (2.14), (2.15). □

Analogously to (2.11) one finds, after an elementary calculation, that

$$\sum_{j=1}^{n+1} \frac{1+\hat{\tau}_j}{n+\dfrac{4-2\hat{\tau}_j}{5-4\hat{\tau}_j}} = 1. \tag{2.18}$$

3. Jacobi weight functions

For more general weight functions, in particular the Jacobi weight function $w(t) = w^{(\alpha,\beta)}(t)$, where $w^{(\alpha,\beta)}(t) = (1-t)^\alpha(1+t)^\beta$, we have only conjectural results based on numerical experimentation. We are especially interested in cases where the nodes $\{\tau_i\}$ and $\{\hat{\tau}_j\}$ interlace,

$$\hat{\tau}_{n+1} < \tau_n < \hat{\tau}_n < \tau_{n-1} < \cdots < \hat{\tau}_2 < \tau_1 < \hat{\tau}_1. \tag{3.1}$$

We shall assume in this section (in slight contrast to §2) that the polynomials π_n and $\hat{\pi}_{n+1}$ are monic.

3.1. **Quadrature weights for interlacing nodes.** We assume, as in §2, that n is given and fixed. Our computations are based on the following theorem.

Theorem 3.1. *Let w be any (positive) weight function for which the nodes $\{\tau_i\}$, $\{\hat{\tau}_j\}$ (defined in §2) interlace. Then the quadrature weights λ_i and μ_j in (2.1), (2.2) are all positive if and only if*

$$\lambda_i^G > \frac{\| \pi_n \|_w^2}{|\pi_n'(\tau_i)\hat{\pi}_{n+1}(\tau_i)|}, \qquad i = 1,2,\dots,n, \tag{3.2}$$

where λ_i^G are the Christoffel numbers of the n-point Gaussian quadrature rule for the weight function w, and $\| \pi_n \|_w^2 = \int_{-1}^1 \pi_n^2(t)w(t)dt$.

Proof. We first show that the interlacing property implies $\mu_j > 0$. It is clear from (3.1) that

$$\pi_n(\hat{\tau}_j)\hat{\pi}_{n+1}'(\hat{\tau}_j) > 0, \qquad j = 1,2,\dots,n+1. \tag{3.3}$$

Thus the constant in the denominator of (2.2) is positive. In the integral that remains, the integrand is a monic polynomial of degree $2n$. Its $(2n)$th derivative divided by $(2n)!$ is therefore constant equal to 1, and the n-point Gauss formula with remainder term yields

$$\int_{-1}^1 \frac{\pi_n(t)\hat{\pi}_{n+1}(t)}{t-\hat{\tau}_j} w(t)dt = \sum_{k=1}^n \lambda_k^G \frac{\pi_n(\tau_k)\hat{\pi}_{n+1}(\tau_k)}{\tau_k - \hat{\tau}_j} + \int_{-1}^1 \pi_n^2(t)w(t)dt = \| \pi_n \|_w^2,$$

since $\pi_n(\tau_k) = 0$ for all k. Therefore,

$$\mu_j = \frac{\| \pi_n \|_w^2}{\pi_n(\hat{\tau}_j)\hat{\pi}_{n+1}'(\tau_j)}, \qquad j = 1,2,\dots,n+1, \tag{3.4}$$

and the positivity of the μ_j follows from (3.3).

Similarly, for the λ_i we have

$$\int_{-1}^{1} \frac{\pi_n(t)}{t-\tau_i}\,\hat{\pi}_{n+1}(t)w(t)dt = \lambda_i^G \pi_n'(\tau_i)\hat{\pi}_{n+1}(\tau_i) + \| \pi_n \|_w^2,$$

so that from (2.1)

$$\lambda_i = \lambda_i^G + \frac{\| \pi_n \|_w^2}{\pi_n'(\tau_i)\hat{\pi}_{n+1}(\tau_i)}, \qquad i = 1, 2, \ldots, n. \tag{3.5}$$

Now, however, interlacing implies

$$\pi_n'(\tau_i)\hat{\pi}_{n+1}(\tau_i) < 0, \qquad i = 1, 2, \ldots, n,$$

so that $\lambda_i > 0$ for all i if and only if (3.2) holds. □

3.2. **Numerical results for the Jacobi weight function.** For $w(t) = w^{(\alpha,\beta)}(t)$ it suffices to consider $\beta \geq \alpha > -1$, since an interchange of α and β only changes the sign of the argument t in $\pi_n(t)$ and $\hat{\pi}_{n+1}(t)$, hence the signs of the zeros τ_i and $\hat{\tau}_j$, and the weights λ_i and μ_j in (2.1), (2.2) remain the same, as is easily seen.

In order to check the positivity of the weights λ_i and μ_j numerically, we used Theorem 3.1 and examined, first of all, whether interlacing of the zeros holds, and if so, whether or not the inequalities (3.2) are valid for all n up to some large limit (below we take $n \leq 160$). For computational purposes we found it convenient to write these inequalities in the form

$$\lambda_i^G > \frac{\beta_0\beta_1\cdots\beta_n}{\prod_{\substack{k=1\\k\neq i}}^{n} |\tau_i - \tau_k| \prod_{j=1}^{n+1} |\tau_i - \hat{\tau}_j|}, \qquad i = 1, 2, \ldots, n, \tag{3.2'}$$

where the β's are the coefficients in the recurrence relation

$$\begin{aligned} \pi_{\nu+1}(t) &= (t-\alpha_\nu)\pi_\nu(t) - \beta_\nu\pi_{\nu-1}(t), \qquad \nu = 0, 1, 2, \ldots, \\ \pi_0(t) &= 1, \ \pi_{-1}(t) = 0 \end{aligned} \tag{3.6}$$

for the polynomials $\pi_\nu(\,\cdot\,) = \pi_\nu(\,\cdot\,; w^{(\alpha,\beta)})$. To generate these coefficients, and with them the polynomial π_n and its zeros τ_i, we used the routines `recur` and `gauss` in [4]. Similarly for the polynomial $\hat{\pi}_{n+1}$, where we used the routines `indp` and `gauss`. All calculations were done in double precision on a Sun SPARCstation IPX, using Fortran Version 2.0, for $n = 1(1)160$. We found that interlacing and/or positivity fails for $\alpha < -\frac{1}{2}$ and $\alpha > 1$, and also for $-\frac{1}{2} \leq \alpha \leq 1$ and $\beta > 1$. On the other hand, there is strong evidence for both interlacing and positivity to hold if $|\alpha| \leq \frac{1}{2}$, $|\beta| \leq \frac{1}{2}$. Both may even hold for somewhat larger values of α and β, as suggested in Fig. 3.1, where they seem to hold in the triangular-like region, and its reflection with respect to the diagonal $\alpha = \beta$, bounded on the left by the line $\alpha = -\frac{1}{2}$, below by $\alpha = \beta$, and on top by the dashdotted line (for $1 \leq n \leq 40$), the dashed line (for $1 \leq n \leq 80$), and the solid line (for $1 \leq n \leq 160$). We say "seem to hold" since interlacing and the inequality (3.2′) were verified numerically only for discrete points in the (α, β)-plane spaced apart by .1 in most of the region, and by .001 (in the β-values) near the top of the region. We also verified the failure (for some n) of either interlacing or (3.2′) for $\alpha = -\frac{1}{2} - .01$ and $\beta = -\frac{1}{2}(.1)1$, as well as for $\alpha = \beta = 1(.1)2$. It seems safe, therefore, to state the following conjecture.

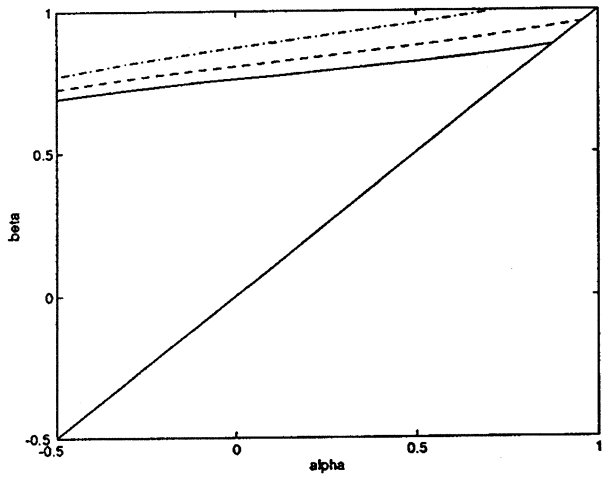


FIGURE 3.1. Positivity of quadrature weights for the Jacobi weight function $w^{(\alpha,\beta)}$

Conjecture 3.1. *For the Jacobi weight function* $w(t) = w^{(\alpha,\beta)}(t)$ *the quadrature weights* λ_i *and* μ_j *in* (2.1), (2.2) *are all positive if* (α, β) *is in the square* $|\alpha| \leq \frac{1}{2}$, $|\beta| \leq \frac{1}{2}$.

The positivity expressed in Conjecture 3.1 has been proved in §2 at the four corner points of the square.

REFERENCES

1. A. Bellen, *Alcuni problemi aperti sulla convergenza in media dell'interpolazione Lagrangiana estesa*, Rend. Ist. Mat. Univ. Trieste **20** (1988), 1–9. MR **92e:**41001
2. P. Erdös and P. Turán, *On interpolation I*, Ann. Math. **38** (1937), 142–155.
3. W. Gautschi, *On mean convergence of extended Lagrange interpolation*, J. Comput. Appl. Math. **43** (1992), 19–35. MR **93j:**41003
4. W. Gautschi, *Algorithm 726: ORTHPOL — A package of routines for generating orthogonal polynomials and Gauss-type quadrature rules*, ACM Trans. Math. Software **20** (1994), 21–62.
5. W. Gautschi and S. Li, *A set of orthogonal polynomials induced by a given orthogonal polynomial*, Aequationes Math. **46** (1993), 174–198. MR **94e:**33012
6. G. Pólya, *Über die Konvergenz von Quadraturverfahren*, Math. Z. **37** (1933), 264–286.

DEPARTMENT OF COMPUTER SCIENCES, PURDUE UNIVERSITY, WEST LAFAYETTE, INDIANA 47907-1398

E-mail address: wxg@cs.purdue.edu

DEPARTMENT OF MATHEMATICS, SOUTHEASTERN LOUISIANA UNIVERSITY, HAMMOND, LOUISIANA 70402

E-mail address: kli@selu.edu

10.9. [165] "Remark: 'Barycentric formulae for cardinal (SINC-) interpolants' by Jean-Paul Berrut"

---

[165] "Remark: 'Barycentric formulae for cardinal (SINC-) interpolants' by Jean-Paul Berrut," *Numer. Math.* **87**, 791–792 (2001).

Numer. Math. (2001) 87: 791–792
Digital Object Identifier (DOI) 10.1007/s002110000232

Numerische Mathematik

Remark

Barycentric formulae for cardinal (SINC-) interpolants by Jean-Paul Berrut

Walter Gautschi

Purdue University, Department of Computer Sciences, West Lafayette, IN 47907-1398, USA

Received February 14, 2000 / Published online October 16, 2000 –

Summary. A formula for the efficient evaluation of the (truncated) cardinal series is known to be numerically unstable near the interpolation abscissae. Here it is shown how the series can be evaluated in an entirely stable manner.

Mathematics Subject Classification (1991): 65D05

As is well known, the (symmetrically truncated) cardinal series of a function f,

$$C_N(f,h)(x) = \sum_{k=-N}^{N} f(kh)\,\mathrm{sinc}\left(\frac{x-kh}{h}\right), \tag{1}$$

where $\mathrm{sinc}(u) = (\sin \pi u)/\pi u$ if $u \neq 0$ and $\mathrm{sinc}(u) = 1$ if $u = 0$, can be written as

$$C_N(f,h)(x) = \frac{h}{\pi} \sin\frac{\pi x}{h} \sum_{k=-N}^{N} \frac{(-1)^k}{x-kh} f(kh). \tag{2}$$

This formula has the advantage of requiring only one value of the sine function to be evaluated, but, as observed in [1, p. 707], the drawback of being unstable when x is very close to one of the interpolation abscissae kh. In that case, one of the terms in the summation of (2) is extremely large in absolute value and "overshadows", i.e., reduces or even eliminates the influence of, all the other terms. As a result, the sum is obtained with low relative accuracy, and even an accurate evaluation of the sine factor cannot salvage the accuracy.

We wish to point out here that the difficulty can be avoided by a simple rearrangement of the computation. Let the integer k_0 and the real number t be such that

$$x = (k_0 + t)h, \quad |t| \le \tfrac{1}{2}. \tag{3}$$

Assume for simplicity that $|k_0| \le N$. When x is close to one of the abscissae kh, then $|t|$ is very small. Since

$$\sin \frac{\pi x}{h} = \sin \pi(k_0 + t) = (-1)^{k_0} \sin \pi t,$$

one obtains

$$C_N(f,h)(x) = (-1)^{k_0} \frac{\sin \pi t}{\pi t} \sum_{|k| \le N} \frac{(-1)^k t}{t + k_0 - k} f(kh).$$

Introducing the new index of summation $\kappa = k - k_0$, and separating out the term with $\kappa = 0$ (which occurs under the assumption $|k_0| \le N$), one finds

$$C_N(f,h)(x) = \frac{\sin \pi t}{\pi t} \left\{ f(k_0 h) + \sum_{\kappa} \frac{(-1)^\kappa t}{t - \kappa} f((k_0 + \kappa)h) \right\}, \tag{4}$$

where the summation is from $-N - k_0$ to $N - k_0$ with $\kappa = 0$ omitted.

Provided that the factor in front of the braced expression is evaluated accurately (and set to 1 if $t = 0$), the formula (4), with k_0 and t as in (3), yields accurate results even if x is very close (or equal!) to one of the abscissae kh.

If $|k_0| > N$, the same formula holds with the first term within braces omitted.

To illustrate (4) numerically, we reproduce the example in [1, p. 707], where $f(x) = x \exp(-x^2)$, $N = 190$, and $h = .1$. The formula (4) in IEEE double precision then yields (absolute) errors as shown below for the values

x	|Error|
$1/6$	2.8×10^{-17}
$.5 + 10^{-5}$	0.0
$.5 + 10^{-10}$	2.8×10^{-16}
$.5 + 10^{-15}$	5.6×10^{-17}

of x indicated. This should be compared with the errors of (2) quoted in [1] to be, respectively, 2.8×10^{-17}, 4.9×10^{-6}, 1.8×10^{-10}, and 3.6×10^{-4}.

References

1. Berrut, J.-P. (1989) Barycentric formulae for cardinal (SINC-) interpolants. Numer. Math. **54**, 703–718. [Erratum in Numer. Math. **55**, 747 (1989)]

10.10. [202] "Experimental Mathematics Involving Orthogonal Polynomials"

---

[202] "Experimental Mathematics Involving Orthogonal Polynomials," in *Approximation and computation — in honor of Gradimir V. Milovanović* (W. Gautschi, G. Mastroianni, and Th. M. Rassias, eds.), 117–134, *Springer Optim. Appl.* **42** (2011).

Experimental Mathematics Involving Orthogonal Polynomials

Walter Gautschi

Dedicated to Gradimir V. Milovanović on his 60th birthday

2000 *Mathematics Subject Classification*: 26D07 33C45 6504 6505 65D32

1 Introduction

In Wikipedia [20], the term "experimental mathematics" is defined as follows: "*Experimental mathematics is an approach to mathematics in which numerical computation is used to investigate mathematical objects and identify properties and patterns.*" The ultimate goal of experimental mathematics is to encourage, and provide direction for, purely mathematical research, in the hope of thereby extending the boundaries of mathematical knowledge. It is in this spirit that we are going to deal here with a few special topics that involve orthogonal polynomials.

A key to experimental mathematics is *numerical computation*, and that presupposes the existence of a body of reliable computational tools that allows us to generate numerically all entities of interest. In the realm of orthogonal polynomials, we are in the fortunate position of having at disposal a number of well-tested computational techniques for this purpose, supported by a package of software in Matlab, OPQ (Orthogonal Polynomials and Quadrature), in the public domain (`http://www.cs.purdue.edu/archives/2002/wxg/codes`). This not only enables but also encourages experimentation in this area of mathematics.

The *mathematical objects* we want to investigate are, on the one hand, Jacobi polynomials and, on the other hand, quadrature formulae. The *properties* and

Walter Gautschi
Department of Computer Sciences, Purdue University, West Lafayette, IN 47907-2066, USA
e-mail: wxg@cs.purdue.edu

W. Gautschi et al. (eds.), *Approximation and Computation: In Honor of Gradimir V. Milovanović*, Springer Optimization and Its Applications 42,
DOI 10.1007/978-1-4419-6594-3_9,

patterns to be identified are, in the former case, inequalities and respective domains of validity–inequalities for zeros of Jacobi polynomials and Bernstein's inequality for Jacobi polynomials–and positivity in the latter case–positivity of Newton–Cotes formulae on zeros of Jacobi polynomials and positivity of generalized Gauss–Radau and Gauss–Lobatto formulae. We will also report on experiments with Gaussian quadrature formulae corresponding to exotic weight functions, for example weight functions exhibiting super-exponential decay at infinity or densely oscillatory behavior at zero. Both are of interest in computing integral transforms that involve modified Bessel functions K_ν of complex order $\nu = \alpha + i\beta$.

2 Inequalities for Zeros of Jacobi Polynomials

We denote the zeros of the Jacobi polynomial $P_n^{(\alpha,\beta)}(x)$, $\alpha > -1$, $\beta > -1$, by

$$x_{n,r}^{(\alpha,\beta)} = \cos\theta_{n,r}^{(\alpha,\beta)}, \quad r = 1,2,\ldots,n, \tag{1}$$

and assume them, as is customary, in decreasing order,

$$0 < \theta_{n,1}^{(\alpha,\beta)} < \theta_{n,2}^{(\alpha,\beta)} < \cdots < \theta_{n,n}^{(\alpha,\beta)} < \pi. \tag{2}$$

We write $x_n^{(\alpha,\beta)} = \cos\theta_n^{(\alpha,\beta)}$ for the largest zero $x_n^{(\alpha,\beta)} = x_{n,1}^{(\alpha,\beta)}$.

It is well known (cf., e.g., [24, Theorem 8.1.2]) that for r fixed, $r \geq 1$, there holds

$$\lim_{n\to\infty} n\theta_{n,r}^{(\alpha,\beta)} = j_{\alpha,r}, \tag{3}$$

where $j_{\alpha,r}$ is the rth positive zero of the Bessel function J_α. (The speed of convergence is $O(n^{-4})$, as follows from a result of Gatteschi [6]; cf. also [18, Sect. 5.6].) The experiments in this section have to do with the pattern of convergence, specifically with monotonicity.

2.1 *Inequalities for the Largest Zero*

In [19], we considered the case $r = 1$ of the largest zero $x_{n,1}^{(\alpha,\beta)}$ of the Jacobi polynomial and tried to computationally determine for which values of α and β convergence in (3) is monotone increasing,

$$n\theta_n^{(\alpha,\beta)} < (n+1)\theta_{n+1}^{(\alpha,\beta)}, \quad n = 1,2,3,\ldots. \tag{4}$$

This requires an accurate computation of the quantities $\theta_n^{(\alpha,\beta)} = \cos^{-1} x_n^{(\alpha,\beta)}$, in particular of the largest zero $x_n^{(\alpha,\beta)}$ of $P_n^{(\alpha,\beta)}$, for arbitrary values of the parameters

$\alpha > -1, \beta > -1$. Since $x_{n,r}^{(\alpha,\beta)}$ are the nodes of the n-point Gauss–Jacobi quadrature formula, this can be readily accomplished by the OPQ routine

$$\texttt{xw} = \texttt{gauss(n,ab)}, \tag{5}$$

which in the first column of the $(n \times 2)$-array xw furnishes the n nodes $x_{n,r}^{(\alpha,\beta)}$ of the quadrature formula (though in increasing order). The second column contains the corresponding quadrature weights, which here can be ignored. The routine (5) is applicable to Gauss quadrature for any weight function whose recurrence coefficients α_k, β_k in the three-term recurrence relation

$$\pi_{k+1}(x) = (x - \alpha_k)\pi_k(x) - \beta_k \pi_{k-1}(x), \quad k = 0, 1, \ldots, n-1, \tag{6}$$

for the respective monic orthogonal polynomials π_k are known (π_{-1} in (6) is to be taken as zero, and β_0, though arbitrary, is assumed to be the zero-order moment of the weight function, i.e., its integral over the interval of orthogonality). In (5), ab is the $(n \times 2)$-array containing in the first column the coefficients $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$, and in the second column the coefficients $\beta_0, \beta_1, \ldots, \beta_{n-1}$. In the case of Jacobi polynomials, these of course are explicitly known (cf. [8, p. 29]), and are furnished by the OPQ routine

$$\texttt{ab} = \texttt{r\_jacobi(n,a,b)}, \tag{7}$$

where a and b play the role of the Jacobi parameters α and β, respectively.

Extensive experimentation with the routines (5) and (7) revealed (cf. also [13] for a small revision) that the domain of validity $\mathcal{D}$ for (4) in the (α,β)-plane is almost the full domain $\{(\alpha,\beta) : \alpha > -1, \beta > -1\}$, except for a small part near the lower left-hand corner, which has to be deleted. In fact, for $-1 < \alpha < 1$, the lower boundary of $\mathcal{D}$ near this corner is made up of two parts, the straight-line segment $\beta = -\alpha - 1$ $(-1 < \alpha < -1/2)$, on which (4) actually holds for all $n \geq 1$, and the curve $\beta = \beta(\alpha)$ $(-1/2 < \alpha < 1)$, where $\beta(\alpha)$ is the solution of the equation (in β)

$$2\arccos\left(\frac{1}{\alpha+\beta+4}\left[-(\alpha-\beta)+2\sqrt{2+\frac{\alpha\beta-2}{\alpha+\beta+3}}\right]\right) + \arccos\frac{\alpha-\beta}{\alpha+\beta+2} - \pi = 0. \tag{8}$$

This equation expresses equality in (4) for $n = 1$; see Fig. 1. The point $\alpha = \beta = -1/2$ must be deleted, since for this point one has equality in (4) for all $n \geq 1$.

It is true that these are all experimental results obtained by computation, but the numerical evidence in support of them seems compelling. It may be interesting to note that the inequalities (4) can be *proved* to hold for all n sufficiently large if $\alpha + \beta + 1 > 0$, and to be false for all n sufficiently large if $\alpha + \beta + 1 < 0$ (cf. [13, Theorem in Sect. 2]).

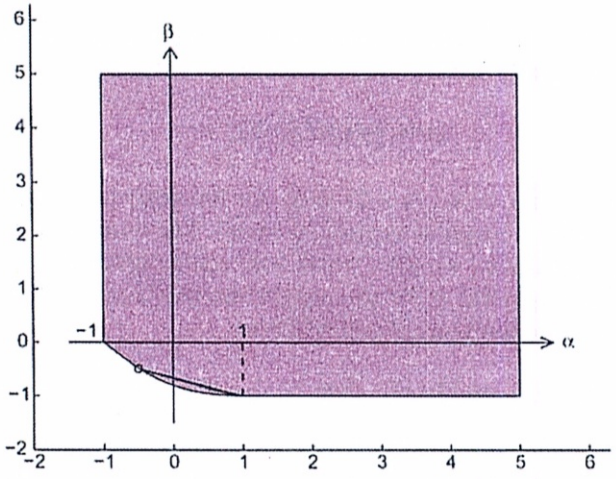


Fig. 1 Conjectured domain of validity for (4)

2.2 *Inequalities for All Zeros*

It is natural to investigate in a similar manner the case of *all* zeros $x_{n,r}^{(\alpha,\beta)} = \cos\theta_{n,r}^{(\alpha,\beta)}$, $r = 1,2,\ldots,n$, and to see to what extent the inequalities (4) continue to hold, that is, for what values of α, β one has

$$n\theta_{n,r}^{(\alpha,\beta)} < (n+1)\theta_{n+1,r}^{(\alpha,\beta)}, \quad r = 1,2,\ldots,n;\ n = 1,2,3\ldots. \tag{9}$$

While it is again true that for *fixed* r the inequalities (9) are valid for all n sufficiently large if $\alpha+\beta+1>0$ (convergence in (3), therefore, being ultimately monotone), this no longer holds if we allow $r = r(n)$ to grow with n. This can be seen already in the special case of ultraspherical polynomials ($\alpha = \beta$) and $r = n$, in which case the inequalities are found to be false for $n \geq n_0$, where $n_0 = n_0(\alpha)$ depends on α. For example, $n_0(2.2009\ldots) = 50$, $n_0(1.0605\ldots) = 100$ (cf. [14, Sect. 2]).

Restricting n in (9) to $n = 1,2,\ldots,N$, one finds [14] that the domain of validity for the inequalities, $\mathcal{D}_N$, depends on N and is bounded above by an ascending, slightly concave, curve B_N, on the left by the vertical segment at $\alpha = -1$ between B_N and the α-axis, and below by the diagonally descending line segment $\mathrm{K} = \{(\alpha,\beta) : \beta = -\alpha-1, -1<\alpha<-1/2\}$ followed by a descending, slightly convex, curve C_N (see Fig. 2).

It is suggestive to conclude from Fig. 2 that as $N \to \infty$, the domain of validity $\mathcal{D} = \mathcal{D}_\infty$ of *all* inequalities in (9) is the horizontal strip $\{(\alpha,\beta) : \alpha > -1, |\beta| \leq 1/2\}$ with the lower left-hand corner cut off by the diagonal segment K (see Fig. 3). This in fact was reinforced by the validity of (9) for $n \leq N = 500$, and for 100 points selected randomly in the strip $\{(\alpha,\beta) : -1/2 \leq \alpha \leq 20, |\beta| \leq 1/2\}$.

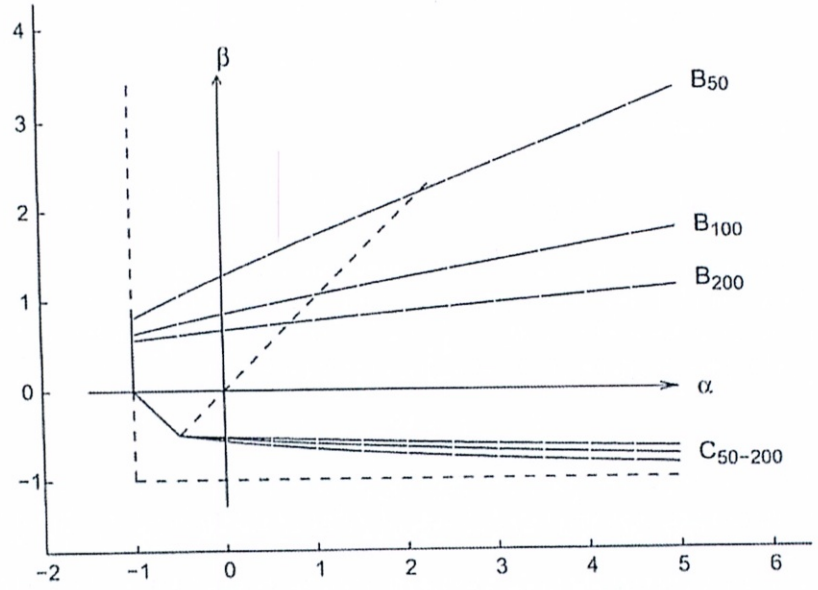


Fig. 2 Domains of validity of the inequalities in (9) when $n = 1, 2, \ldots, N$

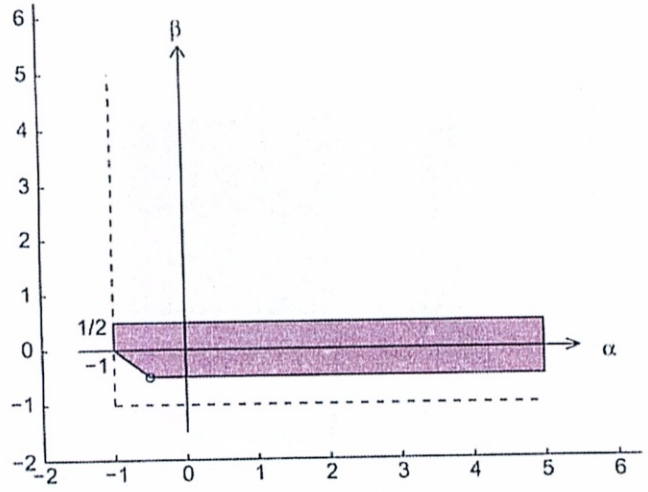


Fig. 3 Domain of validity of (9)

2.3 *Modified Inequalities for All Zeros*

In [1], Richard Askey suggested to the author to try replacing the factor n in (9) by the factor $n + (\alpha + \beta + 1)/2$, which is often more natural. This led us to consider the modified inequalities

$$(n+(\alpha+\beta+1)/2)\,\theta_{n,r}^{(\alpha,\beta)} > (n+(\alpha+\beta+3)/2)\,\theta_{n+1,r}^{(\alpha,\beta)}, \qquad r=1,2,\ldots,n;\ n=1,2,3,\ldots. \tag{10}$$

Clearly, for any *fixed* r, we have

$$\lim_{n\to\infty}((n+(\alpha+\beta+1)/2)\,\theta_{n,r}^{(\alpha,\beta)} = \lim_{n\to\infty}\frac{n+(\alpha+\beta+1)/2}{n}\,n\theta_{n,r}^{(\alpha,\beta)} = j_{\alpha,r}, \tag{11}$$

so that, if (10) holds, convergence in (11) is monotone decreasing.

It is shown in [15] that (10), for fixed r, is true for n sufficiently large if $\alpha^2+3\beta^2>1$, and false for n sufficiently large if $\alpha^2+3\beta^2<1$. Thus, if (10) is to hold for all n and r, then the point (α,β) must lie outside the ellipse

$$\mathcal{E}: \quad \alpha^2+3\beta^2=1$$

or possibly on it. Note, however, that the four points with $|\alpha|=|\beta|=1/2$, which are all on the ellipse, must be excluded, since for them equality holds in (10) for all n and r.

Using the same software as in the preceding sections, we determined by extensive numerical computation that (10) holds in the four rectangular regions (shown in Fig. 4 and extending to infinity in both the α- and β- directions) with corners at

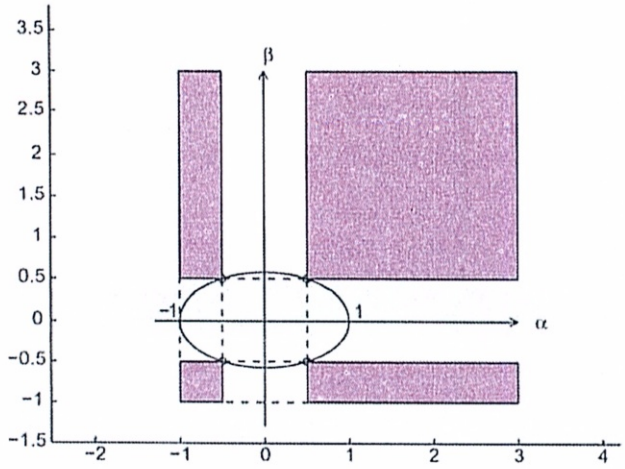


Fig. 4 (Partial) domains of validity of (10)

the four points $|\alpha|=|\beta|=1/2$ on $\mathcal{E}$. The same inequalities also hold in the vertical strips on top and bottom of the ellipse (with base interval $-1/2<\alpha<1/2$), possibly with equality near the ellipse. Inequalities in either direction are observed in the horizontal strip to the right of the ellipse and in the small remaining pieces to the left thereof.

With regard to the region inside the ellipse $\mathcal{E}$, it was found that the inequalities (10) hold with $>$ replaced by $<$ (or by $\leq$ in the two caps on the left and right, where $|\alpha| > 1/2$), and the inequality may occur in both directions in the small caps on top and bottom (where $|\beta| > 1/2$). The inequality on the upward diagonal of the square $|\alpha| < 1/2$, $|\beta| < 1/2$, called "remarkable" by Szegö, has been proved by the Sturm comparison theorem ([24, Sect. 6.3(5)]).

3 Bernstein's Inequality for Jacobi Polynomials

Originally (in 1931, see [2]), Bernstein formulated his inequality for Legendre polynomials,

$$(\sin\theta)^{1/2}|P_n(\cos\theta)| < \sqrt{2/\pi}\, n^{-1/2}, \quad 0 \leq \theta \leq \pi,$$

and showed that the constant $\sqrt{2/\pi}$ is best possible. In the 1980s, the inequality has been slightly sharpened (for example, by replacing n on the right by $n + 1/2$) and generalized by various authors to ultraspherical and, eventually, to Jacobi polynomials. A definitive form for the latter was given by Chow, Gatteschi, and Wong in [4], and reads

$$(\sin\tfrac{1}{2}\theta)^{\alpha+1/2}(\cos\tfrac{1}{2}\theta)^{\beta+1/2}|P_n^{(\alpha,\beta)}(\cos\theta)| \leq \frac{\Gamma(q+1)}{\Gamma(1/2)}\binom{n+q}{n}N^{-q-1/2}, \qquad (12)$$
$$N = n + (\alpha+\beta+1)/2, \ 0 \leq \theta \leq \pi,$$

where $|\alpha| \leq 1/2$, $|\beta| \leq 1/2$, and $q = \max(\alpha,\beta)$. Here also, the constant $\Gamma(q+1)/\Gamma(1/2)$ is best possible [16, Sect. 2].

A matter of some interest is to measure, and compute, the degree of sharpness of the inequality (12) on some domain $\mathcal{D}(n,\alpha,\beta,q)$ (where q may depend on α, β, but such that $q(\alpha,\beta) = q(\beta,\alpha)$, or may be an independent parameter). Another objective is to extend the inequality to more general regions $\mathcal{R}_s = \{(\alpha,\beta) : -1/2 \leq \alpha \leq s, -1/2 \leq \beta \leq s\}$ in the (α,β)-plane, the original region being $\mathcal{R}_{1/2}$.

3.1 Sharpness of Bernstein's Inequality

Upon dividing both sides of (12) by the expression on the right, and letting $x = \cos\theta$, the inequality (12) may be given the form

$$f_n(x;\alpha,\beta,q) \leq 1, \quad -1 \leq x \leq 1. \qquad (13)$$

Following [16], given a domain $\mathcal{D}(n,\alpha,\beta,q)$, we define the "local magnitude" of f_n by

$$\rho_n = \rho_n(\alpha,\beta,q) = \|f_n(\cdot;\alpha,\beta,q)\|_\infty, \qquad (14)$$

where the infinity norm is taken over the interval $[-1,1]$; the globally largest and smallest magnitude are then defined, respectively, by

$$\rho_{\mathcal{D}}^{+} = \max_{\mathcal{D}} \rho_n, \quad \text{and} \quad \rho_{\mathcal{D}}^{-} = \min_{\mathcal{D}} \rho_n. \tag{15}$$

If, on the one hand, $\rho_{\mathcal{D}}^{+} \le 1$, i.e., the inequality (13) holds on $\mathcal{D}$, the quantities in (15) may be interpreted as follows: on a scale from 0 to 1, the *best and worst degree of sharpness on* $\mathcal{D}$ is $\rho_{\mathcal{D}}^{+}$ and $\rho_{\mathcal{D}}^{-}$. If, on the other hand, $\rho_{\mathcal{D}}^{+} > 1$, i.e., the inequality does not hold on $\mathcal{D}$, we can "correct" it by dividing f_n by $\rho_{\mathcal{D}}^{+}$ and by considering the *modified inequality* on $\mathcal{D}$,

$$\hat{f}_n(x;\alpha,\beta,q) \le 1, \quad -1 \le x \le 1, \tag{16}$$

where

$$\hat{f}_n(x;\alpha,\beta,q) = \frac{1}{\rho_{\mathcal{D}}^{+}} f_n(x;,\alpha,\beta,q). \tag{17}$$

By construction, the best degree of sharpness on $\mathcal{D}$ of the modified inequality is $\hat{\rho}_{\mathcal{D}}^{+} = 1$, and the worst degree of sharpness

$$\hat{\rho}_{\mathcal{D}}^{-} = \frac{\rho_{\mathcal{D}}^{-}}{\rho_{\mathcal{D}}^{+}}. \tag{18}$$

To compute the quantities in (15), one takes the maximum and minimum (if necessay) on a sufficiently fine grid of $\mathcal{D}$. This then leaves us with the problem of computing the infinity norm in (14). In the case of Bernstein's inequality this can conveniently be done by computing local extrema of $f_n(x;\alpha,\beta,q)$ in $-1 < x < 1$ and comparing them with the boundary values at $x = \pm 1$. The two routines in (7) and (5) are again heavily engaged in this endeavor, together with a routine for computing the relative extrema; cf. Sects. 4 and 5 of [16] and the Appendix therein for a Matlab script. Because of the reflection formula for Jacobi polynomials, it suffices to consider $\beta \ge \alpha \ge -1/2$, the last inequality by virtue of the fact that $\|f_n\|_\infty = \infty$ if $\alpha < -1/2$.

Table 1 Degree of sharpness of Bernstein's inequality on $\mathcal{R}_{1/2}$

$q \to$	q^+	q^-	1	0.5	0	−0.5
$\rho_{\mathcal{D}}^{+}$	1.000	1.000	0.999	1.000	1.023	1.000
$\rho_{\mathcal{D}}^{-}$	0.998	0.998	0.813	0.909	0.975	0.917

On the original domain $\mathcal{R}_{1/2} = \{(\alpha,\beta) : |\alpha| \le 1/2, |\beta| \le 1/2\}$, taking $\mathcal{D} = \{[5\ 10\ 20\ 50\ 100], \mathcal{R}_{1/2}, q\}$, one obtains the results in Table 1 for selected values of q. Evidently, the choices $q^+ = \max(\alpha,\beta)$, $q^- = \min(\alpha,\beta)$ are by far the best with regard to overall sharpness.

3.2 Bernstein's Inequality on Larger Domains

We now proceed to the larger domains

$$\mathcal{R}_s = \{(\alpha,\beta) : -1/2 \le \alpha \le s, -1/2 \le \beta \le s\}, \quad s > 1/2,$$

and let $\mathcal{D}_s = \{[5\ 10\ 20\ 50\ 100], \mathcal{R}_s, q\}$. Experimentation (with $q = q^+$) revealed that

$$\rho^-_{\mathcal{D}_s} = \rho^-_{\mathcal{D}_{1/2}} \quad \text{for all } s > 1/2, \tag{19}$$

i.e., the minimum of ρ_n on $\mathcal{D}_s$ is always attained in $\mathcal{D}_{1/2}$. Also, the maximum of ρ_n on $\mathcal{D}_s$ is found to be always attained in the upper right-hand corner of $\mathcal{D}_s$,

$$\rho^+_{\mathcal{D}_s} = \max_{\mathcal{D}_s} \rho_n(\alpha,\beta,q^+) = \max_n \rho_n(s,s,s). \tag{20}$$

Using this property, we were able to compute for many values of s both $\rho^+_{\mathcal{D}_s}$ and $\hat{\rho}^-_{\mathcal{D}_s}$ for the modified (in the sense of (16) and (17)) Bernstein's inequality. An extract of the results is given in Table 2. As can be seen from Table 2, the degrees of

Table 2 Degree of sharpness on $\mathcal{D}_s$ of the modified Bernstein's inequality

s	$\rho^+_{\mathcal{D}_s}$	$\hat{\rho}^-_{\mathcal{D}_s}$
1	1.039	0.961
2	1.120	0.891
5	1.495	0.667
10	3.047	0.327

sharpness on $\mathcal{D}_s$ of the modified Bernstein's inequality, even for s as large as 10, are well within one decimal order of magnitude. Also remarkable is the experimentally observed fact that exactly the same results are obtained if we let n go from 5 up to 200, so that the results in Table 2 are likely to be valid for all $n \ge 5$.

4 Quadrature Formulae

Our interest in this section is in the positivity of quadrature formulae. Classical Newton–Cotes formulae of moderate to large order are notorious not only for their lack of positivity, but also for their wildly oscillating weights. Newton–Cotes formulae with nonuniformly distributed nodes, however, may well exhibit positivity. A well-known example is Fejér's quadrature formula using Chebyshev nodes of the first and second kind. A Newton–Cotes formula using both kinds of nodes simultaneously, even in a more general setting, has been proposed by Milovanović and conjectured to be positive. This will be further elaborated on in Sect. 4.1. In

Sects. 4.2–4.4, we look at generalized Gauss–Radau and Gauss–Lobatto formulae, for which positivity has been conjectured by us in the past, and has been proved very recently by Joulak and Beckermann.

4.1 Positivity of Weighted Newton–Cotes Formulae

Let w be a (positive) weight function on the interval I, and X_n a set of n distinct points x_k in I. A *weighted Newton–Cotes formula* is an interpolatory quadrature formula of the form

$$\int_I w(x)f(x)\mathrm{d}x = \sum_{x_k \in X_n} w_k f(x_k), \quad f \in \mathbb{P}_{n-1}. \tag{21}$$

Definition We write $(w, X_n) \in \mathrm{NC}_+$ if and only if in (21) there holds

$$w_k > 0, \quad \text{all } k. \tag{22}$$

The quadrature rule (21) is then called a *positive* weighted Newton–Cotes formula.

A well-known example is Fejér's quadrature rule [5], for which $w(x) = 1$ on $I = [-1, 1]$, and $X_n = X_n^T$ or X_n^U, the zeros of the Chebyshev polynomial T_n of the first kind resp. U_n of the second kind. In particular, there holds

$$(w \equiv 1, X_{2n-1}^U) \in \mathrm{NC}_+. \tag{23}$$

Noting that $U_{2n-1} = 2T_n U_{n-1}$, we can write

$$X_{2n-1}^U = X_n^T \cup X_{n-1}^U. \tag{24}$$

This is the motivation for the following conjecture.

Conjecture of Milovanović ([23], [22, Sect. 5.1.2]) There holds

$$(w, X_{2n-1}) \in \mathrm{NC}_+, \tag{25}$$

where

$$w(x) = (1-x)^{\alpha+1/2}(1+x)^{\beta+1/2} \quad \text{on } [-1, 1], \tag{26}$$

and

$$X_{2n-1} = X_n^{P(\alpha,\beta)} \cup X_{n-1}^{P(\alpha+1,\beta+1)}. \tag{27}$$

Here, α and β are arbitrary real numbers greater than -1, and $X_n^{P(\alpha,\beta)}$ the zeros of the Jacobi polynomial $P_n^{(\alpha,\beta)}$, and $X_{n-1}^{P(\alpha+1,\beta+1)}$ the zeros of the Jacobi polynomial $P_{n-1}^{(\alpha+1,\beta+1)}$.

Note that (23), (24) are the special case $\alpha = \beta = -1/2$ of the conjecture.

Testing the conjecture requires a reliable and stable procedure for generating the weighted Newton–Cotes formula (21), that is, the weights w_k, given n and the nodes $x_k \in X_n$. Clearly,

$$w_k = \int_I \ell_k^{(n)}(x) w(x) \mathrm{d}x, \quad k = 1, 2, \ldots, n, \tag{28}$$

where $\ell_k^{(n)}$ are the elementary Lagrange interpolation polynomials belonging to the nodes x_k,

$$\ell_k^{(n)}(x) = \prod_{\substack{\ell=1 \\ \ell \neq k}}^{n} \frac{x - x_\ell}{x_k - x_\ell}. \tag{29}$$

As already suggested in [7], the integral in (28) can be computed by $\lfloor \frac{n+1}{2} \rfloor$-point Gaussian quadrature relative to the weight function w, and the Lagrange polynomials $\ell_k^{(n)}$ by means of the barycentric formula

$$\ell_k^{(n)}(x) = \frac{\lambda_k^{(n)}/(x - x_k)}{\sum_{\ell=1}^{n} \lambda_\ell^{(n)}/(x - x_\ell)}, \quad x \neq x_k. \tag{30}$$

Here, $\lambda_k^{(n)}$ are auxiliary quantities that are readily calculated by the recursive scheme (written in pseudoMatlab)

$$\begin{aligned}
&\lambda_1^{(1)} = 1; \\
&\texttt{for k} = 2 : \texttt{n} \\
&\quad \texttt{for l} = 1 : \texttt{k} - 1 \\
&\quad\quad \lambda_\ell^{(k)} = \lambda_\ell^{(k-1)}/(x_\ell - x_k); \\
&\quad \texttt{end} \\
&\quad \lambda_k^{(k)} = \textstyle\prod_{\ell=1}^{k-1} 1/(x_k - x_\ell); \\
&\texttt{end}
\end{aligned}$$

In contrast to [7, Sect. 2.1], where $\lambda_k^{(k)}$ was computed by a sum, causing potentially severe cancellation problems, here, following [3, Sect. 3], we compute it more stably by a product. This is implemented in the OPQ routine `NewtCotes.m`, which calls on the routine `gauss.m` to do the integration.

We tested Milovanović's conjecture for $\alpha = -.75 : .25 : 5$, $\beta = \alpha : .25 : 5$, and $n = 2 : 100$ (by symmetry, it suffices to take $\beta \geq \alpha$). We also examined, in part already in [7], the ranges $\alpha = -.9 : .1 : 1.0$, $\beta = \alpha : .1 : 1.0$, and $n = 2 : 100$. In all these tests, the conjecture, without exception, was confirmed.

4.2 *Positivity of Generalized Gauss–Radau Formulae*

Generalized Gauss–Radau formulae are quadrature formulae of Gauss type, i.e., of maximum algebraic degree of exactness, that involve a boundary point of arbitrary

multiplicity $r \geq 2$ (those with $r = 1$ being the ordinary Gauss–Radau formulae). They are thus of the form

$$\int_a^\infty f(x)\mathrm{d}\lambda(x) = \sum_{\rho=0}^{r-1} \lambda_0^{(\rho)} f^{(\rho)}(a) + \sum_{\nu=1}^{n} \lambda_\nu^R f(\tau_\nu^R), \quad f \in \mathbb{P}_{2n-1+r}, \tag{31}$$

where $\mathrm{d}\lambda$ is a positive measure whose support may be bounded or unbounded. (In the former case, the upper limit of the integral could be some b with $a < b < \infty$.) The interior nodes τ_ν^R and weights λ_ν^R are easily obtained (and computed by our routine `gauss.m`) from the n-point Gaussian quadrature formula relative to the modified measure

$$\mathrm{d}\lambda^{[r]}(x) = (x-a)^r \mathrm{d}\lambda(x). \tag{32}$$

The major difficulty in computing generalized Gauss–Radau formulae lies in the boundary weights $\lambda_0^{(\rho)}$, and has been addressed only recently in [9] (see also [17] for additional difficulties when n is very large, of the order of magnitude $n \approx 400$). The method proposed in [9], and implemented in the OPQ routine `gradau.m`, is based on the solution of an $(r \times r)$ upper triangular system of linear algebraic equations,

$$Ax = b, \tag{33}$$

where the matrix A is expressible in a somewhat complicated manner in terms of the monic nth-degree polynomial $\pi_{n,r}$ orthogonal with respect to the measure (32), but the diagonal elements, and also the element in position $(r-1, r)$, are more simply expressible as

$$a_{ii} = (i-1)!\pi_{n,r}^2(a),\ i = 1,2,\ldots,r; \quad a_{r-1,r} = -2a_{rr} \sum_{\nu=1}^{n} (\tau_\nu^R - a)^{-1}. \tag{34}$$

The vector $x = [x_j]$ of unknowns in (33), and the right-hand vector $b = [b_i]$, are given by

$$x_j = \lambda_0^{(j-1)}, \quad b_i = \int_a^\infty (x-a)^{i-1} \pi_{n,r}^2(x)\mathrm{d}\lambda(x), \quad i,j = 1,2,\ldots,r. \tag{35}$$

With regard to the positivity of (31), it is clear that all interior weights λ_ν^R are positive by definition. When $r = 2$, the same is true for the boundary weights. This follows from the positivity of a_{ii} and b_i, and from $x_2 = b_2/a_{22} > 0$, $x_1 = (b_1 - a_{12}x_2)/a_{11} > 0$, since $a_{12} < 0$ by (34). In the general case $r > 2$, however, positivity of x was left open in [9], but was conjectured to hold, based on extensive computation. Today we know that positivity in fact is a proven theorem; cf. Sect. 4.4.

4.3 Positivity of Generalized Gauss–Lobatto Formulae

Generalized Gauss–Lobatto formulae are similar to generalized Gauss–Radau formulae, except that the interval of integration is necessarily bounded, and there are boundary points of multiplicity $r \geq 2$ at either end of the interval. Thus,

$$\int_a^b f(x)\mathrm{d}\lambda(x) = \sum_{\rho=0}^{r-1} \lambda_0^{(\rho)} f^{(\rho)}(a) + \sum_{\nu=1}^{n} \lambda_\nu^L f(\tau_\nu^L) + \sum_{\rho=0}^{r-1} (-1)^\rho \lambda_{n+1}^{(\rho)} f^{(\rho)}(b), \quad f \in \mathbb{P}_{2n-1+2r}. \tag{36}$$

The signs $(-1)^\rho$ in the last summation are included in anticipation of the fact that $\lambda_0^{(\rho)} = \lambda_{n+1}^{(\rho)}$ in case of symmetry (i.e., $a+b=0$ and $\mathrm{d}\lambda(-x) = \mathrm{d}\lambda(x)$).

For computational methods, which are similar to those for generalized Gauss–Radau formulae indicated in Sect. 4.2, we refer to [9] and [17]. They are implemented in the OPQ routine `globatto.m`.

Positivity of (36) is here understood to mean

$$\lambda_0^{(\rho)} > 0,\ \lambda_{n+1}^{(\rho)} > 0,\ \rho = 0,1,\ldots,r-1; \quad \lambda_\nu^L > 0,\ \nu = 1,2,\ldots,n. \tag{37}$$

The interior weights λ_ν^L, for reasons similar as in Sect. 4.2, are all positive, and so are $\lambda_0^{(\rho)}$, $\lambda_{n+1}^{(\rho)}$ if $r = 2$. For general $r > 2$, positivity of (36) has been conjectured in [9], again on the basis of extensive computation. In the meantime, it has been proven; cf. Sect. 4.4.

4.4 Positivity of Most General Gauss–Radau/Lobatto Formulae

The question of positivity regarding generalized Gauss–Radau and Gauss–Lobatto formulae has been settled very recently by H. Joulak and B. Beckermann, even for more general formulae of the form

$$\int_a^b f(x)\mathrm{d}\lambda(x) = \sum_{\rho=0}^{r-1} \lambda_0^{(\rho)} f^{(\rho)}(a) + \sum_{\nu=1}^{n} \lambda_\nu f(\tau_\nu) + \sum_{\sigma=0}^{s-1} (-1)^\sigma \lambda_{n+1}^{(\sigma)} f^{(\sigma)}(b), \quad f \in \mathbb{P}_{2n-1+r+s}, \tag{38}$$

where possibly $b = \infty$ if $s = 0$ or $a = -\infty$ if $r = 0$. In fact, we have the following theorem.

Theorem ([21]) *For any positive* $d\lambda$, *and* $r \geq 0$, $s \geq 0$, *there holds*

$$\begin{aligned} &\lambda_\nu > 0 \ \ (1 \leq \nu \leq n); \\ &\lambda_0^{(\rho)} > 0 \ \ (0 \leq \rho \leq r-1); \ \ \lambda_{n+1}^{(\sigma)} > 0 \ \ (0 \leq \sigma \leq s-1). \end{aligned} \tag{39}$$

The proof of the theorem given in [21] is based on the fact that certain elementary Hermite interpolation polynomials associated with the points a, τ_ν, b of multiplicities $r, 2, s$, respectively, are nonnegative on (a,b).

5 Gauss Quadrature with Exotic Weight Functions

In certain integral transforms involving modified Bessel functions of complex order, the integrand exhibits behavior at infinity, and at zero, that is highly unusual. To properly account for this behavior, it is necessary to develop Gaussian quadrature formulae having weight functions that mimic this behavior. This in turn requires generating the necessary orthogonal polynomials, specifically the coefficients in their three-term recurrence relation. The behavior at infinity is characterized by super-exponential decay, the one at zero by dense oscillation. In the former case, we use a discretized Stieltjes procedure to generate the necessary recurrence coefficients, in the latter case the classical Chebyshev algorithm, executed in symbolic variable-precision computation to counteract the underlying severe ill-conditioning.

5.1 Weight Function Decaying Super-Exponentially at Infinity

The real and imaginary parts of the Macdonald function (or modified Bessel function) $K_\nu(s)$ with complex order $\nu = \alpha + i\beta$ and $s > 0$ are known to be representable by integral transforms,

$$\begin{aligned} \operatorname{Re} K_{\alpha+i\beta}(s) &= \int_0^\infty e^{-s\cosh x} \cosh \alpha x \cos \beta x \, dx, \\ \operatorname{Im} K_{\alpha+i\beta}(s) &= \int_0^\infty e^{-s\cosh x} \sinh \alpha x \sin \beta x \, dx. \end{aligned} \tag{40}$$

In both, the integrand decays extremely rapidly at infinity, owing to the factor $\cosh x$ in the exponent. The essence of this behavior is captured by the weight function

$$w(x) = \exp(-e^x), \quad 0 \leq x < \infty, \tag{41}$$

not depending on s. It becomes relevant after a suitable change of variables (cf. [11, Sect. 3]).

The recurrence coefficients for the polynomials orthogonal with respect to the weight function (41) can be generated by a multiple-component discretization procedure, decomposing the interval $[0,\infty]$ into four subintervals and using Fejér quadrature rules as a general-purpose device of discretization (cf. [11, Sect. 2]). The relevant OPQ routine is `mcdis.m`, and the first 100 recurrence coefficients generated by it are listed in the file `abmacdonald`. It can be downloaded from the web site cited above in Sect. 1 by clicking on MCD. The Matlab commands

```
load -ascii abmacdonald;
xw=gauss(n,abmacdonald);
```

then produce the n-point Gauss quadrature rule relative to the weight function (41), with the first column of `xw` containing the nodes x_ν and the second column the weights w_ν.

To give an example, consider the integral

$$I = \int_0^\infty \frac{\exp(-e^x)}{1+x}\,dx \tag{42}$$

and its approximation $I_n^G = \sum_{\nu=1}^n w_\nu/(1+x_\nu)$ by the n-point Gaussian quadrature rule. In Table 3, we compare it with the n-point Gauss–Laguerre approximation I_n^L of

$$I = e^{-1}\int_0^\infty \frac{1}{(1+\ln(1+t))(1+t)}\,e^{-t}dt.$$

As can be seen, our special Gauss quadrature rule, already for $n = 13$ yields 14 correct decimal places, whereas Gauss–Laguerre manages to produce only four.

Table 3 Gauss and Gauss–Laguerre quadrature of the integral (42)

n	I_n^G	I_n^L
1	0.15171877142494	0.108637...
2	0.15936463844634	0.140436...
3	0.15987602667503	0.151468...
4	0.15991604862904	0.155900...
⋮	⋮	⋮
12	0.15991988389384	0.159868...
13	0.15991988389391	0.159886...
14	0.15991988389391	0.159897...

5.2 *Weight Functions Densely Oscillating at Zero*

Integral transforms in which K_ν, $\nu = \alpha + i\beta$, acts as a kernel, called Kontorovich–Lebedev transforms (when $\alpha = 0$ or 1/2), yield peculiar behavior of the integrand

at zero, caused by the behavior of $K_\nu(x)$ near $x = 0$. If $\alpha = 0$, for example, the transform is

$$F(\beta) = \int_0^\infty K_{i\beta}(x) f(x) \mathrm{d}x,$$

which is real-valued for real-valued f. The behavior of $K_{i\beta}$ near zero is (cf. [12, Sect. 2])

$$K_{i\beta}(x) \sim \sqrt{\frac{\pi}{\beta \sinh(\pi\beta)}} \sin(\beta \ln(2/x) + \gamma), \quad \gamma = \arg \Gamma(1 + i\beta), \quad x \downarrow 0,$$

showing that $K_{i\beta}(x)$ is densely oscillating near $x = 0$. This prompted us in [12] to consider Gauss quadrature on $[0, 1]$ with the nonnegative weight function

$$w_\beta(x) = 1 + \sin(\beta \ln(1/x) + \gamma), \quad 0 < x \leq 1.$$

But how do we find the necessary orthogonal polynomials?

The only way we knew how to do this is by applying the classical Chebyshev algorithm that allows us to generate the required recurrence coefficients directly from the moments of the weight function. The problem is that this approach via moments is quite ill-conditioned. We therefore used a symbolic version `schebyshev.m` of the OPQ routine `chebyshev.m` to generate the recurrence coefficients in variable-precision arithmetic (cf. [12, Example 3.5]). This symbolic routine can also be downloaded from the web site mentioned in Sect. 1, by clicking on SOPQ[1].

For illustration, we show here the simpler (but not less challenging!) example of the weight function

$$w(x) = 1 + \sin(1/x) \quad \text{on } [0, 1], \tag{43}$$

taken from [10, Sect. 2.1]. Here the moments $m_k = \int_0^1 x^k w(x) \mathrm{d}x$ are computed by

$$m_k = \frac{1}{k+1} + m_k^0, \quad k = 0, 1, 2, \ldots, \tag{44}$$

where m_k^0 are the "core moments" $m_k^0 = \int_0^1 x^k \sin(1/x) \mathrm{d}x$ that can be generated recursively by

$$m_{-1}^0 = \int_1^\infty \frac{\sin t}{t} \mathrm{d}t = \tfrac{\pi}{2} - \mathrm{Si}(1),$$

$$m_0^0 = \int_1^\infty \frac{\sin t}{t^2} \mathrm{d}t = \sin 1 - \mathrm{Ci}(1),$$

and

$$m_{k+1}^0 = \frac{1}{k+2} \left[\frac{1}{k+1} (\cos 1 - m_{k-1}^0) + \sin 1 \right], \quad k = 0, 1, 2, \ldots.$$

[1] SOPQ is a symbolic counterpart to the package OPQ, but far from complete. A worthwhile project for anyone familiar with the symbolic toolbox of Matlab would be to transcribe the entire package OPQ into symbolic Matlab.

To give a numerical example, suppose we want to compute the integral

$$I = \int_0^1 f(x)\sin(1/x)\mathrm{d}x,\ f(x) = \tan((\tfrac{1}{2}\pi - \delta)x), \quad 0 < \delta < \tfrac{1}{2}\pi. \tag{45}$$

Table 4 Numerical results for the integral (45) for $\delta = 0.1$

n	I_n
4	1.2716655036125
8	1.2957389942560
12	1.2961790099686
⋮	⋮
32	1.2961861708636
36	1.2961861708636

We write this in the form

$$I = \int_0^1 f(x)[1 + \sin(1/x)]\mathrm{d}x - \int_0^1 f(x)\mathrm{d}x,$$

and compute the first integral by the special Gauss formula for the weight function (43), and the second integral by Gauss–Legendre quadrature on $[0, 1]$. The results I_n for $\delta = 0.1$, using n-point Gauss formulae, are shown in Table 4.

Even the special Gauss formula here has some trouble converging, the reason being a pole close to the upper limit of the integral when δ is small.

Other densely oscillating integrals, and also integrals of rapidly decaying functions like $e^{-1/x}$ on $[0, 1]$, or $\exp(-1/x - x)$ on $[0, \infty]$, are treated in [10] similarly and with similar success.

References

1. Askey, Richard. Email of May 13, 2008.
2. Bernstein, Serge. Sur les polynomes orthogonaux relatifs a un segment fini. J. Math. 10 (1931), 219–286.
3. Berrut, Jean-Paul and Lloyd N. Trefethen. Barycentric Lagrange interpolation. SIAM Rev. 46 (2004)(3), 501–517.
4. Chow, Yunshyong, L. Gatteschi, and R. Wong. A Bernstein-type inequality for the Jacobi polynomial. Proc. Am. Math. Soc. 121 (1994)(3), 703–709.
5. Fejér, L. Mechanische Quadraturen mit positiven Cotesschen Zahlen. Math. Z. 37 (1933), 287–309.
6. Gatteschi, Luigi. On the zeros of Jacobi polynomials and Bessel functions. In: International conference on special functions: theory and computation (Turin, 1984). Rend. Sem. Mat. Univ. Politec. Torino (Special Issue), pp. 149–177 (1985)
7. Gautschi, Walter. Moments in quadrature problems. Approximation theory and applications. Comput. Math. Appl. 33 (1997)(1–2), 105–118.

8. GAUTSCHI, WALTER. *Orthogonal polynomials: computation and approximation*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 2004.
9. GAUTSCHI, WALTER. Generalized Gauss–Radau and Gauss–Lobatto formulae, BIT Numer. Math. 44 (2004)(4), 711–720.
10. GAUTSCHI, WALTER. Computing polynomials orthogonal with respect to densely oscillating and exponentially decaying weight functions and related integrals, J. Comput. Appl. Math. 184 (2005)(2), 493–504.
11. GAUTSCHI, WALTER. Numerical quadrature computation of the Macdonald function for complex orders, BIT 45 (2005)(3), 593–603.
12. GAUTSCHI, WALTER. Computing the Kontorovich–Lebedev integral transforms and their inverses, BIT 46 (2006)(1), 21–40.
13. GAUTSCHI, WALTER. On a conjectured inequality for the largest zero of Jacobi polynomials, Numer. Algorithm 49 (2008)(1–4), 195–198.
14. GAUTSCHI, WALTER. On conjectured inequalities for zeros of Jacobi polynomials, Numer. Algorithm 50 (2009)(1), 93–96.
15. GAUTSCHI, WALTER. New conjectured inequalities for zeros of Jacobi polynomials, Numer. Algorithm 50 (2009)(3), 293–296.
16. GAUTSCHI, WALTER. How sharp is Bernstein's inequality for Jacobi polynomials?, Electr. Trans. Numer. Anal. 36 (2009), 1–8.
17. GAUTSCHI, WALTER. High-order generalized Gauss–Radau and Gauss–Lobatto formulae for Jacobi and Laguerre weight functions, Numer. Algorithm 51 (2009)(2), 143–149.
18. GAUTSCHI, WALTER AND CARLA GIORDANO. Luigi Gatteschi's work on aymptotics of special functions and their zeros, Numer. Algorithm 49 (2008)(1–4), 11–31.
19. GAUTSCHI, WALTER AND PAUL LEOPARDI. Conjectured inequalities for Jacobi polynomials and their largest zeros, Numer. Algorithm 45(1–4)(2007), 217–230.
20. `http://en.Wikipedia.org/wiki/Experimental_mathematics`
21. JOULAK, HÉDI AND BERNHARD BECKERMANN. On Gautschi's conjecture for generalized Gauss–Radau and Gauss–Lobatto formulae, J. Comput. Appl. Math. 233 (2009)(3), 768–774.
22. MASTROIANNI, G. AND G.V. MILOVANOVIĆ. *Interpolation processes: basic theory and applications*, Springer Monographs in Mathematics, Springer, Berlin, 2009.
23. MILOVANOVIĆ, GRADIMIR V. Personal communication, December 1993.
24. SZEGÖ, GABOR. *Orthogonal polynomials*, 4th ed., American Mathematical Society, Colloquium Publications, Vol. 23, Amer. Math. Soc., Providence, RI, 1975.