Contemporary Mathematicians

Claude Brezinski
Ahmed Sameh
Editors

Walter Gautschi

Selected Works with Commentaries

Volume 3

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 3

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 978-1-4614-7131-8 ISBN 978-1-4614-7132-5 (eBook)
DOI 10.1007/978-1-4614-7132-5
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, 2009
hiking in the Swiss alps

Contents

Part II Reprints

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

Part I

Commentaries

In all commentaries, reference numbers preceded by "GA" refer to the numbers in the list of Gautschi's publications; see Section 4, Vol. 1. Numbers in boldface type indicate that the respective papers are included in these selected works.

21

Linear recurrence relations

Lisa Lorentzen

Walter Gautschi is a giant in the field of linear recurrence relations. His concern is with stability in computing solutions $\{y_n\}_{n=0}^{\infty}$ of such equations. Suppose the recurrence relation is of the form

$$y_{n+1} + a_n y_n + b_n y_{n-1} = 0 \qquad \text{for} \quad n = 1, 2, 3, \ldots. \tag{21.1}$$

It seems so deceivingly natural to start with values or expressions for y_0 and y_1, and then compute y_2, $y_3, \ldots$ successively from (21.1). However, this does not always work. Yet, in every new generation of mathematicians or users of mathematics, along come some incorrigible optimists with a naive trust in this method. We are happy, of course, for every new optimist in the field; mathematicians do not get far without optimism, stamina, creativity, and enthusiasm. But the new ones can definitely benefit from some sensible guidance. And what they should do, is to start with Walter Gautschi's *SIAM Review* paper [GA**29**] on three-term recurrence relations from 1967. This is what most people do, and this is what I did when I started my study of continued fractions. Continued fractions and recurrence relations indeed share a substantial intersection which, however, calls for some degree of alertness.

So what can go wrong if one computes a solution as described above? Several things, says the Master. But the worst scenario occurs if one tries to compute a solution $\{f_n\}_{n=0}^{\infty}$ of (21.1) which happens to be *minimal*. A sequence $\{f_n\}$ is a minimal solution if (21.1) has a second solution $\{y_n\}$ for which $f_n/y_n \to 0$. This second solution is then called a *dominant* solution. The solution space of (21.1) is obviously a two-dimensional vector space, so a small error in the initial data, for example a rounding error, changes $\{f_n\}$ to some dominant solution $\{\alpha f_n + \beta y_n\}$, $\beta \neq 0$, with totally different asymptotic behavior. The discrepancy between f_n and $\alpha f_n + \beta y_n$ may be catastrophic after only a few computational steps, as so convincingly demonstrated by Gautschi.

Not every such recurrence relation has a minimal solution, and one may think that the subspace of minimal solutions is so small – if it exists at all – that the chance

C. Brezinski and A. Sameh (eds.), *Walter Gautschi, Volume 3: Selected Works with Commentaries*, Contemporary Mathematicians, DOI 10.1007/978-1-4614-7132-5_1,

of encountering one is also minimal. But that is not at all the case. On the contrary, as so often in mathematics, special cases are often the most interesting ones. A number of important sequences of special functions are indeed minimal solutions of linear recurrence relations. And here we are at the heart of the problem: how can we compute minimal solutions stably and efficiently?

For recurrence relations of the form (21.1) the answer can be found in continued fraction theory: the continued fraction

$$\frac{-b_1}{-a_1-}\,\frac{b_2}{-a_2-}\,\frac{b_3}{-a_3-}\,\cdots = \frac{b_1}{a_1-}\,\frac{b_2}{a_2-}\,\frac{b_3}{a_3-}\,\cdots \tag{21.2}$$

has approximants

$$\frac{b_1}{a_1-}\,\frac{b_2}{a_2-}\,\cdots\frac{b_n}{a_n} = \frac{A_n}{B_n},$$

where $\{A_{n-1}\}_{n=0}^{\infty}$ and $\{B_{n-1}\}_{n=0}^{\infty}$ are solutions of (21.1) with initial conditions

$$A_{-1} = 1,\ A_0 = 0; \quad B_{-1} = 0,\ B_0 = 1.$$

Gautschi observes the following connection between the continued fraction (21.2) and minimal solutions of (21.1), and attributes it to Pincherle, who proved it in an obscure 1894 paper written in Italian: there exists a minimal solution $\{f_n\}$ of (21.1) satisfying $f_0 \neq 0$ if and only if the continued fraction (21.2) converges to a finite limit. In that case, moreover,

$$r_n := \frac{f_n}{f_{n-1}} = \frac{-b_n}{a_n-}\,\frac{b_{n+1}}{a_{n+1}-}\,\cdots, \quad n = 1, 2, 3, \ldots, \tag{21.3}$$

provided $f_n \neq 0$ for all n.

This immediately suggests a stable way to compute minimal solutions, namely to compute the continued fractions $r_n, r_{n-1}, \ldots, r_1$ in (21.3) and then f_n from

$$f_n = r_n r_{n-1} \cdots r_1 f_0,$$

assuming f_0 is known. For more details, see also Section 11.1, Vol. 2.

But things are not always as easy as they may look on paper. It took a Walter Gautschi to sort out the problems and work this simple idea into useful, reliable algorithms. As always, it is the stability analysis, controlling the error, that takes ingenuity. Via some very nice twists and tricks — see, e.g., Gautschi's treatment in [GA**29**, Sec. 7] and [GA**35**] of the three-term recurrence relation satisfied by Jacobi polynomials of purely imaginary parameters and argument — his algorithms work like a dream; these are not just algorithms on paper.

But what if f_0 is unknown? Also this problem was handled by Gautschi: he replaced the condition "f_0 known" by "$\sum_{n=0}^{\infty} \lambda_n y_n$ known", with known coefficients λ_n – a situation one often meets in the theory of special functions. Also this was incorporated into his algorithms. Of course, Walter Gautschi has also treated linear

recurrence relations of other forms (for example, see [GA**150**]) with the same care, and he has applied them to compute important sequences of special functions, orthogonal polynomials and interesting integrals. What is so very nice about his algorithms is that they come with such a very careful and convincing stability analysis. He has forever changed the way one looks at recurrence relations and continued fractions.

People do not only read his books and papers – they really use his results. His contributions to the *Handbook of Mathematical Functions* by Abramowitz and Stegun are frequently consulted, both his Chapter 7 on the error functions and Fresnel integrals and Chapter 5 which he wrote with W.F. Cahill on the exponential integral and related functions. Not to mention his algorithms for the complex error function, the incomplete gamma functions, the Fresnel integrals etc. in the NAG-library and other places (cf. Section 6.1, Vol. 1). To me , the very fact that so many people talk with ease about minimal solutions and stability analysis as if they had known about it all their lives, is particularly gratifying. And this happens not only in conferences on recurrence relations, but on special functions, orthogonal polynomials, continued fractions, and applied mathematics, to mention just a few.

You know your ideas have made a deep impression when fellow mathematicians begin to name concepts after you. And in the literature one finds references to the "Gautschi algorithm" number so and so, the "Gautschi method" for stability analysis, and even (more amusingly) the "Gautschi-type method" as if there were some people out there of "Gautschi-type". I think one would have a hard time finding anyone like Walter Gautschi. After the very sad death of his twin brother, Walter is unique. His clear mind and his creativity penetrate all his work, and also his oral as well as written presentations. So I end this short exposition with a serious advice: dig in and enjoy.

22

Ordinary differential equations

John Butcher

These days everyone talks about "impact" as something that can be measured in terms of citations within a year or two, but the impact of many important contributions to science can be looked at in other, more perceptive, ways. I believe this is especially true of [GA**14**]. This paper is forward-looking to the extent that its importance has become recognised more and more as time has passed. In my opinion the impact of this contribution has been tremendous. Over the years it has become known as a pioneering paper in the fitted type of approach to the solution of initial value problems. It has been referenced directly soon after its publication but even more so in recent years. It is related to exponential integration, to exponential fitting, and to modern approaches to the solution of highly-oscillatory problems. The ideas and results in the original paper have been rediscovered independently by later authors, but the depth and scholarship in Gautschi's exposition are unmatched. Here are the key definitions near the start of the paper.

A linear functional L in $C^s[a,b]$ is said to be of *algebraic order* p if

$$Lt^r = 0 \qquad (r = 0, 1, \ldots, p);$$

it is said to have *trigonometric order* p*, relative to period* T, if

$$L1 = L\cos\left(r\tfrac{2\pi}{T}t\right) = L\sin\left(r\tfrac{2\pi}{T}t\right) = 0 \qquad (r = 1, 2, \ldots, p).$$

On this foundation, the paper goes on to analytical questions concerned with the existence of trigonometric methods, the actual construction of methods, especially of Adams and Störmer types, numerical investigations, and the sensitivity of numerical results to the value of T in relation to the exact period.

The chapter [GA15] from S*urvey of numerical analysis*, McGraw-Hill, New York (1962), written in collaboration with H. A. Antosiewicz, surveys the state of knowledge, at the time, of numerical methods for ordinary differential equations. This work set the standard for theoretical expositions on this subject, appearing as it did, a short time prior to the monograph of P. Henrici. Although the work of

C. Brezinski and A. Sameh (eds.), *Walter Gautschi, Volume 3: Selected Works with Commentaries*, Contemporary Mathematicians, DOI 10.1007/978-1-4614-7132-5_2,

Curtiss and Hirschfelder had appeared several years earlier, it was not yet known and appreciated in the mathematical community. However, a cautionary example problem,

$$\frac{dy}{dx} = \begin{pmatrix} 0 & 1 \\ 10a^2 & 9a \end{pmatrix} y, \qquad y(0) = \begin{pmatrix} 1 \\ -a \end{pmatrix},$$

is presented which, for $a > 0$, leads to approximations to the solution $\exp(-ax)y(0)$ being eventually, but inevitably, overshadowed by terms which grow like $\exp(10ax)$. After stiffness had become a recognised phenomenon, it would have become more illuminating to consider $a < 0$; in this case the difficulty would not have been that the required solution is buried amongst dominant alternative solutions, but that the required solution has now become dominant even though its dominance is lost in computations with classical explicit methods.

Looking now at [GA**54**], we are reminded of a crucial time in the history of Runge–Kutta methods. This review paper acknowledged recent work, by Fehlberg and others, in constructing embedded methods for the purpose of step-size control. It appeared at a time when Henrici's monograph was becoming recognised as a model for exposition in numerical analysis and took the rigorous mathematical style a step further. But global error bounds based on very reasonable assumptions, such as the Lipschitz condition, do not necessarily give tight error bounds. This beautiful paper viewed retrospectively, encapsulates all these ideas.

Paper [GA56] contains short and elegant proofs of the asymptotic behaviour of the coefficients in Adams and other integration formulae.

For a linear k-step method (ρ, σ), where ρ is given, with zeros satisfying $1 = \zeta_1 \geq |\zeta_2| \geq |\zeta_3| \geq \cdots \geq |\zeta_k|$, there is a unique choice of σ to give order $p = k+1$. The aim of the paper [GA**73**] is to determine the method for which $|\zeta_i| \leq \gamma$, $i = 2, 3, \ldots, k$, $0 \leq \gamma < 1$, that has minimal global error constant. It is shown that in the optimal solution, $\zeta_i = -\gamma$, $i = 2, 3, \ldots, k$. Ramifications of the result are studied in detail.

Somewhere between the appearance of the first and last paper surveyed here, I met Walter Gautschi in person. I was once his guest at Purdue and met him from time to time at conferences. I have come to know him as a kind and courteous person as well as a scholarly, knowledgeable, and original mathematician.

23

Computer algorithms and software packages

Gradimir V. Milovanović

During the preparation of the *Handbook of Mathematical Functions*, under the direction of Milton Abramowitz at the Bureau of Standards (now the "National Institute of Standards and Technology"), Walter Gautschi, then a young research mathematician, joined this project in 1956. This was the starting point of a period of intense work with special functions. During the 1960s, in addition to theoretical work in several domains of special functions (see Section 6, Vol. 1), Walter developed a number of computer algorithms evaluating special functions: the gamma function and incomplete beta function ratios [GA22], Bessel functions of the first kind [GA23], Legendre functions [GA24], derivatives of e^x/x, $\cos(x)/x$, and $\sin(x)/x$ [GA27], [GA38], regular Coulomb wave functions [GA28], [GA33], the complex error function [GA36], repeated integrals of the coerror function [GA60], and incomplete gamma functions [GA69].

In 1968 Gautschi began to write computer algorithms for Gaussian quadrature formulas, the first being the one in [GA32]. This opened the door for extensive work on orthogonal polynomials and their applications (see Sections 11, 12, 14, 15 in Vol. 2), but also for developing related software. The first major software package, `ORTHPOL`, appeared in 1994 as Algorithm 726 in [GA**141**]. It contains routines, written in FORTRAN, that produce the coefficients in the three-term recurrence relation for arbitrary orthogonal polynomials as well as nodes and weights of Gauss-type quadrature rules. A more specialized package, `GQRAT` [GA159], produced Gauss quadrature rules which are exact for a combination of polynomials and rational functions. They are useful for integrating functions that have poles outside the interval of integration.

The package `ORTHPOL`, as well as the subsequent package `OPQ` of MATLAB routines, both made available on the internet (`http://www.cs.purdue.edu/archives/2002/wxg/codes`), led to a significant boost in the computational use and application of orthogonal polynomials. The companion package `SOPQ`, also available on the internet, contains symbolic versions of some of the more important routines in `OPQ`. They can be used for high-precision work in orthogonal polynomials and Gaussian

C. Brezinski and A. Sameh (eds.), *Walter Gautschi, Volume 3: Selected Works with Commentaries*, Contemporary Mathematicians, DOI 10.1007/978-1-4614-7132-5_3,

quadrature. A similar package in MATHEMATICA is `OrthogonalPolynomials` [1] (see also [2]).

A very comprehensive account of computational methods and software in MATLAB is provided in [GA**179**]. It illustrates the use of the `OPQ` routines in an elegant, interesting, and methodical way.

References

[1] Aleksandar S. Cvetković and Gradimir V. Milovanović. The Mathematica package "OrthogonalPolynomials". *Facta Univ. Ser. Math. Inform.*, 19:17–36, 2004.

[2] Gradimir V. Milovanović and Aleksandar S. Cvetković. Special classes of orthogonal polynomials and corresponding quadratures of Gaussian type. *Math. Balkanica*, 26(1–2):169–184, 2012.

24

History and biography

Gerhard Wanner

24.1. Euler

The ICIAM Congress 2007, held in Zürich, happened to be in the year of Euler's 300th anniversary. It was then clear to the organizers, that one of the invited talks should be dedicated to Euler and Euler's work. Fortunately, Walter Gautschi accepted this invitation and presented a fascinating talk on Euler's life, his personality, an overview of his work and some selected topics in more detail. This took place in the largest lecture hall (the "Turnhalle"), filled up to the last seat. I still remember the total silence in the audience, when Gautschi ran a video of an Euler gear transmission, turning, as he said, "without any noise". An expanded version of this talk [GA**187**] was prepared for the proceedings of the congress and, by mutual agreement between the publishers, also appeared in *SIAM Review* 2008, followed by a Chinese translation. Two particular items from this talk, Euler's treatment of slowly converging series and Euler's discovery of the convergence to a wrong limit of interpolatory polynomials for the logarithm, a phenomenon which 100 years later became known as q-theory, led to two separate publications, [GA**183**] and [GA**186**].

24.2. The Bieberbach conjecture

An extraordinary story is told in [GA**101**], where Gautschi, who had worked all his life on numerical analysis, quadrature, and orthogonal polynomials, suddenly had the occasion to complete, in a couple of days, Louis de Branges's proof of a long-standing conjecture in pure mathematics. This conjecture, an inequality for the Taylor coefficients of a 1-1 holomorphic mapping from a circle to a simply connected domain, was formulated by Bieberbach in 1916 during his early work on the Riemann mapping theorem. During many decades, this conjecture had resisted the efforts of the foremost experts in complex analysis. Louis de Branges finally managed to reduce this conjecture to inequalities for integrals of Jacobi polynomials and thought that Walter Gautschi, with his algorithms and computers, could help to

C. Brezinski and A. Sameh (eds.), *Walter Gautschi, Volume 3: Selected Works with Commentaries*, Contemporary Mathematicians, DOI 10.1007/978-1-4614-7132-5_4,

verify them. Gautschi not only did a lot of computer computations, but eventually found out that the inequalities had been proved a decade earlier by R. Askey and G. Gasper. I remember that P. Henrici, who lectured on this proof in January of 1985 in Stockholm on the occasion of Dahlquist's 60th anniversary, concluded his talk with the observation that a mathematician cannot know everything, but that "it is always important to know where to ask".

24.3. Survey articles

Walter Gautschi, with his broad knowledge of numerical analysis and his many personal contacts with leading experts, was (and is) in excellent position to write extraordinarily clear survey articles. Even when he wrote on a particular scientist, his narrative always turned into a beautiful and clear exposition of the underlying mathematics. We therefore collect them together: the article [GA**74**] on Gauss-Christoffel quadrature, the article [GA**143**] on Philip Rabinowitz and numerical integration, the papers [GA**144**] on 2d-iterations and numerical quadrature and [GA**189**] on asymptotics and estimation of zeros of special functions summarizing work of Luigi Gatteschi, and finally [GA**170**], the interplay between classical analysis and numerical linear algebra as a special tribute to Gene H. Golub. The same subject is dealt with in Gautschi's commentary [GA**184**], written for the edition of the selected works of Gene H. Golub.

Finally, in [GA**201**], Gautschi tells the story of how he came into scientific contact with G. V. Milovanović (we all have experienced, as referees, receiving a paper which immediately could be simplified and improved; authors then often react angrily, but in other situations such as the one described here, this was the starting point of a long friendship and collaboration). Gautschi's paper then continues with a description of Milovanović's work on Gaussian integration with unusual weight functions, and moment-preserving spline approximation.

24.4. Biography

The biography, which Gautschi wrote, was for his esteemed teacher Alexander M. Ostrowski [GA**196**], one of the great mathematicians of the 20th century. This paper is an extended version of an earlier paper [GA171] (not reproduced in these volumes) written in Italian. This account of Ostrowski's life and work, carefully written by one of his last students, is highly interesting and needs no further comment.

25

Miscellanea

Martin J. Gander

Here, five "miscellaneous" papers of Walter Gautschi are commented on, [GA**96**, GA**124**, GA**125**, GA**175**, GA**197**], preceded by some personal reminiscences.

I encountered Walter Gautschi's work several years before I encountered him in person. I was a PhD student at Stanford and taking a course given by Gene Golub on orthogonal polynomials and quadrature. Several faculty members were also taking this course, among them Andrew Stuart, who became my PhD supervisor, and Alan Karp. During the lectures, Alan Karp posed an interesting problem of computing Gauss quadrature nodes and weights for difficult weight functions arising in radiative transfer. I immediately put to work what I had learned in class, and failed, since all the methods we had seen were becoming rapidly unstable, and it was not possible to compute the recurrence coefficients of the required orthogonal polynomials to sufficiently high accuracy. So I started to search the literature and came across a paper of Walter Gautschi, [GA**141**], which describes precisely the problems I was working on, and also proposes an ingenious discretization procedure, which allowed me to replace the unstable approaches I tried before by orthogonal transformations, which are naturally numerically stable. This procedure allowed us to compute very effectively high-order Gauss quadrature rules for all important weight functions in this application, and led to the short paper [3].

I met Walter Gautschi for the first time on Sunday, April 26, 1998, when he came for a seminar to the École Polytechnique in Paris, where I was doing my postdoc. We hit it off immediately, and when our twins were born in Montreal, this added a further common bond, since Walter Gautschi also had a twin brother, Werner Gautschi, a very talented mathematician as well, who unfortunately passed away too early in life. When I moved to Geneva for a full professorship, I invited Walter Gautschi to give a talk at our mathematics colloquium, and, happily, he agreed to come. He gave a very well-received talk about "The spiral of Theodorus, numerical analysis, and special functions". To my delight, I found this talk again in

C. Brezinski and A. Sameh (eds.), *Walter Gautschi, Volume 3: Selected Works with Commentaries*, Contemporary Mathematicians, DOI 10.1007/978-1-4614-7132-5_5,

one of the papers I was assigned to study more closely in this tremendous enterprise of commenting on the selected works of Walter Gautschi. I will do this, however, in chronological order, so the Theodorus paper will come last.

25.1. The FG algorithm

This paper, [GA**96**], which is joint work with Bernard Flury from the University of Bern, appeared when I was still in high school! It is very atypical for the work of Walter Gautschi I am familiar with, dealing with a topic from numerical linear algebra. For a given set of symmetric positive definite matrices $A_1, A_2, \ldots, A_k$, the authors present an iterative algorithm to compute an orthogonal transformation B such that the matrices $B^T A_1 B, B^T A_2 B, \ldots, B^T A_k B$ are as close to diagonal as possible. In order to measure this "closeness", they introduce (and motivate) the function

$$\Phi(A_1, A_2, \ldots, A_k; n_1, \ldots, n_k) := \prod_{i=1}^{k} [\det(\mathrm{diag} A_i)]^{n_i} / [\det(A_i)]^{n_i},$$

where the n_i are given numbers. The best choice of B is one for which

$$\Phi(B^T A_1 B, B^T A_2 B, \ldots, B^T A_k T; n_1, \ldots, n_k) \longrightarrow \min .$$

In order to compute an approximate minimizer, the authors introduce the FG(Flury–Gautschi) algorithm, which consists of an outer iteration F and an inner iteration G. The algorithm is described in pseudocode, and the authors prove convergence of the algorithm. In the case $k = 1$, their algorithm reduces to the Jacobi method. In addition to the convergence of the two procedures, the authors also analyze under which conditions the solution is unique, and they give several hints for improving the algorithm.

Unfortunately, there was no implementation of the algorithm given in the paper[1]. Because of my interest in the algorithm, and since several details of the implementations were only addressed by comments, I decided to implement the algorithm myself in Matlab (see `http:/www.unige.ch/~gander/FG.php`)[2]. The algorithm was tested on the same example as given in the paper. It took quite a while to obtain the same results, because the implementation of the stopping criterion, based, as it was, on a comparison of eigenvectors becoming close, is tricky since normalized eigenvectors are only unique up to a sign and also come numerically in an arbitrary order. The current implementation now faithfully reproduces the authors' Fortran results. Their implementation on a CDC 170/855, in 1986, took 0.07 seconds of CPU time for this example to be executed. In Matlab on my Thinkpad T60, in

[1] With the help of Walter, we later found the Fortran implementation in [2].

[2] Many thanks to Hui Zhang, who also implemented the algorithm independently, so we could compare.

2012, the same example takes 0.03 seconds of CPU time. One wonders where all the computing power has gone these days[3].

Another test, which illustrates why the identity matrix as an initial guess of B can fail in the F-algorithm, is to simultaneously diagonalize a stiffness and a mass matrix (where this is actually possible)[4]. Specifically, the matrices

$$A_1 = \begin{bmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 4 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 4 \end{bmatrix}$$

give rise to an infinite loop when the initial guess of B in the F-algorithm is the identity matrix, and one needs to use an alternative random initial guess.

I could imagine that such an algorithm would find many users if it were generally available in Matlab, since the simultaneous diagonalization of matrices is an important task.

25.2. Slowly convergent series

The relevant paper on this topic, [GA**124**], as well as the paper [GA**125**] in the next subsection, are more in the core area —numerical quadrature — of Walter Gautschi's research interests. The problem is to sum the series

$$S_0 = \sum_{k=1}^{\infty} k^{\nu-1} r(k), \quad S_1 = \sum_{k=1}^{\infty} (-1)^{k-1} k^{\nu-1} r(k),$$

where $r(k)$ is a rational function. By using a preliminary partial fraction decomposition, Walter shows that it suffices to consider r of the form

$$r(s) = \frac{1}{(s+a)^m}, \qquad \Re a \geq 0, \quad m \geq 1.$$

Such series can be transformed into integrals by writing the fraction as a Laplace transform and then changing the order of summation and integration. The result is a weighted integral of an entire function; it then remains to determine Gauss quadrature rules for the respective weight function. With the hand of the master, Walter determines the three-term recurrence coefficients for the required orthogonal polynomials, which, as I experienced myself, are not always easy to compute to high precision. From these, one can easily obtain the required Gauss quadrature rules. He then illustrates the resulting fast summation procedure in the case of five infinite series, of which the first was communicated to Walter by Professor P. J. Davis who came upon it in his study of spirals, a topic we will again encounter in the fifth paper.

[3] Compilation would make this certainly much faster.

[4] Many thanks to Ivan Graham for suggesting this useful example during a conference in Urümqi in August 2012.

25.3. Slowly convergent series occurring in plate contact problems

This paper is a continuation of the previous paper, and it appeared in the same journal, right after the previous one. The subject is again the fast summation of infinite series, this time of the form

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

where z is complex with $|z| \leq 1$ and $p = 2, 3$, and also of the more difficult forms

$$\sum_{k=0}^{\infty}(2k+1)^{-p}\,\frac{\cosh((2k+1)x)}{\cosh((2k+1)b)}\,, \quad \sum_{k=0}^{\infty}(2k+1)^{-p}\,\frac{\sinh((2k+1)x)}{\sinh((2k+1)b)}\,,$$

where $0 \leq x \leq b$. Such series occur in the mathematical treatment of unilateral plate contact problems. After treating some special cases, Walter again uses the device of introducing a Laplace transform, but now only for part of the general term of the series. Interchanging summation and integration, as in the earlier paper, leads to a weighted integral with a weight function similar to the one in the previous paper. There are, however, cases for the parameters where Gauss quadrature is no longer effective, and Walter shows how a further transformation leads to an integral which can be effectively evaluated using a backward recursion scheme. Faithful to his working style, he gives the needed recurrence coefficients to high accuracy, and then shows two fully worked out examples to illustrate the technique.

25.4. The Hardy–Littlewood function

In the short 6-page note [GA**175**], Walter Gautschi gives a summary of his conference presentation at the birthday conference for Olav Njåstad. The topic was the summation of the series

$$H(x) = \sum_{k=1}^{\infty}\sin(x/k)/k, \tag{25.1}$$

which is important in the study of the polygamma functions. Walter first shows how the summation can be performed using orthogonal polynomials and polynomial/rational Gauss quadrature (cf. Section 15.4, Vol. 2), again applying the Laplace transform device. In a first approach, he obtains a formulation in terms of modified Bessel functions of order zero, the power series expansion of which, however, is only suitable for relatively small positive values of x, because otherwise severe cancellation errors make the approach numerically useless. As an alternative, Walter rewrites the expression obtained by using an integral representation of the Bessel function, in which case the trapezoidal rule can be used effectively and without cancellation. He

then also uses rational Gauss–Laguerre quadrature directly in the original formulation, and with this approach the range of x-values can be substantially enlarged before cancellation problems set in. Walter finally shows a completely different approach, based on direct summation of the first $n \approx x$ terms combined with an acceleration procedure, which is very effective for large values of x.

As it turned out, this short paper became the major inspiration for a recent publication by Kuznetsov [5] on asymptotic approximations to the Hardy–Littlewood function. Kuznetsov's goal was to find a value of x for which $H(x)$ in (25.1) satisfies $H(x) < -\pi/2$, in order to provide an explicit counterexample to a conjecture of Clark and Ismail. (The value of x found was extremely large, of the order 10^{21}!) Kuznetsov in his paper says "This turns out to be a surprisingly hard problem", and then goes on to use and extend the techniques introduced by Walter in order to solve it.

25.5. The spiral of Theodorus

On May 22, 2003, Walter Gautschi visited us in the Section of Mathematics at the University of Geneva, and gave a colloquium lecture precisely on the topic of the paper [GA**197**]. It was a fascinating lecture, I remember it very well. Like the paper, it started with an intriguing spiral, the spiral of Theodorus, shown in Figure 25.1. As one can see, the spiral is constructed starting at the point $(1, 0)$ by always moving in the direction orthogonal to the current position vector, and going precisely a distance of length 1. This gives for the second point $(1, 1)$, with a distance $\sqrt{2}$ from the origin (just use Pythagoras), for the third point a location with distance $\sqrt{2+1}$ from the origin (use Pythagoras again), for the fourth point a distance $\sqrt{3+1}$, the

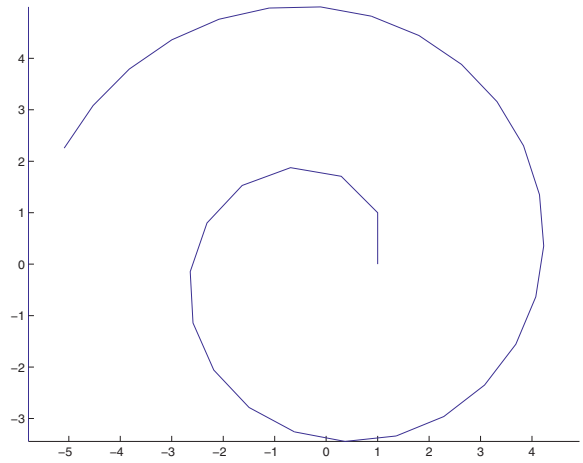


Fig. 25.1. The spiral of Theodorus

general point numbered n having a distance $\sqrt{n}$ from the origin. The distribution of the angles in the spiral of Theodorus has interesting number-theoretic properties (see [4], where the spiral is given the name "Quadratwurzelschnecke"[5]).

Using complex variables, one can also describe this spiral for $\alpha = 1, 2, \ldots$ by the recurrence relation

$$T(\alpha + 1) = \left(1 + \frac{i}{\sqrt{\alpha}}\right) T(\alpha), \quad T(1) = 1, \tag{25.2}$$

which gives $T(2) = 1 + i$, $T(3) = (1 + \frac{i}{\sqrt{2}})(1 + i) = 1 - \frac{1}{\sqrt{2}} + i(1 + \frac{1}{\sqrt{2}})$, etc. The spiral of Theodorus is thus obtained by applying a Forward Euler Method (with step 1) to the differential equation

$$T'(\alpha) = \frac{i}{\sqrt{\alpha}} T(\alpha), \tag{25.3}$$

which has as a solution the circle, the dynamics of which, however, slows down more and more as one moves along the circle.

The problem treated by Walter Gautschi, however, is a different one. Professor Davis [1, p. 33ff] had been wondering if it is possible to interpolate the spiral of Theodorus by a smooth, if possible analytic, curve. This problem is similar to a problem Euler faced when he tried to interpolate the factorial function, which led to his discovery of the gamma function. Davis, inspired by Euler's work, found the following interpolant:

$$T(\alpha) = \prod_{k=1}^{\infty} \frac{1 + i/\sqrt{k}}{1 + i/\sqrt{k + \alpha - 1}}, \quad \alpha \geq 0.$$

This product also satisfies the recurrence relation (25.2), and can be evaluated for any value $\alpha \geq 0$. It therefore produces a continuous (in fact, analytic) version of the Theodorus spiral.

Unfortunately, the product is very slowly convergent, and thus not suitable for numerical evaluation. This is where Walter Gautschi comes in: using logarithmic differentiation, he derives a polar representation for the continuous spiral of Theodorus, in which there now appears a slowly convergent series. For a particular point on the spiral (where it crosses the positive real axis for the first time), the series is given by

$$\sum_{k=1}^{\infty} \frac{1}{k^{3/2} + k^{1/2}},$$

the so-called Theodorus constant, and it is with this series that Davis had aroused Walter's interest in this problem. Using again Laplace transforms (cf. Section 25.2), Walter shows how the summation of the series can be transformed into a problem

[5] square-root snail

of integration, which can be solved very effectively by Gaussian quadrature — "an absolute gem of numerical analysis" according to Davis [1, p. 42].

With regard to identifying $T(\alpha)$ in terms of known special functions, however, Davis writes [1, pp. 41/42]: "Computation is one thing, and the identification of $T(\alpha)$ is another matter, and it still eluded me. The Spirit of Euler infused me constantly, but contributed nothing toward the solution. The mistake I made was that I had been consulting the wrong Swiss mathematician. I should have consulted the Swiss-born-and-trained American mathematician, Walter Gautschi, who ... in the course of this work ... also identified $T(\alpha)$."

The analytic Theodorus spiral can also be continued backward into a second sheet of the Riemann surface, as was proposed by J. Waldvogel [6], and Walter concludes with a figure of what he calls the twin-spiral of Theodorus, a very well-chosen name, given the context, and one which I will later also explain to my children.

One could ask what the differential equation might be that describes this twin spiral. It is certainly not equation (25.3), since this one only gives a circle. Something to think about!

25.6. Epilogue

My most recent meeting with Walter Gautschi was at the conference in honor of Claude Brezinski's 70th birthday in Sardinia, in the fall of 2011. As always, we had very nice discussions, Walter and I, and Walter gave a lovely presentation about a real problem from applications [GA**204**], solved in a very elegant way, how could it be different, using Gauss quadrature. I hope we will meet many more times in the future.

References

[1] Philip J. Davis. *Spirals: from Theodorus to chaos*. With contributions by Walter Gautschi and Arieh Iserles. A K Peters, Wellesley, MA, 1993. x+237 pp. ISBN: 1-56881-010-5.

[2] Bernard N. Flury and Gregory Constantine. The F-G diagonalization algorithm. Algorithm AS 211. *Applied Statistics* 34:177–183, 1985

[3] Martin J. Gander and Alan H. Karp. Stable computation of high order Gauss quadrature rules using discretization for measures in radiation transfer. *J. Quantitative Spectroscopy Radiative Transfer*, 68(2):213–223, 2001.

[4] Edmund Hlawka. Gleichverteilung und Quadratwurzelschnecke. *Monatsh. Math.*, 89(1):19–44, 1980. (Excerpted English translation in [1, pp. 157-167].)

[5] A. Kuznetsov. Asymptotic approximations to the Hardy–Littlewood function. *J. Comput. Appl. Math.*, 237(1):603–613, 2013.

[6] Jörg Waldvogel. Analytic continuation of the Theodorus spiral. Preliminary version at *http://www.sam.math.ethz.ch/~waldvoge/Papers/theopaper.html*

Part II

Reprints

26

Papers on Linear Recurrence Relations

---

26 Computation of successive derivatives of $f(z)/z$, Math. Comp. 20, 209–214 (1966)

29 Computational aspects of three-term recurrence relations, SIAM Rev. 9, 24–82 (1967)

35 An application of minimal solutions of three-term recurrences to Coulomb wave functions, Aequationes Math. 2, 171–176 (1969)

37 (with B. J. Klein) Recursive computation of certain derivatives — a study of error propagation, Comm. ACM 13, 7–9 (1970)

135 Is the recurrence relation for orthogonal polynomials always stable?, BIT 33, 277–284 (1993)

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 (1997)

26.1. [26] "Computation of Successive Derivatives of *f(z)/z*"

---

[26] "Computation of Successive Derivatives of *f(z)/z*," *Math. Comp*. **20**, 209–214 (1966).

Computation of Successive Derivatives of $f(z)/z$*

By Walter Gautschi†

1. Introduction. It is sometimes necessary to calculate derivatives of the form

$$d_n(z) = \frac{d^n}{dz^n}\left(\frac{f(z)}{z}\right) \qquad (n = 0, 1, 2, \cdots), \tag{1.1}$$

where f is a function whose derivatives can be formed readily. Analytic differentiation in (1.1), while elementary, is obviously tedious, and the resulting expressions are of doubtful practical value. In the following we present a simple and effective recursive algorithm to generate these derivatives. As an example, we consider the cases where $f(z) = e^z$, $f(z) = \cos z$, and $f(z) = \sin z$.

Our main observation may be paraphrased in the following surprising way. The calculation of a large number of derivatives (1.1) at a fixed point z is a stable process if the function $g(\zeta) = f(\zeta)/\zeta$ has a pole at $\zeta = 0$, and an unstable process if $g(\zeta)$ is regular at $\zeta = 0$.

2. The Recurrence Relation. Let $z \neq 0$ be arbitrary complex, and let $f(\zeta)$ be analytic in the circle $|\zeta - z| \leqq r$, $r > |z|$, which includes the origin $\zeta = 0$. Our point of departure is the identity

$$\frac{f(z) - f(0)}{z} = \int_0^1 f'(tz)\, dt.$$

Differentiating n times gives

$$d_n(z) - (-1)^n \frac{n!}{z^{n+1}} f(0) = \int_0^1 t^n f^{(n+1)}(tz)\, dt. \tag{2.1}$$

Denoting the integral on the right by I_n, integration by parts yields

$$I_n + \frac{n}{z} I_{n-1} = \frac{f^{(n)}(z)}{z},$$

hence, together with (2.1), the recurrence relation

$$d_n(z) + \frac{n}{z} d_{n-1}(z) = \frac{f^{(n)}(z)}{z} \qquad (n = 1, 2, 3, \cdots). \tag{2.2}$$

We note that (2.2) represents a linear inhomogeneous first-order difference equation for d_n. Computational aspects of such difference equations were discussed at length in [1]. It was noted there, that a naive application of (2.2) in the forward direction is accompanied by an undesirable build-up of rounding errors whenever the quantity

$$\rho_n = \frac{d_0 h_n}{d_n}$$

Received September 10, 1965.
* Work performed under the auspices of the U. S. Atomic Energy Commission.
† Present address: Purdue University.

becomes large in absolute value for some n. Here, h_n denotes the solution (normalized by $h_0 = 1$) of the homogeneous difference equation that corresponds to (2.2), i.e.

$$h_n = (-1)^n \frac{n!}{z^n}.$$

Numerical instability is particularly prominent if $\lim_{n\to\infty} |\rho_n| = \infty$, or, equivalently, if

$$\lim_{n\to\infty} \frac{d_n}{h_n} = 0. \tag{2.3}$$

By (2.1) we have

$$z\frac{d_n}{h_n} = f(0) + (-1)^n \frac{z^{n+1}}{n!}\int_0^1 t^n f^{(n+1)}(tz)\,dt. \tag{2.4}$$

The second term on the right, disregarding the sign, we recognize as being the nth remainder (in integral form) of the Taylor expansion of $f(0)$ about z. Because of the analyticity assumption made at the beginning of this section, this remainder tends to zero, as $n \to \infty$, and so

$$\lim_{n\to\infty} \frac{d_n}{h_n} = \frac{f(0)}{z}. \tag{2.5}$$

In particular, if $f(0) = 0$, then (2.3) holds, and we have numerical instability. On the other hand, if $f(0) \neq 0$, then

$$\lim_{n\to\infty} \rho_n = \frac{f(z)}{f(0)},$$

and $|\rho_n|$ is bounded for all n, provided $d_n(z)$ does not vanish for some n. Hence, no serious numerical difficulties should attend the use of (2.2), unless $|f(z)/f(0)|$ is very large, or $|\rho_n|$ reaches a large peak prior to converging to the limiting value $|f(z)/f(0)|$.

An alternate proof of (2.5) can be given using Cauchy's formula for the nth derivative of an analytic function,

$$d_n(z) = \frac{n!}{2\pi i}\oint_C \frac{f(\zeta)\,d\zeta}{(\zeta - z)^{n+1}\zeta}.$$

If $f(0) = 0$, we may take for C a circle about z containing the origin and contained in the circle of analyticity of f. If $f(0) \neq 0$, we must add to C a small contour C_0 encircling the origin in the negative direction. Taking for C_0 a small circle, and letting its radius tend to zero, we arrive at

$$d_n(z) = (-1)^n \frac{n!}{z^{n+1}} f(0) + \frac{n!}{2\pi i}\oint_C \frac{f(\zeta)\,d\zeta}{(\zeta - z)^{n+1}\zeta}.$$

Hence,

$$z\frac{d_n}{h_n} = f(0) + \frac{(-1)^n}{2\pi i}\oint_C \left(\frac{z}{\zeta - z}\right)^{n+1} \frac{f(\zeta)}{\zeta}\,d\zeta. \tag{2.6}$$

Since $f(\zeta)/\zeta$ is bounded on C, and

$$\left|\frac{z}{\zeta - z}\right| \leqq q < 1,$$

it is clear that the integral in (2.6) tends to zero, as $n \to \infty$, and so we again obtain (2.5).

We may summarize as follows: *Let f be analytic in a circle about z which includes the origin in its interior. Then the generation of a large number of derivatives* (1.1), *using forward recursion by* (2.2), *is in general numerically stable if $f(0) \neq 0$, but highly unstable if $f(0) = 0$.*

We observe, however, that forward recursion by (2.2), even in the case $f(0) = 0$, may still be adequate, if only a relatively small number of derivatives are required. In fact, the recursion should be adequate as long as $n \leqq |z|$.

3. Recursive Algorithm in the Case $f(0) = 0$. We take advantage of a remark made on p. 25 of [1]. Since $|\rho_n| \to \infty$, we may apply the recursion (2.2) in the backward direction, starting with $n = \nu$ sufficiently large, and using zero initial value,

$$(3.1)\quad d_{n-1}^{[\nu]} = (f^{(n)}(z) - z d_n^{[\nu]})/n \qquad (n = \nu, \nu - 1, \cdots, 1), \qquad d_\nu^{[\nu]} = 0.$$

Then, for $n \geqq 0$ in any bounded set, we will have

$$d_n^{[\nu]} \to d_n \quad \text{as} \quad \nu \to \infty.$$

Moreover, the relative error of $d_n^{[\nu]}$ is given by

$$(3.2)\qquad \frac{d_n^{[\nu]} - d_n}{d_n} = \frac{\rho_n}{\rho_\nu}.$$

It remains to estimate a reasonable starting value ν for n, given, say, that the results for $n = 0, 1, 2, \cdots, N$ are to be accurate to S significant digits. According to (3.2), we must require that $|\rho_n/\rho_\nu| \leqq \epsilon$ for all $0 \leqq n \leqq N$, where

$$\epsilon = \tfrac{1}{2}\, 10^{-S},$$

that is,

$$(3.3)\qquad \frac{n!}{\nu!}\,|z|^{\nu-n}\left|\frac{d_\nu}{d_n}\right| \leqq \epsilon \qquad (n = 0, 1, 2, \cdots, N).$$

In addition to the analyticity assumption introduced earlier, we now *assume that $f^{(n)}$ is uniformly bounded, and bounded away from zero on the segment from 0 to z as $n \to \infty$*. Then it is clear from (2.1), where now $f(0) = 0$, that $|d_\nu/d_n| < 1$ for ν sufficiently large. Hence, it appears reasonable to replace $|d_\nu/d_n|$ in (3.3) by 1, and to require

$$(3.4)\qquad \frac{n!}{\nu!}\,|z|^{\nu-n} \leqq \epsilon \qquad (n = 0, 1, 2, \cdots, N).$$

Denote the expression on the left by p_n. Clearly, $\{p_n\}$ is a sequence of positive numbers which initially decrease, until n is near $|z|$, and from then on increase rapidly to ∞. (The case $|z| < 1$, in which p_n increases from the beginning, is of

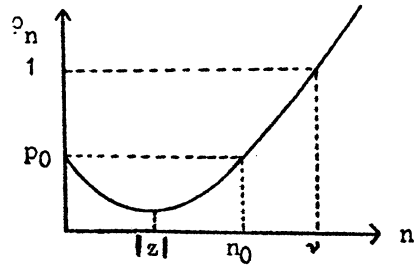


FIGURE 1. Behavior of $p_n = n! \mid z \mid^{\nu-n}/\nu!$

little consequence for the following.) Denote by n_0 the integer $n > 0$ for which p_n is near to p_0 "for the second time" (see Figure 1), hence $\mid z \mid^n/n!$ near 1. Then, (3.4) is implied by $p_0 \leqq \epsilon$, if $N \leqq n_0$, and by $p_N \leqq \epsilon$ if $N > n_0$. We may replace (3.4) therefore by

$$\frac{\mid z \mid^\nu}{\nu!} \leqq \epsilon \qquad (N \leqq n_0), \qquad \frac{N!}{\nu!} \mid z \mid^{\nu-N} \leqq \epsilon \qquad (N > n_0).$$

Using Stirling's formula, these conditions are adequately approximated by

$$\left(\frac{e \mid z \mid}{\nu}\right)^\nu \leqq \epsilon \qquad (N \leqq n_0), \qquad \left(\frac{e \mid z \mid}{\nu}\right)^\nu \left(\frac{N}{e \mid z \mid}\right)^N \leqq \epsilon \qquad (N > n_0).$$

We note, incidentally, that again by Stirling's formula,

$$n_0 \approx [e \mid z \mid], \qquad e = 2.71828 \cdots .$$

The first inequality, upon taking logarithms, can be written in the form

$$\frac{\nu}{e \mid z \mid} \ln \left(\frac{\nu}{e \mid z \mid}\right) \geqq \frac{s}{e \mid z \mid}, \tag{3.5}$$

where

$$s = S \ln 10 + \ln 2.$$

Similarly, the second inequality amounts to

$$\nu \ln \left(\frac{\nu}{e \mid z \mid}\right) - N \ln \left(\frac{N}{e \mid z \mid}\right) \geqq s,$$

which can be written in the form

$$\left(\frac{\nu}{N} - 1\right) \ln \left(\frac{N}{e \mid z \mid}\right) + \frac{\nu}{N} \ln \frac{\nu}{N} \geqq \frac{s}{N}. \tag{3.6}$$

Since certainly $\nu > N$, and moreover $N \geqq e \mid z \mid$ (N now being larger than n_0, and $n_0 \approx e \mid z \mid$), the first term on the left is $\geqq 0$. Hence, (3.6) will be satisfied if we require

$$\frac{\nu}{N} \ln \frac{\nu}{N} \geqq \frac{s}{N}. \tag{3.7}$$

Both conditions (3.5), (3.7) now have the form $t \ln t \geqq c$. Denoting by $t(y)$ the inverse function of $y = t \ln t$ $(t \geqq 1)$, we obtain our final estimate of ν in the form

$$(3.8) \quad \nu \geqq e\,|z|\,t\left(\frac{s}{e\,|z|}\right) \qquad (N \leqq n_0), \qquad \nu \geqq Nt\left(\frac{s}{N}\right) \qquad (N > n_0).$$

We note that in (3.8) the function $t(y)$ need only be available to low accuracy. Formulas giving 1% accuracy, or better, may be found in [2].

The algorithm just described may still be unsatisfactory, numerically, if $|z|$ is relatively large. The recursion (3.1) then is likely to suffer from loss of accuracy, due to cancellation of digits, particularly for n near 1. For such n, indeed, z/n in (3.1) will have large absolute value, yet $d_{n-1}^{[\nu]}$ has normally the same order of magnitude as $d_n^{[\nu]}$. The difficulty may be resolved by *applying* (2.2) *in forward direction as long as* $n \leqq |z|$, *and using the backward recurrence algorithm described above for the remaining* n *with* $|z| < n \leqq N$.

4. Examples. Consider first $f(z) = e^z$, and let

$$d_n(z) = \frac{d^n}{dz^n}\left(\frac{e^z}{z}\right).$$

Then (2.2) gives immediately

$$(4.1) \qquad d_n(z) + \frac{n}{z}\,d_{n-1}(z) = \frac{e^z}{z} \qquad (n = 1, 2, 3, \cdots).$$

Our theory of Sections 2 and 3 clearly applies. Since $f(0) = 1$, it follows that (4.1) is numerically stable in the forward direction. We note, incidentally, that

$$(4.2) \qquad d_n(z) = (-1)^n \frac{n!}{z^{n+1}}\,e^z e_n(-z),$$

where

$$(4.3) \qquad e_n(z) = \sum_{k=0}^{n} \frac{z^k}{k!}$$

is the nth partial sum of the exponential series.

Likewise, if $f(z) = \cos z$, and

$$c_n(z) = \frac{d^n}{dz^n}\left(\frac{\cos z}{z}\right),$$

we obtain

$$(4.4) \qquad c_n(z) + \frac{n}{z}\,c_{n-1}(z) = \tau_n(z) \qquad (n = 1, 2, 3, \cdots),$$

where $\{\tau_n(z)\}_{n=1}^{\infty} = \{-\sin z, -\cos z, \sin z, \cos z, \cdots\}$. Like the previous recursion, (4.4) is numerically stable. On the other hand, if $f(z) = \sin z$, and

$$s_n(z) = \frac{d^n}{dz^n}\left(\frac{\sin z}{z}\right),$$

then

$$s_n(z) + \frac{n}{z} s_{n-1}(z) = \sigma_n(z) \qquad (n = 1, 2, 3, \cdots), \tag{4.5}$$

$\{\sigma_n(z)\}_{n=1}^{\infty} = \{\cos z, -\sin z, -\cos z, \sin z, \cdots\}$, is numerically unstable, and the algorithm of Section 3 should be applied, including the device mentioned at the end of Section 3.

In terms of (4.3), we may also write

$$c_n(z) = \frac{(-1)^n n!}{2z^{n+1}} [e^{iz} e_n(-iz) + e^{-iz} e_n(iz)],$$

$$s_n(z) = \frac{(-1)^n n!}{2iz^{n+1}} [e^{iz} e_n(-iz) - e^{-iz} e_n(iz)],$$

as follows readily from (4.2) and Euler's formula.

The functions $s_n(x)$ have found wide applications in diffraction theory, and are extensively tabulated (see [4]). The generation of d_n, c_n, and s_n, may also be useful for the analytic continuation of the exponential-, cosine-, and sine-integrals, respectively. ALGOL procedures generating $d_n(x)$, $c_n(x)$, and $s_n(x)$ for real x may be found in [3].

Argonne National Laboratory
Argonne, Illinois

1. W. GAUTSCHI, "Recursive computation of certain integrals," *J. Assoc. Comput. Mach.*, v. 8, 1961, pp. 21–40.
2. W. GAUTSCHI, "Algorithm 236—Bessel functions of the first kind," *Comm. Assoc. Comput. Mach.*, v. 7, 1964, pp. 479–480.
3. W. GAUTSCHI, "Algorithm 282—derivatives of e^x/x, cos $(x)/x$, and sin $(x)/x$," *Comm. Assoc. Comput. Mach.* (v. 9, 1966.)
4. *Tables of the Function* (sin ϕ)/ϕ *and of its First Eleven Derivatives*, The Annals of the Computation Laboratory of Harvard University, Harvard Univ. Press, Cambridge, Mass., 1949. MR 11, 692.

26.2. [29] "COMPUTATIONAL ASPECTS OF THREE-TERM RECURRENCE RELATIONS"

---

[29] "Computational Aspects of Three-Term Recurrence Relations," *SIAM Rev*. **9**, 24–82 (1967).

SIAM REVIEW
Vol. 9, No. 1, January, 1967

COMPUTATIONAL ASPECTS OF THREE-TERM RECURRENCE RELATIONS*

WALTER GAUTSCHI†

Introduction. Recurrence relations are one of the basic mathematical tools of computation. There is hardly a computational task which does not rely on recursive techniques, at one time or another. The widespread use of recurrence relations can be ascribed to their intrinsic constructive quality, and the great ease with which they are amenable to mechanization. On the other hand, like most recursive processes, recurrence relations are susceptible to error growth. Each cycle of a recursive process not only generates its own rounding errors, but also inherits the rounding errors committed in all the previous cycles. If conditions are unfavorable, the resulting propagation of error may well be disastrous. It is this aspect of recurrence relations—the possibility and the prevention of numerical instability—that will be of concern to us.

The problem of numerical instability has been studied extensively for difference equations arising in the numerical solution of ordinary and partial differential equations. In the seemingly much simpler context of a single linear difference equation, the problem has received only sporadic attention, even though such difference equations, particularly of the second order, occur prominently in many branches of pure and applied mathematics. We mention, e.g., the recurrence relations satisfied by large classes of special functions of mathematical physics and statistics, the three-term recurrence relations that lie at the heart of continued fraction theory and the theory of orthogonal polynomials, and the miscellaneous recurrence relations one encounters when constructing series expansions, asymptotic or otherwise, to solutions of linear differential equations. We believe, therefore, that a systematic review of some of the computational problems attending recurrence relations might be of value. In the following we attempt to present such a survey, restricting attention to the special case of three-term recurrence relations.

The kind of instability we are concerned with, may be described as follows. Consider a three-term recurrence relation of the form

$$y_{n+1} + a_n y_n + b_n y_{n-1} = 0, \qquad n = 1, 2, 3, \cdots, \tag{0.1}$$

where a_n , b_n are given sequences of real or complex numbers, and $b_n \neq 0$. The general solution of (0.1) can be spanned by any pair f_n , g_n of linearly independent solutions. We are interested in the special case where there exists such a pair having the property

$$\lim_{n\to\infty} \frac{f_n}{g_n} = 0. \tag{0.2}$$

* Received by the editors February 17, 1966, and in revised form July 18, 1966.

† Computer Sciences Department, Purdue University, Lafayette, Indiana, and Argonne National Laboratory, Argonne, Illinois. This work was performed in part under the auspices of the United States Atomic Energy Commission.

Serious problems then arise if one attempts to compute the solution f_n or any constant multiple of f_n.

To see this, we first note that (0.2) implies

$$\lim_{n\to\infty} \frac{f_n}{y_n} = 0 \tag{0.3}$$

for any solution y_n not proportional to f_n. Such a solution, indeed, is representable in the form $y_n = af_n + bg_n$, with $b \neq 0$, and therefore

$$\lim_{n\to\infty} \frac{f_n}{y_n} = \lim_{n\to\infty} \frac{f_n/g_n}{b + a(f_n/g_n)} = 0.$$

If we now generate f_n by (0.1), using only approximate initial values $y_0 \doteq f_0$, $y_1 \doteq f_1$ (due to rounding, for example), but recurring with infinite precision, we obtain a solution y_n which, in general, is linearly independent of f_n. Therefore, by (0.3), we will have

$$\left|\frac{y_n - f_n}{f_n}\right| \to \infty \quad \text{as} \quad n \to \infty,$$

i.e., the relative error of y_n, the intended approximation to f_n, becomes arbitrarily large. Therefore, this straightforward method of computing f_n is utterly ineffective.

Observe that the set of all solutions f_n of (0.1) having the property indicated in (0.3) forms a one-dimensional subspace of the space of all solutions. (There can be no two linearly independent solutions f_n, f_n' enjoying this property, since, otherwise, f_n/f_n' and f_n'/f_n would both have the limit zero, as $n \to \infty$, which is absurd.) We call the solutions of this subspace *minimal at infinity*, or briefly *minimal*.[1] A nonminimal solution will be referred to as *dominant*. Each dominant solution is asymptotically proportional to g_n. Note that, in contrast to dominant solutions, a minimal solution is uniquely determined by one initial value.

To illustrate the difficulty of calculating minimal solutions, consider the problem of generating Bessel functions of the first kind, $J_n(x)$, for fixed x, and $n = 0, 1, 2, \cdots$.[2] As is well-known, these functions (of n) obey the three-term recurrence relation

$$y_{n+1} - \frac{2n}{x} y_n + y_{n-1} = 0. \tag{0.4}$$

From tables of Bessel functions we find, e.g., that for $x = 1, J_0(1) = .7651976866$, $J_1(1) = .4400505857$, accurate to ten figures. Generating the next 99 values of $J_n(1)$ on a digital computer by straightforward recursion, we obtain the results

[1] The notion of a minimal solution appears to have first been introduced by Pincherle in connection with his generalization of continued fractions [44]. Pincherle called it "distinguished" solution (soluzione distinta). In the theory of linear differential equations the term "principal" solution is also in use [24]. The minimal solution can often be identified with a solution "of type II" in a terminology of Schäfke [51].

[2] This example is well known, and has received considerable attention in the literature. See the references in §5.

TABLE 1

n	"$J_n(1)$"	n	"$J_n(1)$"
0	7.651976866 (−1)	9	−4.645246881 (−4)
1	4.400505857 (−1)	10	−8.332374506 (−3)
2	1.149034848 (−1)	11	−1.661829654 (−1)
3	1.956335358 (−2)	12	−3.647692865 (0)
4	2.476636684 (−3)	13	−8.737844579 (1)
5	2.497398891 (−4)	...	...
6	2.076220699 (−5)	20	−2.818590869 (12)
7	−5.934052751 (−7)	...	...
8	−2.906988084 (−5)	100	−2.586550446 (175)

shown in Table 1.[3] (The numbers in parentheses denote powers of 10 by which the preceding numbers have to be multiplied.) Obviously, there is little resemblance with the true values of $J_n(1)$, which are known to decrease steadily with increasing n, and to approach zero very rapidly as $n \to \infty$. In fact, since $J_7(1)$ came out to be negative, all digits shown, for $n \geqq 7$, including the sign and the exponent, must be illusory.

The disastrous build-up of errors, in this example, is due to the fact that with $f_n = J_n(x)$, also $g_n = Y_n(x)$, the Bessel function of the second kind, is a solution of (0.4) and, moreover,

$$\frac{f_n}{g_n} \sim -\frac{(x/2)^{2n}}{2(n!)^2} \quad \text{as} \quad n \to \infty.$$

Therefore, $J_n(x)$ is indeed highly minimal at infinity.

Methods of calculating minimal solutions of three-term recurrence relations, including applications, constitute the main theme of this paper. In §1 we begin with a brief survey of continued fractions, emphasizing computational methods. The relevance of continued fractions is contained in a result due to Pincherle which expresses ratios of a minimal solution in terms of continued fractions. In §2 we recall some classical results from the asymptotic theory of linear difference equations which will find repeated use in the later parts of the paper. §3 brings a first algorithm for calculating a minimal solution, based on the result of Pincherle. The problem considered is to calculate a minimal solution f_n known to satisfy

$$\sum_{m=0}^{\infty} \lambda_m f_m = s, \qquad s \neq 0. \tag{0.5}$$

The special case $\lambda_0 = 1$, $\lambda_m = 0$, $m > 0$, amounts to prescribing f_0. Consideration of an infinite series (0.5) has the distinct advantage that the resulting algorithm does not require the computation of f_n for any value of n. Our first algorithm is mathematically (though not computationally) equivalent to the

[3] Computation was performed on the CDC 3600 computer, which in floating point arithmetic allows precision of about 12 decimal digits.

backward recurrence algorithm of J. C. P. Miller. While this algorithm is widely regarded as just a "trick of the trade," our presentation will show that it derives naturally from rather elegant results of classical analysis. In §4 two alternate algorithms are described which are more flexible, but more elaborate, than the first algorithm. The remaining paragraphs discuss a number of applications, mostly to the computation of higher transcendental functions such as Bessel functions (§5), associated Legendre functions (§6), regular Coulomb wave functions (§7), and other miscellaneous functions (§§8–10). §11 contains an application to a Sturm-Liouville boundary value problem on an infinite interval.

The extent to which our algorithms are affected by rounding errors will not be discussed in detail. Our experience seems to indicate, however, that none of the algorithms is sensitive to rounding, unless the series (0.5) is subject to cancellation of terms. The rigorous analysis of error propagation is an interesting, though difficult, outstanding problem in this area. A significant contribution in this direction is due to Olver, who recently analyzed the error accumulation in Miller's algorithm [38].

In principle, there are other stable procedures that could be used to calculate minimal solutions of (0.1). For example, we could set up the boundary value problem of finding the solution y_n of (0.1) which satisfies

$$y_0 = f_0, \qquad y_N = f_N$$

for some sufficiently large N.[4] Clearly, this amounts to solving the linear system of equations

$$\begin{pmatrix} a_1 & 1 & & & & & \\ b_2 & a_2 & 1 & & & 0 & \\ & b_3 & a_3 & 1 & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot & \\ & 0 & & & \cdot & a_{N-2} & 1 \\ & & & & & b_{N-1} & a_{N-1} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \cdot \\ \cdot \\ \cdot \\ y_{N-1} \end{pmatrix} = \begin{pmatrix} -b_1 f_0 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ -f_N \end{pmatrix}, \tag{0.6}$$

whose matrix is tridiagonal. Any of the standard methods, such as triangular decomposition methods, may be used to solve (0.6). Unfortunately, the procedure requires two values, f_0 and f_N, of the desired solution to be known in advance. Either one may be difficult, or time-consuming, to obtain.

The problem of computing minimal solutions is clearly not peculiar to three-term recurrence relations. It may equally arise in connection with other functional equations, such as linear homogeneous difference and differential equations of arbitrary order, and systems of such equations. Whenever the space S of all solutions is the direct sum $S = S_1 \oplus S_2$ of two subspaces S_1 and S_2, and every solution $s_1 \in S_1$ dominates, in some appropriate sense, over every solution $s_2 \in S_2$, we may consider S_2 as the set of minimal solutions with respect to

[4] The author is indebted to Dr. M. E. Rose for pointing this out.

the decomposition $S = S_1 \oplus S_2$. (There may be several such decompositions.) The problem of computing minimal solutions in this sense has not been thoroughly studied, though the work of Clenshaw [7] and Schäfke [51] suggests that effective computational methods may exist also in this more general context.

1. Three-term recursion and continued fractions. It is well-known that the concepts of three-term recursion and continued fraction are closely related. To every continued fraction we may in fact associate a three-term recurrence relation, namely the fundamental recurrence formula for the numerators and denominators. Vice versa, every three-term recurrence relation may be interpreted as the fundamental recurrence formula for some continued fraction. The first point of view is useful for computing continued fractions, the second for computing the minimal solution. We begin by considering several methods of calculating a continued fraction.

Suppose we are given the continued fraction

$$\frac{a_1}{b_1+}\,\frac{a_2}{b_2+}\,\frac{a_3}{b_3+}\cdots, \tag{1.1}$$

where the partial numerators a_n and partial denominators b_n are real or complex numbers. Denote its nth numerator and nth denominator by A_n and B_n, respectively, so that

$$\frac{a_1}{b_1+}\,\frac{a_2}{b_2+}\cdots\frac{a_n}{b_n} = \frac{A_n}{B_n}. \tag{1.2}$$

The value of the continued fraction (1.1), if it exists, is defined as the limit $\lim_{n\to\infty} A_n/B_n$. The quantities A_n, B_n satisfy the fundamental recurrence formulas (see, e.g., [59, p. 15])

$$\begin{aligned} A_n &= b_nA_{n-1} + a_nA_{n-2}, \\ B_n &= b_nB_{n-1} + a_nB_{n-2}, \end{aligned} \qquad n = 1, 2, 3, \cdots, \tag{1.3}$$

where

$$A_{-1} = 1, \quad A_0 = 0; \quad B_{-1} = 0, \quad B_0 = 1. \tag{1.4}$$

This shows that $\alpha_n = A_{n-1}$ and $\beta_n = B_{n-1}$ constitute a pair of linearly independent solutions of the three-term recurrence relation

$$y_{n+1} - b_ny_n - a_ny_{n-1} = 0, \qquad n = 1, 2, 3, \cdots. \tag{1.5}$$

A first method of computation flows directly from these fundamental recurrence relations. Thus, one generates the A's and B's recursively, by means of (1.3) and (1.4), and concurrently the ratios A_n/B_n, until the latter converge within the required tolerance. As A_n and B_n are likely to grow rapidly with n, some care must be exercised if this method is used on a digital computer. Initial scaling, and possibly repeated subsequent scaling, may be necessary to avoid overflow.

A second method, which avoids the necessity of scaling, consists in evaluating

the finite continued fraction in (1.2) "from tail to head." Thus, formally, we set

$$f_k^{(n)} = \frac{a_k}{b_k+}\,\frac{a_{k+1}}{b_{k+1}+}\,\cdots\,\frac{a_n}{b_n}, \qquad 1 \leqq k \leqq n, \tag{1.6}$$

and generate these quantities recursively by

$$f_k^{(n)} = \frac{a_k}{b_k + f_{k+1}^{(n)}}, \qquad k = n, n-1, \cdots, 1, \tag{1.7}$$

using as initial value

$$f_{n+1}^{(n)} = 0. \tag{1.8}$$

Then, $f_1^{(n)} = A_n/B_n$. To obtain the value of the continued fraction, the backward recursion (1.7) will have to be carried out repeatedly, with increasing values of n, until successive values of $f_1^{(n)}$ agree within the accuracy desired. While certainly an inconvenience, the repetitive nature of this process nevertheless provides some self-checking features not possessed by the previous method.

A third method of computation, finally, exploits the connection between continued fractions and infinite series, expressed by the relation

$$\frac{A_n}{B_n} = \sum_{k=1}^{n} \rho_1\rho_2\cdots\rho_k,$$

where

$$1 + \rho_{k+1} = \frac{1}{1 + (a_{k+1}/b_k b_{k+1})(1+\rho_k)}, \qquad k = 2, 3, \cdots, n-1,$$

$$\rho_1 = a_1/b_1, \qquad 1 + \rho_2 = \frac{1}{1 + (a_2/b_1 b_2)}.$$

(This result may be obtained from Theorem 2.1 and formula (2.6) in [59], by an appropriate equivalence transformation. See also [56]; the formula defining ρ_k in this reference contains a typographical error.) Clearly, these relations can be modelled into a recursive algorithm to generate successive approximants of a continued fraction. Let, indeed,

$$u_1 = 1, \qquad u_k = 1 + \rho_k, \qquad k \geqq 2,$$

$$v_k = \rho_1\rho_2\cdots\rho_k, \qquad k \geqq 1,$$

$$w_k = \sum_{i=1}^{k} v_i, \qquad k \geqq 1,$$

so that $w_k = A_k/B_k$. Then

$$\begin{aligned} u_{k+1} &= \frac{1}{1 + \dfrac{a_{k+1}}{b_k b_{k+1}} u_k}, \\ v_{k+1} &= v_k(u_{k+1} - 1), \\ w_{k+1} &= w_k + v_{k+1}, \end{aligned} \qquad k = 1, 2, 3, \cdots, \tag{1.9}$$

the initial values being

$$u_1 = 1, \qquad v_1 = w_1 = \frac{a_1}{b_1}. \tag{1.10}$$

None of the disadvantages noted in the previous two methods are present here.

We have seen that the continued fraction (1.1) leads naturally to the three-term recursion (1.5). Suppose now, conversely, that we are given a three-term recurrence relation

$$y_{n+1} + a_n y_n + b_n y_{n-1} = 0, \qquad b_n \neq 0,\ n = 1, 2, 3, \cdots. \tag{1.11}$$

Define α_n, β_n to be the special solutions of (1.11) with initial values

$$\alpha_0 = 1, \quad \alpha_1 = 0; \qquad \beta_0 = 0, \quad \beta_1 = 1. \tag{1.12}$$

Then, evidently, $A_n = \alpha_{n+1}$ and $B_n = \beta_{n+1}$ are the numerators and denominators, respectively, of the continued fraction

$$\frac{-b_1}{-a_1-}\ \frac{b_2}{-a_2-}\ \frac{b_3}{-a_3-}\ \cdots, \tag{1.13}$$

which is equivalent to the continued fraction

$$\frac{b_1}{a_1-}\ \frac{b_2}{a_2-}\ \frac{b_3}{a_3-}\ \cdots. \tag{1.14}$$

We may formally arrive at this continued fraction also in the following way. Let us introduce the ratios

$$r_n = \frac{y_{n+1}}{y_n}, \qquad n = 0, 1, 2, \cdots.$$

Dividing (1.11) by y_n then gives

$$r_n + a_n + \frac{b_n}{r_{n-1}} = 0,$$

from which

$$r_{n-1} = \frac{-b_n}{a_n + r_n}.$$

Applying this formula repeatedly, with n successively increasing, we get

$$r_{n-1} = \frac{y_n}{y_{n-1}} = \frac{-b_n}{a_n-}\ \frac{b_{n+1}}{a_{n+1}-}\ \frac{b_{n+2}}{a_{n+2}-}\ \cdots. \tag{1.15}$$

In particular, when $n = 1$,

$$\frac{y_1}{y_0} = \frac{-b_1}{a_1-}\ \frac{b_2}{a_2-}\ \frac{b_3}{a_3-}\ \cdots.$$

This derivation indicates that the continued fraction (1.14), and similarly

the continued fractions in (1.15), are related to ratios of consecutive values for some solution y_n. The argument, however, neither insures us of the convergence of these continued fractions, nor does it tell us for what particular solution the ratios are to be formed. These matters are clarified by the following theorem.

THEOREM 1.1 (Pincherle [45]). *The continued fraction* (1.14) *converges if and only if the recurrence relation* (1.11) *possesses a minimal solution* f_n, *with* $f_0 \neq 0$. *In case of convergence, moreover, one has*

$$\frac{f_n}{f_{n-1}} = \frac{-b_n}{a_n -} \frac{b_{n+1}}{a_{n+1} -} \frac{b_{n+2}}{a_{n+2} -} \cdots, \qquad n = 1, 2, 3, \cdots, \tag{1.16}$$

provided $f_n \neq 0$ *for* $n = 0, 1, 2, \cdots$.

Proof. (a) Assume the continued fraction in (1.14) converges. Then so does the equivalent continued fraction (1.13). Therefore

$$\lim_{n\to\infty} \frac{\alpha_n}{\beta_n} = c,$$

where α_n, β_n are the solutions of (1.11) defined by the initial values (1.12), and c is some constant. Let

$$f_n = \alpha_n - c\beta_n . \tag{1.17}$$

Take any other solution of (1.11), say $y_n = a\alpha_n + b\beta_n$. Then $ac + b \neq 0$, and

$$\lim_{n\to\infty} \frac{f_n}{y_n} = \lim_{n\to\infty} \frac{\alpha_n - c\beta_n}{a\alpha_n + b\beta_n} = \lim_{n\to\infty} \frac{(\alpha_n/\beta_n) - c}{a(\alpha_n/\beta_n) + b} = 0.$$

This shows that the solution f_n defined in (1.17) is a minimal solution of (1.11). Moreover, $f_0 = \alpha_0 \neq 0$.

(b) Assume now that (1.11) possesses a minimal solution, f_n say, for which $f_0 \neq 0$. Then

$$f_n = f_0\alpha_n + f_1\beta_n , \qquad n \geqq 0.$$

We note that β_n is not a constant multiple of f_n, since $f_0 \neq 0$. Therefore, f_n being minimal,

$$\lim_{n\to\infty} \frac{f_n}{\beta_n} = f_0 \lim_{n\to\infty} \frac{\alpha_n}{\beta_n} + f_1 = 0,$$

and so

$$\lim_{n\to\infty} \frac{\alpha_n}{\beta_n} = -\frac{f_1}{f_0}.$$

This establishes convergence of the continued fraction (1.13), and thus of that in (1.14), and also proves (1.16) for $n = 1$.

To prove (1.16) for general $n > 1$, we need only observe that $z_m = f_{n+m-1}$, considered as a function of m, is a minimal solution of

$$z_{m+1} + a_{n+m-1}z_m + b_{n+m-1}z_{m-1} = 0, \qquad m = 1, 2, 3, \cdots .$$

Since by assumption, $z_0 = f_{n-1} \neq 0$, the portion of Theorem 1.1 already proved yields

$$\frac{z_1}{z_0} = \frac{f_n}{f_{n-1}} = \frac{-b_n}{a_n -} \frac{b_{n+1}}{a_{n+1} -} \frac{b_{n+2}}{a_{n+2} -} \cdots$$

as asserted. This completes the proof of Theorem1.1.

Consider again the three-term recurrence relation

$$y_{n+1} + a_n y_n + b_n y_{n-1} = 0, \qquad b_n \neq 0, \tag{1.18}$$

but assume, for simplicity, that the coefficients a_n, b_n are defined, and (1.18) holds, for all integers $n = 0, \pm 1, \pm 2, \cdots$. Let ν be an arbitrary integer, and let $\eta_n^{(\nu)}$ denote the solution of (1.18) having starting values

$$\eta_\nu^{(\nu)} = 1, \qquad \eta_{\nu+1}^{(\nu)} = 0 \tag{1.19}$$

at $n = \nu$ and $n = \nu + 1$, respectively. Then the following duality theorem holds.

THEOREM 1.2. *The function $\eta_n^{(\nu)}$ satisfies, for fixed ν and variable n, the three-term recurrence relation*

$$\eta_{n+1}^{(\nu)} + a_n \eta_n^{(\nu)} + b_n \eta_{n-1}^{(\nu)} = 0, \qquad n = 0, \pm 1, \pm 2, \cdots, \tag{1.20}$$

and for fixed n and variable ν, the three-term recurrence relation

$$\eta_n^{(\nu)} + \frac{a_\nu}{b_\nu} \eta_n^{(\nu-1)} + \frac{1}{b_{\nu-1}} \eta_n^{(\nu-2)} = 0, \qquad \nu = 0, \pm 1, \pm 2, \cdots. \tag{1.21}$$

Proof. The first part of the theorem follows from the definition of $\eta_n^{(\nu)}$. To prove the second part, we first observe that (1.21) holds true for $\nu = n - 1$, n, $n + 1$. For example, when $\nu = n$, using (1.19) and (1.20), we have

$$\eta_n^{(n)} + \frac{a_n}{b_n} \eta_n^{(n-1)} + \frac{1}{b_{n-1}} \eta_n^{(n-2)} = 1 + \frac{1}{b_{n-1}} \left(-a_{n-1} \eta_{n-1}^{(n-2)} - b_{n-1} \eta_{n-2}^{(n-2)} \right)$$

$$= 1 + 0 - 1 = 0.$$

The verification for $\nu = n \pm 1$ is analogous. Assume now (1.21) to be true for all integers n, ν satisfying $|n - \nu| \leqq k$, where $k \geqq 1$ is some integer. We show that (1.21) then also holds for $|n - \nu| = k + 1$. We consider the two cases $n - \nu = k + 1$, $n - \nu = -(k + 1)$ separately. In the first case, we use (1.20) in the form

$$\eta_n^{(\nu)} = -a_{n-1} \eta_{n-1}^{(\nu)} - b_{n-1} \eta_{n-2}^{(\nu)},$$

and observe that (1.21) can be applied to both terms on the right, since $|n - 1 - \nu| = k$, and $|n - 2 - \nu| = k - 1 < k$. We obtain

$$\eta_n^{(\nu)} = -a_{n-1} \left(-\frac{a_\nu}{b_\nu} \eta_{n-1}^{(\nu-1)} - \frac{1}{b_{\nu-1}} \eta_{n-1}^{(\nu-2)} \right) - b_{n-1} \left(-\frac{a_\nu}{b_\nu} \eta_{n-2}^{(\nu-1)} - \frac{1}{b_{\nu-1}} \eta_{n-2}^{(\nu-2)} \right)$$

$$= -\frac{a_\nu}{b_\nu} \left(-a_{n-1} \eta_{n-1}^{(\nu-1)} - b_{n-1} \eta_{n-2}^{(\nu-1)} \right) - \frac{1}{b_{\nu-1}} \left(-a_{n-1} \eta_{n-1}^{(\nu-2)} - b_{n-1} \eta_{n-2}^{(\nu-2)} \right)$$

$$= -\frac{a_\nu}{b_\nu} \eta_n^{(\nu-1)} - \frac{1}{b_{\nu-1}} \eta_n^{(\nu-2)},$$

having again used (1.20). The second case is verified similarly, using (1.20) in the form

$$\eta_n^{(\nu)} = -\frac{1}{b_{n+1}}\,(a_{n+1}\eta_{n+1}^{(\nu)} + \eta_{n+2}^{(\nu)}).$$

Since we already established (1.21) for $|\, n - \nu \,| \leqq 1$, it now follows by induction that the result holds for $|\, n - \nu \,| = k$, $k = 0, 1, 2, 3, \cdots$, that is, for all integers n, ν. Theorem 1.2 is proved.

We note that relation (1.21), for $\nu > n$, can also be obtained from the known fact (cf. [43, vol. I, p. 3]) that $\eta_{n-1}^{(\nu)}$ and $\eta_n^{(\nu)}$ are the numerators and denominators, respectively, of the continued fraction

$$b_{n-1}' + \frac{a_n'}{b_n'+}\,\frac{a_{n+1}'}{b_{n+1}'+}\cdots\frac{a_{\nu-1}'}{b_{\nu-1}'},$$

where

$$b_{m-1}' = -\frac{a_m}{b_m}, \qquad a_m' = -\frac{1}{b_m}.$$

Alternatively, Theorem 1.2 may be obtained, as a special case, from the known result that "multipliers" of a linear difference equation satisfy the adjoint difference equation (cf. [35, §12.6]).

2. Some results from the asymptotic theory of linear second order difference equations. In applications of Theorem 1.1, it is in general easier to recognize a given solution of a three-term recurrence relation to be minimal than to establish convergence of the corresponding continued fraction. One is aided in this by classical results from the asymptotic theory of difference equations, notably by a theorem of Poincaré, and by refinements and extensions thereof due to Perron and Kreuser. For convenience of the reader, we are recalling here these theorems for the special case of a second-order difference equation

$$y_{n+1} + a_n y_n + b_n y_{n-1} = 0, \qquad n = 1, 2, 3, \cdots. \tag{2.1}$$

We assume, throughout, that

$$b_n \neq 0, \qquad n = 1, 2, 3, \cdots. \tag{2.2}$$

We begin with the case where the coefficients a_n and b_n in (2.1) have finite limits

$$a_n \to a, \qquad b_n \to b, \qquad n \to \infty, \tag{2.3}$$

not excluding that $b = 0$. One then calls (2.1) a *Poincaré difference equation*, and calls

$$\Phi(t) = t^2 + at + b \tag{2.4}$$

the *characteristic polynomial* of (2.1). As may be expected, the solutions of (2.1) behave similarly, for large n, to the solutions of the difference equation (2.1) with *constant* coefficients $a_n = a$, $b_n = b$. This is borne out by the following two theorems.

THEOREM 2.1 (Poincaré [46]). *If the characteristic polynomial* (2.4) *of* (2.1) *has zeros* t_1, t_2 *of distinct moduli,*

$$|t_1| > |t_2|, \tag{2.5}$$

then for every nontrivial solution y_n *of* (2.1) *we have*

$$\lim_{n\to\infty} \frac{y_{n+1}}{y_n} = t_r, \qquad r = 1, \text{ or } r = 2. \tag{2.6}$$

THEOREM 2.2 (Perron [41]). *Under the assumption of Theorem* 2.1 *there exist two linearly independent solutions* $y_{n,1}$ *and* $y_{n,2}$ *of* (2.1) *such that*

$$\lim_{n\to\infty} \frac{y_{n+1,r}}{y_{n,r}} = t_r, \qquad r = 1, 2. \tag{2.7}$$

Theorem 2.2 implies that

$$f_n = y_{n,2}$$

is a minimal solution of (2.1). To see this, choose τ_1 and τ_2 such that

$$|t_2| < \tau_2 < \tau_1 < |t_1|,$$

which under the assumption (2.5) is certainly possible. By (2.7) we then have, for n sufficiently large,

$$\left|\frac{y_{n+1,1}}{y_{n,1}}\right| \geqq \tau_1, \qquad \left|\frac{y_{n+1,2}}{y_{n,2}}\right| \leqq \tau_2, \qquad n \geqq n_0.$$

Hence

$$|y_{n,1}| \geqq \tau_1^{n-n_0} |y_{n_0,1}|, \qquad |y_{n,2}| \leqq \tau_2^{n-n_0} |y_{n_0,2}|,$$

and

$$\left|\frac{y_{n,2}}{y_{n,1}}\right| \leqq \left(\frac{\tau_2}{\tau_1}\right)^{n-n_0} \left|\frac{y_{n_0,2}}{y_{n_0,1}}\right|, \qquad n \geqq n_0.$$

This shows that

$$\lim_{n\to\infty} \frac{y_{n,2}}{y_{n,1}} = 0,$$

from which the assertion follows.

We also note that in (2.6) one has $r = 2$ for the minimal solution, and $r = 1$ for any other solution.

We shall require a generalization of Theorem 2.2 relating to a difference equation (2.1) whose coefficients satisfy

$$a_n \sim an^\alpha, \qquad b_n \sim bn^\beta, \qquad ab \neq 0; \quad \alpha, \beta \text{ real}; \quad n \to \infty. \tag{2.8}$$

The asymptotic structure of the solutions now depends on the *Newton-Puiseux* diagram formed with the points $P_0(0, 0)$, $P_1(1, \alpha)$, $P_2(2, \beta)$. This is the broken line $\overline{P_0P_1P_2}$, if P_1 is above the straight line joining P_0 with P_2; otherwise it

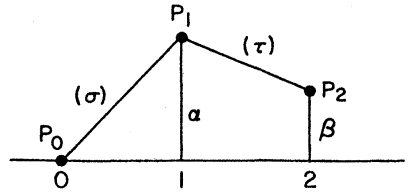


FIG. 1. *Newton-Puiseux diagram for difference equation* (2.1), (2.8)

is the line segment $\overline{P_0P_2}$. We denote by σ the slope of $\overline{P_0P_1}$, and by τ the slope of $\overline{P_1P_2}$ (Fig. 1), so that $\sigma = \alpha, \tau = \beta - \alpha$.

THEOREM 2.3 (Perron [42], Kreuser [29]). (a) *If the point P_1 is above the line segment $\overline{P_0P_2}$ (i.e., $\sigma > \tau$), the difference equation* (2.1) *has two linearly independent solutions, $y_{n,1}$ and $y_{n,2}$, for which*

$$\frac{y_{n+1,1}}{y_{n,1}} \sim -an^{\sigma}, \qquad \frac{y_{n+1,2}}{y_{n,2}} \sim -\frac{b}{a}n^{\tau}, \qquad n \to \infty. \tag{2.9}$$

(b) *If the points P_0, P_1, P_2 are collinear (i.e., $\sigma = \tau = \alpha$), let t_1, t_2 be the roots of $t^2 + at + b = 0$, and $|t_1| \geqq |t_2|$. Then* (2.1) *has two linearly independent solutions, $y_{n,1}$ and $y_{n,2}$, such that*

$$\frac{y_{n+1,1}}{y_{n,1}} \sim t_1 n^{\alpha}, \qquad \frac{y_{n+1,2}}{y_{n,2}} \sim t_2 n^{\alpha}, \qquad n \to \infty, \tag{2.10}$$

provided $|t_1| > |t_2|$. If $|t_1| = |t_2|$ (in particular, if t_1, t_2 are complex conjugates) then

$$\limsup_{n\to\infty} \left[\frac{|y_n|}{(n!)^{\alpha}}\right]^{1/n} = |t_1| \tag{2.11}$$

for all nontrivial solutions of (2.1).

(c) *If the point P_1 lies below the line segment $\overline{P_0P_2}$ then*

$$\limsup_{n\to\infty} \left[\frac{|y_n|}{(n!)^{\beta/2}}\right]^{1/n} = \sqrt{|b|} \tag{2.12}$$

for all nontrivial solutions of (2.1).

An argument similar to the one following Theorem 2.2 will show that in both case (a) and the first part of case (b) the solution $f_n = y_{n,2}$ is a minimal solution of (2.1). Furthermore, in the first part of case (b),

$$\lim_{n\to\infty} \frac{y_{n+1}}{n^{\alpha}y_n} = t_r, \qquad r = 1, \text{ or } r = 2, \tag{2.13}$$

where $r = 2$ for the minimal solution, and $r = 1$ for any other solution.

The second part of (b), and part (c) of Theorem 2.3 are somewhat inconclusive for our purposes, as they do not permit distinguishing two solutions with distinct asymptotic properties. In this connection, the example given later in §9 is of interest.

Proofs of Poincaré's theorem may be found, e.g., in [21], [35], [37]. An elegant

proof of Perron's theorem is given in [14], and reproduced in [34]. Far-reaching generalizations, and simplified proofs, of all these theorems, including Kreuser's theorem, were recently obtained in [51].

3. A first algorithm for computing the minimal solution. We assume now that the recurrence relation

$$y_{n+1} + a_n y_n + b_n y_{n-1} = 0, \qquad n = 1, 2, 3, \cdots, \tag{3.1}$$

has a nonvanishing[5] minimal solution, f_n. We wish to calculate f_n for $n = 0, 1, 2, \cdots, N$. In order to specify f_n uniquely, we can impose *one* condition, for example prescribe the value of f_0. For later applications, we consider the more general normalization

$$\sum_{m=0}^{\infty} \lambda_m f_m = s, \qquad s \neq 0, \tag{3.2}$$

where s and $\lambda_0, \lambda_1, \cdots$ are given quantities, and the series is known to converge. We do not exclude that $\lambda_m = 0$ for all $m > 0$, in which case (3.2) amounts to prescribing f_0.

In a sense, (3.2) represents the most general linear condition that may be imposed. A class of nonlinear conditions will also be considered briefly.

To introduce the algorithm, let

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

Suppose first that r_n, s_n are known for some value $n = \nu \geqq N$. The desired solution f_n, $n = 0(1)N$, can then be obtained as follows.

From Theorem 1.1 we know that

$$r_{n-1} = \frac{-b_n}{a_n -} \frac{b_{n+1}}{a_{n+1} -} \frac{b_{n+2}}{a_{n+2} -} \cdots, \qquad n = 1, 2, 3, \cdots. \tag{3.4}$$

Hence, we can generate the ratios r_n for $0 \leqq n < \nu$ as in (1.6)–(1.8) by

$$r_{n-1} = \frac{-b_n}{a_n + r_n}, \qquad n = \nu, \nu - 1, \cdots, 1. \tag{3.5}$$

Similarly, we have

$$\begin{aligned} s_{n-1} &= \frac{1}{f_{n-1}} \sum_{m=n}^{\infty} \lambda_m f_m = \frac{1}{f_{n-1}} \left(\lambda_n f_n + \sum_{m=n+1}^{\infty} \lambda_m f_m \right) \\ &= \lambda_n r_{n-1} + r_{n-1} \left(\frac{1}{f_n} \sum_{m=n+1}^{\infty} \lambda_m f_m \right), \end{aligned}$$

so that

$$s_{n-1} = r_{n-1}(\lambda_n + s_n), \qquad n = \nu, \nu - 1, \cdots, 1. \tag{3.6}$$

Hence, also the quantities s_n for $0 \leqq n < \nu$, and thus in particular s_0, can be ob-

[5] The assumption of f_n to be nonvanishing is no serious restriction from the practical point of view. This is further discussed at the end of this paragraph.

tained recursively. Using (3.2) we now have

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

and so

$$f_0 = \frac{s}{\lambda_0 + s_0}.$$

This gives us the initial value of the desired solution. The remaining values can now be obtained immediately from

$$f_n = r_{n-1}f_{n-1}, \qquad n = 1, 2, \cdots, N.$$

The actual algorithm follows this procedure very closely, except that for the infinite continued fraction, and the infinite series, representing r_{n-1} and s_n, respectively, we now substitute truncated continued fractions, and truncated series. More precisely, we define

$$r_\nu^{(\nu)} = 0, \qquad r_{n-1}^{(\nu)} = \frac{-b_n}{a_n -}\,\frac{b_{n+1}}{a_{n+1} -}\,\cdots\,\frac{b_\nu}{a_\nu}, \qquad 1 \leqq n \leqq \nu, \tag{3.7}$$

and

$$s_\nu^{(\nu)} = 0, \qquad s_n^{(\nu)} = \sum_{m=n+1}^{\nu}\lambda_m r_n^{(\nu)} r_{n+1}^{(\nu)}\cdots r_{m-1}^{(\nu)}, \qquad 0 \leqq n < \nu. \tag{3.8}$$

One then verifies readily that the formulas (3.5), (3.6) continue to hold if r_n is replaced by $r_n^{(\nu)}$, and s_n by $s_n^{(\nu)}$ throughout. Hence the following set of recursions arises naturally,

$$\begin{aligned} &r_\nu^{(\nu)} = 0, \qquad r_{n-1}^{(\nu)} = \frac{-b_n}{a_n + r_n^{(\nu)}}, \\ &s_\nu^{(\nu)} = 0, \qquad s_{n-1}^{(\nu)} = r_{n-1}^{(\nu)}(\lambda_n + s_n^{(\nu)}), \qquad n = \nu, \nu-1, \cdots, 1, \\ &f_0^{(\nu)} = \frac{s}{\lambda_0 + s_0^{(\nu)}}, \qquad f_n^{(\nu)} = r_{n-1}^{(\nu)} f_{n-1}^{(\nu)}, \qquad n = 1, 2, \cdots, N. \end{aligned} \tag{3.9}$$

While our initial procedure gave us the exact values f_n of the minimal solution, the quantities $f_n^{(\nu)}$ now derived are at best approximations to f_n. It remains to successively improve $f_n^{(\nu)}$ by repeating (3.9) for a sequence of increasing values of ν. The complete algorithm for computing the minimal solution may thus be defined as follows:

Step 1: Select an integer $\nu \geqq N$, and let $\phi_n^{(\nu)} = 0$, $n = 0, 1, \cdots, N$.

Step 2: Calculate $f_n^{(\nu)}$, $n = 0, 1, \cdots, N$, according to the formulas in (3.9).

Step 3: If the $N + 1$ values of $f_n^{(\nu)}$ obtained in Step 2 do not agree with the current values of $\phi_n^{(\nu)}$ to within the desired accuracy, then redefine $\phi_n^{(\nu)}$ by $\phi_n^{(\nu)} = f_n^{(\nu)}$, $n = 0, 1, \cdots, N$, increase ν by some fixed integer, say 5, and repeat Step 2; otherwise accept $f_n^{(\nu)}$ as the final approximations to f_n, $n = 0, 1, \cdots, N$.

We note that in the special case $\lambda_0 = 1, \lambda_1 = \lambda_2 = \cdots = 0$, all $s_n^{(\nu)}$ vanish, so that the recursion for $s_{n-1}^{(\nu)}$ in (3.9) may be omitted. Moreover, $s = f_0$, and there-

fore $f_0^{(\nu)} = f_0$. In this case, the value of f_0 must be known before the algorithm (3.9) can be applied. The use of an infinite series (3.2), instead, has the remarkable advantage of not requiring any value of f_n to be known in advance.

Our derivation of (3.9) also demonstrates that $f_n^{(\nu)} = f_n$ if instead of zero initial values in the first two recursions we select initial values $r_\nu^{(\nu)} = r_\nu$, $s_\nu^{(\nu)} = s_\nu$. While these quantities in general are not known beforehand, they may sometimes be approximated closely when ν is large. This suggests to modify (3.9) by defining

$$r_\nu^{(\nu)} = \rho_\nu, \qquad s_\nu^{(\nu)} = \sigma_\nu, \tag{3.10}$$

where ρ_ν and σ_ν are suitable approximations to r_ν and s_ν, respectively. The better these approximations are, the faster we expect our algorithm to converge. We return to this point later.

We may give (3.9) a somewhat different interpretation as follows. Consider the solution $\eta_n^{(\nu)}$ of the difference equation (3.1), defined by "initial" values

$$\eta_\nu^{(\nu)} = 1, \qquad \eta_{\nu+1}^{(\nu)} = 0 \tag{3.11}$$

at $n = \nu$ and $n = \nu + 1$, respectively. The values of $\eta_n^{(\nu)}$ for $0 \leqq n \leqq \nu$ may be obtained by applying (3.1) in the backward direction, starting at $n = \nu$. Then we assert that

$$f_n^{(\nu)} = \frac{s}{\sum_{m=0}^{\nu} \lambda_m \eta_m^{(\nu)}} \eta_n^{(\nu)}, \qquad 0 \leqq n \leqq N. \tag{3.12}$$

To verify this, we observe, first of all, that the quantities $r_{n-1}^{(\nu)}$ defined in (3.7) are consecutive ratios of the solution $\eta_n^{(\nu)}$,

$$r_{n-1}^{(\nu)} = \frac{\eta_n^{(\nu)}}{\eta_{n-1}^{(\nu)}}, \qquad 1 \leqq n \leqq \nu + 1. \tag{3.13}$$

This is trivial for $n = \nu + 1$, and for $n \leqq \nu$ follows from the fact that the ratio $\eta_n^{(\nu)}/\eta_{n-1}^{(\nu)}$ satisfies the same nonlinear recursion (3.5) satisfied by $r_{n-1}^{(\nu)}$. Inserting (3.13) into (3.8), we find

$$s_n^{(\nu)} = \frac{1}{\eta_n^{(\nu)}} \sum_{m=n+1}^{\nu} \lambda_m \eta_m^{(\nu)},$$

and using this for $n = 0$, we obtain

$$f_0^{(\nu)} = \frac{s}{\lambda_0 + s_0^{(\nu)}} = \frac{s\eta_0^{(\nu)}}{\lambda_0 \eta_0^{(\nu)} + \eta_0^{(\nu)} s_0^{(\nu)}} = \frac{s}{\lambda_0 \eta_0^{(\nu)} + \sum_{m=1}^{\nu} \lambda_m \eta_m^{(\nu)}} \eta_0^{(\nu)}.$$

This proves our assertion (3.12) for $n = 0$. To prove it for $n > 0$, we need only observe that in view of (3.13), the quantities $f_n^{(\nu)}$ in (3.12) satisfy $f_n^{(\nu)}/f_{n-1}^{(\nu)} = r_{n-1}^{(\nu)}$, as required by (3.9).

The algorithm of generating the $\eta_n^{(\nu)}$ and using (3.12) is often referred to as *Miller's backward recurrence algorithm*. It was first proposed as a computational scheme by J. C. P. Miller in connection with the tabulation of Bessel functions (see [5, p. xvii]). An error analysis has recently been given by Olver [38].

While algorithm (3.9) and Miller's algorithm are mathematically equivalent, they have different computational characteristics. In many cases, e.g., the quantities $\eta_n^{(\nu)}$ grow rapidly as ν increases, and may cause "overflow" on a digital computer. In contrast to this, the quantity $r_n^{(\nu)}$ in (3.9) converges to a finite limit as $\nu \to \infty$, and so does $s_n^{(\nu)}$ if the algorithm converges at all.

We now use (3.12) to discuss convergence as $\nu \to \infty$ of the algorithm (3.9). Let g_n denote any solution of the difference equation (3.1) other than f_n, so that

$$\lim_{n\to\infty} \frac{f_n}{g_n} = 0. \tag{3.14}$$

Clearly,

$$\eta_n^{(\nu)} = a^{(\nu)} f_n + b^{(\nu)} g_n,$$

for some constants $a^{(\nu)}$, $b^{(\nu)}$. By (3.11), we must have

$$a^{(\nu)} f_{\nu+1} + b^{(\nu)} g_{\nu+1} = 0,$$
$$a^{(\nu)} f_\nu + b^{(\nu)} g_\nu = 1.$$

The first of these relations gives $b^{(\nu)} = -(f_{\nu+1}/g_{\nu+1})a^{(\nu)}$, so that

$$\eta_n^{(\nu)} = a^{(\nu)} \left(f_n - \frac{f_{\nu+1}}{g_{\nu+1}} g_n \right).$$

Substituting in (3.12), and simplifying, we obtain

$$f_n^{(\nu)} = \frac{f_n \left(1 - \dfrac{f_{\nu+1}}{g_{\nu+1}} \dfrac{g_n}{f_n} \right)}{1 - \dfrac{1}{s} \displaystyle\sum_{m=\nu+1}^{\infty} \lambda_m f_m - \dfrac{f_{\nu+1}}{s g_{\nu+1}} \sum_{m=0}^{\nu} \lambda_m g_m}. \tag{3.15}$$

In view of (3.14) and the convergence of the infinite series in (3.2), it is clear that $\lim_{\nu\to\infty} f_n^{(\nu)} = f_n$ if and only if

$$\lim_{\nu\to\infty} \frac{f_{\nu+1}}{g_{\nu+1}} \sum_{m=0}^{\nu} \lambda_m g_m = 0. \tag{3.16}$$

We have proved the following theorem.

THEOREM 3.1. *Suppose the recurrence relation* (3.1) *has a nonvanishing minimal solution,* f_n, *for which* (3.2) *holds. Let* g_n *be any other solution of* (3.1). *Then the algorithm* (3.9) *converges in the sense*

$$\lim_{\nu\to\infty} f_n^{(\nu)} = f_n$$

if and only if (3.16) *is satisfied.*

Condition (3.16) holds, e.g., if the λ's are uniformly bounded, and

$$\frac{g_{\nu+1}}{g_\nu} \to t_1, \qquad \frac{f_{\nu+1}}{f_\nu} \to t_2, \qquad \nu \to \infty,$$
$$|t_1| > |t_2|, \qquad |t_2| < 1.$$

If all but a finite number of the λ's are zero, then (3.16) is a consequence of (3.14). Theorem 3.1, in this case, has been noted previously in [16].

It is useful to observe that convergence of the algorithm (3.9), in the sense of Theorem 3.1, implies that

$$r_n^{(\nu)} \to r_n\,, \qquad s_n^{(\nu)} \to s_n\,, \qquad \nu \to \infty\,, \tag{3.17}$$

where r_n, s_n are the quantities defined in (3.3). The first of these relations follows directly from (3.4) and (3.7). The second follows by induction on n. Indeed, if $n = 0$, we have from the third line in (3.9),

$$s_0^{(\nu)} = \frac{s - \lambda_0 f_0^{(\nu)}}{f_0^{(\nu)}} \to \frac{s - \lambda_0 f_0}{f_0} = s_0\,, \qquad \nu \to \infty.$$

Assuming now $s_{n-1}^{(\nu)} \to s_{n-1}$, we get from the second line in (3.9), and from (3.6), that

$$s_n^{(\nu)} = \frac{s_{n-1}^{(\nu)}}{r_{n-1}^{(\nu)}} - \lambda_n \to \frac{s_{n-1}}{r_{n-1}} - \lambda_n = s_n\,, \qquad \nu \to \infty.$$

In case of convergence of the algorithm (3.9), we may obtain from (3.15) the following approximate expression for the relative error, valid for ν sufficiently large,

$$\frac{f_n^{(\nu)} - f_n}{f_n} \doteq \frac{1}{s}\sum_{m=\nu+1}^{\infty} \lambda_m f_m + \frac{f_{\nu+1}}{s g_{\nu+1}} \sum_{m=0}^{\nu} \lambda_m g_m - \frac{f_{\nu+1}}{g_{\nu+1}} \frac{g_n}{f_n}\,. \tag{3.18}$$

It is interesting to examine what effect the modification (3.10) of algorithm (3.9) will have upon the relative error (3.18). We assume that

$$\rho_\nu = r_\nu(1 + \epsilon_\nu), \qquad \sigma_\nu = s_\nu(1 + \eta_\nu),$$

where r_ν, s_ν are defined by (3.3), and ϵ_ν, η_ν are small numbers. Then a simple computation will show that in place of (3.15) we now have

$$f_n^{(\nu)} = f_n \frac{1 + \epsilon_\nu \dfrac{f_{\nu+1}}{g_{\nu+1}} \dfrac{g_n}{f_n} \left(1 - \rho_\nu \dfrac{g_\nu}{g_{\nu+1}}\right)^{-1}}{1 + \dfrac{\eta_\nu}{s} \displaystyle\sum_{m=\nu+1}^{\infty} \lambda_m f_m + \dfrac{\epsilon_\nu f_{\nu+1}}{s g_{\nu+1}} \left(1 - \rho_\nu \dfrac{g_\nu}{g_{\nu+1}}\right)^{-1} \sum_{m=0}^{\nu} \lambda_m g_m}\,.$$

Since $|\, \rho_\nu g_\nu / g_{\nu+1} \,|$ is usually substantially smaller than 1 (at least for large ν) we see that the modification (3.10) reduces the relative error of $f_n^{(\nu)}$ effectively by a factor of $|\, \epsilon_\nu \,|$, or $|\, \eta_\nu \,|$, whichever is larger. Hence, our statement made earlier that the convergence of $f_n^{(\nu)}$ to f_n is faster the better ρ_ν approximates r_ν, and σ_ν approximates s_ν, is clearly vindicated.

It is tempting to try a substitution of the type

$$F_n = c_n f_n\,, \qquad c_n \neq 0, \tag{3.19}$$

to exert influence upon the convergence criteria (3.14) and (3.16). We note, however, that *these criteria are invariant with respect to any linear substitution of the form* (3.19).

We now briefly consider the case in which condition (3.2) is replaced by a nonlinear condition of the form

$$\sum_{m=0}^{\infty} \lambda_m f_m{}^p = s, \qquad s \neq 0, \tag{3.2p}$$

where p is some real number. It must be noted that this condition specifies the minimal solution only to within a constant factor c satisfying $c^p = 1$.

Algorithm (3.9) extends readily to the case of general p, if we define $r_{n-1}^{(\nu)}$ as before, and let

$$s_\nu{}^{(\nu)} = 0, \qquad s_n{}^{(\nu)} = \sum_{m=n+1}^{\nu} \lambda_m [r_n{}^{(\nu)} r_{n+1}^{(\nu)} \cdots r_{m-1}^{(\nu)}]^p, \qquad 0 \leqq n < \nu.$$

We obtain

$$\begin{aligned} &r_\nu{}^{(\nu)} = 0, \qquad r_{n-1}^{(\nu)} = \frac{-b_n}{a_n + r_n{}^{(\nu)}}, && n = \nu, \nu-1, \cdots, 1, \\ &s_\nu{}^{(\nu)} = 0, \qquad s_{n-1}^{(\nu)} = [r_{n-1}^{(\nu)}]^p (\lambda_n + s_n{}^{(\nu)}), && \\ &f_0{}^{(\nu)} = \left[\frac{s}{1 + s_0{}^{(\nu)}}\right]^{1/p}, \qquad f_n{}^{(\nu)} = r_{n-1}^{(\nu)} f_{n-1}^{(\nu)}, && n = 1, 2, \cdots, N. \end{aligned} \tag{3.9p}$$

The nonuniqueness of f_n is reflected in the multivalued definition of $f_0{}^{(\nu)}$.

As in the proof of Theorem 3.1, one shows that (3.9p) converges as $\nu \to \infty$ if

$$\lim_{\nu \to \infty} h_\nu{}^{(i)} = 0, \qquad i = 1, 2, \cdots, p,$$

where

$$h_\nu{}^{(i)} = \left(\frac{f_{\nu+1}}{g_{\nu+1}}\right)^i \sum_{m=0}^{\nu} \lambda_m g_m{}^i f_m{}^{p-i}.$$

We conclude this paragraph with some practical remarks concerning the algorithm (3.9).

The effectiveness of the algorithm is clearly enhanced if good estimates of the initial value of ν are available. Such estimates can sometimes be obtained from (3.18), and from known asymptotic properties of the solutions f_n and g_n. (See §§5, 7 for examples.)

It is worth noting that the storage requirements on a digital computer do not depend on ν. It suffices to store permanently only those N quantities $r_n{}^{(\nu)}$ which are needed to build up the final results $f_n{}^{(\nu)}$. All the other $r_n{}^{(\nu)}$, as well as the $s_n{}^{(\nu)}$, can be generated in temporary storage cells.

The assumption

$$f_{n-1} \neq 0, \qquad n = 1, 2, 3, \cdots,$$

in Theorem 3.1 is ordinarily fulfilled in practice, if for no other reason than rounding errors. Nevertheless, one might think, in view of $\lim_{\nu \to \infty} r_{n-1}^{(\nu)} = f_n/f_{n-1}$, that the case of f_{n-1} nearly equal to zero for some $n \geqq 1$ might cause numerical diffi-

culties. By the following, admittedly superficial, considerations we wish to show that the presence, or proximity, of such zeros need be of no great concern.

Suppose, indeed, that f_{n-1} is very small in modulus, compared to f_n. For definiteness, let $n > 1$. Then, by (3.3), $|\, r_{n-1}\,|$ is very large, and so is $|\, r_{n-1}^{(\nu)}\,|$, when ν is sufficiently large. From the first line in (3.9) it follows that $|\, a_n + r_n^{(\nu)}|$ must be very small compared to $|\, b_n\,|$. Since neither a_n nor $r_n^{(\nu)}$ will normally be small, this means that many digits will cancel when the sum $a_n + r_n^{(\nu)}$ is formed, and so $r_{n-1}^{(\nu)}$ is not only very large, but also very inaccurate in terms of significant digits. Consequently, $r_{n-2}^{(\nu)}$ will be very small, and also inaccurate. However, $r_{n-3}^{(\nu)} = -b_{n-2}/(a_{n-2} + r_{n-2}^{(\nu)})$ (if $n > 2$) will again be accurate, since a_{n-2} in the denominator picks up lost accuracy, $r_{n-2}^{(\nu)}$ being normally much smaller than a_{n-2}. Later on, in the formation of the final results, $f_{n-1}^{(\nu)} = r_{n-2}^{(\nu)} f_{n-2}^{(\nu)}$ will come out very small and inaccurate, as one must expect. The really questionable point is the computation of $f_n^{(\nu)} = r_{n-1}^{(\nu)} f_{n-1}^{(\nu)}$, since $r_{n-1}^{(\nu)}$ is large and $f_{n-1}^{(\nu)}$ is small, and both are inaccurate. We note, however, that

$$f_n^{(\nu)} = r_{n-1}^{(\nu)} r_{n-2}^{(\nu)} f_{n-2}^{(\nu)} = r_{n-1}^{(\nu)} \frac{-b_{n-1} f_{n-2}^{(\nu)}}{a_{n-1} + r_{n-1}^{(\nu)}} = \frac{-b_{n-1} f_{n-2}^{(\nu)}}{1 + (a_{n-1}/r_{n-1}^{(\nu)})},$$

which shows that the largeness of $r_{n-1}^{(\nu)}$ saves $f_n^{(\nu)}$ from becoming inaccurate, even though $r_{n-1}^{(\nu)}$ is. A similar reasoning applies to $s_{n-1}^{(\nu)}$, $s_{n-2}^{(\nu)}$.

More serious is a possible loss of accuracy in the calculation of $f_0^{(\nu)}$, as this would affect all subsequent $f_n^{(\nu)}$. It could indeed occur that $|\, \lambda_0 + s_0^{(\nu)}|$ is small in comparison with $|\, \lambda_0\,|$, so that many digits cancel when $\lambda_0 + s_0^{(\nu)}$ is formed. The resulting value of $f_0^{(\nu)}$ would then be quite inaccurate. The same difficulty might arise if $\lambda_0 = 0$. Suppose, indeed, that λ_p $(p > 0)$ is the first nonvanishing coefficient in the series (3.2),

$$\lambda_p \neq 0, \qquad \lambda_m = 0, \qquad 0 \leqq m < p,$$

and that $|\lambda_p + s_p^{(\nu)}|$ happens to be very small compared to $|\, \lambda_p\,|$. Then $s_{p-1}^{(\nu)}$ is necessarily inaccurate, and this inaccuracy will be transmitted to all subsequent $s_{n-1}^{(\nu)}$, and finally to $f_0^{(\nu)}$, in view of the relations $s_{n-1}^{(\nu)} = r_{n-1}^{(\nu)} s_n^{(\nu)}$, $n = p - 1$, $p - 2, \cdots, 1$, and $f_0^{(\nu)} = s/s_0^{(\nu)}$.

Now for large ν, and $p \geqq 0$, we have

$$\lambda_p + s_p^{(\nu)} \doteq \lambda_p + \frac{1}{f_p} \sum_{m=p+1}^{\infty} \lambda_m f_m = \lambda_p + \frac{1}{f_p}(s - \lambda_p f_p) = \frac{s}{f_p},$$

so that $|(\lambda_p + s_p^{(\nu)})/\lambda_p\,|$ is small if $|\, s/(\lambda_p f_p)|$ is small. Hence, *dangerous cancellation occurs when s is small in absolute value compared to the first nonvanishing term* $\lambda_p f_p$ *in* (3.2), i.e., when cancellation occurs in the series (3.2) itself. For this reason, some care must be exercised in the selection of the identity (3.2).

4. Second and third algorithm for computing the minimal solution. The effectiveness of our first algorithm (3.9) is somewhat limited if no reasonable estimate of the starting value ν of n is known a priori. The recursions in (3.9) must then be repeated with increasing values of ν, until sufficient agreement is obtained between successive results $f_n^{(\nu)}$, for all $n = 0, 1, \cdots, N$. This disadvantage can

be removed, at the expense of a more complex algorithm, by making use of the duality theorem 1.2, or, alternatively, by evaluating a sequence of continued fractions (3.4). The corresponding algorithms will now be developed. The first of these, though not in the form given here, is due to Shintani [52].

As was noted in the previous paragraph, we can obtain r_n , s_n recursively, for $0 \leqq n < N$, and hence also f_n for $0 \leqq n \leqq N$, once r_N , s_N are known. In the following we derive a method for calculating r_N , s_N recursively. If f_0 is known, the s_n will not be required, and the algorithm then reduces to one suggested by G. Blanch ([4, p. 405 ff]) in connection with Bessel functions.

As for r_N , we may simply evaluate the continued fraction

$$r_N = \frac{-b_{N+1}}{a_{N+1}-} \frac{b_{N+2}}{a_{N+2}-} \frac{b_{N+3}}{a_{N+3}-} \cdots, \tag{4.1}$$

by either the first, or third method described in §1. In the first case we have

$$r_N = \lim_{k\to\infty} \frac{A_k}{B_k}, \tag{4.2}$$

where

$$\begin{aligned} &A_{-1} = 1, \quad A_0 = 0; \quad B_{-1} = 0, \quad B_0 = 1; \\ &A_k = a_{N+k}A_{k-1} - b_{N+k}A_{k-2}, \\ &B_k = a_{N+k}B_{k-1} - b_{N+k}B_{k-2}, \end{aligned} \qquad k = 1, 2, 3, \cdots. \tag{4.3}$$

In the second case we have

$$r_N = \lim_{k\to\infty} w_k, \tag{4.4}$$

where the w's are generated as follows:

$$\begin{aligned} &u_1 = 1, \quad v_1 = w_1 = -\frac{b_{N+1}}{a_{N+1}}, \\ &u_{k+1} = \frac{1}{1 - (b_{N+k+1}/a_{N+k}a_{N+k+1})u_k}, \\ &v_{k+1} = v_k(u_{k+1} - 1), \\ &w_{k+1} = w_k + v_{k+1}, \end{aligned} \qquad k = 1, 2, 3, \cdots. \tag{4.5}$$

For the computation of s_N , we make use of the fact that

$$s_N = \lim_{\nu\to\infty} s_N^{(\nu)}, \tag{4.6}$$

where $s_N^{(\nu)}$ is defined by (3.8). The quantities $s_N^{(\nu)}$, $\nu \geqq N$, may be obtained recursively as follows. From the definition (3.8) of $s_n^{(\nu)}$, and from (3.13), we note that

$$\eta_N^{(\nu)} s_N^{(\nu)} = \sum_{m=N+1}^{\nu} \lambda_m \eta_m^{(\nu)}. \tag{4.7}$$

Hence, using (1.21), we can write

$$\eta_N^{(\nu)} s_N^{(\nu)} = \sum_{m=N+1}^{\nu} \lambda_m \left[-\frac{a_\nu}{b_\nu} \eta_m^{(\nu-1)} - \frac{1}{b_{\nu-1}} \eta_m^{(\nu-2)} \right]$$

$$= -\frac{a_\nu}{b_\nu} (\eta_N^{(\nu-1)} s_N^{(\nu-1)} + \lambda_\nu \eta_\nu^{(\nu-1)})$$

$$- \frac{1}{b_{\nu-1}} (\eta_N^{(\nu-2)} s_N^{(\nu-2)} + \lambda_{\nu-1} \eta_{\nu-1}^{(\nu-2)} + \lambda \eta^{(\cdot\cdot)})$$

Since

$$\eta_\nu^{(\nu-1)} = \eta_{\nu-1}^{(\nu-2)} = 0,$$

$$\eta_\nu^{(\nu-2)} = -a_{\nu-1}\eta_{\nu-1}^{(\nu-2)} - b_{\nu-1}\eta_{\nu-2}^{(\nu-2)} = -b_{\nu-1},$$

we get

$$\eta_N^{(\nu)} s_N^{(\nu)} = -\frac{a_\nu}{b_\nu} \eta_N^{(\nu-1)} s_N^{(\nu-1)} - \frac{1}{b_{\nu-1}} \eta_N^{(\nu-2)} s_N^{(\nu-2)} + \lambda_\nu, \tag{4.8}$$

or, alternatively,

$$s_N^{(\nu)} = -\rho_N^{(\nu-1)} \left[\frac{a_\nu}{b_\nu} s_N^{(\nu-1)} + \frac{1}{b_{\nu-1}} \rho_N^{(\nu-2)} s_N^{(\nu-2)} \right] + \frac{\lambda_\nu}{\eta_N^{(\nu)}},$$

where we have set $\rho_N^{(\nu)} = \eta_N^{(\nu)}/\eta_N^{(\nu+1)}$. Taking into account the recursive relations for $\rho_N^{(\nu)}$, $\eta_N^{(\nu)}$, which follow from (1.21), we arrive at the following algorithm for generating $s_N^{(\nu)}$:

$$\rho_N^{(N-1)} = 0, \qquad \eta_N^{(N)} = 1, \qquad \eta_N^{(N-1)} = 0,$$

$$s_N^{(N)} = s_N^{(N-1)} = 0,$$

$$\rho_N^{(\nu-1)} = -\frac{1}{\dfrac{a_\nu}{b_\nu} + \dfrac{1}{b_{\nu-1}} \rho_N^{(\nu-2)}}$$

$$\eta_N^{(\nu)} = -\frac{a_\nu}{b_\nu} \eta_N^{(\nu-1)} - \frac{1}{b_{\nu-1}} \eta_N^{(\nu-2)} \qquad \nu = N+1, N+2, \cdots. \tag{4.9}$$

$$s_N^{(\nu)} = -\rho_N^{(\nu-1)} \left[\frac{a_\nu}{b_\nu} s_N^{(\nu-1)} + \frac{\rho_N^{(\nu-2)}}{b_{\nu-1}} s_N^{(\nu-2)} \right] + \frac{\lambda_\nu}{\eta_N^{(\nu)}}$$

This, together with (4.2) [or (4.4)] and the remarks at the beginning of this paragraph constitutes our second algorithm for computing the minimal solution of (3.1).

As noted previously, the quantities $\eta_N^{(\nu)}$ may grow rapidly, as ν increases, and may cause overflow on a computer. However, if $\eta_N^{(\nu)}$ is large, just short of overflowing, it is normally permissible to replace the term $\lambda_\nu/\eta_N^{(\nu)}$ in the last relation of (4.9) by zero, and to continue the recursion for $s_N^{(\nu)}$ in the truncated form.

To develop the third algorithm, let

$$q_n = \frac{1}{f_0} \sum_{m=0}^{n} \lambda_m f_m = \sum_{m=0}^{n} \lambda_m r_0 r_1 \cdots r_{m-1}, \tag{4.12}$$

where as before $r_{n-1} = f_n/f_{n-1}$. Denoting the product of the first n of the r's by p_n, we obtain

$$\begin{aligned} p_0 &= 1, \qquad p_n = r_{n-1}p_{n-1}, \\ q_0 &= \lambda_0, \qquad q_n = q_{n-1} + \lambda_n p_n, \end{aligned} \qquad n = 1, 2, 3, \cdots. \tag{4.13}$$

Each r_{n-1} in (4.13) will be computed from the continued fraction

$$r_{n-1} = \frac{-b_n}{a_n -} \frac{b_{n+1}}{a_{n+1} -} \frac{b_{n+2}}{a_{n+2} -} \cdots$$

by applying either (4.2), (4.3), or (4.4), (4.5), with N replaced by $n - 1$. From (4.12), and the identity (3.2), it follows that

$$q = \lim_{n\to\infty} q_n = \frac{s}{f_0}.$$

Hence we continue generating the q_n in (4.13) until they meet some specific criterion of convergence. Thereafter, we may obtain as many of the final answers as desired by means of

$$f_0 = s/q, \qquad f_n = p_n f_0, \qquad n = 1, 2, 3, \cdots. \tag{4.14}$$

If the q_n converge too rapidly, it may occur, of course, that some of the later p_n required in (4.14) are not yet available, and must be generated by continuing the first recursion in (4.13). It should also be noted that the q-recursion in (4.13) can be omitted if f_0 is known in advance.

An obvious disadvantage of the third algorithm is the fact that a rather large number of continued fractions have to be evaluated, in contrast to just one continued fraction in the first two algorithms. Even though some of these continued fractions (especially the later ones) may converge quite rapidly, the expenditure of computation in the third algorithm is in general higher than in the first and second algorithm.

In spite of these shortcomings, there might be situations in which the third algorithm is more convenient than the others. Suppose, e.g., that we are to evaluate an infinite series

$$\sum_{m=0}^{\infty} \alpha_m f_m .$$

Not knowing the number of terms required, for given accuracy, one normally accumulates terms until, say, for the first time

$$| \alpha_n f_n | \leqq \epsilon \left| \sum_{m=0}^{n-1} \alpha_m f_m \right| .$$

Since this is equivalent to

$$| \alpha_n p_n | \leqq \epsilon \left| \sum_{m=0}^{n-1} \alpha_m p_m \right| ,$$

we could make use of this condition to terminate (4.13) at the proper time.

We also observe that the third algorithm converges under the sole condition that f_n be minimal; no additional condition, such as (3.16), is required.

5. Bessel functions of the first kind. Bessel functions $J_\alpha(z)$ of the first kind, and Bessel functions $Y_\alpha(z)$ of the second kind, obey the same recurrence relation

$$y_{\alpha+1} - \frac{2\alpha}{z} y_\alpha + y_{\alpha-1} = 0. \tag{5.1}$$

It was the computation of modified Bessel functions $I_n(x)$ that led J. C. P. Miller to invent his backward recurrence algorithm [5, p. xvii].[6] Various authors, since then, observed that this algorithm can be used effectively to generate other Bessel functions as well, including Bessel functions of the second kind ([15], [54], [47], [22], [26], [39 §9.12, Exps. 1 and 7], [2] [32], [33]). To our knowledge the use of ratios of Bessel functions, and thus of a procedure resembling closely our algorithm (3.9), was first suggested by C. W. Jones [27], and is further described in [9], [40], [10]. The ideas involved are extended here in a natural way to Bessel functions of a complex argument. Some new technical details are also included, such as the estimation of the initial value of ν in our first algorithm.

Consider

$$f_n = J_{a+n}(z), \qquad g_n = Y_{a+n}(z), \qquad n = 0, 1, 2, \cdots, \tag{5.2}$$

where $0 \leqq a < 1$, and $z = x + iy$ is a complex number not on the negative real axis. Since $J_{a+n}(\bar{z}) = \overline{J_{a+n}(z)}$, we may assume $y \geqq 0$. As follows directly from (5.1), both functions in (5.2) satisfy the three-term recurrence relation

$$y_{n+1} - \frac{2(a+n)}{z} y_n + y_{n-1} = 0, \qquad n = 1, 2, 3, \cdots. \tag{5.3}$$

However, their asymptotic behavior for large n is quite different. We have, in fact,

$$J_{a+n}(z) \sim \frac{e^{-a}}{\sqrt{2\pi n}} \left(\frac{ez}{2n}\right)^{a+n}, \quad Y_{a+n}(z) \sim -e^a \sqrt{\frac{2}{\pi n}} \left(\frac{2n}{ez}\right)^{a+n}, \quad n \to \infty. \tag{5.4}$$

Therefore, f_n is the minimal solution of (5.3), and the dominance of every other solution over f_n is extremely pronounced: f_n/g_n tends to zero about as rapidly as $|z|^{2n}/(2n)!$, when $n \to \infty$.

It may be noted that this behavior also follows from the general asymptotic results of §2. In fact, the Newton-Puiseux diagram (see Fig. 2) for equation (5.3) has two sides with slopes $+1$ and -1, respectively. Hence, by Theorem 2.3(a), there are two solutions, $y_{n,1}$ and $y_{n,2}$, of (5.3) with different asymptotic behavior, viz.,

$$\frac{y_{n+1,1}}{y_{n,1}} \sim \frac{2n}{z}, \qquad \frac{y_{n+1,2}}{y_{n,2}} \sim \frac{z}{2n}, \qquad n \to \infty.$$

[6] As pointed out by Logan [30], the idea of reversing recurrence schemes to control the propagation of errors can be traced back to Lord Rayleigh, who already recommended that spherical Bessel functions be calculated in the direction of decreasing order [48, p. 38ff].

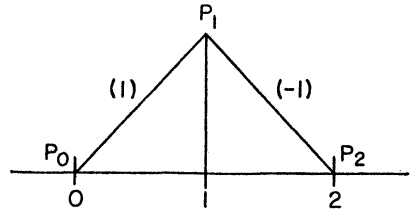


FIG. 2. *Newton-Puiseux diagram for* (5.3)

Since $\lim_{n\to\infty} J_{a+n}(z) = 0$ for any fixed z, we may readily identify $y_{n,2} = J_{a+n}(z)$ and $y_{n,1} = Y_{a+n}(z)$.

In view of the marked predominance of Y_{a+n} over J_{a+n}, it is virtually impossible to generate J_{a+n} directly by means of (5.3). Algorithm (3.9), however, appears to be very effective. In fact, various infinite series of the form (3.2) are available for bypassing the calculation of initial values. Moreover, rather close estimates can be derived for the initial value ν of the recursion index n, thus eliminating the need for many repetitions of the backward recurrence process, as well as the risk of doing too much unnecessary computing.

We first discuss the selection of a suitable infinite series (3.2). We may choose from a family of candidates furnished by Sonine's formula [13, p. 64], which may be written in the form

$$\sum_{m=0}^{\infty} i^m \frac{a+m}{a} C_m^a(\gamma) J_{a+m}(z) = \frac{(z/2)^a e^{i\gamma z}}{\Gamma(1+a)}. \tag{5.5}$$

The parameter γ will presently be specified to suit our purpose; $C_m^a(\gamma)$ are the Gegenbauer polynomials, i.e., the coefficients in the expansion

$$(1 - 2\gamma t + t^2)^{-a} = \sum_{m=0}^{\infty} C_m^a(\gamma) t^m.$$

It is readily seen that

$$\begin{aligned} C_m^a(-\gamma) &= (-1)^m C_m^a(\gamma), \\ C_m^a(1) &= \frac{\Gamma(2a+m)}{m!\Gamma(2a)}, \\ C_{2m-1}^a(0) &= 0, \qquad C_{2m}^a(0) = (-1)^m \frac{\Gamma(a+m)}{m!\Gamma(a)}, \qquad m > 0, \end{aligned} \tag{5.6}$$

while, of course, $C_0^a(\gamma) = 1$.

In accordance with our remark at the end of §3 we should try to select γ in such a way that

$$\frac{s}{f_0} = \frac{(z/2)^a e^{i\gamma z}}{\Gamma(1+a) J_a(z)}$$

cannot become very small in absolute value. Now, if $|z|$ is small, then $J_a(z) \sim (z/2)^a/\Gamma(1+a)$, so that $|s/f_0| \sim 1$. For large $|z|$, we have $J_a(z) \sim (\pi z/2)^{-1/2} \cos(z - a\pi/2 - \pi/2)$, and again, $|s/f_0|$ cannot be small if z is real.

However, if $z = x + iy$, and $y > 0$ is large, then $|\cos(z - a\pi/2 - \pi/2)| \sim e^y/2$, and so

$$\left|\frac{s}{f_0}\right| \sim \frac{2\sqrt{\pi}}{\Gamma(1+a)}\left(\frac{|z|}{2}\right)^{a+1/2} e^{-(1+\gamma)y}.$$

To prevent this from becoming exponentially small, we must require $\gamma \leqq -1$. For convenience, we choose $\gamma = -1$. In view of the first two relations in (5.6), identity (5.5) then becomes

$$\sum_{m=0}^{\infty} (-i)^m \frac{a+m}{a} \frac{\Gamma(2a+m)}{m!\Gamma(2a)} J_{a+m}(z) = \frac{(z/2)^a e^{-iz}}{\Gamma(1+a)},$$

or finally, noting that $a\Gamma(2a) = \Gamma(1+2a)/2$,

$$J_a(z) + 2\sum_{m=1}^{\infty} (-i)^m \frac{(a+m)\Gamma(2a+m)}{m!\Gamma(1+2a)} J_{a+m}(z) = \frac{(z/2)^a e^{-iz}}{\Gamma(1+a)}. \tag{5.7}$$

The coefficients

$$\lambda_m = 2(-i)^m \frac{(a+m)\Gamma(2a+m)}{m!\Gamma(1+2a)}, \qquad m = 1, 2, 3, \cdots, \tag{5.8}$$

are best obtained recursively as follows,

$$\begin{aligned} l_1 &= 1, \\ l_{m+1} &= \frac{m+2a}{m+1} l_m, \qquad m = 1, 2, 3, \cdots, \\ \lambda_m &= 2(-i)^m (a+m) l_m. \end{aligned}$$

In the special case $a = 0$, we simply have $\lambda_m = 2(-i)^m$.

If $z = x$ is real and positive we could choose the real or imaginary part of (5.7) as our normalization identity. We find it more convenient, however, to use (5.5) with $\gamma = 0$. By virtue of the last relations in (5.6), this identity can be written in the form

$$J_a(x) + \sum_{m=1}^{\infty} \frac{(a+2m)\Gamma(a+m)}{m!\Gamma(1+a)} J_{a+2m}(x) = \frac{(x/2)^a}{\Gamma(1+a)}. \tag{5.9}$$

The coefficients

$$\lambda_{2m} = (a+2m)\frac{\Gamma(a+m)}{m!\Gamma(1+a)}, \qquad m = 1, 2, 3, \cdots, \tag{5.10}$$

are obtained recursively by means of

$$\begin{aligned} l_1 &= 1, \\ l_{m+1} &= \frac{m+a}{m+1} l_m, \qquad m = 1, 2, 3, \cdots, \\ \lambda_{2m} &= (a+2m) l_m. \end{aligned}$$

Again, if $a = 0$, the expression for λ_{2m} simplifies to $\lambda_{2m} = 2$. In this case, one could also use the second algorithm in its simplified form (without the s-recursions), if one computes $J_0(x)$ from an appropriate rational approximation. This would probably result in a more efficient algorithm to generate Bessel functions of integer order, than the use of (5.9).

We also note that in the special case of modified Bessel functions

$$I_{a+n}(x) = e^{-i(a+n)\pi/2} J_{a+n}(ix), \qquad x > 0,$$

the recurrence relation (5.3) assumes the form

$$y_{n+1} + \frac{2(a+n)}{x} y_n - y_{n-1} = 0, \qquad n = 1, 2, 3, \cdots,$$

and relation (5.7) the form

$$I_a(x) + 2\sum_{m=1}^{\infty} \frac{(a+m)\Gamma(2a+m)}{m!\Gamma(1+2a)} I_{a+m}(x) = \frac{(x/2)^a e^x}{\Gamma(1+a)}.$$

It is now an easy matter to verify that algorithm (3.9), whether the λ_m be defined by (5.8), or by (5.10), converges as $\nu \to \infty$, provided $J_{a+n}(z) \neq 0$ for $n = 0, 1, 2, \cdots$. By Theorem 3.1 we need only show that

$$h_\nu = \frac{f_{\nu+1}}{g_{\nu+1}} \sum_{m=0}^{\nu} \lambda_m g_m$$

has the limit zero. Now in the case of (5.8), since $0 \leqq a < 1, \Gamma(1+2a) > .88$, we clearly have

$$|\lambda_m| = \frac{2}{\Gamma(1+2a)} \frac{a+m}{m} \frac{\Gamma(2a+m)}{\Gamma(m)}$$

$$< \frac{2}{\Gamma(1+2a)} \frac{m+1}{m} \frac{\Gamma(m+2)}{\Gamma(m)} < 2.3(m+1)^2.$$

Therefore, if ν is already so large that $|g_\nu| \geqq |g_m|$ for $0 \leqq m < \nu$, we shall have

$$|h_\nu| \leqq 2.3(\nu+1)^3 \left|\frac{f_{\nu+1} g_\nu}{g_{\nu+1}}\right| = O(\nu^2 f_{\nu+1}),$$

hence $\lim_{\nu\to\infty} h_\nu = 0$, by virtue of (5.4). A similar argument applies to (5.10).

We proceed now to estimate the initial value of ν to be used in algorithm (3.9), given the number of significant digits desired. Such an estimate may be found from the estimate (3.18) for the relative errors. For definiteness, we assume z complex, and assume identity (5.7) in the role of (3.2).

If ν is large, the infinite series

$$\sum_{m=\nu+1}^{\infty} \lambda_m f_m$$

in (3.18) may roughly be approximated by its first term, $\lambda_{\nu+1} f_{\nu+1}$, and similarly

$$\sum_{m=0}^{\nu} \lambda_m g_m$$

may be approximated by the last term $\lambda_\nu g_\nu$. Then

$$\frac{f_n^{(\nu)} - f_n}{f_n} \doteq \frac{1}{s}\lambda_{\nu+1}f_{\nu+1}\left(1 + \frac{\lambda_\nu}{\lambda_{\nu+1}}\frac{g_\nu}{g_{\nu+1}}\right) - \frac{f_{\nu+1}}{g_{\nu+1}}\frac{g_n}{f_n}$$
$$\doteq \frac{1}{s}\lambda_{\nu+1}f_{\nu+1} - \frac{f_{\nu+1}}{g_{\nu+1}}\frac{g_n}{f_n}. \tag{5.11}$$

Our aim is to find an upper bound for the moduli of these expressions, valid for $n = 0, 1, 2, \cdots, N$. Since $|g_n/f_n|$ ultimately grows rapidly with n, it is plausible to expect that a bound which holds for $n = N$ will also be a valid bound when $n < N$, particularly if N is large. We therefore consider the simplified problem of bounding the modulus of the last member in (5.11), when $n = N$. As a further simplification we assume N, and thus ν, so large that the asymptotic expressions in (5.4) are reasonably accurate. In particular, then, $2N > e|z|$. Under these assumptions a short calculation gives

$$\left|\frac{f_n^{(\nu)} - f_n}{f_n}\right| \lesssim e^{-y}\left(\frac{e|z|}{2\nu}\right)^\nu + \left(\frac{e|z|}{2}\right)^{2(\nu-N)} N^{2N}\nu^{-2\nu}, \qquad y = \operatorname{Im} z,$$

where a few unimportant coefficients have been omitted. For $f_n^{(\nu)}$ to be an approximation of f_n to d significant digits, we are led to require, simultaneously,

$$e^{-y}\left(\frac{e|z|}{2\nu}\right)^\nu \leqq \frac{1}{4}\cdot 10^{-d}, \quad \left(\frac{e|z|}{2}\right)^{2(\nu-N)} N^{2N}\nu^{-2\nu} \leqq \frac{1}{4}\cdot 10^{-d}. \tag{5.12}$$

In the case of real arguments $z = x > 0$, and using relation (5.9) in place of (5.7), our reasoning must be slightly modified, but the conclusion is the same as in (5.12), with $y = 0$.

Now the first inequality in (5.12), after taking logarithms and multiplying by $-2/(e|z|)$, gives

$$\frac{2\nu}{e|z|}\ln\frac{2\nu}{e|z|} \geqq \frac{2(D-y)}{e|z|}, \tag{5.13}$$

where

$$D = d\ln 10 + \ln 4.$$

Similarly, the second inequality gives

$$\nu\ln\frac{2\nu}{e|z|} \geqq N\ln\left(\frac{2N}{e|z|}\right) + \frac{1}{2}D,$$

which may be rewritten in the form

$$\left(\frac{\nu}{N} - 1\right)\ln\left(\frac{2N}{e|z|}\right) + \frac{\nu}{N}\ln\frac{\nu}{N} \geqq \frac{D}{2N}.$$

Since $\nu > N$, and $2N > e|z|$, this is certainly satisfied if we require

$$\frac{\nu}{N}\ln\frac{\nu}{N} \geqq \frac{D}{2N}. \tag{5.14}$$

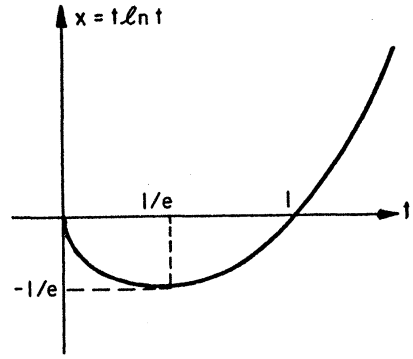


FIG. 3. *Graph of* $x = t \ln t$

Both conditions (5.13) and (5.14) have now the form $t \ln t \geqq c$. Since this is equivalent to $t \geqq t(c)$ with $t(x)$ the inverse function of $x = t \ln t$ in the region $t \geqq 1/e$ (see Fig. 3), our conditions may be given the final form

$$\nu \geqq \frac{e|z|}{2} t\left[\frac{2(D-y)}{e|z|}\right], \quad \text{if} \quad 0 \leqq y < D + \frac{|z|}{2}, \tag{5.15}$$

$$\nu \geqq Nt\left(\frac{D}{2N}\right). \tag{5.16}$$

Low-accuracy approximations to the function $t(x)$ are not hard to obtain. In the interval $1/e \leqq t \leqq 1$ we may first approximate the graph of $t \ln t$ by a quadratic parabola passing through the points $(1/e, -1/e)$, $(1, 0)$, and having zero slope at the first of these points:

$$t \ln t \doteq -\frac{1}{e} + \frac{e}{(e-1)^2}\left(t - \frac{1}{e}\right)^2.$$

Taking then the inverse function of the right-hand member to approximate $t(x)$, we obtain

$$t(x) \doteq \frac{1}{e} + \frac{e-1}{\sqrt{e}}\left(x + \frac{1}{e}\right)^{1/2} \doteq .36788 + 1.0422(x + .36788)^{1/2}, \quad -1/e \leqq x \leqq 0.$$

The accuracy of this approximation is about 4%, or better.

In the region $0 \leqq x \leqq 10$, we truncated the expansion of $t(x)$ in Chebyshev polynomials, having determined the first few expansion coefficients by numerical integration. We so obtained

$$t(x) \doteq 1.0125 + .8577x - .129013x^2 + .0208645x^3 - .00176148x^4 + .000057941x^5,$$

with a maximum percentage error of about 1%.

For larger values of x, we first observe that

$$t(x) \sim x/\ln x, \qquad x \to \infty.$$

In fact, $[t(x) \ln x]/x = (\ln x)/\ln t(x)$, and using the rule of Bernoulli-

L'Hospital, we find

$$\lim_{x\to\infty}\frac{\ln x}{\ln t(x)} = \lim_{x\to\infty}\frac{\frac{1}{x}}{\frac{1}{t(x)}\cdot\frac{1}{1+\ln t(x)}} = \lim_{x\to\infty}\frac{x+t(x)}{x} = 1.$$

Unfortunately, the asymptotic expression so obtained does not give sufficient accuracy, unless x is very large. Applying, however, one step of Newton's method to the equation $t \ln t = x$, with $x/\ln x$ as initial approximation, we get

$$t(x) \doteq \frac{x}{\ln x}\,\frac{1}{1-\frac{\ln\ln x}{1+\ln x}}.$$

This approximation now appears to be in error by less than 1% for $x \geqq 2$. As $x \to \infty$, the relative error clearly tends to zero.

An alternate method of selecting ν in the case $z = x > 0$, $a = 0$, was derived by W. Kahan [28], using Olver's error analysis [38]. Let ϵ be the largest relative error tolerated in the final results, $J_0(z), J_1(x), \cdots, J_N(x)$. Let K be the integer

$$K = \max(N, [x]),$$

and, with $\beta > 0$ arbitrary (though small, in practice), define

$$y_K = 0, \qquad y_{K+1} = \beta,$$

$$y_{n+1} = \frac{2n}{x}y_n - y_{n-1}, \qquad n = K+1, K+2, \cdots.$$

Then ν may be taken to be the smallest n for which $y_{n+1} \geqq \beta/\epsilon$.

We have seen that Bessel functions $J_{a+n}(z)$ of positive orders can be computed entirely from their recurrence relation. This remains true, to a certain extent, for Bessel functions

$$y_n = J_{a-n}(z), \qquad n = 1, 2, 3, \cdots; \quad 0 < a < 1, \tag{5.17}$$

of negative orders. They satisfy the recurrence relation

$$y_{n+1} + \frac{2(n-a)}{z}y_n + y_{n-1} = 0, \qquad n = 2, 3, 4, \cdots, \tag{5.18}$$

which has the same Newton-Puiseux diagram as (5.3). The solution (5.17), however, is now a dominant solution, the minimal solution being $f_n = (-1)^n J_{n-a}(z)$. It appears therefore safe to generate $J_{a-n}(z)$ by means of (5.18) in the ordinary fashion. Moreover, the recursion may be started with $n = 0$, and the initial values $y_{-1} = J_{a+1}(z)$, $y_0 = J_a(z)$ obtained by the methods previously discussed.

The assumption $a > 0$ is of course essential. If $a = 0$, the two solutions y_n and f_n above are the same (minimal) solution of (5.18), and forward recursion by (5.18) is doomed to fail. The same must be expected if a is close to zero, and indeed if a is close to one.

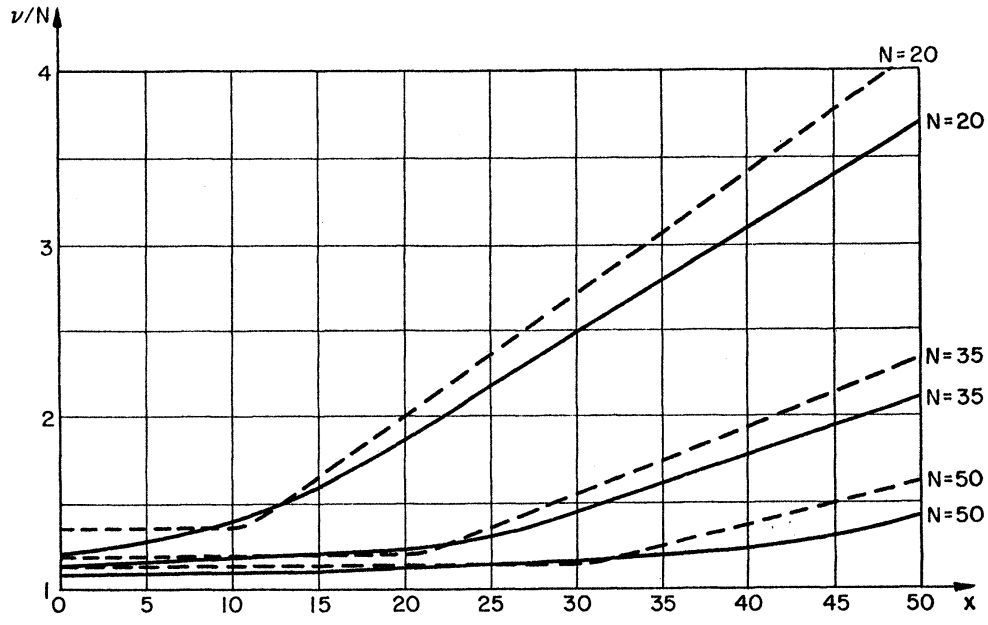


FIG. 4. ESTIMATED ν/N / EMPIRICAL ν/N FOR BESSEL FUNCTIONS $J_n(x)$, $n = 0(1)N$.

We present now a few numerical results concerning the first algorithm for computing $J_{a+n}(z)$. The performance of this algorithm was found to be quite insensitive to changes of a in the interval $0 \leqq a < 1$, so that the results given for $a = 0$ may be considered as representative.

Our main concern was to determine the quality of the estimate of ν given above in (5.15), (5.16). We compared this estimate with the smallest value of ν empirically observed to yield $J_n(z)$, $n = 0(1)N$, to six significant digits.[7] For real $z = x$, the results are shown in Fig. 4, while for complex $z = re^{i\phi}$ they are depicted in Fig. 5. Both figures show that agreement between estimated and actual ν is rather satisfactory on the whole, even though for larger values of $|z|$ it is worsening. Remarkable is also the relative smallness of ν/N over an extended region of the complex plane.

ALGOL procedures based on the methods of this paragraph may be found in [18].

6. Legendre functions. A further class of special functions amenable to the methods of §§3 and 4 are the associated Legendre functions of the first and second kind, $P_\alpha^m(z)$ and $Q_\alpha^m(z)$. We assume that m is a nonnegative[8] integer, z a complex number outside the interval $(0, 1)$, with $\mathrm{Re}\, z > 0$, and α arbitrary real or

[7] More precisely, algorithm (3.9) was run with $\nu = N + 2, N + 4, N + 6, \cdots$ until for the first time the $N + 1$ values $f_n^{(\nu)}$, $n = 0, 1, \cdots, N$, agreed to six significant digits with the respective values of $f_n^{(\nu-2)}$.

[8] If α is an integer $\geqq m$, or nonintegral, then $P_\alpha^{-m}(z) = [\Gamma(\alpha - m + 1)/\Gamma(\alpha + m + 1)]P_\alpha^m(z)$, and the restriction to nonnegative integers m is not essential. Similarly, $Q_\alpha^{-m}(z) = [\Gamma(\alpha - m + 1)/\Gamma(\alpha + m + 1)]Q_\alpha^m(z)$.

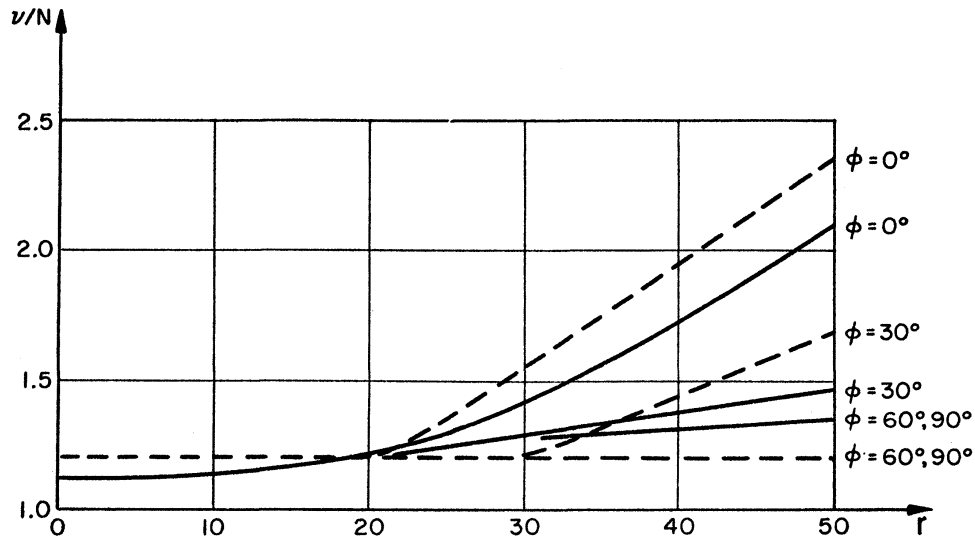


FIG. 5. ESTIMATED ν/N, EMPIRICAL ν/N } FOR BESSEL FUNCTIONS $J_n(re^{i\phi})$, n = 0(1)N, WHERE N = 35.

complex, but $\alpha \neq -1, -2, -3, \cdots$. The Legendre functions of the first kind are then representable by a definite integral,

$$P_\alpha^m(z) = \frac{\Gamma(\alpha + m + 1)}{\pi\Gamma(\alpha + 1)} \int_0^\pi [z + (z^2 - 1)^{1/2} \cos t]^\alpha \cos mt \, dt.$$

A similar representation holds for Legendre functions of the second kind,

$$Q_\alpha^m(z) = (-1)^m \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha - m + 1)} \int_0^\infty \frac{\cosh mt}{[z + (z^2 - 1)^{1/2} \cosh t]^{\alpha+1}} \, dt,$$

provided Re $(\alpha - m) > -1$. In both these formulas the meaning of the expressions $(z - 1)^{1/2}$, $(z + 1)^{1/2}$ is as usual obtained by continuity in the complex plane, cut along the interval $(-\infty, 1)$, assuming them real for $z > 1$. A similar remark applies to the other fractional powers.

It is well known that P_α^m and Q_α^m satisfy identical three-term recurrence relations, both with respect to order m and degree α. (See, e.g., [12, p. 160].) The fact that backward recursion techniques are applicable to obtain Legendre functions of integral order and argument greater than unity was already mentioned in [10]. The use of Miller's algorithm (cf. §3) for calculating toroidal functions of the second kind is described in [50]. No mention is made, in this reference, of the usefulness of infinite series for normalization purposes, which makes this algorithm even more attractive.

We begin with considering the recurrence with respect to order m. Both $P_\alpha^m(z)$ and $Q_\alpha^m(z)$, as functions of m, are solutions of

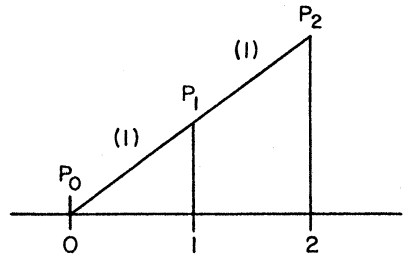


FIG. 6. *Newton-Puiseux diagram for* (6.1)

$$y_{m+1} + \frac{2mz}{(z^2 - 1)^{1/2}} y_m + (m + \alpha)(m - \alpha - 1)y_{m-1} = 0, \tag{6.1}$$

$$m = 1, 2, 3, \cdots.$$

We first assume that α is not an integer. The case of integral α will be dealt with later.

The Newton-Puiseux diagram (see Fig. 6) for the difference equation (6.1) is a straight line segment with slope 1, and thus case (b) of Theorem 2.3 applies. The characteristic equation is

$$t^2 + \frac{2z}{(z^2 - 1)^{1/2}} t + 1 = 0,$$

which has the roots

$$t_1 = -\left(\frac{z + 1}{z - 1}\right)^{1/2}, \qquad t_2 = t_1^{-1}.$$

Since $\operatorname{Re} z > 0$, it is readily seen that

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

By Theorem 2.3, and the remarks following it, the difference equation (6.1) thus possesses a minimal solution, $y_{m,2}$, for which

$$\lim_{m\to\infty} \frac{y_{m+1,2}}{m y_{m,2}} = t_2;$$

for any other solution the corresponding limit is t_1. Let

$$f_m = \frac{P_\alpha^m(z)}{\Gamma(\alpha + m + 1)} = \frac{1}{\pi\Gamma(\alpha + 1)} \int_0^\pi [z + (z^2 - 1)^{1/2} \cos t]^\alpha \cos mt \, dt, \tag{6.2}$$

so that

$$\frac{f_{m+1}}{f_m} \sim \frac{P_\alpha^{m+1}(z)}{m P_\alpha^m(z)}, \qquad m \to \infty.$$

The second member of this relation, as was just observed, has a finite limit as

$m \to \infty$, which is either t_1 or t_2. Were it t_1, then $|f_m|$ would tend to ∞, since $|t_1| > 1$. This, however, is impossible, since f_m by (6.2) are essentially the Fourier coefficients of a smooth function, and thus $\lim_{m\to\infty} f_m = 0$. Therefore, the limit is t_2, and $P_\alpha^m(z)$ is indeed the minimal solution of (6.1), while $Q_\alpha^m(z)$ belongs among the dominant solutions.

It follows that $P_\alpha^m(z)$, $m = 0, 1, 2, \cdots$, can be obtained by the algorithms of §§3 and 4. As will be seen shortly, an infinite series can be used for normalization, so that no values of $P_\alpha^m(z)$ need to be known in advance. The functions $Q_\alpha^m(z)$, $m = 0, 1, 2, \cdots$, on the other hand, can safely be generated by forward use of (6.1); this requires two initial values for $m = 0$ and $m = 1$ to be available. In the important special case $\alpha = -\frac{1}{2} + n$, where n is an integer, these initial values may also be obtained by the aforementioned algorithms, applied to the recurrence with respect to degree (cf. below).

It is more convenient, computationally, to deal with f_m defined in (6.2), rather than P_α^m, since then we not only avoid excessively large numbers, but also obtain a very simple identity for normalization. It is well known, indeed, that (see [12, p. 166])[9]

$$P_\alpha(z) + 2\sum_{m=1}^{\infty} \frac{\Gamma(\alpha+1)}{\Gamma(\alpha+m+1)} P_\alpha^m(z) = [z + (z^2-1)^{1/2}]^\alpha, \tag{6.3}$$

valid for $\operatorname{Re} z > 0$ and arbitrary α. Hence,

$$f_0 + 2\sum_{m=1}^{\infty} f_m = \frac{[z + (z^2-1)^{1/2}]^\alpha}{\Gamma(\alpha+1)}, \tag{6.4}$$

which may serve in the capacity of condition (3.2), with

$$s = \frac{[z + (z^2-1)^{1/2}]^\alpha}{\Gamma(\alpha+1)}, \qquad \lambda_0 = 1, \lambda_m = 2, m > 0.$$

The convergence of the first algorithm then follows from the remark made after Theorem 3.1.

To insure numerical stability, the ratio

$$\frac{s}{f_0} = \frac{[z + (z^2-1)^{1/2}]^\alpha}{P_\alpha(z)} \tag{6.5}$$

should not be allowed to become excessively small (cf. §3). While it is difficult to check the magnitude of this function for the full range of z and α, we shall at least look into the behavior of this function near the singular points $z = -1$, $z = +1$, $z = \infty$.

As z tends to $+1$, or -1, in the plane cut from $-\infty$ to 1, we have $P_\alpha(z) \to 1$, and so $|s/f_0| \to 1$.

To study the behavior at infinity, we make use of the following facts (see [49, §54]): If $\alpha \neq -\frac{1}{2} + n$, where n is an integer, we have, as $z \to \infty$,

$$P_\alpha(z) \sim A_\alpha(2z)^{-(\alpha+1)} + B_\alpha(2z)^\alpha,$$

[9] As is customary, we write $P_\alpha(z)$ for $P_\alpha^0(z)$.

where

$$A_\alpha = \frac{\Gamma(-\alpha - \frac{1}{2})}{\sqrt{\pi}\Gamma(-\alpha)}, \qquad B_\alpha = \frac{\Gamma(\alpha + \frac{1}{2})}{\sqrt{\pi}\Gamma(\alpha + 1)}.$$

Otherwise, when $\alpha = -\frac{1}{2} + n$, then

$$P_{-(1/2)+n}(z) \sim \begin{cases} \dfrac{\sqrt{2}}{\pi} z^{-1/2} \ln z, & \text{if } n = 0, \\ \dfrac{1}{\sqrt{\pi}} \dfrac{\Gamma(|n|)}{\Gamma(|n| + \frac{1}{2})} (2z)^{|n|-1/2}, & \text{if } n \neq 0. \end{cases}$$

Hence, in the former case,

$$\frac{s}{f_0} \sim \frac{(2z)^\alpha}{A_\alpha(2z)^{-(\alpha+1)} + B_\alpha(2z)^\alpha} = \frac{1}{A_\alpha(2z)^{-(2\alpha+1)} + B_\alpha},$$

which becomes small in modulus only if $\operatorname{Re}(2\alpha + 1) < 0$, i.e., $\operatorname{Re}\alpha < -\frac{1}{2}$. In the case $\alpha = -\frac{1}{2} + n$, we have

$$\frac{s}{f_0} \sim \begin{cases} \dfrac{\pi}{2 \ln z}, & \text{if } n = 0, \\ \sqrt{\pi} \dfrac{\Gamma(n + \frac{1}{2})}{\Gamma(n)}, & \text{if } n > 0, \\ \sqrt{\pi} \dfrac{\Gamma(|n| + \frac{1}{2})}{\Gamma(|n|)} (2z)^{-2|n|}, & \text{if } n < 0. \end{cases}$$

Here, the third case ($n < 0$) is critical, and also the first, but to a much lesser degree.

For all practical purposes, then, (6.5) will be small in modulus only if $\operatorname{Re}\alpha < -\frac{1}{2}$. This can easily be avoided by employing the relation

$$P_\alpha^m(z) = P_{-\alpha-1}^m(z), \tag{6.6}$$

when necessary. If $\operatorname{Re}\alpha < -\frac{1}{2}$, then indeed $\operatorname{Re}(-\alpha - 1) > -\frac{1}{2}$.

Restricting α to have real part $-\frac{1}{2}$ one obtains *Mehler's conical functions* $P_{-(1/2)+i\tau}^m(z)$, where τ is real. Since $P_{-(1/2)+i\tau}^m(z) = P_{-(1/2)-i\tau}^m(z)$, by (6.6), these functions are real when z is real. It suffices, moreover, to consider nonnegative values of τ. We shall assume $z = x > 1$, which is a case of practical interest.

Since $\Gamma(\alpha + m + 1)$ is now complex, the scaled functions (6.2) used previously are not as convenient anymore. To maintain the computational advantages noted before, we consider

$$f_m = \frac{1}{m!} P_{-(1/2)+i\tau}^m(x). \tag{6.7}$$

As follows from (6.1) and our previous discussion, f_m so defined is a minimal solution of

$$y_{m+1} + \frac{2mx}{(m+1)(x^2-1)^{1/2}} y_m + \frac{(m - \frac{1}{2})^2 + \tau^2}{m(m+1)} y_{m-1} = 0, \tag{6.8}$$

$$m = 1, 2, 3, \cdots.$$

To arrive at a normalizing identity for f_m, involving real quantities only, we write down (6.3) (with $z = x$) once for $\alpha = -\frac{1}{2} + i\tau$, and once for $\alpha = -\frac{1}{2} - i\tau$, and then form the arithmetic mean of the two identities. Noting (6.7), we then obtain

$$f_0 + \sum_{m=1}^{\infty} \lambda_m f_m = [x + (x^2 - 1)^{1/2}]^{-1/2} \cos(\tau \ln [x + (x^2 - 1)^{1/2}]),$$

where

$$\lambda_m = u_m + \bar{u}_m, \qquad u_m = \frac{m!\Gamma(\frac{1}{2} + i\tau)}{\Gamma(\frac{1}{2} + i\tau + m)}.$$

The λ's are best obtained from a three-term recurrence relation. We clearly have

$$u_{m+1} = \frac{(m+1)u_m}{m + \frac{1}{2} + i\tau} = \frac{(m+1)(m + \frac{1}{2} - i\tau)}{(m + \frac{1}{2})^2 + \tau^2} u_m.$$

For notational simplicity, let

$$\alpha_m = m + \frac{1}{2}, \qquad \beta_m = \frac{(m + \frac{1}{2})^2 + \tau^2}{m + 1}. \tag{6.9}$$

Then

$$\beta_m u_{m+1} = (\alpha_m - i\tau)u_m,$$
$$\beta_m \bar{u}_{m+1} = (\alpha_m + i\tau)\bar{u}_m.$$

Adding, and subtracting, we get

$$\begin{aligned} \beta_m \lambda_{m+1} &= \alpha_m \lambda_m - \tau \mu_m, \\ \beta_m \mu_{m+1} &= \alpha_m \mu_m + \tau \lambda_m, \end{aligned} \tag{6.10}$$

where $\mu_m = i(u_m - \bar{u}_m)$. Eliminating the μ's, we find

$$\lambda_{m+1} - \frac{\alpha_{m-1} + \alpha_m}{\beta_m} \lambda_m + \frac{\alpha_{m-1}^2 + \tau^2}{\beta_{m-1}\beta_m} \lambda_{m-1} = 0,$$

or, with the values (6.9) inserted,

$$\lambda_{m+1} - \frac{2m(m+1)}{(m + \frac{1}{2})^2 + \tau^2} \lambda_m + \frac{m(m+1)}{(m + \frac{1}{2})^2 + \tau^2} \lambda_{m-1} = 0, \quad m = 2, 3, \cdots. \tag{6.11}$$

The initial values are

$$\lambda_1 = \frac{1}{\frac{1}{4} + \tau^2}, \qquad \lambda_2 = \frac{3 - 4\tau^2}{(\frac{1}{4} + \tau^2)(\frac{9}{4} + \tau^2)}. \tag{6.12}$$

We observe that the recursion (6.11) belongs to case (b) of Theorem 2.3, the characteristic equation being $(t-1)^2 = 0$. Because of the double root $t_1 = t_2 = 1$, Theorem 2.3 does not permit us to decide whether the recursion in (6.11), (6.12) is numerically stable. We observe, however, that another solution of (6.11) is μ_m, as follows by eliminating the λ's in (6.10). Therefore, Re u_m and Im u_m are a pair

of linearly independent solutions of (6.11). Using Stirling's formula, and disregarding constant factors, we find

$$u_m \sim m e^{-i r \ln m}, \qquad m \to \infty,$$

so that both solutions oscillate, for large m, with linearly increasing amplitudes. Therefore, numerical instability cannot arise.

A further interesting special case is obtained by assuming α a nonnegative integer, $\alpha = p$. Then, in fact,

$$P_p^m(z) = \frac{(z^2-1)^{m/2}}{2^p p!} \frac{d^{p+m}}{dz^{p+m}} (z^2-1)^p.$$

This shows that

$$P_p^m(z) \equiv 0 \quad \text{if} \quad m > p.$$

We note that Theorem 1.1 with $f_m = P_p^m(z)$ is no longer applicable, since the assumption $f_m \neq 0$ (all m) does not hold. Neither apply the asymptotic results of §2, the assumption (2.2) now being violated.

Nevertheless, f_m still satisfies the recurrence relation (6.1) (with $\alpha = p$) for *all* values of m, thus in particular for $m = p, p-1, \cdots, 1$, whereby $f_{p+1} = 0$. The algorithm described at the beginning of §3 becomes applicable, and it follows that the r-recursion in our algorithm (3.9), if started with $\nu = p$, furnishes the *exact* ratios $r_{m-1} = f_m/f_{m-1}$, apart from rounding errors. The same is true for the s-recursion, which yields exact values of

$$s_{m-1} = \sum_{r=m}^{p} \lambda_r f_r / f_{m-1},$$

the infinite series in (6.3) reducing to a finite sum, when $\alpha = p$. In short, (3.9) *with $\nu = p$ now represents the complete algorithm for computing $f_m = P_p^m(z)$, $m = 0, 1, 2, \cdots, p$, and no iteration on ν is required.*

We now proceed to the recurrence relation with respect to degree. Let a, m, and z be fixed, and consider $P_{a+n}^m(z)$, $Q_{a+n}^m(z)$ as functions of n. They both obey the relation

$$\begin{aligned}(n+a-m+1)y_{n+1} - (2n+2a+1)zy_n \\ + (n+a+m)y_{n-1} = 0, \quad n = 0, 1, 2, \cdots.\end{aligned} \tag{6.13}$$

This is a Poincaré difference equation whose characteristic equation is

$$t^2 - 2zt + 1 = 0.$$

The roots are

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

and it is readily verified that for $\operatorname{Re} z > 0$,

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

From Theorem 2.3(b), and the remarks following this theorem, we conclude that (6.13) has a minimal solution f_n for which $\lim_{n\to\infty} f_{n+1}/f_n = t_2$, while the limit is t_1 for every other solution. Now it is known (see, e.g., [12, p. 162]) that

$$Q_{a+n}^m(z) \sim (-1)^m \sqrt{\frac{\pi}{2}}\, n^{m-1/2}(z^2-1)^{-1/4} t_2^{a+n+1/2}, \qquad n \to \infty,$$

for z outside the cut from $-\infty$ to 1, thus in particular for those z which we are considering here. It follows immediately, therefore, that the minimal solution is $f_n = Q_{a+n}^m(z)$, and that $g_n = P_{a+n}^m(z)$ is now a dominant solution.

The computation of $P_{a+n}^m(z)$ for $n = 0, 1, 2, \cdots$ can proceed using (6.13) in the normal fashion. The required initial values $P_a{}^m(z)$, $P_{a+1}^m(z)$ may be obtained by the methods discussed above. These functions are thus again computable entirely from their recurrence relations. On the other hand, $Q_{a+n}^m(z)$, as the minimal solution of (6.13), is amenable to the algorithms of §§3 and 4.

Unfortunately, no simple infinite series involving the $f_n = Q_{a+n}^m(z)$ for arbitrary a exists, which would be convergent in the region considered here. Normalization of f_n, therefore, has to be accomplished by computing the initial value $Q_a{}^m(z)$. In the special case of *toroidal functions* $Q_{-(1/2)+n}^m(z)$, however, we have the following relation [12, p. 166]:

$$Q_{-1/2}^m(z) + 2\sum_{n=1}^{\infty} Q_{-(1/2)+n}^m(z) = (-1)^m \sqrt{\frac{\pi}{2}}\,\Gamma\left(m+\frac{1}{2}\right)(z-1)^{-1/2}\left(\frac{z+1}{z-1}\right)^{m/2},$$

which lends itself well for normalization, unless z is complex and near the singular point -1.

We wish now to give some additional numerical information concerning the algorithms described in this paragraph.

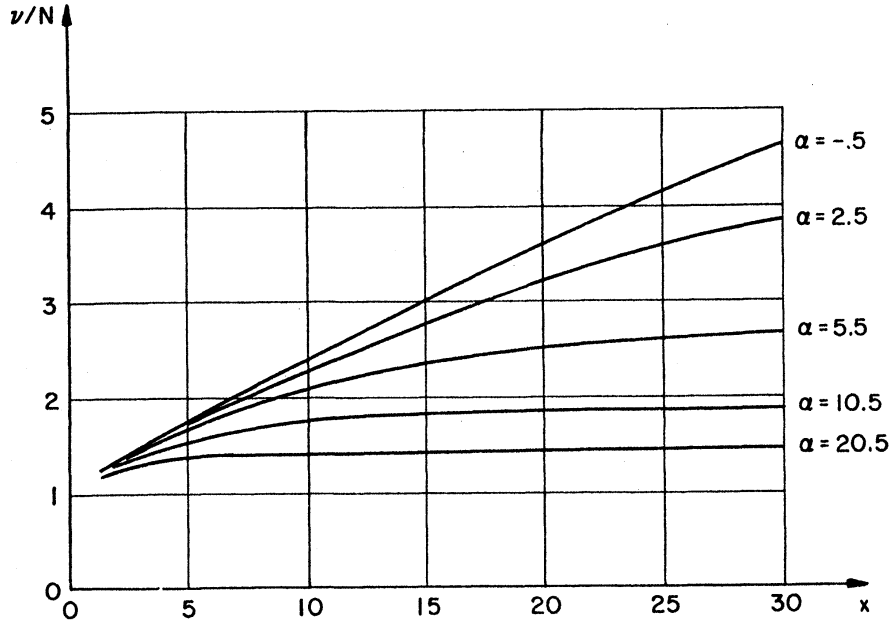


FIG. 7. *Empirical ν/N for Legendre functions $P_a{}^n(x)$, $n = 0(1)N$, where $N = 50$*

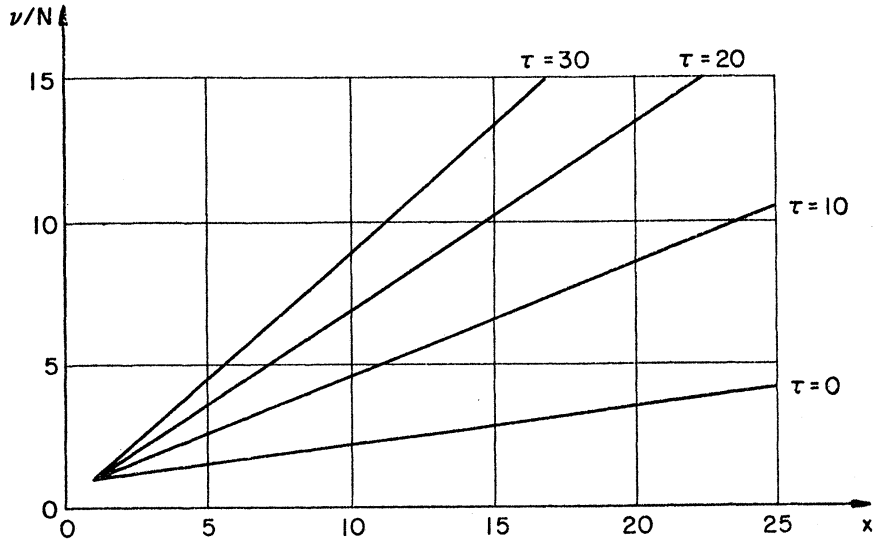


FIG. 8. *Empirical ν/N for conical functions $P^n_{-(1/2)+i\tau}(x)$, $n = 0(1)N$, where $N = 50$*

Of foremost interest is again the determination of ν/N in our first algorithm. A derivation of an estimate by analytical means appears to be out of question. We tried, therefore, to determine the behavior of ν/N empirically, as a function of the various parameters involved. To simplify the task, we assumed a fixed accuracy requirement of six significant digits. Moreover, we decided to consider a fixed value of N. Since ν/N was found to decrease with N, we deemed it desirable to select a relatively large value of N as representative, namely, $N = 50$. If we would not do so, we would considerably overestimate ν/N, and pay heavily for this in cases where N is actually large. To compensate for a possible underestimation in cases where N is small, we suggest that a relatively large increment of ν, say 10, or even 20, be used in the iteration process of the first algorithm. Having thus disposed of two parameters, we are still left with two in each case.

In the case of Legendre functions $f_n = P_\alpha{}^n(x)/\Gamma(\alpha + n + 1)$, where $x > 1$, $\alpha \geqq -\frac{1}{2}$, the value of ν/N found empirically for $N = 50$ is depicted in Fig. 7 as a function of x and α. A reasonably good approximation to these curves was obtained in the form

$$\frac{\nu}{N} \doteq \frac{37.26 + .1283(\alpha + 38.26)x}{37.26 + .1283(\alpha + 1)x}.$$

For the conical functions $P^n_{-(1/2)+i\tau}(x)/n!$, where $x > 1$, $\tau \geqq 0$, the empirical value of ν/N as a function of x and τ is shown in Fig. 8. The curves were fitted by a function which is linear in both x and τ, viz.,

$$\frac{\nu}{N} \doteq 1 + (.140 + .0246\tau)(x - 1).$$

As the graphs in Fig. 8 show, the conical functions are by far the hardest to compute. As ν/N becomes large, considerable accumulation of rounding errors must be expected.

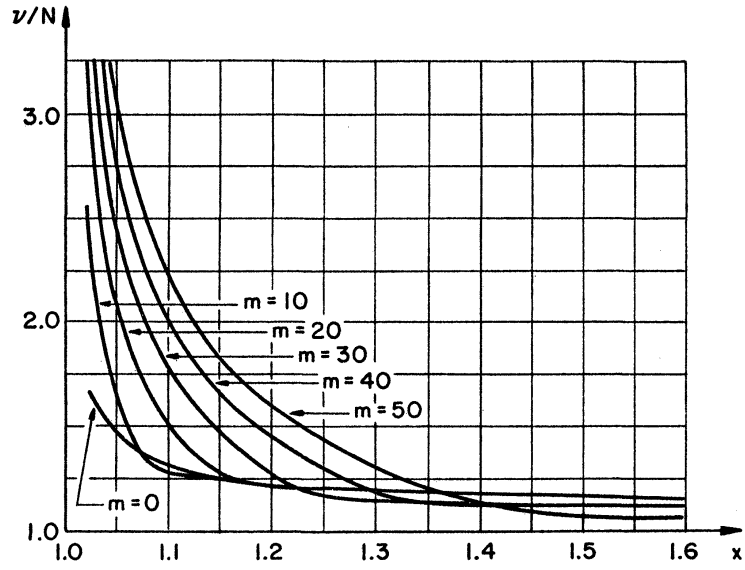


FIG. 9. *Empirical ν/N for toroidal functions $Q^m_{-(1/2)+n}(x)$, $n = 0(1)N$, where $N = 50$*

Finally, in the case of toroidal functions $Q^m_{-(1/2)+n}(x)$, where $x > 1$, $m \geqq 0$, the behavior of ν/N as a function of x and m is shown in Fig. 9, and is roughly approximated by

$$\frac{\nu}{N} \doteq 1.15 + \frac{.0146 + .00122m}{x - 1}.$$

ALGOL procedures based on the methods of this paragraph are available in [19].

7. Coulomb wave functions. Coulomb wave functions are of importance in the study of nuclear interactions. They arise when Schrödinger's equation for a charged particle in the Coulomb field of a fixed charge is separated in polar coordinates. The radial component then satisfies the differential equation

$$\frac{d^2y}{d\rho^2} + \left[1 - \frac{2\eta}{\rho} - \frac{L(L+1)}{\rho^2}\right] y = 0, \tag{7.1}$$

where η is a real parameter, L a nonnegative integer, and $\rho > 0$. Physically, η depends on the relative charges. If both are of equal sign, then $\eta > 0$, otherwise, $\eta < 0$. The variable ρ is a radial distance, suitably scaled, while L is the orbital angular-momentum quantum number of the particle.

The origin $\rho = 0$ is a regular singular point of (7.1), with indicial equation

$$\lambda(\lambda - 1) = L(L + 1).$$

Since the roots of this equation are $\lambda_1 = L + 1$, $\lambda_2 = -L$, the differential equation (7.1) has a solution corresponding to λ_1 which is regular at $\rho = 0$, admitting an expansion of the form

$$y_1(\rho) = \rho^{L+1} \sum_{n=0}^{\infty} c_n \rho^n.$$

In quantum mechanics it is customary to normalize this solution to have sinusoidal behavior as $\rho \to \infty$, with amplitude equal to 1. So normalized, the solution is called the *regular Coulomb wave function*, and denoted by $F_L(\eta, \rho)$. The solution corresponding to λ_2, on the other hand, will contain a logarithmic term, since λ_1 differs from λ_2 by a positive integer. If normalized similarly as F_L, it is called the *irregular Coulomb wave function*, and denoted by $G_L(\eta, \rho)$.

The line

$$\rho = 2\eta, \qquad \eta > 0,$$

which separates regions of different asymptotic behavior of the solutions of (7.1) as $\rho \to \infty$ and $\eta \to \infty$, is called the *transition line*.

In terms of Whittaker's function $M_{\kappa,\mu}(z)$ (see [12] for notation), we have

$$F_L(\eta, \rho) = (2i)^{-(L+1)} C_L(\eta) M_{i\eta, L+1/2}(2i\rho), \tag{7.2}$$

where

$$C_L(\eta) = \frac{2^L e^{-\pi\eta/2} \, |\, \Gamma(L + 1 + i\eta)|}{(2L + 1)!}. \tag{7.3}$$

We note for later use,

$$C_0(\eta) = \left(\frac{2\pi\eta}{e^{2\pi\eta} - 1}\right)^{1/2}, \qquad C_L(\eta) = \frac{(L^2 + \eta^2)^{1/2}}{L(2L + 1)} C_{L-1}(\eta), \tag{7.4}$$

$$L = 1, 2, 3, \cdots.$$

As functions of L, both the regular and irregular Coulomb wave function satisfy the three-term recurrence relation

$$L[(L + 1)^2 + \eta^2]^{1/2} y_{L+1} - (2L + 1)\left[\eta + \frac{L(L + 1)}{\rho}\right] y_L + (L + 1)[L^2 + \eta^2]^{1/2} y_{L-1} = 0, \qquad L = 1, 2, 3, \cdots. \tag{7.5}$$

This difference equation has the same Newton-Puiseux diagram as the recurrence relation for the Bessel functions (see Fig. 2). Hence, there are two solutions of (7.5) with markedly distinct asymptotic properties as $L \to \infty$. These, in fact, are precisely the regular and irregular Coulomb wave functions, since for fixed η and ρ, it is known that

$$F_L(\eta, \rho) \sim C_L(\eta) \rho^{L+1}, \qquad G_L(\eta, \rho) \sim \frac{1}{2LC_L(\eta)\rho^L}, \qquad L \to \infty, \tag{7.6}$$

and furthermore,

$$C_L(\eta) \sim \frac{1}{e\sqrt{2}} e^{-\pi\eta/2} \left(\frac{e}{2L}\right)^{L+1}, \qquad L \to \infty. \tag{7.7}$$

In particular, F_L is the minimal solution of (7.5). Therefore, F_L may be obtained by the algorithm in (3.9), provided a suitable infinite series can be found for normalization. The proper selection of this series is a rather crucial matter, and will be discussed next.

For computational convenience we first let

$$f_L = \frac{2^L L!}{(2L)!C_L(\eta)} F_L(\eta, \rho). \tag{7.8}$$

Among other things (relatively slow rate of growth of the coefficients λ_L in (7.11) below), this effectively removes square roots in (7.5). In fact, using (7.4), one finds that f_L is the minimal solution of

$$\frac{L[(L+1)^2+\eta^2]}{(L+1)(2L+3)} y_{L+1} - \left[\eta + \frac{L(L+1)}{\rho}\right] y_L + \frac{L(L+1)}{2L-1} y_{L-1} = 0. \tag{7.9}$$

The following expansion is known (see, e.g., [6, p. 131, formula (16β)]),

$$z^{(1+\mu)/2} e^{\alpha z/2} = \sum_{n=0}^{\infty} \frac{\Gamma(\mu+n)}{\Gamma(\mu+2n)} P_n^{((\mu-1)/2+\kappa,(\mu-1)/2-\kappa)}(\alpha) M_{\kappa,\mu/2+n}(z), \tag{7.10}$$

where $P_n^{(\alpha,\beta)}(z)$ is the Jacobi polynomial of degree n. (For notation, see [55].) Letting $\mu = 1$, $\kappa = i\eta$, $z = 2i\rho$, $\alpha = -i\omega$, and writing L for n, this becomes in view of (7.2), (7.8),

$$\rho e^{\omega\rho} = \sum_{L=0}^{\infty} \lambda_L f_L, \qquad \lambda_L = i^L P_L^{(i\eta,-i\eta)}(-i\omega). \tag{7.11}$$

If $\omega = i$, then one easily shows that (7.11) reduces to a result attributed in [53] to P. Henrici. As one of several alternatives, it was suggested in this reference to apply Miller's backward recurrence algorithm to (7.9), using Henrici's series for normalization. Unfortunately, the process suffers from severe loss of accuracy when η and ρ are positive and large. We show that by selecting ω judiciously, the loss of accuracy can be kept under control.

We recall (cf. the end of §3) that loss of accuracy due to cancellation occurs if

$$\frac{s}{\lambda_0 f_0} = \frac{\rho e^{\omega\rho}}{f_0} \tag{7.12}$$

is very small in absolute value. Let

$$\tau = \frac{\rho}{2\eta},$$

so that the point (η, ρ) is above or below the transition line, depending on whether $\tau > 1$ or $0 < \tau < 1$, respectively, and $\eta < 0$ if $-\infty < \tau < 0$. In each of these three cases, f_0 will behave differently as $|\eta| \to \infty$ and τ is held fixed. In fact, using general asymptotic results for Whittaker functions due to Buchholz (see [6, p. 101 ff, formulae (7), (11), (16a)]), one obtains from (7.2), after some computation,[10]

$$f_0 \sim \frac{1}{\sqrt{2\pi\eta}} \left(\frac{\tau}{\tau-1}\right)^{1/4} e^{\pi\eta} \cos\left\{2\eta[\sqrt{\tau(\tau-1)} - \ln(\sqrt{\tau} + \sqrt{\tau-1})] - \frac{\pi}{4}\right\},$$

$$\tau > 1, \eta \to \infty,$$

[10] In the cited formula (7) of [6], the factor $\exp(\mp\kappa i(\kappa - (1+\mu)/2))$ should read $\exp(\mp\pi i(\kappa - (1+\mu)/2))$.

$$f_0 \sim \frac{1}{2\sqrt{2\pi\eta}}\left(\frac{\tau}{1-\tau}\right)^{1/4} \exp\{\eta[\pi - 2\arccos\sqrt{\tau} + 2\sqrt{\tau(1-\tau)}]\},$$

$$0 < \tau < 1, \eta \to \infty,$$

$$f_0 \sim \frac{1}{\sqrt{2\pi|\eta|}}\left(\frac{|\tau|}{|\tau|+1}\right)^{1/4} \sin\left\{2|\eta|[\sqrt{|\tau|(|\tau|+1)} + \ln(\sqrt{|\tau|} + \sqrt{|\tau|+1})] - \frac{\pi}{4}\right\}, \qquad -\infty < \tau < 0, \eta \to -\infty.$$

To prevent the quantity in (7.12) from becoming exponentially small, as $\eta \to \infty$, we are led to choose

$$\omega \geqq \begin{cases} \dfrac{\pi}{2\tau}, & \tau \geqq 1, \\ \dfrac{1}{2\tau}[\pi - 2\arccos\sqrt{\tau} + 2\sqrt{\tau(1-\tau)}], & 0 < \tau < 1, \\ 0, & \tau < 0. \end{cases} \tag{7.13}$$

Since for reasons which become clear later, small values of ω are to be preferred, equality in (7.13) is suggested. The parameter ω so defined then depends continuously on τ in the interval $(0, \infty)$, and decreases monotonically from ∞ to 0. Clearly, as long as η is small, say <1, the choice $\omega = 0$ is entirely satisfactory.

Other series expansions obtained by letting $\mu = 2L_0 + 1$, $\alpha = 0$ in (7.10) have also been suggested for normalization [57], whereby the integer L_0 is adjusted empirically to control the loss of accuracy.

The normalization identity now completely determined by (7.11) and (7.13) (with equality sign), we proceed with a discussion of the resulting algorithm (3.9).

We first observe that the coefficients λ_L in (7.11) satisfy

$$\lambda_{L+1} = \frac{2L+1}{L+1}\omega\lambda_L + \frac{L^2+\eta^2}{L(L+1)}\lambda_{L-1}, \qquad L = 1, 2, 3, \cdots, \tag{7.14}$$

$$\lambda_0 = 1, \qquad \lambda_1 = \omega - \eta, \tag{7.15}$$

as follows readily from the well-known recurrence relation for Jacobi polynomials. In particular, they are all real. Using (7.14), (7.15) to generate the λ_L, algorithm (3.9) becomes

$$r_\nu^{(\nu)} = 0, \qquad r_{L-1}^{(\nu)} = \frac{1}{(2L-1)}\left\{\eta/(L(L+1)) + 1/\rho - \left[1 + \left(\frac{\eta}{L+1}\right)^2\right] r_L^{(\nu)}/(2L+3)\right\}^{-1}, \qquad L = \nu, \nu-1, \cdots, 1,$$

$$s_\nu^{(\nu)} = 0, \qquad s_{L-1}^{(\nu)} = r_{L-1}^{(\nu)}(s_L^{(\nu)} + \lambda_L), \tag{7.16}$$

$$f_0^{(\nu)} = \frac{\rho e^{\omega\rho}}{1 + s_0^{(\nu)}}, \qquad f_L^{(\nu)} = r_{L-1}^{(\nu)} f_{L-1}^{(\nu)}, \qquad L = 1, 2, \cdots, L_{\max}.$$

The final results F_L are readily obtained from (7.8), with the help of (7.4).

It is worthwhile to examine more closely the three-term recurrence relation (7.14). We note that it is a difference equation of the Poincaré type, with characteristic equation

$$t^2 - 2\omega t - 1 = 0.$$

Since the roots are

$$t_1 = \omega + \sqrt{\omega^2 + 1}, \qquad t_2 = \omega - \sqrt{\omega^2 + 1},$$

it follows from Theorem 2.2 that (7.14) for $\omega \neq 0$ has a minimal solution λ_L' for which

$$\frac{\lambda_{L+1}'}{\lambda_L'} \sim \omega - \sqrt{\omega^2 + 1}, \qquad L \to \infty, \tag{7.17}$$

while all other solutions behave according to

$$\frac{\lambda_{L+1}}{\lambda_L} \sim \omega + \sqrt{\omega^2 + 1}, \qquad L \to \infty. \tag{7.18}$$

To convince ourselves that the solution λ_L defined by (7.15) is *not* a minimal solution, we make use of the asymptotic formula[11]

$$P_n^{(\alpha,\beta)}(z) \sim (z-1)^{-\alpha/2}(z+1)^{-\beta/2}[(z+1)^{1/2} + (z-1)^{1/2}]^{\alpha+\beta} \times (2\pi n)^{-1/2}(z^2-1)^{-1/4}[z + (z^2-1)^{1/2}]^{n+1/2}, \qquad n \to \infty,$$

where z is outside the segment $[-1, 1]$. It follows, by a simple computation, that

$$\lambda_L = i^L P_L^{(i\eta,-i\eta)}(-i\omega) \sim e^{-\eta\phi}(2\pi L)^{-1/2}(1+\omega^2)^{-1/4}[\omega + \sqrt{\omega^2+1}]^{L+1/2}, \tag{7.19}$$
$$\omega \neq 0, \qquad L \to \infty,$$

where

$$\phi = \arctan\frac{1}{\omega}, \tag{7.20}$$

so that indeed (7.18) holds, rather than (7.17).

It may appear, therefore, that the use of (7.14) in forward direction is numerically "safe." Unfortunately, and surprisingly, it was observed by computation [applying algorithm (3.9) to (7.14)] that λ_L *defined by* (7.14), (7.15) *approaches a minimal solution*, in the sense

$$\lambda_1 \to \lambda_1',$$

$\{\lambda_L'\}$ being normalized by $\lambda_0' = 1$, as either η, or ρ, or both, become large.

[11] See [55, p. 194], where the result is stated for real α, β. The derivation by the method of steepest descent, however, is valid for arbitrary complex values of α, β.

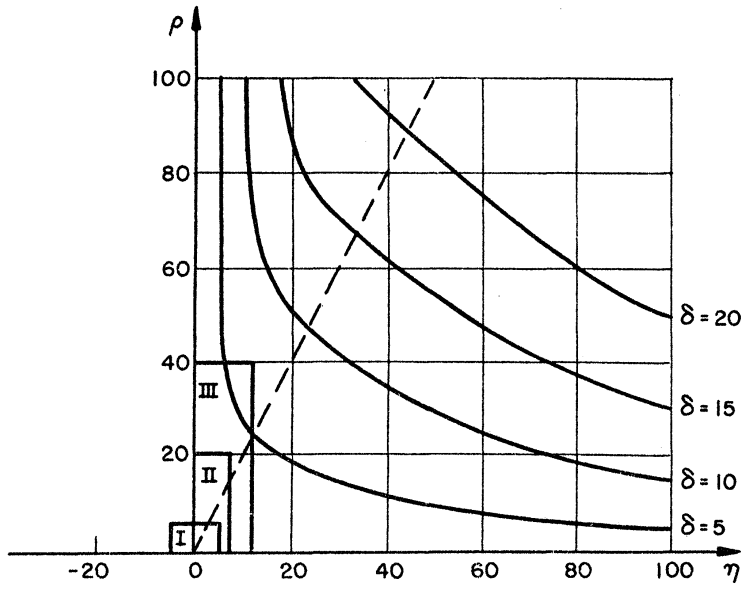


FIG. 10. *Degree of minimality of* $\{\lambda_L\}$. *The regions I, II, III indicate coverage of the tables* [36], [31], [58], *respectively, for* $F_L(\eta, \rho)$.

(Recall that ω is equal to the right-hand expressions in (7.13), and is thus a function of ρ and η.) To describe this phenomenon more precisely, let

$$\delta = -\log | \lambda_1 - \lambda_1' |,$$

which may be considered a measure of the "degree of minimality" of the solution λ_L. (We expect, roughly speaking, that the generation of λ_L by (7.14), (7.15) involves a loss of about δ decimal digits due to cancellation.) Fig. 10 shows the behavior of δ as a function of η and ρ. In particular, it can be seen that no serious cancellation problems arise in the regions (marked, I II, III) which are commonly of interest in applications. However, in special applications which involve large values of η and ρ, the loss of accuracy may indeed be quite substantial.

An obvious way to counteract this phenomenon is to generate the λ_L in double precision arithmetic. However, this may not be very efficient, considering that L, in the region in question, may assume values as large as 100, or more. We suggest the following alternative.

Let

$$\epsilon = \lambda_1 - \lambda_1' = \omega - \eta - \lambda_1', \tag{7.21}$$

a quantity that can be calculated to any degree of accuracy (in double precision, if necessary) without too much effort, using algorithm (3.9) for λ_1'. Let furthermore λ_L'' be the solution of (7.14) defined by

$$\lambda_0'' = -\lambda_1', \qquad \lambda_1'' = 1. \tag{7.22}$$

Then, using elementary facts from the theory of linear difference equations, one finds that

$$\lambda_L = \lambda_L' + \frac{\epsilon}{1 + \lambda_1'^2}(\lambda_L'' + \lambda_1'\lambda_L'). \tag{7.23}$$

Having determined ϵ accurately, we may now use (7.23) to calculate λ_L. This requires the computation of the minimal solution λ_L' by algorithm (3.9), and the computation of λ_L'' by (7.14), (7.22), but all of this can safely be done in single precision. Thus, double precision arithmetic will only be required in the computation of ϵ from (7.21).

For later use, we note the analogue of (7.19) for $\omega = 0$. In this case we use (cf. [55, p. 194] and footnote[11])

$$P_n^{(\alpha,\beta)}(0) \sim \frac{1}{\sqrt{\pi n}} 2^{(\alpha+\beta+1)/2} \cos\left([n + (\alpha + \beta + 1)/2]\frac{\pi}{2} - \left(\alpha + \frac{1}{2}\right)\frac{\pi}{2}\right), \qquad n \to \infty,$$

and find that

$$\lambda_L = i^L P_L^{(i\eta,-i\eta)}(0) \sim \frac{1}{\sqrt{2\pi L}}[(-1)^L e^{\pi\eta/2} + e^{-\pi\eta/2}], \qquad L \to \infty. \tag{7.24}$$

The starting value ν in (7.16) may be estimated similarly as for Bessel functions. Using (5.11), we may approximate the relative error of $f_L^{(\nu)}$ by

$$\frac{1}{\rho} e^{-\omega\rho}\lambda_{\nu+1} f_{\nu+1} - \frac{f_{\nu+1}}{g_{\nu+1}}\frac{g_L}{f_L}, \tag{7.25}$$

where $g_L = 2^L L!G_L(\eta, \rho)/((2L)!C_L(\eta))$. We wish to bound this for $L = L_{\max}$, assuming $L_{\max}$ and $\nu > L_{\max}$ large. By (7.6), (7.7), we have for large L,

$$\frac{f_L}{g_L} \sim \frac{e^{-\pi\eta}}{2e}\left(\frac{e\rho}{2L}\right)^{2L+1}.$$

Hence, the second term in (7.25), for large ν and L, may be estimated by

$$\frac{f_{\nu+1}}{g_{\nu+1}}\frac{g_L}{f_L} \sim \frac{\rho^2 L}{4\nu^3}\left(\frac{e\rho}{2}\right)^{2(\nu-L)} L^{2L}\nu^{-2\nu}. \tag{7.26}$$

To estimate the first term in (7.25) we observe from (7.19), (7.24), (7.6), and (7.8), that for L large, and $\omega \geqq 0$,

$$|\lambda_L f_L| \lesssim \frac{\rho A(\eta)(1 + \omega^2)^{-1/4}[B(\omega)]^{1/2}}{\sqrt{4\pi L}}\left(\frac{e\rho B(\omega)}{2L}\right)^L,$$

where

$$A(\eta) = \begin{cases} 2\cosh(\pi\eta/2), & \omega = 0, \\ e^{-\eta\phi}, & \omega > 0, \end{cases}$$

$$B(\omega) = \omega + \sqrt{\omega^2 + 1},$$

ϕ being defined in (7.20) The total relative error (7.25) will thus be $\leqq \frac{1}{2}\cdot 10^{-d}$, if we require that

$$e^{-\omega\rho}A(\eta)(1+\omega^2)^{-1/4}[B(\omega)]^{1/2}\left(\frac{e\rho B(\omega)}{2(\nu+1)}\right)^{\nu+1} \leqq \frac{1}{4}\cdot 10^{-d},$$

$$\left(\frac{e\rho}{2}\right)^{2(\nu-L)} L^{2L}\nu^{-2\nu} \leqq \frac{1}{4}\cdot 10^{-d},\quad L = L_{\max}.$$

From here on, the analysis proceeds as for Bessel functions. Assuming (without loss of generality) that $L_{\max} > e\rho/2$, the result is that ν must satisfy both of the following conditions:

$$(7.27)\qquad \begin{aligned} \nu &\geqq \frac{e\rho B(\omega)}{2}\, t\left(\frac{2}{e\rho B(\omega)}\,[D - \omega\rho + \ln\,(A(\eta)\sqrt{B(\omega)}(1+\omega^2)^{-1/4})]\right), \\ \nu &\geqq L_{\max}\, t\left(\frac{D}{2L_{\max}}\right), \end{aligned}$$

where we recall that $D = d \ln 10 + \ln 4$, and $t(x)$ is the inverse function of $x = t \ln t$.

An ALGOL procedure for the computation of $F_L(\eta, \rho)$, using the methods described in this paragraph, may be found in [20].

8. Incomplete beta and gamma function. The incomplete beta function is defined by the integral

$$(8.1)\qquad B_x(p, q) = \int_0^x t^{p-1}(1-t)^{q-1}\,dt, \quad p > 0, q > 0, 0 \leqq x \leqq 1.$$

The complete beta function is obtained when $x = 1$, and can be expressed in terms of gamma functions,

$$(8.2)\qquad B_1(p, q) = \int_0^1 t^{p-1}(1-t)^{q-1}\,dt = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}.$$

For large p or large q Laplace's method (see [11, p. 37]) yields the asymptotic formulae,

$$(8.3)\qquad B_x(p, q) \sim (1-x)^{q-1}\frac{x^p}{p}, \qquad p \to \infty, q \text{ fixed},$$

$$(8.4)\qquad B_x(p, q) \sim \Gamma(p)q^{-p}, \qquad q \to \infty, p \text{ fixed}.$$

In probability distribution theory the following ratio of beta functions is important,

$$(8.5)\qquad I_x(p, q) = \frac{B_x(p, q)}{B_1(p, q)}.$$

Recurrence relations hold in both variables p and q (see [3]):[12]

$$(8.6)\qquad \begin{aligned} pI_x(p+1, q) - [(p+q-1)x + p]I_x(p, q) \\ + (p+q-1)xI_x(p-1, q) = 0, \end{aligned}$$

[12] Formula (14) in [3] contains a misprint: the last term on the left should have the factor q, not p.

$$qI_x(p, q+1) - [(p+q-1)(1-x)+q]I_x(p, q) + (p+q-1)(1-x)I_x(p, q-1) = 0. \tag{8.7}$$

It also follows from (8.5) that

$$I_x(q, p) = 1 - I_{1-x}(p, q). \tag{8.8}$$

The calculation of $I_x(p, q)$ presents no difficulty when both p and q are small or moderately large. Expansion of $(1-t)^{q-1}$ into the binomial series then leads to a rapidly convergent series for $B_x(p, q)$, especially since by (8.8) we can always arrange to have x in the interval $0 \leqq x \leqq \frac{1}{2}$. Moreover, the gamma functions in (8.2) are rapidly calculated by reducing the arguments to some standard interval for which rational approximations are available [60]. When p or q is large, however, it may be more efficient to make use of the recursions (8.6) or (8.7).

Consider then, first,

$$f_n = I_x(p+n, q), \qquad n = 0, 1, 2, \cdots; \quad 0 < p \leqq 1, \quad q > 0.$$

By (8.6) this is a solution of

$$y_{n+1} - \left(1 + \frac{n+p+q-1}{n+p}x\right)y_n + \frac{n+p+q-1}{n+p}xy_{n-1} = 0, \tag{8.9}$$

again a Poincaré difference equation. The characteristic equation $t^2 - (1+x)t + x = 0$ has the roots

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

By inspection (8.9) has the solution $y_n \equiv 1$, which clearly corresponds to the root t_1. On the other hand, from (8.3) and (8.5), we find

$$f_{n+1}/f_n \sim x, \qquad n \to \infty,$$

so that f_n corresponds to the root t_2. Therefore, f_n is the minimal solution of (8.9).

While our methods of §§3 and 4 again apply, it must be noted that in contrast to the previous examples the dominant solution is now bounded. Forward recursion by means of (8.9) should therefore cause no difficulties if the f_n are to be obtained to a fixed number of decimals after the decimal point. If a given number of *significant* digits is required, however, it is more appropriate to employ the algorithms in §§3 and 4. The initial value $f_0 = I_x(p, q)$ needed in these algorithms may be obtained by first reducing q modulo 1 to q_0, where $0 < q_0 \leqq 1$, then calculating $I_x(p, q_0)$, $I_x(p, q_0+1)$ by series expansion, and finally applying the second recursion (8.7) to connect with $I_x(p, q)$.

Consider next

$$g_n = I_x(p, q+n), \qquad n = 0, 1, 2, \cdots; \quad p > 0, \quad 0 < q \leqq 1.$$

From (8.7) we now get the difference equation

$$y_{n+1} - \left[1 + \frac{n+p+q-1}{n+q}(1-x)\right] y_n + \frac{n+p+q-1}{n+q}(1-x) y_{n-1} = 0, \tag{8.10}$$

which may also be obtained from (8.9) by interchanging p with q and, simultaneously, x with $1 - x$. Therefore (8.10) has the solutions g_n and $I_{1-x}(q+n, p)$, of which the latter is again the minimal solution. We see that g_n is among the dominant solutions, and no problem of numerical instability arises.

For a detailed description of these algorithms we refer to [17].

The incomplete gamma function is defined by

$$P(a, x) = \frac{1}{\Gamma(a)} \int_0^x e^{-t} t^{a-1}\, dt, \qquad a > 0, x > 0. \tag{8.11}$$

It satisfies the well-known recurrence relation

$$P(a, x) = P(a-1, x) - \frac{x^{a-1} e^{-x}}{\Gamma(a)},$$

which, by elimination of the inhomogeneous term, can be brought into the form

$$aP(a+1, x) - (x+a)P(a, x) + xP(a-1, x) = 0.$$

Letting $f_n = P(a+n, x)$, we therefore find that f_n is a solution of

$$(a+n)y_{n+1} - (x+a+n)y_n + xy_{n-1} = 0, \quad n = 1, 2, 3, \cdots. \tag{8.12}$$

This again is a Poincaré difference equation, whose characteristic equation $t^2 - t = 0$ has the roots $t_1 = 1$, $t_2 = 0$. The solution of (8.12) corresponding to t_1 is clearly $y_n \equiv 1$. The solution corresponding to t_2 is f_n, since

$$\frac{f_{n+1}}{f_n} \sim \frac{x}{n}, \qquad n \to \infty,$$

as follows from the well-known asymptotic formula

$$P(a, x) \sim x^a e^{-x}/\Gamma(a+1), \qquad a \to \infty.$$

(See, e.g., [13, p. 140].) Consequently, f_n is a minimal solution of (8.12).

To obtain an infinite series in f_n, we multiply f_m by

$$\lambda_m = \frac{\Gamma(a+m)}{m!\Gamma(a)}, \tag{8.13}$$

and sum over m. We get

$$\sum_{m=0}^{\infty} \lambda_m f_m = \frac{1}{\Gamma(a)} \sum_{m=0}^{\infty} \frac{1}{m!} \int_0^x e^{-t} t^{a+m-1}\, dt$$

$$= \frac{1}{\Gamma(a)} \int_0^x e^{-t} t^{a-1} \sum_{m=0}^{\infty} \frac{t^m}{m!}\, dt$$

$$= \frac{1}{\Gamma(a)} \int_0^x t^{a-1}\, dt = \frac{x^a}{a\Gamma(a)},$$

and therefore,

$$\sum_{m=0}^{\infty} \lambda_m f_m = \frac{x^a}{\Gamma(a+1)}. \tag{8.14}$$

The coefficients λ_m can easily be obtained from the recursion

$$\lambda_0 = 1, \qquad \lambda_m = \frac{a+m-1}{m}\lambda_{m-1}, \qquad m = 1, 2, 3, \cdots. \tag{8.15}$$

Our algorithms may now be applied to (8.12), (8.14) to compute $P(a+n, x)$ for $n = 0, 1, 2, \cdots, N$.

9. Repeated integrals of the error function. In problems of heat conduction the complementary error function

$$\operatorname{erfc} z = \frac{2}{\sqrt{\pi}} \int_z^{\infty} e^{-t^2}\, dt$$

and its repeated integrals frequently occur. Following Hartree [25] we denote

$$\mathrm{i}^0 \operatorname{erfc} z = \operatorname{erfc} z,$$

$$\mathrm{i}^n \operatorname{erfc} z = \int_z^{\infty} \mathrm{i}^{n-1} \operatorname{erfc} t\, dt, \qquad n = 1, 2, 3, \cdots.$$

It is also convenient to define

$$\mathrm{i}^{-1} \operatorname{erfc} z = \frac{2}{\sqrt{\pi}} e^{-z^2}.$$

Expressed as a single integral, we have

$$\mathrm{i}^n \operatorname{erfc} z = \frac{2}{\sqrt{\pi}} \int_z^{\infty} \frac{(t-z)^n}{n!} e^{-t^2}\, dt.$$

Writing

$$\mathrm{i}^{n+1} \operatorname{erfc} z = \frac{2}{\sqrt{\pi}} \left(\frac{1}{n+1} \int_z^{\infty} \frac{(t-z)^n}{n!} t e^{-t^2}\, dt - \frac{z}{n+1} \int_z^{\infty} \frac{(t-z)^n}{n!} e^{-t^2}\, dt \right),$$

and evaluating the first integral by parts, one finds

$$\mathrm{i}^{n+1} \operatorname{erfc} z + \frac{z}{n+1} \mathrm{i}^n \operatorname{erfc} z - \frac{1}{2(n+1)} \mathrm{i}^{n-1} \operatorname{erfc} z = 0, \quad n = 0, 1, 2, \cdots.$$

Consider now

$$f_n = e^{z^2} \mathrm{i}^n \operatorname{erfc} z, \qquad n = -1, 0, 1, 2, \cdots,$$

which clearly is a solution of

$$y_{n+1} + \frac{z}{n+1} y_n - \frac{1}{2(n+1)} y_{n-1} = 0, \qquad n = 0, 1, 2, \cdots. \tag{9.1}$$

To this difference equation case (c) of Theorem 2.3 could be applied with the result that all solutions behave "similarly" as $n \to \infty$, viz.,

$$\limsup_{n\to\infty} (|\, y_n \,| \sqrt{n!})^{1/n} = \frac{1}{\sqrt{2}}.$$

This conclusion is somewhat deceiving, as in fact f_n is the minimal solution of (9.1).

To see this, we make use of the result that for any fixed z, as $n \to \infty$,

$$\mathrm{i}^n \operatorname{erfc} z \sim \frac{e^{-(1/2)z^2}}{2^n \Gamma\left(\frac{n}{2}+1\right)} \exp\left(-\sqrt{2n}\, z\right). \tag{9.2}$$

[See [13, p. 123] and also recall that the repeated integrals of the error function are related to parabolic cylinder functions $D_\nu(z)$ by

$$\mathrm{i}^n \operatorname{erfc} z = (e^{-z^2}/2^{n-1}\pi)^{1/2} D_{-n-1}(z\sqrt{2}).]$$

By inspection, moreover, one sees that

$$g_n = (-1)^n e^{z^2} \mathrm{i}^n \operatorname{erfc}(-z)$$

also satisfies the recurrence relation (9.1). Applying (9.2) to both f_n and g_n, we find

$$(-1)^n \frac{f_n}{g_n} \sim e^{-2\sqrt{2n}z}, \qquad n \to \infty. \tag{9.3}$$

This shows that f_n is indeed the minimal solution of (9.1) whenever $\operatorname{Re} z > 0$. Otherwise, when $\operatorname{Re} z < 0$, g_n is the minimal solution.

Our algorithms of §§3 and 4 for computing f_n are particularly simple, in this case, since the initial value is known to be

$$f_{-1} = 2/\sqrt{\pi}.$$

From (9.3) it is evident that convergence of the first algorithm is better the further away z is from the imaginary axis.

The application of Miller's backward recurrence algorithm in this connection was first suggested by M. Abramowitz [1], and is further analyzed in [16].

We note, incidentally, that Theorem 1.1 gives us the identity

$$\frac{f_n}{f_{n-1}} = \cfrac{\cfrac{1}{2(n+1)}}{\cfrac{z}{n+1} + \cfrac{\cfrac{1}{2(n+2)}}{\cfrac{z}{n+2} + \cfrac{\cfrac{1}{2(n+3)}}{\cfrac{z}{n+3} + \cdots}}},$$

which by an equivalence transformation can be brought into the form

$$\frac{\mathrm{i}^n \operatorname{erfc} z}{\mathrm{i}^{n-1} \operatorname{erfc} z} = \cfrac{1/2}{z+} \cfrac{(n+1)/2}{z+} \cfrac{(n+2)/2}{z+} \cdots, \qquad \operatorname{Re} z > 0.$$

For $n = 0$, this reduces to the well-known result

$$2e^{z^2}\int_z^\infty e^{-t^2}\,dt = \frac{1}{z+}\,\frac{1/2}{z+}\,\frac{1}{z+}\,\frac{3/2}{z+}\cdots.$$

10. An example arising in the numerical computation of Fourier coefficients. Let $f(t)$ be a function defined and continuous on the closed interval $[0, 2\pi]$, and let

$$(10.1)\qquad a_p = \int_0^{2\pi} f(t)\cos pt\,dt, \qquad b_p = \int_0^{2\pi} f(t)\sin pt\,dt, \quad p = 0, 1, 2, \cdots,$$

denote its Fourier coefficients. The computation of Fourier coefficients of high order (p large) is notoriously difficult because of two reasons. Firstly, if one attempts to apply standard integration techniques, such as the trapezoidal rule, one is forced into a rather fine subdivision of the interval $[0, 2\pi]$ in order to cover adequately the many oscillations of the trigonometric factors in (10.1). Secondly, even if one adopts a sufficiently fine subdivision, substantial cancellation of digits will occur in the summation associated with the integration formula. Indeed, by Riemann's lemma, both a_p and b_p tend to zero when $p \to \infty$, whereas the individual terms of the integration formula need not be small at all. In matter of fact, cancellation will be more prominent the smoother the function f is!

In order to circumvent these difficulties, it has been suggested to use Gauss type integration methods, treating the troublesome trigonometric factors as weight functions [61], [62]. As the general theory of Gaussian quadrature requires nonnegative weight functions, one first writes

$$(10.2)\qquad a_p = \int_0^{2\pi} f(t)\cos pt\,dt = \int_0^{2\pi} f(t)\,dt - \int_0^{2\pi} f(t)(1-\cos pt)\,dt,$$

and similarly for b_p. Then Gaussian integration is applied to the second integral, while the first integral is evaluated by some standard technique. Both integrals may have to be evaluated to high accuracy, since for large p, they are nearly equal. Thus, our cancellation problem is not entirely eliminated, but appears to be under better control.

Gaussian quadrature formulae of possibly various orders have to be obtained for each value of p. While this is a formidable task in itself, it appears feasible on current high-speed computers. One would presumably start from the moments

$$(10.3)\qquad c_n = \int_0^{2\pi} t^n(1-\cos pt)\,dt, \qquad s_n = \int_0^{2\pi} t^n(1-\sin pt)\,dt, \qquad n = 0, 1, 2, \cdots,$$

and use these to construct either the associated orthogonal polynomials, or the continued fractions associated with the formal power series

$$\sum_{n=0}^{\infty} c_n z^{-n-1}, \qquad \sum_{n=0}^{\infty} s_n z^{-n-1}.$$

The abscissae and weights of the desired quadrature formula then follow readily. Because of the inherent sensitivity of these quantities with respect to perturbations of the moments, it is rather important that the moments (10.3) be obtained as accurately as possible. Our concern here will be with a stable generation of these moments.

We assume p a positive integer. Integrating by parts, we have[13]

$$c_{n+1} = \int_0^{2\pi} t^{n+1}(1 - \cos pt)\, dt = \left[t^{n+1}\left(t - \frac{\sin pt}{p}\right)\right]_0^{2\pi} - \int_0^{2\pi} (n+1)t^n \left(t - \frac{\sin pt}{p}\right) dt$$

$$= (2\pi)^{n+2} - (n+1)\int_0^{2\pi} t^{n+1}\, dt + \frac{n+1}{p}\int_0^{2\pi} t^n \sin pt\, dt$$

$$= (2\pi)^{n+2} - \frac{n+1}{n+2}(2\pi)^{n+2} - \frac{n+1}{p}\int_0^{2\pi} t^n(1 - \sin pt)\, dt + \frac{n+1}{p}\int_0^{2\pi} t^n\, dt$$

$$= (2\pi)^{n+2} - \frac{n+1}{n+2}(2\pi)^{n+2} + \frac{1}{p}(2\pi)^{n+1} - \frac{n+1}{p}s_n,$$

hence,

$$c_{n+1} = -\frac{n+1}{p}s_n + (2\pi)^{n+1}\left(\frac{1}{p} + \frac{2\pi}{n+2}\right), \quad n = 0, 1, 2, \cdots. \tag{10.4}$$

Similarly, one obtains

$$s_{n+1} = \frac{n+1}{p}c_n + \frac{(2\pi)^{n+2}}{n+2}, \quad n = 0, 1, 2, \cdots. \tag{10.5}$$

Replacing n by $n - 1$ in (10.5), and inserting the result in (10.4), one gets

$$c_{n+1} = -\frac{n(n+1)}{p^2}c_{n-1} + \frac{(2\pi)^{n+2}}{n+2}, \quad n = 1, 2, 3, \cdots. \tag{10.6}$$

Eliminating similarly the c's from (10.4) and (10.5), one gets

$$s_{n+1} = -\frac{n(n+1)}{p^2}s_{n-1} + (2\pi)^n\left(\frac{n+1}{p^2} + \frac{2\pi}{p} + \frac{4\pi^2}{n+2}\right), \quad n = 1, 2, 3, \cdots. \tag{10.7}$$

Writing down (10.6) once with n increased by unity, and once with n decreased

[13] In principle, c_n and s_n could be evaluated in closed form. However recursive generation of these quantities is more effective. Alternatively, we could integrate the additive term t^n in closed form, and compute $\int_0^{2\pi} t^n \cos pt\, dt$ and $\int_0^{2\pi} t^n \sin pt\, dt$ recursively. No substantial simplification would result, however.

by unity, and eliminating the inhomogeneous terms, one finally obtains

$$c_{n+2} + \left[\frac{(n+1)(n+2)}{p^2} - \frac{4\pi^2(n+1)}{n+3}\right] c_n - 4\pi^2 \frac{(n-1)n(n+1)}{p^2(n+3)} c_{n-2} = 0. \tag{10.8}$$

Similarly,

$$s_{n+2} + \left[\frac{(n+1)(n+2)}{p^2} - \sigma_n\right] s_n - \sigma_n \frac{(n-1)n}{p^2} s_{n-2} = 0, \tag{10.9}$$

where

$$\sigma_n = 4\pi^2 \frac{n + 2 + 2\pi p + \dfrac{4\pi^2 p^2}{n+3}}{n + 2\pi p + \dfrac{4\pi^2 p^2}{n+1}}.$$

The recurrence relations (10.8), (10.9) are valid for $n \geqq 2$.

It is clear that (10.8), (10.9) permit, in principle, all moments of even order to be obtained from those of order 0 and 2,

$$\begin{aligned} c_0 &= 2\pi, \qquad c_2 = 4\pi\left(\frac{2\pi^2}{3} - \frac{1}{p^2}\right), \\ s_0 &= 2\pi, \qquad s_2 = 4\pi^2\left(\frac{2\pi}{3} + \frac{1}{p}\right), \end{aligned} \tag{10.10}$$

and all moments of odd order from those of order 1 and 3,

$$\begin{aligned} c_1 &= 2\pi^2, \qquad c_3 = 4\pi^2\left(\pi^2 - \frac{3}{p^2}\right), \\ s_1 &= 2\pi\left(\pi + \frac{1}{p}\right), \qquad s_3 = 4\pi\left(\pi^3 + \frac{2\pi^2}{p} - \frac{3}{p^3}\right). \end{aligned} \tag{10.11}$$

As it happens, however, the moments are minimal solutions of (10.8) and (10.9), respectively. Therefore, straightforward recursion, as indicated, is highly unstable. We expect the algorithms of §§3 and 4 to be rather more effective, especially since the first relations in (10.10), (10.11) can be used for normalization.

To establish the minimal character of the moments, let us first write

$$c_{2n+h} = C_n, \qquad s_{2n+h} = S_n,$$

where h is either zero or one. Then C_n and S_n are both solutions of three-term recurrence relations of the standard form

$$y_{n+1} + a_n y_n + b_n y_{n-1} = 0, \qquad n = 1, 2, 3, \cdots, \tag{10.12}$$

the Newton-Puiseux diagram in both cases having the form shown in Fig. 11.

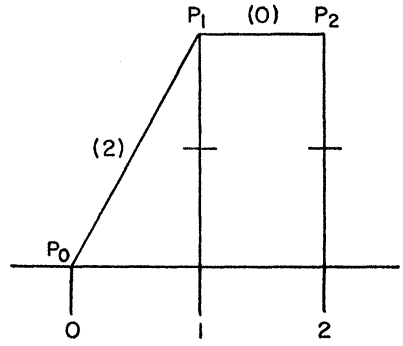


FIG. 11. *Newton-Puiseux diagram for* (10.8), (10.9)

Moreover,

$$a_n \sim \frac{4}{p^2} n^2, \qquad b_n \sim -\frac{16\pi^2}{p^2} n^2, \qquad n \to \infty.$$

It follows from part (a) of Theorem 2.3 that (10.12) has a pair of fundamental solutions, $y_{n,1}$ and $y_{n,2}$, for which

$$\frac{y_{n+1,1}}{y_{n,1}} \sim -\frac{4}{p^2} n^2, \qquad \frac{y_{n+1,2}}{y_{n,2}} \sim 4\pi^2, \qquad n \to \infty. \tag{10.13}$$

Both solutions thus tend with n to infinity, but the first one much more rapidly than the second.

On the other hand, applying Laplace's method [11, p. 37] to the integrals in (10.3), one finds readily that for $n \to \infty$,

$$C_n \sim p^2(\pi/n)^3(2\pi)^{2n+h}, \qquad S_n \sim \frac{\pi}{n}(2\pi)^{2n+h}, \qquad h = 1, 2.$$

The C's and S's, therefore, exhibit the same asymptotic behavior as $y_{n,2}$ in (10.13). Consequently, they are both minimal solutions of the respective equation (10.12).

11. A Sturm-Liouville boundary value problem. Consider the Sturm-Liouville boundary value problem with one boundary condition at infinity,

$$(p(t)y')' + q(t)y = 0, \qquad y(0) = 1, \qquad y(\infty) = 0. \tag{11.1}$$

We assume that p and q are real-valued continuous functions in $[0, \infty)$, with $p(t) > 0$, $q(t) \leqq 0$, and in addition that

$$\int^{\infty} \frac{dt}{p(t)} = \infty, \qquad -\int^{\infty} q(t)\left(\int^{t} \frac{ds}{p(s)}\right) dt = \infty. \tag{11.2}$$

Then the boundary value problem (11.1) has an unique solution which is minimal in the continuous sense [24, p. 357 ff]. When solving the problem numerically, by a method of finite differences, we expect the approximate solution to be minimal in the discrete sense. We wish to illustrate this in the case of a simple finite difference scheme.

Consider mesh points $t_n = nh$, $n = 0, 1, 2, \cdots$, where $h > 0$ is small, but fixed, and let y_n designate approximations at t_n to the solution $y(t)$ of (11.1),

$$y_n \doteq y(t_n), \qquad n = 0, 1, 2, \cdots.$$

Such approximations may be obtained by first rewriting (11.1) as a system of two first-order differential equations, letting $z = p(t)y'$,

$$z' + q(t)y = 0,$$

$$y' - \frac{1}{p(t)} z = 0,$$

and then replacing derivatives by central difference quotients. We get

$$\frac{z_{n+1/2} - z_{n-1/2}}{h} + q_n y_n = 0,$$

$$\frac{y_{n+1/2} - y_{n-1/2}}{h} - \frac{1}{p_n} z_n = 0,$$

where $p_n = p(t_n)$, $q_n = q(t_n)$. Eliminating the z's, we obtain the following discrete analogue of (11.1),

$$y_{n+1} - \frac{p_{n+1/2} + p_{n-1/2} - h^2 q_n}{p_{n+1/2}} y_n + \frac{p_{n-1/2}}{p_{n+1/2}} y_{n-1} = 0, \quad n = 1, 2, 3, \cdots, \tag{11.3}$$

$$y_0 = 1, \qquad \lim_{n\to\infty} y_n = 0. \tag{11.4}$$

It appears to be an open question whether under the assumptions (11.2), or some discrete analogue thereof, the difference equation (11.3) possesses a minimal solution satisfying (11.4), if h is suitably restricted. The answer, however, is in the affirmative, if we make the stronger assumptions

$$\lim_{t\to\infty} p(t) = p > 0, \qquad \lim_{t\to\infty} q(t) = q < 0. \tag{11.5}$$

Then, indeed, (11.3) is a Poincaré difference equation having the characteristic equation

$$t^2 - \left(2 - h^2 \frac{q}{p}\right) t + 1 = 0.$$

Since $p > 0$, $q < 0$, the roots t_1, t_2 of this equation are real and distinct for all $h > 0$. In fact,

$$t_1 = 1 - h^2 \frac{q}{2p} + h \sqrt{\frac{-q}{p}\left(1 - h^2 \frac{q}{4p}\right)}, \qquad t_2 = t_1^{-1},$$

so that $t_1 > 1$, $0 < t_2 < 1$. The solution of (11.3) corresponding to t_2 therefore is a minimal solution, for arbitrary h, and can be normalized to satisfy the first condition in (11.4). The second condition (at infinity) is insured, since by Theorem 2.2,

$$\frac{y_{n+1}}{y_n} \sim t_2, \qquad n \to \infty,$$

for any minimal solution of (11.3).

Clearly, algorithm (3.9) applies in its simplified form (without the s-recursion), since y_0 is given to be 1.

By way of an example, consider

$$y'' = \frac{1+t}{2+t}\, y, \qquad y(0) = 1, \qquad y(\infty) = 0. \tag{11.6}$$

(This may be interpreted as a heat conduction problem for an infinite rod; cf. [8, p. 150]). Here,

$$p(t) \equiv 1, \qquad q(t) = -\frac{1+t}{2+t},$$

and (11.5) is satisfied with $p = 1$, $q = -1$. The discrete analogue of (11.6) takes the form

$$y_{n+1} - \left(2 + h^2 \frac{1+nh}{2+nh}\right) y_n + y_{n-1} = 0,$$

$$y_0 = 1, \qquad \lim_{n\to\infty} y_n = 0.$$

Applying algorithm (3.9), we obtain approximations $y_n^{(\nu)}$ to y_n from

$$\begin{aligned} &r_\nu^{(\nu)} = 0, \qquad r_{n-1}^{(\nu)} = \frac{1}{2 + h^2 \dfrac{1+nh}{2+nh} - r_n^{(\nu)}}, \qquad n = \nu, \nu-1, \cdots, 1, \\ &y_0^{(\nu)} = 1, \qquad y_n^{(\nu)} = r_{n-1}^{(\nu)}\, y_{n-1}^{(\nu)}, \qquad n = 1, 2, \cdots, N. \end{aligned} \tag{11.7}$$

Here, N is determined by the length of the interval in which the solution $y(t)$ is sought.

Table 2 displays selected numerical results for integrating (11.6) by (11.7) on the interval [0, 5]. The first column shows the number N of subintervals, the second column the corresponding value of h ($= 5/N$), the third column the

TABLE 2

Approximate solution $y_n^{(\nu)}$ of the boundary value problem (11.6) *by means of* (11.7), *for* $n = kN/5$, $k = 0(1)5$

N	h	ν	t					
			0	1	2	3	4	5
5	1	13	1.0	.446887	.191699	.080285	.033098	.013494
10	.5	25	1.0	.443648	.187645	.077222	.031219	.012465
50	.1	116	1.0	.442753	.186395	.076251	.030620	.012137
250	.02	511	1.0	.442729	.186352	.076217	.030598	.012124

smallest integer ν for which six significant digits are achieved. The remaining columns contain the approximations $y_n^{(\nu)}$ corresponding to $t = 1(1)5$.

Acknowledgment. Parts of this paper were written while the author was with the Applied Mathematics Division of the Argonne National Laboratory as a resident research associate during the summers of 1964 and 1965. The author wishes to express his gratitude to Drs. W. F. Miller and J. W. Givens for making this association possible, and for providing him access to the Division's excellent computing facilities. He is also indebted to Dr. H. C. Thacher, Jr., for reading the entire manuscript, and for suggesting many improvements, both in form and substance. The author received valuable comments from Dr. F. W. J. Olver.

REFERENCES

[1] M. Abramowitz, *Review 58*, Math. Tables Aids Comput., 10 (1956), p. 176.
[2] R. Babushkova, *Über numerische Stabilität einiger Rekursionsformeln*, Apl. Mat., 9 (1964), pp. 186–193.
[3] T. A. Bancroft, *Some recurrence formulae in the incomplete beta function ratio*, Ann. Math. Statist., 20 (1949), pp. 451–455.
[4] G. Blanch, *Numerical evaluation of continued fractions*, this Review, 6 (1964), pp. 383–421.
[5] British Association for the Advancement of Science, *Mathematical Tables, vol. X, Bessel functions*, Part II, Functions of positive integer order, Cambridge University Press, 1952.
[6] H. Buchholz, *Die konfluente hypergeometrische Funktion*, Ergebnisse der angewandten Mathematik, vol. 2, Springer, Berlin, 1953.
[7] C. W. Clenshaw, *The numerical solution of linear differential equations in Chebyshev series*, Proc. Cambridge Philos. Soc., 53 (1957), pp. 134–149.
[8] L. Collatz, *The numerical treatment of differential equations*, 3rd ed., Springer, Berlin, 1960.
[9] F. J. Corbató, *On the computation of auxiliary functions for two-center integrals by means of a high-speed computer*, J. Chem. Phys., 24 (1956), pp. 452–453.
[10] F. J. Corbató and J. L. Uretsky, *Generation of spherical Bessel functions in digital computers*, J. Assoc. Comput. Mach., 6 (1959), pp. 366–375.
[11] A. Erdélyi, *Asymptotic expansions*, Dover, New York, 1956.
[12] A. Erdélyi et al., *Higher Transcendental Functions*, vol. I, McGraw-Hill, New York, 1953.
[13] ———, *Higher Transcendental Functions*, vol. II, McGraw-Hill, New York, 1953.
[14] M. A. Evgrafov, *A new proof of a theorem of Perron*, Izv. Akad. Nauk SSSR Ser. Mat., 17 (1953), pp. 77–82.
[15] L. Fox, *A short table for Bessel functions of integer orders and large arguments*, Royal Society Shorter Mathematical Tables, No. 3, Cambridge University Press, 1954.
[16] W. Gautschi, *Recursive computation of the repeated integrals of the error function*, Math. Comput., 15 (1961), pp. 227–232.
[17] ———, *Algorithm 222—Incomplete beta function ratios*, Comm. ACM, 7 (1964), pp. 143–144; *Certification of Algorithm 222*, Ibid., p. 244.
[18] ———, *Algorithm 236—Bessel functions of the first kind*, Ibid., 7 (1964), pp. 479–480; *Certification of Algorithm 236*, Ibid., 8 (1965), pp. 105–106.
[19] ———, *Algorithm 259—Legendre functions for arguments larger than one*, Ibid., 8 (1965), pp. 488–492.
[20] ———, *Algorithm 292—Regular Coulomb wave functions*, Ibid., 9 (1966), pp. 793–795.
[21] A. O. Gel'fond, *Calculus of Finite Differences*, 2nd ed., Gosud. Izdat. Fiz.-Mat. Lit.,

Moscow, 1959. (German and French translations of the first edition are available.)

[22] M. GOLDSTEIN AND R. M. THALER, *Recurrence techniques for the calculation of Bessel functions*, Math. Tables Aids Comput., 13 (1959), pp. 102–108.

[23] M. C. GRAY, *Bessel functions of integral order and complex argument*, Comm. ACM, 4 (1961), p. 169.

[24] P. HARTMAN, *Ordinary Differential Equations*, John Wiley, New York, 1964.

[25] D. R. HARTREE, *Some properties and applications of the repeated integrals of the error function*, Mem. Proc. Manchester Lit. Philos. Soc., 80 (1936), pp. 85–102.

[26] S. HITOTUMATU, *Note on the computation of Bessel functions through recurrence formula*, J. Math. Soc. Japan, 15 (1963), pp. 353–359.

[27] C. W. JONES, *A short table for the Bessel functions* $I_{n+1/2}(x)$, $(2/\pi)K_{n+1/2}(x)$, Royal Society Shorter Mathematical Tables, No. 1, Cambridge University Press, 1952.

[28] W. KAHAN, *Note on bounds for generating Bessel functions by recurrence*, unpublished.

[29] P. KREUSER, *Über das Verhalten der Integrale homogener linearer Differenzengleichungen im Unendlichen*, Dissertation, University of Tübingen, Leipzig, 1914.

[30] N. A. LOGAN, *Survey of some early studies of the scattering of plane waves by a sphere*, Proc. IEEE, 53 (1965), pp. 773–785.

[31] A. V. LUK'YANOV, I. B. TEPLOV AND M. K. AKIMOV, *Tablicy volnovih kulonovskih funkciĭ*, Izdat. Akad. Nauk SSSR., Moscow, 1961. (English translation by D. E. Brown, Pergamon Press, New York, 1965).

[32] S. MAKINOUCHI, *Note on the recurrence techniques for the calculation of Bessel functions* $J_\nu(x)$, Technology Reports Osaka University, 15 (1965), pp. 185–201.

[33] ———, *Note on the recurrence techniques for the calculation of Bessel functions* $I_\nu(x)$, Ibid., 15 (1965), pp. 203–216.

[34] H. MESCHKOWSKI, *Differenzengleichungen*, Vandenhoeck and Ruprecht, Göttingen, 1959.

[35] L. M. MILNE-THOMSON, *The Calculus of Finite Differences*, Macmillan, London, 1933.

[36] NATIONAL BUREAU OF STANDARDS, *Tables of Coulomb wave functions*, vol. I, Appl. Math. Ser. 17, 1952.

[37] N. E. NÖRLUND, *Vorlesungen über Differenzenrechnung*, Springer, Berlin, 1924.

[38] F. W. J. OLVER, *Error analysis of Miller's recurrence algorithm*, Math. Comput., 18 (1964), pp. 65–74.

[39] ———, *Bessel functions of integer order*, Handbook of Mathematical Functions, NBS Appl. Math. Ser. 55, 1964, Chap. 9.

[40] M. ONOE, *Tables of modified quotients of Bessel functions of the first kind for real and imaginary arguments*, Columbia University Press, New York, 1958.

[41] O. PERRON, *Über einen Satz des Herrn Poincaré*, J. Reine Angew. Math., 136 (1909), pp. 17–37.

[42] ———, *Über lineare Differenzengleichungen*, Acta Math., 34 (1911), pp. 109–137.

[43] ———, *Die Lehre von den Kettenbrüchen*, B. G. Teubner, Stuttgart, vol. I, 1954, vol. II, 1957.

[44] S. PINCHERLE, *Sur la génération de systèmes récurrents au moyen d'une équation linéaire différentielle*, Acta Math., 16 (1892), pp. 341–363.

[45] ———, *Delle funzioni ipergeometriche e di varie questioni ad esse attinenti*, Giorn. Mat. Battaglini, 32 (1894), pp. 209–291, esp. Ch. III. § 15. (Also in: *Opere Scelte*, vol. 1, pp. 273–357.)

[46] H. POINCARÉ, *Sur les équations linéaires aux différentielles ordinaires et aux différences finies*, Amer. J. Math., 7 (1885), pp. 203–258. (Also in: *Oeuvres Henri Poincaré*, vol. 1, pp. 226–289.)

[47] J. B. RANDELS AND R. F. REEVES, *Note on empirical bounds for generating Bessel functions*, Comm. ACM, 1 (1958), pp. 3–5.

[48] LORD RAYLEIGH (J. W. STRUTT), *The incidence of light upon a transparent sphere of dimensions comparable with the wave-length*, Proc. Roy. Soc. London Ser. A, 84

(1910), pp. 25–46. (Also in: *Scientific Papers by John William Strutt, Baron Rayleigh*, vol. 5, pp. 547–568.)

[49] L. ROBIN, *Fonctions sphériques de Legendre et fonctions sphéroïdales*, vol. II, Gauthier-Villars, Paris, 1958.

[50] A. ROTENBERG, *The calculation of toroidal harmonics*, Math. Comput., 14 (1960), pp. 274–276.

[51] F. W. SCHÄFKE, *Lösungstypen von Differenzengleichungen und Summengleichungen in normierten abelschen Gruppen*, Math. Z., 88 (1965), pp. 61–104.

[52] H. SHINTANI, *Note on Miller's recurrence algorithm*, J. Sci. Hiroshima Univ. Ser. A-I Math., 29 (1965), pp. 121–133.

[53] I. A. STEGUN AND M. ABRAMOWITZ, *Generation of Coulomb wave functions by means of recurrence relations*, Phys. Rev., 98 (1955), pp. 1851–1852.

[54] ———, *Generation of Bessel functions on high speed computers*, Math. Tables Aids Comput., 11 (1957), pp. 255–257.

[55] G. SZEGÖ, *Orthogonal Polynomials*, American Mathematical Society Colloquium Publications, vol. 23, American Mathematical Society, New York, 1959.

[56] D. TEICHROEW, *Use of continued fractions in high speed computing*, Math. Tables Aids Comput., 6 (1952), pp. 127–133.

[57] E. THIELEKER, *Generating functions of confluent hypergeometric functions*, unpublished.

[58] A. TUBIS, *Tables of nonrelativistic Coulomb wave functions*, Los Alamos Scientific Laboratory, LA-2150, Los Alamos, New Mexico, 1958.

[59] H. S. WALL, *Analytic Theory of Continued Fractions*, D. van Nostrand, New York, 1948.

[60] H. WERNER AND R. COLLINGE, *Chebyshev approximations for the gamma function*, Math. Comput., 15 (1961), pp. 195–197.

[61] D. J. WHEELER, *Personal communication*, 1960.

[62] I. ZAMFIRESCU, *An extension of Gauss' formula of integrating improper integrals*, Acad. R. P. Romine Studii Cercetari Matem., 4 (1963), pp. 615–631.

26.3. [35] "An Application of Minimal Solutions of Three-Term Recurrences to Coulomb Wave Functions"

---

[35] "An Application of Minimal Solutions of Three-Term Recurrences to Coulomb Wave Functions," *Aequationes Math*. **2**, 171–176 (1969).

An Application of Minimal Solutions of Three-Term Recurrences to Coulomb Wave Functions

WALTER GAUTSCHI (La Fayette, Indiana)[1])

Dedicated to Professor ALEXANDER OSTROWSKI on his 75th birthday

1. In a recent article [3] we dealt with various recurrence algorithms for computing minimal solutions of linear second-order difference equations–solutions, that is, which grow more slowly than any other linearly independent solution. Such solutions (if they exist) are uniquely determined up to a multiplicative constant. The value of this constant may be determined by specifying *one* initial value, or, more generally, by specifying the value of an infinite series in this solution. In the latter case, it is possible to obtain the respective solution without reference to any initial data.

Among several examples we considered in particular the regular Coulomb wave functions $F_L(\eta,\varrho)$ (see [1] for notations). If we let

$$f_L = \frac{2^L L!}{(2L)!\, C_L(\eta)} F_L(\eta,\varrho), \qquad C_L(\eta) = \frac{2^L e^{-\pi\eta/2}|\Gamma(L+1+i\eta)|}{(2L+1)!}, \tag{1.1}$$

then f_L is a minimal solution of

$$\frac{L[(L+1)^2+\eta^2]}{(L+1)(2L+3)} y_{L+1} - \left[\eta + \frac{L(L+1)}{\varrho}\right] y_L + \frac{L(L+1)}{2L-1} y_{L-1} = 0 \qquad (L=1,2,3,\ldots), \tag{1.2}$$

and we have the following infinite series relation [3],

$$\sum_{L=0}^{\infty} \lambda_L f_L = \varrho\, e^{\omega\varrho}, \qquad \lambda_L = i^L P_L^{(i\eta,-i\eta)}(-i\omega). \tag{1.3}$$

Here, $P_n^{(\alpha,\beta)}(x)$ denotes the Jacobi polynomial of degree n. The parameters η,ϱ,ω are assumed to be real, with $\varrho>0$, $\omega\geqslant 0$. Provided ω is chosen appropriately, the algorithms mentioned above lead to effective schemes of computing f_L over an extended range of the parameters η, ϱ, and for as many values of L as are desired [3, § 7], [4]. An advantage of this approach is the absence of any need to compute $F_0(\eta,\varrho)$, which is known to be tedious, calling for a variety of methods in different regions of the parameters [2].

As one proceeds to large values of η, however, the generation of the coefficients λ_L becomes subject to serious loss of accuracy due to cancellation errors. The reason

[1]) Computer Sciences Department, Purdue University, Lafayette, Indiana, and Argonne National Laboratory, Argonne, Illinois. This work was performed in part under the auspices of the United States Atomic Energy Commission.

Received November 29, 1967.

for this can be traced to a peculiar phenomenon associated with the recurrence relation for the Jacobi polynomials of purely imaginary parameters and variable. The phenomenon, already observed in [3], but left unexplained there, is briefly described in section 2. In section 3 we further elucidate this phenomenon and, at the same time, provide a simple scheme to eliminate the cancellation problem which it causes. The algorithm that so results proves to be effective for an almost unlimited region of the parameters η, ϱ, and L. The only factor restricting its use on a digital computer appears to be the possible occurrence of 'overflow' when $|\eta|$ is very large. These matters will be discussed in section 4.

2. From the well-known recurrence relation for Jacobi polynomials, one finds that λ_L satisfies

$$\lambda_{L+1} = \frac{2L+1}{L+1}\,\omega\,\lambda_L + \frac{L^2+\eta^2}{L(L+1)}\,\lambda_{L-1} \qquad (L=1,2,3,\ldots), \tag{2.1}$$

$$\lambda_0 = 1\,, \qquad \lambda_1 = \omega - \eta\,. \tag{2.2}$$

(In particular, all λ_L are real.) It is readily seen, that (2.1) possesses a minimal solution, whenever $\omega \neq 0$, which we denote by λ'_L, assuming $\lambda'_0 = 1$. The desired solution λ_L is known to be nonminimal [3]. It would appear, therefore, that (2.1) and (2.2) lend themselves conveniently for the accurate generation of λ_L. This is indeed the case as long as η is not too large. As $\eta \to \infty$, it was observed, however, that λ_L 'approaches' the minimal solution λ'_L in the sense that $\lambda_1 - \lambda'_1 \to 0$. Therefore, the initial values of λ_L, as η becomes large, will ultimately be indistinguishable (in finite arithmetic) from those of λ'_L, even though for large L the two solutions behave quite differently. In fact, $\lambda_L \to \infty$ as $L \to \infty$, while $\lambda'_L \to 0$ as $L \to \infty$. It is clear, therefore, that the solution λ_L cannot be determined accurately from initial values, when η is large, unless one resorts to multiple-precision arithmetic.

If it were possible to compute

$$\varepsilon = \lambda_1 - \lambda'_1 \tag{2.3}$$

accurately, then the following device can be used [3].

Let λ''_L be the solution of (2.1) defined by

$$\lambda''_0 = -\lambda'_1\,, \qquad \lambda''_1 = 1\,. \tag{2.4}$$

Then

$$\lambda_L = \lambda'_L + \frac{\varepsilon}{1+\lambda'^2_1}\,(\lambda''_L + \lambda'_1\,\lambda'_L), \tag{2.5}$$

which shows that for small ε the solution λ_L initially follows closely λ'_L until the dominance of λ''_L outweighs the smallness of ε. All terms in (2.5) can be computed

accurately: λ_L' by the algorithms mentioned at the beginning of section 1, ε by assumption, and λ_L'' by straightforward application of (2.1) and (2.4). It may be noted, in this respect, that relative errors δ_0, δ_1 in the initial values λ_0'', λ_1'' give rise to comparable relative errors in λ_L'', when L is large, namely errors approximately equal to $\delta_0\lambda_1'^2/(1+\lambda_1'^2)$ and $\delta_1/(1+\lambda_1'^2)$, respectively. This indicates that the solution λ_L'' is computationally well defined.

3. We proceed now to derive an explicit expression for ε defined in (2.3). We may assume $\eta>0$, in which case $\omega>0$ [cf. (4.2) below].

First of all we note that a minimal solution of (2.1) is given by

$$\lambda_L^{min} = i^{L+1} Q_L^{(i\eta,-i\eta)}(-i\omega),$$

where $Q_n^{(\alpha,\beta)}(x)$ denotes the Jacobi function of the second kind. This follows from the asymptotic formula[2]

$$(x-1)^\alpha (x+1)^\beta Q_n^{(\alpha,\beta)}(x) \sim n^{-1/2}[x-(x^2-1)^{1/2}]^{n+1}\phi(x) \quad (n\to\infty), \tag{3.1}$$

in which x is a real or complex number outside the segment $[-1, 1]$, α and β are real or complex with $\mathrm{Re}\,\alpha>-1$, $\mathrm{Re}\,\beta>-1$, and $\phi(x)\neq 0$ is regular outside of $[-1, 1]$ and independent of n. It is understood in (3.1) that one takes that branch of $x-(x^2-1)^{1/2}$ for which $|x-(x^2-1)^{1/2}|<1$. Letting $x=-i\omega$, $\alpha=i\eta$, $\beta=-i\eta$, one readily obtains

$$\lambda_{L+1}^{min}/\lambda_L^{min} \sim \omega-(\omega^2+1)^{1/2} \quad (L\to\infty).$$

On the other hand, it is known [3, p. 66] that

$$\lambda_{L+1}/\lambda_L \sim \omega+(\omega^2+1)^{1/2} \quad (L\to\infty),$$

showing that λ_L^{min} is indeed minimal. It follows, therefore, that

$$\lambda_L' = \frac{\lambda_L^{min}}{\lambda_0^{min}} = i^L \frac{Q_L^{(i\eta,-i\eta)}(-i\omega)}{Q_0^{(i\eta,-i\eta)}(-i\omega)}.$$

In order to evaluate λ_0^{min}, we make use of [5, p. 75]

$$Q_0^{(\alpha,\beta)}(x) = 2^{\alpha+\beta}\frac{\Gamma(\alpha+1)\,\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)}(x-1)^{-\alpha-1}(x+1)^{-\beta} \times F\left(\alpha+1, 1; \alpha+\beta+2; \frac{2}{1-x}\right),$$

where $F(a, b; c; x)$ denotes the hypergeometric function. Assuming $\alpha+\beta=0$ (as is the

[2] See [5, p. 223], where the result is obtained for real $\alpha>-1$, $\beta>-1$. The derivation by the method of steepest descent, however, is valid also when α and β are complex, with $\mathrm{Re}\,\alpha>-1$, $\mathrm{Re}\,\beta>-1$.

case in our context), we have

$$F\left(\alpha+1, 1; 2; \frac{2}{1-x}\right) = \sum_{v=0}^{\infty} \frac{(\alpha+1)(\alpha+2)\cdots(\alpha+v)}{(v+1)!}\left(\frac{2}{1-x}\right)^v$$
$$= \sum_{v=0}^{\infty} \frac{(-1)^v}{v+1}\binom{-\alpha-1}{v}\left(\frac{2}{1-x}\right)^v,$$

which, on applying

$$\sum_{v=0}^{\infty}(-1)^v\binom{-\alpha-1}{v}\frac{z^v}{v+1} = \frac{1}{z}\int_0^z (1-t)^{-\alpha-1}\,dt = \frac{1}{\alpha z}\left[(1-z)^{-\alpha}-1\right],$$

becomes

$$F\left(\alpha+1, 1; 2; \frac{2}{1-x}\right) = \frac{1-x}{2\alpha}\left[\left(1-\frac{2}{1-x}\right)^{-\alpha}-1\right].$$

Therefore,

$$Q_0^{(\alpha,-\alpha)}(x) = \Gamma(\alpha+1)\,\Gamma(-\alpha+1)\,(x-1)^{-\alpha-1}(x+1)^{\alpha}\frac{1-x}{2\alpha}\left[\left(1-\frac{2}{1-x}\right)^{-\alpha}-1\right]$$
$$= -\frac{\Gamma(\alpha+1)\,\Gamma(-\alpha+1)}{2\alpha}\left[1-(x+1)^{\alpha}(x-1)^{-\alpha}\right],$$

provided that

$$-\pi < \arg(x-1) + \arg\left(1-\frac{2}{1-x}\right) \leqslant \pi.$$

Letting $x=-i\omega$ (which satisfies this condition), and $\alpha=i\eta$, we obtain

$$\lambda_0^{min} = iQ_0^{(i\eta,-i\eta)}(-i\omega) = -\frac{\Gamma(1+i\eta)\,\Gamma(1-i\eta)}{2\eta}\left[1-(1-i\omega)^{i\eta}(-1-i\omega)^{-i\eta}\right].$$

By an elementary computation one finds that

$$(1-i\omega)^{i\eta}(-1-i\omega)^{-i\eta} = e^{-2\eta\phi},$$

where

$$\phi = \arctan\frac{1}{\omega}.$$

Since, furthermore, $\Gamma(1+i\eta)\Gamma(1-i\eta)=\pi\eta/\sinh(\pi\eta)$, we finally obtain

$$\lambda_0^{min} = -\frac{\pi}{2\sinh(\pi\eta)}(1-e^{-2\eta\phi}). \tag{3.2}$$

To determine λ_1^{min}, we use [5, p. 80]

$$Q_1^{(\alpha,\beta)}(x) = \tfrac{1}{2}[(\alpha+\beta+2)\,x+\alpha-\beta]\,Q_0^{(\alpha,\beta)}(x) - 2^{\alpha+\beta-1}(\alpha+\beta+2)\frac{\Gamma(\alpha+1)\,\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)}(x-1)^{-\alpha}(x+1)^{-\beta},$$

and find

$$\lambda_1^{min} = -\,Q_1^{(i\eta,\,-i\eta)}(-i\omega) = (\omega-\eta)\,\lambda_0^{min} + \frac{\pi\eta}{\sinh(\pi\eta)}\,e^{-2\eta\phi}. \tag{3.3}$$

Combining (3.2) and (3.3), we get

$$\lambda_1' = \frac{\lambda_1^{min}}{\lambda_0^{min}} = \omega - \eta - \frac{2\eta}{e^{2\eta\phi}-1}. \tag{3.4}$$

Therefore, in view of (2.2), (2.3), the desired expression for ε is

$$\varepsilon = \lambda_1 - \lambda_1' = \frac{2\eta}{e^{2\eta\phi}-1}, \qquad \phi = \arctan\frac{1}{\omega}. \tag{3.5}$$

This result both explains the phenomenon described in section 2, and provides a simple formula to compute ε, and thus λ_L by means of (2.5).

4. The values of the exponentials in (1.1), (1.3), and (3.5), for large $|\eta|$, may become so large (or so small) as to exceed the range of permissible floating point numbers on a particular computer. If this range is given by $[10^{-R}, 10^{R}]$, such 'overflow' will occur in any of the following three cases,

$$e^{\pi|\eta|/2} > 10^R, \quad e^{\omega\varrho} > 10^R, \quad e^{2|\eta|\phi} > 10^R. \tag{4.1}$$

By definition of ω [3, p. 65], we have

$$\omega\varrho = \begin{cases} \pi\eta & (\tau \geqslant 1), \\ \eta\left[\pi - 2\arccos\sqrt{\tau} + 2\sqrt{\tau(1-\tau)}\right] & (0<\tau<1), \\ 0 & (\tau<0), \end{cases} \tag{4.2}$$

where $\tau = \varrho/2\eta$. Since $0 \leqslant \pi - 2\arccos\sqrt{\tau} + 2\sqrt{\tau(1-\tau)} \leqslant \pi$ for $0 \leqslant \tau \leqslant 1$, it follows that $0 \leqslant \omega\varrho \leqslant \pi|\eta|$. Moreover, $2|\eta|\phi = 2|\eta|\arctan(1/\omega) \leqslant \pi|\eta|$. Therefore, none of the cases in (4.1) will arise if $|\eta|$ is restricted to satisfy

$$e^{\pi|\eta|} \leqslant 10^R, \quad \text{i.e.} \quad |\eta| \leqslant \frac{R\ln 10}{\pi} = (.7329\ldots)\,R. \tag{4.3}$$

On the CDC 3600, e.g., one has $R = 308$, so that on this computer the restriction (4.3) amounts to $|\eta| \leqslant 225.7\ldots$.

Another place where overflow may occur is in the generation of the quantities λ_L'' by (2.1), (2.4), when $\eta>0$. As $L\to\infty$, one finds

$$\lambda_L'' \sim \frac{1+\lambda_1'^2}{2\eta}(e^{2\eta\phi}-1)\lambda_L,$$

and in view of the known asymptotic behavior of λ_L (cf. [3, p. 66]), and (3.4),

$$\lambda_L'' \simeq \frac{1+(\omega-\eta)^2}{2\eta}(2\pi L)^{-1/2}(1+\omega^2)^{-1/4}e^{\eta\phi}[\omega+\sqrt{\omega^2+1}]^{L+1/2},$$

having assumed $\exp(2\eta\phi)\gg 1$. Roughly, then, $\lambda_L''\simeq\exp(\eta\phi)\,[\omega+\sqrt{\omega^2+1}]^L$, and to avoid overflow we should have

$$e^{\eta\phi}[\omega+\sqrt{\omega^2+1}]^L \leqslant 10^R.$$

Letting v denote the largest value of L for which λ_L'' is required (an estimate for v may be found in [3, p. 69]) the last inequality is satisfied if

$$\eta \arctan\left(\frac{1}{\omega}\right)+v\ln(\omega+\sqrt{\omega^2+1})\leqslant R\ln 10. \tag{4.4}$$

We note that v depends not only on the parameters η, ϱ, and L, but also on the desired accuracy for $F_L(\eta,\varrho)$. If six significant digits are required, for example, it was found that (4.4) holds true for $0\leqslant\eta\leqslant 100$, $0.1\leqslant\varrho\leqslant 200$, $0\leqslant L\leqslant 100$, if $R=308$ as before.

The region in which our recurrence algorithm is applicable (using standard floating point artithmetic) is thus delineated by the two inequalities (4.3) and (4.4).

REFERENCES

[1] Abramowitz, M., *Coulomb Wave Functions*, in: *Handbook of Mathematical Functions* (edited by M. Abramowitz and I.A. Stegun), NBS Applied Mathematics Series, No. 55 (Washington, D.C. 1964), Chapt. 14.

[2] Fröberg, C.E., *Numerical Treatment of Coulomb Wave Functions*, Rev. Modern Phys. *27*, 399–411 (1955).

[3] Gautschi, W., *Computational Aspects of Three-Term Recurrence Relations*, SIAM Rev. *9*, 24–82 (1967).

[4] Gautschi, W., *Algoritm 292, Regular Couloumb wave functions*. Comm. ACM *9*, 793–795 (1966).

[5] Szegő, G., *Orthogonal Polynomials*, Revised ed., AMS Colloquium Publ., Vol. 23, (New York 1959.)

Purdue University

26.4. [37] (with B. J. Klein) "Recursive Computation of Certain Derivatives — A Study of Error Propagation"

---

[37] (with B. J. Klein) "Recursive Computation of Certain Derivatives — A Study of Error Propagation," *Comm. ACM* **13**, 7–9 (1970).

Recursive Computation of Certain Derivatives—A Study of Error Propagation

WALTER GAUTSCHI
Purdue University, Lafayette, Indiana*

AND

BRUCE J. KLEIN
Virginia Polytechnic Institute,† Blacksburg, Virginia

A brief study is made of the propagation of errors in linear first-order difference equations. The recursive computation of successive derivatives of e^x/x and $(\cos x)/x$ is considered as an illustration.

KEY WORDS AND KEY PHRASES: recursive computation, successive derivatives, error propagation
CR CATEGORIES: 5.11, 5.12

1. Introduction

In [2] one of us published algorithms for computing successive derivatives of e^x/x, $(\cos x)/x$, and $(\sin x)/x$. It was brought to our attention [5] that the first two of these algorithms are subject to substantial loss of accuracy if x (or $|x|$ in the case of the second algorithm) is large and n, the order of derivative, is larger than $|x|$. In the following we examine the reasons responsible for this difficulty and suggest ways in which it may be overcome. Revised algorithms implementing the results of this article appear as Remark on Algorithm 282 in the Algorithms section of this issue (see footnote).

Although hardly more than an isolated example,[1] the question discussed here well illustrates the pitfalls inherent in the indiscriminate use of recurrence relations. It may also serve to remind us of the computational limitations of analytic formula manipulation systems.

Consider, for example, the derivatives

$$d_n(x) = \frac{d^n}{dx^n}\left(\frac{e^x}{x}\right), \qquad n = 0, 1, 2, \ldots . \tag{1.1}$$

Work supported by the National Aeronautics and Space Administration (NASA) under Grant NGR 15-005-039. This paper gives the theoretical background of Remark on Algorithm 282 "Derivatives of e^x/x, cos $(x)/x$, and sin $(x)/x$" by the same authors, which appears on pages 53–54.

* Department of Computer Sciences.
† College of Arts and Sciences.

[1] We note, however, that the function d_n in (1.1) is of some relevance in molecular structure calculations by virtue of $A_n(1,\alpha) = -d_n(-\alpha)$, $A_n(-1,\alpha) = (-1)^n d_n(\alpha)$, where $A_n(\sigma,\alpha) = \int_\sigma^\infty e^{-\alpha t} t^n\,dt$ are auxiliary "molecular integrals" (cf. [4, 6]).

Analytic differentiation yields

$$d_n(x) = (-1)^n \frac{n!}{x^{n+1}} e^x e_n(-x), \tag{1.2}$$

where

$$e_n(z) = \sum_{k=0}^{n} \frac{z^k}{k!}. \tag{1.3}$$

Formula manipulation systems most likely would deal with (1.1) by effectively evaluating the expression in (1.2). Note, however, that for x positive and large, and $n \gg x$, the dominant term in the sum for $e_n(-x)$ has the order of magnitude $e^x/\sqrt{(2\pi x)}$, while the sum itself is close to e^{-x}. For such values of x and n, the evaluation of (1.2) thus involves considerable cancellation of leading digits, the resulting loss of accuracy amounting to about $\log_{10} e^{2x} = (.868\ldots)x$ decimal digits.

Alternatively, one might try to compute the desired derivatives recursively, as in [2], using

$$d_n(x) = -\frac{n}{x} d_{n-1}(x) + \frac{e^x}{x}, \qquad n = 1, 2, 3, \ldots, \quad d_0(x) = \frac{e^x}{x}. \tag{1.4}$$

While, technically speaking, this recursion is stable, it will be seen that the cancellation problem reappears with the same devastating force.

2. Error Propagation in Linear First-order Difference Equations

The recurrence relation (1.4) is an example of a first-order linear difference equation

$$y_n = a_n y_{n-1} + b_n, \qquad n = 1, 2, 3, \ldots, \quad a_n \neq 0. \tag{2.1}$$

We consider solutions on the set $\mathfrak{N}$ of nonnegative integers n. Given a particular solution $\{f_n\}$ of (2.1) to be computed, we wish to examine the influence of a single error at $m \in \mathfrak{N}$ upon the value of f_n at any other $n \in \mathfrak{N}$. Since the solution $\{f_n\}$ may vary considerably in magnitude, it is appropriate to consider *relative* errors and restrict attention to the subset $\mathfrak{N}_0 \subset \mathfrak{N}$ on which $f_n \neq 0$. Assuming for simplicity that $f_0 \neq 0$, the question can easily be answered as follows (cf. [1]).

Let $\{\tilde{f}_n\}$ denote the "perturbed" solution of (2.1) corresponding to the starting value $\tilde{f}_m = f_m(1+\epsilon)$, $m \in \mathfrak{N}_0$. Then for any $n \in \mathfrak{N}_0$ we have

$$\tilde{f}_n = f_n\left(1 + \frac{\rho_n}{\rho_m}\epsilon\right), \tag{2.2}$$

where[2]

$$\rho_n = \frac{f_0 h_n}{f_n}, \qquad h_n = a_n a_{n-1} \ldots a_1. \tag{2.3}$$

[2] The factor f_0 in the definition of ρ_n is included only for the purpose of normalization, making $\rho_0 = 1$.

A relative error ϵ introduced at m thus induces a relative error $(\rho_n/\rho_m)\epsilon$ at n. In particular, the error is magnified if $|\rho_n| > |\rho_m|$ and damped if $|\rho_n| < |\rho_m|$. The quantities ρ_n will be referred to as "amplification factors."

The behavior of the function $\{|\rho_n|\}$ clearly determines the error propagation pattern associated with the particular solution $\{f_n\}$ of (2.1). If there is any choice of direction in which the recursion (2.1) can be employed, then the direction in which $|\rho_n|$ decreases (or has a tendency to decrease) is generally the one to be preferred. Following this direction, errors introduced at each step of the recursion (due to rounding, for example) have a tendency to be consistently damped out. Proceeding in direction of increasing $|\rho_n|$ is tolerable only if the maximum error amplification remains within acceptable limits.

3. Successive derivatives of e^x/x

From (1.2) and (2.3) we find that the amplification factors ρ_n associated with the solution (1.1) of the difference equation (1.4) are given by

$$\rho_n(x) = \frac{1}{e_n(-x)}. \tag{3.1}$$

If $x < 0$, then $|\rho_n|$ decreases monotonically from 1 to $e^{-|x|}$. In this case the recursion (1.4) is properly applied in the forward direction for all $n > 0$. If $x > 0$, the behavior of $|\rho_n|$ is as shown in Figure 1. Disregarding relatively small values of x (for which $|\rho_n|$ remains within acceptable limits for all $n \geq 0$), it is seen that $|\rho_n|$ initially decreases until it reaches a minimum value near $n_0 = [x]$, and from then on increases, reaching the limit $|\rho_\infty| = e^x$ rather abruptly. The recursion (1.4) is now properly applied in the forward direction on the interval $0 < n \leq n_0$, and in the backward direction on $n_0 < n < \infty$, unless an error amplification of $|\rho_\infty/\rho_{n_0}|$ is tolerable, in which case forward recursion may be used on the whole interval $0 < n < \infty$.

We note that $|e_n(-n)| \sim e^n/2\sqrt{(2\pi n)}$ as $n \to \infty$, from which it follows that the maximum error amplification is approximately $e^{2x}/2\sqrt{(2\pi x)}$, when x is large.

The graphs in Figure 1 may be interpreted as follows. Writing $d_n(x)$ in the form

$$d_n(x) = (-1)^n \frac{n!}{x^{n+1}} + \int_0^1 t^n e^{xt}\, dt \tag{3.2}$$

[by using the remainder term of the exponential series in (1.2)] and assuming $x > 0$ large, one observes that the integral on the right of (3.2) initially dominates, until n is large enough to make the first term of comparable magnitude. From this point on, the first term quickly becomes the dominant term. As long as the integral dominates, $d_n(x)$ varies relatively slowly with n, so that by (2.3) $|\rho_n|$ is approximately proportional to $|h_n| = n!\,x^{-n}$. Once the first term takes over, $|\rho_n|$ becomes constant, equal to e^x. Therefore, the curves in Figure 1, up to a scale factor, are essentially those for $n!\,x^{-n}$, levelled off at the value of n for which the integral in (3.2) becomes negligible.

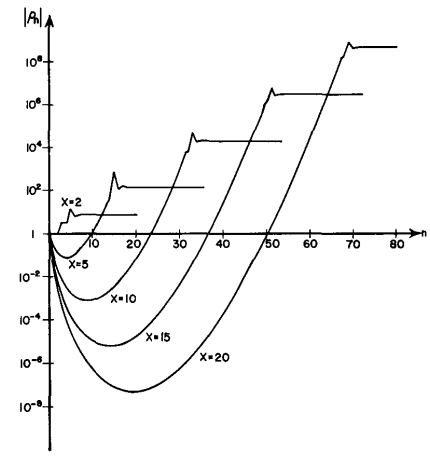


FIG. 1. Amplification factors $|\rho_n(x)|$ of (3.1), for $0 \leq n \leq 80$, $x = 2, 5, 10, 15, 20$

It remains to consider the question of computing an appropriate starting value in cases where backward recurrence is called for. From the remarks just made, it is clear that $d_n(x)$ can be approximated by

$$q_n(x) = (-1)^n \frac{n!}{x^{n+1}} \tag{3.3}$$

to any degree of accuracy, if n is taken sufficiently large. To analyze this more carefully, observe that the integral in (3.2) is bounded by $e^x/(n+1)$, and that $n! > (n/e)^n$ for every integer $n \geq 1$. Therefore,

$$\left|\frac{d_n - q_n}{q_n}\right| = \frac{x^{n+1}}{n!}\int_0^1 t^n e^{xt}\, dt < \frac{x^{n+1}}{(n+1)!}e^x < \left(\frac{ex}{n+1}\right)^{n+1} e^x,$$

from which it follows that $|(d_n - q_n)/q_n| \leq \delta$ $(0 < \delta < 1)$, and consequently $|(d_n - q_n)/d_n| \leq \delta/(1-\delta)$, as soon as n is large enough to satisfy

$$\left(\frac{ex}{n+1}\right)^{n+1} e^x \leq \delta. \tag{3.4}$$

In particular, q_n approximates d_n to s significant digits if (3.4) holds with $\delta = \frac{1}{2}10^{-s}$. Taking logarithms, this condition amounts to

$$\frac{n+1}{ex}\ln\frac{n+1}{ex} \geq \frac{x + s\ln 10 + \ln 2}{ex},$$

which in turn is equivalent to

$$n + 1 \geq ex\, t\left(\frac{x + s\ln 10 + \ln 2}{ex}\right), \tag{3.5}$$

where $t(y)$ denotes the inverse function of $y = t\ln t$. (Low-accuracy approximations to $t(y)$ are obtained in another context in [3, p. 51].) Thus, if n^0 is the smallest integer n satisfying (3.5), then $q_n(x)$ in (3.3) may be used to approximate $d_n(x)$ (to s significant digits) for $n \geq n^0$, while backward recursion in (1.4) may be used to obtain $d_n(x)$ for $n_0 \leq n < n^0$.

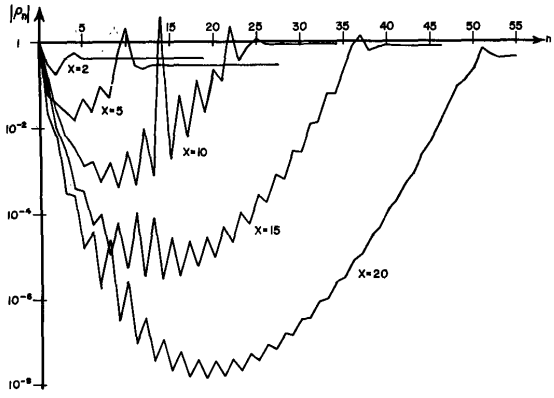


FIG. 2. Amplification factors $|\rho_n(x)|$ of (4.3), for $0 \leq n \leq 55$, $x = 2, 5, 10, 15, 20$

4. Successive Derivatives of (cos x)/x and (sin x)/x

The derivatives

$$c_n(x) = \frac{d^n}{dx^n}\left(\frac{\cos x}{x}\right) \tag{4.1}$$

satisfy the difference equation

$$c_n(x) = -\frac{n}{x}c_{n-1}(x) + \frac{1}{x}\mathrm{Re}(i^n e^{ix}), \qquad n = 1, 2, 3, \cdots, \tag{4.2}$$

and the associated amplification factors ρ_n are now

$$\rho_n(x) = \frac{\cos x}{\mathrm{Re}\,[e^{ix}e_n(-ix)]}. \tag{4.3}$$

Clearly, $\rho_n(-x) = \rho_n(x)$. The behavior of $|\rho_n|$ is shown in Figure 2. The graphs are basically the same as those in Figure 1, except that they are leveled off at an earlier stage (due to the limiting value now being $\rho_\infty = \cos x$) and are not nearly as smooth.

The recurrence (4.2) is again properly applied in the forward direction for $0 < n \leq n_0$ ($n_0 = [|x|]$), and should be used in this backward direction for $n_0 < n < \infty$, unless the maximum error amplification $|1/\rho_{n_0}|$ (now approximately half as large as in the case of $d_n(x)$) is within tolerable limits. Due to the fluctuations in $|\rho_n|$, occasional losses of significant digits must be expected, even if the recursion is used in the proper direction. Loss of significance is apt to occur for those values of n for which $|c_n(x)|$ is exceptionally small.

The identity

$$c_n(x) = \frac{(-1)^n n!}{x^{n+1}} + \int_0^1 t^n \,\mathrm{Re}\,[i^{n+1}e^{ixt}]\,dt \tag{4.4}$$

permits us to interpret the graphs of Figure 2 in a similar manner as we did previously for the graphs of Figure 1. It also follows from (4.4) that $q_n(x)$ in (3.3) can be used to approximate $c_n(x)$ to s significant digits for all n satisfying

$$n + 1 \geq e|x|\,t\left(\frac{s \ln 10 + \ln 2}{e|x|}\right).$$

Replacing "Re" by "Im" in (4.2) and (4.3), and "cos x" by "sin x" in (4.3), one obtains the difference equation and associated amplification factors for the derivatives $s_n(x) = (d^n/dx^n)(\sin x/x)$. The graphs of $|\rho_n|$ in this case resemble those of Figure 2, except that no leveling-off occurs, since $\mathrm{Im}[e^{ix}e_n(-ix)] \to 0$ as $n \to \infty$.

RECEIVED MAY, 1969

REFERENCES

1. GAUTSCHI, W. Recursive computation of certain integrals. *J. ACM 8* (1961), 21–40.
2. ——. Algorithm 282—Derivatives of e^x/x, cos $(x)/x$, and sin $(x)/x$. *Comm. ACM 9*, 4 (Apr. 1966), 272.
3. ——. Computational aspects of three-term recurrence relations. *SIAM Rev. 9* (1967), 24–82.
4. KOTANI, M., ET AL. Tables of molecular integrals. Maruzen Co., Tokyo, 1963.
5. MATUSKA, WALTER, JR. Personal communication, 1967.
6. PREUSS, H. Integraltafeln zur Quantenchemie. Springer, Berlin, 1956.

Lowe—*cont'd from page 6*

is in preparation, and more research is required in that area. An important topic for future investigation is a comparison of performance improvement and cost of segmentation for Boolean and probabilistic methods. Such an investigation could well include empirical testing.

RECEIVED FEBRUARY, 1969; REVISED JULY, 1969

REFERENCES

1. LOWE, THOMAS C. The influence of data base charecteristics and usage on direct access file organization. *J. ACM 15*, 4 (Oct. 1968), 535–548.
2. LOWRY, EDWARD S., AND MEDLOCK, C. W. Object code optimization. *Comm. ACM* 12, 1 (Jan. 1969), 13–22.
3. LOWE, THOMAS C. Analysis of Boolean models for time-shared paged environments. *Comm. ACM 12*, 4 (Apr. 1969), 199–205.
4. RAMAMOORTHY, C. V. Analysis of graphs by connectivity considerations. *J. ACM 13*, 2 (Apr. 1966), 211–222.
5. RAMAMOORTHY, C. V. The analytic design of a dynamic look ahead and program segmenting scheme for multiprogrammed computers. Proc. ACM 21st Nat. Conf., 1966, Thompson Book Co., Washington, D.C., pp. 229–239.
6. PROSSER, R. T. Applications of Boolean matrices to the analysis of flow diagrams. Proc. Eastern Joint Comp. Conf., Vol. 16, Dec. 1959, Spartan Books, New York, pp. 133–138.
7. ROSENBLATT, DAVID. On the graphs and asymptotic forms of finite Boolean relation matrices and stochastic matrices. *Naval Res. Logist. Quart. 4* (June 1967), 32–37.
8. LOWE, THOMAS C. An algorithm for rapid calculation of products of Boolean matrices. *Software Age 2* (Mar. 1968), 36–37.

26.5. [135] "IS THE RECURRENCE RELATION FOR ORTHOGONAL POLYNOMIALS ALWAYS STABLE?"

---

[135] "Is the Recurrence Relation for Orthogonal Polynomials Always Stable?," *BIT* **33**, 277–284 (1993).

BIT 33 (1993), 277–284.

IS THE RECURRENCE RELATION FOR ORTHOGONAL POLYNOMIALS ALWAYS STABLE?*

WALTER GAUTSCHI

Dept. of Computer Sciences, Purdue University, West Lafayette, IN 47907-1398, USA

Abstract.

Attention is drawn to a phenomenon of "pseudostability" in connection with the three-term recurrence relation for discrete orthogonal polynomials. The computational implications of this phenomenon are illustrated in the case of discrete Legendre and Krawtchouk polynomials. The phenomenon also helps to explain a form of instability in Stieltjes's procedure for generating recursion coefficients of discrete orthogonal polynomials.

AMS(MOS) Subject classification: 33-04, 35C50, 39A11, 65D20.

1. It is our experience, and the experience of many others, that the basic three-term recurrence relation for orthogonal polynomials is generally an excellent means of computing these polynomials, both within the interval of orthogonality and outside of it. The same recurrence relation, on the other hand, is known to become unstable if one attempts to use it for computing other solutions, for example, the minimal solution when the argument is outside the interval of orthogonality (cf. [4]), or the Hilbert transform of Jacobi polynomials when one of the Jacobi parameters is large and the argument close to 1 (cf. [8, §4]). Here we wish to point out instances of "pseudostability" in connection with the computation of discrete orthogonal polynomials.

Our discussion sheds new light on a hitherto unexplained phenomenon of instability that afflicts the Stieltjes procedure for generating the recursion coefficients of discrete orthogonal polynomials (cf. [6, §8]).

2. The (monic) orthogonal polynomials $\{\pi_n(x; d\lambda)\}$ corresponding to a positive measure $d\lambda$ on the real line are known to satisfy a three-term recurrence relation

$$y_{k+1} = (x - \alpha_k)y_k - \beta_k y_{k-1}, \qquad k = 0, 1, 2, \ldots, \tag{2.1}$$

where $\alpha_k = \alpha_k(d\lambda) \in \mathbb{R}$, $\beta_k = \beta_k(d\lambda) > 0$ are coefficients uniquely determined by the measure $d\lambda$. We are interested in the stability of this recurrence relation with respect to initial values y_0, y_1. That is, letting $\{y_n^*\}$ denote the solution of (2.1) corresponding

* Work supported in part by the National Science Foundation under grant DMS-9023403.
Received September 1992.

to slightly perturbed initial values $y_0^* = y_0(1 + \varepsilon_0)$, $y_1^* = y_1(1 + \varepsilon_1)$, we like to know how much y_n^* differs from y_n for values of n larger than 1. This is an elementary exercise in the theory of linear difference equations. The answers is

$$y_n^* - y_n = \frac{(y_0 y_n z_1 - y_0 y_1 z_n)\varepsilon_0 - (y_1 y_n z_0 - y_0 y_1 z_n)\varepsilon_1}{y_0 z_1 - y_1 z_0}, \tag{2.2}$$

where $\{z_n\}$ is an arbitrary solution of (2.1) linearly independent of $\{y_n\}$. The factors multiplying ε_0 and ε_1 on the right of (2.2), or more precisely, their moduli, determine the extent of error amplification in the absolute error $y_n^* - y_n$. Normally, if $y_n \neq 0$, we prefer to consider relative errors $(y_n^* - y_n)/y_n$. Appropriate amplification factors are then given by

$$\omega_n(x) = \begin{cases} \dfrac{|y_0 z_1 - y_0 y_1 (z_n/y_n)| + |y_1 z_0 - y_0 y_1 (z_n/y_n)|}{|y_0 z_1 - y_1 z_0|} & \text{if } y_n \neq 0, \\ 2\,|y_0 y_1 z_n| / |y_0 z_1 - y_1 z_0| & \text{if } y_n = 0. \end{cases} \tag{2.3}$$

We say that the recurrence relation (2.1) is *unstable* for the solution $\{y_n\}$ if $\omega_n \to \infty$ as $n \to \infty$. In view of (2.3), if $y_n \neq 0$ for n sufficiently large, this is equivalent to $\lim_{n\to\infty} |z_n/y_n| = \infty$, i.e., to $\{y_n\}$ being a minimal solution of (2.1). There are various "backward recurrence" algorithms (see, e.g., [3], [10]) that can be used to compute minimal solutions. A more perfidious predicament (since there are no easy countermeasures) is *pseudostability*; by this we mean that ω_n is uniformly bounded as $n \to \infty$, but the bound is extremely large. We refer to pseudostability also in the case (of particular interest here) where n can assume only a finite number of values, and some of the ω_n are extremely large. (Isolated large values of ω_n may be due to "near zeros", $y_n \approx 0$, and may well be harmless in practice.)

In the case of orthogonal polynomials $y_n = \pi_n(x; d\lambda)$, we have $y_{-1} = 0$, $y_0 = 1$, and we may choose for z_n the solution of (2.1) satisfying $z_{-1} = 1$, $z_0 = 0$. The amplification factor ω_n in (2.3) then simplifies to

$$\omega_n(x) = \begin{cases} \left|1 - \dfrac{y_1}{z_1}\dfrac{z_n}{y_n}\right| + \left|\dfrac{y_1}{z_1}\dfrac{z_n}{y_n}\right| & \text{if } y_n \neq 0, \\ 2\left|\dfrac{y_1}{z_1} z_n\right| & \text{if } y_n = 0, \end{cases} \qquad y_n = \pi_n(x; d\lambda). \tag{2.4}$$

The quantities ω_n in (2.3) and (2.4) characterize stability with respect to *initial values* y_0, y_1. A more complete picture of stability is provided by the following stability measure relative to arbitrary *starting values* y_m, y_{m+1}:

$$\omega_{m\to n}(x) = \begin{cases} \dfrac{|y_m z_{m+1} - y_m y_{m+1}(z_n/y_n)| + |y_{m+1} z_m - y_m y_{m+1}(z_n/y_n)|}{|y_m z_{m+1} - y_{m+1} z_m|} & \text{if } y_n \neq 0, \\ \dfrac{2|y_m y_{m+1} z_n|}{|y_m z_{m+1} - y_{m+1} z_m|} & \text{if } y_n = 0. \end{cases} \tag{2.5}$$

This number indicates to what extent errors committed at $k = m$ and $k = m + 1$ are amplified at $k = n$. We may have $n \geq m$ or $n < m$; clearly, $\omega_{m\to m} = \omega_{m\to m+1} = 1$ if $y_m y_{m+1} \neq 0$, and $\omega_n = \omega_{0\to n}$.

3. We now apply the tools of §2 to discrete orthogonal polynomials. Here, $d\lambda = d\lambda_N$ is a discrete Dirac measure

$$d\lambda_N(x) = \sum_{v=1}^{N} \omega_v \delta(x - x_v)\, dx, \tag{3.1}$$

where

$$x_1 < x_2 < x_3 < \cdots < x_N, \quad \omega_v > 0, \quad v = 1, 2, \ldots, N. \tag{3.2}$$

In this case there are exactly N orthogonal polynomials, $\pi_k(\cdot, d\lambda_N)$, $k = 0, 1, \ldots, N-1$, and the same number of associated recursion coefficients $\alpha_k(d\lambda_N)$ and $\beta_k(d\lambda_N)$, $k = 0, 1, \ldots, N-1$. We present two examples, believed to be representative for a wide class of discrete orthogonal polynomials, exhibiting phenomena of pseudostability. A third example illustrates a case of almost perfect stability. All our computations were done on the Cyber 205, which has machine precisions of 7.11×10^{-15} and 5.05×10^{-29} in single, resp. double precision.

EXAMPLE 3.1. Equally spaced and equally weighted measure $d\lambda_N$: $x_v = -1 + 2(v-1)/(N-1)$, $\omega_v = 2/N$, $v = 1, 2, \ldots, N$.

Here, the recursion coefficients are explicitly known:

$$\alpha_k = 0, \quad k = 0, 1, \ldots, N-1; \tag{3.3}$$

$$\beta_0 = 2, \quad \beta_k = \left(1 + \frac{1}{N-1}\right)^2 \left(1 - \left(\frac{k}{N}\right)^2\right)\left(4 - \frac{1}{k^2}\right)^{-1}, \quad k = 1, 2, \ldots, N-1.$$

For fixed k, and $N \to \infty$, they converge to the respective recursion coefficients for monic Legendre polynomials.

It turns out that in this example the recurrence relation (2.1) applied with $x = x_v$ is generally pseudostable, particularly so if $v \ll N/2$ and N is large. (There is of course symmetry with respect to the midpoint of $[x_1, x_N]$.) We illustrate this in Figure 3.1, which depicts the amplification factor $\omega_n(x)$ of (2.4) on a logarithmic scale for $1 \leq n \leq N-1$, $N = 40$, $x = x_v$, $v = 1, 5, 10, 20$. There is clearly a trend of rapidly increasing $\omega_n(x)$ as n approaches $N-1$ when x is near the ends of the interval $[x_1, x_N]$. Near the center of the interval, the recurrence is quite stable.

The graphs of Figure 3.1 are also indicative of stability with regard to starting values other than y_0, y_1, as is shown in Table 3.1. (Integers in parentheses denote decimal exponents.) Here, the quantity

$$\Omega(x_v) = \max_{0 \leq m < n \leq N-1} \omega_{m\to n}(x_v) \tag{3.4}$$

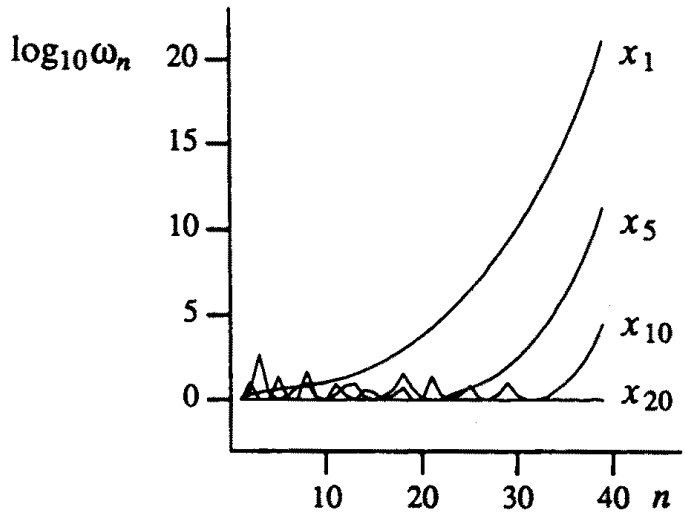


Fig. 3.1. *Amplification factors for Example* 3.1.

is tabulated for selected values of v along with the integers $m = m_v$ and $n = n_v$ for which the maximum in (3.4) is attained, and the maximum relative single-precision error observed in the recurrence.

Table 3.1. *Pseudostability of discrete Legendre polynomials.*

v	$\Omega(x_v)$	m_v	n_v	max err
1	3.771(21)	4	39	1.0310(8)
5	4.148(11)	22	39	3.4959(−2)
10	6.912(4)	32	39	3.2338(−8)
20	3.715(0)	25	38	1.1081(−12)

EXAMPLE 3.2. Krawtchouk polynomials: $x_v = v - 1, \quad \omega_v = \binom{N-1}{v-1} p^{v-1} q^{N-v}$, $v = 1, 2, \ldots, N$, with $p > 0$, $q > 0$, and $p + q = 1$.

Here, too, the recursion coefficients are known explicitly (see, e.g., [1, Eq. (3.5) on p. 161 and Eq. (3.2) on p. 176]),

$$\begin{aligned} &\alpha_k = qk + p(N - 1 - k), \quad k = 0, 1, \ldots, N - 1; \\ &\beta_0 = 1, \ \beta_k = k(N - k)pq, \quad k = 1, 2, \ldots, N - 1. \end{aligned} \tag{3.5}$$

Figure 3.2 shows severe cases of pseudostability when $p = 0.1$, $q = 0.9$, $N = 40$, and the recurrence formula (2.1) is applied for $x = x_1, x_5, x_{10}$ and x_{20}. Unlike the previous example, Figure 3.2 does not indicate the full extent of pseudostability, especially not in the case $x = x_{20}$. Indeed, the more general stability measure $\omega_{m \to n}$ in (2.5) reveals considerable additional error amplification. This can be seen from

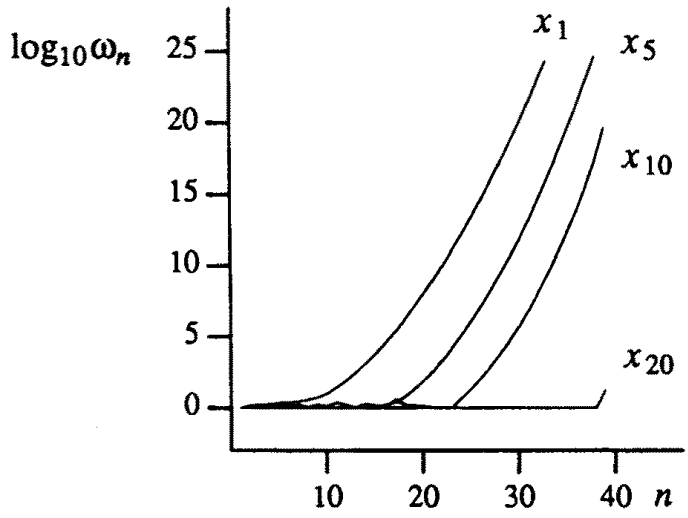


Fig. 3.2. *Amplification factors for Example* 3.2.

Table 3.2, which displays the analogous information as Table 3.1. If p increases, the severity of pseudostability diminishes, the lowest level being attained for $p = q = 1/2$. In this case the quantities in the second and fifth column of Table 3.2 become 7.266(10), 5.797(5), 4.743(2), 5.173(0) and 4.702(−4), 1.217(−8), 1.382(−11), 1.401(−12), respectively.

Table 3.2. *Pseudostability of Krawtchouk polynomials with* $p = 0.1$, $q = 0.9$.

v	$\Omega(x_v)$	m_v	n_v	max err
1	8.931(25)	4	39	4.859(11)
5	2.053(26)	13	39	4.339(12)
10	5.041(20)	20	39	2.741(7)
20	6.115(8)	30	39	4.995(−5)

The occurrence of pseudostability in Example 3.1 and 3.2 may be due, at least in part, to the equispacing of the abscissae x_v. Choosing as abscissae the Chebyshev points on $[-1, 1]$ indeed may lead to perfectly stable recurrences. This is shown in the next example.

EXAMPLE 3.3. The Fejér measure.

This is the Dirac measure (3.1) underlying the Féjer quadrature rule, i.e., $x_v = \cos\left(\dfrac{2v-1}{2N}\right)$ are the Chebyshev points, and ω_v the Cotes numbers for the corresponding (interpolatory) quadrature rule. The latter are known to be all

positive. This example is of some interest in connection with Stieltjes's procedure (cf. §4).

We computed the recursion coefficients β_k (all $\alpha_k = 0$) in double precision by an orthogonal reduction method using Lanczos's algorithm (cf., e.g., [9], [6, §7]). Applying the recurrence relation (2.1) for each $x = x_\nu$, $\nu = 1, 2, \ldots, N/2$, we then determined (again in double precision) the maximum of all amplification factors in (2.5),

$$\Omega_N = \max_{1 \le \nu \le N/2} \max_{0 \le m < n \le N-1} \omega_{m \to n}(x_\nu). \tag{3.6}$$

The results are summarized in Table 3.3, where ν_N is the integer ν for which the maximum in (3.6) is attained. In the last column we also show the maximum single-precision error observed. Compared with the previous two examples, the recurrence relation is now remarkably stable.

Table 3.3. *Stability of the recurrence relation for Fejér's measure.*

N	Ω_N	ν_N	max err
20	1.098(2)	2	9.234(−12)
40	1.465(3)	2	2.148(−10)
80	2.958(4)	3	5.554(−9)
160	8.094(4)	21	3.636(−8)

4. Discrete orthogonal polynomials are an important tool in least squares curve fitting. In this context, a common procedure to generate the required recursion coefficients consists in combining the recurrence relation (2.1) with the well-known formulae

$$\alpha_k = \frac{\sum_{\nu=1}^N \omega_\nu x_\nu \pi_k^2(x_\nu)}{\sum_{\nu=1}^N \omega_\nu \pi_k^2(x_\nu)}, \quad k = 0, 1, \ldots, N-1;$$

$$\beta_0 = \sum_{\nu=1}^N \omega_\nu, \quad \beta_k = \frac{\sum_{\nu=1}^N \omega_\nu \pi_k^2(x_\nu)}{\sum_{\nu=1}^N \omega_\nu \pi_{k-1}^2(x_\nu)}, \quad k = 1, 2, \ldots, N-1. \tag{4.1}$$

Since $\pi_0 = 1$, one begins by using (4.1) with $k = 0$ to compute α_0, β_0. Then (2.1) is used with $k = 0$ and $x = x_\nu$, $\nu = 1, 2, \ldots, N$, to generate all quantities $\pi_1(x_\nu)$ needed to compute α_1, β_1 from (4.1). Returning to (2.1) with $k = 1$ then yields (for $x = x_\nu$) the quantities $\pi_2(x_\nu)$, which in turn allow us to compute α_2, β_2, etc. In this way, all coefficients α_k, β_k, $k = 0, 1, \ldots, N-1$, can be progressively computed, by alternating between (4.1) and (2.1). We have attributed this algorithm to Stieltjes, and called it *Stieltjes's procedure* in [5]. The same procedure has been developed in the 1950's by various authors; see, e.g., Forsythe [2].

Since Stieltjes's procedure relies substantially on the recurrence relation for

discrete orthogonal polynomials, it will necessarily begin to deteriorate, once the recurrence relation starts developing the ill effects of pseudostability. This can be nicely illustrated with the discrete polynomials of Examples 3.1 and 3.2. Using $N = 40, 80, 160$ and 320, we applied Stieltjes's algorithm in single-precision arithmetic and compared the computed coefficients with the known ones in (3.3) and (3.5). The respective relative errors (absolute errors, if $\alpha_k = 0$) are shown in Table 4.1 for Example 3.1. (This is a shortened version of Table 4.1 in [7, §4].) The error growth is not as dramatic as Figure 3.1 would suggest. The reason is that for $x = x_\nu$ near the endpoints of $[x_1, x_N]$ (where error growth is most severe), the values of the polynomials π_k at $x = x_\nu$ appearing in (4.1), when k is large, are much smaller than further inside the interval, so that their errors do not contribute as much to the sums in (4.1) as the errors of the more significant terms. Still, there is substantial deterioration of Stieltjes's algorithm after some point (depending on N).[1] The analogous results for Krawtchouk polynomials are shown in Table 4.2 (where err α_k are relative errors).

Table 4.1. *Accuracy of Stieltjes's procedure for Example* 3.1.

N	k	err α_k	err β_k	N	k	err α_k	err β_k
40	≤35	≤1.91(−13)	≤7.78(−13)	160	≤ 76	≤2.98(−13)	≤7.61(−13)
	37	6.93(−11)	3.55(−10)		94	1.25(−4)	1.17(−3)
	39	1.93(−7)	9.58(−7)		112	2.35(−3)	1.16(0)
80	≤53	≤2.04(−13)	≤6.92(−13)	320	≤106	≤8.65(−13)	≤7.39(−13)
	61	3.84(−7)	9.35(−7)		128	2.46(−6)	4.67(−6)
	69	1.87(−1)	6.14(0)		150	1.15(−3)	2.18(−2)

Table 4.2. *Accuracy of Stieltjes's procedure for Example* 3.2.

N	k	err α_k	err β_k	N	k	err α_k	err β_k
40	≤26	≤5.71(−13)	≤5.83(−13)	160	≤ 54	≤8.00(−13)	≤1.29(−12)
	31	3.27(−6)	3.38(−6)		63	4.96(−7)	5.81(−7)
	36	9.63(−2)	5.07(0)		72	2.06(−1)	1.16(0)
80	≤37	≤2.75(−13)	≤7.11(−13)	320	≤ 84	9.25(−13)	≤2.52(−12)
	43	1.23(−7)	1.35(−7)		95	4.17(−7)	5.26(−7)
	49	2.41(−1)	3.61(−1)		106	2.00(−1)	6.42(−1)

For the Fejér measure, we compared single-precision results furnished by the Stieltjes procedure with double-precision results produced by the Lanczos algorithm. The maximum (absolute) error in the α's and the maximum (relative) error in the β's are shown in Table 4.3. The results confirm the remarkable stability of Stieltjes's algorithm in this case.

[1] This has already been observed in [5, Example 4.1], but was incorrectly attributed to the ill-conditioning of an underlying map, the map H_n of Eq. (3.4) in [5]. (The discussion of the condition of H_n in [5, §3.1] is incomplete inasmuch it does not take into account the dependence of the polynomials π_k on the abscissae τ_ν and weights λ_ν.)

Table 4.3. *Accuracy of Stieltjes's procedure for Example* 3.3.

N	max err α	max err β
40	1.35(−13)	5.19(−13)
80	2.34(−13)	1.80(−12)
160	5.21(−13)	3.14(−12)
320	5.37(−13)	6.05(−12)

Stieltjes's procedure becomes relevant also in connection with absolutely continuous measures $d\lambda$ if one adopts the following idea (cf. [5, §2.2]). Approximate $d\lambda$ by a discrete measure $d\lambda_N$ such that $\alpha_k(d\lambda_N) \to \alpha_k(d\lambda)$ and $\beta_k(d\lambda_N) \to \beta_k(d\lambda)$ as $N \to \infty$, for fixed k. The discretization $d\lambda \approx d\lambda_N$ can often be accomplished by applying a suitable N-point quadrature rule to the inner product associated with $d\lambda$. (In this connection, Example 3.3 suggests the use of Fejér's quadrature rule as especially appropriate.) Possible occurrences of pseudostability, in such applications, are usually of no concern, since convergence is realized for a value of N that is considerably larger than the maximum value of k for which the α_k, β_k are desired. The onset of pseudostability is thereby avoided; see [6, §8] for a numerical illustration. The same is true in the curve fitting context, where the number of data points, N, is usually much larger than the degree k of the least squares approximant.

REFERENCES

1. T. S. Chihara, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.
2. G. E. Forsythe, *Generation and use of orthogonal polynomials for data-fitting with a digital computer*, J. Soc. Indust. Appl. Math. 5 (1957), 74–88.
3. W. Gautschi, *Computational aspects of three-term recurrence relations*, SIAM Rev. 9 (1967), 24–82.
4. W. Gautschi, *Minimal solutions of three-term recurrence relations and orthogonal polynomials*, Math. Comp. 36 (1981), 547–554.
5. W. Gautschi, *On generating orthogonal polynomials*, SIAM J. Sci. Statist. Comput. 3 (1982), 289–317.
6. W. Gautschi, *Computational problems and applications of orthogonal polynomials*, in *Orthogonal Polynomials and Their Applications* (C. Brezinski et al., eds.), pp. 61–71, IMACS Annals on Computing and Applied Mathematics, vol. 9, J. C. Baltzer AG, Basel, 1991.
7. W. Gautschi, *Algorithm* xxx – *ORTHPOL: A package of routines for generating orthogonal polynomials and Gauss-type quadrature rules*, submitted for publication.
8. W. Gautschi and J. Wimp, *Computing the Hilbert transform of a Jacobi weight function*, BIT 27 (1987), 203–215.
9. W. B. Gragg and W. J. Harrod, *The numerically stable reconstruction of Jacobi matrices from spectral data*, Numer. Math. 44 (1984), 317–335.
10. F. W. J. Olver, *Numerical solution of second-order linear difference equations*, J. Res. Nat. Bur. Standards 71B (1967), 111–129.

26.6. [150] "THE COMPUTATION OF SPECIAL FUNCTIONS BY LINEAR DIFFERENCE EQUATIONS"

---

[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 (1997).

THE COMPUTATION OF SPECIAL FUNCTIONS BY LINEAR DIFFERENCE EQUATIONS*

WALTER GAUTSCHI
Department of Computer Sciences, Purdue University, West Lafayette, IN 47907-1398 USA

Abstract The use of linear difference equations for the computation of special functions is discussed, especially with regard to numerical stability. The emphasis is on difference equations of the first and second order. Phenomena of instability and pseudostability are exhibited along with numerical algorithms to deal with them.

1. INTRODUCTION

Difference equations are a popular means of computing special functions and can indeed be quite effective if proper attention is given to the possible occurrence of instabilities. A vast majority of special functions in practical use satisfy *linear* difference equations, either of first order, or, more often, homogeneous of order two. We shall restrict ourselves, therefore, to linear first-order and homogeneous second-order difference equations and want to show, largely by examples, how they can be used to compute special functions that satisfy them. For simplicity we consider only special functions of *real* variables, although the techniques we shall discuss are applicable also to functions of complex variables. Our intent is not to develop complete general-purpose routines of computing special functions; for this, we refer to software

*Work supported in part by the National Science Foundation under grant DMS-9305430.

exhaustively referenced in [16]. Our aim is more modest: we merely illustrate one particular approach toward computing special functions — the one based on linear difference equations — which may typically constitute part of a more comprehensive computational algorithm.

Readers interested in a more exhaustive treatment of numerical aspects of difference equations, including linear equations of higher order and nonlinear equations, may wish to consult Wimp's monograph [22].

2. DIFFERENCE EQUATIONS OF ORDER ONE

The gamma function, defined by Euler's integral, is arguably one of the most fundamental special functions. Not only is its occurrence pervasive in the theory of special functions, and crucial even in important branches of physics, but it also has significantly partaken in the development of many ideas in real and complex analysis. A masterly account of Euler's integral in historical perspective, from the time of Euler to the present, can be found in the essay of P.J. Davis [3].

It seems appropriate, therefore, to start, in §2.1, with the gamma function and related functions and the very simple difference equations satisfied by them. Not surprisingly, they are essentially unproblematic (at least in the real domain), although in the case of the logarithm of the gamma function, and to a lesser degree, the digamma function, improper use of the equations can lead to numerical instabilities. In §2.2 we then look at the incomplete gamma function and its difference equation and discover a first instance of genuine numerical instability. This will prompt us, in §2.3, to investigate more systematically the numerical properties of general first-order difference equations and to develop a simple theory of numerical stability and pseudostability based on amplification factors. Equipped with this theory, we return in §2.4 to the examples involving the gamma function and, in §§2.5–2.7, present additional examples illustrating different stability phenomena and computational algorithms to deal with them.

2.1. The Gamma Function. The gamma function

$$\Gamma(a) = \int_0^\infty e^{-t} t^{a-1} dt \tag{2.1}$$

was introduced by the young Euler (then 22 years of age) in response to a letter of Christian Goldbach, who sought an analytic expression of a

function interpolating the factorials when a is an integer, $n! = \Gamma(n+1)$, $n = 0, 1, \ldots$. It satisfies the identity

$$\Gamma(a+1) = a\Gamma(a) \tag{2.2}$$

for all positive a. This equation is certainly one of the most basic difference equations in analysis. It may be appropriate, therefore, to use it as the starting point in our discussion of numerical aspects of linear difference equations.

The numerical properties of (2.2), at least for real a, are almost self-evident, since the only operation involved is multiplication — a numerically benign operation — regardless of whether (2.2) is applied in forward or in backward direction. This is illustrated in Table 2.1, where the column headed by "err↑" shows relative errors in forward recursion, and the one headed by "err↓" those in backward recursion initiated with exact starting values. Computations, here and in the sequel, are done on a Sun SPARC station IPX in double precision (machine precision eps $\simeq 1.1 \times 10^{-16}$) and in quadruple precision to ascertain errors. Numbers in parentheses are decimal exponents.

TABLE 2.1. Recurrence (forward and backward) for the gamma function

n	err↑	$\Gamma(n+1)$	err↓
2	0.0000(0)	0.20000000000000(1)	0.8882(−15)
40	0.6196(−16)	0.81591528324790(48)	0.7335(−15)
80	0.3278(−15)	0.71569457046264(119)	0.1729(−15)
120	0.4670(−15)	0.66895029134491(199)	0.9406(−17)
160	0.4994(−15)	0.47147236359921(285)	0.6090(−16)

If there is any problem with the difference equation (2.2), it is the rapid growth of the solution itself, which on many computers quickly leads to "overflow". (In single precision, we could not have gone beyond $n = 34$.) Working with the logarithm of the gamma function alleviates this problem but introduces others. The difference equation indeed becomes

$$\ln\Gamma(a+1) = \ln\Gamma(a) + \ln a, \tag{2.3}$$

which requires the evaluation of a logarithm, $\ln a$, in each step, and is thus considerably more expensive than (2.2). Also, multiplication has been replaced by addition, which is a potentially dangerous operation. It is benign when both terms to be added are of the same sign, which is the case when $a \geq 1$ and the recursion (2.3) is applied in forward

direction. If it is applied in the backward direction, the two terms are of opposite sign and there is the potential danger of "cancellation errors". In the case of (2.3), this indeed can be a problem, as is seen in Table 2.2, which exhibits results in a format similar to the one in Table 2.1. In the range $2 \leq n \leq 10^5$ shown, the results of backward recursion, while comparable to those in forward recursion over much of the range, significantly deteriorate near the end of the recursion.

TABLE 2.2. Recurrence (forward and backward) for the logarithm of the gamma function

n	err↑	$\ln \Gamma(n+1)$	err↓
2	0.3346(−16)	0.69314718055995(0)	0.4562(−08)
20000	0.5066(−14)	0.17807562173720(6)	0.1275(−13)
40000	0.6129(−14)	0.38387160658183(6)	0.2211(−14)
60000	0.1238(−14)	0.60013241046210(6)	0.3806(−14)
80000	0.5321(−14)	0.82318911692301(6)	0.1644(−14)
100000	0.2973(−14)	0.10512992218991(7)	0.9343(−16)

Similar results are observed for the digamma function $\psi(a) = \Gamma'(a)/\Gamma(a)$, which satisfies

$$\psi(a+1) = \psi(a) + \frac{1}{a}, \tag{2.4}$$

again, like (2.2), an inexpensive recursion. Its numerical properties in the terminal phase of backward recursion turn out to be rather more favorable than those for the logarithm of the gamma function. The reason for this will become apparent later in §2.4.

2.2. The Incomplete Gamma Function. We now make in (2.1) what appears to be an innocuous change, extending the integration to a finite positive limit x instead of infinity,

$$\gamma(a,x) = \int_0^x e^{-t} t^{a-1} dt. \tag{2.5}$$

This gives rise to (one form of) the incomplete gamma function. Integration by parts immediately yields the difference equation

$$\gamma(a+1,x) = a\gamma(a,x) - x^a e^{-x}, \qquad a > 0. \tag{2.6}$$

When $x \to \infty$, it reduces to (2.2) as it should. One would expect, therefore, that for large x the two difference equations (2.2) and (2.6)

have similar numerical behavior. This is only partially true, however, and less so the smaller x; see Table 2.3. What is happening is that initially, for relatively small $a = n$, the recursion (2.6) does indeed

TABLE 2.3. Recurrence (forward and backward) for the incomplete gamma function

	n	err↑	$\gamma(n+1,x)$	err↓	$\nu(60)$
$x = 20$	0	0.4129(−16)	0.99999999793885(0)	0.1808(−15)	87
	10	0.2073(−15)	0.35895664347218(7)	0.1819(−15)	
	20	0.7525(−15)	0.10726845372436(19)	0.8274(−16)	
	30	0.2381(−13)	0.35741977445572(31)	0.2469(−16)	
	40	0.1256(−10)	0.20745721643958(44)	0.6625(−16)	
	50	0.6616(− 7)	0.14686167183265(57)	0.2385(−16)	
	60	0.2319(− 2)	0.11461624033194(70)	0.1543(−16)	
$x = 15$	0	0.2749(−16)	0.99999969409768(0)	0.8353(−16)	83
	10	0.1689(−15)	0.31989163434435(07)	0.2329(−16)	
	20	0.4382(−14)	0.20186009363578(18)	0.5648(−16)	
	30	0.1818(−11)	0.52337877233360(29)	0.8649(−16)	
	40	0.1531(− 7)	0.19120505355427(41)	0.3119(−17)	
	50	0.1357(− 2)	0.80383949798391(52)	0.6448(−16)	
$x = 10$	0	0.5891(−17)	0.99995460007024(0)	0.2279(−15)	79
	10	0.6166(−15)	0.15130653544997(7)	0.1103(−17)	
	20	0.1616(−12)	0.38640825537424(16)	0.6301(−16)	
	30	0.3215(− 8)	0.21177243656947(26)	0.8637(−17)	
	40	0.1444(− 2)	0.14501602968492(36)	0.1870(−16)	
$x = 1$	0	0.1966(−16)	0.63212055882856(0)	0.1966(−16)	70
	10	0.3362(− 8)	0.36461334624107(−1)	0.4260(−16)	
	20	0.4479(4)	0.18350467697256(−1)	0.4673(−17)	

behave like the one in (2.2), but from some n on, the results begin to deteriorate and eventually become completely meaningless. While the critical changeover point increases with increasing x, it cannot be avoided and will eventually be surpassed (unless overflow comes to the rescue!). On the other hand, if we recur in the backward direction, taking an arbitrary 0 as the initial value at the integer $\nu(60)$ shown in the last column, we obtain all answers for $0 \leq n \leq 60$ accurately to full machine precision. Clearly, this calls for analysis!

2.3. Numerical Stability of First-order Difference Equations; Amplification Factors. Much insight is provided into the numerical behavior of difference equations by analyzing the effect of a small (relative) error at some starting value $n = s$ upon a terminal value at $n = t$. This is easily done, for a general first-order difference equation

$$y_n = a_n y_{n-1} + b_n, \quad n = 1,2,3,\dots; \quad a_n \neq 0. \tag{2.7}$$

Indeed, if $\{f_n\}$ is the desired solution, the general solution is known to be $y_n = f_n + ch_n$, where $\{h_n\}$ is the solution of the homogeneous equation,

$$h_n = a_n h_{n-1}, \quad n = 1,2,3,\dots; \quad h_0 = 1, \tag{2.8}$$

and c an arbitrary constant. The latter is uniquely determined by $y_s = f_s(1+\varepsilon)$ and yields $y_t = f_t\left(1 + \frac{h_t}{f_t}\frac{f_s}{h_s}\varepsilon\right)$, provided $f_t \neq 0$. Thus, a relative error ε at $n = s$ gives rise to a relative error $\frac{h_t}{f_t}\frac{f_s}{h_s}\varepsilon$ at $n = t$, assuming exact arithmetic (except for the initial error). Here, t can be larger or smaller than s, the former case relating to forward recursion, the latter to backward recursion. With ε_t denoting the relative error induced at $n = t$ by a relative error ε_s at $n = s$, we have

$$\varepsilon_t = \frac{\rho_t}{\rho_s}\,\varepsilon_s, \tag{2.9}$$

where[†]

$$\rho_n = \frac{f_0 h_n}{f_n}, \quad h_n = a_n a_{n-1}\cdots a_0 \ \ (a_0 = 1). \tag{2.10}$$

This suggests defining *amplification factors*

$$\omega_{s\to t} := \left|\frac{\rho_t}{\rho_s}\right| \tag{2.11}$$

for the recursion from s to t, which measure the amplification of error involved. Clearly, $\omega_{s\to s} = 1$ and $\omega_{t\to s} = 1/\omega_{s\to t}$. The behavior of the quantities $|\rho_n| = \omega_{0\to n}$ in (2.10) completely determines the numerical stability of the recursion (2.7). If, for example, $|\rho_n|$ increases monotonically in the range $n_0 \le n \le n_1$, then for any s,t in this range, $\omega_{s\to t} > 1$ if $t > s$, and $\omega_{s\to t} < 1$ if $t < s$, which means that in forward recursion, errors are consistently enlarged, whereas in backward recursion, they are consistently diminished.

[†]Here, f_0 is included merely for the purpose of normalization and could, in fact must (if $f_0 = 0$), be omitted.

The following definition is thus immediate.

Definition 2.1. The difference equation (2.7) is said to be *unstable* for computing the solution $\{f_n\}$ (in forward direction) if

$$\lim_{n\to\infty} |\rho_n| = \infty. \tag{2.12}$$

Technically speaking, we may call (2.7) *stable* if

$$\sup_{n\geq 0} |\rho_n| = C < \infty. \tag{2.13}$$

In practice, however, stability in this sense may be misleading. It could well be that (2.13) holds with a constant C which is very large (many decimal orders in magnitude, for example). In this case, initial errors will be magnified by many orders of magnitude, which may completely distort the terminal values. We then speak of *pseudostability.*

It is interesting to note that in the case of instability, we can compute f_N for any fixed N, in principle, as accurately as we wish by starting the recurrence at some sufficiently large $n = \nu > N$, with $y_\nu = 0$, and recurring from $n = \nu$ down to $n = N$:

$$y_{n-1} = \frac{1}{a_n}(y_n - b_n), \qquad n = \nu, \nu-1, \ldots, N+1; \qquad y_\nu = 0. \tag{2.14}$$

Then the initial (relative) error is $\varepsilon_\nu = 1$, and (2.9) tells us that

$$\varepsilon_N = \frac{\rho_N}{\rho_\nu}. \tag{2.15}$$

To obtain f_N to a relative error ε, it thus suffices to take ν so large that $|\rho_N/\rho_\nu| \leq \varepsilon$. However, this is foolproof only if $|\rho_n|$ increases monotonically for $n \geq N$. Then all rounding errors introduced *in the course of* the recursion are consistently attenuated in downward direction. If, on the other hand, $|\rho_n|$ decreases significantly for $n \geq N$ before turning around and tending to ∞ as $n \to \infty$, then in the terminal phase of backward recursion, rounding errors could be significantly amplified. Whether this is tolerable or not depends on several factors: first on the magnitude of the quantity $R = \sup_{N\leq t<s} |\rho_t/\rho_s|$, secondly on the accuracy desired, and finally on the machine precision eps used. If $R \cdot \text{eps}$ is still within the relative accuracy desired, then backward recursion as in (2.14) is permissible; otherwise, it is not. Examples illustrating these matters as well as phenomena of pseudostability will be given in §§2.5–2.7.

2.4. The Gamma and Incomplete Gamma Function Revisited. We now re-examine the numerical examples of §§2.1–2.2 in the light of what we learned in §2.3.

Since the difference equation (2.2) for the gamma function is homogeneous, the associated amplification factors (2.10) are simply $\rho_n = 1$, and there is neither amplification nor attenuation of errors. The recursion is entirely stable in either direction. The results in Table 2.1 attest to that.

For the logarithm of the gamma function, the difference equation (2.3) with $a = n$ (and $n \geq 2$ to avoid zero initial values) has

$$\rho_n = \frac{\ln 2}{\ln \Gamma(n+1)} \sim \frac{\ln 2}{n \ln n}, \quad n \to \infty. \tag{2.16}$$

Here, ρ_n decreases monotonically, so errors are damped out in forward recursion. There are, nevertheless, rounding errors, which, if randomly distributed, can be expected, after n steps, to amount to about $\sqrt{n}$ times the machine precision. For $n = 10^5$, this is about 7×10^{-14}. Yet, the relative error in Table 2.2 for $n = 10^5$ is seen to be only 3×10^{-15}, which is more than a decimal order smaller; evidently, this is the result of consistent error damping. In contrast, for backward recursion from $n = 10^5$ down to 2, one expects an amplification of the initial error by the amount of $\dfrac{1}{\rho_n} \approx 2 \times 10^6$, whereas the amplification observed in Table 2.2 is about 5×10^7, again more than a decimal order larger. This time, the discrepancy is due to all intermediate rounding errors being consistently amplified. Note, however, that going from $s = 10^5$ to $t = 2 \times 10^4$, there is hardly any amplification, since $\rho_t/\rho_s \approx 6$, and the observed mild deterioration of accuracy is entirely due to rounding errors. On the other hand, when $s = 2 \times 10^4$ and $t = 2$, then $\rho_t/\rho_s \approx 3 \times 10^5$, and one loses about five orders of accuracy, as confirmed in Table 2.2.

For the digamma function $\psi(n+1)$, one obtains from (2.4) that

$$\rho_n = \frac{\gamma}{\psi(n+1)} \sim \frac{\gamma}{\ln n}, \quad n \to \infty, \tag{2.17}$$

where γ is Euler's constant $\gamma = .5772\ldots$. This still decreases monotonically with n, but at a much slower rate. Accordingly, also the error growth in backward recursion is much less than before. One finds, indeed, that in place of the error $.4562 \times 10^{-8}$ in the last column of Table 2.2, one now has only $.8058 \times 10^{-13}$.

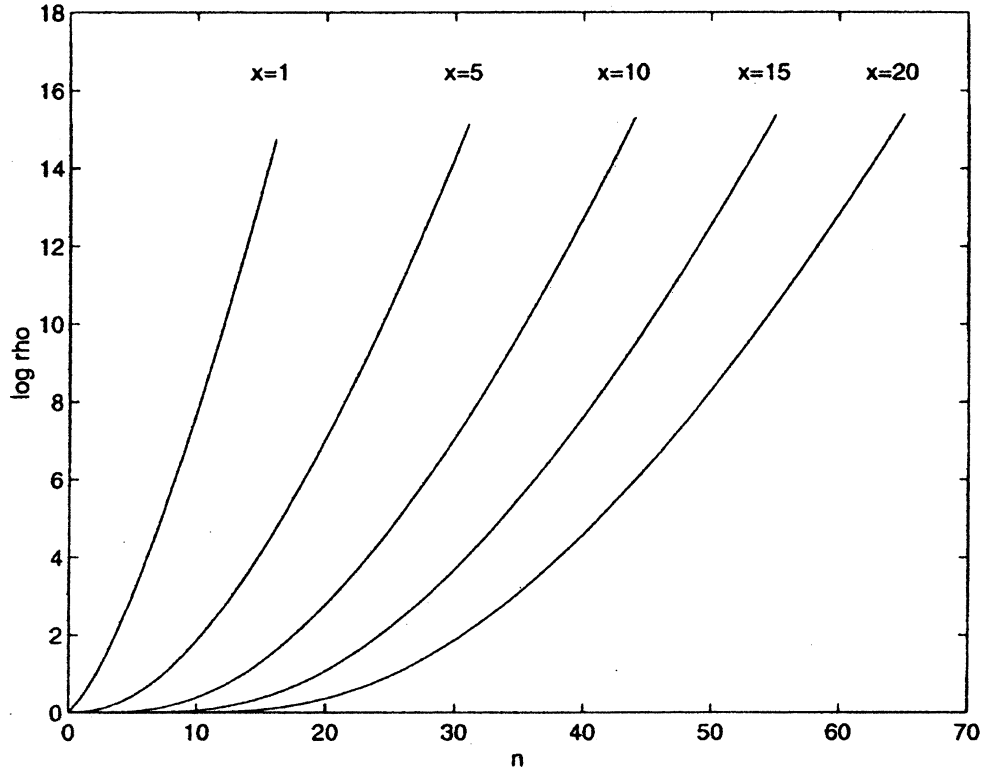


FIGURE 2.1. Amplification factors $\omega_{0\to n} = \rho_n$ for the incomplete gamma function (on a logarithmic scale)

To explain the results of Table 2.3 for the incomplete gamma function, it is useful to plot

$$\rho_n = \frac{(1 - e^{-x})n!}{\gamma(n+1, x)}, \qquad n = 0, 1, 2, \ldots,$$

as a function of n for various (positive) values of x. This is shown in Figure 2.1. It is evident (and can easily be proved from the recurrence relation (2.6)) that ρ_n increases monotonically for every x. Figure 2.1 also shows the rate of growth increasing for decreasing x. This explains the more rapid loss of accuracy as one goes down in Table 2.3 from larger to smaller values of x. This type of behavior of ρ_n makes the difference equation (2.6) ideally suited for backward recursion à la (2.14), at least when x is not excessively large.

2.5. The Remainders of the Exponential Series. A more elementary example, similar in its numerical behavior to the incomplete

gamma function, is provided by

$$r_n(x) = n![e^x - e_n(x)], \quad e_n(x) = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!}, \tag{2.18}$$

the scaled remainders of the exponential series. One easily verifies that $\{r_n\}$ is a solution of the difference equation

$$y_n = ny_{n-1} - x^n, \quad n = 1, 2, 3, \ldots . \tag{2.19}$$

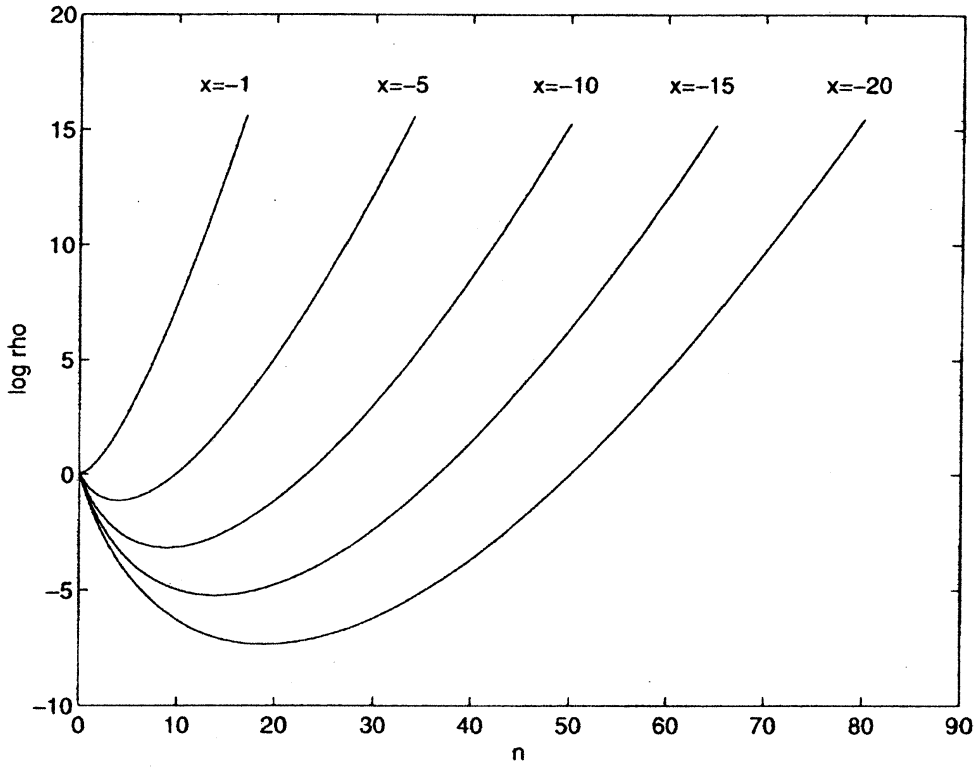


FIGURE 2.2. Amplification factors $|\rho_n|$ for the remainders of the exponential series, when $x < 0$ (on a logarithmic scale)

Since $h_n = n!$ is the solution of the associated homogeneous equation, one obtains from (2.10)

$$\rho_n = \frac{e^x - 1}{e^x - e_n(x)} . \tag{2.20}$$

As $n \to \infty$, the denominator tends to zero, so that $|\rho_n| \to \infty$, showing that (2.19) is unstable for computing $\{r_n\}$. In fact, for $x > 0$,

we have monotonic growth of ρ_n, just as in the case of the incomplete gamma function. When $x < 0$, the situation becomes a bit more complicated. The amplification factors (2.20) then behave in modulus as shown in Figure 2.2. There is a significant downward dip of $|\rho_n|$ (note the logarithmic scale of the vertical axis!) before its journey to infinity. Backward recursion therefore loses accuracy in its final stage, as explained at the end of §2.3. Since the minimum of $|\rho_n|$ is attained at $n \approx |x|$, the remedy is rather simple: recur forward for all $n < |x|$, and backward otherwise. In this way one always enjoys a regime of consistent error damping.

The following two examples involve phenomena of pseudostability.

2.6. Successive Derivatives of e^x/x. The nth derivative $d_n(x) = \frac{d^n}{dx^n}(e^x/x)$ satisfies the difference equation

$$y_n = \frac{1}{x}\,(-ny_{n-1} + e^x), \qquad n = 1, 2, 3, \ldots, \tag{2.21}$$

which is most easily obtained by applying Leibniz's rule of differentiation to the product $x \cdot e^x/x = e^x$. The corresponding homogeneous equation has the solution $h_n = (-1)^n n! x^{-n}$, which, combined with the identity $d_n(x) = (-1)^n \dfrac{n!}{x^{n+1}} e^x e_n(-x)$, gives rise to the amplification factors

$$\rho_n = \frac{1}{e_n(-x)} \tag{2.22}$$

with $e_n(\cdot)$ as defined in (2.18). For $x < 0$, therefore, ρ_n decreases monotonically to $e^{-|x|}$ as $n \to \infty$, and forward recursion is completely satisfactory. When $x > 0$, the behavior of $|\rho_n|$ is rather bizarre, as indicated in Figure 2.3. It is clear for one thing that $\rho_n \to e^x$ as $n \to \infty$, so that the difference equation (2.21) is stable for d_n, but pseudostable if $x > 0$ is large. What is striking is the abruptness with which the limit is attained. An explanation is provided by the identity

$$d_n(x) = (-1)^n \frac{n!}{x^{n+1}} + \int_0^1 t^n e^{xt} dt, \tag{2.23}$$

in which, for large $x > 0$, the first term or the second term is dominant depending on whether n is relatively large or not. As long as the integral dominates, it varies slowly with n so that $|\rho_n| = |d_0 h_n / d_n|$ is approximately proportional to $|h_n| = n! x^{-n}$. As soon as the first

term becomes dominant — and this happens rather quickly — then ρ_n becomes practically constant equal to e^x.

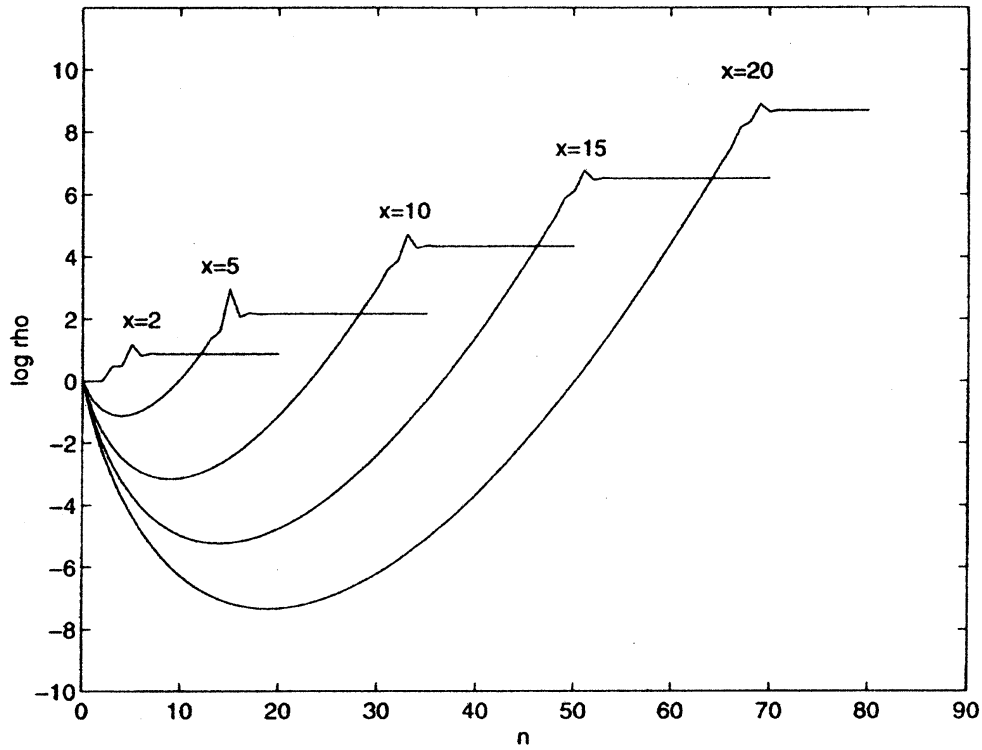


FIGURE 2.3. Amplification factors for the derivatives of e^x/x, $x > 0$ (on a logarithmic scale)

This last observation also provides a clue as to an appropriate method to calculate $d_n(x)$ accurately when $x > 0$. Figure 2.3 suggests that for $n > x$ backward recursion is indicated. But we cannot start with an arbitrary zero initial values for $n = \nu$ sufficiently large, as in (2.14), since ρ_ν does not tend to ∞. Instead, we have to start with an accurate starting value d_ν. From the discussion above it is clear that an appropriate choice is $d_\nu(x) \approx (-1)^\nu \nu!/x^{\nu+1}$. It is possible, indeed, to estimate precisely how large ν must be taken for this approximation to ensure any prescribed relative accuracy (cf. [13]).

2.7. Exponential Integrals. The exponential integrals, defined by

$$E_n(x) = \int_1^\infty e^{-xt} t^{-n} dt, \quad x > 0, \tag{2.24}$$

are important in many physical applications, such as transport theory and radiative transfer. For negative integer values of n they are known

as molecular integrals in quantum chemistry and are in fact equal to $(-1)^n d_{|n|}(x)$ (cf. §2.6). Here we consider positive integer values of n only, although the discussion extends easily to arbitrary positive n.

Integration by parts shows that $f_n = E_{n+1}(x)$ satisfies

$$y_n = \frac{1}{n}\,(e^{-x} - xy_{n-1}), \qquad n = 1, 2, 3, \ldots\ . \tag{2.25}$$

For the corresponding amplification factors

$$\rho_n = \frac{x^n E_1(x)}{n! E_{n+1}(x)}, \qquad n = 0, 1, 2, \ldots, \tag{2.26}$$

one finds that, when $\rho_1 \leq 1$, i.e., $x \leq .61006\ldots$, they are monotonically decreasing from 1 to 0, making the recursion (2.25) particularly effective for the computation of f_n. This, of course, requires $f_0 = E_1(x)$, which for such small values of x, however, is easily calculated by Taylor expansion. For larger values of x, the ρ_n initially increase to some

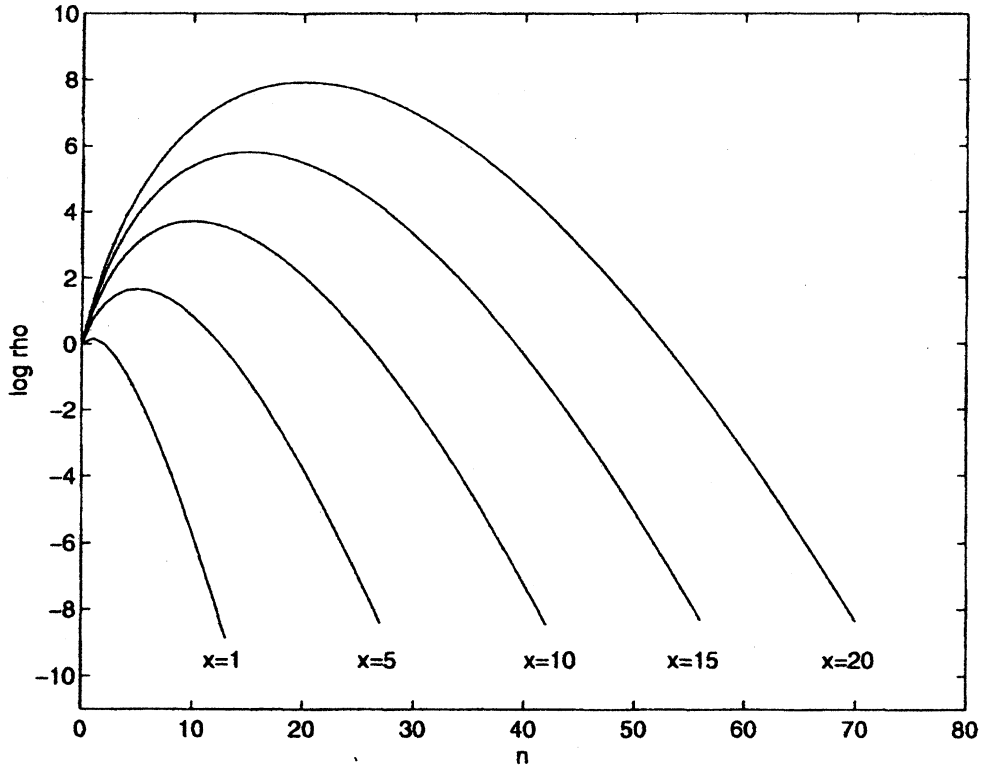


FIGURE 2.4. Amplification factors for exponential integrals (on a logarithmic scale)

maximum value (which can be significantly large) before turning around and decreasing monotonically to zero; cf. Figure 2.4. Clearly, $|\rho_n|$ is uniformly bounded for each fixed $x > 0$, and (2.25) thus stable for computing $f_n = E_{n+1}(x)$; but for large x the extremely large value of $\max_n |\rho_n|$ renders (2.25) pseudostable. Since this maximum occurs near x, the proper procedure of computing $f_n = E_{n+1}(x)$ is to recur backward from $n_1 = \langle x \rangle$ (the integer closest to x) down to n, if $n < n_1$, and to recur forward from n_1 to n, if $n > n_1$. In this way one again enjoys the benefit of consistent error attenuation. The starting value $E_{n_1+1}(x)$ can be computed effectively from a continued fraction. This indeed is a procedure used as early as 1960 by G.F. Miller [17, pp. 4–5] and also adopted in our algorithm[‡] [7].

3. DIFFERENCE EQUATIONS OF ORDER TWO

A great majority of special functions satisfy difference equations of the second order, and they are often linear and homogeneous. It is of interest, therefore, to consider the class of linear homogeneous difference equations of order two,

$$y_{n+1} + a_n y_n + b_n y_{n-1} = 0, \quad n = 1, 2, 3, \ldots; \quad b_n \neq 0, \tag{3.1}$$

where the coefficients a_n, b_n may typically depend on additional parameters. The numerical characteristics of (3.1) can be described, similarly as in §2.3, in terms of amplification factors, but there are now two starting values that need to be considered and studied as to their effect on terminal values of the solution. This will be discussed in §3.1. The phenomenon of instability will be seen in §3.2 to be tied to the presence of a minimal solution of (3.1), which in turn can be characterized in terms of the convergence of a continued fraction. This leads to useful computational algorithms. Some representative examples will be discussed in §3.3, involving Bessel functions, Coulomb wave functions, and repeated integrals of the coerror function. Another vast area in which difference equations of the type (3.1) play a fundamental role are

[‡]If $n < n_1$, we could use the continued fraction to compute $E_{n+1}(x)$ directly and dispense with the backward recurrence from n_1 to n. This is accomplished by inserting the statement "**if** $n1 > nmax$ **then** $n1 := nmax$;" immediately after the statement "$n1 := entier(x + .5)$;" on p. 763 of [7], and accordingly delete the clause "**if** $n1 \leq nmax$ **then**" (but not the assignment statement following it!) in the two occurrences near the end of the algorithm.

orthogonal polynomials. Here, as shown in §3.4.1, minimal solutions occur when the variable is outside the interval of orthogonality, which is a case of interest, for example, in the study of Gaussian quadrature remainders for analytic functions. Finally, a phenomenon of pseudostability is shown in §3.4.2 to arise in connection with discrete orthogonal polynomials.

3.1. Numerical Stability of Second-order Difference Equations; Amplification Factors. Suppose that $\{f_n\}$ is the solution of (3.1) to be computed. For simplicity assume that $f_n \neq 0$ for all n, so we can talk about relative errors. Let $\{g_n\}$ be an arbitrary second solution of (3.1), linearly independent of $\{f_n\}$, and assume $g_0 \neq 0$. In analogy to the discussion in §2.3, we consider the problem of error propagation: given that relative errors ε_s and ε_{s+1} are committed at some starting indices $n = s$ and $n = s+1$, what is the resulting relative error at the "terminal" index $n = t$? The problem amounts to identifying the solution $\{y_n\}$ of (3.1) satisfying

$$y_s = f_s(1 + \varepsilon_s), \quad y_{s+1} = f_{s+1}(1 + \varepsilon_{s+1}), \tag{3.2}$$

and comparing y_t with f_t. This is an elementary exercise in the theory of linear difference equations. We know that the general solution of (3.1) is a linear combination of two linearly independent solutions, say, $y_n = c_1 f_n + c_2 g_n$. The two conditions (3.2) then serve to fix the constants c_1 and c_2, and hence the solution $\{y_n\}$, which can then be compared at $n = t$ with f_t. The result is conveniently expressed in terms of the quantities[§]

$$\rho_n = \frac{f_0 g_n}{g_0 f_n}, \quad n = 0, 1, 2, \ldots, \tag{3.3}$$

in the form

$$\frac{y_t - f_t}{f_t} = \frac{(\rho_{s+1} - \rho_t)\varepsilon_s - (\rho_s - \rho_t)\varepsilon_{s+1}}{\rho_{s+1} - \rho_s}. \tag{3.4}$$

This suggests to define amplification factors as follows:

$$\omega_{s \to t} := \frac{|\rho_{s+1} - \rho_t| + |\rho_s - \rho_t|}{|\rho_{s+1} - \rho_s|}. \tag{3.5}$$

[§]The factor f_0/g_0 is included in (3.3) solely for aesthetic reasons, namely to make $\rho_0 = 1$. It is not essential, however, and could be removed from the subsequent expressions in (3.4) and (3.5) by dividing it out, both in the numerator and denominator. This *must* be done if $g_0 = 0$. Also, if $f_t = 0$, the expressions in (3.4), (3.5) must be similarly modified.

They tell us the amount by which initial (relative) errors ε_s, ε_{s+1} at $n = s$, $n = s+1$, are amplified at $n = t$. Obviously, $\omega_{s\to s} = \omega_{s\to s+1} = 1$. Also, the quantity $\omega_{s\to t}$, having an intrinsic meaning, does not depend on the choice of $\{g_n\}$ in (3.3), so long as $\{g_n\}$ is linearly independent of $\{f_n\}$.

If for fixed s we have that $\omega_{s\to t}$ tends to infinity as $t \to \infty$, then forward recursion is unstable for computing $\{f_n\}$. This happens precisely if $|\rho_n| \to \infty$ as $n \to \infty$. Thus:

Definition 3.1. The difference equation (3.1) is said to be *unstable* for computing $\{f_n\}$ (in forward direction) if

$$\lim_{n\to\infty} |\rho_n| = \infty, \tag{3.6}$$

where ρ_n is defined by (3.3). The difference equation is called *stable* if

$$\sup_{n\geq 0} |\rho_n| = C < \infty. \tag{3.7}$$

Again, we must be prepared to deal with *pseudostability*, i.e., with the case in which the constant C in (3.7) is unacceptably large (though finite).

It is worth observing that if one puts $\varepsilon_{s+1} = -1$, which in view of (3.2) means $y_{s+1} = 0$, then as a consequence of (3.4) one gets

$$\frac{f_0 y_t}{y_0} = f_t \frac{1 - \frac{\rho_t}{\rho_{s+1}}}{1 - \frac{1}{\rho_{s+1}}}, \tag{3.8}$$

independently of the error ε_s. Thus, in the case of instability, since $\rho_{s+1} \to \infty$ as $s \to \infty$, if we recur backward, starting from some sufficiently large $n = s$ with starting values $y_s = 1$, $y_{s+1} = 0$, the quantity on the left of (3.8) approximates f_t arbitrarily well for any fixed t. This is the basis of J.C.P. Miller's backward recurrence algorithm (cf. §3.2).

In contrast to first-order equations, we may now also consider boundary value problems. Thus, we may be given the values of $\{f_n\}$ at $n = 0$ and $n = N$ and are to obtain the intermediate values f_n for $1 \leq n \leq N-1$. We are then interested in the relative errors in these intermediate values caused by relative errors ε_0, ε_N in the boundary values. An analysis similar to the one above will show that in this context,

$$\frac{y_n - f_n}{f_n} = \frac{(\rho_N - \rho_n)\varepsilon_0 + (\rho_n - 1)\varepsilon_N}{\rho_N - 1}, \tag{3.9}$$

with the ρ's defined as before by (3.3). The appropriate error amplification measure is now

$$\omega_{0:N} := \frac{\max_{0\le n\le N}(|\rho_N - \rho_n| + |\rho_n - 1|)}{|\rho_N - 1|}. \tag{3.10}$$

In effect, we are dealing here with the solution of a system of linear equations in the unknowns $y_1, y_2, \ldots, y_{N-1}$ having a tridiagonal coefficient matrix $A \in \mathbb{R}^{(N-1)\times(N-1)}$ and special right-hand vector $b \in \mathbb{R}^{N-1}$,

$$A = \begin{bmatrix} a_1 & 1 & & & 0 \\ b_2 & a_2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & b_{N-2} & a_{N-2} & 1 \\ 0 & & & b_{N-1} & a_{N-1} \end{bmatrix}, \quad b = \begin{bmatrix} -b_1 f_0 \\ 0 \\ \vdots \\ 0 \\ -f_N \end{bmatrix}.$$

The quantity (3.10), in a sense, is the condition number for this special linear system. It may well be that the difference equation (3.1) is unstable for initial value problems (i.e., $\rho_N \to \infty$ as $N \to \infty$), but stable for boundary value problems (i.e., $\omega_{0:N}$ only moderately larger than 1, or even equal to 1). This is particularly evident if we are interested only in relative errors (3.9) for $0 \le n \le n_0$, with $n_0 < N$ fixed, and accordingly consider

$$\omega_{0:n_0:N} := \frac{\max_{0\le n\le n_0}(|\rho_N - \rho_n| + |\rho_n - 1|)}{|\rho_N - 1|}. \tag{3.10$_0$}$$

Then indeed $\omega_{0:n_0:N} \to 1$ as $N \to \infty$.

3.2. Minimal Solutions and Continued Fractions; Backward Recurrence Algorithm. Instability, by definition (3.6) and (3.3), means that the solution $\{f_n\}$ of (3.1) has the property that

$$\lim_{n\to\infty} \frac{f_n}{g_n} = 0 \tag{3.11}$$

for some solution $\{g_n\}$ linearly independent of f_n. It is easily seen that (3.11) then holds for *any* linearly independent solution g_n of (3.1). A solution $\{f_n\}$ which has this property is called a *minimal solution* of (3.1). A minimal solution is determined, if it exists, up to a constant factor; in other words, the set of minimal solutions, if not empty, is

a one-dimensional subspace of the linear space of all solutions. To determine $\{f_n\}$ uniquely thus requires *one* condition, not two, as for other solutions.

To remove the arbitrary factor inherent in a minimal solution, it is natural to consider ratios

$$r_n = \frac{f_{n+1}}{f_n}, \quad n = 0, 1, 2, \ldots . \tag{3.12}$$

From the difference equation (3.1), dividing it by f_n, one then immediately obtains

$$r_n + a_n + \frac{b_n}{r_{n-1}} = 0,$$

hence

$$r_{n-1} = \frac{-b_n}{a_n + r_n}.$$

Iterating this equation indefinitely yields formally the continued fraction

$$r_{n-1} = \frac{f_n}{f_{n-1}} = \frac{-b_n}{a_n -} \frac{b_{n+1}}{a_{n+1} -} \frac{b_{n+2}}{a_{n+2} -} \cdots . \tag{3.13}$$

For $n = 1$, in particular,

$$\frac{f_1}{f_0} = \frac{-b_1}{a_1 -} \frac{b_2}{a_2 -} \frac{b_3}{a_3 -} \cdots . \tag{3.13$_1$}$$

A result of Pincherle now tells us that the continued fraction on the right of (3.13$_1$) converges precisely if (3.1) has a minimal solution $\{f_n\}$, and the limit then is expressible in terms of it by the left-hand side of (3.13$_1$). More precisely, we have

Theorem 3.1 ([21, Ch. III, §15]). *The continued fraction on the right of* (3.13$_1$) *converges if and only if the difference equation* (3.1) *possesses a minimal solution* $\{f_n\}$ *with* $f_0 \neq 0$. *In case of convergence, moreover,* (3.13) *holds for each* $n = 1, 2, 3, \ldots$, *provided* $f_n \neq 0$ *for all* n.

For a proof, see, e.g., [5, p. 31].

To compute a minimal solution $\{f_n\}$ of (3.1), it would be unreasonable to employ (3.1) in forward direction. Not only would this require *two* starting values (whereas *one* would be enough to identify it), but also unavoidable rounding errors would activate other solutions of the equation, which by their dominance (cf. (3.6)) would eventually completely overshadow the desired solution. Theorem 3.1 provides a

more natural means of computation. Suppose, indeed, that the minimal solution is specified by a general linear condition of the form

$$\sum_{m=0}^{\infty} \lambda_m f_m = s, \qquad s \neq 0, \tag{3.14}$$

where λ_m and s are known, and the series is known to converge. (A special case of (3.14) is $\lambda_0 = 1$, $\lambda_m = 0$ for $m > 0$, which specifies $f_0 = s$.) Assume, moreover, that we want to compute f_n for $n = 0, 1, 2, \ldots, N$. Define r_n as in (3.12) and s_n by

$$s_n = \frac{1}{f_n} \sum_{m=n+1}^{\infty} \lambda_m f_m \,. \tag{3.15}$$

Then, if r_ν and s_ν were known for some $\nu > N$, we could proceed as follows:

$$\left.\begin{aligned} r_{n-1} &= \frac{-b_n}{a_n + r_n} \\ s_{n-1} &= r_{n-1}(\lambda_n + s_n) \end{aligned}\right\} \quad n = \nu, \nu - 1, \ldots, 1, \tag{3.16}$$

$$f_0 = \frac{s}{\lambda_0 + s_0}, \quad f_n = r_{n-1} f_{n-1}, \quad n = 1, 2, \ldots, N.$$

These formulae follow easily from the definition of r_n and s_n in (3.12) and (3.15), respectively. It turns out that choosing $r_\nu = 0$, $s_\nu = 0$ in (3.16) yields a viable algorithm for computing the minimal solution $\{f_n\}$:

Theorem 3.2 ([5, p. 39]). *If* (3.1) *has a nonvanishing minimal solution* $\{f_n\}$ *satisfying* (3.14), *and* $r_{n-1}^{(\nu)}$, $s_{n-1}^{(\nu)}$, $f_n^{(\nu)}$ *are the quantities generated in* (3.16) *using* $r_\nu = s_\nu = 0$, *then*

$$\lim_{\nu \to \infty} f_n^{(\nu)} = f_n, \qquad n = 0, 1, 2, \ldots, N, \tag{3.17}$$

if and only if

$$\lim_{\nu \to \infty} \frac{f_{\nu+1}}{g_{\nu+1}} \sum_{m=0}^{\nu} \lambda_m g_m = 0 \tag{3.18}$$

for some solution $\{g_n\}$ *of* (3.1) *linearly independent of* $\{f_n\}$.

The speed of convergence in (3.17) is determined by the speed of convergence in (3.18) and by how fast the infinite series in (3.14)

converges (cf. [5, Eq. (3.15)]). If only a finite number of the λ_m are nonzero, then (3.18) follows trivially from (3.6).

An alternative interpretation of the algorithm in Theorem 3.2 is in terms of the solution $\{y_n^{(\nu)}\}$ of (3.1) defined by

$$y_\nu = 1, \quad y_{\nu+1} = 0 \quad (\nu > N). \tag{3.19}$$

It can be shown, indeed, that ([5, p. 38])

$$f_n^{(\nu)} = \frac{s}{\sum_{m=0}^{\nu} \lambda_m y_m^{(\nu)}} y_n^{(\nu)}, \quad n = 0, 1, 2, \ldots, N. \tag{3.20}$$

In other words, $\{f_n^{(\nu)}\}$ is the solution $\{y_n^{(\nu)}\}$ of (3.1) obtained by backward recursion, using the starting values (3.19), "normalized" by the factor $s/\sum_{m=0}^{\nu} \lambda_m y_m^{(\nu)}$. In this form, the algorithm is called *Miller's backward recurrence algorithm.* It was first proposed by J.C.P. Miller as a means of computing Bessel functions [1, p. xvii] and has since found applications to many other special functions. Nevertheless, the quantities $y_n^{(\nu)}$ generated in this algorithm can become quite large and on many computers may produce "overflow". No such problems are present in the "continued fraction algorithm" (3.16).

While the algorithm (3.16) is based on backward recurrence (cf. (3.5) with $s = \nu$ and $t = n$), there are also algorithms based on boundary value techniques (cf. (3.10)). The best known is *Olver's algorithm* ([18], [20], [8, §2.2.2(iv)]), which has a built-in feature of estimating an appropriate value of ν to ensure any prescribed accuracy. Realistic error bounds for difference equations of oscillatory and monotone type are provided in [19]. All known recurrence algorithms can be interpreted and unified in terms of triangular matrix factorization and numerical linear algebra techniques; for this, and also for extensions to systems of difference equations, see [2].

3.3. Examples. We begin with the example of Bessel functions, which gave rise to Miller's algorithm[¶].

Example 3.1. Bessel functions $J_n(x)$, $n = 0, 1, 2, \ldots$; $x > 0$.

The difference equation here is

$$y_{n+1} - \frac{2n}{x} y_n + y_{n-1} = 0, \quad n = 1, 2, 3, \ldots . \tag{3.21}$$

[¶]As pointed out in [5, p. 46], the idea of backward recurrence in connection with spherical Bessel functions can be traced back at least to Lord Rayleigh (1910).

The Bessel functions of the first kind, $\{J_n(x)\}$, are a minimal solution, a second solution being $\{Y_n(x)\}$, the Bessel functions of the second kind. Their dominance is rather pronounced, since

$$\frac{Y_n(x)}{J_n(x)} \sim -2\left(\frac{2n}{ex}\right)^{2n}, \quad n \to \infty, \tag{3.22}$$

but starts "taking hold" only once n exceeds x. For extremely large values of x, backward recurrence thus will become expensive, and other techniques may be more appropriate. (For modified Bessel functions $I_n(x)$, there is an alternative continued fraction due to Perron, which is particularly effective to calculate $r_\nu = I_{\nu+1}(x)/I_\nu(x)$ when x is very large; see [14]).

There are many infinite series in the Bessel functions J_n that can be used for normalization. One that was found particularly convenient, for real $x > 0$, is

$$J_0(x) + 2\sum_{m=1}^{\infty} J_{2m}(x) = 1.$$

The algorithm (3.16) then becomes (we write r_{n-1} instead of $r_{n-1}^{(\nu)}$, etc.)

$$\left.\begin{aligned} r_\nu &= 0, \quad r_{n-1} = \frac{x}{2n - xr_n} \\ s_\nu &= 0, \quad s_{n-1} = r_{n-1}[1 + (-1)^n + s_n] \end{aligned}\right\} \quad n = \nu, \nu-1, \ldots, 1,$$

$$f_0 = \frac{1}{1+s_0}, \quad f_n = r_{n-1}f_{n-1}, \qquad n = 1, 2, \ldots, N. \tag{3.23}$$

For Bessel functions with integer order, it may actually be more efficient to simply evaluate r_N from the continued fraction (3.13), use the recursion in the first line of (3.23) to obtain the ratios r_{n-1} for $n = N, N-1, \ldots, 1$, compute $f_0 = J_0(x)$ from some known rational approximation, and finally use the last relation in (3.23) to obtain $f_n = J_n(x)$ for $n = 1, 2, \ldots, N$. The technique outlined above is more important for Bessel functions J_{a+n}, $0 < a < 1$, of arbitrary positive orders, or for Bessel functions of real order and complex argument. In either case, appropriate series are again available for normalization (see, e.g., [5, Eqs. (5.9) and (5.7)]).

Example 3.2. Coulomb wave functions $F_L(\eta, \rho)$, $L = 0, 1, 2, \ldots$.

Coulomb wave functions, of interest in the study of nuclear interactions between charged particles in a Coulomb field, behave in many

ways similarly to Bessel functions. Corresponding to Bessel functions of the first and second kind, there are now the so-called regular and irregular Coulomb wave functions, denoted by $F_L(\eta,\rho)$ and $G_L(\eta,\rho)$, respectively. Here, L is a nonnegative integer, the orbital angular-momentum quantum number of the particle, η a real nonzero parameter depending on the relative charges, and $\rho > 0$ a scaled radial distance. Both functions satisfy the difference equation

$$\begin{aligned} &L[(L+1)^2+\eta^2]^{1/2}y_{L+1} - (2L+1)\left[\eta + \frac{L(L+1)}{\rho}\right] y_L \\ &+(L+1)[L^2+\eta^2]^{1/2}y_{L-1} = 0, \qquad L = 1,2,3,\ldots, \end{aligned} \tag{3.24}$$

with $\{F_L(\eta,\rho)\}$ being the minimal solution, and

$$\frac{G_L(\eta,\rho)}{F_L(\eta,\rho)} \sim 2e^{1+\pi\eta}\left(\frac{2L}{e\rho}\right)^{2L+1}, \qquad L \to \infty. \tag{3.25}$$

Compare (3.25) with (3.22) to see the analogy with Bessel functions.

For computational purposes, it is more convenient to deal with

$$f_L = \frac{2^L L!}{(2L)!C_L(\eta)}\, F_L(\eta,\rho), \tag{3.26}$$

where

$$C_L(\eta) = \frac{2^L e^{-\pi\eta/2}|\Gamma(L+1+i\eta)|}{(2L+1)!}, \qquad L = 0,1,2,\ldots .$$

The factors introduced in (3.26) are easily computed, since

$$C_0(\eta) = \left(\frac{2\pi\eta}{e^{2\pi\eta}-1}\right)^{1/2}, \quad C_L(\eta) = \frac{(L^2+\eta^2)^{1/2}}{L(2L+1)}\, C_{L-1}(\eta), \quad L = 1,2,3,\ldots .$$

In effect this removes square roots from the difference equation (3.24), which now (for f_L) assumes the form

$$\frac{L[(L+1)^2+\eta^2]}{(L+1)(2L+3)}\, y_{L+1} - \left[\eta + \frac{L(L+1)}{\rho}\right] y_L + \frac{L(L+1)}{2L-1}\, y_{L-1} = 0. \tag{3.27}$$

The choice of an appropriate infinite series (3.14) for (3.26) is a rather delicate matter and depends on the value of

$$\tau = \frac{\rho}{2\eta}.$$

For reasons explained in [5, pp. 64–65], we are led to use

$$\sum_{L=0}^{\infty} \lambda_L f_L = \rho e^{\omega\rho}, \qquad \lambda_L = i^L P_L^{(i\eta,-i\eta)}(-i\omega), \tag{3.28}$$

where $P_L^{(i\eta,-i\eta)}(z)$ is the Jacobi polynomial of degree L with imaginary parameters $\alpha = i\eta$, $\beta = -i\eta$, and ω is defined by

$$\omega = \begin{cases} \dfrac{\pi}{2\tau} & \text{if } \tau \geq 1, \\ \dfrac{1}{2\tau}\left[\pi - 2\cos^{-1}\sqrt{\tau} + 2\sqrt{\tau(1-\tau)}\right] & \text{if } 0 < \tau < 1, \\ 0 & \text{if } \tau < 0. \end{cases}$$

As τ increases from 0 to ∞, the function ω decreases from ∞ to 0, hence is positive for all $\tau > 0$.

Equally intriguing as the choice of (3.28) for the normalizing series is the computation of its coefficients λ_L. From the three-term recurrence relation for Jacobi polynomials, they can be seen to satisfy the difference equation

$$\lambda_{L+1} - \frac{2L+1}{L+1}\,\omega\lambda_L - \frac{L^2+\eta^2}{L(L+1)}\,\lambda_{L-1} = 0, \quad L = 1, 2, 3, \ldots, \tag{3.29}$$

with initial values

$$\lambda_0 = 1, \qquad \lambda_1 = \omega - \eta. \tag{3.30}$$

(In particular, the λ_L are all real.) Moreover, it can be shown that $\{\lambda_L\}$ is *not* a minimal solution of (3.29). It would appear, therefore, that forward recurrence in (3.29), with starting values (3.30), is the method of choice. Curiously, this is only conditionally true, namely only when $\eta > 0$ is not too large. (Note that for $\eta < 0$ we have $\omega = 0$, in which case (3.29), (3.30) pose no computational problems.) The difference equation (3.29) actually does have a minimal solution, λ'_L, which, when normalized by $\lambda'_0 = 1$, satisfies

$$\lambda_0 = \lambda'_0, \quad \lambda_1 - \lambda'_1 \to 0 \text{ as } \eta \to \infty.$$

Therefore, for η large enough, the initial values λ_0 and λ_1 of $\{\lambda_L\}$ are indistinguishable (in machine arithmetic) from those of the minimal solution $\{\lambda'_L\}$, in spite of λ_L going to $+\infty$ and λ'_L to 0 as $L \to \infty$.

Forward recurrence in (3.29), therefore, is doomed to fail when η is large!

A way out of this dilemma is to define $\{\lambda_L''\}$ as the solution of (3.29) with initial values

$$\lambda_0'' = -\lambda_1', \quad \lambda_1'' = 1$$

("orthogonal" to those of $\{\lambda_L'\}$), and to observe that

$$\lambda_L = \lambda_L' + \frac{\varepsilon}{1+\lambda_1'^2}(\lambda_L'' + \lambda_1'\lambda_L'), \tag{3.31}$$

where

$$\varepsilon = \lambda_1 - \lambda_1'. \tag{3.32}$$

When ε is small, the relation (3.31) shows that λ_L initially behaves like the minimal solution $\{\lambda_L'\}$ but starts deviating from it once the dominance of λ_L'' begins to outweigh the smallness of ε. The trick now is to compute ε not by (3.32) (which would cause large cancellation errors), but explicitly by (cf. [6, Eq. (3.5)])

$$\varepsilon = \frac{2\eta}{e^{2\eta\phi}-1}, \quad \phi = \tan^{-1}\frac{1}{\omega} \quad (\omega > 0). \tag{3.33}$$

Then the computation of λ_L from (3.31) (using our continued fraction algorithm for λ_L') is completely stable, and we are ready to apply the algorithm of Theorem 3.2 to the difference equation (3.27) and normalizing series (3.28).

Example 3.3. Repeated integrals of the coerror function $\mathrm{i}^n \operatorname{erfc} x$, $n = -1, 0, 1, 2, \ldots$; $x > 0$.

These are defined by

$$\mathrm{i}^{-1}\operatorname{erfc} x = \frac{2}{\sqrt{\pi}} e^{-x^2}, \quad \mathrm{i}^n \operatorname{erfc} x = \int_x^\infty \mathrm{i}^{n-1}\operatorname{erfc} t\, dt, \quad n = 0, 1, 2, \ldots .$$

Let

$$f_n = \mathrm{i}^{n-1}\operatorname{erfc} x, \quad n = 0, 1, 2, \ldots .$$

Then $\{f_n\}$ is a solution of the difference equation

$$y_{n+1} + \frac{x}{n} y_n - \frac{1}{2n} y_{n-1} = 0, \quad n = 1, 2, 3, \ldots . \tag{3.34}$$

A second solution is given by

$$g_n = (-1)^n \mathrm{i}^{n-1}\operatorname{erfc}(-x),$$

and since, for any fixed z (real or complex),

$$\mathrm{i}^n \operatorname{erfc} z = \frac{e^{-\frac{1}{2}z^2}}{2^n \Gamma\left(\frac{n}{2}+1\right)} e^{-\sqrt{2n}\, z} \left[1 + O\left(\frac{1}{\sqrt{n}}\right)\right], \quad n \to \infty \tag{3.35}$$

(cf. [4, Eq. (3.5)]), one finds that for $x > 0$

$$(-1)^{n+1} \frac{g_{n+1}}{f_{n+1}} = \frac{\mathrm{i}^n \operatorname{erfc}(-x)}{\mathrm{i}^n \operatorname{erfc} x} \sim e^{2\sqrt{2n}\, x}, \quad n \to \infty, \tag{3.36}$$

showing that $\{f_n\}$ is a minimal solution of (3.34). In this case, no normalizing series (3.14) is required, since we know $f_0 = \frac{2}{\sqrt{\pi}} e^{-x^2}$ and we can take $\lambda_0 = 1$, $\lambda_m = 0$ for $m > 0$, and $s = f_0$ in the algorithm of Theorem 3.2 (i.e., all $s_{n-1}^{(\nu)} = 0$).

It can be shown that ([4, §4])

$$|\rho_n| = \left|\frac{\mathrm{i}^{n-1} \operatorname{erfc}(-x)}{\mathrm{i}^{n-1} \operatorname{erfc} x}\right|, \quad n = 0, 1, 2, \ldots, \tag{3.37}$$

is monotonically increasing for any fixed $x > 0$. Moreover, from (3.8) (with $s = \nu$, $t = n$) and (3.20), we see that the approximations $f_n^{(\nu)}$ produced by the algorithm (3.16) have relative errors

$$\frac{f_n^{(\nu)} - f_n}{f_n} = \frac{1 - \rho_n}{\rho_{\nu+1} - 1}, \quad n = 0, 1, \ldots, N+1,$$

so that, up to an additive term of $O(\rho_{\nu+1}^{-2})$,

$$\left|\frac{f_n^{(\nu)} - f_n}{f_n}\right| \le \frac{1 + |\rho_n|}{|\rho_{\nu+1}|} \le \frac{1 + |\rho_{N+1}|}{|\rho_{\nu+1}|} \le \frac{2|\rho_{N+1}|}{|\rho_{\nu+1}|}.$$

To guarantee a relative error ε for all $f_n^{(\nu)}$, $0 \le n \le N+1$, it suffices, therefore, to choose ν such that

$$\frac{2|\rho_{N+1}|}{|\rho_{\nu+1}|} \le \varepsilon. \tag{3.38}$$

Assuming N (and hence ν) large enough for the O-term in (3.35) to be negligible, we obtain from (3.38) and (3.37)

$$e^{-2\sqrt{2\nu}\, x} \le \frac{\varepsilon}{2e^{2\sqrt{2N}x}},$$

that is,

$$\nu \geq \left(\frac{\ln \frac{1}{\varepsilon} + 2\sqrt{2N}x + \ln 2}{2\sqrt{2}\,x} \right)^2 .$$

Since the dominance of the second solution g_n, by (3.36), becomes weaker as x decreases, one may get away with forward recursion for x sufficiently small and N not too large. This can be carefully analyzed and implemented in a general-purpose procedure to compute f_n, or a suitably normalized version of f_n that avoids over- or underflow as much as possible (cf. [9, 10]).

3.4. Orthogonal Polynomials. An important source of linear second-order difference equations are orthogonal polynomials relative to some (nonnegative) mass distribution $d\sigma$ supported on a finite or infinite interval, or on a finite set of points on the real line. In the former case, if all moments

$$m_n = \int_{\mathbb{R}} x^n d\sigma(x), \qquad n = 0, 1, 2, \ldots,$$

exist and are finite, there are infinitely many orthogonal polynomials $p_n(x) = p_n(x; d\sigma)$, $n = 0, 1, 2, \ldots$, in the latter case exactly N of them, $p_0, p_1, \ldots, p_{N-1}$, where N is the number of support points of $d\sigma = d\sigma_N$. In either case, if assumed monic, they satisfy the difference equation

$$y_{n+1} = (x - a_n)y_n - b_n y_{n-1}, \qquad n = 0, 1, 2, \ldots, \tag{3.39}$$

with starting values

$$y_{-1} = 0, \quad y_0 = 1 \qquad (y_n = p_n(\,\cdot\,; d\sigma)).$$

The coefficients a_n, b_n are uniquely determined by $d\sigma$, except for b_0, which is conveniently defined to be the total mass, $b_0 = \int_{\mathbb{R}} d\sigma(x)$.

Generally speaking, all solutions of (3.39) behave similarly if x is located on the support interval of $d\sigma$. (Chebyshev polynomials on [-1,1] are a case in point!) This is no longer true if x is outside the support interval, as will be seen in the next subsection. There is also a phenomenon of pseudostability that may occur in connection with discrete orthogonal polynomials (corresponding to a discrete mass distribution). This will be discussed in §3.4.2.

3.4.1. *Associated functions.* There is an interesting set of functions associated with $d\sigma$ and the orthogonal polynomials $\{p_n(\,\cdot\,; d\sigma)\}$, defined

by

$$q_n(z) = q_n(z; d\sigma) = \int_{\mathbb{R}} \frac{p_n(x; d\sigma)}{z - x}\, d\sigma(x), \qquad n = 0, 1, 2, \ldots,$$

where z is real or complex and assumed outside the support interval of $d\sigma$. These satisfy exactly the same difference equation (3.39) (with b_0 as defined above) as the one for orthogonal polynomials, but with starting values

$$y_{-1} = 1, \quad y_0 = \int_{\mathbb{R}} \frac{d\sigma(x)}{z - x} \quad (y_n = q_n(z; d\sigma)), \tag{3.40}$$

and with x in (3.39) replaced by z. Moreover, they constitute a minimal solution of (3.39), inasmuch as

$$\lim_{n\to\infty} \frac{q_n(z; d\sigma)}{p_n(z; d\sigma)} = 0, \qquad z \in \mathbb{C} \setminus \operatorname{supp} d\sigma, \tag{3.41}$$

at least for the class of distributions $d\sigma$ which give rise to a determined moment problem [11]. This includes *all* distributions $d\sigma$ supported on a finite interval, and many others with unbounded support.

Since we know the initial value for $n = -1$, Theorem 3.2 provides a simple algorithm to compute $q_n(z)$ for $0 \le n \le N$:

$$\begin{aligned} &r_\nu^{(\nu)} = 0, \quad r_{n-1}^{(\nu)} = \frac{b_n}{z - a_n - r_n^{(\nu)}}, \qquad n = \nu, \nu - 1, \ldots, 1, 0, \\ &q_{-1}^{(\nu)} = 1, \quad q_n^{(\nu)} = r_{n-1}^{(\nu)} q_{n-1}^{(\nu)}, \quad n = 0, 1, \ldots, N. \end{aligned} \tag{3.42}$$

Under the assumptions above, we have

$$\lim_{\nu\to\infty} q_n^{(\nu)} = q_n(z), \qquad n = 0, 1, \ldots, N.$$

The algorithm is of interest even in the case $N = 0$, as it allows us to compute the "Cauchy transform" of the distribution $d\sigma$ (cf. y_0 in (3.40)). Another interesting application is to Gaussian quadrature (with weight distribution $d\sigma$ supported on a finite interval), since the ratio in (3.41) is nothing but the kernel in the remainder term

$$R_n(f) = \frac{1}{2\pi i} \oint_\Gamma \frac{q_n(z; d\sigma)}{p_n(z; d\sigma)}\, f(z) dz$$

of the quadrature rule [15]. Here, Γ is a contour surrounding the support interval of $d\sigma$ and f is assumed analytic in the domain enclosed by Γ.

3.4.2. *Discrete orthogonal polynomials.* Numerical difficulties with discrete orthogonal polynomials have been known for some time, but were recognized only recently [12] to be attributable to a phenomenon of pseudostability. Pseudostability is relevant also in cases, such as here, where the difference equation (3.39) holds only for a finite number of n-values, $n = 0, 1, \ldots, N-1$ (where $n = N-1$ yields $p_N(\cdot\,; d\sigma_N)$, which, though well defined, is no longer orthogonal, since it vanishes at all N support points of $d\sigma_N$). It simply means that the amplification factors $\omega_{s\to t}$ defined in (3.5), as s and t vary over $0 \le s < t \le N-1$, may become very large in parts of this region. (Isolated large values may be due to near zeros of $p_n(\cdot)$ and need not be of any concern.)

We illustrate the phenomenon in the simplest case of discrete Legendre polynomials, i.e., the polynomials $p_n(\cdot\,; d\sigma_N)$ orthogonal with respect to the discrete N-point distribution $d\sigma_N$ having support points x_j and jumps w_j, where

$$x_j = -1 + 2\,\frac{j-1}{N-1}, \qquad w_j = \frac{2}{N}, \qquad j = 1, 2, \ldots, N.$$

The respective recursion coefficients a_n, b_n are known explicitly:

$$a_n = 0, \qquad n = 0, 1, \ldots, N-1;$$

$$b_0 = 2, \quad b_n = \left(1 + \tfrac{1}{N-1}\right)^2 \frac{1 - \left(\frac{n}{N}\right)^2}{4 - \frac{1}{n^2}}, \qquad n = 1, 2, \ldots, N-1.$$

It is thus easy to generate the solutions of (3.39) with starting values $y_{-1} = 0$, $y_0 = 1$ (producing $f_n = p_n(\cdot\,; d\sigma_N)$) and $y_{-1} = 1$, $y_0 = 0$ (producing a linearly independent solution, g_n). The respective amplification factors $\omega_{s\to t}$ in (3.5) are then readily computed (paying attention to footnote (§)).

In Figure 3.1(a)–(d) are shown two-dimensional plots of $\omega_{s\to t}$ on a logarithmic vertical scale, for $N = 40$ and $x = x_j$, $j = 1, 5, 10, 20$. (By symmetry, this covers also the cases $x = x_k$ $k = N - j + 1$.) It can be seen that pseudostability starts developing for $t > s$ as t approaches N. It is rather pronounced when x is one of the lateral support points (cf. (a), (b) of Figure 3.1) and much less so as x moves toward the center of the interval [-1,1] (cf. (c), (d) of Figure 3.1).

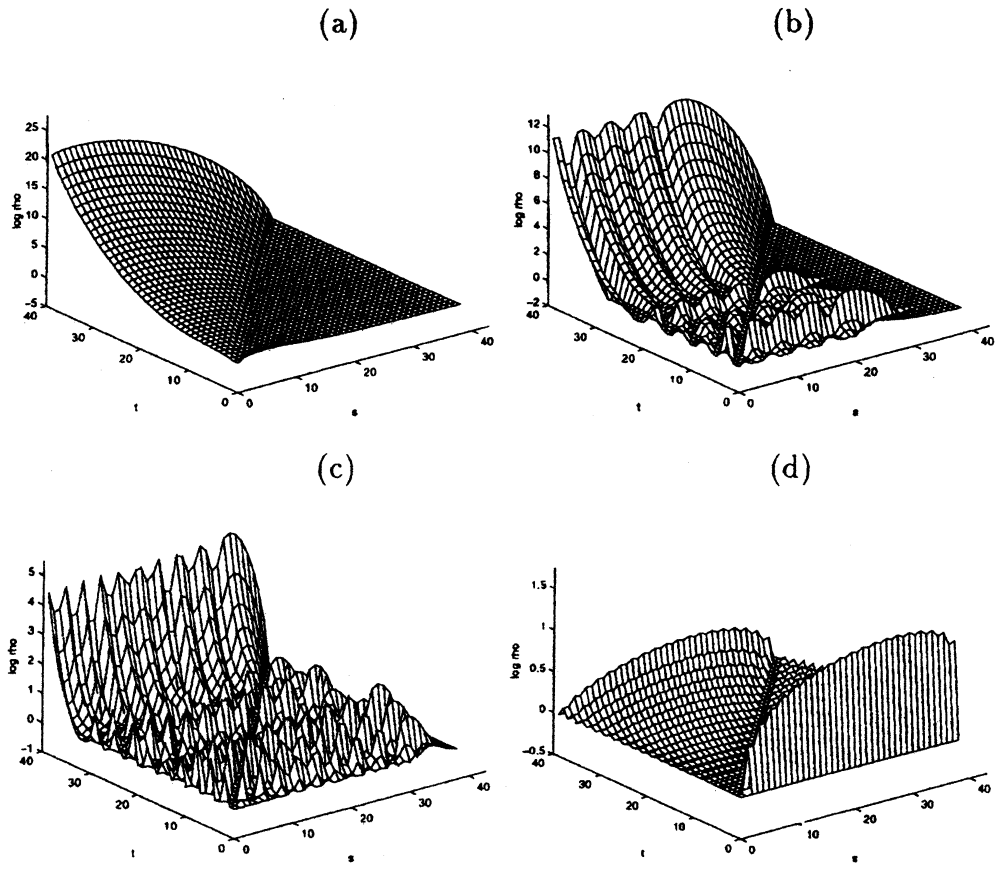


FIGURE 3.1. Amplification factors $\omega_{s\to t}$ for discrete Legendre polynomials, $N = 40$
(a) $x = x_1$ (b) $x = x_5$ (c) $x = x_{10}$ (d) $x = x_{20}$

Similar phenomena can be observed for other discrete orthogonal polynomials with equally spaced support points, e.g., the Krawtchouk polynomials ([12, Example 3.2]). Nonequally spaced points, such as Chebyshev points, on the other hand, seem to stay clear from such problems of pseudostability ([12, Example 3.3]).

REFERENCES

[1] British Association for the Advancement of Science, *Mathematical Tables, vol. X, Bessel functions, Part II, Functions of positive integer order*, Cambridge University Press, 1952.

[2] J.R. Cash and R.V.M. Zahar, A unified approach to recurrence algorithms, in *Approximation and Computation* (R.V.M. Zahar, ed.), Birkhäuser, Boston, 1994, pp. 97–120.

[3] P.J. Davis, Leonhard Euler's integral: A historical profile of the gamma function, *Amer. Math. Monthly* **66** (1959), 849–869.

[4] W. Gautschi, Recursive computation of the repeated integrals of the error function, *Math. Comp.* **15** (1961), 227–232.

[5] W. Gautschi, Computational aspects of three-term recurrence relations, *SIAM Rev.* **9** (1967), 24–82.

[6] W. Gautschi, An application of minimal solutions of three-term recurrences to Coulomb wave functions, *Aequationes Math.* **2** (1969), 171–176.

[7] W. Gautschi, Algorithm 471 — Exponential integrals, *Comm. ACM* **16** (1972), 761–763.

[8] W. Gautschi, Computational methods in special functions — a survey, in *Theory and Application of Special Functions* (R.A. Askey, ed.), Academic Press, New York, 1975, pp. 1–98.

[9] W. Gautschi, Evaluation of the repeated integrals of the coerror function, *ACM Trans. Math. Software* **3** (1977), 240–252.

[10] W. Gautschi, Algorithm 521 — Repeated integrals of the coerror function, *ACM Trans. Math. Software* **3** (1977), 301–302.

[11] W. Gautschi, Minimal solutions of three-term recurrence relations and orthogonal polynomials, *Math. Comp.* **36** (1981), 547–554.

[12] W. Gautschi, Is the recurrence relation for orthogonal polynomials always stable?, *BIT* **33** (1993), 277–284.

[13] W. Gautschi and B.J. Klein, Recursive computation of certain derivatives — a study of error propagation, *Comm. ACM* **13** (1970), 7–9.

[14] W. Gautschi and J. Slavik, On the computation of modified Bessel function ratios, *Math. Comp.* **32** (1978), 865–875.

[15] W. Gautschi and R.S. Varga, Error bounds for Gaussian quadrature of analytic functions, *SIAM J. Numer. Anal.* **20** (1983), 1170–1186.

[16] 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* (W. Gautschi, ed.), American Mathematical Society, Providence, 1994, pp. 79–125.

[17] G.F. Miller, *Tables of generalized exponential integrals*, Mathematical Tables, vol. 3, National Physical Laboratory, Her Majesty's Stationery Office, London, 1960.

[18] F.W.J. Olver, Numerical solution of second-order linear difference equations, *J. Res. Nat. Bur. Standards* **71B** (1967), 111–129.

[19] F.W.J. Olver, Error bounds for linear recurrence relations, *Math. Comp.* **50** (1988), 481–499.

[20] F.W.J. Olver and D.J. Sookne, Note on backward recurrence algorithms, *Math. Comp.* **26** (1972), 941–947.

[21] S. Pincherle, Delle funzioni ipergeometriche e di varie questioni ad esse attinenti, *Giorn. Mat. Battaglini* **32** (1894), 209–291. [Opere Scelte, vol. 1, pp. 273–357.]

[22] J. Wimp, *Computation with Recurrence Relations*, Pitman, Boston, 1984.

27

Papers on Ordinary Differential Equations

---

14 Numerical integration of ordinary differential equations based on trigonometric polynomials, Numer. Math. 3, 381–397 (1961)

54 Global error estimates in "one-step" methods for ordinary differential equations, Rend. Mat. (2) 8, 601–617 (1975) (translated from Italian)

73 (with M. Montrone) Multistep methods with minimum global error coefficient, Calcolo 17, 67–75 (1980) (translated from Italian)

27.1. [14] "Numerical integration of ordinary differential equations based on trigonometric polynomials"

---

[14] "Numerical integration of ordinary differential equations based on trigonometric polynomials," *Numer. Math*. **3**, 381–397 (1961).

Numerische Mathematik 3, 381–397 (1961)

Numerical integration of ordinary differential equations based on trigonometric polynomials

By

WALTER GAUTSCHI*

There are many numerical methods available for the step-by-step integration of ordinary differential equations. Only few of them, however, take advantage of special properties of the solution that may be known in advance. Examples of such methods are those developed by BROCK and MURRAY [2], and by DENNIS [4], for exponential type solutions, and a method by URABE and MISE [5] designed for solutions in whose Taylor expansion the most significant terms are of relatively high order. The present paper is concerned with the case of periodic or oscillatory solutions where the frequency, or some suitable substitute, can be estimated in advance. Our methods will integrate exactly appropriate trigonometric polynomials of given order, just as classical methods integrate exactly algebraic polynomials of given degree. The resulting methods depend on a parameter, $v=h\omega$, where h is the step length and ω the frequency in question, and they reduce to classical methods if $v\to 0$. Our results have also obvious applications to numerical quadrature. They will, however, not be considered in this paper.

1. Linear functionals of algebraic and trigonometric order

In this section $[a, b]$ is a finite closed interval and $C^s[a, b]$ $(s\geqq 0)$ denotes the linear space of functions $x(t)$ having s continuous derivatives in $[a, b]$. We assume $C^s[a, b]$ normed by

$$\|x\| = \sum_{\sigma=0}^{s} \max_{a\leqq t\leqq b} |x^{(\sigma)}(t)|. \tag{1.1}$$

A linear functional L in $C^s[a, b]$ is said to be of *algebraic order* p, if

$$Lt^r = 0 \qquad (r = 0, 1, \ldots, p); \tag{1.2}$$

it is said to be of *trigonometric order* p, *relative to period* T, if

$$L1 = L\cos\left(r\frac{2\pi}{T}t\right) = L\sin\left(r\frac{2\pi}{T}t\right) = 0 \qquad (r = 1, 2, \ldots, p). \tag{1.3}$$

Thus, a functional L is of algebraic order p if it annihilates all algebraic polynomials of degree $\leqq p$, and it is of trigonometric order p, relative to period T, if it annihilates all trigonometric polynomials of order $\leqq p$ with period T.

Functionals of trigonometric order p are comparable with those of algebraic order $2p$, in the sense that both involve the same number of conditions. The

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

relationship turns out to be much closer if we let L depend on the appropriate number of parameters. In fact, consider functionals of the form

$$L x = \beta_1 L_1 x + \cdots + \beta_{2p} L_{2p} x + L_{2p+1} x, \tag{1.4}$$

where L_λ ($\lambda \geqq 1$) are fixed linear continuous functionals in $C^s[a, b]$ and β_λ real parameters. Then the following theorem holds.

Theorem 1. *Let the functionals L_λ in* (1.4) *satisfy the following conditions:*

(i) $L_\lambda 1 = 0 \quad (\lambda = 1, 2, \ldots, 2p+1)$.

(ii) *There is a unique set of parameters, $\beta_\lambda = \beta_\lambda^0$, such that the functional L in* (1.4) *is of algebraic order $2p$, that is to say,*

$$\det(L_\lambda t^\varkappa) \neq 0 \qquad \begin{pmatrix} \varkappa \text{ row index, } \lambda \text{ column index} \\ \varkappa, \lambda = 1, 2, \ldots, 2p \end{pmatrix}. \tag{1.5}$$

Then, for T sufficiently large, there is also a unique set of parameters, $\beta_\lambda = \beta_\lambda(T)$, such that L is of trigonometric order p relative to period T. Furthermore,

$$\beta_\lambda(T) \to \beta_\lambda^0 \quad \text{as} \quad T \to \infty. \tag{1.6}$$

Proof. The main difficulty in the proof is the fact that in the limit, as $T \to \infty$, equations (1.3) degenerate into one single equation, $L 1 = 0$. We therefore transform (1.3) into an equivalent set of equations from which the behavior of the solution at $T = \infty$ can be studied more easily.

In this connection the following trigonometric identities are helpful,

$$\sin^{2r} \frac{x}{2} = \sum_{\varrho=1}^{r} \sigma_{r\varrho} (1 - \cos \varrho x) \qquad (r = 1, 2, 3, \ldots), \tag{1.7}$$

where $\sigma_{r\varrho}$ are suitable real numbers and $\sigma_{rr} \neq 0$. The existence of such numbers is obvious, if one observes that $\sin^{2r} \frac{x}{2} = [(1 - \cos x)/2]^r$ can be written as a cosine-polynomial of exact order r. Differentiating both sides in (1.7) gives also

$$\sin^{2r-1} \frac{x}{2} \cos \frac{x}{2} = \sum_{\varrho=1}^{r} \tau_{r\varrho} \sin \varrho x \qquad (r = 1, 2, 3, \ldots), \tag{1.8}$$

where $\tau_{r\varrho} = \varrho \sigma_{r\varrho}/r$, and in particular $\tau_{rr} = \sigma_{rr} \neq 0$.

Equations (1.3) are equivalent to

$$L 1 = 0,$$

$$L\left(1 - \cos r \frac{2\pi}{T} t\right) = L \sin r \frac{2\pi}{T} t = 0 \qquad (r = 1, 2, \ldots, p).$$

Because of assumption (i) the first of these equations is automatically satisfied. The remaining equations are equivalent to

$$\sum_{\varrho=1}^{r} \sigma_{r\varrho} L\left(1 - \cos \varrho \frac{2\pi}{T} t\right) = \sum_{\varrho=1}^{r} \tau_{r\varrho} L \sin \varrho \frac{2\pi}{T} t = 0 \qquad (r = 1, 2, \ldots, p). \tag{1.9}$$

Using (1.7) and the linearity of L we have

$$\sum_{\varrho=1}^{r} \sigma_{r\varrho} L\left(1-\cos\varrho\frac{2\pi}{T}t\right) = L\sum_{\varrho=1}^{r}\sigma_{r\varrho}\left(1-\cos\varrho\frac{2\pi}{T}t\right) = L\left[\sin^{2r}\left(\frac{\pi}{T}t\right)\right].$$

Similarly, using (1.8), we find

$$\sum_{\varrho=1}^{r} \tau_{r\varrho} L\sin\varrho\frac{2\pi}{T}t = L\left[\sin^{2r-1}\left(\frac{\pi}{T}t\right)\cos\frac{\pi}{T}t\right].$$

Therefore, letting

$$u=\frac{\pi}{T}, \tag{1.10}$$

we can write (1.9), after suitable multiplications, as follows:

$$\begin{aligned} L\left[\left(\frac{\sin u t}{u}\right)^{2r-1}\cos u t\right] &= 0 \\ L\left[\left(\frac{\sin u t}{u}\right)^{2r}\right] &= 0 \end{aligned} \qquad (r=1,2,\ldots,p). \tag{1.11}$$

This represents a system of $2p$ linear algebraic equations in the same number of unknowns β_λ, the coefficient matrix and known vector of which both depend on the parameter u. We show that in the limit as $u\to 0$ the system (1.11) goes over into the system of equations $Lt^r=0$ $(r=1,2,\ldots,2p)$.

In fact, it is readily seen, by expansion or otherwise, that for any integers $\sigma\geqq 0$, $r\geqq 1$, as $u\to 0$,

$$\begin{aligned} \frac{d^\sigma}{dt^\sigma}\left[\left(\frac{\sin u t}{u}\right)^{2r-1}\cos u t\right] &\to \frac{d^\sigma}{dt^\sigma}t^{2r-1}, \\ \frac{d^\sigma}{dt^\sigma}\left(\frac{\sin u t}{u}\right)^{2r} &\to \frac{d^\sigma}{dt^\sigma}t^{2r}, \end{aligned}$$

the convergence being uniform with respect to t in any finite interval. In particular,

$$\begin{aligned} \left\|\left(\frac{\sin u t}{u}\right)^{2r-1}\cos u t - t^{2r-1}\right\| &\to 0 \\ \left\|\left(\frac{\sin u t}{u}\right)^{2r} - t^{2r}\right\| &\to 0 \end{aligned} \qquad (u\to 0),$$

so that, by the continuity of the L_λ, also

$$\begin{aligned} L_\lambda\left[\left(\frac{\sin u t}{u}\right)^{2r-1}\cos u t\right] &\to L_\lambda t^{2r-1} \\ L_\lambda\left(\frac{\sin u t}{u}\right)^{2r} &\to L_\lambda t^{2r} \end{aligned} \qquad (u\to 0).$$

From this our assertion follows immediately.

Since the limiting system, by assumption, has a unique solution, β_λ^0, the matrix of the system (1.11) is nonsingular for $u=0$, and hence remains so for u sufficiently small. It follows that for sufficiently large T there is a unique solution, $\beta_\lambda(T)$, of (1.11), satisfying (1.6). Theorem 1 is proved.

Remark 1. Assumption (i) in Theorem 1 is not essential, but convenient for some of the applications made later. The theorem holds without the assumption (i) if the functional L in (1.4) is made to depend on $2p+1$ parameters,

$$L x = \beta_0 L_0 x + \beta_1 L_1 x + \cdots + \beta_{2p} L_{2p} x + L_{2p+1} x, \tag{1.4'}$$

and assumption (1.5) is modified, accordingly, to

$$\det(L_\lambda t^\varkappa) \neq 0 \qquad \begin{pmatrix} \varkappa \text{ row index, } \lambda \text{ column index} \\ \varkappa, \lambda = 0, 1, \ldots, 2p \end{pmatrix}. \tag{1.5'}$$

The proof remains the same.

Remark 2. For particular choices of the L_λ it may happen that the functional L can be made of higher algebraic order than the number of parameters would indicate. Even if the excess in order is a multiple of 2, this does not mean necessarily that a similar increase in trigonometric order is possible. For example,

$$L x = \beta\, x(0) + x(1) - \tfrac{1}{2} x'(0) - \tfrac{1}{2} x'(1), \qquad \beta = -1$$

if of algebraic order 2, but in general cannot be made of trigonometric order 1, since

$$L\left[\frac{\sin u t}{u} \cos u t\right] = \frac{\sin 2u}{2u} - \frac{1}{2}(1 + \cos 2u) > 0 \quad \left(0 < u < \frac{\pi}{2}\right).$$

2. Linear multi-step methods

Linear functionals in C^1 play an important rôle in the numerical solution of first order differential equations

$$x' = f(t, x), \qquad x(t_0) = x_0, \tag{2.1}$$

in that they provide the natural mathematical setting for a large class of numerical methods, the so-called linear multi-step methods. These are methods which define approximations x_m to values $x(t_0 + mh)$ of the desired solution by a relation of the following form

$$\begin{aligned} x_{n+1} + \alpha_1 x_n + \cdots + \alpha_k x_{n+1-k} &= h(\beta_0 x'_{n+1} + \beta_1 x'_n + \cdots + \beta_k x'_{n+1-k}) \\ &(n = k-1, k, k+1, \ldots), \end{aligned} \tag{2.2}$$

where

$$x'_m = f(t_0 + mh, x_m).$$

Once k "starting" values $x_0, x_1, \ldots, x_{k-1}$ are known, (2.2) is used to obtain successively all approximations x_m $(m \geqq k)$ desired.

The integer $k > 0$ will be called the *index* of the multi-step method, assuming, of course, that not both α_k and β_k vanish. (2.2) is called an *extrapolation method* if $\beta_0 = 0$, and an *interpolation method* if $\beta_0 \neq 0$. Interpolation methods require the solution of an equation at each stage, because x'_{n+1} in (2.2) is itself a function of the new approximation x_{n+1}.

It is natural to associate with (2.2) the linear functional

$$L x = \sum_{\lambda=0}^{k} [\alpha_\lambda x(t_0 + (n+1-\lambda)h) - h\beta_\lambda x'(t_0 + (n+1-\lambda)h)] \qquad (\alpha_0 = 1). \tag{2.3}$$

The multi-step method (2.2) is called of algebraic order p, if its associated linear functional (2.3) is of algebraic order p; similarly one defines trigonometric order of a multi-step method.

Since any linear transformation $t'=at+b$ $(a\neq 0)$ of the independent variable transforms an algebraic polynomial of degree $\leqq p$ into one of the same kind, it is clear that (2.2) is of algebraic order p if and only if the functional

$$L^1 x=\sum_{\lambda=0}^{k}\left[\alpha_\lambda x(k-\lambda)-\beta_\lambda x'(k-\lambda)\right] \tag{2.4}$$

is of algebraic order p. Here, the parameter h has dropped out, so that the coefficients α_λ, β_λ of a multi-step method of algebraic order do not depend on h. The situation is somewhat different in the trigonometric case, where a linear transformation other than a translation (or reflexion) changes the period of a trigonometric polynomial. By a translation, however, it is seen that (2.2) is of trigonometric order p, relative to period T, if and only if

$$L^h x=\sum_{\lambda=0}^{k}\left\{\alpha_\lambda x\left[(k-\lambda)h\right]-h\beta_\lambda x'\left[(k-\lambda)h\right]\right\} \tag{2.5}$$

is of trigonometric order p relative to period T.

For a multi-step method to be useful it must be numerically stable, which above all imposes certain restrictions on the coefficients α_λ (see, e.g., [*1*, sec. 9]). In view of this we shall consider the α_λ as prescribed numbers satisfying the conditions of stability. Also they shall satisfy

$$\sum_{\lambda=0}^{k}\alpha_\lambda=0 \qquad (\alpha_0=1) \tag{2.6}$$

to insure algebraic and trigonometric order $p=0$.

It is then well known ([*1*, sec. 6]) that to any given set of $k+1$ coefficients α_λ satisfying (2.6) there corresponds a unique extrapolation method with index k and algebraic order k. Letting therefore $k=2p$ we can apply Theorem 1 to $L=L^h$, identifying

$$L_\lambda x=-h x'\left[(2p-\lambda)h\right] \quad (1\leqq\lambda\leqq 2p), \qquad L_{2p+1}x=\sum_{\lambda=0}^{2p}\alpha_\lambda x\left[(2p-\lambda)h\right]. \tag{2.7}$$

It follows that there exists a unique extrapolation method with *even* index $k=2p$ and trigonometric order p relative to any sufficiently large period T. Again, as is well known, given $k+1$ coefficients α_λ, there corresponds a unique interpolation method with index k and algebraic order $k+1$. Letting now $k+1=2p$, a similar application of Theorem 1 shows the existence, for T sufficiently large, of an interpolation method with *odd* index $k=2p-1$ and trigonometric order p relative to period T. Furthermore, in the limit as $T\to\infty$, the resulting methods of trigonometric order p reduce to those of algebraic order $2p$.

The essential parameter is actually not T, but h/T, as is seen if the conditions (1.11) of trigonometric order p are written down for the functional L^h. Since

$$\frac{d}{dt}\left[\left(\frac{\sin u t}{u}\right)^{2r-1}\cos u t\right]=\left(\frac{\sin u t}{u}\right)^{2r-2}(2r\cos^2 u t-1),$$

$$\frac{d}{dt}\left[\left(\frac{\sin u t}{u}\right)^{2r}\right]=2r\left(\frac{\sin u t}{u}\right)^{2r-1}\cos u t$$

one finds[1]

$$h\sum_{\lambda=0}^{k}\beta_\lambda\left(\frac{\sin[u(k-\lambda)h]}{u}\right)^{2r-2}\left(2r\cos^2[u(k-\lambda)h]-1\right)$$
$$=\sum_{\lambda=0}^{k}\alpha_\lambda\left(\frac{\sin[u(k-\lambda)h]}{u}\right)^{2r-1}\cos[u(k-\lambda)h],$$

$$2rh\sum_{\lambda=0}^{k}\beta_\lambda\left(\frac{\sin[u(k-\lambda)h]}{u}\right)^{2r-1}\cos[u(k-\lambda)h]=\sum_{\lambda=0}^{k}\alpha_\lambda\left(\frac{\sin[u(k-\lambda)h]}{u}\right)^{2r}.$$

Dividing the first relation by h^{2r-1}, and the second relation by h^{2r}, and letting

$$v=2uh=\frac{2\pi}{T}h,$$

one gets[1]

$$\sum_{\lambda=0}^{k}\beta_\lambda\left(\frac{2\sin[\frac{1}{2}(k-\lambda)v]}{v}\right)^{2r-2}\left(2r\cos^2\left[\frac{1}{2}(k-\lambda)v\right]-1\right)$$
$$=\sum_{\lambda=0}^{k}\alpha_\lambda\left(\frac{2\sin[\frac{1}{2}(k-\lambda)v]}{v}\right)^{2r-1}\cos\left[\frac{1}{2}(k-\lambda)v\right], \tag{2.8}$$
$$2r\sum_{\lambda=0}^{k}\beta_\lambda\left(\frac{2\sin[\frac{1}{2}(k-\lambda)v]}{v}\right)^{2r-1}\cos\left[\frac{1}{2}(k-\lambda)v\right]=\sum_{\lambda=0}^{k}\alpha_\lambda\left(\frac{2\sin[\frac{1}{2}(k-\lambda)v]}{v}\right)^{2r}$$
$$(r=1,2,\ldots,p).$$

We summarize our findings in the following

Theorem 2. *In correspondence to each set of coefficients α_λ with zero sum there exist unique sets of coefficients $\beta_\lambda(v)$, $\beta_\lambda^*(v)$ depending on the parameter*

$$v=2\pi h/T,$$

such that for v sufficiently small,

$$x_{n+1}+\alpha_1 x_n+\cdots+\alpha_{2p}x_{n+1-2p}=h[\beta_1(v)x_n'+\cdots+\beta_{2p}(v)x_{n+1-2p}'] \tag{2.9}$$

is an extrapolation method of trigonometric order p relative to period T, and

$$x_{n+1}+\alpha_1 x_n+\cdots+\alpha_{2p-1}x_{n+2-2p}$$
$$=h[\beta_0^*(v)x_{n+1}'+\beta_1^*(v)x_n'+\cdots+\beta_{2p-1}^*(v)x_{n+2-2p}'] \tag{2.10}$$

is an interpolation method of trigonometric order p relative to period T. The $\beta_\lambda(v)$ solve the system of linear algebraic equations (2.8) *with $k=2p$, $\beta_0=0$, the $\beta_\lambda^*(v)$ solve the same system with $k=2p-1$ and with no restrictions on the β's. As $v\to 0$ the multi-step methods* (2.9) *and* (2.10) *reduce to those of algebraic order $2p$, respectively.*

3. Existence criterion for trigonometric multi-step methods

Theorem 2 establishes the existence of trigonometric multi-step methods only for $v=2\pi h/T$ sufficiently small. A more precise condition on v is furnished by the following

[1] If $r=1$ the coefficient of β_k in the first relation, to be meaningful, must be defined as unity.

Theorem 3. *Multi-step methods* (2.9) *and* (2.10) *of trigonometric order* p, *relative to period* T, *exist if*

$$|v| < \min\left(v_p, \frac{2\pi}{2p-1}\right) \qquad (v = 2\pi h/T), \tag{3.1}$$

where v_p *is the smallest positive zero of the cosine-polynomial*

$$C_p(v) = \begin{cases} \sum_{n=1}^{(p^2+1)/2} v_p\left(p^2 - \frac{1}{2}p + \frac{1}{2} - n\right)\cos(2n-1)\frac{v}{2} & (p \text{ odd}) \\ \frac{1}{2}v_p\left(p^2 - \frac{1}{2}p\right) + \sum_{n=1}^{p^2/2} v_p\left(p^2 - \frac{1}{2}p - n\right)\cos n v & (p \text{ even}). \end{cases} \tag{3.2}$$

Here, $v_p(m)$ *denotes the number of combinations of* p *nonnegative*[2] *integers not exceeding* $2p-1$ *which have the sum* m.

Proof. The linear functional associated with the extrapolation method (2.9) is

$$L x = \sum_{\lambda=1}^{2p} \beta_\lambda L_\lambda x + L_{2p+1} x,$$

where $L_\lambda x = -h x'[(2p-\lambda)h]$ $(1 \leqq \lambda \leqq 2p)$ and L_{2p+1} is given such that $L_{2p+1} 1 = 0$. Similarly,

$$L^* x = \sum_{\lambda=0}^{2p-1} \beta_\lambda^* L_\lambda^* x + L_{2p}^* x$$

with $L_\lambda^* = L_{\lambda+1}$, $L_{2p}^* 1 = 0$, is the functional associated with the interpolation method (2.10). It is apparent, therefore, that the conditions (1.3) of trigonometric order for these particular functionals give rise to a system of $2p$ linear algebraic equations in the unknowns β_λ and β_λ^*, respectively, the matrix of which in either case is given by

$$B(v) = \begin{pmatrix} v\sin(2p-1)v & v\sin(2p-2)v & \dots & v\sin v & 0 \\ -v\cos(2p-1)v & -v\cos(2p-2)v & \dots & -v\cos v & -v \\ 2v\sin 2(2p-1)v & 2v\sin 2(2p-2)v & \dots & 2v\sin 2v & 0 \\ -2v\cos 2(2p-1)v & -2v\cos 2(2p-2)v & \dots & -2v\cos 2v & -2v \\ \dots & \dots & \dots & \dots & \dots \\ p\,v\sin p(2p-1)v & p\,v\sin p(2p-2)v & \dots & p\,v\sin p\,v & 0 \\ -p\,v\cos p(2p-1)v & -p\,v\cos p(2p-2)v & \dots & -p\,v\cos p\,v & -p\,v \end{pmatrix}.$$

The instance $v=0$ (in which B is singular) is sufficiently dealt with by Theorem 2. Theorem 3 will therefore be proved if it is shown that $B(v)$ is nonsingular for all nonvanishing values of v satisfying (3.1).

Replacing the trigonometric functions in $B(v)$ by Euler's expressions and applying a few obvious elementary operations on rows and columns of the

[2] In terms of partitions (more commonly used in combinatorial analysis) which involve *positive* integers with given sum, we have

$$v_p(m) = \pi_{p-1}(2p-1, m) + \pi_p(2p-1, m),$$

where $\pi_k(l, m)$ denotes the number of partitions of m into k unequal parts not exceeding l.

resulting matrix, one shows that the determinant of B is equal to

$$\det B(v) = (p!)^2\, 2^{-p}\, i^p\, v^{2p}\, e^{-p^2(2p-1)\,iv} \begin{vmatrix} w_{2p-1}^{2p} & w_{2p-2}^{2p} & \dots & w_1^{2p} & w_0^{2p} \\ \dots & \dots & \dots & \dots & \dots \\ w_{2p-1}^{p+1} & w_{2p-2}^{p+1} & \dots & w_1^{p+1} & w_0^{p+1} \\ w_{2p-1}^{p-1} & w_{2p-2}^{p-1} & \dots & w_1^{p-1} & w_0^{p-1} \\ \dots & \dots & \dots & \dots & \dots \\ w_{2p-1} & w_{2p-2} & \dots & w_1 & w_0 \\ 1 & 1 & \dots & 1 & 1 \end{vmatrix} \quad (w_\lambda = e^{i\lambda v}).$$

The last determinant is a minor of the Vandermonde determinant

$$\begin{vmatrix} u^{2p} & w_{2p-1}^{2p} & \dots & w_1^{2p} & w_0^{2p} \\ \dots & \dots & \dots & \dots & \dots \\ u^{p} & w_{2p-1}^{p} & \dots & w_1^{p} & w_0^{p} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 1 & \dots & 1 & 1 \end{vmatrix} = \prod_{\sigma=0}^{2p-1} (u - w_\sigma) \prod_{0\le\sigma<\varrho\le 2p-1} (w_\varrho - w_\sigma),$$

namely, up to the sign $(-1)^p$, the coefficient of u^p in the expansion along the first column. From the right-hand side it is seen that this coefficient is equal to

$$(-1)^p\, \sigma_p(w_0, w_1, \dots, w_{2p-1}) \prod_{0\le\sigma<\varrho\le 2p-1} (w_\varrho - w_\sigma),$$

where σ_p denotes the p-th elementary symmetric function in $2p$ variables. Therefore,

$$\det B(v) = (p!)^2\, 2^{-p}\, i^p\, v^{2p}\, e^{-p^2(2p-1)\,iv}\, \sigma_p(w_0, w_1, \dots, w_{2p-1}) \prod_{0\le\sigma<\varrho\le 2p-1} (w_\varrho - w_\sigma) \quad (w_\lambda = e^{i\lambda v}). \tag{3.3}$$

For the product in (3.3) we have

$$\begin{aligned} \prod_{\sigma<\varrho} (w_\varrho - w_\sigma) &= \prod_{\sigma<\varrho} e^{\frac{1}{2}(\varrho+\sigma)\,iv} \prod_{\sigma<\varrho} \left[e^{\frac{1}{2}(\varrho-\sigma)\,iv} - e^{-\frac{1}{2}(\varrho-\sigma)\,iv}\right] \\ &= (2i)^{p(2p-1)}\, e^{\frac{1}{2}p(2p-1)^2 iv} \prod_{\sigma<\varrho} \sin\tfrac{1}{2}(\varrho-\sigma)\, v. \end{aligned}$$

Also,

$$\sigma_p(w_0, w_1, \dots, w_{2p-1}) = \sum e^{(\lambda_1+\lambda_2+\cdots+\lambda_p)\,iv},$$

where the sum extends over all combinations $(\lambda_1, \lambda_2, \dots, \lambda_p)$ of p nonnegative integers not greater than $2p-1$. Thus,

$$\begin{aligned} \det B(v) = (-1)^p\, (p!)^2\, 2^{2p(p-1)}\, v^{2p} \times \\ \times \left[e^{-\frac{1}{2}p(2p-1)\,iv} \sum e^{(\lambda_1+\cdots+\lambda_p)\,iv}\right] \prod_{0\le\sigma<\varrho\le 2p-1} \sin\tfrac{1}{2}(\varrho-\sigma)\, v. \end{aligned} \tag{3.4}$$

It is seen from this that $B(v)$ for $v \neq 0$ is singular if and only if either the expression in brackets or the product following this expression vanishes.

As regards the first expression we can write it in the form

$$e^{-\frac{1}{2}p(2p-1)\,iv} \sum_{n=p(p-1)/2}^{p(3p-1)/2} \nu_p(n)\, e^{inv} = \sum_{n=p(p-1)/2}^{p(3p-1)/2} \nu_p(n)\, e^{[n-\frac{1}{2}p(2p-1)]\,iv},$$

with $v_p(n)$ as defined in Theorem 3. Consider first the case p even. Then, by a shift of the summation index, the last sum is seen to be

$$\sum_{n=-p^2/2}^{p^2/2} v_p(p^2 - \tfrac{1}{2}p + n)\, e^{n\,i\,v}.$$

Since the determinant (3.4) is real, this sum must be real too, which is only possible if

$$v_p(p^2 - \tfrac{1}{2}p + n) = v_p(p^2 - \tfrac{1}{2}p - n) \quad (p \text{ even}).$$

Our sum then becomes

$$v_p(p^2 - \tfrac{1}{2}p) + 2\sum_{n=1}^{p^2/2} v_p(p^2 - \tfrac{1}{2}p - n)\cos n\,v \quad (p \text{ even}). \tag{3.5}$$

Analogously, if p is odd, the sum in question is

$$\left[\sum_{n=p(p-1)/2}^{(p(2p-1)-1)/2} + \sum_{n=(p(2p-1)+1)/2}^{p(3p-1)/2}\right] v_p(n)\, e^{[n-\frac{1}{2}p(2p-1)]\,i\,v}$$

$$= \sum_{n=1}^{(p^2+1)/2} \left[v_p(p^2 - \tfrac{1}{2}p + \tfrac{1}{2} - n)\, e^{-(2n-1)\,i\,v/2} + v_p(p^2 - \tfrac{1}{2}p - \tfrac{1}{2} + n)\, e^{(2n-1)\,i\,v/2}\right].$$

Since this again must be real we also have

$$v_p(p^2 - \tfrac{1}{2}p + \tfrac{1}{2} - n) = v_p(p^2 - \tfrac{1}{2}p - \tfrac{1}{2} + n) \quad (p \text{ odd}),$$

and our sum becomes

$$2\sum_{n=1}^{(p^2+1)/2} v_p\left(p^2 - \frac{1}{2}p + \frac{1}{2} - n\right)\cos(2n-1)\,\frac{v}{2} \quad (p \text{ odd}). \tag{3.6}$$

Substituting (3.5) and (3.6) for the bracketed expression in (3.4) we finally obtain

$$\det B(v) = (-1)^p\,(p!)^2\,2^{2p^2-2p+1}\,v^{2p}\,C_p(v) \prod_{0\leqq\sigma<\varrho\leqq 2p-1} \sin\tfrac{1}{2}(\varrho-\sigma)\,v, \tag{3.7}$$

with $C_p(v)$ as defined in (3.2).

Now, $C_p(v) \neq 0$ for $0 < |v| < v_p$ if v_p is the smallest positive zero of C_p. Also, the sine-product in (3.7) is certainly nonvanishing for $0 < |v| < 2\pi/(2p-1)$. Therefore, $\det B(v)$ is nonvanishing for

$$0 < |v| < \min\left(v_p, \frac{2\pi}{2p-1}\right),$$

which proves our theorem.

For reference we list the cosine-polynomials $C_p(v)$ for $p = 1, 2, 3$:

$$C_1(v) = \cos\frac{v}{2},$$

$$C_2(v) = 1 + \cos v + \cos 2v,$$

$$C_3(v) = 3\cos\frac{v}{2} + 3\cos 3\frac{v}{2} + 2\cos 5\frac{v}{2} + \cos 7\frac{v}{2} + \cos 9\frac{v}{2}.$$

One finds easily that

$$v_1 = \pi, \qquad v_2 = v_3 = \frac{\pi}{2}$$

so that the bounds in (3.1) for $p=1, 2, 3$ are respectively π, $\pi/2$, $2\pi/5$

We also note from (3.2) that

$$0 < |v| < \frac{\pi}{p^2} \tag{3.8}$$

is a sufficient condition for nonvanishing of $\det B(v)$.

4. Trigonometric extrapolation and interpolation methods of Adams' type

Multi-step methods with

$$\alpha_0 = -\alpha_1 = 1, \qquad \alpha_\lambda = 0 \qquad (\lambda > 1)$$

and maximal algebraic order for fixed index are called Adams methods. In this section we list methods of trigonometric order that correspond to Adams' extrapolation and interpolation methods in the sense of Theorem 2. The coefficients $\beta_\lambda(v)$ and $\beta_\lambda^*(v)$ are obtained as the power series solution of the appropriate system of equations (2.8) where coefficient matrix and known vector are expanded into their Taylor series.

Adams extrapolation methods of trigonometric order p

$$x_{n+1} = x_n + h\sum_{\lambda=1}^{2p}\beta_{p\lambda}(v)\, x'_{n+1-\lambda} \qquad (v = 2\pi h/T)$$

$$\beta_{11} = \frac{3}{2}\left(1 - \frac{1}{4}v^2 + \frac{1}{120}v^4 + \cdots\right), \qquad \beta_{12} = -\frac{1}{2}\left(1 + \frac{1}{12}v^2 + \frac{1}{120}v^4 + \cdots\right);$$

$$\beta_{21} = \frac{55}{24}\left(1 - \frac{95}{132}v^2 + \frac{79}{792}v^4 + \cdots\right), \qquad \beta_{22} = -\frac{59}{24}\left(1 - \frac{923}{708}v^2 + \frac{15647}{21240}v^4 + \cdots\right),$$

$$\beta_{23} = \frac{37}{24}\left(1 - \frac{421}{444}v^2 + \frac{1921}{13320}v^4 + \cdots\right), \quad \beta_{24} = -\frac{9}{24}\left(1 + \frac{1}{4}v^2 + \frac{11}{120}v^4 + \cdots\right);$$

$$\beta_{31} = \frac{4277}{1440}\left(1 - \frac{5257}{3666}v^2 + \frac{196147}{439920}v^4 + \cdots\right),$$

$$\beta_{32} = -\frac{7923}{1440}\left(1 - \frac{48607}{15846}v^2 + \frac{2341619}{633840}v^4 + \cdots\right),$$

$$\beta_{33} = \frac{9982}{1440}\left(1 - \frac{107647}{29946}v^2 + \frac{2791381}{513360}v^4 + \cdots\right),$$

$$\beta_{34} = -\frac{7298}{1440}\left(1 - \frac{69473}{21894}v^2 + \frac{10276973}{2627280}v^4 + \cdots\right),$$

$$\beta_{35} = \frac{2877}{1440}\left(1 - \frac{10433}{5754}v^2 + \frac{20683}{32880}v^4 + \cdots\right),$$

$$\beta_{36} = -\frac{475}{1440}\left(1 + \frac{55}{114}v^2 + \frac{1015}{2736}v^4 + \cdots\right);$$

. . . . . .

Adams interpolation methods of trigonometric order p

$$x_{n+1} = x_n + h \sum_{\lambda=0}^{2p-1} \beta^*_{p\lambda}(v)\, x'_{n+1-\lambda} \qquad (v = 2\pi h/T)$$

$$\beta^*_{10} = \beta^*_{11} = \frac{1}{2}\left(1 + \frac{1}{12} v^2 + \frac{1}{120} v^4 + \cdots\right);$$

$$\beta^*_{20} = \frac{9}{24}\left(1 + \frac{1}{4} v^2 + \frac{11}{120} v^4 + \cdots\right), \qquad \beta^*_{21} = \frac{19}{24}\left(1 - \frac{43}{228} v^2 + \frac{13}{360} v^4 + \cdots\right),$$

$$\beta^*_{22} = -\frac{5}{24}\left(1 - \frac{1}{12} v^2 - \frac{7}{72} v^4 + \cdots\right), \qquad \beta^*_{23} = \frac{1}{24}\left(1 + \frac{11}{12} v^2 + \frac{193}{360} v^4 + \cdots\right);$$

$$\beta^*_{30} = \frac{475}{1440}\left(1 + \frac{55}{114} v^2 + \frac{500267}{22800} v^4 + \cdots\right),$$

$$\beta^*_{31} = \frac{1427}{1440}\left(1 - \frac{5149}{8562} v^2 - \frac{15139837}{342480} v^4 + \cdots\right),$$

$$\beta^*_{32} = -\frac{798}{1440}\left(1 - \frac{163}{114} v^2 - \frac{1964441}{10640} v^4 + \cdots\right),$$

$$\beta^*_{33} = \frac{482}{1440}\left(1 - \frac{1697}{1446} v^2 - \frac{20178851}{57840} v^4 + \cdots\right),$$

$$\beta^*_{34} = -\frac{173}{1440}\left(1 + \frac{29}{1038} v^2 - \frac{22688263}{41520} v^4 + \cdots\right),$$

$$\beta^*_{35} = \frac{27}{1440}\left(1 + \frac{13}{6} v^2 - \frac{187111}{240} v^4 + \cdots\right);$$

. . . . . .

As shown in Section 3 the series for $\beta_{p\lambda}$ and $\beta^*_{p\lambda}$ certainly converge for $|v| < r_p$ where $r_1 = \pi$, $r_2 = \pi/2$, $r_3 = 2\pi/5$.

We also note the explicit formulae

$$\beta_{11} = \frac{\sin \frac{3}{2} v}{v \cos \frac{1}{2} v}, \qquad -\beta_{12} = \beta^*_{10} = \frac{\tan \frac{1}{2} v}{v}.$$

5. Trigonometric extrapolation and interpolation methods of Störmer's type

Linear multi-step methods are also used in connection with differential equations of higher order, in particular with second order differential equations in which the first derivative is absent,

$$x'' = f(t, x), \qquad x(t_0) = x_0, \qquad x'(t_0) = x'_0. \tag{5.1}$$

They take here the form

$$x_{n+1} + \alpha_1 x_n + \cdots + \alpha_k x_{n+1-k} = h^2(\beta_0 x''_{n+1} + \beta_1 x''_n + \cdots + \beta_k x''_{n+1-k}), \tag{5.2}$$
$$x''_m = f(t_0 + m h, x_m).$$

The terminology introduced in Section 2 extends in an obvious manner to this new situation. With the multi-step method (5.2) there is now associated the functional

$$L x = \sum_{\lambda=0}^{k} [\alpha_\lambda x(t_0 + (n + 1 - \lambda) h) - h^2 \beta_\lambda x''(t_0 + (n + 1 - \lambda) h)] \qquad (\alpha_0 = 1).$$

Theorem 1 (with the modification mentioned in Remark 1 on p. 384) can then be applied to this functional provided that not all the values of α_λ are fixed in advance. Otherwise our assumption (1.5′) would not hold. Except for this provision, however, the construction of multi-step methods (5.2) of trigonometric order follows the same pattern as outlined in Sections 2 and 4 for first order differential equations.

We content ourselves in this section with listing a few methods that result if one takes

$$\alpha_\lambda = 0 \quad \text{for} \quad \lambda > 2. \tag{5.3}$$

In the algebraic case such methods of maximal order (for given index k) are called Störmer methods (cf., e.g., [3, p. 125]).

Störmer extrapolation methods of trigonometric order p

$$x_{n+1} + \alpha_{p1}(v)\, x_n + \alpha_{p2}(v)\, x_{n-1} = h^2 \sum_{\lambda=1}^{2p-1} \beta_{p\lambda}(v)\, x''_{n+1-\lambda} \qquad (v = 2\pi h/T)$$

$$\alpha_{11} = -2, \qquad \alpha_{12} = 1, \qquad \beta_{11} = 1 - \frac{1}{12}v^2 + \frac{1}{360}v^4 + \cdots;$$

$$\alpha_{21} = -2\left(1 - \frac{1}{6}v^4 + \frac{1}{36}v^6 + \cdots\right), \qquad \alpha_{22} = -\alpha_{21} - 1,$$

$$\beta_{21} = \frac{13}{12}\left(1 - \frac{19}{52}v^2 + \frac{7}{120}v^4 + \cdots\right), \qquad \beta_{22} = -\frac{2}{12}\left(1 - \frac{9}{4}v^2 + \frac{37}{120}v^4 + \cdots\right),$$

$$\beta_{23} = \frac{1}{12}\left(1 + \frac{1}{4}v^2 + \frac{7}{120}v^4 + \cdots\right);$$

$$\alpha_{31} = -2\left(1 - \frac{27}{20}v^6 + \cdots\right), \qquad \alpha_{32} = -\alpha_{31} - 1,$$

$$\beta_{31} = \frac{299}{240}\left(1 - \frac{4315}{5382}v^2 + \frac{7357}{49680}v^4 + \cdots\right),$$

$$\beta_{32} = -\frac{176}{240}\left(1 - \frac{3181}{792}v^2 + \frac{264593}{47520}v^4 + \cdots\right),$$

$$\beta_{33} = \frac{194}{240}\left(1 - \frac{2047}{582}v^2 + \frac{38129}{7760}v^4 + \cdots\right),$$

$$\beta_{34} = -\frac{96}{240}\left(1 - \frac{913}{432}v^2 + \frac{6923}{25920}v^4 + \cdots\right),$$

$$\beta_{35} = \frac{19}{240}\left(1 + \frac{221}{342}v^2 + \frac{17521}{41040}v^4 + \cdots\right);$$

. . . . . .

Störmer interpolation methods of trigonometric order p

$$x_{n+1} + \alpha^*_{p1}(v)\, x_n + \alpha^*_{p2}(v)\, x_{n-1} = h^2 \sum_{\lambda=0}^{2p-2} \beta^*_{p\lambda}\, x''_{n+1-\lambda} \qquad (v = 2\pi h/T)$$

$$\alpha^*_{11} = -2\left(1 + \frac{1}{2}v^2 + \frac{11}{24}v^4 + \frac{301}{720}v^6 + \cdots\right),$$

$$\alpha^*_{12} = -\alpha^*_{11} - 1, \quad \beta^*_{10} = 1 + \frac{11}{12}v^2 + \frac{301}{360}v^4 + \cdots;$$

$$\alpha_{21}^* = -2, \quad \alpha_{22}^* = 1, \quad \beta_{20}^* = \frac{1}{12}\left(1 + \frac{1}{4}v^2 + \frac{7}{120}v^4 + \cdots\right),$$

$$\beta_{21}^* = \frac{10}{12}\left(1 - \frac{1}{20}v^2 + \frac{1}{120}v^4 + \cdots\right), \quad \beta_{22}^* = \frac{1}{12}\left(1 + \frac{1}{4}v^2 + \frac{7}{120}v^4 + \cdots\right);$$

$$\alpha_{31}^* = -2\left(1 + \frac{3}{40}v^6 + \cdots\right), \qquad \alpha_{32}^* = -\alpha_{31}^* - 1,$$

$$\beta_{30}^* = \frac{19}{240}\left(1 + \frac{221}{342}v^2 + \frac{17\,521}{41\,040}v^4 + \cdots\right),$$

$$\beta_{31}^* = \frac{204}{240}\left(1 - \frac{79}{459}v^2 + \frac{11\,039}{110\,160}v^4 + \cdots\right),$$

$$\beta_{32}^* = \frac{14}{240}\left(1 + \frac{95}{42}v^2 - \frac{103}{80}v^4 + \cdots\right), \quad \beta_{33}^* = \frac{4}{240}\left(1 - \frac{16}{9}v^2 - \frac{4711}{2160}v^4 + \cdots\right),$$

$$\beta_{34}^* = -\frac{1}{240}\left(1 + \frac{31}{18}v^2 + \frac{3899}{2160}v^4 + \cdots\right);$$

. . . . . .

The series for $\alpha_{p\lambda}, \beta_{p\lambda}$ converge if $|v| < r_p$ where $r_1 = \infty$, $r_2 = \pi/2$, those for $\alpha_{p\lambda}^*, \beta_{p\lambda}^*$ converge if $|v| < r_p^*$ where $r_1^* = \pi/3$, $r_2^* = \pi/2$. This can be shown by reasonings similar to, but more complicated than, those in Section 3. The values of r_3, r_3^* were not obtained because of the complexity of the calculations required.

We also note the explicit formulae

$$\beta_{11} = \left(\frac{2\sin\frac{1}{2}v}{v}\right)^2, \quad \alpha_{11}^* = -\frac{2\cos v}{2\cos v - 1}, \quad \beta_{10}^* = \frac{2(1-\cos v)}{v^2(2\cos v - 1)}.$$

6. Effect of uncertainty in the choice of T

Multi-step methods of trigonometric order presuppose the knowledge of the period T of the solution, if it is periodic, or of a suitable substitute, if the solution is only oscillatory. Precise knowledge of this kind is usually not available in advance, so that one has to rely on suitable estimates of T. Since T enters only through the parameter $v = 2\pi h/T$ and $T = \infty$ gives the classical multi-step methods, one expects that uncertainties in the value of T should not seriously impair the effectiveness of trigonometric multi-step methods (when applicable) as long as T is not significantly underestimated.

It is instructive to study from this point of view the simple initial value problem

$$\frac{dx}{dt} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} x, \quad x(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{6.1}$$

which has the solution

$$x(t) = \begin{pmatrix} \cos t \\ \sin t \end{pmatrix}.$$

Every multi-step method of trigonometric order $\geqq 1$ relative to period 2π is exact in this case, so that the example allows us to isolate the effect of inaccurately estimating the period.

Let us select Adams' interpolation method of trigonometric order 1, which can be written in the form

$$x_{n+1} = x_n + h\frac{\tan\frac{1}{2}v}{v}(x'_{n+1} + x'_n) \qquad (v = 2\pi h/T). \tag{6.2}$$

The correct choice of T is 2π, giving $v=h$. We consider now T to be some "estimate" of 2π and use

$$\lambda = \frac{2\pi}{T}$$

to measure the quality of the estimate (underestimation, if $\lambda>1$, overestimation, if $\lambda<1$, precise estimate, if $\lambda=1$).

Letting

$$\tau = \frac{1}{\lambda}\tan\frac{\lambda h}{2},$$

application of (6.2) to (6.1) then gives

$$x_{n+1} = x_n + \tau\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}(x_{n+1} + x_n),$$

or else, collecting terms,

$$\begin{pmatrix}1 & \tau\\ -\tau & 1\end{pmatrix}x_{n+1} = \begin{pmatrix}1 & -\tau\\ \tau & 1\end{pmatrix}x_n, \qquad x_{n+1} = \frac{1}{1+\tau^2}\begin{pmatrix}1-\tau^2 & -2\tau\\ 2\tau & 1-\tau^2\end{pmatrix}x_n.$$

If we set

$$\tau = \tan\tfrac{1}{2}\vartheta,$$

we get

$$x_{n+1} = \begin{pmatrix}\cos\vartheta & -\sin\vartheta\\ \sin\vartheta & \cos\vartheta\end{pmatrix}x_n.$$

Obviously,

$$\vartheta = 2\arctan\left(\frac{1}{\lambda}\tan\frac{\lambda h}{2}\right). \tag{6.3}$$

The n-th approximation x_n to the solution of (6.1) is thus obtained by rotating the initial vector $x_0 = \begin{pmatrix}1\\0\end{pmatrix}$ n-times through the angle ϑ, where ϑ is given by (6.3). Therefore

$$x_n = \begin{pmatrix}\cos n\vartheta\\ \sin n\vartheta\end{pmatrix},$$

which shows that the approximations have the correct amplitude, but phase errors

$$\varepsilon_n = n(\vartheta - h) = nh\left\{\frac{2}{h}\arctan\left(\frac{1}{\lambda}\tan\frac{\lambda h}{2}\right) - 1\right\}. \tag{6.4}$$

If $\lambda=1$ then $\varepsilon_n=0$, as we expect. In the limit as $\lambda\to 0$ we obtain the phase error of the method of algebraic order 1, which in our example is the trapezoidal rule. The expression in curled brackets, as function of λ, has a behavior as shown in Figure 1. It is seen from this, in particular, that the error in absolute value

is less than the error at $\lambda=0$ for all λ with $0<\lambda<\lambda_0$ where $\lambda_0>1$. This means that in using the modified trapezoidal rule (6.2) we may overestimate the period as much as we wish, and even underestimate it somewhat, and still get better results than with the ordinary trapezoidal rule. On the other hand, the curve in Figure 1 also shows that the error reduction is not very substantial unless λ is close to 1. If $h=.1$, for example, there is a gain of at least one decimal digit only if the estimated period differs from the true period by 5% or less.

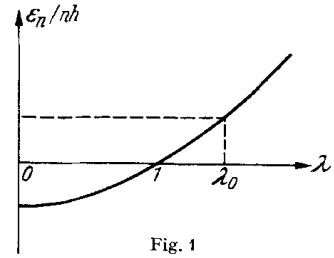


Fig. 1

7. Numerical examples

An important class of differential equations to which trigonometric multi-step methods may advantageously be applied is given by equations of the form

$$x''+P(t)\,x=0, \tag{7.1}$$

where $P(t)$ is a nearly constant nonnegative function,

$$P(t)=P_0[1+p(t)]\geqq 0 \qquad (t\geqq t_0). \tag{7.2}$$

Here, P_0 is a positive constant and $p(t)$ a function which is "small" in some sense for $t\geqq t_0$.

Equation (7.1) may be considered a perturbation of $x''+P_0x=0$, the differential equation of a harmonic oscillator with angular frequency $\sqrt{P_0}$. This suggests the following values of T (and thus of v) as natural choices in methods of trigonometric order,

$$T=2\pi/\sqrt{P_0}, \qquad v=h\sqrt{P_0}. \tag{7.3}$$

If one is willing to select these values anew at each step of integration, one can improve upon (7.3) by using

$$T=T_n=2\pi/\sqrt{P(t_n)}, \qquad v=v_n=h\sqrt{P(t_n)} \tag{7.4}$$

in the computation of x_{n+1}.

Particularly favorable results are expected if t_0 is relatively large and $p(t)$ such that

$$\int^{\infty}|p(t)|\,dt<\infty, \tag{7.5}$$

in which case it is known that $x=c_1\cos\sqrt{P_0}t+c_2\sin\sqrt{P_0}t+o(1)$ (c_1, c_2 constants, $t\to\infty$) for every solution of (7.1). Our first example belongs to this type.

Example 1. $x''+\left(100+\frac{1}{4t^2}\right)x=0, \qquad 0<t_0\leqq t\leqq 10.$

The general solution can be expressed in terms of Bessel functions, $x=c_1\sqrt{t}J_0(10t)+c_2\sqrt{t}Y_0(10t)$. We single out the particular solution $\sqrt{t}\,J_0(10t)$ by choosing the initial values accordingly. Table 1 below shows selected results (every 50th value, using $t_0=1$, $h=.02$) obtained by the Störmer extrapolation methods of algebraic order 2 and 4, and of trigonometric order 1 and 2, in this

order[3]. In the latter two methods the constant value (7.3) of T was used, that is, $T=\pi/5$, $v=.2$.

Table 1 reveals an average increase in accuracy of about three decimal digits in favor of the trigonometric extrapolation methods. This — it should be noted — is at practically no extra cost in computation, since the modified coefficients of the trigonometric methods, if (7.3) is used, need only be computed once, at

Table 1. *Störmer extrapolation method of various algebraic and trigonometric orders. Example 1 with* $t_0=1$

t	alg. ord. $p=2$	alg. ord. $p=4$	trig. ord. $p=1$	trig. ord. $p=2$	exact 7D values
1	−.2459358	−.2459358	−.2459358	−.2459358	−.2459358
2	.2345901	.2354337	.2362055	.2362115	.2362085
3	−.1425368	−.1485247	−.1495871	−.1495966	−.1495937
4	.0018875	.0143880	.0147257	.0147349	.0147338
5	.1393247	.1234167	.1248068	.1248015	.1248002
6	−.2330076	−.2205650	−.2240619	−.2240630	−.2240592
7	.2472935	.2461304	.2511024	.2511101	.2511049
8	−.1773539	−.1924022	−.1972536	−.1972659	−.1972606
9	.0470268	.0771940	.0798806	.0798938	.0798900
10	.0993055	.0620548	.0632097	.0631997	.0632007

the beginning of the computations. If the choice (7.4) is made an additional $\frac{3}{4}$ decimal digit is gained on the average, the amount of computing being somewhat larger than before.

Störmer interpolation methods of algebraic order 2 and of trigonometric order 1, applied to Example 1, gave results which are 10—20 times worse than the corresponding results in Table 1, the trigonometric method being, on the average, more accurate by $2\frac{1}{2}$ decimal digits. The interpolation method of algebraic order 4, however, is almost 100 times better than the corresponding extrapolation method. Nevertheless there is also here an improvement of about $1\frac{1}{2}$ decimal digits in favor of the trigonometric modification.

Larger values of t_0 would put trigonometric methods into an even more favorable light. As t_0 decreases from 1 to 0, trigonometric methods gradually lose their superiority.

In our next example — a Mathieu differential equation — the relation (7.5) is not satisfied any more.

Example 2. $x''+100(1-\alpha\cos 2t)\,x=0$, $t_0=0$, $x_0=1$, $x_0'=0$ $(0<\alpha\leqq 1)$. We integrated this equation for various values of α using the same methods and the same step length $h=.02$ as in Example 1. An independent calculation was done with the help of Nyström's method, which was also used to obtain starting values. Selected results (every 25th value) of the Störmer extrapolation methods, in the case $\alpha=.1$, are displayed in Table 2[3]. Trigonometric order, also in this example, is to be understood relative to period $T=\pi/5$.

[3] Calculations were done on ORACLE in 32 binary bit floating point arithmetic (the equivalent of about 9 significant decimal digits). The final results were rounded to 7 decimal places. — The author takes the opportunity to acknowledge the able assistance of Miss RUTH BENSON in performing these calculations.

The results in Table 2 follow a similar pattern as those above in Table 1, the main difference being a reduction, to roughly half the size, of the improvement of trigonometric methods over the algebraic ones. The average gain in accuracy is now about $1\frac{1}{2}$ decimal digits. The remarks made above on interpolation methods hold true also in Example 2, except for the reduction just mentioned. Obviously, as α decreases to 0, trigonometric methods become increasingly

Table 2. *Störmer extrapolation method of various algebraic and trigonometric orders. Example 2 with* $\alpha = .1$

t	alg. ord. $p=2$	alg. ord. $p=4$	trig. ord. $p=1$	trig. ord. $p=2$	exact 7 D values
0	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000
0.5	.0767165	.0690295	.0685134	.0691273	.0692085
1.0	−.9035098	−.9056448	−.9089870	−.9080120	−.9084179
1.5	−.7105151	−.6908656	−.6942472	−.6938453	−.6939608
2.0	.1985482	.2287643	.2304036	.2311394	.2309590
2.5	.9715966	.9679083	.9764633	.9767822	.9763699
3.0	.2552862	.2045198	.2060842	.2056667	.2057667
3.5	−.9456869	−.9505080	−.9618456	−.9613337	−.9616794
4.0	−.4833155	−.4221211	−.4260400	−.4262622	−.4265317
4.5	.5453242	.5922666	.6026736	.6021053	.6022367
5.0	.9517667	.9263164	.9422702	.9418659	.9417373

superior to algebraic methods. We have experienced only a slight decrease in this superiority when we let α increase from .1 to 1.

It is anticipated that trigonometric methods can be applied, with similar success, also to nonlinear differential equations describing oscillation phenomena.

References

[1] Antosiewicz, H. A., and W. Gautschi: Numerical methods in ordinary differential equations, Chap. 9 of "Survey of numerical analysis" (ed. J. Todd). New York-Toronto-London: McGraw-Hill Book Co. (in press).

[2] Brock, P., and F. J. Murray: The use of exponential sums in step by step integration. Math. Tables Aids Comput. **6**, 63–78 (1952).

[3] Collatz, L.: The numerical treatment of differential equations, 3rd ed. Berlin-Göttingen-Heidelberg: Springer 1960.

[4] Dennis, S. C. R.: The numerical integration of ordinary differential equations possessing exponential type solutions. Proc. Cambridge Philos. Soc. **56**, 240–246 (1960).

[5] Urabe, M., and S. Mise: A method of numerical integration of analytic differential equations. J. Sci. Hiroshima Univ., Ser. A **19**, 307–320 (1955).

Oak Ridge National Laboratory
Mathematics Panel
P. O. Box X
Oak Ridge, Tennessee

(Received September 18, 1961)

27.2. [54] “Global error estimates in “one-step” methods for ordinary differential equations”

---

[54] “Global error estimates in “one-step” methods for ordinary differential equations,” *Rend. Mat*. (2) **8**, 601–617 (1975) (translated from Italian).

Global error estimates in "one-step" methods for ordinary differential equations* **

Walter Gautschi

Dedicated to Professor Mauro Picone on the occasion of his ninetieth birthday

Abstract. We consider one-step methods for the numerical solution of ordinary differential equations and propose to utilize recent progress in local error estimation in order to asymptotically estimate the global error.

1. Introduction

The majority of numerical methods for the solution of systems of ordinary differential equations generates approximations of the solution vector corresponding to a finite sequence of points. By *global error* one usually understands the difference between the vector of approximation and the solution vector at the respective points. The *local error*, on the other hand, is the difference between the approximate solution and the exact one, after a single step of the method initiated with exact data. It is generally agreed that "one-step" methods, in particular methods of Runge–Kutta type, notoriously do not permit an easy and efficient estimate of the local error, not to speak of the global one. The situation, in recent years, has changed a bit after rather efficient schemes have become known for accurately estimating the local error (at least asymptotically for small steps). It is natural, then, to attempt incorporating these schemes in procedures for the estimation of the global error. This is the subject of our work.

* This work has been sponsored in part by the National Science Foundation, research grant GP-36557

** English translation by Walter Gautschi of "Stime dell'errore globale nei metodi "one-step" per equazioni differenziali ordinarie", *Rend. Mat.* (2) **8** (1975), 601–617.

The desired estimates (as, in principle, has been known for some time) can be obtained by integrating the variational differential equation satisfied by the principal part of the global error. This approach requires computing the Jacobian matrix of the differential system evaluated along the solution trajectory and therefore, in practice, may limit the applicability of the procedure to problems of small or medium dimensions. Nevertheless, the occurrence of the Jacobian matrix is quite natural in view of the well-known role it plays in the theory of perturbation. (For procedures not using the Jacobian matrix, see [20]).

In Sections 2–7 we recall some basic concepts for "one-step" methods, including also their properties of stability and convergence ([11], [10], [21]). The implementation, and the theoretical justification, of the procedure for estimating the global error is presented in Sections 8–9. Section 10, finally, contains a numerical example.

2. The differential system

We consider the Cauchy problem

$$dy/dx = f(x,y), \quad a \le x \le b, \quad y(a) = y_a, \tag{2.1}$$

for a system of m ordinary first-order differential equations. We assume f to be defined, and sufficiently regular, in the rectangular domain

$$\mathcal{R}_0 = [a,b] \times \mathcal{D}_0, \quad \mathcal{D}_0 = \{y \in \mathbb{R}^m : \ c_i \le y^i \le d_i, \ i = 1,2,\dots,m\},$$

where y^i denotes the ith component of y. We consider $\mathcal{R}_0$ the *fundamental domain* which is to include not only the exact solution, but also all approximations generated. Later, for various reasons, we will have to enlarge somewhat the domain in which f is defined.

Meanwhile, *we assume, once and for all, that* $y_a \in \mathcal{D}_0$, *and that* (2.1) *has a unique solution* $y(x)$ *on* $[a,b]$ *such that* $y(x) \in \mathcal{D}_0$ *for* $a \le x \le b$.

3. "One-step" methods

A "one-step" method for the calculation of an approximate solution of (2.1) can be identified by a function

$$\Phi : \quad [a,b] \times \mathcal{D}_0 \times [0,h_0] \to \mathbb{R}^m, \tag{3.1}$$

which in some way is connected with the function f in (2.1). By means of

$$y_h = y + h\Phi(x,y;h), \quad 0 < h \le h_0, \tag{3.2}$$

it indicates how to proceed from a generic point (x,y) to the "next" point $(x+h, y_h)$, just as f indicates how to proceed from (x,y) to $(x+dx, y+fdx)$.

In order to obtain a sequence $u_n \approx y(x_n)$ of approximations to the solution of (2.1), the formula (3.2) is used in the following manner:

$$u_{n+1} = u_n + h_{n+1}\Phi(x_n, u_n; h_{n+1}), \quad n = 0, 1, \ldots, N-1, \quad u_0 = y_a, \tag{3.3}$$

where $x_n = a + h_1 + h_2 + \cdots + h_n$, and $x_N = b$. The choice of the "steps" $h_1, h_2, \ldots, h_N$ is part of the *steering mechanism* for (3.3), which, normally, is designed with the intention of keeping the norm of the error, $\|u_n - y(x_n)\|$, sufficiently small. More generally, the steering mechanism may also involve the choice of "one-step" methods varying from step to step.

As indicated in (3.1), we want Φ to be defined in all of $\mathcal{R}_0 \times [0, h_0]$. For some methods this assumption requires that the domain of definition of f be slightly enlarged. For example, if Φ represents the midpoint rule,

$$\Phi(x, y; h) = f(x + \tfrac{1}{2}h, y + \tfrac{1}{2}hf(x, y)),$$

the interval $[a, b]$ should be enlarged to the right by the quantity $\frac{1}{2}h_0$, whereas the sides of $\mathcal{D}_0$ should be extended from both extremes by the quantity $\frac{1}{2}h_0M_0$, where $M_0 = \max_{\mathcal{R}_0} \|f(x, y)\|$.

We assume, once and for all, that $0 < h_{n+1} \leq h_0$ *and* $u_n \in \mathcal{D}_0$ *for each* $n = 0, 1, \ldots, N-1$.

4. Local description of "one-step" methods

There are a few concepts that describe local properties of a method Φ. We begin with the one of truncation error (or "local error").

Given a generic point $(x, y) \in \mathcal{R}_0$, we construct a solution tract of (2.1) emanating therefrom,

$$du/dt = f(t, u), \quad x \leq t \leq x + h_0, \quad u(x) = y. \tag{4.1}$$

We call $u(t)$, $x \leq t \leq x + h_0$, the *reference solution* at the point (x, y), and denote it, if necessary, more completely by $u(t; x, y)$. We assume that $u(t; x, y)$, $x \leq t \leq x + h_0$, is defined for all points $(x, y) \in \mathcal{R}_0$; once again, this assumption requires a slight extension of the domain in which f is defined.

DEFINITION 4.1. For arbitrary $(x, y) \in \mathcal{R}_0$ and $h \in (0, h_0]$, the *truncation error* of Φ at the point (x, y) is defined by

$$t(x, y; h) = h^{-1}[y_h - u(x + h; x, y)]. \tag{4.2}$$

By (3.2), therefore,

$$t(x, y; h) = \Phi(x, y; h) - h^{-1}[u(x + h) - u(x)]. \tag{4.2'}$$

DEFINITION 4.2. The method Φ is called *consistent* if

$$t(x,y;h) \to 0 \quad \text{as } h \to 0,$$

uniformly for $(x,y) \in \mathcal{R}_0$.

By (4.2′) and (4.1), if Φ is consistent, then necessarily

$$\Phi(x,y;0) = f(x,y). \tag{4.3}$$

DEFINITION 4.3. The method Φ is said to have *order* p, if there exists a constant $C > 0$ not depending on x, y and h such that

$$\|t(x,y;h)\| \leq Ch^p \quad \text{for each } (x,y) \in \mathcal{R}_0, \ h \in [0,h_0]. \tag{4.4}$$

Property (4.4) will be expressed more briefly in the form

$$t(x,y;h) = O(h^p), \quad h \to 0. \tag{4.4′}$$

Normally, p is an integer. (See, however, [5]). We call p the *exact order* of Φ if (4.4) does not hold for any larger p. Evidently, $p > 0$ implies consistency of Φ.

DEFINITION 4.4. A function $\tau(x,y)$ on $\mathcal{R}_0$ for which $\tau(x,y) \not\equiv 0$ and

$$t(x,y;h) = \tau(x,y)h^p + O(h^{p+1}), \quad h \to 0, \tag{4.5}$$

is called *principal error function* of the method Φ.

Since $\tau \not\equiv 0$, p in (4.5) is the exact order of Φ.

5. Global description of "one-step" methods

We now examine the global behavior of algorithm (3.3). The set of points

$$\{x_n\}_{n=0}^N, \quad x_n = a + h_1 + h_2 + \cdots + h_n, \ x_N = b$$

will be called a *grid* on the interval $[a,b]$, and we will denote it by $m_h[a,b]$, where h stands for the collection of lengths $h_1, h_2, \ldots, h_N$. The *fineness* of the grid $m_h[a,b]$ is defined by

$$|h| = \max_{1 \leq n \leq N} h_n.$$

A (vector-valued) function defined on the grid $m_h[a,b]$ is called a *grid function*. Any function $y(x)$ defined on $[a,b]$ induces a grid function by restriction.

With the algorithm (3.3) we associate an operator D_h defined by

$$(D_h u)_n = h_{n+1}^{-1}(u_{n+1} - u_n) - \Phi(x_n, u_n; h_{n+1}), \quad n = 0, 1, \ldots, N-1. \tag{5.1}$$

D_h acts on grid functions (with $u_n \in \mathcal{D}_0$) and generates a new grid function defined on the whole grid except the final point x_N. Note that for the exact solution $y(x)$ of (2.1), by virtue of (4.2′),

$$(D_h y)_n = -t(x_n, y(x_n); h_{n+1}). \tag{5.2}$$

DEFINITION 5.1. The method Φ is called *stable* on $[a, b]$ if for any grid $m_h[a, b]$, with $|h|$ arbitrarily small, and for arbitrary grid functions v, w (with $v_n \in \mathcal{D}_0$, $w_n \in \mathcal{D}_0$), there exists a constant $K > 0$ not depending on n and h such that

$$\max_{0 \le n \le N} \|v_n - w_n\| \le K(\|v_0 - w_0\| + \max_{0 \le n \le N-1} \|(D_h v)_n - (D_h w)_n\|). \tag{5.3}$$

We refer to (5.3) as the *stability inequality.* In order to motivate Definition 5.1, let u, w be grid functions for which

$$(D_h u)_n = 0, \quad 0 \le n \le N-1, \quad u_0 = y_a,$$

$$(D_h w)_n = \varepsilon_n, \quad 0 \le n \le N-1, \quad w_0 = y_a + \varepsilon,$$

where ε_n, ε are "small" vectors. We may interpret u as the result of applying algorithm (3.3) in infinite precision, and w the result of applying it in finite precision. The residual vectors ε_n and ε may reflect the presence of rounding errors. Stability, then, implies that

$$\max_{0 \le n \le N} \|u_n - w_n\| \le K(\|\varepsilon\| + \max_{0 \le n \le N-1} \|\varepsilon_n\|),$$

that is, the error of the finite-precision result is of the same order of magnitude as the rounding errors, for any grid, no matter how fine.

It is remarkable that essentially all "one-step" methods are stable.

THEOREM 5.1. *If $\Phi(x, y; h)$ satisfies a Lipschitz condition with respect to y, uniformly on $[a, b] \times \mathcal{D}_0 \times [0, h_0]$, that is,*

$$\|\Phi(x, y; h) - \Phi(x, y^*; h)\| \le M\|y - y^*\|, \quad \text{for each } x \in [a, b],\ y, y^* \in \mathcal{D}_0,\ h \in [0, h_0], \tag{5.4}$$

then the method Φ is stable.

For the proof one takes any two grid functions v, w and verifies that

$$e_n \le (1 + h_n M)e_{n-1} + h_n d, \quad n = 1, 2, \ldots, N,$$

where

$$e_n = \|v_n - w_n\|, \quad d = \max_{0\le n\le N-1} \|(D_h\, v)_n - (D_h\, w)_n\|.$$

It then easily follows that

$$e_n \le e^{(b-a)M} e_0 + e^{(b-a)M} \sum_{k=1}^{n} h_k d \le e^{(b-a)M}\{e_0 + (b-a)d\},$$
$$n = 0, 1, 2, \ldots, N,$$

that is,

$$\max_{0\le n\le N} e_n \le e^{(b-a)M}\{e_0 + (b-a)d\},$$

which is the stability inequality (5.3) with $K = e^{(b-a)M} \max(1, b-a)$.

Theorem 5.1 remains valid for variable-methods algorithms involving a family of "one-step" methods $\{\Phi_n\}$ if each satisfies a Lipschitz condition with constant M not depending on n.

It is useful to note that Φ need not necessarily be continuous in x.

COROLLARY. *Let $m_h[a,b]$ be an arbitrary grid on $[a,b]$ and let A_n, b_n be two grid functions on $m_h[a,b]$, the former matrix-valued, the latter vector-valued, such that*

$$\|A_n\| \le \alpha, \quad \|b_n\| \le \beta \quad \text{for } n = 0, 1, \ldots, N-1, \tag{5.5}$$

where α, β do not depend on n and h. Given any (vector-valued) grid function u on $m_h[a,b]$ satisfying

$$u_{n+1} = u_n + h_{n+1}(A_n u_n + b_n), \quad n = 0, 1, \ldots, N-1, \tag{5.6}$$

there exists a constant $\gamma > 0$ not depending on n and h, and depending only on α, β, and u_0, such that

$$\|u_n\| \le \gamma, \quad n = 0, 1, \ldots, N. \tag{5.7}$$

The corollary follows by letting $A_n = A(x_n)$, $b_n = b(x_n)$ for certain bounded functions $A(x)$, $b(x)$, and by observing that

$$\Phi(x, y; h) = A(x)y + b(x)$$

satisfies a Lipschitz condition (5.4) on $\mathcal{R}_0 = [a,b] \times \mathbb{R}^m$ with constant $M = \alpha$. Taking $v_n = u_n$, $w_n = 0$ in the stability inequality (5.3), we obtain the desired bound (5.7) with $\gamma = K(\|u_0\| + \beta)$. The constant K depends on α.

6. Convergence of "one-step" methods

DEFINITION 6.1. The method Φ is said to be *convergent* on $[a,b]$ if for arbitrary x with $a < x \leq b$ one has

$$\max_{0\leq n\leq N} \|u_n - y(x_n)\| \to 0 \quad \text{as } |h| \to 0, \tag{6.1}$$

where $a = x_0 < x_1 < \cdots < x_N = x$ is a grid on $[a,x]$ with fineness $|h| = \max_{1\leq n\leq N}(x_n - x_{n-1})$, $\{u_n\}$ are the approximation vectors generated on this grid by algorithm (3.3), and $y(x_n)$ is the exact solution vector of (2.1) at the grid point x_n.

The stability inequality (5.3) applied with $v_n = u_n$, $w_n = y(x_n)$, together with (5.2), immediately gives the following result:

THEOREM 6.1. *The method Φ is convergent if it is consistent and stable. Moreover, if Φ has order p, then*

$$\max_{0\leq n\leq N} \|u_n - y(x_n)\| = O(|h|^p), \quad |h| \to 0. \tag{6.2}$$

7. Asymptotic error formula

In what follows, we shall need a refinement of Theorem 6.1, obtained independently by HENRICI [11] and TIHONOV and GORBUNOV [23], [24]. (For more recent alternative results, see RAKITSKIĬ [16]). We assume that

$$h_{n+1} = \vartheta(x_n)h, \quad n = 0,1,\ldots,N-1, \tag{7.1}$$

where $\vartheta(x)$ is piecewise continuous on $[a,b]$ and

$$\theta \leq \vartheta(x) \leq \Theta \quad \text{on } [a,b], \quad 0 < \theta \leq 1 \leq \Theta.$$

In addition, for the "base step" h in (7.1) we require that

$$0 < h \leq h_0\Theta^{-1}$$

so that $h_{n+1} \leq h_0$ in agreement with previous assumptions.

Algorithm (3.3) then becomes

$$\begin{cases} x_{n+1} = x_n + \vartheta(x_n)h, \\ u_{n+1} = u_n + \vartheta(x_n)h\Phi(x_n,u_n;\vartheta(x_n)h), \quad n = 0,1,\ldots,N-1, \\ x_0 = a, \quad u_0 = y_a, \end{cases} \tag{7.2}$$

with N such that $x_N = b$.

THEOREM 7.1. *Assume that*

(i) $\Phi(x,y;h) \in C^2[\mathcal{R}_0 \times [0,h_0]]$,
(ii) Φ *is a method of order* $p \geq 1$ *admitting a principal error function* $\tau(x,y) \in C[\mathcal{R}_0]$,
(iii) $e(x)$ *is the solution of the linear initial value problem*

$$\begin{cases} e' = f_y(x,y(x))e + [\vartheta(x)]^p \tau(x,y(x)), & a \leq x \leq b, \\ e(a) = 0, \end{cases} \tag{7.3}$$

where $f_y = [f^i_{yj}]$ *denotes the Jacobian matrix of* f.

Then

$$\max_{0 \leq n \leq N} \|u_n - y(x_n) - e(x_n)h^p\| = O(h^{p+1}), \quad h \to 0. \tag{7.4}$$

The last relation will be expressed more briefly in the form

$$u_n - y(x_n) = e(x_n)h^p + O(h^{p+1}), \quad 0 \leq n \leq N. \tag{7.4'}$$

8. Global error estimate

In order to estimate the error $u_n - y(x_n)$, neglecting terms of order $O(h^{p+1})$, it suffices, according to (7.4') to obtain $e(x_n)$ with an error of order $O(h)$. This can be achieved by integrating (7.3) with Euler's method, using appropriate approximations of the Jacobian matrix and the principal error function along the solution trajectory.

THEOREM 8.1. *Assume that*

(i) $\Phi(x,y;h) \in C^2[\mathcal{R}_0 \times [0,h_0]]$,
(ii) Φ *is a method of order* $p \geq 1$ *admitting a principal error function* $\tau(x,y) \in C^1[\mathcal{R}_0]$,
(iii) *an estimate* $r(x,y;h) \in C[\mathcal{R}_0 \times [0,h_0]]$ *is available for the truncation error* $t(x,y;h)$ *satisfying*

$$r(x,y;h) = t(x,y;h) + O(h^{p+1}), \quad h \to 0, \tag{8.1}$$

uniformly for $(x,y) \in \mathcal{R}_0$,

(iv) *along with* u_n *we generate the sequence* v_n, $n = 0,1,\ldots,N$, *in the following manner*:

$$\begin{cases} x_{n+1} = x_n + \vartheta(x_n)h, \\ u_{n+1} = u_n + \vartheta(x_n)h\Phi(x_n,u_n;\vartheta(x_n)h), \\ v_{n+1} = v_n + \vartheta(x_n)h[f_y(x_n,u_n)v_n + h^{-p}r(x_n,v_n;\vartheta(x_n)h)], \\ x_0 = a, \quad u_0 = y_a \quad v_0 = 0, \end{cases} \tag{8.2}$$

where $x_N = b$.

Then

$$u_n - y(x_n) = v_n h^p + O(h^{p+1}), \quad 0 \le n \le N. \tag{8.3}$$

PROOF. We note, first of all, that

$$f_y(x_n, u_n) = f_y(x_n, y(x_n)) + O(h^p), \tag{8.4}$$

$$h^{-p} r(x_n, u_n; h) = \tau(x_n, y(x_n)) + O(h). \tag{8.5}$$

Indeed, Eq (8.4) follows from (6.2) and from assumption (i), according to which $f(x, y) = \Phi(x, y; 0) \in C^2[\mathcal{R}_0]$. Moreover, since τ_y is continuous by assumption (ii),

$$\tau(x_n, u_n) = \tau(x_n, y(x_n)) + \tau_y(x_n, \overline{u}_n)(u_n - y(x_n)),$$

where $\overline{u}_n$ is a point on the segment from u_n to $y(x_n)$ (its exact location varies from component to component). Therefore, using again (6.2), we get $\tau(x_n, u_n) = \tau(x_n, y(x_n)) + O(h^p)$, and by assumption (iii) and (4.5),

$$\begin{aligned} r(x_n, u_n; h) &= t(x_n, u_n; h) + O(h^{p+1}) = \tau(x_n, u_n) h^p + O(h^{p+1}) \\ &= \tau(x_n, y(x_n)) h^p + O(h^{2p}) + O(h^{p+1}), \end{aligned}$$

from which (8.5) follows, since $p \ge 1$.

Let now $g(x, y) = f_y(x, y(x))y + [\vartheta(x)]^p \tau(x, y(x))$. Since the equation for v_{n+1} in (8.2) has the form $v_{n+1} = v_n + h_{n+1}(A_n v_n + b_n)$, with A_n bounded matrices and b_n bounded vectors, it follows from the corollary to Theorem 5.1 that

$$v_n = O(1), \quad h \to 0. \tag{8.6}$$

Substituting (8.4), (8.5), and (8.6) into the equation for v_{n+1}, and noting from (8.5) that

$$h^{-p} r(x_n, u_n; \vartheta(x_n) h) = [\vartheta(x_n)]^p \tau(x_n, y(x_n)) + O(h),$$

we find

$$\begin{aligned} v_{n+1} &= v_n + \vartheta(x_n) h \{ f_y(x_n, y(x_n)) v_n + [\vartheta(x_n)]^p \tau(x_n, y(x_n)) + O(h) \} \\ &= v_n + \vartheta(x_n) h g(x_n, v_n) + O(h^2). \end{aligned}$$

Since $v_0 = 0$, this is a $O(h^2)$-perturbation of Euler's method applied to $e' = g(x, e)$, $e(a) = 0$—the "variational equation" (7.3) of Theorem 7.1. Euler's method being stable, we can conclude that $v_n = e(x_n) + O(h)$, from which, by virtue of (7.4′), there follows (8.3). Theorem 8.1 is thus proved.

It is of some interest to note that assumption (ii), concerning τ, can be weakened to $\tau(x,y) \in C[\mathcal{R}_0]$. Since the stronger assumption has been used only to prove (8.5), it suffices to show that (8.5) can be obtained under the weaker assumption.

From the definition (4.2′) of the truncation error, we have

$$t_y(x,y;h) = \Phi_y(x,y;h) - h^{-1}[u_y(x+h) - u_y(x)], \tag{8.7}$$

where $u_y(t)$ is a solution of the initial value problem

$$\begin{cases} du_y/dt = f_y(t,u)u_y, & x \le t \le x + h_0, \\ u_y(x) = I, \end{cases}$$

I being the unit matrix. Moreover, $u_y \in C^2[x, x+h_0]$. Therefore,

$$h^{-1}[u_y(x+h) - u_y(x)] = du_y/dt|_{t=x} + hd^2u_y/dt^2|_{t=\xi} = f_y(x,y) + O(h),$$

where $x < \xi < x + h$ (the exact location of ξ varies from component to component). Using this last relation in (8.7), together with $\Phi_y(x,y;h) = \Phi_y(x,y;0) + h\Phi_{yh}(x,y;\overline{h}) = f_y(x,y) + O(h)$, we obtain

$$t_y(x,y;h) = O(h), \quad h \to 0. \tag{8.8}$$

Now, by (8.1),

$$\begin{aligned} r(x_n,u_n;h) &= t(x_n,u_n;h) + O(h^{p+1}) \\ &= t(x_n,y(x_n);h) + t_y(x_n,\overline{u}_n;h)[u_n - y(x_n)] + O(h^{p+1}), \end{aligned}$$

and therefore, by (8.8) and (6.2),

$$r(x_n,u_n;h) = t(x_n,y(x_n);h) + O(h^{p+1}) = \tau(x_n,y(x_n)))h^p + O(h^{p+1}),$$

which, again, establishes (8.5).

9. Local error estimators

Many estimators $r(x,y;h)$ for the truncation error have been found that satisfy (8.1). The best known, perhaps, is the one based on Richardson extrapolation to zero. Yet, this procedure is rather inefficient in terms of additional function evaluations. More attractive are estimators that use pairs of "one-step" methods. If Φ is the basic method of integration, of order p, and Φ^* any method of order $p^* = p+1$, then

$$r(x,y;h) = \Phi(x,y;h) - \Phi^*(x,y;h) \tag{9.1}$$

is an acceptable estimator. Indeed, from the definition (4.2),

$$\Phi(x, y; h) - h^{-1}[u(x+h) - u(x)] = t(x, y; h),$$

$$\Phi^*(x, y; h) - h^{-1}[u(x+h) - u(x)] = O(h^{p+1}),$$

from which (8.1) follows by subtraction.

Frequently, Φ is an explicit Runge–Kutta process with s stages,

$$\begin{cases} k_1 = f(x, y), \\ k_\sigma = f\left(x + \mu_\sigma h, y + h \sum_{\tau=1}^{\sigma-1} \lambda_{\sigma\tau} k_\tau\right), \quad \sigma = 2, 3, \ldots, s, \\ \Phi(x, y; h) = \sum_{\sigma=1}^{s} \alpha_\sigma k_\sigma(x, y; h). \end{cases}$$

In order to make (9.1) efficient, one chooses for Φ^* an analogous process with s^* stages, where $s^* > s$, in such a way that

$$\mu_\sigma^* = \mu_\sigma, \quad \lambda_{\sigma\tau}^* = \lambda_{\sigma\tau} \quad \text{for } \sigma = 2, 3, \ldots, s.$$

The estimator $r(x, y; h)$ then "costs" only $s^* - s$ additional evaluations of f. If $s^* = s + 1$, one can even try to save another evaluation by choosing (if possible)

$$\mu_{s^*} = 1, \quad \lambda_{s^*\tau} = \alpha_\tau, \quad \text{for } \tau = 1, 2, \ldots, s^* - 1. \tag{9.2}$$

In this case, indeed, k_{s^*} will be identical with k_1 of the next step.

Many pairs of Runge–Kutta formulae of this type have been developed by FEHLBERG [6], [7], [8]. There is considerable freedom in the choice of the parameters μ_σ, $\lambda_{\sigma\tau}$, α_σ. The choices made by Fehlberg were guided by an attempt to reduce the magnitude of the principal error function $\tau(x, y)$ of the method Φ. His formulae correspond to values of p, s, s^* shown below:

p	3	4	5	6	7	8
s	4	5	6	8	11	15
s^*	5	6	8	10	13	17

For $p = 3$ ($p^* = 4$), for example, the formulae satisfy (9.2), and take on the following form:

$$
\begin{aligned}
k_1 &= f(x,y),\\
k_2 &= f(x+\tfrac{2}{7}h, y+\tfrac{2}{7}hk_1),\\
k_3 &= f(x+\tfrac{7}{15}h, y+\tfrac{77}{900}hk_1+\tfrac{343}{900}hk_2),\\
k_4 &= f(x+\tfrac{35}{38}h, y+\tfrac{805}{1444}hk_1-\tfrac{77175}{54872}hk_2+\tfrac{97125}{54872}hk_3),\\
y_h &= y+h(\tfrac{79}{490}k_1+\tfrac{2175}{3626}k_2+\tfrac{2166}{9065}k_4),\\
k_5 &= f(x+h, y_h),\\
y_h^* &= y+h(\tfrac{229}{1470}k_1+\tfrac{1125}{1813}k_3+\tfrac{13718}{81585}k_4+\tfrac{1}{18}k_5),\\
r(x,y;h) &= h^{-1}(y_h-y_h^*).
\end{aligned}
\tag{9.3}
$$

Similar formulae were developed by other authors; see for example, CESCHINO [2], TANAKA [22], BACHMANN [1], SARAFYAN [17], ENGLAND [4]. Estimators that use information on several consecutive steps are given by SHINTANI [18], [19], PROTHERO [15], KIŠ [13], [14], and HUDDLESTON [12].

10. Numerical example

We illustrate Theorem 8.1 by applying Fehlberg's third-order method (9.3) to an example taken from [9], that is

$$
\begin{aligned}
d^2c/dx^2 &= -\pi^2x^2c-\pi s(c^2+s^2)^{-1/2},\\
& \qquad\qquad 2\sqrt{q}\le x\le 2\sqrt{q}+1,\\
d^2s/dx^2 &= -\pi^2x^2s+\pi c(c^2+s^2)^{-1/2}
\end{aligned}
\tag{10.1}
$$

where $q\ge 0$ is an integer. The initial conditions are chosen to be

$$
c=1,\ dc/dx=0,\ s=0,\ ds/dx=2\pi\sqrt{q}\quad \text{for } x=2\sqrt{q},
\tag{10.2}
$$

which (for each $q=0,1,2,\dots$) identify the solution

$$
c(x)=\cos(\tfrac{\pi}{2}x^2),\ s(x)=\sin(\tfrac{\pi}{2}x^2),\ \ 2\sqrt{q}\le x\le 2\sqrt{q}+1.
\tag{10.3}
$$

For the purpose of this illustration, (10.1) is treated as a system of four first-order differential equations for the vector-valued function

$$
y(x)=\begin{bmatrix} c(x)\\ c'(x)\\ s(x)\\ s'(x)\end{bmatrix}.
$$

The length of the interval of integration is kept constant at 1, but the interval itself is moved to the right as q assumes the values $0, 1, 2, \ldots$, thus entering into regions of gradually increasing frequencies. One expects, therefore, that the error estimation becomes more difficult with increasing q.

We choose (arbitrarily) the step-control function to be

$$\vartheta(x) = \begin{cases} 1 \text{ if } 0 \le \xi \le \frac{1}{4}, \\ \frac{1}{2} \text{ if } \frac{1}{4} < \xi \le \frac{1}{2}, \\ 1 \text{ if } \frac{1}{2} < \xi \le \frac{3}{4}, \\ 2 \text{ if } \frac{3}{4} < \xi \le 1, \end{cases}$$

where $\xi = x - 2\sqrt{q}$. Selected numerical results are reported below in Table 1. The first column contains the values of q, the second some selected values x_n of x, the third the observed global errors $\|u_n - y(x_n)\|_\infty$, and the fourth the estimates $\|v_n h^3\|_\infty$ according to (8.3) (where $p = 3$). The column headed by "%" indicates the discrepancy in percents between the actual and estimated errors. The lower part of the table shows only the errors and estimates of maximum discrepancy.

Table 1. Global errors, and their estimates, for the example (10.1), computed by Fehlberg's method (9.3) and the estimation procedure (8.2). (Numbers in parentheses indicate decimal exponents, for example $4.17(-5) = 4.17 \times 10^{-5}$).

		$h = .025$			$h = .0125$			$h = .00625$		
q	x	error	est	%	error	est	%	error	est	%
		$\times 10^7$	$\times 10^7$		$\times 10^8$	$\times 10^8$		$\times 10^9$	$\times 10^9$	
0	.1	.67665	.67638	.04	.84548	.84529	.02	1.0566	1.0565	.01
	.2	1.3532	1.3484	.36	1.6850	1.6818	.19	2.1017	2.0996	.10
	.3	1.7165	1.7037	.75	2.1259	2.1175	.39	2.6442	2.6388	.20
	.4	1.7566	1.7400	.95	2.1663	2.1555	.50	2.6885	2.6815	.26
	.5	1.7439	1.7200	1.37	2.1349	2.1195	.72	2.6395	2.6295	.38
	.6	5.0043	5.1641	3.19	6.5333	6.6252	1.41	8.3289	8.3839	.66
	.7	12.075	12.161	.72	15.551	15.584	.22	19.689	19.704	.07
	.8	69.886	66.962	4.18	89.073	86.832	2.52	111.59	110.06	1.37
	.9	219.68	202.87	7.65	267.98	256.04	4.46	327.51	319.58	2.42
	1.0	417.22	368.32	11.72	476.14	442.74	7.01	559.66	537.85	3.90
0	max	4.17(−5)	3.68(−5)	11.7	4.76(−6)	4.43(−6)	7.01	5.60(−7)	5.38(−7)	3.90
1	max	2.10(−4)	2.63(−4)	25.0	.970(−3)	1.19(−3)	23.1	9.04(−5)	1.05(−4)	16.3
2	max	2.73(−3)	3.69(−3)	34.9	3.95(−4)	5.09(−4)	28.8	2.28(−4)	2.71(−4)	19.0
3	max	7.45(−3)	11.5(−3)	54.4	4.82(−3)	6.67(−3)	38.3	8.43(−4)	1.11(−3)	31.5
4	max	1.45(−2)	2.62(−2)	81.3	1.49(−3)	2.18(−3)	46.8	1.61(−3)	2.19(−3)	36.1
5	max	3.34(−2)	6.94(−2)	107	2.79(−3)	4.50(−3)	61.1	2.88(−3)	4.07(−3)	41.2

As expected, the quality of the estimates worsens with increasing q, but improves with decreasing h. For $q = 0$ the percental discrepancy is about halved each time h is reduced to $h/2$, indicating that the results respect the asymptotic law expressed in (8.3). For $q > 0$, the technique is not yet sufficiently refined, at this point, but the estimates, nevertheless, are rather satisfactory on the whole.

References

[1] BACHMANN, K. H.: *Genäherte Integration von Differentialgleichungssystemen mit Schrittweitensteuerung*, Z. Angew. Math. Mech. **48** (1968), 210–212.

[2] CESCHINO, F.: *Evaluation de l'erreur par pas dans les problèmes différentiels*, Chiffres **5** (1962), 223–229.

[3] CODDINGTON, E. A. AND LEVINSON, N.: *Theory of ordinary differential equations*, McGraw-Hill, New York, 1955.

[4] ENGLAND, R.: *Error estimates for Runge–Kutta type solutions to systems of ordinary differential equations*, Comput. J. **12** (1969/70), 166–170.

[5] ESSER, H. AND SCHERER, K.: *Konvergenzordnungen von Ein- und Mehrschrittverfahren bei gewöhnlichen Differentialgleichungen*, Computing **12** (1974), 127–143.

[6] FEHLBERG, E.: *Classical fifth-, sixth-, seventh-, and eighth-order Runge–Kutta formulas with stepsize control*, NASA Technical Report 287 (1968).

[7] FEHLBERG, E.: *Klassische Runge–Kutta-Formeln fünfter und siebenter Ordnung mit Schrittweiten-Kontrolle*, Computing 4 (1969), 93–106.

[8] FEHLBERG, E.: *Klassische Runge–Kutta-Formeln vierter und niedrigerer Ordnung mit Schrittweiten-Kontrolle und ihre Anwendung auf Wärmeleitungsprobleme*, Computing **6** (1970), 61–71.

[9] FEHLBERG, E.: *Klassische Runge–Kutta–Nyström-Formeln mit Schrittweiten-Kontrolle für Differentialgleichungen* $\ddot{x} = f(t, x)$, Computing **10** (1972), 305–315.

[10] GRIGORIEFF, R. D.: *Numerik gewöhnlicher Differentialgleichungen*, Vol. 1: Einschrittverfahren, Teubner, Stuttgart, 1972.

[11] HENRICI, P.: *Discrete variable methods in ordinary differential equations*, John Wiley, New York and London, 1962.

[12] HUDDLESTON, R. E.: *Variable-step truncation error estimates for Runge–Kutta methods of order* 4 *or less*, J. Math. Anal. Appl. **39** (1972), 636–646.

[13] KIŠ, O.: *On an error estimate for the Runge–Kutta method* (Russian), Studia Sci. Math. Hungar. **5** (1970), 427–432.

[14] KIŠ, O.: *On the Runge–Kutta method* (Russian), Studia Sci. Math. Hungar. **5** (1970), 433–435.

[15] PROTHERO, A.: *Local-error estimates for variable-step Runge–Kutta methods*, Conf. on Numerical Solution of Differential Equations (Dundee, 1969), 228–233. Lecture Notes in Mathematics **109**, Springer, Berlin-Heidelberg-New York, 1969.

[16] RAKITSKIĬ, JU. V.: *Asymptotic error formulae for the solutions of systems of ordinary differential equations by functional numerical methods* (Russian), Dokl. Akad. Nauk SSSR **193** (1970), 40–42.

[17] SARAFYAN, D.: *Error estimation for Runge–Kutta methods through pseudo-iterative formulas*, Riv. Mat. Univ. Parma (2) **9** (1968), 1–42.

[18] SHINTANI, H.: *Approximate computation of errors in numerical integration of ordinary differential equations by one-step methods*, J. Sci. Hiroshima Univ. Ser. A-I Math. **29** (1965), 97–120.

[19] SHINTANI, H.: *On a one-step method of order* 4, J. Sci. Hiroshima Univ. Ser. A-I Math. **30** (1966), 91–107.

[20] STETTER, H. J.: *Local estimation of the global discretization error*, SIAM J. Numer. Anal. **8** (1971), 512–523.

[21] STETTER, H. J.: *Analysis of discretization methods for ordinary differential equations*, Springer, New York - Heidelberg - Berlin, 1973.

[22] TANAKA, M.: *Runge–Kutta formulas with the ability of error estimation*, Rep. Statist. Appl. Res. Un. Japan. Sci. Engrs. **13** (1966), no. 3, 42–62.

[23] TIHONOV, A. N. AND GORBUNOV, A. D.: *Asymptotic error bounds for the Runge–Kutta method* (Russian), Ž. Vyčisl. Mat. i Mat. Fis. **3** (1963), 195–197.

[24] TIHONOV, A. N. AND GORBUNOV, A. D.: *Error estimates for a Runge–Kutta type method and the choice of optimal meshes* (Russian), Ž. Vyčisl. Mat. i Mat. Fis. **4** (1964), 232–241.

27.3. [73] (with M. Montrone) "Multistep methods with minimum global error coefficient"

---

[73] (with M. Montrone) "Multistep methods with minimum global error coefficient," *Calcolo* **17**, 67–75 (1980) (translated from Italian).

Multistep methods with minimum global error coefficient* **

W. GAUTSCHI*** and M. MONTRONE†

Abstract. We consider linear k-step methods of maximal order in which the roots $\zeta_i \neq 1$ of the first characteristic polynomial are constrained to satisfy $|\zeta_i| \leq \gamma$, $0 \leq \gamma < 1$. We find the unique method of this class having minimum global error coefficient.

1. Let us consider a generic k-step method

$$\sum_{s=0}^{k} \alpha_s y_{n+s} = h \sum_{s=0}^{k} \beta_s f_{n+s}, \quad \alpha_k = 1, \tag{1.1}$$

for the solution of the initial value problem

$$y' = f(x, y), \quad y(x_0) = y_0. \tag{1.2}$$

With the method (1.1) one associates the linear functional [1, p. 327]

$$Lu = \sum_{s=0}^{k} [\alpha_s u(s) - \beta_s u'(s)]. \tag{1.3}$$

The method has order p if and only if

$$L\,t^r = 0, \quad r = 0, 1, \ldots, p; \quad L\,t^{p+1} \neq 0.$$

The global error coefficient is [3, p. 223]

* Work carried out under the finalized project "Medicina Preventiva" (MPP 1).

** English translation by Walter Gautschi of "Metodi multistep con minimo coefficiente dell'errore globale", *Calcolo* **17** (1980), 67–75.

*** Department of Computer Sciences, Purdue University, Lafayette, Indiana, U. S. A.

† Istituto di Analisi Matematica, Università di Bari, Collaboratore G. N. I. M.

$$C_{k,p} = \frac{L\,\frac{t^{p+1}}{(p+1)!}}{\sum_{s=0}^{k}\beta_s}, \tag{1.4}$$

where p is the order of the method.

We assume that the characteristic polynomial

$$\rho(\zeta) = \sum_{s=0}^{k}\alpha_s\zeta^s \tag{1.5}$$

has roots ζ_s with

$$\begin{aligned} &1 = \zeta_1 \geq |\zeta_2| \geq |\zeta_3| \geq \cdots \geq |\zeta_k|, \\ &\zeta_s \text{ simple if } |\zeta_s| = 1. \end{aligned} \tag{1.6}$$

These conditions, as is known, are indispensable for the convergence of the method. Moreover, Dahlquist's theory ensures that, given such a polynomial $\rho(\zeta)$ of degree k, it is always possible to determine, correspondingly, a convergent k-step method of order at least $k+1$. Only if k is even, $\zeta_2 = -1$, and the roots ζ_s, $s = 3, \ldots, k$, are distinct and complex of modulus one, can the method have order $k+2$, which is the maximum order possible.

As far as the interval of absolute stability associated with the method is concerned, it is well known, however, that this interval is the larger the further inside the unit circle the roots of $\rho(\zeta)$ different from 1 are located.

In view of these considerations, we fix γ, $0 \leq \gamma < 1$, in this work, and examine the class Δ_γ of characteristic polynomials for which $\zeta_1 = 1$, $|\zeta_s| \leq \gamma$ for $s = 2, 3, \ldots, k$, and, among all polynomials in Δ_γ, we look for the one that minimizes in absolute value the global error coefficient in the corresponding k-step method of order $k+1$.

We find that this minimum is attained for the polynomial having all roots different from 1 concentrated at the point $-\gamma$.

For some values of γ and k we construct the multistep methods associated with this characteristic polynomial, and we determine the respective intervals of absolute stability.

2. Consider the transformation $\zeta = \frac{1+z}{1-z}$, $z = \frac{\zeta-1}{\zeta+1}$, which maps the unit circle into the negative half-plane and, in particular, the point $\zeta = 1$ into $z = 0$; moreover, it maps the circle $\Gamma_\gamma : |\zeta| \leq \gamma < 1$ into a circle C_γ in the same half-plane (see Figure 1).

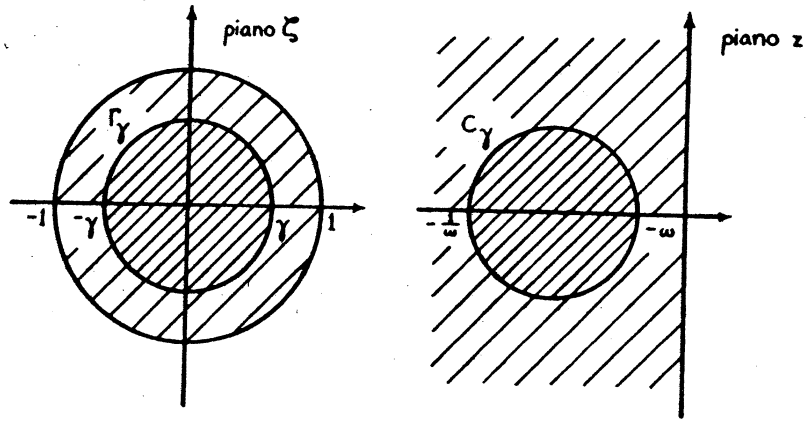


Fig. 1.

The circle C_γ intersects the real axis in the points $-\omega$ and $-\dfrac{1}{\omega}$, where $\omega = \dfrac{1-\gamma}{1+\gamma}$.

Let the transformed characteristic polynomials be

$$r(z) = \left(\frac{1-z}{2}\right)^k \rho\left(\frac{1+z}{1-z}\right), \qquad s(z) = \left(\frac{1-z}{2}\right)^k \sigma\left(\frac{1+z}{1-z}\right), \tag{2.1}$$

where $\sigma(\zeta) = \sum_{s=0}^{k} \beta_s \zeta^s$ is the second characteristic polynomial.

If the polynomial $\rho(\zeta)$ has a root of multiplicity p at $\hat{\zeta}$, then the polynomial $r(z)$ has a root of the same multiplicity at $z = \dfrac{\hat{\zeta}-1}{\hat{\zeta}+1}$, if $\hat{\zeta} \neq -1$, or has degree $k-p$, if $\hat{\zeta} = -1$, and vice versa. Therefore, all the roots of $r(z)$ are contained in the negative half-plane. Furthermore, $r(1) = 1$ and, letting

$$r(z) = a_1 z + a_2 z^2 + \cdots + a_k z^k, \tag{2.2}$$

we have that $a_s > 0$, $s = 1, 2, \ldots, k$, if $\rho \in \Delta_\gamma$.

We denote the zeros of $r(z)$ by z_s, $0 = z_1 < |z_2| \leq |z_3| \leq \cdots \leq |z_k|$.

The class Δ_γ of polynomials $\rho(\zeta)$ transforms into the class D_γ of polynomials $r(z)$ defined by

$$D_\gamma = \{r(z) : \ z_1 = 0, z_s \in C_\gamma, \ s = 2, 3, \ldots, k\}. \tag{2.3}$$

If we put

$$\frac{z}{\ln\frac{1+z}{1-z}}\cdot\frac{r(z)}{z}=b_0+b_1z+b_2z^2+\cdots, \tag{2.4}$$

it is known (cf. [3, p. 230]) that for every polynomial $r(z)$ one obtains the k-step formula of maximum order by letting

$$s(z)=b_0+b_1z+\cdots+b_kz^k. \tag{2.5}$$

Moreover,

$$C_{k,p}=\frac{b_p}{2^pb_0}, \tag{2.6}$$

where p is the order realized in this way. If $r\in D_\gamma$, then $p=k+1$.

The problem at hand, therefore, consists in determining

$$\min_{r\in D_\gamma}|C_{k,k+1}|=\min_{r\in D_\gamma}\left|\frac{b_{k+1}}{2^{k+1}b_0}\right|. \tag{2.7}$$

3. We put

$$\frac{z}{\ln\frac{1+z}{1-z}}=\lambda_0+\lambda_2z^2+\lambda_4z^4+\cdots, \tag{3,1}$$

and recall (cf. [3, p. 231]) that $\lambda_0=\frac{1}{2}$, $\lambda_{2\nu}<0$ for $\nu\geq 1$.

It then follows from (2.4) that

$$b_0=\tfrac{1}{2}a_1,$$

$$b_{k+1}=\begin{cases}\lambda_{k+1}a_1+\lambda_{k-1}a_3+\cdots+\lambda_2a_k, & k \text{ odd},\\ \lambda_ka_2+\lambda_{k-2}a_4+\cdots+\lambda_2a_k, & k \text{ even}.\end{cases} \tag{3.2}$$

We must minimize

$$2^k|C_{k,k+1}|=\begin{cases}|\lambda_{k+1}|+|\lambda_{k-1}|\frac{a_3}{a_1}+\cdots+|\lambda_2|\frac{a_k}{a_1}, & k \text{ odd},\\ |\lambda_k|\frac{a_2}{a_1}+|\lambda_{k-2}|\frac{a_4}{a_1}+\cdots+|\lambda_2|\frac{a_k}{a_1}, & k \text{ even}.\end{cases} \tag{3.3}$$

Putting $u_j=-z_j$, $j=2,3,\ldots,k$, we have

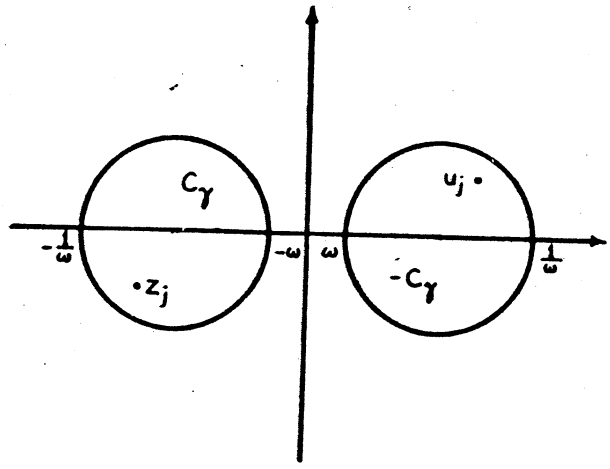


Fig. 2.

$$r(z) = z\,\frac{\prod_{j=2}^{k}(z+u_j)}{\prod_{j=2}^{k}(1+u_j)}. \tag{3.4}$$

The points u_j are located in the circle $-C_\gamma$ (see Figure 2).

Now,

$$\prod_{j=2}^{k}(z+u_j) = \sigma_0(u)z^{k-1} + \sigma_1(u)z^{k-2} + \cdots + \sigma_{k-1}(u), \tag{3.5}$$

where

$$\sigma_0(u) = 1,\ \sigma_1(u) = u_1 + u_2 + \cdots + u_k, \ldots, \sigma_{k-1}(u) = u_1u_2\cdots u_k$$

are the elementary symmetric functions in the variables $u_2, u_3, \ldots, u_k$.

Therefore,

$$\frac{a_s}{a_1} = \frac{\sigma_{k-s}(u)}{\sigma_{k-1}(u)} = \sigma_{s-1}(v), \quad s = 1, 2, \ldots, k, \tag{3.6}$$

where $\sigma_{s-1}(v)$ are the elementary symmetric functions in the variables $v_2, v_3, \ldots, v_k$ with $v_j = \frac{1}{u_j}$. Since the transformation $v = 1/u$ maps the circle $-C_\gamma$ into itself, we have that $v_j \in -C_\gamma$, $j = 2, 3, \ldots, k$.

If among the v_j there are conjugate complex pairs, $v_\mu = \xi_\mu + i\eta_\mu$, $\overline{v}_\mu = \xi_\mu - i\eta_\mu$, $\eta_\mu > 0$, then from the identity

$$\prod_{\lambda}(z+v_\lambda)\prod_{\mu}[(z+\xi_\mu)^2 + \eta_\mu^2] = \sum_{s=1}^{k}\sigma_{s-1}(v)z^{k-s}, \tag{3.7}$$

where v_λ and ξ_μ are positive, it is clear that each $\sigma_{s-1}(v)$ is a nondecreasing function of η_μ.

To minimize $\sigma_{s-1}(v)$, it thus suffices to consider the case in which all v_j are real. In this case, the minimum of $\sigma_{s-1}(v)$ for the v_j varying in $-C_\gamma$ clearly obtains if $v_2 = v_3 = \cdots = v_{k-1} = \omega$, independently of s.

Moreover,

$$\min_{v_j \in -C_\gamma} \sigma_{s-1}(v) = \binom{k-1}{s-1} \omega^{s-1}. \tag{3.8}$$

Therefore, by (3.3) and (3.6), one has

$$\min_{r \in D_\gamma} |C_{k,k+1}| = \frac{1}{2^k} \begin{cases} |\lambda_{k+1}| + \binom{k-1}{2}|\lambda_{k-1}|\omega^2 + \binom{k-1}{4}|\lambda_{k-3}|\omega^4 + \cdots + |\lambda_2|\omega^{k-1} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k \text{ odd}, \\ \binom{k-1}{1}|\lambda_k|\omega + \binom{k-1}{3}|\lambda_{k-2}|\omega^3 + \cdots + |\lambda_2|\omega^{k-1} \quad k \text{ even}. \end{cases}$$

The polynomial realizing the minimum is

$$r(z) = z\left(\frac{z + \frac{1}{\omega}}{1 + \frac{1}{\omega}}\right)^{k-1} \tag{3.9}$$

to which corresponds the charateristic polynomial

$$\rho(\zeta) = (\zeta - 1)(\zeta + \gamma)^{k-1}. \tag{3.10}$$

4. We note, in the case of k odd, that

$$\lim_{\gamma \to 1} \min_{r \in D_\gamma} |C_{k,k+1}| = 2^{-k}|\lambda_{k+1}|. \tag{4.1}$$

Actually, from (3.3) it follows that

$$\inf_{\rho \in \Delta} |C_{k,k+1}| = 2^{-k}|\lambda_{k+1}|, \quad k \text{ odd}, \tag{4.2}$$

where Δ is the class of characteristic polynomials satisfying (1.6). Eq (3.10) suggests (but does not prove!) that, in general, there does not exist a zero-stable method (that is, a polynomial $\rho \in \Delta$) for which $|C_{k,k+1}|$ is equal to the infimum in (4.2). We prove, in fact, as already asserted in [2], that this is true whenever k (odd) ≥ 5, while for $k = 3$, every k-step method with zeros $\zeta_1 = 1$, $\zeta_2 = -1$, $-1 < \zeta_3 < 1$ has minimum coefficient $|C_{3,4}| = \frac{1}{180}$.

It is clear, by (3.3), that $|C_{k,k+1}| = 2^{-k}|\lambda_{k+1}|$ is possible only if $a_3 = a_5 = \cdots = a_k = 0$, that is, if

$$r(z) = a_1 z + a_2 z^2 + a_4 z^4 + \cdots + a_{k-1} z^{k-1}. \tag{4.3}$$

There follows, in particular, that $\zeta = -1$ is a zero of $\rho(\zeta)$.

If $k = 3$, we have

$$r(z) = z(a_1 + a_2 z).$$

By the stability assumption it follows that $r(z)$ has the zeros $z_1 = 0$, $z_2 = \ell$ with $-\infty < \ell < 0$ arbitrary, from which

$$\rho(\zeta) = (\zeta - 1)(\zeta + 1)(\zeta - \lambda), \quad -1 < \lambda < 1.$$

Each of these polynomials thus generates a zero-stable 3-step method, of order 4, having minimum coefficient $|C_{3,4}| = \frac{1}{8}|\lambda_4| = \frac{1}{180}$.

Assume now k (odd) ≥ 5. Since $\rho \in \Delta$, we have $a_{k-1} \neq 0$, and the sum of the zeros z_j of $r(z)$, being equal to $-\frac{a_{k-2}}{a_{k-1}}$, must be zero. In particular,

$$\sum_{j=1}^{k-1} \operatorname{Re} z_j = 0.$$

On the other hand, by the same assumption $\rho \in \Delta$, we have $\operatorname{Re} z_j \leq 0$, from which, necessarily, $\operatorname{Re} z_j = 0$, $j = 1, 2, \ldots, k-1$. Since $z_1 = 0$, the latter is compatible with zero-stability only if $r(z)$ has odd degree, contradicting (4.3).

In an analogous manner one proves (see also [3, p. 286, Problem 37]) that for k (even) ≥ 2,

$$\inf_{\rho \in \Delta} |C_{k,k+2}| = 2^{-k-1}|\lambda_{k+2}|. \tag{4.4}$$

If $k \geq 4$, there does not exist a zero-stable method of order $k+2$ attaining the infimum in (4.4), while for $k = 2$ Milne's method is the unique 2-step method with $C_{2,4} = \frac{1}{8}|\lambda_4| = \frac{1}{189}$.

5. We now examine the methods of Section 3 corresponding to $k = 2$ and $k = 3$.

For $k = 2$ one obtains

$$y_{n+2} - (1-\gamma)y_{n+1} - \gamma y_n = \frac{h}{12}[(5-\gamma)f_{n+2} + 8(1+\gamma)f_{n+1} - (1-5\gamma)f_n] \tag{5.1}$$

with

$$C = -\frac{1}{24}\frac{1-\gamma}{1+\gamma}. \tag{5.2}$$

Note that for $\gamma = 0$ one obtains the Adams–Moulton method with error constant $C = -1/24$. For $\gamma = 1$ one obtains Milne's method.

The interval of absolute stability is [4, p. 74]

$$\left[-6\frac{1-\gamma}{1+\gamma}, 0\right]. \tag{5.3}$$

For $k = 3$ one has

$$\begin{aligned} &y_{n+3} - (1-2\gamma)y_{n+2} - \gamma(2-\gamma)y_{n+1} - \gamma^2 y_n \\ &= \tfrac{h}{24}\left[(9 - 2\gamma + y^2)f_{n+3} + (19 + 26\gamma - 5\gamma^2)f_{n+2}\right. \\ &\left.-(5 - 26\gamma - 19\gamma^2)f_{n+1} + (1 - 2\gamma + 9\gamma^2)f_n\right] \end{aligned} \tag{5.4}$$

with

$$C = -\frac{1}{720}\frac{19 - 22\gamma + 19\gamma^2}{(1+\gamma)^2}. \tag{5.5}$$

The interval of absolute stability is

$$\left[-3\frac{1-\gamma}{1+\gamma}, 0\right]. \tag{5.6}$$

If $\gamma = 0$ one obtains the Adams–Moulton method.

References

[1] Antosiewicz, H. A. and W. Gautschi, *Numerical methods in ordinary differential equations*, Survey of Numerical Analysis (J. Todd, ed.), 314–346, McGraw-Hill, New York, 1962.

[2] Gautschi, W., *On error reducing multistep methods*, Notices Amer. Math. Soc. (65), **10** (1963), 95.

[3] Henrici, P., *Discrete variable methods in ordinary differential equations*, Wiley, New York, 1962.

[4] Lambert, J. D., *Computational methods in ordinary differential equations*, Wiley, London & New York, 1973.

28

Papers on Computer Algorithms and Software Packages

---

141 Algorithm 726: ORTHPOL — a package of routines for generating orthogonal polynomials and Gauss-type quadrature rules, ACM Trans. Math. Software 20, 21–62 (1994); Remark on Algorithm 726, *ibid.* 24, 355 (1998)

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 (2006)

28.1. [141] "Algorithm 726: ORTHPOL — A Package of Routines for Generating Orthogonal Polynomials and Gauss-Type Quadrature Rules"

---

[141] "Algorithm 726: ORTHPOL — A Package of Routines for Generating Orthogonal Polynomials and Gauss-Type Quadrature Rules," *ACM Trans. Math. Software* **20**, 21–62 (1994); Remark on Algorithm 726, *ibid.* **24**, 355 (1998).

Algorithm 726: ORTHPOL—A Package of Routines for Generating Orthogonal Polynomials and Gauss-Type Quadrature Rules

WALTER GAUTSCHI
Purdue University

A collection of subroutines and examples of their uses, as well as the underlying numerical methods, are described for generating orthogonal polynomials relative to arbitrary weight functions. The object of these routines is to produce the coefficients in the three-term recurrence relation satisfied by the orthogonal polynomials. Once these are known, additional data can be generated, such as zeros of orthogonal polynomials and Gauss-type quadrature rules, for which routines are also provided.

Categories and Subject Descriptors: G.1.2 [**Numerical Analysis**]: Approximation; G.1.4 [**Numerical Analysis**]: Quadrature and Numerical Differentiation; G.4 [**Mathematical Software**]

General Terms: Algorithms

Additional Key Words and Phrases: Gauss-type quadrature rules, orthogonal polynomials

1. INTRODUCTION

Classical orthogonal polynomials, such as those of Legendre, Chebyshev, Laguerre, and Hermite, have been used for purposes of approximation in widely different disciplines and over a long period of time. Their popularity is due in part to the ease with which they can be employed and in part to the wealth of analytic results known for them. Widespread use of nonclassical orthogonal polynomials, in contrast, has been impeded by a lack of effective and generally applicable constructive methods. The present set of computer routines has been developed over the past 10 years in the hope of remedying this impediment and of encouraging the use of nonstandard orthogonal polynomials. A number of applications indeed have already been made, for

This work was supported in part by National Research Foundation grants, most recently by grant DMS-9023403.
Author's address: Department of Computer Sciences, Purdue University, West Lafayette, IN 47907-1398.

example, to numerical quadrature (Cauchy principal value integrals with coth-kernel [Gautschi et al. 1987], Hilbert transform of Jacobi weight functions [Gautschi and Wimp 1987], integration over half-infinite intervals [Gautschi 1991c], rational Gauss-type quadrature [Gautschi 1993a; 1993b]), to moment-preserving spline approximation [Gautschi 1984a; Gautschi and Milovanović 1986; Frontini et al. 1987], to the summation of slowly convergent series [Gautschi 1991a, 1991b], and, perhaps most notably, to the proof of the Bieberbach conjecture [Gautschi 1986b].

In most applications, orthogonality is with respect to a positive weight function, w, on a given interval or union of intervals, or with respect to positive weights, w_t, concentrated on a discrete set of points, $\{x_t\}$, or a combination of both. For convenience of notation, we subsume all of these cases under the notion of a positive measure, $d\lambda$, on the real line $\mathbb{R}$. That is, the respective inner product is written as a Riemann–Stieltjes integral,

$$(u, v) = \int_{\mathbb{R}} u(t)v(t)\, d\lambda(t), \tag{1.1}$$

where the function $\lambda(t)$ is the indefinite integral of w for the continuous part, and a step function with jumps w_t at x_t for the discrete part. We assume that (1.1) is meaningful whenever u, v are polynomials. There is then defined a unique set of (monic) orthogonal polynomials,

$$\pi_k(t) = t^k + \text{lower-degree terms}, \qquad k = 0, 1, 2, \ldots,$$
$$(\pi_k, \pi_\ell) = 0 \qquad \text{if} \quad k \neq \ell. \tag{1.2}$$

We speak of "continuous" orthogonal polynomials if the support of $d\lambda$ is an interval or a union of intervals, of "discrete" orthogonal polynomials if the support of $d\lambda$ consists of a discrete set of points, and of orthogonal polynomials of "mixed type" if the support of $d\lambda$ has both a continuous and discrete part. In the first and last cases, there are infinitely many orthogonal polynomials, one for each degree, whereas in the second case, there are exactly N orthogonal polynomials, $\pi_0, \pi_1, \ldots, \pi_{N-1}$, where N is the number of support points. In all cases, we denote the polynomials by $\pi_k(\cdot) = \pi_k(\cdot; d\lambda)$, or $\pi_k(\cdot; w)$, if we want to indicate their dependence on the measure $d\lambda$ or weight function w, and use similar notations for other quantities depending on $d\lambda$ or w.

It is a distinctive feature of orthogonal polynomials, compared to other orthogonal systems, that they satisfy a three-term recurrence relation,

$$\pi_{k+1}(t) = (t - \alpha_k)\pi_k(t) - \beta_k \pi_{k-1}(t), \qquad k = 0, 1, 2, \ldots,$$
$$\pi_0(t) = 1, \qquad \pi_{-1}(t) = 0, \tag{1.3}$$

with coefficients $\alpha_k = \alpha_k(d\lambda) \in \mathbb{R}$, $\beta_k = \beta_k(d\lambda) > 0$ that are uniquely determined by the measure $d\lambda$. By convention, the coefficient β_0, which multiplies

$\pi_{-1} = 0$ in (1.3), is defined by

$$\beta_0 = \beta_0(d\lambda) = \int_{\mathbb{R}} d\lambda(t). \tag{1.4}$$

The knowledge of these coefficients is absolutely indispensable for any sound computational use and application of orthogonal polynomials [Gautschi 1982a, 1990]. One of the principal objectives of the present package is precisely to provide routines for generating these coefficients. Routines for related quantities are also provided, such as Gauss-type quadrature weights and nodes and, hence, also zeros of orthogonal polynomials.

Occasionally (e.g., in Gautschi [1984a], Gautschi and Milovanović [1986], Frontini et al. [1987], and Gautschi [1993a; 1993b]), one needs to deal with indefinite (i.e., sign-changing) measures $d\lambda$. The positivity of the β_k is then no longer guaranteed, indeed not even the existence of all orthogonal polynomials. Nevertheless, our methods can still be formally applied, albeit at the risk of possible breakdowns or instabilities.

There are basically four methods used here to generate recursion coefficients: (1) *Methods based on explicit formulas.* These relate to classical orthogonal polynomials and are implemented in the routine **recur** of Section 2. (2) *Methods based on moment information.* These are dealt with in Section 3 and are represented by a single routine, **cheb**. Its origin can be traced back to work of Chebyshev on discrete least squares approximation. (3) *Bootstrap methods based on inner product formulas for the coefficients, and orthogonal reduction methods.* We have attributed the idea for the former method to Stieltjes, and referred to it in Gautschi [1982a] as the *Stieltjes procedure.* The prototype is the routine **sti** in Section 4, applicable for discrete orthogonal polynomials. An alternative routine is **lancz**, which accomplishes the same purpose, but uses the *method of Lanczos.* Either of these routines can be used in **mcdis**, which applies to continuous as well as to mixed-type orthogonal polynomials. In contrast to all previous routines, **mcdis** uses a discretization process and, thus, furnishes only approximate answers whose accuracies can be controlled by the user. The routine, however, is by far the most sophisticated and flexible routine in this package, one that requires, or can greatly benefit from, ingenuity of the user. The same kind of discretization is also applicable to moment-related methods, yielding the routine **mccheb**. (4) *Modification algorithms.* These are routines generating recursion coefficients for measures modified by a rational factor, utilizing the recursion coefficients of the original measure, which are assumed to be known. They can be thought of as algorithmic implementations of the Christoffel, or generalized Christoffel, theorem and are incorporated in the routines **chri** and **gchri** of Section 5. An important application of all of these routines is made in Section 6, where routines are provided that generate the weights and nodes of quadrature rules of Gauss, Gauss–Radau, and Gauss–Lobatto types.

Each routine has a single-precision and double-precision version with similar names, except for the prefix **d** in double-precision procedures. The latter are generally a straightforward translation of the former. An exception

is the routine **dlga** used in **drecur** for computing the logarithm of the gamma function, which employs a different method than the single-precision companion routine **alga**.

All routines of the package have been checked for ANSI conformance and tested on two computers: the Cyber 205 and a Sun 4/670 MP workstation. The former has machine precisions $\epsilon^s \approx 7.11 \times 10^{-15}$, $\epsilon^d \approx 5.05 \times 10^{-29}$ in single and double precision, respectively, while the latter has $\epsilon^s \approx 5.96 \times 10^{-8}$, $\epsilon^d \approx 1.11 \times 10^{-16}$. The Cyber 205 has a large floating-point exponent range, extending from approximately -8617 to $+8645$ in single as well as in double precision, whereas the Sun 4/670 has the rather limited exponent range $[-38, 38]$ in single precision, but a larger range $[-308, 308]$ in double precision. All output cited relates to work on the Cyber 205.

The package is organized as follows: Section 0 contains (slightly amended) *netlib* routines, namely, **r1mach** and **d1mach**, providing basic machine constants for a variety of computers. Section 1 contains all of the driver routines, named **test1**, **test2**, etc., which are used (and described in the body of this paper) to test the subroutines of the package. The complete output of each test is listed immediately after the driver. Sections 2–6 constitute the core of the package: The single- and double-precision subroutines described in the equally numbered sections of this paper. All single-precision routines are provided with comments and instructions for their use. These, of course, apply to the double-precision routines as well.

2. CLASSICAL WEIGHT FUNCTIONS

Among the most frequently used orthogonal polynomials are the Jacobi polynomials, generalized Laguerre polynomials, and Hermite polynomials, supported, respectively, on a finite interval, half-infinite interval, and the whole real line. The respective weight functions are

$$w(t) = w^{(\alpha,\beta)}(t) = (1-t)^{\alpha}(1+t)^{\beta} \quad \text{on } (-1,1),\ \alpha > -1,\ \beta > -1\text{: Jacobi;} \tag{2.1}$$

$$w(t) = w^{(\alpha)}(t) = t^{\alpha}e^{-t} \quad \text{on } (0,\infty),\ \alpha > -1\text{: Generalized Laguerre;} \tag{2.2}$$

$$w(t) = e^{-t^2} \quad \text{on } (-\infty,\infty)\text{: Hermite.} \tag{2.3}$$

Special cases of the Jacobi polynomials are the Legendre polynomials ($\alpha = \beta = 0$); the Chebyshev polynomials of the first ($\alpha = \beta = -\frac{1}{2}$), second ($\alpha = \beta = \frac{1}{2}$), third ($\alpha = -\beta = -\frac{1}{2}$), and fourth ($\alpha = -\beta = \frac{1}{2}$) kinds; and the Gegenbauer polynomials ($\alpha = \beta = \lambda - \frac{1}{2}$). The Laguerre polynomials are the special case $\alpha = 0$ of the generalized Laguerre polynomials.

For each of these polynomials, the corresponding recursion coefficients $\alpha_k = \alpha_k(w)$, $\beta_k = \beta_k(w)$ are explicitly known (see, e.g., Chihara [1978, pp. 217–221] and are generated in single precision by the routine **recur**. Its calling sequence is

recur(n, ipoly, al, be, a, b, ierr).

On entry,

n is the number of recursion coefficients desired; type integer.

ipoly is an integer identifying the polynomial as follows:

1 = Legendre polynomial on $(-1, 1)$;
2 = Legendre polynomial on $(0, 1)$;
3 = Chebyshev polynomial of the first kind;
4 = Chebyshev polynomial of the second kind;
5 = Chebyshev polynomial of the third kind;
6 = Jacobi polynomial with parameters **al**, **be**;
7 = generalized Laguerre polynomial with parameter **al**; and
8 = Hermite polynomial.

al, be are the input parameters α, β for Jacobi and generalized Laguerre polynomials; type real; they are only used if **ipoly** = 6 or 7, and in the latter case, only **al** is used.

On return,

a, b are real arrays of dimension **n** with $\mathbf{a}(k), \mathbf{b}(k)$ containing the coefficients $\alpha_{k-1}, \beta_{k-1}$, respectively, $k = 1, 2, \ldots, \mathbf{n}$.

ierr is an error flag, where

ierr = 0 on normal return,
ierr = 1 if either **al** or **be** is out of range when **ipoly** = 6 or **ipoly** = 7,
ierr = 2 if there is potential overflow in the evaluation of β_0 when **ipoly** = 6 or **ipoly** = 7; in this case, β_0 is set equal to the largest machine-representable number,
ierr = 3 if **n** is out of range, and
ierr = 4 if **ipoly** is not one of the admissible integers.

No provision has been made for Chebyshev polynomials of the fourth kind, since their recursion coefficients are obtained from those for the third-kind Chebyshev polynomials simply by changing the sign of the α's (and leaving the β's unchanged).

The corresponding double-precision routine is **drecur**; it has the same calling sequence, except for real data types now being double precision.

In the cases of Jacobi polynomials (**ipoly** = 6) and generalized Laguerre polynomials (**ipoly** = 7), the recursion coefficient β_0 (and only this one) involves the gamma function Γ. Accordingly, a function routine, **alga**, is provided that computes the logarithm $\ln \Gamma$ of the gamma function, and a separate routine, **gamma**, computing the gamma function by exponentiating its logarithm. Their calling sequences are

function alga(x)
function gamma(x, ierr),

where **ierr** is an output variable set equal to 2 or 0 depending on whether the gamma function does, or does not, overflow, respectively. The corresponding

double-precision routines have the names **dlga** and **dgamma**. All of these routines require machine-dependent constants for reasons explained below.

The routine **alga** is based on a rational approximation valid on the interval $[\frac{1}{2}, \frac{3}{2}]$. Outside this interval, the argument x is written as

$$x = x_e + m,$$

where

$$x_e = \begin{cases} x - \lfloor x \rfloor + 1 & \text{if } x - \lfloor x \rfloor \le \frac{1}{2}, \\ x - \lfloor x \rfloor & \text{otherwise} \end{cases}$$

is in the interval $(\frac{1}{2}, \frac{3}{2}]$ and where $m \ge -1$ is an integer. If $m = -1$ (i.e., $0 < x \le \frac{1}{2}$), then $\ln \Gamma(x) = \ln \Gamma(x_e) - \ln x$, while for $m > 0$, one computes $\ln \Gamma(x) = \ln \Gamma(x_e) + \ln p$, where $p = x_e(x_e + 1)\cdots(x_e + m - 1)$. If m is so large, say, $m \ge m_0$, that the product p would overflow, then $\ln p$ is computed (at a price!) as $\ln p = \ln x_e + \ln(x_e + 1) + \cdots + \ln(x_e + m - 1)$. It is here where a machine-dependent integer is required, namely, m_0 = smallest integer m such that $1 \cdot 3 \cdot 5 \cdots (2m + 1)/2^m$ is greater than or equal to the largest machine-representable number, R. By Stirling's formula, the integer m_0 is seen to be the smallest integer m satisfying $((m + 1)/e)\ln((m + 1)/e) \ge (\ln R - \frac{1}{2}\ln 8)/e$, hence, equal to $\lfloor e \cdot t((\ln R - \frac{1}{2}\ln 8)/e)\rfloor$, where $t(y)$ is the inverse function of $y = t \ln t$. For our purposes, the low-accuracy approximation of $t(y)$, given in Gautschi [1967b, pp. 51–52], and implemented in the routine **t**, is adequate.

The rational approximation chosen on $[\frac{1}{2}, \frac{3}{2}]$ is one due to W. J. Cody and K. E. Hillstrom, namely, the one labeled $n = 7$ in Table II of Cody and Hillstrom [1967]. It is designed to yield about 16 correct decimal digits (cf. Table I of Cody and Hillstrom [1967]), but because of numerical cancellation furnishes only about 13–14 correct decimal digits.

Since rational approximations for $\ln \Gamma$ having sufficient accuracies for double-precision computation do not seem to be available in the literature, we use a different approach for the routine **dlga**, namely, the asymptotic approximation (cf. eq. 6.1.42 of Abramowitz and Stegun [1964], where the constants B_{2m} are Bernoulli numbers)

$$\ln \Gamma(y) = (y - \tfrac{1}{2})\ln y - y + \tfrac{1}{2}\ln(2\pi) + \sum_{m=1}^{n} \frac{B_{2m}}{2m(2m-1)} y^{-(2m-1)} + R_n(y) \tag{2.4}$$

for values of $y > 0$ large enough to have

$$|R_n(y)| \le \frac{1}{2} 10^{-d}, \tag{2.5}$$

where d is the number of decimal digits carried in double-precision arithmetic, another machine-dependent real number. If (2.5) holds for $y \ge y_0$ and if $x \ge y_0$, we compute $\ln \Gamma(x)$ from the asymptotic expression (2.4) (where

$y = x$ and the remainder term is neglected). Otherwise, we let k_0 be the smallest positive integer k such that $x + k \geq y_0$, and use

$$\ln \Gamma(x) = \ln \Gamma(x + k_0) - \ln(x(x + 1) \cdots (x + k_0 - 1)), \qquad (2.6)$$

where the first term on the right is computed from (2.4) (with $y = x + k_0$). Since, for $y > 0$,

$$|R_n(y)| \leq \frac{|B_{2n+2}|}{(2n + 2)(2n + 1)} y^{-(2n+1)}$$

(cf. Abramowitz and Stegun [1964, eq. 6.1.42]), the inequality (2.5) is satisfied if

$$y \geq \exp\left\{ \frac{1}{2n + 1} \left[d \ln 10 + \ln \frac{2|B_{2n+2}|}{(2n + 2)(2n + 1)} \right] \right\}. \qquad (2.7)$$

In our routine **dlga**, we have selected $n = 9$. For double-precision accuracy on the Cyber 205, we have $d \approx 28.3$, for which (2.7) then gives $y \geq \exp\{.121188 \cdots d + .053905 \cdots\} \approx 32.6$.

For single-precision calculation, we selected the method of rational approximation, rather than the asymptotic formula (2.4) and (2.6), since we found that the former is generally more accurate, by a factor, on the average, of about 20 and as large as 300. Neither method yields full machine accuracy. The former, as already mentioned, loses accuracy in the evaluation of the approximation. The latter suffers loss of accuracy because of cancellation occurring in (2.6), which typically amounts to a loss of 2–5 significant decimal digits in the gamma function itself.

Since these errors affect only the coefficient β_0 (and only if **ipoly** = 6 or 7), they are of no consequence unless the output of the routine **recur** serves as input to another routine, such as **gauss** (cf. Section 6), which makes essential use of β_0. In this case, for maximum single-precision accuracy, it is recommended that β_0 be first obtained in double precision by means of **drecur** with **n** = 1 and then converted to single precision.

3. MOMENT-RELATED METHODS

It is a well-known fact that the first n recursion coefficients $\alpha_k(d\lambda)$, $\beta_k(d\lambda)$, $k = 0, 1, \ldots, n - 1$ (cf. (1.3)), are uniquely determined by the first $2n$ moments μ_k of the measure $d\lambda$,

$$\mu_k = \mu_k(d\lambda) = \int_{\mathbb{R}} t^k \, d\lambda(t), \qquad k = 0, 1, 2, \ldots, 2n - 1. \qquad (3.1)$$

Formulas are known, for example, that express the α's and β's in terms of Hankel determinants in these moments. The trouble is that these formulas become increasingly sensitive to small errors as n becomes large. There is an inherent reason for this: The underlying (nonlinear) map $K_n: \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ has

a condition number, cond K_n, that grows exponentially with n (cf. Gautschi [1982a, sect. 3.2]). Any method that attempts to compute the desired coefficients from the moments in (3.1), therefore, is doomed to fail, unless n is quite small or extended precision is being employed. That goes, in particular, for an otherwise elegant method due to Chebyshev (who developed it for the case of discrete measures $d\lambda$) that generates the α's and β's directly from the moments (3.1), bypassing determinants altogether (cf. Chebyshev [1859] and Gautschi [1982a, sect. 2.3]).

Variants of Chebyshev's algorithm with more satisfactory stability properties have been developed by Sack and Donovan [1972] and by Wheeler [1974] (independently of Chebyshev's work). The idea is to forgo the moments (3.1) as input data and instead depart from so-called *modified moments*. These are defined by replacing the power t^k in (3.1) by an appropriate polynomial $p_k(t)$ of degree k,

$$\nu_k = \nu_k(d\lambda) = \int_{\mathbb{R}} p_k(t)\, d\lambda(t), \qquad k = 0, 1, 2, \ldots, 2n-1. \tag{3.2}$$

For example, p_k could be one of the classical orthogonal polynomials. More generally, we shall assume that $\{p_k\}$ are monic polynomials satisfying a three-term recurrence relation similar to the one in (1.3),

$$\begin{aligned} p_{k+1}(t) &= (t - a_k)p_k(t) - b_k p_{k-1}(t), \qquad k = 0, 1, 2, \ldots, \\ p_0(t) &= 1, \qquad p_{-1}(t) = 0, \end{aligned} \tag{3.3}$$

with coefficients $a_k \in \mathbb{R}$, $b_k \geq 0$ that are known. (In the special case $a_k = 0$, $b_k = 0$, we are led back to powers and ordinary moments.) There now exists an algorithm, called the *modified Chebyshev algorithm* in Gautschi [1982a, sect. 2.4], which takes the $2n$ modified moments in (3.2) and the $2n-1$ coefficients $\{a_k\}_{k=0}^{2n-2}$, $\{b_k\}_{k=0}^{2n-2}$ in (3.3), and from them generates the n desired coefficients $\alpha_k(d\lambda)$, $\beta_k(d\lambda)$, $k = 0, 1, \ldots, n-1$. It generalizes Chebyshev's algorithm, which can be recovered (if need be) by putting $a_k = b_k = 0$. The modified Chebyshev algorithm is embodied in the subroutine **cheb**, which has the calling sequence

```
cheb(n, a, b, fnu, alpha, beta, s, ierr, s0, s1, s2)
dimension a(*), b(*), fnu(*), alpha(n), beta(n), s(n),
  s0(*), s1(*), s2(*)
```

On entry,

n is the number of recursion coefficients desired; type integer.

a, b are arrays of dimension $2 \times \mathbf{n} - 1$ holding the coefficients $\mathbf{a}(k) = a_{k-1}$, $\mathbf{b}(k) = b_{k-1}$, $k = 1, 2, \ldots, 2n-1$.

fnu is an array of dimension $2 \times \mathbf{n}$ holding the modified moments $\mathbf{fnu}(k) = \nu_{k-1}$, $k = 1, 2, \ldots, 2 \times \mathbf{n}$.

On return,

alpha, beta are arrays of dimension **n** containing the desired recursion coefficients **alpha**$(k) = \alpha_{k-1}$, **beta**$(k) = \beta_{k-1}$, $k = 1, 2, \ldots,$ **n**.

s is an array of dimension **n** containing the numbers **s**$(k) = \int_{\mathbb{R}} \pi_{k-1}^2 \, d\lambda$, $k = 1, 2, \ldots,$ **n**.

ierr is an error flag, equal to 0 on normal return, equal to 1 if $|\nu_0|$ is less than the machine zero, equal to 2 if **n** is out of range, equal to $-(k-1)$ if **s**(k), $k = 1, 2, \ldots,$ **n**, is about to underflow, and equal to $+(k-1)$ if it is about to overflow.

The arrays **s0**, **s1**, **s2** of dimension $2 \times$ **n** are needed for working space.

There is again a map $K_n: \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ underlying the modified Chebyshev algorithm, namely, the map taking the $2n$ modified moments into the n pairs of recursion coefficients. The condition of the map K_n (actually of a somewhat different, but closely related, map) has been studied in [Gautschi 1982a, sect. 3.3; 1986a] in the important case where the polynomials p_k defining the modified moments are themselves orthogonal polynomials, $p_k(\cdot) = p_k(\cdot; d\mu)$, with respect to a measure $d\mu$ (e.g., one of the classical ones) for which the recursion coefficients α_k, b_k are known. The upshot of the analysis then is that the condition of K_n is characterized by a certain positive polynomial $g_n(\cdot; d\lambda)$ of degree $4n - 2$, depending only on the target measure $d\lambda$, in the sense that

$$\operatorname{cond} K_n = \int_{\mathbb{R}} g_n(t; d\lambda) \, d\mu(t). \tag{3.4}$$

Thus, the numerical stability of the modified Chebyshev algorithm is determined by the magnitude of g_n on the support of $d\mu$.

The occurrence of underflow (overflow) in the computation of the α's and β's, especially on computers with limited exponent range, can often be avoided by multiplying all modified moments by a sufficiently large (small) scaling factor before entering the routine. On exit, the coefficient β_0 (and only this one!) then has to be divided by the same scaling factor. (There may occur harmless underflow of auxiliary quantities in the routine **cheb**, which is difficult to avoid since some of these quantities actually are expected to be zero.)

Example 3.1 $d\lambda_\omega(t) = [(1 - \omega^2 t^2)(1 - t^2)]^{-1/2} \, dt$ on $(-1, 1)$, $0 \le \omega < 1$. This example is of some historical interest, in that it has already been considered by Christoffel [1877, example 6]; see also Rees [1945]. Computationally, the example is of interest as there are empirical reasons to believe that for the choice $d\mu(t) = (1 - t^2)^{-1/2} \, dt$ on $(-1, 1)$, which appears rather natural, the modified Chebyshev algorithm is exceptionally stable, uniformly in n, in the sense that in (3.4) one has $g_n \le 1$ on $\operatorname{supp} d\mu$ for all n (cf. Gautschi [1984b, example 5.2]). With the above choice of $d\mu$, the polynomials p_k are clearly the Chebyshev polynomials of the first kind, $p_0 = T_0$, $p_k =$

$2^{-(k-1)}T_k$, $k \geq 1$, and the modified moments are given by

$$\nu_0 = \int_{-1}^{1} d\lambda_\omega(t), \qquad \nu_k = \frac{1}{2^{k-1}} \int_{-1}^{1} T_k(t)\, d\lambda_\omega(t), \qquad k = 1,2,3,\ldots. \tag{3.5}$$

They are expressible in terms of the Fourier coefficients $C_r(\omega^2)$ in

$$(1 - \omega^2 \sin^2\theta)^{-1/2} = C_0(\omega^2) + 2\sum_{r=1}^{\infty} C_r(\omega^2)\cos 2r\theta \tag{3.6}$$

by means of (cf. Gautschi [1982a, example 3.3])

$$\left.\begin{aligned} \nu_0 &= \pi C_0(\omega^2), \\ \nu_{2m} &= \frac{(-1)^m \pi}{2^{2m-1}} C_m(\omega^2) \\ \nu_{2m-1} &= 0 \end{aligned}\right\} \qquad m = 1,2,3,\ldots. \tag{3.7}$$

The Fourier coefficients $\{C_r(\omega^2)\}$, in turn, can be accurately computed as the minimal solution of a certain three-term recurrence relation (see Gautschi [1982a, pp. 310–311]).

The ordinary moments

$$\mu_0 = \nu_0, \qquad \mu_k = \int_{-1}^{1} t^k\, d\lambda_\omega(t), \qquad k = 1,2,3,\ldots, \tag{3.8}$$

likewise can be expressed in terms of the Fourier coefficients $C_r(\omega^2)$ by writing t^{2m} as a linear combination of Chebyshev polynomials $T_0, T_2, \ldots, T_{2m}$ (cf. Luke [1975, Eq. 22, p. 454]). The result is

$$\left.\begin{aligned} \mu_{2m} &= \frac{(-1)^m \pi}{2^{2m-1}} \sum_{r=0}^{m} (-1)^r \gamma_r^{(m)} C_{m-r}(\omega^2) \\ \mu_{2m-1} &= 0 \end{aligned}\right\} \qquad m = 1,2,3,\ldots, \tag{3.9}$$

where

$$\begin{aligned} \gamma_0^{(m)} &= 1, \\ \gamma_r^{(m)} &= \frac{2m+1-r}{r}\gamma_{r-1}^{(m)}, \qquad r = 1,2,\ldots,m-1, \\ \gamma_m^{(m)} &= \frac{m+1}{2m}\gamma_{m-1}^{(m)}. \end{aligned} \tag{3.10}$$

The driver routine **test1** (in Section 1 of the package) generates for $\omega^2 = .1(.2).9, .99, .999$ the first n recurrence coefficients $\beta_k(d\lambda_\omega)$ (all $\alpha_k = 0$), both in single and double precision, using modified moments if **modmom** = **.true.** and ordinary moments otherwise. In the former case, $n = 80$; in the latter, $n = 20$. It prints the double-precision values of β_k, together with the relative errors of the single-precision values (computed as the difference of

Table I. Selected Output from **test1** in the Case of Modified Moments

ω^2	k	β_k^{double}	err β_k^{single}
.100	0	3.22488269744043879645983272 5	1.433(−14)
	1	0.50658408063826844751584957 27	1.187(−14)
	5	0.24999999538900319018810282 67	1.109(−14)
	11	0.24999999999999999996365048540	1.454(−18)
	18	0.25000000000000000000000000000	0.000
.500	0	3.70814935460274383686770069 4	9.005(−15)
	1	0.54305341895553637462503337 73	2.431(−14)
	8	0.24999998464317232960837794 80	4.109(−15)
	20	0.24999999999999999978894635584	8.442(−18)
	35	0.25000000000000000000000000000	0.000
.900	0	5.15618422669634637640514154 3	6.950(−15)
	1	0.63497316614524587116224926 13	7.920(−15)
	19	0.24999999569259500946295028 30	1.820(−14)
	43	0.24999999999999998282104100896	6.872(−16)
	79	0.24999999999999999999999999962	1.525(−26)
.999	0	9.68226512110059406067820825 7	1.194(−13)
	1	0.79378214213851769655317195 71	6.311(−14)
	19	0.24990638943982092000474525 37	1.026(−14)
	43	0.24999558226336808258597500 68	8.282(−15)
	79	0.24999984176881578761530692 11	1.548(−15)

the double-precision and single-precision values divided by the double-precision value). In **test1**, as well as in all subsequent drivers, not all error flags are interrogated for possible malfunction. The user is urged, however, to do so as a matter of principle.

The routine

fmm(n, eps, modmom, om2, fnu, ierr, f, f0, rr)

used by the driver computes the first $2 \times \mathbf{n}$ modified (ordinary) moments for $\omega^2 = \mathbf{om2}$, to a relative accuracy **eps** if **modmom** = **.true.** (**.false.**). The results are stored in the array **fnu**. The arrays **f**, **f0**, and **rr** are internal working arrays of dimension **n**, and **ierr** is an error flag. On normal return, **ierr** = 0; otherwise, **ierr** = 1, indicating lack of convergence (within a prescribed number of iterations) of the backward recurrence algorithm for computing the minimal solution $\{C_r(\omega^2)\}$. The latter is likely to occur if ω^2 is too close to 1. The routine **fmm**, as well as its double-precision version **dmm**, is listed immediately after the routine **test1**.

Table I shows selected results from the output of **test1**, when **modmom** = **.true.** (Complete results are given in the package immediately after **test1**.) The values for $k = 0$ are expressible in terms of the complete elliptic integral, $\beta_0 = 2K(\omega^2)$, and were checked, where possible, against the 16S-values in Abramowitz and Stegun [1964, Table 17.1]. In all cases, there was agreement to all 16 digits. The largest relative error observed was 2.43×10^{-13} for $\omega^2 = .999$ and $k = 2$. When $\omega^2 \leq .99$, the error was always less than 2.64×10^{-14}, which confirms the extreme stability of the modified Chebyshev algo-

Table II. Selected Output from **test1** in the Case of Ordinary Moments

ω^2	k	err β_k	ω^2	k	err β_k
.100	1	1.187(−14)	.900	1	3 270(−15)
	7	2.603(−10)		7	4.819(−10)
	13	9.663(−6)		13	1.841(−5)
	19	4.251(−1)		19	6.272(−1)
.500	1	2.431(−14)	.999	1	6.311(−14)
	7	5.571(−10)		7	1.745(−9)
	13	9.307(−6)		13	8.589(−5)
	19	5.798(−1)		19	4.808(0)

rithm in this example. It can be seen (as was to be expected) that for ω^2 not too close to 1, the coefficients converge rapidly to $\frac{1}{4}$.

In contrast, Table II shows selected results (for complete results, see the package) in the case of ordinary moments (**modmom** = **.false.**) and demonstrates the severe instability of the Chebyshev algorithm. Note that the moments themselves are all accurate to essentially machine precision, as has been verified by additional computations.

The next example deals with another weight function for which the modified Chebyshev algorithm performs rather well.

Example 3.2 $d\lambda_\sigma(t) = t^\sigma \ln(1/t)\,dt$ on $(0, 1]$, $\sigma > -1$.

What is nice about this example is that both modified and ordinary moments of $d\lambda_\sigma$ are known in closed form. The latter are obviously given by

$$\mu_k(d\lambda_\sigma) = \frac{1}{(\sigma + 1 + k)^2} \qquad k = 0, 1, 2, \ldots, \tag{3.11}$$

whereas the former, relative to shifted monic Legendre polynomials (**ipoly** = 2 in **recur**), are (cf. Gautschi [1979])

$$\frac{(2k)!}{k!^2}\nu_k(d\lambda_\sigma) = \begin{cases} (-1)^{k-\sigma}\dfrac{\sigma!^2(k-\sigma-1)!}{(k+\sigma+1)!}, & 0 \le \sigma < k, \\ & \sigma \text{ an integer}, \\ \dfrac{1}{\sigma+1}\left\{\dfrac{1}{\sigma+1} + \displaystyle\sum_{r=1}^{k}\left(\dfrac{1}{\sigma+1+r} - \dfrac{1}{\sigma+1-r}\right)\right\} & \\ \cdot \displaystyle\prod_{r=1}^{k}\dfrac{\sigma+1-r}{\sigma+1+r}, & \text{otherwise}. \end{cases} \tag{3.12}$$

The routines **fmm** and **dmm** appended to **test2** in Section 1 of the package, similarly as the corresponding routines in Example 3.1, generate the first

Table III. Selected Output from **test2** in the Case of Modified Moments

σ	k	α_k	β_k
− .5	0	.111111111111111111111111	4.000000000000000000000000
	12	.499497191609463856624220 2	0.062312770828774884775638 86
	24	.499866291232421894380159 2	0.062453725573422426004572 26
	48	.499965263548544580066196 9	0.062488507177486847424336 18
	99	.499991618402435627167078 9	0.062497338230518216369371 56
0	0	.250000000000000000000000	1.000000000000000000000000
	12	.499283180215736131027262 5	0.062383568359535711235603 30
	24	.499806283948614639850153 2	0.062471000844691110016391 28
	48	.499949408379702387935642 4	0.062492812681109674623738 89
	99	.499987799201590328304791 9	0.062498326706169259262048 96
.5	0	.360000000000000000000000	0.444444444444444444444444
	12	.499375573291755564420326 7	0.062370827382807526119608 87
	24	.499832449770639448872272 5	0.062465810119454968835430 89
	48	.499956727522377172779152 1	0.062491153327110271766959 32
	99	.499989693184178978188767 4	0.062497872512816829738256 35

$2 \times \mathbf{n}$ modified moments $\nu_0, \nu_1, \ldots, \nu_{2n-1}$ if **modmom** = **.true.** and the first $2 \times \mathbf{n}$ ordinary moments otherwise. The calling sequence of **fmm** is

fmm(n, modmom, intexp, sigma, fnu).

The logical variable **intexp** is to be set **.true.** if σ is an integer and **.false.** otherwise. In either case, the input variable **sigma** is assumed to be of type **real**.

The routine **test2** generates the first **n** recursion coefficients $\alpha_k(d\lambda_\sigma)$, $\beta_k(d\lambda_\sigma)$ in single and double precision for $\sigma = -\frac{1}{2}, 0, \frac{1}{2}$, where $\mathbf{n} = 100$ for the modified Chebyshev algorithm (**modmom** = **.true.**) and $\mathbf{n} = 12$ for the classical Chebyshev algorithm (**modmom** = **.false.**). Selected double-precision results to 25 significant digits, when modified moments are used, are shown in Table III. (The complete results are given in the package after **test2**.)

The largest relative errors observed, over all $k = 0, 1, \ldots, 99$, were, respectively, 6.211×10^{-11}, 2.237×10^{-12}, and 1.370×10^{-12} for the α's and 1.235×10^{-10}, 4.446×10^{-12}, and 2.724×10^{-12} for the β's, attained consistently at $k = 99$. The accuracy achieved is slightly less than in Example 3.1, for reasons explained in Gautschi [1984b, Example 5.3].

The complete results for $\sigma = -\frac{1}{2}$ are also available in Gautschi [1991b, Appendix, Table 1]. (They differ occasionally by one unit in the last decimal place from those produced here, probably because of a slightly different computation of the modified moments.) The results for $\sigma = 0$ can be checked up to $k = 15$ against the 30S-values given in Stroud and Secrest [1966, p. 92], and for $16 \leq k \leq 19$ against 12S-values in Danloy [1973, Table 3]. There is complete agreement to all 25 digits in the former case and agreement to 12 digits in the latter, although there are occasional end-figure discrepancies of one unit. These are believed to be due to rounding errors committed in Danloy [1973], since similar discrepancies occur also in the range $k \leq 15$. We

Table IV. Selected Output from **test2** in the Case of Ordinary Moments

k	σ	err α_k	err β_k	σ	err α_k	err β_k	σ	err α_k	err β_k
2	−.5	1.8(−13)	7.7(−14)	0	4.2(−13)	7.6(−13)	.5	1.6(−12)	2.6(−13)
5		2.2(−9)	1.2(−9)		4.2(−9)	1.2(−10)		1.3(−8)	6.6(−9)
8		1.1(−5)	5.5(−6)		4.3(−6)	3.8(−6)		6.0(−5)	5.2(−6)
11		2.5(−1)	1.7(−1)		1.3(0)	3.2(−1)		2.2(0)	4.7(−1)

do not know of any tables for $\sigma = \frac{1}{2}$, but a test is given in Section 5, Example 5.1.

The use of ordinary moments (**modmom** = **.false.**) produces predictably worse results, the relative errors of which are shown in Table IV.

4. STIELTJES, ORTHOGONAL REDUCTION, AND DISCRETIZATION PROCEDURES

4.1 The Stieltjes Procedure

It is well known that the coefficients $\alpha_k(d\lambda)$, $\beta_k(d\lambda)$ in the basic recurrence relation (1.3) can be expressed in terms of the orthogonal polynomials (1.2) and the inner product (1.1) as follows:

$$\alpha_k(d\lambda) = \frac{(t\pi_k, \pi_k)}{(\pi_k, \pi_k)}, \qquad k \geq 0;$$

$$\beta_0(d\lambda) = (\pi_0, \pi_0), \qquad \beta_k(d\lambda) = \frac{(\pi_k, \pi_k)}{(\pi_{k-1}, \pi_{k-1})}, \qquad k \geq 1. \tag{4.1}$$

Provided that the inner product can be readily calculated, (4.1) suggests the following "bootstrap" procedure: Compute α_0 and β_0 by the first relations in (4.1) for $k = 0$. Then use the recurrence relation (1.3) for $k = 0$ to obtain π_1. With π_0 and π_1 known, apply (4.1) for $k = 1$ to get α_1, β_1, then again apply (1.3) to obtain π_2, and so on. In this way, alternating between (4.1) and (1.3), we can bootstrap ourselves up to as many of the coefficients α_k, β_k as are desired. We attributed this procedure to Stieltjes and called it *Stieltjes's procedure* in Gautschi [1982a].

In the case of discrete orthogonal polynomials, that is, for inner products of the form

$$(u, v) = \sum_{k=1}^{N} w_k u(x_k) v(x_k), \qquad w_k > 0, \tag{4.2}$$

Stieltjes's procedure is easily implemented; the resulting routine is called **sti** and has the calling sequence

sti(n, ncap, x, w, alpha, beta, ierr, p0, p1, p2).

On entry,

n is the number of recursion coefficients desired; type integer.

ncap is the number of terms, N, in the discrete inner product; type integer.

x, w are arrays of dimension **ncap** holding the abscissas $\mathbf{x}(k) = x_k$ and weights $\mathbf{w}(k) = w_k, k = 1, 2, \ldots,$ **ncap**, of the discrete inner product.

On return,

alpha, beta are arrays of dimension **n** containing the desired recursion coefficients $\mathbf{alpha}(k) = \alpha_{k-1}$, $\mathbf{beta}(k) = \beta_{k-1}$, $k = 1, 2, \ldots,$ **n**.

ierr is an error flag having the value 0 on normal return and the value 1 if **n** is not in the proper range $1 \le n \le N$; if during the computation of a recursion coefficient with index k there is impending underflow or overflow, **ierr** will have the value $-k$ in case of underflow and the value $+k$ in case of overflow. (No error flag is set in case of harmless underflow.)

The arrays **p0, p1, p2** are working arrays of dimension **ncap**. The double-precision routine has the name **dsti**.

Occurrence of underflow (overflow) can be forestalled by multiplying all weights w_k by a sufficiently large (small) scaling factor prior to entering the routine. Upon return, the coefficient β_0 will then have to be readjusted by dividing it by the same scaling factor.

4.2 Orthogonal Reduction Method

Another approach to producing the recursion coefficients α_k, β_k from the quantities x_k, w_k defining the inner product (4.2) is based on the observation (cf. Boley and Golub [1987] and Gautschi [1991d, sect. 7]) that the symmetric tridiagonal matrix of order $N + 1$,

$$J(d\lambda_N) = \begin{bmatrix} 1 & \sqrt{\beta_0} & & & 0 \\ \sqrt{\beta_0} & \alpha_0 & \sqrt{\beta_1} & & \\ & \sqrt{\beta_1} & \alpha_1 & \ddots & \\ & & \ddots & \ddots & \sqrt{\beta_{N-1}} \\ 0 & & & \sqrt{\beta_{N-1}} & \alpha_{N-1} \end{bmatrix} \tag{4.3}$$

(the "extended Jacobi matrix" for the discrete measure $d\lambda_N$ implied in (4.2)), is orthogonally similar to the matrix

$$\begin{bmatrix} 1 & \sqrt{w}^T \\ \sqrt{w} & D_x \end{bmatrix}, \qquad \sqrt{w} = \begin{bmatrix} \sqrt{w_1} \\ \vdots \\ \sqrt{w_N} \end{bmatrix}, \qquad D_x = \begin{bmatrix} x_1 & & 0 \\ & \ddots & \\ 0 & & x_N \end{bmatrix}. \tag{4.4}$$

Hence, the desired matrix $J(d\lambda_N)$ can be obtained by applying Lanczos's algorithm to the matrix (4.4). This is implemented in the routine

lancz(n, ncap, x, w, alpha, beta, ierr, p0, p1),

which uses a judiciously constructed sequence of Givens transformations to accomplish the orthogonal similarity transformation (cf. Rutishauser [1963], de Boor and Golub [1978], Gragg and Harrod [1984], and Boley and Golub [1987]; the routine **lancz** is adapted from the routine RKPW in Gragg and Harrod [1984, p. 328]). The input and output parameters of the routine **lancz** have the same meaning as in the routine **sti**, except that **ierr** can only have the value 0 or 1, while **p0, p1** are again working arrays of dimension **ncap**. The double-precision version of the routine is named **dlancz**.

The routine **lancz** is generally superior to the routine **sti**: The procedure used in **sti** may develop numerical instability from some point on and therefore give inaccurate results for larger values of **n**. It furthermore is subject to underflow and overflow conditions. None of these shortcomings is shared by the routine **lancz**. On the other hand, there are cases where **sti** does better than **lancz** (cf. Example 4.5).

We illustrate the phenomenon of instability (which is explained in Gautschi [1993c]) in the case of the "discrete Chebyshev" polynomials.

Example 4.1 The inner product (4.2) with $x_k = -1 + 2(k-1)/(N-1)$, $w_k = 2/N$, $k = 1, 2, \ldots, N$.

This generates discrete analogues of the Legendre polynomials, which they indeed approach as $N \to \infty$. The recursion coefficients are explicitly known:

$$\alpha_k = 0, \qquad k = 0, 1, \ldots, N-1;$$

$$\beta_0 = 2, \qquad \beta_k = \left(1 + \frac{1}{N-1}\right)^2 \left(1 - \left(\frac{k}{N}\right)^2\right) \left(4 - \frac{1}{k^2}\right)^{-1},$$

$$k = 1, 2, \ldots, N-1. \tag{4.5}$$

To find out how well the routines **sti** and **lancz** generate them (in single precision), when $N = 40$, 80, 160, and 320, we wrote the driver **test3**, which computes the respective absolute errors for the α's and relative errors for the β's.

Selected results for Stieltjes's algorithm are shown in Table V. The gradual deterioration, after some point (depending on N), is clearly visible. Lanczos's method, in contrast, preserves essentially full accuracy; the largest error in the α's is 1.42(−13), 2.27(−13), 4.83(−13), and 8.74(−13) for $N = 40$, 80, 160, and 320, respectively, and 3.38(−13), 6.63(−13), 2.17(−12), and 5.76(−12) for the β's.

4.3 Multiple-Component Discretization Procedure

We now assume a measure $d\lambda$ of the form

$$d\lambda(t) = w(t)\,dt + \sum_{j=1}^{p} y_j \delta(t - x_j)\,dt, \qquad p \geq 0, \tag{4.6}$$

Table V. Errors in the Recursion Coefficients α_k, β_k of (4.5) Computed by Stieltjes's Procedure

N	n	err α	err β	N	n	err α	err β
40	≤ 35	$\leq 1.91(-13)$	$\leq 7.78(-13)$	160	≤ 76	$\leq 2.98(-13)$	$\leq 7.61(-13)$
	36	3.01(−12)	1.48(−11)		85	1.61(−9)	1.57(−8)
	37	6.93(−11)	3.55(−10)		94	1.25(−4)	1.17(−3)
	38	2.57(−9)	1.30(−8)		103	2.64(−3)	1.51(−1)
	39	1.93(−7)	9.58(−7)		112	2.35(−3)	1.16(0)
80	≤ 53	$\leq 2.04(-13)$	$\leq 6.92(-13)$	320	≤ 106	$\leq 8.65(-13)$	$\leq 7.39(-13)$
	57	2.04(−10)	5.13(−10)		117	3.96(−10)	7.73(−10)
	61	3.84(−7)	9.35(−7)		128	2.46(−6)	4.67(−6)
	65	1.94(−3)	4.61(−3)		139	2.94(−2)	6.27(−2)
	69	1.87(−1)	6.14(0)		150	1.15(−3)	2.18(−2)

consisting of a continuous part, $w(t)\,dt$, and (if $p > 0$) a discrete part written in terms of the Dirac δ-function. The support of the continuous part is assumed to be an interval or a finite union of disjoint intervals, some of which may extend to infinity. In the discrete part, the abscissas x_j are assumed pairwise distinct, and the weights positive, $y_j > 0$. The inner product (1.1), therefore, has the form

$$(u, v) = \int_{\mathbb{R}} u(t)v(t)w(t)\,dt + \sum_{j=1}^{p} y_j u(x_j)v(x_j). \tag{4.7}$$

The basic idea of the discretization procedure is rather simple: One approximates the continuous part of the inner product, that is, the integral in (4.7), by a sum, using a suitable quadrature scheme. If the latter involves N terms, this replaces the inner product (4.7) by a discrete inner product $(\cdot, \cdot)_{N+p}$ consisting of $N + p$ terms, the N "quadrature terms," and the p original terms. In effect, the measure $d\lambda$ in (4.6) is approximated by a discrete $(N + p)$-point measure $d\lambda_{N+p}$. We then compute the desired recursion coefficients from the formulas (4.1), in which the inner product $(\cdot, \cdot)$ is replaced, throughout, by $(\cdot, \cdot)_{N+p}$. Thus, in effect, we approximate

$$\alpha_k(d\lambda) \approx \alpha_k(d\lambda_{N+p}), \qquad \beta_k(d\lambda) \approx \beta_k(d\lambda_{N+p}). \tag{4.8}$$

The quantities on the right can be computed by the methods in Section 4.1 or 4.2, that is, employing the routines **sti** or **lancz**.

The difficult part of this approach is to find a discretization that results in rapid convergence, as $N \to \infty$, of the approximations on the right of (4.8) to the exact values on the left, even in cases where the weight function w in (4.6) exhibits singular behavior. (The speed of convergence, of course, is unaffected by the discrete part of the inner product (4.7).) To be successful in this endeavor often requires considerable inventiveness on the part of the user. Our routines, **mcdis** and **dmcdis**, which implement this idea in single (resp., double) precision, however, are designed to be flexible enough to promote the use of effective discretization procedures.

Indeed, if the support of the weight function w in (4.7) is contained in the (finite or infinite) interval (a, b), it is often useful to first decompose that

interval into a finite number of subintervals,

$$\operatorname{supp} w \subset [a, b] = \bigcup_{i=1}^{m} [a_i, b_i], \qquad m \geq 1, \tag{4.9}$$

and to approximate the inner product separately on each subinterval $[a_i, b_i]$, using an appropriate weighted quadrature rule. Thus, the integral in (4.7) is written as

$$\int_{\mathbb{R}} u(t)v(t)w(t)\,dt = \sum_{i=1}^{m} \int_{a_i}^{b_i} u(t)v(t)w_i(t)\,dt, \tag{4.10}$$

where w_i is an appropriate weight function on $[a_i, b_i]$. The intervals $[a_i, b_i]$ are not necessarily disjoint. For example, the weight function w may be the sum $w = w_1 + w_2$ of two weight functions on $[a, b]$, which we may want to treat individually (cf. Example 4.2). In that case, one would take $[a_1, b_1] = [a_2, b_2] = [a, b]$ and w_1 on the first interval, and w_2 on the other. Alternatively, we may simply want to use a composite quadrature rule to approximate the integral, in which case (4.9) is a partition of $[a, b]$ and $w_i(t) = w(t)$ for each i. Still another example is a weight function w that is already supported on a union of disjoint intervals; in this case, (4.9) would be the same union, or possibly a refined union where some of the subintervals are further partitioned.

In whichever way (4.9) and (4.10) are constructed, each integral on the right of (4.10) is now approximated by an appropriate quadrature rule,

$$\int_{a_i}^{b_i} u(t)v(t)w_i(t)\,dt \approx Q_i(uv), \tag{4.11}$$

where

$$Q_i f = \sum_{r=1}^{N_i} w_{r,i} f(x_{r,i}). \tag{4.12}$$

This gives rise to the approximate inner product

$$\begin{aligned} (u, v)_{N+p} &= \sum_{i=1}^{m} \sum_{r=1}^{N_i} w_{r,i} u(x_{r,i}) v(x_{r,i}) + \sum_{j=1}^{p} y_j u(x_j) v(x_j), \\ N &= \sum_{i=1}^{m} N_i. \end{aligned} \tag{4.13}$$

In our routine **mcdis**, we have chosen, for simplicity, all N_i to be the same integer N_0,

$$N_i = N_0, \qquad i = 1, 2, \ldots, m, \tag{4.14}$$

so that $N = mN_0$. Furthermore, if n is the number of α_k and the number of β_k desired, we have used the following iterative procedure to determine the coefficients α_k, β_k to a prescribed (relative) accuracy ϵ: Let N_0 be increased through a sequence $\{N_0^{[s]}\}_{s=0,1,2,\ldots}$ of integers, for each s use Stieltjes's (or

Lanczos's) algorithm to compute $\alpha_k^{[s]} = \alpha_k(d\lambda_{mN_0^{[s]}+p})$, $\beta_k^{[s]} = \beta_k(d\lambda_{mN_0^{[s]}+p})$, $k = 0, 1, \ldots, n-1$, and stop the iteration for the first $s \geq 1$ for which all inequalities

$$|\beta_k^{[s]} - \beta_k^{[s-1]}| \leq \epsilon\beta_k^{[s]}, \qquad k = 0, 1, \ldots, n-1, \tag{4.15}$$

are satisfied. An error flag is provided if within a preset range $N_0^{[s]} \leq N_0^{\max}$ the stopping criterion (4.15) cannot be satisfied. Note that the latter is based solely on the β-coefficients. This is because, unlike the α's, they are known to be always positive, so that it makes sense to insist on relative accuracy. (In our routine we actually replaced $\beta_k^{[s]}$ on the right of (4.15) by its absolute value to ensure proper termination in cases of sign-changing measures $d\lambda$.)

In view of formulas (4.1), it is reasonable to expect, and indeed has been observed in practice, that satisfaction of (4.15) entails sufficient absolute accuracy for the α's if they are zero or small, and relative accuracy otherwise.

Through a bit of experimentation, we have settled on the following sequence of integers $N_0^{[s]}$:

$$\begin{aligned} N_0^{[0]} &= 2n, & N_0^{[s]} &= N_0^{[s-1]} + \Delta s, & s &= 1, 2, \ldots, \\ \Delta_1 &= 1, & \Delta_s &= 2^{\lfloor s/5 \rfloor} \cdot n, & s &= 2, 3 \ldots . \end{aligned} \tag{4.16}$$

Note that if the quadrature formula (4.11) is exact for each i, whenever $u \cdot v$ is a polynomial of degree $\leq 2n-1$ (which is the maximum degree occurring in the inner products of (4.1), when $k \leq n-1$), then our procedure converges after the very first iteration step! Therefore, if each quadrature rule Q_i has (algebraic) degree of exactness $\geq d(N_0)$ and if $d(N_0)/N_0 = \delta + O(N_0^{-1})$ as $N_0 \to \infty$, then we let $N_0^{[0]} = 1 + \lfloor (2n-1)/\delta \rfloor$ in an attempt to get exact answers after one iteration. Normally, $\delta = 1$ (for interpolatory rules) or $\delta = 2$ (for Gauss-type rules).

The calling sequence of the multiple-component discretization routine is as follows:

```
mcdis(n, ncapm, mc, mp, xp, yp, quad, eps, iq, idelta, irout,
  finl, finr, endl, endr, xfer, wfer, alpha, beta, ncap,
  kount, ierr, ie, be, x, w, xm, wm, p0, p1, p2)
dimension xp(*), yp(*), endl(mc), endr(mc), xfer(ncapm),
  wfer(ncapm), alpha(n), beta(n), be(n), x(ncapm),
  w(ncapm), xm(*), wm(*), p0(*), p1(*), p2(*)
logical finl, finr
```

On entry,

n is the number of recursion coefficients desired; type integer.

ncapm is the integer $N_0^{\max}$ above, that is, the maximum integer N_0 allowed (**ncapm** = 500 will usually be satisfactory).

mc is the number of component intervals in the continuous part of the spectrum; type integer.

mp is the number of points in the discrete part of the spectrum; type integer; if the measure has no discrete part, set **mp** = 0.

xp, yp are arrays of dimension **mp** containing the abscissas and the jumps of the point spectrum.

quad is a subroutine determining the discretization of the inner product on each component interval, or a dummy routine if **iq** $\neq 1$ (see below); specifically, **quad(n, x, w, i, ierr)** produces the abscissas $\mathbf{x}(r) = x_{r,i}$ and weights $\mathbf{w}(r) = w_{r,i}$, $r = 1, 2, \ldots, n$, of the n-point discretization of the inner product on the interval $[a_i, b_i]$ (cf. (4.13)); an error flag **ierr** is provided to signal the occurrence of an error condition in the quadrature process.

eps is the desired relative accuracy of the nonzero recursion coefficients; type real.

iq is an integer selecting a user-supplied quadrature routine **quad** if **iq** $= 1$ or the ORTHPOL routine **qgp** (see below) otherwise.

idelta is a nonzero integer, typically 1 or 2, inducing fast convergence in the case of special quadrature routines; the default value is **idelta** $= 1$.

irout is an integer selecting the routine for generating the recursion coefficients from the discrete inner product; specifically, **irout** $= 1$ selects the routine **sti**, and **irout** $\neq 1$ selects the routine **lancz**.

The logical variables **finl, finr** and the arrays **endl, endr, xfer, wfer** are input variables to the subroutine **qgp** and are used (and, hence, need to be properly dimensioned) only if **iq** $\neq 1$.

On return,

alpha, beta are arrays of dimension **n** holding the desired recursion coefficients $\mathbf{alpha}(k) = \alpha_{k-1}$, $\mathbf{beta}(k) = \beta_{k-1}$, $k = 1, 2, \ldots,$ **n**.

ncap is the integer N_0 yielding convergence.

kount is the number of iterations required to achieve convergence.

ierr is an error flag, equal to 0 on normal return, equal to -1 if **n** is not in the proper range, equal to i if there is an error condition in the discretization on the ith interval, and equal to **ncapm** if the discretized Stieltjes procedure does not converge within the discretization resolution specified by **ncapm**.

ie is an error flag inherited from the routine **sti** or **lancz** (whichever is used).

The arrays **be, x, w, xm, wm, p0, p1, p2** are used for working space, the last five having dimension **mc** $\times$ **ncapm** + **mp**.

A general-purpose quadrature routine, **qgp**, is provided for cases in which it may be difficult to develop special discretizations that take advantage of the structural properties of the weight function w at hand. The routine

assumes the same setup (4.9)–(4.14) used in **mcdis**, with *disjoint* intervals $[a_i, b_i]$, and provides for Q_i in (4.12) the Fejér quadrature rule, suitably transformed to the interval $[a_i, b_i]$, with the same number $N_i = N_0$ of points for each i. Recall that the N-point Fejér rule on the standard interval $[-1, 1]$ is the interpolatory quadrature formula

$$Q_N^F f = \sum_{r=1}^{N} w_r^F f(x_r^F), \tag{4.17}$$

where $x_r^F = \cos((2r - 1)\pi/2N)$ are the Chebyshev points. The weights are all positive and can be computed explicitly in terms of trigonometric functions (cf., e.g., Gautschi [1967a]). The rule (4.17) is now applied to the integral in (4.11) by transforming the interval $[-1, 1]$ to $[a_i, b_i]$ via some monotone function ϕ_i (a linear function if $[a_i, b_i]$ is finite) and letting $f = uvw_i$:

$$\int_{a_i}^{b_i} u(t)v(t)w_i(t)\,dt = \int_{-1}^{1} u(\phi_i(\tau))v(\phi_i(\tau))w_i(\phi_i(\tau))\phi_i'(\tau)\,d\tau$$

$$\approx \sum_{r=1}^{N} w_r^F w_i\big(\phi_i(x_r^F)\big)\phi_i'(x_r^F) \cdot u\big(\phi_i(x_r^F)\big)v\big(\phi_i(x_r^F)\big).$$

Thus, in effect, we take in (4.13)

$$x_{r,i} = \phi_i(x_r^F), \qquad w_{r,i} = w_r^F w_i\big(\phi_i(x_r^F)\big)\phi_i'(x_r^F), \qquad i = 1, 2, \ldots, m. \tag{4.18}$$

If the interval $[a_i, b_i]$ is half-infinite, say, of the form $[0, \infty]$, we use $\phi_i(t) = (1 + t)/(1 - t)$, and similarly for intervals of the form $[-\infty, b]$ and $[a, \infty]$. If $[a_i, b_i] = [-\infty, \infty]$, we use $\phi_i(t) = t/(1 - t^2)$.

The routine **qgp** has the following calling sequence:

```
subroutine qgp(n, x, w, i, ierr, mc, finl, finr, endl, endr, xfer, wfer)
dimension x(n), w(n), endl(mc), endr(mc), xfer(*), wfer(*)
logical finl, finr
```

On entry,

n is the number of terms in the Fejér quadrature rule.

i indexes the interval $[a_i, b_i]$ for which the quadrature rule is desired; an interval that extends to $-\infty$ has to be indexed by 1, and one that extends to $+\infty$ by **mc**.

mc is the number of component intervals; type integer.

finl is a logical variable to be set **.true.** if the extreme left interval is finite and **.false.** otherwise.

finr is a logical variable to be set **.true.** if the extreme right interval is finite and **.false.** otherwise.

endl is an array of dimension **mc** containing the left endpoints of the component intervals; if the first of these extends to $-\infty$, **endl**(1) is not being used by the routine.

endr is an array of dimension **mc** containing the right endpoints of the component intervals; if the last of these extends to $+\infty$, **endr(mc)** is not being used by the routine.

xfer, wfer are working arrays holding the standard Fejér nodes and weights, respectively; they are dimensioned in the routine **mcdis**.

On return,

x, w are arrays of dimension **n** holding the abscissas and weights (4.18) of the discretized inner product for the ith component interval.

ierr has the integer value 0.

The routine calls on the subroutines **fejer**, **symtr** and **tr**, which are appended to the routine **qgp** in Section 4 of the package. The first generates the Fejér quadrature rule; the others perform variable transformations. The user has to provide his or her own function routine **wf(x, i)** to calculate the weight function $w_i(x)$ on the ith component interval.

Example 4.2 Chebyshev weight plus a constant: $w^c(t) = (1 - t^2)^{-1/2} + c$, $c > 0$, $-1 < t < 1$.

It would be difficult here to find a single quadrature rule for discretizing the inner product and to obtain fast convergence. However, using in (4.9) $m = 2$, $[a_1, b_1] = [a_2, b_2] = [-1, 1]$, and $w_1(t) = (1 - t^2)^{-1/2}$, $w_2(t) = c$ in (4.11), and taking for Q_1 the Gauss–Chebyshev, and for Q_2 the Gauss–Legendre n-point rule (the latter multiplied by c), yield convergence to $\alpha_k(w^c)$, $\beta_k(w^c)$, $k = 0, 1, \ldots, n - 1$, in one iteration (provided δ is set equal to 2)! Actually, we need $N_0 = n + 1$, in order to test for convergence; cf. (4.15). The driver **test4** implements this technique and calculates the first $n = 80$ beta-coefficients to a relative accuracy of $5000 \times \epsilon^s$ for $c = 1, 10, 100$. (All α_k are zero.) Attached to the driver is the quadrature routine **qchle** used in this example. It, in turn, calls for the Gauss quadrature routine **gauss**, to be described in Section 6. Anticipating convergence after one iteration, we put **ncapm** = 81.

The weight function of Example 4.2 provides a continuous link between the Chebyshev polynomials ($c = 0$) and the Legendre polynomials ($c = \infty$); the recursion coefficients $\beta_k(w^c)$ indeed converge (except for $k = 0$) to those of the Legendre polynomials, as $c \to \infty$.

Selected results of **test4** (where **irout** in **mcdis** can be arbitrary) are shown in Table VI. The output variable **kount** is 1 in each case, confirming convergence after one iteration. The coefficients $\beta_0(w^c)$ are easily seen to be $\pi + 2c$.

Example 4.3 Jacobi weight with one mass point at the left endpoint: $w^{(\alpha,\beta)}(t; y) = [\mu_0^{(\alpha,\beta)}]^{-1}(1 - t)^\alpha(1 + t)^\beta + y\delta(t + 1)$ on $(-1, 1)$, $\mu_0^{(\alpha,\beta)} = 2^{\alpha+\beta+1}\Gamma(\alpha + 1)\Gamma(\beta + 1)/\Gamma(\alpha + \beta + 2)$, $\alpha > -1$, $\beta > -1$, $y > 0$.

Table VI. Selected Recursion Coefficients $\beta_k(w^c)$ for $c = 1, 10, 100$

k	$\beta_k(w^1)$	$\beta_k(w^{10})$	$\beta_k(w^{100})$
0	5.1415926540	23.14159265	203.1415927
1	0.4351692451	0.3559592080	0.3359108398
5	0.2510395775	0.2535184776	0.2528129500
12	0.2500610870	0.2504824840	0.2505324193
25	0.2500060034	0.2500682357	0.2501336338
51	0.2500006590	0.2500082010	0.2500326887
79	0.2500001724	0.2500021136	0.2500127264

The recursion coefficients α_k, β_k are known explicitly (see Chihara [1985, Eqs. 6.23, 3.5][1]) and can be expressed, with some effort, in terms of the recursion coefficients α_k^J, β_k^J for the Jacobi weight $w^{(\alpha,\beta)}(\cdot) = w^{(\alpha,\beta)}(\cdot; 0)$. The formulas are

$$\alpha_0 = \frac{\alpha_0^J - y}{1+y}, \qquad \beta_0 = \beta_0^J + y,$$

$$\alpha_k = \alpha_k^J + \frac{2k(\alpha+k)}{(\alpha+\beta+2k)(\alpha+\beta+2k+1)}(c_k - 1) + \frac{2(\beta+k+1)(\alpha+\beta+k+1)}{(\alpha+\beta+2k+1)(\alpha+\beta+2k+2)}\left(\frac{1}{c_k} - 1\right),$$

$$\beta_k = \frac{c_k}{c_{k-1}}\beta_k^J, \qquad k = 1, 2, 3, \ldots, \tag{4.19}$$

where

$$c_0 = 1 + y, \qquad c_k = \frac{1 + \dfrac{(\beta+k+1)(\alpha+\beta+k+1)}{k(\alpha+k)} y d_k}{1 + y d_k}, \qquad k = 1, 2, \ldots, \tag{4.20}$$

and

$$d_1 = 1,$$
$$d_k = \frac{(\beta+k)(\alpha+\beta+k)}{(\alpha+k-1)(k-1)} d_{k-1}, \qquad k = 2, 3, \ldots. \tag{4.21}$$

Again, it is straightforward with **mcdis** to get exact results (modulo rounding) after one iteration, by using the Gauss–Jacobi quadrature rule (see **gauss** in Section 6) to discretize the continuous part of the measure. The driver **test5** generates in this manner the first $n = 40$ recursion coefficients α_k, β_k, $k = 0, 1, \ldots, n-1$, to a relative accuracy of $5000 \times \epsilon^s$, for $y = \frac{1}{2}, 1, 2,$

[1] In Chihara [1985] the interval is taken to be [0, 2], rather than [−1, 1]. There is a typographical error in the first formula of (6.23), which should have the numerator $2\beta + 2$ instead of $2\beta + 1$.

4, and 8. For each $\alpha = -.8(.2)1.$ and $\beta = -.8(.2)1.$, it computes the maximum relative errors (absolute error, if $\alpha_k \approx 0$) of the α_k, β_k by comparing them with the exact coefficients. These have been computed in double precision by a straightforward implementation of formulas (4.19)–(4.21).

As expected, the output of **test5** reveals convergence after one iteration, the variable **kount** having consistently the value 1. The maximum relative error in the α_k is found to lie generally between 2×10^{-8} and 3×10^{-8}, the one in the β_k between 7.5×10^{-12} and 8×10^{-12}; they are attained for k at or near 39. The discrepancy between the errors in the α_k and those in the β_k is due to the α_k being considerably smaller than the β_k, by 3–4 orders of magnitude. Replacing the routine **sti** in **mcdis** by **lancz** yields very much the same error picture.

It is interesting to note that the addition of a second mass point at the other endpoint makes an analytic determination of the recursion coefficients intractable (cf. Chihara [1985, p. 713]). Numerically, however, it makes no difference whether there are two or more mass points and whether they are located inside, outside, or on the boundary of the support interval. It was observed, however, that if at least one mass point is located outside the interval $[-1, 1]$ the procedure **sti** used in **mcdis** becomes severely unstable[2] and *must* be replaced by **lancz**.

Example 4.4 Logistic density function: $w(t) = e^{-t}/(1 + e^{-t})^2$ on $(-\infty, \infty)$. In this example we illustrate a slight variation of the discretization procedure (4.9)–(4.13), which ends up with a discrete inner product of the same type as in (4.13) (and thus implementable by the routine **mcdis**), but derived in a somewhat different manner. The idea is to integrate functions with respect to the density w by splitting the integral into two parts, one from $-\infty$ to 0 and the other from 0 to ∞, changing variables in the first part, and thus obtaining

$$\int_{-\infty}^{\infty} f(t)w(t)\,dt = \int_0^{\infty} f(-t)\frac{e^{-t}}{(1+e^{-t})^2}\,dt + \int_0^{\infty} f(t)\frac{e^{-t}}{(1+e^{-t})^2}\,dt. \quad (4.22)$$

Since $(1 + e^{-t})^{-2}$ quickly tends to 1 as $t \to \infty$, a natural discretization of both integrals is provided by the Gauss–Laguerre quadrature rule applied to the product $f(\pm t) \cdot (1 + e^{-t})^{-2}$. This amounts to taking, in (4.13), $m = 2$ and

$$x_{r,1} = -x_r^L, \qquad x_{r,2} = x_r^L; \qquad w_{r,1} = w_{r,2} = \frac{w_r^L}{(1+e^{-x_r^L})^2},$$
$$r = 1, 2, \ldots, N,$$

where x_r^L, w_r^L are the Gauss–Laguerre N-point quadrature nodes and weights.

[2] This has also been observed in a similar example [Gautschi 1982a, Example 4.8], but was incorrectly attributed to a phenomenon of ill-conditioning. Indeed, the statement made at the end of Example 4.8 can now be retracted: Stable methods *do* exist, namely, the method embodied by the routine **mcdis** in combination with **lancz**.

Table VII. Selected Output from **test6**

k	β_k	err α_k	err β_k
0	1.000000000000000000000000	4.572(−13)	1.918(−13)
1	3.289868133696452872944830	1.682(−13)	5.641(−13)
6	89.447603523150188817832	2.187(−12)	2.190(−12)
15	555.7827839879296775066697	1.732(−13)	2.915(−12)
26	1668.580222268668421827788	3.772(−12)	4.112(−12)
39	3753.534025194898387722354	2.482(−11)	4.533(−12)

The driver **test6** incorporates this discretization into the routines **mcdis** and **dmcdis**, runs them for $n = 40$ with error tolerances $5000 \times \epsilon^s$ and $1000 \times \epsilon^d$, respectively, and prints the absolute errors in the α's ($\alpha_k = 0$, in theory) and the relative errors in the β's. (We used the default value $\delta = 1$.) Also printed are the number of iterations #it (= **kount**) in (4.15) and the corresponding final value N_0^f (= **ncap**). In single precision we found that #it = 1, $N_0^f = 81$, and in double precision, #it = 5, $N_0^f = 281$. Both routines returned with the error flags equal to 0, indicating a normal course of events. A few selected double-precision values[3] of the coefficients β_k along with absolute errors in the α's and relative errors in the β's are shown in Table VII. The results are essentially the same no matter whether **sti** or **lancz** is used in **mcdis**. The maximum errors observed are 2.482×10^{-11} for the α's and 4.939×10^{-12} for the β's, which are well within the single-precision tolerance $\epsilon = 5000 \times \epsilon^s$.

On computers with limited exponent range, convergence difficulties may arise, both with **sti** and **lancz**, owing to underflow in many of the Laguerre quadrature weights. This seems to perturb the problem significantly enough to prevent the discretization procedure from converging.

Example 4.5 Half-range Hermite measure: $w(t) = e^{-t^2}$ on $(0, \infty)$.

This is an example of a measure for which there do not seem to exist natural discretizations other than those based on composite quadrature rules. Therefore, we applied our general-purpose routine **qgp** (and its double-precision companion **dqgp**), using, after some experimentation, the partition $[0, \infty] = [0, 3] \cup [3, 6] \cup [6, 9] \cup [9, \infty]$. The driver **test7** implements this, with $n = 40$ and an error tolerance $50 \times \epsilon^s$ in single precision, and $1000 \times \epsilon^d$ in double precision.

The single-precision routine **mcdis** (using the default value $\delta = 1$) converged after one iteration, returning **ncap** = 81, whereas the double-precision routine **dmcdis** took four iterations to converge and returned **ncapd** = 201. Selected results (where err α_k and err β_k both denote relative errors) are shown in Table VIII. The maximum error err α_k occurred at $k = 10$ and had the value 1.038×10^{-12}, whereas $\max_k$ err $\beta_k = 3.180 \times 10^{-13}$ is attained at $k = 0$. The latter is within the error tolerance ϵ, the former only slightly

[3] *Note added in proof:* Alphonse Magnus, in an email message, May 5, 1993, kindly pointed out to the author that the β-coefficients are known explicitly: $\beta_k = k^4\pi^2/(4k^2 - 1)$, $k = 1, 2, \ldots$.

Table VIII. Selected Output from **test7**

k	α_k and err α_k	β_k and err β_k
0	0.56418958354775628694807 95	0.88622692545275801364908 37
	1.096(−13)	3.180(−13)
1	0.98842539284680028548706 34	0 18169011381620932846223 25
	1.514(−13)	7.741(−14)
6	2.08062033640083322481762 2	1.00234785101101084222453 8
	1.328(−13)	5.801(−14)
15	3.21427063607112822744891 4	2.50092791713370266995432 1
	2.402(−14)	8.186(−14)
26	4.20304857887200195266027 7	4.33386790122995044360443 0
	1.415(−13)	7.878(−14)
39	5.13153288689429651931969 2	6.50035623770713293803515 5
	6.712(−13)	1.820(−14)

larger. Comparison of the double-precision results with Table I on the microfiche supplement to Galant [1969] revealed agreement to all 20 decimal digits given there, for all k in the range $0 \le k \le 19$. Interestingly, the routine **sti** in **mcdis** did consistently better than **lancz** on the β's, by a factor as large as 235 (for $k = 33$), and is comparable with **lancz** (sometimes better, sometimes worse) on the α's.

Without composition, that is, using **mc** = 1 in **mcdis**, it takes 8 iterations ($N_0^f = 521$) in single precision and 10 iterations ($N_0^f = 761$) in double precision to satisfy the much weaker error tolerances $\epsilon = \frac{1}{2}\,10^{-6}$ and $\epsilon^d = \frac{1}{2}\,10^{-12}$, respectively. All single-precision results, however, turn out to be accurate to about 12 decimal places. (This is because of the relatively large final increment $\Delta_8 = 2n = 80$ in N_0 (cf. (4.16)) that forces convergence.)

4.4 Discretized Modified Chebyshev Algorithm

The whole apparatus of discretization (cf. (4.9)–(4.14)) can also be employed in connection with the modified Chebyshev algorithm (cf. Section 3), if one discretizes modified moments rather than inner products. Thus, one approximates (cf. (4.14), (4.16))

$$\nu_k(d\lambda) \approx \nu_k(d\lambda_{m N_0^{|s|}+p}) \tag{4.23}$$

and iterates the modified Chebyshev algorithm with $s = 0, 1, 2, \ldots$ until the convergence criterion (4.15) is satisfied. (It would be unwise to test convergence on the modified moments, for reasons explained in Gautschi [1982a, sect. 2.5].) This is implemented in the routine **mccheb**, whose calling sequence is as follows:

mccheb(n, ncapm, mc, mp, xp, yp, quad, eps, iq, idelta, finl, finr, endl, endr, xfer, wfer, a, b, fnu, alpha, beta, ncap, kount, ierr, be, x, w, xm, wm, s, s0, s1, s2)

Its input and output parameters have the same meaning as in the routine **mcdis**. In addition, the arrays **a**, **b** of dimension $2 \times \mathbf{n} - 1$ are to be supplied

with the recursion coefficients $\mathbf{a}(k) = a_{k-1}, \mathbf{b}(k) = b_{k-1}, k = 1, 2, \ldots, 2 \times \mathbf{n} - 1$, defining the modified moments. The arrays **be, x, w, xm, wm, s, s0, s1, s2** are used for working space. The double-precision version of the routine has the name **dmcheb**.

The discretized modified Chebyshev algorithm must be expected to behave similarly as its close relative, the modified Chebyshev algorithm. In particular, if the latter suffers from ill-conditioning, so does the former.

Example 4.6 (Example 3.1, revisited).

We recompute the $n = 40$ first recursion coefficients α_k, β_k of Example 3.1 to an accuracy of $100 \times \epsilon^s$ in single precision, using the routine **mccheb** instead of the routine **cheb**. For the discretization of the modified moments, we employed the Gauss–Chebyshev quadrature rule:

$$\int_{-1}^{1} f(t)(1 - \omega^2 t^2)^{-1/2}(1 - t^2)^{-1/2}\, dt \approx \frac{\pi}{N} \sum_{r=1}^{N} f(x_r)(1 - \omega^2 x_r^2)^{-1/2}, \tag{4.24}$$

where $x_r = \cos((2r - 1)\pi/2N)$ are the Chebyshev points. This is accomplished by the driver **test8**. The results of this test (shown in the package) agree to all 10 decimal places with those of **test1**. The routine **mccheb** converged in one iteration, with **ncap** = 81, for $\omega^2 = .1, .3, .5, .7, .9$; in 4 iterations, with **ncap** = 201, for $\omega^2 = .99$; and in 8 iterations, with **ncap** = 521, for $\omega^2 = .999$. A double-precision version of **test8** was also run with $\epsilon = \frac{1}{2} \times 10^{-20}$ (not shown in the package) and produced correct results to 20 decimals in one iteration (**ncap** = 81) for $\omega^2 = .1, .3, .5, .7$; in 3 iterations (**ncap** = 161) for $\omega^2 = .9$; in 6 iterations (**ncap** = 361) for $\omega^2 = .99$; and in 11 iterations (**ncap** = 921) for $\omega^2 = .999$.

5. MODIFICATION ALGORITHMS

Given a positive measure $d\lambda(t)$ supported on the real line, and two polynomials $u(t) = \pm \prod_{\rho=1}^{r}(t - u_\rho)$, $v(t) = \prod_{\sigma=1}^{s}(t - v_\sigma)$ whose ratio is finite on the support of $d\lambda$, we may ask for the recursion coefficients $\hat{\alpha}_k = \alpha_k(d\hat{\lambda})$, $\hat{\beta}_k = \beta_k(d\hat{\lambda})$ of the modified measure

$$d\hat{\lambda}(t) = \frac{u(t)}{v(t)}\, d\lambda(t), \qquad t \in \operatorname{supp}(d\lambda), \tag{5.1}$$

assuming known the recursion coefficients $\alpha_k = \alpha_k(d\lambda)$, $\beta_k = \beta_k(d\lambda)$ of the given measure. Methods that accomplish the passage from the α's and β's to the $\hat{\alpha}$'s and $\hat{\beta}$'s are called *modification algorithms*. The simplest case $s = 0$ (i.e., $v(t) \equiv 1$) and u positive on $\operatorname{supp}(d\lambda)$ has already been considered by Christoffel [1858], who represented the polynomial $u(\cdot)\hat{\pi}_k(\cdot) = u(\cdot)\pi_k(\cdot; d\hat{\lambda})$ in determinantal form in terms of the polynomials $\pi_j(\cdot) = \pi_j(\cdot; d\lambda)$, $j = k, k + 1, \ldots, k + r$. This is now known as *Christoffel's theorem*. Christoffel, however, did not address the problem of how to generate the new coefficients $\hat{\alpha}_k$, $\hat{\beta}_k$ in terms of the old ones. For the more general modification (5.1), Christoffel's theorem has been generalized by Uvarov [1959; 1969]. The coefficient prob-

lem stated above, in this general case, has been treated in Gautschi [1982b], and previously by Galant [1971] in the special case $v(t) \equiv 1$.

The passage from $d\lambda$ to $d\hat{\lambda}$ can be carried out in a sequence of elementary steps involving real linear factors $t - x$ or real quadratic factors $(t - x)^2 + y^2$, either in $u(t)$ or in $v(t)$. The corresponding elementary steps in the passage from the α's and β's to the $\hat{\alpha}$'s and $\hat{\beta}$'s can all be performed by means of certain nonlinear recurrences. Some of these, however, when divisions of the measure $d\lambda$ are involved, are liable to instabilities. An alternative method can then be used, which appeals to the modified Chebyshev algorithm supplied with appropriate modified moments. These latter are of independent interest and find application, for example, in evaluating the kernel in the contour integral representation of the Gauss quadrature remainder term.

5.1 Nonlinear Recurrence Algorithms

The routine that carries out the elementary modification steps is called **chri** and has the calling sequence

chri(n,iopt,a,b,x,y,hr,hi,alpha,beta,ierr).

On entry,

n is the number of recursion coefficients desired; type integer.

iopt is an integer identifying the type of modification as follows:
(1) $d\hat{\lambda}(t) = (t - x)\,d\lambda(t)$.
(2) $d\hat{\lambda}(t) = ((t - x)^2 + y^2)\,d\lambda(t)$, $y > 0$.
(3) $d\hat{\lambda}(t) = (t^2 + y^2)\,d\lambda(t)$ with $d\lambda(t)$ and $\text{supp}(d\lambda)$ assumed symmetric with respect to the origin and $y > 0$.
(4) $d\hat{\lambda}(t) = d\lambda(t)/(t - x)$.
(5) $d\hat{\lambda}(t) = d\lambda(t)/((t - x)^2 + y^2)$, $y > 0$.
(6) $d\hat{\lambda}(t) = d\lambda(t)/(t^2 + y^2)$ with $d\lambda(t)$ and $\text{supp}(d\lambda)$ assumed symmetric with respect to the origin and $y > 0$.
(7) $d\hat{\lambda}(t) = (t - x)^2 d\lambda(t)$.

a, b are arrays of dimension **n** + 1 holding the recursion coefficients $\mathbf{a}(k) = \alpha_{k-1}(d\lambda)$, $\mathbf{b}(k) = \beta_{k-1}(d\lambda)$, $k = 1, 2, \ldots, \mathbf{n} + 1$.

x, y are real parameters defining the linear and quadratic factors (or divisors) of $d\lambda$.

hr, hi are the real and imaginary part, respectively, of $\int_{\mathbb{R}} d\lambda(t)/(z - t)$, where $z = x + iy$; the parameter **hr** is used only if **iopt** = 4 or 5, and the parameter **hi** only if **iopt** = 5 or 6.

On return,

alpha, beta are arrays of dimension **n** containing the desired recursion coefficients $\mathbf{alpha}(k) = \alpha_{k-1}(d\hat{\lambda})$, $\mathbf{beta}(k) = \beta_{k-1}(d\hat{\lambda})$, $k = 1, 2, \ldots, \mathbf{n}$.

ierr is an error flag, equal to 0 on normal return, equal to 1 if **n** $\leq$ 1 (the routine assumes that **n** is larger than or equal to 2), and equal to 2 if the integer **iopt** is inadmissible.

It should be noted that in the cases **iopt** = 1 and **iopt** = 4, the modified measure $d\hat{\lambda}$ is positive (negative) definite if x is to the left (right) of the support of $d\lambda$, but indefinite otherwise. Nevertheless, it is permissible to have x inside the support of $d\lambda$ (or inside its convex hull), provided the resulting measure $d\hat{\lambda}$ is still quasi-definite (cf. Gautschi [1982b]).

For **iopt** = $1, 2, \ldots, 6$, the methods used in **chri** are straightforward implementations of the nonlinear recurrence algorithms, respectively, in Eqs. (3.7), (4.7), (4.8), (5.1), (5.8), and (5.9) of Gautschi [1982b]. The only minor modification required concerns $\hat{\beta}_0 = \beta_0(d\hat{\lambda})$. In Gautschi [1982b] this constant was taken to be 0, whereas here it is defined to be $\hat{\beta}_0 = \int_{\mathbb{R}} d\hat{\lambda}(t)$. Thus, for example, if **iopt** = 2,

$$\hat{\beta}_0 = \int_{\mathbb{R}} \left((t-x)^2 + y^2\right) d\lambda(t) = \int_{\mathbb{R}} \left((t-\alpha_0+\alpha_0-x)^2 + y^2\right) d\lambda(t)$$

$$= \int_{\mathbb{R}} \left((t-\alpha_0)^2 + (\alpha_0-x)^2 + y^2\right) d\lambda(t),$$

since $\int_{\mathbb{R}}(t-\alpha_0)\,d\lambda(t) = \int_{\mathbb{R}} \pi_1(t)\,d\lambda(t) = 0$. Furthermore (cf. (4.1)),

$$\int_{\mathbb{R}} (t-\alpha_0)^2\, d\lambda(t) = \beta_0\beta_1,$$

so that the formula to be used for $\hat{\beta}_0$ is

$$\hat{\beta}_0 = \beta_0\left(\beta_1 + (\alpha_0 - x)^2 + y^2\right) \qquad (\textbf{iopt} = 2).$$

Similar calculations need to be made in the other cases.

The case **iopt** = 7 incorporates a QR step with shift x, following Kautsky and Golub [1983], and uses an adaptation of the algorithm in Wilkinson [1965, Eq. 67.11, p. 567], to carry out the QR step. The most significant modification made is the replacement of the test $c \neq 0$ by $|c| > \epsilon$, where $\epsilon = 5 \times \epsilon^s$ is a quantity close to, but slightly larger than, the machine precision. (Without this modification, the algorithm could fail.)

The methods used in **chri** are believed to be quite stable when the measure $d\lambda$ is modified multiplicatively (**iopt** = 1, 2, 3, and 7). When divisions are involved (**iopt** = 4, 5, and 6), however, the algorithms rapidly become unstable as the point $z = x + iy \in \mathbb{C}$ moves away from the support interval of $d\lambda$. (The reason for this instability is not well understood at present; see, however, Galant [1992].) For such cases there is an alternative routine, **gchri** (see Section 5.2), that can be used.

Example 5.1 Checking the results (for $\sigma = \frac{1}{2}$) of **test2**.

We apply **chri** (and the corresponding double-precision routine **dchri**) with **iopt** = 1, $x = 0$, to $d\lambda_\sigma(t) = t^\sigma \ln(1/t)$ on $(0, 1)$ with $\sigma = -\frac{1}{2}$, to recompute the results of **test2** for $\sigma = \frac{1}{2}$. This can be done by a minor modification, named **test9**, of **test2**. Selected results from it, showing the relative discrepancies between the single-precision values α_k, β_k (resp. double-precision values α_k^d, β_k^d), computed by the modified Chebyshev algorithm and the modification algorithm, are shown in Table IX (cf. Table III). The maximum errors occur consistently for the last value of $k(= 98)$.

Table IX. Comparison between Modified Chebyshev Algorithm and Modification Algorithm in Example 5.1 (cf. Example 3.2)

σ	k	err α_k	err β_k	err α_k^d	err β_k^d
.5	0	7.895(−14)	4.796(−14)	2.805(−28)	7.952(−28)
	12	3.280(−12)	6.195(−12)	8.958(−26)	1.731(−25)
	24	7.648(−12)	1.478(−11)	2.065(−25)	3.985(−25)
	48	2.076(−11)	4.088(−11)	5.683(−25)	1.121(−24)
	98	6.042(−11)	1.201(−10)	1.504(−24)	2.987(−24)

Example 5.2 Induced orthogonal polynomials.

Given an orthogonal polynomial $\pi_m(\cdot; d\lambda)$ of fixed degree $m \geq 1$, the sequence of orthogonal polynomials $\hat{\pi}_{k,m}(\cdot) = \pi_k(\cdot; \pi_m^2\, d\lambda)$, $k = 0, 1, 2, \ldots$, has been termed *induced orthogonal polynomials* in Gautschi and Li [1993]. Since their measure $d\hat{\lambda}_m$ modifies the measure $d\lambda$ by a product of quadratic factors,

$$d\hat{\lambda}_m(t) = \prod_{\mu=1}^{m} (t - x_\mu)^2 \cdot d\lambda(t), \tag{5.2}$$

where x_μ are the zeros of π_m, we can apply the routine **chri** (with **iopt** = 7) m times to compute the n recursion coefficients $\hat{\alpha}_{k,m} = \alpha_k(d\hat{\lambda}_m)$, $\hat{\beta}_{k,m} = \beta_k(d\hat{\lambda}_m)$, $k = 0, 1, \ldots, n-1$, from the $n + m$ coefficients $\alpha_k = \alpha_k(d\lambda)$, $\beta_k = \beta_k(d\lambda)$, $k = 0, 1, \ldots, n - 1 + m$. The subroutines **indp** and **dindp** in the driver **test10** implement this procedure in single (resp., double) precision. The driver itself uses them to compute the first $n = 20$ recursion coefficients of the induced Legendre polynomials with $m = 0, 1, \ldots, 11$. It also computes the maximum absolute errors in the $\hat{\alpha}$'s ($\hat{\alpha}_{k,m} = 0$ for all m) and the maximum relative errors in the $\hat{\beta}$'s by comparing single-precision with double-precision results.

An excerpt of the output of **test10** is shown in Table X. It already suggests a high degree of stability of the procedure employed by **indp**. This is reinforced by an additional test (not shown in the package) generating $n = 320$ recursion coefficients $\hat{\alpha}_{k,m}$, $\hat{\beta}_{k,m}$, $0 \leq k \leq 319$, for $m = 40$, 80, 160, 320 and $d\lambda$ being the Legendre, the first-kind Chebyshev, the Laguerre, and the Hermite measure. Table XI shows the maximum absolute error in the $\hat{\alpha}_{k,m}$, $0 \leq k \leq 319$ (relative error in the Laguerre case), and the maximum relative error in the $\hat{\beta}_{k,m}$, $0 \leq k \leq 319$.

5.2 Methods Based on the Modified Chebyshev Algorithm

As was noted earlier, the procedure **chri** becomes unstable for modified measures involving division of $d\lambda(t)$ by $t - x$ or $(t - x)^2 + y^2$ as $z = x + iy \in \mathbb{C}$ moves away from the "support interval" of $d\lambda$, that is, from the smallest interval containing the support of $d\lambda$. We now develop a procedure that works better the further away z is from that interval.

The idea is to use modified moments of $d\hat{\lambda}$ relative to the polynomials $\pi_k(\cdot; d\lambda)$ to generate the desired recursion coefficients $\hat{\alpha}_k$, $\hat{\beta}_k$ via the modified Chebyshev algorithm (cf. Section 3). The modified moments in question

Table X. Induced Legendre Polynomials

k	$m = 0, \hat{\beta}_{k,m}$	$m = 2, \hat{\beta}_{k,m}$	$m = 6, \hat{\beta}_{k,m}$	$m = 11, \hat{\beta}_{k,m}$
0	2.0000000000	0.1777777778	0.0007380787	0.0000007329
1	0.3333333333	0.5238095238	0.5030303030	0.5009523810
6	0.2517482517	0.1650550769	0.2947959861	0.2509913424
12	0.2504347826	0.2467060415	0.2521022519	0.1111727541
19	0.2501732502	0.2214990335	0.2274818789	0.2509466619
err $\hat{\alpha}$	0.000(0)	1.350(−13)	9.450(−13)	1.357(−12)
err $\hat{\beta}$	1.737(−14)	2.032(−13)	2.055(−12)	3.748(−12)

Table XI. Accuracy of the Recursion Coefficients for Some Classical Induced Polynomials

	Legendre		Chebyshev		Laguerre		Hermite	
m	err $\hat{\alpha}$	err $\hat{\beta}$	err $\hat{\alpha}$	err $\hat{\beta}$	err $\hat{\alpha}$	err $\hat{\beta}$	err $\hat{\alpha}$	err $\hat{\beta}$
40	3.4(−11)	1.5(−10)	1.9(−9)	7.9(−10)	3.0(−10)	6.0(−10)	1.8(−9)	2.7(−10)
80	1.4(−10)	5.4(−10)	2.1(−9)	2.2(−9)	5.8(−10)	9.2(−10)	7.9(−9)	9.2(−10)
160	1.5(−9)	5.1(−9)	9.5(−9)	1.1(−8)	7.8(−10)	1.4(−9)	1.1(−8)	6.8(−10)
320	3.3(−9)	2.1(−8)	9.6(−9)	2.1(−8)	3.9(−9)	7.2(−9)	2.5(−8)	1.1(−9)

are

$$\nu_k = \nu_k(x; d\lambda) = \int_{\mathbb{R}} \frac{\pi_k(t; d\lambda)}{t - x}\, d\lambda(t), \qquad k = 0, 1, 2, \ldots, \tag{5.3}$$

for linear divisors and

$$\nu_k = \nu_k(x, y; d\lambda) = \int_{\mathbb{R}} \frac{\pi_k(t; d\lambda)}{(t - x)^2 + y^2}\, d\lambda(t), \qquad k = 0, 1, 2, \ldots, \tag{5.4}$$

for quadratic divisors. Both can be expressed in terms of the integrals

$$\rho_k = \rho_k(z; d\lambda) = \int_{\mathbb{R}} \frac{\pi_k(t; d\lambda)}{z - t}\, d\lambda(t), \qquad z \in \mathbb{C} \setminus \operatorname{supp}(d\lambda), \quad k = 0, 1, 2, \ldots, \tag{5.5}$$

the first by means of

$$\nu_k(x; d\lambda) = -\rho_k(z; d\lambda), \qquad z = x, \tag{5.6}$$

and the others by means of

$$\nu_k(x, y; d\lambda) = -\frac{\operatorname{Im} \rho_k(z; d\lambda)}{\operatorname{Im} z}, \qquad z = x + iy. \tag{5.7}$$

The point to observe is that $\{\rho_k(z; d\lambda)\}$ is a minimal solution of the basic recurrence relation (1.3) for the orthogonal polynomials $\{\pi_k(\cdot; d\lambda)\}$ (cf. Gautschi [1981]). The quantities $\rho_k(z; d\lambda)$, $k = 0, 1, \ldots, n$, therefore, can be computed accurately by a backward recurrence algorithm [Gautschi 1981, sect. 5], which, for $\nu > n$, produces approximations $\rho_k^{[\nu]}(z; d\lambda)$ converging to $\rho_k(z; d\lambda)$ when $\nu \to \infty$, for any fixed k,

$$\rho_k^{[\nu]}(z; d\lambda) \to \rho_k(z; d\lambda), \qquad \nu \to \infty. \tag{5.8}$$

The procedure is implemented in the routine

knum(n, nu0, numax, z, eps, a, b, rho, nu, ierr, rold),

which computes $\rho_k(z; d\lambda)$ for $k = 0, 1, \ldots,$ **n** to a relative precision **eps**. The results are stored as **rho**$(k) = \rho_{k-1}(z; d\lambda)$, $k = 1, 2, \ldots,$ **n** + 1, in the complex array **rho** of dimension $n + 1$. The user has to provide a starting index **nu0** $= \nu_0 > n$ for the backward recursion, which the routine then increments by units of 5 until convergence to within **eps** is achieved. If the requested accuracy **eps** cannot be realized for some $\nu \leq$ **numax**, the routine exits with **ierr** = **numax**. Likewise, if $\nu_0 >$ **numax**, the routine exits immediately, with the error flag **ierr** set equal to **nu0**. Otherwise, the value of ν for which convergence is obtained is returned as output variable **nu**. The arrays **a**, **b** of dimension **numax** are to hold the recursion coefficients **a**$(k) = \alpha_{k-1}(d\lambda)$, **b**$(k) = \beta_{k-1}(d\lambda)$, $k = 1, 2, \ldots,$ **numax**, for the given measure $d\lambda$. The complex array **rold** of dimension $n + 1$ is used for working space. In the interest of rapid convergence, the routine should be provided with a realistic estimate of ν_0. For classical measures, such estimates are known (cf. Gautschi [1981, sect. 5]) and are implemented here by the function routines

nu0jac(n, z, eps), nu0lag(n, z, al, eps), nu0her(n, z, eps).

The first is for Jacobi measures, the second is for generalized Laguerre measures with parameter **al** $= \alpha > -1$, and the last is for the Hermite measure. Note that ν_0 for Jacobi measures does not depend on the weight parameters α, β, in contrast to ν_0 for the generalized Laguerre measure.

The name **knum** comes from the fact that $\rho_n(z; d\lambda)$ in (5.5) is the numerator in the kernel

$$K_n(z; d\lambda) = \frac{\rho_n(z; d\lambda)}{\pi_n(z; d\lambda)} \tag{5.9}$$

of the remainder term of the n-point Gaussian quadrature rule for analytic functions (cf., e.g., Gautschi and Varga [1983]). For the sequence of kernels $K_0, K_1, \ldots, K_n$, we have the following routine:

```
subroutine kern(n, nu0, numax, z, eps, a, b, ker, nu, ierr, rold)
complex z, ker, rold, p0, p, pm1
dimension a(numax), b(numax), ker(*), rold(*)
call knum(n, nu0, numax, z, eps, a, b, ker, nu, ierr, rold)
if(ierr.ne.0) return
p0 = (0., 0.)
p = (1., 0.)
do 10 k = 1, n
  pm1 = p0
  p0 = p
  p = (z - a(k))*p0 - b(k)*pm1
  ker(k + 1) = ker(k + 1) / p
10 continue
return
end
```

The meaning of the input and output parameters is the same as in **knum**. The double-precision version of the routine is named **dkern**.

All of the ingredients are now in place to describe the workings of **gchri**, the alternative routine to **chri** when the latter is unstable. First, the routine **knum** is used to produce the first $2n$ modified moments $\nu_k(x; d\lambda)$ (resp., $\nu_k(x, y; d\lambda)$), $k = 0, 1, \ldots, 2n-1$. These are then supplied to the routine **cheb** along with the recursion coefficients $\alpha_k(d\lambda)$, $\beta_k(d\lambda)$ (needed anyhow for the computation of the ν_k), which produces the desired coefficients $\alpha_k(d\hat{\lambda})$, $\beta_k(d\hat{\lambda})$, $k = 0, 1, \ldots, n-1$. The routine has the following calling sequence:

gchri(n, iopt, nu0, numax, eps, a, b, x, y, alpha, beta, nu, ierr, ierrc, fnu, rho, rold, s, s0, s1, s2).

On entry,

n is the number of recursion coefficients desired; type integer.

iopt is an integer identifying the type of modification as follows:
(1) $d\hat{\lambda}(t) = d\lambda(t)/(t-x)$, where x is assumed to be outside of the smallest interval containing $\text{supp}(d\lambda)$.
(2) $d\hat{\lambda}(t) = d\lambda(t)/((t-x)^2 + y^2)$, $y > 0$.

nu0 is an integer $\nu_0 \geq 2n$ estimating the starting index for the backward recursion to compute the modified moments; if no other choices are available, take **nu0** = 3 × **n**.

numax is an integer used to terminate backward recursion in case of nonconvergence; a conservative choice is **numax** = 500.

eps is a relative error tolerance; type real.

a, b are arrays of dimension **numax** to be supplied with the recursion coefficients $\mathbf{a}(k) = \alpha_{k-1}(d\lambda)$, $\mathbf{b}(k) = \beta_{k-1}(d\lambda)$, $k = 1, 2, \ldots,$ **numax**.

x, y are real parameters defining the linear and quadratic divisors of $d\lambda$.

On return,

alpha, beta are arrays of dimension **n** containing the desired recursion coefficients $\mathbf{alpha}(k) = \hat{\alpha}_{k-1}$, $\mathbf{beta}(k) = \hat{\beta}_{k-1}$, $k = 1, 2, \ldots,$ **n**.

nu is the index ν for which the error tolerance **eps** is satisfied for the first time; if it is never satisfied, **nu** will have the value **numax**.

ierr is an error flag, where
ierr = 0 on normal return,
ierr = 1 if **iopt** is inadmissible,
ierr = **nu0** if **nu0** > **numax**,
ierr = **numax** if the backward recurrence algorithm does not converge, and
ierr = −1 if **n** is not in range.

ierrc is an error flag inherited from the routine **cheb**.

The real arrays **fnu,s,s0,s1,s2** are working space, all of dimension $2 \times \mathbf{n}$, except **s**, which has dimension **n**. The complex arrays **rho**, **rold** are also working space, both of dimension $2n$. The routine calls on the subroutines **knum** and **cheb**. The double-precision version of **gchri** has the name **dgchri**.

Since the routine **gchri** is based on the modified Chebyshev algorithm, it shares with the latter its proneness to ill-conditioning, particularly in cases of measures supported on an infinite interval. On finitely supported measures, however, it can be quite effective, as seen in the next example.

Example 5.3 The performance of **chri** and **gchri**.

To illustrate the severe limitations of the routine **chri** in situations where divisions of the measure $d\lambda$ are involved, and at the same time to document the effectiveness of **gchri**, we ran both routines with $n = 40$ for Jacobi measures $d\lambda^{(\alpha,\beta)}$ with parameters $\alpha, \beta = -.8(.4).8, \beta \geq \alpha$. This is done in **test11**.

The routine **test11** first tests division by $t - x$, where $x = -1.001, -1.01, -1.04, -1.07$, and -1.1. Both routines **chri** and **gchri** are run in single and double precision, the latter with $\epsilon = 10 \times \epsilon^s$ and $\epsilon = 100 \times \epsilon^d$, respectively. The double-precision results are used to determine the absolute errors in the $\hat{\alpha}$'s and the relative errors in the $\hat{\beta}$'s for each routine. The required coefficients α_k, β_k, $0 \leq k \leq \nu_{\max} - 1$ ($\nu_{\max} = 500$ for single precision and 800 for double precision) are supplied by **recur** and **drecur** with **ipoly** = 6. The routine **nu0jac** is used to provide the starting recurrence index ν_0 (resp., ν_0^d). In Tables XII and XIII, relating, respectively, to linear and quadratic divisors, we give only the results for the Legendre measure ($\alpha = \beta = 0$). The first line in each three-line block of Table XII shows x, ν_0, ν_0^d, and the maximum (over k, $0 \leq k \leq 39$) errors in the $\hat{\alpha}_k$ and $\hat{\beta}_k$ for **gchri**, followed by the analogous information (except the ν_0's) for **chri**. The recurrence index ν yielding convergence was found (not shown in **test11**) to be $\nu = \nu_0 + 5$ and $\nu^d = \nu_0^d + 5$, without exception.

It can be seen from the leading lines in Table XII that **chri** rapidly loses accuracy as x moves away from the interval $[-1, 1]$, all single-precision accuracy being gone by the time x reaches -1.1. Similar, if not more rapid, erosion of accuracy is observed for the other parameter values of α, β. The next two lines in each three-line block show "reconstruction errors," that is, the maximum errors in the α's and β's if the $\hat{\alpha}$'s and $\hat{\beta}$'s produced by **gchri**, **chri** and **dgchri**, **dchri** are fed back to the routines **chri** and **dchri** with **iopt** = 1 to recover the original recursion coefficients in single and double precision. The first of these two lines shows the errors in reconstructing these coefficients from the output of **gchri** (resp., **dgchri**), and the second from the output of **chri** (resp., **dchri**). Rather remarkably, the coefficients are recovered to essentially full accuracy, even when the input coefficients (produced by **chri** and **dchri**) are very inaccurate! This is certainly a phenomenon that deserves further study. It can also be seen from Table XII (and the more complete results in Section 1 of the package) that **gchri** consistently produces accurate results, some slight deterioration occurring only very close to $x = -1$, where the routine has to work harder.

Table XII. Performance of **gchri** and **chri** for Elementary Divisors $t - x$ of the Legendre Measure $d\lambda(t)$

			gchri		**chri**	
x	ν_0	ν_0^d	err $\hat{\alpha}$	err $\hat{\beta}$	err $\hat{\alpha}$	err $\hat{\beta}$
−1.001	418	757	8.000(−14)	1.559(−13)	1.013(−13)	1.647(−13)
			8.527(−14)*	1.705(−13)	1.010(−27)	2.423(−27)
			1.421(−14)*	5.329(−14)	2.019(−28)	1.211(−27)
−1.010	187	294	4.016(−14)	6.907(−14)	1.396(−10)	2.424(−10)
			3.553(−14)	9.946(−14)	6.058(−28)	1.211(−27)
			7.105(−15)	4.262(−14)	1.515(−28)	9.080(−28)
−1.040	133	187	3.590(−14)	4.759(−14)	5.944(−6)	8.970(−6)
			2.842(−14)	7.103(−14)	5.554(−28)	1.312(−27)
			7.105(−15)	4.263(−14)	1.010(−28)	9.080(−28)
−1.070	120	161	2.194(−14)	4.850(−14)	5.334(−3)	7.460(−3)
			2.842(−14)	7.104(−14)	6.058(−28)	1.211(−27)
			7.105(−15)	4.263(−14)	1.010(−28)	7.062(−28)
−1.100	114	148	2.238(−14)	4.359(−14)	4.163(0)	4.959(+1)
			2.132(−14)	5.683(−14)	3.534(−28)	1.009(−27)
			1.549(−12)	1.833(−12)	1.010(−28)	6.057(−28)

*The second two lines of each three-line block show reconstruction errors.

Table XIII. Performance of **gchri** and **chri** for Elementary Divisors $(t - x)^2 + y^2$ of the Legendre Measure $d\lambda(t)$ with $z = x + iy$ on $\mathscr{E}_\rho$

			gchri		**chri**	
ρ	$\bar{\nu}_0$	$\bar{\nu}_0^d$	$\overline{\text{err}}\ \hat{\alpha}$	$\overline{\text{err}}\ \hat{\beta}$	$\overline{\text{err}}\ \hat{\alpha}$	$\overline{\text{err}}\ \hat{\beta}$
1.050	390	700	7.879(−13)	1.440(−12)	7.685(−14)	1.556(−13)
			7.814(−13)*	1.433(−12)	1.786(−26)	3.042(−26)
			2.024(−14)*	8.442(−14)	3.016(−28)	1.742(−27)
1.275	142	204	6.252(−14)	1.287(−13)	4.562(−7)	6.162(−7)
			6.554(−14)	1.279(−13)	1.541(−27)	3.061(−27)
			2.295(−14)	8.970(−14)	3.579(−28)	1.646(−27)
1.500	117	154	3.991(−14)	7.966(−14)	4.906(−1)	2.339(0)
			4.207(−14)	9.064(−14)	6.932(−28)	1.676(−27)
			3.805(−14)	8.971(−14)	4.351(−28)	1.744(−27)

*The second two lines of each three-line block show reconstruction errors.

The second half of **test11** tests division by $(t - x)^2 + y^2$, where $z = x + iy$ is taken along the upper half of the ellipse

$$\mathscr{E}_\rho = \left\{ z \in \mathbb{C}: z = \frac{1}{2}\left(\rho e^{i\vartheta} + \frac{1}{\rho} e^{-i\vartheta} \right), 0 \le \vartheta \le 2\pi \right\}, \qquad \rho > 1, \quad (5.10)$$

which has foci ± 1 and sum of the semiaxes equal to ρ. (These ellipses are contours of constant ν_0 for Jacobi measures.) We generated information analogous to the one in Table XII, for $\rho = 1.05$, 1.1625, 1.275, 1.3875, and 1.5,

except that all quantities are averaged over 19 equally spaced points on $\mathscr{E}_\rho$ corresponding to $\vartheta = j\pi/20$, $j = 1, 2, \ldots, 19$. Selected results (bars indicate averaging), again for the Legendre case, are shown in Table XIII. They reveal a behavior very similar to the one in Table XII for linear divisors.

6. GAUSS-TYPE QUADRATURE RULES

One of the important uses of orthogonal polynomials is in the approximation of integrals involving a positive measure $d\lambda$ by quadrature rules of maximum, or nearly maximum, algebraic degree of exactness. In this context, it is indispensable to know the recursion coefficients for the respective orthogonal polynomials $\{\pi_k(\cdot; d\lambda)\}$, since they allow us to generate the desired quadrature rules accurately and effectively via eigenvalue techniques. The software developed in the previous sections thus finds here a vast area of application.

6.1 Gaussian Quadrature

Given the (positive) measure $d\lambda$ (having an infinite number of support points), there exists, for each $n \in \mathbb{N}$, a quadrature rule

$$\int_{\mathbb{R}} f(t)\, d\lambda(t) = \sum_{k=1}^{n} w_k f(x_k) + R_n(f) \tag{6.1}$$

having algebraic degree of exactness $2n - 1$, that is, zero error, $R_n(f) = 0$, whenever f is a polynomial of degree $\leq 2n - 1$. The nodes x_k indeed are the zeros of the nth-degree orthogonal polynomial $\pi_n(\cdot; d\lambda)$, and the weights w_k, which are all positive, are also expressible in terms of the same orthogonal polynomials. Alternatively, and more importantly for computational purposes, the nodes x_k are the eigenvalues of the nth-order Jacobi matrix

$$J_n(d\lambda) = \begin{bmatrix} \alpha_0 & \sqrt{\beta_1} & & & 0 \\ \sqrt{\beta_1} & \alpha_1 & \sqrt{\beta_2} & & \\ & \sqrt{\beta_2} & \ddots & \ddots & \\ & & \ddots & & \sqrt{\beta_{n-1}} \\ 0 & & & \sqrt{\beta_{n-1}} & \alpha_{n-1} \end{bmatrix}, \tag{6.2}$$

where $\alpha_k = \alpha_k(d\lambda)$, $\beta_k = \beta_k(d\lambda)$ are the recurrence coefficients for the (monic) orthogonal polynomials $\{\pi_k(\cdot; d\lambda)\}$, and the weights w_k are expressible in terms of the associated eigenvectors. Specifically, if

$$J_n(d\lambda)v_k = x_k v_k, \qquad v_k^T v_k = 1, \qquad k = 1, 2, \ldots, n, \tag{6.3}$$

that is, if v_k is the normalized eigenvector of $J_n(d\lambda)$ corresponding to the eigenvalue x_k, then

$$w_k = \beta_0 v_{k,1}^2, \qquad k = 1, 2, \ldots, n, \tag{6.4}$$

where $\beta_0 = \beta_0(d\lambda)$ is defined in (1.4) and $v_{k,1}$ is the first component of v_k (cf. Golub and Welsch [1969]). There are well-known and efficient algorithms, such as the QR algorithm, to compute eigenvalues and (part of the) eigenvectors of symmetric tridiagonal matrices. These are used in the routine **gauss**,[4] whose calling sequence is as follows:

gauss(n, alpha, beta, eps, zero, weight, ierr, e).

On entry,

n is the number of terms in the Gauss formula; type integer.

alpha, beta are arrays of dimension **n** assumed to hold the recursion coefficients **alpha**$(k) = \alpha_{k-1}$, **beta**$(k) = \beta_{k-1}$, $k = 1, 2, \ldots,$ **n**.

eps is a relative error tolerance, for example, the machine precision.

On return,

zero, weight are arrays of dimension **n** containing the nodes (in increasing order) and the corresponding weights of the Gauss formula, **zero**$(k) = x_k$, **weight**$(k) = w_k$, $k = 1, 2, \ldots,$ **n**.

ierr is an error flag equal to 0 on normal return, equal to i if the QR algorithm does not converge within 30 iterations on evaluating the ith eigenvalue, equal to -1 if **n** is not in range, and equal to -2 if one of the β's is negative.

The array **e** of dimension **n** is used for working space. The double-precision routine has the name **dgauss**.

We refrain here from giving numerical examples, since the use of the routine **gauss** and the routines yet to be described is straightforward. Some use of **gauss** and **dgauss** has already been made in Examples 4.2–4.4 and 5.2.

6.2 Gauss–Radau Quadrature

We now assume that $d\lambda$ is a measure whose support is either bounded from below, bounded from above, or both. Let x_0 be either the infimum or the supremum of supp $d\lambda$, so long as it is finite. (Typically, if supp $d\lambda = [-1, 1]$, then x_0 could be either -1 or $+1$; if supp $d\lambda = [0, \infty]$, then x_0 would have to be 0; etc.). By *Gauss–Radau quadrature* we then mean a quadrature rule of maximum degree of exactness that contains among the nodes the point x_0. It thus has the form

$$\int_{\mathbb{R}} f(t)\, d\lambda(t) = w_0 f(x_0) + \sum_{k=1}^{n} w_k f(x_k) + R_n(f) \tag{6.5}$$

[4]This routine was kindly supplied to the author by Professor G. H. Golub.

and, as is well known, can be made to have a degree of exactness $2n$, that is, $R_n(f) = 0$ for all polynomials of degree $\leq 2n$. Interestingly, all nodes $x_0, x_1, \ldots, x_n$ and weights $w_0, w_1, \ldots, w_n$ can again be interpreted in terms of eigenvalues and eigenvectors, exactly as in the case of Gaussian quadrature rules, but now relative to the matrix (cf. Golub [1973])

$$J_{n+1}^*(d\lambda) = \begin{bmatrix} \alpha_0 & \sqrt{\beta_1} & & & 0 \\ \sqrt{\beta_1} & \alpha_1 & \ddots & & \\ & \ddots & \ddots & \sqrt{\beta_{n-1}} & \\ & & \sqrt{\beta_{n-1}} & \alpha_{n-1} & \sqrt{\beta_n} \\ 0 & & & \sqrt{\beta_n} & \alpha_n^* \end{bmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}, \quad (6.6)$$

where $\alpha_k = \alpha_k(d\lambda)(0 \leq k \leq n-1)$, $\beta_k = \beta_k(d\lambda)(1 \leq k \leq n)$ as before, but

$$\alpha_n^* = \alpha_n^*(d\lambda) = x_0 - \beta_n \frac{\pi_{n-1}(x_0; d\lambda)}{\pi_n(x_0; d\lambda)}. \quad (6.7)$$

Hence, we can use the routine **gauss** to generate the Gauss–Radau formula. This is done in the following subroutine:

```
      subroutine radau(n, alpha, beta, end, zero, weight, ierr, e, a, b)
      dimension alpha(*), beta(*), zero(*), weight(*), e(*), a(*), b(*)
c
c  The arrays alpha, beta, zero, weight, e, a, b are assumed to have
c  dimension n + 1.
c
      epsma = r1mach(3)
c
c  epsma is the machine single precision.
c
      np1 = n + 1
      do 10 k = 1, np1
        a(k) = alpha(k)
        b(k) = beta(k)
   10 continue
      p0 = 0.
      p1 = 1.
      do 20 k = 1, n
        pm1 = p0
        p0 = p1
        p1 = (end - a(k))*p0 - b(k)*pm1
   20 continue
      a(np1) = end - b(np1)*p0 / p1
      call gauss(np1, a, b, epsma, zero, weight, ierr, e)
      return
      end
```

The input variables are **n**, **alpha**, **beta**, and **end**, representing, respectively, n; two arrays of dimension $n+1$ containing the $\alpha_k(d\lambda)$, $\beta_k(d\lambda)$, $k = 0, 1, 2, \ldots, n$; and the "endpoint" x_0. The nodes (in increasing order) of the

Gauss–Radau formula are returned in the array **zero**, and the corresponding weights in the array **weight**. The arrays **e**, **a**, **b** are working space, and **ierr** is an error flag inherited from the routine **gauss**. The double-precision routine has the name **dradau**.

We remark that x_0 could also be outside the support of $d\lambda$, in which case the routine would generate a "Christoffel-type" quadrature rule.

6.3 Gauss - Lobatto Quadrature

Assuming now the support of $d\lambda$ bounded on either side, we let $x_0 = \inf \operatorname{supp}(d\lambda)$ and $x_{n+1} = \sup \operatorname{supp}(d\lambda)$ and consider a quadrature rule of the type

$$\int_{\mathbb{R}} f(t)\, d\lambda(t) = w_0 f(x_0) + \sum_{k=1}^{n} w_k f(x_k) + w_{n+1} f(x_{n+1}) + R_n(f) \quad (6.8)$$

having maximum degree of exactness $2n + 1$. This is called the *Gauss–Lobatto quadrature rule*. Its nodes $x_0, x_1, \ldots, x_{n+1}$ and weights $w_0, w_1, \ldots, w_{n+1}$ again admit the same spectral representation as in the case of the Gauss and Gauss–Radau formulas, only this time the matrix in question has order $n + 2$ and is given by (cf. Golub [1973])

$$J_{n+2}^*(d\lambda) = \begin{bmatrix} \alpha_0 & \sqrt{\beta_1} & & & & 0 \\ \sqrt{\beta_1} & \alpha_1 & \sqrt{\beta_2} & & & \\ & \sqrt{\beta_2} & \ddots & & & \\ & & \ddots & \ddots & \sqrt{\beta_n} & \\ & & & \sqrt{\beta_n} & \alpha_n & \sqrt{\beta_{n+1}^*} \\ 0 & & & & \sqrt{\beta_{n+1}^*} & \alpha_{n+1}^* \end{bmatrix}. \quad (6.9)$$

Here, as before, $\alpha_k = \alpha_k(d\lambda)(0 \le k \le n)$, $\beta_k = \beta_k(d\lambda)(1 \le k \le n)$, and α_{n+1}^*, β_{n+1}^* are the unique solution of the linear 2×2 system

$$\begin{bmatrix} \pi_{n+1}(x_0; d\lambda) & \pi_n(x_0; d\lambda) \\ \pi_{n+1}(x_{n+1}; d\lambda) & \pi_n(x_{n+1}; d\lambda) \end{bmatrix} \begin{bmatrix} \alpha_{n+1}^* \\ \beta_{n+1}^* \end{bmatrix} = \begin{bmatrix} x_0 \pi_{n+1}(x_0; d\lambda) \\ x_{n+1} \pi_{n+1}(x_{n+1}; d\lambda) \end{bmatrix}. \quad (6.10)$$

Hence, we have the following routine for generating the Gauss–Lobatto formulas:

```
      subroutine lob(n, alpha, beta, aleft, right, zero, weight, ierr, e, a, b)
      dimension alpha(*), beta(*), zero(*), weight(*), e(*), a(*), b(*)
c
c  The arrays alpha, beta, zero, weight, e, a, b are assumed to have
c  dimension n + 2.
c
      epsma = r1mach(3)
c
c  epsma is the machine single precision.
c
```

```
      np1 = n + 1
      np2 = n + 2
      do 10 k = 1, np2
        a(k) = alpha(k)
        b(k) = beta(k)
   10 continue
      p0l = 0.
      p0r = 0.
      p1l = 1.
      p1r = 1.
      do 20 k = 1, np1
        pm1l = p0l
        p0l = p1l
        pm1r = p0r
        p0r = p1r
        p1l = (aleft - a(k))*p0l - b(k)*pm1l
        p1r = (right - a(k))*p0r - b(k)*pm1r
   20 continue
      det = p1l*p0r - p1r*p0l
      a(np2) = (aleft*p1l*p0r - right*p1r*p0l) / det
      b(np2) = (right - aleft)*p1l*p1r / det
      call gauss(np2, a, b, epsma, zero, weight, ierr, e)
      return
      end
```

The meaning of the input and output variables is as in the routine **radau**, the variable **aleft** standing for x_0 and **right** for x_{n+1}. The double-precision routine is named **dlob**.

A remark analogous to the one after the routine **radau** applies to the routine **lob**.

ACKNOWLEDGMENTS

The author gratefully acknowledges a number of suggestions from two anonymous referees and from the associate editor, Dr. Ronald F. Boisvert, for improving the code of the package.

REFERENCES

ABRAMOWITZ, M., AND STEGUN, I. A., EDS. 1964. Handbook of mathematical functions. NBS Appl. Math. Ser. 55, U.S. Government Printing Office, Washington, D.C.

BOLEY, D., AND GOLUB, G. H. 1987. A survey of matrix inverse eigenvalue problems. *Inverse Problems 3*, 4, 595–622.

CHEBYSHEV, P. L. 1859. Sur l'interpolation par la méthode des moindres carrés. *Mém. Acad. Impér. Sci. St. Pétersbourg (7) 1*, 15, 1–24. (*Oeuvres I*, pp. 473–498.)

CHIHARA, T. S. 1978. *An Introduction to Orthogonal Polynomials*. Gordon and Breach, New York.

CHIHARA, T. S. 1985. Orthogonal polynomials and measures with end point masses. *Rocky Mountain J. Math. 15*, 3, 705–719.

CHRISTOFFEL, E. B. 1858. Über die Gaußische Quadratur und eine Verallgemeinerung derselben. *J. Reine Angew. Math 55*, 61–82. (*Ges. Math. Abhandlungen I*, pp. 65–87.)

CHRISTOFFEL, E. B. 1877. Sur une classe particulière de fonctions entières et de fractions continues. *Ann. Mat. Pura Appl. Ser. 2*, vol. 8, 1–10. (*Ges. Math. Abhandlungen II*, pp. 42–50.)

CODY, W. J., AND HILLSTROM, K. E. 1967. Chebyshev approximations for the natural logarithm of the gamma function. *Math. Comput 21*, 98 (Apr.), 198–203.

DANLOY, B. 1973. Numerical construction of Gaussian quadrature formulas for $\int_0^1(-\mathrm{Log}\ x)\cdot x^\alpha\cdot f(x)\cdot dx$ and $\int_0^\infty E_m(x)\cdot f(x)\cdot dx$. *Math. Comput. 27*, 124 (Oct.), 861–869.

DE BOOR, C., AND GOLUB, G. H. 1978. The numerically stable reconstruction of a Jacobi matrix from spectral data. *Linear Algebra Appl. 21*, 3 (Sept.), 245–260.

FRONTINI, M., GAUTSCHI, W., AND MILOVANOVIĆ, G. V. 1987. Moment-preserving spline approximation on finite intervals. *Numer. Math. 50*, 5 (Mar.), 503–518.

GALANT, D. 1969. Gauss quadrature rules for the evaluation of $2\pi^{-1/2}\int_0^\infty \exp(-x^2)f(x)\,dx$, review 42. *Math. Comput. 23*, 107 (July), 676–677. (Loose microfiche suppl. E.)

GALANT, D. 1971. An implementation of Christoffel's theorem in the theory of orthogonal polynomials. *Math. Comput. 25*, 113 (Jan.), 111–113.

GALANT, D. 1992. Algebraic methods for modified orthogonal polynomials. *Math. Comput. 59*, 200 (Oct.), 541–546.

GAUTSCHI, W. 1967a. Numerical quadrature in the presence of a singularity. *SIAM J. Numer. Anal. 4*, 3 (Sept.), 357–362.

GAUTSCHI, W. 1967b. Computational aspects of three-term recurrence relations. *SIAM Rev. 9*, 1 (Jan.), 24–82.

GAUTSCHI, W. 1979. On the preceding paper "A Legendre polynomial integral" by James L. Blue. *Math. Comput. 33*, 146 (Apr.), 742–743.

GAUTSCHI, W. 1981. Minimal solutions of three-term recurrence relations and orthogonal polynomials. *Math. Comput. 36*, 154 (Apr.), 547–554.

GAUTSCHI, W. 1982a. On generating orthogonal polynomials. *SIAM J. Sci. Stat. Comput. 3*, 3 (Sept.), 289–317.

GAUTSCHI, W. 1982b. An algorithmic implementation of the generalized Christoffel theorem. In *Numerical Integration*, G. Hämmerlin, Ed. International Series of Numerical Mathematics, vol. 57. Birkhäuser, Basel, pp. 89–106.

GAUTSCHI, W. 1984a. Discrete approximations to spherically symmetric distributions. *Numer. Math. 44*, 1 (June), 53–60.

GAUTSCHI, W. 1984b. Questions of numerical condition related to polynomials. In *Studies in Mathematics*, vol. 24. G. H. Golub, Ed. Studies in Numerical Analysis. Mathematical Association of America, Washington, D.C., pp. 140–177.

GAUTSCHI, W. 1986a. On the sensitivity of orthogonal polynomials to perturbations in the moments. *Numer. Math. 48*, 4 (Apr.), 369–382.

GAUTSCHI, W. 1986b. Reminiscences of my involvement in de Branges's proof of the Bieberbach conjecture. In *The Bieberbach Conjecture*. Mathematical Surveys and Monographs, no. 21. American Mathematical Society, Providence, R.I., pp. 205–211.

GAUTSCHI, W. 1990. Computational aspects of orthogonal polynomials. In *Orthogonal Polynomials—Theory and Practice*, P. Nevai, Ed. NATO ASI Series, Series C: Mathematical and Physical Sciences, vol. 294. Kluwer, Dordrecht, pp. 181–216.

GAUTSCHI, W. 1991a. A class of slowly convergent series and their summation by Gaussian quadrature. *Math. Comput. 57*, 195 (July), 309–324.

GAUTSCHI, W. 1991b. On certain slowly convergent series occurring in plate contact problems. *Math. Comput. 57*, 195 (July), 325–338.

GAUTSCHI, W. 1991c. Quadrature formulae on half-infinite intervals. *BIT 31*, 3, 438–446.

GAUTSCHI, W. 1991d. Computational problems and applications of orthogonal polynomials. In *Orthogonal Polynomials and Their Applications*. IMACS Annals on Computing and Applied Mathematics, vol. 9. Baltzer, Basel, pp. 61–71.

GAUTSCHI, W. 1993a. Gauss-type quadrature rules for rational functions. In *Numerical Integration IV*, H. Brass and G. Hämmerlin, Eds. International Series of Numerical Mathematics, vol. 112. Birhäuser, Basel, 111–130.

GAUTSCHI, W. 1993b. On the computation of generalized Fermi–Dirac and Bose–Einstein integrals. *Comput. Phys. Commun. 74*, 2 (Feb.), 233–238.

GAUTSCHI, W. 1993c. Is the recurrence relation for orthogonal polynomials always stable? *BIT 33*, 2, 277–284.

GAUTSCHI, W., AND LI, S. 1993. A set of orthogonal polynomials induced by a given orthogonal polynomial. *Aequationes Math. 46*, 1/2 (Aug.), 174–198.

GAUTSCHI, W., AND MILOVANOVIĆ, G. V. 1985. Gaussian quadrature involving Einstein and Fermi functions with an application to summation of series. *Math Comput. 44*, 169 (Jan.), 177–190.

GAUTSCHI, W., AND MILOVANOVIĆ, G V. 1986. Spline approximations to spherically symmetric distributions. *Numer. Math. 49*, 2/3 (July), 111–121.

GAUTSCHI, W., AND VARGA, R. S. 1983. Error bounds for Gaussian quadrature of analytic functions. *SIAM J. Numer. Anal. 20*, 6 (Dec.), 1170–1186.

GAUTSCHI, W., AND WIMP, J. 1987. Computing the Hilbert transform of a Jacobi weight function. *BIT 27*, 2, 203–215.

GAUTSCHI, W., KOVAČEVIĆ, M. A., AND MILOVANOVIĆ, G. V. 1987. The numerical evaluation of singular integrals with coth-kernel. *BIT 27*, 3, 389–402.

GOLUB, G. H. 1973. Some modified matrix eigenvalue problems. *SIAM Rev. 15*, 2 (Apr.), 318–334.

GOLUB, G. H., AND WELSCH, J. H. 1969. Calculation of Gauss quadrature rules. *Math. Comput. 23*, 106 (Apr.), 221–230.

GRAGG, W. B., AND HARROD, W. J. 1984. The numerically stable reconstruction of Jacobi matrices from spectral data. *Numer. Math. 44*, 3 (Sept.), 317–335.

KAUTSKY, J., AND GOLUB, G. H. 1983. On the calculation of Jacobi matrices. *Linear Algebra Appl. 52-53* (July), 439–455.

LUKE, Y. L. 1975. *Mathematical Functions and Their Approximations*. Academic Press, New York.

REES, C. J. 1945. Elliptic orthogonal polynomials. *Duke Math J. 12*, 173–187.

RUTISHAUSER, H. 1963. On Jacobi rotation patterns. In *Experimental Arithmetic, High Speed Computing and Mathematics*. Proceedings of Symposia in Applied Mathematics, vol. *15*. American Mathematical Society, Providence, R.I., pp. 219–239.

SACK, R. A., AND DONOVAN, A. F. 1972. An algorithm for Gaussian quadrature given modified moments. *Numer. Math. 18*, 5 (Mar.), 465–478.

STROUD, A. H., AND SECREST, D 1966. *Gaussian Quadrature Formulas* Prentice-Hall, Englewood Cliffs, N.J.

UVAROV, V. B. 1959. Relation between polynomials orthogonal with different weights. *Dokl. Akad. Nauk SSSR 126*, 1, 33–36. (In Russian.)

UVAROV, V. B. 1969. The connection between systems of polynomials that are orthogonal with respect to different distribution functions. *Ž Vyčisl. Mat. i Mat. Fiz. 9*, 6, 1253–1262. (In Russian.)

WHEELER, J. C. 1974. Modified moments and Gaussian quadrature. *Rocky Mountain J. Math. 4*, 2, 287–296.

WILKINSON, J. H. 1965. *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford.

Received February 1992; revised December 1992 and March 1993; accepted March 1993

28.2. [179] "Orthogonal Polynomials, Quadrature, and Approximation: Computational Methods and Software (in Matlab)"

---

[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 (2006).

Orthogonal Polynomials, Quadrature, and Approximation: Computational Methods and Software (in Matlab)

Walter Gautschi

Department of Computer Sciences, Purdue University,
West Lafayette, IN 47907, USA
e-mail: wxg@cs.purdue.edu

Summary. One of the main problems in the constructive theory of orthogonal polynomials is the computation of the coefficients, if not known explicitly, in the three-term recurrence relation satisfied by orthogonal polynomials. Two classes of methods are discussed: those based on moment information, and those using discretization of the underlying inner product. Other computational problems considered are the computation of Cauchy integrals of orthogonal polynomials, and the problem of modification, i.e., of ascertaining the effect on the recurrence coefficients of multiplying the weight function by a (positive) rational function. Moment-based methods and discretization algorithms are also available for generating Sobolev orthogonal polynomials, i.e., polynomials orthogonal with respect to an inner product involving derivatives. Of particular interest here is the computation of their zeros.

Important applications of orthogonal polynomials are to the development of quadrature rules of maximum algebraic degree of exactness, most notably Gauss-type quadrature rules, but also Gauss-Kronrod and Gauss-Turán quadratures. Modification algorithms and discretization methods find application to constructing quadrature rules exact not only for polynomials, but also for rational functions with prescribed poles. Gauss-type quadrature rules are applicable also for computing Cauchy principal value integrals. Gaussian quadrature sums are expressible in terms of the related Jacobi matrix, which has interesting applications to generating orthogonal polynomials on several intervals and to the estimation of matrix functionals.

In the realm of approximation, the classical use of orthogonal polynomials, including Sobolev orthogonal polynomials, is to least squares approximation to which interpolatory constraints may be added. Among other uses considered are moment-preserving spline approximation and the summation of slowly convergent series.

All computational methods and applications considered are supported by a software package, called `OPQ`, of Matlab routines which are downloadable individually from the internet. Their use is illustrated throughout.

1 Introduction

Orthogonal polynomials, unless they are classical, require special techniques for their computation. One of the central problems is to generate the coefficients in the basic three-term recurrence relation they are known to satisfy. There are two general approaches for doing this: methods based on moment information, and discretization methods. In the former, one develops algorithms that take as input given moments, or modified moments, of the underlying measure and produce as output the desired recurrence coefficients. In theory, these algorithms yield exact answers. In practice, owing to rounding errors, the results are potentially inaccurate depending on the numerical condition of the mapping from the given moments (or modified moments) to the recurrence coefficients. A study of related condition numbers is therefore

of practical interest. In contrast to moment-based algorithms, discretization methods are basically approximate methods: one approximates the underlying inner product by a discrete inner product and takes the recurrence coefficients of the corresponding discrete orthogonal polynomials to approximate those of the desired orthogonal polynomials. Finding discretizations that yield satisfactory rates of convergence requires a certain amount of skill and creativity on the part of the user, although general-purpose discretizations are available if all else fails.

Other interesting problems have as objective the computation of new orthogonal polynomials out of old ones. If the measure of the new orthogonal polynomials is the measure of the old ones multiplied by a rational function, one talks about modification of orthogonal polynomials and modification algorithms that carry out the transition from the old to the new orthogonal polynomials. This enters into a circle of ideas already investigated by Christoffel in the 1850s, but effective algorithms have been obtained only very recently. They require the computation of Cauchy integrals of orthogonal polynomials — another interesting computational problem.

In the 1960s, a new type of orthogonal polynomials emerged — the so-called Sobolev orthogonal polynomials — which are based on inner products involving derivatives. Although they present their own computational challenges, moment-based algorithms and discretization methods are still two of the main working tools. The computation of zeros of Sobolev orthogonal polynomials is of particular interest in practice.

An important application of orthogonal polynomials is to quadrature, specifically quadrature rules of the highest algebraic degree of exactness. Foremost among them is the Gaussian quadrature rule and its close relatives, the Gauss–Radau and Gauss–Lobatto rules. More recent extensions are due to Kronrod, who inserts $n + 1$ new nodes into a given n-point Gauss formula, again optimally with respect to degree of exactness, and to Turán, who allows derivative terms to appear in the quadrature sum. When integrating functions having poles outside the interval of integration, quadrature rules of polynomial/rational degree of exactness are of interest. Poles inside the interval of integration give rise to Cauchy principal value integrals, which pose computational problems of their own. Interpreting Gaussian quadrature sums in terms of matrices allows interesting applications to orthogonal polynomials on several intervals, and to the computation of matrix functionals.

In the realm of approximation, orthogonal polynomials, especially discrete ones, find use in curve fitting, e.g. in the least squares approximation of discrete data. This indeed is the problem in which orthogonal polynomials (in substance if not in name) first appeared in the 1850s in work of Chebyshev. The presence of interpolatory constraints can be handled by a modification algorithm relative to special quadratic factors. Sobolev orthogonal polynomials also had their origin in least squares approximation, when one tries to fit simultaneously functions together with some of their derivatives. Physically motivated are approximations by spline functions that preserve as many

moments as possible. Interestingly, these also are related to orthogonal polynomials via Gauss and generalized Gauss-type quadrature formulae. Slowly convergent series whose sum can be expressed as a definite integral naturally invite the application of Gauss-type quadratures to speed up their convergence. An example are series whose general term is expressible in terms of the Laplace transform or its derivative of a known function. Such series occur prominently in plate contact problems.

A comprehensive package, called `OPQ`, of Matlab routines is available that can be used to work with orthogonal polynomials. It resides at the web site

`http://www.cs.purdue.edu/archives/2002/wxg/codes/`

and all its routines are downloadable individually.

2 Orthogonal Polynomials

2.1 Recurrence Coefficients

Background and Notation

Orthogonality is defined with respect to an inner product, which in turn involves a measure of integration, $d\lambda$. An *absolutely continuous* measure has the form

$$d\lambda(t) = w(t)dt \quad \text{on } [a,b], \qquad -\infty \le a < b \le \infty,$$

where w is referred to as a *weight function*. Usually, w is positive on (a,b), in which case $d\lambda$ is said to be a *positive measure* and $[a,b]$ is called the *support* of $d\lambda$. A *discrete measure* has the form

$$d\lambda_N(t) = \sum_{k=1}^{N} w_k \delta(t - x_k)dt, \qquad x_1 < x_2 < \cdots < x_N,$$

where δ is the Dirac delta function, and usually $w_k > 0$. The support of $d\lambda_N$ consists of its N *support points* $x_1, x_2, \ldots, x_N$. For absolutely continuous measures, we make the standing assumption that all *moments*

$$\mu_r = \int_{\mathbb{R}} t^r d\lambda(t), \qquad r = 0, 1, 2, \ldots,$$

exist and are finite. The *inner product* of two polynomials p and q relative to the measure $d\lambda$ is then well defined by

$$(p,q)_{d\lambda} = \int_{\mathbb{R}} p(t)q(t)d\lambda(t),$$

and the *norm* of a polynomial p by

$$\|p\|_{d\lambda} = \sqrt{(p,p)_{d\lambda}}.$$

Orthogonal polynomials relative to the (positive) measure $d\lambda$ are defined by

$$\pi_k(\cdot) = \pi_k(\cdot\,; d\lambda) \text{ a polynomial of exact degree } k,\ k = 0, 1, 2, \ldots,$$

$$(\pi_k, \pi_\ell)_{d\lambda} \begin{cases} = 0, & k \neq \ell, \\ > 0, & k = \ell. \end{cases}$$

There are infinitely many, if $d\lambda$ is absolutely continuous, and they are uniquely defined up to the leading coefficient. If all leading coefficients are equal to 1, they are said to be *monic*. For a discrete measure $d\lambda_N$, there are exactly N orthogonal polynomials $\pi_0, \pi_1, \ldots, \pi_{N-1}$. *Orthonormal polynomials* are defined and denoted by

$$\tilde{\pi}_k(\cdot\,; d\lambda) = \frac{\pi_k(\cdot\,; d\lambda)}{\|\pi_k\|_{d\lambda}}, \quad k = 0, 1, 2, \ldots.$$

They satisfy

$$(\tilde{\pi}_k, \tilde{\pi}_\ell)_{d\lambda} = \delta_{k,\ell} = \begin{cases} 0, & k \neq \ell, \\ 1, & k = \ell. \end{cases}$$

Examples of measures resp. weight functions are shown in Tables 1 and 2. The former displays the most important "classical" weight functions, the latter the best-known discrete measures.

Three-Term Recurrence Relation

For any n ($< N-1$ if $d\lambda = d\lambda_N$), the first $n+1$ monic orthogonal polynomials satisfy a three-term recurrence relation

$$\begin{gathered} \pi_{k+1}(t) = (t - \alpha_k)\pi_k(t) - \beta_k \pi_{k-1}(t), \quad k = 0, 1, \ldots, n-1, \\ \pi_{-1}(t) = 0, \quad \pi_0(t) = 1, \end{gathered} \tag{2.1}$$

where the *recurrence coefficients* $\alpha_k = \alpha_k(d\lambda)$, $\beta_k = \beta_k(d\lambda)$ are real and positive, respectively. The coefficient β_0 in (2.1) multiplies $\pi_{-1} = 0$, and hence can be arbitrary. For later use, it is convenient to define

Table 1. "Classical" weight functions $d\lambda(t) = w(t)dt$

name	$w(t)$	support	comment
Jacobi	$(1-t)^\alpha(1+t)^\beta$	$[-1, 1]$	$\alpha > -1$, $\beta > -1$
Laguerre	$t^\alpha e^{-t}$	$[0, \infty]$	$\alpha > -1$
Hermite	$\lvert t\rvert^{2\alpha} e^{-t^2}$	$[-\infty, \infty]$	$\alpha > -\frac{1}{2}$
Meixner-Pollaczek	$\frac{1}{2\pi} e^{(2\phi-\pi)t} \lvert\Gamma(\lambda + it)\rvert^2$	$[-\infty, \infty]$	$\lambda > 0$, $0 < \phi < \pi$

Table 2. "Classical" discrete measures $\mathrm{d}\lambda(t) = \sum_{k=0}^{M} w_k \delta(t-k)\mathrm{d}t$

name	M	w_k	comment
discrete Chebyshev	$N-1$	1	
Krawtchouk	N	$\binom{N}{k} p^k (1-p)^{N-k}$	$0 < p < 1$
Charlier	∞	$\mathrm{e}^{-a} a^k / k!$	$a > 0$
Meixner	∞	$\frac{c^k}{\Gamma(\beta)} \frac{\Gamma(k+\beta)}{k!}$	$0 < c < 1, \beta > 0$
Hahn	N	$\binom{\alpha+k}{k}\binom{\beta+N-k}{N-k}$	$\alpha > -1, \beta > -1$

$$\beta_0 = \beta_0(\mathrm{d}\lambda) = \int_{\mathbb{R}} \mathrm{d}\lambda(t). \tag{2.2}$$

The proof of (2.1) is rather simple if one expands $\pi_{k+1}(t) - t\pi_k(t) \in \mathbb{P}_k$ in orthogonal polynomials $\pi_0, \pi_1, \ldots, \pi_k$ and observes orthogonality and the obvious, but important, property $(tp, q)_{\mathrm{d}\lambda} = (p, tq)_{\mathrm{d}\lambda}$ of the inner product. As a by-product of the proof, one finds the formulae of Darboux,

$$\begin{aligned} \alpha_k(\mathrm{d}\lambda) &= \frac{(t\pi_k, \pi_k)_{\mathrm{d}\lambda}}{(\pi_k, \pi_k)_{\mathrm{d}\lambda}}, \quad k = 0, 1, 2, \ldots, \\ \beta_k(\mathrm{d}\lambda) &= \frac{(\pi_k, \pi_k)_{\mathrm{d}\lambda}}{(\pi_{k-1}, \pi_{k-1})_{\mathrm{d}\lambda}}, \quad k = 1, 2, \ldots . \end{aligned} \tag{2.3}$$

The second yields

$$\|\pi_k\|_{\mathrm{d}\lambda}^2 = \beta_0 \beta_1 \cdots \beta_k. \tag{2.4}$$

Placing the coefficients α_k on the diagonal, and $\sqrt{\beta_k}$ on the two side diagonals of a matrix produces what is called the *Jacobi matrix* of the measure $\mathrm{d}\lambda$,

$$\boldsymbol{J}(\mathrm{d}\lambda) = \begin{bmatrix} \alpha_0 & \sqrt{\beta_1} & & \mathbf{0} \\ \sqrt{\beta_1} & \alpha_1 & \sqrt{\beta_2} & \\ & \sqrt{\beta_2} & \alpha_2 & \ddots \\ & & \ddots & \ddots \\ \mathbf{0} & & & \end{bmatrix}. \tag{2.5}$$

It is a real, symmetric, tridiagonal matrix of infinite order, in general. Its principal minor matrix of order n will be denoted by

$$\boldsymbol{J}_n(\mathrm{d}\lambda) = \boldsymbol{J}(\mathrm{d}\lambda)_{[1:n,1:n]}. \tag{2.6}$$

Noting that the three-term recurrence relation for the ortho*normal* polynomials is

$$\sqrt{\beta_{k+1}}\tilde{\pi}_{k+1}(t) = (t-\alpha_k)\tilde{\pi}_k(t) - \sqrt{\beta_k}\tilde{\pi}_{k-1}(t), \quad k = 0, 1, 2, \ldots,$$
$$\tilde{\pi}_{-1}(t) = 0, \quad \tilde{\pi}_0(t) = 1/\sqrt{\beta_0}, \tag{2.7}$$

or, in matrix form, with $\tilde{\boldsymbol{\pi}}(t) = [\tilde{\pi}_0(t), \tilde{\pi}_1(t), \ldots, \tilde{\pi}_{n-1}(t)]^{\mathrm{T}}$,

$$t\tilde{\boldsymbol{\pi}}(t) = \boldsymbol{J}_n(\mathrm{d}\lambda)\tilde{\boldsymbol{\pi}}(t) + \sqrt{\beta_n}\tilde{\pi}_n(t)\boldsymbol{e}_n, \tag{2.8}$$

one sees that *the zeros* τ_ν *of* $\tilde{\pi}_n(\,\cdot\,;\mathrm{d}\lambda)$ *are precisely the eigenvalues of* $\boldsymbol{J}_n(\mathrm{d}\lambda)$, *and* $\tilde{\boldsymbol{\pi}}(\tau_\nu)$ *corresponding eigenvectors.* This is only one of many reasons why knowledge of the Jacobi matrix, i.e. of the recurrence coefficients, is of great practical interest. For classical measures as the ones in Tables 1 and 2, all recurrence coefficients are explicitly known (cf. [10, Tables 1.1 and 1.2]). In most other cases, they must be computed numerically.

In the `OPQ` package, routines generating recurrence coefficients have the syntax `ab=r_`*name*`(N)`, where *name* identifies the name of the orthogonal polynomial and `N` is an input parameter specifying the number of α_k and of β_k desired. There may be additional input parameters. The αs and βs are stored in the `N`×2 array `ab`:

α_0	β_0
α_1	β_1
$\vdots$	$\vdots$
α_{N-1}	β_{N-1}

$N \in \mathbb{N}$.

For example, `ab=r_jacobi(N,a,b)` generates the first `N` recurrence coefficients of the (monic) Jacobi polynomials with parameters α=`a`, β=`b`.

Demo#1 The first ten recurrence coefficients for the Jacobi polynomials with parameters $\alpha = -\frac{1}{2}$, $\beta = \frac{3}{2}$.

The Matlab command, followed by the output, is shown in the box below.

```
>> ab=r_jacobi(10,-.5,1.5)
ab =
    6.666666666666666e-01 4.712388980384690e+00
    1.333333333333333e-01 1.388888888888889e-01
    5.714285714285714e-02 2.100000000000000e-01
    3.174603174603174e-02 2.295918367346939e-01
    2.020202020202020e-02 2.376543209876543e-01
    1.398601398601399e-02 2.417355371900826e-01
    1.025641025641026e-02 2.440828402366864e-01
    7.843137254901961e-03 2.455555555555556e-01
    6.191950464396285e-03 2.465397923875433e-01
    5.012531328320802e-03 2.472299168975069e-01
```

2.2 Modified Chebyshev Algorithm

The first $2n$ moments $\mu_0, \mu_1, \ldots, \mu_{2n-1}$ of a measure $\mathrm{d}\lambda$ uniquely determine the first n recurrence coefficients $\alpha_k(\mathrm{d}\lambda)$ and $\beta_k(\mathrm{d}\lambda)$, $k = 0, 1, \ldots, n-1$. However, the corresponding moment map $\mathbb{R}^{2n} \mapsto \mathbb{R}^{2n} : [\mu_k]_{k=0}^{2n-1} \mapsto [\alpha_k, \beta_k]_{k=0}^{n-1}$ is severely ill-conditioned when n is large. Therefore, other moment maps must be sought that are better conditioned. One that has been studied extensively in the literature is based on *modified moments*

$$m_k = \int_{\mathbb{R}} p_k(t)\mathrm{d}\lambda(t), \quad k = 0, 1, 2, \ldots, \tag{2.9}$$

where $\{p_k\}$, $p_k \in \mathbb{P}_k$, is a given system of polynomials chosen to be close in some sense to the desired polynomials $\{\pi_k\}$. We assume that p_k, like π_k, satisfies a three-term recurrence relation

$$\begin{aligned} p_{k+1}(t) &= (t - a_k)p_k(t) - b_k\pi_{k-1}(t), \quad k = 0, 1, 2, \ldots, \\ p_{-1}(t) &= 0, \quad p_0(t) = 1, \end{aligned} \tag{2.10}$$

but with coefficients $a_k \in \mathbb{R}$, $b_k \geq 0$, that are known. The case $a_k = b_k = 0$ yields powers $p_k(t) = t^k$, hence ordinary moments μ_k, which however, as already mentioned, is not recommended.

The modified moment map

$$\mathbb{R}^{2n} \mapsto \mathbb{R}^{2n} : \quad [m_k]_{k=0}^{2n-1} \mapsto [\alpha_k, \beta_k]_{k=0}^{n-1} \tag{2.11}$$

and related maps have been well studied from the point of view of conditioning (cf. [10, §2.1.5 and 2.1.6]). The maps are often remarkably well-conditioned, especially for measures supported on a finite interval, but can still be ill-conditioned otherwise.

An algorithm that implements the map (2.11) is the *modified Chebyshev algorithm* (cf. [10, §2.1.7]), which improves on Chebyshev's original algorithm based on ordinary moments. To describe it, we need the *mixed moments*

$$\sigma_{k\ell} = \int_{\mathbb{R}} \pi_k(t; \mathrm{d}\lambda)p_\ell(t)\mathrm{d}\lambda(t), \quad k, \ell \geq -1, \tag{2.12}$$

which by orthogonality are clearly zero if $\ell < k$.

Algorithm 1 Modified Chebyshev algorithm

initialization:

$$\begin{aligned} &\alpha_0 = a_0 + m_1/m_0, \quad \beta_0 = m_0, \\ &\sigma_{-1,\ell} = 0, \quad \ell = 1, 2, \ldots, 2n-2, \\ &\sigma_{0,\ell} = m_\ell, \quad \ell = 0, 1, \ldots, 2n-1 \end{aligned}$$

continuation (if $n > 1$): for $k = 1, 2, \ldots, n-1$ do

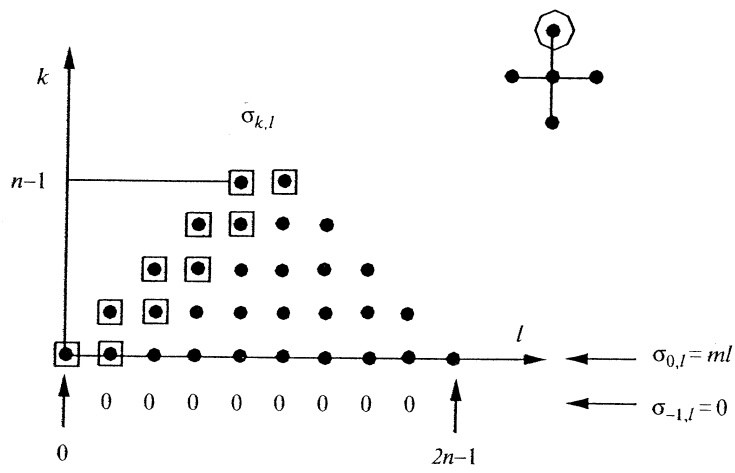


Fig. 1. Modified Chebyshev algorithm, schematically

$$\sigma_{k\ell} = \sigma_{k-1,\ell+1} - (\alpha_{k-1} - a_\ell)\sigma_{k-1,\ell} - \beta_{k-1}\sigma_{k-2,\ell} + b_\ell\sigma_{k-1,k-1}, \quad \ell = k, k+1, \ldots, 2n-k-1,$$

$$\alpha_k = a_k + \frac{\sigma_{k,k+1}}{\sigma_{kk}} - \frac{\sigma_{k-1,k}}{\sigma_{k-1,k-1}}, \quad \beta_k = \frac{\sigma_{kk}}{\sigma_{k-1,k-1}}.$$

If $a_k = b_k = 0$, Algorithm 1 reduces to Chebyshev's original algorithm.

Fig. 1 depicts the trapezoidal array of the mixed moments and the computing stencil indicating that the circled entry is computed in terms of the four entries below. The entries in boxes are used to compute the αs and βs.

The `OPQ` Matlab command that implements the modified Chebyshev algorithm has the form `ab=chebyshev(N,mom,abm)`, where `mom` is the 1×2`N` array of the modified moments, and `abm` the $(2$`N`$-1)\times 2$ array of the recurrence coefficients a_k, b_k from (2.10) needed in Algorithm 1:

m_0	m_1	m_2	$\cdots$	m_{2N-1}

`mom`

a_0	b_0
a_1	b_1
$\vdots$	$\vdots$
a_{2N-2}	b_{2N-2}

`abm`

If the input parameter `abm` is omitted, the routine assumes $a_k = b_k = 0$ and implements Chebyshev's original algorithm.

Demo#2 "Elliptic" orthogonal polynomials.
These are orthogonal relative to the measure

$$d\lambda(t) = [(1-\omega^2 t^2)(1-t^2)]^{-1/2} dt \quad \text{on } [-1,1], \quad 0 \le \omega < 1.$$

To apply the modified Chebyshev algorithm, it seems natural to employ Chebyshev moments (i.e. p_k = the monic Chebyshev polynomial of degree k)

$$m_0 = \int_{-1}^{1} d\lambda(t), \quad m_k = \frac{1}{2^{k-1}} \int_{-1}^{1} T_k(t) d\lambda(t), \quad k \ge 1.$$

Their computation is not entirely trivial (cf. [10, Example 2.29]), but a stable algorithm is available as `OPQ` routine `mm_ell.m`, which for given `N` generates the first `2N` modified moments of $d\lambda$ with ω^2 being input via the parameter `om2`. The complete Matlab routine is as follows:

```
function ab=r_elliptic(N,om2)
   abm=r_jacobi(2*N-1,-1/2);
   mom=mm_ell(N,om2);
   ab=chebyshev(N,mom,abm)
```

For `om2=.999` and `N=40`, results produced by the routine are partially shown in the box below.

```
ab =
     0 9.682265121100620e+00
     0 7.937821421385184e-01
     0 1.198676724605757e-01
     0 2.270401183698990e-01
     0 2.410608787266061e-01
     0 2.454285325203698e-01
     0 2.473016530297635e-01
     0 2.482587060199245e-01
     : :
     0 2.499915376529289e-01
     0 2.499924312667191e-01
     0 2.499932210069769e-01
```

Clearly, $\beta_k \to \frac{1}{4}$ as $k \to \infty$, which is consistent with the fact that $d\lambda$ belongs to the Szegő class (cf. [10, p. 12]). Convergence, in fact, is monotone for $k \ge 2$.

2.3 Discrete Stieltjes and Lanczos Algorithm

Computing the recurrence coefficients of a discrete measure is a prerequisite for discretization methods to be discussed in the next section. Given the measure

$$d\lambda_N(t) = \sum_{k=1}^{N} w_k \delta(t - x_k) dt, \tag{2.13}$$

the problem is to compute $\alpha_{\nu,N} = \alpha_\nu(d\lambda_N)$, $\beta_{\nu,N} = \beta_\nu(d\lambda_N)$ for all $\nu \leq n-1$, $n \leq N$, which will provide access to the discrete orthogonal polynomials of degrees up to n, or else, to determine the Jacobi matrix $\boldsymbol{J}_N(d\lambda_N)$, which will provide access to all discrete orthogonal polynomials. There are two methods in use, a discrete Stieltjes procedure and a Lanczos-type algorithm.

Discrete Stieltjes Procedure

Since the inner product for the measure (2.13) is a finite sum,

$$(p, q)_{d\lambda_N} = \sum_{k=1}^{N} w_k p(x_k) q(x_k), \tag{2.14}$$

Darboux's formulae (2.3) seem to offer attractive means of computing the desired recurrence coefficients, since all inner products appearing in these formulae are finite sums. The only problem is that we do not yet know the orthogonal polynomials $\pi_k = \pi_{k,N}$ involved. For this, however, we can make use of an idea already expressed by Stieltjes in 1884: combine Darboux's formulae with the basic three-term recurrence relation. Indeed, when $k = 0$ we know that $\pi_{0,N} = 1$, so that Darboux's formula for $\alpha_0(d\lambda_N)$ can be applied, and $\beta_0(d\lambda_N)$ is simply the sum of the weights w_k. Now that we know $\alpha_0(d\lambda_N)$, we can apply the recurrence relation (2.1) for $k = 0$ to compute $\pi_{1,N}(t)$ for $t = x_k$, $k = 1, 2, \ldots, N$. We then have all the information at hand to reapply Darboux's formulae for $\alpha_{1,N}$ and $\beta_{1,N}$, which in turn allows us to compute $\pi_{2,N}(t)$ for all $t = x_k$ from (2.1). In this manner we proceed until all $\alpha_{\nu,N}$, $\beta_{\nu,N}$, $\nu \leq n-1$, are determined. If $n = N$, this will yield the Jacobi matrix $\boldsymbol{J}_N(d\lambda_N)$.

The procedure is quite effective, at least when $n \ll N$. As n approaches N, instabilities may develop, particularly if the support points x_k of $d\lambda_N$ are equally, or nearly equally, spaced.

The `OPQ` routine implementing Stieltjes's procedure is called by `ab=stieltjes(n,xw)`, where `n` $\leq N$, and `xw` is an $N \times 2$ array containing the support points and weights of the inner product,

x_1	w_1
x_2	w_2
$\vdots$	$\vdots$
x_N	w_N

`xw`

As usual, the recurrence coefficients $\alpha_{\nu,N}$, $\beta_{\nu,N}$, $0 \leq \nu \leq n-1$, are stored in the `n×2` array `ab`.

Lanczos-type Algorithm

Lanczos's algorithm is a general procedure to orthogonally tridiagonalize a given symmetric matrix $\boldsymbol{A}$. Thus, it finds an orthogonal matrix $\boldsymbol{Q}$ and a symmetric tridiagonal matrix $\boldsymbol{T}$ such that $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{T}$. Both $\boldsymbol{Q}$ and $\boldsymbol{T}$ are uniquely determined by the first column of $\boldsymbol{Q}$.

Given the measure (2.13), it is known that an orthogonal matrix $\boldsymbol{Q} \in \mathbb{R}^{(N+1)\times(N+1)}$ exists, with the first column being $\boldsymbol{e}_1 = [1, 0, \ldots, 0]^{\mathrm{T}} \in \mathbb{R}^{N+1}$, such that (see [10, Corollary to Theorem 3.1])

$$\boldsymbol{Q}^{\mathrm{T}} \begin{bmatrix} 1 & \sqrt{w_1} & \sqrt{w_2} & \cdots & \sqrt{w_N} \\ \sqrt{w_1} & x_1 & 0 & \cdots & 0 \\ \sqrt{w_2} & 0 & x_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sqrt{w_N} & 0 & 0 & \cdots & x_N \end{bmatrix} \boldsymbol{Q} = \begin{bmatrix} 1 & \sqrt{\beta_0} & 0 & \cdots & 0 \\ \sqrt{\beta_0} & \alpha_0 & \sqrt{\beta_1} & \cdots & 0 \\ 0 & \sqrt{\beta_1} & \alpha_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \alpha_{N-1} \end{bmatrix}, \tag{2.15}$$

where $\alpha_k = \alpha_{k,N}$, $\beta_k = \beta_{k,N}$. We are thus in the situation described above, where $\boldsymbol{A}$ is the matrix displayed on the left and $\boldsymbol{T}$ the matrix on the right, the desired Jacobi matrix $\boldsymbol{J}_N(\mathrm{d}\lambda_N)$ bordered by a first column and a first row containing β_0. The computation can be arranged so that only the leading principal minor matrix of order $n+1$ is obtained.

Lanczos's algorithm in its original form (published in 1950) is numerically unstable, but can be stabilized using ideas of Rutishauser (1963). An algorithm and pseudocode of Gragg and Harrod [14], using a sequence of Givens rotations to construct the matrix $\boldsymbol{Q}$ in (2.15), forms the basis for the `OPQ` Matlab code `ab=lanczos(n,xw)`, where the input and output parameters have the same meaning as in the routine `stieltjes.m`.

This routine enjoys good stability properties but may be considerably slower than Stieltjes's procedure.

2.4 Discretization Methods

The basic idea is to discretize the given measure $\mathrm{d}\lambda$, i.e. approximate it by a discrete measure

$$\mathrm{d}\lambda(t) \approx \mathrm{d}\lambda_N(t), \tag{2.16}$$

and then use the recurrence coefficients $\alpha_k(\mathrm{d}\lambda_N)$, $\beta_k(\mathrm{d}\lambda_N)$ of the discrete measure to approximate $\alpha_k(\mathrm{d}\lambda)$, $\beta_k(\mathrm{d}\lambda)$. The former are computed by either Stieltjes's procedure or a Lanczos-type algorithm. The effectiveness of the method is crucially tied to the quality of the discretization. We illustrate this by a simple, yet instructive, example.

Example 1. Chebyshev weight function plus a constant,

$$w(t) = (1-t^2)^{-1/2} + c \quad \text{on } [-1,1], \quad c > 0.$$

It suffices to approximate the inner product for the weight function w. This can always be done by using appropriate quadrature formulae. In the case at

hand, it is natural to treat the two parts of the weight function separately, indeed to use Gauss–Chebyshev quadrature for the first part and Gauss–Legendre quadrature for the second,

$$(p,q)_w = \int_{-1}^{1} p(t)q(t)(1-t^2)^{-1/2}\mathrm{d}t + c\int_{-1}^{1} p(t)q(t)\mathrm{d}t$$
$$\approx \sum_{k=1}^{M} w_k^{\mathrm{Ch}} p(x_k^{\mathrm{Ch}})q(x_k^{\mathrm{Ch}}) + c\sum_{k=1}^{M} w_k^{\mathrm{L}} p(x_k^{\mathrm{L}})q(x_k^{\mathrm{L}}). \qquad (2.17)$$

Here, x_k^{Ch}, w_k^{Ch} are the nodes and weights of the M-point Gauss–Chebyshev quadrature rule, and x_k^{L}, w_k^{L} those of the Gauss–Legendre quadrature rule. The discrete measure implied by (2.17) is $\mathrm{d}\lambda_N$ with $N = 2M$ and

$$\mathrm{d}\lambda_N(t) = \sum_{k=1}^{M} w_k^{\mathrm{Ch}}\delta(t - x_k^{\mathrm{Ch}}) + c\sum_{k=1}^{M} w_k^{\mathrm{L}}\delta(t - x_k^{\mathrm{L}}). \qquad (2.18)$$

What is attractive about this choice is the fact that the approximation in (2.17) is actually an equality whenever the product $p \cdot q$ is a polynomial of degree $\leq 2M-1$. Now if we are interested in computing $\alpha_k(w)$, $\beta_k(w)$ for $k \leq n-1$, then the products $p \cdot q$ that occur in Darboux's formulae are all of degree $\leq 2n-1$. Therefore, we have equality in (2.17) if $n \leq M$. It therefore suffices to take $M = n$ in (2.17) to obtain the first n recurrence coefficients exactly.

In general, the quadrature rules will not produce exact results, and M will have to be increased through a sequence of integers until convergence occurs.

Example 1 illustrates the case of a 2-component discretization. In a general *multiple-component discretization*, the support $[a,b]$ of $\mathrm{d}\lambda$ is decomposed into s intervals,

$$[a,b] = \bigcup_{j=1}^{s} [a_j, b_j], \qquad (2.19)$$

where the intervals $[a_j, b_j]$ may or may not be disjoint. The measure $\mathrm{d}\lambda$ is then discretized on each interval $[a_j, b_j]$ using either a tailor-made M-point quadrature (as in Example 1), or a general-purpose quadrature. For the latter, a Fejér quadrature rule on $[-1,1]$, suitably transformed to $[a_j, b_j]$, has been found useful. (The Fejér rule is the interpolatory quadrature formula based on Chebyshev points.) If the original measure $\mathrm{d}\lambda$ has also a discrete component, this component is simply added on. Rather than go into details (which are discussed in [10, §2.2.4]), we present the Matlab implementation, another illustrative example, and a demo.

The `OPQ` routine for the multiple-component discretization is `ab=mcdis(N,epsO,quad,Mmax)`, where in addition to the variables `ab` and `n`, which have the usual meaning, there are three other parameters, `epsO`: a prescribed accuracy tolerance, `quad`: the name of a quadrature routine carrying out

the discretization on each subinterval if tailor-made (otherwise, `quadgp.m`, a general-purpose quadrature routine can be used), `Mmax`: a maximal allowable value for the discretization parameter M. The decomposition (2.19) is input via the `mc`×2 array

$$\texttt{AB}=\left[\begin{array}{c|c} a_1 & b_1 \\ a_2 & b_2 \\ \vdots & \vdots \\ a_{\texttt{mc}} & b_{\texttt{mc}} \end{array}\right]$$

where `mc` is the number of components (the s in (2.19)). A discrete component which may possibly be present in $\mathrm{d}\lambda$ is input via the array

$$\texttt{DM}=\left[\begin{array}{c|c} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_{\texttt{mp}} & y_{\texttt{mp}} \end{array}\right]$$

with the first column containing the support points, and the second column the associated weights. The number of support points is `mp`. Both `mc` and `mp`, as well as `AB` and `DM`, are global variables. Another global variable is `iq`, which has to be set equal to 1 if the user provides his or her own quadrature routine, and equal to 0 otherwise.

Example 2. The normalized Jacobi weight function plus a discrete measure.

This is the measure

$$\mathrm{d}\lambda(t) = (\beta_0^{\mathrm{J}})^{-1}(1-t)^{\alpha}(1+t)^{\beta}\mathrm{d}t + \sum_{j=1}^{p} w_j \delta(t-x_j)\mathrm{d}t \text{ on } [-1,1],$$

where

$$\beta_0^{\mathrm{J}} = \int_{-1}^{1} (1-t)^{\alpha}(1+t)^{\beta}\mathrm{d}t, \quad \alpha > -1,\ \beta > -1.$$

Here, one single component suffices to do the discretization, and the obvious choice of quadrature rule is the Gauss–Jacobi M-point quadrature formula to which the discrete component is added on. Similarly as in Example 1, taking $M = n$ yields the first n recurrence coefficients $\alpha_k(\mathrm{d}\lambda)$, $\beta_k(\mathrm{d}\lambda)$, $k \le n-1$, exactly. The global parameters in Matlab are here `mc`=1, `mp`=p, `iq`=1, and

$$\texttt{AB}=\left[\begin{array}{c|c} -1 & 1 \end{array}\right] \qquad \texttt{DM}=\left[\begin{array}{c|c} x_1 & w_1 \\ x_2 & w_2 \\ \vdots & \vdots \\ x_p & w_p \end{array}\right]$$

Demo#3 Logistic density function,

$$d\lambda(t) = \frac{e^{-t}}{(1+e^{-t})^2}\,dt, \quad t \in \mathbb{R}.$$

The discretization is conveniently effected by the quadrature rule

$$\int_{\mathbb{R}} p(t)d\lambda(t) = \int_0^\infty \frac{p(-t)}{(1+e^{-t})^2}\,e^{-t}dt + \int_0^\infty \frac{p(t)}{(1+e^{-t})^2}\,e^{-t}dt$$
$$\approx \sum_{k=1}^{M} \lambda_k^{\mathrm{L}} \frac{p(-\tau_k^{\mathrm{L}}) + p(\tau_k^{\mathrm{L}})}{(1+e^{-\tau_k^{\mathrm{L}}})^2},$$

where τ_k^{L}, λ_k^{L} are the nodes and weights of the M-point Gauss–Laguerre quadrature formula. This no longer produces exact results for $M = n$, but converges rapidly as $M \to \infty$. The exact answers happen to be known,

$$\alpha_k(d\lambda) = 0 \quad \text{by symmetry,}$$
$$\beta_0(d\lambda) = 1, \quad \beta_k(d\lambda) = \frac{k^4\pi^2}{4k^2-1}, \quad k \geq 1.$$

Numerical results produced by `mcdis.m` with `N`=40, `eps0`=$10^3\times$`eps`, along with errors (absolute errors for α_k, relative errors for β_k) are shown in the box below. The two entries in the bottom row are the maximum errors taken over $0 \leq n \leq 39$.

n	β_n	err α	err β
0	1.0000000000(0)	7.18(−17)	3.33(−16)
1	3.2898681337(0)	1.29(−16)	2.70(−16)
6	8.9447603523(1)	4.52(−16)	1.43(−15)
15	5.5578278399(2)	2.14(−14)	0.00(+00)
39	3.7535340252(3)	6.24(−14)	4.48(−15)
		6.24(−14)	8.75(−15)

2.5 Cauchy Integrals of Orthogonal Polynomials

The Jacobi Continued Fraction

The *Jacobi continued fraction* associated with the measure $d\lambda$ is

$$\mathcal{J} = \mathcal{J}(t; d\lambda) = \frac{\beta_0}{t-\alpha_0-}\,\frac{\beta_1}{t-\alpha_1-}\,\frac{\beta_2}{t-\alpha_2-}\cdots, \tag{2.20}$$

where $\alpha_k = \alpha_k(d\lambda)$, $\beta_k = \beta_k(d\lambda)$. From the theory of continued fractions it is readily seen that the nth convergent of $\mathcal{J}$ is

$$\frac{\beta_0}{z-\alpha_0-}\,\frac{\beta_1}{z-\alpha_1-}\cdots\frac{\beta_{n-1}}{z-\alpha_{n-1}} = \frac{\sigma_n(z; d\lambda)}{\pi_n(z; d\lambda)}, \quad n = 1, 2, 3, \ldots, \tag{2.21}$$

where π_n is the monic orthogonal polynomial of degree n, and σ_n a polynomial of degree $n-1$ satisfying the same basic three-term recurrence relation as π_n, but with different starting values,

$$\begin{aligned} \sigma_{k+1}(z) &= (z-\alpha_k)\sigma_k(z) - \beta_k\sigma_{k-1}(z), \quad k = 1,2,3,\dots, \\ \sigma_0(z) &= 0, \quad \sigma_1(z) = \beta_0. \end{aligned} \tag{2.22}$$

Recall that $\beta_0 = \int_{\mathbb{R}} d\lambda(t)$. If we define $\sigma_{-1} = -1$, then (2.22) holds also for $k = 0$. We have, moreover,

$$\sigma_n(z) = \int_{\mathbb{R}} \frac{\pi_n(z) - \pi_n(t)}{z-t}\, d\lambda(t), \quad n = 0,1,2,\dots, \tag{2.23}$$

as can be seen by showing that the integral on the right also satisfies (2.22). If we define

$$F(z) = F(z; d\lambda) = \int_{\mathbb{R}} \frac{d\lambda(t)}{z-t} \tag{2.24}$$

to be the *Cauchy transform* of the measure $d\lambda$, and more generally consider

$$\rho_n(z) = \rho_n(z; d\lambda) = \int_{\mathbb{R}} \frac{\pi_n(t)}{z-t}\, d\lambda(t), \tag{2.25}$$

the *Cauchy integral* of the orthogonal polynomial π_n, we can give (2.23) the form

$$\sigma_n(z) = \pi_n(z)F(z) - \rho_n(z), \tag{2.26}$$

and hence

$$\frac{\sigma_n(z)}{\pi_n(z)} = F(z) - \frac{\rho_n(z)}{\pi_n(z)}. \tag{2.27}$$

An important result from the theory of the moment problem tells us that, whenever the moment problem for $d\lambda$ is determined, then

$$\lim_{n\to\infty} \frac{\sigma_n(z)}{\pi_n(z)} = F(z) \quad \text{for } z \in \mathbb{C}\backslash[a,b], \tag{2.28}$$

where $[a,b]$ is the support of the measure $d\lambda$. If $[a,b]$ is a finite interval, then the moment problem is always determined, and (2.28) is known as *Markov's theorem.*

Note from (2.26) that, since $\sigma_{-1} = -1$, we have

$$\rho_{-1}(z) = 1, \tag{2.29}$$

and the sequence $\{\rho_n\}_{n=-1}^{\infty}$ satisfies the same three-term recurrence relation as $\{\pi_n\}_{n=-1}^{\infty}$. As a consequence of (2.27) and (2.28), however, it behaves quite differently at infinity,

$$\lim_{n\to\infty} \frac{\rho_n(z)}{\pi_n(z)} = 0, \tag{2.30}$$

which implies that $\{\rho_n(z)\}$ is the *minimal solution* of the three-term recurrence relation having the initial value (2.29). It is well known that a minimal solution of a three-term recurrence relation is uniquely determined by its starting value, and, moreover, that

$$\frac{\rho_n(z)}{\rho_{n-1}(z)} = \frac{\beta_n}{z-\alpha_n-}\,\frac{\beta_{n+1}}{z-\alpha_{n+1}-}\,\frac{\beta_{n+2}}{z-\alpha_{n+2}-}\cdots, \tag{2.31}$$

i.e. the successive ratios of the minimal solution are the successive *tails* of the Jacobi continued fraction (Pincherle's theorem). In particular, by (2.31) for $n = 0$, and (2.21), (2.28) and (2.29),

$$\rho_0(z) = F(z), \tag{2.32}$$

i.e., ρ_0 is the Cauchy transform of the measure.

We remark that (2.25) is meaningful also for real $z = x$ in (a, b), if the integral is interpreted as a *Cauchy principal value integral* (cf. (4.36))

$$\rho_n(x) = \int_{\mathbb{R}} \frac{\pi_n(t; \mathrm{d}\lambda)}{x-t}\,\mathrm{d}\lambda(t), \quad x \in (a, b), \tag{2.33}$$

and the sequence $\{\rho_n(x)\}$ satisfies the basic three-term recurrence relation with initial values

$$\rho_{-1}(x) = 1, \quad \rho_0(x) = \int_{\mathbb{R}} \frac{\mathrm{d}\lambda(t)}{x-t}, \tag{2.34}$$

but is no longer minimal.

Continued Fraction Algorithm

This is an algorithm for computing the minimal solution $\rho_n(z)$, $z \in \mathbb{C}\backslash[a, b]$, of the basic three-term recurrence relation. Denote the ratio in (2.31) by

$$r_{n-1} = \frac{\rho_n(z)}{\rho_{n-1}(z)}. \tag{2.35}$$

Then, clearly,

$$r_{n-1} = \frac{\beta_n}{z-\alpha_n-r_n}. \tag{2.36}$$

If, for some $\nu \geq N$, we knew r_ν, we could apply (2.36) for $r = \nu, \nu-1, \ldots, 0$, and then obtain

$$\rho_n(z) = r_{n-1}\rho_{n-1}(z), \quad n = 0, 1, \ldots, N. \tag{2.37}$$

The *continued fraction algorithm* is precisely this algorithm, except that r_ν is replaced by 0. All quantities generated then depend on ν, which is indicated by a superscript.

Algorithm 2 Continued fraction algorithm

backward phase; $\nu \geq N$:

$$r_\nu^{[\nu]} = 0, \quad r_{n-1}^{[\nu]} = \frac{\beta_n}{z-\alpha_n-r_n^{[\nu]}}, \quad n = \nu, \nu-1, \ldots, 0.$$

forward phase:

$$\rho_{-1}^{[\nu]}(z) = 1, \quad \rho_n^{[\nu]}(z) = r_{n-1}^{[\nu]}\rho_{n-1}^{[\nu]}(z), \quad n = 0, 1, \ldots, N.$$

It can be shown that, as a consequence of the minimality of $\{\rho_n(z)\}$ (cf. [10, pp. 114–115]),

$$\lim_{\nu\to\infty} \rho_n^{[\nu]}(z) = \rho_n(z), \quad n = 0, 1, \ldots, N, \quad \text{if } z \in \mathbb{C}\backslash[a, b]. \tag{2.38}$$

Convergence is faster the larger $\operatorname{dist}(z, [a, b])$. To compute $\rho_n(z)$, it suffices to apply Algorithm 2 for a sequence of increasing values of ν until convergence is achieved to within the desired accuracy.

The `OPQ` command implementing this algorithm is

```
[rho,r,nu]=cauchy(N,ab,z,eps0,nu0,numax)
```

where the meanings of the output variables `rho`, `r` and input variable `ab` are as shown below.

rho	r	ab	
$\rho_0(z)$	$r_0(z)$	α_0	β_0
$\rho_1(z)$	$r_1(z)$	α_1	β_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\rho_N(z)$	$r_N(z)$	α_{numax}	β_{numax}

The input variable `eps0` is an error tolerance, the variable `nu0` a suitable starting value of ν in Algorithm 2, which is incremented in steps of, say 5, until the algorithm converges to the accuracy `eps0`. If convergence does not occur within $\nu \leq$ `numax`, an error message is issued, otherwise the value of ν yielding convergence is output as `nu`.

2.6 Modification Algorithms

By "modification" of a measure $d\lambda$, we mean here multiplication of $d\lambda$ by a rational function r which is positive on the support $[a, b]$ of $d\lambda$. The *modified measure* thus is

$$d\hat{\lambda}(t) = r(t)d\lambda(t), \quad r \text{ rational and } r > 0 \text{ on } [a, b]. \tag{2.39}$$

We are interested in determining the recurrence coefficients $\hat{\alpha}_k$, $\hat{\beta}_k$ for $d\hat{\lambda}$ in terms of the recurrence coefficients α_k, β_k of $d\lambda$. An algorithm that carries out the transition from α_k, β_k to $\hat{\alpha}_k$, $\hat{\beta}_k$ is called a *modification algorithm*. While the passage from the orthogonal polynomials relative to $d\lambda$ to those relative to $d\hat{\lambda}$ is classical (at least in the case when r is a polynomial), the transition in terms of recurrence coefficients is more recent. It was first treated for linear factors in 1971 by Galant.

Example 3. Linear factor $r(t) = s(t-c)$, $c \in \mathbb{R}\backslash[a,b]$, $s = \pm 1$.

Here, s is a sign factor to make $r(t) > 0$ on (a,b). Galant's approach is to determine the Jacobi matrix of $\mathrm{d}\hat{\lambda}$ from the Jacobi matrix of $\mathrm{d}\lambda$ by means of one step of the symmetric, shifted LR algorithm: by the choice of s, the matrix $s[\boldsymbol{J}_{n+1}(\mathrm{d}\lambda) - c\boldsymbol{I}]$ is symmetric positive definite, hence admits a Cholesky decomposition

$$s[\boldsymbol{J}_{n+1}(\mathrm{d}\lambda) - c\boldsymbol{I}] = \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}},$$

where $\boldsymbol{L}$ is lower triangular. The Jacobi matrix $\boldsymbol{J}_n(\mathrm{d}\hat{\lambda})$ is now obtained by reversing the order of the product on the right, adding back the shift c, and then discarding the last row and column,[1]

$$\boldsymbol{J}_n(\mathrm{d}\hat{\lambda}) = \left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} + c\boldsymbol{I}\right)_{[1:n,1:n]}.$$

Since the matrices involved are tridiagonal, the procedure can be implemented by simple nonlinear recurrence relations. These can also be obtained more systematically via Christoffel's theorem and its generalizations.

Generalized Christoffel's Theorem

We write

$$\mathrm{d}\hat{\lambda}(t) = \frac{u(t)}{v(t)}\,\mathrm{d}\lambda(t), \quad u(t) = \pm\prod_{\lambda=1}^{\ell}(t-u_\lambda),\ v(t) = \prod_{\mu=1}^{m}(t-v_\mu), \tag{2.40}$$

where u_λ and v_μ are real numbers outside the support of $\mathrm{d}\lambda$. The sign of $u(t)$ is chosen so that $\mathrm{d}\hat{\lambda}$ is a positive measure. Christoffel's original theorem (1858) relates to the case $v(t) = 1$, i.e. $m = 0$. The generalization to arbitrary v is due to Uvarov (1969). It has a different form depending on whether $m \le n$ or $m > n$. In the first case, it states that

$$u(t)\pi_n(t;\mathrm{d}\hat{\lambda}) = \text{const} \times \begin{vmatrix} \pi_{n-m}(t) & \cdots & \pi_{n-1}(t) & \pi_n(t) & \cdots & \pi_{n+\ell}(t) \\ \pi_{n-m}(u_1) & \cdots & \pi_{n-1}(u_1) & \pi_n(u_1) & \cdots & \pi_{n+\ell}(u_1) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \pi_{n-m}(u_\ell) & \cdots & \pi_{n-1}(u_\ell) & \pi_n(u_\ell) & \cdots & \pi_{n+\ell}(u_\ell) \\ \rho_{n-m}(v_1) & \cdots & \rho_{n-1}(v_1) & \rho_n(v_1) & \cdots & \rho_{n+\ell}(v_1) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{n-m}(v_m) & \cdots & \rho_{n-1}(v_m) & \rho_n(v_m) & \cdots & \rho_{n+\ell}(v_m) \end{vmatrix}, \tag{2.41}$$

where

$$\rho_k(z) = \int_{\mathbb{R}} \frac{\pi_k(t;\mathrm{d}\lambda)}{z-t}\,\mathrm{d}\lambda(t), \quad k = 0,1,2,\ldots,$$

[1] See, e.g. [9], where it is also shown how a quadratic factor $(t-c_1)(t-c_2)$ can be dealt with by one step of the QR algorithm; see in particular §3.2 and 3.3

are the Cauchy integrals of the orthogonal polynomials π_k. They occur only if $m > 0$. To get monic polynomials, the constant in (2.41) must be taken to be the reciprocal of the (signed) cofactor of the element $\pi_{n+\ell}(t)$.

If $m > n$, the generalized Christoffel theorem has the form

$$u(t)\pi_n(t;\mathrm{d}\hat{\lambda}) = \text{const} \times \begin{vmatrix} 0 & 0 & \cdots & 0 & \pi_0(t) & \cdots & \pi_{n+\ell}(t) \\ 0 & 0 & \cdots & 0 & \pi_0(u_1) & \cdots & \pi_{n+\ell}(u_1) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & \pi_0(u_\ell) & \cdots & \pi_{n+\ell}(u_\ell) \\ 1 & v_1 & \cdots & v_1^{m-n-1} & \rho_0(v_1) & \cdots & \rho_{n+\ell}(v_1) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & v_m & \cdots & v_m^{m-n-1} & \rho_0(v_m) & \cdots & \rho_{n+\ell}(v_m) \end{vmatrix}. \tag{2.42}$$

Both versions of the theorem remain valid for complex u_λ, v_μ if orthogonality is understood in the sense of *formal orthogonality*.

Linear Factor

Generalizing Example 3 to arbitrary complex shifts, we let

$$\mathrm{d}\hat{\lambda}(t) = (t - z)\mathrm{d}\lambda(t), \quad z \in \mathbb{C}\backslash[a, b]. \tag{2.43}$$

Using Christoffel's theorem, letting $\hat{\pi}_n(\,\cdot\,) = \pi_n(\,\cdot\,;\mathrm{d}\hat{\lambda})$, we have

$$(t - z)\hat{\pi}_n(t) = \frac{\begin{vmatrix} \pi_n(t) & \pi_{n+1}(t) \\ \pi_n(z) & \pi_{n+1}(z) \end{vmatrix}}{-\pi_n(z)} = \pi_{n+1}(t) - r_n\pi_n(t), \tag{2.44}$$

where

$$r_n = \frac{\pi_{n+1}(z)}{\pi_n(z)}. \tag{2.45}$$

Following Verlinden [17], we write $(t - z)t\hat{\pi}_k(t)$ in two different ways: in the first, we use the three-term recurrence relation for π_k to obtain

$$\begin{aligned}(t - z)t\hat{\pi}_k(t) &= t\pi_{k+1}(t) - r_k \cdot t\pi_k(t) \\ &= \pi_{k+2}(t) + (\alpha_{k+1} - r_k)\pi_{k+1}(t) + (\beta_{k+1} - r_k\alpha_k)\pi_k(t) - r_k\beta_k\pi_{k-1}(t);\end{aligned}$$

in the second, we use the three-term recurrence relation directly on $\hat{\pi}_k$, and then apply (2.44), to write

$$\begin{aligned}(t - z)t\hat{\pi}_k(t) &= (t - z)[\hat{\pi}_{k+1} + \hat{\alpha}_k\hat{\pi}_k(t) + \hat{\beta}_k\hat{\pi}_{k-1}(t)] \\ &= \pi_{k+2}(t) + (\hat{\alpha}_k - r_{k+1})\pi_{k+1}(t) + (\hat{\beta}_k - r_k\hat{\alpha}_k)\pi_k(t) - r_{k-1}\hat{\beta}_k\pi_{k-1}(t).\end{aligned}$$

Since orthogonal polynomials are linearly independent, the coefficients in the two expressions obtained must be the same. This yields

$$\hat{\alpha}_k - r_{k+1} = \alpha_{k+1} - r_k, \quad r_{k-1}\hat{\beta}_k = r_k\beta_k,$$

hence the following algorithm.

Algorithm 3 Modification by a linear factor $t - z$

initialization:

$$r_0 = z - \alpha_0, \quad r_1 = z - \alpha_1 - \beta_1/r_0,$$
$$\hat{\alpha}_0 = \alpha_1 + r_1 - r_0, \quad \hat{\beta}_0 = -r_0\,\beta_0.$$

continuation (if $n > 1$): for $k = 1, 2, \ldots, n-1$ do

$$r_{k+1} = z - \alpha_{k+1} - \beta_{k+1}/r_k,$$
$$\hat{\alpha}_k = \alpha_{k+1} + r_{k+1} - r_k,$$
$$\hat{\beta}_k = \beta_k r_k/r_{k-1}.$$

Note that this requires α_n, β_n in addition to the usual n recurrence coefficients α_k, β_k for $k \leq n-1$. Algorithm 3 has been found to be numerically stable.

The `OPQ` Matlab command implementing Algorithm 3 is

```
ab=chri1(N,ab0,z)
```

where `ab0` is an `(N+1)×2` array containing the recurrence coefficients α_k, β_k, $k = 0, 1, \ldots, N$.

Quadratic Factor

We consider (real) quadratic factors $(t-x)^2 + y^2 = (t-z)(t-\overline{z})$, $z = x + \mathrm{i}y$, $y > 0$. Christoffel's theorem is now applied with $u_1 = z$, $u_2 = \overline{z}$ to express $(t-z)(t-\overline{z})\hat{\pi}_n(t)$ as a linear combination of π_n, π_{n+1}, and π_{n+2},

$$(t-z)(t-\overline{z})\hat{\pi}_n(t) = \pi_{n+2}(t) + s_n\pi_{n+1}(t) + t_n\pi_n(t), \tag{2.46}$$

where

$$s_n = -\left(r'_{n+1} + \frac{r''_{n+1}}{r''_n}\,r'_n\right), \quad t_n = \frac{r''_{n+1}}{r''_n}\,|r_n|^2. \tag{2.47}$$

Here we use the notation

$$r'_n = \mathrm{Re}\,r_n(z),\ r''_n = \mathrm{Im}\,r_n(z),\ |r_n|^2 = |r_n(z)|^2, \quad n = 0, 1, 2, \ldots, \tag{2.48}$$

where $r_n(z)$ continues to be the quantity defined in (2.45). The same technique used before can be applied to (2.46): express $(t-z)(t-\overline{z})t\hat{\pi}_k(t)$ in two different ways as a linear combination of $\pi_{k+3}, \pi_{k+2}, \ldots, \pi_{k-1}$ and compare the respective coefficients. The result gives rise to the following algorithm.

Algorithm 4 Modification by a quadratic factor $(t-z)(t-\overline{z})$, $z=x+\mathrm{i}y$

initialization:

$$r_0 = z-\alpha_0,\ r_1 = z-\alpha_1-\beta_1/r_0,\ r_2 = z-\alpha_2-\beta_2/r_1,$$

$$\hat{\alpha}_0 = \alpha_2 + r_2' + \frac{r_2''}{r_1''}r_1' - \left(r_1' + \frac{r_1''}{r_0''}r_0'\right),$$

$$\hat{\beta}_0 = \beta_0(\beta_1 + |r_0|^2).$$

continuation (if $n>1$): for $k=1,2,\ldots,n-1$ do

$$r_{k+2} = z-\alpha_{k+2}-\beta_{k+2}/r_{k+1},$$

$$\hat{\alpha}_k = \alpha_{k+2} + r_{k+2}' + \frac{r_{k+2}''}{r_{k+1}''}r_{k+1}' - \left(r_{k+1}' + \frac{r_{k+1}''}{r_k''}r_k'\right),$$

$$\hat{\beta}_k = \beta_k \frac{r_{k+1}''r_{k-1}''}{[r_k'']^2}\left|\frac{r_k}{r_{k-1}}\right|^2.$$

Note that this requires α_k, β_k for k up to $n+1$. Algorithm 4 is also quite stable, numerically.

The `OPQ` routine for Algorithm 4 is

```
ab=chri2(N,ab0,x,y)
```

with obvious meanings of the variables involved.

Since any real polynomial can be factored into a product of real linear and quadratic factors of the type considered, Algorithms 3 and 4 can be applied repeatedly to deal with modification by an arbitrary polynomial which is positive on the support $[a,b]$.

Linear Divisor

In analogy to (2.43), we consider

$$\mathrm{d}\hat{\lambda}(t) = \frac{\mathrm{d}\lambda(t)}{t-z}, \quad z\in\mathbb{C}\backslash[a,b]. \tag{2.49}$$

Now the *generalized* Christoffel theorem (with $\ell=0$, $m=1$) comes into play, giving

$$\hat{\pi}_n(t) = \frac{\begin{vmatrix}\pi_{n-1}(t) & \pi_n(t)\\ \rho_{n-1}(z) & \rho_n(z)\end{vmatrix}}{-\rho_{n-1}(z)} = \pi_n(t) - r_{n-1}\pi_{n-1}(t), \tag{2.50}$$

where now

$$r_n = \frac{\rho_{n+1}(z)}{\rho_n(z)}. \tag{2.51}$$

Similarly as before, we express $t\hat{\pi}_k(t)$ in two different ways as a linear combination of $\pi_{k+1}, \pi_k, \ldots, \pi_{k-2}$ and compare coefficients. By convention,

$$\hat{\beta}_0 = \int_{\mathbb{R}} \mathrm{d}\hat{\lambda}(t) = \int_{\mathbb{R}} \frac{\mathrm{d}\lambda(t)}{t-z} = -\rho_0(z).$$

The result is:

Algorithm 5 Modification by a linear divisor

initialization:

$$\hat{\alpha}_0 = \alpha_0 + r_0, \quad \hat{\beta}_0 = -\rho_0(z).$$

continuation (if $n > 1$): for $k = 1, 2, \ldots, n-1$ do

$$\hat{\alpha}_k = \alpha_k + r_k - r_{k-1},$$
$$\hat{\beta}_k = \beta_{k-1} r_{k-1} / r_{k-2}.$$

Note that here no coefficient α_k, β_k beyond $k \leq n-1$ is needed, not even β_{n-1}.

The ratios r_k of Cauchy integrals that appear in Algorithm 5 can be precomputed by Algorithm 2, where only the backward phase is relevant, convergence being tested on the $r_k^{[\nu]}$. Once converged, the algorithm also provides $\rho_0(z) = r_{-1}^{[\infty]}$.

As z approaches the support interval $[a, b]$, the strength of minimality of the Cauchy integrals $\{\rho_k(z)\}$ weakens and ceases altogether when $z = x \in [a, b]$. For z very close to $[a, b]$, Algorithm 2 therefore converges very slowly. On the other hand, since minimality is very weak, one can generate ρ_k with impunity, if n is not too large, by forward application of the basic three-term recurrence relation, using the initial values $\rho_{-1}(z) = 1$ and $\rho_0(z)$.

All of this is implemented in the `OPQ` routine

```
[ab,nu]=chri4(N,ab0,z,eps0,nu0,numax,rho0,iopt)
```

where all variables except `rho` and `iopt` have the same meaning as before. The parameter `rho` is $\rho_0(z)$, whereas `iopt` controls the method of computation for r_k: Algorithm 2 if `iopt`=1, and forward recursion otherwise.

Quadratic Divisor

We now consider

$$\mathrm{d}\hat{\lambda}(t) = \frac{\mathrm{d}\lambda(t)}{(t-z)(t-\overline{z})} = \frac{\mathrm{d}\lambda(t)}{(t-x)^2+y^2}, \quad z = x + \mathrm{i}y,\ x \in \mathbb{R},\ y > 0. \quad (2.52)$$

Here we have

$$\hat{\alpha}_0 = \frac{\int_{\mathbb{R}} t\mathrm{d}\lambda(t)/|t-z|^2}{\int_{\mathbb{R}} \mathrm{d}\lambda(t)/|t-z|^2} = x + y\,\frac{\operatorname{Re}\rho_0(z)}{\operatorname{Im}\rho_0(z)}, \quad \hat{\beta}_0 = -\frac{1}{y}\operatorname{Im}\rho_0(z). \tag{2.53}$$

We are in the case $\ell = 0$, $m = 2$ of the generalized Christoffel theorems (2.41) and (2.42), which give respectively

$$\hat{\pi}_n(t) = \frac{\begin{vmatrix} \pi_{n-2}(t) & \pi_{n-1}(t) & \pi_n(t) \\ \rho_{n-2}(z) & \rho_{n-1}(z) & \rho_n(z) \\ \rho_{n-2}(\overline{z}) & \rho_{n-1}(\overline{z}) & \rho_n(\overline{z}) \end{vmatrix}}{\begin{vmatrix} \rho_{n-2}(z) & \rho_{n-1}(z) \\ \rho_{n-2}(\overline{z}) & \rho_{n-1}(\overline{z}) \end{vmatrix}}, \ n \geq 2; \quad \hat{\pi}_1(t) = \frac{\begin{vmatrix} 0 & \pi_0(t) & \pi_1(t) \\ 1 & \rho_0(z) & \rho_1(z) \\ 1 & \rho_0(\overline{z}) & \rho_1(\overline{z}) \end{vmatrix}}{\begin{vmatrix} 1 & \rho_0(z) \\ 1 & \rho_0(\overline{z}) \end{vmatrix}}. \tag{2.54}$$

This becomes

$$\hat{\pi}_n(t) = \pi_n(t) + s_n\pi_{n-1}(t) + t_n\pi_{n-2}(t), \quad n \geq 1, \tag{2.55}$$

where

$$s_n = -\left(r'_{n-1} + \frac{r''_{n-1}}{r''_{n-2}}\,r'_{n-2}\right), \ n \geq 1; \quad t_n = \frac{r''_{n-1}}{r''_{n-2}}\,|r_{n-2}|^2, \ n \geq 2, \tag{2.56}$$

with r_n as defined in (2.51) and notation as in (2.48). Exactly the same procedure used to obtain Algorithm 5 yields

Algorithm 6 Modification by a quadratic divisor
initialization:

$$\begin{aligned} &\hat{\alpha}_0 = x + \rho'_0 y/\rho''_0, \quad \hat{\beta}_0 = -\rho''_0/y, \\ &\hat{\alpha}_1 = \alpha_1 - s_2 + s_1, \quad \hat{\beta}_1 = \beta_1 + s_1(\alpha_0 - \hat{\alpha}_1) - t_2, \\ &\hat{\alpha}_2 = \alpha_2 - s_3 + s_2, \quad \hat{\beta}_2 = \beta_2 + s_2(\alpha_1 - \hat{\alpha}_2) - t_3 + t_2. \end{aligned}$$

continuation (if $n > 3$): for $k = 3, 4, \ldots, n-1$ do

$$\hat{\alpha}_k = \alpha_k - s_{k+1} + s_k, \quad \hat{\beta}_k = \beta_{k-2}\,t_k/t_{k-1}.$$

The `OPQ` routine for Algorithm 6 is

```
[ab,nu]=chri5(N,ab0,z,eps0,nu0,numax,rho0,iopt)
```

where the input and output variables have the same meaning as in the routine `chri4.m`.

Just like Algorithms 3 and 4, also Algorithms 5 and 6 can be applied repeatedly to deal with more general polynomial divisors.

Exercises to §2 (Stars indicate more advanced exercises.)

1. Explain why, under the assumptions made about the measure $d\lambda$, the inner product $(p, q)_{d\lambda}$ of two polynomials p, q is well defined.
2. Show that monic orthogonal polynomials relative to an absolutely continuous measure are uniquely defined. {*Hint*: Use Gram-Schmidt orthogonalization.} Discuss the uniqueness in the case of discrete measures.
3. Supply the details of the proof of (2.1). In particular, derive (2.3) and (2.4).
4. Derive the three-term recurrence relation (2.7) for the orthonormal polynomials.
5. (a) With $\tilde{\pi}_k$ denoting the orthonormal polynomials relative to a measure $d\lambda$, show that

$$\int_{\mathbb{R}} t\tilde{\pi}_k(t)\tilde{\pi}_\ell(t)d\lambda(t) = \begin{cases} 0 & \text{if } |k-\ell| > 1, \\ \sqrt{\beta_{k+1}} & \text{if } |k-\ell| = 1, \\ \alpha_k & \text{if } k = \ell, \end{cases}$$

where $\alpha_k = \alpha_k(d\lambda)$, $\beta_k = \beta_k(d\lambda)$.

(b) Use (a) to prove

$$\boldsymbol{J} = \boldsymbol{J}_n(d\lambda) = \int_{\mathbb{R}} t\boldsymbol{p}(t)\boldsymbol{p}^{\mathrm{T}}(t)d\lambda(t),$$

where $\boldsymbol{p}^{\mathrm{T}}(t) = [\tilde{\pi}_0(t), \tilde{\pi}_1(t), \ldots, \tilde{\pi}_{n-1}(t)]$.

(c) With notation as in (b), prove

$$t\boldsymbol{p}(t) = \boldsymbol{J}\boldsymbol{p}(t) + \sqrt{\beta_n}\tilde{\pi}_n(t)\boldsymbol{e}_n,$$

where $\boldsymbol{e}_n = [0, 0, \ldots, 1]^{\mathrm{T}} \in \mathbb{R}^n$.

6. Let $d\lambda(t) = w(t)dt$ be symmetric on $[-a, a]$, $a > 0$, that is, $w(-t) = w(t)$ on $[-a, a]$. Show that $\alpha_k(d\lambda) = 0$, all $k \geq 0$.

7*. Symmetry of orthogonal polynomials.
Let $d\lambda(t) = w(t)dt$ be symmetric in the sense of Exercise 6.

(a) Show that

$$\pi_{2k}(t; d\lambda) = \pi_k^+(t^2), \quad \pi_{2k+1}(t; d\lambda) = t\pi_k^-(t^2),$$

where $\pi_k^{\pm}$ are the monic polynomials orthogonal on $[0, a^2]$ with respect to $d\lambda^{\pm}(t) = t^{\mp 1/2}w(t^{1/2})dt$.

(b) Let (cf. Exercise 6)

$$\pi_{k+1}(t) = t\pi_k(t) - \beta_k\pi_{k-1}(t), \quad k = 0, 1, 2, \ldots,$$
$$\pi_{-1}(t) = 0, \quad \pi_0(t) = 1$$

be the recurrence relation for $\{\pi_k(\cdot\,; d\lambda)\}$, and let $\alpha_k^{\pm}$, $\beta_k^{\pm}$ be the recurrence coefficients for $\{\pi_k^{\pm}\}$. Show that

$$\left.\begin{aligned} \beta_1 &= \alpha_0^+ \\ \beta_{2k} &= \beta_k^+/\beta_{2k-1} \\ \beta_{2k+1} &= \alpha_k^+ - \beta_{2k} \end{aligned}\right\} \quad k = 1, 2, 3, \ldots .$$

(c) Derive relations similar to those in (b) which involve α_0^+ and α_k^-, β_k^-.

(d) Write a Matlab program that checks the numerical stability of the nonlinear recursions in (b) and (c) when $\{\pi_k\}$ are the monic Legendre polynomials.

8. The recurrence relation, in Matlab, of the Chebyshev polynomials of the second kind.

(a) Using Matlab, compute $U_k(x)$ for $1 \le k \le N$ either by means of the three-term recurrence relation $U_{n+1}(x) = 2xU_n(x) - U_{n-1}(x)$ for $n = 0, 1, \ldots, N-1$ (where $U_{-1}(x) = 0$, $U_0(x) = 1$), or else by putting $n = 1 : N$ in the explicit formula $U_n(\cos\theta) = \sin(n+1)\theta/\sin\theta$, where $x = \cos\theta$. For selected values of x and N, determine which of the two methods, by timing each, is more efficient.

(b) Using Matlab, compute the single value $U_N(x)$ either by use of the three-term recurrence relation, or by direct computation based on the trigonometric formula for $U_N(\cos\theta)$. For selected values of x and N, determine which of the two methods, by timing each, is more efficient.

9*. Orthogonality on two separate (symmetric) intervals.

Let $0 < \xi < 1$ and consider orthogonal polynomials π_k relative to the weight function

$$w(t) = \begin{cases} |t|^\gamma (1-t^2)^\alpha (t^2-\xi^2)^\beta, & t \in [-1,\xi] \cup [\xi, 1], \\ 0, & \text{otherwise.} \end{cases}$$

Here, $\gamma \in \mathbb{R}$ and $\alpha > -1$, $\beta > -1$. Evidently, w is a symmetric weight function (in the sense of Exercise 6). Define $\pi_k^\pm$ as in Exercise 7(a).

(a) Transform the polynomials $\pi_k^\pm$ orthogonal on $[\xi^2, 1]$ to orthogonal polynomials $\mathring{\pi}_k^\pm$ on the interval $[-1, 1]$ and obtain the respective weight function $\mathring{w}^\pm$.

(b) Express β_{2k} and β_{2k+1} in terms of $\mathring{\gamma}_r^\pm$, the leading coefficient of the orthonormal polynomial of degree r relative to the weight function $\mathring{w}^\pm$ on $[-1, 1]$. {*Hint*: Use $\beta_r = \|\pi_r\|^2/\|\pi_{r-1}\|^2$ (cf. eqn (2.4)) and relate this to the leading coefficients γ_k, $\gamma_k^\pm$, and $\mathring{\gamma}_k^\pm$, with obvious notations.}

(c) Prove that

$$\lim_{k\to\infty} \beta_{2k} = \frac{1}{4}(1-\xi)^2, \quad \lim_{k\to\infty} \beta_{2k+1} = \frac{1}{4}(1+\xi)^2.$$

{*Hint*: Use the result of (b) in combination with the asymptotic equivalence

$$\mathring{\gamma}_k^{\pm} \sim 2^k \mathring{\gamma}^{\pm}, \quad \mathring{\gamma}^{\pm} = \pi^{-1/2} \exp\left\{ -\frac{1}{2\pi} \int_{-1}^{1} \ln \mathring{w}^{\pm}(x)(1-x^2)^{-1/2} dx \right\}, \quad \text{as } k \to \infty$$

(cf. [16, eqn (12.7.2)]). You may also want to use

$$\int_0^1 \ln(1-a^2x^2)(1-x^2)^{-1/2} dx = \pi \ln \frac{1+(1-a^2)^{1/2}}{2}, \quad a^2 < 1$$

(see [13, eqn 4.295.29]).}

(d) Prove that

$$\lim_{k\to\infty} \alpha_k^{\pm} = \frac{1+\xi^2}{2}, \quad \lim_{k\to\infty} \beta_k^{\pm} = \left(\frac{1-\xi^2}{4}\right)^2.$$

{*Hint*: Express $\alpha_k^{\pm}$, $\beta_k^{\pm}$ in terms of $\mathring{\alpha}_k^{\pm}$, $\mathring{\beta}_k^{\pm}$, and use the fact that the weight function $\mathring{w}^{\pm}$ is in the Szegő class.}

(e) The recurrence coefficients $\{\beta_k\}$ must satisfy the two nonlinear recursions of Exercise 7(b),(c). Each of them can be interpreted as a pair of fixed-point iterations for the even-indexed and for the odd-indexed subsequence, the fixed points being respectively the limits in (c). Show that, asymptotically, both fixed points are "attractive" for the recursion in 7(b), and "repelling" for the one in 7(c). Also show that in the latter, the fixed points become attractive if they are switched. What are the numerical implications of all this?

(f) Consider the special case $\gamma = \pm 1$ and $\alpha = \beta = -\frac{1}{2}$. In the case $\gamma = 1$, use Matlab to run the nonlinear recursion of Exercise 7(b) and compare the results with the known answers

$$\beta_{2k} = \frac{1}{4}(1-\xi)^2 \frac{1+\eta^{2k-2}}{1+\eta^{2k}}, \quad k = 1, 2, 3, \ldots, \quad 0 \le t \le 1$$

and

$$\beta_1 = \frac{1}{2}(1+\xi^2), \quad \beta_{2k+1} = \frac{1}{4}(1+\xi)^2 \frac{1+\eta^{2k+2}}{1+\eta^{2k}}, \quad k = 1, 2, 3, \ldots,$$

where $\eta = (1-\xi)/(1+\xi)$ (see [10, Example 2.30]). Likewise, in the case $\gamma = -1$, run the nonlinear recursion of Exercise 7(c) and compare the results with the exact answers

$$\beta_2 = \tfrac{1}{2}(1-\xi)^2, \quad \beta_{2k} = \tfrac{1}{4}(1-\xi)^2, \quad k = 2, 3, \ldots,$$

and

$$\beta_1 = \xi, \quad \beta_{2k+1} = \tfrac{1}{4}(1+\xi)^2, \quad k = 1, 2, 3, \ldots.$$

Comment on what you observe.

10. Prove the validity of Algorithm 1.
    (a) Verify the initialization part.
    (b) Combine $\sigma_{k+1,k-1} = 0$ with the three-term recurrence relation for π_k to prove the formula for β_k in the continuation part.
    (c) Combine $\sigma_{k+1,k} = 0$ with the three-term recurrence relation for both, π_k and p_k, and use the result of (b), to prove the formula for α_k in the continuation part.

11*. Orthogonal polynomials $\{\pi_k(\cdot\,; w)\}$ relative to the weight function ("hat function")

$$w(t) = \begin{cases} 1+t & \text{if } -1 \le t \le 0, \\ 1-t & \text{if } 0 \le t \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

    (a) Develop a modified Chebyshev algorithm for generating the first n recurrence coefficients $\beta_k(w)$, $k = 0, 1, \ldots, n-1$ (all $\alpha_k(w) = 0$; why?). Define modified moments with respect to a suitable system of (monic) orthogonal polynomials.
    (b) What changes in the routine are required if one wants $\{\pi_k(\cdot\,; 1-w)\}$, or $\{\pi_k(\cdot\,; w(1-w))\}$, or $\{\pi_k(\cdot\,; w^p)\}$ where $p > -1$?
    (c) Download from `OPQ` the routine `chebyshev.m`, write a routine `mom.m` for the modified moments to be used in conjunction with `chebyshev.m` to implement (a), and write a Matlab driver to produce results for selected values of n.
    (d) Devise a 2-component discretization scheme for computing the first n recurrence coefficients $\beta_k(w)$, $k = 0, 1, 2, \ldots, n-1$, which uses an n-point discretization of the inner product on each component interval and is to yield exact answers (in the absence of rounding errors).
    (e) Same as (b).
    (f) Download from `OPQ` the routine `mcdis.m`, write a quadrature routine `qhatf.m` necessary to implement (d), and append a script to the driver of (c) that produces results of the discretization procedure for selected values of n. Download whatever additional routines you need. Run the procedure with `irout` $= 1$ and `irout` $\neq 1$ and observe the respective timings and the maximum discrepancy between the two sets of answers. Verify that the routine "converges" after one iteration if `idelta` is properly set. Compare the results with those of (a).
    (g) Use the routines `acondG.m` and `rcondG.m` to print the absolute and relative condition numbers of the relevant map G_n. Do any of these correlate well with the numerical results obtained in (c)? If not, why not?

12*. Orthogonal polynomials $\{\pi_k(\cdot\,; w)\}$ relative to the weight function ("exponential integral")

$$w(t) = E_1(t), \quad E_1(t) = \int_1^\infty \frac{e^{-ts}}{s}\,\mathrm{d}s \quad \text{on } [0, \infty].$$

These are of interest in the theory of radiative transfer (Chandrasekhar [2, Chapter II, §23]).

(a) Develop and run a multiple-component discretization routine for generating the first n recurrence coefficients $\alpha_k(w)$, $\beta_k(w)$, $k = 0, 1, \ldots, n-1$. Check your results for $n = 20$ against [3, Table 3]. {*Hint*: Decompose the interval $[0, \infty]$ into two subintervals $[0, 2]$ and $[2, \infty]$ (additional subdivisions may be necessary to implement the developments that follow) and incorporate the behavior of $E_1(t)$ near $t = 0$ and $t = \infty$ to come up with appropriate discretizations. For $0 \le t \le 2$, use the power series

$$E_1(t) - \ln(1/t) = -\gamma - \sum_{k=1}^{\infty} \frac{(-1)^k t^k}{k k!},$$

where $\gamma = .57721566490153286\ldots$ is Euler's constant, and for $t > 2$ the continued fraction (cf. [1, eqn 5.1.22])

$$te^t E_1(t) = \frac{1}{1+}\,\frac{a_1}{1+}\,\frac{a_2}{1+}\,\frac{a_3}{1+}\,\frac{a_4}{1+}\cdots, \quad a_k = \lceil k/2 \rceil / t.$$

Evaluate the continued fraction recursively by (cf. [7, §2])

$$\frac{1}{1+}\,\frac{a_1}{1+}\,\frac{a_2}{1+}\cdots = \sum_{k=0}^{\infty} t_k,$$

where

$$t_0 = 1, \quad t_k = \rho_1 \rho_2 \cdots \rho_k, \quad k = 1, 2, 3, \ldots,$$

$$\rho_0 = 0, \quad \rho_k = \frac{-a_k(1 + \rho_{k-1})}{1 + a_k(1 + \rho_{k-1})}, \quad k = 1, 2, 3, \ldots.$$

Download the array `abjaclog(101:200,:)` to obtain the recurrence coefficients `ab` for the logarithmic weight function $\ln(1/t)$.}

(b) Do the same for

$$w(t) = E_2(t), \quad E_2(t) = \int_1^{\infty} \frac{e^{-ts}}{s^2}\, ds \quad \text{on } [0, \infty].$$

Check your results against the respective two- and three-point Gauss quadrature formulae in Chandrasekhar [2, Table VI].

(c) Do the same for

$$w(t) = E_m(t) \quad \text{on } [0, c], \quad 0 < c < \infty, \quad m = 1, 2.$$

Check your results against the respective two-point Gauss quadrature formulae in Chandrasekhar [2, Table VII].

13. Let $C = b_0 + \frac{a_1}{b_1+} \frac{a_2}{b_2+} \frac{a_3}{b_3+} \cdots$ be an infinite continued fraction, and $C_n = b_0 + \frac{a_1}{b_1+} \cdots \frac{a_n}{b_n} = \frac{A_n}{B_n}$ its nth convergent. From the theory of continued fractions, it is known that

$$\left.\begin{aligned} A_n &= b_n A_{n-1} + a_n A_{n-2} \\ B_n &= b_n B_{n-1} + a_n B_{n-2} \end{aligned}\right\} \quad n = 1, 2, 3, \ldots,$$

where

$$A_{-1} = 1, \; A_0 = b_0; \quad B_{-1} = 0, \; B_0 = 1.$$

Use this to prove (2.21) and (2.22).

14. Prove (2.23).
15. Show that (2.30) implies $\lim_{n\to\infty} \frac{\rho_n}{y_n} = 0$, where y_n is *any* solution of the three-term recurrence relation (satisfied by ρ_n and π_n) which is linearly independent of ρ_n. Thus, $\{\rho_n\}$ is indeed a minimal solution.
16. Show that the minimal solutions of a three-term recurrence relation form a one-dimensional manifold.
17. (a) Derive (2.47).
    (b) Supply the details for deriving Algorithm 4.
18. Supply the details for deriving Algorithm 5.
19. (a) Prove (2.53).
    (b) Prove (2.55), (2.56).
    (c) Supply the details for deriving Algorithm 6.

3 Sobolev Orthogonal Polynomials

3.1 Sobolev Inner Product and Recurrence Relation

In contrast to the orthogonal polynomials considered so far, the inner product here involves not only function values, but also successive derivative values, all being endowed with their own measures. Thus,

$$\begin{aligned} (p,q)_S = &\int_{\mathbb{R}} p(t)q(t)\mathrm{d}\lambda_0(t) + \int_{\mathbb{R}} p'(t)q'(t)\mathrm{d}\lambda_1(t) \\ &+ \cdots + \int_{\mathbb{R}} p^{(s)}(t)q^{(s)}(t)\mathrm{d}\lambda_s(t), \quad s \geq 1. \end{aligned} \tag{3.1}$$

If all the measures $\mathrm{d}\lambda_\sigma$ are positive, the inner product (3.1) has associated with it a sequence of (monic) polynomials $\pi_k(\,\cdot\,; S)$, $k = 0, 1, 2, \ldots$, orthogonal in the sense

$$(\pi_k, \pi_\ell)_S \begin{cases} = 0, & k \neq \ell, \\ > 0, & k = \ell. \end{cases} \tag{3.2}$$

These are called *Sobolev orthogonal polynomials*. We cannot expect them to satisfy a three-term recurrence relation, since the inner product no longer

has the shift property $(tp, q) = (p, tq)$. However, like any sequence of monic polynomials of degrees $0, 1, 2, \ldots$, orthogonal or not, they must satisfy an extended recurrence relation of the type

$$\pi_{k+1}(t) = t\pi_k(t) - \sum_{j=0}^{k} \beta_j^k \pi_{k-j}(t), \quad k = 0, 1, 2, \ldots . \tag{3.3}$$

Associated with it is the upper Hessenberg *matrix of recurrence coefficients*

$$\boldsymbol{H}_n = \begin{bmatrix} \beta_0^0 & \beta_1^1 & \beta_2^2 & \cdots & \beta_{n-2}^{n-2} & \beta_{n-1}^{n-1} \\ 1 & \beta_0^1 & \beta_1^2 & \cdots & \beta_{n-3}^{n-2} & \beta_{n-2}^{n-1} \\ 0 & 1 & \beta_0^2 & \cdots & \beta_{n-4}^{n-2} & \beta_{n-3}^{n-1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \beta_0^{n-2} & \beta_1^{n-1} \\ 0 & 0 & 0 & \cdots & 1 & \beta_0^{n-1} \end{bmatrix} . \tag{3.4}$$

In the case $s = 0$ (of ordinary orthogonal polynomials) there holds $\beta_j^k = 0$ for $j > 1$, and the matrix $\boldsymbol{H}_n$ is tridiagonal. If symmetrized by a (real) diagonal similarity transformation, it becomes the Jacobi matrix $\boldsymbol{J}_n(\mathrm{d}\lambda_0)$. When $s > 0$, however, symmetrization of $\boldsymbol{H}_n$ is no longer possible, since $\boldsymbol{H}_n$ may well have complex eigenvalues (see Example 6).

3.2 Moment-Based Algorithm

There are now $s + 1$ sets of modified moments, one set for each measure $\mathrm{d}\lambda_\sigma$,

$$m_k^{(\sigma)} = \int_{\mathbb{R}} p_k(t)\mathrm{d}\lambda_\sigma, \quad k = 0, 1, 2, \ldots; \quad \sigma = 0, 1, \ldots, s. \tag{3.5}$$

The first $2n$ modified moments of all the sets will uniquely determine the matrix $\boldsymbol{H}_n$ in (3.4), i.e. there is a well-determined map

$$[m_k^{(\sigma)}]_{k=0}^{2n-1}, \ \sigma = 0, 1, \ldots, s \ \mapsto \boldsymbol{H}_n, \tag{3.6}$$

called *modified moment map* for Sobolev orthogonal polynomials. In the case where the polynomials p_k in (3.5) satisfy a three-term recurrence relation with known coefficients, and for $s = 1$, an algorithm has been developed that implements the map (3.6). It very much resembles the modified Chebyshev algorithm for ordinary orthogonal polynomials, but is technically much more elaborate (see [12]). The algorithm, however, is implemented in the `OPQ` routine

```
B=chebyshev_sob(N,mom,abm)
```

which produces the `N`×`N` upper triangular matrix `B` of recurrence coefficients, with β_j^k, $0 \le j \le k$, $0 \le k \le$`N`-1, occupying the position $(j + 1, k + 1)$ in the matrix. The input parameter `mom` is the $2 \times$ (`2N`) array of modified moments $m_k^{(\sigma)}$, $k = 0, 1, \ldots,$ `2N`-1; $\sigma = 0, 1$, of the two measures $\mathrm{d}\lambda_0$ and $\mathrm{d}\lambda_1$, and `abm`

the $(2N-1) \times 2$ array of coefficients a_k, b_k, $k = 0, 1, \ldots, 2N-2$, defining the polynomials p_k.

Example 4. Althammer's polynomials (1962).
These are the Sobolev polynomials relative to the measures $d\lambda_0(t) = dt$, $d\lambda_1(t) = \gamma dt$ on $[-1, 1]$, $\gamma > 0$.

A natural choice of modified moments are the Legendre moments, i.e. $p_k(t)$ is the monic Legendre polynomial of degree k. By orthogonality of the Legendre polynomials, all modified moments $m_k^{(0)}$ and $m_k^{(1)}$ are zero for $k > 0$, while $m_0^{(0)} = 2$ and $m_0^{(1)} = 2\gamma$. The following Matlab routine, therefore, can be used to generate the Althammer polynomials.

```
mom=zeros(2,2*N);
mom(1,1)=2; mom(2,1)=2*g;
abm=r_jacobi(2*N-1);
B=chebyshev_sob(N,mom,abm);
```

3.3 Discretization Algorithm

Taking the inner product of both sides of (3.3) with π_{k-j} gives

$$0 = (\pi_{k+1}, \pi_{k-j})_S = (t\pi_k, \pi_{k-j})_S - \beta_j^k (\pi_{k-j}, \pi_{k-j})_S, \quad j = 0, 1, \ldots, k,$$

hence

$$\beta_j^k = \frac{(t\pi_k, \pi_{k-j})_S}{(\pi_{k-j}, \pi_{k-j})_S}, \quad j = 0, 1, \ldots, k; \; k = 0, 1, \ldots, n-1. \tag{3.7}$$

These are the analogues of Darboux's formulae for ordinary orthogonal polynomials, and like these, can be combined with the recurrence relation (3.3) to successively build up the recurrence coefficients β_j^k in the manner of Stieltjes's procedure. The technical details, of course, are more involved, since we must generate not only the polynomials π_k, but also their derivatives, in order to be able to compute the Sobolev inner products in (3.7). This all is implemented, for arbitrary $s \geq 1$, in the Matlab routine `stieltjes_sob.m`. The basic assumption in the design of this routine is the availability, for each measure $d\lambda_\sigma$, of an n_σ-point quadrature rule

$$\int_{\mathbb{R}} p(t)\, d\lambda_\sigma(t) = \sum_{k=1}^{n_\sigma} w_k^{(\sigma)} p(x_k^{(\sigma)}), \quad p \in \mathbb{P}_{2(n-\sigma)-1}, \quad \sigma = 0, 1, \ldots, s, \tag{3.8}$$

that is exact for polynomials p of degree $\leq 2(n-\sigma) - 1$. These are typically Gaussian quadrature rules, possibly with discrete components (present in $d\lambda_\sigma$) added on. The information is supplied to the routine via the $1 \times (s+1)$ array

$$\mathtt{nd} = [n_0, n_1, \ldots, n_s]$$

and the md×(2s + 2) array

$$\texttt{xw}=\left[\begin{array}{ccc|ccc} x_1^{(0)} & \cdots & x_1^{(s)} & w_1^{(0)} & \cdots & w_1^{(s)} \\ x_2^{(0)} & \cdots & x_2^{(s)} & w_2^{(0)} & \cdots & w_2^{(s)} \\ \vdots & & \vdots & \vdots & & \vdots \\ x_{\mathrm{md}}^{(0)} & \cdots & x_{\mathrm{md}}^{(s)} & w_{\mathrm{md}}^{(0)} & \cdots & w_{\mathrm{md}}^{(s)} \end{array}\right]$$

where `md=max(nd)`. In each column of `xw` the entries after $x_{n_\sigma}^{(\sigma)}$ resp. $w_{n_\sigma}^{(\sigma)}$ (if any) are not used by the routine. Two more input parameters are needed; the first is `a0`, the coefficient $\alpha_0(\mathrm{d}\lambda_0)$, which allows us to initialize the matrix of recurrence coefficients,

$$\beta_0^0 = \frac{(t,1)_S}{(1,1)_S} = \frac{(t,1)_{\mathrm{d}\lambda_0}}{(1,1)_{\mathrm{d}\lambda_0}} = \alpha_0(\mathrm{d}\lambda_0).$$

The other, `same`, is a logical variable set equal to 1 if all quadrature rules have the same set of nodes, and equal to 0 otherwise. The role of this parameter is to switch to a simplified, and thus faster, procedure if `same=1`. A call to the routine, therefore, has the form

```
B=stieltjes_sob(N,s,nd,xw,a0,same)
```

Example 5. Althammer's polynomials, revisited.

Here, the obvious choice of the quadrature rule for $\mathrm{d}\lambda_0$ and $\mathrm{d}\lambda_1$ is the n-point Gauss–Legendre rule. This gives rise to the following routine:

```
s=1; nd=[N N];
a0=0; same=1;
ab=r_jacobi(N);
zw=gauss(N,ab);
xw=[zw(:,1) zw(:,1) ...
     zw(:,2) g*zw(:,2)];
B=stieltjes_sob(N,s,nd,xw,a0,same);
```

The results are identical with those obtained in Example 4.

3.4 Zeros

If we let $\boldsymbol{\pi}^{\mathrm{T}}(t) = [\pi_0(t), \pi_1(t), \ldots, \pi_{n-1}(t)]$, where π_k are the Sobolev orthogonal polynomials, then the recurrence relation (3.3) can be written in matrix form as

$$t\boldsymbol{\pi}^{\mathrm{T}}(t) = \boldsymbol{\pi}^{\mathrm{T}}(t)\boldsymbol{H}_n + \pi_n(t)\boldsymbol{e}_n^{\mathrm{T}} \tag{3.9}$$

in terms of the matrix $\boldsymbol{H}_n$ in (3.4). This immediately shows that the zeros τ_ν of π_n are the eigenvalues of $\boldsymbol{H}_n$ and $\boldsymbol{\pi}^{\mathrm{T}}(\tau_\nu)$ corresponding left eigenvectors. Naturally, there is no guarantee that the eigenvalues are real; some may well be

complex. Also, if n is large, there is a good chance that some of the eigenvalues are ill-conditioned.

The `OPQ` routine for the zeros of π_n is

```
z=sobzeros(n,N,B)
```

where `B` is the `N`×`N` matrix returned by `chebyshev_sob.m` or `stieltjes_sob.m`, and `z` the n-vector of the zeros of π_n, $1 \le n \le N$.

Example 6. Sobolev orthogonal polynomials with only a few real zeros (Meijer, 1994).

The Sobolev inner product in question is

$$(u,v)_S = \int_{-1}^{3} u(t)v(t)\,\mathrm{d}t + \gamma \int_{-1}^{1} u'(t)v'(t)\,\mathrm{d}t + \int_{1}^{3} u'(t)v'(t)\,\mathrm{d}t, \quad \gamma > 0. \tag{3.10}$$

Meijer proved that for $n(\text{even}) \ge 2$ and γ sufficiently large, the polynomial $\pi_n(\,\cdot\,;S)$ has exactly two real zeros, one in $[-3,-1]$ and the other in $[1,3]$. If $n(\text{odd}) \ge 3$, there is exactly one real zero, located in $[1,3]$, if γ is sufficiently large. We use the routine `stieltjes_sob.m` and `sobzeros.m` to illustrate this for $n = 6$ and $\gamma = 44{,}000$. (The critical value of γ above which Meijer's theorem takes hold is about $\gamma = 43{,}646.2$; see [10, Table 2.30].)

The inner product corresponds to the case $s = 1$ and

$$\mathrm{d}\lambda_0(t) = \mathrm{d}t \text{ on } [-1,3], \quad \mathrm{d}\lambda_1(t) = \begin{cases} \gamma \mathrm{d}t \text{ if } t \in [-1,1], \\ \mathrm{d}t \text{ if } t \in (1,3]. \end{cases}$$

Thus, we can write, with suitable transformations of variables,

$$\int_{-1}^{3} p(t)\,\mathrm{d}\lambda_0(t) = 2\int_{-1}^{1} p(2x+1)\,\mathrm{d}x,$$
$$\int_{-1}^{3} p(t)\,\mathrm{d}\lambda_1(t) = \int_{-1}^{1} [\gamma p(x) + p(x+2)]\,\mathrm{d}x$$

and apply n-point Gauss–Legendre quadrature to the integrals on the right. This will produce the matrix $\boldsymbol{H}_n$ exactly. The parameters in the routine `stieltjes_sob.m` have to be chosen as follows:

$$\mathtt{nd} = [n, 2n], \quad \mathtt{xw} = \begin{bmatrix} 2\tau_1^G + 1 & \tau_1^G & 2\lambda_1^G & \gamma\lambda_1^G \\ \vdots & \vdots & \vdots & \vdots \\ 2\tau_n^G + 1 & \tau_n^G & 2\lambda_n^G & \gamma\lambda_n^G \\ & \tau_1^G + 2 & & \lambda_1^G \\ & \vdots & & \vdots \\ & \tau_n^G + 2 & & \lambda_n^G \end{bmatrix} \in \mathbb{R}^{2n\times 4},$$

where τ_ν^G, λ_ν^G are the nodes and weight of the n-point Gauss–Legendre quadrature rule. Furthermore, a0=1 and same=0. The complete program, therefore, is as follows:

```
N=6; s=1; a0=1; same=0; g=44000; nd=[N 2*N];
ab=r_jacobi(N); zw=gauss(N,ab);
xw=zeros(2*N,2*(s+1));
xw(1:N,1)=2*zw(:,1)+1; xw(1:N,2)=zw(:,1);
xw(1:N,3)=2*zw(:,2); xw(1:N,4)=g*zw(:,2);
xw(N+1:2*N,2)=zw(:,1)+2; xw(N+1:2*N,4)=zw(:,2);
B=stieltjes_sob(N,s,nd,xw,a0,same);
z=sobzeros(N,N,B)
```

It produces the output

```
z =

 -4.176763898909848e-01 - 1.703657992747233e-01i
 -4.176763898909848e-01 + 1.703657992747233e-01i
  8.453761089539369e-01 - 1.538233952529940e-01i
  8.453761089539369e-01 + 1.538233952529940e-01i
 -1.070135059563751e+00
  2.598402134930250e+00
```

confirming Meijer's theorem for $n = 6$. A more detailed numerical study, also in the case of odd values of n, has been made in [10, Table 2.30].

Exercises to §3

1. Show that a Sobolev inner product does not satisfy the shift property $(tp, q) = (p, tq)$.
2. Prove (3.3).
3. The Sobolev inner product (3.1) is called *symmetric* if each measure $\mathrm{d}\lambda_\sigma$ is symmetric in the sense of Problem 6, §2. For symmetric Sobolev inner products,
   (a) show that $\pi_k(-t; S) = (-1)^k \pi_k(t; S)$;
   (b) show that $\beta_{2r}^k = 0$ for $r = 0, 1, \ldots, \lfloor k/2 \rfloor$.
4. Consider a Sobolev inner product with $s = 1$ and

$$\mathrm{d}\lambda_0 = \mathrm{d}\lambda, \quad \mathrm{d}\lambda_1 = \gamma \mathrm{d}\lambda \ (\gamma > 0),$$

   and $\mathrm{d}\lambda$ a symmetric measure. Use the routines `chebyshev_sob.m` and `sobzeros.m` to check numerically whether or not the positive zeros of the Sobolev orthogonal polynomials are monotonically increasing with γ. Experiment in turn with $\mathrm{d}\lambda(t) = \mathrm{d}t$ on $[-1, 1]$ (Althammer polynomials), $\mathrm{d}\lambda(t) = (1 - t^2)^\alpha$ on $[-1, 1]$, $\alpha > -1$, and $\mathrm{d}\lambda(t) = \exp(-t^2)$ on $\mathbb{R}$. Identify any computational problems and how to deal with them.
5. Special Sobolev orthogonal polynomials.

(a) Consider the special Sobolev orthogonal polynomials that have an absolutely continuous (or, possibly, discrete) ground measure $d\lambda_0$ and all $d\lambda_\sigma$, $1 \le \sigma \le s$, identically zero except for $d\lambda_{r_k}$, $k = 1, 2, \ldots, K$, where $1 \le r_1 < r_2 < \cdots < r_K \le s$, which are atomic measures located at the points c_k and having masses m_k. Assuming that the ground measure is given by the array `ab` of recurrence coefficients, write a Matlab routine `specsob.m` that uses the `OPQ` routine `stieltjes_sob.m` to compute the recurrence matrix `B` of the special Sobolev orthogonal polynomials.

(b) Use your routine together with the `OPQ` routine `sobzeros.m` to check Tables 2–4 in [8], relating to the Hermite measure $d\lambda_0(t) = \exp(-t^2)dt$ and a single atomic measure involving the rth derivative. In the cited reference, the results were obtained by a different method.

(c) In the case of the Laguerre measure $d\lambda_0(t) = \exp(-t)dt$ on $\mathbb{R}_+$ and $r_k = k$, $c_k = 0$, $m_k = 1$, it may be conjectured that any complex zero that occurs has negative real part. Use your routine and `sobzeros.m` to check out this conjecture.

(d) For $d\lambda_0$, r_k, c_k, m_k as in (c), determine the pattern of occurrence of complex zeros. Cover the range $1 \le s \le 10$, $1 \le n \le 40$.

(e) Repeat (c) and (d) with $d\lambda_0$ the Laguerre measure plus an atomic measure with mass 1 at the origin.

4 Quadrature

4.1 Gauss-Type Quadrature Formulae

Gauss Formula

Given a positive measure $d\lambda$, the n-point *Gaussian quadrature formula* associated with the measure $d\lambda$ is

$$\int_{\mathbb{R}} f(t)d\lambda(t) = \sum_{\nu=1}^{n} \lambda_\nu^G f(\tau_\nu^G) + R_n^G(f), \tag{4.1}$$

which has maximum algebraic degree of exactness $2n-1$,

$$R_n^G(f) = 0 \quad \text{if } f \in \mathbb{P}_{2n-1}. \tag{4.2}$$

It is well known that the nodes τ_ν^G are the zeros of $\pi_n(\cdot\,; d\lambda)$, and hence the eigenvalues of the Jacobi matrix $\boldsymbol{J}_n(d\lambda)$; cf. §2.1. Interestingly, the weights λ_ν^G, too, can be expressed in terms of spectral data of $\boldsymbol{J}_n(d\lambda)$; indeed, they are (Golub and Welsch, 1969)

$$\lambda_\nu^G = \beta_0 v_{\nu,1}^2, \tag{4.3}$$

where $v_{\nu,1}$ is the first component of the normalized eigenvector v_ν corresponding to the eigenvalue τ_ν^G,

$$J_n(\mathrm{d}\lambda)v_\nu = \tau_\nu^G v_\nu, \quad v_\nu^{\mathrm{T}} v_\nu = 1, \tag{4.4}$$

and, as usual, $\beta_0 = \int_{\mathbb{R}} \mathrm{d}\lambda(t)$. This is implemented in the `OPQ` Matlab routine

```
xw=gauss(N,ab)
```

where `ab`, as in all previous routines, is the `N`×2 array of recurrence coefficients for $\mathrm{d}\lambda$, and `xw` the `N`×2 array containing the nodes τ_ν^G in the first column, and the weights λ_ν^G in the second.

We remark, for later purposes, that the Gauss quadrature sum, for f sufficiently regular, can be expressed in matrix form as

$$\sum_{\nu=1}^{n} \lambda_\nu^G f(\tau_\nu^G) = \beta_0 e_1^{\mathrm{T}} f(J_n(\mathrm{d}\lambda))e_1, \quad e_1 = [1, 0, \ldots, 0]^{\mathrm{T}}. \tag{4.5}$$

This is an easy consequence of (4.3) and the spectral decomposition of J_n,

$$J_n(\mathrm{d}\lambda)V = VD_\tau, \quad D_\tau = \mathrm{diag}(\tau_1^G, \tau_2^G, \ldots, \tau_n^G),$$

where $V = [v_1, v_2, \ldots, v_n]$.

Example 7. Zeros of Sobolev orthogonal polynomials of Gegenbauer type (Groenevelt, 2002).

The polynomials in question are those orthogonal with respect to the Sobolev inner product

$$(u, v)_S = \int_{-1}^{1} u(t)v(t)(1-t^2)^{\alpha-1}\mathrm{d}t + \gamma \int_{-1}^{1} u'(t)v'(t)\frac{(1-t^2)^\alpha}{t^2+y^2}\,\mathrm{d}t.$$

Groenevelt proved that in the case $\gamma \to \infty$ the Sobolev orthogonal polynomials of even degrees $n \geq 4$ have complex zeros if y is sufficiently small. By symmetry, they must in fact be purely imaginary, and by the reality of the Sobolev polynomials, must occur in conjugate complex pairs. As we illustrate this theorem, we have an opportunity to apply not only the routine `gauss.m`, but also a number of other routines, specifically the modification algorithm embodied in the routine `chri6.m`, dealing with the special quadratic divisor $t^2 + y^2$ in the second integral, and the routine `stieltjes_sob.m` generating the recurrence matrix of the Sobolev orthogonal polynomials:

```
s=1; same=0; eps0=1e-14; numax=250; nd=[N N];
ab0=r_jacobi(numax,alpha);
z=complex(0,y);
nu0=nu0jac(N,z,eps0); rho0=0; iopt=1;
ab1=chri6(N,ab0,y,eps0,nu0,numax,rho0,iopt);
zw1=gauss(N,ab1);
ab=r_jacobi(N,alpha-1); zw=gauss(N,ab);
xw=[zw(:,1) zw1(:,1) zw(:,2) gamma*zw1(:,2)];
a0=ab(1,1); B=stieltjes_sob(N,s,nd,xw,a0,same);
z=sobzeros(N,N,B)
```

Demo#4 The case N=12, $\alpha = \frac{1}{2}$, and $\gamma = 1$ of Example 7.

Applying the above routine for $y = .1$ and $y = .09$ yields the following zeros (with positive imaginary parts; the other six zeros are the same with opposite signs):

y	zeros	y	zeros
.1	.027543282225	.09	.011086169153 i
	.284410786673		.281480077515
	.541878443180		.540697645595
	.756375307278		.755863108617
	.909868274113		.909697039063
	.989848649239		.989830182743

The numerical results (and additional tests) suggest that Groenevelt's theorem also holds for finite, not necessarily large, values of γ, and, when $\gamma = 1$, that the critical value of y below which there are complex zeros must be betweeen .09 and .1.

Gauss–Radau Formula

If there is an interval $[a, \infty]$, $-\infty < a$, containing the support of $d\lambda$, it may be desirable to have an $(n+1)$-point quadrature rule of maximum degree of exactness that has a as a prescribed node,

$$\int_{\mathbb{R}} f(t)d\lambda(t) = \lambda_0^a f(a) + \sum_{\nu=1}^{n} \lambda_\nu^a f(\tau_\nu^a) + R_n^a(f). \tag{4.6}$$

Here, $R_n^a(f) = 0$ for all $f \in \mathbb{P}_{2n}$, and τ_ν^a are the zeros of $\pi_n(\cdot\,; d\lambda_a)$, $d\lambda_a(t) = (t-a)d\lambda(t)$. This is called the *Gauss–Radau formula*. There is again a symmetric, tridiagonal matrix, the *Jacobi–Radau matrix*

$$\boldsymbol{J}_{n+1}^{R,a}(d\lambda) = \begin{bmatrix} \boldsymbol{J}_n(d\lambda) & \sqrt{\beta_n}\boldsymbol{e}_n \\ \sqrt{\beta_n}\boldsymbol{e}_n^T & \alpha_n^R \end{bmatrix}, \quad \alpha_n^R = a - \beta_n \frac{\pi_{n-1}(a)}{\pi_n(a)}, \tag{4.7}$$

where $\boldsymbol{e}_n = [0, 0, \ldots, 1]^{\mathrm{T}} \in \mathbb{R}^n$, $\beta_n = \beta_n(d\lambda)$, and $\pi_k(\cdot) = \pi_k(\cdot\,; d\lambda)$, which allows the Gauss–Radau formula to be characterized in terms of eigenvalues and eigenvectors: all nodes of (4.6), including the node a, are the eigenvalues of (4.7), and the weights λ_ν^a expressible as in (4.3) in terms of the corresponding normalized eigenvectors $\boldsymbol{v}_\nu$ of (4.7),

$$\lambda_\nu^a = \beta_0 v_{\nu,1}^2, \quad \nu = 0, 1, 2, \ldots, n. \tag{4.8}$$

As in (4.5), this implies that the Gauss–Radau quadrature sum, for smooth f, can be expressed as $\beta_0 \boldsymbol{e}_1^{\mathrm{T}} f(\boldsymbol{J}_{n+1}^{R,a})\boldsymbol{e}_1$, where $\boldsymbol{e}_1 = [1, 0, \ldots, 0] \in \mathbb{R}^{n+1}$.

Naturally, if the support of $d\lambda$ is contained in an interval $[-\infty, b]$, $b < \infty$, there is a companion formula to (4.6) which has the prescribed node b,

$$\int_{\mathbb{R}} f(t)\,\mathrm{d}\lambda(t) = \sum_{\nu=1}^{n} \lambda_\nu^b f(\tau_\nu^b) + \lambda_{n+1}^b f(b) + R_n^b(f). \tag{4.9}$$

The eigenvalue/vector characterization also holds for (4.9) if in the formula for α_n^R in (4.7), the variable a, at every occurrence, is replaced by b.

The remainder terms of (4.6) and (4.9), if $f \in C^{2n+1}[a,b]$, have the useful property

$$R_n^a(f) > 0, \quad R_n^b(f) < 0 \quad \text{if } \operatorname{sgn} f^{(2n+1)} = 1 \text{ on } [a,b], \tag{4.10}$$

with the inequalities reversed if $\operatorname{sgn} f^{(2n+1)} = -1$.

For Jacobi resp. generalized Laguerre measures with parameters α, β resp. α, the quantity α_n^R is explicitly known (cf. [10, Examples 3.4 and 3.5]). For example, if $a = -1$ (in the case of Jacobi measures),

$$\alpha_n^R = -1 + \frac{2n(n+\alpha)}{(2n+\alpha+\beta)(2n+\alpha+\beta+1)}, \quad \alpha_n^R = n, \tag{4.11}$$

whereas for $a = 1$, the sign of α_n^R must be changed and α and β interchanged.

The respective `OPQ` Matlab routines are

```
xw=radau(N,ab,end0)
xw=radau_jacobi(N,iopt,a,b)
xw=radau_laguerre(N,a)
```

In the first, `ab` is the `(N+1)×2` array of recurrence coefficients for $\mathrm{d}\lambda$, and `end0` either a (for (4.6)) or b (for (4.9)). The last two routines make use of the explicit formulae for α_n^R in the case of Jacobi resp. Laguerre measures, the parameters being α=`a`, β=`b`. The parameter `iopt` chooses between the two Gauss–Radau formulae: the left-handed, if `iopt`=1, the right-handed otherwise.

Gauss–Lobatto Formula

If the support of $\mathrm{d}\lambda$ is contained in the finite interval $[a,b]$, we may wish to prescribe two nodes, the points a and b. Maximizing the degree of exactness subject to these constraints yields the *Gauss–Lobatto formula*

$$\int_a^b f(t)\mathrm{d}\lambda(t) = \lambda_0^L f(a) + \sum_{\nu=1}^{n} \lambda_\nu^L f(\tau_\nu^L) + \lambda_{n+1}^L f(b) + R_n^{a,b}(f), \tag{4.12}$$

which we write as an $(n+2)$-point formula; we have $R_n^{a,b}(f) = 0$ for $f \in \mathbb{P}_{2n+1}$. The internal nodes τ_ν^L are the zeros of the polynomial $\pi_n(\,\cdot\,;\mathrm{d}\lambda_{a,b})$, $\mathrm{d}\lambda_{a,b}(t) = (t-a)(b-t)\mathrm{d}\lambda(t)$. All nodes and weights can be expressed in terms of eigenvalues and eigenvectors exactly as in the two preceding subsections, except that the matrix involved is the *Jacobi–Lobatto matrix*

$$J_{n+2}^L(\mathrm{d}\lambda) = \begin{bmatrix} J_{n+1}(\mathrm{d}\lambda) & \sqrt{\beta_{n+1}^L}\, e_{n+1} \\ \sqrt{\beta_{n+1}^L}\, e_{n+1}^{\mathrm{T}} & \alpha_{n+1}^L \end{bmatrix}, \tag{4.13}$$

where α_{n+1}^L and β_{n+1}^L are the solution of the 2×2 system of linear equations

$$\begin{bmatrix} \pi_{n+1}(a) & \pi_n(a) \\ \pi_{n+1}(b) & \pi_n(b) \end{bmatrix} \begin{bmatrix} \alpha_{n+1}^L \\ \beta_{n+1}^L \end{bmatrix} = \begin{bmatrix} a\pi_{n+1}(a) \\ b\pi_{n+1}(b) \end{bmatrix}. \tag{4.14}$$

For smooth f, the quadrature sum is expressible as $\beta_0 e_1^{\mathrm{T}} f(J_{n+2}^L) e_1$. For $f \in C^{2n+2}[a,b]$ with constant sign on $[a,b]$, the remainder $R_n^{a,b}(f)$ satisfies

$$R_n^{a,b}(f) < 0 \quad \text{if } \operatorname{sgn} f^{(2n+2)} = 1 \text{ on } [a,b], \tag{4.15}$$

with the inequality reversed if $\operatorname{sgn} f^{(2n+2)} = -1$.

The quantities α_{n+1}^L, β_{n+1}^L for Jacobi measures on $[-1,1]$ with parameters α, β and $a = -b = -1$ are explicitly known (cf. [10, Example 3.8]),

$$\begin{aligned} \alpha_{n+1}^L &= \frac{\alpha - \beta}{2n + \alpha + \beta + 2}, \\ \beta_{n+1}^L &= 4\frac{(n+\alpha+1)(n+\beta+1)(n+\alpha+\beta+1)}{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)}. \end{aligned} \tag{4.16}$$

The OPQ Matlab routines are

```
xw=lobatto(N,ab,endl,endr)
xw=lobatto_jacobi(N,a,b)
```

with the meaning of `ab`, `a`, `b` the same as in the Gauss–Radau routines, and `endl`=a, `endr`=b.

We remark that both Gauss–Radau and Gauss–Lobatto formulae can be generalized to include boundary points of multiplicity $r > 1$. The internal (simple) nodes and weights are still related to orthogonal polynomials, but the boundary weights require new techniques for their computation; see Exercises 12–13.

4.2 Gauss–Kronrod Quadrature

In an attempt to estimate the error of the n-point Gauss quadrature rule, Kronrod in 1964 had the idea of inserting $n+1$ additional nodes and choosing them, along with all $2n+1$ weights, in such a way as to achieve maximum degree of exactness. The resulting quadrature rule can be expected to yield much higher accuracy than the Gauss formula, so that the difference of the two provides an estimate of the error in the Gauss formula. The extended formula thus can be written in the form

$$\int_{\mathbb{R}} f(t)\mathrm{d}\lambda(t) = \sum_{\nu=1}^{n} \lambda_\nu^K f(\tau_\nu^G) + \sum_{\mu=1}^{n+1} \lambda_\mu^{*K} f(\tau_\mu^K) + R_n^{GK}(f), \tag{4.17}$$

and having $3n+2$ free parameters λ_ν^K, λ_μ^{*K}, τ_μ^K at disposal, one ought to be able to achieve degree of exactness $3n+1$,

$$R_n^{GK}(f) = 0 \quad \text{for } f \in \mathbb{P}_{3n+1}. \tag{4.18}$$

A quadrature formula (4.17) that satisfies (4.18) is called a *Gauss–Kronrod formula*. The nodes τ_μ^K, called *Kronrod nodes*, are the zeros of the polynomial π_{n+1}^K of degree $n+1$ which is orthogonal to all polynomials of lower degree in the sense

$$\int_{\mathbb{R}} \pi_{n+1}^K(t)p(t)\pi_n(t;\mathrm{d}\lambda)\,\mathrm{d}\lambda(t) = 0 \quad \text{for all } p \in \mathbb{P}_n. \tag{4.19}$$

Note that the measure of orthogonality here is $\pi_n(t;\mathrm{d}\lambda)\mathrm{d}\lambda(t)$ and thus oscillates on the support of $\mathrm{d}\lambda$. Stieltjes (1894) was the first to consider polynomials π_{n+1}^K of this kind (for $\mathrm{d}\lambda(t) = \mathrm{d}t$); a polynomial π_{n+1}^K satisfying (4.19) is therefore called a *Stieltjes polynomial*. Stieltjes conjectured (in the case $\mathrm{d}\lambda(t) = \mathrm{d}t$) that all zeros of π_{n+1}^K are real and interlace with the n Gauss nodes— a highly desirable configuration! This has been proved only later by Szegő (1935) not only for Legendre measures, but also for a class of Gegenbauer measures. The study of the reality of the zeros for more general measures is an interesting and ongoing activity.

The computation of Gauss–Kronrod formulae is a challenging problem. An elegant solution has been given recently by Laurie (1997), at least in the case when a Gauss–Kronrod formula exists with real nodes and positive weights. It can be computed again in terms of eigenvalues and eigenvectors of a symmetric tridiagonal matrix, just like the previous Gauss-type formulae. The relevant matrix, however, is the *Jacobi–Kronrod matrix*

$$\boldsymbol{J}_{2n+1}^K(\mathrm{d}\lambda) = \begin{bmatrix} \boldsymbol{J}_n(\mathrm{d}\lambda) & \sqrt{\beta_n}\boldsymbol{e}_n & \boldsymbol{0} \\ \sqrt{\beta_n}\boldsymbol{e}_n^{\mathrm{T}} & \alpha_n & \sqrt{\beta_{n+1}}\boldsymbol{e}_1^{\mathrm{T}} \\ \boldsymbol{0} & \sqrt{\beta_{n+1}}\boldsymbol{e}_1 & \boldsymbol{J}_n^* \end{bmatrix}. \tag{4.20}$$

Here, $\alpha_n = \alpha_n(\mathrm{d}\lambda)$, $\beta_n = \beta_n(\mathrm{d}\lambda)$, etc, and $\boldsymbol{J}_n^*$ (which is partially known) can be computed by *Laurie's algorithm* (cf. [10, §3.1.2.2]). Should some of the eigenvalues of (4.20) turn out to be complex, this would be an indication that a Gauss–Kronrod formula (with real nodes) does not exist.

There are two routines in `OPQ`,

```
ab=r_kronrod(N,ab0)
xw=kronrod(n,ab)
```

that serve to compute Gauss–Kronrod formulae. The first generates the Jacobi-Kronrod matrix of order `2N+1`, the other the nodes and weights of

the quadrature formula, stored respectively in the first and second column of the (2N+1)×2 array xw. The recurrence coefficients of the given measure $d\lambda$ are input via the $\lceil 3N/2+1 \rceil \times 2$ array ab0.

4.3 Gauss–Turán Quadrature

The idea of allowing derivatives to appear in a Gauss-type quadrature formula is due to Turán (1950). He considered the case where each node has the same multiplicity $r \geq 1$, that is,

$$\int_{\mathbb{R}} f(t)\,d\lambda(t) = \sum_{\nu=1}^{n} [\lambda_\nu f(\tau_\nu) + \lambda'_\nu f'(\tau_\nu) + \cdots + \lambda_\nu^{(r-1)} f^{(r-1)}(\tau_\nu)] + R_n(f). \tag{4.21}$$

This is clearly related to Hermite interpolation. Indeed, if all nodes were prescribed and distinct, one could use Hermite interpolation to obtain a formula with degree of exactness $rn-1$ (there are rn free parameters). Turán asked, like Gauss before him, whether one can do better by choosing the nodes τ_ν judiciously. The answer is yes; more precisely, we can get degree of exactness $rn-1+k$, $k>0$, if and only if

$$\int_{\mathbb{R}} \omega_n^r(t) p(t)\,d\lambda(t) = 0 \quad \text{for all } p \in \mathbb{P}_{k-1}, \tag{4.22}$$

where $\omega_n(t) = \prod_{\nu=1}^{n}(t-\tau_\nu)$ is the *node polynomial* of (4.21). We have here a new type of orthogonality: the rth power of ω_n, not ω_n, must be orthogonal to all polynomials of degree $k-1$. This is called *power orthogonality*. It is easily seen that r must be odd,

$$r = 2s+1, \quad s \geq 0, \tag{4.23}$$

so that (4.21) becomes

$$\int_{\mathbb{R}} f(t)\,d\lambda(t) = \sum_{\nu=1}^{n} \sum_{\sigma=0}^{2s} \lambda_\nu^{(\sigma)} f^{(\sigma)}(\tau_\nu) + R_{n,s}(f). \tag{4.24}$$

Then in (4.22), necessarily $k \leq n$, and $k=n$ is optimal. The maximum possible degree of exactness, therefore, is $(2s+2)n-1$, and is achieved if

$$\int_{\mathbb{R}} [\omega_n(t)]^{2s+1} p(t)\,d\lambda(t) = 0 \quad \text{for all } p \in \mathbb{P}_{n-1}. \tag{4.25}$$

The polynomial $\omega_n = \pi_{n,s}$ satisfying (4.25) is called *s-orthogonal*. It exists uniquely and has distinct simple zeros contained in the support interval of $d\lambda$. The formula (4.24) is the *Gauss-Turán formula* if its node polynomial ω_n satisfies (4.25) and the weights $\lambda_\nu^{(\sigma)}$ are obtained by Hermite interpolation.

The computation of Gauss-Turán formulae is not as simple as in the case of ordinary Gauss-type formulae. The basic idea, however, is to consider the positive measure $d\lambda_{n,s}(t) = [\pi_{n,s}(t)]^{2s} d\lambda(t)$ and to note that $\pi_{n,s}$ is the nth-degree polynomial orthogonal relative to $d\lambda_{n,s}$. The difficulty is that this defines $\pi_{n,s}$ implicitly, since $\pi_{n,s}$ already occurs in the measure $d\lambda_{n,s}$. Nevertheless, the difficulty can be surmounted, but at the expense of having to solve a system of nonlinear equations; for details, see [10, §3.1.3.2]. The procedure is embodied in the `OPQ` routine

```
xw=turan(n,s,eps0,ab0,hom)
```

where the nodes are stored in the first column of the `n`×(2`s`+2) array `xw`, and the successive weights in the remaining 2`s`+1 columns. The input parameter `eps0` is an error tolerance used in the iterative solution of the nonliner system of equations, and the measure $d\lambda$ is specified by the ((`s`+1)`n`)×2 input array `ab0` of its recurrence coefficients. Finally, `hom=1` or `hom`$\neq$`1` depending on whether or not a certain homotopy in the variable s is used to facilitate convergence of Newton's method for solving the system of nonlinear equations.

4.4 Quadrature Formulae Based on Rational Functions

All quadrature formulae considered so far were based on polynomial degree of exactness. This is meaningful if the integrand is indeed polynomial-like. Not infrequently, however, it happens that the integrand has poles outside the interval of integration. In this case, exactness for appropriate rational functions, in addition to polynomials, is more natural. We discuss this for the simplest type of quadrature rule,

$$\int_{\mathbb{R}} g(t) d\lambda(t) = \sum_{\nu=1}^{n} \lambda_\nu g(\tau_\nu) + R_n(g). \tag{4.26}$$

The problem, more precisely, is to determine λ_ν, τ_ν such that $R_n(g) = 0$ if $g \in \mathbb{S}_{2n}$, where $\mathbb{S}_{2n}$ is a space of dimension $2n$ consisting of rational functions and polynomials,

$$\begin{aligned} &\mathbb{S}_{2n} = \mathbb{Q}_m \oplus \mathbb{P}_{2n-m-1}, \quad 0 \le m \le 2n, \\ &\mathbb{P}_{2n-m-1} = \text{polynomials of degree } \le 2n-m-1, \\ &\mathbb{Q}_m = \text{rational functions with prescribed poles.} \end{aligned} \tag{4.27}$$

Here, m is an integer of our choosing, and

$$\mathbb{Q}_m = \text{span}\left\{ r(t) = \frac{1}{1+\zeta_\mu t}, \quad \mu = 1, 2, \ldots, m \right\}, \tag{4.28}$$

where

$$\zeta_\mu \in \mathbb{C}, \quad \zeta_\mu \neq 0, \quad 1 + \zeta_\mu t \neq 0 \text{ on } \operatorname{supp}(d\lambda). \tag{4.29}$$

The idea is to select the poles $-1/\zeta_\mu$ of the rational functions in $\mathbb{Q}_m$ to match the pole(s) of g closest to the support interval of $d\lambda$.

In principle, the solution of the problem is rather simple: put $\omega_m(t) = \prod_{\mu=1}^m (1 + \zeta_\mu t)$ and construct, if possible, the n-point (poynomial) Gauss formula

$$\int_{\mathbb{R}} g(t) \frac{d\lambda(t)}{\omega_m(t)} = \sum_{\nu=1}^n \lambda_\nu^G g(\tau_\nu^G), \quad g \in \mathbb{P}_{2n-1}, \tag{4.30}$$

for the modified measure $d\hat{\lambda}(t) = d\lambda(t)/\omega_m(t)$. Then

$$\tau_\nu = \tau_\nu^G, \quad \lambda_\nu = \omega_m(\tau_\nu^G)\lambda_\nu^G, \quad \nu = 1, 2, \ldots, n, \tag{4.31}$$

are the desired nodes and weights in (4.26).

We said "if possible", since in general ω_m is complex-valued, and the existence of a Gauss formula for $d\hat{\lambda}$ is not guaranteed. There is no problem, however, if $\omega_m \geq 0$ on the support of $d\lambda$. Fortunately, in many instances of practical interest, this is indeed the case.

There are a number of ways the formula (4.30) can be constructed: a discretization method using Gauss quadrature relative to $d\lambda$ to do the discretization; repeated application of modification algorithms involving linear or quadratic divisors; special techniques to handle "difficult" poles, that is, poles very close to the support interval of $d\lambda$. Rather than going into details (which can be found in [10, §3.1.4]), we present an example taken from solid state physics.

Example 8. Generalized Fermi–Dirac integral.

This is the integral

$$F_k(\eta, \theta) = \int_0^\infty \frac{t^k \sqrt{1 + \theta t/2}}{e^{-\eta + t} + 1} \, dt,$$

where $\eta \in \mathbb{R}$, $\theta \geq 0$, and k is the Boltzmann constant ($=\frac{1}{2}, \frac{3}{2}$, or $\frac{5}{2}$). The ordinary Fermi–Dirac integral corresponds to $\theta = 0$.

The integral is conveniently rewritten as

$$F_k(\eta, \theta) = \int_0^\infty \frac{\sqrt{1 + \theta t/2}}{e^{-\eta} + e^{-t}} \, d\lambda^{[k]}(t), \quad d\lambda^{[k]}(t) = t^k e^{-t} dt, \tag{4.32}$$

which is of the form (4.26) with $g(t) = \sqrt{1 + \theta t/2}/(e^{-\eta} + e^{-t})$ and $d\lambda = d\lambda^{[k]}$ a generalized Laguerre measure. The poles of g evidently are $\eta + \mu i\pi$, $\mu = \pm 1, \pm 3, \pm 5, \ldots$, and all are "easy", that is, at a comfortable distance from the interval $[0, \infty]$. It is natural to take m even, and to incorporate the first $m/2$ pairs of conjugate complex poles. An easy computation then yields

$$\omega_m(t) = \prod_{\nu=1}^{m/2} [(1 + \xi_\nu t)^2 + \eta_\nu t^2], \quad 2 \leq m(\text{even}) \leq 2n, \tag{4.33}$$

where

$$\xi_\nu = \frac{-\eta}{\eta^2 + (2\nu-1)^2\pi^2}, \quad \eta_\nu = \frac{(2\nu-1)\pi}{\eta^2 + (2\nu-1)^2\pi^2}. \tag{4.34}$$

Once the nodes and weights τ_ν, λ_ν have been obtained according to (4.31), the rational/polynomial quadrature approximation is given by

$$F_k(\eta,\theta) \approx \sum_{n=1}^{N} \lambda_n \frac{\sqrt{1+\theta\tau_n/2}}{\mathrm{e}^{-\eta} + \mathrm{e}^{-\tau_n}}. \tag{4.35}$$

It is computed in the `OPQ` routine

```
xw=fermi_dirac(N,m,eta,theta,k,eps0,Nmax)
```

where `eps0` is an error tolerance, `Nmax` a limit on the discretization parameter, and the other variables having obvious meanings.

4.5 Cauchy Principal Value Integrals

When there is a (simple) pole inside the support interval $[a,b]$ of the measure $\mathrm{d}\lambda$, the integral must be taken in the sense of a *Cauchy principal value integral*

$$(\mathcal{C}f)(x;\mathrm{d}\lambda) := \int_a^b \frac{f(t)}{x-t}\,\mathrm{d}\lambda(t) = \lim_{\epsilon\downarrow 0}\left(\int_a^{x-\epsilon} + \int_{x+\varepsilon}^b\right)\frac{f(t)}{x-t}\,\mathrm{d}\lambda(t), \quad x\in(a,b). \tag{4.36}$$

There are two types of quadrature rules for Cauchy principal value integrals: one in which x occurs as a node, and one in which it does not. They have essentially different character and will be considered separately.

Modified Quadrature Rule

This is a quadrature rule of the form

$$(\mathcal{C}f)(x;\mathrm{d}\lambda) = c_0(x)f(x) + \sum_{\nu=1}^{n} c_\nu(x) f(\tau_\nu) + R_n(f;x). \tag{4.37}$$

It can be made "Gaussian", that is, $R_n(f;x) = 0$ for $f \in \mathbb{P}_{2n}$, by rewriting the integral in (4.36) as

$$(\mathcal{C}f)(x;\mathrm{d}\lambda) = f(x)\int_{\mathbb{R}} \frac{\mathrm{d}\lambda(t)}{x-t} - \int_{\mathbb{R}} \frac{f(x)-f(t)}{x-t}\,\mathrm{d}\lambda(t) \tag{4.38}$$

and applying the n-point Gauss formula for $\mathrm{d}\lambda$ to the second integral. The result is

$$(\mathcal{C}f)(x;\mathrm{d}\lambda) = \frac{\rho_n(x)}{\pi_n(x)} f(x) + \sum_{\nu=1}^{n} \lambda_\nu^G \frac{f(\tau_\nu^G)}{x-\tau_\nu^G} + R_n(f;x), \tag{4.39}$$

where $\rho_n(x)$ is the Cauchy principal value integral (2.33) and τ_ν^G, λ_ν^G are the Gauss nodes and weights for $d\lambda$.

Formula (4.39) is not without numerical difficulties. The major one occurs when x approaches one of the Gauss nodes τ_ν^G, in which case two terms on the right go to infinity, but with opposite signs. In effect, this means that for x near a Gaussian node severe cancellation must occur.

The problem can be avoided by expanding the integral (4.36) in Cauchy integrals $\rho_k(x)$. Let $p_n(f;\cdot)$ be the polynomial of degree n interpolating f at the n Gauss nodes τ_ν^G and at x. The quadrature sum in (4.39) is then precisely the Cauchy integral of p_n,

$$(\mathcal{C}f)(x;d\lambda) = \int_a^b \frac{p_n(f;t)}{x-t}\,d\lambda(t) + R_n(f;x). \tag{4.40}$$

Expanding p_n in the orthogonal polynomials π_k,

$$p_n(f;t) = \sum_{k=0}^{n} a_k\pi_k(t), \quad a_k = \frac{1}{\|\pi_k\|^2}\int_a^b p_n(f;t)\pi_k(t)d\lambda(t), \tag{4.41}$$

and integrating, one finds

$$(\mathcal{C}f)(x;d\lambda) = \sum_{k=0}^{n} a_k\rho_k(x) + R_n(f;x), \tag{4.42}$$

where

$$a_k = \frac{1}{\|\pi_k\|^2}\sum_{\nu=1}^{n}\lambda_\nu^G f(\tau_\nu^G)\pi_k(\tau_\nu^G), \ k<n; \quad a_n = \sum_{\nu=1}^{n}\frac{f(x)-f(\tau_\nu^G)}{(x-\tau_\nu^G)\pi_n'(\tau_\nu^G)}. \tag{4.43}$$

The Cauchy integrals $\rho_k(x)$ in (4.42) can be computed in a stable manner by forward recursion; cf. the paragraph surrounding (2.33) and (2.34). This requires $\rho_0(x)$, which is either explicitly known or can be computed by the continued fraction algorithm. Some care must be exercised in computing the divided difference of f in the formula for a_n.

The procedure is inplemented in the `OPQ` routine

```
cpvi=cauchyPVI(N,x,f,ddf,iopt,ab,rho0)
```

with `iopt`$\neq$1, which produces the (`N`+1)-term approximation (4.42) where $R_n(f;x)$ is neglected. The input parameter `ddf` is a routine for computing the divided difference of f in a stable manner. It is used only if `iopt`$\neq$1. The meaning of the other parameters is obvious.

Quadrature Rule in the Strict Sense

This rule, in which the node $t = x$ is absent, is obtained by interpolating f at the n Gauss nodes τ_ν^G by a polynomial $p_{n-1}(f;\cdot)$ of degree $n-1$,

$$f(t) = p_{n-1}(f;t) + E_{n-1}(f;t), \quad p_{n-1}(f;t) = \sum_{\nu=1}^{n} \frac{\pi_n(t)}{(t-\tau_\nu^G)\pi_n'(\tau_\nu^G)} f(\tau_\nu^G),$$

where E_{n-1} is the interpolation error, which vanishes identically if $f \in \mathbb{P}_{n-1}$. The formula to be derived, therefore, will have degree of exactness $n-1$ (which can be shown to be maximum possible). Integrating in the sense of (4.36) yields

$$(\mathcal{C}f)(x;\mathrm{d}\lambda) = \sum_{\nu=1}^{n} \frac{\rho_n(x) - \rho_n(\tau_\nu^G)}{(x-\tau_\nu^G)\pi_n'(\tau_\nu^G)} f(\tau_\nu^G) + R_n^*(f;x), \tag{4.44}$$

where $R_n^*(f;x) = \int_a^b E_{n-1}(f;t)\mathrm{d}\lambda(t)/(x-t)$.

This formula, too, suffers from severe cancellation errors when x is near a Gauss node. The resolution of this problem is similar (in fact, simpler) as in the previous subsection: expand $p_{n-1}(f;\cdot)$ in the orthogonal polynomials π_k to obtain

$$\begin{gathered}(\mathcal{C}f)(x;\mathrm{d}\lambda) = \sum_{k=0}^{n-1} a_k' \rho_k(x) + R_n^*(f;x), \\ a_k' = \frac{1}{\|\pi_k\|^2} \int_a^b p_{n-1}(f;t)\pi_k(t)\mathrm{d}\lambda(t).\end{gathered} \tag{4.45}$$

It turns out that

$$a_k' = a_k, \quad k = 0, 1, \ldots, n-1, \tag{4.46}$$

where a_k, $k < n$, is given by (4.43). This is implemented in the `OPQ` routine `cauchyPVI.m` with `iopt=1`.

4.6 Polynomials Orthogonal on Several Intervals

We are given a finite set of intervals $[c_j, d_j]$, which may be disjoint or not, and on each interval a positive measure $\mathrm{d}\lambda_j$. Let $\mathrm{d}\lambda$ be the "composite" measure

$$\mathrm{d}\lambda(t) = \sum_j \chi_{[c_j,d_j]}(t)\mathrm{d}\lambda_j(t), \tag{4.47}$$

where $\chi_{[c_j,d_j]}$ is the characteristic function of the interval $[c_j, d_j]$. Assuming known the Jacobi matrices $\boldsymbol{J}^{(j)} = \boldsymbol{J}_n(\mathrm{d}\lambda_j)$ of the component measures $\mathrm{d}\lambda_j$, we now consider the problem of determining the Jacobi matrix $\boldsymbol{J} = \boldsymbol{J}_n(\mathrm{d}\lambda)$ of the composite measure $\mathrm{d}\lambda$. We provide two solutions, one based on Stieltjes's procedure, and one based on the modified Chebyshev algorithm.

Solution by Stieltjes's Procedure

The main problem in applying Stieltjes's procedure is to compute the inner products $(t\pi_k, \pi_k)_{\mathrm{d}\lambda}$ and $(\pi_k, \pi_k)_{\mathrm{d}\lambda}$ for $k = 0, 1, 2, \ldots, n-1$. This can be done by using n-point Gaussian quadrature on each component interval,

$$\int_{c_j}^{d_j} p(t)\mathrm{d}\lambda_j(t) = \sum_{\nu=1}^{n} \lambda_\nu^{(j)} p(\tau_\nu^{(j)}), \quad p \in \mathbb{P}_{2n-1}. \tag{4.48}$$

Here we use (4.5) to express the quadrature sum in terms of the Jacobi matrix $\boldsymbol{J}^{(j)}$,

$$\int_{c_j}^{d_j} p(t)\mathrm{d}\lambda_j(t) = \beta_0^{(j)} \boldsymbol{e}_1^{\mathrm{T}} p(\boldsymbol{J}^{(j)})\boldsymbol{e}_1, \quad \beta_0^{(j)} = \int_{c_j}^{d_j} \mathrm{d}\lambda_j(t). \tag{4.49}$$

Then, for the inner products $(t\pi_k, \pi_k)_{\mathrm{d}\lambda}$, $k \le n-1$, we get

$$\begin{aligned}(t\pi_k, \pi_k)_{\mathrm{d}\lambda} &= \int_{\mathbb{R}} t\pi_k^2(t)\mathrm{d}\lambda(t) = \sum_j \int_{c_j}^{d_j} t\pi_k^2(t)\mathrm{d}\lambda_j(t) \\ &= \sum_j \beta_0^{(j)} \boldsymbol{e}_1^{\mathrm{T}} \boldsymbol{J}^{(j)} [\pi_k(\boldsymbol{J}^{(j)})]^2 \boldsymbol{e}_1 \\ &= \sum_j \beta_0^{(j)} \boldsymbol{e}_1^{\mathrm{T}} [\pi_k(\boldsymbol{J}^{(j)})]^{\mathrm{T}} \boldsymbol{J}^{(j)} \pi_k(\boldsymbol{J}^{(j)})\boldsymbol{e}_1\end{aligned}$$

and for $(\pi_k, \pi_k)_{\mathrm{d}\lambda}$ similarly (in fact, simpler)

$$(\pi_k, \pi_k)_{\mathrm{d}\lambda} = \sum_j \beta_0^{(j)} \boldsymbol{e}_1^{\mathrm{T}} [\pi_k(\boldsymbol{J}^{(j)})]^{\mathrm{T}} \pi_k(\boldsymbol{J}^{(j)})\boldsymbol{e}_1.$$

This can be conveniently expressed in terms of the vectors

$$\boldsymbol{\zeta}_k^{(j)} := \pi_k(\boldsymbol{J}^{(j)})\boldsymbol{e}_1, \quad \boldsymbol{e}_1 = [1, 0, \ldots, 0]^T,$$

which, as required in Stieltjes's procedure, can be updated by means of the basic three-term recurrence relation. This leads to the following algorithm.

Algorithm 7 Stieltjes procedure for polynomials orthogonal on several intervals

initialization:

$$\begin{aligned}&\boldsymbol{\zeta}_0^{(j)} = \boldsymbol{e}_1, \quad \boldsymbol{\zeta}_{-1}^{(j)} = \boldsymbol{0} \quad (\text{all } j), \\ &\alpha_0 = \frac{\sum_j \beta_0^{(j)} \boldsymbol{e}_1^{\mathrm{T}} \boldsymbol{J}^{(j)} \boldsymbol{e}_1}{\sum_j \beta_0^{(j)}}, \quad \beta_0 = \sum_j \beta_0^{(j)}.\end{aligned}$$

continuation (if $n > 1$): for $k = 0, 1, \ldots, n-2$ do

$$\begin{aligned}&\boldsymbol{\zeta}_{k+1}^{(j)} = (\boldsymbol{J}^{(j)} - \alpha_k \boldsymbol{I})\boldsymbol{\zeta}_k^{(j)} - \beta_k \boldsymbol{\zeta}_{k-1}^{(j)} \quad (\text{all } j), \\ &\alpha_{k+1} = \frac{\sum_j \beta_0^{(j)} \boldsymbol{\zeta}_{k+1}^{(j)\mathrm{T}} \boldsymbol{J}^{(j)} \boldsymbol{\zeta}_{k+1}^{(j)}}{\sum_j \beta_0^{(j)} \boldsymbol{\zeta}_{k+1}^{(j)\mathrm{T}} \boldsymbol{\zeta}_{k+1}^{(j)}}, \quad \beta_{k+1} = \frac{\sum_j \beta_0^{(j)} \boldsymbol{\zeta}_{k+1}^{(j)\mathrm{T}} \boldsymbol{\zeta}_{k+1}^{(j)}}{\sum_j \beta_0^{(j)} \boldsymbol{\zeta}_k^{(j)\mathrm{T}} \boldsymbol{\zeta}_k^{(j)}}.\end{aligned}$$

In Matlab, this is implemented in the `OPQ` routine

```
ab=r_multidomain_sti(N,abmd)
```

where abmd is the array containing the (α, β)-coefficients of the measures $d\lambda^{(j)}$.

Example 9. Example 1, revisited.

This is the case of two identical intervals $[-1, 1]$ and two measures $d\lambda^{(j)}$ on $[-1, 1]$, one a multiple c of the Legendre measure, the other the Chebyshev measure. This was solved in Example 1 by a 2-component discretization method. The solution by the 2-domain algorithm of this subsection, in Matlab, looks as follows:

```
ab1=r_jacobi(N); ab1(1,2)=2*c;
ab2=r_jacobi(N,-.5);
abmd=[ab1 ab2];
ab=r_multidomain_sti(N,abmd)
```

It produces results identical with those produced by the method of Example 1.

Solution by the Modified Chebyshev Algorithm

The quadrature procedure used in the previous subsection to compute inner products can equally be applied to compute the first $2n$ modified moments of $d\lambda$,

$$m_k = \sum_j \int_{c_j}^{d_j} p_k(t) d\lambda_j(t) = \sum_j \beta_0^{(j)} e_1^{\mathrm{T}} p_k(\boldsymbol{J}^{(j)}) e_1. \tag{4.50}$$

The relevant vectors are now

$$z_k^{(j)} := p_k(\boldsymbol{J}^{(j)}) e_1, \quad e_1 = [1, 0, \ldots, 0]^{\mathrm{T}},$$

and the computation proceeds as in

Algorithm 8 Modified moments for polynomials orthogonal on several intervals

initialization

$$z_0^{(j)} = e_1, \; z_{-1}^{(j)} = \mathbf{0} \text{ (all } j), \quad m_0 = \sum_j \beta_0^{(j)}.$$

continuation: for $k = 0, 1, \ldots, 2n-2$ do

$$z_{k+1}^{(j)} = (\boldsymbol{J}^{(j)} - a_k \boldsymbol{I}) z_k^{(j)} - b_k z_{k-1}^{(j)} \quad \text{(all } j),$$
$$m_{k+1} = \sum_j \beta_0^{(j)} z_{k+1}^{(j)\mathrm{T}} e_1.$$

With these moments at hand, we can apply Algorithm 1 to obtain the desired recurrence coefficients. This is done in the OPQ routine

```
ab=r_multidomain_cheb(N,abmd,abmm)
```

The array `abmd` has the same meaning as in the routine `r_multidomain_sti.m`, and `abmm` is a $((2N-1)\times 2)$ array of the recurrence coefficients a_k, b_k generating the polynomials p_k.

Applied to Example 9, the Matlab program, using Legendre moments (p_k the monic Legendre polynomials), is as follows:

```
abm=r_jacobi(2*N-1);
ab1=abm(1:N,:); ab1(1,2)=2*c;
ab2=r_jacobi(N,-.5);
abmd=[ab1 ab2];
ab=r_multidomain_cheb(N,abmd,abm)
```

It produces results identical with those obtained previously, but takes about three times as long to run.

4.7 Quadrature Estimates of Matrix Functionals

The problem to be considered here is to find lower and upper bounds for the quadratic form

$$\boldsymbol{u}^{\mathrm{T}} f(\boldsymbol{A})\boldsymbol{u}, \quad \boldsymbol{u}\in\mathbb{R}^N, \ \|\boldsymbol{u}\|=1, \tag{4.51}$$

where $\boldsymbol{A}\in\mathbb{R}^{N\times N}$ is a symmetric, positive definite matrix, f a smooth function (for which $f(\boldsymbol{A})$ makes sense), and $\boldsymbol{u}$ a given vector. While this looks more like a linear algebra problem, it can actually be solved, for functions f with derivatives of constant sign, by applying Gauss-type quadrature rules. The connecting link is provided by the *spectral resolution* of $\boldsymbol{A}$,

$$\boldsymbol{A}\boldsymbol{V}=\boldsymbol{V}\boldsymbol{\Lambda}, \ \boldsymbol{\Lambda}=\operatorname{diag}(\lambda_1,\lambda_2,\ldots,\lambda_N), \ \boldsymbol{V}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_N], \tag{4.52}$$

where λ_k are the eigenvalues of $\boldsymbol{A}$ (which for simplicity are assumed distinct), and $\boldsymbol{v}_k$ the normalized eigenvectors of $\boldsymbol{A}$. If we put

$$\boldsymbol{u}=\sum_{k=1}^{N}\rho_k\boldsymbol{v}_k=\boldsymbol{V}\boldsymbol{\rho}, \ \boldsymbol{\rho}=[\rho_1,\rho_2,\ldots,\rho_N]^{\mathrm{T}}, \tag{4.53}$$

and again for simplicity assume $\rho_k\neq 0$, all k, then

$$\begin{aligned}\boldsymbol{u}^{\mathrm{T}} f(\boldsymbol{A})\boldsymbol{u}&=\boldsymbol{\rho}^{\mathrm{T}}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}f(\boldsymbol{\Lambda})\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\rho}=\boldsymbol{\rho}^{\mathrm{T}}f(\boldsymbol{\Lambda})\boldsymbol{\rho},\\&=\sum_{k=1}^{N}\rho_k^2 f(\lambda_k)=:\int_{\mathbb{R}_+}f(t)\mathrm{d}\rho_N(t).\end{aligned} \tag{4.54}$$

This shows how the matrix functional is related to an integral relative to a discrete positive measure. Now we know from (4.10) and (4.15) how Gauss–Radau or Gauss–Lobatto rules (and for that matter also ordinary Gauss rules,

in view of $R_n^G = [f^{(2n)}(\tau)/(2n)!] \int_a^b [\pi_n(t; \mathrm{d}\lambda)]^2 \mathrm{d}\lambda(t)$, $a < \tau < b$) can be applied to obtain two-sided bounds for (4.54) when some derivative of f has constant sign. To generate these quadrature rules, we need the orthogonal polynomials for the measure $\mathrm{d}\rho_N$, and for these the Jacobi matrix $\boldsymbol{J}_N(\mathrm{d}\rho_N)$. The latter, in principle, could be computed by the Lanczos-type algorithm of §2.3. However, in the present application this would require knowledge of the eigenvalues λ_k and expansion coefficients ρ_k, which are too expensive to compute. Fortunately, there is an alternative way to implement Lanczos's algorithm that works directly with the matrix $\boldsymbol{A}$ and requires only multiplications of $\boldsymbol{A}$ into vectors and the computation of inner products.

Lanczos Algorithm

Let ρ_k be as in (4.54) and $h_0 = \sum_{k=1}^N \rho_k \boldsymbol{v}_k$ $(= \boldsymbol{u})$, $\|h_0\| = 1$, as in (4.53).

Algorithm 9 Lanczos algorithm

initialization:

$$\boldsymbol{h}_0 \text{ prescribed with } \|\boldsymbol{h}_0\| = 1, \quad \boldsymbol{h}_{-1} = \boldsymbol{0}.$$

continuation: for $j = 0, 1, \ldots, N-1$ do

$$\begin{aligned}
\alpha_j &= \boldsymbol{h}_j^{\mathrm{T}} \boldsymbol{A} \boldsymbol{h}_j, \\
\tilde{\boldsymbol{h}}_{j+1} &= (\boldsymbol{A} - \alpha_j \boldsymbol{I}) \boldsymbol{h}_j - \gamma_j \boldsymbol{h}_{j-1}, \\
\gamma_{j+1} &= \|\tilde{\boldsymbol{h}}_{j+1}\|, \\
\boldsymbol{h}_{j+1} &= \tilde{\boldsymbol{h}}_{j+1} / \gamma_{j+1}.
\end{aligned}$$

While γ_0 in Algorithm 9 can be arbitrary (it multiplies $\boldsymbol{h}_{-1} = \boldsymbol{0}$), it is convenient to define $\gamma_0 = 1$. The vectors $\boldsymbol{h}_0, \boldsymbol{h}_1, \ldots, \boldsymbol{h}_N$ generated by Algorithm 9 are called *Lanczos vectors*. It can be shown that α_k generated by the Lanczos algorithm is precisely $\alpha_k(\mathrm{d}\rho_N)$, and $\gamma_k = \sqrt{\beta_k(\mathrm{d}\rho_N)}$, for $k = 0, 1, 2, \ldots, N-1$. This provides us with the Jacobi matrix $\boldsymbol{J}_N(\mathrm{d}\rho_N)$. It is true that the algorithm becomes unstable as j approaches N, but in the applications of interest here, only small values of j are needed.

Examples

Example 10. Error bounds for linear algebraic systems.

Consider the system

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \quad \boldsymbol{A} \text{ symmetric, positive definite.} \tag{4.55}$$

Given an approximation $\boldsymbol{x}^* \approx \boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$ to the exact solution $\boldsymbol{x}$, and the residual vector $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^*$, we have $\boldsymbol{x} - \boldsymbol{x}^* = \boldsymbol{A}^{-1}\boldsymbol{b} + \boldsymbol{A}^{-1}(\boldsymbol{r} - \boldsymbol{b}) = \boldsymbol{A}^{-1}\boldsymbol{r}$, thus

$$\|\boldsymbol{x}-\boldsymbol{x}^*\|^2=(\boldsymbol{A}^{-1}\boldsymbol{r})^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{r}=\boldsymbol{r}^{\mathrm{T}}\boldsymbol{A}^{-2}\boldsymbol{r},$$

and therefore

$$\|\boldsymbol{x}-\boldsymbol{x}^*\|^2=\|\boldsymbol{r}\|^2\cdot\boldsymbol{u}^{\mathrm{T}}f(\boldsymbol{A})\boldsymbol{u}, \tag{4.56}$$

where $\boldsymbol{u}=\boldsymbol{r}/\|\boldsymbol{r}\|$ and $f(t)=t^{-2}$. All derivatives of f are here of constant sign on $\mathbb{R}_+$,

$$f^{(2n)}(t)>0,\quad f^{(2n+1)}(t)<0\quad\text{for } t\in\mathbb{R}_+. \tag{4.57}$$

By (4.54), we now have

$$\|\boldsymbol{x}-\boldsymbol{x}^*\|=\|\boldsymbol{r}\|^2\int_{\mathbb{R}_+}t^{-2}\mathrm{d}\rho_N(t). \tag{4.58}$$

The n-point Gauss quadrature rule applied to the integral on the right of (4.58), by the first inequality in (4.57), yields a *lower bound* of $\|\boldsymbol{x}-\boldsymbol{x}^*\|$, without having to know the exact support interval of $\mathrm{d}\rho_N$. If, on the other hand, we know that the support of $\mathrm{d}\rho_N$ is contained in some interval $[a,b]$, $0<a<b$, we can get a lower bound also from the right-handed $(n+1)$-point Gauss–Radau formula, and *upper bounds* from the left-handed $(n+1)$-point Gauss–Radau formula on $[a,b]$, or from the $(n+2)$-point Gauss–Lobatto formula on $[a,b]$.

Example 11. Diagonal elements of $\boldsymbol{A}^{-1}$.

Here, trivially

$$\boldsymbol{u}^{\mathrm{T}}f(\boldsymbol{A})\boldsymbol{u}=\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{e}_i, \tag{4.59}$$

where $f(t)=t^{-1}$ and $\boldsymbol{e}_i$ is the ith coordinate vector. Using n-point Gauss quadrature in (4.54), with $n<N$, yields

$$(\boldsymbol{A}^{-1})_{ii}=\int_{\mathbb{R}_+}t^{-1}\,\mathrm{d}\rho_N(t)>\boldsymbol{e}_1^{\mathrm{T}}\boldsymbol{J}_n^{-1}\boldsymbol{e}_1,\quad \boldsymbol{e}_1^{\mathrm{T}}=[1,0,\ldots,0]\in\mathbb{R}^n. \tag{4.60}$$

Suppose we take $n=2$ steps of Algorithm 9 to compute

$$\boldsymbol{J}_2=\begin{bmatrix}\alpha_0 & \gamma_1\\ \gamma_1 & \alpha_1\end{bmatrix}.$$

We get

$$\begin{aligned}
&\alpha_0=a_{ii},\\
&\tilde{\boldsymbol{h}}_1=(\boldsymbol{A}-\alpha_0\boldsymbol{I})\boldsymbol{e}_i=[a_{1i},\ldots,a_{i-1,i},0,a_{i+1,i},\ldots,a_{Ni}]^{\mathrm{T}},\\
&\gamma_1=\sqrt{\sum_{k\neq i}a_{ki}^2}=:s_i,\\
&\boldsymbol{h}_1=\tilde{\boldsymbol{h}}_1/s_i,\\
&\alpha_1=\frac{1}{s_i^2}\tilde{\boldsymbol{h}}_1^{\mathrm{T}}\boldsymbol{A}\tilde{\boldsymbol{h}}_1=\frac{1}{s_i^2}\sum_{k\neq i}\sum_{\ell\neq i}a_{k\ell}a_{ki}a_{\ell i}.
\end{aligned} \tag{4.61}$$

But

$$\boldsymbol{J}_2^{-1} = \frac{1}{\alpha_0\alpha_1 - \gamma_1^2}\begin{bmatrix} \alpha_1 & -\gamma_1 \\ -\gamma_1 & \alpha_0 \end{bmatrix}, \quad \boldsymbol{e}_1^{\mathrm{T}}\boldsymbol{J}_2^{-1}\boldsymbol{e}_1 = \frac{\alpha_1}{\alpha_0\alpha_1 - \gamma_1^2},$$

so that by (4.60) with $n = 2$, and (4.61),

$$(\boldsymbol{A}^{-1})_{ii} > \frac{\sum_{k\neq i}\sum_{\ell\neq i} a_{k\ell}a_{ki}a_{\ell i}}{a_{ii}\sum_{k\neq i}\sum_{\ell\neq i} a_{k\ell}a_{ki}a_{\ell i} - \left(\sum_{k\neq i} a_{ki}^2\right)^2}. \tag{4.62}$$

Simpler bounds, both lower and upper, can be obtained by the 2-point Gauss-Radau and Gauss-Lobatto formulae, which however require knowledge of an interval $[a, b]$, $0 < a < b$, containing the spectrum of $\boldsymbol{A}$.

Exercises to §4 (Stars indicate more advanced exercises.)

1. Prove (4.5).
2. Prove that complex zeros of the Sobolev orthogonal polynomials of Example 7 must be purely imaginary.

3*. Circle theorems for quadrature weights (cf. [4]).

(a) Gauss–Jacobi quadrature

Let $w(t) = (1-t)^\alpha(1+t)^\beta$ be the Jacobi weight function. It is known [16, eqn (15.3.10)] that the nodes τ_ν and weights λ_ν of the n-point Gauss–Jacobi quadrature formula satisfy

$$\lambda_\nu \sim \frac{\pi}{n} w(\tau_\nu)\sqrt{1-\tau_\nu^2}, \quad n \to \infty,$$

for τ_ν on any compact interval contained in $(-1, 1)$. Thus, suitably normalized weights, plotted against the nodes, lie asymptotically on the unit circle. Use Matlab to demonstrate this graphically.

(b) Gauss quadrature for the logarithmic weight $w(t) = t^\alpha \ln(1/t)$ on $[0, 1]$ (cf. [10, Example 2.27]).

Try, numerically, to find a circle theorem in this case also, and experiment with different values of the parameter $\alpha > -1$. (Use the `OPQ` routine `r_jaclog.m` to generate the recurrence coefficients of the orthogonal polynomials for the weight function w.)

(c) Gauss–Kronrod quadrature.

With w as in (a), the analogous result for the $2n + 1$ nodes τ_ν and weights λ_ν of the $(2n + 1)$-point Gauss–Kronrod formula is expected to be

$$\lambda_\nu \sim \frac{\pi}{2n} w(\tau_\nu)\sqrt{1-\tau_\nu^2}, \quad n \to \infty.$$

That this indeed is the case, when $\alpha, \beta \in [0, \frac{5}{2})$, follows from Theorem 2 in [15]. Use Matlab to illustrate this graphically.

(d) Experiment with the Gauss–Kronrod formula for the logarithmic weight function of (b), when $\alpha = 0$.

4. Discrete orthogonality.
Let $\pi_k(\cdot;\mathrm{d}\lambda)$, $k = 0, 1, 2, \ldots$, be the orthogonal polynomials relative to an absolutely continuous measure. Show that for each $N \geq 2$, the first N of them are orthogonal with respect to the discrete inner product

$$(p, q)_N = \sum_{\nu=1}^{N} \lambda_\nu^G p(\tau_\nu^G) q(\tau_\nu^G),$$

where τ_ν^G, λ_ν^G are the nodes and weights of the N-point Gauss formula for $\mathrm{d}\lambda$. Moreover, $\|\pi_k\|_N^2 = \|\pi_k\|_{\mathrm{d}\lambda}^2$ for $k \leq N-1$.

5. (a) Consider the Cauchy integral

$$\rho_n(z) = \rho_n(z;\mathrm{d}\lambda) = \int_a^b \frac{\pi_n(t;\mathrm{d}\lambda)}{z-t}\,\mathrm{d}\lambda(t),$$

where $[a, b]$ is the support of $\mathrm{d}\lambda$. Show that

$$\rho_n(z) = O(z^{-n-1}) \quad \text{as } z \to \infty.$$

{*Hint*: Expand the integral defining $\rho_n(z)$ in descending powers of z.}
(b) Show that

$$\int_a^b \frac{\mathrm{d}\lambda(t)}{z-t} - \frac{\sigma_n(z)}{\pi_n(z)} = \frac{\rho_n(z)}{\pi_n(z)} = O(z^{-2n-1}) \quad \text{as } z \to \infty.$$

{*Hint*: Use (2.27).}
(c) Consider the partial fraction decomposition

$$\frac{\sigma_n(z)}{\pi_n(z)} = \sum_{\nu=1}^{n} \frac{\lambda_\nu}{z-\tau_\nu^G}$$

of $\sigma_n(z)/\pi_n(z)$ in (2.27). Use (b) to show that $\lambda_\nu = \lambda_\nu^G$ are the weights of the n-point Gaussian quadrature formula for $\mathrm{d}\lambda$. In particular, show that

$$\lambda_\nu^G = \frac{\sigma_n(\tau_\nu^G)}{\pi_n'(\tau_\nu^G)}.$$

(d) Discuss what happens if $z \to x$, $x \in (a, b)$.

6. Characterize the nodes τ_ν^b in (4.9) as zeros of an orthogonal polynomial of degree n, and identify the appropriate Jacobi–Radau matrix for (4.9).

7. Prove (4.10). {*Hint*: Use the fact that both formulae (4.6) and (4.9) are interpolatory.}

8. (a) Prove the first formula in (4.11). {*Hint*: Use the relation between the Jacobi polynomials $P_k = P_k^{(\alpha,\beta)}$ customarily defined and the monic Jacobi polynomials $\pi_k = \pi_k^{(\alpha,\beta)}$, expressed by $P_k(t) = 2^{-k}\binom{2k+\alpha+\beta}{k}\pi_k(t)$. You also need $P_k(-1) = (-1)^k\binom{k+\beta}{k}$ and the β-coefficient for Jacobi polynomials, $\beta_n^J = 4n(n+\alpha)(n+\beta)(n+\alpha+\beta)/(2n+\alpha+\beta)^2(2n+\alpha+\beta+1)(2n+\alpha+\beta-1)$.}

(b) Prove the second formula in (4.11). {*Hint*: With $\pi_k^{(\alpha)}$ and $L_k^{(\alpha)}$ denoting the monic resp. conventional generalized Laguerre polynomials, use $L_k^{(\alpha)}(t) = \big((-1)^k/k!\big)\,\pi_k^{(\alpha)}(t)$. You also need $L_k^{(\alpha)}(0) = \binom{k+\alpha}{k}$, and $\beta_n^L = n(n+\alpha)$.}

9. Prove (4.16). {*Hint*: With notation as in the hint to Exercise 8(a), use $P_k(1) = \binom{k+\alpha}{k}$ in addition to the information provided there.}
10. The (left-handed) *generalized Gauss–Radau formula* is

$$\int_a^\infty f(t)\,\mathrm{d}\lambda(t) = \sum_{\rho=0}^{r-1} \lambda_0^{(\rho)} f^{(\rho)}(a) + \sum_{\nu=1}^{n} \lambda_\nu^R f(\tau_\nu^R) + R_{n,r}^R(f),$$

where $r > 1$ is the multiplicity of the end point $\tau_0 = a$, and $R_{n,r}^R(f) = 0$ for $f \in \mathbb{P}_{2n-1+r}$. Let $\mathrm{d}\lambda^{[r]}(t) = (t-a)^r \mathrm{d}\lambda(t)$ and $\tau_\nu^{[r]}$, $\lambda_\nu^{[r]}$, $\nu = 1, 2, \ldots, n$, be the nodes and weights of the n-point Gauss formula for $\mathrm{d}\lambda^{[r]}$.

(a) Show that

$$\tau_\nu^R = \tau_\nu^{[r]}, \quad \lambda_\nu^R = \frac{\lambda_\nu^{[r]}}{(\tau_\nu^R - a)^r}, \quad \nu = 1, 2, \ldots, n.$$

(b) Show that not only the internal weights λ_ν^R are all positive (why?), but also the boundary weights λ_0, λ_0' if $r = 2$.

11. The *generalized Gauss–Lobatto formula* is

$$\int_a^b f(t)\,\mathrm{d}\lambda(t) = \sum_{\rho=0}^{r-1} \lambda_0^{(\rho)} f^{(\rho)}(a) + \sum_{\nu=1}^{n} \lambda_\nu^L f(\tau_\nu^L) + \sum_{\rho=0}^{r-1} (-1)^\rho \lambda_{n+1}^{(\rho)} f^{(\rho)}(b) + R_{n,r}^L(f),$$

where $r > 1$ is the multiplicity of the end points $\tau_0 = a$, $\tau_{n+1} = b$, and $R_{n,r}^L(f) = 0$ for $f \in \mathbb{P}_{2n-1+2r}$. Let $\mathrm{d}\lambda^{[r]}(t) = [(t-a)(b-t)]^r \mathrm{d}\lambda(t)$ and $\tau_\nu^{[r]}$, $\lambda_\nu^{[r]}$, $\nu = 1, 2, \ldots, n$, be the nodes and weights of the n-point Gauss formula for $\mathrm{d}\lambda^{[r]}$.

(a) Show that

$$\tau_\nu^L = \tau_\nu^{[r]}, \quad \lambda_\nu^L = \frac{\lambda_\nu^{[r]}}{[(\tau_\nu^L - a)(b - \tau_\nu^L)]^r}, \quad \nu = 1, 2, \ldots, n.$$

(b) Show that not only the internal weights λ_ν^L are all positive (why?), but also the boundary weights λ_0, λ_0' and λ_{n+1}, λ_{n+1}' if $r = 2$.

(c) Show that $\lambda_0^{(\rho)} = \lambda_{n+1}^{(\rho)}$, $\rho = 0, 1, \ldots, r-1$, if the measure $\mathrm{d}\lambda$ is symmetric.

12*. Generalized Gauss-Radau quadrature.
   (a) Write a Matlab routine `gradau.m` for generating the generalized Gauss-Radau quadrature rule of Exercise 10 for a measure $d\lambda$ on $[a,\infty]$, having a fixed node a of multiplicity r, $r > 1$. {*Hint*: To compute the boundary weights, set up an (upper triangular) system of linear equations by applying the formula in turn with $\pi_n^2(t)$, $(t-a)\pi_n^2(t), \ldots, (t-a)^{r-1}\pi_n^2(t)$, where $\pi_n(t) = \prod_{\nu=1}^n (t-\tau_\nu^R)$.}
   (b) Check your routine against the known formulae with $r = 2$ for the Legendre and Chebyshev measures (see [10, Examples 3.10 and 3.11]). Devise and implement a check that works for arbitrary $r \geq 1$ and another, in particular, for $r = 1$.
   (c) Use your routine to explore positivity of the boundary weights and see whether you can come up with any conjectures.

13*. Generalized Gauss-Lobatto quadrature.
   (a) Write a Matlab routine `globatto.m` for generating the generalized Gauss-Lobatto rule of Exercise 11 for a measure $d\lambda$ on $[a,b]$, having fixed nodes at a and b of multiplicity r, $r > 1$. For simplicity, start with the case $r \geq 2$ even; then indicate the changes necessary to deal with odd values of r. {*Hint*: Similar to the hint in Exercise 12(a).}
   (b) Check your routine against the known formulae with $r = 2$ for the Legendre and Chebyshev measures (see [10, Examples 3.13 and 3.14]). Devise and implement a check that works for arbitrary $r \geq 1$ and another, in particular, for $r = 1$.
   (c) Explore the positivity of the boundary weights $\lambda_0^{(\rho)}$ and the quantities $\lambda_{n+1}^{(\rho)}$ in the quadrature formula.

14. Show that the monic Stieltjes polynomial π_{n+1}^K in (4.19) exists uniquely.

15. (a) Let $d\lambda$ be a positive measure. Use approximation theory to show that the minimum of $\int_{\mathbb{R}} |\pi(t)|^p d\lambda(t)$, $1 < p < \infty$, extended over all monic polynomials π of degree n is uniquely determined.
   (b) Show that the minimizer of the extremal problem in (a), when $p = 2s+2$, $s \geq 0$ an integer, is the s-orthogonal polynomial $\pi = \pi_{n,s}$. {*Hint*: Differentiate the integral partially with respect to the variable coefficients of π.}

16. (a) Show that r in (4.22) has to be odd.
   (b) Show that in (4.22) with r as in (4.23), one cannot have $k > n$.

17. Derive (4.33) and (4.34).

18. Derive (4.39) from (4.38) and explain the meaning of $R_n(f;x)$. {*Hint*: Use Exercise 5(c) and (2.27).}

19. Show that $p_n(f;t)$ in (4.40) is

$$p_n(f;t) = \frac{\pi_n(t)}{\pi_n(x)} f(x) + \sum_{\nu=1}^n \frac{(t-x)\pi_n(t)}{(t-\tau_\nu^G)(\tau_\nu^G - x)\pi_n'(\tau_\nu^G)} f(\tau_\nu^G),$$

and thus prove (4.40). {*Hint*: Use Exercise 5(c).}

20. Derive (4.42) and (4.43). {*Hint*: For $k < n$, use Gauss quadrature, and for $k = n$ insert the expression for $p_n(f;t)$ from Exercise 19 into the formula for a_n in (4.41). Also use the fact that the elementary Lagrange interpolation polynomials sum up to 1.}
21. Derive (4.44).
22. Prove (4.46). {*Hint*: Use Exercise 4.}
23. (a) Prove that the Lanczos vectors are mutually orthonormal.
    (b) Show that the vectors $\{\boldsymbol{h}_j\}_{j=0}^{n}$, $n < N$, form an orthonormal basis of the *Krylov space*

$$\mathcal{K}_n(\boldsymbol{A}, \boldsymbol{h}_0) = \operatorname{span}(\boldsymbol{h}_0, \boldsymbol{A}\boldsymbol{h}_0, \ldots, \boldsymbol{A}^n\boldsymbol{h}_0).$$

    (c) Prove that

$$\boldsymbol{h}_j = p_j(\boldsymbol{A})\boldsymbol{h}_0, \quad j = 0, 1, \ldots, N,$$

    where p_j is a polynomial of degree j satisfying the three-term recurrence relation

$$\gamma_{j+1}p_{j+1}(\lambda) = (\lambda - \alpha_j)p_j(\lambda) - \gamma_j p_{j-1}(\lambda), \qquad j = 0, 1, \ldots, N-1,$$
$$p_0(\lambda) = 1, \quad p_{-1}(\lambda) = 0.$$

    {*Hint*: Use mathematical induction.}
24. Prove that the polynomial p_k of Exercise 23(c) is equal to the orthonormal polynomial $\tilde{\pi}_k(\cdot\,; \mathrm{d}\rho_N)$. {*Hint*: Use the spectral resolution of $\boldsymbol{A}$ and Exercises 23(a) and (c).}
25. Derive the bounds for $(\boldsymbol{A}^{-1})_{ii}$ hinted at in the last sentence of Example 11.

5 Approximation

5.1 Polynomial Least Squares Approximation

Classical Least Squares Problem

We are given N data points (t_k, f_k), $k = 1, 2, \ldots, N$, and wish to find a polynomial $\hat{p}_n$ of degree $\leq n$, $n < N$, such that a weighted average of the squared errors $[p(t_k) - f_k]^2$ is as small as possible among all polynomials p of degree n,

$$\sum_{k=1}^{N} w_k[\hat{p}_n(t_k) - f_k]^2 \leq \sum_{k=1}^{N} w_k[p(t_k) - f_k]^2 \quad \text{for all } p \in \mathbb{P}_n. \tag{5.1}$$

Here, $w_k > 0$ are positive weights, which allow placing more emphasis on data points that are reliable, and less emphasis on others, by choosing them larger

resp. smaller. If the quality of the data is uniformly the same, then equal weights, say $w_k = 1$, are appropriate.

The problem as formulated suggests a discrete N-point measure

$$d\lambda_N(t) = \sum_{k=1}^{N} w_k \delta(t - t_k), \quad \delta = \text{Dirac delta function}, \tag{5.2}$$

in terms of which the problem can be written in the compact form

$$\|\hat{p}_n - f\|^2_{d\lambda_N} \leq \|p - f\|^2_{d\lambda_N} \quad \text{for all } p \in \mathbb{P}_n. \tag{5.3}$$

The polynomials $\pi_k(\,\cdot\,) = \pi_k(\,\cdot\,; d\lambda_N)$ orthogonal (not necessarily monic) with respect to the discrete measure (5.2) provide an easy solution: one writes

$$p(t) = \sum_{i=0}^{n} c_i \pi_i(t), \quad n < N, \tag{5.4}$$

and obtains for the squared error, using the orthogonality of π_k,

$$\begin{aligned} E_n^2 &= \left(\sum_{i=0}^{n} c_i \pi_i - f, \sum_{j=0}^{n} c_j \pi_j - f \right) = \sum_{i,j=0}^{n} c_i c_j (\pi_i, \pi_j) - 2 \sum_{i=0}^{n} c_i (f, \pi_i) + \|f\|^2 \\ &= \sum_{i=0}^{n} \left(\|\pi_i\| c_i - \frac{(f, \pi_i)}{\|\pi_i\|} \right)^2 + \|f\|^2 - \sum_{i=0}^{n} \frac{(f, \pi_i)^2}{\|\pi_i\|^2}. \end{aligned} \tag{5.5}$$

(All norms and inner products are understood to be relative to the measure $d\lambda_N$.) Evidently, the minimum is attained for $c_i = \hat{c}_i(f)$, where

$$\hat{c}_i(f) = \frac{(f, \pi_i)}{\|\pi_i\|^2}, \quad i = 0, 1, \ldots, n, \tag{5.6}$$

are the "Fourier coefficients" of f relative to the orthogonal system $\pi_0, \pi_1, \ldots, \pi_{N-1}$. Thus,

$$\hat{p}_n(t) = \sum_{i=0}^{n} \hat{c}_i(f) \pi_i(t; d\lambda_N). \tag{5.7}$$

In Matlab, the procedure is implemented in the `OPQ` routine

```
[phat,c]=least_squares(n,f,xw,ab,d)
```

The given function values f_k are input through the `N`×1 array `f`, the abscissae t_k and weights w_k through the `N`×2 array `xw`, and the measure $d\lambda_N$ through the (`N`+1)×2 array `ab` of recurrence coefficients (the routine determines `N` automatically from the size of `xw`). The 1×(`n`+1) array `d` is the vector of leading coefficients of the orthogonal polynomials. The procedure returns as output the `N`×(`n`+1) array `phat` of the values $\hat{p}_\nu(t_k)$, $0 \leq \nu \leq n$, $1 \leq k \leq N$, and the (`n`+1)×1 array `c` of the Fourier coefficients.

Example 12. Equally weighted least squares approximation on $N = 10$ equally spaced points on $[-1, 1]$.

Matlab program:

```
N=10; k=(1:N)'; d=ones(1,N);
xw(k,1)=-1+2*(k-1)/(N-1); xw(:,2)=2/N;
ab=r_hahn(N-1); ab(:,1)=-1+2*ab(:,1)/(N-1);
ab(:,2)=(2/(N-1))^2*ab(:,2); ab(1,2)=2;
[phat,c]=least_squares(N-1,f,xw,ab,d);
```

Demo#5 The program is applied to the function $f(t) = \ln(2+t)$ on $[-1, 1]$, and selected least squares errors $\hat{E}_n$ are compared in the table below with maximum errors E_n^∞ (taken over 100 equally spaced points on $[-1, 1]$).

n	$\hat{E}_n$	E_n^∞
0	4.88(–01)	6.37(–01)
3	2.96(–03)	3.49(–03)
6	2.07(–05)	7.06(–05)
9	1.74(–16)	3.44(–06)

If $n = N - 1$, the least squares error $\hat{E}_{N-1}$ is zero, since the N data points can be interpolated exactly by a polynomial of degree $\leq N - 1$. This is confirmed in the first tabular entry for $n = 9$. The infinity errors are only slightly larger than the least squares errors, except for $n = 9$.

Constrained Least Squares Approximation

It is sometimes desirable to impose constraints on the least squares approximation, for example to insist that at certain points s_j the error should be exactly zero. Thus, the polynomial $p \in \mathbb{P}_n$ is subject to the constraints

$$p(s_j) = f_j, \quad j = 1, 2, \ldots, m; \quad m \leq n, \tag{5.8}$$

but otherwise is freely variable. For simplicity we assume that none of the s_j equals one of the support points t_k. (Otherwise, the procedure to be described requires some simple modifications.)

In order to solve the constrained least squares problem, let

$$p_m(f; t) = p_m(f; s_1, \ldots, s_m; t), \quad \sigma_m(t) = \prod_{j=1}^{m} (t - s_j), \tag{5.9}$$

be respectively the polynomial of degree $m - 1$ interpolating f at the points s_j and the constraint polynomial of degree m. We then write

$$p(t) = p_m(f; t) + \sigma_m(t) q(t). \tag{5.10}$$

This clearly satisfies the constraints (5.8), and q is a polynomial of degree $n - m$ that can be freely varied. The problem is to minimize the squared error

$$\|f - p_m(f;\cdot) - \sigma_m q\|^2_{\mathrm{d}\lambda_N} = \int_{\mathbb{R}} \left[\frac{f(t) - p_m(f;t)}{\sigma_m(t)} - q(t)\right]^2 \sigma_m^2(t)\mathrm{d}\lambda_N(t)$$

over all polynomials q of degree $n-m$. This is an *unconstrained* least squares problem, but for a new function f^* and a new measure $\mathrm{d}\lambda_N^*$,

$$\text{minimize}: \quad \|f^* - q\|_{\mathrm{d}\lambda_N^*}, \quad q \in \mathbb{P}_{n-m}, \tag{5.11}$$

where

$$f^*(t) = \frac{f(t) - p_m(f;t)}{\sigma_m(t)}, \quad \mathrm{d}\lambda_N^*(t) = \sigma_m^2(t)\mathrm{d}\lambda_N(t). \tag{5.12}$$

If $\hat{q}_{n-m}$ is the solution of (5.11), then

$$\hat{p}_n(t) = p_m(f;t) + \sigma_m(t)\hat{q}_{n-m}(t) \tag{5.13}$$

is the solution of the constrained least squares problem. The function f^*, incidentally, can be given the form of a divided difference,

$$f^*(t) = [s_1, s_2, \ldots, s_m, t]f, \quad t \in \operatorname{supp} \mathrm{d}\lambda_N^*,$$

as follows from the theory of interpolation. Note also that the discrete orthogonal polynomials $\pi_k(\cdot;\mathrm{d}\lambda_N^*)$ needed to solve (5.11) can be obtained from the polynomials $\pi_k(\cdot;\mathrm{d}\lambda_N)$ by m modifications of the measure $\mathrm{d}\lambda_N(t)$ by quadratic factors $(t-s_j)^2$.

Example 13. Bessel function $J_0(t)$ for $0 \le t \le j_{0,3}$.

Here, $j_{0,3}$ is the third positive zero of J_0. A natural constraint is to reproduce the first three zeros of J_0 exactly, that is, $m = 3$ and

$$s_1 = j_{0,1}, \quad s_2 = j_{0,2}, \quad s_3 = j_{0.3}.$$

Demo#6 The constrained least squares approximations of degrees $n = 3, 4, 5$ (that is, $n - m = 0, 1, 2$) using $N = 51$ equally spaced points on $[0, j_{0,3}]$ (end points included) are shown in Fig. 2. The solid curve represents the exact function, the dashdotted, dashed, an dotted curves the approximants for $n = 3, 4$, and 5, respectively. The approximations are not particularly satisfactory and show spurious behavior near $t = 0$.

Example 14. Same as Example 13, but with two additional constraints

$$p(0) = 1, \quad p'(0) = 0.$$

Demo#7 Derivative constraints, as the one in Example 14, can be incorporated similarly as before. In this example, the added constraints are designed to remove the spurious behavior near $t = 0$; they also improve considerably the overall accuracy, as is shown in Fig. 3. For further details on Matlab implementation, see [10, Examples 3.51 and 3.52].

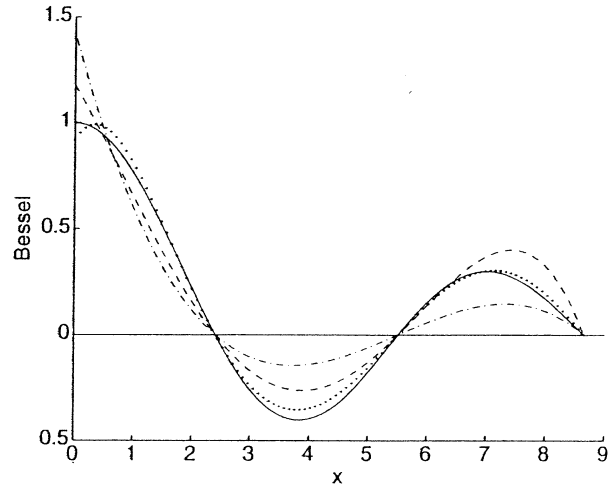


Fig. 2. Constrained least square approximation of the Bessel function

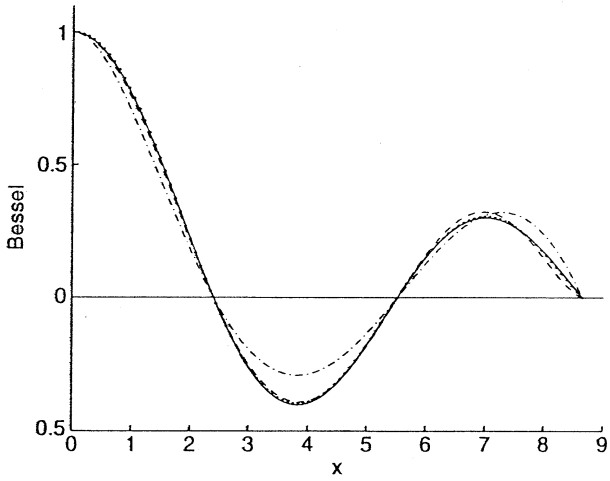


Fig. 3. Derivative-constrained least squares approximation of the Bessel function

Least Squares Approximation in Sobolev Spaces

The task now is to approximate simultaneously functions and some of their first derivatives. More precisely, we want to minimize

$$\sum_{\sigma=0}^{s}\sum_{k=1}^{N} w_k^{(\sigma)}[p^{(\sigma)}(t_k) - f_k^{(\sigma)}]^2$$

over all polynomials $p \in \mathbb{P}_n$, where $f_k^{(\sigma)}$, $\sigma = 0, 1, \ldots, s$, are given function and derivative values, and $w_k^{(\sigma)} > 0$ appropriate weights for each derivative.

These are often chosen such that

$$w_k^{(\sigma)} = \gamma_\sigma w_k, \quad \gamma_\sigma > 0, \quad k = 1, 2, \ldots, N,$$

in terms of one set of positive weights w_k. Evidently,, the problem, analogously to (5.3), can be written in terms of the Sobolev inner product and norm

$$(u, v)_S = \sum_{\sigma=0}^{s} \sum_{k=1}^{N} w_k^{(\sigma)} u^{(\sigma)}(t_k) v^{(\sigma)}(t_k), \quad \|u\|_S = \sqrt{(u, u)_S} \tag{5.14}$$

in the compact form

$$\text{minimize :} \quad \|p - f\|_S^2 \text{ for all } p \in \mathbb{P}_n. \tag{5.15}$$

The solution is entirely analogous to the one provided earlier,

$$\hat{p}_n(t) = \sum_{i=0}^{n} \hat{c}_i(f) \pi_i(t), \quad \hat{c}_i(f) = \frac{(f, \pi_i)_S}{\|\pi_i\|_S^2}, \tag{5.16}$$

where $\{\pi_i\}$ are the orthogonal polynomials of Sobolev type. In Matlab, the procedure is

```
[phat,c]=least_squares_sob(n,f,xw,B)
```

The input parameter `f` is now an $N \times (s+1)$ array containing the N values of the given function and its first s derivatives at the points t_k. The abscissae t_k and the $s+1$ weights $w_k^{(\sigma)}$ of the Sobolev inner product are input via the $N \times (s+2)$ array `xw` (the routine determines N and s automatically from the size of the array `xw`). The user also has to provide the $N \times N$ upper triangular array `B` of the recurrence coefficients for the Sobolev orthogonal polynomials, which for $s = 1$ can be generated by the routine `chebyshev_sob.m` and for arbitrary s by the routine `stieltjes_sob.m`. The output `phat` is an array of dimension `(n+1)`$\times$`(N*(s+1))` containing the N values of the derivative of order σ of the νth-degree approximant $\hat{p}_\nu, \nu \le n$, in positions $(\nu+1, \sigma+1 : s+1 : N*(s+1))$ of the array `phat`. The Fourier coefficients $\hat{c}_i$ are output in the `(n+1)`$\times$`1` vector `c`.

Example 15. The complementary error function on $[0, 2]$.

This is the function

$$f(t) = e^{t^2} \operatorname{erfc} t = \frac{2}{\sqrt{\pi}} e^{t^2} \int_t^\infty e^{-u^2} du, \quad 0 \le t \le 2,$$

whose derivatives are easily calculated.

Demo#8 The routine `least_squares_sob.m` is applied to the function f of Example 15 with $s = 2$ and `N`=5 equally spaced points t_k on $[0, 2]$. All weights are chosen to be equal, $w_k^{(\sigma)} = 1/N$ for $\sigma = 0, 1, 2$. The table below, in the top half, shows selected results for the Sobolev least squares error $\hat{E}_n$

s	n	$\hat{E}_n$	$E^{\infty}_{n,0}$	$E^{\infty}_{n,1}$	$E^{\infty}_{n,2}$
2	0	1.153(+00)	4.759(−01)	1.128(+00)	2.000(+00)
	2	7.356(−01)	8.812(−02)	2.860(−01)	1.411(+00)
	4	1.196(−01)	1.810(−02)	5.434(−02)	1.960(−01)
	9	2.178(−05)	4.710(−06)	3.011(−05)	3.159(−04)
	14	3.653(−16)	1.130(−09)	1.111(−08)	1.966(−07)
0	0	2.674(−01)	4.759(−01)	1.128(+00)	2.000(+00)
	2	2.245(−02)	3.865(−02)	3.612(−01)	1.590(+00)
	4	1.053(−16)	3.516(−03)	5.160(−02)	4.956(−01)

and the maximum errors $E^{\infty}_{n,0}$, $E^{\infty}_{n,1}$, $E^{\infty}_{n,2}$ (over 100 equally spaced points on $[0,2]$) for the function and its first two derivatives. In the bottom half are shown the analogous results for ordinary least squares approximation ($s = 0$) when $n \leq N-1$. (It makes no sense to consider $n > N-1$.) Note that the Sobolev least squares error $\hat{E}_{3N-1}$ is essentially zero, reflecting the fact that the Hermite interpolation polynomial of degree $3N-1$ interpolates the data exactly. In contrast, $\hat{E}_n = 0$ for $n \geq N-1$ in the case of ordinary least squares.

As expected, the table shows rather convincingly that Sobolev least squares approximation approximates the derivatives decidedly better than ordinary least squares approximation, when applicable, and even the function itself, when n is sufficiently large.

5.2 Moment-Preserving Spline Approximation

There are various types of approximation: those that control the maximum pointwise error; those that control some average error (like least squares error); and those, often motivated by physical considerations, that try to preserve the moments of the given function, or at least as many of the first moments as possible. It is this last type of approximation that we now wish to study. We begin with piecewise constant approximation on the whole positive real line $\mathbb{R}_+$, then proceed to spline approximation on $\mathbb{R}_+$, and end with spline approximation on a compact interval.

Piecewise Constant Approximation on $\mathbb{R}_+$

The piecewise constant approximants to be considered are

$$s_n(t) = \sum_{\nu=1}^{n} a_\nu H(t_\nu - t), \quad t \in \mathbb{R}_+, \tag{5.17}$$

where $a_\nu \in \mathbb{R}$, $0 < t_1 < t_2 < \cdots < t_n$, and H is the Heaviside function

$$H(u) = \begin{cases} 1 & \text{if } u \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

The problem is, for given $f \in C^1(\mathbb{R}_+)$, to find, if possible, the a_ν and t_ν such that

$$\int_0^\infty s_n(t)t^j \mathrm{d}t = \mu_j, \quad j = 0, 1, \ldots, 2n-1, \tag{5.18}$$

where

$$\mu_j = \int_0^\infty f(t)t^j \mathrm{d}t, \quad j = 0, 1, \ldots, 2n-1, \tag{5.19}$$

are the *moments* of f, assumed to exist.

The solution can be formulated in terms of Gauss quadrature relative to the measure

$$\mathrm{d}\lambda(t) = -tf'(t)\mathrm{d}t \quad \text{on } \mathbb{R}_+. \tag{5.20}$$

Indeed, if $f(t) = o(t^{-2n})$ as $t \to \infty$, then the problem has a unique solution if and only if $\mathrm{d}\lambda$ in (5.20) admits an n-point Gauss quadrature formula

$$\int_0^\infty g(t)\mathrm{d}\lambda(t) = \sum_{\nu=1}^n \lambda_\nu^G g(\tau_\nu^G), \quad g \in \mathbb{P}_{2n-1}, \tag{5.21}$$

satisfying $0 < \tau_1^G < \tau_2^G < \cdots < \tau_n^G$. If that is the case, then the desired knots t_ν and coefficients a_ν are given by

$$t_\nu = \tau_\nu^G, \quad a_\nu = \frac{\lambda_\nu^G}{\tau_\nu^G}, \quad \nu = 1, 2, \ldots, n. \tag{5.22}$$

A Gauss formula (5.21) always exists if $f' < 0$ on $\mathbb{R}_+$, that is, $\mathrm{d}\lambda(t) \geq 0$.

For the proof, we use integration by parts,

$$\int_0^T f(t)t^j \mathrm{d}t = \frac{1}{j+1} t^{j+1} f(t)\Big|_0^T - \frac{1}{j+1} \int_0^T f'(t)t^{j+1}\mathrm{d}t, \quad j \leq 2n-1,$$

and let $T \to \infty$. The integrated part on the right goes to zero by assumption on f, and the left-hand side converges to the jth moment of f, again by assumption. Therefore, the last term on the right also converges, and since $-tf'(t) = \mathrm{d}\lambda(t)$, one finds

$$\mu_j = \frac{1}{j+1} \int_0^\infty t^j \mathrm{d}\lambda(t), \quad j = 0, 1, \ldots, 2n-1.$$

This shows in particular that the first $2n$ moments of $\mathrm{d}\lambda$ exist, and therefore, if $\mathrm{d}\lambda \geq 0$, also the Gauss formula (5.21).

On the other hand, the approximant s_n has moments

$$\int_0^\infty s_n(t)t^j \mathrm{d}t = \sum_{\nu=1}^n a_\nu \int_0^{t_\nu} t^j \mathrm{d}t = \frac{1}{j+1} \sum_{\nu=1}^n a_\nu t_\nu^{j+1},$$

so that the first $2n$ moments μ_j of f are preserved if and only if

$$\sum_{\nu=1}^{n}(a_\nu t_\nu)t_\nu^j = \int_0^\infty t^j \mathrm{d}\lambda(t), \quad j = 0, 1, \ldots, 2n-1.$$

This is equivalent to saying that the knots t_ν are the Gauss nodes in (5.21), and $a_\nu t_\nu$ the corresponding weights.

Example 16. Maxwell distribution $f(t) = \mathrm{e}^{-t^2}$ on $\mathbb{R}_+$.

Here,

$$\mathrm{d}\lambda(t) = 2t^2\mathrm{e}^{-t^2}\mathrm{d}t \quad \text{on } \mathbb{R}_+,$$

which is a positive measure obtained (up to the factor 2) by twice modifying the half-range Hermite measure by a linear factor t. The first $n+2$ recurrence coefficients of the half-range Hermite measure can be computed by a discretization method. Applying to these recurrence coefficients twice the routine `chri1.m`, with zero shift, then yields the recurrence coefficients $\alpha_k(\mathrm{d}\lambda)$, $\beta_k(\mathrm{d}\lambda)$, $k \leq n-1$, and hence the required n-point Gauss quadrature rule (5.21) for $\mathrm{d}\lambda$. The result for $n = 5$ is depicted in Fig. 4.

Spline Approximation on $\mathbb{R}_+$

The approximant s_n in (5.17) can be interpreted as a spline function of degree 0. We now consider spline functions $s_{n,m}$ of degree $m > 0$,

$$s_{n,m}(t) = \sum_{\nu=1}^{n} a_\nu (t_\nu - t)_+^m, \quad t \in \mathbb{R}_+, \tag{5.23}$$

where u_+^m is the truncated power $u_+^m = u^m$ if $u \geq 0$, and $u_+^m = 0$ if $u < 0$. Given the first $2n$ moments (5.19) of f, the problem again is to determine $a_\nu \in \mathbb{R}$ and $0 < t_1 < t_2 < \cdots < t_n$ such that

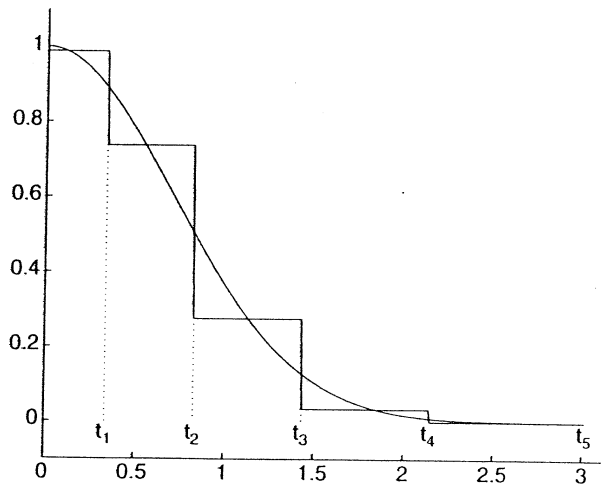


Fig. 4. Piecewise constant approximation of the Maxwell distribution

$$\int_0^\infty s_{n,m}(t)t^j \mathrm{d}t = \mu_j, \quad j = 0, 1, \ldots, 2n-1. \tag{5.24}$$

By a reasoning similar to the one in the previous subsection, but more complicated, involving m integrations by part, one proves that for $f \in C^{m+1}(\mathbb{R}_+)$ and satisfying $f^{(\mu)}(t) = o(t^{-2n-\mu})$ as $t \to \infty$, $\mu = 0, 1, \ldots, m$, the problem has a unique solution if and only if the measure

$$\mathrm{d}\lambda^{[m]}(t) = \frac{(-1)^{m+1}}{m!} t^{m+1} f^{(m+1)}(t)\mathrm{d}t \quad \text{on } \mathbb{R}_+ \tag{5.25}$$

admits an n-point Gauss quadrature formula

$$\int_0^\infty g(t)\,\mathrm{d}\lambda^{[m]}(t) = \sum_{\nu=1}^{n} \lambda_\nu^G g(\tau_\nu^G) \quad \text{for all } g \in \mathbb{P}_{2n-1} \tag{5.26}$$

satisfying $0 < \tau_1^G < \tau_2^G < \cdots < \tau_n^G$. If that is the case, the knots t_ν and coefficients a_ν are given by

$$t_\nu = \tau_\nu^G, \quad a_\nu = \frac{\lambda_\nu^G}{[\tau_\nu^G]^{m+1}}, \quad \nu = 1, 2, \ldots, n. \tag{5.27}$$

Note that $\mathrm{d}\lambda^{[m]}$ in (5.25) is a positive measure, for each $m \geq 0$, and hence (5.26) exists, if f is completely monotonic on $\mathbb{R}_+$, that is, $(-1)^\mu f^{(\mu)}(t) > 0$, $t \in \mathbb{R}_+$, for $\mu = 0, 1, 2, \ldots$.

Example 17. Maxwell distribution $f(t) = \mathrm{e}^{-t^2}$ on $\mathbb{R}_+$, revisited.

We now have

$$\mathrm{d}\lambda^{[m]}(t) = \frac{1}{m!} t^{m+1} H_{m+1}(t)\mathrm{e}^{-t^2}\,\mathrm{d}t \quad \text{on } \mathbb{R}_+,$$

where H_{m+1} is the Hermite polynomial of degree $m+1$. Here, $\mathrm{d}\lambda^{[m]}$ if $m > 0$ is no longer of constant sign on $\mathbb{R}_+$, and hence the existence of the Gauss rule (5.26) is in doubt. Numerical exploration, using discretization methods, yields the situation shown in the table below, where a dash indicates the presence of a negative Gauss node τ_ν^G, and an asterisk the presence of a pair

n	$m=1$	$m=2$	$m=3$	n	$m=1$	$m=2$	$m=3$
1	6.9(−2)	1.8(−1)	2.6(−1)	11	—	1.1(−3)	1.1(−4)
2	8.2(−2)	—	2.3(−1)	12	—	—	*
3	—	1.1(−2)	2.5(−3)	13	7.8(−3)	6.7(−4)	*
4	3.5(−2)	6.7(−3)	2.2(−3)	14	8.3(−3)	5.6(−4)	8.1(−5)
5	2.6(−2)	—	1.6(−3)	15	7.7(−3)	—	7.1(−5)
6	2.2(−2)	3.1(−3)	*	16	—	4.9(−4)	7.8(−5)
7	—	2.4(−3)	*	17	—	3.8(−4)	3.8(−5)
8	1.4(−2)	—	3.4(−4)	18	5.5(−3)	3.8(−4)	*
9	1.1(−2)	1.7(−3)	2.5(−4)	19	5.3(−3)	—	*
10	9.0(−3)	1.1(−3)	—	20	5.4(−3)	3.1(−4)	*

of conjugate complex Gauss nodes. In all cases computed, there were never more than one negative Gauss node, or more than one pair of complex nodes. The numbers in the table represent the maximum errors $\|s_{n,m} - f\|_\infty$, the maximum being taken over 100 equally spaced points on $[0, \tau_n^G]$.

Spline Approximation on a Compact Interval

The problem on a compact interval, say $[0,1]$, is a bit more involved than the problem on $\mathbb{R}_+$. For one, the spline function $s_{n,m}$ may now include a polynomial p of degree m, which was absent before since no moment of p exists on $\mathbb{R}_+$ unless $p \equiv 0$. Thus, the spline approximant has now the form

$$s_{n,m}(t) = p(t) + \sum_{\nu=1}^{n} a_\nu (t_\nu - t)_+^m, \quad p \in \mathbb{P}_m, \quad 0 \le t \le 1, \tag{5.28}$$

where $a_\nu \in \mathbb{R}$ and $0 < t_1 < t_2 < \cdots < t_n < 1$. There are two problems of interest:

Problem I. Find $s_{n,m}$ such that

$$\int_0^1 s_{n,m}(t) t^j \mathrm{d}t = \mu_j, \quad j = 0, 1, \ldots, 2n + m. \tag{5.29}$$

Since we have $m+1$ additional parameters at our disposal (the coefficients of p), we can impose $m+1$ additional moment conditions.

Problem II. Rather than matching more moments, we use the added degree of freedom to impose $m+1$ "boundary conditions" at the end point $t = 1$. More precisely, we want to find $s_{n,m}$ such that

$$\int_0^1 s_{n,m}(t) t^j \mathrm{d}t = \mu_j, \quad j = 0, 1, \ldots, 2n - 1 \tag{5.30}$$

and

$$s_{n,m}^{(\mu)}(1) = f^{(\mu)}(1), \quad \mu = 0, 1, \ldots, m. \tag{5.31}$$

It is still true that a solution can be given in terms of quadrature formulae, but they are now respectively generalized Gauss–Lobatto and generalized Gauss–Radau formulae relative to the measure (see [5, 6])

$$\mathrm{d}\lambda^{[m]}(t) = \frac{(-1)^{m+1}}{m!} f^{(m+1)}(t) \mathrm{d}t \quad \text{on } [0,1]. \tag{5.32}$$

Problem I, in fact, has a unique solution if and only if the *generalized Gauss–Lobatto formula*

$$\begin{aligned} \int_0^1 g(t) \mathrm{d}\lambda^{[m]}(t) = &\sum_{\mu=0}^{m} [\lambda_0^{(\mu)} g^{(\mu)}(0) + (-1)^\mu \lambda_{n+1}^{(\mu)} g^{(\mu)}(1)] \\ &+ \sum_{\nu=1}^{n} \lambda_\nu^L g(\tau_\nu^L), \quad g \in \mathbb{P}_{2n+2m+1}, \end{aligned} \tag{5.33}$$

exists with $0 < \tau_1^L < \cdots < \tau_n^L < 1$. In this case,

$$t_\nu = \tau_\nu^L, \quad a_\nu = \lambda_\nu^L, \quad \nu = 1, 2, \ldots, n, \tag{5.34}$$

and p is uniquely determined by

$$p^{(\mu)}(1) = f^{(\mu)}(1) + (-1)^m m! \lambda_{n+1}^{(m-\mu)}, \quad \mu = 0, 1, \ldots, m. \tag{5.35}$$

Similarly, Problem II has a unique solution if and only if the *generalized Gauss–Radau formula*

$$\int_0^1 g(t) \mathrm{d}\lambda^{[m]}(t) = \sum_{\mu=0}^{m} \lambda_0^{(\mu)} g^{(\mu)}(0) + \sum_{\nu=1}^{n} \lambda_\nu^R g(\tau_\nu^R), \quad g \in \mathbb{P}_{2n+m}, \tag{5.36}$$

exists with $0 < \tau_1^R < \cdots < \tau_n^R < 1$. Then

$$t_\nu = \tau_\nu^R, \quad a_\nu = \lambda_\nu^R, \quad \nu = 1, 2, \ldots, n, \tag{5.37}$$

and (trivially)

$$p(t) = \sum_{\mu=0}^{m} \frac{f^{(\mu)}(1)}{\mu!} (t-1)^\mu. \tag{5.38}$$

In both cases, complete monotonicity of f implies $\mathrm{d}\lambda \geq 0$ and the existence of the respective quadrature formulae. For their construction, see Exercises 12 and 13 of §4.

5.3 Slowly Convergent Series

Standard techniques of accelerating the convergence of slowly convergent series are based on linear or nonlinear sequence transformations: the sequence of partial sums is transformed somehow into a new sequence that converges to the same limit, but a lot faster. Here we follow another approach, more in the spirit of these lectures: the sum of the series is represented as a definite integral; a sequence of quadrature rules is then applied to this integral which, when properly chosen, will produce a sequence of approximations that converges quickly to the desired sum.

An easy way (and certainly not the only one) to obtain an integral representation presents itself when the general term of the series, or part thereof, is expressible in terms of the Laplace transform (or some other integral transform) of a known function. Several instances of this will now be described.

Series Generated by a Laplace Transform

The series

$$S = \sum_{k=1}^{\infty} a_k \tag{5.39}$$

to be considered first has terms a_k that are the Laplace transform

$$(\mathcal{L}f)(s) = \int_0^\infty \mathrm{e}^{-st} f(t)\mathrm{d}t$$

of some known function f evaluated at $s = k$,

$$a_k = (\mathcal{L}f)(k), \quad k = 1, 2, 3, \ldots . \tag{5.40}$$

In this case,

$$\begin{aligned} S &= \sum_{k=1}^{\infty} \int_0^\infty \mathrm{e}^{-kt} f(t)\mathrm{d}t \\ &= \int_0^\infty \sum_{k=1}^{\infty} \mathrm{e}^{-(k-1)t} \cdot \mathrm{e}^{-t} f(t)\mathrm{d}t \\ &= \int_0^\infty \frac{1}{1-\mathrm{e}^{-t}} \mathrm{e}^{-t} f(t)\mathrm{d}t \end{aligned}$$

that is,

$$S = \int_0^\infty \frac{t}{1-\mathrm{e}^{-t}} \frac{f(t)}{t} \mathrm{e}^{-t}\mathrm{d}t. \tag{5.41}$$

There are at least three different approaches to evaluate this integral numerically: one is Gauss–Laguerre quadrature of $(t/(1-\mathrm{e}^{-t}))f(t)/t$ with $\mathrm{d}\lambda(t) = \mathrm{e}^{-t}\mathrm{d}t$ on $\mathbb{R}_+$; another is rational/polynomial Gauss–Laguerre quadrature of the same function; and a third Gauss–Einstein quadrature of the function $f(t)/t$ with $\mathrm{d}\lambda(t) = t\mathrm{d}t/(\mathrm{e}^t - 1)$ on $\mathbb{R}_+$. In the last method, the weight function $t/(\mathrm{e}^t - 1)$ is widely used in solid state physics, where it is named after Einstein (coming from the Einstein-Bose distribution). It is also, incidentally, the generating function of the Bernoulli polynomials.

Example 18. The Theodorus constant

$$S = \sum_{k=1}^{\infty} \frac{1}{k^{3/2} + k^{1/2}} = 1.860025\ldots .$$

This is a universal constant introduced by P.J. Davis (1993) in connection with a spiral attributed to the ancient mathematician Theodorus of Cyrene.

Here we note that

$$\frac{1}{s^{3/2} + s^{1/2}} = s^{-1/2} \frac{1}{s+1} = \left(\mathcal{L}\frac{1}{\sqrt{\pi t}} * \mathrm{e}^{-t}\right)(s),$$

where the star stands for convolution. A simple computation yields (5.40) with

$$f(t) = \frac{2}{\sqrt{\pi}} F(\sqrt{t}),$$

where

$$F(x) = \mathrm{e}^{-x^2} \int_0^x \mathrm{e}^{t^2} \mathrm{d}t$$

is Dawson's integral.

Demo#9 To make $f(t)$ regular at $t = 0$, we divide by $\sqrt{t}$ and write

$$\begin{aligned} S &= \frac{2}{\sqrt{\pi}} \int_0^\infty \frac{t}{1-\mathrm{e}^{-t}} \frac{F(\sqrt{t})}{\sqrt{t}} t^{-1/2} \mathrm{e}^{-t} \mathrm{d}t \\ &= \frac{2}{\sqrt{\pi}} \int_0^\infty \frac{F(\sqrt{t})}{\sqrt{t}} t^{-1/2} \frac{t}{\mathrm{e}^t - 1} \mathrm{d}t. \end{aligned}$$

To the first integral we apply Gauss–Laguerre quadrature with $\mathrm{d}\lambda(t) = t^{-1/2}\mathrm{e}^{-t}\mathrm{d}t$ on $\mathbb{R}_+$, or rational Gauss–Laguerre with the same $\mathrm{d}\lambda$, and to the second integral Gauss–Einstein quadrature (modified by the factor $t^{-1/2}$). The errors committed in these quadrature methods are shown in the table below.

n	Gauss-Laguerre	rational Gauss-Laguerre	Gauss-Einstein
1	9.6799(–03)	1.5635(–02)	1.3610(–01)
4	5.5952(–06)	1.1893(–08)	2.1735(–04)
7	4.0004(–08)	5.9689(–16)	3.3459(–07)
10	5.9256(–10)		5.0254(–10)
15	8.2683(–12)		9.4308(–15)
20	8.9175(–14)		4.7751(–16)
	timing: 10.8	timing: 8.78	timing: 10.4

The clear winner is rational Gauss–Laguerre, both in terms of accuracy and run time.

Example 19. The Hardy–Littlewood function

$$H(x) = \sum_{k=1}^{\infty} \frac{1}{k} \sin \frac{x}{k}, \quad x > 0.$$

It can be shown that

$$a_k := \frac{1}{k} \sin \frac{x}{k} = (\mathcal{L}f(t;x))(k),$$

where

$$f(t;x) = \frac{1}{2\mathrm{i}} \left[I_0(2\sqrt{\mathrm{i}xt}) - I_0(2\sqrt{-\mathrm{i}xt}) \right]$$

and I_0 is the modified Bessel function. This gives rise to the two integral representations

$$H(x) = \int_0^\infty \frac{t}{1-\mathrm{e}^{-t}} \frac{f(t;x)}{t} \mathrm{e}^{-t} \mathrm{d}t = \int_0^\infty \frac{f(t;x)}{t} \frac{t}{\mathrm{e}^t - 1} \mathrm{d}t.$$

Among the three quadrature methods, Gauss–Einstein performs best, but all suffer from internal cancellation of terms in the quadrature sum. The problem becomes more prominent as the number n of terms increases. In this case, other methods can be applied, using the Euler-Maclaurin formula [11].

Fig. 5 shows the behavior of $H(x)$ in the range $0 \leq x \leq 100$.

"Alternating" Series Generated by a Laplace Transform

These are series in which the general terms are Laplace transforms with alternating signs of some function f, that is, series (5.39) with

$$a_k = (-1)^{k-1}(\mathcal{L}f)(k), \quad k = 1, 2, 3, \ldots . \tag{5.42}$$

An elementary computation similar to the one carried out in the previous subsection will show that

$$S = \int_0^\infty \frac{1}{1+\mathrm{e}^{-t}} f(t)\mathrm{e}^{-t}\mathrm{d}t = \int_0^\infty f(t)\frac{1}{\mathrm{e}^t+1}\,\mathrm{d}t. \tag{5.43}$$

We can again choose between three quadrature methods: Gauss–Laguerre quadrature of the function $f(t)/(1+\mathrm{e}^{-t})$ with $\mathrm{d}\lambda(t) = \mathrm{e}^{-t}\mathrm{d}t$, rational/polynomial Gauss–Laguerre of the same function, and Gauss-Fermi quadrature of $f(t)$ with $\mathrm{d}\lambda(t) = \mathrm{d}t/(\mathrm{e}^t+1)$ involving the Fermi function $1/(\mathrm{e}^t+1)$ (also used in solid state physics).

Example 20. The series

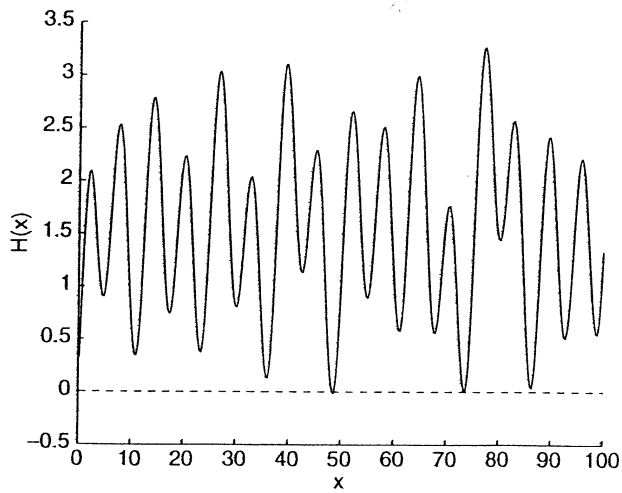


Fig. 5. The Hardy-Littlewood function

$$S = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} \mathrm{e}^{-1/k}.$$

One can show that the function f in question here is $f(t) = J_0(2\sqrt{t})$, with J_0 the Bessel function of order zero. Errors obtained by the three quadrature methods are displayed in the table below, showing the clear superiority of Gauss–Fermi quadrature.

n	Gauss-Laguerre	rational Gauss-Laguerre	Gauss-Fermi
1	1.6961(–01)	1.0310(–01)	5.6994(–01)
4	4.4754(–03)	4.6605(–05)	9.6454(–07)
7	1.7468(–04)	1.8274(–09)	9.1529(–15)
10	3.7891(–06)	1.5729(–13)	2.8163(–16)
15	2.6569(–07)	1.5490(–15)	
20	8.6155(–09)		
40	1.8066(–13)		
	timing: 12.7	timing: 19.5	timing: 4.95

Series Generated by the Derivative of a Laplace Transform

These are series (5.39) in which

$$a_k = -\frac{\mathrm{d}}{\mathrm{d}s}(\mathcal{L}f)(s)\Big|_{s=k}, \quad k = 1, 2, 3, \ldots. \tag{5.44}$$

In this case one finds

$$S = \int_0^{\infty} \frac{t}{1 - \mathrm{e}^{-t}} f(t)\mathrm{e}^{-t}\mathrm{d}t = \int_0^{\infty} f(t)\frac{t}{\mathrm{e}^t - 1}\,\mathrm{d}t, \tag{5.45}$$

and Gauss–Laguerre, rational/polynomial Gauss–Laguerre, and Gauss–Einstein quadrature are again options as in Examples 18 and 19.

Example 21. The series

$$S = \sum_{k=1}^{\infty} (\tfrac{3}{2}k + 1)k^{-2}(k+1)^{-3/2}.$$

The relevant function f is calculated to be

$$f(t) = \frac{\operatorname{erf}\sqrt{t}}{\sqrt{t}} \cdot t^{1/2},$$

where erf is the error function $\operatorname{erf} x = (2/\sqrt{\pi}) \int_0^x \mathrm{e}^{-t^2}\mathrm{d}t$. Numerical results analogous to those in the two previous tables are shown below.

n	Gauss-Laguerre	rational Gauss-Laguerre	Gauss-Einstein
1	4.0125(–03)	5.1071(–02)	8.1715(–02)
4	1.5108(–05)	4.5309(–08)	1.6872(–04)
7	4.6576(–08)	1.3226(–13)	3.1571(–07)
10	3.0433(–09)	1.2087(–15)	5.4661(–10)
15	4.3126(–11)		1.2605(–14)
20	7.6664(–14)		
30	3.4533(–16)		
	timing: 6.50	timing: 10.8	timing: 1.58

The run time is best for Gauss–Einstein quadrature, though the error is worse than for the closest competitor, rational Gauss–Laguerre.

Series Occurring in Plate Contact Problems

The series of interest here is

$$R_p(z) = \sum_{k=0}^{\infty} \frac{z^{2k+1}}{(2k+1)^p}, \quad z \in \mathbb{C},\ |z| \le 1,\ p = 2 \text{ or } 3. \tag{5.46}$$

Rather than expressing the whole general term of the series as a Laplace transform, we do this only for the coefficient,

$$\frac{1}{(k+\frac{1}{2})^p} = (\mathcal{L}f)(k), \quad f(t) = \frac{1}{(p-1)!}\, t^{p-1}\mathrm{e}^{-t/2}. \tag{5.47}$$

Then

$$\begin{aligned} R_p(z) &= \frac{z}{2^p} \sum_{k=0}^{\infty} \frac{z^{2k}}{(k+\frac{1}{2})^p} \\ &= \frac{z}{2^p} \sum_{k=0}^{\infty} z^{2k} \int_0^{\infty} \mathrm{e}^{-kt} \cdot \frac{t^{p-1}\mathrm{e}^{-t/2}}{(p-1)!}\, \mathrm{d}t \\ &= \frac{z}{2^p(p-1)!} \int_0^{\infty} \sum_{k=0}^{\infty} (z^2\mathrm{e}^{-t})^k \cdot t^{p-1}\mathrm{e}^{-t/2}\mathrm{d}t \\ &= \frac{z}{2^p(p-1)!} \int_0^{\infty} \frac{1}{1-z^2\mathrm{e}^{-t}}\, t^{p-1}\mathrm{e}^{-t/2}\mathrm{d}t, \end{aligned}$$

that is,

$$R_p(z) = \frac{z}{2^p(p-1)!} \int_0^{\infty} \frac{t^{p-1}\mathrm{e}^{t/2}}{\mathrm{e}^t - z^2}\, \mathrm{d}t, \quad z^{-2} \in \mathbb{C}\backslash[0,1]. \tag{5.48}$$

The case $z = 1$ can be treated directly by using the connection with the zeta function, $R_p(1) = (1-2^{-p})\zeta(p)$. Assume therefore $z \neq 1$. When $|z|$ is close to 1, the integrand in (5.48) is rather ill-behaved near $t = 0$, exhibiting a steep

boundary layer. We try to circumvent this by making the change of variables $e^{-t} \mapsto t$ to obtain

$$R_p(z) = \frac{1}{2^p(p-1)!z} \int_0^1 \frac{t^{-1/2}[\ln(1/t)]^{p-1}}{z^{-2}-t}\,dt.$$

This expresses $R_p(z)$ as a Cauchy integral of the measure

$$d\lambda^{[p]}(t) = t^{-1/2}[\ln(1/t)]^{p-1}dt.$$

Since by assumption, z^{-2} lies outside the interval $[0,1]$, the integral can be evaluated by the continued fraction algorithm, once sufficiently many recurrence coefficients for $d\lambda^{[p]}$ have been precomputed. For the latter, the modified Chebyshev algorithm is quite effective. The first 100 coefficients are available for $p = 2$ and $p = 3$ in the `OPQ` files `absqm1log1` and `absqm1log2` to 25 resp. 20 decimal digits.

Example 22.

$$R_p(x),\ p = 2 \text{ and } 3,\ x = .8, .9, .95, .99, .999 \text{ and } 1.0.$$

Numerical results are shown in he table below and are accurate to all digits

x	$R_2(x)$	$R_3(x)$
.8	0.87728809392147	0.82248858052014
.9	1.02593895111111	0.93414857586540
.95	1.11409957792905	0.99191543992243
.99	1.20207566477686	1.03957223187364
.999	1.22939819733	1.05056774973
1.000	1.233625	1.051795

shown. Full acuracy cannot be achieved for $x \geq .999$ using only 100 recurrence coefficients of $d\lambda^{[p]}$.

Example 23.

$$R_p(e^{i\alpha}),\ p = 2 \text{ and } 3,\ \alpha = \omega\pi/2,\ \omega = .2, .1, .05, .01, .001 \text{ and } 0.0.$$

Numerical results are shown in the table below.

p	ω	$\mathrm{Re}(R_p(z))$	$\mathrm{Im}(R_p(z))$
2	.2	0.98696044010894	0.44740227008596
3		0.96915102126252	0.34882061265337
2	0.1	1.11033049512255	0.27830297928558
3		1.02685555765937	0.18409976778928
2	0.05	1.17201552262936	0.16639152396897
3		1.04449441539672	0.09447224926029
2	0.01	1.22136354463481	0.04592009281744
3		1.05140829197388	0.01928202831056
2	0.001	1.232466849	0.006400460
3		1.051794454	0.001936923
2	0.000	1.2336	0.0000
3		1.0518	0.0000

Here, too, full accuracy is not attainable for $\omega \leq 0.001$ with only 100 recurrence coefficients. Curiously, the continued fraction algorithm seems to converge also when $z = 1$, albeit slowly.

Series Involving Ratios of Hyperbolic Functions

More of a challenge are series of the type

$$T_p(x;b) = \sum_{k=0}^{\infty} \frac{1}{(2k+1)^p} \frac{\cosh(2k+1)x}{\cosh(2k+1)b}, \quad 0 \leq x \leq b, \ b > 0, \ p = 2,3, \quad (5.49)$$

which also occur in plate contact problems. Here, we first expand the ratio of hyperbolic cosines into an infinite series,

$$\begin{aligned} &\frac{\cosh(2k+1)x}{\cosh(2k+1)b} \\ &\quad = \sum_{n=0}^{\infty} (-1)^n \left\{ \mathrm{e}^{-(2k+1)[(2n+1)b-x]} + \mathrm{e}^{-(2k+1)[(2n+1)b+x]} \right\}, \end{aligned} \quad (5.50)$$

insert this in (5.49) and apply the Laplace transform technique of the previous subsection. This yields, after an elementary computation (using an interchange of the summations over k and n),

$$T_p(x,b) = \frac{1}{2^p(p-1)!} \sum_{n=0}^{\infty} (-1)^n \mathrm{e}^{(2n+1)b} [\varphi_n(-x) + \varphi_n(x)], \quad (5.51)$$

where

$$\varphi_n(s) = \mathrm{e}^s \int_0^1 \frac{\mathrm{d}\lambda^{[p]}(t)}{\mathrm{e}^{2[(2n+1)b+s]} - t}, \quad -b \leq s \leq b. \quad (5.52)$$

The integral on the right is again amenable to the continued fraction algorithm for $\mathrm{d}\lambda^{[p]}$, which for large n converges almost instantaneously. Convergence of the series (5.51) is geometric with ratio e^{-b}.

Exercises to §5 (Stars indicate more advanced exercises.)

1. With $\pi_0, \pi_1, \ldots, \pi_{N-1}$ denoting the discrete orthogonal polynomials relative to the measure $\mathrm{d}\lambda_N$, and $\hat{c}_i(f)$ the Fourier coefficients of f with respect to these orthogonal polynomials, show that

$$\sum_{i=0}^{n} |\hat{c}_i(f)|^2 \|\pi_i\|^2 \leq \|f\|^2, \quad n < N,$$

with equality holding for $n = N - 1$.

2. Prove the following alternative form for the Fourier coefficients,

$$\hat{c}_i(f) = \frac{1}{\|\pi_i\|^2}\left(f - \sum_{j=0}^{i-1} \hat{c}_j(f)\pi_j, \pi_i\right), \quad i = 0, 1, \ldots, n,$$

and discuss its possible advantages over the original form.

3. Discuss the modifications required in the constrained least squares approximation when ν $(0 \le \nu \le m)$ of the points s_j are equal to one of the support points t_k.
4. What are $p_m(f; \cdot)$, f^*, and σ_m in Example 13?
5. Calculate the first and second derivative of the complementary error function of Example 15.

6*. Prove the unique solvability of the problem (5.24) under the conditions stated in (5.25)–(5.26), and, in the affirmative case, derive (5.27).

7. Derive the measure $d\lambda^{[m]}$ for the Maxwell distribution of Example 17.
8. Derive the formula for f in Example 18.
9. Derive the formula for f in Example 19.
10. Derive (5.43).
11. Derive the formula for f in Example 20.
12. Derive (5.45).
13. Derive the formula for f in Example 21.
14. Supply the details for deriving (5.51).

References

1. M. Abramowitz and I.A. Stegun (eds), *Handbook of Mathematical Functions*, Dover Publications, New York, 1992.
2. S. Chandrasekhar, *Radiative Transfer*, Oxford University Press, 1950.
3. B. Danloy, *Numerical construction of Gaussian quadrature formulas for* $\int_0^1 -\text{Log } x\, x^\alpha f(x)\,dx$ *and* $\int_0^\infty E_m(x)f(x)\,dx$, Math. Comp. **27** (1973), 861–869.
4. P.J. Davis and P. Rabinowitz, *Some geometrical theorems for abscissas and weights of Gauss type*, J. Math. Anal. Appl. **2** (1961), 428–437.
5. M. Frontini, W. Gautschi, G.V. Milovanović, *Moment-preserving spline approximation on finite intervals*, Numer. Math. **50** (1987), 503–518.
6. M. Frontini, W. Gautschi, G.V. Milovanović, *Moment-preserving spline approximation on finite intervals and Turán quadratures*, Facta Univ. Ser. Math. Inform. **4** (1989), 45–56.
7. W. Gautschi, *Anomalous convergence of a continued fraction for ratios of Kummer functions*, Math. Comp. **31** (1977), 994–999.
8. W. Gautschi, *On the computation of special Sobolev-type orthogonal polynomials*, Ann. Numer. Math. **4** (1997), 329–342.
9. W. Gautschi, *The interplay between classical analysis and (numerical) linear algebra—a tribute to Gene H. Golub*, Electr. Trans. Numer. Anal. **13** (2002), 119–147.

10. W. Gautschi, *Orthogonal Polynomials: Computation and Approximation*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 2004.
11. W. Gautschi, *The Hardy-Littlewood function: an exercise in slowly convergent series*, J. Comput. Appl. Math. **179** (2005), 249–254.
12. W. Gautschi and M. Zhang, *Computing orthogonal polynomials in Sobolev spaces*, Numer. Math. **71** (1995), 159–183.
13. I.S. Gradshteyn and I.M. Ryzhik, *Tables of Integrals, Series, and Products* (6th edn), Academic Press, San Diego, CA, 2000.
14. W.B. Gragg and W.J. Harrod, *The numerically stable reconstruction of Jacobi matrices from spectral data*, Numer. Math. **44** (1984), 317–335.
15. F. Peherstorfer, K. Petras, *Stieltjes polynomials and Gauss-Kronrod quadrature for Jacobi weight functions*, Numer. Math. **95** (2003), 689–706.
16. G. Szegő, *Orthogonal Polynomials* (4th edn), AMS Colloq. Publ. 23, Amer. Math. Soc., Providence, RI, 1975.
17. P. Verlinden, *Stable rational modification of a weight*, Numer. Algorithms **22** (1999), 183–192.

29

Papers on History and Biography

---

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 (1981)

91 (with J. Wimp) In memoriam: Yudell L. Luke June 26, 1918 – May 6, 1983, Math. Comp. 43, 349–352 (1984)

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 (1986)

143 The work of Philip Rabinowitz on numerical integration, Numer. Algorithms 9, 199–222 (1995)

144 Luigi Gatteschi's work on special functions and numerical analysis, in *Special Functions* (G. Allasia, ed.), Annals Numer. Math. 2, 3–19 (1995)

170 The interplay between classical analysis and (numerical) linear algebra — a tribute to Gene H. Golub, Electron. Trans. Numer. Anal. 13, 119–147 (2002)

183 Leonhard Eulers Umgang mit langsam konvergenten Reihen, Elem. Math. 62, 174–183 (2007)

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 (2007)

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 (2008)

187 Leonhard Euler: his life, the man, and his work, SIAM Rev. 50, 3–33 (2008). [Also published in *ICIAM 07, 6$^{\text{th}}$ 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* (2–3) (2008).]

189 (with C. Giordano) Luigi Gatteschi's work on asymptotics of special functions tions and their zeros, in *A collection of essays in memory of Luigi Gatteschi* (G. Allasia, C. Brezinski, and M. Redivo-Zaglia, eds.), *Numer. Algorithms* 49, 11–31 (2008)

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 (2010)

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 (2011)

29.1. [74] "A Survey of Gauss–Christoffel Quadrature Formulae"

---

[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 (1981).

E. B. Christoffel
Ed. by P. L. Butzer and F. Fehér, Aachen
Birkhäuser Verlag, Basel 1981

A Survey of Gauss–Christoffel Quadrature Formulae

Walter Gautschi
Department of Computer Sciences, Purdue University, West Lafayette, Ind. (USA)

We present a historical survey of Gauss–Christoffel quadrature formulae, beginning with Gauss' discovery of his well-known method of approximate integration and the early contributions of Jacobi and Christoffel, but emphasizing the more recent advances made after the emergence of powerful digital computing machinery. One group of inquiry concerns the development of the quadrature formula itself, e.g. the inclusion of preassigned nodes and the admission of multiple nodes, as well as other generalizations of the quadrature sum. Another is directed towards the widening of the class of integrals made accessible to Gauss–Christoffel quadrature. These include integrals with nonpositive measures of integration and singular principal value integrals. An account of the error and convergence theory will also be given, as well as a discussion of modern methods for generating Gauss–Christoffel formulae, and a survey of numerical tables.

Introduction

Gauss' famous method of approximate integration, almost immediately after its discovery and throughout the 19th century, attracted the attention of some of the leading mathematicians of the time. It first inspired Jacobi to provide an elegant alternative derivation. Christoffel then significantly generalized the method and subsequently extended it to arbitrary measures of integration. Stieltjes established the legitimacy of the method, by proving its convergence, while Markov endowed it with an error term. Thus, by the end of the 19th century, the Gauss–Christoffel integration method became firmly entrenched in the repertoire of numerical methods of approximation.

Whether or not the Gauss–Christoffel method had actually been widely used in practice is a matter of some doubt, since the method requires the evaluation of functions at irrational arguments, hence tedious interpolation. All this changed when powerful digital computers entered the scene, which generated a climate of renewed interest in Gauss–Christoffel quadrature. The formulae began to be routinely applied, and increased usage, in turn, led to important new theoretical developments. The state of the art, including applications and extensive numerical tables, has been summarized in the book by Stroud & Secrest [1966]. Here we wish to present an extensive historical survey of Gauss–Christoffel quadrature formulae, covering the period from the early beginnings to the most recent developments, emphasizing, however, the progress made in the last 10–15 years.

We begin in Section 1 with a brief outline of the discovery of Gauss and

the early contributions of Jacobi, Christoffel, and others. In Section 2 we describe how the Gauss–Christoffel quadrature rule has been extended in various directions, first by Christoffel, who introduced preassigned nodes, then much later by Turán and others, who introduced derivative values in addition to function values. Closely related to Christoffel's work is Kronrod's extension of Gauss–Christoffel quadrature rules, which leads to practical schemes of implementation. Further miscellaneous extensions of the idea of Gauss continue to be made. Section 3 is devoted to various efforts of extending the scope of applications of Gauss–Christoffel formulae. Thus, applications to more general types of integrals are considered, including integrals with nonpositive measures of integration and singular principal value type integrals. In Section 4 we review work on the remainder term and related questions of convergence. Section 5, finally, will deal with constructive methods of generating Gauss–Christoffel formulae and also contains a review of available numerical tables.

Although an effort has been made to make this survey reasonably complete, it was not possible to include all topics of interest. Perhaps the most important omission is the extension of Gauss–Christoffel quadrature formulae to multiple integrals. While it is not entirely clear what constitutes a Gauss–Christoffel formula for a multiple integral, various interpretations are possible. A full discussion of these, however, would go beyond the scope of this review. Indefinite integrals, likewise, have been omitted from consideration. Numerous applications of Gauss–Christoffel's quadrature formula have been, and continue to be made, both within the fields of numerical analysis and outside of it. It was not feasible to survey them all, and we restricted ourselves to mentioning only a few selected applications, as the occasion permits. Special properties of zeros of orthogonal polynomials and of Christoffel numbers, and composite Gauss–Christoffel formulae, are additional topics left out from consideration.

1. Genesis of Gaussian Quadrature and Early History

The story of Gaussian quadrature begins with Newton and Cotes. Newton, in 1676, was the first to suggest a truly general method of approximate integration. Cotes, independently, arrived at similar methods, and brought them into workable form after learning of Newton's ideas. In 1814, Gauss takes the work of Newton and Cotes as a point of departure, combines it with his own work on the hypergeometric series, and develops his famous new method of integration which significantly improves upon the earlier method of Newton and Cotes. Gauss' work in turn was simplified by Jacobi and further developed through much of the 19th century by Mehler, Christoffel, and others. Eventu-

ally, there emerged a coherent theory which received its first systematic expositions by Christoffel [1877], Radau [1880], and Heine [1881] in his book on spherical functions.

In this section, we can only give a bare outline of the developments that took place in this period of approximately 200 years. A very detailed historical account can be found in Runge & Willers [1915], and a German edition of the four principal memoirs (of Newton, Cotes, Gauss and Jacobi) in Kowalewski [1917]. The important contribution of Christoffel [1858], which also falls in this period but points into new directions, will be discussed later in Section 2.1.1.

1.1. Newton–Cotes quadrature formulae

One of Newton's early accomplishments (which he already alluded to in a letter to Leibniz, dated October 24, 1676, and published later in 1687 as Lemma 5 in the third book of the "Principia") was his "... expeditious method of passing a parabolic curve through given points". In modern terminology, given a function f and n pairwise distinct points τ_ν, Newton constructs the unique polynomial $p_{n-1}(f;\cdot)$ of degree $\leq n-1$ which at the points τ_ν assumes the same values as f,

$$p_{n-1}(f;\tau_\nu) = f(\tau_\nu), \qquad \nu = 1, 2, \ldots, n, \quad p_{n-1} \in \mathbf{P}_{n-1}.$$

Newton ingeniously expresses this interpolation polynomial in terms of divided differences. Here we find it more convenient to express it in the form given much later (1795) by Lagrange,

$$p_{n-1}(f;t) = \sum_{\nu=1}^{n} l_\nu(t) f(\tau_\nu), \tag{1.1}$$

where $l_\nu \in \mathbf{P}_{n-1}$ are the special polynomials with $l_\nu(\tau_\nu) = 1$ and $l_\nu(\tau_\mu) = 0$, $\mu \neq \nu$. If we write

$$f(t) = p_{n-1}(f;t) + r_n(f;t), \tag{1.2}$$

where $r_n(f;\cdot)$ denotes the interpolation error, we then have, by the uniqueness of the interpolation polynomial,

$$r_n(f;\cdot) = 0, \qquad \text{all } f \in \mathbf{P}_{n-1}. \tag{1.3}$$

Newton, in the "Principia", already hints at the possibility that "... the area under the curve can be found, since the quadrature of a parabolic curve can be effected". Indeed, if

$$I(f) = \int_a^b f(t)\,dt, \tag{1.4}$$

where $a < b$ are finite numbers, integration of (1.2) yields

$$I(f) = Q_n(f) + R_n(f), \quad Q_n(f) = \sum_{\nu=1}^{n} \lambda_\nu f(\tau_\nu) \tag{1.5}$$

where, by (1.1),

$$\lambda_\nu = I(l_\nu), \qquad \nu = 1, 2, \ldots, n, \quad R_n(f) = I(r_n(f; \cdot)). \tag{1.6}$$

(The quantities λ_ν, as well as the points τ_ν, depend on n; for simplicity we suppress this dependence in our notation, both here and in subsequent discussions.) One calls (1.5) an n-point *quadrature formula*, $Q_n(f)$ the *quadrature sum*, and $R_n(f)$ the *remainder*. The points τ_ν are also referred to as *nodes*, while the numbers λ_ν are called the *weights* of the quadrature formula. The quadrature sum is expected to approximate the integral, the error being given by the remainder. The latter, by virtue of (1.3), satisfies

$$R_n(f) = 0, \qquad \text{all } f \in \mathbf{P}_{n-1}. \tag{1.7}$$

Following Radau [1880], one expresses (1.7) by saying that the quadrature rule Q_n has *degree of exactness* $n-1$, and we write $d(Q_n) = n-1$. (Q_n then also has degree of exactness k for any integer k with $0 \le k < n$.) It is easily seen that a quadrature formula Q_n has degree of exactness $n-1$ if and only if it is obtained via interpolation, as described. Hence Q_n, for which (1.7) holds, is also called *interpolatory*.

In the case of equally spaced points τ_ν, the numbers λ_ν in (1.6) can be computed once and for all. They are called *Cotes numbers*, in recognition of Roger Cotes who first computed them for $n \le 11$. (For a history of these numbers, see Johnson [1915].) In the case of arbitrary (distinct) nodes τ_ν, the Cotes numbers can be expressed in terms of the *node polynomial*

$$\omega_n(t) = \prod_{\nu=1}^{n} (t - \tau_\nu). \tag{1.8}$$

Indeed, from (1.6), using Lagrange's formula for l_ν, one gets

$$\lambda_\nu = I\left[\frac{\omega_n(\cdot)}{\omega_n'(\tau_\nu)(\cdot - \tau_\nu)}\right], \qquad \nu = 1, 2, \ldots, n. \tag{1.9}$$

The formula (1.5), with λ_ν given by (1.9), is called the *Newton–Cotes quadrature formula*. It includes as special cases the trapezoidal formula, Simpson's formula, and many other formulas that were known prior to Newton's time. It will serve here as a basis on which to build the more advanced quadrature formulae to be discussed later in this survey.

1.2. The discovery of Gauss

If we let the nodes τ_ν in the Newton–Cotes formula (1.5) vary freely, and for each set of (distinct) nodes compute the weights λ_ν in accordance with (1.9),

what is the maximum degree of exactness that can be achieved? And how are the nodes τ_ν to be selected in order to realize this optimum? These were questions raised by Gauss [1814], and answered most elegantly by means of his theory of continued fractions associated with hypergeometric series.

To begin with, one easily conjectures that $\max_{\tau_\nu,\lambda_\nu} d(Q_n) = 2n-1$, since there are $2n$ unknowns to be found, and $2n$ conditions imposed. To verify the conjecture, Gauss starts from the characteristic function of the "monomial errors",

$$R_n\left(\frac{1}{z-\cdot}\right) = \sum_{k=0}^{\infty} \frac{R_n(t^k)}{z^{k+1}},$$

where z is a formal parameter (intended to be large). The problem then amounts to determine τ_ν, λ_ν such that

$$R_n\left(\frac{1}{z-\cdot}\right) = O\left(\frac{1}{z^{2n+1}}\right), \qquad z \to \infty.$$

Observe now that for the integral in (1.4), where $b = -a = 1$,

$$(1.10)' \qquad I\left(\frac{1}{z-\cdot}\right) = \ln\frac{1+1/z}{1-1/z} = \frac{2}{z-}\,\frac{1/3}{z-}\,\frac{2\cdot 2/3\cdot 5}{z-}\,\frac{3\cdot 3/5\cdot 7}{z-}\cdots.$$

The continued fraction on the right was well known to Gauss, being a special case of his general continued fraction for ratios of hypergeometric functions (Gauss [1812]). He also knew well that the n-th convergent — a rational function $R_{n-1,n}$ with numerator degree $n-1$ and denominator degree n — if expanded in reciprocal powers of z, approximates the function on the left up to terms of order z^{-2n-1},

$$I\left(\frac{1}{z-\cdot}\right) = R_{n-1,n}(z) + O\left(\frac{1}{z^{2n+1}}\right), \qquad z \to \infty.$$

Gauss now decomposes $R_{n-1,n}$ in partial fractions and takes the residues and poles to be the weights and nodes in the quadrature formula (1.5),

$$R_{n-1,n}(z) = \sum_{\nu=1}^{n} \frac{\lambda_\nu}{z-\tau_\nu} =: Q_n\left(\frac{1}{z-\cdot}\right).$$

It then follows immediately that

$$\begin{aligned} R_n\left(\frac{1}{z-\cdot}\right) &= I\left(\frac{1}{z-\cdot}\right) - Q_n\left(\frac{1}{z-\cdot}\right) \\ &= I\left(\frac{1}{z-\cdot}\right) - R_{n-1,n}(z) = O\left(\frac{1}{z^{2n+1}}\right), \qquad z \to \infty, \end{aligned}$$

hence $d(Q_n) = 2n-1$, as desired. Gauss proceeds to express the denominator

and numerator polynomials of $R_{n-1,n}$ (now known as Legendre polynomials of the first and second kind) in terms of his hypergeometric series.

Gauss' discovery must be rated as one of the most significant events of the 19th century in the field of numerical integration and perhaps in all of numerical analysis. The result not only has great beauty and power, but also influenced many later developments in computing and approximation. It soon inspired contemporaries, such as Jacobi and Christoffel, to perfect Gauss' method and to develop it into new directions. Towards the end of the century, it inspired Heun [1900] to generalize Gauss' idea to ordinary differential equations, which in turn led to significant developments in the numerical solution of differential equations, notably the discovery of the Runge–Kutta method (Kutta [1901]). Gauss' influence continues into the 20th century and is still felt today, as we shall have ample occasion to document in subsequent chapters of this survey.

1.3. The contribution of Jacobi

The continued fraction (1.10) and its close association with the integral $I(1/(z-\cdot))$ is seen by Gauss to be the true source of his new method of integration. Jacobi [1826], on the other hand, with characteristic clarity and simplicity, derives Gauss' result purely on the basis of polynomial divisibility arguments. The central concept that emerges in Jacobi's work is *orthogonality*. (The name "orthogonal" for function systems came into use only later, probably first in E. Schmidt's 1905 Göttingen dissertation; see also Schmidt [1907, p. 439]. For polynomials, the term appears in the early writings of Szegö (Szegö [1918], [1919]. Murphy [1835] uses the term "reciprocal".) In effect, Jacobi shows that, *given any integer* k, *with* $0 \le k \le n$, *the quadrature rule* Q_n *in* (1.5) *has degree of exactness* $d(Q_n) = n - 1 + k$ *if and only if the following two conditions are satisfied*:

(i) Q_n *is interpolatory*

(ii) $I(\omega_n p) = 0$, *all* $p \in \mathbf{P}_{k-1}$.

Here ω_n is the node polynomial (1.8). Condition (ii) requires ω_n to be orthogonal to all polynomials of degree $\le k-1$. (If $k = 0$, a polynomial of degree -1 is understood to be identically zero.) It is seen, therefore, that each additional degree of exactness, over and beyond what is possible with the Newton–Cotes formula, requires orthogonality of ω_n to one additional power. In particular, $k \le n$, since ω_n cannot be orthogonal to itself, so that the maximum possible degree of exactness is indeed $2n-1$.

Jacobi's argumentation is extremely transparent; it goes as follows: Clearly, (i) is necessary. The necessity of (ii) is a consequence of $I(\omega_n p) = Q_n(\omega_n p) = 0$, the degree of exactness of Q_n being $n-1+k$ and ω_n vanishing at all the nodes τ_ν. For the sufficiency, let p be an arbitrary polynomial of degree $\le n-1+k$. Divide p by ω_n,

$$p = q\omega_n + r, \qquad q \in \mathbf{P}_{k-1},\ r \in \mathbf{P}_{n-1}.$$

Then

$$\begin{aligned} I(p) &= I(q\omega_n) + I(r) \\ &= I(r) \qquad \text{[by (ii)]} \\ &= Q_n(r) \qquad \text{[by (i)]} \\ &= Q_n(p) - Q_n(q\omega_n) = Q_n(p), \end{aligned}$$

i.e., Q_n has degree of exactness $n - 1 + k$.

The case $k = n$, of course, is of particular interest, as it leads to the Gauss formula of maximum degree of exactness. In this case, ω_n must be orthogonal to all lower degree polynomials, i.e. ω_n is the n-th degree Legendre polynomial (if the interval $[a, b]$ is standardized to $[-1, 1]$),

$$\omega_n(t) = \pi_n(t), \quad I(\pi_k \pi_l) = 0 \qquad \text{for } k \neq l.$$

Jacobi then proceeds to obtain the "Rodrigues formula"

$$\pi_n(t) = \text{const} \cdot D^n(t^2 - 1)^n, \qquad D = d/dt, \tag{1.11}$$

from which he concludes that all nodes τ_ν are real, simple, and contained in the interior of $[-1, 1]$. (The simplicity of the nodes is already pointed out by Gauss. The fact that all weights λ_ν are positive seems to have escaped both Gauss and Jacobi, although Jacobi's result $-1 < \tau_\nu < 1$, combined with an observation of Gauss (Gauss [1814, §21]) indeed yields positivity).

The analogue of (1.11) for general $0 \leq k \leq n$,

$$\omega_n(t) = \text{const} \cdot D^k[(t^2 - 1)^k p(t)], \qquad p \in \mathbf{P}_{n-k},$$

where p has exact degree $n - k$, but is otherwise arbitrary, is due to Radau [1880].

1.4. Gauss–Christoffel quadrature formulae

After the work of Jacobi, the matter of Gaussian quadrature, except for Christoffel's 1858 memoir which we discuss later, remained dormant for nearly forty years. Then, in 1864, Mehler, and others after him, began to introduce weighted integrals, i.e. integrals over $[-1, 1]$ with respect to a measure $d\lambda(t) = \omega(t)dt$ with $\omega \not\equiv 1$. This development soon led Posse [1875], and Christoffel [1877], to consider the case of a *general* (nonnegative and integrable) *weight function* ω on a finite interval $[a, b]$. Christoffel, in particular, systematically generalizes the Gauss–Jacobi theory to arbitrary weighted integrals, and in the process establishes (what is now called) the Christoffel–Darboux formula for an arbitrary weight function (anticipating

Darboux [1878, p. 413] by one year). Following Stieltjes [1894] we will consider, somewhat more generally, integrals of the form

$$I(f) = \int_a^b f(t)d\lambda(t), \tag{1.12}$$

where $d\lambda(t)$ is a (positive) Stieltjes measure on the finite or infinite interval $[a, b]$. We assume that $\lambda(t)$ has infinitely many points of increase, and $d\lambda(t)$ finite moments of all orders. It seems appropriate, then, in view of Christoffel's work, to call the n-point quadrature formula (1.5) for the weighted integral $I(f)$ in (1.12) a *Gauss–Christoffel quadrature formula*, if it has maximum degree of exactness $2n-1$. The weights λ_ν, as has long been customary, will be called the *Christoffel numbers* for $d\lambda$.

With the measure $d\lambda$ there is associated a unique system of (monic) *orthogonal polynomials* $\pi_k(t) = \pi_k(t; d\lambda)$,

$$\deg \pi_k = k, \qquad k = 0, 1, 2, \ldots,$$

$$\int_a^b \pi_k(t)\pi_l(t)d\lambda(t) = 0, \qquad \text{all } k \neq l.$$

They are known to satisfy a three-term recurrence relation (Christoffel [1877], Darboux [1878], Stieltjes [1884a])

$$\begin{aligned} \pi_{k+1}(t) &= (t-\alpha_k)\pi_k(t) - \beta_k\pi_{k-1}(t), \qquad k = 0, 1, 2, \ldots, \\ \pi_{-1}(t) &= 0, \quad \pi_0(t) = 1, \end{aligned} \tag{1.13}$$

where the coefficients α_k, β_k are real, and $\beta_k > 0$ for $k > 0$. (β_0 is arbitrary.) The generalized Gauss–Jacobi theory rests precisely on these orthogonal polynomials.

To begin with, there is a verbatim extension of Jacobi's argument: The quadrature rule Q_n in (1.5) has degree of exactness $d(Q_n) = n - 1 + k$ if and only if λ_ν is given by (1.9) and the node polynomial ω_n is orthogonal (with respect to $d\lambda$) to all polynomials of degree $\leqq k-1$. The last condition can be expressed equivalently in the form

$$\omega_n(t) = \pi_n(t) - c_1\pi_{n-1}(t) - \cdots - c_{n-k}\pi_k(t),$$

where c_r are arbitrary real constants. If $k = n$, we get uniquely $\omega_n(t) = \pi_n(t; d\lambda)$, i.e. *the nodes τ_ν of the n-point Gauss–Christoffel formula are the zeros of the n-th degree orthogonal polynomial π_n*. They are all real, simple, contained in (a, b), and separated by the zeros of π_{n-1} (Christoffel [1877]). Equally interesting is the case $k \geqq n-1$, which leads to "quasi-orthogonal" polynomials $\omega_n = \pi_n - c\pi_{n-1}$ with c arbitrary real. These were introduced by M. Riesz [1922/23] in connection with the moment problem, and were shown to have only real and simple zeros, at least $n-1$ of which are in (a, b). Combining this with a remark of Stieltjes [1884a, pp. 384–85], it follows that

(1.14) $$\lambda_\nu = I(l_\nu^2) > 0, \qquad \nu = 1, 2, \ldots, n,$$

whenever (1.5) has degree of exactness $\geq 2n-2$. In particular, *all Christoffel numbers are positive.* The case $k \geq n-2$ is studied in detail by Micchelli & Rivlin [1973a].

To generalize the approach of Gauss, it is convenient to introduce the three functions

$$L(z) = \int_a^b \frac{d\lambda(t)}{z-t}, \quad \rho_n(z) = \int_a^b \frac{\pi_n(t)}{z-t} d\lambda(t), \qquad z \notin [a, b],$$

$$\sigma_n(z) = \int_a^b \frac{\pi_n(z) - \pi_n(t)}{z-t} d\lambda(t),$$

which figure prominently in the work of Christoffel [1877], and have previously been used by Christoffel [1858] and Jacobi [1859] in special cases. Clearly,

$$\pi_n(z) L(z) = \sigma_n(z) + \rho_n(z).$$

Since σ_n is a polynomial of degree $n-1$, it represents the "entire part" of $\pi_n L$, while ρ_n, containing only negative powers in its power series expansion, is the "remainder". We have, in fact,

$$\rho_n(z) = \sum_{k=0}^{\infty} \frac{r_k}{z^{k+1}}, \quad r_k = \int_a^b t^k \pi_n(t) d\lambda(t),$$

which shows that $\rho_n(z) = O(z^{-n-1})$, by virtue of π_n being the orthogonal polynomial of degree n. Therefore,

(1.15) $$L(z) - \frac{\sigma_n(z)}{\pi_n(z)} = \frac{\rho_n(z)}{\pi_n(z)} = O\left(\frac{1}{z^{2n+1}}\right), \qquad z \to \infty.$$

One now defines the weights λ_ν and the nodes τ_ν of Q_n, as Gauss did previously in the case $d\lambda(t) = dt$, by means of the partial fraction decomposition of σ_n / π_n,

$$\frac{\sigma_n(z)}{\pi_n(z)} = \sum_{\nu=1}^{n} \frac{\lambda_\nu}{z - \tau_\nu} =: Q_n\left(\frac{1}{z-\cdot}\right),$$

which, incidentally, yields

(1.16) $$\lambda_\nu = \frac{\sigma_n(\tau_\nu)}{\pi_n'(\tau_\nu)}, \qquad \nu = 1, 2, \ldots, n.$$

It then follows that

$$L(z) - \frac{\sigma_n(z)}{\pi_n(z)} = I\left(\frac{1}{z-\cdot}\right) - Q_n\left(\frac{1}{z-\cdot}\right)$$

$$= R_n\left(\frac{1}{z-\cdot}\right) = \sum_{k=0}^{\infty} \frac{R_n(t^k)}{z^{k+1}},$$

which, combined with (1.15), shows that $R_n(t^k) = 0$ for $0 \le k \le 2n-1$, i.e. (1.5) is the desired Gauss–Christoffel formula. At the same time we recognize

$$\frac{\rho_n(z)}{\pi_n(z)} = \sum_{k=2n}^{\infty} \frac{R_n(t^k)}{z^{k+1}}$$

to be the *generating function of the monomial errors.*

Eq. (1.15) gives rise to another important observation. If both L and σ_n/π_n are expanded in descending powers, the two expansions must agree through the first $2n$ terms. This identifies the rational function σ_n/π_n as the *n-th convergent of the continued fraction* associated with the integral L,

$$L(z) \sim \frac{\beta_0}{z-\alpha_0-} \frac{\beta_1}{z-\alpha_1-} \frac{\beta_2}{z-\alpha_2-} \cdots. \tag{1.17}$$

The characterization of π_n as the denominator of this convergent indeed is the way orthogonal polynomials were generally viewed throughout the 19th century. (See, however, Murphy [1835].) The recurrence relation (1.13) (with $t = z$) is nothing but the recurrence relation that σ_n and π_n, as numerators and denominators of the continued fraction (1.17), must satisfy. The coefficients α_k, β_k in (1.13) are therefore the same as those in (1.17) (where $\beta_0 = \int_a^b d\lambda(t)$), and we now recognize σ_n as being a second solution of (1.13); its initial values are $\sigma_{-1} = -1$, $\sigma_0 = 0$.

Our developments up to this point already yielded several explicit formulas for the Christoffel numbers λ_ν; cf. (1.9), (1.14), (1.16). Among the many others that are known, we mention only the elegant formula

$$\lambda_\nu = \frac{1}{\sum_{k=0}^{n-1} [\pi_k^*(\tau_\nu)]^2}, \qquad \nu = 1, 2, \ldots, n,$$

due to Shohat [1929], which expresses λ_ν in terms of the ortho*normal* polynomials $\pi_k^* = h_k^{-1/2}\pi_k$, $h_k = \int_a^b \pi_k^2(t)d\lambda(t)$. In principle, as has recently been observed (Billauer [1974]), one could dispense with Christoffel numbers altogether if one writes the Gauss–Christoffel quadrature sum in the form

$$Q_n(f) = \frac{[\tau_1, \tau_1, \ldots, \tau_n](f/\pi_{n+1})}{[\tau_1, \tau_2, \ldots, \tau_n](1/\pi_{n+1})} \int_a^b d\lambda(t),$$

where $[\tau_1, \tau_2, \ldots, \tau_n]g$ denotes the $(n-1)$-st order divided difference of g, and $\pi_{n+1} = \pi_{n+1}(\cdot\,; d\lambda)$.

Still another approach to the Gauss–Christoffel formula, due to Markov [1885], is via Hermite interpolation. One replaces the integral over f by the integral over $q_{2n-1}(f;\cdot)$, the Hermite interpolation polynomial of degree $\le 2n-1$ interpolating both f and its derivative f' at the nodes τ_ν. By requiring the weights of the derivative terms $f'(\tau_\nu)$ in the quadrature sum to be all equal to zero, one again is led to choose τ_ν as the zeros of the orthogonal polynomial

$\pi_n(\cdot; d\lambda)$. Markov's derivation has the advantage of yielding, via the remainder term of Hermite interpolation, an explicit expression for the remainder R_n in (1.5), namely

$$R_n(f) = \int_a^b \pi_n^2(t)[\tau_1, \tau_1, \ldots, \tau_n, \tau_n, t]f\, d\lambda(t),$$

where $[\tau_1, \tau_1, \ldots, \tau_n, \tau_n, t]f$ denotes the $2n$-th divided difference of f formed with the nodes τ_ν (each taken twice) and t. If $f \in C^{2n}[a, b]$, then, alternatively,

$$R_n(f) = \frac{h_n}{(2n)!} f^{(2n)}(\bar{t}_n), \quad h_n = \int_a^b \pi_n^2(t) d\lambda(t), \tag{1.18}$$

where $\bar{t}_n \in (a, b)$ is an (unknown) intermediate value.

The remainder in (1.18) can be further expanded in the manner of Euler–Maclaurin, as is proposed by Bilharz [1951] for the Gauss–Legendre formula and discussed in Krylov [1959, Ch. 11, §3] for arbitrary quadrature rules.

Another form of the remainder, valid for holomorphic functions f,

$$R_n(f) = \frac{1}{2\pi i} \oint_\Gamma \frac{\rho_n(z)}{\pi_n(z)} f(z) dz, \tag{1.19}$$

where Γ is a contour encircling the interval $[a, b]$, follows from a contour integral representation of the error in polynomial interpolation, given by Darboux [1878] and Heine [1881].

The development of Gauss–Christoffel formulae, as already mentioned, began with Mehler [1864], who considered $d\lambda(t) = (1-t)^\alpha (1+t)^\beta dt$ on $[-1, 1]$ with arbitrary $\alpha > -1$, $\beta > -1$. The resulting quadrature formula is now named after Jacobi, who studied the corresponding orthogonal polynomials (Jacobi [1859]). Particularly noteworthy is the special case $\alpha = \beta = -1/2$, already discussed by Mehler, which yields as orthogonal polynomials the Chebyshev polynomials of the first kind, $\pi_n(\cos\theta) = 2^{1-n} \cos n\theta$. Its zeros τ_ν are given explicitly by $\tau_\nu = \cos((2\nu - 1)\pi/2n)$, and all weights λ_ν turn out to be equal, $\lambda_\nu = \pi/n$. Posse [1875] indeed proves that this Gauss–Chebyshev formula is the only Gauss–Christoffel formula having equal weights. Other Gauss–Christoffel formulae for which the nodes and weights can be expressed explicitly in terms of trigonometric functions are those for $d\lambda(t) = (1-t^2)^{1/2} dt$ and $d\lambda(t) = [(1-t)/(1+t)]^{1/2} dt$ on $[-1, 1]$, obtained independently by Posse [1875] and Stieltjes [1884b]. Gauss–Christoffel quadrature rules on infinite intervals appear first in Radau [1883], who considers $d\lambda(t) = e^{-t} dt$ on $[0, \infty]$, and in Gourier [1883], who considers $d\lambda(t) = e^{-t^2/2} dt$ on $[-\infty, \infty]$. The former are named after Laguerre, who earlier discussed the relevant orthogonal polynomials (Laguerre [1879]), the latter after Hermite [1864], who studied the orthogonal polynomials for $d\lambda(t) = e^{-t^2} dt$ on $[-\infty, \infty]$. (These attributions may

not be entirely justified, historically, as Laguerre polynomials were already used by Lagrange [1762–1765, pp. 534–539] and were treated in unpublished work, dated 1826, by Abel [Oeuvres 2, p. 284] and again, later, by Murphy [1835, pp. 146–148] and Chebyshev [1859b]. Likewise, Hermite polynomials were used by Laplace [1810/11] in his work on probability and studied by Chebyshev [1859b].) The more general orthogonal polynomials relative to the measure $d\lambda(t) = t^{\alpha}e^{-t}dt$ on $[0,\infty]$ occur in the work of Sohockiĭ [1873] and Sonin [1880, pp. 41–43], the corresponding Gauss–Christoffel formula in Deruyts [1886].

We remark that Gauss–Christoffel quadrature formulae can also be interpreted as quadrature rules of given degree of exactness and *minimal number of nodes* (Chakalov [1930/31]). Indeed, given an integer $d \geq 1$, any quadrature rule (1.5) of degree of exactness d, having distinct (real or complex) nodes τ_ν and (real or complex) weights λ_ν, must have more than $d/2$ nodes, i.e. $n > d/2$. If d is odd, the Gauss–Christoffel formula is the unique quadrature rule Q_n with $d(Q_n) = d$ having the minimum number $n = (d+1)/2$ of nodes. If d is even, there are infinitely many Q_n with $d(Q_n) = d$ that have the minimum number $n = (d/2)+1$ of nodes. They all can be obtained by taking as node polynomial ω_n any polynomial $\pi_n - c\pi_{n-1}$ having distinct zeros, and by defining λ_ν in the usual manner by (1.9).

2. Extensions of the Gauss–Christoffel Quadrature Formula

2.1. Gaussian quadrature with preassigned nodes

While in Gauss–Christoffel quadrature formulae there is no freedom in the choice of the nodes, all being uniquely determined by the measure $d\lambda$, there may be situations in which employment of certain preassigned nodes is highly desirable. The question then arises as to how the remaining (free) nodes, and all weights (including those for the preassigned nodes), are to be chosen in order to maximize the degree of exactness. Christoffel [1858] was the first to consider this problem and, in the case $d\lambda(t) = dt$ on $[-1, 1]$, to give a complete solution under the assumption that all preassigned nodes are *outside* the open interval $(-1, 1)$. An interesting case of preassigned nodes *inside* the interval of integration is considered only recently by Kronrod [1964 a, b] in connection with a practical implementation of the Gaussian integration scheme.

2.1.1. *Christoffel's work and related developments.* Although Christoffel, in his 1858 paper, considers only integrals with constant weight function and finite interval $[-1, 1]$, his work extends easily to general weighted integrals over a finite or half-infinite interval $[a, b]$,

$$I(f) = \int_a^b f(t)d\lambda(t), \qquad d\lambda(t) \geq 0. \tag{2.1}$$

He proposes to approximate this integral by means of a quadrature formula of the form

$$(2.2) \qquad I(f) = Q_n(f) + R_n(f), \quad Q_n(f) = \sum_{\nu=1}^{n} \lambda_\nu f(\tau_\nu) + \sum_{\lambda=1}^{l} \mu_\lambda f(u_\lambda),$$

where $l \geq 1$, and $u_1, u_2, \ldots, u_l$ are given real numbers not in the open interval (a, b). We call u_λ the *fixed nodes* of Q_n, and τ_ν the *free nodes*. All weights λ_ν and μ_λ are assumed real and freely variable. We continue to use the notation ω_n for the node polynomial of the free nodes, and let u denote the one for the fixed nodes,

$$\omega_n(t) = \prod_{\nu=1}^{n} (t - \tau_\nu), \quad u(t) = \pm \prod_{\lambda=1}^{l} (t - u_\lambda).$$

By assumption, u preserves its sign on $[a, b]$, and the plus or minus sign is taken so as to render $u(t) \geq 0$ on $[a, b]$.

Following the approach of Newton–Cotes (cf. Section 1.1), it is clearly possible to make (2.2) *interpolatory*, i.e. to achieve degree of exactness $n - 1 + l$. On the other hand, by Jacobi's argument with obvious changes (replace ω_n by $u\omega_n$ in Section 1.3), one finds that $d(Q_n) = n - 1 + l + k$, $0 \leq k \leq n$, *if and only if* Q_n *is interpolatory and* $I(u\omega_n p) = 0$ *for all* $p \in \mathbf{P}_{k-1}$. Thus, ω_n must be orthogonal to all polynomials of degree $\leq k - 1$ with respect to the measure $d\sigma(t) = u(t)d\lambda(t)$. Since $d\sigma$ is a positive measure, we are back to the situation discussed in Section 1.4. In particular, the quadrature formula (2.2) will have maximum degree of exactness $d(Q_n) = 2n - 1 + l$ precisely if $\omega_n = \pi_n(\cdot\,; ud\lambda)$ and the weights are obtained by interpolation,

$$(2.3) \qquad \lambda_\nu = I\left[\frac{u(\cdot)\omega_n(\cdot)}{u(\tau_\nu)\omega_n'(\tau_\nu)(\cdot - \tau_\nu)}\right], \quad \mu_\lambda = I\left[\frac{u(\cdot)\omega_n(\cdot)}{u'(u_\lambda)\omega_n(u_\lambda)(\cdot - u_\lambda)}\right],$$

$$\nu = 1, 2, \ldots, n, \quad \lambda = 1, 2, \ldots, l.$$

The formula (2.2), with τ_ν the zeros of $\pi_n(\cdot\,; ud\lambda)$, and λ_ν, μ_λ given by (2.3), is called the *Christoffel quadrature formula*. Christoffel proceeds to express $\pi_n(\cdot\,; ud\lambda)$ explicitly in terms of the polynomials $p_\tau(\cdot\,; d\lambda)$ orthogonal with respect to $d\lambda$,

$$(2.4) \qquad u(t)\pi_n(t; ud\lambda) = \text{const} \cdot \begin{vmatrix} p_n(t) & p_{n+1}(t) & \cdots & p_{n+l}(t) \\ p_n(u_1) & p_{n+1}(u_1) & \cdots & p_{n+l}(u_1) \\ \cdots & \cdots & \cdots & \cdots \\ p_n(u_l) & p_{n+1}(u_l) & \cdots & p_{n+l}(u_l) \end{vmatrix}.$$

This is now commonly referred to as *Christoffel's theorem*. It shows that $u\pi_n$ is a linear combination of p_n, p_{n+1}, p_{n+l}. The theorem is useful in many applications, e.g. in the asymptotic theory of orthogonal polynomials (Shohat [1928])

and in studying the qualitative behavior of the zeros and weights of π_n (Bezikovič [1937]).

It is interesting to read how Christoffel motivates his formula (Christoffel [1858, p. 74]): "... Gegenwärtige Methode gewährt demnach die Möglichkeit, bei der angenäherten Integration einer *gegebenen* Funktion alle Vortheile zu vereinigen, welche einerseits aus der Berücksichtigung des numerischen Verlaufs dieser Funktion, und andererseits aus der Anwendung der Gaußischen Methode entspringen können. Man wird nämlich jene n willkürlichen Wurzeln so wählen, daß für sie die Funktion $F(x)$ besonders einfache, oder auch solche Werthe annimmt, von denen ein großer Einfluß auf den Werth des gesuchten Integrals zu erwarten ist." [His "F" is our "f", and his "n" is our "l".] One of the "... especially simple values" of f that Christoffel had in mind, undoubtedly, was $f(u_\lambda) = 0$, in which case the corresponding term $\mu_\lambda f(u_\lambda)$ in the quadrature sum $Q_n(f)$ disappears, and the high degree of accuracy is retained with one fewer quadrature node. In the extreme case where *all* u_λ are zeros of f, one effectively gets an n-point formula with degree of exactness $2n - 1 + l$.

Christoffel's new quadrature formula, and the companion theorem of Christoffel, is but one of several important discoveries contained in Christoffel's remarkable 1858 memoir. Among the others is the discrete orthogonality property for Legendre polynomials, obtained by Christoffel independently of Chebyshev, who introduced discrete orthogonal polynomials in his least-squares approximation method (Chebyshev [1855]). From the discrete orthogonality relation Christoffel then derives the "Christoffel–Darboux" summation formula for Legendre polynomials. Equally remarkable is the fact that Christoffel was already preoccupied with the question of convergence of Gaussian quadrature rules, and with the related question of convergence of series expansions in Legendre polynomials. Evidently convinced, but unable to prove, that his summation formula holds the key to convergence, he writes in a letter to Borchardt, dated December 3, 1857, that for convergence "... scheint nun die Formel 44. [his summation formula] wie geschaffen...".

Christoffel's quadrature formula (2.2), (2.3) allows an interesting alternative interpretation, already pointed out by Christoffel [1858, p. 76] and recently rediscovered (Esser [1971a], [1972]) in a more general context (multiple fixed nodes). Observe, first of all, that

$$Q_n^*(f) = \sum_{\nu=1}^{n} \lambda_\nu f(\tau_\nu)$$

is a quadrature rule in its own right. It has the property that $Q_n^*(up) = I(up)$ for all $p \in \mathbf{P}_{2n-1}$. Let P be the interpolation operator $Pf = p_{l-1}(f; \cdot)$, where $p_{l-1}(f; \cdot)$ is the polynomial of degree $\leq l - 1$ interpolating f at the fixed nodes $u_1, u_2, \ldots, u_l$. Then, with E denoting the identity operator, the quadrature sum in (2.2) can be written equivalently in the form

$$Q_n(f) = I(Pf) + Q_n^*((E-P)f). \tag{2.5}$$

Christoffel indeed interprets the interpolation term $I(Pf)$ on the right as a "first approximation" to $I(f)$, and the second term as "the correction to be applied in order to obtain a more accurate value". More importantly, (2.5) yields a new formula for the weights μ_λ,

$$\mu_\lambda = (I - Q_n^*)\left[\frac{u(\cdot)}{u'(u_\lambda)(\cdot - u_\lambda)}\right], \qquad \lambda = 1, 2, \ldots, l, \tag{2.6}$$

which lends itself more easily for the study of convergence of Christoffel's quadrature rule (Esser [1971a], [1972]). The function on which $I - Q_n^*$ acts in (2.6) is just the elementary Lagrange polynomial l_λ for the nodes $u_1, u_2, \ldots, u_l$ (cf. Section 1.1).

Another interesting use of (2.5) is made by Krylov [1959, Ch. 9, §3], who considers $d\lambda(t) = \omega(t)dt$ in (2.1), where ω is a function that is positive at $t = b$, and changes sign exactly at the nodes $u_1, u_2, \ldots, u_l$, which are now assumed *interior* points of $[a, b]$. Thus, $I(f)$ is an integral with an oscillatory weight function. Since $I(up) = \int_a^b p(t)u(t)\omega(t)dt$, and $u(t)\omega(t) \geq 0$ on $[a, b]$, the nodes τ_ν of the quadrature rule Q_n^* are just the zeros of $\pi_n(\cdot\,; u\omega dt)$, and $u(\tau_\nu)\lambda_\nu$ the corresponding Christoffel numbers. $Q_n(f)$ in (2.5) then approximates $I(f)$ in (2.1) with degree of exactness $2n - 1 + l$. The required function $(E - P)f$ can be represented in terms of the divided difference of f as $(E - P)f = u(\cdot)[u_1, u_2, \ldots, u_l, \cdot]f$.

The special Christoffel formula for $d\lambda(t) = dt$ on $[-1, 1]$, with $l = 2$, $u_1 = -u_2 = 1$, has already been obtained by Lobatto [1852, §§207–210]. It is customary, therefore, to call (2.2), (2.3), when $l = 2$, $u_1 = a$, $u_2 = b$ (hence $[a, b]$ is finite), a *Gauss–Lobatto formula*. The same formula, together with the simpler one with $l = 1$, $u_1 = a$ or $u_1 = b$, was also discussed (for $d\lambda(t) = dt$) by Radau [1880]; the latter is commonly referred to as the *Gauss–Radau formula*. All weights of a Gauss–Radau and a Gauss–Lobatto formula are necessarily positive. Shohat [1929] has a systematic study of the Gauss–Legendre–Lobatto formula. The two Gauss–Radau formulae for the Chebyshev measure $d\lambda(t) = (1 - t^2)^{-1/2}dt$ on $[-1, 1]$, as well as the respective Gauss–Lobatto formula, can be expressed explicitly in terms of trigonometric functions (Markov [1885]). All three formulas have equal coefficients associated with all interior nodes. For the last one, this equicoefficient property is proved by Gatteschi, Monegato & Vinardi [1976] to be characteristic among Gauss–Lobatto formulae with Jacobi weight function, even if multiple fixed nodes are admitted. Gauss–Radau and Gauss–Lobatto formulae for some classical measures $d\lambda$ are reviewed extensively in Bouzitat [1952], where in particular one finds explicit constructions of these formulae for all measures with square root singularities at one or both endpoints. For other examples see Ljaščenko & Oleĭnik [1974], [1975].

There are various generalizations of Christoffel's theorem. It is a simple matter, e.g., to observe that the theorem remains valid for *multiple nodes* u_λ, if the appropriate rows in the determinant of (2.4) are replaced by rows of derivatives (Szegö [1921]). A more substantial generalization is due to Uvarov [1959], [1969], who considers $d\sigma(t) = [u(t)/v(t)]d\lambda(t)$, where

$$u(t) = \pm \prod_{\lambda=1}^{l} (t - u_\lambda), \quad v(t) = \prod_{\mu=1}^{m} (t - v_\mu)$$

are such that the measure $d\sigma$ they generate is a positive Stieltjes measure on $[a, b]$. Assuming the roots u_λ pairwise distinct, and the same for the v_μ, Uvarov establishes the *generalized Christoffel's theorem*

$$u(t)\pi_n\left(t; \frac{u}{v}d\lambda\right) =$$

$$\text{(2.7)} \qquad \text{const} \cdot \begin{vmatrix} p_{n-m}(t) & p_{n-m+1}(t) & \cdots & p_{n+l}(t) \\ p_{n-m}(u_1) & p_{n-m+1}(u_1) & \cdots & p_{n+l}(u_1) \\ \cdots & \cdots & \cdots & \cdots \\ p_{n-m}(u_l) & p_{n-m+1}(u_l) & \cdots & p_{n+l}(u_l) \\ r_{n-m}(v_1) & r_{n-m+1}(v_1) & \cdots & r_{n+l}(v_1) \\ \cdots & \cdots & \cdots & \cdots \\ r_{n-m}(v_m) & r_{n-m+1}(v_m) & \cdots & r_{n+l}(v_m) \end{vmatrix}, \qquad m \leq n,$$

and another similar theorem in the case $m > n$. Here, $p_k = p_k(\cdot\,; d\lambda)$ are the orthogonal polynomials with respect to the measure $d\lambda$, and

$$r_k(z) = \int_a^b \frac{p_k(t)}{z - t} d\lambda(t), \qquad k = 0, 1, 2, \ldots .$$

The case of confluent zeros u_λ or v_μ is handled similarly as in the classical Christoffel theorem. Interestingly, Christoffel [1877] already has an example of (2.7), namely the case $d\lambda(t) = dt$ on $[-1, 1]$ and $u(t) = 1$, $v(t) = t^2 + a^2$, but he refrains from giving any explanation. Kumar [1974a, b] and Price [1979], apparently unaware of Uvarov's result, discuss further examples. See also Szegö [1922, Kap. II], Grinšpun [1966].

2.1.2. *Kronrod's extension of quadrature rules.* Motivated by a desire to provide a practical means of estimating the error in numerical integration, Kronrod [1964a, b] initiates a study of *pairs* (Q_1, Q_2) *of quadrature rules*,

$$Q_i(f) = \sum_{\nu=1}^{n_i} \lambda_{\nu,i} f(\tau_{\nu,i}), \qquad i = 1, 2.$$

The intent here is to use the more accurate of the two, say Q_2, to estimate the error of the other, Q_1, the integral to be approximated being the weighted integral in (2.1). One defines the degree of exactness of the pair (Q_1, Q_2) by

$$d(Q_1, Q_2) = \min(d(Q_1), d(Q_2))$$

(cf. Section 1.1). The number of distinct points that are either nodes of Q_1 or nodes of Q_2 is denoted by $n(Q_1, Q_2)$. Assuming Q_1 not identical with Q_2, one has

$$n(Q_1, Q_2) \geq d(Q_1, Q_2) + 2, \qquad Q_1 \neq Q_2,$$

since otherwise $d(Q_1, Q_2) > n(Q_1, Q_2) - 2$, hence $d(Q_1) \geq n(Q_1, Q_2) - 1$ and $d(Q_2) \geq n(Q_1, Q_2) - 1$, which would imply $Q_1 = Q_2$, both Q_1 and Q_2 being interpolatory on the set of (distinct) nodes of Q_1 and Q_2.

Following Kronrod, we pose the following problem: Given an interpolatory quadrature rule Q_1 for the integral I in (2.1), find a quadrature rule $Q_2 \neq Q_1$ such that $d(Q_2)$ is as large as possible, subject to

$$d(Q_2) \geq d(Q_1), \quad n(Q_1, Q_2) = d(Q_1, Q_2) + 2 = d(Q_1) + 2. \tag{2.8}$$

In other words, we wish to maximize the degree of exactness of Q_2 under the condition that the pair (Q_1, Q_2) have maximum degree of exactness [the first condition in (2.8)] and the minimum number of nodes [the second condition in (2.8)]. We consider this optimum Q_2, if it exists, to have $n(Q_1, Q_2)$ nodes, those in Q_1 or Q_2, although some of the weights, conceivably, could be zero. We call Q_2 the *minimum node Kronrod extension* of Q_1. The same problem can be posed with $n(Q_1, Q_2)$ prescribed arbitrarily, $n(Q_1, Q_2) \geq d(Q_1) + 2$. We then call Q_2 simply a *Kronrod extension* of Q_1. Since the quadrature rule Q_2 contains among its nodes all those of Q_1, we may think of the Kronrod extension as a quadrature formula with *preassigned nodes* (those of Q_1). What differs from Christoffel's theory (cf. Section 2.1.1) is the fact that the preassigned nodes are now located *within* the interval $[a, b]$, and the corresponding node polynomial is no longer of constant sign.

Since Q_1 has n_1 nodes, and is interpolatory, $d(Q_1) \geq n_1 - 1$, and therefore $n(Q_1, Q_2) \geq n_1 + 1$. Let ω_{Q_1} denote the node polynomial (of degree n_1) of Q_1. Then the following can be shown (which generalizes slightly a result of Kronrod [1964b, Thms. 5 and 6]): *If there exists a monic polynomial ω of degree $n(Q_1, Q_2) - n_1$, orthogonal with respect to $\omega_{Q_1} d\lambda$ to all polynomials of lower degree,*

$$\int_a^b \omega(t) t^k \omega_{Q_1}(t) d\lambda(t) = 0, \qquad k = 0, 1, 2, \ldots, n(Q_1, Q_2) - n_1 - 1,$$

then there exists a unique Kronrod extension Q_2 of Q_1, having degree of exactness $d(Q_2) \geq 2n(Q_1, Q_2) - n_1 - 1$. The extension is the unique interpolatory quadrature rule that has as nodes the n_1 nodes of Q_1 and the $n(Q_1, Q_2) - n_1$ zeros of ω. [If one of the latter happens to coincide with a node of Q_1, the quadrature rule $Q_2(f)$ also involves the derivative of f at that node.] In general, there is no assurance that the nodes of Q_2 are real and contained in $[a, b]$.

Perhaps the most interesting case is the one originally considered by Kronrod: Q_1 is the n-point Gauss–Christoffel quadrature rule. In this case the minimum node Kronrod extension Q_2 has degree of exactness $d(Q_2) \geq 2(2n+1) - n - 1 = 3n + 1$, and the new nodes to be inserted in Q_1 are the zeros of the (monic) polynomial ω_{n+1} of degree $n+1$ satisfying

$$\int_a^b \omega_{n+1}(t) t^k \pi_n(t) d\lambda(t) = 0, \qquad k = 0, 1, 2, \ldots, n.$$

Here π_n is the node polynomial of Q_1, i.e. $\pi_n = \pi_n(\cdot\,; d\lambda)$. If $d\lambda(t) = (1-t^2)^{\mu-1/2}dt$ on $[-1, 1]$, $0 \leq \mu \leq 2$, the polynomial ω_{n+1} exists uniquely and has $n+1$ distinct zeros in $[-1, 1]$ which are separated by the zeros of π_n (cf. Section 3.1.2). Rabinowitz [1980] in this case (and similarly for the Kronrod extension of Gauss–Lobatto formulae) proves that $d(Q_2)$ equals, but does not exceed, $2[(3n+3)/2] - 1$, except when $\mu = 0$ or $\mu = 1$, in which cases $d(Q_2)$ is larger (if $n \geq 4$). Monegato [1978a] shows that all weights of Q_2 are positive if $0 \leq \mu \leq 1$. In the case $\mu = 1/2$ (i.e., $d\lambda(t) = dt$), tables of Q_2, accurate to 16 decimal digits, have been computed by Kronrod [1964b] for $n = 1(1)40$. Baratella [1979] has tables for Kronrod extensions of Gauss–Radau formulae.

Nothing, in principle, prevents us from repeating the process of extension and generating a sequence $Q_1, Q_2, Q_3, \ldots$ of successively extended quadrature rules. Whether indeed this is possible, and yields rules Q_i with all nodes real, has not been proved, not even in the case $d\lambda(t) = dt$.

Kronrod extensions of Gauss–Lobatto formulae, as well as repeated extensions of a low-order Gauss–Legendre rule, have been computed by Patterson [1968]. The latter are used in an automatic integration routine of Cranley & Patterson [1971] and Patterson [1973]. Piessens [1973a] uses a Kronrod pair (Q_1, Q_2) for $d\lambda(t) = dt$, $n = 10$, for similar purposes.

Particularly simple are the (minimum node) Kronrod extensions of the Gauss–Chebyshev rules, with $d\lambda(t) = (1-t^2)^{\pm 1/2}dt$ and $d\lambda(t) = [(1-t)/(1+t)]^{1/2}dt$ on $[-1, 1]$, which can be written down explicitly and extended infinitely often (Mysovskih [1964], Monegato [1976]). Weight distributions $d\lambda(t)$ with infinite support, on the other hand, seem to resist satisfactory Kronrod extension. Kahaner & Monegato [1978], e.g., prove that minimum node extensions of the n-point (generalized) Gauss–Laguerre rule, with $d\lambda(t) = t^\alpha e^{-t}dt$ on $[0, \infty]$, $-1 < \alpha \leq 1$, do not exist for $n \geq 23$ if one requires that all nodes be real and all coefficients positive. Moreover, the ordinary Gauss–Laguerre formula ($\alpha = 0$) cannot be so extended if $n > 1$, nor can the Gauss–Hermite formula, unless $n = 1, 2$, or 4, confirming earlier empirical results of Ramskiĭ [1974]. Further remarks on the difficulties of Kronrod extension can be found in Monegato [1979].

Computational methods for generating Kronrod extensions of Gauss and Lobatto rules are discussed by Patterson [1968], Piessens & Branders [1974], and Monegato [1978b].

2.2. Gaussian quadrature with multiple nodes

Gauss' principle applied to quadrature sums involving derivative values in addition to function values not only uncovers new theoretical foundations, but also yields formulae of considerable practical value in situations where derivatives are readily accessible. The breakthrough came in 1950, through the work of Turán, and has led to intensive further developments, particularly in Romanian and Italian schools of numerical analysis.

2.2.1. *The quadrature rule of Turán.* The quadrature rule for the integral I in (2.1), considered by Turán [1950], has multiple nodes τ_ν, each having the same multiplicity $r \geq 1$,

$$
\begin{aligned}
&I(f) = Q_n(f) + R_n(f), \\
&Q_n(f) = \sum_{\nu=1}^{n} [\lambda_\nu f(\tau_\nu) + \lambda'_\nu f'(\tau_\nu) + \cdots + \lambda_\nu^{(r-1)} f^{(r-1)}(\tau_\nu)].
\end{aligned} \tag{2.9}
$$

We continue to use ω_n to denote the node polynomial $\omega_n(t) = \prod_{\nu=1}^{n}(t - \tau_\nu)$. The appropriate interpolation process for the Newton–Cotes approach is now *Hermite interpolation*, which, given any set of (distinct) nodes τ_ν, will yield a degree of exactness $rn - 1$ for (2.9). We therefore call (2.9) *interpolatory* if $d(Q_n) = rn - 1$. Jacobi's theory is easily adapted (replace ω_n by ω_n^r in Section 1.3) to show that (2.9) *has degree of exactness* $d(Q_n) = rn - 1 + k$, $0 \leq k \leq n$, *if and only if* (2.9) *is interpolatory and* $I(\omega_n^r p) = 0$ *for all* $p \in \mathbf{P}_{k-1}$. Thus, it is now the r-th power of ω_n, not ω_n, which must be orthogonal to all polynomials of degree $\leq k - 1$. We call this new type of orthogonality *power orthogonality* or, specifically, *r-th power orthogonality*. Unless $k = 0$, the power r must be *odd*, since otherwise $I(\omega_n^r) > 0$, and ω_n^r could not be orthogonal to constants, let alone to $\mathbf{P}_{k-1}$. We assume, therefore, that

$$
r = 2s + 1, \qquad s \geq 0.
$$

We then have $k \leq n$, since otherwise $p = \omega_n$ would yield $I(\omega_n^{r+1}) = 0$, which is obviously impossible. We see, therefore, that (2.9) has maximum degree of exactness $d(Q_n) = (r + 1)n - 1$ precisely if

$$
\int_a^b [\omega_n(t)]^{2s+1} t^k \, d\lambda(t) = 0, \qquad k = 0, 1, \ldots, n - 1. \tag{2.10}
$$

In the special case $s = 0$ one recovers the Gauss–Christoffel formula.

Turán [1950] proves that there exists a unique polynomial $\omega_n = \pi_{n,s}(\cdot\,; d\lambda)$ for which (2.10) is satisfied. Moreover, $\pi_{n,s}$ has n distinct real zeros which are all contained in the open interval (a, b). The same is proved independently, and by entirely different methods, by Ossicini [1966]. Turán furthermore identifies $\pi_{n,s}$ as the solution of the extremal problem

$$\int_a^b [\omega(t)]^{2s+2} d\lambda(t) = \min, \tag{2.11}$$

where the minimum is sought among all monic polynomials ω of degree n. (The system (2.10) is formally obtained from (2.11) by differentiating the objective function in (2.11) with respect to all coefficients of ω.) The special case $s = 0$ in (2.11) expresses a well-known extremal property of the orthogonal polynomial $\omega = \pi_n(\cdot\,; d\lambda)$. Many essential features of the Gauss–Christoffel theory are thus seen to generalize naturally to quadrature rules (2.9) with multiple nodes. One important feature, however, the positivity of the weights, does not completely carry over. For the case $r = 3$, Turán observes that $\lambda_\nu^{(2)} > 0$, while for general r, Ossicini & Rosati [1978] prove $\lambda_\nu^{(\rho)} > 0$ whenever $\rho \geq 0$ is even. The weights $\lambda_\nu^{(\rho)}$, for ρ odd, may have either sign, in general. This is always true for symmetric integrals, but happens also in other cases, e.g., when $d\lambda(t) = e^{-t}dt$ on $[0, \infty]$ and $n \geq 3$ (cf. the tables in Stroud & Stancu [1965]).

The Chebyshev measure $d\lambda(t) = (1-t^2)^{-1/2}dt$, as always, provides for interesting examples. Bernstein [1930] indeed proves that for each $s \geq 0$ the extremal polynomial in (2.11) is precisely the Chebyshev polynomial $\omega = 2^{1-n}T_n$. Therefore, the Chebyshev points $\tau_\nu = \cos((2\nu - 1)\pi/2n)$, $\nu = 1, 2, \ldots, n$, serve as nodes for *all* Turán formulas (2.9) with $r = 1, 3, 5, \ldots$, i.e. there are weights $\lambda_\nu^{(\rho)}$ such that

$$\int_{-1}^{1} \frac{f(t)}{(1-t^2)^{1/2}} dt = \sum_{\nu=1}^{n} \sum_{\rho=0}^{2s} \lambda_\nu^{(\rho)} f^{(\rho)}\left(\cos\left(\frac{2\nu-1}{2n}\pi\right)\right) + R_n(f), \tag{2.12}$$

with $R_n(f) = 0$ for all $f \in \mathbf{P}_{2(s+1)n-1}$ (Turán [1950]). Equivalently, there exist weights $\mu_\nu^{(\rho)}$ such that

$$\int_{-\pi}^{\pi} g(t)dt = \sum_{\nu=1}^{n} \sum_{\rho=0}^{2s} \mu_\nu^{(\rho)} g^{(\rho)}\left(\frac{2\nu-1}{2n}\pi\right) + R_n(g), \tag{2.12'}$$

where $R_n(g) = 0$ for all even trigonometric polynomials g of degree $\leq 2(s+1)n - 1$. The coefficients $\mu_\nu^{(\rho)}$ in (2.12′) admit simple explicit expressions, already obtained by Kis [1957], and rediscovered repeatedly (Rosati [1968], Riess [1976]). Micchelli & Rivlin [1972] generalize (2.12) to

$$\int_{-1}^{1} \frac{f(t)}{(1-t^2)^{1/2}} dt = \frac{\pi}{n}\left\{\sum_{\nu=1}^{n} f(\tau_\nu) + \sum_{\sigma=1}^{\infty} \alpha_\sigma [\tau_1^{2\sigma}, \tau_2^{2\sigma}, \ldots, \tau_n^{2\sigma}] f'\right\},$$
$$\alpha_\sigma = (-1)^\sigma \binom{-1/2}{\sigma} \Big/ (2\sigma \cdot 4^{(n-1)\sigma}), \tag{2.13}$$

where $[\tau_1^{2\sigma}, \ldots, \tau_n^{2\sigma}]f'$ denotes the divided difference of f' formed with the nodes τ_ν, each taken with multiplicity 2σ, and f is holomorphic. If $f \in \mathbf{P}_{2(s+1)n-1}$ then (2.13) reduces to (2.12). The case $s = 1$ is easily worked out; for $s = 2$ the

formulas are given in Riess [1975]. Micchelli & Rivlin also obtain the "Lobatto analogue" of (2.13).

The remainder R_n in Turán's formula (2.9) is studied by Ionescu [1967] and Ossicini [1968], the remainder in (2.12) by Pavel [1967]. It is shown, in particular, that the Peano kernel $K_{(r+1)n}$ (cf. Section 4.2) is positive, a fact that follows also from earlier work of Chakalov [1954] concerning more general quadrature rules (those of Section 2.2.2). For finite intervals $[a, b]$, and holomorphic functions f, Ossicini & Rosati [1975] find the contour integral representation

$$R_n(f) = \frac{1}{2\pi i} \oint_\Gamma \frac{\rho_{n,s}(z)}{[\pi_{n,s}(z)]^{2s+1}} f(z)dz, \quad \rho_{n,s}(z) = \int_a^b \frac{[\pi_{n,s}(t)]^{2s+1}}{z-t} d\lambda(t),$$

where $\pi_{n,s} = \pi_{n,s}(\cdot\,; d\lambda)$ is the s-orthogonal polynomial for the measure $d\lambda$ (cf. Section 2.2.3). This reduces to a classical formula, when $s = 0$ (cf. Section 4.1.1, Eq. (4.1)).

Convergence of Turán's quadrature formula, in the case of a finite interval $[a, b]$ and $f \in C^{2s}[a, b]$, is established by Ossicini & Rosati [1978]. Roghi [1978] estimates the rate of convergence.

2.2.2. *Arbitrary multiplicities and preassigned nodes.* Chakalov [1954], [1957] and Popoviciu [1955], independently, generalize Turán's work to quadrature rules having nodes with arbitrary multiplicities, hence quadrature sums of the form

$$Q_n(f) = \sum_{\nu=1}^{n} \sum_{\rho=0}^{r_\nu - 1} \lambda_\nu^{(\rho)} f^{(\rho)}(\tau_\nu), \qquad r_\nu \geq 1. \tag{2.14}$$

It is important, now, to assume the nodes ordered, say

$$a \leq \tau_1 < \tau_2 < \cdots < \tau_n \leq b, \tag{2.15}$$

so that r_1 refers to the multiplicity of the first node, r_2 to that of the second, etc. (A permutation of the multiplicities $r_1, r_2, \ldots, r_n$, with the nodes held fixed, in general yields a new quadrature rule, a point emphasized only recently by Ghizzetti & Ossicini [1975].)

The maximum possible degree of exactness can again be determined by a simple adaptation of Jacobi's theory (cf. Section 1.3). One finds

$$\max d(Q_n) = 2\sum_{\nu=1}^{n} \left[\frac{r_\nu + 1}{2}\right] - 1, \tag{2.16}$$

so that multiplicities r_ν that are even do not contribute toward an increase in the degree of exactness. For this reason one normally assumes all r_ν to be *odd* integers,

$$r_\nu = 2s_\nu + 1.$$

The maximum degree of exactness (2.16) is then attained if and only if

$$\int_a^b \prod_{\nu=1}^{n} (t - \tau_\nu)^{2s_\nu + 1} t^k d\lambda(t) = 0, \qquad k = 0, 1, \ldots, n-1. \tag{2.17}$$

Interestingly enough, there again exists a unique set of ordered nodes τ_ν for which (2.17) is satisfied; all nodes, moreover, are contained in the open interval (a, b). The existence is proved by

Chakalov [1954], Popoviciu [1955], and Morelli & Verna [1969], existence and uniqueness (subject to (2.15)) by Ghizzetti & Ossicini [1975]. Karlin & Pinkus [1976a] prove the latter also for Stancu's generalization of (2.14) [see (2.19) below]. An extremal property analogous to (2.11) holds also for (2.14),

$$\int_a^b \prod_{\nu=1}^{n} (t-\tau_\nu)^{2s_\nu+2} d\lambda(t) = \min. \tag{2.18}$$

Once the nodes τ_ν are obtained, either from (2.17) or from (2.18), the quadrature rule (2.14) can be constructed in the usual way by Hermite interpolation. The weights $\lambda_\nu^{(\rho)}$, for which Chakalov [1954] and others have explicit expressions, normally vary in sign. Examples (with only one multiple node) in which all weights are positive, however, have been constructed; see Richert [1979].

The positivity of the Peano kernel $K_{d(Q_n)+1}$ for the error functional R_n is again secured, as is shown by Chakalov [1954]. See also Pavel [1968a].

In a series of papers, Stancu [1957a, b], [1959] generalizes the formula of Chakalov and Popoviciu in the same way as Christoffel generalized Gauss' formula. Thus, the quadrature sum is now

$$Q_n(f) = \sum_{\nu=1}^{n} \sum_{\rho=0}^{r_\nu-1} \lambda_\nu^{(\rho)} f^{(\rho)}(\tau_\nu) + \sum_{\lambda=1}^{l} \sum_{\kappa=0}^{k_\lambda-1} \mu_\lambda^{(\kappa)} f^{(\kappa)}(u_\lambda), \tag{2.19}$$

where u_λ are preassigned nodes such that

$$u(t) = \pm \prod_{\lambda=1}^{l} (t-u_\lambda)^{k_\lambda} \geq 0 \qquad \text{on } [a,b].$$

The theory of Chakalov and Popoviciu, including their discussion of the remainder term, extends readily to this more general situation, the results pertaining to (2.17) and (2.18) remaining in full force if $d\lambda(t)$ is replaced by $u(t)d\lambda(t)$ throughout. (The remainder is also discussed by Pavel [1968b].) Special cases of (2.19), in part supplemented by numerical tables, are further considered by Stancu & Stroud [1963], Stroud & Secrest [1966, Tables 13, 14], Ossicini [1968], Morelli [1967/68], and Rebolia [1973]. The case $r_1 = r_2 = \cdots = r_n = 1$, which (for $l \leq 2$) includes generalized Radau and Lobatto formulae, is of particular interest and is studied by Ionescu [1951], Gatteschi [1963/64], Ramskiĭ [1968], Esser [1972], Maskell & Sack [1974], and Porath & Wenzlaff [1976].

Further generality can be introduced by imposing constraints on the weights, e.g., that some selected weights, either λ's or μ's, be equal to zero. Maximizing the degree of exactness under such constraints is more difficult, the underlying interpolation process now being of the Birkhoff–Hermite type. The special case $r_1 = r_2 = \cdots = r_n = r$, $\lambda_\nu^{(\rho)} = 0$ for $0 \leq \rho < r-1$, and $l = 1$, $u_1 = 0$, $k_1 = 2[(r-1)/2] + 1$, with $d\lambda(t) = dt$ on $[-1, 1]$, is considered by Hammer & Wicke [1960] and leads to interesting nonclassical orthogonal polynomials, for which Struble [1960] has numerical tables. See also Patterson [1969] for a similar example. The case of simple nodes τ_ν, and zero constraints on some of the μ-weights, is treated by Micchelli & Rivlin [1973b, Thm. 4]. Lorentz & Riemenschneider [1978] and Dyn [1979] discuss the general Birkhoff–Hermite case. For generalizations to nonpolynomial quadrature rules, see Section 2.3.3.

2.2.3. *Power-orthogonal polynomials.* The condition (2.10) gives rise to a sequence of (monic) polynomials $\pi_{n,s}(\cdot\,; d\lambda)$ of degree n, $n = 0, 1, 2, \ldots$, each having the property that its $(2s+1)$-st power is orthogonal to all polynomials of lower degree. Thus, in particular,

$$\int_a^b [\pi_{n,s}(t)]^{2s+1} \pi_{k,s}(t) d\lambda(t) = 0, \qquad \text{all } k < n.$$

The polynomials $\pi_{n,s}$ are called *s-orthogonal polynomials* (Ghizzetti & Ossicini [1967], [1970, p. 74f]); they reduce to ordinary orthogonal polynomials when $s = 0$.

More generally, given a sequence of arbitrary nonnegative integers, $\sigma = \{s_1, s_2, s_3, \dots\}$, $s_i \geq 0$, the condition (2.17) defines a sequence of polynomials

$$\pi_{n,\sigma}(t) = \prod_{\nu=1}^{n} (t - \tau_\nu^{(n)}), \qquad a < \tau_1^{(n)} < \tau_2^{(n)} < \cdots < \tau_n^{(n)} < b, \qquad n = 0, 1, 2, \dots,$$

such that

$$\int_a^b \prod_{\nu=1}^{n} (t - \tau_\nu^{(n)})^{2s_\nu+1} \cdot \pi_{k,\sigma}(t) d\lambda(t) = 0, \qquad \text{all } k < n.$$

These are called *σ-orthogonal polynomials* (Ghizzetti & Ossicini [1974/75]). The s-orthogonal polynomials correspond to the special sequence $\sigma = \{s, s, s, \dots\}$.

Very little is known about such power-orthogonal polynomials. From Bernstein's observation, leading to (2.12), one knows that the Chebyshev polynomials T_n are s-orthogonal on $[-1, 1]$ for each $s = 0, 1, 2, \dots$, with respect to the Chebyshev measure $d\lambda(t) = (1 - t^2)^{-1/2} dt$. Three other measures $d\lambda$ are presently known (they all depend on s) for which the s-orthogonal polynomials can be identified (Ossicini & Rosati [1975]). Except for Rodrigues' formula, which has an analogue for σ-orthogonal polynomials (Ghizzetti & Ossicini [1974], [1974/75]), no general theory is currently available.

2.2.4. *Constructive aspects and applications.* Quadrature rules such as (2.14) are usually computed in two steps: First, one generates the nodes τ_ν, either by solving directly the associated extremal problem (2.18), or by solving a system of nonlinear equations which derives from the orthogonality condition (2.17). Then, one determines the weights $\lambda_\nu^{(\rho)}$ of the quadrature rule, usually by solving a linear system of equations expressing the interpolatory character of the rule.

The first step is clearly the more critical one, computationally. In view of the many powerful optimization techniques currently available, however, it is reasonable to expect that the minimum problem (2.18) can be solved more or less routinely, provided the objective function and its gradient can be computed accurately and efficiently. In this connection, observe that an N-point Gauss–Christoffel quadrature rule, relative to the measure $d\lambda$ (or $ud\lambda$ in the case of quadrature rules (2.19)), where $N = n + 1 + \Sigma_{\nu=1}^{n} s_\nu$, will evaluate the objective function and its gradient exactly, except for rounding errors. The required quadrature rules, on the other hand, may be obtained by the methods discussed in Section 5.

When setting up the linear system for the weights $\lambda_\nu^{(\rho)}$, some care must be exercised in the selection of the polynomial basis functions. One wants the system to be reasonably well-conditioned and sparse. A choice that offers some of these advantages is that of the Newton polynomials $1, t - \tau_1, \dots, (t - \tau_1)^{r_1}, (t - \tau_1)^{r_1}(t - \tau_2), \dots, (t - \tau_1)^{r_1}(t - \tau_2)^{r_2} \cdots (t - \tau_n)^{r_n - 1}$, which leads to a triangular system. Note that the right-hand vector of the linear system can again be computed by Gauss–Christoffel integration.

Quadrature rules of the Turán type have been applied by Micchelli & Rivlin [1972] to the calculation of Fourier coefficients, and by Kastlunger & Wanner [1972] to the construction of implicit Runge–Kutta formulas for integrating ordinary differential equations. These turn out to be "A-stable", if $r = 1$ and $r = 3$, hence are useful for stiff differential equations, but are only $A(\alpha)$-stable, for some $\alpha < \pi/2$, when $r \geq 5$.

2.3. Further miscellaneous extensions

2.3.1. *Product-type quadrature rules.* When integrating a product of two functions it may be desirable to sample the two functions independently on two different sets of points, "at their own speed" as it were. This leads naturally to product-type quadrature rules of the form

$$\int_a^b f(t) g(t) d\lambda(t) = \sum_{\mu=1}^{m} \sum_{\nu=1}^{n} f(\tau_\mu) \lambda_{\mu,\nu} g(\sigma_\nu) + R_{m,n}(f, g), \tag{2.20}$$

first introduced and studied by Boland & Duris [1971]. We assume that $d\lambda(t)$ is a positive measure, and $\{\tau_\mu\}$, $\{\sigma_\nu\}$ two sets of pairwise distinct real nodes. We denote the node polynomials by

$$\omega_m(t) = \prod_{\mu=1}^{m}(t-\tau_\mu), \quad \chi_n(t) = \prod_{\nu=1}^{n}(t-\sigma_\nu).$$

One says that the quadrature rule $Q(f, g)$ in (2.20) has *joint degree of exactness* $d(Q) = (k, l)$ if $R_{m,n}(f, g) = 0$ whenever $f \in \mathbf{P}_k$ and $g \in \mathbf{P}_l$. The formula (2.20) is called *interpolatory* if it has joint degree of exactness $d(Q) = (m-1, n-1)$. Equivalently, (2.20) is interpolatory if the quadrature sum in (2.20) is the result of integrating the product of two interpolation polynomials, one of degree $m-1$ interpolating f at the nodes τ_μ, the other of degree $n-1$ interpolating g at the nodes σ_ν. The quadrature weights are then given by

$$\lambda_{\mu\nu} = \int_a^b \frac{\omega_m(t)}{\omega_m'(\tau_\mu)(t-\tau_\mu)} \frac{\chi_n(t)}{\chi_n'(\sigma_\nu)(t-\sigma_\nu)} d\lambda(t), \qquad 1 \le \mu \le m, \quad 1 \le \nu \le n.$$

Interpolatory product-type quadrature rules are the analogues of Newton–Cotes formulas for ordinary integrals. They are uniquely determined by the nodes τ_μ and σ_ν.

Given m and n, where for definiteness we assume $m \ge n$, it is of interest to determine the domain of all possible joint degrees of exactness. The most complete answer is due to Gribble [1977]. Let C, G_1 and G_2 be disjoint subsets of the Gaussian integers defined by $C = \{(k, l): 0 \le k \le m-1, 0 \le l \le n-1\}$, $G_1 = \{(k, l): k \ge 0, l \ge n, k + l \le 2n-1\}$, $G_2 = \{(k, l): 0 \le l \le n-1, k \ge m, k + l \le 2m-1\}$ (see Figure 2.1). Then a product-type quadrature rule Q can have joint degree of exactness $d(Q) = (k, l)$ if and only if $(k, l) \in C \cup G_1 \cup G_2$. Those rules Q with

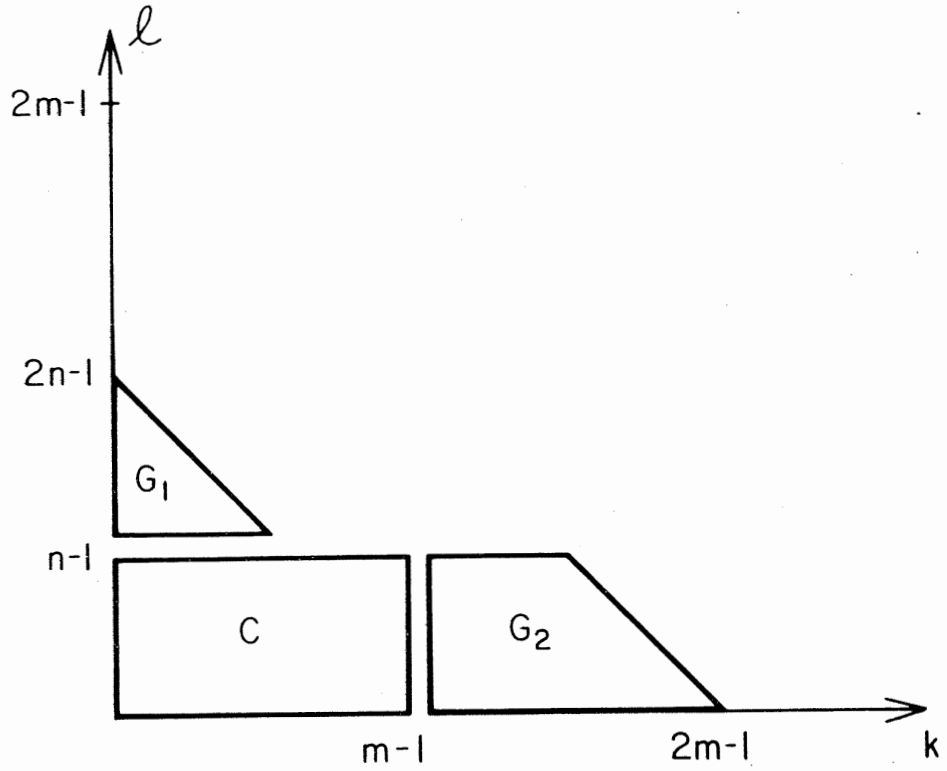


Fig. 2.1. Degrees of exactness for product-type quadrature rules

$d(Q) = (m-1, n-1)$, hence also $d(Q) \in C$, are precisely the interpolatory quadrature rules, which, as already observed, can be constructed for arbitrary nodes τ_μ, σ_ν. It is natural to consider the quadrature rules Q with $d(Q) \in G_1 \cup G_2$ as being "of Gaussian type". They exist only if some of their nodes are suitably restricted. Indeed, for a quadrature rule Q to have $d(Q) \in G_1$ it is necessary and sufficient that the nodes σ_ν be such that

$$\int_a^b \chi_n(t) t^s d\lambda(t) = 0 \qquad \text{for } s = 0, 1, \dots, k+l-n. \tag{2.21}$$

The nodes τ_μ can then be selected arbitrarily. Since $k + l - n \le n - 1$ for $(k, l) \in G_1$, the condition (2.21) can always be satisfied with pairwise distinct real nodes σ_ν, and uniquely so, if $k + l = 2n - 1$. Similarly, a quadrature rule Q with $d(Q) \in G_2$ exists if and only if

$$\int_a^b \omega_m(t) t^r d\lambda(t) = 0 \qquad \text{for } r = 0, 1, \dots, k+l-m,$$

in which case the nodes σ_ν can be chosen arbitrarily. In particular, if $m = n$, it follows (Boland [1973]) that the only quadrature rule Q for which simultaneously $d(Q) = (n-1, n)$ and $d(Q) = (n, n-1)$ is the ordinary Gauss–Christoffel quadrature rule, in which $\tau_\nu = \sigma_\nu$, $\nu = 1, 2, \dots, n$, are the Gaussian nodes, λ_ν the Christoffel numbers, and $\lambda_{\mu\nu} = 0$ for $\mu \ne \nu$.

The error term $R_{m,n}(f, g)$, and convergence results, are discussed in Boland & Duris [1971] and Boland [1972].

2.3.2. *Gaussian quadrature involving interval functionals.* All quadrature sums considered so far involve point evaluation functionals, i.e. the values of a function (and perhaps some of its derivatives) at certain well-determined points. In physical applications it is not uncommon that no such function values are accessible, but only certain averages

$$I(u_k, v_k; f) = \frac{1}{\operatorname{meas}[u_k, v_k]} \int_{u_k}^{v_k} f(t) d\mu_k(t), \quad \operatorname{meas}[u_k, v_k] = \int_{u_k}^{v_k} d\mu_k(t), \qquad k = 1, 2, \dots, n,$$

taken over small intervals $[u_k, v_k]$, $u_k < v_k$. In such cases, it is meaningful to employ quadrature rules of the type

$$\int_a^b f(t) d\lambda(t) = \sum_{k=1}^n \lambda_k I(u_k, v_k; f) + R_n(f). \tag{2.22}$$

Ordinary quadrature rules are contained in (2.22) as the limit case $u_k \to \tau_k$, $v_k \to \tau_k$, $k = 1, 2, \dots, n$.

The study of quadrature rules (2.22) involving interval functionals was initiated independently by Omladič, Pahor & Suhadolc [1975/76] and Pittnauer & Reimer [1976]. They showed, first of all, that the theory of interpolatory quadrature rules (Newton–Cotes formulae) carries over completely: Given any n nonoverlapping intervals $[u_k, v_k]$ (some possibly degenerate), one can construct a unique quadrature rule (2.22) which is exact for all polynomials of degree $\le n - 1$. In the special case $d\lambda(t) = d\mu_k(t) = dt$ on $[-1, 1]$, Pittnauer & Reimer [1976], [1979a] also extend the theory of Gaussian quadrature. In particular, they establish the following interesting extremal characterization of Gauss–Legendre formulae. For given numbers $\sigma_k > 1$, $k = 1, 2, \dots, n$, consider the function

$$G_n(u, v) = \int_{-1}^{1} \Omega_n(t; u, v) dt - \sum_{k=1}^n \sigma_k \int_{u_k}^{v_k} \Omega_n(t; u, v) dt,$$

where $u \in \mathbf{R}^n$, $v \in \mathbf{R}^n$ are points in the closed polyhedron P: $-1 \le u_1 \le v_1 \le \cdots \le u_n \le v_n \le 1$, and

$$\Omega_n(t; u, v) = \prod_{k=1}^n (t - u_k)(t - v_k).$$

Then the minimum of $G_n(u, v)$ on P is necessarily attained at an interior point of P. If (u, v) is such an interior minimum point, and $\lambda_k = (v_k - u_k)\sigma_k$, then (2.22) is a Gauss–Legendre quadrature formula, i.e. exact for all $f \in \mathbf{P}_{2n-1}$. Higher degree of exactness, when (u, v) is an interior point of P, is unattainable.

Every choice of numbers $\sigma_k > 1$ will produce a Gauss–Legendre formula of the type (2.22). Uniqueness, therefore, no longer holds, but the positivity of the weights λ_k is still guaranteed.

We remark that the property just described, when subject to the constraint $u = v$, yields the classical characterization of the Legendre polynomial $\pi_n(t; dt)$ as the monic n-th degree polynomial of minimum L_2-norm.

Peano estimates of the remainder (cf. Section 4.2), as well as a convergence theory for quadrature rules (2.22), are developed in Pittnauer & Reimer [1979b].

2.3.3. *Nonpolynomial Gaussian quadrature.* Gauss' principle can be extended in a natural way to nonpolynomial functions. Thus, given a system of linearly independent functions

$$u_1(t), u_2(t), u_3(t), \dots, \qquad a \le t \le b, \tag{2.23}$$

usually chosen to be complete in some suitable function space, the quadrature rule

$$\int_a^b f(t)d\lambda(t) = \sum_{\nu=1}^{n} \lambda_\nu f(\tau_\nu) + R_n(f) \tag{2.24}$$

is to be constructed in such a way as to integrate exactly as many successive functions in (2.23) as possible. If the first $2n$ functions are integrated exactly, one calls the rule (2.24) *Gaussian with respect to the system* (2.23).

Gaussian formulae, indeed also Gauss–Radau formulae, for the system $u_\nu(t) = t^{\alpha_\nu}$, $0 \le \alpha_1 < \alpha_2 < \cdots$, on $[0, 1]$ are already established by Stieltjes [1884c]. Trigonometric functions $u_1(t) = 1$, $u_2(t) = \cos t$, $u_3(t) = \sin t$, $u_4(t) = \cos 2t, \dots$ yield quadrature rules exact for trigonometric polynomials up to a certain degree. Assuming $d\lambda(t) = dt$ on $[0, 2\pi]$, and $0 \le \tau_1 < \tau_2 < \cdots < \tau_n < 2\pi$, Schmidt [1947] shows that the maximum possible degree is $n - 1$, and is attained precisely if $\tau_\nu = \nu(2\pi/n) - \gamma$, $0 \le \gamma < 2\pi/n$, and $\lambda_\nu = 2\pi/n$. This elevates the trapezoidal rule to a Gaussian formula for trigonometric functions. The case of an arbitrary finite interval $[a, b]$, in the context of trigonometric (and also exponential) systems, is discussed by Crout [1929/30, §16], Newbery [1969] and Knight & Newbery [1970]; integrals with arbitrary positive measures $d\lambda(t)$ on $[0, 2\pi]$ by Turečkiĭ [1959], [1960] and Keda [1961a]. Keda [1961b] and Rosati [1968] obtain trigonometric Gauss formulae with multiple nodes. For Gauss formulae with respect to spline functions, see Schoenberg [1958] and Micchelli & Pinkus [1977]. Harris & Evans [1977/78] have Gauss formulae for systems (2.23) that include algebraic powers together with functions exhibiting endpoint singularities.

Nonpolynomial Gaussian formulae can sometimes be obtained via ordinary Gaussian formulae through suitable transformations. Thus, for example, the n-point formula for $d\lambda(t) = t^\alpha e^{-t}dt$ on $[0, \infty]$, Gaussian with respect to the system $u_r(t) = (t+1)^{-r}$, $r = 0, 1, 2, \dots$ (Krylov, Korolev & Skoblja [1959], Pal'cev & Skoblja [1965]), is simply related to the n-point Gauss–Christoffel formula with measure $(t+1)^{-2n}d\lambda(t)$ on $[0, \infty]$. A similar example, involving Fourier transforms, is discussed in Kruglikova & Krylov [1961]. See also Stroud & Secrest [1966, §3.2].

Ghizzetti [1954/55], inspired by work of Radon on the remainder term (cf. Section 4.2), constructs a very general class of Gauss formulae (2.24) which are exact for all solutions of a linear homogeneous differential equation $Lf = 0$ of order $2n$. The existence of such formulae depends on the homogeneous n-point boundary value problem

$$Ly = 0, \quad y(\tau_\nu) = 0, \qquad \nu = 1, 2, \dots, n. \tag{2.25}$$

If (2.25) has exactly q linearly independent eigensolutions $y_r(t)$, $n \le q \le 2n-1$, Gauss formulae exist if and only if

$$\int_a^b y_r(t)\,d\lambda(t) = 0, \qquad r = 1, 2, \dots, q, \tag{2.26}$$

and then, in fact, ∞^{q-n} many. The classical case corresponds to $L = D^{2n}$, $D = d/dt$, where (2.25) has exactly n linearly independent solutions $y_r(t) = t^{r-1}\prod_{\nu=1}^{n}(t-\tau_\nu)$, $r = 1, 2, \dots, n$, and (2.26) expresses the usual orthogonality criterion for $\omega_n(t) = \prod_{\nu=1}^{n}(t-\tau_\nu)$. Since $q = n$ in this case, the formula is unique.

Gauss formulae for harmonic functions have been proposed in connection with the Dirichlet problem for Laplace's equation (Stroud [1974]). If D is a bounded, simply connected two-dimensional open domain, with rectifiable boundary ∂D, the solution of $\Delta^2 u = 0$ in D, with u prescribed on ∂D, has the known representation $u(P) = -\int_{\partial D}(\partial G/\partial n)(P, Q)u(Q)ds$, where G is the Green's function of D and $\partial G/\partial n$ its normal derivative (known to be nonpositive). Treating $-\partial G/\partial n$ as a weight function it is natural to seek an approximation of the form $u(P) \approx \sum_{\nu=1}^{n} g_\nu u(Q_\nu)$, where $g_\nu \in \mathbf{R}$ and $Q_\nu \in \partial D$ depend on P, and try to make the formula exact for harmonic polynomials of as high a degree as possible. Since there are $2n$ free parameters and $2n-1$ linearly independent harmonic polynomials of degree $\le n-1$, one expects that one parameter, say Q_n, can be selected arbitrarily on ∂D and all others determined such that the formula has harmonic degree of exactness $n-1$. Barrow & Stroud [1976] indeed show that this is possible by proving the existence of at least one Gaussian formula of harmonic degree $n-1$. Their proof is based on degree theory for mappings and homotopy arguments. Numerical procedures for computing such formulae are discussed in Stroud [1974]. Johnson & Riess [1979] construct formulas for circular regions. Similar ideas are pursued in Barrow [1976], [1977] in connection with the heat equation and other parabolic equations.

A generalization in another direction is due to Engels [1972], [1973], who extends Markov's derivation of Gaussian quadrature rules (cf. Section 1.4) in the sense that the underlying Hermite interpolation operator, though still linear, need no longer be polynomial. It turns out that a number of known quadrature rules, e.g. the optimal quadrature rule of Wilf [1964] and more general optimal quadrature rules (Engels [1977]), become Gaussian in this generalized sense. A further extension of this theory to quadrature rules with prescribed (simple or double) nodes is given in Engels [1974].

The existence of a Gaussian quadrature rule (2.24) with respect to the system (2.23) is always guaranteed if the first $2n$ functions of this system form a Chebyshev system on $[a, b]$. In the language of moment spaces, the Gauss formula corresponds to the unique *lower principal representation* of the measure $d\lambda(t)$ (see, e.g., Karlin & Studden [1966, §3]). Gaussian quadrature rules with multiple nodes, based on extended Chebyshev systems, are established in Karlin & Pinkus [1976a, b] and Barrow [1978] (cf. also Section 2.2.2).

3. Extension of Integrals Accessible to Gauss–Christoffel Quadrature

3.1. Nonpositive integrals

The positivity $d\lambda(t) \ge 0$ of the measure of integration is a sufficient, but by no means a necessary, condition for the existence and uniqueness of the orthogonal polynomial $\pi_n(\cdot\,; d\lambda)$, and hence for the unique existence of the n-point Gauss–Christoffel quadrature formula (cf. Section 1.4). Viewing the orthogonality conditions $\int_a^b \pi_n(t)t^k d\lambda(t) = 0$, $k = 0, 1, \dots, n-1$, as a system of linear algebraic equations for the coefficients of π_n, one finds, indeed, that a necessary and sufficient condition is merely

$$(3.1) \qquad \Delta_n = \det \begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_{n-1} \\ \mu_1 & \mu_2 & \cdots & \mu_n \\ \cdot & \cdot & \cdots & \cdot \\ \mu_{n-1} & \mu_n & \cdots & \mu_{2n-2} \end{bmatrix} \neq 0,$$

$$\mu_r = \int_a^b t^r d\lambda(t), \qquad r = 0, 1, 2, \ldots .$$

If also $\Delta_{n+1} \neq 0$, then $\int_a^b \pi_n^2(t) d\lambda(t) \neq 0$, and the degree of exactness of the Gauss–Christoffel formula cannot exceed $2n-1$.

If $d\lambda(t) \geq 0$ then (3.1) is certainly true, even in the strengthened form $\Delta_n > 0$, all $n \geq 1$, as is known from the theory of the moment problem (Wall [1948, p. 325]). If $d\lambda$ is an arbitrary measure, the condition (3.1) may still be valid, but some of the familiar properties of orthogonal polynomials may cease to hold. Thus, the zeros of π_n need no longer be real, let alone contained in (a, b), and the Christoffel numbers need no longer be positive. Concerning the latter, all one can say is that (for real-valued $d\lambda$ and real nodes) the number of positive [negative] Christoffel numbers equals the number of positive [negative] eigenvalues of the Hankel matrix in (3.1) (Stroud [1963]).

While there is some general theory concerning orthogonal polynomials with sign-variable weight functions (Struble [1963], Monegato [1980]), we will consider here only a few examples of nonpositive (including complex-valued) measures $d\lambda$ that are of interest in applications.

3.1.1. *Odd and even weight functions on symmetric intervals.* Let $\omega(t)$ be an odd function on a symmetric interval $[-a, a]$, $a > 0$, and $d\lambda(t) = \omega(t)dt$. Assume further that $n = 2m$ is even. Then the determinant in (3.1) has a checkerboard pattern of zero and nonzero elements, from which it follows, by Laplace expansion, that

$$(3.2) \qquad \Delta_n = (-1)^n \det \begin{bmatrix} \mu_1 & \mu_3 & \cdots & \mu_{n-1} \\ \mu_3 & \mu_5 & \cdots & \mu_{n+1} \\ \cdot & \cdot & \cdots & \cdot \\ \mu_{n-1} & \mu_{n+1} & \cdots & \mu_{2n-3} \end{bmatrix}^2 .$$

There is, therefore, a unique n-point Gauss–Christoffel quadrature formula if n is even and the determinant in (3.2) is different from zero. The latter is certainly true if ω is nonnegative on $[0, a]$, since then

$$\mu_{2r-1} = \int_{-a}^{a} t^{2r-1}\omega(t)dt = \int_0^{a^2} t^{r-1}\omega(\sqrt{t})dt, \qquad r = 1, 2, 3, \ldots,$$

are moments of the nonnegative measure $d\sigma(t) = \omega(\sqrt{t})dt$ on $[0, a^2]$. The desired Gauss–Christoffel formula indeed can be constructed in terms of the m-point Gauss–Christoffel formula for $d\sigma$ (Radau [1880, p. 317f], Piessens

[1970a]). For a similar construction in the case of multiple nodes see also Levin [1974]. If n is odd, the Gauss–Christoffel formula does not exist, as is already observed by Christoffel [1877] in the special case $d\lambda(t) = tdt$ on $[-1, 1]$.

Among examples of odd weight functions that have received attention are $d\lambda(t) = t^{2s+1}dt$ on $[-1, 1]$ (Rothmann [1961]) and $d\lambda(t) = \sin t\, dt$ on $[-\pi, \pi]$ (Piessens [1970a]). Another interesting example, $d\lambda(t) = \ln((1+t)/(1-t))dt$ on $[-1, 1]$, arises in the evaluation of a certain two-dimensional Cauchy principal value integral describing the aerodynamical load on a lifting body (Song [1969]). For this example, Piessens, Chawla & Jayarajan [1976] have numerical tables.

If $\omega(t)$ is even, but not necessarily of constant sign, and $n = 2m + 1$ is odd, then a unique n-point Gauss–Christoffel formula exists if

$$\det \begin{bmatrix} \mu_0 & \mu_2 & \cdots & \mu_{n-1} \\ \mu_2 & \mu_4 & \cdots & \mu_{n+1} \\ \cdots & \cdots & \cdots & \cdots \\ \mu_{n-1} & \mu_{n+1} & \cdots & \mu_{2n-2} \end{bmatrix} \cdot \det \begin{bmatrix} \mu_2 & \mu_4 & \cdots & \mu_{n-1} \\ \mu_4 & \mu_6 & \cdots & \mu_{n+1} \\ \cdots & \cdots & \cdots & \cdots \\ \mu_{n-1} & \mu_{n+1} & \cdots & \mu_{2n-4} \end{bmatrix} \neq 0.$$

It can be constructed in terms of the m-point Gauss–Radau formula for the weight function $\omega(\sqrt{t})/\sqrt{t}$ on $[0, a^2]$ or, equivalently, the m-point Gauss–Christoffel formula for $d\sigma(t) = \sqrt{t}\omega(\sqrt{t})dt$ on $[0, a^2]$ (Piessens [1972a]). As always, in such cases, there is no assurance that all nodes are real. An example of interest in Fourier analysis is $d\lambda(t) = \cos t\, dt$ on $[-\pi, \pi]$ (Piessens [1972a]).

3.1.2. *Oscillatory weight functions.* An interesting example of an oscillating weight function is $d\lambda(t) = \pi_m(t; d\sigma)d\sigma(t)$ for some positive measure $d\sigma$ on $[a, b]$. Here, $\mu_r = \int_a^b t^r \pi_m(t)d\sigma(t) = 0$ if $r < m$, and $\mu_m > 0$, so that (3.1) cannot hold unless $n \geq m + 1$. The case $n = m + 1$, already considered by Stieltjes in his last letter to Hermite (Baillaud & Bourget [1905, Vol. II, p. 439]), is of particular interest in connection with Kronrod's extension of Gauss–Christoffel quadrature rules (cf. Section 2.1.2). In this case,

$$\Delta_n = \det \begin{bmatrix} 0 & 0 & \cdots & 0 & \mu_m \\ 0 & 0 & \cdots & \mu_m & \mu_{m+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mu_m & \mu_{m+1} & \cdots & \mu_{2m-1} & \mu_{2m} \end{bmatrix} = \mu_m^n > 0 \quad (n = m + 1),$$

showing that $\pi_n(\cdot\,; d\lambda)$ exists uniquely. Stieltjes conjectures that the zeros of $\pi_n(\cdot\,; d\lambda)$ are all real, simple, contained in (a, b), and separated by the zeros of $\pi_m(\cdot\,; d\sigma)$. Unfortunately, the conjecture is not true in this generality, but has been proved by Szegö [1935] for ultraspherical polynomials, $d\sigma(t) = (1-t^2)^{\mu-1/2}dt$, with $0 < \mu \leq 2$. The special case $\mu = 0$ (of the Chebyshev polynomial $\pi_m(\cdot\,; d\sigma) = 2^{1-m}T_m$) yields $\pi_n(t; d\lambda) = 2^{1-m}(t^2 - 1)U_{m-1}(t)$, where

U_{m-1} is the Chebyshev polynomial of the second kind. The corresponding Gauss–Christoffel quadrature formula, interestingly enough, has degree of exactness $3m-1$ if $m>1$, not $2m+1$, as one might expect (Micchelli & Rivlin [1972], Riess & Johnson [1974]).

3.1.3. *Complex-valued weight functions.* Gauss–Christoffel quadrature rules with a complex weight function were first introduced by Salzer [1955], [1961] in connection with the inversion of the Laplace transform. The integral of interest here is the Bromwich integral

$$I(f)=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} e^{\zeta}\zeta^{-s}f(\zeta)d\zeta, \qquad s>0,$$

where f is assumed holomorphic in a half-plane containing the contour $\operatorname{Re}\zeta=c$, and bounded as $\zeta\to\infty$ in $|\arg\zeta|<\pi/2$. Salzer [1955], in the case $s=1$, and Skoblja [1961], Krylov & Skoblja [1961], Wellekens [1970], Piessens [1971a, c], in the case of general $s>0$, approximate $I(f)$ by a (complex) quadrature sum

$$Q_n(f)=\sum_{\nu=1}^{n} c_\nu f(\zeta_\nu) \tag{3.3}$$

which is Gaussian in the sense that $Q_n(f)=I(f)$ whenever f is a polynomial of degree $\leq 2n-1$ in $1/\zeta$. This calls for polynomials π_r in the variable $1/\zeta$ satisfying the orthogonality condition

$$\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} e^{\zeta}\zeta^{-s}\pi_k(1/\zeta)\pi_l(1/\zeta)d\zeta=0, \qquad k\neq l.$$

Such polynomials $\pi_r=\pi_{r,s}(z)$ exist uniquely. Indeed, $\Delta_n\neq 0$ for all $n\geq 1$ (Krylov & Skoblja [1974, p. 94ff]), where Δ_n is the determinant in (3.1) formed with the moments

$$\mu_r=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} e^{\zeta}\zeta^{-s-r}d\zeta=\frac{1}{\Gamma(s+r)}, \qquad r=0,1,2,\ldots .$$

It turns out that $\pi_{n,s}(z)=y_n(z;s,-1)$, where $y_n(z;a,b)$ is the generalized Bessel polynomial of Krall & Frink [1949]. (Bessel polynomials have a long history, and many interesting applications; see Grosswald [1978] for a recent exposition.) The nodes ζ_ν in (3.3) — the reciprocals of the zeros of $\pi_{n,s}(z)$ — are all contained in the right half-plane, when $s\geq 1$ (Wimp [1965]). Extensive numerical tables (Krylov & Skoblja [1968], Piessens [1969a]) in fact suggest that $\operatorname{Re}\zeta_\nu>0$ even for $s>0$. This has been proved by Martinez [1977]. More precise results concerning the zeros of $\pi_{n,s}$ can be found in de Bruin, Saff & Varga [to appear].

The convergence $Q_n(f)\to I(f)$ as $n\to\infty$ is discussed in Luke [1969, p. 254].

The quadrature rule (3.3) may be constructed by the method of Golub & Welsch (cf. Section 5.1), since the three-term recurrence relation for Bessel polynomials is known explicitly. For a discussion of this, see Luvison [1974] and Piessens [1975]. Piessens [1973b] has a Fortran program for generating Q_n, which uses the Newton–Raphson method.

Instead of applying (3.3) directly to f, Salzer [1976] proposes to apply Q_n to a Lagrange or Hermite interpolation polynomial of degree $2n-1$ based on interpolation points on the real line. This obviates the need of evaluating f in the complex plane and still often produces satisfactory results (Pexton [1976]).

Kronrod extensions of Q_n in (3.3) (cf. Section 2.1.2) are discussed by Piessens [1969b], [1971b], who also constructs Radau type formulas with the prescribed point at infinity.

Gauss–Christoffel quadrature rules with other complex weight functions, in particular Jacobi weight functions $(1-t)^{\alpha}(1+t)^{\beta}$ with complex parameters α, β, satisfying $\operatorname{Re}\alpha > -1$, $\operatorname{Re}\beta > -1$, are used in atomic scattering theory by Nuttal & Wherry [1978], and in elasticity theory by Theocaris & Ioakimidis [1977]. Jacobi measures in which α, β are no longer subject to the restriction $\operatorname{Re}\alpha > -1$, $\operatorname{Re}\beta > -1$, and correspondingly the integral is to be interpreted as an appropriate loop integral, are discussed by Maskell & Sack [1974].

3.2. Cauchy principal value integrals

Quadrature rules can be adapted to deal with Cauchy-type singular integrals extended over segments of the real line, or over the circle, or over more general curves in the complex plane. We consider here only Cauchy principal value integrals of the form

$$I(f)(x) = \int\!\!\!\!\!\!-_a^b \frac{f(t)}{x-t}\, d\lambda(t), \qquad x \in (a,b), \tag{3.4}$$

where $[a,b]$ is a finite or infinite interval and $d\lambda(t) = \omega(t)dt$ a measure of integration that admits Gauss–Christoffel quadrature formulae and is such that the integral in (3.4) is meaningful. (Hölder continuity of f on $[a,b]$ usually suffices.) Singular integrals over the circle, which give rise to principal value integrals $\int_0^{2\pi} f(t)\cot((x-t)/2)dt$ with Hilbert kernel (and 2π-periodic functions f) are best treated by trigonometric interpolation at equally spaced points. For this, see Gaier [1964, Ch. 2, §2], Korneĭčuk [1964], Gabdulhaev [1976]. Rabinowitz [1978] has a survey of numerical methods for evaluating Cauchy principal value integrals.

We distinguish between two types of quadrature rules for (3.4). In the first type, the parameter x enters as a node,

$$Q_n(f)(x) = c_0(x)f(x) + \sum_{\nu=1}^{n} c_\nu(x) f(\tau_\nu); \tag{3.5}$$

in the other, it does not,

$$Q_n^*(f)(x) = \sum_{\nu=1}^{n} c_\nu^*(x) f(\tau_\nu). \tag{3.5*}$$

All nodes τ_ν are assumed independent of x. We will call (3.5*) a *quadrature rule in the strict sense*, and (3.5) a *modified quadrature rule*. The two quadrature rules have essentially different character: (3.5) can be made "Gaussian", i.e. of degree of exactness $2n$, whereas (3.5*) cannot. The degree of exactness of (3.5*), indeed, cannot exceed $n-1$ (Sanikidze [1970a]), since otherwise $I(f)(x) \equiv 0$ when $f(t) = \Sigma_{\nu=1}^{n}(t-\tau_\nu)$, which contradicts well-known inversion formulas for Cauchy singular integrals (Gahov [1958, §42.3], Mushelišvili [1946, §86]).

3.2.1. *Modified Gauss–Christoffel quadrature formulae.* Let $\{\pi_k\}$ denote the (monic) orthogonal polynomials belonging to $d\lambda$, and let

$$G_n(f) = \sum_{\nu=1}^{n} \lambda_\nu f(\tau_\nu) \tag{3.6}$$

denote the n-point Gauss–Christoffel quadrature rule for the measure $d\lambda$. In analogy to the Gauss–Christoffel theory (cf. Section 1.4) we define

$$L(x) = \int_a^b \frac{d\lambda(t)}{x-t}, \qquad \rho_n(x) = \int_a^b \frac{\pi_n(t)}{x-t}\, d\lambda(t),$$

$$\sigma_n(x) = \int_a^b \frac{\pi_n(x) - \pi_n(t)}{x-t}\, d\lambda(t).$$

Clearly,

$$\pi_n(x) L(x) = \sigma_n(x) + \rho_n(x),$$

and, the integrand of σ_n being a polynomial of degree $\leq n-1$ in the variable t,

$$\sigma_n(x) = G_n\left[\frac{\pi_n(x) - \pi_n(\cdot)}{x-\cdot}\right] = \pi_n(x) G_n\left[\frac{1}{x-\cdot}\right].$$

Consequently,

$$L(x) - G_n\left[\frac{1}{x-\cdot}\right] = \frac{\rho_n(x)}{\pi_n(x)}. \tag{3.7}$$

Now in order to approximate the integral $I(f)(x)$ in (3.4), we write

$$I(f)(x) = f(x) \int_a^b \frac{d\lambda(t)}{x-t} - \int_a^b \frac{f(x)-f(t)}{x-t}\, d\lambda(t), \tag{3.8}$$

and observe that the second integral is integrated exactly by the rule G_n whenever $f \in \mathbf{P}_{2n}$. Therefore,

$$I(f)(x) = f(x)\left\{L(x) - G_n\left[\frac{1}{x-\cdot}\right]\right\} + G_n\left[\frac{f(\cdot)}{x-\cdot}\right] + R_n(f),$$

or, by virtue of (3.7),

$$I(f)(x) = \frac{\rho_n(x)}{\pi_n(x)} f(x) + \sum_{\nu=1}^{n} \lambda_\nu \frac{f(\tau_\nu)}{x-\tau_\nu} + R_n(f). \tag{3.9}$$

Here $R_n(f) = 0$ for all $f \in \mathbf{P}_{2n}$. We call (3.9) — a quadrature rule of the type (3.5) — the *modified Gauss–Christoffel quadrature formula* for $I(f)(x)$.

We remark that (3.9) is valid for any interpolatory quadrature rule G_n, if π_n is understood to be the node polynomial of G_n. The degree of accuracy, of course, will be correspondingly smaller. In particular, we may construct modified versions of the Gauss–Radau, Gauss–Lobatto, etc., formulae for $I(f)(x)$. A simple limit argument will show that, for any quadrature rule G_n,

$$\lambda_\nu = -\frac{\rho_n(\tau_\nu)}{\pi_n'(\tau_\nu)}, \qquad \nu = 1, 2, \ldots, n. \tag{3.10}$$

If f is holomorphic in a neighborhood of the interval $[a, b]$, the formula (3.9) can also be obtained by applying the residue theorem to the function $f(\zeta)\dot{\pi}_n(t)/[(x-\zeta)(\zeta-t)\pi_n(\zeta)]$ and subsequent integration over the variable t. This yields the useful contour integral representation of the remainder,

$$R_n(f)(x) = \frac{1}{2\pi i} \oint_\Gamma \frac{\rho_n(z)}{(x-z)\pi_n(z)} f(z)\,dz, \tag{3.11}$$

where Γ is a contour encircling the interval $[a, b]$.

Particularly noteworthy is the special case in which x is a root of $\rho_n(x) = 0$. Then (3.9) becomes

$$I(f)(x) = \sum_{\nu=1}^{n} \lambda_\nu \frac{f(\tau_\nu)}{x-\tau_\nu} + R_n(f) \qquad (\rho_n(x) = 0), \tag{3.9°}$$

and we get a formula that looks like what would have been obtained had we simply applied G_n to the integral in (3.4), treating the principal value integral as if it were an ordinary integral. If G_n is a Gauss–Christoffel formula, then again $R_n(f) = 0$ for $f \in \mathbf{P}_{2n}$. Korneĭčuk [1964] appears to be the first who noted the simple and elegant formula (3.9°). He also observes that between any two zeros of π_n there is at least one zero of ρ_n, if all $\lambda_\nu > 0$. (This was already noted by Stieltjes [1883] through an examination of the behavior of $\rho_n(x)/\pi_n(x)$ on (a, b) and taking note of (3.10).) Inevitably, the formula (3.9°) has been rediscovered many times (see, e.g., Lebedev & Baburin [1965], Delves [1967/68], Piessens [1970c], Stark [1971], Erdogan & Gupta [1971/72], Krenk [1975/76]).

As x approaches a node τ_ν, the quadrature sum in (3.9) tends to a finite value, even though one term tends to $+\infty$ and another to $-\infty$. The limit, in fact, is

$$(3.12)\quad I(f)(\tau_\nu) = \left[\frac{\rho_n'(\tau_\nu)}{\pi_n'(\tau_\nu)} + \frac{1}{2}\lambda_\nu \frac{\pi_n''(\tau_\nu)}{\pi_n'(\tau_\nu)}\right] f(\tau_\nu) + G_n^\nu\left[\frac{f(\cdot)}{\tau_\nu - \cdot}\right] - \lambda_\nu f'(\tau_\nu) + R_n(f),$$

where

$$G_n^\nu(f) = \sum_{\substack{\mu=1 \\ \mu \neq \nu}}^{n} \lambda_\mu f(\tau_\mu),$$

or, equivalently,

$$(3.12')\quad I(f)(\tau_\nu) = \left\{\rho_0(\tau_\nu) - G_n^\nu\left[\frac{1}{\tau_\nu - \cdot}\right]\right\} f(\tau_\nu) + G_n^\nu\left[\frac{f(\cdot)}{\tau_\nu - \cdot}\right] - \lambda_\nu f'(\tau_\nu) + R_n(f).$$

Although the limits (3.12), (3.12′) are well-determined, the evaluation of $I(f)(x)$ in (3.9), when x is close to one of the nodes τ_ν, is subject to severe cancellation errors. To avoid them, one must reorganize the computation in a different way, as will be discussed in Section 3.2.3.

When $[a, b]$ is a finite interval, say $[-1, 1]$, and $d\lambda(t) = dt$, the integral (3.4) can always be transformed into the form

$$\int_{-1}^{1} \frac{f(t)}{t}\, dt,$$

with a new f, for example by a linear fractional transformation. Using as base rule (3.6) the Gauss–Legendre formula, for which $\rho_n(x) = (1/2)Q_n(x)$ is the Legendre function of the second kind, one finds $\rho_n(0) = 0$ if n is even, so that (3.9°) becomes applicable with $x = 0$. This gives (Price [1960], Lebedev & Baburin [1965], Piessens [1970c])

$$\int_{-1}^{1} \frac{f(t)}{t}\, dt = \sum_{\nu=1}^{n} \frac{\lambda_\nu}{\tau_\nu} f(\tau_\nu) + R_n(f), \qquad n \text{ even},$$

which is exact for $f \in \mathbf{P}_{2n}$. An analogous formula holds if $d\lambda(t) = \omega(t)dt$ is an even measure on a symmetric interval.

If n is odd, no such Gaussian formula exists (cf. Section 3.1.1). However, in this case π_n vanishes for $\tau_\nu = 0$, so that (3.12) becomes applicable (Hunter [1972]),

$$\int_{-1}^{1} \frac{f(t)}{t}\, dt = \lambda_\nu f'(0) + \sum_{\substack{\mu=1 \\ \mu \neq \nu}}^{n} \frac{\lambda_\mu}{\tau_\mu} f(\tau_\mu) + R_n(f), \qquad n \text{ odd}, \quad \tau_\nu = 0.$$

This formula, too, is exact for $f \in \mathbf{P}_{2n}$.

The quadrature rule (3.9) can be generalized to incorporate poles, either on or outside of $[a, b]$, in addition to, or in place of the pole at x, with an appropriate extension of the remainder formula in (3.11); see Hunter [1972], Chawla & Ramakrishnan [1974], Chawla & Jayarajan [1975], Ioakimidis &

Theocaris [1977b], Lether [1977]. Remainder expressions of the Markov type (see Section 1.4) have been obtained in the special case (3.9°) by Železnova, Korneĭčuk & Markov [1965], and in the general case (3.9) [and (3.13) below] by Elliott & Paget [1979].

For the Jacobi measure $d\lambda(t) = (1-t)^\alpha(1+t)^\beta dt$, $\alpha > -1$, $\beta > -1$, Tsamasphyros & Theocaris [1977] claim convergence of (3.9) for functions f which are Hölder continuous with exponent μ, $0 < \mu \leq 1$. Convergence in the case of an arbitrary weight function on a compact interval $[a, b]$ is proved for functions $f \in C^1[a, b]$ by Elliott & Paget [1979] and discussed for Hölder continuous functions by Elliott [1979].

3.2.2. *Gauss–Christoffel quadrature formulae in the strict sense.* An alternative use of the quadrature formula (3.6) in (3.8) can be made as follows: Use (3.6) to approximate the second integral on the right of (3.8) and, at the same time, approximate the factor $f(x)$ multiplying the first integral by the interpolation polynomial of degree $n-1$ based on the nodes τ_ν of the quadrature rule (3.6). The result is a quadrature formula for $I(f)(x)$ of the type (3.5*), namely

$$I(f)(x) = \sum_{\nu=1}^{n} \left(\frac{\rho_n(x)}{\pi_n'(\tau_\nu)(x-\tau_\nu)} + \frac{\lambda_\nu}{x-\tau_\nu} \right) f(\tau_\nu) + R_n(f), \tag{3.13}$$

or, equivalently, by virtue of (3.10),

$$I(f)(x) = \sum_{\nu=1}^{n} \frac{\rho_n(x) - \rho_n(\tau_\nu)}{\pi_n'(\tau_\nu)(x-\tau_\nu)} f(\tau_\nu) + R_n(f). \tag{3.13'}$$

As pointed out earlier, this formula has degree of exactness at most equal to $n-1$, unless x is a zero of $\rho_n(x)$, in which case (3.13) reduces to (3.9°) and has degree of exactness $d(G_n)+1$. The formula (3.13) can also be obtained more directly by replacing f in (3.4) by the polynomial of degree $\leq n-1$ interpolating f at the nodes τ_ν of (3.6). Korneĭčuk [1964], taking for G_n the Gauss–Christoffel formula, appears to be the first to obtain (3.13). The special case of Jacobi weight functions is considered by Sanikidze [1970a] and Šeško [1976], and interpolatory quadrature rules based on Chebyshev points of the first and second kind, with $d\lambda(t) = dt$, are used by Sanikidze [1968], [1970c], [1970d], Chawla & Jayarajan [1975], Šeško [1976] and Chawla & Kumar [1978]. Sanikidze [1970b] also discusses interpolatory formulae based on the zeros of two consecutive orthogonal polynomials. Many convergence criteria and error estimates can be found in the work of Sanikidze. Paget & Elliott [1972] also have error estimates based on contour integration. Perhaps the most remarkable convergence results are due to Elliott & Paget [1975], [1976a] and Šeško [1976], who, independently, in the case of the Gauss–Jacobi formula for $d\lambda(t) = (1-t)^\alpha(1+t)^\beta dt$, $\alpha > -1$, $\beta > -1$, prove convergence of (3.13) for all functions f that are Hölder continuous on $[-1, 1]$ with exponent μ, $0 < \mu \leq 1$.

Šeško [1976] indeed proves uniform convergence if $\alpha > 0$, $\beta > 0$, not only for the Gauss–Jacobi formula, but also for the interpolatory formula based on Chebyshev points. Analogous results for $d\lambda(t) = (1-t)^{\alpha}(1+t)^{\beta}\ln((1-t)/(1+t))dt$ can be found in Šeško & Jakimenko [1980]. Sanikidze [1972] has similar convergence results for the Kronrod extension (cf. Section 2.1.2) of (3.13) in the case of the Gauss–Chebyshev formula. See also Chawla & Kumar [1978], [1979].

Formulas of the type (3.13) for infinite intervals and Hermite measure $d\lambda(t) = e^{-t^2}dt$, including their convergence, are discussed by Kas'janov [1977]. Velev, Semanov & Soliev [1977] use Hermite interpolation processes to derive quadrature rules with multiple nodes for the approximation of singular integrals (3.4) with $d\lambda(t) = (1-t^2)^{-1/2}dt$.

For the use of Gauss-type quadrature rules to approximate Cauchy principal value integrals in higher dimensions, see Gabdulhaev [1975], Gabdulhaev & Onegov [1976], Velev, Semenov & Soliev [1977], Šeško [1979], Tsamasphyros & Theocaris [1979] and Theocaris, Ioakimidis & Kazantzakis [1980].

3.2.3. *Computational considerations.* Although the quadrature rules (3.9) and (3.13) are numerically unstable when x is near one of the nodes τ_ν, a device already used by Korneĭčuk [1964] allows us to evaluate the quadrature sums in a stable manner for arbitrary $x \in (a, b)$. We describe the procedure for the formula (3.13), assuming that the underlying quadrature formula is a Gauss–Christoffel formula.

We represent the polynomial $p_{n-1}(f;\cdot)$ of degree $\leq n-1$ interpolating f at the zeros τ_ν of $\pi_n(\cdot\,; d\lambda)$ in the form

$$p_{n-1}(f;t) = \sum_{k=0}^{n-1} a_k \pi_k(t), \tag{3.14}$$

where, by virtue of the discrete orthogonality property of orthogonal polynomials,

$$a_k = h_k^{-1} \sum_{\nu=1}^{n} \lambda_\nu \pi_k(\tau_\nu) f(\tau_\nu), \qquad k = 0, 1, \ldots, n-1, \tag{3.15}$$

with $h_k = \int_a^b \pi_k^2(t)\,d\lambda(t)$. Integrating (3.14) yields (3.13) in the form

$$I(f)(x) = \sum_{k=0}^{n-1} a_k \rho_k(x) + R_n(f). \tag{3.16}$$

The polynomials $\{\pi_k(x)\}$ and functions $\{\rho_k(x)\}$ required in (3.15) and (3.16) both satisfy the recurrence relation (cf. Section 1.4)

$$y_{k+1} = (x - \alpha_k)y_k - \beta_k y_{k-1}, \qquad k = 0, 1, 2, \ldots, \tag{3.17}$$

the initial values being $\pi_{-1}(x) = 0$, $\pi_0(x) = 1$ for $\{\pi_k(x)\}$, and

$$\rho_{-1}(x) = 1, \quad \rho_0(x) = \int_a^b \frac{d\lambda(t)}{x - t}, \tag{3.18}$$

for $\{\rho_k(x)\}$. (We assume that $\beta_0 = \int_a^b d\lambda(t)$ in (3.17).) The computation of $\rho_n(x)$ by means of (3.17), (3.18) is quite stable if x is in the interior of $[a, b]$. The only nontrivial computation, therefore, is that of $\rho_0(x)$ in (3.18). For many of the standard measures $d\lambda$, however, $\rho_0(x)$ can be expressed, and thus evaluated, in terms of known special functions. The sum in (3.16) is most effectively evaluated by Clenshaw's algorithm (Paget & Elliott [1972]).

A similar procedure applies to the quadrature rule (3.9) (Elliott & Paget [1979]). The approach, indeed, is capable of dealing with a much wider class of integrals, for example

$$I(f)(x) = \int_a^b K(x,t)f(t)d\lambda(t),$$

where $K(x,t)$ is a singular (or weakly singular) kernel, or a kernel that otherwise exhibits unpleasant behavior. For work along these lines see Bahvalov & Vasil'eva [1968], Piessens & Poleunis [1971], Branders & Piessens [1975], Patterson [1976/77], Elliott & Paget [1976b], [1978], Sloan [1978], and Smith & Sloan [1980].

An adaptive automatic integration routine for singular integrals (3.4) (with $d\lambda(t) = dt$) is developed in Piessens, VanRoy–Branders & Mertens [1976].

3.2.4. *Applications to singular integral equations.* The quadrature rules developed in Sections 3.2.1 and 3.2.2 are widely used for the approximate solution of singular integral equations in problems of elasticity theory, fluid flow, aerodynamics and electromagnetic scattering. In one of its simpler forms, the problem consists in finding a solution $y(t)$ of an integral equation of the first kind,

$$\text{(3.19)} \qquad \int\!\!\!\!\!\!-_{-1}^{1} \frac{y(t)}{x-t}\,dt + \int_{-1}^{1} k(x,t)y(t)dt = f(x), \qquad -1 < x < 1,$$

where k and f are given, usually smooth, functions. Depending on whether or not one seeks a solution that is bounded at one or both of the endpoints, there may be no solution (unless a compatibility condition is fulfilled), a unique solution, or infinitely many solutions. When solutions exist, they will be of the form

$$\text{(3.20)} \qquad y(t) = u(t)\omega(t)$$

where ω exhibits square root singularities at the endpoints, and u is smooth, if k and f are. The exact type of singularity of ω is well-determined, once the boundedness characteristics of y have been defined.

For the numerical solution of (3.19) one now substitutes (3.20) into (3.19), applies the quadrature rule (3.9) with the appropriate $d\lambda(t) = \omega(t)dt$ to the first integral in (3.19) and the parent quadrature rule G_n in (3.6) to the second. If one further chooses for x the roots x_i of $\rho_n(x) = 0$, there results a system of linear equations

$$\text{(3.21)} \qquad \sum_{\nu=1}^{n} \lambda_\nu \left[\frac{1}{x_i - \tau_\nu} + k(x_i, \tau_\nu) \right] u_\nu = f(x_i)$$

for the unknowns u_ν which approximate $u(\tau_\nu)$. If there are fewer than n zeros of $\rho_n(x)$, additional equations — usually physically meaningful ones — can be adjoined.

Similarly one deals with integral equations of the second kind,

$$a(x)y(x) + b(x)\int\!\!\!\!\!\!-_{-1}^{1} \frac{y(t)}{x-t}\,dt + \int_{-1}^{1} k(x,t)y(t)dt = f(x), \qquad -1 < x < 1,$$

the solution of which again admits representations of the form (3.20), but with ω now a more general Jacobi type weight function. To again arrive at a linear system of the type (3.21), the "collocation points" x_i must now be chosen as roots of the equation (Theocaris [1976], Ioakimidis & Theocaris [1978a])

$$a(x)\omega(x) + b(x)\frac{\rho_n(x)}{\pi_n(x)} = 0.$$

Best results (when u in (3.20) is smooth) can be expected from the employment of the appropriate Gauss–Jacobi quadrature rule (3.6). This indeed has been the choice in the work of Stark [1971], Erdogan & Gupta [1971/72], Krenk [1975/76], Theocaris & Ioakimidis [1978a] and others. In many applications the value of u at one or both endpoints is physically meaningful, and

indeed may be the only quantity of interest. In such cases the Gauss–Radau and Gauss–Lobatto rules are appropriate and are the preferred choice in the work of Ioakimidis & Theocaris [1977a], [1978a, b], Theocaris & Ioakimidis [1977/78], [1978b], Krenk [1978], and Theocaris & Tsamasphyros [1979]. Occasionally, other types of singularity arise in singular integral equations and must be dealt with accordingly. Theocaris & Ioakimidis [1977], [1979a], for example, consider a problem with complex Jacobi-type singularities, Theocaris, Chrysakis & Ioakimidis [1979] one with a logarithmic singularity, while Cohen [1978] considers problems on an infinite interval. Interpolation schemes that allow $u(t)$ to be obtained for arbitrary $t \neq \tau_\nu$, with an accuracy comparable to the one of the approximations u_ν, are discussed in Theocaris & Ioakimidis [1979b]. For convergence results, see Ioakimidis & Theocaris [1980].

Similar methods can also be applied to singular integro-differential equations; see Ioakimidis & Theocaris [1979].

4. The Remainder Term and Convergence

The analysis of the remainder of a quadrature rule has a long and extensive history and continues to be an active topic of research. There are three major areas of concern: The representation of the remainder in some form or another, the estimation of its magnitude, and conclusions concerning the convergence behavior of the quadrature rule. We only review work that relates specifically to Gauss–Christoffel quadrature rules.

One of the early representations of the remainder, Markov's formula for $R_n(f)$ in terms of the $2n$-th derivative of f (cf. Section 1.4), while widely quoted, is of limited practical value, as it stands. For one, high-order derivatives are usually difficult to estimate. Then the formula cannot be applied to functions of low-order continuity. And finally, it does not lend itself easily for a comparison with other quadrature rules which may have different degrees of exactness. For these reasons, other representations are being used, notably representations valid for functions that can be extended holomorphically into the complex plane, and others valid for real functions of a given continuity class.

4.1. The remainder term for holomorphic functions

There are several approaches for estimating the remainder $R_n(f)$ when f is holomorphic. Among the oldest is the method of contour integration. More recent approaches use tools of functional analysis and approximation theory. Whatever the approach, the results are often quite comparable.

4.1.1. *Estimates based on contour integration.* For simplicity we assume that $[a, b]$ is a finite interval, which we standardize to $[-1, 1]$. The use of contour integration to represent the remainder $R_n(f)$ can be traced back at least to Heine [1881] whose work immediately yields (cf. Section 1.4)

$$R_n(f) = \frac{1}{2\pi i} \oint_\Gamma \frac{\rho_n(z)}{\pi_n(z)} f(z)\,dz. \tag{4.1}$$

It is assumed in (4.1) that f is single-valued holomorphic in a domain D which includes the interval $[-1, 1]$ in its interior, Γ is a contour in D surrounding $[-1, 1]$, π_n the appropriate orthogonal polynomial, and

$$\rho_n(z) = \int_{-1}^{1} \frac{\pi_n(t)}{z - t} d\lambda(t).$$

Two choices of Γ are most frequently made: $\Gamma = C_r$, the circle $|z| = r$, $r > 1$, and $\Gamma = \mathscr{E}_\rho$, the ellipse with foci at ± 1 and sum of its semiaxes equal to ρ, $\rho > 1$. The parameters r and ρ can be varied in certain intervals $1 < r \leq r^*$, $1 < \rho \leq \rho^*$ determined by the domain of holomorphy of f. Circles, of course, can only be used if D is sufficiently large so as to contain a circle C_r for some $r > 1$. In this respect, ellipses $\mathscr{E}_\rho$ have the advantage of shrinking to the interval $[-1, 1]$ when $\rho \to 1$, which makes them suitable to deal with functions which are analytic on the segment $[-1, 1]$. Having families of contours Γ at disposal provides for flexibility and gives an opportunity for optimization in the estimates of $|R_n(f)|$ to be made. These estimates follow directly from (4.1), and have the form

$$|R_n(f)| \leq \frac{1}{2\pi} \gamma_n l(\Gamma) \max_{z \in \Gamma} |f(z)|, \tag{4.2}$$

where $l(\Gamma)$ is the length of Γ and γ_n either a strict upper bound for $|\rho_n(z)/\pi_n(z)|$ on Γ, or an asymptotic estimate valid for $n \to \infty$. In the latter case, (4.2) is only an approximate relation. Strict error bounds are obtained in this manner, for some of the classical Gauss–Christoffel formulae, by McNamee [1964], Chawla [1967], [1968], Kambo [1970], [1970/71], Donaldson [1973], Kumar [1974a, b], Porath & Wenzlaff [1976], asymptotic estimates by Fock [1932], Barrett [1960/61], Chawla & Jain [1968a, b], Donaldson & Elliott [1972], Ramakrishnan [1973] and Smith [1977]. As pointed out in some of these references, the method can be extended to infinite intervals, and is easily adapted to incorporate poles and other singularities of f.

An equivalent form of (4.1) is (cf. Section 1.4)

$$R_n(f) = \frac{1}{2\pi i} \oint_\Gamma R_n\left(\frac{1}{z - \cdot}\right) f(z) dz. \tag{4.1'}$$

In this form the remainder is studied extensively by Takahasi & Mori [1970], [1971], who display many revealing contour maps of $|R_n(1/(z - \cdot))|$ for Gauss–Legendre and other quadrature rules. Lether [1980] expands $(z - \cdot)^{-1}$ in (4.1') in a series of Chebyshev polynomials of the second kind and obtains an estimate of $R_n(f)$ for arbitrary measure $d\lambda(t)$ on a finite interval.

Freud [1973], [1975a, b] establishes the new representation

$$R_n(f) = \sum_{\nu=0}^{\infty} \frac{h_\nu}{2\pi i} \oint_{\Gamma_\nu} \frac{f(z)dz}{\pi_\nu(z)\pi_{\nu+1}(z)}, \tag{4.3}$$

valid for arbitrary measures $d\lambda(t)$, where π_k are the (monic) orthogonal polynomials, $h_\nu = \int_a^b \pi_\nu^2(t)d\lambda(t)$, and Γ_ν are contours enclosing all zeros of π_ν and $\pi_{\nu+1}$. In the first paper, Freud combines (4.3) with asymptotic results for orthogonal polynomials to derive asymptotic estimates for $|R_n(f)|$ under the assumption that $d\lambda(t)$ has support in $[-1,1]$ and is such that $\ln \lambda'(\cos\theta)$ is Lebesgue integrable on $[-\pi, \pi]$. In the subsequent papers these estimates are further developed into strict upper bounds. (Cf. also Section 4.1.3, in particular v. Sydow [1977/78].)

Assuming $d\lambda(t) = \omega(t)dt$ on $[-1,1]$, where ω is even, positive, and Lebesgue integrable, and assuming f holomorphic in $|z|<1$ and continuous on $|z|=1$, Stenger [1966] uses (4.1) to derive the expansion

$$R_n(f) = \sum_{\nu=0}^{\infty} a_{2n+2\nu} r_{n,\nu}, \tag{4.4}$$

where a_{2k} are the coefficients in the Maclaurin series of f. The quantities $r_{n,\nu} = R_n(t^{2n+2\nu})$ are shown to be positive, and to satisfy $r_{n,\nu+1} - r_{n+1,\nu} > 0$, for all $n \geq 1$, $\nu \geq 0$. This has the interesting consequence that $R_n(f) \geq R_{n+1}(f) \geq 0$ for all $n \geq n_0$, whenever $a_{2k} \geq 0$ for all $k \geq n_0$. (See also Section 4.3, in particular Brass [1978].) Upper bounds for $|R_n(f)|$ can be obtained from (4.4) by applying Schwarz's or Hölder's inequality (if $\{r_{n,\nu}\}_{\nu=0}^{\infty} \in l_p$, $1 \leq p \leq \infty$), and by using Cauchy's inequality to estimate $|a_{2k}|$.

Representations of $R_n(f)$, similar to the one in (4.4), in terms of other expansions are obtained for various special Gauss–Christoffel formulae by Chawla [1970a], [1971a], Kambo [1971] and Jayarajan [1974], who use expansions in Chebyshev or Legendre polynomials. This again yields error bounds if one suitably estimates the expansion coefficients. Luke [1975], for arbitrary measure $d\lambda(t)$, expands f in orthogonal polynomials $\pi_k(t; d\lambda)$ and works out the corresponding expansion for $R_n(f)$. This is extended in Luke, Ting & Kemp [1975] to the case of Christoffel quadratures (with preassigned nodes).

4.1.2. *Hilbert space norm estimates.* The idea of using Hilbert space methods to estimate linear functionals that are important in approximation (such as the error functional $R_n(f)$ in a quadrature rule) was first introduced by Davis [1953]. Here the scenario calls for a Hilbert space $\mathcal{H} = \mathcal{H}(D)$ of functions which are single-valued holomorphic in a domain D that contains the interval $[-1,1]$. If, then, R_n is a bounded linear functional in $\mathcal{H}$, one gets immediately

$$|R_n(f)| \leq \sigma_n \|f\|, \tag{4.5}$$

where $\sigma_n = \|R_n\|$ is the norm of the error functional R_n and $\|f\|$ the norm of f in the Hilbert space $\mathcal{H}$. The former depends only on the quadrature rule in question, the latter only on the function to which the rule is applied. Indeed, if $\{p_k\}$ is a complete orthonormal system in $\mathcal{H}$, then

$$\sigma_n^2 = \sum_{k=0}^{\infty} |R_n(p_k)|^2. \tag{4.6}$$

Davis [1953] originally, and Stetter [1968], Riess [1971], Haber [1971], [1971/72], Kofron [1972], Hämmerlin [1972] subsequently, use circular domains bounded by C_r, $r > 1$, and equip $\mathcal{H}$ with the inner product $(f, g) = \int_{C_r} f(z)\overline{g(z)}ds$. The orthonormal system then consists of powers, $p_k(z) = (2\pi r)^{-1/2}(z/r)^k$. The norm of f in (4.5) can be further estimated to yield

$$|R_n(f)| \leq \tau_n \sup_{z \in C_r} |f(z)|, \qquad \tau_n = \sigma_n \sqrt{2\pi r}.$$

For reasons already indicated in the previous section, domains bounded by an ellipse $\mathcal{E}_\rho$, $\rho > 1$ (with semimajor axis a and semiminor axis b, $a + b = \rho$) are the preferred choice of many authors. They are used, e.g., by Davis & Rabinowitz [1954], Davis [1962], Barnhill [1968], Chawla [1969], Riess & Johnson [1969], Haber [1971/72], in conjunction with the double integral inner product $(f, g) = \iint_{\text{int}(\mathcal{E}_\rho)} f(z)\overline{g(z)}dxdy$. This yields estimates of the form

$$|R_n(f)| \leq \tau_n \sup_{z \in \mathcal{E}_\rho} |f(z)|, \quad \tau_n = \sigma_n \sqrt{\pi ab}, \tag{4.7}$$

where σ_n can be computed (or estimated) from (4.6), the p_k being essentially Chebyshev polynomials of the second kind. For a number of quadrature rules, including Gaussian rules, the quantities σ_n are tabulated for selected values of a (or ρ) in Lo, Lee & Sun [1965] and Stroud & Secrest [1966]. (Earlier tables in Davis & Rabinowitz [1954] and Davis [1962] contain a systematic error.) Somewhat sharper bounds result through the use of the line integral inner product $(f, g) = \int_{\mathcal{E}_\rho} f(z)\overline{g(z)}|1 - z^2|^{-1/2}ds$, as is shown in Chawla [1968/69], [1969] and Rabinowitz & Richter [1970], or through the use of $(f, g) = \int_{\mathcal{E}_\rho} f(z)\overline{g(z)}|\omega(z)|ds$, where $d\lambda(t) = \omega(t)dt$ (Chawla [1970b]). The orthonormal functions in the former case are Chebyshev polynomials of the first kind. Nearly identical results are derived by other means in Chawla [1971b]. Knauff [1976/77] uses Banach space methods to obtain estimates of the type (4.7) for Gauss–Chebyshev quadratures. Indeed, there are many other ways such estimates can be derived; Rabinowitz [1969] compares five of them in the case of Gauss–Legendre formulae.

Nicholson, Rabinowitz, Richter & Zeilberger [1971], and Curtis & Rabinowitz [1972] study the error of Gauss–Legendre, Radau and Lobatto formulae when applied to Chebyshev polynomials. In view of (4.6), this yields information on the error norms σ_n in the respective Hilbert spaces $\mathcal{H}(\mathcal{E}_\rho)$.

4.1.3. *Estimates via approximation theory.* If $[a, b]$ is a finite interval, f continuous on $[a, b]$, and if p^*_{2n-1} achieves the best uniform approximation $E_{2n-1}(f)$ of f by polynomials of degree $\leq 2n - 1$,

$$E_{2n-1}(f) = \inf_{p \in \mathbf{P}_{2n-1}} \max_{a \le t \le b} |f(t) - p(t)| = \|f - p^*_{2n-1}\|_\infty,$$

then it is a simple matter to observe that for any Gauss–Christoffel formula, $|R_n(f)| = |R_n(f - p^*_{2n-1})| \le 2\mu_0 \|f - p^*_{2n-1}\|_\infty$, hence

$$|R_n(f)| \le 2\mu_0 E_{2n-1}(f), \quad \mu_0 = \int_a^b d\lambda(t). \tag{4.8}$$

This was already noted by Bernstein [1918], who combines (4.8) with his own estimates of $E_{2n-1}(f)$ for holomorphic functions. The same observation, similarly applied to holomorphic functions, is made by Stenger [1966] and Locher & Zeller [1968]. The strongest result is due to v. Sydow [1977/78], who proves

$$|R_n(f)| \le 4\mu_0(1-\rho^{-2})^{-1}\rho^{-2n} \cdot \max_{z \in \mathscr{E}_\rho} |f(z)| \tag{4.9}$$

for arbitrary measure $d\lambda(t)$ and functions f that are holomorphic in the interior of $\mathscr{E}_\rho$ and continuous on the boundary. The formula (4.9) is typical for many results obtained previously in special cases by the methods of Sections 4.1.1 and 4.1.2. Locher [1974] makes a somewhat different use of (4.8).

4.2. The Peano representation of the remainder

Kronecker's stern dictum "... ohne Restglied ist es keine Formel!" (Kronecker [1894]) has lost much of its punch since Peano [1913], [1914] showed that essentially every linear functional that annihilates polynomials up to a certain degree automatically generates its own remainder term. Thus, for a quadrature rule over a finite interval $[a, b]$, if the error functional R_n satisfies $R_n(p) = 0$ for all $p \in \mathbf{P}_{s-1}$ and f has a piecewise continuous derivative of order s on $[a, b]$ (or, less restrictively, $f^{(s-1)}$ is absolutely continuous on $[a, b]$), then

$$R_n(f) = \int_{-\infty}^{\infty} K_s(t) f^{(s)}(t)\, dt, \tag{4.10}$$

where

$$K_s(t) = R_n\left[\frac{(\cdot - t)_+^{s-1}}{(s-1)!}\right]. \tag{4.11}$$

Here the plus sign on the right is the "cutoff" symbol, indicating that the function on which it acts is to be set equal to zero if the argument is negative. K_s in (4.11) is called the s-th *Peano kernel* of R_n; it is a spline function of degree $s-1$, with knots at the quadrature nodes and compact support $[a, b]$. (The integral in (4.10) could therefore be extended over $[a, b]$.) The formula (4.10) simplifies if K_s has constant sign on $[a, b]$, in which case

$$R_n(f) = c_s f^{(s)}(\bar{t}), \quad c_s = \int_{-\infty}^{\infty} K_s(t)\, dt = R_n\left(\frac{t^s}{s!}\right), \tag{4.10'}$$

where $\bar{t}$ is some (unknown) intermediate value in $[a, b]$.

A quadrature rule which has degree of exactness d (but not $d+1$) thus generates exactly $d+1$ Peano kernels $K_1, K_2, \ldots, K_{d+1}$. We have $d = 2n-1$ for the n-point Gauss–Christoffel formula, $d = 2n-2$ for the Radau formula, etc.

Peano's representation (4.10) can be used in different ways to estimate the remainder. For example,

$$|R_n(f)| \leq e_s \max_{a \leq t \leq b} |f^{(s)}(t)|, \tag{4.12}$$

where

$$e_s = \int_{-\infty}^{\infty} |K_s(t)|\, dt, \qquad s = 1, 2, \ldots, d+1. \tag{4.13}$$

The numbers e_s are often referred to as the *Peano constants* of R_n. (Their dependence on n is suppressed in the notation). Equality in (4.12) can be attained for special f. Note also that for Gauss–Christoffel formulae, according to Markov (cf. Section 1.4), $e_{2n} = [(2n)!]^{-1} \int_a^b \pi_n^2(t) d\lambda(t)$. Alternatively, if $f^{(s)}$ is of bounded variation, $|R_n(f)| \leq \mathrm{Var}(f^{(s)}) \max_t |K_{s+1}(t)|$. Still another use of the Peano representation is made by Cosma Cagnazzi [1970] who for quadrature rules with positive coefficients derives estimates of the form $|R_n(f)| \leq e_s^* o_s$, where $o_s = \max_{a \leq t \leq b} f^{(s)}(t) - \min_{a \leq t \leq b} f^{(s)}(t)$ is the oscillation of $f^{(s)}$ on $[a, b]$, and $e_s^* = (s!)^{-1} \int_a^b (t-a)^s d\lambda(t)$ are certain constants depending only on s, but not on the specific quadrature rule under consideration.

Once the first few Peano constants are known, (4.12) is especially useful for estimating the quadrature error in cases where only low-order derivatives of f exist, or are accessible. The importance of this point was already stressed by v. Mises [1933], who in fact, apparently unaware of Peano's work, constructs the Peano kernels by repeated integration (v. Mises [1936]). v. Mises also observes that the Peano kernel K_s of the n-point Gauss–Legendre formula has exactly $2n - s$ sign changes in $[-1, 1]$ (hence none if $s = 2n$), a fact noted later again by Roghi [1967]. Similar statements hold for Gauss–Radau and Gauss–Lobatto formulae (cf., e.g., Brass [1977, Satz 82]).

Stroud [1966] makes the point that the Peano estimate (4.12), for functions of low-order continuity, often compares favorably with other estimates of the same form obtained by approximation-theoretic means. See Rabinowitz [1968], Riess & Johnson [1969], Chui [1972], for estimates of the latter kind.

The Peano constants provide a convenient means of measuring the quadrature error for functions of a given continuity class. This allows comparisons of different quadrature rules on a common basis. It is remarkable, in this respect, that the Gauss–Legendre formula, even for functions of low continuity, compares favorably with other common integration rules, such as Romberg integration, which use the same number of points (Stroud [1965]). According to Stroud & Secrest [1966], the first two Peano constants indeed are only marginally larger than the corresponding constants for the best quadrature

rules (which minimize the integral in (4.13)). It appears therefore, contrary to widespread belief, that Gauss–Christoffel formulae are not only effective for highly regular functions, but also handle functions of low-order continuity at least as well as other common quadrature rules.

Selected Peano constants e_s, $s = 1,2,4,8,\ldots$, including e_{2n}, are tabulated in Stroud & Secrest [1966] for many Gauss–Christoffel and related quadrature rules. Their computation, particularly for large s, is quite difficult because of severe cancellation problems.

Peano-type error estimates in the case of infinite intervals $[a, b]$, particularly for Gauss–Laguerre formulae, are obtained by Stroud & Chen [1972].

Radon [1935], Rémès [1940] and Milne [1949] generalize Peano's theory to functionals R that do not annihilate polynomials, but instead annihilate all solutions of a linear homogeneous differential equation of order s. If L is the associated linear differential operator, and $g(\tau, t)$ the Green's function of the initial value problem, then

$$Rf = \int_{-\infty}^{\infty} K(t)(Lf)(t)dt,$$

where the Peano kernel is now given by

$$K(t) = R[g_+(\cdot, t)].$$

Here, $g_+(\tau, t) = g(\tau, t)$ if $\tau < t$, and $g_+(\tau, t) = 0$ otherwise. For Peano kernels of constant sign,

$$Rf = (Lf)(\bar{t}) \cdot Rw, \qquad a < \bar{t} < b,$$

where w is any solution of $Lw = 1$. Still further generalizations are due to Sard [1948], who also makes precise the class of functionals R to which Peano's theory applies.

4.3. Convergence

The convergence theory for quadrature rules of the form

$$\int_a^b f(t)d\lambda(t) = \sum_{\nu=1}^{n} \lambda_\nu^{(n)} f(\tau_\nu^{(n)}) + R_n(f), \qquad n = 1,2,3,\ldots, \tag{4.14}$$
$$a \le \tau_n^{(n)} < \tau_{n-1}^{(n)} < \cdots < \tau_1^{(n)} \le b,$$

is particularly simple if $[a, b]$ is a *finite* interval. By a theorem of Steklov [1916] and Pólya [1933] the quadrature rule in (4.14) converges for every continuous function,

$$\lim_{n\to\infty} R_n(f) = 0, \qquad f \in C[a, b],$$

precisely if

$$(4.15)\qquad \begin{cases} \lim\limits_{n\to\infty} R_n(p) = 0 & \text{for every polynomial } p \\ \text{and} \quad \sum\limits_{\nu=1}^{n} |\lambda_\nu^{(n)}| \le K & \text{for all } n = 1, 2, 3, \ldots, \end{cases}$$

where $K > 0$ is a constant not depending on n. If (4.14) are Gauss–Christoffel quadrature formulae then the first condition in (4.15) is trivially true, $R_n(p)$ being zero if $2n$ exceeds the degree of p, and the second follows from the positivity of the Christoffel numbers,

$$\sum_{\nu=1}^{n} |\lambda_\nu^{(n)}| = \sum_{\nu=1}^{n} \lambda_\nu^{(n)} = \int_a^b d\lambda(t).$$

Thus, Gauss–Christoffel quadrature rules on a finite interval always converge for every continuous function. We can see this also directly from (4.8) and Weierstrass's approximation theorem. Stieltjes [1884a], in a beautiful memoir, indeed proves convergence for every function that is Riemann–Stieltjes integrable.

Mindful, however, of the fact (Lipow & Stenger [1972]) that for every quadrature rule which converges on $C[a, b]$ there is an $f \in C[a, b]$ for which convergence is arbitrarily slow, one ought to be less concerned with convergence as such, and more with the quality of convergence. In this regard the estimates discussed in Sections 4.1.1–4.1.3 provide useful insights. Bernstein's estimate (4.8), e.g., when used in conjunction with results of Jackson and Bernstein, yields $R_n(f) = o(n^{-s})$ if $f \in C^s[a, b]$, $\limsup |R_n(f)|^{1/n} < 1$ if f is analytic on $[a, b]$, and $|R_n(f)|^{1/n} = o(1)$ if f is entire. Similarly, the bound in (4.9) assures us of geometric convergence in the case of holomorphic functions and tells us how the convergence rate increases with the size of the domain of holomorphy. For results on convergence rates in terms of the r-th modulus of continuity see Butzer, Scherer & Westphal [1973] and Butzer [1979/80].

The convergence of Gauss–Christoffel formulae on *infinite* intervals is a more subtle question. It is intimately related to the determinacy of the moment problem for $d\lambda(t)$. That such a connection exists, in the case of a half-infinite interval $[0, \infty]$, is suggested by a result of Stieltjes according to which the moment problem is determined if and only if the continued fraction corresponding to the integral $\int_0^\infty d\lambda(t)/(z - t)$ converges. The determinacy of the moment problem therefore implies convergence of the Gauss–Christoffel rule for $f(t) = (z - t)^{-1}$, $z \notin [0, \infty]$ (cf. Section 5, Eq. (5.7)). For more general functions f the theory gradually evolved through the work of Uspensky [1916], [1928], Shohat [1927], Jouravsky [1928] and Shohat & Tamarkin [1943]. Assuming that $\int_{-\infty}^{\infty} f(t) d\lambda(t)$ exists as an (improper) Riemann–Stieltjes integral and that the moment problem on $[-\infty, \infty]$ is determined for $d\lambda(t)$, the Gauss–Christoffel quadrature rule converges if $|f(t)| \le A + Bt^{2s}$ for all real t, where A, B are positive constants and $s \ge 1$ an integer (Freud [1971, Ch. 3,

Thm. 1.1]). In fact, this is true for every sequence of positive quadrature rules (i.e., $\lambda_\nu^{(n)} \geq 0$ in (4.14)) which converge on polynomials. For the determinacy of the moment problem one has well-known criteria due to Carleman, M. Riesz and others (see, e.g., Shohat & Tamarkin [1943, p. 19f]). More general theorems of this type, for arbitrary intervals $[a, b]$, are known in which the condition on f is replaced by the condition that for suitable functions G_a, G_b the limits $\lim_{t \downarrow a} f(t)/G_a(t)$, $\lim_{t \uparrow b} f(t)/G_b(t)$ be zero (Ivanova [1955], Freud [1971, Ch. 3, Thms. 1.6, 1.6a, 1.6b]), or at least finite (Esser [1971b]). If $[a, b]$ is compact, these conditions allow f to become singular at one or both endpoints. Concrete theorems of this kind for classical Gauss–Christoffel formulae are summarized in Freud [1971, p. 96].

The Steklov–Pólya criterion (4.15) can be extended to quadrature rules that have multiple nodes with arbitrary multiplicities, so long as the multiplicities do not exceed a fixed integer s for all n. If the criterion is fulfilled one gets convergence for all $f \in C^s[a, b]$ (Bandemer [1966], [1967]). In particular, all Christoffel–Stancu quadrature rules on finite intervals converge in this sense (Filippi & Esser [1970], Esser [1971a], [1972]). Convergence theorems for Gauss–Radau formulae on infinite intervals are included in Freud [1971, Ch. 3, Thm. 1.4]; for the Gauss–Laguerre measure, see also Krylov & Fedenko [1962].

Another aspect of convergence is monotonicity. While monotone convergence cannot hold for *all* $f \in C[a, b]$ (Filippi & Esser [1970, Satz 9]), Brass [1978], for the quadrature rule $Q_n(f)$ in (4.14), shows $Q_n(f) \leq Q_m(f)$ for all $m > n$, if $f^{(2n)}$ is continuous on $[a, b]$ and nonnegative. If this condition holds for each n, convergence is monotone. For an alternative proof, see Locher [1980].

Finally, attempts may be made to speed up the convergence of quadrature rules through appropriate acceleration techniques. The use of the ε-algorithm for this purpose is studied empirically by Chisholm, Genz & Rowlands [1972].

5. Computation of Gauss–Christoffel Quadrature Formulae; Numerical Tables

Generating Gauss–Christoffel quadrature rules is closely related to the problem of generating orthogonal polynomials (see Section 1.4). In principle, this problem was already solved by Chebyshev [1859a] for discrete measures, and by Stieltjes [1884a] for general measures. If the measure in question is $d\lambda(t)$, and $(f, g) = \int_a^b f(t)g(t)d\lambda(t)$ denotes the inner product of f and g, then the (monic) orthogonal polynomials $\{\pi_k\}$ satisfy the three-term recurrence relation

$$\begin{aligned} \pi_{k+1}(t) &= (t - \alpha_k)\pi_k(t) - \beta_k \pi_{k-1}(t), \qquad k = 0, 1, 2, \ldots, \\ \pi_{-1}(t) &= 0, \quad \pi_0(t) = 1, \end{aligned} \tag{5.1}$$

where the coefficients α_k, β_k are given by

$$\alpha_k = \frac{(t\pi_k, \pi_k)}{(\pi_k, \pi_k)}, \qquad k = 0, 1, 2, \ldots, \tag{5.2}$$

$$\beta_k = \frac{(\pi_k, \pi_k)}{(\pi_{k-1}, \pi_{k-1})}, \qquad k = 1, 2, 3, \ldots \tag{5.3}$$

(Darboux [1878], Stieltjes [1884a]). If $d\lambda(t) \geq 0$, as we shall assume, then (5.3) shows that $\beta_k > 0$ for $k \geq 1$. Since π_0 is known, and β_0 is arbitrary, we obtain α_0 from (5.2), whence π_1 from (5.1). Knowing π_0 and π_1, we now compute α_1 and β_1 from (5.2) and (5.3), and then again π_2 from (5.1). Continuing in this manner, we can generate as many polynomials, and therefore as many of the coefficients α_k, β_k, as are desired. This is the *procedure of Stieltjes.*

While Stieltjes' procedure is very elegant, it leaves an important point unanswered: How are we to compute the inner products in (5.2), (5.3)? The manner in which the recursion coefficients α_k, β_k are determined, indeed, turns out to be rather critical for the numerical stability of the procedure. We find it convenient, therefore, to first assume that all coefficients α_k, β_k are explicitly known. (This is true for "classical" orthogonal polynomials.) An efficient algorithm for computing Gauss–Christoffel formulae can then be based on the associated Jacobi matrix. This is discussed in Section 5.1. The more difficult situation in which the coefficients α_k, β_k must be generated along with the polynomials π_k is deferred to Sections 5.2 and 5.3. In Section 5.4 we review numerical tables available for Gauss-type formulae.

5.1. Methods based on the Jacobi matrix

Suppose we wish to generate the n-point Gauss–Christoffel quadrature rule

$$\int_a^b f(t)d\lambda(t) = \sum_{\nu=1}^{n} \lambda_\nu f(\tau_\nu) + R_n(f). \tag{5.4}$$

We associate with the measure $d\lambda(t)$ the symmetric tridiagonal matrix of order n,

$$J_n = \begin{bmatrix} \alpha_0 & \sqrt{\beta_1} & & & 0 \\ \sqrt{\beta_1} & \alpha_1 & \sqrt{\beta_2} & & \\ & \sqrt{\beta_2} & \ddots & \ddots & \\ & & \ddots & \ddots & \sqrt{\beta_{n-1}} \\ 0 & & & \sqrt{\beta_{n-1}} & \alpha_{n-1} \end{bmatrix}, \tag{5.5}$$

where α_k, β_k are the recursion coefficients in (5.1). We refer to J_n as the n-th *Jacobi matrix* of $d\lambda(t)$. The polynomial π_n, then, is precisely the characteristic polynomial of J_n. The nodes τ_ν, being the zeros of π_n, are therefore the eigenvalues of J_n. Denoting by v_ν the (normalized) eigenvector corresponding to τ_ν,

$$J_n v_\nu = \tau_\nu v_\nu, \quad v_\nu^T v_\nu = 1, \qquad \nu = 1, 2, \ldots, n,$$

Wilf [1962, Ch. 2, Exercise 9] (and Goertzel around 1954 before him (Wilf [1980])) observes that the Christoffel numbers λ_ν are expressible in terms of the first components $v_{\nu,1}$ of v_ν by means of

$$\lambda_\nu = \mu_0 v_{\nu,1}^2, \qquad \nu = 1, 2, \ldots, n, \quad \mu_0 = \int_a^b d\lambda(t).$$

Obtaining the n-point Gauss–Christoffel formula (5.4), therefore, amounts to calculating the eigenvalues and first components of the corresponding eigenvectors of the symmetric tridiagonal matrix J_n (a fact noted also by Gordon [1968]).

This is accomplished most effectively by Francis' QR algorithm (Golub & Welsch [1969], Wilkinson & Reinsch [1971, p. 241ff], Sack & Donovan [1971/72], Gautschi [1979b]), or by Rutishauser's LR algorithm (Sack & Donovan [1971/72], Capovani, Ghelardoni & Lombardi [1976a, b]), both executed with appropriate shift strategies. These methods are indeed extremely fast. According to Capovani et al. [1976b], e.g., it takes only 7.24 seconds of machine time on an IBM 370/168 to generate a 1000-point Gauss–Hermite formula! Alternative procedures based on the Newton–Raphson method, or other rootfinding methods, which compute τ_ν as zeros of π_n, not only require considerable care in the selection of initial approximations (Stroud & Secrest [1966], Laurie [1977], Laurie & Rolfes [1979]), but also tend to be slower (Gautschi [1979b]).

For special measures, such as the Legendre measure $d\lambda(t) = dt$, faster methods can be obtained by combining high-order rootfinding procedures with a judicious choice of initial approximations; see, e.g., Lether [1978], Gatteschi [1979], Gautschi [1979b].

The eigenvalue method described for Gauss–Christoffel formulae can be modified to produce Gauss–Radau and Gauss–Lobatto formulae; for this, see Golub [1973].

In the case of measures supported on the non-negative real axis an ingenious algorithm is due to Rutishauser [1962a, b]. It departs from Stieltjes' integral and its corresponding continued fraction ("S-fraction"),

$$\int_a^b \frac{d\lambda(t)}{z-t} \sim \frac{\mu_0}{z-} \frac{q_1}{1-} \frac{e_1}{z-} \frac{q_2}{1-} \frac{e_2}{z-} \cdots. \tag{5.6}$$

(This continued fraction is not to be confused with the associated continued fraction, the "J-fraction", already used by Gauss, which is a contraction of the S-fraction. Accordingly, (5.6) is valid only for $0 \leq a < b \leq \infty$; see Perron [1957, Satz 4.1].) The coefficients q_k, e_k are all positive, and are readily obtained from the recursion coefficients α_k, β_k in (5.1), by virtue of

$$\left.\begin{aligned} \alpha_0 &= q_1 \\ \alpha_k &= e_k + q_{k+1} \\ \beta_k &= e_k q_k \end{aligned}\right\} \qquad k = 1, 2, 3, \ldots .$$

The connection between the S-fraction in (5.6) and Gauss–Christoffel formulae is expressed by the relation

$$\sum_{\nu=1}^{n} \frac{\lambda_\nu}{z - \tau_\nu} = \frac{\mu_0}{z -} \; \frac{q_1}{1 -} \; \frac{e_1}{z -} \; \frac{q_2}{1 -} \; \cdots \; \frac{e_{n-1}}{z -} \; \frac{q_n}{1}, \tag{5.7}$$

i.e., the Gauss–Christoffel nodes τ_ν are the poles, and the Christoffel numbers λ_ν the corresponding residues, of the $2n$-th convergent of the continued fraction in (5.6). Rutishauser now computes the poles of this convergent in a Graeffe-like manner, generating a sequence of finite continued fractions, all of the same form as in (5.7), each having as poles the squares of the poles of the preceding continued fraction. The process converges quadratically, and yields the poles τ_ν and residues λ_ν simultaneously.

5.2. Generation of the Jacobi matrix

Given the first $2n$ moments

$$\mu_k = \int_a^b t^k d\lambda(t), \qquad k = 0, 1, 2, \ldots, 2n-1, \tag{5.8}$$

it is possible to generate the Jacobi matrix (i.e., the coefficients α_k, β_k, $k = 0, 1, \ldots, n-1$) by means of Stieltjes' procedure. It suffices to represent each polynomial $\pi_k(t)$ explicitly in terms of powers of t and to compute the inner products in (5.2), (5.3) by "multiplying out" term by term. In this manner each α_k, β_k is obtained as a ratio of two quadratic forms in the coefficients of π_k and π_{k-1}, the matrices involved being Hankel matrices in the moments (5.8).

In terms of modern digital computation, however, the procedure is subject to two major criticisms: In the first place, the algorithm is highly unstable, especially for finite intervals $[a, b]$. This is ultimately a manifestation of the fact that the Gauss–Christoffel nodes τ_ν and weights λ_ν, considered as functions of the moments μ_k, become progressively more ill-conditioned (i.e., more sensitive to small perturbations in the moments) as n increases (Gautschi [1968a], [1978]). Secondly, the procedure is unnecessarily expensive, requiring, as it

does, of the order $O(n^3)$ arithmetic operations. Both these deficiencies can be alleviated.

The numerical stability is greatly enhanced (Sack & Donovan [1971/72]) if instead of the moments μ_k one employs the "modified moments"

$$\nu_k = \int_a^b p_k(t)d\lambda(t), \qquad k = 0, 1, \ldots, 2n-1, \tag{5.9}$$

where $\{p_k\}$ is a suitable system of polynomials (usually orthogonal on $[a, b]$ with respect to some other, classical, measure $ds(t)$). The resulting improvement in the numerical condition is analyzed in Gautschi [1970a].

To arrive at an efficient algorithm, assume that $\{p_k\}$ satisfies a three-term recurrence relation analogous to the one in (5.1),

$$p_{k+1}(t) = (t - a_k)p_k(t) - b_k p_{k-1}(t), \qquad k = 0, 1, \ldots, 2n-1,$$
$$p_{-1}(t) = 0, \quad p_0(t) = 1,$$

with coefficients a_k, b_k that are known explicitly. (If they are all zero, then $p_k(t) = t^k$, and the modified moments reduce to ordinary moments.) The desired recursion coefficients α_k, β_k can then be obtained via the "mixed moments"

$$\sigma_{k,l} = \int_a^b \pi_k(t)p_l(t)d\lambda(t), \qquad k, l \geq -1,$$

in the following manner. One initializes

$$\begin{cases} \sigma_{-1,l} = 0, & l = 1, 2, \ldots, 2n-2, \\ \sigma_{0,l} = \nu_l, & l = 0, 1, \ldots, 2n-1, \\ \alpha_0 = a_0 + \dfrac{\nu_1}{\nu_0}, & \beta_0 = 0, \end{cases} \tag{5.10°}$$

and then continues, for $k = 1, 2, \ldots, n-1$, with

$$\begin{cases} \sigma_{k,l} = \sigma_{k-1,l+1} - (\alpha_{k-1} - a_l)\sigma_{k-1,l} - \beta_{k-1}\sigma_{k-2,l} + b_l\sigma_{k-1,l-1} \\ \qquad\qquad l = k, k+1, \ldots, 2n-k-1, \\ \alpha_k = a_k - \dfrac{\sigma_{k-1,k}}{\sigma_{k-1,k-1}} + \dfrac{\sigma_{k,k+1}}{\sigma_{k,k}}, \qquad \beta_k = \dfrac{\sigma_{k,k}}{\sigma_{k-1,k-1}}. \end{cases} \tag{5.10^k}$$

The algorithm (5.10) not only furnishes the coefficients α_k, β_k, $k \leq n-1$, hence the orthogonal polynomials $\{\pi_k\}_{k=0}^n$, but also, at the same time, the normalization factors $\sigma_{k,k} = \int_a^b \pi_k^2(t)d\lambda(t)$, $k \leq n-1$. The number of arithmetic operations is clearly of the order $O(n^2)$, one order less than what is required in Stieltjes' procedure.

In the special case of ordinary moments ($a_k = b_k = 0$), and discrete

measures $d\lambda(t)$, the algorithm (5.10) reduces to one of Chebyshev [1859a]. The general case is due to Sack & Donovan [1971/72] who obtained an algorithm equivalent to (5.10). In the form (5.10) it was given, independently, by Wheeler [1974]. A derivation can also be found in Gautschi [1978]. Earlier algorithms of Golub & Welsch [1969] and Gautschi [1970a] are not competitive, as they too require $O(n^3)$ operations.

The performance of (5.10) appears to be most satisfactory if the interval $[a, b]$ is finite and $\{p_k\}$ are orthogonal on $[a, b]$ with respect to some (standard) weight function. For infinite intervals, a certain degree of ill-conditioning unfortunately persists (Gautschi [1970a]). The success of the algorithm, moreover, depends critically on the ability to compute the modified moments (5.9) accurately. This is often possible through a judicious use of recurrence relations, as for example in the case of Chebyshev and Gegenbauer moments (Piessens & Branders [1973], Branders [1976], Luke [1977], Lewanowicz [1979]). In other cases, closed form expressions can be obtained (Gautschi [1970a, examples (i), (iii)], Wheeler & Blumstein [1972], Blue [1979], Gautschi [1979a], Gatteschi [1980]).

As an application of the algorithm (5.10) we show how Christoffel's theorem (cf. Section 2.1.1) can be implemented in algorithmic form. Thus, we seek polynomials $\{\pi_k\}$ orthogonal on $[a, b]$ with respect to the measure $d\lambda(t) = u(t)ds(t)$, where u is a polynomial of some fixed degree m. Assuming that $ds(t)$ has a set of known orthogonal polynomials, we use these as the polynomials p_k in the modified moments (5.9). Writing u in the form

$$u(t) = \sum_{k=0}^{m} c_k p_k(t), \tag{5.11}$$

we find

$$\nu_k = \begin{cases} c_k \int_a^b p_k^2(t)ds(t), & k \le m \\ 0, & \text{otherwise.} \end{cases}$$

Applying now the algorithm (5.10) immediately yields the recursion coefficients α_k, β_k for the desired polynomials π_k. Note that algorithm (5.10) requires only $O(n)$ operations in this case, since $\nu_k = 0$ for all $k > m$.

In some applications, e.g. to Christoffel quadrature rules with preassigned nodes (cf. Section 2.1.1), one is not given the coefficients c_k in (5.11), but rather the zeros of u. An algorithmic implementation of Christoffel's theorem for this situation is given in Galant [1971].

Branders [1976], Laurie [1977], and Laurie & Rolfes [1979] implement Stieltjes' algorithm by expanding π_k in Chebyshev polynomials and by taking advantage of special properties, notably formulae for the product of two

Chebyshev polynomials, to carry out the computations. This approach relies on the Chebyshev moments of $d\lambda(t)$ and therefore represents but another realization of algorithm (5.10).

5.3. A discretization method

An approximative method for computing Gauss–Christoffel formulae, based on discrete orthogonal polynomials, is proposed by Gautschi [1968a]. It is applicable whenever the weight function $\omega(t)$ in $d\lambda(t) = \omega(t)dt$ can be evaluated for arbitrary t. We now describe a variant of this method which incorporates algorithm (5.10) and the method of Golub & Welsch.

Assuming first $[a, b]$ a finite interval, let $\{d\lambda_N(t)\}_{N=1}^{\infty}$ be a sequence of discrete N-point measures on $[a, b]$, approximating $d\lambda(t)$ in the sense that

$$\lim_{N\to\infty} \int_a^b p(t)d\lambda_N(t) = \int_a^b p(t)d\lambda(t) \tag{5.12}$$

for every polynomial p. The Jacobi matrix $J_{n,N}$ of order n, belonging to $d\lambda_N(t)$, then converges to J_n, the desired Jacobi matrix in (5.5), as $N \to \infty$,

$$\lim_{N\to\infty} J_{n,N} = J_n.$$

The following procedure, therefore, suggests itself: Select a suitable system of classical orthogonal polynomials $\{p_k\}$ and compute the corresponding modified moments

$$\nu_{k,N} = \int_a^b p_k(t)d\lambda_N(t), \qquad k = 0, 1, \ldots, 2n-1. \tag{5.13}$$

(These are easily obtained, since the integral in (5.13) is now a finite sum.) Apply algorithm (5.10) to generate the elements $\alpha_{k,N}$, $\beta_{k,N}$ of $J_{n,N}$. Increase N until $J_{n,N}$ sufficiently approximates J_n. Then obtain the desired Gauss–Christoffel formula from $J_{n,N} \approx J_n$, using the method described in Section 5.1.

The quality of this procedure depends crucially on the choice of the discretization $d\lambda_N(t)$ of $d\lambda(t)$. If, as in many applications, $d\lambda(t) = \omega(t)dt$, where ω is continuous and positive in the open interval (a, b), and integrable at both endpoints (although possibly singular there), then a discrete measure $d\lambda_N(t)$ may be obtained by applying a suitable N-point quadrature rule Q_N to the integral on the right of (5.12),

$$\int_a^b p(t)d\lambda_N(t) = Q_N(p\omega).$$

The condition (5.12) requires that Q_N be convergent when applied to $p\omega$, i.e. convergent even in the possible presence of endpoint singularities. Fortunately, most quadrature rules have this property, at least if the singularity is monotone,

or can be majorized by a monotone singularity (Bezikovič [1939], Rabinowitz [1967], [1970], [1977], [1979], Gautschi [1967], Feldstein & Miller [1971], Miller [1971], el-Tom [1971]; see also Freud [1971, Ch. 3, Thm. 1.6(b)], Mikloško [1970b], Esser [1971b]). A specific quadrature rule recommended by Gautschi [1968a] is the Fejér quadrature formula, i.e. the interpolatory quadrature rule based on the Chebyshev points on $[a, b]$. This often yields satisfactory convergence rates.

If the interval $[a, b]$ is infinite, it can be reduced to a finite interval by means of a suitable transformation of variables, whereupon the procedure described again applies (Gautschi [1968a]). For reasons of numerical stability, however, it is now advisable to compute the approximate Jacobi matrix $J_{n,N}$ by Stieltjes' procedure.

5.4. Numerical tables

A large number of numerical tables of Gauss-type quadrature rules have been prepared to assist the occasional user. They are summarized below in Tables 1–6. Early tables, later superseded by more extensive and more accurate ones, are not included in this summary. For convenience we divide the Gauss–Christoffel formulae into four groups (Tables 1–4), in accordance with the type of weight function involved. Gauss–Radau and Gauss–Lobatto formulae are collected in Table 5, where "R" in column 1 stands for "Radau", and "L" for "Lobatto". The letter "n" in the heading denotes the number of free nodes. Turán formulae are listed in Table 6. Here "n" means the number of distinct nodes, while "r" refers to the multiplicity of each. Throughout these tables we use the notation "$a(h)b$" to indicate the sequence of numbers $a, a+h, a+2h, \ldots, b$. If the step h is not constant, we write "var" in place of "h". The accuracy of the tables is indicated in terms of the number of significant digits (S) or the number of decimal digits after the decimal point (D), as appropriate.

Gauss–Christoffel formulae for Jacobi measures $d\lambda(t) = (1-t)^{\alpha}(1+t)^{\beta}dt$ with $\alpha = \pm 1/2$, $\beta = \pm 1/2$ are explicitly known in terms of trigonometric functions, hence need not be tabulated (cf. Section 1.4). The same is true for Gauss–Radau and Gauss–Lobatto formulae with Chebyshev measure $d\lambda(t) = (1-t^2)^{-1/2}dt$; see, e.g. Bouzitat [1952].

Tables of Gauss–Lobatto formulae having double nodes at the endpoints are given for $d\lambda(t) = dt$ in Gatteschi [1963/64].

Extensive tabulations for weight functions depending on a parameter can sometimes be avoided by expanding the nodes and weights in suitable series in that parameter, or by using other curve fitting procedures. It then suffices to tabulate the coefficients in the respective expansions or approximations. King & Dupuis [1976] adopt this approach for the measure $d\lambda(t) = e^{-xt^2}dt$ on

$[-1, 1]$, which is of interest in quantum mechanics, while Lambin & Vigneron [1979] provide series in Chebyshev polynomials for the Laguerre measure $d\lambda(t) = t^{\alpha}e^{-t}dt$ on $[0, \infty]$, $-1 < \alpha \leq 1$.

Ultimately, however, it is more productive to have high-quality computer software available for generating arbitrary Gauss-type formulae. Although, at the present time, this is still an elusive goal, computer programs with various degrees of generality and efficacy have been published; see, e.g., Rutishauser [1962b], Stroud & Secrest [1966], Gautschi [1968b], Golub & Welsch [1969], Davis & Rabinowitz [1975] and Laurie & Rolfes [1979]. A computer algorithm for the complex weight function $e^{\zeta}\zeta^{-s}$ on $[c - i\infty, c + i\infty]$ (cf. Section 3.1.3) is given in Piessens [1973b].

Acknowledgment.

The author gratefully acknowledges advice from Professor R.A. Askey on historical matters.

TABLE 1. Gauss–Jacobi formulae.

Weight Function	$[a, b]$	n	Accuracy	Reference
1	$[-1, 1]$	2(1)64(4)96(8)168 256, 384, 512	30S	Stroud & Secrest [1966]
$(1-t^2)^{\alpha}$ $\alpha = -1/2(1/2)3/2$ $(\alpha \neq 0)$	$[-1, 1]$	2(1)20	30S	"
$(1+t)^{\beta}$, $\beta = 1$	$[-1, 1]$	2(1)30	30S	"
$\beta = 2(1)4$	$[-1, 1]$	2(1)20	30S	"
$\lvert t\rvert^{\alpha}$, $\alpha = 1(1)4$	$[-1, 1]$	2(1)20	30S	"
t^{α} and $t^{\alpha}(1-t)^{\alpha}$ $\alpha = -.9(.1)3.$ $\alpha = -2/3(1/3)8/3$ $(\alpha \neq 0, 1, 2)$ $\alpha = -3/4(1/2)11/4$	$[0, 1]$	1(1)15	20S	Krylov & Vorob'eva [1971]
$t^{q-1}(1-t)^{p-q}$ $q = .1(.1)1.$ $p = (2q-1)(.1)(q+1)$	$[0, 1]$	2(1)15	15–16S	Glonti [1971]
t^2	$[0, 1]$	1(1)20	15D	Sprung & Hughes [1965]
$t^{\alpha-1}(1-t)^{\beta-1}$ $\alpha, \beta = 1/2$(var.)$3/2$	$[0, 1]$	2(1)12	12S	Boujot & Maroni [1968]
t^{α}, $\alpha = 0(1)5$	$[0, 1]$	1(1)8	12D	Fishman [1957]
$t^{\beta}(1-t)^{\alpha}$ $\alpha, \beta = -.9(.1)3.$, $\beta \leq \alpha$	$[0, 1]$	1(1)8	8S	Krylov et al. [1963]
$\lvert t\rvert^{\alpha}$, $\alpha = -3/4$(var.)$- 1/4$	$[-1, 1]$	1(1)8	8S	Bertova et al. [1953]
t^s, $s = 0(2)10$	$[-1, 1]$	2(1)4	7S	Rothmann [1961]
$s = 1(2)11$	$[-1, 1]$	2, 4	7S	"

Table 2. Gauss–Laguerre and Gauss–Hermite formulae.

Weight Function	$[a, b]$	n	Accuracy	Reference
e^{-t}	$[0, \infty]$	2(1)32(4)68	30S	Stroud & Secrest [1966]
e^{-t}	$[0, \infty]$	100, 150, 200, 300	24S	Berger & Danson [1968]
e^{-t}	$[0, \infty]$	400(100)900	23–24S	Berger et al. [1969]
e^{-t^2}	$[-\infty, \infty]$	2(1)64(4)96(8)136	30S	Stroud & Secrest [1966]
e^{-t^2} [1])	$[-\infty, \infty]$	300	15S	Afshar et al. [1973]
$\lvert t\rvert^{\alpha} e^{-t^2}$, $\alpha = 1, 2, 3$	$[-\infty, \infty]$	2(1)20	30S	Stroud & Secrest [1966]
$t^{\alpha} e^{-t}$, $\alpha = -.5(.5)10.$	$[0, \infty]$	4(4)16(8)32(16)64(32)128	25S	Shao et al. [1964b]
$\lvert t\rvert^{2\lambda} e^{-t^2}$, $\lambda = 0(1)10$	$[-\infty, \infty]$	8(8)32(16)64(32)128(64)256	25S	"
$t^{\alpha} e^{-t}$, $\alpha = -.5(1)3.5$	$[0, \infty]$	4, 8, 16, 32	25S	Shao et al. [1964a]
$\lvert t\rvert^{2\lambda} e^{-t^2}$, $\lambda = 0(1)4$	$[-\infty, \infty]$	8, 16, 32, 64	25S	"
$t^{s} e^{-t}$, $s = 1(1)5$	$[0, \infty]$	4(4)16	18S	Rabinowitz & Weiss [1959]
$t^{\alpha} e^{-t}$, $\alpha = 0(-.01)-.99$	$[0, \infty]$	2(1)16	15S	Dekanosidze [1966]
$\alpha = -.75, -.5, -.25$	$[0, \infty]$	1(1)15	15–17S	Concus et al. [1963]
$\alpha = -1/3, -2/3$	$[0, \infty]$	1(1)15	15–17S	Concus [1964]
$t^{s} e^{-t}$	$[0, \infty]$	1(1)15	8S	Aĭzenštat et al. [1962]
$s = -.9(.02)0(.05)3.$				
$s = -.75, -.25$				
$s = -2/3(1/3)8/3$				
$(s \neq 0, 1, 2)$				

[1]) This table gives the modified Christoffel numbers, i.e. the Christoffel numbers divided by the weight function evaluated at the respective node.

Table 3. Gauss–Christoffel formulae for power and logarithmic singularities.

Weight Function	$[a, b]$	n	Accuracy	Reference
$\ln(1/t)$	$[0, 1]$	2(1)16	30S	Stroud & Secrest [1966]
$t^{\alpha}[\ln(1/t)]^{m}$, $\alpha = 0, -.5$, $m = 1, 2, 3$	$[0, 1]$	2(1)20	30S	Kutt [1976]
$t^{\alpha} \ln(1/t)$	$[0, 1]$	3(var.)50	25S	Piessens & Branders [1975]
$\alpha = -1/2, -1/3, -1/4, -1/5, 1/3, 1/2$				
$\alpha = 0, -1/2$	$[0, 1]$	5, 10(10)100	20S	Branders & Piessens [1971]
$(1-t)^{\alpha} t^{\beta} \ln(1/t)$	$[0, 1]$	3(var.)50	25S	Piessens & Branders [1975]
$\alpha, \beta = -1/2, -1/3, -1/4, -1/5, 1/3, 1/2$				

$\ln(1/(1-t^2))$	$[-1,1]$	1(1)30	25*S*	Laurie & Rolfes [1977]
$\ln[(1+t)/(1-t)]$	$[-1,1]$	2(2)18	20*S*	Piessens et al. [1976]
$t^\alpha \ln(e/t)$, $\alpha = -.9(.01)0(.1)5$.	$[0,1]$	1(1)10	15*S*	Krylov & Pal'cev [1967]
$t^\alpha \ln(e/t)\ln(e/(1-t))$, $\alpha = 0(1)5$	$[0,1]$	1(1)10	15*S*	"
$t^\alpha e^{-t} \ln(1+1/t)$, $\alpha = 0(1)5$	$[0,\infty]$	1(1)10	15*S*	"
$t^{\alpha-1}[\ln(1/t)]^{\beta-1}$ $\alpha = 1/2(1/2)5/2$, $\beta = 1/2(\text{var.})2$	$[0,1]$	2(1)12	12*S*	Boujot & Maroni [1968]
$t^{-1}(1-t)^{\beta-1}/(\pi^2+\ln^2(t^{-1}-1))$ $\beta = 0, 1$	$[0,1]$	2(1)12	12*S*	"

TABLE 4. Miscellaneous Gauss–Christoffel formulae.

Weight function	$[a,b]$	n	Accuracy	Reference
$(t-a)^\alpha$ $a = -1.1, -1.01, -1.001$, $\alpha = -.5, -1, -2$	$[-1,1]$	3(var.)50	25*S*	Piessens & Branders [1975]
$(t^2+a^2)^\alpha$ $a = 1, .1, .01, .001$, $\alpha = -.5, -1, -2$	$[-1,1]$	3(var.)50	25*S*	"
$t^\gamma(1-t^\alpha)^\beta$ $\alpha = 3, 4, 6, 8$, $\beta = \pm 1/2$, $\gamma = 0, \pm 1/2$	$[0,1]$	2(2)8(4)16, 24	25*S*	Byrd & Galant [1970]
$\alpha = 2$, $\beta = -3/4, -2/3$, $\gamma = 0$				
$(1+t^2)^{-k-1}$, $k = 3(1)10^2)$	$[-\infty,\infty]$	4	10*D*	Harper [1962]
$k = 5(1)10$	$[-\infty,\infty]$	6	10*D*	"
$(1+t^2)^{-1}$	$[-1,1]$	2(1)7	7*S*	Reiz [1950b]
$(1-t^2)^{-1/2}(1+t^2)^{-1}$	$[-1,1]$	1(1)4	8*D*	Kumar [1974a]
$t(1+t)^{-13\ 3)}$	$[0,\infty]$	1(1)5	8*S*	Kumar & Jain [1974]
$1-\sqrt{t}$	$[0,1]$	1(1)10	5–15*S*	Struble [1960]
$(1-\sqrt{t})^2/2\sqrt{t}$	$[0,1]$	1(1)10	5–15*S*	"
$\cos t$ $\sin t$	$[-\pi,\pi]$	3(var.)50	25*S*	Piessens & Branders [1975]
$\sin t$	$[-\pi,\pi]$	2(2)18	16*S*	Piessens [1970b]
$\cos t$	$[-\pi/2,\pi/2]$	1(1)4	12*D*	Piessens [1970a]
$1+\frac{\cos}{\sin}(2\pi kt)$, $k = 1,2,3,5$	$[0,1]$	6, 8, 11, 13	10–15*D*	Mikloško [1970a]
$\frac{1}{2}\left(1+\frac{\cos}{\sin} m\pi t\right)$, $m = 1(1)12$	$[-1,1]$	1(1)8, 16, 32	12*D*	Gautschi [1970b]
$1-\frac{\cos}{\sin} kt$, $k = 1, 1024$	$[0,2\pi]$	1(1)4	6–7*S*	Zamfirescu [1963]

TABLE 4. (continued)

Weight Function	$[a, b]$	n	Accuracy	Reference
$\left(1+\begin{matrix}\cos\\ \sin\end{matrix} t\right)(1+t)^{-(2n-1+s)}$ $s = 1.05(.05)4.$	$[0, \infty]$	1(1)10	10S	Krylov & Kruglikova [1968]
$\lvert t\rvert^{\alpha} e^{-\lvert t\rvert}$, $\alpha = 1, 2, 3$	$[-\infty, \infty]$	2(1)20	30S	Stroud & Secrest [1966]
$t^{\alpha} e^{-at}$, $a = 1, 2, 5$, $\alpha = -.5, 0, .5$	$[0, 1]$	3(var.)50	25S	Piessens & Branders [1975]
e^{-at^2}, $a = 1, 2, 5, 10$	$[-1, 1]$	3(var.)50	25S	"
	$[0, 1]$	3(var.)50	25S	"
e^{-kt} $k = 2, 7$	$[-1, 1]$	2, 7	15S	Cecchi [1967]
$k = 2(1)16$	$[-1, 1]$	2(1)10	15S	Cecchi [no date]
$2\pi^{-1/2} e^{-t^2}$	$[0, \infty]$	1(1)20	20S	Galant [1969]
e^{-t^2}	$[0, \infty]$	2(1)15	15S	Steen et al. [1969]
	$[0, 1]$	2(1)10	15S	"
e^{-xt^2}, $x = 0, .5, 10$	$[-1, 1]$	10	20S	King & Dupuis [1976]
$t^{-m} e^{-t}$, $m = 0(1)10$	$[1, \infty]$	2(1)10	16S	Olson [1969]
$t^{\alpha}(t+1)^{-2n} e^{-t}$, $\alpha = -.5(.5)5$	$[0, \infty]$	1(1)10	10S	Pal'cev & Skoblja [1965]
$E_1(t)$ (exponential integral)	$[0, \infty]$	10, 20	12S	Danloy [1973]
$E_m(t)$, $m = 1(1)5$	$[0, \infty]$	2, 3	6–8S	Reiz [1950a]
$m = 1(1)3$	$[0, \infty]$	4	7–8S	"
$E_m(t)$, $m = 1, 2$	$[0, \tau]$ $\tau = .1(\text{var.})\infty$	3, 4	4–6S	Kegel [1962]
erfc t	$[0, \infty]$	2(1)12	12–16S	Vigneron & Lambin [1980]
$(-1)^s J_m(t)$ (Bessel function) $m = 0(1)2$, $s = 1(1)20$	$[j_{m,s-1}, j_{m,s}]$	2(2)8	14D	Piessens [1972b]
const $\cdot\, t^{-2/3} e^{-t} Ai((3t/2)^{2/3})$	$[0, \infty]$	1, 2, 4, 6	17S	Schulten et al. [1979]
$N(i, k; t)$, $k = 2, 4$, $i = 1(1)k$ (normalized B-spline of degree $k-1$)	$[-1, 1]$	1(1)17	14S	Phillips & Hanson [1974]
$(2\pi i)^{-1} p^{-1} e^{p}$	$[c - i\infty, c + i\infty]$	2(1)24	30S	Stroud & Secrest [1966]
$(2\pi i)^{-1} p^{-s} e^{p}$, $s = 1(1)5$	$[c - i\infty, c + i\infty]$	1(1)15	20S	Krylov & Skoblja [1968]
$s = .01(.01)3.(s \neq 1, 2, 3)$	$[c - i\infty, c + i\infty]$	1(1)10	7–8S	"
$s = .1(\text{var.})4$	$[c - i\infty, c + i\infty]$	6(1)12	16S	Piessens [1969a]

[2]) The orthogonal polynomial system is finite in this case. The Gauss–Christoffel nodes and weights are expressible in terms of Jacobi nodes and weights; see Haber [1964].

[3]) The orthogonal polynomial system is finite in this case.

TABLE 5. Gauss–Radau and Gauss–Lobatto formulae.

Weight Function	$[a, b]$	n	Accuracy	Reference
1 (R)	$[-1, 1]$	2(1)19(4)47	30*S*	Stroud & Secrest [1966]
e^{-t} (R)	$[0, \infty]$	3(1)5	20*S*	Stancu & Stroud [1963]
$t^s e^{-t}$, $s = 0, -1/3, -1/2, -2/3$ (R)	$[0, \infty]$	1(1)15	16*S*	Cassity [1965]
$s = -.99$(var.)10 (R)	$[0, \infty]$	1(1)15	16*S*	Cassity & Hopper [1964]
1 (L)	$[-1, 1]$	2(1)32(4)96	30*S*	Stroud & Secrest [1966]
1 (L)	$[-1, 1]$	1(1)14(8)46(16)94	20*D*	Michels [1963]
1 (L)	$[-1, 1]$	3(4)23(8)47, 63	19*D*	Rabinowitz [1960]
$\sqrt{t}$ (R)	$[0, 1]$	1(1)5	8*D*	Akkerman [1959]
t and t^2 (L)	$[0, 1]$	1(1)4	8*D*	"

TABLE 6. Turán formulae.

Weight Function	$[a, b]$	n	r	Accuracy	Reference
1	$[-1, 1]$	2(1)7	3, 5	20*S*	Stroud & Stancu [1965]
1 [4])	$[-1, 1]$	2(1)9	3	11*S*	Lo Cascio [1973]
		2(1)7	5		
		2(1)5	7		
1 (L)	$[-1, 1]$	4(1)7	3, 5	12–16*S*	Rebolia [1973]
e^{-t}	$[0, \infty]$	1(1)3	3, 5	20*S*	Stroud & Stancu [1965]
e^{-t}	$[0, \infty]$	1	3(2)23	12*S*	Verna [1969]
e^{-t^2}	$[-\infty, \infty]$	2(1)7	3, 5	20*S*	Stroud & Stancu [1965]
e^{-t^2}	$[-\infty, \infty]$	2(1)3	3(2)7	12*S*	Verna [1969]

[4]) Only the nodes are tabulated. Some of the tabular entries are inaccurate.

BIBLIOGRAPHY

Afshar, R., Mueller, F.M. and Shaffer, J.C. [1973]: *Hilbert transformation of densities of states using Hermite functions.* J. Computational Phys. *11*, 190–209.

Aĭzenštat, V.S., Krylov, V.I. and Metel'skiĭ, A.S. [1962]: *Tables for the Numerical Laplace Transform and the Evaluation of Integrals of the Form* $\int_0^\infty x^s e^{-x} f(x)dx$ (Russian). Izdat. Akad. Nauk BSSR, Minsk.

Akkerman, R.B. [1959]: *Quadrature formulas of the type of Markov formulas* (Russian). Trudy Mat. Inst. Steklov *53*, 5–15.

Bahvalov, N.S. and Vasil'eva, L.G. [1968]: *The calculation of integrals of oscillatory functions by interpolation at the Gaussian nodes* (Russian). Ž. Vyčisl. Mat. i Mat. Fiz. *8*, 175–181.

Baillaud, B. and Bourget, H. [1905]: *Correspondance d'Hermite et de Stieltjes* I, II. Gauthier-Villars, Paris.

Bandemer, H. [1966]: *Erweiterung der Stekloffschen Sätze über mechanische Quadraturverfahren.* Wiss. Z. Techn. Hochsch. Karl-Marx-Stadt *8*, 205–208.

—— [1967]: *Über verallgemeinerte Interpolationsquadratur.* Math. Nachr. *34*, 379–387.

Baratella, P. [1979]: *Un'estensione ottimale della formula di quadratura di Radau.* Rend. Sem. Mat. Univ. e Politec. Torino *37*, 147–158.

Barnhill, R.E. [1968]: *Asymptotic properties of minimum norm and optimal quadratures.* Numer. Math. *12*, 384–393.

Barrett, W. [1960/61]: *Convergence properties of Gaussian quadrature formulae.* Comput. J. *3*, 272–277.

Barrow, D.L. [1976]: *Existence of Gauss interpolation formulas for the one-dimensional heat equation.* Math. Comp. *30*, 24–34.

——— [1977]: *Gauss interpolation formulas and totally positive kernels.* Math. Comp. *31*, 984–993.

——— [1978]: *On multiple node Gaussian quadrature formulae.* Math. Comp. *32*, 431–439.

——, Stroud, A.H. [1976]: *Existence of Gauss harmonic interpolation formulas.* SIAM J. Numer. Anal. *13*, 18–26.

Berger, B.S. and Danson, R. [1968]: *Tables of zeros and weights for Gauss–Laguerre quadrature.* Math. Comp. *22*, 458–459.

——, ——, Carpenter, R. [1969]: *Tables of zeros and weights for Gauss–Laguerre quadrature to 24S for N = 400, 500 and 600, and tables of zeros and weights for Gauss–Laguerre quadrature to 23S for N = 700, 800 and 900.* Math. Comp. *23*, Review 60, 882.

Bernstein, S. [1918]: *Quelques remarques sur l'interpolation.* Math. Ann. *79*, 1–12.

——— [1930]: *Sur les polynomes orthogonaux relatifs à un segment fini.* J. Math. Pures Appl. (9) *9*, 127–177.

Bertova, E.I., Kuznecov, Ja.T., Natanson, I.P. and Caregradskiĭ, H.A. [1953]: *Approximate computation of definite integrals by means of the multiplicative method of isolating the singularity* (Russian). Prikl. Mat. i Meh. *17*, 639–644.

Bezikovič, Ja.S. [1937]: *On formulas of mechanical quadrature with n ordinates, exact for polynomials of degree not higher than 2n − 2 and 2n − 3* (Russian). Trudy Leningr. Indust. Inst. No. 4, 1–18.

——— [1939]: *Process of mechanical quadratures for improper integrals* (Russian). Leningrad. Gos. Univ. Uč. Zap. Ser. Mat. Nauk *6*, 36–42.

Bilharz, H. [1951]: *Über die Gaußsche Methode zur angenäherten Berechnung bestimmter Integrale.* Math. Nachr. *6*, 171–192.

Billauer, A. [1974]: *On Gaussian quadrature by divided differences of a modified function.* BIT *14*, 359–361.

Blue, J.L. [1979]: *A Legendre polynomial integral.* Math. Comp. *33*, 739–741.

Boland, W.R. [1972]: *The convergence of product-type quadrature formulas.* SIAM J. Numer. Anal. *9*, 6–13.

——— [1973]: *Properties of product-type quadrature formulas.* BIT *13*, 287–291.

——, Duris, C.S. [1971]: *Product-type quadrature formulas.* BIT *11*, 139–158.

Boujot, J.-P. and Maroni, P. [1968]: *Algorithme général de construction de tables de Gauss pour les problèmes de quadratures.* Institut Blaise Pascal, Publ. n° NMX/8.1.8/AI, Paris.

Bouzitat, J. [1952]: *Intégration numérique approchée par la méthode de Gauss généralisée et extension de cette méthode.* In: Mineur, H., *Techniques de Calcul Numérique*, pp. 557–605. Beranger, Paris.

Branders, M. [1976]: *Application of Chebyshev polynomials in numerical integration* (Flemish). Dissertation, Catholic University of Leuven, Leuven.

———, Piessens, R. [1971]: *Gaussian quadrature formulas for integrals with logarithmic singularity.* Report TW8, Appl. Math. Progr. Div., Catholic University of Leuven, Leuven.

———, ——— [1975]: *An extension of Clenshaw-Curtis quadrature.* J. Comput. Appl. Math. *1*, 55–65.

Brass, H. [1977]: *Quadraturverfahren.* Vandenhoeck & Ruprecht, Göttingen.

——— [1978]: *Monotonie bei den Quadraturverfahren von Gauss und Newton–Cotes.* Numer. Math. *30*, 349–354.

de Bruin, M.G., Saff, E.B. and Varga, R.S. [to appear]: *On the zeros of generalized Bessel polynomials.* Indag. Math.

Butzer, P.L. [1979/80]: *The Banach–Steinhaus theorem with rates, and applications to various branches of analysis.* In: General Inequalities II (Ed. by Beckenbach, E.F.), Birkhäuser, Basel.

———, Scherer, K. and Westphal, U. [1973]: *On the Banach–Steinhaus theorem and application in locally convex spaces.* Acta Sci. Math. (Szeged) *34*, 25–34.

Byrd, P.F. and Galant, D.C. [1970]: *Gauss quadrature rules involving some nonclassical weight functions.* NASA Techn. Note D-5785, Ames Research Center, Moffett Field, California.

Capovani, M., Ghelardoni, G. and Lombardi, G. [1976a]: *Utilizzazione di proprietà delle matrici per lo studio dei polinomi ultrasferici.* Rend. Mat. *9*, 57–69.

———, ———, ——— [1976b]: *Un metodo per il calcolo dei nodi e dei pesi delle formule di quadratura di Gauss–Hermite.* Calcolo *13*, 441–452.

Cassity, C.R. [1965]: *Abscissas, coefficients, and error term for the generalized Gauss–Laguerre quadrature formula using the zero ordinate.* Math. Comp. *19*, 287–296.

——, Hopper, H.R. [1964]: *Abscissas and coefficients for the generalized Gauss–Laguerre quadrature formula using the zero ordinate.* NASA Report TMX-53099, George C. Marshall Space Flight Center, Huntsville, Alabama.

Cecchi, M.M. [1967]: *L'integrazione numerica di una classe di integrali utili nei calcoli quantomecanici.* Calcolo *4*, 363–368.

——— [no date]: *Tavole di nodi e pesi per* $\int_{-1}^{1} e^{-kx} f(x)dx \simeq \Sigma_{i=1}^{m} A_i f(x_i)$. Ist. Elaborazione della Informazione, Felici, Pisa.

Chakalov, L. [1930/31]: *Sur les formules de quadrature à nombre minimum de termes.* Bull. Math. Phys. Bucarest *2*, 160–163.

——— [1954]: *General quadrature formulae of Gaussian type* (Bulgarian). Bulgar. Akad. Nauk Izv. Mat. Inst. *1*, 67–84.

——— [1957]: *Formules générales de quadrature mécanique du type de Gauss.* Colloq. Math. *5*, 69–73.

Chawla, M.M. [1967]: *On the Chebyshev polynomials of the second kind.* SIAM Rev. *9*, 729–733.

——— [1968]: *Error bounds for the Gauss–Chebyshev quadrature formula of the closed type.* Math. Comp. *22*, 889–891.

——— [1968/69]: *Asymptotic estimates for the error of the Gauss–Legendre quadrature formula.* Comput. J. *11*, 339–340.

——— [1969]: *On Davis' method for the estimation of errors of Gauss–Chebyshev quadratures.* SIAM J. Numer. Anal. *6*, 108–117.

——— [1970a]: *Estimation of errors of Gauss–Chebyshev quadratures.* Comput. J. *13*, 107–109.

——— [1970b]: *Hilbert spaces for estimating errors of quadratures for analytic functions.* BIT *10*, 145–155.

——— [1971a]: *Asymptotic Gauss quadrature errors as Fourier coefficients of the integrand.* J. Austral. Math. Soc. *12*, 315–322.

——— [1971b]: *Estimating errors of numerical approximation for analytic functions.* Numer. Math. *16*, 370–374.

——, Jain, M.K. [1968a]: *Error estimates for Gauss quadrature formulas for analytic functions.* Math. Comp. *22*, 82–90.

———, ——— [1968b]: *Asymptotic error estimates for the Gauss quadrature formula.* Math. Comp. *22*, 91–97.

———, Jayarajan, N. [1975]: *Quadrature formulas for Cauchy principal value integrals.* Computing *15*, 347–355.

———, Kumar, S. [1978]: *Convergence of Fejér type quadrature formulas for Cauchy principal value integrals.* J. Math. Phys. Sci. *12*, 49–59.

———, ——— [1979]: *Convergence of quadratures for Cauchy principal value integrals.* Computing *23*, 67–72.

———, Ramakrishnan, T.R. [1974]: *Modified Gauss–Jacobi quadrature formulas for the numerical evaluation of Cauchy-type singular integrals.* BIT *14*, 14–21.

Chebyshev, P.L. [1855]: *On continued fractions* (Russian). Učen. Zap. Imp. Akad. Nauk *3*, 636–664. [French translation: J. Math. Pures Appl. (2) *3* (1858), 289–323. Oeuvres I, 203–230.]

——— [1859a]: *Sur l'interpolation par la méthode des moindres carrés.* Mém. Acad. Impér. Sci. St. Pétersbourg (7) *1*, no. 15, 1–24. [Oeuvres I, 473–498.]

——— [1859b]: *Sur le développement des fonctions à une seule variable.* Bull. Phys.-Math. Acad. Imp. Sci. St. Pétersbourg *1*, 193–200. [Oeuvres I, 501–508].

Chisholm, J.S.R., Genz, A. and Rowlands, G.E. [1972]: *Accelerated convergence of sequences of quadrature approximations.* J. Computational Phys. *10*, 284–307.

Christoffel, E.B. [1858]: *Über die Gaußische Quadratur und eine Verallgemeinerung derselben.* J. Reine Angew. Math. *55*, 61–82. [Ges. Math. Abhandlungen I, 65–87.]

——— [1877]: *Sur une classe particulière de fonctions entières et de fractions continues.* Ann. Mat. Pura Appl. (2) *8*, 1–10 [Ges. Math. Abhandlungen II, 42–50.]

Chui, C.K. [1972]: *Concerning Gaussian–Chebyshev quadrature errors.* SIAM J. Numer. Anal. *9*, 237–240.

Cohen, H. [1978]: *Accurate numerical solutions of integral equations with kernels containing poles.* J. Computational Phys. *26*, 257–276.

Concus, P. [1964]: *Additional tables for the evaluations of* $\int_0^\infty x^\beta e^{-x} f(x)dx$ *by Gauss–Laguerre quadrature.* Math. Comp. *18*, Review 81, 523.

———, Cassatt, D., Jaehnig, G., Melby, E. [1963]: *Tables for the evaluation of* $\int_0^\infty x^\beta e^{-x} f(x)dx$ *by Gauss–Laguerre quadrature.* Math. Comp. *17*, 245–256.

Cosma Cagnazzi, L. [1970]: *Su un nuovo tipo di formule di maggiorazione del resto di una formula di quadratura.* Rend. Mat. (6) *3*, 203–210.

Cranley, R. and Patterson, T.N.L. [1971]: *On the automatic numerical evaluation of definite integrals.* Comput. J. *14*, 189–198.

Crout, P.D. [1929/30]: *The approximation of functions and integrals by a linear combination of functions.* J. Math. Phys. (M.I.T.) *9*, 278–314.

Curtis, A.R. and Rabinowitz, P. [1972]: *On the Gaussian integration of Chebyshev polynomials.* Math. Comp. *26*, 207–211.

Danloy, B. [1973]: *Numerical construction of Gaussian quadrature formulas for* $\int_0^1 (-\mathrm{Log}\, x)x^\alpha f(x)dx$ *and* $\int_0^\infty E_m(x)f(x)dx$. Math. Comp. *27*, 861–869.

Darboux, G. [1878]: *Mémoire sur l'approximation des fonctions de très-grands nombres et sur une classe étendue de développements en série.* J. Math. Pures Appl. (3) *4*, 5–56, 377–416.

Davis, P.J. [1953]: *Errors of numerical approximation for analytic functions.* J. Rational Mech. Anal. *2*, 303–313.

——— [1962]: *Errors of numerical approximation for analytic functions.* In: Survey of Numerical Analysis (Ed. by Todd, J.), 468–484. McGraw-Hill, New York.

———, Rabinowitz, P. [1954]: *On the estimation of quadrature errors for analytic functions.* Math. Tables Aids Comput. *8*, 193–203.

———, ——— [1975]: *Methods of Numerical Integration.* Academic Press, New York.

Dekanosidze, E.N. [1966]: *Tables of Roots and Weight-factors of Generalized Laguerre Polynomials* (Russian). Vyčisl. Centr. Akad. Nauk SSSR, Moscow.

Delves, L.M. [1967/68]: *The numerical evaluation of principal value integrals.* Comput. J. *10*, 389–391.

Deruyts, J. [1886]: *Sur le calcul approché de certaines intégrales définies.* Bull. Acad. Roy. Belgique (3) *11*, 307–311.

Donaldson, J.D. [1973]: *Estimates of upper bounds for quadrature errors.* SIAM J. Numer. Anal. *10*, 13–22.

———, Elliott, D. [1972]: *A unified approach to quadrature rules with asymptotic estimates of their remainders.* SIAM J. Numer. Anal. *9*, 573–602.

Dyn, N. [1979]: *On the existence of Hermite–Birkhoff quadrature formulas of Gaussian type.* MRC Tech. Summary Rep. # 1978, Mathematics Research Center, University of Wisconsin, Madison, Wisconsin.

Elliott, D. [1979]: *On the convergence of Hunter's quadrature rule for Cauchy principal value integrals.* BIT *19*, 457–462.

———, Paget, D.F. [1975]: *On the convergence of a quadrature rule for evaluating certain Cauchy principal value integrals.* Numer. Math. *23*, 311–319.

———, ——— [1976a]: *On the convergence of a quadrature rule for evaluating certain Cauchy principal value integrals: an addendum.* Numer. Math. *25*, 287–289.

———, ——— [1976b]: *Product-integration rules and their convergence.* BIT *16*, 32–40.

———, ——— [1978]: *The convergence of product integration rules.* BIT *18*, 137–141.

———, ——— [1979]: *Gauss type quadrature rules for Cauchy principal value integrals.* Math Comp. *33*, 301–309.

Engels, H. [1972]: *Über allgemeine Gaußsche Quadraturen.* Computing *10*, 83–95.

——— [1973]: *Spezielle Interpolationsquadraturen vom Gauss'schen Typ.* In: Numerische, insbesondere approximationstheoretische Behandlung von Funktionalgleichungen (Ed. by Ansorge, R. and Törnig, W.), 54–68. Lecture Notes in Math. *333*, Springer, Berlin.

——— [1974]: *Positive Interpolationsquadraturen mit teilweise vorgeschriebenen Stützstellen.* Computing *13*, 121–141.

——— [1977]: *Eine Familie interpolatorischer Quadraturformeln mit ableitungsfreien Fehlerschranken.* Numer. Math. *28*, 49–58.

Erdogan, F. and Gupta, G.D. [1971/72]: *On the numerical solution of singular integral equations.* Quart. Appl. Math. *29*, 525–534.

Esser, H. [1971a]: *Konvergenz von Quadraturverfahren vom Radau-Typ.* Computing *7*, 254–263.

——— [1971b]: *Bemerkungen zu einem Satz von G. Freud über Quadraturkonvergenz.* Computing *8*, 216–220.

——— [1972]: *Konvergenz von Quadraturverfahren vom Gaußschen Typ.* Studia Sci. Math. Hungar. *7*, 375–378.

Feldstein, A. and Miller, R.K. [1971]: *Error bounds for compound quadrature of weakly singular integrals.* Math. Comp. *25*, 505–520.

Filippi, S. and Esser, H. [1970]: *Darstellungs- und Konvergenzsätze für Quadraturverfahren auf C und C^m.* Forschungsber. des Landes Nordrhein-Westfalen No. 2137. Westdeutscher Verlag, Köln.

Fishman, H. [1957]: *Numerical integration constants.* Math. Tables Aids Comput. *11*, 1–9.

Fock, V. [1932]: *On the remainder term of certain quadrature formulae* (Russian). Bull. Acad. Sci. Leningrad (7), 419–448.

Freud, G. [1971]: *Orthogonal Polynomials.* Pergamon Press, New York.

——— [1973]: *Error estimates for Gauss–Jacobi quadrature formulae.* In: Topics in Numerical Analysis (Ed. by Miller, J.J.H.), 113–121. Academic Press, London.

——— [1975a]: *Numerical estimates for the error of Gauss–Jacobi quadrature formulae.* In: Topics in Numerical Analysis II (Ed. by Miller, J.J.H.), 43–50. Academic Press, London.

——— [1975b]: *Error estimates for Gauss–Jacobi quadrature formulae and some applications of them* (Hungarian). Alkalmaz. Mat. Lapok *1*, 23–36.

Gabdulhaev, B.G. [1975]: *Cubature formulae for multidimensional singular integrals II* (Russian). Izv. Vyss. Učebn. Zaved. Matematika, no. 4 (155), 3–13.

——— [1976]: *Quadrature formulae with multiple nodes for singular integrals* (Russian). Dokl. Akad. Nauk SSSR *227*, 531–534.

———, Onegov, L.A. [1976]: *Cubature formulae for singular integrals* (Russian). Izv. Vyss. Učebn. Zaved. Matematika, no. 7 (170), 100–105.

Gahov, R.D. [1958]: *Boundary Problems* (Russian). Izdat. Fiz. -Mat. Lit., Moscow. [2nd ed., ibid., 1963. English translation: Pergamon Press, Oxford, 1966.]

Gaier, D. [1964]: *Konstruktive Methoden der konformen Abbildung*. Springer, Berlin.

Galant, D. [1969]: *Gauss quadrature rules for the evaluation of* $2\pi^{-1/2}\int_0^\infty \exp(-x^2)f(x)dx$. Math. Comp. *23*, Review 42, 676–677. Loose microfiche suppl. E.

——— [1971]: *An implementation of Christoffel's theorem in the theory of orthogonal polynomials*. Math. Comp. *25*, 111–113.

Gatteschi, L. [1963/64]: *Su una formula di quadratura "quasi gaussiana". Tabulazione delle ascisse d'integrazione e delle relative costanti di Christoffel*. Atti Accad. Sci. Torino Cl. Sci. Fis. Mat. Natur. *98*, 641–661.

——— [1979]: *On the construction of some Gaussian quadrature rules*. In: Numerische Integration. ISNM *45* (Ed. by Hämmerlin, G.), 138–146. Birkhäuser, Basel.

——— [1980]: *On some orthogonal polynomial integrals*. Math. Comp. *35*, 1291–1298.

———, Monegato, G. and Vinardi, G. [1976]: *Alcuni problemi relativi alle formule di quadratura del tipo di Tchebycheff*. Calcolo *13*, 79–104.

Gauss, C.F. [1812]: *Disquisitiones generales circa seriem infinitam* $1+((\alpha\cdot\beta)/(1\cdot\gamma))x+\cdots$ *etc.* Commentationes Societatis Regiae Scientarium Gottingensis Recentiores *2*. [Werke III, 123–162.]

——— [1814]: *Methodus nova integralium valores per approximationem inveniendi*. Commentationes Societatis Regiae Scientarium Gottingensis Recentiores *3*. [Werke III, 163–196.]

Gautschi, W. [1967]: *Numerical quadrature in the presence of a singularity*. SIAM J. Numer. Anal. *4*, 357–362.

——— [1968a]: *Construction of Gauss–Christoffel quadrature formulas*. Math. Comp. *22*, 251–270.

——— [1968b]: *Algorithm 331 — Gaussian quadrature formulas*. Comm. ACM *11*, 432–436.

——— [1970a]: *On the construction of Gaussian quadrature rules from modified moments*. Math. Comp. *24*, 245–260.

——— [1970b]: *Tables of Gaussian quadrature rules for the calculation of Fourier coefficients*. Math. Comp. *24*, no. 110, loose microfiche suppl. A–D.

——— [1978]: *Questions of numerical condition related to polynomials*. In: Recent Advances in Numerical Analysis (Ed. by de Boor, C. and Golub, G.H.), 45–72. Academic Press, New York.

——— [1979a]: *On the preceding paper "A Legendre polynomial integral" by James L. Blue*. Math. Comp. *33*, 742–743.

——— [1979b]: *On generating Gaussian quadrature rules*. In: Numerische Integration. ISNM *45* (Ed. by Hämmerlin, G.), 147–154. Birkhäuser, Basel.

Ghizzetti, A. [1954/55]: *Sulle formule di quadratura*. Rend. Sem. Mat. Fis. Milano *26*, 1–16.

———, Ossicini, A. [1967]: *Su un nuovo tipo di sviluppo di una funzione in serie di polinomi*. Rend. Accad. Naz. Lincei (8) *43*, 21–29.

———, ——— [1970]: *Quadrature Formulae*. Academic Press, New York.

———, ——— [1974]: *Polinomi s-ortogonali e sviluppi in serie ad essi collegati*. Mem. Accad. Sci. Torino Cl. Sci. Fis. Mat. Nat. (4), no. 18, 1–16.

———, ——— [1974/75]: *Generalizzazione dei polinomi s-ortogonali e dei corrispondenti sviluppi in serie*. Atti Accad. Sci. Torino *109*, 371–379.

———, ——— [1975]: *Sull' esistenza e unicità delle formule di quadratura gaussiane*. Rend. Mat. (6) *8*, 1–15.

Glonti, È.N. [1971]: *Tables of Roots and Quadrature Coefficients of Jacobi Polynomials* (Russian). Vyčisl. Centr Akad. Nauk SSSR, Moscow.

Golub, G.H. [1973]: *Some modified matrix eigenvalue problems.* SIAM Rev. *15*, 318–334.

———, Welsch, J. H. [1969]: *Calculation of Gauss quadrature rules.* Math. Comp. *23*, 221–230. Loose microfiche suppl. A1–A10.

Gordon, R.G. [1968]: *Error bounds in equilibrium statistical mechanics.* J. Mathematical Phys. *9*, 655–663.

Gourier, G. [1883]: *Sur une méthode capable de fournir une valeur approchée de l'intégrale* $\int_{-\infty}^{\infty} F(x)dx$. C.R. Acad. Sci. Paris *97*, 79–82.

Gribble, J.D. [1977]: *Further properties of inner product quadrature formulas.* BIT *17*, 392–408. [Erratum, ibid. *20* (1980), 260.]

Grinšpun, Z.S. [1966]: *On a class of orthogonal polynomials* (Russian). Vestnik Leningrad. Univ. *21*, no. 19, 147–149.

Grosswald, E. [1978]: *Bessel Polynomials.* Lecture Notes in Math. *698*, Springer, Berlin.

Haber, S. [1964]: *A note on some quadrature formulas for the interval* $(-\infty, \infty)$. Math. Comp. *18*, 313–314.

——— [1971]: *On certain optimal quadrature formulas.* J. Res. Nat. Bur. Standards *75B*, 85–88.

——— [1971/72]: *The error in numerical integration of analytic functions.* Quart. Appl. Math. *29*, 411–420.

Hammer, P.C. and Wicke, H.H. [1960]: *Quadrature formulas involving derivatives of the integrand.* Math. Comp. *14*, 3–7.

Hämmerlin, G. [1972]: *Fehlerabschätzung bei numerischer Integration nach Gauss.* Methoden und Verfahren der mathematischen Physik *6*, 153–163. B.I.-Hochschultaschenbücher, No. 725, Bibliographisches Institut, Mannheim.

Harper, W.M. [1962]: *Quadrature formulas for infinite intervals.* Math. Comp. *16*, 170–175.

Harris, C.G. and Evans, W.A.B. [1977/78]: *Extension of numerical quadrature formulae to cater for end point singular behaviours over finite intervals.* Internat. J. Comput. Math. *6*, 219–227.

Heine, E. [1881]: *Anwendungen der Kugelfunctionen und der verwandten Functionen.* 2nd ed., Reimer, Berlin. [I. Theil: Mechanische Quadratur, 1–31.]

Hermite, C. [1864]: *Sur un nouveau développement en série des fonctions.* C.R. Acad. Paris *58*, 93–100, 266–273. [Oeuvres II, 293–308.]

Heun, K. [1900]: *Neue Methode zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen.* Z. Math. Phys. *45*, 23–38.

Hunter, D.B. [1972]: *Some Gauss-type formulae for the evaluation of Cauchy principal values of integrals.* Numer. Math. *19*, 419–424.

Ioakimidis, N.I. and Theocaris, P.S. [1977a]: *On the numerical solution of a class of singular integral equations.* J. Math. Phys. Sci. *11*, 219–235.

———, ——— [1977b]: *On the numerical evaluation of Cauchy principal value integrals.* Rev. Roumaine Sci. Techn. Sér. Méc. Appl. *22*, 803–818.

———, ——— [1978a]: *Numerical solution of Cauchy type singular integral equations by use of the Lobatto–Jacobi numerical integration rule.* Apl. Mat. *23*, 439–452.

———, ——— [1978b]: *The numerical evaluation of a class of generalized stress intensity factors by use of the Lobatto–Jacobi numerical integration rule.* Internat. J. Fracture *14*, 469–484.

———, ——— [1979]: *On the numerical solution of singular integro-differential equations.* Quart. Appl. Math. *37*, 325–331.

———, ——— [1980]: *On convergence of two direct methods for solution of Cauchy type singular integral equations of the first kind.* BIT *20*, 83–87.

Ionescu, D.V. [1951]: *Some formulas of mechanical quadrature* (Romanian). Acad. R.P. Romîne Fil. Cluj. Stud. Cerc. Mat., 16–37.

——— [1967]: *Restes des formules de quadrature de Gauss et de Turán.* Acta Math. Acad. Sci. Hungar. *18*, 283–295.

Ivanova, A.N. [1955]: *On convergence of sequences of quadrature formulas of Gauss type on an infinite interval* (Russian). Dokl. Akad. Nauk SSSR *104*, 169–172.

Jacobi, C.G.J. [1826]: *Ueber Gauß neue Methode, die Werthe der Integrale näherungsweise zu finden.* J. Reine Angew. Math. *1*, 301–308.

——— [1859]: *Untersuchungen über die Differentialgleichung der hypergeometrischen Reihe.* J. Reine Angew. Math. *56*, 149–165. [Math. Werke III, 97–113.]

Jayarajan, N. [1974]: *Error estimates for Gauss–Chebyshev and Clenshaw–Curtis quadrature formulas.* Calcolo *11*, 289–296.

Johnson, L.W. and Riess, R.D. [1979]: *Gauss harmonic interpolation formulas for circular regions.* J. Inst. Math. Appl. *23*, 215–222.

Johnson, W.W. [1915]: *On Cotesian numbers: Their history, computation, and values to* $n = 20$. Quart. J. Pure Appl. Math. *46*, 52–65.

Jouravsky, A. [1928]: *Sur la convergence des formules des quadratures mécaniques dans un intervalle infini.* J. Soc. Phys.-Math. Leningrad *2*, 31–52.

Kahaner, D.K. and Monegato, G. [1978]: *Nonexistence of extended Gauss–Laguerre and Gauss–Hermite quadrature rules with positive weights.* Z. Angew. Math. Phys. *29*, 983–986.

Kambo, N.S. [1970]: *Error of the Newton–Cotes and Gauss–Legendre quadrature formulas.* Math. Comp. *24*, 261–269.

——— [1970/71]: *Error bounds for the Lobatto and Radau quadrature formulas.* Numer. Math. *16*, 383–388.

——— [1971]: *Error of certain Gauss quadrature formulas.* J. Inst. Math. Appl. *7*, 303–307.

Karlin, S. and Pinkus, A. [1976a]: *Gaussian quadrature formulae with multiple nodes.* In: Studies in Spline Functions and Approximation Theory (Ed. by Karlin, S., Micchelli, C.A., Pinkus, A. and Schoenberg, I.J.), 113–141. Academic Press, New York.

———, ——— [1976b]: *An extremal property of multiple Gaussian nodes.* In: Studies in Spline Functions and Approximation Theory (Ed. by Karlin, S., Micchelli, C.A., Pinkus, A. and Schoenberg, I.J.), 143–162. Academic Press, New York.

——, Studden, W.J. [1966]: *Tchebycheff Systems: With Applications in Analysis and Statistics.* Interscience, New York.

Kas'janov, V.I. [1977]: *The finite-dimensional approximation of singular integrals over the real line* (Russian). Izv. Vyšš. Učebn. Zaved. Matematika, no. 11 (186), 27–33.

Kastlunger, K. and Wanner, G. [1972]: *On Turán type implicit Runge–Kutta methods.* Computing *9*, 317–325.

Keda, N.P. [1961a]: *On the theory of quadrature for periodic functions* (Russian). Dokl. Akad. Nauk BSSR *5*, 375–379.

——— [1961b]: *Quadrature formulae with derivatives for periodic functions* (Russian). Vesci Akad. Navuk BSSR Ser. Fiz.-Tehn. Navuk, no. 4, 38–44.

Kegel, W.H. [1962]: *Zur numerischen Berechnung der Integrale* $\int_0^\tau f(x)K_n(x)dx$. Z. Astrophys. *54*, 34–40.

King, H.F. and Dupuis, M. [1976]: *Numerical integration using Rys polynomials.* J. Computational Phys. *21*, 144–165.

Kis, O. [1957]: *Remark on mechanical quadrature* (Russian). Acta Math. Acad. Sci. Hungar. *8*, 473–476.

Knauff, W. [1976/77]: *Fehlernormen zur Quadratur analytischer Funktionen.* Computing *17*, 309–322.

Knight, C.J. and Newbery, A.C.R. [1970]: *Trigonometric and Gaussian quadrature.* Math. Comp. *24*, 575–581.

Kofroň, J. [1972]: *Die ableitungsfreien Fehlerabschätzungen von Quadraturformeln I.* Apl. Mat. *17*, 39–52.

Korneĭčuk, A.A. [1964]: *Quadrature formulae for singular integrals* (Russian). Ž. Vyčisl. Mat. i Mat. Fiz. *4*, no. 4, suppl., 64–74.

Kowalewski, A. [1917]: *Newton, Cotes, Gauss, Jacobi: Vier grundlegende Abhandlungen über Interpolation und genäherte Quadratur.* von Veit, Leipzig.

Krall, H.L. and Frink, O. [1949]: *A new class of orthogonal polynomials: The Bessel polynomials.* Trans. Amer. Math. Soc. *65*, 100–115.

Krenk, S. [1975/76]: *On quadrature formulas for singular integral equations of the first and second kind.* Quart. Appl. Math. *33*, 225–232.

——— [1978]: *Quadrature formulae of closed type for solution of singular integral equations.* J. Inst. Math. Appl. *22*, 99–107.

Kronecker, L. [1894]: *Vorlesungen über die Theorie der einfachen und vielfachen Integrale.* Leipzig.

Kronrod, A.S. [1964a]: *Integration with control of accuracy* (Russian). Dokl. Akad. Nauk SSSR *154*, 283–286.

——— [1964b]: *Nodes and Weights for Quadrature Formulae. Sixteen-place Tables* (Russian). Izdat. "Nauka", Moscow. [English translation: Consultants Bureau, New York, 1965.]

Kruglikova, L.G. and Krylov, V.I. [1961]: *A numerical Fourier transform* (Russian). Dokl. Akad. Nauk BSSR *5*, 279–283.

Krylov, V.I. [1959]: *Approximate Calculation of Integrals* (Russian). Izdat. Fiz.-Mat. Lit., Moscow. [2nd ed., Izdat. "Nauka", Moscow, 1967. English translation of 1st ed.: McMillan, New York, 1962.]

———, Fedenko, N.P. [1962]: *Approximate representation of the integral* $\int_0^\infty x^s e^{-x} f(x)dx$ *by mechanical quadrature containing the value* $f(0)$ (Russian). Vesci Akad. Navuk BSSR Ser. Fiz.-Tehn. Navuk, no. 2, 5–9.

———, Kruglikova, L.G. [1968]: *A Handbook on Numerical Harmonic Analysis* (Russian). Izdat. "Nauka i Tehnika", Minsk. [English translation: Israel Progr. Sci. Transl., Jerusalem, 1969.]

———, Pal'cev, A.A. [1967]: *Tables for the Numerical Integration of Functions with Logarithmic and Exponential Singularities* (Russian). Izdat. "Nauka i Tehnika", Minsk. [English translation: Israel Progr. Sci. Transl., Jerusalem, 1971.]

———, Skoblja, N.S. [1961]: *On the numerical inversion of the Laplace transform* (Russian). Inž.-Fiz. Ž. *4*, 85–101.

———, ——— [1968]: *Handbook on the Numerical Inversion of the Laplace Transform* (Russian). Izdat. "Nauka i Tehnika", Minsk. [English translation: Israel Progr. Sci. Transl., Jerusalem, 1969.]

———, ——— [1974]: *Approximate Methods for the Fourier Transformation and Inversion of the Laplace Transform* (Russian). Izdat. "Nauka", Moscow.

———, Vorob'eva, A.A. [1971]: *Tables for the Computation of the Integrals of Functions with Power Singularities:* $\int_0^1 x^\alpha f(x)dx$, $\int_0^1 x^\alpha (1-x)^\alpha f(x)dx$ (Russian). Izdat. "Nauka i Tehnika", Minsk.

———, Korolev, N.I. and Skoblja, N.S. [1959]: *A remark on computing the integral* $\int_0^\infty x^s e^{-x} f(x)dx$ (Russian). Dokl. Akad. Nauk BSSR *3*, 3–7.

———, Lugin, V.V. and Janovič, L.A. [1963]: *Tables for the Numerical Integration of Functions with Power Singularities:* $\int_0^1 x^\beta (1-x)^\alpha f(x)dx$ (Russian). Izdat. Akad. Nauk Belorussk. SSR, Minsk.

Kumar, R. [1974a]: *A class of quadrature formulas.* Math. Comp. *28*, 769–778.

—— [1974b]: *Certain Gaussian quadratures.* J. Inst. Math. Appl. *14*, 175–182.

———, Jain, M.K. [1974]: *Quadrature formulas for semi-infinite integrals.* Math. Comp. *28*, 499–503.

Kutt, H.R. [1976]: *Gaussian quadrature formulae for improper integrals involving a logarithmic singularity.* CSIR Special Report WISK 232, CSIR, Pretoria, South Africa.

Kutta, W. [1901]: *Beitrag zur näherungsweisen Integration totaler Differentialgleichungen.* Z. Math. Phys. *46*, 435–453.

Lagrange, J.-L. [1762–1765]: *Solution de différents problèmes de calcul intégral.* Miscellanea Taurinensia *3*. [Oeuvres 1, 471–668.]

Laguerre, E.N. [1879]: *Sur l'intégrale* $\int_x^\infty e^{-x}dx/x$. Bull. Soc. Math. France *7*, 72–81. [Oeuvres I, 428–437.]

Lambin, P. and Vigneron, J.P. [1979]: *Tables for the Gaussian computation of* $\int_0^\infty x^\alpha e^{-x} f(x)dx$ *for values of* α *varying continuously between* -1 *and* $+1$. Math. Comp. *33*, 805–811.

Laplace, P.S. [1810/11]: *Mémoire sur les intégrales définies et leur application aux probabilités.* Mém. Acad. Sci. Paris (1) *11*, 279–347. [Oeuvres 12, 357–412.]

Laurie, D.P. [1977]: *Gaussian quadrature rules using Chebyshev expansions.* CSIR Special Report WISK 255, Nat. Res. Inst. Math. Sciences, Pretoria, South Africa.

———, Rolfes, L. [1977]: *A FORTRAN subroutine for computing Gaussian quadrature rules.* NRIMS Note No. 31, Nat. Res. Inst. Math. Sciences, Pretoria, South Africa.

———, ——— [1979]: *Computation of Gaussian quadrature rules from modified moments — Algorithm 015.* J. Comput. Appl. Math. *5*, 235–243.

Lebedev, V.I. and Baburin, O.V. [1965]: *On the computation of principal value integrals, the weights and nodes of Gaussian quadrature formulae* (Russian). Ž. Vyčisl. Mat. i Mat. Fiz. *5*, 454–462.

Lether, F.G. [1977]: *Modified quadrature formulas for functions with nearby poles.* J. Comput. Appl. Math. *3*, 3–9.

——— [1978]: *On the construction of Gauss-Legendre quadrature rules.* J. Comput. Appl. Math. *4*, 47–52.

——— [1980]: *Error estimates for Gaussian quadrature.* Appl. Math. Comput. *7*, 237–246.

Levin, M. [1974]: *A remark on a quadrature formula* (Russian). Tallin. Polütehn. Inst. Toimetised Seer. A, no. 312, 71–74.

Lewanowicz, S. [1979]: *Construction of a recurrence relation for modified moments.* J. Comput. Appl. Math. *5*, 193–206.

Lipòw, P.R. and Stenger, F. [1972]: *How slowly can quadrature formulas converge?* Math. Comp. *26*, 917–922.

Ljaščenko, N.Ja. and Oleĭnik, A.G. [1974]: *Two-sided quadrature formulae of closed type for the computation of the approximate values of iterated integrals with a preassigned number of correct decimal places* (Russian). Vyčisl. Prikl. Mat. (Kiev) Vyp. *23*, 136–148.

———, ——— [1975]: *Two-sided quadrature formulae of semiclosed type for the computation of the approximate values of iterated integrals* (Russian). Vyčisl. Prikl. Mat. (Kiev) Vyp. *27*, 19–27.

Lo, Y.T., Lee, S.W. and Sun, B. [1965]: *On Davis' method of estimating quadrature errors.* Math. Comp. *19*, 133–138.

Lobatto, R. [1852]: *Lessen over de Differentiaal- en Integraal-Rekening.* Part II. *Integraal-Rekening.* Van Cleef, The Hague.

Lo Cascio, M.L. [1973]: *Alcuni risultati numerici sugli s-polinomi di Legendre.* Pubbl. dell'Ist. Mat. Appl. No. 111, Quaderno No. 2, 29–48. Ist. Mat. Appl., Fac. Ingegn., Univ. degli Studi Roma, Rome.

Locher, F. [1974]: *Normschranken für Interpolations- und Quadraturverfahren.* In: Numerische Methoden bei Differentialgleichungen und mit funktionalanalytischen Hilfsmitteln. ISNM *19* (Ed. by Albrecht, J. and Collatz, L.), 159–167. Birkhäuser, Basel.

——— [1980]: *Dividierte Differenzen und Monotonie von Quadraturformeln.* Numer. Math. *34*, 99–109.

———, Zeller, K. [1968]: *Approximationsgüte und numerische Integration.* Math. Z. *104*, 249–251.

Lorentz, G.G. and Riemenschneider, S.D. [1978]: *Birkhoff quadrature matrices.* In: Linear Spaces and Approximation. ISNM *40* (Ed. by Butzer, P.L. and Sz. Nagy, B.), 359–374. Birkhäuser, Basel.

Luke, Y.L. [1969]: *The Special Functions and their Approximations II.* Academic Press, New York.

——— [1975]: *On the error in a certain interpolation formula and in the Gaussian integration formula.* J. Austral. Math. Soc. *19*, 196–209.

——— [1977]: *Algorithms for the Computation of Mathematical Functions.* Academic Press, New York.

———, Ting, B.Y. and Kemp, M.J. [1975]: *On generalized Gaussian quadrature.* Math. Comp. *29*, 1083–1093.

Luvison, A. [1974]: *On the construction of Gaussian quadrature rules for inverting the Laplace transform.* Proc. IEEE *62*, 637–638.

Markov, A. [1885]: *Sur la méthode de Gauss pour le calcul approché des intégrales.* Math. Ann. *25*, 427–432.

Martinez, J.R. [1977]: *Transfer functions of generalized Bessel polynomials.* IEEE CAS *24*, 325–328.

Maskell, S.J. and Sack, R.A. [1974]: *Generalised Lobatto quadrature formulas for contour integrals.* In: Studies in Numerical Analysis (Ed. by Scaife, B.K.P.), 295–310. Academic Press, London.

Mehler, F.G. [1864]: *Bemerkungen zur Theorie der mechanischen Quadraturen.* J. Reine Angew. Math. *63*, 152–157.

Micchelli, C.A. and Pinkus, A. [1977]: *Moment theory for weak Chebyshev systems with applications to monosplines, quadrature formulae and best one-sided L^1-approximation by spline functions with fixed knots.* SIAM J. Math. Anal. *8*, 206–230.

———, Rivlin, T.J. [1972]: *Turán formulae and highest precision quadrature rules for Chebyshev coefficients.* IBM J. Res. Develop. *16*, 372–379.

———, ——— [1973a]: *Numerical integration rules near Gaussian quadrature.* Israel J. Math. *16*, 287–299.

———, ——— [1973b]: *Quadrature formulae and Hermite–Birkhoff interpolation.* Advances in Math. *11*, 93–112.

Michels, H.S. [1963]: *Abscissas and weight coefficients for Lobatto quadrature.* Math. Comp. *17*, 237–244.

Mikloško, J. [1970a]: *Numerical integration with highly oscillating weight functions.* Apl. Mat. *15*, 133–145.

——— [1970b]: *Asymptotic properties and the convergence of numerical quadratures.* Numer. Math. *15*, 234–249.

Miller, R.K. [1971]: *On ignoring the singularity in numerical quadrature.* Math. Comp. *25*, 521–532.

Milne, W.E. [1949]: *The remainder in linear methods of approximation.* J. Res. Nat. Bur. Standards *43*, 501–511.

von Mises, R. [1933]: *Zur mechanischen Quadratur.* Z. Angew. Math. Mech. *13*, 53–56.

——— [1936]: *Über allgemeine Quadraturformeln.* J. Reine Angew. Math. *174*, 56–67.

Monegato, G. [1976]: *A note on extended Gaussian quadrature rules.* Math. Comp. *30*, 812–817.

——— [1978a]: *Positivity of the weights of extended Gauss–Legendre quadrature rules.* Math. Comp. *32*, 243–245.

——— [1978b]: *Some remarks on the construction of extended Gaussian quadrature rules.* Math. Comp. *32*, 247–252.

——— [1979]: *An overview of results and questions related to Kronrod schemes.* In: Numerische Integration. ISNM *45* (Ed. by Hämmerlin, G.), 231–240. Birkhäuser, Basel.

——— [1980]: *On polynomials orthogonal with respect to particular variable-signed weight functions.* Z. Angew. Math. Phys. *31*, 549–555.

Morelli, A. [1967/68]: *Formula di quadratura con valori della funzione e delle sue derivate anche in punti fuori dell'intervallo di integrazione.* Atti Accad. Sci. Torino Cl. Sci. Fis. Mat. Natur. *102*, 569–579.

———, Verna, I. [1969]: *Formula di quadratura in cui compaiono i valori della funzione e delle derivate con ordine massimo variabile da nodo a nodo.* Rend. Circ. Mat. Palermo (2) *18*, 91–98.

Murphy, R. [1835]: *Second memoir on the inverse method of definite integrals.* Trans. Cambridge Philos. Soc. *5*, 113–148.

Mushelišvili, N.I. [1946]: *Singular Integral Equations* (Russian). OGIZ, Moscow–Leningrad. [3rd ed., Izdat. "Nauka", Moscow, 1968. English translation of the 1st ed.: Noordhoff, Groningen, 1953. German translation of the 2nd ed.: Akademie-Verlag, Berlin, 1965.]

Mysovskih, I.P. [1964]: *A special case of quadrature formulae containing preassigned nodes* (Russian). Vesci Akad. Navuk BSSR Ser. Fiz.-Tehn. Navuk, No. 4, 125–127.

McNamee, J. [1964]: *Error-bounds for the evaluation of integrals by the Euler–Maclaurin formula and by Gauss-type formulae.* Math. Comp. *18*, 368–381.

Newbery, A.C.R. [1969]: *Some extensions of Legendre quadrature.* Math. Comp. *23*, 173–176.

Nicholson, D., Rabinowitz, P., Richter, N. and Zeilberger, D. [1971]: *On the error in the numerical integration of Chebyshev polynomials.* Math. Comp. *25*, 79–86.

Nuttal, J. and Wherry, C.J. [1978]: *Gaussian integration for complex weight functions.* J. Inst. Math. Appl. *21*, 165–170.

Olson, A.P. [1969]: *Gaussian quadratures for* $\int_1^\infty \exp(-x)f(x)dx/x^m$ *and* $\int_1^\infty g(x)dx/x^m$. Math. Comp. *23*, 447.

Omladič, M., Pahor, S. and Suhadolc, A. [1975/76]: *On a new type of quadrature formulas.* Numer. Math. *25*, 421–426.

Ossicini, A. [1966]: *Costruzione di formule di quadratura di tipo Gaussiano.* Ann. Mat. Pura Appl. (4) *72*, 213–237.

——— [1968]: *Le funzioni di influenza nel problema di Gauss sulle formule di quadratura.* Matematiche (Catania) *23*, 7–30.

———, Rosati, F. [1975]: *Funzioni caratteristiche nelle formule di quadratura gaussiane con nodi multipli.* Boll. Un. Mat. Ital. (4) *11*, 224–237.

———, ——— [1978]: *Sulla convergenza dei funzionali ipergaussiani.* Rend. Mat. (6) *11*, 97–108.

Paget, D.F. and Elliott, D. [1972]: *An algorithm for the numerical evaluation of certain Cauchy principal value integrals.* Numer. Math. *19*, 373–385.

Pal'cev, A.A. and Skoblja, N.S. [1965]: *The integration of bounded functions with a Laguerre weight* (Russian). Vescī Akad. Navuk BSSR Ser. Fīz.-Mat. Navuk, no. 3, 15–23.

Patterson, T.N.L. [1968]: *The optimum addition of points to quadrature formulae.* Math. Comp. *22*, 847–856. Loose microfiche suppl. C1–C11. [Errata: ibid. *23* (1969), 892.]

——— [1969]: *Integration formulae involving derivatives.* Math. Comp. *23*, 411–412.

——— [1973]: *Algorithm 468 — Algorithm for automatic numerical integration over a finite interval.* Comm. ACM *16*, 694–699.

——— [1976/77]: *On high precision methods for the evaluation of Fourier integrals with finite and infinite limits.* Numer. Math. *27*, 41–52.

Pavel, P. [1967]: *On the remainder of some Gaussian formulae.* Studia Univ. Babeş-Bolyai Ser. Math.-Phys. *12*, no. 2, 65–70.

——— [1968a]: *On some quadrature formulae of Gaussian type* (Romanian). Studia Univ. Babeş-Bolyai Ser. Math.-Phys. *13*, no. 1, 51–58.

——— [1968b]: *On the remainder of certain quadrature formulae of Gauss–Christoffel type* (Romanian). Studia Univ. Babeş-Bolyai Ser. Math.-Phys. *13*, no. 2, 67–72.

Peano, G. [1913]: *Resto nelle formule di quadratura, espresso con un integrale definito.* Atti R. Accad. Lincei (5) Rend. Cl. Sci. Fis. Mat. Nat. *22*, 562–569.

——— [1914]: *Residuo in formulas de quadratura.* Mathesis (4) *4* (*34*), 5–10.

Perron, O. [1957]: *Die Lehre von den Kettenbrüchen II.* 3rd ed., Teubner, Stuttgart.

Pexton, R.L. [1976]: *Application of Salzer's inverse Laplace transform algorithm.* J. Computational Phys. *20*, 492–494.

Phillips. J.L. and Hanson, R.J. [1974]: *Gauss quadrature rules with B-spline weight functions.* Math. Comp. *28*, Review 18, 666. Loose microfiche suppl. A1–C4.

Piessens, R. [1969a]: *Gaussian quadrature formulas for the numerical integration of Bromwich's integral and the inversion of the Laplace transform*, Rep. TW1, Inst. Appl. Math., University of Leuven, Leuven.

——— [1969b]: *New quadrature formulas for the numerical inversion of the Laplace transform.* BIT *9*, 351–361.

——— [1970a]: *Gaussian quadrature formulas for the integration of oscillating functions.* Z. Angew. Math. Mech. *50*, 698–700.

——— [1970b]: *Gaussian quadrature formulas for the integration of oscillating functions.* Math. Comp. *24*, Review 24, 478–479. Loose microfiche suppl. E.

——— [1970c]: *Numerical evaluation of Cauchy principal values of integrals.* BIT *10*, 476–480.

——— [1971a]: *Gaussian quadrature formulas for the numerical integration of Bromwich's integral and the inversion of the Laplace transform.* J. Engrg. Math. *5*, 1–9.

——— [1971b]: *Some aspects of Gaussian quadrature formulae for the numerical inversion of the Laplace transform.* Comput. J. *14*, 433–436.

——— [1971c]: *On a numerical method for the calculation of transient responses.* J. Franklin Inst. *292*, 57–64.

——— [1972a]: *Gaussian quadrature formulas for the evaluation of Fourier-cosine coefficients.* Z. Angew. Math. Mech. *52*, 56–58.

——— [1972b]: *Gaussian quadrature formulae for integrals involving Bessel functions.* Math. Comp. *26*, Review 44, 1016–1017. Loose microfiche suppl. D1–D14, E1–E14, F1–F5.

——— [1973a]: *An algorithm for automatic integration.* Angew. Informatik, Heft 9, 399–401.

——— [1973b]: *Algorithm 453 — Gaussian quadrature formulas for Bromwich's integral.* Comm. ACM *16*, 486–487.

——— [1975]: *Comments on: "On the construction of Gaussian quadrature rules for inverting the Laplace transform" by A. Luvison.* Proc. IEEE *63*, 817–818.

———, Branders, M. [1973]: *The evaluation and application of some modified moments.* BIT *13*, 443–450.

———, ——— [1974]: *A note on the optimal addition of abscissas to quadrature formulas of Gauss and Lobatto type.* Math. Comp. *28*, 135–139. Suppl., ibid., 344–347.

———, ——— [1975]: *Tables of Gaussian quadrature formulas.* Appl. Math. Progr. Div., University of Leuven, Leuven.

———, Poleunis, F. [1971]: *A numerical method for the integration of oscillatory functions.* BIT *11*, 317–327.

———, Chawla, M.M. and Jayarajan, N. [1976]: *Gaussian quadrature formulas for the numerical calculation of integrals with logarithmic singularity.* J. Computational Phys. *21*, 356–360.

———, Van Roy-Branders, M. and Mertens, I. [1976]: *The automatic evaluation of Cauchy principal value integrals.* Angew. Informatik, Heft 1, 31–35.

Pittnauer, F. and Reimer, M. [1976]: *Interpolation mit Intervallfunktionalen.* Math. Z. *146*, 7–15.

———, ——— [1979a]: *Intervallfunktionale vom Gauss-Legendre Typ.* Math. Nachr. *87*, 239–248.

———, ——— [1979b]: *Zur Fehler- und Konvergenztheorie der Integration mit Intervallfunktionalen.* Rev. Roumaine Math. Pures Appl. *24*, 1105–1115.

Pólya, G. [1933]: *Über die Konvergenz von Quadraturverfahren.* Math. Z. *37*, 264–286.

Popoviciu, T. [1955]: *Sur une généralisation de la formule d'intégration numérique de Gauss.* Acad. R.P. Romîne Fil. Iaşi Stud. Cerc. Şti. *6*, 29–57.

Porath, G. and Wenzlaff, G. [1976]: *Über eine verallgemeinerte Lobattosche Quadraturformel.* Beiträge Numer. Math. *5*, 147–156.

Posse, C. [1875]: *Sur les quadratures.* Nouv. Ann. Math. (2) *14*, 49–62.

Price, J.F. [1960]: *Discussion of quadrature formulas for use on digital computers.* Math. Note No. 217, Math. Res. Lab., Boeing Scientific Res. Laboratories.

Price, T.E., Jr. [1979]: *Orthogonal polynomials for nonclassical weight functions.* SIAM J. Numer. Anal. *16*, 999–1006.

Rabinowitz, P. [1960]: *Abscissas and weights for Lobatto quadrature of high order.* Math. Comp. *14*, 47–52.

——— [1967]: *Gaussian integration in the presence of a singularity.* SIAM J. Numer. Anal. *4*, 191–201.

——— [1968]: *Error bounds in Gaussian integration of functions of low-order continuity.* Math. Comp. *22*, 431–434.

——— [1969]: *Rough and ready error estimates in Gaussian integration of analytic functions.* Comm. ACM *12*, 268–270.

——— [1970]: *Gaussian integration of functions with branch point singularities.* Internat. J. Comput. Math. *2*, 297–306.

——— [1977]: *Ignoring the singularity in numerical integration.* In: Topics in Numerical Analysis III (Ed. by Miller, J.J.H.), 361–368. Academic Press, London.

——— [1978]: *The numerical evaluation of Cauchy principal value integrals.* Proc. Sympos. Numerical Analysis, Durban, South Africa, 1–29.

——— [1979]: *On avoiding the singularity in the numerical integration of proper integrals.* BIT *19*, 104–110.

——— [1980]: *The exact degree of precision of generalized Gauss–Kronrod integration rules.* Math. Comp. *35*, 1275–1283.

———, Richter, N. [1970]: *New error coefficients for estimating quadrature errors for analytic functions.* Math. Comp. *24*, 561–570.

———, Weiss, G. [1959]: *Tables of abscissas and weights for numerical evaluation of integrals of the form* $\int_0^\infty e^{-x} x^n f(x)dx$. Math. Tables Aids Comput. *13*, 285–294. [Errata by Shao, T.S. and Chen, T.C., ibid. *18* (1964), 177.]

Radau, R. [1880]: *Étude sur les formules d'approximation qui servent à calculer la valeur numérique d'une intégrale définie.* J. Math. Pures Appl. (3) *6*, 283–336.

———[1883]: *Remarque sur le calcul d'une intégrale définie.* C.R. Acad. Sci. Paris *97*, 157–158.

Radon, J. [1935]: *Restausdrücke bei Interpolations und Quadraturformeln durch bestimmte Integrale.* Monatsh. Math. Phys. *42*, 389–396.

Ramakrishnan, T.R. [1973]: *Asymptotic estimates for the error of Gauss–Jacobi quadrature formulas.* Calcolo *10*, 233–244.

Ramskiĭ, Ju.S. [1968]: *Quadrature formulae of A.A. Markov's type for functions of class* $\bar{C}_n^{(r)}[a, b]$ (Russian). Vyčisl. Prikl. Mat. (Kiev) Vyp. *6*, 73–81.

——— [1974]: *The improvement of a certain quadrature formula of Gauss type* (Russian). Vyčisl. Prikl. Mat. (Kiev) Vyp. *22*, 143–146.

Rebolia, L. [1973]: *Formule di quadratura "chiuse" di tipo Gaussiano: Tabulazione dei nodi e dei coefficienti.* Calcolo *10*, 245–256.

Reiz, A. [1950a]: *Quadrature formulae for the numerical calculation of mean intensities and fluxes in a stellar atmosphere.* Ark. Astr. *1*, 147–153.

——— [1950b]: *On a special case of a quadrature formula of Christoffel.* Math. Tables Aids Comput. *4*, 181–185.

Rémès, E.J. [1940]: *Sur les termes complémentaires de certaines formules d'analyse approximative.* C.R. (Doklady) Acad. Sci. URSS *26*, 129–133.

Richert, W.R. [1979]: *Gauss-Quadraturformeln mit mehrfachen Knoten.* In: Numerische Integration. ISNM *45* (Ed. by Hämmerlin, G.), 241–244. Birkhäuser, Basel.

Riess, R.D. [1971]: *A note on error bounds for Gauss–Chebyshev quadrature.* SIAM J. Numer. Anal. *8*, 509–511.

——— [1975]: *Gauss–Turán quadratures of Chebyshev type and error formulae.* Computing *15*, 173–179.

——— [1976]: *A Gaussian approach to equally spaced quadratures.* J. Inst. Math. Appl. *17*, 261–265.

———, Johnson, L.W. [1969]: *Estimating Gauss–Chebyshev quadrature errors.* SIAM J. Numer. Anal. *6*, 557–559.

———, ——— [1974]: *On the determination of quadrature formulae of highest degree of precision for approximating Fourier coefficients.* J. Inst. Math. Appl. *13*, 345–351.

Riesz, M. [1922/23]: *Sur le problème des moments III.* Ark. Mat. Astr. Fys. *17*, no. 16, 1–52.

Roghi, G. [1967]: *Sul resto delle formule di quadratura di tipo gaussiano.* Matematiche (Catania) *22*, 143–159.

——— [1978]: *Alcune osservazioni sulla convergenza di formule di quadratura ipergaussiane.* Pubbl. Ist. Mat. Appl. Fac. Ingr. Univ. degli Studi Roma, Quaderno No. 12, 15–20.

Rosati, F. [1968]: *Problemi di Gauss e Tchebychef relativi a formule di quadratura esatte per polinomi trigonometrici.* Matematiche (Catania) *23*, 31–49.

Rothmann, H.A. [1961]: *Gaussian quadrature with weight function x^n on the interval* $(-1, 1)$. Math. Comp. *15*, 163–168.

Runge, C. and Willers, F.A. [1915]: *Numerische und graphische Integration.* Encyklopädie der Math. Wiss., vol. 2, sec. 3, 47–176, Leipzig.

Rutishauser, H. [1962a]: *On a modification of the QD-algorithm with Graeffe-type convergence.* Z. Angew. Math. Phys. *13*, 493–496.

——— [1962b]: *Algorithm 125 — Weightcoeff.* Comm. ACM *5*, 510–511.

Sack, R.A. and Donovan, A.F. [1971/72]: *An algorithm for Gaussian quadrature given modified moments.* Numer. Math. *18*, 465–478.

Salzer, H.E. [1955]: *Orthogonal polynomials arising in the numerical evaluation of inverse Laplace transforms.* Math. Tables Aids Comput. *9*, 164–177.

——— [1961]: *Additional formulas and tables for orthogonal polynomials originating from inversion integrals.* J. Math. Phys. *40*, 72–82.

——— [1976]: *Inverse Laplace transforms of osculatory and hyperosculatory interpolation polynomials.* J. Computational Phys. *20*, 480–491.

Sanikidze, D.G. [1968]: *The convergence of certain quadrature processes* (Russian). Sakharth. SSR Mecn. Akad. Moambe *52*, 577–581.

——— [1970a]: *Quadrature formulae, for singular integrals, that have the highest algebraic degree of accuracy* (Russian). Sakharth. SSR Mecn. Akad. Moambe *57*, 281–284.

——— [1970b]: *The approximate computation of singular integrals with summable density by methods of mechanical quadrature* (Russian). Ukrain. Mat. Ž. *22*, 106–144.

——— [1970c]: *The convergence of a quadrature process for certain singular integrals* (Russian). Ž. Vyčisl. Mat. i Mat. Fiz. *10*, 189–196.

——— [1970d]: *The order of approximation of certain singular operators by quadrature sums* (Russian). Izv. Akad. Nauk Armjan. SSR Ser. Mat. *5*, no. 4, 371–384.

——— [1972]: *Quadrature processes for Cauchy type integrals* (Russian). Mat. Zametki *11*, 517–526.

Sard, A. [1948]: *Integral representations of remainders.* Duke Math. J. *15*, 333–345.

Schmidt, E. [1907]: *Zur Theorie der linearen und nichtlinearen Integralgleichungen, I. Theil: Entwicklung willkürlicher Funktionen nach System vorgeschriebener.* Math. Ann. *63*, 433–476.

Schmidt, R. [1947]: *Mechanische Quadratur nach Gauss für periodische Funktionen.* Bayer. Akad. Wiss. Math.-Natur. Kl. S.-B., 155–173.

Schoenberg, I.J. [1958]: *Spline functions, convex curves and mechanical quadrature.* Bull. Amer. Math. Soc. *64*, 352–357.

Schulten, Z., Anderson, D.G.M. and Gordon, R.G. [1979]: *An algorithm for the evaluation of the complex Airy functions.* J. Computational Phys. *31*, 60–75.

Šeško, M.A. [1976]: *Convergence of quadrature processes for singular integrals* (Russian). Izv. Vyss. Učebn. Zaved. Matematika, no. 12 (175), 108–118.

——— [1979]: *On the convergence of cubature processes for a two-dimensional singular integral* (Russian). Dokl. Akad. Nauk BSSR *23*, 293–296.

———, Jakimenko, T.S. [1980]: *On the convergence of a quadrature process for a singular integral with power-logarithmic singularity* (Russian). Izv. Vyss. Učebn. Zaved. Matematika, no. 1 (212), 82–84.

Shao, T.S., Chen, T.C. and Frank, R.M. [1964a]: *Tables of zeros and Gaussian weights of certain*

associated Laguerre polynomials and the related generalized Hermite polynomials. Math. Comp. *18*, 598–616. [Errata by Danloy, B., ibid. *26* (1972), 813.]

———, ———, ——— [1964b]: *Tables of zeros and Gaussian weights of certain associated Laguerre polynomials and the related generalized Hermite polynomials.* Tech. Rep. 00.1100, Development Laboratory, Data Systems Div., IBM Corp., Poughkeepsie, New York.

Shohat, J.A. [1927]: *Sur les quadratures mécaniques et sur les zeros des polynomes de Tchebycheff dans un intervalle infini.* C.R. Acad. Sci. Paris *185*, 597–598.

——— [1928]: *On the asymptotic properties of a certain class of Tchebycheff polynomials.* Proc. Intern. Math. Congress Toronto 1924, 611–618. The University of Toronto Press, Toronto.

——— [1929]: *On a certain formula of mechanical quadratures with non-equidistant ordinates.* Trans. Amer. Math. Soc. *31*, 448–463.

———, Tamarkin, J.D. [1943]: *The Problem of Moments.* Mathem. Surveys, No. 1, Amer. Math. Soc., New York.

Skoblja, N.S. [1961]: *On the calculation of Mellin's integral* (Russian). Dokl. Akad. Nauk BSSR *5*, 142–145.

Sloan, I.H. [1978]: *On the numerical evaluation of singular integrals.* BIT *18*, 91–102.

Smith, H.V. [1977]: *Error estimates for Gauss–Legendre quadrature of integrands possessing Dirichlet series expansions.* BIT *17*, 108–112.

Smith, W.E. and Sloan, I.H. [1980]: *Product-integration rules based on the zeros of Jacobi polynomials.* SIAM J. Numer. Anal. *17*, 1–13.

Sohockiĭ, Ju. [1873]: *On definite integrals and functions used in series expansions* (Russian). St. Pétersbourg.

Song, C.C.S. [1969]: *Numerical integration of a double integral with Cauchy-type singularity.* AIAA J. *7*, 1389–1390.

Sonin, N.J. [1880]: *Recherches sur les fonctions cylindriques et le développement des fonctions continues en séries.* Math. Ann. *16*, 1–80.

Sprung, D.W.L. and Hughes, D.J. [1965]: *Gauss weights and ordinates for* $\int_0^1 f(x)x^2dx$. Math. Comp. *19*, 139–142.

Stancu, D.D. [1957a]: *Generalization of the quadrature formula of Gauss–Christoffel* (Romanian). Acad. R.P. Romîne Fil. Iaşi Stud. Cerc. Şti. Mat. *8*, no. 1, 1–18.

——— [1957b]: *On a class of orthogonal polynomials and on some general quadrature formulae with minimum number of terms* (Romanian). Bull. Math. Soc. Sci. Math. Phys. R.P. Roumaine (N.S.) *1*, no. 49, 479–498.

——— [1959]: *Sur quelques formules générales de quadrature du type Gauss–Christoffel.* Mathematica (Cluj) *1* (24), 167–182.

———, Stroud, A.H. [1963]: *Quadrature formulas with simple Gaussian nodes and multiple fixed nodes.* Math. Comp. *17*, 384–394.

Stark, V.J.E. [1971]: *A generalized quadrature formula for Cauchy integrals.* AIAA J. *9*, 1854–1855.

Steen, N.M., Byrne, G.D. and Gelbard, E.M. [1969]: *Gaussian quadratures for the integrals* $\int_0^\infty \exp(-x^2)f(x)dx$ *and* $\int_0^b \exp(-x^2)f(x)dx$. Math. Comp. *23*, 661–671.

Steklov, V.A. [1916]: *On the approximate calculation of definite integrals by means of formulae of mechanical quadrature* (Russian). Izv. Akad. Nauk SSSR (6) *10*, 169–186.

Stenger, F. [1966]: *Bounds on the error of Gauss-type quadratures.* Numer. Math. *8*, 150–160.

Stetter, F. [1968]: *Error bounds for Gauss–Chebyshev quadrature.* Math. Comp. *22*, 657–659.

Stieltjes, T.J. [1883]: *Sur l'évaluation approchée des intégrales.* C.R. Acad. Sci. Paris *97*, 740–742, 798–799. [Oeuvres I, 314–316, 317–318.]

——— [1884a]: *Quelques recherches sur la théorie des quadratures dites mécaniques.* Ann. Sci. Éc. Norm. Paris, Sér. 3, *1*, 409–426. [Oeuvres I, 377–396.]

——— [1884b]: *Note sur quelques formules pour l'évaluation de certaines intégrales.* Bul. Astr. Paris *1*, 568. [Oeuvres I, 426–427.]

——— [1884c]: *Sur une généralisation de la théorie des quadratures mécaniques.* C.R. Acad. Sci. Paris *99*, 850–851. [Oeuvres I, 428–429.]

——— [1894]: *Recherches sur les fractions continues.* Ann. Fac. Sci. Toulouse *8*, 1–122; ibid. *9* (1895), 1–47. [Oeuvres II, 402–566.]

Stroud, A.H. [1963]: *Coefficients in quadrature formulas.* Math. Comp. *17*, 289–291.

——— [1965]: *Error estimates for Romberg quadrature.* J. Soc. Indust. Appl. Math. Ser. B Numer. Anal. *2*, 480–488.

——— [1966]: *Estimating quadrature errors for functions with low continuity.* SIAM J. Numer. Anal. *3*, 420–424.

——— [1974]: *Gauss harmonic interpolation formulas.* Comm. ACM *17*, 471–475.

———, Chen, Kwan-wei [1972]: *Peano error estimates for Gauss–Laguerre quadrature formulas.* SIAM J. Numer. Anal. *9*, 333–340.

———, Secrest, D. [1966]: *Gaussian Quadrature Formulas.* Prentice-Hall, Englewood Cliffs, N.J.

———, Stancu, D.D. [1965]: *Quadrature formulas with multiple Gaussian nodes.* J. Soc. Indust. Appl. Math. Ser. B Numer. Anal. *2*, 129–143.

Struble, G. [1960]: *Tables for use in quadrature formulas involving derivatives of the integrand.* Math. Comp. *14*, 8–12.

——— [1963]: *Orthogonal polynomials: Variable-signed weight functions.* Numer. Math. *5*, 88–94.

von Sydow, B. [1977/78]: *Error estimates for Gaussian quadrature formulae.* Numer. Math. *29*, 59–64.

Szegö, G. [1918]: *Ein Beitrag zur Theorie der Polynome von Laguerre und Jacobi.* Math. Z. *1*, 341–356.

——— [1919]: *Über Orthogonalsysteme von Polynomen.* Math. Z. *4*, 139–151.

——— [1921]: *Über die Entwickelung einer analytischen Funktion nach den Polynomen eines Orthogonalsystems.* Math. Ann. *82*, 188–212.

——— [1922]: *Über die Entwicklung einer willkürlichen Funktion nach den Polynomen eines Orthogonalsystems.* Math. Z. *12*, 61–94.

——— [1935]: *Über gewisse orthogonale Polynome die zu einer oszillierenden Belegungsfunktion gehören.* Math. Ann. *110*, 501–513.

Takahasi, H. and Mori, M. [1970]: *Error estimation in the numerical integration of analytic functions.* Rep. Comput. Centre Univ. Tokyo *3*, 41–108.

———, ——— [1971]: *Estimation of errors in the numerical quadrature of analytic functions.* Applicable Anal. *1*, 201–229.

Theocaris, P.S. [1976]: *On the numerical solution of Cauchy-type singular integral equations.* Serdica *2*, 252–275.

———, Ioakimidis, N.I. [1977]: *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*, 1085–1098.

———, ——— [1977/78]: *Numerical integration methods for the solution of singular integral equations.* Quart. Appl. Math. *35*, 173–187.

———, ——— [1978a]: *On the Gauss–Jacobi numerical integration method applied to the solution of singular integral equations.* Bull. Calcutta Math. Soc. *71*, 29–43.

———, ——— [1978b]: *Application of the Gauss, Radau and Lobatto numerical integration rules to the solution of singular integral equations.* Z. Angew. Math. Mech. *58*, 520–522.

———, ——— [1979a]: *A method of numerical solution of Cauchy-type singular equations with generalized kernels and arbitrary complex singularities.* J. Computational Phys. *30*, 309–323.

———, ——— [1979b]: *A remark on the numerical solution of singular integral equations and the determination of stress-intensity factors.* J. Engrg. Math. *13*, 213–222.

———, ———, Kazantzakis, J.G. [1980]: *On the numerical evaluation of two-dimensional principal value integrals.* Internat. J. Numer. Methods Engrg. *15*, 629–634.

———, Tsamasphyros, G. [1979]: *Numerical solution of systems of singular integral equations with variable coefficients.* Applicable Anal. *9*, 37–52.

———, Chrysakis, A.C. and Ioakimidis, N.I. [1979]: *Cauchy-type integrals and integral equations with logarithmic singularities.* J. Engrg. Math. *13*, 63–74.

el-Tom, M.E.A. [1971]: *On ignoring the singularity in approximate integration.* SIAM J. Numer. Anal. *8*, 412–424.

Tsamasphyros, G.J. and Theocaris, P.S. [1977]: *On the convergence of a Gauss quadrature rule for evaluation of Cauchy type singular integrals.* BIT *17*, 458–464.

———, ——— [1979]: *Cubature formulas for the evaluation of surface singular integrals.* BIT *19*, 368–377.

Turán, P. [1950]: *On the theory of the mechanical quadrature.* Acta Sci. Math. Szeged *12*, 30–37.

Tureckiĭ, A.H. [1959]: *On quadrature formulas exact for trigonometric polynomials* (Russian). Učen. Zap. Belorussk. Univ., no. 1 (49), 31–54.

——— [1960]: *Quadrature formulae with an even number of interpolation points accurate for trigonometric polynomials* (Russian). Dokl. Akad. Nauk BSSR *4*, 365–368.

Uspensky, J.V. [1916]: *On the convergence of mechanical quadratures between infinite limits* (Russian). Izv. Akad. Nauk SSSR *10*, 851–866.

——— [1928]: *On the convergence of quadrature formulas related to an infinite interval.* Trans. Amer. Math. Soc. *30*, 542–559.

Uvarov, V.B. [1959]: *Relation between polynomials orthogonal with different weights* (Russian). Dokl. Akad. Nauk SSSR *126*, 33–36.

——— [1969]: *The connection between systems of polynomials that are orthogonal with respect to different distribution functions* (Russian). Ž. Vyčisl. Mat. i Mat. Fiz. *9*, 1253–1262.

Velev, G.D., Semenov, I.P. and Soliev, Ju. [1977]: *Interpolatory quadrature and cubature formulae with multiple nodes for some singular integrals* (Russian). Izv. Vyss. Učebn. Zaved. Matematika, no. 2 (177), 10–20.

Verna, I. [1969]: *Formule di quadratura gaussiana su intervalli infiniti.* Rend. Mat. (6) *2*, 409–424.

Vigneron, J.P. and Lambin, P. [1980]: *Gaussian quadrature of integrands involving the error function.* Math. Comp. *35*, 1299–1307.

Wall, H.S. [1948]: *Analytic Theory of Continued Fractions.* Van Nostrand, New York.

Wellekens, C.J. [1970]: *Generalisation of Vlach's method for the numerical inversion of the Laplace transform.* Electron. Lett. *6*, 741–743.

Wheeler, J.C. [1974]: *Modified moments and Gaussian quadratures.* Rocky Mountain J. Math. *4*, 287–296.

———, Blumstein, C. [1972]: *Modified moments for harmonic solids*, Phys. Rev. *B6*, 4380–4382.

Wilf, H.S. [1962]: *Mathematics for the Physical Sciences.* Wiley, New York.

——— [1964]: *Exactness conditions in numerical quadrature.* Numer. Math. *6*, 315–319.

——— [1980]: personal communication.

Wilkinson, J.H. and Reinsch, C. [1971]: *Linear Algebra.* Handbook for Automatic Computation, Vol. II, Springer, New York.

Wimp, J. [1965]: *On the zeros of a confluent hypergeometric function.* Proc. Amer. Math. Soc. *16*, 281–283.

Zamfirescu, I. [1963]: *An extension of Gauss' method for the calculation of improper integrals* (Romanian). Acad. R.P. Romîne Stud. Cerc. Mat. *14*, 615–631.

Železnova, K.M., Korneĭčuk, A.A. and Markov, A.S. [1965]: *Approximate calculation of singular integrals* (Russian). In: Proc. Conf. Math. Methods of Solution of Problems in Nuclear Phys., 38–40. Ob"edin. Inst. Jadernyh Issled., Dubna.

Received: February 25, 1980.

29.2. [91] (with J. Wimp) "In memoriam YUDELL L. LUKE June 26, 1918 – May 6, 1983"

---

[91] (with J. Wimp) "In memoriam: Yudell L. Luke June 26, 1918 – May 6, 1983," *Math. Comp*. **43**, 349–352 (1984).

In memoriam YUDELL L. LUKE June 26, 1918–May 6, 1983

Yudell Luke was born in Kansas City, Mo., on June 26, 1918, the son of David and Sarah Luke. The household was very traditionally Jewish; in fact, his father served as a sexton in a synagogue. In 1937 Yudell graduated from Kansas City Missouri Junior College and two years later received a B. S. degree with honors from the University of Illinois, followed by an M. S. degree in 1940. While there, he met Laverne Podoll from Chicago who was to become his wife. He taught briefly at the University of Illinois, then served as a full lieutenant in the U.S. Navy from 1942 to 1946, stationed in Hawaii. Upon being discharged, he returned to Kansas City and was immediately hired as the head of the Mathematical Analysis Section of the newly formed Midwest Research Institute. There, Yudell was able to attract and keep together a group of young mathematicians, some of whom later on became researchers in their own right. He advanced to Senior Advisor for Mathematics in 1961, and to Principal Advisor in 1967. After the mathematics group at MRI was dissolved abruptly in 1971, he was appointed to a professorship at the University of Missouri in Kansas City and, in 1978, was given the distinction of Curator's Professor, a position he held until his untimely death.

At the beginning of his career, Yudell was heavily involved in problems of applied mechanics: stress, beam vibrations, aerodynamic lag, supersonic flutter. It was during this early preoccupation with applied problems that Yudell saw the potential usefulness of special function theory and the pressing need to make advanced special functions—integrals of Bessel functions at that time—accessible to the practicing scientist. He began to study these functions in the framework of generalized hypergeometric functions, and his involvement with the latter soon turned into a love affair that was to last throughout his life.

Yudell's main concern was approximation. Foremost in his mind were rational approximations, and he developed a great number of them, not only for specific functions, such as the gamma and incomplete gamma function, elliptic integrals, and elementary functions, but also for general hypergeometric and confluent hypergeometric functions. He used a variety of techniques, most notably his own extension of Lanczos' τ-method, where as forcing term in the differential equation he took not only a multiple of a Chebyshev polynomial, as did Lanczos, but also multiples of more general Jacobi polynomials. He was able to show that in many cases there result approximations of Padé type, specifically those on the main diagonal of the Padé table. A distinguished feature of Yudell's work in this area is his persistent effort of providing detailed information about the error term, either in the form of analytic representations or asymptotic estimates for large degrees. He did not hesitate to develop the necessary asymptotics himself if it was not available in the literature. His results permitted him not only to give unusually sharp a priori error

estimates, valid in large domains of the complex plane, but also, on several occasions, to obtain important convergence statements for Padé approximants along all columns, rows, and diagonals of the Padé table. More recently, his work assumed a computer-oriented flavor, for example, when he developed FORTRAN routines for generating $[n, n]$-type rational approximations to hypergeometric and confluent hypergeometric functions. These either furnish approximations at fixed (complex) argument z and for $n = 1, 2, 3, \ldots$ (until the error is sufficiently small), or yield for given n the coefficients of the desired numerator and denominator polynomials. In a similar vein, he looked at various approximation schemes that have the same complexity (defined by Yudell in his own pragmatic way) and compared them with regard to accuracy attained. Is it better, for example, to apply Padé approximation to convergence factors, rather than to the whole series? Are there any advantages to be gained from using Kummer's transformation for hypergeometric functions prior to the application of the approximation process? These and other questions are answered by meticulous analysis.

Having developed a great deal of expertise in practical rational approximation, it was only natural for Yudell to look around for interesting applications. He was led, in this way, to contribute to questions of univalence for Gauss' error function, to the accurate computation of a technical constant in the theory of trigonometric series, to rational predictor-corrector formulae for nonlinear ordinary differential equations, and eventually was able to interpret many of his rational approximations in terms of summability processes. More importantly, perhaps, he got involved in Padé approximation of the exponential function, a problem of considerable interest in the numerical solution of differential equations, both ordinary and partial. Yudell's contribution, characteristically, consists in providing representations of the error in the approximants on the main and first two subdiagonals in terms of modified Bessel functions as well as related asymptotics. Earlier, already, he obtained asymptotic error bounds for Padé approximants to $\exp(A)$, where A is a bounded linear operator in Banach space. In joint work with G. P. Barker this eventually led to interesting remarks on asymptotic series for matrix functions, in particular, to an extension of Watson's lemma from functions of a complex variable to functions in a matrix argument. By studying the sign of the error term in Padé approximants, and by a variety of other techniques, Yudell also enriched the field of analytic inequalities, deriving a large number of rational inequalities on various intervals of the real line for many of the important special functions. As he rightly points out, such inequalities appear infrequently in the literature.

Series expansion is another important source of approximations in which Yudell took an active interest. One owes to him expansions of the confluent hypergeometric function, and of many of its special cases, in series of Bessel functions, and more generally expansions of hypergeometric functions in other hypergeometric functions. Still more general are the expansions of Meijer's G-function in other G-functions which he developed together with Jet Wimp. These contain, among others, expansions of hypergeometric functions in Jacobi, Laguerre and Hermite polynomials and expansions of the sine and cosine integrals in squares of Bessel functions. Of considerable practical interest are expansions in Chebyshev polynomials, for which he gave many examples, both numerically and in analytic form. There are peculiar

computational difficulties associated with the generation of the desired expansion coefficients, owing to the fact that they often represent solutions of second and higher order linear difference equations possessing minimum or intermediate growth properties. The development of stable algorithms for computing such solutions is an interesting branch of computational mathematics, to which Yudell contributed a useful variant of J. C. P. Miller's backward recurrence algorithm, partly in collaboration with Jet Wimp. Functions of several variables can be expanded in multiple Chebyshev series. Yudell was one of the first to provide numerical tables of associated coefficients in the case of Bessel functions of a real argument and real order between 0 and 1.

In addition to rational approximation and series expansion, there is a third large area—numerical quadrature—that has attracted Yudell's interest. Already in the early 50's he was developing interpolatory quadrature rules with equally spaced nodes for integrals and iterated integrals exhibiting singular or oscillatory weight functions. He provided not only relevant numerical data, but also discussed the error terms, in particular their Peano kernels, in his characteristically pragmatic, but effective way. The work on Filon-type formulae for oscillatory integrals is perhaps his best-known contribution from that period. Furthermore, following Poisson, Turing, Goodwin and others, he helped popularizing the exceptional qualities of the composite trapezoidal and midpoint rules as a means of evaluating many special functions, such as Bessel functions, complete elliptic integrals, the error function, and others. More recently, he turned to general quadrature rules of Gaussian type and related interpolation processes. Among other things, he discovered a novel expansion of the error term, in which intervene the coefficients in the expansion of the integrand (or, rather, the smooth factor multiplying the weight function in the integrand) in the appropriate system of orthogonal polynomials and certain quantities depending only on these orthogonal polynomials. He applied this to Stieltjes-type integrals and generalizations thereof, and in particular to integral representations of the hypergeometric function, thereby opening up yet another approach for its numerical evaluation.

Yudell's early work on special functions is summarized in his book "Integrals of Bessel Functions", published in 1962 by McGraw-Hill. Rather modestly titled, this is actually a comprehensive compendium not only of the functions in the title, but also of Bessel functions themselves and of generalized hypergeometric functions, of which Bessel functions and their integrals are, or can be expressed in terms of, special cases. Yudell also contributed a chapter of the same title to the famous "Handbook of Mathematical Functions" edited by M. Abramowitz and I. A. Stegun. His life work, however, culminated in the two volumes of "The Special Functions and their Approximations", published in 1969 by Academic Press, and the follow-up volumes "Mathematical Functions and their Approximations" of 1975 and "Algorithms for the Computation of Mathematical Functions" of 1977, both also with Academic Press. These works contain an amazing wealth of information, theoretical as well as practical, pertaining to special functions, summarizing and systematizing to a large extent Yudell's own research and that of his collaborators, without neglecting, however, relevant work of others. The "Mathematical Functions" is currently being translated into Russian.

His intellectual interests were never limited to mathematics alone. He loved opera, philosophy, baseball, among other things. While at MRI he gave an extensive series of lectures on the history of philosophy, focusing especially on Spinoza, whose work, he believed, contains the most meaningful elements of those ethical and intellectual ideals which alone can provide a personal bedrock in an uncertain, frenetically changing world. He ended the last lecture with a quotation from Spinoza's book of Ethics, "That which is noble is as difficult as it is rare". It had the force of a personal credo.

Yudell's 1975 book was dedicated to his daughters Molly, Janis, Linda, and Debra and their husbands, and his book of 1977 to his present and future grandchildren. Undoubtedly, the loving support of his family greatly fostered his mathematical growth, and it is natural and indicative of Yudell's personality that he wished the fruits of his life-long research dedicated to them. In his surviving family, and in his grateful students and colleagues, his values will endure.

WALTER GAUTSCHI & JET WIMP

29.3. [101] "Reminiscences of My Involvement in de Branges's Proof of the Bieberbach Conjecture"

---

[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 (1986).

Reminiscences of My Involvement in de Branges's Proof of the Bieberbach Conjecture

WALTER GAUTSCHI

Around February 3, 1984 (I can't remember the exact date), Louis de Branges came to my office and asked whether he could talk to me for a minute about some work he was doing; perhaps I could be of help. I distinctly remember the first thought that ran through my mind: "Me? Helping de Branges?" We hardly knew each other, never engaged in any mathematical conversation in all the 20 or so years we were at Purdue, and—so I believed—had interests diametrically apart. He sat down and told me that he had a way of proving the Bieberbach conjecture, but needed to establish certain inequalities involving hypergeometric functions. He felt it would be worthwhile, as a first step, to check as many of these inequalities as possible on the computer. Could I do this for him?

Now this was a time when I happened to be under all sorts of pressures. I was expected (and very much wanted) to write a paper for *BIT* to honor Germund Dahlquist on his 60th birthday. Through some mix-up the invitation had reached me just a few days earlier (on February 1)—way past the deadline of December 31, 1983—but I was graciously given an extension through February 29. So I had less than four weeks in which to produce worthwhile results and a publishable paper. At the same time I was in the midst of rewriting a chapter of a survey article for the *MAA Studies in Numerical Analysis* in order to be ready to incorporate the new version on the galley proofs that were to arrive at any time. Also, I was scheduled to leave for Europe on March 7 for lectures in Italy, Yugoslavia, and Germany. As if this were not enough, I had, as the newly appointed managing editor of *Mathematics of Computation*, to deal with a constant stream of manuscripts for this journal. And classes had to be taught also, department committee meetings attended to, etc., etc.

I didn't, of course, tell Louis all these things, but they weighed heavily on my mind when I replied that I would probably not have the time to do anything for him right away. He then told me that he was soon going to give a seminar

The work described in this article was supported, in part, by the National Science Foundation under grant DCR-8320561.

on the subject and asked me to at least attend the seminar and in that way get some more concrete ideas of what was involved.

The seminar took place on February 7, and I managed to attend. I was immediately struck by the clarity, freshness, and elegance of Louis' talk and began to appreciate how those inequalities came about. To my delight, they could be written in terms of orthogonal polynomials—currently a subject very much on my mind. What was needed was to show that for any positive integer n the set of n inequalities

$$F_{n,k}(x) := \int_0^1 t^{n-k-1/2} P_k^{(2n-2k,1)}(1-2tx)\,dt > 0, \tag{1}$$

$$0 < x < 1,\ k = 0, 1, 2, \ldots, n-1$$

is valid, where $P_k^{(\alpha,\beta)}$ is the Jacobi polynomial of degree k with parameters α, β. (For $k = 0$, the inequality is trivially true.) Louis' theory in fact states that the validity of (1) for some n implies the Bieberbach conjecture for the $(n+1)$st coefficient (but not vice versa). Louis concluded the lecture by showing how he evaluated $F_{n,k}$—a polynomial of degree k—explicitly for the first few values of n (for $n \leq 4$, I believe) and how he could verify the correctness of (1) in these cases. Unfortunately, they did not include the largest value of n for which Bieberbach's conjecture had already been proven.

I saw right away how (1) could be verified computationally using Gauss-Jacobi quadrature (with weight function $t^{-1/2}$ on $[0, 1]$; but see (2) below), and I pointed this out during the discussion, claiming, with zest, that it would be easy for me to go as far up with n as $n = 100$, if that should be necessary. I was clearly fired up by now and was determined to carry out the computations immediately, no matter what. Having developed reliable software for orthogonal polynomials and Gaussian quadrature during the past few years, I knew that it shouldn't take too much time for me to write the necessary programs.

To begin with, I noted by a simple symmetry argument that one needed only the classical Gauss-Legendre quadrature rule on $[-1, 1]$. If $\tau_\nu^{(2m)}, \lambda_\nu^{(2m)}$, $\nu = 1, 2, \ldots, 2m$, are the nodes and weights, respectively, of the $2m$-point Gaussian quadrature rule, with $1 > \tau_1^{(2m)} > \tau_2^{(2m)} > \cdots > \tau_{2m}^{(2m)} > -1$, then in fact

$$\int_0^1 t^{-1/2} p(t)\,dt = 2\sum_{\nu=1}^{m} \lambda_\nu^{(2m)} p([\tau_\nu^{(2m)}]^2) \tag{2}$$

for any $p \in \mathsf{P}_{2m-1}$, $m = 1, 2, 3, \ldots$. Since the integral in (1) is of the form (2), with p a polynomial of degree n, it suffices to take $2m-1 \geq n$ in (2), for example, $m = [n/2] + 1$. The Gauss formula involved in (2) can easily be generated for any value of m (this indeed is done by one of the easier parts of my software package), and the polynomial $P_k^{(\alpha,\beta)}$ in (1) is readily and accurately generated by the well-known three-term recurrence relation. I actually found it slightly more convenient to compute

$$f_{n,k}(x) = \int_0^1 t^{n-k-1/2} \pi_k^{(2n-2k,1)}(1-2tx)\,dt, \tag{3}$$

for $0 \leq x \leq 1$, $k = 1, 2, \ldots, n-1$, where $\pi_k^{(\alpha,\beta)}$ is the monic Jacobi polynomial.

My first program ran the next day, on February 8, and "verified" (1) for all $n \leq 18$. It cost me \$3.69 in computer time on the CDC 6500. The program, of course, was still fairly primitive; I simply evaluated $f_{n,k}$ for up to 400 equally spaced points on the interval $[0, 1]$ and printed the minimum value and corresponding x-value to see whether the minimum was positive (or a "machine zero" when $x = 1$ and k is odd). I took this simple-minded approach because I was fairly sure that I was going to hit a negative minimum for some early value of n, which would render Louis' argument inconclusive for that value of n, and I could quit and go on with my own work. It didn't work out that way!

After this first piece of positive evidence, I began to improve the program, incorporated error-monitoring devices, compared double precision with single precision results, determined all minima and maxima of $f_{n,k}$ on $[0, 1]$ to full machine precision using Newton's method, and checked between any two extrema for possible additional oscillations that I may have missed. I then pushed this improved version of the program up to $n = 30$ and found the validity of (1) confirmed in every case. The most expensive run (for $27 \leq n \leq 30$) still cost me only \$10.84.

At this point I was convinced that (1) is true for all n. I began to play with the idea of writing up this work in a short note entitled, tentatively, "Numerical evidence in support of a conjecture of L. de Branges." (I didn't dare yet to bring Bieberbach into the title!) I even prepared neat photoready printout on our Diablo printer that could be reproduced, together with the program listing, in the microfiche or supplements section of the journal. (I had in mind my own *Mathematics of Computation.*) A brief excerpt from these tables is shown on the next page.

Happy about this encouraging development, I called on February 13 my good friend Luigi Gatteschi at the University of Turin and asked him whether I could possibly give a second lecture in Turin (one had already been scheduled); the title: "La congettura di Bieberbach è (probabilmente) vera." He readily agreed (though I seemed to detect a skeptical tone in his voice) and subsequently arranged the first lecture to be given at the University and the second at the Polytechnic.

Still not completely satisfied with the strictly computational nature of my work, I began to develop complicated analytic criteria that would insure, mathematically, that $f_{n,k}$ could not have any zeros on any given sufficiently small subinterval of $[0, 1]$. By applying these criteria in a judicious manner, one could then in principle prove (again with the help of the computer) that the whole interval $[0, 1)$ is free of zeros. I spent about a week on efforts along these lines, but did not get very far, since the program I wrote turned out to be extremely slow. About this time, on February 20, Professor Jack Schwartz from the Courant Institute at NYU visited our Computer Sciences Department. I requested beforehand ten minutes of his time to talk to him briefly about this computational

n	k	x	f(x)	df(x)	ddf(x)
25	23	0	2.992383247e-04	-2.85e-02	2.439457320e+00
		3.205597026e-01	2.341983715e-08	-7.58e-20	7.210458286e-06
		3.417860015e-01	2.391757633e-08	9.68e-20	-5.828107204e-06
		4.340919006e-01	1.072194076e-08	-1.27e-20	3.934615048e-06
		4.662672133e-01	1.136217104e-08	2.70e-20	-3.184816919e-06
		5.522705925e-01	5.785863359e-09	2.36e-20	2.457239610e-06
		5.912254258e-01	6.401691048e-09	-1.98e-20	-2.091412963e-06
		6.679768905e-01	3.545491509e-09	1.17e-20	1.891163295e-06
		7.096583223e-01	4.124040146e-09	-2.57e-21	-1.744348249e-06
		7.744958058e-01	2.406779734e-09	-3.54e-21	1.841974435e-06
		8.146887759e-01	2.974169399e-09	-1.11e-20	-1.903521612e-06
		8.656751077e-01	1.775935986e-09	-1.01e-20	2.376104557e-06
		9.002022063e-01	2.375351261e-09	8.51e-21	-2.907485610e-06
		9.362182846e-01	1.395724928e-09	-7.17e-21	4.540772788e-06
		9.612703112e-01	2.103525418e-09	-1.40e-20	-7.509062965e-06
		9.819556336e-01	1.112364664e-09	-3.14e-20	1.838703966e-05
		9.944763566e-01	2.166255736e-09	1.27e-19	-7.519729386e-05
		1.000000000e+00	4.784809710e-19	-1.02e-06	-3.376022124e-04
25	24	0	1.583271559e-05	-2.13e-03	2.534501112e-01
		8.821620539e-02	8.890073589e-08	6.86e-19	9.519682273e-05
		1.007789679e-01	9.102976075e-08	1.47e-18	-6.626872232e-05
		1.654376237e-01	3.432305315e-08	-4.90e-20	2.621243315e-05
		1.895837701e-01	3.655406944e-08	4.55e-19	-1.841404435e-05
		2.621430762e-01	1.721658961e-08	-6.88e-21	9.824593756e-06
		2.961071425e-01	1.896011080e-08	5.88e-21	-7.338968761e-06
		3.726369664e-01	1.015730680e-08	-9.74e-21	4.916658213e-06
		4.142485505e-01	1.154094862e-08	-6.56e-21	-3.940854645e-06
		4.905874739e-01	6.707130829e-09	-4.35e-20	3.131842409e-06
		5.370536552e-01	7.869709830e-09	1.27e-21	-2.707663255e-06
		6.092538964e-01	4.817924151e-09	-2.91e-21	2.475361689e-06
		6.572727170e-01	5.859388113e-09	-2.44e-21	-2.328490771e-06
		7.218483349e-01	3.696588620e-09	-2.38e-21	2.418952894e-06
		7.678171738e-01	4.693289289e-09	-1.50e-20	-2.516816547e-06
		8.219130195e-01	2.988555425e-09	-4.24e-22	2.987795528e-06
		8.621912336e-01	4.012828437e-09	2.54e-21	-3.552522979e-06
		9.036814275e-01	2.511845675e-09	1.49e-20	4.988334162e-06
		9.348846218e-01	3.660165547e-09	2.51e-20	-7.304448233e-06
		9.624003854e-01	2.141929923e-09	-4.80e-20	1.360061023e-05
		9.817059823e-01	3.622132535e-09	1.46e-19	-3.126465152e-05
		9.945898427e-01	1.627629589e-09	-4.94e-19	1.385542984e-04
		1.000000000e+00	5.412894218e-09	1.82e-06	6.092753732e-04

Extrema of $f_{n,k}(x)$ on $[0,1]$, with first and second derivatives, for $n = 25$ and $k = 23, 24$. (The computer printout is in floating-point E-format, so that $e-xx$ is to be read as 10^{-xx}. Note also that the derivatives at the interior extrema are not exactly zero, but approximately $\varepsilon \cdot f_{n,k}(0)$, where $\varepsilon = 3.55 \times 10^{-15}$ is the machine precision of the CDC computer.)

work on the Bieberbach conjecture. He showed polite interest in the matter, but didn't say much. Only at the end of our brief meeting he casually asked why not use Sturm sequences. I remember how this question caught me by surprise and how I wondered why I hadn't thought of it myself. After all, I used Sturm sequences in a similar setting some eight years ago in connection with Chebyshev-type quadrature rules. On second thought, however, I could understand why Sturm didn't enter my mind: The polynomial $F_{n,k}$ in (1) is given

in the form of an integral, and it was not immediately obvious how to generate Sturm sequences in rational form.

It soon occurred to me, however, that the explicit power representation of $F_{n,k}$ can be obtained (in rational form) rather easily by substituting the representation

$$(4)\quad P_k^{(\alpha,\beta)}(u) = \frac{\Gamma(\alpha+k+1)}{k!\Gamma(\alpha+\beta+k+1)} \sum_{m=0}^{k} \binom{k}{m} \frac{\Gamma(\alpha+\beta+k+m+1)}{2^m\Gamma(\alpha+m+1)}(u-1)^m$$

into (1). Indeed, with $u = 1 - 2tx$, one gets $u - 1 = -2tx$, and the integral in (1), using (4), is easily evaluated, yielding a polynomial with coefficients in the form of ratios of factorials (in fact, a ${}_3F_2$). So the work to be done from now on was clearly mapped out for me: Apply the Sturm sequence algorithm to (1) [or alternatively, to (3)] on the interval $[0, 1]$ in rational arithmetic—for example, using the MACSYMA system—and in this way show compellingly, once and for all, and for as many n as desired, that $F_{n,k}$ cannot vanish on $[0, 1)$ and therefore, since $F_{n,k}(0) > 0$, that it remains positive on $(0, 1)$. Time, however, was getting short, and I decided to postpone this work until after my return from Europe. In the meantime, I programmed the Sturm sequence algorithm for (1) in double precision FORTRAN in order to check out the algorithmic details and to make it easier for me, upon my return, to transcribe the program into the MACSYMA language (a system I still had to get better acquainted with). By February 26 (a Sunday) I had the program running satisfactorily on the CDC computer and producing results as expected.

A week earlier, incidentally, I managed to complete my paper for Dahlquist and sent it off to the editors.

My departure for Europe was becoming imminent. Since my verification work seemed well under way, and in good shape now, I let it rest for a while and turned my attention to the lectures I was to give in Europe. Just to set my mind at ease, I wanted to make sure, however, that the inequality (1) was not by chance already known in the literature. Actually, looking at the rather delicate behavior of $f_{n,k}(x)$ for x near 1 and k near n, as exemplified in the short table above, I rather doubted that analytical techniques could be sharp enough to provide a proof of (1). But it didn't hurt to check. I knew there was only a handful of mathematicians in the world who could possibly be familiar with a result of the type (1) and even come up with a proof of it, among them Dick Askey at the University of Wisconsin, whom I knew best. So I called him on February 29 and told him of the inequality (1) and what it implied according to de Branges. He immediately interrupted me with an emphatic: "I don't believe it!" and recounted some rather outrageous claims that had been made in the past by a number of people. I countered that de Branges was a serious mathematician and that we were dealing here with first-rate work. Even if Louis' implication should not hold tight, I said, the inequalities (1) were quite interesting in their own right and ought to be scrutinized further. Besides, I was fully convinced of their validity. Dick now agreed to look into it.

I was working late at home on my lectures, that same night, when the phone rang and I heard Dick Askey's triumphant voice on the other end of the line: "The inequality is not a conjecture—it's a theorem!" He then pointed out a result in a joint 1976 paper with George Gasper that contains (1) as a special case. I was, of course, delighted to hear this incredible news, but also disappointed, realizing that all my hectic work had been in vain. After I checked and confirmed the result myself, I saw Louis the next morning and told him the good news. He replied, rather matter-of-factly: "Well, that proves Bieberbach's conjecture."

Immediately after I talked to Louis, I called Luigi Gatteschi in Turin and asked him to change the title of my second lecture. There was no point anymore to talk about numerical evidence for a conjecture that had turned into a theorem, and I proposed, instead, to talk about the work I did in the paper for Dahlquist. I told him that I would explain everything when I was in Turin. He agreed to send out a change of title notice, and he scheduled this second lecture to be held also at the University on March 13.

It was during the first ten minutes of this lecture that I first apologized for the change of subject and then briefly announced the validity of the Bieberbach conjecture subject to verification of de Branges's work. This was probably the first time that the word got out in Europe, but it was a small audience, consisting largely of graduate students and only a few faculty members. A week later, I attended a conference in Munich celebrating the 25th anniversary of the journal *Numerische Mathematik*. There I saw another good friend of mine, Dick Varga, and told him informally of de Branges's proof of the Bieberbach conjecture. I knew he was going to give a talk himself about a number of conjectures, including the Riemann hypothesis. At the end of the discussion period he turned towards me and put me on the spot with: "Speaking of conjectures, how about Bieberbach's conjecture, Professor Gautschi?" So I went to the blackboard and announced again de Branges's proof of the conjecture and the role played by the inequality of Askey and Gasper. But this time, it was before a large international audience of experts, and I felt the enormous impact of my brief presentation. The word now spread to different parts of Europe.

Looking back at this episode, I cannot help concluding with a few philosophical remarks. 1. *The computer is an important aid in theorem proving.* In our case, the computer could have been used to prove Bieberbach's conjecture (using Sturm sequences in MACSYMA) if not for all n, then at least for as many n as desirable and practicable. Equally importantly, my computer results gave Louis confidence in his overall proof strategy; his approach indeed seemed capable of proving the complete Bieberbach conjecture. 2. *The availability of high-quality mathematical software is of the utmost importance in scientific research.* While this statement is undisputed among computer scientists, it deserves to be better understood and appreciated by the mathematical community. Had I not had available my own software package on orthogonal polynomials, I would probably not have undertaken these computations, given the severe time constraints under

which I was operating. 3. *One should never underestimate the usefulness of a result in pure mathematics.* No one in his wildest dreams, least of all the authors, could have imagined that the Askey/Gasper nonnegativity result would provide a critical link in the proof of the Bieberbach conjecture. *Inequalities*, in particular, *are always potentially useful.* I am fortunate to have inherited a love for inequalities from my teacher, Professor Alexander Ostrowski, who was a master at them, and from Professor Mauro Picone, who openly confessed to me his fondness for inequalities. Perhaps it is an auspicious omen that the Bieberbach conjecture itself—now de Branges's theorem—consists of a set of inequalities.

29.4. [143] "The work of Philip Rabinowitz on numerical integration"

---

[143] "The work of Philip Rabinowitz on numerical integration," *Numer. Algorithms* **9**, 199–222 (1995).

Numerical Algorithms 9(1995)199–222

The work of Philip Rabinowitz on numerical integration*

Walter Gautschi

Department of Computer Sciences, Purdue University, West Lafayette, IN 47907-1398, USA

Received 4 October 1994
Communicated by C. Brezinski

A summary is given of Philip Rabinowitz's contributions to numerical analysis with emphasis on his work on integration.

Keywords: Work of P. Rabinowitz, numerical integration, bibliography.

AMS subject classification: 01A65, 01A70, 65-03, 65D30.

1. The age of SEAC

The span of Rabinowitz's active professional life roughly coincides with the age of modern electronic computers. His early work was done at the National Bureau of Standards (NBS) where he had access to the SEAC (Standards Eastern Automatic Computer), a machine that averaged 2500 additions and 200 multiplications per second and had a high-speed memory of 1024 words of 44 binary digits each. This period at NBS also marks the beginning of a close and long-standing collaboration with P.J. Davis.

Although SEAC was extremely slow and limited, by today's standards, the machine was put to good use, and Rabinowitz became involved in some path-breaking applications. One was the first automatic computation of nerve excitation along and across a nerve fiber [4], which required the solution of a system of four nonlinear ordinary differential equations – the Hodgkin–Huxley equations – by the Runge–Kutta method. (I myself was later involved with a two-dimensional model of this problem, described by a partial differential equation which, as I recall, caused a great deal of difficulties on account of instability.) Another first was the automatic computation of high-order Gauss–Legendre quadrature rules (with up to 96 points) [11,14] using Newton's method, and later of Gauss–Laguerre [91] and Gauss–Lobatto rules [49]. The closest ever

* Lecture presented at the conference "Constructive Approximation and Its Applications" held on May 17–20, 1994, in Tel Aviv honoring Professor P. Rabinowitz on the occasion of his retirement.

done before, by laborious hand computation, was a table of Gauss formulae up to 16 points, with an accuracy of 15 digits, compiled during the Depression under the Work Projects Administration for New York City [36]. Still other work involved interesting experiments with the Monte Carlo method [12], which came into vogue just a few years earlier, in which the use of pseudorandom numbers was compared with the use of number-theoretic (equidistributed) sequences when computing the volume of the n-dimensional hypercube for $n = 2, 3, \ldots, 12$.

Rabinowitz's most significant project, however, was the development, jointly with P. Davis, of a general-purpose code of orthogonalization [9] and its application to a vast array of applied problems. The task at hand, basically, was to generate an orthogonal sequence of elements in an inner product space by applying the Gram-Schmidt procedure to a given sequence of linearly independent elements and to provide quantities of interest such as residual errors and various expansion coefficients. Applications originally envisaged, and eventually carried out [13,15,43,50,51], involved not only the standard applications to least squares approximation and curve fitting, but also the approximate solution of ordinary and partial differential equations by L_2 approximations over the boundary or over the domain, and orthogonal polynomials relative to a simply connected domain in the complex plane and their application to conformal mapping. Here the inner product could be either a line integral over the boundary of the domain, giving rise to Szegö kernel functions, or a double integral over the domain, giving rise to Bergman kernel functions. In practice, all integrals need to be discretized by appropriate quadrature and cubature formulae, which may explain in part Rabinowitz's lifelong interest in numerical integration. There is also a substantial ingredient of approximation theory in these applications, involving the L_2 norm, and it was only natural for Rabinowitz to take an interest in approximation problems in other norms such as the L_∞ and L_1 norms [53,25]. The techniques, of course, are then rather different, resting as they do on linear and nonlinear programming [56]. Finally, there is a small step from the ideas of orthogonalization to other Hilbert space applications in numerical analysis, such as the estimation of linear functionals – especially error functionals – by means of appropriate norms. This idea in fact was pioneered by P. Davis [8] as early as 1953, and its further development by Rabinowitz and his collaborators will be the topic of the next section.

2. Error estimates for analytic functions

Remainder terms in numerical analysis are usually expressed in terms of a derivative of some appropriate order. P. Davis [8] was the first to propose interpreting the remainder as a bounded linear functional in a Hilbert space of analytic functions and then to estimate it in terms of the norm of the functional and the norm of the function to which it is applied. This yields a derivative-free error bound but requires values of the function in the complex plane.

Thus, if E is the functional in question, acting on functions f in the Hilbert space $\mathscr{H}$, then

$$|E(f)| \leq \|E\| \|f\|_{\mathscr{H}}, \quad f \in \mathscr{H}. \tag{2.1}$$

If $\{\pi_r\}_{r=0}^{\infty}$ is a complete orthonormal system in $\mathscr{H}$, then Hilbert space theory tells us that

$$\|E\|^2 = \sum_{r=0}^{\infty} |E(\pi_r)|^2. \tag{2.2}$$

In connection with quadrature rules on a finite interval, say $[-1,1]$, and functions analytic on $[-1,1]$, and hence on a sufficiently small closed domain containing $[-1,1]$ in its interior, there are two Hilbert spaces $\mathscr{H}$ that are particularly pertinent in which to construct estimates of the type (2.1). Both involve functions analytic on an ellipse $\mathscr{E}_\rho$ bounded by

$$\partial\mathscr{E}_\rho = \{z \in \mathbb{C}: z = \tfrac{1}{2}(\rho e^{i\vartheta} + \rho^{-1}e^{-i\vartheta}), 0 \leq \vartheta \leq 2\pi\}, \quad \rho > 1, \tag{2.3}$$

having foci at ± 1 and semiaxes $a = \frac{1}{2}(\rho + \rho^{-1})$, $b = \frac{1}{2}(\rho - \rho^{-1})$. (If $\rho \downarrow 1$, the ellipse $\mathscr{E}_\rho$ shrinks to the interval $[-1,1]$, while for $\rho \to \infty$ it inflates to larger and larger circle-like regions.) The first space, $\mathscr{H} = L^2(\mathscr{E}_\rho)$, adopted for example in [10], consists of those functions f satisfying

$$\int\int_{\mathscr{E}_\rho} |f(z)|^2 dxdy < \infty, \tag{2.4}$$

and the other, $\mathscr{H} = L^2(\partial\mathscr{E}_\rho)$, used in [84], of those satisfying

$$\int_{\partial\mathscr{E}_\rho} |f(z)|^2 |1 - z^2|^{-1/2} |dz| < \infty. \tag{2.5}$$

The respective inner products are

$$(u,v) = \int\int_{\mathscr{E}_\rho} u(z)\overline{v(z)}dxdy, \quad u,v \in L^2(\mathscr{E}_\rho), \tag{2.6}$$

and

$$(u,v) = \int_{\partial\mathscr{E}_\rho} u(z)\overline{v(z)}|1 - z^2|^{-1/2}|dz|, \quad u,v \in L^2(\partial\mathscr{E}_\rho). \tag{2.7}$$

A complete orthonormal system relative to the inner product (2.6) is given by Chebyshev polynomials of the second kind,

$$\pi_r(z) = \frac{2}{\sqrt{\pi}}\left(\frac{r+1}{\rho^{2r+2} - \rho^{-2r-2}}\right)^{1/2} U_r(z), \quad r = 0,1,2,\ldots \text{(in } L^2(\mathscr{E}_\rho)), \tag{2.8}$$

while Chebyshev polynomials of the first kind form a complete orthonormal system relative to the inner product (2.7),

$$\pi_r(z) = \sqrt{\frac{2}{\pi}}\left(\frac{1}{\rho^{2r} + \rho^{-2r}}\right)^{1/2} T_r(z), \quad r = 0,1,2,\ldots \text{(in } L^2(\partial\mathscr{E}_\rho)). \tag{2.9}$$

So, if we are interested, for example, in the quadrature error

$$E_n(f) = \int_{-1}^{1} f(x)dx - Q_n(f), \quad Q_n(f) = \sum_{\nu=1}^{n} w_\nu f(x_\nu), \tag{2.10}$$

we can apply (2.2) immediately either to (2.8) or to (2.9) and obtain

$$\|E_n\|^2_{L^2(\mathscr{E}_\rho)} = \frac{4}{\pi}\sum_{r=0}^{\infty} \frac{r+1}{\rho^{2r+2} - \rho^{-2r-2}} \left\{ \frac{1+(-1)^r}{r+1} - Q_n(U_r) \right\}^2, \tag{2.11}$$

$$\|E_n\|^2_{L^2(\partial\mathscr{E}_\rho)} = \frac{2}{\pi}\sum_{r=0}^{\infty} \frac{1}{\rho^{2r} + \rho^{-2r}} \left\{ \frac{1+(-1)^r}{1-r^2} - Q_n(T_r) \right\}^2, \tag{2.12}$$

where in (2.12) the first term between braces is zero when $r = 1$. If the rule Q_n has polynomial degree of exactness d, then it suffices to start the summation in (2.11) and (2.12) with $r = d + 1$, although in practice, to account for rounding errors, it may be better to use (2.11), (2.12) as written [54].

There are various ways the norm $\|f\|_{\mathscr{H}}$ in (2.1) may be estimated for the two Hilbert spaces $\mathscr{H}$ considered, the simplest being to replace $|f(z)|$ in (2.4) and (2.5) by its maximum value on $\partial\mathscr{E}_\rho$ (hence, on $\mathscr{E}_\rho$),

$$\|f\|_{L^2(\mathscr{E}_\rho)} \le \sqrt{\pi ab}\, M_\rho(f), \quad \|f\|_{L^2(\partial\mathscr{E}_\rho)} \le \sqrt{2\pi}\, M_\rho(f), \tag{2.13}$$

where

$$M_\rho(f) = \max_{z\in\partial\mathscr{E}_\rho} |f(z)|. \tag{2.14}$$

It turns out that the second bound in (2.13) used in conjunction with (2.12) often gives sharper estimates than the first in conjunction with (2.11), at least asymptotically as $\rho \to \infty$ (cf. [84, p. 565]).

The bound (2.1) for $\mathscr{H} = L^2(\mathscr{E}_\rho)$ and $\mathscr{H} = L^2(\partial\mathscr{E}_\rho)$ holds for all $1 < \rho < \rho^*$, where ρ^* is determined by the location of the singularities of f. There is room, therefore, for optimization with respect to ρ on the interval $(1, \rho^*)$.

For specific quadrature rules Q_n one has the practical problem of evaluating, or estimating, the errors $E_n(U_r)$ and $E_n(T_r)$ in (2.11), (2.12). For the composite trapezoidal and Simpson's rules, this has been done in [84], for Gauss-type rules in [42,7], and for Gauss–Legendre rules more recently by K. Petras in [47].

It is natural to try to optimize the quadrature rule (2.10), either for fixed nodes x_ν or otherwise, in the sense of minimizing the error norms (2.11) or (2.12) for given ρ. In the case of (2.11), this has been done by Barnhill and Wixom [3], and for (2.12) by Rabinowitz and Richter [84]. From (2.11) and (2.12) it is plausible that as $\rho \to \infty$ the optimal rule will be the one that annihilates as many of the initial terms in the infinite series as possible, that is, the Gauss rule. For finite values of ρ, the optimal rules have to be computed numerically. Their behavior as $\rho \downarrow 1$ is rather more complicated and is discussed in [85]. A similar study has been made in [86] for Chebyshev-type quadrature rules, where the limiting case $\rho \to \infty$ is

particularly interesting, as the optimal rules (the same for both norms (2.11) and (2.12)) can be characterized (and computed) algebraically. For these rules, and extensions thereof, see also [32,2,31].

The ideas introduced by Davis and Rabinowitz in the area of Hilbert space methods applied to numerical analysis have generated a considerable amount of interest, and an extensive literature evolved pursuing various ramifications of them, including the idea of contour integral representations of quadrature errors for analytic functions. For a recent survey, see [30], and for important new developments, [93].

3. Ignoring or avoiding the singularity

If there is a proverbial red thread running through Rabinowitz's work, then it is his persistent study of integration in the presence of a singularity. He started this line of inquiry in 1965 jointly with P.J. Davis, but came back to it repeatedly – alone or with others – throughout his career, the last time as recently as 1992. The question here is how quadrature rules behave when applied to functions that have a monotonic and integrable singularity ξ, either at one of the endpoints or in the interior of the interval of integration, but are otherwise continuous. One can either *ignore* the singularity, i.e., proceed as if there were none, replacing the value at ξ by zero (or any other finite number) should the quadrature rule call for one; or one can *avoid* it, i.e., remove one or several terms of the quadrature sum in the neigborhood of the singularity. One reason for considering not only endpoint singularities, but also interior ones, is to be able to deal with the case where the location of the singularity may not be known or too difficult to compute, so that the simple expedient of breaking up the integral into two pieces may not be feasible. A good reason for ignoring or avoiding the singularity, apart from the appealing simplicity of the procedure, is that not only the location, but also the nature of the singularity, may be unknown.

The groundwork of this theory is laid in two papers, one by Davis and Rabinowitz [17], and the other by Rabinowitz [52]. The former deals with integrals

$$I(f) = \int_0^1 f(x)dx \tag{3.1}$$

and considers composite quadrature rules Q_n over a partition of [0,1] into n subintervals of equal length, whereby a fixed elementary, positive, quadrature rule, suitably transformed, is applied to each subinterval. The interest is in convergence, or lack thereof, as $n \to \infty$, and it is assumed that the singularity is ignored. If $0 < \xi \leq 1$, then a necessary and sufficient condition for convergence is found to be

$$\lim_{n\to\infty} \frac{1}{n} f(\xi(n)) = 0, \tag{3.2}$$

where $\xi(n)$ is the largest abscissa in Q_n which is less than ξ. The condition (3.2) is always satisfied if $\xi = 1$ is an endpoint; it is also true if $\xi < 1$ is an interior rational point, provided all abscissae of Q_n are rational or, as subsequently observed [52], irrational but algebraic of bounded degree. However, if ξ is irrational, convergence may no longer hold, as is shown by the rectangular rule applied to $f(x) = |\xi - x|^{-\alpha}$, $1/2 \leq \alpha < 1$. It is interesting how number theory – in particular, Diophantine approximation – enters into the analysis of this example. Equally interesting is the fact that convergence is restored if $0 < \alpha < 1/2$ and ξ is irrational but algebraic.

In the second paper [52], mostly endpoint singularities, say at $\xi = 1$, are considered, but more general sequences of (positive) quadrature rules are allowed, as well as weighted integrals,

$$\int_0^1 f(x)w(x)dx = Q_n(f) + E_n(f), \quad Q_n(f) = \sum_{\nu=1}^{n} w_\nu^{(n)} f(x_\nu^{(n)}), \tag{3.3}$$

where w is a positive integrable weight function and

$$0 \leq x_n^{(n)} < x_{n-1}^{(n)} < \cdots < x_1^{(n)} < 1, \quad w_\nu^{(n)} > 0.$$

If $Q_n(g) \to \int_0^1 gwdx$ for every $g \in C[0,1]$, then a sufficient condition for convergence is

$$\frac{w_\nu^{(n)}}{w(x_\nu^{(n)})} \leq c(x_{\nu-1}^{(n)} - x_\nu^{(n)}) \tag{3.4}$$

for all n sufficiently large and all ν such that $x_\nu^{(n)}$ is in some neighborhood of $\xi = 1$. This is true, for example, for Gauss–Jacobi quadratures with parameters $|\alpha| \leq 1/2$, $|\beta| \leq 1/2$. I myself [26] verified (3.4) for Fejér quadratures, i.e., interpolatory rules (3.3) with $w(x) \equiv 1$ and $x_\nu^{(n)}$ the Chebyshev points in [0,1], and used the result to justify the discretized Stieltjes procedure for generating Gaussian quadrature rules and orthogonal polynomials [27], [28,§2.2]. The criterion (3.4), under mild additional conditions on the growth of the integrand, applies also to weighted integrals over a half-infinite interval with the singularity located at the finite endpoint. It thus covers, for example, generalized Gauss–Laguerre quadratures.

Later, in [59], some of the assumptions are relaxed; for example, the positivity of the quadrature rules is dropped, and it suffices to assume (3.4) with $w_\nu^{(n)}$ replaced by $|w_\nu^{(n)}|$. Also, similar results hold if the singularity is avoided instead of ignored. There is an interesting discussion, again using number theory, of the behavior of composite quadrature rules when applied to functions such as $|x - \xi|^{-\alpha} \log^{\beta}|x - \xi|$ with ξ irrational in (0,1). Further extensions are discussed in [62], where partitions in nonequal subintervals are allowed, or uniform partitions but elementary quadrature rules differing from one subinterval to the next; and in [87] to product rules of integration, i.e., rules of the form (3.3) where w need not be a positive weight function, but can be any function $w = k$ with $k \in L_1[0,1]$. For Gauss–Jacobi and related rules, see also [69] and [89,90].

While convergence may be reassuring, it would be more informative to have rates of convergence. These are discussed at great length in [37] for singularities ξ in $[-1,1]$ and Gaussian rules on $[-1,1]$ relative to smooth bounded weight functions. It is shown, for example, that n-point rules applied to $f(x) = |x-\xi|^{-\alpha}$ yield, as $n \to \infty$, an error of $O(n^{-2+2\alpha})$ if $\xi = \pm 1$ and of $O(n^{-1+\alpha})$ if $-1 < \xi < 1$ when the singularity is avoided. Ignoring it gives an error of $O(n^{-1+2\alpha}(\log n)^{\alpha}(\log\log n)^{\alpha(1+\epsilon)})$ for almost all ξ, where $\epsilon > 0$ can be chosen arbitrarily small. Generalized Markov–Stieltjes inequalities are the required tools in this analysis. Similar results are obtained in [64] for Gauss–Lobatto and Gauss–Radau formulae with Jacobi weight function, and, in the case of endpoint singularities, for generalized Jacobi weight functions (cf. (4.6)). The above, of course, are very slow rates of convergence, but this is the price one has to pay if one chooses, or is forced, to disregard singularities.

4. Product rules of integration

Product integration rules employing orthogonal polynomials, and their convergence, have been studied extensively in the late 1970s and early 1980s by Elliott and Paget [22,23], Smith and Sloan [95], Sloan and Smith [94], and others. Starting in 1986, and ever since, Rabinowitz pursued this subject as well. The concern is with integrals of the form

$$I(kf) = \int_{-1}^{1} k(x)f(x)dx, \tag{4.1}$$

where f is smooth and k is absolutely integrable but not necessarily of constant sign. Indeed, k may exhibit difficult behavior, for example, be highly oscillatory, and may also depend on additional parameters, as for example in the numerical solution of linear integral equations, where k would be the kernel of the integral operator. One can distinguish between two approaches, one interpolatory and the other noninterpolatory.

4.1. *Interpolatory product integration rules*

Given n distinct points $x_\nu^{(n)}$ in $[-1,1]$, one approximates f by its Lagrange interpolation polynomial $L_n f$ of degree $n-1$ relative to the points $x_\nu^{(n)}$, so that

$$I(kf) = Q_n(f,k) + E_n(f,k), \quad Q_n(f,k) = \int_{-1}^{1} k(x)(L_n f)(x)dx. \tag{4.2}$$

Usually, the $x_\nu^{(n)}$ are chosen to be zeros of an orthogonal polynomial with respect to some positive weight function w on $[-1,1]$, possibly adjoined with one or both endpoints ± 1. Questions of interest are: convenient algorithms for evaluating $Q_n(f,k)$, and the convergence to zero of $E_n(f,k)$ as $n \to \infty$.

With regard to algorithms, assume for simplicity that $x_\nu^{(n)}$ are the zeros of the nth-degree orthonormal polynomial $\pi_n(\cdot; w)$ relative to the weight function w. (Including $+1$ or -1, or both, among the points requires only minor modifications.) A first expression for Q_n is obtained immediately by integrating the Lagrange polynomial,

$$Q_n(f,k) = \sum_{\nu=1}^{n} w_\nu^{(n)} f(x_\nu^{(n)}), \quad w_\nu^{(n)} = \int_{-1}^{1} k(x) l_\nu(x; w)dx, \tag{4.3}$$

where l_ν are the elementary Lagrange polynomials associated with the points $x_\mu^{(n)}$. This expresses the quadrature rule directly in terms of the function values, which may be useful for analytical purposes. For computation, it is often more convenient to proceed as follows. Expand k/w in the given orthonormal polynomials,

$$\frac{k(x)}{w(x)} = \sum_{r=0}^{\infty} m_r \pi_r(x; w), \quad m_r = m_r(k) = \int_{-1}^{1} k(x)\pi_r(x; w)dx. \tag{4.4}$$

Here, $m_r(k)$ are "modified moments" of k, which we assume can be computed accurately. (There are many important instances where this is the case; for references up to 1984, see [29, p. 169]; also see [48].) Then clearly,

$$\begin{aligned} w_\nu^{(n)} &= \int_{-1}^{1} \frac{k(x)}{w(x)} l_\nu(x; w)w(x)dx \\ &= \int_{-1}^{1} \sum_{r=0}^{\infty} m_r \pi_r(x; w) l_\nu(x; w)w(x)dx \\ &= \sum_{r=0}^{n-1} m_r \int_{-1}^{1} \pi_r(x; w) l_\nu(x; w)w(x)dx, \end{aligned}$$

where orthogonality has been used in the last equation. The last integral, $\pi_r l_\nu$ being a polynomial of degree at most $2n-2$, can be evaluated exactly by the n-point Gauss formula. If $c_\mu^{(n)}$ are the respective Christoffel numbers, one gets

$$\begin{aligned} w_\nu^{(n)} &= \sum_{r=0}^{n-1} m_r \sum_{\mu=1}^{n} c_\mu^{(n)} \pi_r(x_\mu^{(n)}; w) l_\nu(x_\mu^{(n)}; w) \\ &= c_\nu^{(n)} \sum_{r=0}^{n-1} m_r \pi_r(x_\nu^{(n)}; w), \end{aligned}$$

that is,

$$w_\nu^{(n)} = c_\nu^{(n)} S_{n-1}(x_\nu^{(n)}; k/w), \tag{4.5}$$

where $S_{n-1}(\cdot; k/w)$ is the nth partial sum of the Fourier expansion of k/w in the orthonormal polynomials π_r. This can be computed conveniently by Clenshaw's algorithm based on the recurrence relation for the orthogonal polynomials π_r.

In [65] the weight function w is chosen to be a generalized smooth Jacobi weight function,

$$w(x) = \psi(x)(1-x)^{\alpha}(1+x)^{\beta}\prod_{j=1}^{J}|x-t_j|^{\gamma_j}, \tag{4.6}$$

where $-1 < t_J < t_{J-1} < \cdots < t_1 < 1$, the exponents α, β and γ_j are all larger than -1, and $\psi \in C[-1,1]$ is positive and of Dini type, i.e., its modulus of continuity ω_ψ on $[-1,1]$ satisfies

$$\int_0^2 \frac{\omega_\psi(t)}{t}\,dt < \infty. \tag{4.7}$$

By using a result of Nevai [41] on mean convergence of Lagrange interpolation, Rabinowitz shows that $\lim_{n\to\infty} E_n(f,k) = 0$ for all $f \in C[-1,1]$, if k vanishes at most on a set of measure zero and $\int_{-1}^{1} |k(x)|\log^+|k(x)|\,dx < \infty$, and provided w in (4.6) is such that $(1-x^2)^{1/4}w^{1/2}$ and $k(x)(1-x^2)^{-1/4}w^{-1/2}$ are both in $L_1[-1,1]$. It is assumed here that $x_\nu^{(n)}$ are the n Gauss points for the weight function w. Similar results hold for Gauss–Radau and Gauss–Lobatto points. This generalizes and unifies earlier results of Sloan and Smith and others.

Convergence for all $f \in R[-1,1]$, the Riemann integrable functions, and at the same time the convergence $|Q_n|(f,k) \to I(|k|f)$ as $n \to \infty$, where $|Q_n|(f,k) = \sum_{\nu=1}^{n} |w_\nu^{(n)}| f(x_\nu^{(n)})$, is shown in [66,88] under appropriately strengthened assumptions. In particular, this implies convergence of the "stability constant", $\sum_{\nu=1}^{n} |w_\nu^{(n)}| \to I(|k|)$.

While the use of Lagrange interpolation in (4.2) appears most natural and leads to simple algorithms, other interpolation processes could be used instead. Hermite–Fejér-type interpolation, for example, is studied in [72]. The increased complexity in the resulting algorithms and inherent limitations in convergence rates, however, suggest that their use will be advantageous only in exceptional circumstances. For convergence results of Hermite and Hermite–Fejér interpolation on infinite intervals and their implications for product integration on the real line, see [38,80].

4.2. *Noninterpolatory product integration rules*

Instead of interpolating f on the whole interval, one can do local interpolation on n subintervals (not necessarily equal) of a partition of $[-1,1]$. To still obtain convergence for all $f \in R[-1,1]$ as $n \to \infty$ and as the partition is made infinitely fine, it is necessary to impose some restrictions on the local interpolation points: First of all, there should be no more than a finite number of them in each subinterval, where this number is independent of n. More importantly, in each subinterval the separation of the interpolation points should be larger than a fixed fraction (independent of n) of the length of the subinterval. Under these conditions, Rabinowitz shows in [70] that the desired convergence indeed takes place, but

has a counterexample if the last condition is not met. Fewer restrictions are needed to obtain the same kind of convergence if one uses noninterpolatory spline approximants as in [74].

Another type of noninterpolatory approach relies on the Fourier expansion of f (not of k/w, as in section 4.1),

$$f(x) = \sum_{r=0}^{\infty} f_r \pi_r(x; w), \quad f_r = \int_{-1}^{1} f(x)\pi_r(x; w)w(x)dx, \tag{4.8}$$

which yields

$$I(kf) = \sum_{r=0}^{\infty} f_r m_r(k), \tag{4.9}$$

where again $m_r(k)$ are the modified moments of k (cf. (4.4)). To implement this, one needs, on the one hand, to truncate the infinite series in (4.9), and on the other hand, approximate the integral defining f_r by a finite sum, say, the Gauss quadrature sum. Thus,

$$I(kf) \approx Q_n^N(f), \quad Q_n^N(f) = \sum_{r=0}^{N} f_r^{(n)} m_r(k), \tag{4.10}$$

where

$$f_r^{(n)} = \sum_{\nu=1}^{n} c_\nu^{(n)} f(x_\nu^{(n)}) \pi_r(x_\nu^{(n)}; w). \tag{4.11}$$

One can now study two limiting processes: either n and N tend to infinity independently, or one first lets $n \to \infty$ for fixed N, and then lets $N \to \infty$. Both are analyzed in [70], but the latter is easier. Thus, if $f_r^{(n)} \to f_r$ as $n \to \infty$ for all $f \in \mathscr{F}[-1, 1]$ (some appropriate class of functions which could also be singular), then

$$\lim_{N\to\infty} \lim_{n\to\infty} Q_n^N(f) = I(kf),$$

provided $k/w \in L_{2,w}[-1, 1]$ and $f \in L_{2,w}[-1, 1] \cap \mathscr{F}[-1, 1]$. More specialized results, under weaker conditions on k, hold for generalized smooth Jacobi weight functions (cf. (4.6)).

The procedure described is not restricted to finite intervals $[-1,1]$ but applies equally well, at least formally, to infinite intervals.

5. Embedded quadrature rules

The idea of "embedding" a quadrature rule is to adjoin to the points of the given rule a set of new points and assign new weights to all points to produce an extended quadrature rule, usually in such a way as to attain maximum polynomial degree of exactness. The motivation for this is entirely practical: obtain a better

approximation (perhaps for the purpose of estimating the error of the given one) by making maximum use of information (function values) already on hand. "Reverse embedding" (first proposed by Patterson [45]) consists of deleting points from a quadrature rule and redefining in a suitable way the weights of the surviving points. The best-known example of an embedded quadrature rule is the Gauss–Kronrod rule.

5.1. *Gauss–Kronrod quadrature rules*

The history behind these rules is rather curious. Kronrod [35] in 1964 had the idea of embedding the n-point Gauss–Legendre rule into a $(2n+1)$-point extended formula by adding $n+1$ new points and redefining the weights to achieve maximum degree of exactness $3n+1$ (at least). He found that the $n+1$ points to be inserted must be the zeros of the polynomial π_{n+1}^* of degree $n+1$ which satisfies

$$\int_{-1}^{1} \pi_{n+1}^*(x)p(x)\pi_n(x)dx = 0, \quad \text{all } p \in \mathbb{P}_n, \tag{5.1}$$

where π_n is the Legendre polynomial of degree n. In other words, π_{n+1}^* must be orthogonal to all lower-degree polynomials relative to the oscillating measure $\pi_n(x)dx$. Kronrod in fact computes these polynomials and their zeros for n up to 40 along with the weights of the extended quadrature formula. It so happens that Stieltjes already in 1894 (in his last letter to Hermite), apparently out of pure curiosity and certainly not motivated by quadrature concerns, arrived at the same polynomial π_{n+1}^* via manipulations involving the Legendre function of the second kind. He conjectured that all zeros of π_{n+1}^* are real, simple, contained in $[-1,1]$, and interlace with the zeros of π_n. It was Szegö [97] who in 1935 took up Stieltjes's conjecture and proved it, not only for Legendre polynomials, but also for Gegenbauer polynomials $\pi_n(\cdot) = \pi_n(\cdot; w_\gamma)$ with $0 < \gamma \leq 2$ (orthogonal with respect to the weight function $w_\gamma(x) = (1-x^2)^{\gamma-1/2}$). Here, (5.1) becomes

$$\int_{-1}^{1} \pi_{n+1}^*(x; w_\gamma)p(x)\pi_n(x; w_\gamma)w_\gamma(x)dx = 0, \quad p \in \mathbb{P}_n. \tag{5.2}$$

Szegö's analysis, like Stieltjes's, relies heavily on the Gegenbauer function of the second kind and, in addition, on higher monotonicity properties of expansion coefficients associated with it. The existence of Szegö's work (and hence of Stieltjes's) and its relevance to Kronrod extended quadrature rules was pointed out by Mysovskih in 1964 and, independently, by Barrucand in 1970.

Rabinowitz's contributions to Gauss–Kronrod quadrature are theoretical as well as empirical. In [61] he proves that for Gegenbauer weight functions w_γ, $0 < \gamma \leq 2$, the extended quadrature rule has polynomial degree of exactness precisely equal to $3n+1$ if n is even, and $3n+2$ if n is odd, unless $\gamma = 1$, in which case it was known that the degree is $4n+1$. The basic idea of the proof rests on

the Fourier expansion

$$\pi_n(x; w_\gamma)\pi^*_{n+1}(x; w_\gamma) = \sum_{r=0}^{n} c_r \pi_{n+1+r}(x; w_\gamma). \tag{5.3}$$

One notes indeed that the extended quadrature sum, applied to

$$f_r(x) = \pi_n(x; w_\gamma)\pi^*_{n+1}(x; w_\gamma)\pi_{n+1+r}(x; w_\gamma), \quad r = 0, 1, 2, \ldots, \tag{5.4}$$

is always zero, so that the error E_n of the extended rule is given by

$$E_n(f_r) = \int_{-1}^{1} f_r(x) w_\gamma(x) dx.$$

Therefore, by (5.3), if $0 \le r \le n$, one gets

$$E_n(f_r) = c_r \|\pi_{n+1+r}\|^2. \tag{5.5}$$

It then follows easily that the precise degree of exactness is $d_n = 3n + 1 + r_0$, where r_0 is the smallest index r for which $c_r \neq 0$. Rabinowitz then shows, on the basis of Szegö's analysis, that $r_0 = 0$ if n is even, and $r_0 = 1$ otherwise.

The other two theoretical contributions of Rabinowitz are of a negative type. The first [67] concerns the definiteness character of Gegenbauer-type Gauss–Kronrod quadrature rules, i.e., whether or not they admit error terms of the form const $\cdot f^{(d_n+1)}(\xi)$. He answers this in the negative for $0 < \gamma < 1$ and $n \ge 2$. In fact, it is known from work of Akrivis and Förster [1] that an *open* quadrature rule of precise degree d is nondefinite if there exists a function $f \in C^{d+1}[-1, 1]$ for which $f^{(d+1)} \ge 0, f \not\equiv 0$ on $[-1,1]$ and $E_n(f) < 0$. Now again, based on Szegö's theory, Rabinowitz is able to show that, in the notation above, $c_{r_0} < 0$, so that $E_n(f_{r_0}) < 0$ by (5.5) and obviously $f_{r_0}^{(d_n+1)} > 0$. Moreover, the Gauss–Kronrod formula is open if $0 < \gamma < 1$. Therefore, the Akrivis–Förster result applies with $f = f_{r_0}$. For $1 < \gamma \le 2$, the question is still open, even though numerical evidence in [33] (with regard to the Kronrod nodes lying in $(-1,1)$) would suggest nondefiniteness also in this case, at least for $n \le 40$. The other negative result [63, p. 75] is that the n-point Gauss–Jacobi formula does not admit a Kronrod extension with all nodes in $[-1,1]$ when n is even and the Jacobi parameters are $\alpha = -1/2$, $-1/2 < \beta < 1/2$, or when n is odd and $\alpha = -1/2$, $1/2 < \beta \le 3/2$.

The empirical work, done jointly with Elhay and Kautsky [79], seeks to determine numerically the feasibility of extending a Gauss–Kronrod formula once more in the manner of Patterson [44, §3.2]. In particular, this is examined for Gegenbauer weight functions w_γ, following the approach of [33].

5.2. *Reverse embedding of quadrature rules*

For deleting one quadrature point at a time, one has the following elegant result [81]: Let

$$Q_n(f) = \sum_{\nu=1}^{n} w_\nu f(x_\nu), \quad w_\nu > 0,$$

be a positive interpolatory quadrature rule for $I(kf) = \int_a^b k(x)f(x)dx$, $k \in L_1[a,b]$, $-\infty \le a < b \le \infty$,

$$I(kf) = Q_n(f) + E_n(f), \quad E_n(\mathbb{P}_{n-1}) = 0. \tag{5.6}$$

Then there exists a $\mu \in \{1,2,\ldots,n\}$ and a rule

$$Q'_{n-1}(f) = \sum_{\substack{\nu=1 \\ \nu\neq\mu}}^{n} w'_\nu f(x_\nu), \quad w'_\nu \ge 0,$$

which is exact for $f \in \mathbb{P}_{n-2}$. Here is the proof: Let $D_n f$ denote the $(n-1)$st divided difference of f with respect to the points $x_1, x_2, \ldots, x_n$,

$$D_n f := [x_1, x_2, \ldots, x_n]f = \sum_{\nu=1}^{n} \delta_\nu f(x_\nu).$$

Clearly,

$$D_n \mathbb{P}_{n-2} = 0 \tag{5.7}$$

and, in particular, $D_n 1 = 0$, i.e.,

$$\sum_{\nu=1}^{n} \delta_\nu = 0. \tag{5.8}$$

Since none of the δ_ν vanishes, there must be positive as well as negative ones. Define (one is reminded of the simplex method in linear programming!)

$$\min_{\delta_\nu > 0} \frac{w_\nu}{\delta_\nu} = \frac{w_\mu}{\delta_\mu} = m. \tag{5.9}$$

Then

$$Q'_{n-1}(f) = Q_n(f) - mD_n f \tag{5.10}$$

is the desired rule. Indeed,

$$\begin{aligned} Q'_{n-1}(f) &= \sum_{\nu=1}^{n} w_\nu f(x_\nu) - m\sum_{\nu=1}^{n} \delta_\nu f(x_\nu) \\ &= \sum_{\nu=1}^{n} (w_\nu - m\delta_\nu) f(x_\nu) \\ &= \sum_{\nu=1}^{n} w'_\nu f(x_\nu), \end{aligned}$$

and $w_\mu - m\delta_\mu = 0$ while $w_\nu - m\delta_\nu \ge 0$ trivially if $\delta_\nu \le 0$, and by (5.9) otherwise. Further,

$$Q'_{n-1}(f) - I(kf) = [Q_n(f) - I(kf)] - mD_n f,$$

which vanishes for any $f \in \mathbb{P}_{n-2}$, the first (bracketed) term by assumption (cf. (5.6)), and the other by (5.7).

The result can be used to generate finitely many positive embedded quadrature rules $Q_n \supset Q_{n_1} \supset Q_{n_2} \supset \cdots$, where $n > n_1 > n_2 > \cdots$, the precision decreasing by 1 (at least) each time.

A similar result holds for multidimensional integration rules which integrate exactly the first N polynomials of a given sequence; cf. [81, §3].

The only thing that matters for reverse embedding is the first N (linearly independent) polynomials for which a given N-point rule is exact. Hence, if one starts with an efficient integration rule (one that does more than a plain interpolating rule, e.g., a Gauss rule), one can expect to obtain a sequence with a large gap between the degrees of the first and second rule in the sequence. By choosing the high-degree rule judiciously, one can even obtain in this way an optimal pair of embedded rules, at least in two dimensions.

6. Cauchy principal value integrals

There is a dazzling array of papers in which the techniques described in the preceding sections are applied to Cauchy principal value integrals

$$I(kf;\lambda) = \int_{-1}^{1} k(x)\frac{f(x)}{x-\lambda}dx, \quad -1 < \lambda < 1, \tag{6.1}$$

where k and f are suitable functions such that (6.1) exists. In some of the papers, $k = w$ is a nonnegative weight function, in others a more general, sign-variable, function. The latter choice is of interest in connection with singular integral equations. The question of existence is discussed in several of these papers. It suffices, e.g., that k and f are locally (near λ) of Dini type and $k \in L_1[-1,1]$, $f \in R[-1,1]$ (bounded Riemann integrable) [73].

Many methods for evaluating $I(kf;\lambda)$ involve product integration rules (cf. section 4), either interpolatory or noninterpolatory, and Gauss-type points as interpolation abscissas. Thus, [73] analyzes the convergence of interpolatory rules based on Gauss–Radau and Gauss–Lobatto points relative to a generalized smooth Jacobi weight w; cf. (4.6). (This generalizes earlier work for $k = w$ and for straight Gauss points.) One advantage of using Radau and Lobatto points accrues in Nyström's method for solving integral equations, where they yield directly approximations of the solution at one or both endpoints. These are often the values of most physical interest.

The noninterpolatory product integration technique of (4.8)–(4.10) is adapted to Cauchy principal value integrals (6.1) in [82] and [76]. In the first of these papers, $k = w$ is assumed a weight function and f is expanded in orthogonal polynomials with respect to this weight function. This yields in place of (4.10) the approximation

$$I(wf;\lambda) \approx Q_n^N(f;\lambda), \quad Q_n^N(f;\lambda) = \sum_{r=0}^{N} f_r^{(n)} q_r(\lambda;w), \tag{6.2}$$

where q_r are the functions of the second kind,

$$q_r(\lambda; w) = \int_{-1}^{1} \frac{\pi_r(x; w)}{x - \lambda} w(x)dx, \quad r = 0, 1, 2, \ldots, \tag{6.3}$$

and $f_r^{(n)}$ some quadrature approximations of the Fourier coefficients. These need not necessarily be the Gauss formulae, as in (4.11), but could, e.g., be Gauss–Kronrod formulae as in [68], providing higher accuracy. The functions q_r in (6.3) satisfy the same recurrence relation as the orthogonal polynomials π_r, but with different starting values, $q_{-1} = -1$, $q_0 = q_0(\lambda; w)$. The convergence of Q_n^N for $f \in R[-1, 1]$ as n and N tend to infinity is studied in [82] via appropriate Christoffel–Darboux formulae for π_r and q_r.

In the second paper [76] the same expansion of f (in orthogonal polynomials $\pi_r(\cdot; w)$) is used, but in the more general integral (6.1) with $k \neq w$. This now requires integrals

$$q_r(\lambda; k, w) = \int_{-1}^{1} k(x) \frac{\pi_r(x; w)}{x - \lambda} dx, \quad r = 0, 1, 2, \ldots, \tag{6.4}$$

which happen to satisfy the same three-term recurrence relation as before, but with inhomogeneous terms involving the modified moments $m_r(k) = \int_{-1}^{1} k(x)\pi_r(x; w)dx$. Yet another interesting use of modified moments! The convergence theory for k can be reduced to the previous one for w, following ideas of Criscuolo and Mastroianni [6].

Piecewise linear approximation of f in (6.1), or of $[f(x) - f(\lambda)]/(x - \lambda)$ in the usual decomposition

$$I(kf; \lambda) = \int_{-1}^{1} k(x) \frac{f(x) - f(\lambda)}{x - \lambda} dx + f(\lambda) I(k; \lambda),$$

is studied in [71], and more general spline approximation (in the case $k = w$) in [77].

Kronrod extensions of Gauss and Gauss–Lobatto rules for (6.1) (with $k = w$) have also been derived [63], but they are not without numerical difficulties.

7. Multidimensional integration

There are two papers of Rabinowitz (one jointly with N. Richter [83], the other with F. Mantel [39]) which deal with integration over two- and three-dimensional domains. In applications, such as integral equations or inner product evaluations in Gram–Schmidt orthogonalization procedures, where the ultimate goal is much larger than just computing integrals, it is important to keep the number of integration points as low as possible while still maintaining a sufficient degree of accuracy. The construction of such minimum-point cubature rules is greatly simplified (but still a formidable task!) if the domain of integration is fully

symmetric, i.e., closed under sign changes and permutations of the coordinates and the search is restricted to cubature rules exhibiting the same symmetry. Typical examples of fully symmetric domains are the cube, the sphere, and the entire space. In the latter case, there is usually a spherically symmetric, positive weight function involved. These are also the domains of most importance in practice.

For fully symmetric domains, it is natural to employ fully symmetric cubature rules. These are rules whose points consist of the union of fully symmetric sets, each set having a single weight associated with it. That is,

$$\int\int_D f(x,y)dxdy = \sum_{n=1}^{N} w_n \sum_{FS} f(g_n) + R(f), \tag{7.1}$$

where the points g_n are called the "generators" and the inner sum is extended over the fully symmetric set generated by g_n. For a "good" formula (7.1) one wants $g_n \in \text{int}(D)$, and $w_n > 0$, for all n.

Fully symmetric formulae for the square with a minimum number of points, which are exact up to degree 7, have already been given by Hammer and Stroud [34] in 1958. In [83] fully symmetric, odd-degree formulae are constructed which are exact up to degree 15, not only for the square, but also for the circle and for the entire plane with weight functions $\exp(-r^2)$ and $\exp(-r)$, where $r = \sqrt{x^2 + y^2}$. For those formulae that are not good in the above sense, additional generators are judiciously added to make them so. The determination of the structure of formulae (7.1) having the minimum number of points for a given degree of exactness is an intricate problem, not to speak of the actual solution of the moment equations that it entails.

The procedure of constructing fully symmetric cubature formulae of arbitrary (odd) degree, with a low number of points, in two and three dimensions is placed on a firm foundation in the important paper [39]. The construction involves several phases: First, one needs to determine the structure of the formula, i.e., the type of generators and the number of generators of each type, before one can set up the nonlinear (moment) equations expressing the exactness condition. Therefore, conditions need to be worked out insuring the consistency of the equations, which take on the form of inequality constraints. At this point one is ready to attack the main problem: minimizing the number of evaluation points subject to the consistency constraints. This is formulated and solved by an integer programming problem. Having thus determined all consistent structures of fully symmetric cubature rules with the same minimum number of points, one then faces the arduous task of numerically solving, whenever possible, the nonlinear moment equations. Finally, if one insists on good rules, one has to reexamine those that turn out not to be so by the above procedure. All this has been carried out for the three prototype domains described above, now both in two and three dimensions, and has helped in bringing new order and classification into the multitude of fully symmetric cubature formulae. Many old ones have been thus recovered, and new ones discovered, such as 9th-degree 3-dimensional formulae for the three domains considered.

8. For the connoisseur

Any extensive body of work, such as Rabinowitz's, is bound to contain some precious pearls. One such was already mentioned in section 5.2; I now describe two more.

In [19] the question is raised as to whether there exists a quadrature formula of the type

$$\int_0^\infty f(x)dx = \sum_{\nu=1}^{n} w_\nu f(x_\nu) + cf^{(k)}(\xi), \quad 0 < \xi < \infty, \tag{8.1}$$

valid for all $f \in L_1[0,\infty) \cap C^k(0,\infty)$, where c is some constant and k a suitable integer. It is shown that the answer is "no", even if one allowed the error term to consist of a finite number of terms $c_i f^{(k_i)}(\xi_i)$. Here is the proof: One has

$$\int_0^\infty f(x)dx = r\int_0^\infty f(rx)dx, \quad r > 0. \tag{8.2}$$

If (8.1) were true, then (8.1), (8.2) would imply

$$\int_0^\infty f(x)dx = r\sum_{\nu=1}^{n} w_\nu f(rx_\nu) + cr^{k+1}f^{(k)}(\bar{\xi}), \quad 0 < \bar{\xi} < \infty.$$

Choose any f which together with $f^{(k)}$ is bounded on $[0,\infty)$ (for example, $f(x) = (1+x^2)^{-1}$) and let $r \downarrow 0$ to get a contradiction! Note that the integral in (8.1) is a simple integral with weight function $w \equiv 1$. For weighted integrals over $(0,\infty)$, the result clearly does not hold, as the Gauss–Laguerre formula shows.

The second example of delightful work concerns geometric properties of Gauss quadrature rules [16]. Classical asymptotics for Jacobi and Laguerre polynomials can be applied to the respective Gaussian nodes and weights and the results reinterpreted in geometric terms. Thus, in the case of Jacobi weight functions $w^{(\alpha,\beta)}(x) = (1-x)^\alpha(1+x)^\beta$, $\alpha > -1$, $\beta > -1$, the following asymptotic equivalence is shown for the nodes $x_\nu^{(n)}$ and weights $w_\nu^{(n)}$ of the n-point Gauss–Jacobi quadrature formula, as $n \to \infty$:

$$\frac{nw_\nu^{(n)}}{\pi w^{(\alpha,\beta)}(x_\nu^{(n)})} \sim \sqrt{1-(x_\nu^{(n)})^2}, \quad \nu = 1,2,\ldots,n. \tag{8.3}$$

In other words, suitably normalized Christoffel numbers, if plotted over the respective Gauss nodes, lie on a circle, asymptotically as $n \to \infty$. This is illustrated in figure 1, where I plotted the points $(\sqrt{1-x_\nu^2}, nw_\nu/(\pi w^{(\alpha,\beta)}(x_\nu)))$, $\nu = 1,2,\ldots,n$, for α, $\beta = -0.75(0.25)1.0$, $\beta \geq \alpha$, on the left for $n = 20(5)40$, and on the right for $n = 50(15)80$. The same circle theorem holds for Gauss–Jacobi–Lobatto quadrature, and it is conjectured, "with meager numerical evidence at hand", that it also holds for the Radau formula. Plots analogous to those in figure 1 indeed confirm that. I was also intrigued by the authors' suggestion that a circle theorem may hold for a much wider class of weight functions on $[-1,1]$. This is indeed the case

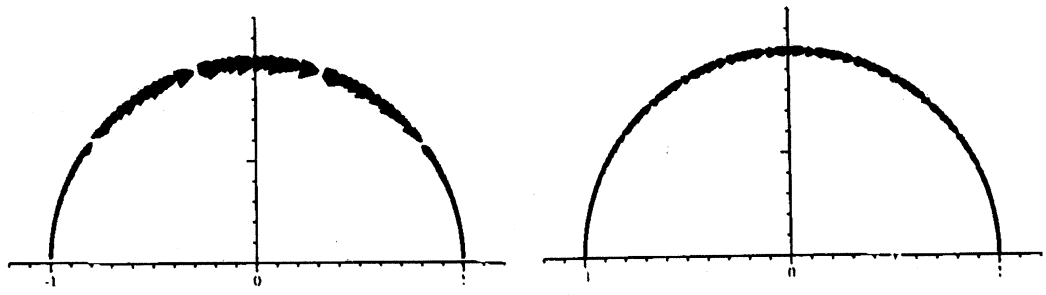


Figure 1. The circle theorem for Gauss–Jacobi quadrature.

and follows from more recent asymptotic results of Nevai [40, theorem 6]. (I am indebted to Professor Nevai for this remark.) For example, (8.3) will hold for generalized smooth Jacobi weight functions (cf. (4.6)) for all ν and n such that $x_\nu^{(n)}$ is contained in a compact subset of the punctured interval $[-1,1]$ (with the singular points removed). Before I knew about Nevai's result, I was experimenting with another set of Gauss formulae, namely those belonging to the numerator polynomials of order 1 associated with the Jacobi weight function. (The respective orthogonal polynomials are easily obtained from the three-term recurrence relation of the Jacobi polynomials by shifting down the indices of the coefficients by 1.) I prepared plots analogous to those of figure 1, but made the mistake of dividing on the left of (8.3) not by the true weight function w (which would be difficult to compute), but simply by the Jacobi weight function $w^{(\alpha,\beta)}$. What I got was the picture in figure 2. True, the picture is erroneous, but pretty nevertheless! The correct picture would actually be similar to the one in figure 1, as follows from Nevai's result mentioned above.

For weight functions on the half-infinite interval, specifically the generalized

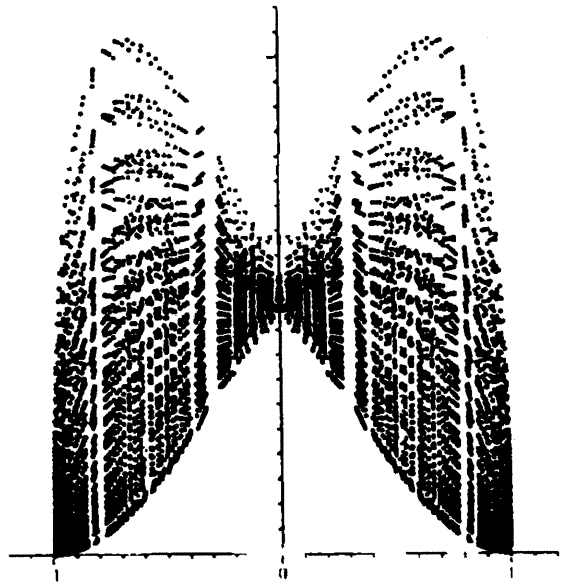


Figure 2. A butterfly "theorem" for associated Jacobi weights.

Laguerre weight $w^{(\alpha)}(x) = x^\alpha e^{-x}$ on $[0,\infty)$, there holds a "parabola theorem" [16, equation (15)], in the sense that

$$\frac{\sqrt{n}w_\nu^{(n)}}{\pi w^{(\alpha)}(x_\nu^{(n)})} \sim \sqrt{x_\nu^{(n)}}, \quad \nu = 1,2,\ldots,n;\ n \to \infty. \tag{8.4}$$

Here again, it is likely that this result extends to a more general class of weight functions.

Another theorem in [16], called the "trapezoid theorem", states for the same weight functions as before that, asymptotically as $n \to \infty$,

$$\frac{w_\nu^{(n)}}{w(x_\nu^{(n)})} \sim \frac{1}{2}|x_{\nu-1}^{(n)} - x_{\nu+1}^{(n)}|, \tag{8.5}$$

as long as the nodes involved remain in the interior of the basic interval in question. For improvements of (8.5), see [24, equation (2.31)], [46, §4].

9. Expository writings

The work that stands out among all expository writings of Rabinowitz is the monograph on numerical integration, written jointly with P.J. Davis. This first appeared in 1967 as a modest text of some 230 pages [18], but immediately gained wide appeal. It was thoroughly overhauled in 1975, when it appeared, doubled in size, under a new title [20]. A second enlarged edition [21] came out in 1984. The book has taken its place among the leading reference works on numerical integration, both one- and multidimensional. To describe its qualities, I can do no better than quote from my review of the 1975 edition in *Mathematical Reviews* (MR 56 #7119): "Its outstanding features continue to be the thorough coverage of the research literature, the balanced treatment of many diverse points of view, both theoretical and practical, the excellent choice of illustrative numerical examples, the strong orientation toward computer implementation, and the inimitable delightful prose of the authors, which, although informal at times, is always informative and enlightening." If anything, the second 1984 edition reinforces these qualities, and with its extensive bibliography of over 1500 items has become indispensable as a reference work to students and researchers alike.

Among the text books one must mention the second edition of Ralston's book on Numerical Analysis [92], which was written jointly with Rabinowitz. It has become one of the standard introductory texts in Numerical Analysis. There is also a book on Nonlinear Equations [57] edited by Rabinowitz, to which he contributed a useful bibliography [58]. A number of survey articles were written or coauthored by Rabinowitz; some of the early ones on orthogonalization [15], linear programming [53] and approximation [56] have already been mentioned.

Others are [60] on Cauchy principal value integrals and [78] on recent progress in extrapolation methods. Finally, there is an extremely useful and scholarly compilation of cubature formulae [5] for cubes, spheres, simplices, and the entire space, which updates and complements the listings in the 1971 book of Stroud [96].

Acknowledgements

The author gratefully acknowledges helpful comments by Professors P. Rabinowitz and R. Cools, and by Dr. J.N. Lyness, on earlier drafts of this paper.

References

[1] G. Akrivis and K.-J. Förster, On the definiteness of quadrature formulae of Clenshaw-Curtis type, Computing 33 (1984) 363–366.
[2] L.A. Anderson and W. Gautschi, Optimal weighted Chebyshev-type quadrature formulas, Calcolo 12 (1975) 211–248.
[3] R.E. Barnhill and J.A. Wixom, Quadratures with remainders of minimum norm. I, II, Math. Comp. 21 (1967) 66–75 and 382–387.
[4] K.S. Cole, H.A. Antosiewicz and P. Rabinowitz, Automatic computation of nerve excitation, J. Soc. Ind. Appl. Math. 3 (1955) 153–172. [Correction, ibid. 6 (1958) 196–197.]
[5] R. Cools and P. Rabinowitz, Monomial cubature rules since "Stroud": a compilation, J. Comp. Appl. Math. 48 (1993) 309–326.
[6] G. Criscuolo and G. Mastroianni, On the convergence of an interpolatory product rule for evaluating Cauchy principal value integrals, Math. Comp. 48 (1987) 725–735.
[7] A.R. Curtis and P. Rabinowitz, On the Gaussian integration of Chebyshev polynomials, Math. Comp. 26 (1972) 207–211.
[8] P. Davis, Errors of numerical approximation for analytic functions, J. Rat. Mech. Anal. 2 (1953) 303–313.
[9] P. Davis and P. Rabinowitz, A multiple purpose orthonormalizing code and its uses, J. ACM 1 (1954) 183–191.
[10] P. Davis and P. Rabinowitz, On the estimation of quadrature errors for analytic functions, Math. Tables Aids Comp. 8 (1954) 193–203.
[11] P.J. Davis and P. Rabinowitz, Abscissas and weights for Gaussian quadratures of high order, J. Res. Nat. Bur. Standards 56 (1956) 35–37.
[12] P.J. Davis and P. Rabinowitz, Some Monte Carlo experiments in computing multiple integrals, Math. Tables Aids Comp. 10 (1956) 1–8.
[13] P. Davis and P. Rabinowitz, Numerical experiments in potential theory using orthonormal functions, J. Washington Acad. Sci. 46 (1956) 12–17.
[14] P. Davis and P. Rabinowitz, Additional abscissas and weights for Gaussian quadratures of high order. Values for $n = 64$, 80, and 96, J. Res. Nat. Bur. Standards 60 (1958) 613–614.
[15] P. Davis and P. Rabinowitz, Advances in orthonormalizing computation, in: *Advances in Computers*, Vol. 2 (Academic Press, New York, 1961) pp. 55–133.
[16] P.J. Davis and P. Rabinowitz, Some geometrical theorems for abscissas and weights of Gauss type, J. Math. Anal. Appl. 2 (1961) 428–437.
[17] P.J. Davis and P. Rabinowitz, Ignoring the singularity in approximate integration, SIAM J. Numer. Anal. B 2 (1965) 367–383.

[18] P.J. Davis and P. Rabinowitz, *Numerical Integration* (Blaisdell, Waltham, MA, 1967).
[19] P.J. Davis and P. Rabinowitz, On the nonexistence of simplex integration rules for infinite integrals, Math. Comp. 26 (1972) 687–688.
[20] P.J. Davis and P. Rabinowitz, *Methods of Numerical Integration* (Academic Press, New York, 1975).
[21] P.J. Davis and P. Rabinowitz, *Methods of Numerical Integration*, 2nd ed. (Academic Press, Orlando, FL, 1984).
[22] D. Elliott and D.F. Paget, Product-integration rules and their convergence, BIT 16 (1976) 32–40.
[23] D. Elliott and D.P. Paget, The convergence of product integration rules, BIT 18 (1978) 137–141.
[24] K.-J. Förster and K. Petras, On estimates for the weights in Gaussian quadrature in the ultraspherical case, Math. Comp. 55 (1990) 243–264.
[25] J.H. Freilich and P. Rabinowitz, Asymptotic approximation by polynomials in the L_1 norm, J. Approx. Theory 8 (1973) 304–314.
[26] W. Gautschi, Numerical quadrature in the presence of a singularity, SIAM J. Numer. Anal. 4 (1967) 357–362.
[27] W. Gautschi, Construction of Gauss–Christoffel quadrature formulas, Math. Comp. 22 (1968) 251–270.
[28] W. Gautschi, On generating orthogonal polynomials, SIAM J. Sci. Statist. Comp. 3 (1982) 289–317.
[29] W. Gautschi, Questions of numerical condition related to polynomials, in: *Studies in Numerical Analysis*, ed. G.H. Golub, Studies in Mathematics vol. 24 (The Mathematical Association of America, 1984) pp. 140–177.
[30] W. Gautschi, Remainder estimates for analytic functions, in: *Numerical Integration: Recent Developments, Software and Applications*, eds. O. Espelid and A. Genz, NATO ASI Series, Series C: Mathematical and Physical Sciences, Vol. 357 (Kluwer, Dordrecht, 1992) pp. 133–145.
[31] W. Gautschi and G. Monegato, On optimal Chebyshev-type quadratures, Numer. Math. 28 (1977) 59–67.
[32] W. Gautschi and H. Yanagiwara, On Chebyshev-type quadratures, Math. Comp. 28 (1974) 125–134.
[33] W. Gautschi and S.E. Notaris, An algebraic study of Gauss–Kronrod quadrature formulae for Jacobi weight functions, Math. Comp. 51 (1988) 231–248.
[34] P.C. Hammer and A.H. Stroud, Numerical evaluation of multiple integrals. II, Math. Tables Aids Comp. 12 (1958) 272–280.
[35] A.S. Kronrod, *Nodes and Weights for Quadrature Formulae. Sixteen-place Tables* (Russian) (Izdat. "Nauka", Moscow, 1964). [English transl.: Consultants Bureau, New York, 1965.]
[36] A.N. Lowan, N. Davids and A. Levenson, Table of the zeros of the Legendre polynomials of order 1–16 and the weight coefficients for Gauss' mechanical quadrature formula, Bull. Amer. Math. Soc. 48 (1942) 739–743.
[37] D.S. Lubinsky and P. Rabinowitz, Rates of convergence of Gaussian quadrature for singular integrands, Math. Comp. 43 (1984) 219–242.
[38] D.S. Lubinsky and P. Rabinowitz, Hermite and Hermite–Fejér interpolation and associated product integration rules on the real line: The L_1 theory, Canad. J. Math. 44 (1992) 561–590.
[39] F. Mantel and P. Rabinowitz, The application of integer programming to the computation of fully symmetric integration formulas in two and three dimensions, SIAM J. Numer. Anal. 14 (1977) 391–425.
[40] P.G. Nevai, Orthogonal polynomials, Mem. Amer. Math. Soc. 18, no. 213 (1979).
[41] P. Nevai, Mean convergence of Lagrange interpolation. III, Trans. Amer. Math. Soc. 282 (1984) 669–698.
[42] D. Nicholson, P. Rabinowitz, N. Richter and D. Zeilberger, On the error in the numerical integration of Chebyshev polynomials, Math. Comp. 25 (1971) 79–86.

[43] J. Nowinski and P. Rabinowitz, The method of the kernel function in the theory of elastic plates, J. Appl. Math. Phys. 13 (1962) 26–42.
[44] T.N.L. Patterson, The optimum addition of points to quadrature formulae, Math. Comp. 22 (1968) 847–856. [Errata, ibid. 23, 892.]
[45] T.N.L. Patterson, On some Gauss and Lobatto based integration formulae, Math. Comp. 22 (1968) 877–881.
[46] K. Petras, Gaussian quadrature formulae – second Peano kernels, nodes, weights and Bessel functions, Calcolo 30 (1993) 1–27.
[47] K. Petras, Gaussian integration of Chebyshev polynomials and analytic functions, *Proc. on Special Functions* (dedicated to Luigi Gatteschi on his seventieth birthday), Ann. Numer. Math. 3 (1995), to appear.
[48] R. Piessens, Modified Clenshaw–Curtis integration and applications to numerical computation of integral transforms, in: *Numerical Integration: Recent Developments, Software and Applications*, eds. P. Keast and G. Fairweather, NATO ASI Series, Series C: Mathematical and Physical Sciences, Vol. 203 (Reidel, Dordrecht, 1987) pp. 35–51.
[49] P. Rabinowitz, Abscissas and weights for Lobatto quadrature of high order, Math. Comp. 14 (1960) 47–52.
[50] P. Rabinowitz, Numerical experiments in conformal mapping by the method of orthonormal polynomials, J. ACM 13 (1966) 296–303.
[51] P. Rabinowitz, Calculations of the conductivity of a medium containing cylindrical inclusions by the method of orthogonalized particular solutions, J. Appl. Phys. 37 (1966) 557–560.
[52] P. Rabinowitz, Gaussian integration in the presence of a singularity, SIAM J. Numer. Anal. 4 (1967) 191–201.
[53] P. Rabinowitz, Applications of linear programming to numerical analysis, SIAM Rev. 10 (1968) 121–159.
[54] P. Rabinowitz, Practical error coefficients for estimating quadrature errors for analytic functions, Commun. ACM 11 (1968) 45–46.
[55] P. Rabinowitz, Rough and ready error estimates in Gaussian integration of analytic functions, Commun. ACM 12 (1969) 268–270.
[56] P. Rabinowitz, Mathematical programming and approximation, in: *Approximation Theory*, ed. A. Talbot (Academic Press, London, 1970) pp. 217–231.
[57] P. Rabinowitz (ed.), *Numerical Methods for Nonlinear Algebraic Equations* (Gordon and Breach, London, 1970).
[58] P. Rabinowitz, A short bibliography on solution of systems of nonlinear algebraic equations, in [57, pp. 195–199].
[59] P. Rabinowitz, Ignoring the singularity in numerical integration, in: *Topics in Numerical Analysis*, ed. J.J.H. Miller (Academic Press, London, 1977) pp. 361–368.
[60] P. Rabinowitz, The numerical evaluation of Cauchy principal value integrals, *Proc. 4th Symp. on Numerical Mathematics*, University of Natal, Durban, South Africa (1978) pp. 54–82.
[61] P. Rabinowitz, The exact degree of precision of generalized Gauss–Kronrod integration rules, Math. Comp. 35 (1980) 1275–1283.
[62] P. Rabinowitz, Generalized composite integration rules in the presence of a singularity, Calcolo 20 (1983) 231–238.
[63] P. Rabinowitz, Gauss–Kronrod integration rules for Cauchy principal value integrals, Math. Comp. 41 (1983) 63–78 [Corrigenda, ibid. 45 (1985) 277; 50 (1988) 655.]
[64] P. Rabinowitz, Rates of convergence of Gauss, Lobatto, and Radau integration rules for singular integrals, Math. Comp. 47 (1986) 625–638.
[65] P. Rabinowitz, The convergence of interpolatory product integration rules, BIT 26 (1986) 131–134.
[66] P. Rabinowitz, On the convergence of interpolatory product integration rules based on Gauss, Radau and Lobatto points, Israel J. Math. 56 (1986) 66–74.

[67] P. Rabinowitz, On the definiteness of Gauss–Kronrod integration rules, Math. Comp. 46 (1986) 225–227.
[68] P. Rabinowitz, A stable Gauss–Kronrod algorithm for Cauchy principal-value integrals, Comp. Math. Appl. Part B 12 (1986) 1249–1254.
[69] P. Rabinowitz, Numerical integration in the presence of an interior singularity, J. Comp. Appl. Math. 17 (1987) 31–41.
[70] P. Rabinowitz, The convergence of noninterpolatory product integration rules, in: *Numerical Integration: Recent Developments, Software and Applications*, eds. P. Keast and G. Fairweather, NATO ASI Series, Series C: Mathematical and Physical Sciences, Vol. 203 (Reidel, Dordrecht, 1987) pp. 1–16.
[71] P. Rabinowitz, Convergence results for piecewise linear quadratures for Cauchy principal value integrals, Math. Comp. 51 (1988) 741–747.
[72] P. Rabinowitz, Product integration based on Hermite–Fejér interpolation, J. Comp. Appl. Math. 28 (1989) 85–101.
[73] P. Rabinowitz, On an interpolatory product rule for evaluating Cauchy principal value integrals, BIT 29 (1989) 347–355.
[74] P. Rabinowitz, Numerical integration based on approximating splines, J. Comp. Appl. Math. 33 (1990) 73–83.
[75] P. Rabinowitz, Numerical evaluation of Cauchy principal value integrals with singular integrands, Math. Comp. 55 (1990) 265–276.
[76] P. Rabinowitz, Generalized noninterpolatory rules for Cauchy principal value integrals, Math. Comp. 54 (1990) 217–279.
[77] P. Rabinowitz, Uniform convergence of Cauchy principal value integrals of interpolating splines, Israel Math. Conf. Proc. 4 (1991) 225–231.
[78] P. Rabinowitz, Extrapolation methods in numerical integration, Numer. Algor. 3 (1992) 17–28.
[79] P. Rabinowitz, S. Elhay and J. Kautsky, Empirical mathematics: the first Patterson extension of Gauss–Kronrod rules, Int. J. Comp. Math. 36 (1990) 119–129.
[80] P. Rabinowitz and L. Gori, L_1-norm convergence of Hermite–Fejér interpolation based on the Laguerre and Hermite abscissas, Rend. Mat. Appl. (7) 14 (1994) 159–176.
[81] P. Rabinowitz, J. Kautsky, S. Elhay and J.C. Butcher, On sequences of imbedded integration rules, in: *Numerical Integration: Recent Developments, Software and Applications*, eds. P. Keast and G. Fairweather, NATO ASI Series, Series C: Mathematical and Physical Sciences, Vol. 203 (Reidel, Dordrecht, 1987) pp. 113–139.
[82] P. Rabinowitz and D.S. Lubinsky, Noninterpolatory integration rules for Cauchy principal value integrals, Math. Comp. 53 (1989) 279–295.
[83] P. Rabinowitz and N. Richter, Perfectly symmetric two-dimensional integration formulas with minimal number of points, Math. Comp. 23 (1969) 765–779.
[84] P. Rabinowitz and N. Richter, New error coefficients for estimating quadrature errors for analytic functions, Math. Comp. 24 (1970) 561–570.
[85] P. Rabinowitz and N. Richter, Asymptotic properties of minimal integration rules, Math. Comp. 24 (1970) 593–609.
[86] P. Rabinowitz and N. Richter, Chebyshev-type integration rules of minimum norm, Math. Comp. 24 (1970) 831–845.
[87] P. Rabinowitz and I.H. Sloan, Product integration in the presence of a singularity, SIAM J. Numer. Anal. 21 (1984) 149–166.
[88] P. Rabinowitz and W.E. Smith, Interpolatory product integration for Riemann-integrable functions, J. Austral. Math. Soc. Ser. B 29 (1987) 195–202.
[89] P. Rabinowitz and W.E. Smith, Interpolatory product integration in the presence of singularities: L_2 theory, J. Comp. Appl. Math. 39 (1992) 79–87.
[90] P. Rabinowitz and W.E. Smith, Interpolatory product integration in the presence of singularities: L_p theory, in: *Numerical Integration: Recent Developments, Software and Applications*, eds. T.O.

Espelid and A. Genz, NATO ASI Series, Series C: Mathematical and Physical Sciences, Vol. 357 (Kluwer, Dordrecht, 1992) pp. 93–109.

[91] P. Rabinowitz and G. Weiss, Tables of abscissas and weights for numerical evaluation of integrals of the form $\int_0^\infty e^{-x}x^n f(x)dx$, Math. Tables Aids Comp. 13 (1959) 285–294.

[92] A. Ralston and P. Rabinowitz, *A First Course in Numerical Analysis* (McGraw-Hill, New York, 1978).

[93] T. Schira, Ableitungsfreie Fehlerabschätzungen bei numerischer Integration holomorpher Funktionen, Dissertation, University of Karlsruhe (1994).

[94] I.H. Sloan and W.E. Smith, Properties of interpolatory product integration rules, SIAM J. Numer. Anal. 19 (1982) 427–442.

[95] W.E. Smith and I.H. Sloan, Product-integration rules based on the zeros of Jacobi polynomials, SIAM J. Numer. Anal. 17 (1980) 1–13.

[96] A.H. Stroud, *Approximate Calculation of Multiple Integrals* (Prentice-Hall, Englewood Cliffs, NJ, 1971).

[97] G. Szegö, Über gewisse orthogonale Polynome, die zu einer oszillierenden Belegungsfunktion gehören, Math. Ann. 110 (1935) 501–513. [Collected Papers (ed. R. Askey), Vol. 2, 545–557.]

29.5. [144] "Luigi Gatteschi's work on special functions and numerical analysis"

---

[144] "Luigi Gatteschi's work on special functions and numerical analysis," in *Special Functions* (G. Allasia, ed.), *Annals Numer. Math.* **2**, 3–19 (1995).

Luigi Gatteschi's work on special functions and numerical analysis

Walter Gautschi

Department of Computer Sciences, Purdue University, West Lafayette, IN 47907-1398, USA

A summary is given of Luigi Gatteschi's contributions to iterative algorithms generalizing the arithmetic-geometric mean algorithm, and to Chebyshev- and Gauss-type quadrature formulae.

1. It is customary at honorary meetings such as this to give a talk on one's recent research accomplishments. I have decided not to follow this practice and, instead, use this auspicious occasion to reflect on the work of the honoree, Luigi Gatteschi. After all, it is *him* that we honor, and it seemed appropriate to me to let *his* accomplishments take center stage.

It is hardly possible, of course, for anyone to be completely familiar with someone else's lifetime achievements, especially if they are as varied at Gatteschi's, and even less possible to summarize these in the short span of one hour. I will therefore concentrate on Gatteschi's work on Special Functions and Numerical Analysis, with which I am more familiar, and regretfully have to leave aside his voluminous work on asymptotics and estimation of zeros of special functions.[1] This is not to say that Gatteschi's work on these latter topics is completely unrelated to his work in numerical analysis. On the contrary, there is a close bond between these two topics, each supporting, and being motivated by, the other.

2. When one looks at a mathematician's life-time work, one cannot ignore the roots where he is coming from: What was the academic environment in which he grew up and in which his views on, and attitudes about, mathematics were shaped? Who were the major mathematical figures who had a direct and lasting influence on him, be it in the acquisition of the mathematical craft, in the formation of mathematical taste, or in the development of research areas? I think it is relatively easy to answer these questions in the case of Gatteschi. He was definitely a product of the famous Tuscan school of mathematics, of its strong tradition in analysis fostered in Pisa by such great masters as Luigi Bianchi and Ulisse Dini, and continued in Florence by Giovanni Sansone. It was here in Florence, where

[1] Some of this work, however, is being reviewed by R. Wong [52].

Gatteschi acquired a solid foundation in classical analysis and a deep appreciation for clarity of thought and simplicity of exposition, which were to become the trademarks of his writing. It is here also, where he eventually fell under the spell of Sansone, who was to guide him in his first research activities. These happen to concern some problems in number theory, in which Sansone was interested at the time. Much later, in 1982, when Gatteschi wrote a commemorative account of Sansone [26], he commented with modesty that he was able to make only little progress on the problems Sansone suggested to him. No wonder, because one of them was to prove the irreducibility over the rationals of Legendre polynomials (excepting the trivial cases of odd-degree polynomials), a problem which, to the best of my knowledge, is still open today!

After a brief stay at Stanford, where he encountered Gabor Szegö and followed a seminar of Johannes Van der Corput, and after three more years in Bari, Gatteschi moved to Turin, where he entered another renowned school of analysis, one that was founded by Lagrange and at the time had Francesco Tricomi as one of its prominent exponents. Tricomi, an analyst of unusual versatility, that included a strong interest in special functions, had just returned from a stint in the United States, where he collaborated on the Bateman project. He must have brought back with him a sense of the utility of special functions and of the importance of actually being able to compute them, with rigorous and realistic bounds for the error. I think it was Tricomi who instilled in Gatteschi an awareness of the constructive, algorithmic, and utilitarian side of mathematics. The desire to produce mathematics that is at the same time rigorous and applicable has certainly been one of the driving forces behind Gatteschi's work.

I will concentrate on two general areas in which Gatteschi has been active: Iteration and Numerical Quadrature. Special functions, as we will see, are intimately tied to both these areas, a feature that may well have attracted Gatteschi to these problem areas in the first place.

3. Iterations, say of the form

$$x_{n+1} = f(x_n, y_n), \quad y_{n+1} = g(x_n, y_n), \quad n = 0, 1, 2, \ldots, \tag{1}$$

have exerted a great deal of fascination among mathematicians and scientists alike, not only because of their importance in dynamical systems, but also because of their constructive, computational qualities. What is especially intriguing is the enormous diversity of phenomena they are able to describe: from completely chaotic behavior and patterns exhibiting fractal-like structures, to extremely regular behavior characterized by rapid convergence.

Among iterations (1) in which f and g are both homogeneous of degree 1, the best known is Gauss's algorithm of the arithmetic-geometric mean,

$$x_{n+1} = \tfrac{1}{2}(x_n + y_n), \quad y_{n+1} = \sqrt{x_n y_n}, \quad x_0 > 0,\ y_0 > 0 \tag{2}$$

(actually already studied by Lagrange in 1784). Special functions – indeed elliptic functions – make their appearance when one tries to study the convergence of the

iteration. Assuming $0 < y_0 < x_0 = 1$, which by symmetry and homogeneity does not restrict generality, Gauss in fact has shown that

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} y_n = \frac{\pi}{2}\frac{1}{K(\sqrt{1-y_0^2})}, \tag{3}$$

where

$$K(\omega) = \int_0^{\pi/2} (1-\omega^2\sin^2\phi)^{-1/2}d\phi, \quad 0 < \omega < 1, \tag{4}$$

is the complete elliptic integral of the first kind. A beautiful connection, indeed, between iteration and special functions! The algorithm is also of considerable interest to numerical analysts, as it converges quadratically and hence provides a fast and powerful algorithm to compute elliptic integrals.

What is remarkable about this algorithm is the fact that very minor changes in (2) may produce significant changes in the behavior of the iteration, echoing perhaps its propensity to chaotic behavior. One such minor change is to replace x_n in the geometric mean of (2) by x_{n+1}. This gives rise to what is now called Borchardt's algorithm,

$$x_{n+1} = \tfrac{1}{2}(x_n + y_n), \quad y_{n+1} = \sqrt{x_{n+1}y_n}, \quad x_0 \geq 0,\ y_0 \geq 0, \tag{5}$$

which, incidentally, was discussed by Borchardt in a letter of 1880 addressed to the Italian geometer Cremona to honor another Italian geometer, Chelini. (Interestingly, as was pointed out by Gatteschi in an article describing Fubini's juvenile mathematical work [25], Gauss in 1800 was already aware of this possible modification and wrote to Pfaff about it in a letter that has not been preserved. Pfaff's reply, however, eventually became widely known after publication of Volume 10 of Gauss's works in 1917, that is, 37 years after Borchardt's letter.) The iteration (5) still has a common limit, which now, however, involves elementary functions, for example, if $0 \leq x_0 < y_0$,

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} y_n = \frac{\sqrt{y_0^2 - x_0^2}}{\cos^{-1}(x_0/y_0)}, \tag{6}$$

and a similar limit involving the inverse hyperbolic cosine if $0 \leq y_0 < x_0$. Moreover, convergence is no longer quadratic, but only linear.

The algorithm (5) has a nice geometric illustration in terms of regular polygons. (Borchardt already mentioned this connection.) Let r be the radius of the polygon and a its apothem, i.e., the length of the perpendicular dropped from the center to any one of the sides (see figure 1), and let r', a' be the analogous quantities for the polygon with twice as many sides but the same perimeter. Then

$$a' = \tfrac{1}{2}(a+r), \quad r' = \sqrt{a'r},$$

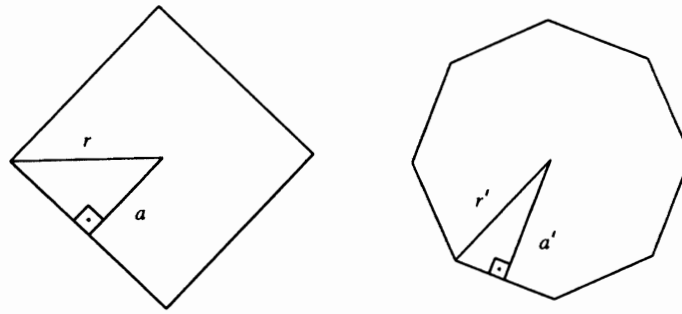


Figure 1. Geometric illustration of Borchardt's algorithm.

as can be verified by elementary geometry. These are exactly the relations underlying (5). If we take the common perimeter to be equal to 2 and start with the 2-gon, for which $a = 0$, $r = \frac{1}{2}$, we arrive at the algorithm (5) with $x_0 = 0$, $y_0 = \frac{1}{2}$, hence with the common limit equal to $\frac{1}{2}/\frac{1}{2}\pi = 1/\pi$. (Although this sequence converges to π^{-1} only linearly, there are other iterations which are known to converge quadratically to π; see, e.g., [8]. Euler's constant γ, in contrast, seems to be a more difficult constant [46] in that no iterations are known that would converge quadratically to γ; indeed, it is not even known whether γ is irrational [2].)

4. In [16], Gatteschi asks what would happen if in Gauss's arithmetic-geometric mean algorithm one were to replace the arithmetic mean by a more general weighted mean:

$$x_{n+1} = \frac{1}{k}(x_n + (k-1)y_n), \quad y_{n+1} = \sqrt{x_n y_n}, \quad n = 0, 1, 2, \ldots, k \geq 1. \tag{7}$$

This seems to be an intractable problem, and even today, not much is known about the limit, if $k \neq 2$. Gatteschi, on the other hand, discovers that a slight modification "à la Borchardt", which was suggested to him by Tricomi, namely

$$x_{n+1} = \frac{1}{k}(x_n + (k-1)y_n), \quad y_{n+1} = \sqrt{x_{n+1} y_n}, \quad n = 0, 1, 2, \ldots, \tag{8}$$

becomes indeed manageable. But not without effort! What is required are entirely new transcendental functions that generalize the trigonometric functions cosine and sine.

We saw in §3, equation (6), that the cosine function governs Borchardt's algorithm, which is (8) for $k = 2$. Now $\cos\sqrt{2}x$ is a solution of the functional equation

$$\phi(2x) = 2\phi^2(x) - 1, \tag{9}$$

indeed the unique solution that is analytic and even, and satisfies

$$\phi(x) = 1 - x^2 + \cdots. \tag{10}$$

The transcendental function that Gatteschi needs, to deal with (8) for $k \neq 2$, is likewise a solution of a functional equation, namely

$$\phi(\sqrt{2k}x) = k\phi^2(x) + 1 - k, \quad k \geq 1, \tag{11}$$

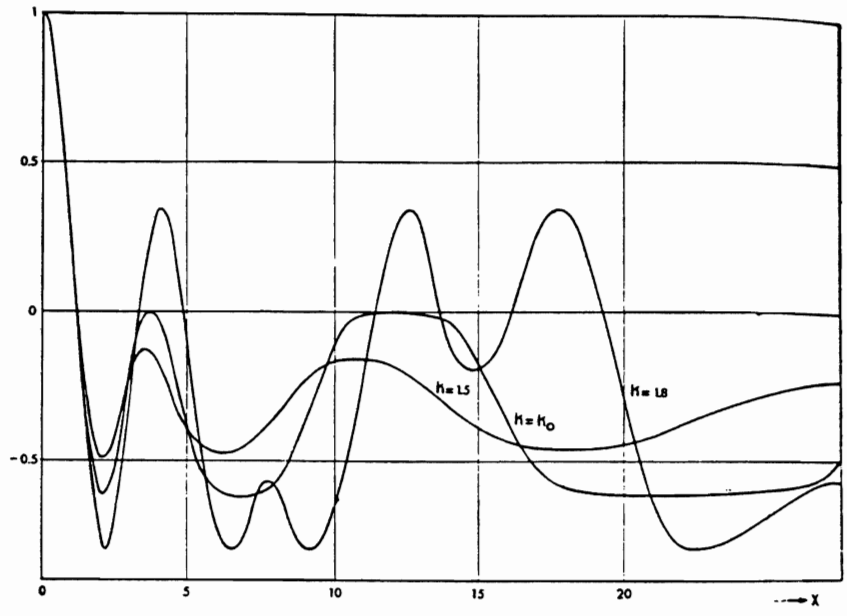


Figure 2. Gatteschi's function Φ_k, $k = 1.5$, $k = k_0$, $k = 1.8$.

and again the unique solution that is analytic, even, and satisfies (10). He denotes this solution by $\Phi_k(x)$ and proves that it is an entire function, just like $\Phi_2(x) = \cos\sqrt{2}x$. He also introduces the companion function $\Psi_k(x) = -\Phi_k'(x)/\sqrt{2}$, which for $k = 2$ becomes $\Psi_2(x) = \sin\sqrt{2}x$. Because of the importance of these functions in connection with (8), Gatteschi in [16] prepares tables for them and graphs, one of which is reproduced in figure 2. Subsequently, in [17], he gave a detailed analysis of their properties for real x and real $k > \frac{1}{2}$. (Note that $k = \frac{1}{2}$ yields the trivial solution $\phi \equiv 1$ of (11).) If $\frac{1}{2} < k \leq 1$, the function $\Phi_k(x)$ decreases monotonically for $x > 0$, but remains positive. If $k > 1$, the function becomes eventually oscillatory in a rather complicated way, the character of oscillation depending on whether $1 < k < k_0$, $k = k_0$, $k_0 < k < 2$, or $k > 2$; here, $k_0 = (1 + \sqrt{5})/2$ – yet another unexpected appearance of the golden ratio!

5. Returning to the iteration (8), Gatteschi assumes that

$$y_0 > 0, \quad x_0 > (1 - k)y_0. \tag{12}$$

It is clear from (10) that $\Phi_k(x)$ initially decreases, and Gatteschi in fact proves that for $k > 1$ there is a $\mu > 0$ such that $\Phi_k(x)$ decreases on $0 < x < \mu$ and has a local minimum at $x = \mu$, where $\Phi_k(\mu) = 1 - k$.

Now suppose first that $x_0 < y_0$. Then, by (12), one has $1 - k < x_0/y_0 < 1$. Therefore, from what was just said about Φ_k, we can write

$$\frac{x_0}{y_0} = \Phi_k(\alpha_0), \quad 0 < \alpha_0 < \mu, \tag{13}$$

where α_0 in the interval shown is uniquely determined. Gatteschi then shows that (8) has the common limit

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} y_n = y_0 \frac{\Psi_k(\alpha_0)}{\sqrt{2}\alpha_0} \quad (x_0 < y_0). \tag{14}$$

If, on the other hand, $y_0 < x_0$, then Φ_k and Ψ_k in (13), (14) have to be replaced by the "hyperbolic" functions $\Phi_k^*(x) = \Phi_k(ix)$ and $\Psi_k^*(x) = (1/i)\Psi_k(ix)$, respectively.

A further natural generalization of (8) is to introduce "weights" also in the geometric mean, that is, to consider

$$x_{n+1} = \frac{1}{k}(x_n + (k-1)y_n), \quad y_{n+1} = x_{n+1}^p y_n^{1-p}, \quad n = 0, 1, 2, \ldots, \tag{15}$$

where $k > 1$ and $0 < p < 1$. Gatteschi observes that this iteration can be treated analogously as above, the appropriate functional equation now being

$$\phi\left(\left(\frac{k}{1-p}\right)^p x\right) = k\phi^{1/(1-p)}(x) + 1 - k, \tag{16}$$

defining a function $\Phi_{k,p}(x)$ of two parameters.

One cannot help wondering how these new transcendental functions fit into the framework of classical special function theory. Are they subsumable to hypergeometric functions or generalized hypergeometric functions? My guess is that they are not. Do they satisfy an algebraic differential equation or are they like the gamma function, deprived of any such equation? I don't know.

6. An interesting application of iterations of the kind described is given by Gatteschi in [18], where he proposes to use them for computing infinite products

$$\prod_{n=0}^{\infty}(1 - aq^n), \quad |q| < 1 \tag{17}$$

(where, for convergence, it is assumed that $1 - aq^n \neq 0$ for all $n \geq 0$). These products are fundamental in the theory of generalized hypergeometric functions, just like the gamma function is fundamental in the theory of hypergeometric functions.

By an ingenious modification of the elementary arithmetic-harmonic mean algorithm

$$x_{n+1} = \frac{1}{2}(x_n + y_n), \quad y_{n+1} = \frac{2}{\frac{1}{x_n} + \frac{1}{y_n}}, \quad n = 0, 1, 2, \ldots \tag{18}$$

(which is known to converge for any $x_0 > 0, y_0 > 0$ to the geometric mean $\sqrt{x_0 y_0}$), Gatteschi arrives at a family of iterations depending on an arbitrary parameter $\xi \neq 1$: take x_0, y_0 such that

$$\frac{x_0}{y_0} = 1 - \frac{a}{1-\xi} \tag{19}$$

and iterate according to

$$\left.\begin{aligned} x_{n+1} &= x_n \frac{\xi y_n + (1-\xi)x_n}{y_n} \\ y_{n+1} &= x_{n+1} \frac{y_n}{qx_n + (1-q)y_n} \end{aligned}\right\} \quad n = 0, 1, 2, \ldots, \tag{20}$$

where a and q are the parameters in (17). It turns out that both sequences again have a common limit,

$$X = \lim_{n\to\infty} x_n = \lim_{n\to\infty} y_n = x_0 \prod_{n=0}^{\infty} (1 - aq^n). \tag{21}$$

Thus, if $x_0 \neq 0$, the infinite product (17) can be computed as X/x_0.

But how should the parameter ξ be chosen? Gatteschi's answer is one that one would expect of a numerical analyst: Arrange things so that the limit X in (21) is about halfway between x_n and y_n, at least for n sufficiently large. That led Gatteschi to investigate the ratio $(X - x_n)/(y_n - X)$, and he was able to prove, using a theorem of Cesàro on the equivalence of two limits, that

$$\lim_{n\to\infty} \frac{X - x_n}{y_n - X} = \frac{\xi - 1}{2 - q - \xi}.$$

Putting this equal to 1 yields the desired value of ξ,

$$\xi = \frac{3-q}{2}. \tag{22}$$

The initialization (19) then becomes

$$\frac{x_0}{y_0} = \frac{1 + 2a - q}{1 - q}, \tag{23}$$

and $x_0 \neq 0$ requires that $1 + 2a - q \neq 0$. This, however, is no serious restriction, since, if $1 + 2a - q = 0$, then $a^* = aq$ satisfies $1 + 2a^* - q = (1-q)^2 \neq 0$, and one simply uses

$$\prod_{n=0}^{\infty} (1 - aq^n) = \prod_{n=0}^{\infty} (1 - a^* q^{n-1}) = (1-a) \prod_{n=0}^{\infty} (1 - a^* q^n).$$

By design, the average $t_n = \frac{1}{2}(x_n + y_n)$ is a better approximation to X than either x_n or y_n.

When both a and q are real, say in the interval $(-1, 1)$ (which is no essential restriction, since $aq^n \in (-1, 1)$ for n sufficiently large), then Gatteschi shows that the sequences $\{x_n\}$, $\{y_n\}$ defined by (20)–(23) enjoy simple monotonicity properties, asymptotically for large n, provided the signs of x_0 and y_0 are chosen appropriately.

Gatteschi's iteration (20) has been used by Allasia and Bonardo [1] in 1980 to produce extensive 20-digit tables of the infinite product (17) and related products.

7. Gauss, with his arithmetic-geometric mean algorithm, provided the primary motivation for Gatteschi's work on iteration, although there were stimuli coming from other mathematicians such as Borchardt and, above all, Tricomi, that helped him along. It so happens that Gauss is also the prime inspirer of Gatteschi's work on quadrature, with two other eminent mathematicians – Chebyshev and Bernstein – close behind. I am referring, of course, to Gaussian quadrature rules and Chebyshev quadrature formulae. As I see it, Gatteschi's work on the latter is more substantial and important, and I therefore start with it.

8. The concern here is with quadrature formulae of the form

$$\int_a^b f(x)w(x)dx = c_n \sum_{\nu=1}^{n} f(x_\nu^{(n)}) + R_n^C(f) \tag{24}$$

on a finite or infinite interval (a,b) involving a nonnegative weight function w and having, as shown, all weights equal to c_n. It is furthermore required that all nodes $x_\nu^{(n)}$ be simple and real, and contained in $[a,b]$, say

$$a \le x_n^{(n)} < x_{n-1}^{(n)} < \cdots < x_2^{(n)} < x_1^{(n)} \le b; \tag{25}$$

one usually assumes that the degree of exactness is n, i.e.,

$$R_n^C(f) = 0 \text{ whenever } f \in \mathbb{P}_n. \tag{26}$$

(Here, $\mathbb{P}_n$ denotes the class of polynomials of degree $\le n$.) Occasionally, other degrees d of exactness – both $d < n$ and $d > n$ – are also considered. When (a,b) is infinite, one must assume, of course, that the integral on the left of (24) exists, e.g., that w has finite moments of all orders.

Note that if (24) has degree of exactness $d \ge 0$, in particular if (26) holds, then taking $f \equiv 1$ in (24) immediately gives

$$c_n = \frac{\mu_0}{n}, \quad \mu_0 = \int_a^b w(x)dx. \tag{27}$$

Chebyshev's work [9] in 1874 was inspired by what is today known as the Gauss-Chebyshev formula – the Gaussian quadrature on $[-1, 1]$ with weight function $w(x) = (1-x^2)^{-1/2}$ – which Chebyshev ascribed to Hermite and which indeed is a formula of the type (24) with $c_n = \pi/n$, having degree of exactness $d = 2n-1$. Can something similar be done in the case $w(x) \equiv 1$ on $[-1, 1]$, hence $c_n = 2/n$? Chebyshev found that this is indeed the case when $n = 1, 2, \ldots, 7$, constructing a formula (24) for which (25), (26) is satisfied. He even computed the nodes $x_\nu^{(n)}$ to six digits, but did not say why he stopped at $n = 7$. His motivation for being interested in equal weights was one of numerical stability: The influence of random errors in the function values upon the quadrature sum is minimized when all weights are equal. It is probably more accurate to say that this is a good excuse for studying such formulae! Indeed, nearly equal weights would do as well, for all practical purposes, but would deprive the theory of Chebyshev quadrature from much of its mathematical charm. Today, these formulae, particularly also

their analogues in higher dimensions, are more important in combinatorics because of their relevance to spherical designs [50,41]. They also play a role in electrostatics [40].

I mentioned that Chebyshev stopped at $n = 7$ without explanation. It was Radau [51] who six years later showed that when $n = 8$ the relations (24) (with $w(x) \equiv 1$ on $[-1, 1]$) and (25), (26) are incompatible, in that some of the nodes are necessarily complex. He did discover, however, a valid formula for $n = 9$. It took some fifty years until Bernstein [6] in 1937 proved that the formulae found by Chebyshev and Radau are in fact the only ones possible.

A large part of the theory of Chebyshev quadrature, therefore, is concerned with questions of existence and nonexistence, for one weight function w or another. It is easier to prove nonexistence, since it suffices to derive necessary conditions for (24)–(26) to hold; violation of a necessary condition implies nonexistence.

One such necessary condition was put forward by Bernstein, who compares the n-point Chebyshev formula (24) with the m-point Gauss formula for the same weight function w,

$$\int_a^b f(x)w(x)dx = \sum_{\mu=1}^{m} \gamma_\mu^{(m)} f(\xi_\mu^{(m)}) + R_m^G(f), \quad R_m^G(\mathbb{P}_{2m-1}) = 0. \tag{28}$$

Here the nodes $\xi_\mu^{(m)}$, ordered again decreasingly as

$$a < \xi_m^{(m)} < \xi_{m-1}^{(m)} < \cdots < \xi_2^{(m)} < \xi_1^{(m)} < b, \tag{29}$$

are the zeros of the mth-degree orthogonal polynomial relative to w, and $\gamma_\mu^{(m)}$ the corresponding Christoffel numbers (known to be all positive). Then Bernstein's necessary condition is the following: If (24) has degree of exactness $d = 2m - 1$, $m < n$, then

$$c_n \leq \gamma_1^{(m)}. \tag{30}$$

(Actually, Bernstein derived this condition only in the case $w(x) \equiv 1$ on $[-1, 1]$, but his method extends trivially to arbitrary nonnegative weight functions, yielding the more accurate condition $c_n \leq \min(\gamma_m^{(m)}, \gamma_1^{(m)})$.)

It is readily understood why a condition such as (30) would interest Gatteschi: it calls for a deep study of Christoffel numbers, and hence indirectly of zeros of orthogonal polynomials. In particular, one needs sharp inequalities for these quantities if one wants to refute (30) and arrive at realistic nonexistence results. Gatteschi's expertise on such matters thus finds here a wide-open field of application! He indeed has made profound use of the inequality (30) and also extended it to deal with more general quadrature formulae, as we will see shortly.

9. To begin with, it is natural to expect that something similar to what Bernstein proved for $w(x) = 1$ should hold also for ultraspherical weight functions $w(x) = w_\lambda(x) = (1 - x^2)^{\lambda - 1/2}$ on $[-1, 1]$. Gatteschi in [13] indeed uses Bernstein's inequality (30), in combination with delicate monotonicity and limit results of

O. Szász and M.T. Vacca concerning local extrema of ultraspherical polynomials, to show that, at least for $\lambda > 0$ (for $\lambda < 0$, see [10,11]), there is for every λ an integer $n_0(\lambda)$ such that for all $n > n_0(\lambda)$ the Chebyshev quadrature formula (24)–(26) for $w = w_\lambda$ does not exist. Actually, he proves something rather more precise: For the Chebyshev formula to exist, it must be true that

$$n > \frac{\Gamma\left(\left\lfloor \frac{n+2}{2} \right\rfloor + 2\lambda + 1\right)}{\Gamma\left(\left\lfloor \frac{n+2}{2} \right\rfloor\right)} C_\lambda, \tag{31}$$

where C_λ is the constant

$$C_\lambda = \frac{2^\lambda \Gamma^2(\lambda + \frac{1}{2})}{\Gamma(2\lambda+1)} j_{\lambda+1/2}^{1-2\lambda} J_{\lambda-1/2}^2(j_{\lambda+1/2}), \tag{32}$$

with $j_{\lambda+1/2}$ being the first positive zero of the Bessel function $J_{\lambda+1/2}$. Since the ratio of the Γ-functions in (31), as $n \to \infty$, behaves like $\lfloor (n+2)/2 \rfloor^{2\lambda+1}$, and $\lambda > 0$, it is clear that (31) cannot hold for n sufficiently large. It would be easy, from (31), (32) to determine the above $n_0(\lambda)$ explicitly for any given $\lambda > 0$. Gatteschi in fact does this for $\lambda = 1/2$ (i.e., $w_\lambda \equiv 1$) and obtains $n_0(1/2) = 13$ instead of the sharp $n_0(1/2) = 9$ proved by Bernstein.

Fifteen years later, Gatteschi in [34], together with Vinardi, returns to this problem and in the case $0 < \lambda < 1$ brings into play extremely sharp inequalities for zeros of ultraspherical polynomials, which are obtained by an imaginative use of the Sturm comparison theorem. Together with corresponding, equally sharp, inequalities for Christoffel numbers, Gatteschi and Vinardi obtain new necessary conditions for the n-point Chebyshev formula to have degree of extactness d. In the classical case $\lambda = 1/2$, the condition becomes

$$(d+1)(d+3) < \frac{4n}{J_1^2(j_0)} = 14.8415227\ldots n, \tag{33}$$

where J_1 is the Bessel function of order 1 and j_0 the first positive zero of J_0. If one takes, as in (26), by symmetry,

$$d = \begin{cases} n & \text{if } n \text{ is odd,} \\ n+1 & \text{if } n \text{ is even,} \end{cases} \tag{34}$$

then (33) is false for $n = 8$ and for $n \geq 10$, which precisely recovers Bernstein's result. For more general λ and d, in place of (33), they also obtain the necessary condition

$$d + 1 < \rho(\lambda) n^{1/(2\lambda+1)}, \quad 0 < \lambda < 1, \tag{35}$$

where $\rho(\lambda)$ is computable in terms of gamma functions, Bessel functions of order $\lambda + 1/2$, and the first positive zero of the Bessel function $J_{\lambda-1/2}$. This, too, in case $\lambda = 1/2$, reduces to – in fact, sharpens – Bernstein's well-known inequality $d < 4n^{1/2}$, replacing the factor 4 by $2/J_1(j_0) = 3.852469\ldots$.

Nonexistence results for Jacobi weight functions $w_{\alpha,\beta}(x) = (1-x)^{\alpha}(1+x)^{\beta}$ on $[-1, 1]$, have been proven, a few years after Gatteschi's 1963/64 paper, by A. Ossicini [47] for all parameters $\alpha > -1$, $\beta > -1$ outside the square $-1 < \alpha \leq -1/2$, $-1 < \beta \leq -1/2$. For Laguerre and Hermite weights, Gatteschi [14] already in 1964 proves nonexistence for all $n \geq 3$ and $n \geq 4$, respectively. (He has been preceded, however, by Krylov [42], who proved the same in 1958, also using Bernstein's method.)

10. The Chebyshev quadratures so far considered are typically *open* quadrature rules, in that all nodes are contained in the interior of the interval of integration. When, in the mid 1970s, Gatteschi asked what happens when one imposes nodes at the endpoints, perhaps even multiple nodes, and requires that only the interior nodes have constant weights associated with them, he entered completely uncharted territory. While analogous results were to be expected, the technical tools to derive them did not exist and had to be developed from scratch. In particular, the method of Bernstein had to be appropriately generalized.

This was done, in the general case of multiple endpoints, in Gatteschi's paper [21] and applied there to Jacobi weight functions with the expected nonexistence for n large enough proved also for *closed* Chebyshev quadrature formulae. Simple endpoints and the case of ultraspherical weight functions were treated jointly with Monegato and Vinardi in [36], and more definitively, in joint work with Vinardi [34]. There, one finds a particularly elegant extension of Bernstein's inequality, which I would like to briefly describe.

11. The formula to be considered in place of (24) is now

$$\int_{-1}^{1} f(x)w(x)dx = \bar{a}_n[f(-1)+f(1)] + \bar{c}_n \sum_{\nu=1}^{n} f(\bar{x}_\nu^{(n)}) + R_n^{\bar{C}}(f), \tag{36}$$
$$-1 < \bar{x}_n^{(n)} < \bar{x}_{n-1}^{(n)} < \cdots < \bar{x}_2^{(n)} < \bar{x}_1^{(n)} < 1,$$

where, for the sake of definiteness, the interval is taken to be $[-1, 1]$ and $w(x)$ assumed a nonnegative *even* weight function. In analogy to Bernstein's method, one now associates with (36) the $(m+2)$-point Gauss-Lobatto formula

$$\int_{-1}^{1} f(x)w(x)dx = \alpha_m[f(-1)+f(1)] + \sum_{\mu=1}^{m} \gamma_\mu^{(m)} f(\xi_\mu^{(m)}) + R_m^{GL}(f), \tag{37}$$
$$-1 < \xi_m^{(m)} < \xi_{m-1}^{(m)} < \cdots < \xi_2^{(m)} < \xi_1^{(m)} < 1,$$

for which

$$R_m^{GL}(f) = 0, \quad \text{all } f \in \mathbb{P}_{2m+1}. \tag{38}$$

It is well known that the nodes $\xi_\mu^{(m)}$ in (37) are the zeros of the polynomial of degree m which is orthogonal relative to the weight function $w(x)(1-x^2)$, and the weights $\gamma_\mu^{(m)}$ are expressible in terms of the corresponding Christoffel numbers divided by

$1 - [\xi_\mu^{(m)}]^2$. It is much less obvious how Bernstein's necessary condition (30) has to be adapted to (36) and (37). Gatteschi and Vinardi [34], however, using interesting techniques (among which a result of Erdös and Turán on interpolation), succeed in establishing the simple and elegant necessary condition

$$2\bar{a}_n + \bar{c}_n(1 + \xi_1^{(m)}) < 2\alpha_m + \gamma_1^{(m)}(1 + \xi_1^{(m)}), \tag{39}$$

assuming that (36) has degree of exactness $d = 2m + 1$, $m < n$,

$$R_n^{\bar{C}}(f) = 0, \quad \text{all } f \in \mathbb{P}_{2m+1}, \quad m < n. \tag{40}$$

For (36) to be a closed Chebyshev formula requires, by symmetry,

$$R_n^{\bar{C}}(f) = 0 \quad \text{if} \begin{cases} f \in \mathbb{P}_{n+2}, & n \text{ odd}, \\ f \in \mathbb{P}_{n+3}, & n \text{ even}. \end{cases} \tag{41}$$

Thus, in (40), one takes $m = (n + 1)/2$ if n is odd, and $m = (n + 2)/2$ if n is even, to arrive at a necessary condition for the existence of a closed Chebyshev formula. This is carried out by Gatteschi and Vinardi in the case of ultraspherical weight functions w_λ with $0 < \lambda < 1$. The same very sharp inequalities for zeros and Christoffel numbers that led to the condition (33) are again brought to bear on this new problem. They produce results which, while not simple, nevertheless allow the conclusion that, in the classical case $\lambda = 1/2$, the condition (39) is violated if $n = 24$ and $n \geq 26$. Combined with numerical tests, this then implies that the formula (36) for $w(x) \equiv 1$, satisfying (41), can only exist if $1 \leq n \leq 11$. The respective formulae, indeed, have previously been calculated in joint work with Monegato and Vinardi [36].

Half-open Chebyshev quadratures with, say, one node at -1 and n nodes interior to $[-1, 1]$, have been considered in [22] for constant weight function $w(x) \equiv 1$. Here, Gatteschi shows that, if one stipulates a nonnegative boundary weight and a positive common weight at the interior points, then a formula having degree of exactness $d = n + 1$ cannot exist if $n = 10$ and $n \geq 12$. Baratella [3], shortly thereafter, actually shows that one has existence only for $1 \leq n \leq 6$, and that $n = 2, 4, 6$ give the classical Chebyshev formulae (i.e., with zero boundary weight) and $n = 1$ the Gauss-Radau formula. The two remaining ones she calculates to 10 decimals.

Bernstein's technique and the various extensions of it due to Gatteschi and his collaborators establish a strong link between Chebyshev quadrature rules and Gaussian rules, although not relative to the same number of nodes. One can ask, on the other hand, whether a Gaussian quadrature formula can at the same time be a Chebyshev formula, i.e., have all weights equal. An answer to this was already given in 1875 – one year after Chebyshev's original work – by the Russian mathematician K.A. Posse [49], who proved that the Gauss-Chebyshev formula, which served as inspiration for Chebyshev's work, is the only one (up to a linear transformation) which has this equal-weight property *for all n*. In 1984, however, I pointed out [38] that if this property is required only for even n, then there are other weight functions that fit the bill, e.g., weight functions supported on two

separate intervals. Peherstorfer [48], indeed, showed that given any proper subsequence of the positive integers, one can find explicitly all weight functions which admit equally-weighed n-point Gauss formulae for all n of that subsequence.

Analogous questions for closed Chebyshev formulae of the type (36), but allowing for endpoints with multiplicity $r \geq 1$, have been partially answered in [36], where it is shown by a tour-de-force argument that the analogue of Posse's result holds within the class of Jacobi weight functions. There is room here for further research along the lines of [38] and [48].

12. Straight Gauss-type quadrature formulae, i.e., formulae with maximum degree of exactness, on the interval $[-1, 1]$, that have endpoints of multiplicity 2, have been considered by Gatteschi [12] already in 1963 in the case of a constant weight function $w(x) \equiv 1$. The motivation given by Gatteschi for studying such formulae is typically of a practical nature: If one composes them over m subintervals of some given interval $[a, b]$, then ordinary Gauss, Gauss-Lobatto (simple endpoints), and his generalized Gauss-Lobatto rule, all adjusted to have the same degree of exactness $2n - 1$, require respectively nm, $nm + 1$, and $nm + 3 - m$ function evaluations, although the last requires, in addition, two derivative values, at the endpoints of $[a, b]$. Thus, for $m > 3$, Gatteschi's generalized rule is the most efficient one in terms of function evaluations. Such formulae, more recently, have found use in spectral methods [5].

The principal merit of Gatteschi's work on this particular generalization is the derivation of explicit formulae for the quadrature weights: those associated with the endpoints are given as rational functions of n, the number of interior nodes, while those associated with the interior nodes are expressed in terms of the ultraspherical polynomial $P_{n+1}^{(5/2)}$ evaluated at the zeros of $P_n^{(5/2)}$. Explicit formulae and bounds for the remainder term are also given. Finally, Gatteschi provides 12-digit tables of nodes and weights for $n = 1(1)16$, which he computes by Newton's method, taking as initial approximations (for the interior nodes) – what else? – asymptotic approximations.

Analogous formulae for all four Chebyshev weight functions are given by Li and myself in [39], and by Bernardi and Maday in [4] for ultraspherical weights and endpoints of arbitrary multiplicity.

13. The development of sharp asymptotic approximations and inequalities for zeros of orthogonal polynomials and Christoffel numbers indeed has preoccupied Gatteschi ever since. Thus, in the case of ultraspherical polynomials, the approximations given in [23] are so sharp that they allow one to compute the respective Gauss nodes to 20 correct decimals in just one iteration, if one applies a high-order iterative method of Lether [43]. Equally sharp approximations, by a resourceful use of the Sturm comparison theorem, are developed in [27] for zeros of Jacobi polynomials with parameters $|\alpha| \leq 1/2$, $|\beta| \leq 1/2$, in [29,30] for zeros of generalized Laguerre polynomials, and in [31] – just three years ago – for zeros of confluent hypergeometric functions. Gatteschi indeed shows that many of these approximations, when the O-term is removed, turn into sharp inequalities

for the zeros in question. Similar techniques, incidentally, were also applied by Gatteschi and Laforgia [32] to obtain interesting estimates for the first positive zero and the abscissa of the first maximum of the Bessel function J_ν of order $\nu > 0$.

Sharp asymptotic approximations for Christoffel numbers in the case of arbitrary Jacobi weight functions are derived in [28].

Curiously, in one of Gatteschi's early work [15], asymptotic formulae for Legendre polynomials, and more generally, for ultraspherical polynomials, that involve Bessel functions, are turned around to approximate, and hence compute, Bessel functions in terms of Legendre and ultraspherical polynomials. A neat idea, not entirely ineffective, but one that is unlikely to be found in modern software for computing Bessel functions!

14. There is more recent work of Gatteschi, however, that is highly relevant to modern software for computing Gaussian quadrature formulae with nonstandard weight functions w on some interval (a, b). This often requires the computation of "modified moments" $\nu_n = \int_a^b \pi_n(x)w(x)dx$, $n = 0, 1, 2, \ldots$, of w relative to a system of classical orthogonal polynomials $\{\pi_n\}$. Gatteschi takes up this problem in [24], where he looks at two particular weight functions. The first is $w(x) = x^\rho(1-x)^\alpha \ln(1/x)$ on (0, 1), where $\rho > -1$, $\alpha > -1$. Taking for π_n the shifted Jacobi polynomial with parameters α, β (the same α as in the weight function), he succeeds in computing ν_n explicitly in terms of the gamma function and its logarithmic derivative. This continues a line of research started by Blue [7] and myself [37] a year earlier. The second weight function is $w(x) = e^{-x}x^\rho \ln^p x$ on $(0, \infty)$, with $\rho > -1$ and $p = 1$ and 2. The choice of generalized Laguerre polynomials with parameter α for $\{\pi_n\}$ then yields results similar to those for the first weight function. A special case gives modified Hermite moments for $w(x) = e^{-x^2} \ln^p x$ on $(0, \infty)$.

15. Gauss quadrature rules have found many applications, both inside and outside of numerical analysis. In joint work with Lyness, Gatteschi makes two such applications. The first, in [33], is to weighted integrals $\int_{-\pi}^{\pi} w(\theta)f(\theta)d\theta$ of 2π-periodic functions f extended over a full period. The standard approach for constructing quadrature formulae having maximum trigonometric degree of exactness is to use orthogonal trigonometric polynomials. Lyness and Gatteschi, instead, use transformations of variables to reduce the problem to an ordinary ("algebraic") Gauss quadrature problem, albeit for a rather tricky weight function on $(-\infty, \infty)$ given by $w(2\tan^{-1} t)/(1+t^2)^{d+1}$, where $w(\cdot)$ is the weight function in the given integral, and d the optimal trigonometric degree. In [44,45], they apply Gauss and Gauss-Radau formulae to construct optimal $(d+1)$-point quadrature rules that are exact for all functions $t^\lambda(\sqrt{1+t^2})^\mu$ with nonnegative integers λ, μ satisfying $\lambda + \mu \leq d$. Such integration problems arise in cubature over a triangle when the functions to be integrated exhibit certain singular behavior at a vertex or along the sides of the triangle.

16. Any account of Gatteschi's work on special functions and numerical analysis

would be incomplete without mentioning his books on the subject. The one on special functions [20] is an impressive treatise covering the gamma function, hypergeometric and confluent hypergeometric functions, orthogonal polynomials, Legendre and Bessel functions and other assorted functions, not including, however, elliptic functions and integrals. This is a most lucid and beautiful account of special function theory, and it is a pity that the book – written in Italian – has never been translated into English and made accessible to a larger audience.

There are two books by Gatteschi on Numerical Analysis, [35] and [19], the first written jointly with T. Zeuli. While both texts reflect the state of the subject at the time they were written, they still make for useful reading because of the richness in numerical examples and the close attention given to rigorous error estimates.

References

[1] G. Allasia and F. Bonardo, On the numerical evaluation of two infinite products, Math. Comp. 35 (1980) 917–931.

[2] D. Bailey, Numerical results on the transcendence of constants involving π, e, and Euler's constant, Math. Comp. 50 (1988) 275–281.

[3] P. Baratella, Formule di quadratura alla Tchebycheff di tipo semichiuso, Atti Accad. Sci. Torino Cl. Sci. Fis. Mat. Natur. 112 (1978) 179–184.

[4] C. Bernardi and Y. Maday, Some spectral approximations of one-dimensional fourth-order problems, in *Progress in Approximation Theory*, eds. P. Nevai and A. Pinkus (Academic Press, Boston, 1991) pp. 43–116.

[5] C. Bernardi, G. Coppoletta and Y. Maday, Some spectral approximations of two-dimensional fourth-order problems, Math. Comp. 59 (1992) 63–76.

[6] S.N. Bernstein, Sur les formules de quadrature de Cotes et Tchebycheff, C.R. Acad. Sci. URSS 14 (1937) 323–326. [*Collected Works*, vol. 2 (Izdat. Akad. Nauk SSSR, Moscow, 1954) pp. 200–204 (in Russian).]

[7] J.L. Blue, A Legendre polynomial integral, Math. Comp. 33 (1979) 739–741.

[8] J.M. Borwein and P.B. Borwein, *Pi and the AGM* (Wiley, New York, 1987).

[9] P.L. Chebyshev, Sur les quadratures, J. Math. Pures Appl. (2) 19 (1874) 19–34. [*Œuvres*, vol. 2 (Chelsea, New York, 1962) pp. 165–180.]

[10] K.-J. Förster, On Chebyshev quadrature for ultraspherical weight functions, Calcolo 23 (1986) 355–381.

[11] K.-J. Förster, Problem 5: A problem in Chebyshev quadrature for ultraspherical weight functions, in *Numerical Integration IV*, eds. H. Brass and G. Hämmerlin, Internat. Ser. Numer. Math. 112 (Birkhäuser, Basel, 1993) pp. 378–379.

[12] L. Gatteschi, Su una formula di quadratura "quasi gaussiana". Tabulazione delle ascisse d'integrazione e delle relative costanti di Christoffel, Atti Accad. Sci. Torino Cl. Sci. Fis. Mat. Natur. 98 (1963–64) 641–661.

[13] L. Gatteschi, Su di un problema connesso alle formule di quadratura di Tschebyscheff, Rend. Sem. Mat. Univ. Politec. Torino 23 (1963–64) 75–87.

[14] L. Gatteschi, Sulla non esistenza di certe formule di quadratura, Rend. Sem. Mat. Univ. Politec. Torino 24 (1964–65) 157–172.

[15] L. Gatteschi, Su un metodo di calcolo numerico delle funzioni di Bessel di prima specie, Rend. Sem. Mat. Univ. Politec. Torino 25 (1965–66) 109–120.

[16] L. Gatteschi, Su una generalizzazione dell' algoritmo iterativo di Borchardt, Mem. Accad. Sci. Torino Cl. Sci. Fis. Mat. Natur. (4) n. 4 (1966) 3–18.

[17] L. Gatteschi, Su di una equazione funzionale generalizzante quella del coseno, Rend. Sem. Mat. Univ. Politec. Torino 26 (1966–67) 65–86.
[18] L. Gatteschi, Procedimenti iterativi per il calcolo numerico di due prodotti infiniti, Rend. Sem. Mat. Univ. Politec. Torino 29 (1969–70) 187–201.
[19] L. Gatteschi, *Lezioni di Analisi Numerica* (Levrotto & Bella, Torino, 1971).
[20] L. Gatteschi, *Funzioni Speciali* (Unione Tipografico-Editrice Torinese, 1973).
[21] L. Gatteschi, Il problema di Tchebycheff per formule di quadratura di tipo chiuso, Boll. Un. Mat. Ital. (4) 11, Suppl. fasc. 3 (1975) 641–653.
[22] L. Gatteschi, Alcuni risultati sulle formule di quadratura del tipo di Tchebycheff, Rend. Mat. (6) 10 (1977) 523–533.
[23] L. Gatteschi, On the construction of some Gaussian quadrature rules, in *Numerische Integration*, ed. G. Hämmerlin, Internat. Ser. Numer. Math. 45 (Birkhäuser, Basel, 1979) pp. 138–146.
[24] L. Gatteschi, On some orthogonal polynomial integrals, Math. Comp. 35 (1980) 1291–1298.
[25] L. Gatteschi, Il contributo di Guido Fubini agli algoritmi iterativi, Atti Acc. Sci. Torino Cl. Sci. Fis. Mat. Natur., Suppl. 115 (1982) 61–70.
[26] L. Gatteschi, Giovanni Sansone (1888–1979), Atti Accad. Sci. Torino Cl. Sci. Fis. Mat. Natur. 116 (1982) 483–488.
[27] L. Gatteschi, New inequalities for the zeros of Jacobi polynomials, SIAM J. Math. Anal. 18 (1987) 1549–1562.
[28] L. Gatteschi, Uniform approximation of Christoffel numbers for Jacobi weight, in *Numerical Integration III*, eds. H. Brass and G. Hämmerlin, Internat. Ser. Numer. Math. 85 (Birkhäuser, Basel, 1988) pp. 49–59.
[29] L. Gatteschi, Some new inequalities for the zeros of Laguerre polynomials, in *Numerical Methods and Approximation Theory III*, ed. G.V. Milovanović (University of Niš, Niš, 1988) pp. 23–38.
[30] L. Gatteschi, Uniform approximations for the zeros of Laguerre polynomials, in *Numerical Mathematics Singapore 1988*, eds. R.P. Agarwal, Y.M. Chow and S.J. Wilson, Internat. Ser. Numer. Math. 86 (Birkhäuser, Basel, 1988) pp. 137–148.
[31] L. Gatteschi, New inequalities for the zeros of confluent hypergeometric functions, in *Asymptotic and Computational Analysis*, ed. R. Wong, Lecture Notes Pure Appl. Math. 124 (Dekker, New York, 1990) pp. 175–192.
[32] L. Gatteschi and A. Laforgia, Nuove disuguaglianze per il primo zero ed il primo massimo della funzione di Bessel $J_\nu(x)$, Rend. Sem. Mat. Univ. Politec. Torino 34 (1975–76) 411–424.
[33] L. Gatteschi and J.N. Lyness, An indirect approach to trigonometric quadrature rules, Calcolo 20 (1983) 191–210.
[34] L. Gatteschi and G. Vinardi, Sul grado di precisione di formule di quadratura del tipo di Tchebycheff, Calcolo 15 (1978) 59–85.
[35] L. Gatteschi and T. Zeuli, *Introduzione alla Analisi Numerica* (Editrice Tirrenia, Torino, 1966).
[36] L. Gatteschi, G. Monegato and G. Vinardi, Alcuni problemi relativi alle formule di quadratura del tipo di Tchebycheff, Calcolo 13 (1976) 79–104.
[37] W. Gautschi, On the preceding paper "A Legendre polynomial integral" by James L. Blue, Math. Comp. 33 (1979) 742–743.
[38] W. Gautschi, On some orthogonal polynomials of interest in theoretical chemistry, BIT 24 (1984) 473–483.
[39] W. Gautschi and S. Li, Gauss-Radau and Gauss-Lobatto quadratures with double end points, J. Comp. Appl. Math. 34 (1991) 343–360.
[40] J. Korevaar and J.L.H. Meyers, Spherical Faraday cage for the case of equal point charges and Chebyshev-type quadrature on the sphere, Integral Transforms and Special Functions 1 (1993) 105–117.
[41] J. Korevaar and J.L.H. Meyers, Chebyshev-type quadrature on multidimensional domains, J. Approx. Theory 79 (1994) 144–164.

[42] V.I. Krylov, Mechanical quadratures with equal coefficients for the integrals $\int_0^\infty e^{-x} f(x)dx$ and $\int_{-\infty}^\infty e^{-x^2} f(x)dx$ (Russian), Dokl. Akad. Nauk BSSR 2 (1958) 187–192.
[43] F.G. Lether, On the construction of Gauss-Legendre quadrature rules, J. Comp. Appl. Math. 4 (1978) 47–51.
[44] J.N. Lyness and L. Gatteschi, A note on cubature over a triangle of a function having specified singularities, in *Numerical Integration*, ed. G. Hämmerlin, Internat. Ser. Numer. Math. 57 (Birkhäuser, Basel, 1982) pp. 164–169.
[45] J.N. Lyness and L. Gatteschi, On quasi degree quadrature rules, Numer. Math. 39 (1982) 259–267.
[46] J. Nunemacher, On computing Euler's constant, Math. Mag. 65 (1992) 313–322.
[47] A. Ossicini, Sulle formule di quadratura di Tschebyscheff, Pubbl. Ist. Naz. Appl. Calcolo, n. 660, quad. 7 (1966) 43–59.
[48] F. Peherstorfer, Gauss-Tchebycheff quadrature formulas, Numer. Math. 58 (1990) 273–286.
[49] K.A. Posse, Sur les quadratures, Nouv. Ann. Math. (2) 14 (1875) 49–62.
[50] P. Rabau and B. Bajnok, Bounds for the number of nodes in Chebyshev type quadrature formulas, J. Approx. Theory 67 (1991) 199–214.
[51] R. Radau, Etude sur les formules d'approximation qui servent à calculer la valeur numérique d'une intégrale définie, J. Math. Pures Appl. (3) 6 (1880) 283–336.
[52] R. Wong, Error bounds for asymptotic approximations of special functions, Ann. Numer. Math. 2 (1995), this volume.

29.6. [170] "THE INTERPLAY BETWEEN CLASSICAL ANALYSIS AND (NUMERICAL) LINEAR ALGEBRA — A TRIBUTE TO GENE H. GOLUB"

---

[170] "The Interplay Between Classical Analysis and (Numerical) Linear Algebra — A Tribute to Gene H. Golub," *Electron. Trans. Numer. Anal*. **13**, 119–147 (2002).

Electronic Transactions on Numerical Analysis.
Volume 13, pp. 119-147, 2002.

ISSN 1068-9613.

ETNA
Kent State University
etna@mcs.kent.edu

THE INTERPLAY BETWEEN CLASSICAL ANALYSIS AND (NUMERICAL) LINEAR ALGEBRA — A TRIBUTE TO GENE H. GOLUB*

WALTER GAUTSCHI†

Dedicated in friendship, and with high esteem, to Gene H. Golub on his 70th birthday

Abstract. Much of the work of Golub and his collaborators uses techniques of linear algebra to deal with problems in analysis, or employs tools from analysis to solve problems arising in linear algebra. Instances are described of such interdisciplinary work, taken from quadrature theory, orthogonal polynomials, and least squares problems on the one hand, and error analysis for linear algebraic systems, element-wise bounds for the inverse of matrices, and eigenvalue estimates on the other hand.

Key words. Gauss-type quadratures, eigenvalue/vector characterizations, orthogonal polynomials, modification algorithms, polynomials orthogonal on several intervals, least squares problem, Lanczos algorithm, bounds for matrix functionals, iterative methods.

AMS subject classifications. 65D32, 33C45, 65D10, 15A45, 65F10.

1. Introduction.

It has been a privilege for me to have known Gene Golub for so many years and to have been able to see his very extensive work unfold. What intrigues me most about his work — at least the part I am familiar with — is the imaginative use made of linear algebra in problems originating elsewhere. Much of Golub's work, indeed, can be thought of as lying on the interface between classical analysis and linear algebra. The interface, to be sure, is directional: a problem posed in analysis may be solved with the help of linear algebra, or else, a linear algebra problem solved with tools from analysis. Instances of the former type occur in quadrature problems, orthogonal polynomials, and least squares problems, while examples of the latter type arise in error estimates for the solution of linear algebraic systems, element-wise bounds for the inverse of a matrix, and in eigenvalue estimates of interest in iterative methods.

It will not be possible here to pursue all the ramifications of this interesting interplay between different disciplines, but we try to bring across some of the main ideas and will refer to the literature for variations and extensions.

2. Quadrature.

Integration with respect to some given measure $d\lambda$ on the real line $\mathbb{R}$ is certainly a topic that belongs to analysis, and so is the evaluation or approximation of integrals $\int_{\mathbb{R}} f(t)d\lambda(t)$. If one follows Gauss, one is led to orthogonal polynomials relative to the measure $d\lambda$, which is another vast area of classical analysis. How does linear algebra enter in all of this? It was in 1969 when the connection between Gauss quadrature rules and the algebraic eigenvalue problem was, if not discovered, then certainly exploited in the now classical and widely cited paper [33]. We begin with giving a brief account of this work, and then discuss various extensions thereof made subsequently.

*Received August 14, 2002. Accepted for publication August 20, 2002. Recommended by L. Reichel. Expanded version of a lecture presented in a special session honoring Professor Gene H. Golub at the Latsis Symposium 2002 on *Iterative Solvers for Large Linear Systems, celebrating 50 years of the conjugate gradient method*, held at the Swiss Federal Institute of Technology in Zurich, February 18–21, 2002.

†Department of Computer Sciences, Purdue University, West Lafayette, Indiana 47907-1398. E-mail: wxg@cs.purdue.edu

ETNA
Kent State University
etna@mcs.kent.edu

2.1. Gauss quadrature. Assume $d\lambda$ is a positive measure on $\mathbb{R}$, all (or sufficiently many) of whose moments

$$\mu_r = \int_{\mathbb{R}} t^r d\lambda(t), \quad r = 0, 1, 2, \dots, \tag{2.1}$$

exist with $\mu_0 > 0$. The n-point *Gauss quadrature rule* for $d\lambda$ is

$$\int_{\mathbb{R}} f(t) d\lambda(t) = \sum_{\nu=1}^{n} \lambda_\nu f(\tau_\nu) + R_n(f), \tag{2.2}$$

where $\lambda_\nu = \lambda_\nu^{(n)}$, $\tau_\nu = \tau_\nu^{(n)}$ depend on n and $d\lambda$, and $R_n(f) = 0$ whenever f is a polynomial of degree $\leq 2n-1$,

$$R_n(f) = 0, \quad f \in \mathbb{P}_{2n-1}. \tag{2.3}$$

This is the maximum degree possible. If $f^{(2n)}$ is continuous on the support of $d\lambda$ and has constant sign, then

$$R_n(f) > 0 \quad \text{if } \operatorname{sgn} f^{(2n)} = 1, \tag{2.4}$$

with the inequality reversed if $\operatorname{sgn} f^{(2n)} = -1$.

The connection between Gauss quadrature and orthogonal polynomials is well known. If $\pi_k(\cdot) = \pi_k(\cdot\,; d\lambda)$, $k = 0, 1, 2, \dots$, denotes the system of (monic) polynomials orthogonal with respect to the measure $d\lambda$,

$$\int_{\mathbb{R}} \pi_k(t)\pi_\ell(t) d\lambda(t) \begin{cases} = 0 & \text{if } k \neq \ell, \\ > 0 & \text{if } k = \ell, \end{cases} \tag{2.5}$$

then $\tau_1, \tau_2, \dots, \tau_n$ are the zeros of $\pi_n(\cdot\,; d\lambda)$, and the λ_ν can be expressed in terms of the orthogonal polynomials as well. The former are all distinct and contained in the interior of the support interval of $d\lambda$, the latter all positive. What is important here is the well-known fact that the orthogonal polynomials satisfy a three-term recurrence relation,

$$\begin{gathered} \pi_{k+1}(t) = (t - \alpha_k)\pi_k(t) - \beta_k \pi_{k-1}(t), \quad k = 0, 1, 2, \dots, \\ \pi_{-1}(t) = 0, \quad \pi_0(t) = 1, \end{gathered} \tag{2.6}$$

with well-determined real coefficients $\alpha_k = \alpha_k(d\lambda)$ and $\beta_k = \beta_k(d\lambda) > 0$. In terms of these, one defines the *Jacobi matrix*

$$\boldsymbol{J}(d\lambda) = \begin{bmatrix} \alpha_0 & \sqrt{\beta_1} & & & 0 \\ \sqrt{\beta_1} & \alpha_1 & \sqrt{\beta_2} & & \\ & \sqrt{\beta_2} & \alpha_2 & \sqrt{\beta_3} & \\ & & \ddots & \ddots & \ddots \\ 0 & & & & \end{bmatrix}, \tag{2.7}$$

in general an infinite symmetric tridiagonal matrix. Its leading principal minor matrix of order n will be denoted by

$$\boldsymbol{J}_n(d\lambda) = [\boldsymbol{J}(d\lambda)]_{[1:n,1:n]}. \tag{2.8}$$

ETNA
Kent State University
etna@mcs.kent.edu

Here, then, is the connection between the Gauss quadrature formula (2.2) and the algebraic eigenvalue problem: *the Gauss nodes τ_ν are the eigenvalues of $\boldsymbol{J}_n(\mathrm{d}\lambda)$, whereas the Gauss weights λ_ν are*

$$\lambda_\nu = \mu_0 v_{\nu,1}^2, \tag{2.9}$$

where $v_{\nu,1}$ is the first component of the normalized eigenvector v_ν corresponding to the eigenvalue τ_ν. The eigenvalue charaterization of the nodes τ_ν is an easy consequence of the recurrence relation (2.6) and has been known for some time prior to the 1960s. The characterization (2.9) of the weights λ_ν is more intricate and seems to have first been observed in 1962 by Wilf [45, Ch.2, Exercise 9], or even previously, around 1954, by Goertzel [46]; it has also been used by the physicist Gordon in [34, p. 658]. The merit of Golub's work in [33] is to have clearly realized the great computational potential of this result and in fact to have developed a stable and efficient computational procedure based on the QL algorithm.

It is useful to note that the quadrature sum in (2.2) for smooth functions f can be written in terms of $\boldsymbol{J}_n = \boldsymbol{J}_n(\mathrm{d}\lambda)$ as

$$\sum_{\nu=1}^{n} \lambda_\nu f(\tau_\nu) = \mu_0 \boldsymbol{e}_1^T f(\boldsymbol{J}_n)\boldsymbol{e}_1, \quad \boldsymbol{e}_1^T = [1, 0, \ldots, 0] \in \mathbb{R}^n. \tag{2.10}$$

This follows readily from the spectral decomposition of $\boldsymbol{J}_n$ and (2.9). Also, for the remainder $R_n(f)$ in (2.2) one has (cf. [21, p. 291, (vii)])

$$|R_n(f)| \le \frac{\|f^{(2n)}\|_\infty}{(2n)!} \int_{\mathbb{R}} \pi_n^2(t)\mathrm{d}\lambda(t) = \frac{\|f^{(2n)}\|_\infty}{(2n)!}\, \beta_0\beta_1 \cdots \beta_n, \tag{2.11}$$

provided $f^{(2n)}$ is continuous on the support $\mathrm{supp}(\mathrm{d}\lambda)$ of $\mathrm{d}\lambda$. The ∞-norm of $f^{(2n)}$ is the maximum of $|f^{(2n)}|$ on $\mathrm{supp}(\mathrm{d}\lambda)$, and (2.11) holds regardless of whether or not $f^{(2n)}$ has constant sign.

The Jacobi matrix $\boldsymbol{J}_n(\mathrm{d}\lambda)$, and with it the Gauss quadrature rule, is uniquely determined by the first $2n$ moments $\mu_0, \mu_1, \ldots, \mu_{2n-1}$ of the measure $\mathrm{d}\lambda$. The *Chebyshev algorithm* (cf. [21, §2.3]) is a vehicle for passing directly from these $2n$ moments to the $2n$ recursion coefficients α_k, β_k, $k = 0, 1, \ldots, n-1$. Although numerically unstable, the procedure can be carried out in symbolic computation to arbitrary precision. (A Maple 5 script named `cheb.mws` can be found on the internet at `http://www.cs.purdue.edu/archives/2001/wxg/codes`.)

2.2. Gauss-Radau and Gauss-Lobatto quadrature. If the support of $\mathrm{d}\lambda$ is a finite interval $[a, b]$, the Gauss quadrature formula can be modified by requiring that one or both of the endpoints of $[a, b]$ be quadrature nodes. This gives rise to *Gauss-Radau* resp. *Gauss-Lobatto* formulae. Interestingly, both these formulae allow again a characterization in terms of eigenvalue problems; this was shown by Golub in [23].

2.2.1. Gauss-Radau quadrature. If $\tau_0 = a$ is the prescribed node, the $(n+1)$-point Gauss-Radau formula is

$$\int_a^b f(t)\mathrm{d}\lambda(t) = \lambda_0^a f(a) + \sum_{\nu=1}^{n} \lambda_\nu^a f(\tau_\nu^a) + R_n^a(f), \tag{2.12}$$

where the remainder now vanishes for polynomials of degree $\le 2n$,

$$R_n^a(f) = 0, \quad f \in \mathbb{P}_{2n}. \tag{2.13}$$

ETNA
Kent State University
etna@mcs.kent.edu

Define a *modified Jacobi matrix* of order $n+1$ by

$$\boldsymbol{J}_{n+1}^{R,a}(\mathrm{d}\lambda) = \begin{bmatrix} \boldsymbol{J}_n(\mathrm{d}\lambda) & \sqrt{\beta_n}\boldsymbol{e}_n \\ \sqrt{\beta_n}\boldsymbol{e}_n^T & \alpha_n^R \end{bmatrix}, \tag{2.14}$$

where $\boldsymbol{J}_n(\mathrm{d}\lambda)$ is the same matrix as in (2.8), $\beta_n = \beta_n(\mathrm{d}\lambda)$ as in (2.6),

$$\alpha_n^R = a - \beta_n \frac{\pi_{n-1}(a)}{\pi_n(a)}, \tag{2.15}$$

with $\pi_m(\,\cdot\,) = \pi_m(\,\cdot\,;\mathrm{d}\lambda)$, and $\boldsymbol{e}_n^T = [0, 0, \ldots, 1]$ the nth canonical basis vector of $\mathbb{R}^n$. Then, *the nodes in* (2.12) (including $\tau_0 = a$) *are the eigenvalues of* $\boldsymbol{J}_{n+1}^{R,a}(\mathrm{d}\lambda)$, *and the weights* λ_ν^a, $\nu = 0, 1, \ldots, n$, *are again given by* (2.9) *in terms of the respective normalized eigenvectors.* An analogous result holds for the Gauss-Radau formula with prescribed node $\tau_{n+1} = b$,

$$\int_a^b f(t)\mathrm{d}\lambda(t) = \sum_{\nu=1}^{n} \lambda_\nu^b f(\tau_\nu^b) + \lambda_{n+1}^b f(b) + R_n^b(f). \tag{2.16}$$

The only change is replacing a in (2.15) by b, giving rise to a modified Jacobi matrix $\boldsymbol{J}_{n+1}^{R,b}(\mathrm{d}\lambda)$. Both quadrature sums in (2.12) and (2.16) allow a matrix representation analogous to (2.10), with $\boldsymbol{J}_n$ replaced by $\boldsymbol{J}_{n+1}^{R,a}$ resp. $\boldsymbol{J}_{n+1}^{R,b}$ and the dimension of $\boldsymbol{e}_1$ increased by 1.

The remainders R_n^a, R_n^b of the two Gauss-Radau formulae have the useful property

$$R_n^a(f) > 0, \quad R_n^b(f) < 0 \quad \text{if } \operatorname{sgn} f^{(2n+1)} = 1 \text{ on } [a, b], \tag{2.17}$$

with the inequalities reversed if $\operatorname{sgn} f^{(2n+1)} = -1$. This means that one of the two Gauss-Radau approximations is a lower bound, and the other an upper bound for the exact value of the integral.

It now takes $2n+1$ moments $\mu_0, \mu_1, \ldots, \mu_{2n}$ to obtain $\boldsymbol{J}_{n+1}^{R,a}(\mathrm{d}\lambda)$, $\boldsymbol{J}_{n+1}^{R,b}(\mathrm{d}\lambda)$ and the $(n+1)$-point Gauss-Radau formulae. Chebyshev's algorithm will provide the recursion coefficients needed to generate $\boldsymbol{J}_n(\mathrm{d}\lambda)$ in (2.14), β_n, and the ratio of orthogonal polynomials in (2.15).

The case of a discrete measure $\mathrm{d}\lambda_N$ supported on N points t_k with $a \le t_1 < t_2 < \cdots < t_N \le b$, and having positive jumps w_k^2 at t_k,

$$\int_{\mathbb{R}} f(t)\mathrm{d}\lambda_N(t) := \sum_{k=1}^{N} w_k^2 f(t_k), \tag{2.18}$$

is of some interest in applications. For one thing, the Gauss-Radau formulae (2.12), (2.16) (and, for that matter, the Gauss formula (2.2) as well), provide "compressions" of the sum $S = \sum_{k=1}^{N} w_k^2 f(t_k)$, i.e., approximations of S by a sum with fewer terms if $n < N-1$. When f is a polynomial of degree $\le 2n$, the compressed sums in fact have the same value as the original sum. More importantly, the formula (2.12) with $n < N$ together with the companion formula (2.16) furnish upper and lower bounds of S if $f^{(2n+1)} < 0$ on $[a, b]$. Applications of this will be made in §§5.2–5.4. Chebyshev's algorithm can again be used to generate (2.12) and (2.16) from the moments of $\mathrm{d}\lambda_N$. There is also a numerically stable alternative to Lanczos's algorithm (cf. §5.1), due to Gragg and Harrod [35], generating the Jacobi matrix $\boldsymbol{J}_n(\mathrm{d}\lambda_N)$ directly from the quantities w_k and t_k in (2.18).

ETNA
Kent State University
etna@mcs.kent.edu

2.2.2. Gauss-Lobatto quadrature. Written as an $(n+2)$-point formula, the Gauss-Lobatto quadrature rule is

$$\int_a^b f(t)\mathrm{d}\lambda(t) = \lambda_0 f(a) + \sum_{\nu=1}^{n} \lambda_\nu f(\tau_\nu) + \lambda_{n+1} f(b) + R_n^{a,b}(f), \tag{2.19}$$

and has the exactness property

$$R_n^{a,b}(f) = 0, \quad f \in \mathbb{P}_{2n+1}, \tag{2.20}$$

and the sign property

$$R_n^{a,b}(f) < 0 \quad \text{if } \operatorname{sgn} f^{(2n+2)} = 1 \text{ on } [a,b], \tag{2.21}$$

with the inequality reversed if $\operatorname{sgn} f^{(2n+2)} = -1$. The appropriate modification of the Jacobi matrix is

$$\boldsymbol{J}_{n+2}^L(\mathrm{d}\lambda) = \begin{bmatrix} \boldsymbol{J}_{n+1}(\mathrm{d}\lambda) & \sqrt{\beta_{n+1}^L}\,\boldsymbol{e}_{n+1} \\ \sqrt{\beta_{n+1}^L}\,\boldsymbol{e}_{n+1}^T & \alpha_{n+1}^L \end{bmatrix}, \tag{2.22}$$

with notations similar as in (2.14). Here, α_{n+1}^L, β_{n+1}^L are defined as the solution of the 2×2 linear system

$$\begin{bmatrix} \pi_{n+1}(a) & \pi_n(a) \\ \pi_{n+1}(b) & \pi_n(b) \end{bmatrix} \begin{bmatrix} \alpha_{n+1}^L \\ \beta_{n+1}^L \end{bmatrix} = \begin{bmatrix} a\pi_{n+1}(a) \\ b\pi_{n+1}(b) \end{bmatrix}. \tag{2.23}$$

Then *the nodes of* (2.19) (including $\tau_0 = a$ and $\tau_{n+1} = b$) *are the eigenvalues of* $\boldsymbol{J}_{n+2}^L(\mathrm{d}\lambda)$ *and the weights* λ_ν, $\nu = 0, 1, \ldots, n, n+1$, *once again are given by* (2.9) *in terms of the respective normalized eigenvectors*. Hence, (2.10) again holds, with $\boldsymbol{J}_n$ replaced by $\boldsymbol{J}_{n+2}^L$ and $\boldsymbol{e}_1$ having dimension $n+2$.

2.3. Gauss quadrature with multiple nodes. The Gauss-Radau and Gauss-Lobatto formulae may be generalized by allowing an arbitrary number of prescribed nodes, even of arbitrary multiplicities, outside, or on the boundary, of the support interval of $\mathrm{d}\lambda$. (Those of even multiplicities may also be inside the support interval.) The remaining "free" nodes are either simple or of *odd* multiplicity. The quadrature rule in question, therefore, has the form

$$\int_{\mathbb{R}} f(t)\mathrm{d}\lambda(t) = \sum_{\nu=1}^{n} \sum_{\sigma=0}^{2s_\nu} \lambda_\nu^{(\sigma)} f^{(\sigma)}(\tau_\nu) + \sum_{\mu=1}^{m} \sum_{\rho=0}^{r_\mu - 1} \kappa_\mu^{(\rho)} f^{(\rho)}(u_\mu) + R_{n,m}(f), \tag{2.24}$$

where τ_ν are the free nodes and u_μ the prescribed ones, and the formula is required to have maximum degree of exactness $2\,(n + \sum_\nu s_\nu) + \sum_\mu r_\mu - 1$. This has a long history, going back to Christoffel (all $s_\nu = 0$ and $r_\mu = 1$) and including among its contributors Turán ($m = 0$, $s_\nu = s$ for all ν), Chakalov, Popoviciu, and Stancu (cf. [19, §2.2]).

The prescribed nodes u_μ give rise to the polynomial

$$u(t) = \omega \prod_{\mu=1}^{m} (t - u_\mu)^{r_\mu},$$

ETNA
Kent State University
etna@mcs.kent.edu

where $\omega = \pm 1$ is chosen such that $u(t) \geq 0$ for t on the support of $\mathrm{d}\lambda$. For the formula (2.24) to have maximum algebraic degree of exactness, the free nodes τ_ν ("Gauss nodes") must be chosen to satisfy

$$\int_{\mathbb{R}} \prod_{\nu=1}^{n} (t-\tau_\nu)^{2s_\nu+1} t^k u(t)\mathrm{d}\lambda(t) = 0, \quad k = 0, 1, \ldots, n-1.$$

By far the simplest scenario is the one in which $s_\nu = 0$ for all ν. In this case, τ_ν are the zeros of the polynomial $\pi_n(\,\cdot\,; u\mathrm{d}\lambda)$ of degree n orthogonal with respect to the (positive) measure $u\mathrm{d}\lambda$. This gives rise to the *problem of modification*: given the Jacobi matrix of the measure $\mathrm{d}\lambda$, find the Jacobi matrix of the modified measure $u\mathrm{d}\lambda$. An elegant solution of this problem involves genuine techniques from linear algebra; this will be described in §3. The weights $\lambda_\nu = \lambda_\nu^{(0)}$ are computable similarly as in (2.9) for ordinary Gauss quadrature, namely [27, §6]

$$\lambda_\nu = \mu_0 v_{\nu,1}^2 / u(\tau_\nu), \quad \nu = 1, 2, \ldots, n,$$

where $\mu_0 = \int_{\mathbb{R}} u(t)\mathrm{d}\lambda(t)$ and $v_{\nu,1}$ is the first component of the normalized eigenvector of $\boldsymbol{J}_n(u\mathrm{d}\lambda)$ corresponding to the eigenvalue τ_ν. For the computation of the remaining weights $\kappa_\mu^{(\rho)}$ in (2.24), see [37].

The case of multiple Gauss nodes ($s_\nu > 0$) is a good deal more complicated, requiring the iterative solution of a system of nonlinear equations for the τ_ν and the solution of linear algebraic systems for the weights $\lambda_\nu^{(\sigma)}$, $\kappa_\mu^{(\rho)}$; see, e.g., [27, §5] and [22].

2.4. Gauss-Kronrod quadrature. The quadrature rules discussed so far are products of the 19th century (except for the multiple-node Gauss rules). Let us turn now to a truly 20th-century product — the *Gauss-Kronrod formula*

$$\int_{\mathbb{R}} f(t)\mathrm{d}\lambda(t) = \sum_{\nu=1}^{n} \lambda_\nu^K f(\tau_\nu^G) + \sum_{\mu=1}^{n+1} \lambda_\mu^{*K} f(\tau_\mu^K) + R_n^K(f), \tag{2.25}$$

where τ_ν^G are the nodes of the n-point Gauss formula for $\mathrm{d}\lambda$, and the $n+1$ remaining nodes, called *Kronrod nodes*, as well as all $2n+1$ weights λ_ν^K, λ_μ^{*K} are determined by requiring maximum degree of exactness $3n+1$, i.e.,

$$R_n^K(f) = 0, \quad f \in \mathbb{P}_{3n+1}. \tag{2.26}$$

This was proposed by Kronrod [39] in the 1960s in the special case $\mathrm{d}\lambda(t) = \mathrm{d}t$ on $[-1, 1]$ as an economical way of estimating the error of the n-point Gauss-Legendre quadrature rule. The formula (2.25) nowadays is widely used in automatic and adaptive quadrature routines ([43], [17]).

Remarkably enough, there is an eigenvalue/vector characterization similar to those in §§2.1, 2.2 also for Gauss-Kronrod quadrature rules. This was discovered in 1997 by Laurie [40]. He assumes that there exists a positive Gauss-Kronrod formula (i.e., $\lambda_\nu^K > 0$, $\lambda_\mu^{*K} > 0$, and $\tau_\nu^K \in \mathbb{R}$), which need not be the case in general. (Indeed, the Kronrod nodes and all weights may well be complex.) The modified Jacobi matrix is now a symmetric tridiagonal matrix of order $2n+1$ and has the form

$$\boldsymbol{J}_{2n+1}^K(\mathrm{d}\lambda) = \begin{bmatrix} \boldsymbol{J}_n(\mathrm{d}\lambda) & \sqrt{\beta_n}\boldsymbol{e}_n & \boldsymbol{0} \\ \sqrt{\beta_n}\boldsymbol{e}_n^T & \alpha_n & \sqrt{\beta_{n+1}}\boldsymbol{e}_1^T \\ \boldsymbol{0} & \sqrt{\beta_{n+1}}\boldsymbol{e}_1 & \boldsymbol{J}_n^* \end{bmatrix} \tag{2.27}$$

ETNA
Kent State University
etna@mcs.kent.edu

with notation similar as before and $\boldsymbol{J}_n^*$ a symmetric tridiagonal matrix. The structure of $\boldsymbol{J}_n^*$ differs according as n is even or odd. For definiteness, suppose that n is even. Then

$$\boldsymbol{J}_n^* = \begin{bmatrix} \boldsymbol{J}_{[n+1:3n/2]}(\mathrm{d}\lambda) & \sqrt{\beta_n^*}\boldsymbol{e}_{n/2} \\ \sqrt{\beta_n^*}\boldsymbol{e}_{n/2}^T & \boldsymbol{J}^K_{[(3n+2)/2:2n]} \end{bmatrix} \quad (n \text{ even}), \tag{2.28}$$

where $\boldsymbol{J}_{[p:q]}(\mathrm{d}\lambda)$ denotes the principal minor matrix of $\boldsymbol{J}(\mathrm{d}\lambda)$ that has diagonal elements $\alpha_p, \alpha_{p+1}, \ldots, \alpha_q$, and similarly for $\boldsymbol{J}^K_{[p:q]}$. Thus, the upper left square block of $\boldsymbol{J}_n^*$ (of order $n/2$) may be assumed known, and the rest, including the constant β_n^*, is to be determined. Laurie devised an algorithm that determines the unknown elements of $\boldsymbol{J}_n^*$ in such a way that *the Gauss nodes τ_ν^G and Kronrod nodes τ_μ^K are the eigenvalues of $\boldsymbol{J}^K_{2n+1}(\mathrm{d}\lambda)$ and the weights are given by*

$$\begin{aligned} \lambda_\nu^K &= \mu_0[\boldsymbol{u}^K_{\nu,1}]^2, \quad \nu = 1, \ldots, n; \\ \lambda_\mu^{*K} &= \mu_0[\boldsymbol{u}^K_{n+\mu,1}]^2, \quad \mu = 1, \ldots, n, n+1, \end{aligned} \tag{2.29}$$

where $\boldsymbol{u}_1^K, \boldsymbol{u}_2^K, \ldots, \boldsymbol{u}_{2n+1}^K$ are the normalized eigenvectors of $\boldsymbol{J}^K_{2n+1}(\mathrm{d}\lambda)$ corresponding to the eigenvalues $\tau_1^G, \ldots, \tau_n^G$; $\tau_1^K, \ldots, \tau_{n+1}^K$, and $\boldsymbol{u}^K_{1,1}, \boldsymbol{u}^K_{2,1}, \ldots, \boldsymbol{u}^K_{2n+1,1}$ their first components. Moreover, $\boldsymbol{J}_n^*$ *in* (2.27) *has the same eigenvalues as* $\boldsymbol{J}_n(\mathrm{d}\lambda)$, i.e., the Gauss nodes $\tau_1^G, \ldots, \tau_n^G$.

If the Gauss nodes τ_ν^G are already known, as is often the case, there is some redundancy in Laurie's algorithm, inasmuch as it regenerates them all. In a joint paper with Calvetti, Gragg, and Reichel [8], Golub removes this redundancy by focusing directly on the Kronrod nodes. The basic idea is to observe that the trailing matrix $\boldsymbol{J}_n^*$ in (2.27) as well as the leading matrix $\boldsymbol{J}_{[n+1:3n/2]}(\mathrm{d}\lambda)$ in (2.28) (again with n assumed even) have their own sets of orthogonal polynomials and respective Gauss quadrature rules, the measures of which, however, are unknown. Since the eigenvalues of $\boldsymbol{J}_n^*$ are the same as those of $\boldsymbol{J}_n(\mathrm{d}\lambda)$, the former Gauss rule has nodes τ_ν^G, $\nu = 1, 2, \ldots, n$, and positive weights λ_ν^*, say, while the latter has certain nodes $\tilde{\tau}_\kappa$ and weights $\tilde{\lambda}_\kappa$, $\kappa = 1, 2, \ldots, n/2$. Let the matrices of normalized eigenvectors of $\boldsymbol{J}_n(\mathrm{d}\lambda)$ and $\boldsymbol{J}_n^*$ be $\boldsymbol{v} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n]$ and $\boldsymbol{v}^* = [\boldsymbol{v}_1^*, \boldsymbol{v}_2^*, \ldots, \boldsymbol{v}_n^*]$, respectively. The new algorithm will make use of the *last* components $\boldsymbol{v}_{1,n}, \boldsymbol{v}_{2,n}, \ldots, \boldsymbol{v}_{n,n}$ of the eigenvectors in $\boldsymbol{v}$ (assumed known) and the *first* components $\boldsymbol{v}_{1,1}^*, \boldsymbol{v}_{2,1}^*, \ldots, \boldsymbol{v}_{n,1}^*$ of those in $\boldsymbol{v}^*$. The latter, according to (2.9), are related to the Gauss weights λ_ν^* through

$$[\boldsymbol{v}_{\nu,1}^*]^2 = \lambda_\nu^*, \quad \nu = 1, 2, \ldots, n,$$

where the underlying measure is assumed normalized to have total mass 1, and one computes

$$\lambda_\nu^* = \sum_{\kappa=1}^{n/2} \ell_\nu(\tilde{\tau}_\kappa)\tilde{\lambda}_\kappa, \quad \nu = 1, 2, \ldots, n,$$

in terms of the second Gauss rule (for $\boldsymbol{J}_{[n+1:3n/2]}(\mathrm{d}\lambda)$) and the elementary Lagrange interpolation polynomials ℓ_ν associated with the nodes $\tau_1^G, \tau_2^G, \ldots, \tau_n^G$. Therefore,

$$\boldsymbol{v}_{\nu,1}^* = \sqrt{\lambda_\nu^*}, \quad \nu = 1, 2, \ldots, n. \tag{2.30}$$

Now let

$$\boldsymbol{J}_n^* = \boldsymbol{v}^* \boldsymbol{D} \boldsymbol{v}^{*T}, \quad \boldsymbol{D} = \operatorname{diag}(\tau_1^G, \tau_2^G, \ldots, \tau_n^G) \tag{2.31}$$

ETNA
Kent State University
etna@mcs.kent.edu

be the spectral decomposition of $\boldsymbol{J}_n^*$, and define

$$\boldsymbol{V} = \begin{bmatrix} v & 0 & 0 \\ 0^T & 1 & 0^T \\ 0 & 0 & v^* \end{bmatrix}, \tag{2.32}$$

a matrix of order $2n+1$. From (2.27), one gets

$$\boldsymbol{V}^T(\boldsymbol{J}_{2n+1}^K(\mathrm{d}\lambda) - \lambda \boldsymbol{I})\boldsymbol{V} = \begin{bmatrix} \boldsymbol{D} - \lambda \boldsymbol{I} & \sqrt{\beta_n}\boldsymbol{v}^T\boldsymbol{e}_n & 0 \\ \sqrt{\beta_n}\boldsymbol{e}_n^T\boldsymbol{v} & \alpha_n - \lambda & \sqrt{\beta_{n+1}}\boldsymbol{e}_1^T\boldsymbol{v}^* \\ 0 & \sqrt{\beta_{n+1}}\boldsymbol{v}^{*T}\boldsymbol{e}_1 & \boldsymbol{D} - \lambda \boldsymbol{I} \end{bmatrix},$$

where the matrix on the right is a diagonal matrix plus a Swiss cross containing the known elements $\boldsymbol{e}_n^T\boldsymbol{v}$ and the elements $\boldsymbol{e}_1^T\boldsymbol{v}^*$ that were computed in (2.30). A further (cosmetic) orthogonal similarity transformation involving a permutation and a sequence of Givens rotations can be applied to yield

$$\tilde{\boldsymbol{V}}^T(\boldsymbol{J}_{2n+1}^K(\mathrm{d}\lambda) - \lambda \boldsymbol{I})\tilde{\boldsymbol{V}} = \left[\begin{array}{c|cc} \boldsymbol{D} - \lambda \boldsymbol{I} & 0 & 0 \\ \hline 0 & \boldsymbol{D} - \lambda \boldsymbol{I} & \boldsymbol{c} \\ 0^T & \boldsymbol{c}^T & \alpha_n - \lambda \end{array}\right], \tag{2.33}$$

where $\tilde{\boldsymbol{V}}$ is the transformed matrix $\boldsymbol{V}$ and $\boldsymbol{c}$ a vector containing the entries in positions $n+1$ to $2n$ of the transformed vector $[\sqrt{\beta_n}\boldsymbol{e}_n^T\boldsymbol{v}, \sqrt{\beta_{n+1}}\boldsymbol{e}_1^T\boldsymbol{v}^*, \alpha_n]$. Eq. (2.33) now reveals that one set of eigenvalues of $\boldsymbol{J}_{2n+1}^K(\mathrm{d}\lambda)$ is $\{\tau_1^G, \tau_2^G, \ldots, \tau_n^G\}$, while the remaining eigenvalues are those of the trailing block in (2.33). From

$$\begin{bmatrix} \boldsymbol{D} - \lambda \boldsymbol{I} & \boldsymbol{c} \\ \boldsymbol{c}^T & \alpha_n - \lambda \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & 0 \\ \boldsymbol{c}^T(\boldsymbol{D} - \lambda \boldsymbol{I})^{-1} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{D} - \lambda \boldsymbol{I} & \boldsymbol{c} \\ 0^T & -f(\lambda) \end{bmatrix},$$

where

$$f(\lambda) = \lambda - \alpha_n + \sum_{\nu=1}^{n} \frac{c_\nu^2}{\tau_\nu^G - \lambda}, \quad \boldsymbol{c}^T = [c_1, c_2, \ldots, c_n], \tag{2.34}$$

it follows that the remaining eigenvalues, i.e., the Kronrod nodes, are the zeros of $f(\lambda)$. It is evident from (2.34) that they interlace with the Gauss nodes τ_ν^G. The normalized eigenvectors $\boldsymbol{u}_1^K, \boldsymbol{u}_2^K, \ldots, \boldsymbol{u}_{2n+1}^K$ required to compute the weights λ_ν^K, λ_μ^{*K} via (2.29) can be computed from the columns of $\boldsymbol{V}$ by keeping track of the orthogonal transformations.

3. Orthogonal polynomials. The connection between orthogonal polynomials and Jacobi matrices (cf. §2.1) gives rise to several interesting problems:

(a) Given the Jacobi matrix for the measure $\mathrm{d}\lambda$, find the Jacobi matrix for the modified measure $\mathrm{d}\lambda_{\mathrm{mod}} = r\mathrm{d}\lambda$, where r is either a polynomial or a rational function.

(b) Given the Jacobi matrices for two measures $\mathrm{d}\lambda_1$ and $\mathrm{d}\lambda_2$, find the Jacobi matrix for $\mathrm{d}\lambda = \mathrm{d}\lambda_1 + \mathrm{d}\lambda_2$.

ETNA
Kent State University
etna@mcs.kent.edu

(c) Let $[c_j, d_j]$ be a finite set of intervals, disjoint or not, and $d\lambda_j$ a positive measure on $[c_j, d_j]$. Let $d\lambda(t) = \sum_j \chi_{[c_j,d_l]}(t) d\lambda_j(t)$, where $\chi_{[c_j,d_j]}$ is the characteristic function of the interval $[c_j, d_j]$,

$$\chi_{[c_j,d_j]}(t) = \begin{cases} 1 \text{ if } t \in [c_j, d_j], \\ 0 \text{ otherwise.} \end{cases}$$

Knowing the Jacobi matrices $\boldsymbol{J}^{(j)}$ for $d\lambda_j$, find the Jacobi matrix for $d\lambda$.

The problem in (b) is discussed in [13], where three algorithms are developed for its solution. We do not attempt to describe them here, since they are rather technical and not easily summarized. Suffice it to say that linear algebra figures prominently in all three of these algorithms. A special case of Problem (a) — modification of the measure by a polynomial factor — a problem discussed in [27] and [38], is considered in §§3.2, 3.3. It is related to a classical theorem of Christoffel (cf., e.g., [19, p. 85]), which expresses the orthogonal polynomials for the modified measure in determinantal form in terms of the orthogonal polynomials of the original measure. For algorithmic and computational purposes, however, the use of Jacobi matrices is vastly superior. The case of rational r, in particular $r(t) = (t - x)^{-1}$ with real x outside the support of $d\lambda$, and $r(t) = [(t - x)^2 + y^2]^{-1}$, $y > 0$, is treated in [15], where algorithms are developed that are similar to those in [20] but are derived in a different manner. Problem (c) is dealt with in §3.4. Note that Problem (b) is a special case of Problem (c).

We begin with an integral representation of Jacobi matrices and then in turn describe modification of the measure by a linear, quadratic, and higher-degree polynomial, and solution procedures for Problem (c).

3.1. Integral representation of the Jacobi matrix. Let $\tilde{\pi}_0, \tilde{\pi}_1, \tilde{\pi}_2, \ldots$ be the system of *orthonormal* polynomials with respect to to the measure $d\lambda$, that is,

$$\tilde{\pi}_k(t) = \frac{\pi_k(t)}{\|\pi_k\|}, \quad \|\pi_k\|^2 = \int_{\mathbb{R}} \pi_k^2(t) d\lambda, \tag{3.1}$$

with π_k as in (2.5), (2.6). They satisfy the recurrence relation

$$\begin{gathered} \sqrt{\beta_{k+1}}\tilde{\pi}_{k+1}(t) = (t - \alpha_k)\tilde{\pi}_k(t) - \sqrt{\beta_k}\tilde{\pi}_{k-1}(t), \quad k = 0, 1, 2, \ldots, \\ \tilde{\pi}_{-1}(t) = 0, \quad \tilde{\pi}_0(t) = 1/\sqrt{\beta_0}, \end{gathered} \tag{3.2}$$

with recursion coefficients α_k, β_k as in (2.6) and $\beta_0 = \int_{\mathbb{R}} d\lambda(t)$ $(= \mu_0)$. From (3.2) and the orthonormality of the polynomials $\tilde{\pi}_k$ one easily checks that

$$\int_{\mathbb{R}} t\tilde{\pi}_k(t)\tilde{\pi}_\ell(t) d\lambda(t) = \begin{cases} 0 & \text{if } |k - \ell| > 1, \\ \sqrt{\beta_{k+1}} & \text{if } |k - \ell| = 1, \\ \alpha_k & \text{if } k = \ell. \end{cases} \tag{3.3}$$

This allows us to represent the Jacobi matrix $\boldsymbol{J} = \boldsymbol{J}_n(d\lambda)$ of order n (cf. (2.8)) in integral form as

$$\boldsymbol{J} = \int_{\mathbb{R}} t\boldsymbol{p}(t)\boldsymbol{p}^T(t) d\lambda(t), \tag{3.4}$$

where

$$\boldsymbol{p}^T(t) = [\tilde{\pi}_0(t), \tilde{\pi}_1(t), \ldots, \tilde{\pi}_{n-1}(t)]. \tag{3.5}$$

ETNA
Kent State University
etna@mcs.kent.edu

Orthogonality, on the other hand, is expressible as

$$\int_{\mathbb{R}} \boldsymbol{p}(t)\boldsymbol{p}^T(t)\mathrm{d}\lambda(t) = \boldsymbol{I}, \tag{3.6}$$

where $\boldsymbol{I} = \boldsymbol{I}_n$ is the unit matrix of order n, and the first n recurrence relations in (3.2) can be given the form

$$t\boldsymbol{p}(t) = \boldsymbol{J}\boldsymbol{p}(t) + \sqrt{\beta_n}\tilde{\pi}_n(t)\boldsymbol{e}_n, \tag{3.7}$$

where $\boldsymbol{e}_n = [0, 0, \ldots, 1]^T \in \mathbb{R}^n$.

3.2. Modification by a linear factor. The problem to be studied is the effect on the Jacobi matrix $\boldsymbol{J}$ of modifying the (positive) measure $\mathrm{d}\lambda$ into a measure $\mathrm{d}\lambda_{\mathrm{mod}}$ defined by

$$\mathrm{d}\lambda_{\mathrm{mod}}(t) = \omega(t-c)\mathrm{d}\lambda(t), \tag{3.8}$$

where c is a real constant outside, or on the boundary, of the support interval of $\mathrm{d}\lambda$ and $\omega = \pm 1$ chosen such that the measure $\mathrm{d}\lambda_{\mathrm{mod}}$ is again positive. A solution of this problem has been given already by Galant [16] and was taken up again, and simplified, in [27].

The symmetric matrix $\omega(\boldsymbol{J} - c\boldsymbol{I})$ is positive definite since by the assumtions made regarding (3.8) all its eigenvalues are positive. It thus admits a Cholesky decomposition

$$\omega(\boldsymbol{J} - c\boldsymbol{I}) = \boldsymbol{L}\boldsymbol{L}^T, \tag{3.9}$$

where $\boldsymbol{L}$ is lower triangular and bidiagonal. By (3.4), (3.6), and (3.8) one has

$$\omega(\boldsymbol{J} - c\boldsymbol{I}) = \omega\int_{\mathbb{R}} (t-c)\boldsymbol{p}(t)\boldsymbol{p}^T(t)\mathrm{d}\lambda(t) = \int_{\mathbb{R}} \boldsymbol{p}(t)\boldsymbol{p}^T(t)\mathrm{d}\lambda_{\mathrm{mod}}(t).$$

This may be written as

$$\omega(\boldsymbol{J} - c\boldsymbol{I}) = \boldsymbol{L}\int_{\mathbb{R}} \boldsymbol{L}^{-1}\boldsymbol{p}(t)\boldsymbol{p}^T(t)\boldsymbol{L}^{-T}\mathrm{d}\lambda_{\mathrm{mod}}(t)\boldsymbol{L}^T,$$

which, since $\boldsymbol{L}^{-1}\omega(\boldsymbol{J} - c\boldsymbol{I})\boldsymbol{L}^{-T} = \boldsymbol{I}$ by (3.9), implies

$$\int_{\mathbb{R}} \boldsymbol{p}_{\mathrm{mod}}(t)\boldsymbol{p}_{\mathrm{mod}}^T(t)\mathrm{d}\lambda_{\mathrm{mod}}(t) = \boldsymbol{I}, \tag{3.10}$$

where

$$\boldsymbol{p}_{\mathrm{mod}}(t) = \boldsymbol{L}^{-1}\boldsymbol{p}(t). \tag{3.11}$$

This means that $\boldsymbol{p}_{\mathrm{mod}}$ are the orthonormal polynomials with respect to the measure $\mathrm{d}\lambda_{\mathrm{mod}}$. What is the corresponding Jacobi matrix $\boldsymbol{J}_{\mathrm{mod}}$?

First observe that from (3.7) one has

$$(t-c)\boldsymbol{p}(t) = (\boldsymbol{J} - c\boldsymbol{I})\boldsymbol{p}(t) + \sqrt{\beta_n}\tilde{\pi}_n(t)\boldsymbol{e}_n. \tag{3.12}$$

Using the analogues of (3.4), (3.6), one has, by (3.11) and (3.8),

$$\begin{aligned} \boldsymbol{J}_{\mathrm{mod}} - c\boldsymbol{I} &= \int_{\mathbb{R}} (t-c)\boldsymbol{p}_{\mathrm{mod}}(t)\boldsymbol{p}_{\mathrm{mod}}^T(t)\mathrm{d}\lambda_{\mathrm{mod}}(t) \\ &= \omega\boldsymbol{L}^{-1}\int_{\mathbb{R}} (t-c)^2\boldsymbol{p}(t)\boldsymbol{p}^T(t)\mathrm{d}\lambda(t)\boldsymbol{L}^{-T}. \end{aligned} \tag{3.13}$$

ETNA
Kent State University
etna@mcs.kent.edu

Multiplying (3.12) with its transpose, one gets

$$(t-c)^2\boldsymbol{p}(t)\boldsymbol{p}^T(t) = [(\boldsymbol{J}-c\boldsymbol{I})\boldsymbol{p}(t) + \sqrt{\beta_n}\tilde{\pi}_n(t)\boldsymbol{e}_n][\boldsymbol{p}^T(t)(\boldsymbol{J}-c\boldsymbol{I}) + \sqrt{\beta_n}\tilde{\pi}_n(t)\boldsymbol{e}_n^T],$$

and observing that by (3.6)

$$\int_{\mathbb{R}} (\boldsymbol{J}-c\boldsymbol{I})\boldsymbol{p}(t)\boldsymbol{p}^T(t)(\boldsymbol{J}-c\boldsymbol{I})\mathrm{d}\lambda(t) = (\boldsymbol{J}-c\boldsymbol{I})^2$$

and by orthonormality

$$\int_{\mathbb{R}} (\boldsymbol{J}-c\boldsymbol{I})\boldsymbol{p}(t)\tilde{\pi}_n(t)\boldsymbol{e}_n^T\mathrm{d}\lambda(t) = 0, \quad \int_{\mathbb{R}} \tilde{\pi}_n^2(t)\mathrm{d}\lambda(t) = 1,$$

one finds

$$\int_{\mathbb{R}} (t-c)^2\boldsymbol{p}(t)\boldsymbol{p}^T(t)\mathrm{d}\lambda(t) = (\boldsymbol{J}-c\boldsymbol{I})^2 + \beta_n\boldsymbol{e}_n\boldsymbol{e}_n^T.$$

Thus, by (3.13),

$$\boldsymbol{J}_{\mathrm{mod}} - c\boldsymbol{I} = \omega\boldsymbol{L}^{-1}\left((\boldsymbol{J}-c\boldsymbol{I})^2 + \beta_n\boldsymbol{e}_n\boldsymbol{e}_n^T\right)\boldsymbol{L}^{-T}.$$

Substituting from (3.9) yields for the desired Jacobi matrix

$$\boldsymbol{J}_{\mathrm{mod}} = \frac{1}{\omega}\boldsymbol{L}^T\boldsymbol{L} + c\boldsymbol{I} + \omega\beta_n\boldsymbol{L}^{-1}\boldsymbol{e}_n\boldsymbol{e}_n^T\boldsymbol{L}^{-T}$$

and noting that $\boldsymbol{L}^{-1}\boldsymbol{e}_n = \boldsymbol{e}_n/(\boldsymbol{e}_n^T\boldsymbol{L}\boldsymbol{e}_n)$ finally

$$\boldsymbol{J}_{\mathrm{mod}} = \frac{1}{\omega}\boldsymbol{L}^T\boldsymbol{L} + c\boldsymbol{I} + \gamma\boldsymbol{e}_n\boldsymbol{e}_n^T, \quad \gamma = \omega\beta_n/(\boldsymbol{e}_n^T\boldsymbol{L}\boldsymbol{e}_n)^2. \tag{3.14}$$

Equations (3.9) and (3.14) allow the following interpretation: *The matrix* $\boldsymbol{J}_1 := \boldsymbol{J}_{\mathrm{mod}} - \gamma\boldsymbol{e}_n\boldsymbol{e}_n^T$ *is the result of one step of the symmetric LR algorithm with shift* c,

$$\boldsymbol{J} - c\boldsymbol{I} = \frac{1}{\omega}\boldsymbol{L}\boldsymbol{L}^T, \quad \boldsymbol{J}_1 = \frac{1}{\omega}\boldsymbol{L}^T\boldsymbol{L} + c\boldsymbol{I}. \tag{3.15}$$

Note that $\boldsymbol{J}_1$ differs from $\boldsymbol{J}_{\mathrm{mod}}$ only by one element in the lower right-hand corner. We could get rid of it by deleting the last row and last column of $\boldsymbol{J}_1$. This would yield the desired Jacobi matrix of order $n-1$. If we are interested in the Jacobi matrix $\boldsymbol{J}_{n,\mathrm{mod}}$ of order n, we can *apply the symmetric LR algorithm* (with shift c) *to* $\boldsymbol{J}_{n+1}(\mathrm{d}\lambda)$ *and then obtain* $\boldsymbol{J}_{n,\mathrm{mod}}$ *by discarding the last row and last column in the resulting matrix.*

3.3. Modification by a quadratic and higher-degree factor. Modification by a quadratic factor $(t-c_1)(t-c_2)$ essentially amounts to two applications of (3.15),

$$\boldsymbol{J} - c_1\boldsymbol{I} = \frac{1}{\omega_1}\boldsymbol{L}_1\boldsymbol{L}_1^T, \quad \boldsymbol{J}_1 = \frac{1}{\omega_1}\boldsymbol{L}_1^T\boldsymbol{L}_1 + c_1\boldsymbol{I} \tag{3.16}$$

followed by

$$\boldsymbol{J}_1 - c_2\boldsymbol{I} = \frac{1}{\omega_2}\boldsymbol{L}_2\boldsymbol{L}_2^T, \quad \boldsymbol{J}_2 = \frac{1}{\omega_2}\boldsymbol{L}_2^T\boldsymbol{L}_2 + c_2\boldsymbol{I}. \tag{3.17}$$

ETNA
Kent State University
etna@mcs.kent.edu

From the second and third of these equations one gets

$$\boldsymbol{L}_2\boldsymbol{L}_2^T = \omega_2\left(\frac{1}{\omega_1}\boldsymbol{L}_1^T\boldsymbol{L}_1 + (c_1 - c_2)\boldsymbol{I}\right).$$

In particular, if $c_1 = c_2 = c$ and $\omega_2/\omega_1 = 1$, then

$$\boldsymbol{L}_2\boldsymbol{L}_2^T = \boldsymbol{L}_1^T\boldsymbol{L}_1. \tag{3.18}$$

Let

$$\boldsymbol{Q} = \boldsymbol{L}_1^{-T}\boldsymbol{L}_2, \quad \boldsymbol{R} = \boldsymbol{L}_2^T\boldsymbol{L}_1^T. \tag{3.19}$$

Then, using (3.18), one computes

$$\begin{aligned}\boldsymbol{Q}^T\boldsymbol{Q} &= \boldsymbol{L}_2^T\boldsymbol{L}_1^{-1}\boldsymbol{L}_1^{-T}\boldsymbol{L}_2 = \boldsymbol{L}_2^T(\boldsymbol{L}_1^T\boldsymbol{L}_1)^{-1}\boldsymbol{L}_2\\ &= \boldsymbol{L}_2^T(\boldsymbol{L}_2\boldsymbol{L}_2^T)^{-1}\boldsymbol{L}_2 = \boldsymbol{L}_2^T\boldsymbol{L}_2^{-T}\boldsymbol{L}_2^{-1}\boldsymbol{L}_2 = \boldsymbol{I},\end{aligned}$$

so that $\boldsymbol{Q}$ is orthogonal. As a product of two upper triangular matrices, $\boldsymbol{R}$ is upper triangular. Since, again by (3.18), $\boldsymbol{Q}\boldsymbol{R} = \boldsymbol{L}_1^{-T}\boldsymbol{L}_2\boldsymbol{L}_2^T\boldsymbol{L}_1^T = \boldsymbol{L}_1^{-T}\boldsymbol{L}_1^T\boldsymbol{L}_1\boldsymbol{L}_1^T = \boldsymbol{L}_1\boldsymbol{L}_1^T$, the first equation of (3.16) can be written as

$$\boldsymbol{J} - c\boldsymbol{I} = \boldsymbol{Q}\boldsymbol{R} \tag{3.20}$$

and the second of (3.17) similarly as

$$\boldsymbol{J}_2 = \boldsymbol{R}\boldsymbol{Q} + c\boldsymbol{I}. \tag{3.21}$$

Thus, $\boldsymbol{J}_2$ is obtained by one step of the QR algorithm with shift c. It is now clear how the modification

$$\mathrm{d}\lambda_{\mathrm{mod}}(t) = (t-c)^2\mathrm{d}\lambda(t) \tag{3.22}$$

is to be handled: *apply one step of the QR algorithm with shift c to the Jacobi matrix* $\boldsymbol{J}_{n+2}(\mathrm{d}\lambda)$ *of order* $n+2$ *and discard the last two rows and columns of the resulting matrix to obtain* $\boldsymbol{J}_{n,\mathrm{mod}}$.

More generally, a modification $\mathrm{d}\lambda_{\mathrm{mod}}(t) = (t-c)^{2m}\mathrm{d}\lambda(t)$ with an even power can be handled by m steps of the QR algorithm with shift c, discarding the appropriate number of rows and columns, and a modification $\mathrm{d}\lambda_{\mathrm{mod}}(t) = (t-c)^{2m+1}\mathrm{d}\lambda(t)$ by an odd power by means of m shifted QR steps followed by one step of the symmetric LR algorithm as in §3.2. In this way it is possible to accomplish the modification $\mathrm{d}\lambda_{\mathrm{mod}}(t) = r(t)\mathrm{d}\lambda(t)$ for any polynomial r with real roots and $r(t) \geq 0$ for t on the support of $\mathrm{d}\lambda$. Alternative methods, not necessarily requiring knowledge of the roots, are developed in [38].

3.4. Polynomials orthogonal on several intervals. Here we describe two solution procedures for Problem (c) based respectively on Stieltjes's procedure (cf. [21, §2.1]) and modified moments.

3.4.1. Solution by Stieltjes's procedure. Suppose we are interested in generating the Jacobi matrix $\boldsymbol{J} = \boldsymbol{J}_n(\mathrm{d}\lambda)$ of order n for the measure $\mathrm{d}\lambda(t) = \sum_j \chi_{[c_j,d_j]}(t)\,\mathrm{d}\lambda_j(t)$. It is well known that the recursion coefficients $\alpha_k = \alpha_k(\mathrm{d}\lambda)$, $\beta_k = \beta_k(\mathrm{d}\lambda)$ satisfy

$$\alpha_k = \frac{(t\pi_k, \pi_k)_{\mathrm{d}\lambda}}{(\pi_k, \pi_k)_{\mathrm{d}\lambda}}, \quad k = 0, 1, \ldots, n-1, \tag{3.23}$$

ETNA
Kent State University
etna@mcs.kent.edu

$$\beta_k = \frac{(\pi_k, \pi_k)_{d\lambda}}{(\pi_{k-1}, \pi_{k-1})_{d\lambda}}, \quad k = 1, 2, \ldots, n-1, \tag{3.24}$$

where

$$(u, v)_{d\lambda} = \int_{\mathbb{R}} u(t)v(t)d\lambda(t) \tag{3.25}$$

is the inner product associated with $d\lambda$. We also recall the basic recurrence relation (cf. (2.6))

$$\begin{gathered}\pi_{k+1}(t) = (t - \alpha_k)\pi_k(t) - \beta_k\pi_{k-1}(t), \quad k = 0, 1, \ldots, n-1, \\ \pi_{-1}(t) = 0, \quad \pi_0(t) = 1,\end{gathered} \tag{3.26}$$

satisfied by the (monic) orthogonal polynomials $\pi_k(\cdot) = \pi_k(\cdot; d\lambda)$. For convenience, we let, as before,

$$\beta_0 = \int_{\mathbb{R}} d\lambda(t). \tag{3.27}$$

Stieltjes's procedure consists in the following: Compute α_0 from (3.23) with $k = 0$ and β_0 from (3.27). Then use (3.26) with $k = 0$ to generate π_1. Go back to Eqs. (3.23), (3.24) and use them for $k = 1$ to obtain α_1, β_1. Then (3.26) is reapplied with $k = 1$ to get π_2, etc. This procedure, alternating between (3.23), (3.24) and (3.26), is continued until α_{n-1}, β_{n-1} are obtained.

The principal issue in this procedure is the computation of the inner products in (3.23), (3.24). Since they require integrating polynomials of degrees at most $2n-1$, one can use n-point Gauss quadrature

$$\int_{c_j}^{d_j} p(t)d\lambda_j(t) = \sum_{\nu=1}^{n} \lambda_\nu^{(j)} p(\tau_\nu^{(j)}), \quad p \in \mathbb{P}_{2n-1}, \tag{3.28}$$

for the measure $d\lambda_j$ on each constituent interval $[c_j, d_j]$ of $d\lambda$. It has been observed in [14] that the explicit calculation of the Gauss nodes and weights is not required, but only matrix manipulations involving the Jacobi matrices $\boldsymbol{J}^{(j)}$ (of order n) for $d\lambda_j$ (cf. (2.10)).

We illustrate this for the inner product

$$(t\pi_k, \pi_k)_{d\lambda} = \int_{\mathbb{R}} t\pi_k^2(t)d\lambda(t) = \sum_j \int_{c_j}^{d_j} t\pi_k^2(t)d\lambda_j(t). \tag{3.29}$$

Denote $\beta_0^{(j)} = \int_{c_j}^{d_j} d\lambda_j(t)$ and let

$$\boldsymbol{\zeta}_k^{(j)} = \pi_k(\boldsymbol{J}^{(j)})\boldsymbol{e}_1, \quad \boldsymbol{e}_1^T = [1, 0, \ldots, 0] \in \mathbb{R}^n. \tag{3.30}$$

Then, using (3.29), (3.28), and (2.10), one has

$$\begin{aligned}(t\pi_k, \pi_k)_{d\lambda} &= \sum_j \lambda_\nu^{(j)} \tau_\nu^{(j)} \pi_k^2(\tau_\nu^{(j)}) \\ &= \sum_j \beta_0^{(j)} \boldsymbol{e}_1^T \boldsymbol{J}^{(j)} [\pi_k(\boldsymbol{J}^{(j)})]^2 \boldsymbol{e}_1 = \sum_j \beta_0^{(j)} \boldsymbol{e}_1^T \pi_k(\boldsymbol{J}^{(j)}) \boldsymbol{J}^{(j)} \pi_k(\boldsymbol{J}^{(j)}) \boldsymbol{e}_1,\end{aligned}$$

ETNA
Kent State University
etna@mcs.kent.edu

that is,

$$(t\pi_k, \pi_k)_{\mathrm{d}\lambda} = \sum_j \beta_0^{(j)} \boldsymbol{\zeta}_k^{(j)T} \boldsymbol{J}^{(j)} \boldsymbol{\zeta}_k^{(j)}. \tag{3.31}$$

Similarly (in fact a bit simpler), one finds

$$(\pi_k, \pi_k)_{\mathrm{d}\lambda} = \sum_j \beta_0^{(j)} \boldsymbol{\zeta}_k^{(j)T} \boldsymbol{\zeta}_k^{(j)}. \tag{3.32}$$

The updating of the $\boldsymbol{\zeta}_k^{(j)}$ required in Stieltjes's procedure follows immediately from (3.26),

$$\boldsymbol{\zeta}_{k+1}^{(j)} = (\boldsymbol{J}^{(j)} - \alpha_k \boldsymbol{I})\boldsymbol{\zeta}_k^{(j)} - \beta_k \boldsymbol{\zeta}_{k-1}^{(j)}, \tag{3.33}$$

where $\boldsymbol{I}$ is the unit matrix of order n and $\boldsymbol{\zeta}_{-1}^{(j)} = \mathbf{0}$.

3.4.2. Solution by the modified Chebyshev algorithm. The desired recursion coefficients $\alpha_k(\mathrm{d}\lambda)$, $\beta_k(\mathrm{d}\lambda)$, $k = 0, 1, \ldots, n-1$, can also be produced from the first $2n$ *modified moments*

$$m_k = \int_{\mathbb{R}} p_k(t)\mathrm{d}\lambda(t), \quad k = 0, 1, \ldots, 2n-1, \tag{3.34}$$

where $\{p_k\}$ is a system of polynomials satisfying a three-term recurrence relation

$$\begin{gathered} p_{k+1}(t) = (t - a_k)p_k(t) - b_k p_{k-1}(t), \quad k = 0, 1, \ldots, n-1, \\ p_{-1}(t) = 0, \quad p_0(t) = 1 \end{gathered} \tag{3.35}$$

with known coefficients a_k, b_k. A procedure accomplishing this is the *modified Chebyshev algorithm* (cf. [21, §2.4]); this works also if the polynomials $\{p_k\}$ satisfy an extended recurrence relation $p_{k+1}(t) = tp_k(t) - \sum_{j=0}^{k} c_{kj} p_j(t)$, and even if the measure $\mathrm{d}\lambda$ is indefinite (see, e.g., [26]). The computation of the modified moments (3.34) by Gauss quadrature is entirely analogous to the computation of inner products in §3.4.1. Letting now

$$\boldsymbol{z}_k^{(j)} = p_k(\boldsymbol{J}^{(j)})\boldsymbol{e}_1, \tag{3.36}$$

one finds

$$m_k = \sum_j \beta_0^{(j)} \boldsymbol{z}_k^{(j)T} \boldsymbol{e}_1. \tag{3.37}$$

Updating the vectors $\boldsymbol{z}_k^{(j)}$ can again be done via the recurrence relation (3.35),

$$\boldsymbol{z}_{k+1}^{(j)} = (\boldsymbol{J}^{(j)} - a_k \boldsymbol{I})\boldsymbol{z}_k^{(j)} - b_k \boldsymbol{z}_{k-1}^{(j)}, \quad \boldsymbol{z}_{-1}^{(j)} = \mathbf{0}. \tag{3.38}$$

There is yet a third algorithm proposed in [14], which is based on a fast Cholesky decomposition. For this, we refer to the original source.

We remark that the modified Chebyshev algorithm provides an alternative way of solving Problem (a) for polynomial modifications $\mathrm{d}\lambda_{\mathrm{mod}}(t) = r(t)\mathrm{d}\lambda(t)$ (cf. [19, p. 123], [15]). Indeed, if $r \in \mathbb{P}_m$, then r can be expressed in terms of the polynomials p_k as

$$r(t) = \sum_{j=0}^{m} c_j p_j(t). \tag{3.39}$$

ETNA
Kent State University
etna@mcs.kent.edu

If one assumes that $\{p_k\}$ are orthogonal relative to the measure $\mathrm{d}\lambda$, then the modified moments $m_k = \int_{\mathbb{R}} p_k(t)\mathrm{d}\lambda_{\mathrm{mod}}(t)$ are simply

$$m_k = \begin{cases} c_k \int_{\mathbb{R}} p_k^2(t)\mathrm{d}\lambda(t) \text{ if } k \le m, \\ 0 \text{ if } k > m. \end{cases} \tag{3.40}$$

The modified Chebyshev algorithm, if $m < k$, in fact simplifies, owing to the $k-m+1$ zero modified moments in (3.40).

4. The least squares problem. The polynomial least squares problem P_N is as follows: Given N data points (t_k, y_k), $k = 1, 2, \ldots, N$, where $t_1, t_2, \ldots, t_N$ are mutually distinct points on the real line, and N positive weights w_k^2, find a polynomial $q^0 \in \mathbb{P}_{n-1}$, $n \le N$, such that

$$P_N: \quad \sum_{k=1}^{N} w_k^2 \left(y_k - q^0(t_k)\right)^2 \le \sum_{k=1}^{N} w_k^2 \left(y_k - q(t_k)\right)^2 \quad \text{for all } q \in \mathbb{P}_{n-1}.$$

With Problem P_N one associates the discrete inner product

$$(u, v)_{\mathrm{d}\lambda_N} = \int_{\mathbb{R}} u(t)v(t)\mathrm{d}\lambda_N := \sum_{k=1}^{N} w_k^2 u(t_k)v(t_k), \tag{4.1}$$

and the norm $\|u\|_{\mathrm{d}\lambda_N}^2 = (u, u)_{\mathrm{d}\lambda_N}$, in terms of which P_N can be written as

$$\|y - q^0\|_{\mathrm{d}\lambda_N}^2 \le \|y - q\|_{\mathrm{d}\lambda_N}^2 \quad \text{for all } q \in \mathbb{P}_{n-1}.$$

It is well known that the problem allows an elegant solution by means of the orthonormal polynomials $\tilde{\pi}_k(\,\cdot\,) = \tilde{\pi}_k(\,\cdot\,; \mathrm{d}\lambda_N)$. Recall that there are exactly N such polynomials, $\tilde{\pi}_0, \tilde{\pi}_1, \ldots, \tilde{\pi}_{N-1}$; we define

$$\tilde{\pi}_N(t) = \frac{\prod_{k=1}^{N}(t - t_k)}{\|\pi_{N-1}\|}, \tag{4.2}$$

where π_{N-1} is the *monic* orthogonal polynomial.

4.1. Matrix formulation of the least squares problem and its solution. Let $\boldsymbol{J} = \boldsymbol{J}_N(\mathrm{d}\lambda_N)$ be the Jacobi matrix of order N for the measure $\mathrm{d}\lambda_N$ (cf. (2.8)) and $\boldsymbol{p}^T = [\tilde{\pi}_0, \tilde{\pi}_1, \ldots, \tilde{\pi}_{N-1}]$ the vector of the N discrete orthonormal polynomials. Then, similarly as in (3.7),

$$t\boldsymbol{p}(t) = \boldsymbol{J}\boldsymbol{p}(t) + \tilde{\pi}_N(t)\boldsymbol{e}_N, \tag{4.3}$$

where $\tilde{\pi}_N(t)$ is defined as in (4.2). Note by (4.3) and (4.2) that the eigenvalues of $\boldsymbol{J}$ are the knots $t_1, t_2, \ldots, t_N$. Thus, if $\boldsymbol{P} = [\boldsymbol{p}(t_1), \boldsymbol{p}(t_2), \ldots, \boldsymbol{p}(t_N)]$, then

$$\boldsymbol{J}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{\Lambda}, \quad \boldsymbol{\Lambda} = \mathrm{diag}(t_1, t_2, \ldots, t_N). \tag{4.4}$$

As a consequence of dual orthogonality (cf. [41, §2.4.6]), one has

$$\boldsymbol{p}^T(t_k)\boldsymbol{p}(t_k) = w_k^{-2}, \quad k = 1, 2, \ldots, N,$$

so that $w_k\boldsymbol{p}(t_k)$ are the normalized eigenvectors of $\boldsymbol{J}$. Thus, if

$$\boldsymbol{D} = \mathrm{diag}(w_1, w_2, \ldots, w_N), \tag{4.5}$$

ETNA
Kent State University
etna@mcs.kent.edu

the matrix $\boldsymbol{PD}$ is orthogonal, and one has

$$\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{D}^{-2}, \quad \boldsymbol{P}\boldsymbol{D}^2\boldsymbol{P}^T = \boldsymbol{I}. \tag{4.6}$$

Finally,

$$\boldsymbol{e}_1^T\boldsymbol{P}\boldsymbol{D} = [w_1, w_2, \ldots, w_N]\tilde{\pi}_0. \tag{4.7}$$

Now let $\boldsymbol{t} = [t_1, t_2, \ldots, t_N]^T$, $\boldsymbol{y} = [y_1, y_2, \ldots, y_N]^T$, and let $q(t) = \boldsymbol{p}^T(t)\boldsymbol{c}$ be any polynomial of degree $N-1$ with coefficients $\boldsymbol{c} = [c_0, c_1, \ldots, c_{N-1}]^T$ in the basis of the orthonormal polynomials. One checks that $q(\boldsymbol{t}) = \boldsymbol{P}^T\boldsymbol{c}$. In terms of the Euclidean vector norm $\|\cdot\| = \|\cdot\|_{\mathbb{R}^N}$, the squared error in Problem P_N for the polynomial q, in view of (4.5), can be written as

$$\begin{aligned} \|y - q\|^2_{\mathrm{d}\lambda_N} &= \|\boldsymbol{D}\,(\boldsymbol{y} - q(\boldsymbol{t}))\,\|^2 \\ &= \|\boldsymbol{D}(\boldsymbol{y} - \boldsymbol{P}^T\boldsymbol{c})\|^2 = \|\boldsymbol{P}\boldsymbol{D}\cdot\boldsymbol{D}(\boldsymbol{y} - \boldsymbol{P}^T\boldsymbol{c})\|^2 \\ &= \|\boldsymbol{P}\boldsymbol{D}^2\boldsymbol{y} - \boldsymbol{c}\|^2, \end{aligned} \tag{4.8}$$

where the orthogonality of $\boldsymbol{PD}$ and the second relation in (4.6) have been used in the last two equations. Choosing $\boldsymbol{c} = \boldsymbol{P}\boldsymbol{D}^2\boldsymbol{y}$ drives the error to zero and yields the interpolation polynomial of degree $N-1$. The solution of the least squares problem P_N, on the other hand, requires $\boldsymbol{c} = \begin{bmatrix} \boldsymbol{c}_n \\ \boldsymbol{0} \end{bmatrix}$, where $\boldsymbol{c}_n = [c_0, c_1, \ldots, c_{n-1}]^T$, and by (4.8) is equal to

$$q^0(t) = \boldsymbol{p}^T(t)\begin{bmatrix} \boldsymbol{c}_n \\ \boldsymbol{0} \end{bmatrix}, \quad \boldsymbol{c}_n = \boldsymbol{P}_{[1:n]}\boldsymbol{D}^2\boldsymbol{y}. \tag{4.9}$$

Here, $\boldsymbol{P}_{[1:n]}$ is the matrix formed with the first n rows of $\boldsymbol{P}$.

4.2. Updating and downdating the least squares solution. Suppose we adjoin to the N data points considered in §4.1 an additional point (t_{N+1}, y_{N+1}) and give it the weight w^2_{N+1}. How can the solution of the least squares problem P_N for the original N data points be used to obtain the solution of the least squares problem for the augmented set of $N+1$ data points? This is the problem of *updating* the least squares solution. There is an analogous problem of *downdating* whereby a single data point is deleted. An interesting treatment of these problems by matrix methods is given in [12].

We discuss here updating techniques only and refer to the cited reference for similar downdating techniques. In essence, the problem of updating can be considered as solved once we have constructed the Jacobi matrix $\boldsymbol{J}_{\mathrm{up}} = \boldsymbol{J}_{N+1}(\mathrm{d}\lambda_{N+1})$ of order $N+1$ for the augmented measure $\mathrm{d}\lambda_{N+1}$ from the Jacobi matrix $\boldsymbol{J} = \boldsymbol{J}_N(\mathrm{d}\lambda_N)$ for the original measure $\mathrm{d}\lambda_N$, the inner product for $\mathrm{d}\lambda_{N+1}$ being

$$(u, v)_{\mathrm{d}\lambda_{N+1}} = \int_{\mathbb{R}} u(t)v(t)\mathrm{d}\lambda_{N+1}(t) := \sum_{k=1}^{N+1} w_k^2 u(t_k)v(t_k). \tag{4.10}$$

Let (cf. (3.27))

$$\beta_0 = \int_{\mathbb{R}} \mathrm{d}\lambda_N(t), \quad \beta_{0,\mathrm{up}} = \int_{\mathbb{R}} \mathrm{d}\lambda_{N+1}(t), \tag{4.11}$$

so that

$$\tilde{\pi}_0 = 1/\sqrt{\beta_0}, \quad \tilde{\pi}_{0,\mathrm{up}} = 1/\sqrt{\beta_{0,\mathrm{up}}}. \tag{4.12}$$

ETNA
Kent State University
etna@mcs.kent.edu

There is a unique orthogonal matrix Q_{N+1} of order $N+1$ whose first row is prescribed to be

$$e_1^T Q_{N+1} = \frac{1}{\sqrt{\beta_{0,\mathrm{up}}}}(\sqrt{\beta_0}e_1^T + w_{N+1}e_{N+1}^T) \tag{4.13}$$

and which accomplishes a similarity transformation of the matrix $\begin{bmatrix} J & 0 \\ 0^T & t_{N+1} \end{bmatrix}$ to tridiagonal form,

$$Q_{N+1}\begin{bmatrix} J & 0 \\ 0^T & t_{N+1} \end{bmatrix} Q_{N+1}^T = T_{N+1}, \quad T_{N+1} \text{ tridiagonal} \tag{4.14}$$

(cf. [42, p. 113, (7-2-2)]). We claim that

$$J_{\mathrm{up}} = T_{N+1}. \tag{4.15}$$

To see this, recall that, with $Q = PD$ (orthogonal), Eq. (4.4) implies $J = P\Lambda P^{-1} = QD^{-1}\Lambda DQ^{-1}$, hence

$$J = Q\Lambda Q^T. \tag{4.16}$$

By (4.7), in view of the first relation in (4.12), there holds

$$\sqrt{\beta_0}e_1^T Q = [w_1, w_2, \ldots, w_N]. \tag{4.17}$$

The analogous relations for the augmented problem are

$$J_{\mathrm{up}} = Q_{\mathrm{up}}\Lambda_{\mathrm{up}}Q_{\mathrm{up}}^T, \quad \Lambda_{\mathrm{up}} = \begin{bmatrix} \Lambda & 0 \\ 0^T & t_{N+1} \end{bmatrix}$$

and

$$\sqrt{\beta_{0,\mathrm{up}}}e_1^T Q_{\mathrm{up}} = [w_1, w_2, \ldots, w_N, w_{N+1}], \tag{4.18}$$

where e_1 now has dimension $N+1$. Define

$$Q^* = Q_{\mathrm{up}}\begin{bmatrix} Q^T & 0 \\ 0^T & 1 \end{bmatrix}.$$

Then

$$\begin{aligned} Q^*\begin{bmatrix} J & 0 \\ 0^T & t_{N+1} \end{bmatrix} Q^{*T} &= Q_{\mathrm{up}}\begin{bmatrix} Q^T & 0 \\ 0^T & 1 \end{bmatrix}\begin{bmatrix} J & 0 \\ 0^T & t_{N+1} \end{bmatrix}\begin{bmatrix} Q & 0 \\ 0^T & 1 \end{bmatrix} Q_{\mathrm{up}}^T \\ &= Q_{\mathrm{up}}\begin{bmatrix} Q^T JQ & 0 \\ 0^T & t_{N+1} \end{bmatrix} Q_{\mathrm{up}}^T, \end{aligned}$$

hence, by (4.16),

$$Q^*\begin{bmatrix} J & 0 \\ 0^T & t_{N+1} \end{bmatrix} Q^{*T} = Q_{\mathrm{up}}\begin{bmatrix} \Lambda & 0 \\ 0^T & t_{N+1} \end{bmatrix} Q_{\mathrm{up}}^T = Q_{\mathrm{up}}\Lambda_{\mathrm{up}}Q_{\mathrm{up}}^T = J_{\mathrm{up}}.$$

Furthermore, using (4.18),

$$e_1^T Q^* = e_1^T Q_{\mathrm{up}}\begin{bmatrix} Q^T & 0 \\ 0^T & 1 \end{bmatrix} = \frac{1}{\sqrt{\beta_{0,\mathrm{up}}}}[w_1, \ldots, w_N, w_{N+1}]\begin{bmatrix} Q^T & 0 \\ 0^T & 1 \end{bmatrix},$$

ETNA
Kent State University
etna@mcs.kent.edu

which by (4.7) and the first of (4.12) becomes

$$\begin{aligned} \boldsymbol{e}_1^T \boldsymbol{Q}^* &= \frac{1}{\sqrt{\beta_{0,\mathrm{up}}}} [\sqrt{\beta_0}\boldsymbol{e}_1^T \boldsymbol{Q} \;\; w_{N+1}] \begin{bmatrix} \boldsymbol{Q}^T & \boldsymbol{0} \\ \boldsymbol{0}^T & 1 \end{bmatrix} \\ &= \frac{1}{\sqrt{\beta_{0,\mathrm{up}}}} [\sqrt{\beta_0}\boldsymbol{e}_1^T \;\; w_{N+1}] = \frac{1}{\sqrt{\beta_{0,\mathrm{up}}}} [\sqrt{\beta_0}\boldsymbol{e}_1^T + w_{N+1}\boldsymbol{e}_{N+1}^T]. \end{aligned}$$

Thus, $\boldsymbol{Q}^*$ satisfies exactly the properties defining $\boldsymbol{Q}_{N+1}$, showing that indeed $\boldsymbol{J}_{\mathrm{up}} = \boldsymbol{T}_{N+1}$. The analogue of (4.3),

$$t\boldsymbol{p}_{\mathrm{up}}(t) = \boldsymbol{J}_{\mathrm{up}}\boldsymbol{p}_{\mathrm{up}}(t) + \tilde{\pi}_{N+1,\mathrm{up}}(t)\boldsymbol{e}_{N+1},$$

in combination with the second relation of (4.12) can now be used to generate the new discrete orthonormal polynomials, and with them the updated least squares solution by the analogue of (4.9).

Algorithmically, the transformation (4.14) can be implemented by a sequence of appropriate Givens rotations (cf. [12, Eq. (4.7)]). The updating technique described here is not the only possible one; for others, see [*ibid.*, §§4.3–4.7].

Since the solution for the one-point least squares problem P_1 is trivially $\tilde{\pi}_0 = 1/|w_1|$, $\boldsymbol{J} = [t_1]$, $\boldsymbol{c} = [|w_1|y_1]$, one can use the updating technique to build up the least squares solutions of P_N successively for $N = 2, 3, \ldots$ without necessarily having to store the entire data set.

5. Linear algebraic systems. Many linear algebra problems that involve a symmetric positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ can be related to discrete orthogonal polynomials supported on the spectrum of $\boldsymbol{A}$. This provides the link between linear algebra and analysis. It may be appropriate, at this point, to recall that the use of discrete (and other) orthogonal polynomials in the context of linear algebra has been pioneered by Stiefel [44]; see also [36, §14].

For simplicity assume that $\boldsymbol{A}$ has distinct[1] eigenvalues λ_n,

$$0 < \lambda_N < \lambda_{N-1} < \cdots < \lambda_1, \tag{5.1}$$

and denote the respective (orthonormal) eigenvectors by $\boldsymbol{v}_n$,

$$\boldsymbol{A}\boldsymbol{v}_n = \lambda_n \boldsymbol{v}_n, \quad \boldsymbol{v}_n^T \boldsymbol{v}_m = \delta_{nm}, \quad n, m = 1, 2, \ldots, N. \tag{5.2}$$

(There should be no danger of confusing these λ's with the weights of the Gauss quadrature rule in (2.2).) Thus, with $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N]$, $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$, there holds

$$\boldsymbol{A}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{\Lambda}, \quad \boldsymbol{\Lambda} = \boldsymbol{V}^T\boldsymbol{A}\boldsymbol{V}. \tag{5.3}$$

Now consider a discrete measure $\mathrm{d}\rho_N$ defined by

$$\int_{\mathbb{R}_+} f(t)\mathrm{d}\rho_N(t) := \sum_{k=1}^{N} \rho_k^2 f(\lambda_k), \tag{5.4}$$

where ρ_k^2 are positive weights, and assume, temporarily, that the measure $\mathrm{d}\rho_N$ is normalized,

$$\int_{\mathbb{R}_+} \mathrm{d}\rho_N(t) = 1. \tag{5.5}$$

[1] Otherwise, some terms in (5.4) below consolidate, so that N becomes smaller.

ETNA
Kent State University
etna@mcs.kent.edu

It is possible to generate the orthonormal polynomials $\tilde{\pi}_k(\cdot\,; \mathrm{d}\rho_N)$, $k = 0, 1, \ldots, N-1$, resp. the associated Jacobi matrix $\boldsymbol{J}_N = \boldsymbol{J}_N(\mathrm{d}\rho_N)$, entirely by matrix-vector multiplications involving the matrix $\boldsymbol{A}$ and by an initial vector

$$\boldsymbol{h}_0 = \sum_{k=1}^{N} \rho_k \boldsymbol{v}_k, \quad \|\boldsymbol{h}_0\| = 1, \tag{5.6}$$

whose components in the basis of the normalized eigenvectors are the (positive or negative) square roots of the weights ρ_k^2. (Here and in the following, $\|\cdot\|$ denotes the Euclidean vector norm.) A method accomplishing this is the *Lanczos algorithm*, which is briefly described in §5.1. The subsequent sections give applications of this algorithm when combined with quadrature methods.

5.1. The Lanczos algorithm. Let $\boldsymbol{h}_0$ be given as in (5.6), and define $\boldsymbol{h}_{-1} = \boldsymbol{0}$. The Lanczos algorithm is defined as follows:

$$\begin{aligned} &\text{for } j = 0, 1, \ldots, N-1 \text{ do} \\ &\left[\begin{array}{l} \alpha_j = \boldsymbol{h}_j^T \boldsymbol{A} \boldsymbol{h}_j \\ \tilde{\boldsymbol{h}}_{j+1} = (\boldsymbol{A} - \alpha_j \boldsymbol{I})\boldsymbol{h}_j - \gamma_j \boldsymbol{h}_{j-1} \\ \gamma_{j+1} = \|\tilde{\boldsymbol{h}}_{j+1}\| \\ \boldsymbol{h}_{j+1} = \tilde{\boldsymbol{h}}_{j+1}/\gamma_{j+1} \end{array}\right. \end{aligned} \tag{5.7}$$

Note that γ_0 can be arbitrary, but is often defined, in accordance with (5.5), by $\gamma_0 = 1$, or, in accordance with (5.10) below, by $\gamma_0 = \beta_0$.

The vectors $\boldsymbol{h}_j$ so generated are orthonormal, as one checks by induction, and it is evident from (5.7) that $\{\boldsymbol{h}_j\}_{j=0}^{n}$, $n < N$, forms an orthonormal basis for the *Krylov space*

$$\mathcal{K}_n(\boldsymbol{A}, \boldsymbol{h}_0) = \operatorname{span}(\boldsymbol{h}_0, \boldsymbol{A}\boldsymbol{h}_0, \ldots, \boldsymbol{A}^n \boldsymbol{h}_0).$$

One also verifies by induction that

$$\boldsymbol{h}_j = p_j(\boldsymbol{A})\boldsymbol{h}_0, \tag{5.8}$$

where p_j is a polynomial of degree j satisfying the three-term recurrence relation

$$\begin{aligned} \gamma_{j+1} p_{j+1}(\lambda) &= (\lambda - \alpha_j) p_j(\lambda) - \gamma_j p_{j-1}(\lambda), \\ &\qquad j = 0, 1, \ldots, N-1, \\ p_{-1}(\lambda) &= 0, \quad p_0(\lambda) = 1. \end{aligned} \tag{5.9}$$

We claim that $p_k(\cdot) = \tilde{\pi}_k(\cdot\,; \mathrm{d}\rho_N)$. Indeed, from the second relation in (5.3) one has

$$p_n(\boldsymbol{\Lambda}) = \boldsymbol{V}^T p_n(\boldsymbol{A}) \boldsymbol{V},$$

hence, by (5.8),

$$\boldsymbol{h}_n = \boldsymbol{V} p_n(\boldsymbol{\Lambda}) \boldsymbol{V}^T \boldsymbol{h}_0.$$

Orthonormality $\boldsymbol{h}_n^T \boldsymbol{h}_m = \delta_{nm}$ of the Lanczos vectors $\boldsymbol{h}_j$ then yields

$$\boldsymbol{h}_0^T \boldsymbol{V} p_n(\boldsymbol{\Lambda}) \boldsymbol{V}^T \boldsymbol{V} p_m(\boldsymbol{\Lambda}) \boldsymbol{V}^T \boldsymbol{h}_0 = \boldsymbol{h}_0^T \boldsymbol{V} p_n(\boldsymbol{\Lambda}) p_m(\boldsymbol{\Lambda}) \boldsymbol{V}^T \boldsymbol{h}_0 = \delta_{nm},$$

ETNA
Kent State University
etna@mcs.kent.edu

which, since $\boldsymbol{V}^T\boldsymbol{h}_0 = \sum_{k=1}^N \rho_k \boldsymbol{e}_k$ by (5.6), implies

$$\sum_{k,\ell=1}^{N} \rho_k \boldsymbol{e}_k^T \operatorname{diag}\left(p_n(\lambda_1)p_m(\lambda_1), \ldots, p_n(\lambda_N)p_m(\lambda_N)\right) \rho_\ell \boldsymbol{e}_\ell$$
$$= \sum_{k,\ell=1}^{N} \rho_k \rho_\ell \boldsymbol{e}_k^T p_n(\lambda_\ell)p_m(\lambda_\ell)\boldsymbol{e}_\ell = \sum_{k=1}^{N} \rho_k^2 p_n(\lambda_k)p_m(\lambda_k) = \delta_{nm},$$

as claimed.

The recurrence relation (5.9), therefore, must be identical with the one in (3.2), i.e., $\gamma_j = \sqrt{\beta_j}$.

If the measure $\mathrm{d}\rho_N$ is not normalized and, as in (3.27), one puts

$$\beta_0 = \int_{\mathbb{R}_+} \mathrm{d}\rho_N(t), \tag{5.10}$$

then the recurrence relation still holds, except that one must define $p_0(\lambda) = 1/\sqrt{\beta_0}$.

5.2. Bounds for matrix functionals. Given $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ positive definite and f a function analytic on an interval containing the spectrum of $\boldsymbol{A}$, the problem to be considered is finding lower and upper bounds for the bilinear form

$$\boldsymbol{u}^T f(\boldsymbol{A})\boldsymbol{v}, \tag{5.11}$$

where $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^N$ are given vectors. The solution of this problem has many applications; some will be discussed in subsequent sections. For applications to constrained least squares problems for matrices, see [29], and [7] for applications to the evaluation of suitable regularization parameters in Tikhonov regularization. The case $f(t) = (\lambda - t)^{-1}$ with λ outside the spectrum of $\boldsymbol{A}$ is important in physical chemistry and solid state physics applications; for references, see [32, §1].

Let first $\boldsymbol{u} = \boldsymbol{v}$. With $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N]$ and $\boldsymbol{\Lambda}$ as defined in (5.2), (5.3), we let

$$\boldsymbol{u} = \sum_{k=1}^{N} \rho_k \boldsymbol{v}_k \tag{5.12}$$

and for simplicity assume $\rho_k \neq 0$ for all k. Then $\boldsymbol{u} = \boldsymbol{V}\boldsymbol{\rho}$, $\boldsymbol{\rho} = [\rho_1, \rho_2, \ldots, \rho_N]^T$, and $f(\boldsymbol{A}) = \boldsymbol{V} f(\boldsymbol{\Lambda})\boldsymbol{V}^T$. Therefore,

$$\boldsymbol{u}^T f(\boldsymbol{A})\boldsymbol{u} = \boldsymbol{\rho}^T \boldsymbol{V}^T \boldsymbol{V} f(\boldsymbol{\Lambda})\boldsymbol{V}^T \boldsymbol{V}\boldsymbol{\rho} = \boldsymbol{\rho}^T f(\boldsymbol{\Lambda})\boldsymbol{\rho} = \sum_{k=1}^{N} \rho_k^2 f(\lambda_k),$$

that is,

$$\boldsymbol{u}^T f(\boldsymbol{A})\boldsymbol{u} = \int_{\mathbb{R}_+} f(t)\mathrm{d}\rho_N(t), \tag{5.13}$$

where $\mathrm{d}\rho_N$ is the discrete measure defined in (5.4). The desired bounds can be obtained by applying Gauss, Gauss-Radau, or Gauss-Lobatto quadrature to the integral in (5.13), provided the appropriate derivative of f has constant sign (cf. §§2.1,2.2). The Lanczos algorithm (cf. §5.1) applied with $\boldsymbol{h}_0 = \boldsymbol{u}/\|\boldsymbol{u}\|$ furnishes the necessary (discrete) orthogonal polynomials, resp. their recursion coefficients. For Gauss formulae, the quality of the bounds, even

ETNA
Kent State University
etna@mcs.kent.edu

when no specific information is known about the sign of derivatives of f, can be estimated in terms of the absolute values of these derivatives and the quantities $\gamma_j^2 = \beta_j$ generated during the Lanczos process (cf. [6]). One simply makes use of (2.11).

The case $\boldsymbol{u} \neq \boldsymbol{v}$ can be handled by using the polarization identity $\boldsymbol{u}^T f(\boldsymbol{A})\boldsymbol{v} = \frac{1}{4}(\boldsymbol{p}^T f(\boldsymbol{A})\boldsymbol{p} - \boldsymbol{q}^T f(\boldsymbol{A})\boldsymbol{q})$ where $\boldsymbol{p} = \boldsymbol{u} + \boldsymbol{v}$, $\boldsymbol{q} = \boldsymbol{u} - \boldsymbol{v}$ (cf. [2, §3.1.2], [3, p. 426], or, for a similar identity, [32, Eq. (3)]) and applying appropriate bounds to the first and second term of the identity. Alternatively, a "nonsymmetric" Lanczos process can be applied in conjunction with Gauss-Radau quadrature [30].

For the important function $f(t) = t^{-1}$ (see, e.g., (5.19) or (5.22) below), the case of an arbitrary nonsingular matrix $\boldsymbol{A}$ can be reduced to the case of a symmetric positive definite matrix by noting that

$$\boldsymbol{u}^T \boldsymbol{A}^{-1} \boldsymbol{v} = \boldsymbol{u}^T (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{w}, \quad \boldsymbol{w} = \boldsymbol{A}^T \boldsymbol{v} \tag{5.14}$$

(cf. [2, §3.2], [3, p. 427]).

5.3. Error bounds. We consider now the system of linear algebraic equations

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{5.15}$$

with $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ symmetric and positive definite. Given any approximation $\boldsymbol{x}^*$ to the exact solution $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$, the object is to estimate the error $\boldsymbol{x} - \boldsymbol{x}^*$ in some norm. We begin with using the Euclidean vector norm $\| \cdot \|$.

Let $\boldsymbol{r}$ be the *residual* of $\boldsymbol{x}^*$,

$$\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^*. \tag{5.16}$$

Since $\boldsymbol{x} - \boldsymbol{x}^* = \boldsymbol{A}^{-1}\boldsymbol{r}$, we have

$$\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 = \boldsymbol{r}^T \boldsymbol{A}^{-2} \boldsymbol{r}, \tag{5.17}$$

which is (5.11) with $\boldsymbol{u} = \boldsymbol{v} = \boldsymbol{r}$ and $f(t) = t^{-2}$. Here the derivatives are $f^{(2n)}(t) = (2n+1)!\, t^{-(2n+2)}$, $f^{(2n+1)}(t) = -(2n+2)!\, t^{-(2n+3)}$, so that

$$f^{(2n)}(t) > 0, \quad f^{(2n+1)}(t) < 0 \text{ for } t \in \mathbb{R}_+. \tag{5.18}$$

The n-point Gauss formula (with $n < N$) applied to the integral in (5.13) (with $f(t) = t^{-2}$) thus produces a lower bound for the squared error (5.17). If the spectrum of $\boldsymbol{A}$ can be enclosed in an interval $[a, b]$, $0 < a < b$, then the "left-sided" $(n+1)$-point Gauss-Radau formula yields an upper bound, and the "right-sided" formula a lower bound for (5.17). The Lanczos algorithm applied with $\boldsymbol{h}_0 = \boldsymbol{r}/\|\boldsymbol{r}\|$ yields the recursion coefficients for the orthogonal polynomials required for generating these quadrature rules.

If instead of the Euclidean norm one takes the A-norm $\|\boldsymbol{u}\|_A^2 = \boldsymbol{u}^T \boldsymbol{A}\boldsymbol{u}$ (cf. [31]), then

$$\|\boldsymbol{x} - \boldsymbol{x}^*\|_A^2 = \boldsymbol{r}^T \boldsymbol{A}^{-1} \boldsymbol{r}, \tag{5.19}$$

which is (5.11) with $\boldsymbol{u} = \boldsymbol{v} = \boldsymbol{r}$ and $f(t) = t^{-1}$. Since this function satisfies the same inequalities as in (5.18), the Gauss and Gauss-Radau formulae applied to the integral in (5.13) (with $f(t) = t^{-1}$) produce the same kind of bounds as in the case of the Euclidean norm. The difference between the N-point and n-point Gauss quadrature approximation equals $\|\boldsymbol{x} - \boldsymbol{x}_n\|_A^2 / \|\boldsymbol{r}\|^2$, where $\boldsymbol{x}_n$ is the nth iterate of the conjugate gradient method started with $\boldsymbol{r}$ (cf. [32, Eq. (50)]). The conjugate gradient method, in fact, can be used not only as an alternative to the Lanczos algorithm to generate the recursion coefficients of the orthogonal

ETNA
Kent State University
etna@mcs.kent.edu

polynomials, but also to improve the approximation $\boldsymbol{x}^*$. The A-norm of the improved approximation can then be estimated from below and above (see [11, §5]). For analogous error estimates in the Euclidean norm, see [9].

The idea of using Gauss-Radau quadratures in combination with (5.18) to get error bounds for linear systems goes back to Dahlquist, Eisenstat, and Golub [10]. They also suggest a procedure based on linear programming when all eigenvalues are known (cf. [10, §2]). This requires knowledge of the moments μ_m of $\mathrm{d}\rho_N$, which by (5.13) are given by

$$\mu_m = \int_{\mathbb{R}_+} t^m \mathrm{d}\rho_N(t) = \boldsymbol{r}^T \boldsymbol{A}^m \boldsymbol{r}.$$

Thus, computing the first $2n+1$ moments $\mu_0, \mu_1, \ldots, \mu_{2n}$ amounts to generating the *Krylov sequence* $\boldsymbol{r}, \boldsymbol{Ar}, \ldots, \boldsymbol{A}^{2n}\boldsymbol{r}$ and computing the inner products of its members with $\boldsymbol{r}$. In view of

$$\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 = \int_{\mathbb{R}_+} t^{-2} \mathrm{d}\rho_N(t) = \sum_{k=1}^{N} \rho_k^2 \lambda_k^{-2},$$

an upper bound can be found by solving the linear programming problem

$$\text{max!} \sum_{k=1}^{N} \gamma_k \lambda_k^{-2} \tag{5.20}$$

subject to the constraints

$$\begin{aligned} \sum_{k=1}^{N} \gamma_k \lambda_k^m &= \mu_m, \quad m = 0, 1, \ldots, 2n, \\ \gamma_k &\geq 0, \quad k = 1, 2, \ldots, N. \end{aligned} \tag{5.21}$$

Here, n can be any integer $< N$. A lower bound can similarly be obtained by replacing "max" in (5.20) by "min". The same procedure, with λ_k^{-2} in (5.20) replaced by λ_k^{-1}, works for the A-norm.

The ideas outlined above, and still other ideas from the theory of moments, are applied in [10] to obtain upper and lower bounds for the errors in the Jacobi iterative method. Bounds for matrix moments $\mu_m = \boldsymbol{r}^T \boldsymbol{A}^m \boldsymbol{r}$ are similarly obtained in [24].

5.4. The diagonal elements of $\boldsymbol{A}^{-1}$. Given $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ positive definite, the problem is to find bounds for the diagonal elements $(\boldsymbol{A}^{-1})_{ii}$ of $\boldsymbol{A}^{-1}$, $i = 1, 2, \ldots, N$. Here,

$$(\boldsymbol{A}^{-1})_{ii} = \boldsymbol{e}_i^T \boldsymbol{A}^{-1} \boldsymbol{e}_i, \tag{5.22}$$

where $\boldsymbol{e}_i$ is the ith canonical basis vector. This is (5.11) with $\boldsymbol{u} = \boldsymbol{v} = \boldsymbol{e}_i$ and $f(t) = t^{-1}$. As before, f satisfies the inequalities (5.18).

5.4.1. Lower bound from Gauss quadrature. By virtue of (2.4) and the first of (5.18), the n-point Gauss quadrature sum (cf. (2.10), where $\mu_0 = 1$) yields a lower bound for the integral, i.e.,

$$(\boldsymbol{A}^{-1})_{ii} = \int_{\mathbb{R}_+} t^{-1} \mathrm{d}\rho_N(t) > \boldsymbol{e}_1^T \boldsymbol{J}_n^{-1} \boldsymbol{e}_1, \quad \boldsymbol{e}_1^T = [1, 0, \ldots, 0] \in \mathbb{R}^n, \tag{5.23}$$

ETNA
Kent State University
etna@mcs.kent.edu

where $\boldsymbol{J}_n = \boldsymbol{J}_n(\mathrm{d}\rho_N)$. Consider $n = 2$; we apply two steps of the Lanczos algorithm with $\boldsymbol{h}_0 = \boldsymbol{e}_i$ to generate

$$\boldsymbol{J}_2 = \begin{bmatrix} \alpha_0 & \gamma_1 \\ \gamma_1 & \alpha_1 \end{bmatrix}.$$

According to (5.7) we have

$$\begin{aligned} \alpha_0 &= a_{ii}, \\ \tilde{\boldsymbol{h}}_1 &= (\boldsymbol{A} - \alpha_0 \boldsymbol{I})\boldsymbol{e}_i = [a_{1i}, \ldots, a_{i-1,i}, 0, a_{i+1,i}, \ldots, a_{Ni}]^T, \\ \gamma_1 &= \sqrt{\sum_{k \neq i} a_{ki}^2} =: s_i, \\ \boldsymbol{h}_1 &= \tilde{\boldsymbol{h}}_1 / s_i, \\ \alpha_1 &= \frac{1}{s_i^2} \tilde{\boldsymbol{h}}_1^T \boldsymbol{A} \tilde{\boldsymbol{h}}_1 = \frac{1}{s_i^2} \sum_{k \neq i} \sum_{\ell \neq i} a_{k\ell} a_{ki} a_{\ell i}. \end{aligned} \tag{5.24}$$

Since

$$\boldsymbol{J}_2^{-1} = \frac{1}{\alpha_0 \alpha_1 - \gamma_1^2} \begin{bmatrix} \alpha_1 & -\gamma_1 \\ -\gamma_1 & \alpha_0 \end{bmatrix},$$

one has

$$\boldsymbol{e}_1^T \boldsymbol{J}_2^{-1} \boldsymbol{e}_1 = \frac{\alpha_1}{\alpha_0 \alpha_1 - \gamma_1^2}, \tag{5.25}$$

and therefore, by (5.23) and (5.24),

$$(\boldsymbol{A}^{-1})_{ii} > \frac{\displaystyle\sum_{k \neq i} \sum_{\ell \neq i} a_{k\ell} a_{ki} a_{\ell i}}{\displaystyle a_{ii} \sum_{k \neq i} \sum_{\ell \neq i} a_{k\ell} a_{ki} a_{\ell i} - \left(\sum_{k \neq i} a_{ki}^2 \right)^2}. \tag{5.26}$$

It should be noted that this bound, in contrast to those given below in §§5.4.2–5.4.3, does not require any information about the spectrum of $\boldsymbol{A}$.

5.4.2. Upper and lower bounds from Gauss-Radau quadrature. If the spectrum of $\boldsymbol{A}$ can be enclosed in the interval $[a, b]$, $0 < a < b$, then by the second of (5.18) and the first of (2.17) (with the inequality reversed) the "left-sided" $(n+1)$-point Gauss-Radau quadrature sum in (2.12) yields an upper bound, and similarly the "right-sided" quadrature sum in (2.16) a lower bound for the integral. Taking $n = 1$ in (2.14), (2.15), one gets

$$\boldsymbol{J}_2^{R,a}(\mathrm{d}\rho_N) = \begin{bmatrix} \alpha_0 & \gamma_1 \\ \gamma_1 & \alpha_1^R \end{bmatrix}, \quad \alpha_1^R = a + \frac{\gamma_1^2}{\alpha_0 - a},$$

where $\alpha_0 = a_{ii}$, $\gamma_1 = s_i$ from (5.24). Replacing here a by b yields $\boldsymbol{J}_2^{R,b}(\mathrm{d}\rho_N)$. From (5.25), where α_1 is replaced by α_1^R, a simple computation then gives

$$\frac{a_{ii} - b + s_i^2/b}{a_{ii}^2 - a_{ii} b + s_i^2} < (\boldsymbol{A}^{-1})_{ii} < \frac{a_{ii} - a + s_i^2/a}{a_{ii}^2 - a_{ii} a + s_i^2}. \tag{5.27}$$

ETNA
Kent State University
etna@mcs.kent.edu

5.4.3. Upper bound from Gauss-Lobatto quadrature. The $(n+2)$-point Gauss-Lobatto quadrature sum in (2.19), on account of (2.21) and the first of (5.18) (with n replaced by $n+1$), yields an upper bound for the integral. Taking $n=0$ in (2.22), one gets

$$\boldsymbol{J}_2^L(\mathrm{d}\rho_N) = \begin{bmatrix} \alpha_0 & \gamma_1^L \\ \gamma_1^L & \alpha_1^L \end{bmatrix},$$

where by (2.23) the quantities α_1^L and γ_1^L solve the 2×2 system

$$\begin{bmatrix} a-\alpha_0 & 1 \\ b-\alpha_0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1^L \\ (\gamma_1^L)^2 \end{bmatrix} = \begin{bmatrix} a(a-\alpha_0) \\ b(b-\alpha_0) \end{bmatrix}, \quad \alpha_0 = a_{ii}.$$

Carrying out the solution and using (5.25) with α_1, γ_1 replaced by α_1^L, γ_1^L, yields

$$(\boldsymbol{A}^{-1})_{ii} < \frac{a+b-a_{ii}}{ab}. \tag{5.28}$$

The results in §§5.4.1–5.4.3 are from [30, Thm. 5.1]. When $n > 2$, the quadrature sum $\boldsymbol{e}_1^T \boldsymbol{J}_n^{-1} \boldsymbol{e}_1$ can be computed for all three quadrature rules in terms of quantities generated during the course of the Lanczos algorithm; see [30, Thm. 5.3]. For an application to Vičsek fractal Hamiltonian matrices, see [25].

5.4.4. The trace of $\boldsymbol{A}^{-1}$ and the determinant of $\boldsymbol{A}$. In principle, each method described in §§5.4.1–5.4.3 can be used to estimate the trace

$$\mathrm{tr}\,(\boldsymbol{A}^{-1}) = \sum_{i=1}^{N} (\boldsymbol{A}^{-1})_{ii} \tag{5.29}$$

of $\boldsymbol{A}^{-1}$ by applying the method to each term in the sum of (5.29), hence N times. For large sparse matrices there are, however, more efficient estimation procedures based on sampling and a Monte Carlo approach (cf. [2, §4]).

Alternatively, we may note that (cf. [1])

$$\mathrm{tr}\,(\boldsymbol{A}^{-1}) = \sum_{k=1}^{N} \lambda_k^{-1} = \int_{\mathbb{R}_+} t^{-1} \mathrm{d}\rho_N(t), \tag{5.30}$$

where $\mathrm{d}\rho_N$ is the discrete measure (5.4) with $\rho_k = 1$, $k = 1, 2, \ldots, N$. As in §5.4.2, we may then apply Gauss-Radau quadratures on an interval $[a, b]$ containing all eigenvalues λ_k to get lower and upper bounds. The only difference is that now $\mathrm{d}\rho_N$ is no longer normalized, in fact

$$\mu_0 = \int_{\mathbb{R}_+} \mathrm{d}\rho_N(t) = N, \tag{5.31}$$

and the Lanczos algorithm, in accordance with (5.6), is to be started with

$$\boldsymbol{h}_0 = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \boldsymbol{v}_k.$$

Observing that

$$\begin{aligned} \mu_1 &= \int_{\mathbb{R}_+} t \mathrm{d}\rho_N(t) = \sum_{k=1}^{N} \lambda_k = \sum_{i=1}^{N} a_{ii} = \mathrm{tr}\,(\boldsymbol{A}), \\ \mu_2 &= \int_{\mathbb{R}_+} t^2 \mathrm{d}\rho_N(t) = \sum_{k=1}^{N} \lambda_k^2 = \sum_{i,j=1}^{N} a_{ij}^2 = \|\boldsymbol{A}\|_F^2, \end{aligned} \tag{5.32}$$

ETNA
Kent State University
etna@mcs.kent.edu

from (5.7) one gets

$$\alpha_0 = \boldsymbol{h}_0^T \boldsymbol{A} \boldsymbol{h}_0 = \frac{1}{N} \sum_{k=1}^{N} \boldsymbol{v}_k^T \sum_{\ell=1}^{N} \boldsymbol{A} \boldsymbol{v}_\ell = \frac{1}{N} \sum_{k=1}^{N} \lambda_k,$$

that is,

$$\alpha_0 = \frac{1}{N} \mu_1. \tag{5.33}$$

Furthermore,

$$\tilde{\boldsymbol{h}}_1 = (\boldsymbol{A} - \alpha_0 \boldsymbol{I}) \boldsymbol{h}_0 = \frac{1}{\sqrt{N}} (\boldsymbol{A} - \alpha_0 \boldsymbol{I}) \sum_{k=1}^{N} \boldsymbol{v}_k = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} (\lambda_k - \alpha_0) \boldsymbol{v}_k,$$

and

$$\gamma_1^2 = \tilde{\boldsymbol{h}}_1^T \tilde{\boldsymbol{h}}_1 = \frac{1}{N} \sum_{k=1}^{N} (\lambda_k - \alpha_0) \boldsymbol{v}_k^T \sum_{\ell=1}^{N} (\lambda_\ell - \alpha_0) \boldsymbol{v}_\ell.$$

An elementary calculation yields

$$\gamma_1^2 = \frac{1}{N} (\mu_2 - \frac{1}{N} \mu_1^2). \tag{5.34}$$

The rest of the calculation is the same as in §5.4.2, except that, by (5.31), one has to include the factor $\mu_0 = N$ in (5.25). The result is

$$\frac{1}{b} \left(1 - \frac{\frac{1}{N} \mu_1^2 + N b^2}{\mu_2 - b \mu_1} \right) < \frac{1}{N} \operatorname{tr} (\boldsymbol{A}^{-1}) < \frac{1}{a} \left(1 - \frac{\frac{1}{N} \mu_1^2 + N a^2}{\mu_2 - a \mu_1} \right), \tag{5.35}$$

with μ_1, μ_2 given by (5.32). The same inequalities, in a different form, are derived in [1, Eq. (9)] by means of difference equations.

As far as the determinant $\det \boldsymbol{A}$ is concerned, we note that the trace is invariant to similarity transformations, so that by (5.3)

$$\operatorname{tr} (\ln \boldsymbol{A}) = \operatorname{tr} (\boldsymbol{V} \ln \boldsymbol{\Lambda} \boldsymbol{V}^T) = \operatorname{tr} (\ln \boldsymbol{\Lambda}) = \sum_{k=1}^{N} \ln \lambda_k = \ln \prod_{k=1}^{N} \lambda_k.$$

Since $\det \boldsymbol{A} = \prod_k \lambda_k$, this yields

$$\det \boldsymbol{A} = \exp(\operatorname{tr} (\ln \boldsymbol{A})). \tag{5.36}$$

Here, the trace of $\ln \boldsymbol{A}$ can be estimated as described for $\boldsymbol{A}^{-1}$, with the function $f(t) = t^{-1}$ replaced by $f(t) = \ln t$. This latter function has derivatives whose signs are opposite to those in (5.18), which gives rise to bounds whose types are opposite to those obtained in §§5.4.1–5.4.3.

Note that in place of $\boldsymbol{J}_2^{-1}$ in the quadrature sum (5.25), we now require $\ln \boldsymbol{J}_2$. This can be defined by linear interpolation at the eigenvalues $0 < \kappa_2 < \kappa_1$ of $\boldsymbol{J}_2$ (see, e.g., [18, Ch. 5]),

$$\ln \boldsymbol{J}_2 = \frac{1}{\kappa_1 - \kappa_2} [(\boldsymbol{J}_2 - \kappa_2 \boldsymbol{I}) \ln \kappa_1 + (\kappa_1 \boldsymbol{I} - \boldsymbol{J}_2) \ln \kappa_2].$$

In particular, therefore,

$$\boldsymbol{e}_1^T \ln \boldsymbol{J}_2 \boldsymbol{e}_1 = \frac{1}{\kappa_1 - \kappa_2} [(\alpha_0 - \kappa_2) \ln \kappa_1 + (\kappa_1 - \alpha_0) \ln \kappa_2], \tag{5.37}$$

where α_0 is given by the first of (5.24) resp. by (5.33).

ETNA
Kent State University
etna@mcs.kent.edu

5.5. Iterative methods. Consider again the system (5.15) with $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ symmetric and positive definite. Based on the splitting $\boldsymbol{A} = \boldsymbol{M} - \boldsymbol{N}$, where $\boldsymbol{M}$ and $\boldsymbol{N}$ are symmetric and $\boldsymbol{M}$ positive definite, a large class of iterative methods for solving (5.15) is given by

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_{k-1} + \omega_{k+1}(\delta \boldsymbol{z}_k + \boldsymbol{x}_k - \boldsymbol{x}_{k-1}), \quad k = 0, 1, 2, \ldots, \quad \boldsymbol{x}_{-1} = \boldsymbol{0}, \tag{5.38}$$

where

$$\boldsymbol{M}\boldsymbol{z}_k = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k. \tag{5.39}$$

In practice, $\boldsymbol{M}$ is chosen such that linear systems with this matrix as coefficient matrix can be easily solved. Depending on the choice of parameters, the iteration (5.38) includes such methods as the conjugate gradient, the Richardson second-order, and the Chebyshev semi-iterative method. For optimizing the speed of convergence of the two latter methods, it is important to have good estimates of the smallest and largest eigenvalues of $\boldsymbol{M}^{-1}\boldsymbol{N}$. Such estimates can be found via certain discrete orthogonal polynomials and the modified Chebyshev algorithm (cf. §3.4.2) generating them; see [28].

To analyze the speed of convergence of the iteration, there is no loss of generality in assuming, as we do, that $\boldsymbol{b} = \boldsymbol{0}$, and thus considering convergence of $\boldsymbol{x}_k$ resp. $\boldsymbol{z}_k$ to the zero vector.

Substituting $\boldsymbol{x}_k = -\boldsymbol{A}^{-1}\boldsymbol{M}\boldsymbol{z}_k$ obtained from (5.39) into (5.38) yields

$$\boldsymbol{z}_{k+1} = \omega_{k+1}\boldsymbol{B}\boldsymbol{z}_k + (1 - \omega_{k+1})\boldsymbol{z}_{k-1}, \quad \boldsymbol{z}_{-1} = \boldsymbol{0}, \tag{5.40}$$

where

$$\boldsymbol{B} = \boldsymbol{I} - \delta\boldsymbol{M}^{-1}\boldsymbol{A}. \tag{5.41}$$

Since $\boldsymbol{B} = \boldsymbol{I} - \delta\boldsymbol{M}^{-1}(\boldsymbol{M} - \boldsymbol{N}) = (1 - \delta)\boldsymbol{I} + \delta\boldsymbol{M}^{-1}\boldsymbol{N}$, the eigenvalues ν_n of $\boldsymbol{M}^{-1}\boldsymbol{N}$ are related to the eigenvalues λ_n of $\boldsymbol{B}$ by

$$\nu_n = 1 + \frac{1}{\delta}(\lambda_n - 1), \quad n = 1, 2, \ldots, N. \tag{5.42}$$

We may therefore focus attention on the eigenvalues of $\boldsymbol{B}$. Note that the eigenvalue problem $\boldsymbol{B}\boldsymbol{v} = \lambda\boldsymbol{v}$ for $\boldsymbol{B}$ is equivalent to the generalized eigenvalue problem

$$\boldsymbol{A}\boldsymbol{v} = \kappa\boldsymbol{M}\boldsymbol{v}, \quad \kappa = \frac{1 - \lambda}{\delta}. \tag{5.43}$$

Since $\boldsymbol{M}$ is positive definite, the Cholesky decomposition $\boldsymbol{M} = \boldsymbol{L}\boldsymbol{L}^T$ will transform (5.43) into an ordinary eigenvalue problem for the symmetric matrix $\boldsymbol{L}^{-1}\boldsymbol{A}\boldsymbol{L}^{-T}$. It follows that (5.43), and therefore $\boldsymbol{B}$, has real eigenvalues and a complete set of M-orthogonal eigenvectors $\boldsymbol{v}_n$,

$$\boldsymbol{B}\boldsymbol{v}_n = \lambda_n\boldsymbol{v}_n, \quad \boldsymbol{v}_n^T\boldsymbol{M}\boldsymbol{v}_m = \delta_{nm}, \quad n, m = 1, 2, \ldots, N. \tag{5.44}$$

From (5.40), one obtains by induction that

$$\boldsymbol{z}_k = p_k(\boldsymbol{B})\boldsymbol{z}_0, \quad k = 0, 1, 2, \ldots, \tag{5.45}$$

where p_k are polynomials of degree k satisfying

$$\begin{aligned} p_{k+1}(\lambda) &= \omega_{k+1}\lambda p_k(\lambda) + (1 - \omega_{k+1})p_{k-1}(\lambda), \\ p_{-1}(\lambda) &= 0, \quad p_0(\lambda) = 1. \end{aligned} \tag{5.46}$$

With $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N]$ denoting the set of eigenvectors of $\boldsymbol{B}$, one has $\boldsymbol{BV} = \boldsymbol{V\Lambda}$, where $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$, hence $\boldsymbol{V}^T\boldsymbol{BV} = \boldsymbol{\Lambda}$, $\boldsymbol{V}^T p_k(\boldsymbol{B})\boldsymbol{V} = p_k(\boldsymbol{\Lambda})$, and thus

$$p_k(\boldsymbol{B}) = \boldsymbol{V} p_k(\boldsymbol{\Lambda})\boldsymbol{V}^T. \tag{5.47}$$

The speed of convergence $\boldsymbol{z}_k \to \boldsymbol{0}$ in (5.45), therefore, is determined by the absolutely largest of the quantities $p_k(\lambda_n)$, $n = 1, 2, \ldots, N$.

If we expand $\boldsymbol{z}_0$ in the eigenvectors of B,

$$\boldsymbol{z}_0 = \sum_{i=1}^{N} \alpha_i \boldsymbol{v}_i, \tag{5.48}$$

then from (5.45) we get

$$\boldsymbol{z}_k = \sum_{i=1}^{N} \alpha_i p_k(\lambda_i)\boldsymbol{v}_i.$$

By the M-orthonormality (5.44) of the eigenvectors, the M-inner products of the iterates $\boldsymbol{z}_k$ become

$$\begin{aligned}\langle \boldsymbol{z}_n, \boldsymbol{z}_m\rangle_M := \boldsymbol{z}_n^T\boldsymbol{M}\boldsymbol{z}_m &= \sum_{i=1}^{N} \alpha_i p_n(\lambda_i)\boldsymbol{v}_i^T\boldsymbol{M}\sum_{j=1}^{N} \alpha_j p_m(\lambda_j)\boldsymbol{v}_j \\ &= \sum_{i,j=1}^{N} \alpha_i\alpha_j p_n(\lambda_i)p_m(\lambda_j)\boldsymbol{v}_i^T\boldsymbol{M}\boldsymbol{v}_j \\ &= \sum_{i=1}^{N} \alpha_i^2 p_n(\lambda_i)p_m(\lambda_i),\end{aligned}$$

that is,

$$\langle \boldsymbol{z}_n, \boldsymbol{z}_m\rangle_M = \int_{\mathbb{R}} p_n(\lambda)p_m(\lambda)\mathrm{d}\alpha_N(\lambda). \tag{5.49}$$

Here, $\mathrm{d}\alpha_N$ is a discrete measure supported on the eigenvalues λ_i of $\boldsymbol{B}$ and having jumps α_i^2 at λ_i.

Along with the measure $\mathrm{d}\alpha_N$ there come discrete orthogonal polynomials $\{\pi_k\}$,

$$\int_{\mathbb{R}} \pi_n(\lambda)\pi_m(\lambda)\mathrm{d}\alpha_N(\lambda) \begin{cases} = 0 & \text{if } n \neq m, \\ > 0 & \text{if } n = m, \end{cases} \qquad n, m = 0, 1, \ldots, N-1 \tag{5.50}$$

and Jacobi matrices $\boldsymbol{J}_k = \boldsymbol{J}_k(\mathrm{d}\alpha_N)$, $k = 1, 2, \ldots, N$. The extreme eigenvalues of $\boldsymbol{J}_k$, i.e., the extreme zeros of π_k, with increasing k, in general provide good approximations to the extreme eigenvalues of $\boldsymbol{B}$, hence by (5.42), to those of $\boldsymbol{M}^{-1}\boldsymbol{N}$.

In order to generate the matrices $\boldsymbol{J}_k$, one can use the modified Chebyshev algorithm (cf. §3.4.2), defining *modified moments* in terms of the polynomials p_k by

$$m_k = \langle \boldsymbol{z}_k, \boldsymbol{z}_0\rangle_M = \int_{\mathbb{R}} p_k(\lambda)\mathrm{d}\alpha_N(\lambda), \quad k = 0, 1, 2, \ldots. \tag{5.51}$$

The polynomials p_k indeed satisfy a three-term recurrence relation with known coefficients (cf. (5.46)). The first relation in (5.51) is used to compute the modified moments.

ETNA
Kent State University
etna@mcs.kent.edu

While the procedure described requires $2m$ modified moments to obtain $\boldsymbol{J}_m$, that is, $2m$ iterations of (5.38), there are special iterative methods, such as the Chebyshev semi-iterative method, where the same can be accomplished already after m iteration (cf. [28, §3]).

A similar method is developed in [4] to determine a few of the largest singular values of a large sparse matrix and the corresponding left and right singular vectors, and is extended in [5] to estimate complex eigenvalues of a large sparse nonsymmetric matrix in connection with an adaptive Chebyshev iterative method.

Acknowledgments. The author is indebted to Professors L. Reichel and Z. Strakoš for helpful comments on a draft version of this paper.

REFERENCES

[1] Z. Bai and G. H. Golub, *Bounds for the trace of the inverse and the determinant of symmetric positive definite matrices*, in The Heritage of P. L. Chebyshev: A Festschrift in Honor of the 70th Birthday of T. J. Rivlin, C. A. Micchelli, ed., Annals Numer. Math., 4 (1997), pp. 29–38.

[2] Z. Bai, M. Fahey, and G. Golub, *Some large-scale matrix computation problems*, J. Comput. Appl. Math., 74 (1996), pp. 71–89.

[3] M. Benzi and G. H. Golub, *Bounds for the entries of matrix functions with applications to preconditioning*, BIT, 39 (1999), pp. 417–438.

[4] M. Berry and G. Golub, *Estimating the largest singular values of large sparse matrices via modified moments*, Numer. Algorithms, 1 (1991), pp. 353–374.

[5] D. Calvetti, G. H. Golub, and L. Reichel, *An adaptive Chebyshev iterative method for nonsymmetric linear systems based on modified moments*, Numer. Math., 67 (1994), pp. 21–40.

[6] D. Calvetti, G. H. Golub, and L. Reichel, *A computable error bound for matrix functionals*, J. Comput. Appl. Math., 103 (1999), pp. 301–306.

[7] D. Calvetti, G. H. Golub, and L. Reichel, *Estimation of the L-curve via Lanczos bidiagonalization*, BIT, 39 (1999), pp. 603–619.

[8] D. Calvetti, G. H. Golub, W. B. Gragg, and L. Reichel, *Computation of Gauss-Kronrod quadrature rules*, Math. Comp., 69 (2000), 1035–1052.

[9] D. Calvetti, S. Morigi, L. Reichel, and F. Sgallari, *An iterative method with error estimators*, in Numerical Analysis 2000, Vol. 5, Quadrature and Orthogonal Polynomials, W. Gautschi, F. Marcellán, and L. Reichel, eds., J. Comput. Appl. Math., 127 (2001), pp. 93–119.

[10] G. Dahlquist, S. C. Eisenstat and G. H. Golub, *Bounds for the error of linear systems of equations using the theory of moments*, J. Math. Anal. Appl., 37 (1972), pp. 151–166.

[11] G. Dahlquist, G. H. Golub, and S. G. Nash, *Bounds for the error in linear systems*, in Semi-Infinite Programming (Proc. Workshop, Bad Honnef, 1978), R. Hettich, ed., Lecture Notes in Control and Information Sci., Vol. 15, Springer-Verlag, Berlin, 1979, pp. 154–172.

[12] S. Elhay, G. H. Golub, and J. Kautsky, *Updating and downdating of orthogonal polynomials with data fitting applications*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 327–353.

[13] S. Elhay, G. H. Golub, and J. Kautsky, *Jacobi matrices for sums of weight functions*, BIT, 32 (1992), pp. 143–166.

[14] B. Fischer and G. H. Golub, *On generating polynomials which are orthogonal over several intervals*, Math. Comp., 56 (1991), pp. 711–730.

[15] B. Fischer and G. H. Golub, *How to generate unknown orthogonal polynomials out of known orthogonal polynomials*, J. Comput. Appl. Math., 43 (1992), pp. 99–115.

[16] D. Galant, *An implementation of Christoffel's theorem in the theory of orthogonal polynomials*, Math. Comp., 25 (1971), pp. 111–113.

[17] W. Gander and W. Gautschi, *Adaptive quadrature — revisited*, BIT, 40 (2000), pp. 84–101.

[18] F. R. Gantmacher, *The theory of matrices*, Vol. 1, AMS Chelsea Publishing, Providence, RI, 1998. [Translated from the Russian by K. A. Hirsch. Reprint of the 1959 translation.]

[19] W. Gautschi, *A survey of Gauss-Christoffel quadrature formulae*, in E. B. Christoffel (Aachen/Monschau, 1979), P. L. Butzer and F. Fehér, eds., Birkhäuser, Basel, 1981, pp. 72–147.

[20] W. Gautschi, *An algorithmic implementation of the generalized Christoffel theorem*, in Numerical Integration, G. Hämmerlin, ed., Internat. Ser. Numer. Math., Vol. 57, 1982, Birkhäuser, Basel, pp. 89–106.

[21] W. Gautschi, *On generating orthogonal polynomials*, SIAM J. Sci. Statist. Comput., 3 (1982), pp. 289–317.

[22] W. Gautschi and G. V. Milovanović, *s-orthogonality and construction of Gauss-Turán-type quadrature formulae*, J. Comput. Appl. Math., 86 (1997), pp. 205–218.

[23] G. H. Golub, *Some modified matrix eigenvalue problems*, SIAM Rev., 15 (1973), pp. 318–334.

[24] G. H. Golub, *Bounds for matrix moments*, Rocky Mountain J. Math., 4 (1974), pp. 207–211.

ETNA
Kent State University
etna@mcs.kent.edu

[25] G. H. GOLUB, *Matrix computation and the theory of moments*, in Proceedings of the International Congress of Mathematicians (Zürich, 1994), S. D. Chatterji, ed., Vol. 2, Birkhäuser, Basel, 1995, pp. 1440–1448. [See also Bull. Belg. Math. Soc. Simon Stevin 1996, suppl.,1–9.]

[26] G. H. GOLUB AND M. H. GUTKNECHT, *Modified moments for indefinite weight functions*, Numer. Math., 57 (1990), pp. 607–624.

[27] G. H. GOLUB AND J. KAUTSKY, *Calculation of Gauss quadratures with multiple free and fixed knots*, Numer. Math., 41 (1983), pp. 147–163.

[28] G. H. GOLUB AND M. D. KENT, *Estimates of eigenvalues for iterative methods*, Math. Comp., 53 (1989), pp. 619–626.

[29] G. H. GOLUB AND U. VON MATT, *Quadratically constrained least squares and quadratic problems*, Numer. Math., 59 (1991), pp. 561–580.

[30] G. H. GOLUB AND G. MEURANT, *Matrices, moments and quadrature*, in Numerical Analysis 1993 (Dundee, 1993), Pitman Res. Notes Math. Ser., Vol. 303, Longman Sci. Tech., Harlow, 1994.

[31] G. H. GOLUB AND G. MEURANT, *Matrices, moments and quadrature. II. How to compute the norm of the error in iterative methods*, BIT, 37 (1994), pp. 687–705.

[32] G. H. GOLUB AND Z. STRAKOŠ, *Estimates in quadratic formulas*, Numer. Algorithms, 8 (1994), pp. 241–268.

[33] G. H. GOLUB AND J. H. WELSCH, *Calculation of Gauss quadrature rules*, Math. Comp., 23 (1969), pp. 221–230. [Addendum: loose microfiche suppl. A1–A10.]

[34] R. G. GORDON, *Error bounds in equilibrium statistical mechanics*, J. Mathematical Phys., 9 (1968), pp. 655–663.

[35] W. B. GRAGG AND W. J. HARROD, *The numerically stable reconstruction of Jacobi matrices from spectral data*, Numer. Math., 44 (1994), pp. 317–335.

[36] M. R. HESTENES AND E. STIEFEL, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436.

[37] J. KAUTSKY AND S. ELHAY, *Calculation of the weights of interpolatory quadratures*, Numer. Math., 40 (1982), pp. 407–422.

[38] J. KAUTSKY AND G. H. GOLUB, *On the calculation of Jacobi matrices*, Linear Algebra Appl., 52/53 (1983), pp. 439–455.

[39] A. S. KRONROD, *Nodes and Weights of Quadrature Formulas. Sixteen-Place Tables*, Consultants Bureau, New York, 1965. [Authorized translation from the Russian.]

[40] D. P. LAURIE, *Calculation of Gauss-Kronrod quadrature rules*, Math. Comp., 66 (1997), pp. 1133–1145.

[41] A. F. NIKIFOROV, S. K. SUSLOV, AND V. B. UVAROV, *Classical Orthogonal Polynomials of a Discrete Variable*, Springer Series in Computational Physics, Springer-Verlag, Berlin, 1991. [Translated from the Russian.]

[42] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Classics in Applied Mathematics, Vol. 20, SIAM, Philadelphia, PA, 1998. [Corrected reprint of the 1980 original.]

[43] R. PIESSENS, E. DE DONCKER-KAPENGA, C. W. ÜBERBUBER, AND D. K. KAHANER, *QUADPACK. A Subroutine Package for Automatic Integration*, Springer Series in Computational Mathematics, Vol. 1, Springer-Verlag, Berlin, 1983.

[44] E. L. STIEFEL, *Kernel polynomials in linear algebra and their numerical applications*, in Further Contributions to the Solution of Simultaneous Linear Equations and the Determination of Eigenvalues, Nat. Bur. Standards Appl. Math. Ser., Vol. 49, U.S. Government Printing Office, Washington, DC, 1958, pp. 1–22.

[45] H. S. WILF, *Mathematics for the Physical Sciences*, John Wiley, New York, 1962. [Reprinted in 1978 by Dover Publications, New York.]

[46] H. S. WILF, Personal communication, 1980.

29.7. [183] “Leonhard Eulers Umgang mit langsam konvergenten Reihen”

---

[183] “Leonhard Eulers Umgang mit langsam konvergenten Reihen,” *Elem. Math*. **62**, 174–183 (2007).

Elem. Math. 62 (2007) 174 – 183
0013-6018/07/040174-10

Elemente der Mathematik

Leonhard Eulers Umgang mit langsam konvergenten Reihen

Walter Gautschi

1 Das Basler Problem

Eines der brennendsten mathematischen Probleme Anfang des 18. Jahrhunderts, das zwar schon im 17. Jahrhundert von Pietro Mengoli, und auch von John Wallis erwähnt, aber erst durch die fieberhaften, jedoch erfolglosen Anstrengungen der hervorragendsten Gelehrten wie Leibniz, Stirling, de Moivre und allen Bernoullis aktuell geworden ist, bestand darin, die Summe der unendlichen Reihe

$$1+\frac{1}{4}+\frac{1}{9}+\frac{1}{16}+\frac{1}{25}+\cdots \tag{1}$$

durch bekannte Grössen auszudrücken. Ein frustrierter Jakob Bernoulli, damals wohl der geübteste Mathematiker im Umgang mit unendlichen Reihen, stellte das Anliegen [2]: „... sollte jemand das, was unseren Anstrengungen bis jetzt entgangen ist, finden und uns mitteilen, so werden wir ihm sehr dankbar sein". Wohl infolge der grossen diesbezüglichen Bemühungen von Jakob und Johann Bernoulli ist das Problem als „Basler Problem" in die Geschichte der Mathematik eingegangen.

Es ist bekannt, dass Euler schon 1735 das Problem gelöst, und für die fragliche Summe den Wert $\pi^2/6$ angegeben hat (was ihn fast von einem Tag auf den anderen weltberühmt gemacht hat), doch waren dieser Entdeckung – was für Euler typisch ist – numerische Rechnungen vorausgegangen. Diese sind durchaus nicht trivial, da es sich in (1) um eine sehr langsam konvergente Reihe handelt: Für eine Genauigkeit von 10^{-d} braucht man ungefähr 10^d Glieder der Reihe, also für sechs Dezimalstellen eine Million Glieder! Es ist daher interessant zu sehen, wie sich Euler mit dieser Schwierigkeit auseinandergesetzt hat. Wie so oft bei Euler sind aus diesem speziellen Problem Resultate hervorgegangen, die einen sehr allgemeinen und weittragenden Charakter haben. Als Beispiel hat er selbst seine Ideen auf die damals ebenso schwierige Aufgabe angewandt, die sogenannte Eulersche Konstante genau zu berechnen.

2 Eine erste Approximation zur Lösung des Basler Problems

Wir schreiben

$$s = \sum_{\nu=1}^{\infty} \frac{1}{\nu^2}. \tag{2}$$

In §22 von *De summatione innumerabilium progressionum* (Die Summierung einer unendlichen Reihe, E20; OI,14, S. 25–41[1]; eingereicht 1731, veröffentlicht 1738) beginnt Euler mit der Integraldarstellung

$$s = -\int_0^1 \frac{\ln(1-t)}{t}\,\mathrm{d}t,$$

die man leicht durch Taylor Entwicklung von $\ln(1-t)$ und nachfolgende gliedweise Integration bestätigen kann. Mittels der Substitution $t \mapsto 1-t$ kann man auch

$$s = -\int_0^1 \frac{\ln t}{1-t}\,\mathrm{d}t$$

schreiben. Nun zerlegt Euler das letzte Integral in zwei Teile, ein Integral von 0 bis x (mit $0 < x < 1$) und ein Integral von x bis 1, wobei er im letzteren wieder $t \mapsto 1-t$ substituiert. Das gibt

$$s = -\int_0^x \frac{\ln t}{1-t}\,\mathrm{d}t - \int_0^y \frac{\ln(1-t)}{t}\,\mathrm{d}t, \quad y = 1-x.$$

Partielle Integration im ersten Integral und Taylor Entwicklung von $\ln(1-t)$ liefert dann[2]

$$s = \ln x \ln(1-x) + \sum_{\nu=1}^{\infty} \frac{x^\nu + y^\nu}{\nu^2}.$$

Um die Konvergenzgeschwindigkeit der letzten Reihe zu maximieren, nimmt Euler $x = 1/2$, also $y = 1/2$, und erhält

$$s = (\ln 2)^2 + \sum_{\nu=1}^{\infty} \frac{1}{2^{\nu-1}\nu^2}. \tag{3}$$

Wie man sieht, gelang es Euler, einen Faktor $2^{-\nu}$ in die Basler Reihe einzuschmuggeln. Die Reihe in (3) konvergiert daher erheblich schneller als die ursprüngliche Reihe in (2). In der Tat, nimmt man n Glieder der Reihe und bezeichnet die resultierende Approximation von s mit $s^{(n)}$, so hat man das in Tabelle 1 gezeigte Konvergenzverhalten:

[1] Wir fügen den Arbeiten von Euler deren Eneström-Index Zahlen (E-Zahlen) bei, sowie den Band der *Opera omnia*, in dem sie zu finden sind, wo OI,14, z.B. *Opera omnia*, Serie I, Vol. 14 bedeutet. Siehe die Web Seite `http://www.math.dartmouth.edu/~euler` des U.S. Euler Archivs für eine nach den E-Zahlen geordnete kommentierte Liste sämtlicher Werke von Euler.

[2] Hier folgen wir Eulers Vorgehen in §196 der *Institutiones calculi integralis*, Vol. 1, E342, OI,11, und nicht der etwas umständlicheren Herleitung in der zitierten Abhandlung.

n	$s^{(n)}$	Fehler
5	1.643543291695979	1.39×10^{-03}
10	1.644920051673697	1.40×10^{-05}
20	1.644934062865116	3.98×10^{-09}
40	1.644934066848226	8.88×10^{-16}

Tabelle 1: Konvergenzverhalten der Reihe in (3)

Euler benutzt die Formel (3), um s auf sechs Dezimalstellen zu berechnen.

3 Eine zweite Approximation

Der Ausgangspunkt hier ist die bekannte Trapezregel für die Integration einer Funktion f,

$$\int_1^{n+1} f(x)\mathrm{d}x \approx \frac{1}{2}f(1) + f(2) + \cdots + f(n) + \frac{1}{2}f(n+1),$$

die Euler, wie vor ihm schon Gregory, verfeinert, indem er auf der linken Seite die Korrekturglieder

$$\frac{1}{12}[f(n+2) - f(n+1)] - \frac{1}{12}[f(2) - f(1)]$$

hinzufügt. Man erhält so, nach einfacher Umordnung,

$$\sum_{\nu=1}^{n+1} f(\nu) \approx \int_1^{n+1} f(x)\mathrm{d}x + \frac{1}{12}[5f(n+1) + f(n+2)] + \frac{1}{12}[7f(1) - f(2)].$$

Nimmt man an, dass f im Unendlichen verschwindet und ins Unendliche summiert und integriert werden kann, so bekommt man, wenn $n \to \infty$,

$$\sum_{\nu=1}^{\infty} f(\nu) \approx \int_1^{\infty} f(x)\mathrm{d}x + \frac{1}{12}[7f(1) - f(2)]. \tag{4}$$

Mit Bezug auf das Basler Problem hat Euler in §14 von *Methodus universalis serierum convergentium summas quam proxime inveniendi* (Eine allgemeine Methode, Approximationen zu Summen konvergenter Reihen zu finden, E46; OI,14, S. 101–107; eingereicht 1735, veröffentlicht 1741) nun die sehr nützliche Idee, für ein bestimmtes $\nu_0 > 1$ die ersten ν_0 Glieder der Reihe direkt zu summieren,

$$\sum_{\nu=1}^{\nu_0} \frac{1}{\nu^2} = s_0, \tag{5}$$

und dann (4) auf $f(x) = (\nu_0 + x)^{-2}$ anzuwenden. Das gibt

$$s \approx s_0 + \frac{1}{\nu_0 + 1} + \frac{1}{12}\left[\frac{7}{(\nu_0+1)^2} - \frac{1}{(\nu_0+2)^2}\right]. \tag{6}$$

Die Resultate für verschiedene Wahlen von ν_0 sind in Tabelle 2 zusammengestellt:

ν_0	$s \approx$	Fehler
10	1.644919055011046	1.50×10^{-05}
20	1.644932866546282	1.20×10^{-06}
40	1.644933981455983	8.54×10^{-08}
80	1.644934061144287	5.70×10^{-09}
160	1.644934066479512	3.69×10^{-10}

Tabelle 2: Die Approximation (6) in Abhängigkeit von ν_0

Euler wählte $\nu_0 = 10$ und erhielt $s \approx 1.644920$, wo aber die zwei letzten Ziffern 19 statt 20 heissen sollten. Im Vergleich mit der ersten Approximation $s^{(n)}$ von (3) konvergiert diese zweite bedeutend langsamer, enthält aber den Keim einer wesentlich allgemeineren und wirksameren Methode, die im nächsten Abschnitt beschrieben werden soll.

4 Die Euler-Maclaurin Summationsformel

Offensichtlich ging es Euler nicht nur um die Summe aller reziproken Quadrate, sondern viel allgemeiner um irgendeine Funktion f summiert über alle natürlichen Zahlen, $\sum_{\nu=1}^{\infty} f(\nu)$. Dies führte zu einer seiner frühen Glanzleistungen – heute Euler-Maclaurin Formel genannt, weil auch Maclaurin sie sechs Jahre später, unabhängig von Euler, gefunden hat. Euler gibt sie zuerst ohne Beweis in *Methodus generalis summandi progressiones* (Eine allgemeine Methode zur Summierung von Reihen, E25; OI,14, S. 42–72; eingereicht 1732, veröffentlicht 1738) an, und leitet sie in *Inventio summae cuiusque seriei ex dato termino generali* (Bestimmung der Summe einer beliebigen Reihe aus ihrem allgemeinen Term, E47; OI,14, S. 108–123; eingereicht 1735, veröffentlicht 1741) vollständig her. In moderner Schreibweise hat sie die Gestalt

$$\begin{aligned}\frac{1}{2}f(0) + f(1) + \cdots + f(n-1) + \frac{1}{2}f(n) \\ = \int_0^n f(x)\mathrm{d}x + \sum_{\mu=1}^{M} \frac{B_{2\mu}}{(2\mu)!}\left[f^{(2\mu-1)}(n) - f^{(2\mu-1)}(0)\right] + R_M,\end{aligned} \tag{7}$$

wo $B_2, B_4, B_6, \ldots$ die Bernoullischen Zahlen bezeichnen, die Jakob Bernoulli in seiner *Ars conjectandi* eingeführt hat und durch die Entwicklung

$$\frac{z}{e^z - 1} = 1 - \frac{1}{2}z + \sum_{\mu=1}^{\infty} \frac{B_{2\mu}}{(2\mu)!} z^{2\mu}, \quad |z| < 2\pi,$$

definiert sind. Euler gibt nie ein Restglied an, aber es kann hier auf verschiedene Art geschrieben werden, z.B. in der Form (vgl. Stoer und Bulirsch [6, §3.3])

$$R_M = \frac{B_{2M+2}}{(2M+2)!} \sum_{k=0}^{n-1} f^{(2M+2)}(\xi_k), \quad k < \xi_k < k+1. \tag{8}$$

Die Konstanten $B_{2\mu}$ hat Euler rekursiv berechnet und damals noch nicht als Bernoullische Zahlen erkannt.

In (7), (8) wird vorausgesetzt, dass die $(2M+2)$-te Ableitung von f auf $\mathbb{R}_+ = [0,\infty]$ stetig ist. Nimmt man weiterhin an, dass alle Ableitungen von f ungerader Ordnung bis zur Ordnung $2M-1$, und f selbst im Unendlichen verschwinden, und f ins Unendliche integrierbar ist, so folgt aus (7), (8), wenn $n \to \infty$,

$$\sum_{\nu=1}^{\infty} f(\nu) = \int_0^{\infty} f(x)\mathrm{d}x - \frac{1}{2}f(0) - \sum_{\mu=1}^{M} \frac{B_{2\mu}}{(2\mu)!} f^{(2\mu-1)}(0) + R_M, \tag{9}$$

$$R_M = \frac{B_{2M+2}}{(2M+2)!} \sum_{k=0}^{\infty} f^{(2M+2)}(\xi_k), \quad k < \xi_k < k+1. \tag{10}$$

Die unendliche Reihe in (10) konvergiert unter der Voraussetzung, dass $f^{(2M+2)}$ auf $\mathbb{R}_+$ positiv und monoton abnehmend ist, und auch $f^{(2M+1)}$ im Unendlichen verschwindet,

$$f^{(2M+2)}(x) > 0, \quad f^{(2M+3)}(x) < 0, \quad x \in \mathbb{R}_+; \quad f^{(2M+1)}(\infty) = 0.$$

Dann gilt nämlich

$$0 < \sum_{k=0}^{n-1} f^{(2M+2)}(\xi_k) < \sum_{k=0}^{n-1} f^{(2M+2)}(k) = f^{(2M+2)}(0) + \sum_{k=1}^{n-1} f^{(2M+2)}(k)$$

$$< f^{(2M+2)}(0) + \int_0^{n-1} f^{(2M+2)}(x)\mathrm{d}x = f^{(2M+2)}(0) + f^{(2M+1)}(n-1) - f^{(2M+1)}(0),$$

und, weil $f^{(2M+1)}$ monoton nach 0 wächst, für alle n:

$$0 < \sum_{k=0}^{n-1} f^{(2M+2)}(\xi_k) < f^{(2M+2)}(0) - f^{(2M+1)}(0).$$

Insbesondere muss auch $f^{(2M+2)}$ im Unendlichen verschwinden, und man zeigt wie oben, dass die fragliche Reihe das Cauchy-Bolzano Konvergenzkriterium erfüllt.

Es folgt

$$|R_M| < \frac{|B_{2M+2}|}{(2M+2)!}[f^{(2M+2)}(0) - f^{(2M+1)}(0)]. \tag{11}$$

5 Anwendungen

In §§31–32 von E47 wendet Euler die Formel (9) (ohne Restglied) auf das Basler Problem an, und in §§25–26 auch auf die Berechnung der Eulerschen Konstanten.

5.1 Anwendung auf das Basler Problem

Wie schon in (5) summiert Euler die ersten ν_0 (= 10) Glieder der Basler Reihe direkt,

$$s = \sum_{\nu=1}^{\infty} \frac{1}{\nu^2} = s_0 + \sum_{\nu=1}^{\infty} \frac{1}{(\nu_0+\nu)^2} \tag{12}$$

und berechnet die Summe der übrigen Glieder durch Anwendung von (9) auf die Funktion

$$f(x) = \frac{1}{(\nu_0 + x)^2}\,. \tag{13}$$

Diese erfüllt wegen $f^{(m)}(x) = (-1)^m (m+1)!(\nu_0 + x)^{-(m+2)}$ alle in §4 gemachten Voraussetzungen, so dass (9), (11), auf (13) angewandt, Folgendes liefert:

$$\sum_{\nu=1}^{\infty} \frac{1}{(\nu_0+\nu)^2} = \frac{1}{\nu_0} - \frac{1}{2}\frac{1}{\nu_0^2} + \sum_{\mu=1}^{M} \frac{B_{2\mu}}{\nu_0^{2\mu+1}} + R_M, \tag{14}$$

$$|R_M| < \frac{|B_{2M+2}|}{\nu_0^{2M+3}} \left(1 + \frac{2M+3}{\nu_0}\right). \tag{15}$$

Man sieht, dass das Restglied im Absolutbetrag, bis auf den Faktor $(1 + (2M+3)/\nu_0)$, kleiner ist als das erste vernachlässigte Glied der Reihe auf der rechten Seite von (14). Letztes ist ja für alternierende (konvergente) Reihen bekannt; hier allerdings haben wir es mit einer divergenten (asymptotischen) Reihe zu tun. Wegen (vgl. z.B. [1, eq 23.1.15])

$$\frac{2(2M+2)!}{(2\pi)^{2M+2}} < |B_{2M+2}| < \frac{2(2M+2)!}{(2\pi)^{2M+2}} \cdot \frac{1}{1 - 2^{-(2M+1)}}$$

gilt auch

$$|R_M| < \frac{2(2M+2)!}{(2\pi\nu_0)^{2M+2}\,\nu_0} \left(1 + \frac{2M+3}{\nu_0}\right) \Big/ \left(1 - 2^{-(2M+1)}\right), \tag{16}$$

was für grosse M mit (15) praktisch identisch ist.

Beste Genauigkeit erhält man, wenn $M = M_{\text{opt}}$ so gewählt wird, dass die obere Schranke in (16) am kleinsten ist. Mit Eulers Wahl $\nu_0 = 10$ findet man

$$M_{\text{opt}} = 30, \quad |R_{M_{\text{opt}}}| < 1.4966 \times 10^{-26}. \tag{17}$$

Die Euler-Maclaurin Formel (14), zusammen mit (12) für $\nu_0 = 10$, ermöglicht es also, die Basler Reihe s mindestens auf 26 Dezimalstellen genau zu berechnen. In der Tat findet man (mit 50-stelliger Arithmetik) die Approximation

$$s \approx 1.64493\,40668\,48226\,43647\,24151\,6562,$$

mit einem Fehler von 1.030×10^{-27}. Euler hat vermutlich mit $M = 12$ gerechnet (obwohl er (14) nur für $M = 7$ explizit ausschreibt) und so s zu 20 Dezimalstellen genau erhalten. Sehr wahrscheinlich hat dieses genaue Resultat ihm die Identifikation mit $\pi^2/6$ nahegelegt.

5.2 Berechnung der Eulerschen Konstanten

Die Eulersche Konstante ist durch den Grenzwert

$$\gamma = \lim_{n\to\infty} \left(\sum_{\nu=1}^{n} \frac{1}{\nu} - \ln n \right)$$

definiert. Wie zuvor summiert man zunächst die ersten ν_0 ($< n$) Glieder der Reihe direkt,

$$s_0 = \sum_{\nu=1}^{\nu_0} \frac{1}{\nu},$$

und schreibt dann

$$\sum_{\nu=1}^{n} \frac{1}{\nu} = s_0 + s, \quad s = \sum_{\nu=1}^{n-\nu_0} \frac{1}{\nu_0 + \nu}.$$

Auf s kann die Euler-Maclaurin Formel (7) angewandt werden, wo n durch $n - \nu_0$ zu ersetzen ist, und f durch

$$f(x) = \frac{1}{\nu_0 + x}.$$

Man erhält

$$s = \ln n - \ln \nu_0 - \frac{1}{2\nu_0} + \frac{1}{2n} + \sum_{\mu=1}^{M} \frac{B_{2\mu}}{2\mu} \left(-\frac{1}{n^{2\mu}} + \frac{1}{\nu_0^{2\mu}} \right) + R_M, \tag{18}$$

und für den Rest, ähnlich wie in §5.1,

$$|R_M| < \frac{2(2M+2)!}{(2\pi\nu_0)^{2M+2}\nu_0} \left[1 + \frac{\nu_0}{2M+2} \left(1 - \left(\frac{\nu_0}{n} \right)^{2M+2} \right) \right] \Big/ \left(1 - 2^{-(2M+1)} \right).$$

Addiert man $s_0 - \ln n$ auf beiden Seiten von (18), und lässt $n \to \infty$, sowohl in (18) als auch in der Abschätzung des Restglieds, so bekommt man

$$\gamma = s_0 - \ln \nu_0 - \frac{1}{2\nu_0} + \sum_{\mu=1}^{M} \frac{B_{2\mu}}{2\mu} \frac{1}{\nu_0^{2\mu}} + R_M, \tag{19}$$

wo

$$|R_M| < \frac{2(2M+2)!}{(2\pi\nu_0)^{2M+2}\nu_0} \left(1 + \frac{\nu_0}{2M+2} \right) \Big/ \left(1 - 2^{-(2M+1)} \right). \tag{20}$$

Für den optimalen Wert von M erhält man wieder $M_{\text{opt}} = 30$, und

$$|R_{M_{\text{opt}}}| < 2.301 \times 10^{-27}. \tag{21}$$

Euler berechnete γ auf diese Weise, mit $\nu_0 = 10$, zu 16 korrekten Dezimalstellen, wahrscheinlich mit der Wahl $M = 7$, hätte aber mit $M = 30$ mehr als zehn weitere Dezimalzahlen erhalten können, nämlich

$$\gamma = .57721\,56649\,01532\,86060\,65120\,89914,$$

mit einem Fehler von 1.688×10^{-28}.

6 Die Eulersche Reihentransformation

In dem Werk *Institutiones calculi differentialis cum eius usu in analysi finitorum ac doctrina serierum* (Grundlagen des Differentialkalküls mit Anwendungen auf die endliche Analysis und die Lehre der Reihen, E212; OI,10; veröffentlicht 1755) leitet Euler in Part II, Ch. 1: *De transformatione serierum* (Über Reihentransformationen), §3, unter anderem folgende Transformation her,

$$\sum_{\nu=0}^{\infty} a_\nu x^{\nu+1} = \sum_{n=0}^{\infty} \left(\frac{x}{1-x}\right)^{n+1} \Delta^n a_0,$$

wo Δ den Differenzenoperator $\Delta a_\nu = a_{\nu+1} - a_\nu$ bedeutet. Für $x = -1$ geht sie über in

$$\sum_{\nu=0}^{\infty} (-1)^\nu a_\nu = \sum_{n=0}^{\infty} \frac{(-1)^n}{2^{n+1}} \Delta^n a_0, \tag{22}$$

was heute als Eulersche Reihentransformation bekannt ist[3]. Dafür gibt er viele Beispiele, unter anderem auch solche, die divergente Reihen betreffen, z.B. die relativ harmlose Reihe

$$s = 1 - 1 + 1 - 1 + 1 - 1 \pm \cdots,$$

für die $a_\nu = 1$, also $\Delta a_0 = \Delta^2 a_0 = \cdots = 0$, und daher $s = \frac{1}{2}$ ist. Eine waghalsigere Reihe ist

$$s = \sum_{\nu=0}^{\infty} (-1)^\nu (\nu+1)!,$$

für die Euler durch geistreiche Manipulationen $s = .4036524077$ findet. Den exakten Wert kann man durch das Exponentialintegral $E_1(x) = \int_x^\infty e^{-t} dt/t$ ausdrücken[4],

$$s = 1 - eE_1(1) = .4036526376768\ldots,$$

woraus man sieht, dass Euler sich in den letzten vier Ziffern seines Resultats geirrt hat.

Ein klassisches Beispiel (bei Euler in *op. cit.*, §11.I) ist die sehr langsam konvergente Reihe

$$s = \sum_{\nu=0}^{\infty} \frac{(-1)^\nu}{\nu+1} = \ln 2,$$

für welche Eulers Transformation die wesentlich schneller konvergierende Reihe

$$s = \sum_{n=0}^{\infty} \frac{1}{(n+1)2^{n+1}}$$

[3]Laut Otto Spiess [5, §5, Fussnote 1] benutzte Euler diese Transformation bereits 1743 in einem Brief an Goldbach.

[4]Siehe Fussnote 2 des Herausgebers (G. Kowalewski) in *op. cit.*, S. 226.

liefert. Etwas interessanter ist die Leibnizsche Reihe (*ibid.*, §11.II)

$$s = \sum_{\nu=0}^{\infty} \frac{(-1)^\nu}{2\nu+1} = \frac{\pi}{4},$$

für die $a_\nu = 1/(2\nu+1)$ und $\Delta^n a_0 = (-1)^n 2^{2n} n!^2/(2n+1)!$ ist, also

$$s = \sum_{n=0}^{\infty} \frac{2^{n-1} n!^2}{(2n+1)!}.$$

Das allgemeine Glied ist nach der Formel von Sterling für $n \to \infty$ äquivalent mit $\sqrt{\frac{\pi}{2n}}\, 2^{-(n+1)}$, so dass die Konvergenzbeschleunigung hier etwa gleich gross ist wie im vorherigen Beispiel.

Allgemein kann man sagen, dass (22) gültig ist, falls die Reihe auf der linken Seite (die nicht notwendigerweise alternierend, also $a_\nu > 0$, sein muss) konvergiert. Dann konvergiert auch die Reihe auf der rechten Seite, und zwar zum selben Grenzwert, aber nicht notwendigerweise schneller. Man hat Konvergenzbeschleunigung dann, wenn alle $a_\nu > 0$, die Folge $\{a_\nu\}_{\nu=0}^{\infty}$ vollständig monoton, d.h. $(-1)^n \Delta^n a_k > 0$ ist für alle $n, k = 0, 1, 2, \ldots$, und $a_{\nu+1}/a_\nu \geq a > \frac{1}{2}$ gilt. Die Konvergenzbeschleunigung ist in der Tat um so beträchtlicher, je grösser a ist (Knopp [4, Satz 155]; der Operator Δ ist bei Knopp als rückwärtiger Differenzenoperator definiert, also ist er das Negative unseres Operators).

7 Die Lambertsche Reihe

Zum Schluss noch eine kleine Perle aus Eulers Werkzeugkasten für unendliche Reihen, die zwar nichts mit dem Vorhergehenden zu tun hat, aber dennoch einen Einblick in Eulers Einfallsreichtum gestattet. Es handelt sich um die Lambertsche Reihe

$$s(x) = \sum_{\nu=1}^{\infty} \frac{1}{x^\nu - 1}, \quad x > 1, \tag{23}$$

speziell für den Fall, dass $x = 10$, dem Euler im Zusammenhang mit einem missglückten Interpolationsversuch begegnet ist (vgl. [3], wo $s(10) = -S(0)$). Die Reihe tritt an verschiedenen Stellen der Arbeit *Consideratio quarumdam serierum, quae singularibus proprietatibus sunt praeditae* (Betrachtung einiger Reihen, die sich durch spezielle Eigenschaften auszeichnen, E190; OI,14, S. 516–541; eingereicht 1750, veröffentlicht 1753) auf, z.B. in §§28–29. Dort entwickelt Euler jedes Glied der Reihe (23) in eine geometrische Reihe in Potenzen von $1/x$, sammelt dann alle Glieder mit gleicher Potenz und erhält so

$$s(x) = \frac{1}{x} + \frac{2}{x^2} + \frac{2}{x^3} + \frac{3}{x^4} + \frac{2}{x^5} + \frac{4}{x^6} + \frac{2}{x^7} + \frac{4}{x^8} + \frac{3}{x^9} + \cdots.$$

Als Meister im Aufspüren von versteckten regelmässigen Mustern bemerkt Euler nun, dass der Zähler in jedem Bruch genau gleich der Anzahl der Teiler der entsprechenden Potenz

von $1/x$ ist, also z.B. in $4/x^6$ ist 4 gleich der Anzahl der Teiler $1, 2, 3, 6$ von 6. Wenn $x = 10$, kann das Resultat mühelos in Dezimalform hingeschrieben werden, was Euler bis auf 30 Stellen tut:

$$s(10) = .12232\,42434\,26244\,52626\,44283\,44628\ldots .$$

Hier ist die Anzahl der Teiler stets kleiner als 10; wenn sie grösser oder gleich 10 ist, müssen kleine Anpassungen vorgenommen werden. Das ist zum ersten Mal an der 49-ten Dezimalstelle der Fall.

Dank. Für den Vorschlag in Fussnote 2 danke ich dem anonymen Begutachter der Arbeit.

Literatur

[1] Abramowitz, M.; Stegun, I.A.: *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. NBS Appl. Math. Series, vol. 55, Washington, DC, 1964.

[2] Bernoulli, J.: *Positiones arithmeticæ de seriebus infinitis, earumque summa finita*. Basel 1689. [Auch in *Opera Jacobi Bernoulli*, Vol. 1, Geneva 1744, 375–402; bes. 398.]

[3] Gautschi, W.: On Euler's attempt to compute logarithms by interpolation: A commentary to his letter of February 16, 1734 to Daniel Bernoulli. *J. Comp. Appl. Math.*, to appear.

[4] Knopp, K.: *Theorie und Anwendung der unendlichen Reihen*. 4. Aufl., Springer, Berlin 1947.

[5] Spiess, Otto: Die Summe der reziproken Quadratzahlen. *Festschrift zum 60. Geburtstag von Prof. Dr. Andreas Speiser*, 66–86. Orell Füssli, Zürich 1945.

[6] Stoer, J.; Bulirsch, R.: *Introduction to numerical analysis*. Texts in Applied Mathematics, vol. 12, Springer, New York 2002.

Walter Gautschi
Department of Computer Sciences
Purdue University
250 N. University Street
West Lafayette, IN 47907-2066, USA
e-mail: wxg@cs.purdue.edu

29.8. [184] COMMENTARY, BY WALTER GAUTSCHI

---

[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 (2007).

22

COMMENTARY, BY WALTER GAUTSCHI

This group of five papers, especially the first and third, has a distinctly "interdisciplinary" character in the sense that classical analysis problems are recast in terms of, and successfully solved by, techniques of linear algebra and, vice versa, problems that have a linear algebra flavor are approached and solved using tools of classical analysis. A similar intriguing mix of analysis and algebra permeates the remaining three papers.

Calculation of Gauss quadrature rules, by Golub and Welsch [53]

The concern here is with the calculation of the n-point Gaussian quadrature rule

$$\int_a^b \omega(t)f(t)\mathrm{d}t = \sum_{\nu=1}^{n} w_\nu f(\tau_\nu) + R_n(f)$$

for the nonnegative weight function $\omega(t)$ on $[a,b]$, i.e., the calculation of the nodes τ_ν and weights w_ν. The connection of this problem with orthogonal polynomials is classical, thanks to work of Gauss [35], Jacobi [61], Christoffel [22], Stieltjes [86], and others: The Gaussian nodes τ_ν are the zeros of π_n, the nth-degree polynomial orthogonal with respect to the weight function ω, and the Gauss weights w_ν are also expressible, in different ways, in terms of these orthogonal polynomials.

An alternative characterization of the Gauss nodes τ_ν can be derived from the classical fact that the ortho*normal* polynomials $\{\tilde{\pi}_k\}$ satisfy a three-term recurrence relation

$$\sqrt{\beta_{k+1}}\tilde{\pi}_{k+1}(t) = (t-\alpha_k)\tilde{\pi}_k(t) - \sqrt{\beta_k}\tilde{\pi}_{k-1}(t), \quad k = 0,1,2,\ldots,$$
$$\tilde{\pi}_{-1} = 0, \quad \tilde{\pi}_0 = \mu_0^{-1/2},$$

with certain real, resp. positive coefficients α_k, β_k which depend on the weight function ω, and $\mu_0 = \int_a^b \omega(t)\mathrm{d}t$. If $\tilde{\boldsymbol{\pi}}(t) = [\tilde{\pi}_0(t), \tilde{\pi}_1(t), \ldots, \tilde{\pi}_{n-1}(t)]^{\mathrm{T}}$, then indeed,

$$t\tilde{\boldsymbol{\pi}}(t) = \boldsymbol{J}\tilde{\boldsymbol{\pi}}(t) + \sqrt{\beta_n}\tilde{\pi}_n(t)\boldsymbol{e}_n, \quad \boldsymbol{e}_n = [0,0,\ldots,1]^{\mathrm{T}},$$

where $\boldsymbol{J} = \boldsymbol{J}_n$ is the Jacobi matrix of order n for the weight function ω, i.e., the symmetric, tridiagonal matrix having the α_k, $k = 0,1,\ldots,n-1$, on the diagonal,

and the $\sqrt{\beta_k}$, $k = 1, \ldots, n-1$, on the two side diagonals. There follows, for $t = \tau_\nu$, since $\tilde{\pi}_n(\tau_\nu) = 0$,

$$\boldsymbol{J}\tilde{\boldsymbol{\pi}}(\tau_\nu) = \tau_\nu \tilde{\boldsymbol{\pi}}(\tau_\nu), \quad \nu = 1, 2, \ldots, n,$$

so that τ_ν are the eigenvalues of $\boldsymbol{J}$ and $\tilde{\boldsymbol{\pi}}(\tau_\nu)$ corresponding eigenvectors. This is the first important mathematical ingredient of the present paper. The other is an expression for the Gaussian weights,

$$w_\nu = \frac{1}{\tilde{\boldsymbol{\pi}}^{\mathrm{T}}(\tau_\nu)\tilde{\boldsymbol{\pi}}(\tau_\nu)}, \quad \nu = 1, 2, \ldots, n, \tag{22.1}$$

for which the authors refer to Wilf (apparently to [93, Section 2.9, eqn (69) or Ch. 2, Exercise 9]). The formula, however, is older; see Szegö [89, eqn (3.4.8)], where it is attributed to Shohat [85]. The authors re-express this formula in terms of the eigenvectors $\boldsymbol{q}_\nu$ normalized by $\boldsymbol{q}_\nu^{\mathrm{T}}\boldsymbol{q}_\nu = 1$, i.e., in terms of

$$\boldsymbol{q}_\nu = \frac{\tilde{\boldsymbol{\pi}}(\tau_\nu)}{[\tilde{\boldsymbol{\pi}}^{\mathrm{T}}(\tau_\nu)\tilde{\boldsymbol{\pi}}(\tau_\nu)]^{1/2}} = \tilde{\boldsymbol{\pi}}(\tau_\nu) w_\nu^{1/2},$$

by noting that the first component of $\tilde{\boldsymbol{\pi}}(\tau_\nu)$ is $\mu_0^{-1/2}$, hence

$$w_\nu = \mu_0 \boldsymbol{q}_{\nu,1}^2, \quad \nu = 1, 2, \ldots, n,$$

where $\boldsymbol{q}_{\nu,1}$ is the first component of $\boldsymbol{q}_\nu$.

There is a detailed discussion in the paper of how Francis's QR algorithm with appropriate shifts can be adapted to compute the eigenvalues of a symmetric, tridiagonal matrix (the matrix $\boldsymbol{J}$) and the first components of the normalized eigenvectors. Related software in Algol is provided in the microfiche supplement of the paper.

Interestingly, the same eigenvalue/vector characterization of Gauss rules, and even the same numerical method (QR algorithm), have been suggested a year earlier in the physics literature by Gordon [54, eqn (26) and p. 660]. This work has had considerable impact in the physical sciences and engineering, whereas the work of Golub and Welsch has had a wider impact in the areas of computational mathematics and information science. Both works have actually been submitted for publication less than a month apart, the former on October 20, the latter on November 13 of 1967. Rarely have two important and overlapping works, like these, popped up simultaneously in two entirely different venues!

Similar ideas have since been developed for other quadrature rules of Gaussian type. Indeed, Golub himself [45] was the first to derive eigenvalue/vector algorithms for Gauss–Radau and Gauss–Lobatto formulae. Laurie [65] did it for his anti-Gaussian formulae, and Calvetti and Reichel [19] for a symmetric modification thereof. Quadrature rules involving derivative terms of arbitrary orders on the boundary or outside the interval of integration require first the generation of the appropriate Jacobi matrix before the (simple) internal nodes can be calculated from its eigenvalues and the corresponding weights from the

associated eigenvectors; see Golub and Kautsky [47, Section 6] and also Ezzirani and Guessab [32]. This has led to important work on the stable calculation of general interpolatory quadratures (Kautsky and Elhay [63], Elhay and Kautsky [31]). A rather substantial extension is the one to Gauss–Kronrod quadratures due to Laurie [66] (see also the commentary to the last paper). For other types of extended quadrature formulae, see Gout and Guessab [55]. Golub-Welsch type algorithms have been developed also for quadrature rules in the complex plane, for example Gauss-Szegö type formulae on the unit circle (Gragg [57, abstract], [56], Watkins [91, pp. 465–466], Jagels and Reichel [62]), Gauss quadrature on the semicircle (Gautschi and Milovanović [43]), Gauss formulae for the Jacobi weight function with complex parameters (Nuttal and Wherry [78]), or those used to approximate the Bromwich integral in the theory of Laplace transform inversion (Luvison [68], Piessens [79]), and complex Gauss formulae for weighted line integrals in the complex plane (Saylor and Smolarski [84, Section 6]).

There are instances in the area of orthogonal polynomials and quadrature where eigenvalues of more general matrices are of interest, for example banded lower Hessenberg matrices in the case of multiple orthogonal polynomials and related quadrature rules (Coussement and Van Assche [24], Borges [11]), or full-blown upper Hessenberg matrices for zeros of Sobolev orthogonal polynomials (Gautschi and Zhang [44, p. 161]) and also for the Gauss–Szegö quadrature rules mentioned above.

Any advances in improving the QR algorithm for computing eigenvalues and eigenvectors of a symmetric tridiagonal matrix give rise immediately to improved Golub-Welsch algorithms. Some possibilities in this regard are discussed by Laurie [67, Section 2]; for positive definite Jacobi matrices, see also Laurie [67, Section 5] and the references therein.

There still remains, of course, the problem of computing the recurrence coefficients α_k, β_k, if not known explicitly, given the weight function ω. This problem is addressed in Section 4 of the paper, where an algorithm of V.I. Mysovskih is described, which computes these coefficients by a Cholesky decomposition of the Hankel matrix in the moments $\mu_r = \int_a^b t^r \omega(t)\mathrm{d}t$ of the weight function. Any method based on moments, however, is notoriously unstable, owing to severe ill-conditioning (for large n) of the underlying moment map. This was first shown in 1968 by the writer [36]; see also [42, Sections 2.1.4, 2.1.6]. Shortly thereafter, Sack and Donovan, in a technical report [82], introduced the idea of "generalized moments" $m_r = \int_a^b p_r(t)\omega(t)\mathrm{d}t$, where p_r is a polynomial of exact degree r, which, at the suggestion of this writer, they renamed "modified moments" in their formal publication [83]. Under the assumption that the polynomials p_r also satisfy a three-term recurrence relation, but with known coefficients, Sack and Donovan developed an algorithm, later given a more definitive form by Wheeler [92], which computes the desired recurrence coefficients α_k, β_k directly from the modified moments. Wheeler suspected that Chebyshev might already have done something of this nature, which was confirmed by the writer and pinpointed to Chebyshev's 1859 memoir [21], where Wheeler's algorithm indeed appears at the

end of Section 3 in the special case of ordinary moments ($p_r(t) = t^r$) and discrete orthogonal polynomials. The algorithm for ordinary, resp. modified moments was therefore named in [37] the Chebyshev, resp. modified Chebyshev algorithm. The latter is not only more efficient than Mysovskih's algorithm, having complexity $O(n^2)$ instead of $O(n^3)$, but is often also more stable. The condition of the underlying modified moment map has been studied in [37, Section 3.3] and [38]; see also [42, Sections 2.1.5, 2.1.6]. For alternative techniques of computing α_k, β_k, based on discretization, see [42, Section 2.2].

Updating and downdating of orthogonal polynomials with data fitting applications, by Elhay, Golub, and Kautsky [30]

The use, in data fitting applications, of (what today are called) discrete orthogonal polynomials can be traced back to a 1859 memoir of Chebyshev [21]. Forsythe [34], a hundred years later and independently, discussed the same procedure and developed it into a viable computer algorithm. The present paper introduces new ideas of updating and downdating in this context, although similar ideas have previously been applied in connection with the related problem of QR factorization of matrices. Mertens [69] reviews downdating algorithms in statistical applications and in the least squares context attributes the concept of downdating to Legendre and Gauss, the originators of least squares theory.

The problem of data fitting is here understood to be the following weighted least squares problem: Given a set $S_N = \{x_j, y_j, w_j^2\}_{j=1}^N$ of N data points $\{x_j, y_j\}$ and positive weights $\{w_j^2\}$, find the polynomial $\hat{q}_n \in \mathbb{P}_n$ of degree $\leq n$ $(< N)$ such that

$$\sum_{j=1}^{N} w_j^2 [y_j - \hat{q}_n(x_j)]^2 \leq \sum_{j=1}^{N} w_j^2 [y_j - q(x_j)]^2 \quad \text{for all } q \in \mathbb{P}_n.$$

The inner product and norm naturally associated with this problem are

$$[u, v]_N = \sum_{j=1}^{N} w_j^2 u(x_j) v(x_j), \quad \|u\|_N = \sqrt{[u, u]_N},$$

in terms of which the least squares problem is simply $\|y - \hat{q}\|_N^2 \leq \|y - q\|_N^2$, all $q \in \mathbb{P}_n$. The solution is most conveniently expressed in terms of the polynomials $\{\pi_k\}_{k=0}^{N-1}$ orthonormal with respect to the inner product $[\cdot, \cdot]_N$ (the "discrete orthogonal polynomials"), namely as the nth-degree "Fourier polynomial" of y,

$$\hat{q}_n(x) = \sum_{j=0}^{n} c_j \pi_j(x), \quad c_j = [\pi_j, y]_N.$$

With regard to the least squares problem, updating means the following: Determine the solution $\hat{q}_n$ corresponding to the enlarged set $S_{N+1} = S_N \cup \{x_{N+1}, y_{N+1}, w_{N+1}^2\}$ in terms of the solution $\hat{q}_n$ corresponding to the original

set S_N. Downdating, conversely, means the determination of $\hat{q}_n$ for S_N in terms of $\hat{q}_n$ for S_{N+1}.

There is a similar problem of up- and downdating for the orthogonal polynomials, more precisely for their Jacobi matrices $\boldsymbol{J}_n$ (cf. the commentary to the first paper): Knowing $\boldsymbol{J}_n$ for S_N, find $\boldsymbol{J}_n$ for S_{N+1}, and vice versa. An algorithm of Gragg and Harrod [58, Section 3] using a sequence of Givens similarity transformations, attributed essentially to Rutishauser [81], can be thought of as an updating procedure in this sense, since it introduces one data point and weight at a time.

As one would expect from the authors, both problems of up- and downdating are solved (in several different ways) by reformulating them in terms of matrices and then applying appropriate techniques of numerical linear algebra.

An application of the updating procedure for Jacobi matrices is made in [29] to generate Jacobi matrices for sums of weight functions.

Up- and downdating algorithms have subsequently been developed for least squares problems in the complex plane, for general complex nodes, for example, in [12, Section 4], and for nodes on the unit circle in [80, Section 3], [2]. For an updating procedure in connection with orthogonal rational functions, and function vectors, having prescribed poles, see [90, Section 3] and [27, Section 5].

Matrices, moments and quadrature, by Golub and Meurant [48]

One of the central themes here is the estimation of matrix functionals $\varphi(\boldsymbol{A}) = \boldsymbol{u}^{\mathrm{T}} f(\boldsymbol{A})\boldsymbol{v}$, where $\boldsymbol{A}$ is a symmetric (usually positive definite) matrix, f a smooth function for which $f(\boldsymbol{A})$ is meaningful, and $\boldsymbol{u}$, $\boldsymbol{v}$ are column vectors. A prototype example, and one given the most attention in this work, is estimating the (i,j)-entry of the inverse matrix $\boldsymbol{A}^{-1}$, in which case $f(t) = t^{-1}$ and $\boldsymbol{u} = \boldsymbol{e}_i$, $\boldsymbol{v} = \boldsymbol{e}_j$ are coordinate vectors. The problem has been treated previously by physicists in connection with the estimation of resolvents, where $\boldsymbol{A} = z\boldsymbol{I} - \boldsymbol{H}$, z is an energy, and $\boldsymbol{H}$ a Hamiltonian, thus $\boldsymbol{A}^{-1}$ is the resolvent of $\boldsymbol{H}$. Much related work can also be found in the quantum chemistry literature; see, e.g., [51, Introduction] and the examples and references given therein.

There are three basic steps in solving the problem: (i) The functional is written as an integral, $\varphi(\boldsymbol{A}) = \int_a^b f(\lambda)\mathrm{d}\alpha(\lambda)$, where $\mathrm{d}\alpha$ is a discrete measure supported on the spectrum $\sigma(\boldsymbol{A})$ of $\boldsymbol{A}$ and $[a,b]$ an interval containing $\sigma(\boldsymbol{A})$. This is done by a spectral resolution of $\boldsymbol{A}$, and in the important case $\boldsymbol{u} = \boldsymbol{v}$ yields a positive measure $\mathrm{d}\alpha$. (ii) The integral is estimated by quadrature rules, typically Gauss, Gauss–Radau, or Gauss–Lobatto rules. These, with an increasing number of nodes, are capable of providing increasingly sharper upper and lower bounds for the integral, provided the derivatives of f have constant sign on $[a,b]$ (as is the case, for example, when $f(t) = t^{-1}$, $a > 0$) and the measure $\mathrm{d}\alpha$ is positive. Otherwise, they may still yield estimates of increasing quality. (iii) Generating the quadrature rules requires the discrete orthogonal polynomials for $\mathrm{d}\alpha$, in particular the Jacobi matrix $\boldsymbol{J} = \boldsymbol{J}(\mathrm{d}\alpha)$ of the measure $\mathrm{d}\alpha$ (cf. the commentary

to the first paper), which can be obtained by the Lanczos or the conjugate gradient algorithm. An interesting technical detail is the way the quadrature sums are expressed in terms of the $(1,1)$-element of $f(\boldsymbol{J}^0)$, where $\boldsymbol{J}^0$ is closely related (equal, in the case of Gauss formulae) to the Jacobi matrix $\boldsymbol{J}$ or a leading principal minor matrix thereof.

It is possible to generalize these ideas to the "block" case, where $\boldsymbol{u}$ and $\boldsymbol{v}$ are replaced by an $n \times m$ matrix $\boldsymbol{W}$ (typically with $m = 2$), in which case $d\alpha$ becomes a matrix-valued measure and one has to deal with matrix-valued orthogonal polynomials and quadrature rules, as is done in Sections 3.3 and 4.3 of the present work.

When $f(t) = t^s$ is any power, not necessarily $s = -1$, and $\boldsymbol{u} = \boldsymbol{v}$, the procedure has previously been described by Golub in [46], and in the case $s = -2$, of interest in ℓ_2 error bounds for systems of linear equations, even before by Dahlquist *et al.* in [25] and also in [26, Section 3]. In the latter work, improved approximations are obtained by the conjugate gradient method and the respective errors estimated as described. In a sequel [49] to the present work, and already in [51, Section 4], the case $s = -1$ is further applied to obtain error bounds and stopping criteria in iterative methods for solving linear systems, notably the conjugate gradient method; see also [70], [33], and for the preconditioned conjugate gradient method, [71], [9]. Applications to constructing preconditioners can be found in [10]. Similar ideas have been pursued by M. Arioli and coworkers in a variety of application areas involving partial differential equations and their discretizations ([4], [8] , [6], [7], [3], [5]). A valuable exposition of error estimates in the conjugate gradient method is [88], where some of the recent results are traced back to the original work of Hestenes and Stiefel [59], and the influence of rounding errors is given serious attention. For the latter, see also [51, Section 5], [87, Section 4], and [94]. For a recent comprehensive review of these and related matters, see [72], especially Sections 3.3 and 5.3.

Altogether different applications of the work of Golub and Meurant are to highly ill-conditioned linear algebraic systems, specifically to the determination of the Tikhonov regularization parameter [14], [15], [20], or to the determination of upper and lower bounds for the Lagrange multipliers in constrained least squares and quadratic problems [52]. The blur identification problem in image processing [76, Section 6] contains yet another application.

The work of Golub and Meurant has inspired other researchers to develop variants of their techniques for estimating matrix functionals. We mention, for example, Calvetti *et al.* [17], [18], where next to Gauss and Gauss–Radau quadratures also anti-Gauss formulae are used (see the commentary to the first paper) and Calvetti *et al.* [16], where functionals $\boldsymbol{u}^{\mathrm{T}}[f(\boldsymbol{A})]^{\mathrm{T}} g(\boldsymbol{A})\boldsymbol{u}$ are estimated for matrices $\boldsymbol{A}$ that are no longer necessarily symmetric, and the quadrature and antiquadrature rules are therefore based on the Arnoldi rather than the Lanczos process.

A stable numerical method for inverting shape from moments, by Golub, Milanfar, and Varah [50]

The basic problem here is the determination of an n-sided polygon P in the complex plane, having vertices z_j, $j = 1, 2, \ldots, n$, given its first $2n-2$ "harmonic" moments $c_k = \int\int_P z^k \mathrm{d}x\mathrm{d}y$, $k = 0, 1, \ldots, 2n-3$. If the associated "complex" moments are defined by $\tau_0 = \tau_1 = 0$, $\tau_k = k(k-1)c_{k-2}$, $k = 2, 3, \ldots, 2n-1$, the vertex reconstruction amounts to solving the system of $2n$ nonlinear equations

$$\sum_{j=1}^{n} a_j z_j^k = \tau_k, \quad k = 0, 1, \ldots, 2n-1.$$

These are formally identical with the equations for a Gaussian quadrature formula (with nodes z_j, weights a_j, and moments τ_k of the underlying weight function), except that all these quantities are now complex and, moreover, the first two moments vanish. While the classical Prony's method is still applicable (it determines the coefficients of the monic polynomial of degree n having the z_j as its zeros), it is notoriously unstable. The object of this work is to develop a solution procedure which, though not necessarily perfectly stable, is more stable than Prony's method.

This is done essentially by reformulating the problem, implicit already in [89, eqn (2.2.9)], as a generalized eigenvalue problem involving two Hankel matrices in the moments, or better, in transformed moments obtained by appropriate scaling and shifting.

In practice, the number n of vertices is usually not known a priori and must be estimated from the given sequence of moments, which, to complicate matters, may be corrupted by noise.

There are a number of potential application areas for procedures as here described, one, discussed previously, to tomographic reconstruction, and another, described in the present work, to the problem of geophysical inversion from gravimetric measurements.

The theoretical results of sensitivity analysis are nicely corroborated by numerical examples. There remain, however, a number of issues for further study, for example, a sound statistical analysis of procedures for estimating the number of vertices, especially in the presence of noise, and the incorporation of a priori geometrical constraints. Some of these issues have been taken up in the more recent work [28].

Computation of Gauss–Kronrod quadrature rules, by Calvetti, Golub, Gragg, and Reichel [13]

In order to economically estimate the error $R_n(f)$ of the n-point Gauss quadrature rule (cf. the commentary to the first paper), Kronrod [64] in 1964 constructed (for the weight function $\omega = 1$ on $[-1, 1]$) an extended Gauss formula

$$\int_a^b \omega(t)f(t)\mathrm{d}t = \sum_{\nu=1}^{n} \lambda_\nu^K f(\tau_\nu) + \sum_{\mu=1}^{n+1} \lambda_\mu^{*K} f(\tau_\mu^K) + R_n^K(f),$$

now called the Gauss–Kronrod quadrature formula, by adjoining to the n Gauss nodes τ_ν additional $n+1$ nodes τ_μ^K – the Kronrod nodes – and selecting them, and all weights λ_ν^K, λ_μ^{*K}, such as to achieve maximum degree of exactness $3n+1$ (at least). The same idea, in a germinal form, can be traced back to the late 19th century (cf. [40]). It turns out that the Kronrod nodes must be the zeros of the polynomial π_{n+1}^K of degree $n+1$ orthogonal to all lower-degree polynomials with respect to the (sign-changing) weight function $\omega(t)\pi_n(t;\omega)$ on (a,b), where π_n is the orthogonal polynomial of degree n relative to the weight function ω. While the polynomial π_{n+1}^K (considered for $\omega = 1$ already by Stieltjes in 1894 without reference to quadrature) always exists uniquely, its zeros may or may not all be real and contained in $[a,b]$. An extensive literature thus evolved dealing precisely with this question of reality, and also with the question of positivity of all weights λ_ν^K, λ_μ^{*K}. (For surveys on this and other aspects of Gauss–Kronrod formulae, see Monegato [74], [75], Gautschi [39], and Notaris [77].) In comparison, the question of actually computing the Gauss–Kronrod formula, when it exists, i.e., computing its nodes and weights, has received less attention; see, however, the recent survey by Monegato [73].

Among the most remarkable computational advances in this area is the algorithm of Laurie [66] for computing positive Gauss–Kronrod formulae. Laurie recognizes the equivalence of this problem with an inverse eigenvalue problem for a symmetric tridiagonal matrix with prescribed entries on the side diagonal; see also [23, pp. 15–16]. His algorithm much resembles the Golub-Welsch algorithm (cf. the commentary to the first paper) for ordinary Gauss formulae. In the present work by Calvetti *et al.*, this algorithm is modified and simplified in the sense that the Gauss nodes τ_ν need not be recomputed (as they are in Laurie's algorithm) in cases where they are already known. Indeed, not even the full tridiagonal Jacobi–Kronrod matrix of order $2n+1$ needs to be generated. The resulting new algorithm is then used by the authors to compute also Kronrod extensions of Gauss–Radau and Gauss–Lobatto formulae.

Modifications required to deal with nonpositive Gauss–Kronrod rules are developed in [1].

The work is too recent to have had a major impact, but it can be expected to find many applications, most likely in the area of adaptive quadrature. One such application (to the motion of droplets) is already briefly mentioned in [60, p. 63].

Summary

Golub's work described here is characterized, on the one hand, by the imaginative use of linear algebra techniques in problems originating elsewhere, and on the other hand, by bringing tools outside of linear algebra to bear on problems

involving matrices. Both these features of Golub's work are elaborated in greater detail in the recent essay [41].

Acknowledgment. The writer is grateful for comments by D.P. Laurie, L. Reichel, and Z. Strakoš on an earlier draft of these commentaries.

REFERENCES

1. G. S. Ammar, D. Calvetti, and L. Reichel, Computation of Gauss–Kronrod quadrature rules with non-positive weights. *Electron. Trans. Numer. Anal.*, **9**, 26–38 (1999).
2. G. S. Ammar, W. B. Gragg, and L. Reichel, Downdating of Szegö polynomials and data-fitting applications. *Linear Algebra Appl.*, **172**, 315–336 (1992).
3. M. Arioli, A stopping criterion for the conjugate gradient algorithm in a finite element method framework. *Numer. Math.*, **97**(1), 1–24 (2004).
4. Mario Arioli and Lucia Baldini, A backward error analysis of a null space algorithm in sparse quadratic programming. *SIAM J. Matrix Anal. Appl.*, **23**(2), 425–442 (2001).
5. M. Arioli, D. Loghin, and A. J. Wathen, Stopping criteria for iterations in finite element methods. *Numer. Math.*, **99**(3), 381–410 (2005).
6. Mario Arioli and Gianmarco Manzini, A null space algorithm for mixed finite-element approximations of Darcy's equation. *Comm. Numer. Meth. Engrg.*, **18**(9), 645–657 (2002).
7. Mario Arioli and Gianmarco Manzini, Null space algorithm and spanning trees in solving Darcy's equation. *BIT Numer. Math.*, **43**(Suppl.), 839–848 (2003).
8. M. Arioli, E. Nonlard, and A. Russo, Stopping criteria for iterative methods: applications to PDE's. *Calcolo*, **38**(2), 97–112 (2001).
9. Owe Axelsson and Igor Kaporin, Error norm estimation and stopping criteria in preconditioned conjugate gradient iterations. *Numer. Linear Algebra Appl.*, **8**(4), 265–286 (2001).
10. Michele Benzi and Gene H. Golub, Bounds for the entries of matrix functions with applications to preconditionings. *BIT*, **39**(3), 417–438 (1999).
11. Carlos F. Borges, On a class of Gauss-like quadrature rules. *Numer. Math.*, **67**(3), 271–288 (1994).
12. Adhemar Bultheel and Marc Van Barel, Vector orthogonal polynomials and least squares approximation. *SIAM J. Matrix Anal. Appl.*, **16**(3), 863–885 (1995).
13*. D. Calvetti, G. H. Golub, W. B. Gragg, and L. Reichel, Computation of Gauss–Kronrod quadrature rules. *Math. Comp.*, **69**(231), 1035–1052 (2000).
14. D. Calvetti, G. H. Golub, and L. Reichel, Estimation of the L-curve via Lanczos bidiagonalization. *BIT*, **39**(4), 603–619 (1999).
15. D. Calvetti, P. C. Hansen, and L. Reichel, L-curve curvature bounds via Lanczos bidiagonalization. *Electron. Trans. Numer. Anal.*, **14**, 20–35 (2002).
16. Daniela Calvetti, Sun-Mi Kim, and Lothar Reichel, Quadrature rules based on the Arnoldi process. *SIAM J. Matrix Anal. Appl.*, **26**(3), 765–781 (2005).
17. D. Calvetti, S. Morigi, L. Reichel, and F. Sgallari, Computable error bounds and estimates for the conjugate gradient method. *Numer. Algorithms*, **25**(1–4), 75–88 (2000).
18. D. Calvetti, S. Morigi, L. Reichel, and F. Sgallari, An iteration method with error estimators. *J. Comput. Appl. Math.*, **127**(1–2), 93–119 (2001).
19. Daniela Calvetti and Lothar Reichel, Symmetric Gauss–Lobatto and modified anti-Gauss rules. *BIT*, **43**(3), 541–554 (2003).

20. Daniela Calvetti and Lothar Reichel, Tikhonov regularization of large linear problems. *BIT*, **43**(2), 263–283 (2003).
21. P. L. Chebyshev, Sur l'interpolation par la méthode des moindres carrés. *Mém. Acad. Impér. Sci. St. Petersbourg*, (7) **1**(15), 1–24 (1859). [Also in *Œuvres* I, pp. 473–498.]
22. E. B. Christoffel, Sur une classe particulière de fonctions entières et de fractions continues. *Ann. Mat. Pura Appl.*, (2) **8**, 1–10 (1877). [Also in *Ges. Math. Abhandlungen* I, 65–87.]
23. Moody T. Chu and Gene H. Golub, Structured inverse eigenvalue problems, in *Acta Numerica 2002*, Vol. 11, pp. 1–71, Cambridge University Press, Cambridge (2002).
24. Jonathan Coussement and Walter Van Assche, Gaussian quadrature for multiple orthogonal polynomials. *J. Comput. Appl. Math.*, **178**(1–2), 131–145 (2005).
25. Germund Dahlquist, Stanley C. Eisenstat, and Gene H. Golub, Bounds for the error of linear systems of equations using the theory of moments. *J. Math. Anal. Appl.*, **37**(4), 151–166 (1972).
26. Germund Dahlquist, Gene H. Golub, and Stephen G. Nash, Bounds for the error in linear systems, *Lecture Notes in Control and Information Sci.* 15, pp. 154–172, Springer, Berlin (1979).
27. Steven Delvaux and Marc Van Barel, Orthonormal rational function vectors. *Numer. Math.*, **100**(3), 409–440 (2005).
28. Michael Elad, Peyman Milanfar, and Gene H. Golub, Shape from moments—an estimation theory perspective. *IEEE Trans. Signal Process.*, **52**(7), 1814–1829 (2004).
29. Sylvan Elhay, Gene H. Golub, and Jaroslav Kautsky, Jacobi matrices for sums of weight functions. *BIT*, **32**(1), 143–166 (1992).
30. Sylvan Elhay, Gene H. Golub, and Jaroslav Kautsky, Updating and downdating of orthogonal polynomials with data fitting applications. *SIAM J. Matrix Anal. Appl.*, **12**(2), 327–353 (1991).
31. Sylvan Elhay and Jaroslav Kautsky, Algorithm 655—IQPACK: FORTRAN subroutines for the weights of interpolatory quadratures. *ACM Trans. Math. Software*, **13**(4), 399–415 (1987).
32. Abdelkrim Ezzirani and Allal Guessab, A fast algorithm for Gaussian type quadrature formulae with mixed boundary conditions and some lumped mass spectral approximations. *Math. Comp.*, **68**(225), 217–248 (1999).
33. Bernd Fischer and Gene H. Golub, On the error computation for polynomial based iteration methods, in *Recent advances in iterative methods*, IMA Vol. Math. Appl. 60, Springer, New York, pp. 59–67 (1994).
34. G. E. Forsythe, Generation and use of orthogonal polynomials for data-fitting with a digital computer. *J. Soc. Indust. Appl. Math.*, **5**(2), 74–88 (1957).
35. C. F. Gauss, Methodus nova integralium valores per approximationem inveniendi. *Commentationes Societatis Regiae Scientarium Gottingensis Recentiores*, **3** (1814). [Also in *Werke* III, 163–196.]
36. Walter Gautschi, Construction of Gauss–Christoffel quadrature formulas. *Math. Comp.*, **22**(102), 251–270 (1968).
37. Walter Gautschi, On generating orthogonal polynomials. *SIAM J. Sci. Statist. Comput.*, **3**(3), 289–317 (1982).

38. Walter Gautschi, On the sensitivity of orthogonal polynomials to perturbations in the moments. *Numer. Math.*, **48**(4), 369–382 (1986).
39. Walter Gautschi, Gauss-Kronrod quadrature—a survey, In *Numerical Methods and Approximation Theory III*, G. V. Milovanović (ed.), University of Niš, Niš, pp. 39–66 (1988).
40. Walter Gautschi, A historical note on Gauss–Kronrod quadrature. *Numer. Math.*, **100**(3), 483–484 (2005).
41. Walter Gautschi, The interplay between classical analysis and (numerical) linear algebra — a tribute to Gene H. Golub. *Electron. Trans. Numer. Anal.*, **13**, 119–147 (2002).
42. Walter Gautschi, *Orthogonal polynomials: computation and approximation*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford (2004).
43. Walter Gautschi and Gradimir V. Milovanović, Polynomials orthogonal on the semicircle. *J. Approx. Theory*, **46**(3), 230–250 (1986).
44. Walter Gautschi and Minda Zhang, Computing orthogonal polynomials in Sobolev spaces. *Numer. Math.*, **71**(2), 159–183 (1995).
45*. Gene H. Golub, Some modified matrix eigenvalue problems. *SIAM Rev.*, **15**(2), 318–334 (1973).
46. Gene H. Golub, Bounds for matrix moments. *Rocky Mountain J. Math.*, **4**(2), 207–211 (1974).
47. G. H. Golub, and J. Kautsky, Calculation of Gauss quadratures with multiple free and mixed knots. *Numer. Math.*, **41**(2), 147–163 (1983).
48*. Gene H. Golub and Gérard Meurant, Matrices, moments and quadrature, in *Numerical Analysis 1993 (Dundee, 1993)*, Pitman Res. Notes Math. Ser. 303, Longman Sci. Tech., Harlow, pp. 105–156 (1994).
49. G. H. Golub and G. Meurant, Matrices, moments and quadrature II; how to compute the norm of the error in iterative methods. *BIT*, **37**(3), 687–705 (1997).
50. Gene H. Golub, Peyman Milanfar, and James Varah, A stable numerical method for inverting shape from moments. *SIAM J. Sci. Comput.*, **21**(4), 1222–1243 (1999/00) (electronic).
51. Gene H. Golub and Zdeněk Strakoš, Estimates in quadratic formulas. *Numer. Algorithms*, **8**, 241–268 (1994).
52. Gene H. Golub and Urs von Matt, Quadratically constrained least squares and quadratic problems. *Numer. Math.*, **59**(6), 561–580 (1991).
53*. Gene H. Golub and John J. Welsch, Calculation of Gauss quadrature rules. *Math. Comp.*, **23**(106), 221–230 (1969). Microfiche suppl. A1–A10.
54. Roy G. Gordon, Error bounds in equilibrium statistical mechanics. *J. Math. Phys.*, **9**(5), 655–663 (1968).
55. C. Gout and A. Guessab, Extended Lagrange interpolation and nonclassical Gauss quadrature formulae. *Math. Comput. Modelling*, **38**(1–2), 209–228 (2003).
56. William B. Gragg, Positive definite Toeplitz matrices, the Arnoldi process for isometric operators, and Gaussian quadrature on the unit circle. *J. Comput. Appl. Math.*, **46**(1–2), 183–198 (1993).
57. William B. Gragg, The QR algorithm for unitary Hessenberg matrices. *J. Comput. Appl. Math.*, **16**(1), 1–8 (1986).

58. William B. Gragg and William J. Harrod, The numerically stable reconstruction of Jacobi matrices from spectral data. *Numer. Math.*, **44**(3), 317–335 (1984).
59. M. R. Hestenes and E. Stiefel, Methods of conjugate gradients for solving linear syatems. *J. Research Nat. Bur. Standards*, **49**, 409–436 (1952).
60. W. Holländer and S. K. Zaripov, Hydrodynamically interacting droplets at small Reynolds numbers. *Internat. J. Multiphase Flow*, **31**(1), 53–68 (2005).
61. C. G. J. Jacobi, Ueber Gauß neue Methode, die Werthe der Integrale näherungsweise zu finden. *J. Reine Angew. Math.*, **1**, 301–308 (1826).
62. Carl Jagels and Lothar Reichel, Szegö-Lobatto quadrature rules. *Electr. Trans. Numer. Anal.*, (to appear).
63. J. Kautsky and S. Elhay, Calculation of the weights of interpolatory quadratures. *Numer. Math.*, **40**(3), 407–422 (1982).
64. A. S. Kronrod, *Nodes and weights of quadrature formulas. Sixteen-place tables* (Russian), Izdat. Nauka, Moscow (1964). [Authorized translation by Consultants Bureau, New York, 1965.]
65. Dirk P. Laurie, Anti-Gaussian quadrature formulas. *Math. Comp.*, **65**(214), 739–747 (1996).
66. Dirk P. Laurie, Calculation of Gauss–Kronrod quadrature rules. *Math. Comp.*, **66**(219), 1133–1145 (1997).
67. Dirk P. Laurie, Computation of Gauss-type quadrature formulas. *J. Comput. Appl. Math.*, **127**(1–2), 201–217 (2001).
68. Angelo Luvison, On the construction of Gaussian quadrature rules for inverting the Laplace transform. *Proc. IEEE*, **62**, 637–638 (1974).
69. B. J. A. Mertens, Downdating: interdisciplinary research between statistics and computing. *Statistica Neerlandica*, **55**(3), 358–366 (2001).
70. Gérard Meurant, The computation of bounds for the norm of the error in the conjugate gradient algorithm. *Numer. Algorithms*, **16**(1), 77–87 (1997).
71. Gérard Meurant, Numerical experiments in computing bounds for the norm of the error in the preconditioned conjugate gradient algorithm. *Numer. Algorithms*, **23**(3–4), 353–365 (1999).
72. Gérard Meurant and Zdeněk Strakoš, The Lanczos and conjugate gradient algorithms in finite precision arithmetic, in *Acta Numerica 2006*, Vol. 15, (pp. 471–542) Cambridge University Press, Cambridge (2006).
73. Giovanni Monegato, An overview of the computational aspects of Kronrod quadrature rules. *Numer. Algorithms*, **26**(2), 173–196 (2001).
74. Giovanni Monegato, An overview of results and questions related to Kronrod schemes. In *Numerische Integration*, G. Hämmerlin (ed.), Internat. Ser. Numer. Math. 45, Birkhäuser, Basel, pp. 231–240 (1979).
75. Giovanni Monegato, Stieltjes polynomials and related quadrature rules. *SIAM Rev.*, **24**(2), 137–158 (1982).
76. Nhat Nguyen, Peyman Milanfar, and Gene Golub, Efficient generalized cross-validation with applications to parametric image restoration and resolution enhancement. *IEE$_E$ Trans. Image Process.*, **10**(9), 1299–1308 (2001).
77. Sotirios E. Notaris, An overview of results on the existence or nonexistence and the error term of Gauss–Kronrod quadrature formulae, in *Approximation and Computation*, R.V.M. Zahar (ed.), Internat. Ser. Numer. Math. 119, Birkhäuser, Boston, pp. 485–496 (1994).

78. J. Nuttall and C. J. Wherry, Gaussian integration for complex weight functions. *J. Inst. Math. Appl.*, **21**(2), 165–170 (1978).
79. Robert Piessens, Comments on: "On the construction of Gaussian quadrature rules for inverting the Laplace transform" by A. Luvison. *Proc. IEEE*, **63**, 817–818 (1975).
80. L. Reichel, G. S. Ammar, and W. B. Gragg, Discrete least squares approximation by trigonometric polynomials. *Math. Comp.*, **57**(195), 273–289 (1991).
81. H. Rutishauser , On Jacobi rotation patterns. In *Experimental Arithmetic, High Speed Computing and Mathematics*, Proc. Sympos. Appl. Math. 15, Amer. Math. Soc., Providence, RI, pp. 219–239 (1963).
82. R. A. Sack and A. F. Donovan, An algorithm for Gaussian quadrature given generalized moments, Technical report, Dept. Math., University of Salford, Salford, UK (1969).
83. R. A. Sack and A. F. Donovan, An algorithm for Gaussian quadrature given modified moments. *Numer. Math.*, **18**(5), 465–478 (1972).
84. Paul E. Saylor and Dennis C. Smolarski, Why Gaussian quadrature in the complex plane?. *Numer. Algorithms*, **26**(3), 251–280 (2001).
85. J. A. Shohat, On a certain formula of mechanical quadratures with non-equidistant ordinates. *Trans. Amer. Math. Soc.*, **31**, 448–463 (1929).
86. T. J. Stieltjes, Quelques recherches sur la théorie des quadratures dites mécaniques. *Ann. Sci. Éc. Norm. Paris*, (3) **1**, 409–426 (1884). [Also in *Œuvres* I, 377–396.]
87. Z. Strakoš and J. Liesen, On numerical stability in large scale linear algebraic computations. *Z. Angew. Math. Mech.*, **85**(5), 307–325 (2005).
88. Zdeněk Strakoš and Petr Tichý, On error estimation in the conjugate gradient method and why it works in finite precision computations. *Electron. Trans. Numer. Anal.*, **13**, 56–80 (2002).
89. Gabor Szegö, *Orthogonal polynomials*, AMS Colloquium Publications 23, Amer. Math. Soc., New York (1939).
90. Marc Van Barel, Dario Fasino, Luca Gemignani, and Nicola Mastronardi, Orthogonal rational functions and structured matrices. *SIAM J. Matrix Anal. Appl.*, **26**(3), 810–829 (2005).
91. David S. Watkins, Some perspectives on the eigenvalue problem. *SIAM Rev.*, **35**(3), 430–471 (1993).
92. John C. Wheeler, Modified moments and Gaussian quadratures. *Rocky Mountain J. Math.*, **4**(2), 287–296 (1974).
93. Herbert S. Wilf, *Mathematics for the physical sciences*, Wiley, New York (1962).
94. Wolfgang Wülling, The stabilization of weights in the Lanczos and conjugate gradient method, *BIT Numer. Math.*, **45**(2), 395–414 (2005).

An asterisk denotes a paper reprinted in this volume.

29.9. [186] "On Euler's attempt to compute logarithms by interpolation: A commentary to his letter of February 16, 1734 to Daniel Bernoulli"

---

[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 (2008).

Available online at www.sciencedirect.com
ScienceDirect
Journal of Computational and Applied Mathematics 219 (2008) 408–415
JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS
www.elsevier.com/locate/cam

On Euler's attempt to compute logarithms by interpolation: A commentary to his letter of February 16, 1734 to Daniel Bernoulli

Walter Gautschi

Department of Computer Sciences, 250 N. University Street, West Lafayette, IN 47907-2066, USA

Received 21 September 2006

Dedicated to Claude Brezinski on the occasion of his retirement

Abstract

In the letter to Daniel Bernoulli, Euler reports on his attempt to compute the common logarithm $\log x$ by interpolation at the successive powers of 10. He notes that for $x = 9$ the procedure, though converging fast, yields an incorrect answer. The interpolation procedure is analyzed mathematically, and the discrepancy explained on the basis of modern function theory. It turns out that Euler's procedure converges to a q-analogue $S_q(x)$ of the logarithm, where $q = \frac{1}{10}$. In the case of the logarithm $\log_\omega x$ to base $\omega > 1$ (considered by Euler almost twenty years later), the limit of the analogous procedure (interpolating at the successive powers of ω) is $S_q(x)$ with $q = 1/\omega$. It is shown that by taking $\omega > 1$ sufficiently close to 1 and interpolating at sufficiently many points, the logarithm $\log x$ can indeed be approximated arbitrarily closely, although, if x, $1 < x < 10$, is relatively large, extremely high-precision arithmetic is required to overcome severe numerical cancellation. An alternative procedure for computing $\log x$ by interpolation at points in $[1, 10^\omega]$, $\omega > 0$, accumulating at the lower end point, is shown to converge to the desired limit, but also not without numerical complications.

MSC: 01A50; 65−03

Keywords: Euler's correspondence with Daniel Bernoulli; Interpolation series for the logarithm; q-analogue of the logarithm

1. The handwritten original of the letter[1] in question is kept at the University Library of Basel under the signature Ms. L Ia 689 fol. 145–146v and has been published by G. Eneström in [2]. Fig. 1 shows the passage relevant to us, including the rather formal closing phrases "Womit/verbleibe mit schuldigster Hochachtung/Eurer Hochedelgebohrnen/Meines Hochgeehrtesten Herren Professors/gehorsamster und verbundenster/Leonhard Euler". [*Author's translation*: Herewith I remain in most obliged respect your Honorable's and my most highly esteemed Professor's most obedient and indebted Leonhard Euler.]

The mathematical passage reads as follows: "Ich vermeinte neulich, daß nachfolgende *Series*

$$\frac{m-1}{9} - \frac{(m-1)(m-10)}{990} + \frac{(m-1)(m-10)(m-100)}{999\,000} - \frac{(m-1)(m-10)(m-100)(m-1000)}{9\,999\,000\,000} + \text{etc.}$$

E-mail address: wxg@cs.purdue.edu.

[1] The letter is dated in the old style (Julian), since Euler wrote from Petersburg; the corresponding date in the new style (Gregorian) is February 27, 1734.

0377-0427/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2006.11.027

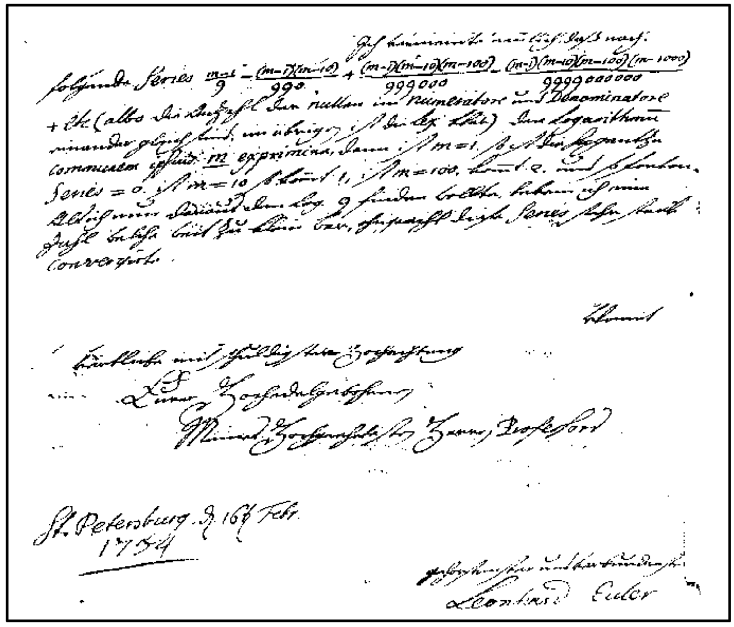


Fig. 1. Excerpt from Euler's letter to D. Bernoulli.

(alwo die Anzahl der nullen im *Numeratore* und *Denominatore* einander gleich sind, im übrigen ist die *Lex* klar) den *Logarithmum communem ipsius m* exprimiere, dann ist $m = 1$, so ist die gantze *Series* $= 0$, ist $m = 10$ so kommt 1, ist $m = 100$, kommt 2, und so fortan. Als ich nun daraus den *Log*[*arithmum*] 9 finden wollte, bekam ich eine Zahl welche weit zu klein war, ohngeacht diese *Series* sehr stark convergirte". [*Author's translation*: I recently thought that the following series

$$\frac{m-1}{9} - \frac{(m-1)(m-10)}{990} + \frac{(m-1)(m-10)(m-100)}{999\,000} - \frac{(m-1)(m-10)(m-100)(m-1000)}{9\,999\,000\,000} + \text{etc.}$$

(where the number of zeros in the numerator and in the denominator is the same—the law, after all, is clear) would represent the common logarithm of m, for, when $m = 1$, the whole series is $=0$, if $m = 10$, it becomes 1, if $m = 100$ it is 2, and so on. Now when I wanted to find from it the logarithm of 9, I obtained a number which is much too small, even though the series converged very strongly.]

2. Euler's intention, in modern terminology, is to compute the common logarithm by interpolating a certain number (ideally infinitely many) of known values of the logarithm. Fearless (even reckless) as so often was the case, Euler takes the simplest values, $\log 10^k = k$, $k = 0, 1, 2, 3, \ldots$, and for $\log x$, $x > 0$, writes down the infinite series

$$S(x) = \sum_{k=1}^{\infty} a_k (x-1)(x-10)\cdots(x-10^{k-1}), \tag{1}$$

whose nth partial sum is Newton's interpolation polynomial of degree n, hence

$$a_k = [x_0, x_1, \ldots, x_k] f$$

the divided difference of order k for the function $f(x) = \log x$ and abscissae $x_r = 10^r$, $r = 0, 1, \ldots, k$. This may have been the way in which Euler determined the first four coefficients

$$a_1 = \frac{1}{9}, \quad a_2 = -\frac{1}{990}, \quad a_3 = \frac{1}{999\,000}, \quad a_4 = -\frac{1}{9\,999\,000\,000}.$$

The "law", which he alludes to, apparently is

$$a_n = \frac{(-1)^{n-1}}{10^{n(n-1)/2}(10^n - 1)}, \quad n = 1, 2, 3, \ldots . \tag{2}$$

We assert, somewhat more generally, that for arbitrary integer valued $r \geqslant 0$,

$$[x_r, x_{r+1}, \ldots, x_{r+n}]f = \frac{(-1)^{n-1}}{10^{rn+n(n-1)/2}(10^n - 1)}. \tag{3}$$

One proves (3) by mathematical induction on n. For $n = 1$, the assertion is evidently true. The validity of (3) for some n and arbitrary $r \geqslant 0$, and a well-known property of divided differences (see, e.g., [4, (2.64)]), then imply

$$\begin{aligned}
&[x_r, x_{r+1}, \ldots, x_{r+n}, x_{r+n+1}]f \\
&\quad = \frac{[x_{r+1}, x_{r+2}, \ldots, x_{r+n+1}]f - [x_r, x_{r+1}, \ldots, x_{r+n}]f}{x_{r+n+1} - x_r} \\
&\quad = \frac{(-1)^{n-1}}{10^{rn+n(n-1)/2}(10^n - 1)} \frac{1 - 10^n}{10^n(10^{r+n+1} - 10^r)} \\
&\quad = \frac{(-1)^n}{10^{rn+n(n-1)/2}10^{n+r}(10^{n+1} - 1)} \\
&\quad = \frac{(-1)^n}{10^{r(n+1)+n(n+1)/2}(10^{n+1} - 1)},
\end{aligned}$$

which is precisely (3) with n replaced by $n + 1$. This proves (3), and therefore also (2).

3. It suffices, of course, to assume $1 \leqslant x < 10$, since every other positive number x' can be written in the form $x' = x \times 10^p$ with some integer $p \neq 0$, and $\log x' = p + \log x$. The series (1) then converges uniformly on [1, 10] and, as Euler remarks, very fast. The nth term $t_n(x)$ of (1), when a_n is given by (2), in fact computes to

$$t_n(x) = -\frac{\prod_{k=0}^{n-1}(1 - x/10^k)}{10^n - 1} = -\frac{q^n}{1 - q^n}(x; q)_n, \quad q = \frac{1}{10}, \tag{4}$$

where

$$(x; q)_n = \prod_{k=0}^{n-1}(1 - xq^k) \tag{5}$$

is the q-shifted factorial (cf. [1, Section 10.2]). There holds, for $1 \leqslant x < 10$ and $n \geqslant 2$,

$$|t_n(x)| < \frac{9}{10^n - 1}\left(1 - \frac{1}{10^{n-1}}\right) < \frac{9}{10^n}, \tag{6}$$

so that the nth partial sum of the series

$$S(x) = \sum_{k=1}^{\infty} t_k(x) \tag{7}$$

approximates its limit up to an error less than $9 \times 10^{-(n+1)}(1 + 10^{-1} + 10^{-2} + \cdots) = 10^{-n}$. For Euler's special value $x = 9$ one so obtains

$$S(9) = 0.897778586588\ldots,$$

a value which is significantly smaller than $\log 9 = 0.954242509439\ldots$; the relative error is 5.92%.

One can speculate what prompted Euler to communicate his computation for the special value $x = 9$. Very likely, he also tried other (integer valued) $x < 9$, but had to observe that the results are then even worse. As a matter of fact, the relative deviation of the limit value from the true value of the logarithm increases monotonically as x decreases over the natural numbers from 9 to 2, and at $x = 2$ is about 10 times as large as at $x = 9$, and at $x = 0$ even 100%.

4. From today's perspective it is not surprising that the series (7) does not converge to the expected value. The nth term of the series, after all, is a polynomial of degree n, thus an analytic function of the complex variable x, and the series itself converges uniformly on each disk $|x| \leqslant R$. In fact,

$$\left|\prod_{k=1}^{n-1}(1 - x/10^k)\right| \leqslant \prod_{k=1}^{n-1}(1 + R/10^k),$$

and the product on the right converges absolutely when $n \to \infty$. Therefore, $C \times \sum_{n=1}^{\infty} 1/(10^n - 1)$, where $C = (R + 1)\prod_{k=1}^{\infty}(1+R/10^k)$, is a convergent majorant of the series. By Weierstraß's double-series theorem, $S(x)$ thus represents a function which is analytic in every domain $|x| \leqslant R$, hence an entire function. Consider now $d(z) = S(z) - \log z$ in the domain $\mathscr{D} = \{z \in \mathbb{C} : |\arg z| < \pi\}$, where log denotes the principal branch of the logarithm. If we had $d(z) = 0$ at infinitely many points which have a point of accumulation in $\mathscr{D}\backslash\{\infty\}$ (for example, in an arbitrarily small interval of the real line), it would follow that $d(z) \equiv 0$ for all $z \in \mathscr{D}$. This evidently is impossible since $d(x) \to \infty$ when $x \downarrow 0$. Consequently, $S(x)$ cannot be identically equal to $\log x$ on any interval, however small.

Interestingly, however, the function $S(x)$ may be interpreted as a q-analogue of the logarithm, where $q = \frac{1}{10}$; cf. Section 5.

5. The motivation for Euler's bold choice $x_r = 10^r$ of the abscissae of interpolation is of course clear: not a single logarithm needs to be computed in order to generate the interpolation data. Almost equally simple would be the choice $x_r = \omega^r$, $\omega > 0$, which requires only one single logarithm, $\log \omega$. It is natural, then, to consider interpolation to the logarithm to base ω, that is, to $\log_\omega x = \log x / \log \omega$. What is the interpolation series[2] in this case and how does it behave?

To analyze the function $S(x; \omega)$ represented by the interpolation series, it is useful to introduce a q-analogue of the logarithm as defined by E. Koelink and W. Van Assche (see [5], where in Section 6 other definitions of the q-logarithm, used in the physics literature, are also discussed),

$$S_q(x) = -\sum_{n=1}^{\infty} \frac{q^n}{1 - q^n}(x; q)_n, \tag{8}$$

where $(x; q)_n$ is the q-shifted factorial (5). One verifies, at least formally, that

$$\lim_{q\to 1}(1 - q)S_q(x) = -\sum_{n=1}^{\infty} \frac{(1 - x)^n}{n} = \ln x, \quad 0 < x < 2 \tag{9}$$

and

$$S_q(q^{-n}) = n, \quad n = 0, 1, 2, \ldots, \tag{10}$$

which motivates the name "q-analogue of the logarithm". On the other hand, in analogy to (4) one obtains

$$S(x; \omega) = \sum_{n=1}^{\infty} t_n(x; \omega), \quad t_n(x; \omega) = -\frac{q^n}{1 - q^n}(x; q)_n, \quad q = \frac{1}{\omega}, \tag{11}$$

[2] Euler returns to this series almost twenty years later in his memoir [3] (where a is written in place of ω). He derives very elegantly the logarithmic nature (10), (12) of $S(x; \omega)$, emphasizing repeatedly that it holds only for positive integer values of n, and he computes (in Section 10) $S_q(q^{-n})$ also for negative n, explicitly for $n \geqslant -5$. He missed, however, the close connection of $\log \omega \cdot S(x; \omega)$ to $\log x$ when $\omega \downarrow 1$ (cf. (13) and (16) below), which in view of the strange numerical behavior of $\log \omega \cdot S(x; \omega)$ as $\omega \downarrow 1$ (cf. Section 6) is easy to understand. Instead, he used the series $S(x; \omega)$ as a springboard to derive all sorts of identities for it, among others two special cases (in Sections 17 and 26) of what today is known as the "q-binomial theorem". He also finds the expansion of $S(x; \omega)$ in powers of x and from known infinite products deduces new infinite series. At the end of the memoir Euler calculates Lambert's series $-S(0; \omega) = \sum_{n=1}^{\infty} 1/(\omega^n - 1)$ for $\omega = 10$ to 30 decimal places, but not before developing a convergence acceleration scheme for the more general series $\sum_{n=1}^{\infty} 1/(\omega^n - z)$.

so that

$$S(x;\omega) = S_{1/\omega}(x). \tag{12}$$

Note that (10) with $q = 1/\omega$ are precisely the interpolation conditions which produced the interpolation series $S(x;\omega)$ in the first place. It is evident from (11) and (5) that when $\omega < 1$, hence $q > 1$, the terms $t_n(x;\omega)$ converge to 1 if $x = 0$, or to infinity in absolute value if $x \neq 0$, so that the series in (11) diverges. This definitely rules out the temptation of choosing $x_r = 10^{-r}$.

Assume, therefore, that $\omega > 1$. By an argument analogous to the one in Section 4 the series $S(x;\omega)$, and hence also $S_{1/\omega}(x)$, is seen to be an entire function (now depending on the parameter ω). It is true that larger values of ω yield faster convergence of the series in (11), but (9) and (12) suggest that better approximations to the logarithm can be expected for values of $\omega > 1$ closer to 1. We now show indeed that $\log x$ can be approximated by Euler's interpolation process as accurately as we wish by taking $\omega > 1$ sufficiently close to 1 and taking sufficiently many terms in the series of (11). We prove this for $0 < x < 2$, and provide numerical evidence for it when $x \geqslant 2$.

Since $\log x = \log\omega \cdot \log_\omega x$, the nth-degree interpolation approximation to the common logarithm $\log x$ is

$$s_n = \log\omega \cdot S_n(x;\omega), \tag{13}$$

where $S_n(x;\omega)$ is the nth partial sum of $S(x;\omega)$. Now

$$\frac{\ln(1/q)}{\ln 10} S_q(x) = \frac{\ln(1/q)}{(1-q)\ln 10} \cdot (1-q) S_q(x),$$

so that as $q \to 1$, since $\ln q^{-1}/(1-q) \to 1$, it follows from (9) that

$$\lim_{q\to 1} \frac{\ln(1/q)}{\ln 10} S_q(x) = \frac{\ln x}{\ln 10} = \log x, \quad 0 < x < 2.$$

Therefore, since $q = 1/\omega$ and using (12), $\lim_{\omega\downarrow 1} \log\omega \cdot S(x;\omega) = \log x$, so that, given any $\varepsilon > 0$, we can choose $\omega > 1$ sufficiently close to 1 to have

$$|\log\omega \cdot S(x;\omega) - \log x| \leqslant \frac{\varepsilon}{2}. \tag{14}$$

On the other hand, n can be taken large enough so that

$$|\log\omega \cdot S_n(x;\omega) - \log\omega \cdot S(x;\omega)| \leqslant \frac{\varepsilon}{2}. \tag{15}$$

Combining (14) and (15) yields

$$\begin{aligned} |s_n - \log x| &= |\log\omega \cdot S_n(x;\omega) - \log\omega \cdot S(x;\omega) + \log\omega \cdot S(x;\omega) - \log x| \\ &\leqslant |\log\omega \cdot S_n(x;\omega) - \log\omega \cdot S(x;\omega)| + |\log\omega \cdot S(x;\omega) - \log x| \\ &\leqslant \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon, \end{aligned} \tag{16}$$

as was to be shown.

6. We have seen that the interpolation procedure converges for $0 < x < 2$ (to the correct value $\log x$) as $\omega \downarrow 1$ and $n \to \infty$. Interestingly, the same seems to persist also for $x \geqslant 2$, but not without considerable numerical obstacles. Before discussing this, we note a simple scheme to evaluate $t_n(x;\omega)$ and thus, by summation, $S(x;\omega)$. Letting $u_n = (\omega^n - 1)t_n(x;\omega)$, one obtains from (11) the recursive procedure

$$\begin{aligned} &t_1(x;\omega) = \frac{x-1}{\omega-1}, \\ &\left.\begin{aligned} u_n &= (1 - x/\omega^{n-1})u_{n-1}, \\ t_n(x;\omega) &= \frac{u_n}{\omega^n - 1}, \end{aligned}\right\} \quad n = 2, 3, \ldots, \end{aligned} \tag{17}$$

Table 1
Largest values of $|\log\omega \cdot t_n(x;\omega)|$

$x\backslash\omega$	1.1	1.05	1.025	1.0125	1.00625
2	0.41	0.42	0.43	0.43	0.43
6	0.11×10^{4}	0.78×10^{7}	0.81×10^{15}	0.17×10^{32}	0.15×10^{65}
10	0.19×10^{8}	0.24×10^{16}	0.74×10^{32}	0.13×10^{66}	0.82×10^{132}

Table 2
Errors achievable by the interpolation process of Section 5

$x\backslash\omega$	1.1	1.05	1.025	1.0125	1.00625
2	0.17×10^{-12}	0.14×10^{-23}	0.95×10^{-46}	0.18×10^{-88}	0.20×10^{-174}
6	0.24×10^{-8}	0.43×10^{-15}	0.22×10^{-28}	0.43×10^{-55}	0.11×10^{-107}
10	0.43×10^{-4}	0.12×10^{-6}	0.17×10^{-11}	0.54×10^{-21}	0.76×10^{-40}
d	40	50	60	100	200
n	100	200	400	800	1500

which needs to be initialized by

$$u_1 = x - 1. \tag{18}$$

To interpolate the common logarithm $\log x$, the initial terms t_1 and u_1 must be multiplied by $\log\omega$.

The "obstacles" referred to above have to do with the fact that for values of ω larger than, but close to 1, the quantities $\log\omega \cdot t_n(x;\omega)$ become extremely large before eventually converging to zero as $n \to \infty$, at least when $x \leqslant 10$ is relatively large. This is illustrated in Table 1 above, which shows $\max_{n\geqslant 1}|\log\omega \cdot t_n(x;\omega)|$ for selected values of x and ω.

Yet, for each fixed ω, the series $S(x;\omega) = \sum_{n=1}^{\infty} t_n(x;\omega)$ converges. Because of the enormous amount of internal cancellation that may take place in this series, however, the computation must be performed in appropriately high precision.

This again is illustrated in Table 2, showing the errors achievable in symbolic/variable-precision computation with d decimal digits and n terms of the series. It should, perhaps, be emphasized that for each fixed ω, increasing d and n beyond the values shown, will not reduce the errors any further; all it does is compute $S(x;\omega)$, and with it, $\log\omega \cdot S(x;\omega) - \log x$, more accurately. This is why we called the errors "achievable".

This somewhat bizarre behavior of the interpolation process, on reflection, is not entirely unexpected: For one, x in (9) already had to be restricted to the interval $(0, 2)$. For another, when $\omega > 1$ is very close to 1, then all $x_r = \omega^r$ initially are almost equal to 1. If they were all equal to 1, then the interpolation series would be Taylor's expansion of $\log x$ at 1, which diverges if $x > 2$. Our interpolation process, for x much larger than 2, thus behaves, initially, as if it would diverge, and only when n becomes large and the points x_r begin to spread out, does it turn around and take on a more reasonable, eventually convergent, demeanor. While there may be some theoretical interest in this kind of approximation process, it has little practical merit because of the excessive computing effort required. (The last five columns of Table 2 take, respectively, 96, 187, 382, 741, and 1493 seconds to compute on the Sun Ultra5 Workstation.)

Nevertheless, if x is restricted to the interval $[1, 5]$, the process is not entirely impractical, since when $\omega = 1.1$, for example, there holds $|\log\omega \cdot t_n(x;\omega)| < 72.2$, and with $n = 100$ terms, one is still able to obtain values of $\log x$, $1 \leqslant x \leqslant 5$, accurate to about 10 decimal digits, even in 14-digit computation. For values of x in the interval $(5, 10]$, one applies the process to $x/2$ and adds $\log 2$ to the result. Better yet, in today's age of technology and binary computer arithmetic, we may restrict x to the interval $[1, 2]$, in which case $|\log\omega \cdot t_n(x;\omega)| < 1$ and $\omega = 1.1$, $n = 20$ generally yields an accuracy of 10 or more decimal digits (nine digits near $x = 1$), while $\omega = 1.05$, $n = 15$ yields 11 or more correct digits.

7. There is still another way in which Euler's ideas can in principle be salvaged and made workable. To begin with, choose as interpolation abscissae $x_r = 10^{\omega/(r+1)}$, $\omega > 0$, $r = 0, 1, 2, \ldots$, so that

$$x_r \in (1, 10^{\omega}] \quad \text{for all } r = 0, 1, 2, \ldots . \tag{19}$$

It is known, in fact (cf., e.g., [4, p. 83]), that for the function f and (arbitrary) abscissae of interpolation, all lying in a finite interval $[a, b]$, the interpolation series converges to $f(x)$ for any x in $[a, b]$, provided f has infinitely many in $[a, b]$ continuous derivatives and, moreover, there holds

$$\lim_{k\to\infty} \frac{(b-a)^k}{k!} M_k = 0, \tag{20}$$

where M_k denotes an upper bound of $|f^{(k)}|$ on $[a, b]$. This easily follows from Cauchy's formula [4, (2.12)] for the error of interpolation. It can also be shown ([4, p. 84]), that (20) is indeed true if f is analytic in a disk with center at the middle of the interval $[a, b]$ and radius $r > \frac{3}{2}(b-a)$.

In our case $f(x) = \log x$, one has $f^{(k)}(x) = (-1)^{k-1}(k-1)!x^{-k}/\ln 10$, and (20) is equivalent to $|(b/a) - 1| < 1$. More precisely, one has at least geometric convergence with ratio q if

$$\left|\frac{b}{a} - 1\right| \leqslant q < 1. \tag{21}$$

Choosing $q = \frac{1}{2}$, one obtains for the interval (19), where $b/a = 10^{\omega}$,

$$\omega \leqslant \log \tfrac{3}{2} = 0.17609\ldots . \tag{22}$$

Thus, in the interval (19), when ω is given by (22), the interpolation series converges (to the correct value) at least geometrically with ratio $\frac{1}{2}$.

Now if x is given with $1 \leqslant x < 10$, one determines the integer $k_0 \geqslant 0$ such that

$$10^{k_0\omega} \leqslant x < 10^{(k_0+1)\omega}. \tag{23}$$

This can easily be achieved (on a computer) by means of a small routine like (in pseudocode)

```
k0 = 0;
while x ⩾ 10^((k0+1)ω)
  k0 = k0+1;
end
```

If then one puts $t = 10^{-k_0\omega}x$, there holds $1 \leqslant t < 10^{\omega}$, and one computes $\log t$ as above, whereupon $\log x = k_0\omega + \log t$.

Here too, however, not everything works as expected. It transpires (apparently because of the crowding of the interpolation abscissae in the lower part of the interval $(1, 10^{\omega}]$), that the algorithm described eventually succumbs to the detrimental effects of rounding errors. The latter progressively affect the computation of the divided differences (no longer explicitly known) to the point of rendering them completely meaningless. In computation with 36 decimal places (quadruple precision in Fortran), for example, the error of the interpolation polynomial is seen to decrease monotonically with increasing degree, but only up to a degree n of about $n = 18$; thereafter, the error increases rapidly. Nevertheless, it is still possible, in this precision, to obtain at least 10 good decimals, generally, however, many more, even as many as 35.

8. We have tried to understand and, following his own ideas, to rehabilitate Euler's unsuccessful computation of the logarithm, but do not want to leave behind the impression that the resulting computational schemes would be competitive with newer methods of approximation theory (see, e.g., [6]). These modern methods, however, are products of the 20th century.

Acknowledgments

I wish to thank Dr. E.A. Fellmann (Basel) for drawing my attention to Euler's letter and for persuading me to write a commentary on it, and to an anonymous referee for the very interesting Ref. [3]. I am also indebted to Erik Koelink and Walter Van Assche for a useful exchange of ideas regarding the q-analogue of the logarithm.

References

[1] G.E. Andrews, R. Askey, R. Roy, Special functions, Encyclopedia of Mathematics and its Applications, vol. 71, Cambridge University Press, Cambridge, 1999.
[2] G. Eneström, Bib. Math. 7(3) 1906–1907, 134–137.
[3] L. Euler, Consideratio quarumdam serierum quae singularibus proprietatibus sunt praeditae, Novi Comment. Acad. Sci. Petropolitanae 3 (1750/51) 1753, 10–12, 86–108. (Also in Leonhardi Euleri Opera Omnia, Ser. I, vol. 14, pp. 516–541, B.G. Teubner, Leipzig and Berlin, 1925. An English translation of this memoir can be downloaded from the E190 page of the Euler Archive at http://www.math.dartmouth.edu/∼euler.)
[4] W. Gautschi, Numerical Analysis: An Introduction, Birkhäuser, Boston, MA, 1997.
[5] E. Koelink, W. Van Assche, Leonhard Euler and a q-analogue of the logarithm, in preparation.
[6] J.M. Muller, Elementary Functions: Algorithms and Implementation, second ed., Birkhäuser, Boston, MA, 2006.

29.10. [187] "Leonhard Euler: His Life, the Man, and His Works"

---

[187] "Leonhard Euler: His Life, the Man, and His Works," *SIAM Rev*. **50**, 3–33 (2008). [Also published in *ICIAM 07, 6th 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* (2–3) (2008).]

SIAM Review
Vol. 50, No. 1, pp. 3–33

Leonhard Euler: His Life, the Man, and His Works*

Walter Gautschi†

Abstract. On the occasion of the 300th anniversary (on April 15, 2007) of Euler's birth, an attempt is made to bring Euler's genius to the attention of a broad segment of the educated public. The three stations of his life—Basel, St. Petersburg, and Berlin—are sketched and the principal works identified in more or less chronological order. To convey a flavor of his work and its impact on modern science, a few of Euler's memorable contributions are selected and discussed in more detail. Remarks on Euler's personality, intellect, and craftsmanship round out the presentation.

Key words. Leonhard Euler, sketch of Euler's life, works, and personality

AMS subject classification. 01A50

DOI. 10.1137/070702710

Seh ich die Werke der Meister an,
So sehe ich, was sie getan;
Betracht ich meine Siebensachen,
Seh ich, was ich hätt sollen machen.
–Goethe, Weimar 1814/1815

1. Introduction. It is a virtually impossible task to do justice, in a short span of time and space, to the great genius of Leonhard Euler. All we can do, in this lecture, is to bring across some glimpses of Euler's incredibly voluminous and diverse work, which today fills 74 massive volumes of the *Opera omnia* (with two more to come). Nine additional volumes of correspondence are planned and have already appeared in part, and about seven volumes of notebooks and diaries still await editing!

We begin in section 2 with a brief outline of Euler's life, going through the three stations of his life: Basel, St. Petersburg (twice), and Berlin. In section 3, we identify in more or less chronological order Euler's principal works and try to convey a flavor and some characteristic features of his work by describing in more detail a few of his many outstanding contributions. We conclude in section 4 with remarks on Euler's personality and intellect, as gained from testimonials of his contemporaries, and on the quality of his craft, and in section 5 with some bibliographic information for further reading.

*Published electronically February 1, 2008. Expanded version of a lecture presented at the 6th International Congress on Industrial and Applied Mathematics in Zürich, Switzerland, on July 18, 2007. For a video of a preliminary version of this lecture, presented on March 7, 2007, at Purdue University, see http://epubs.siam.org/sam-bin/getfile/SIREV/articles/70271_01.avi. By mutual agreement between the editorial boards of the European Mathematical Society and the Society for Industrial and Applied Mathematics, and with the consent of the author, this lecture is being published also in the Proceedings of the International Congress of Industrial and Applied Mathematics, Zürich, July 16–20, 2007, R. Jeltsch and G. Wanner, eds., European Mathematical Society (EMS), Zürich, 2008.

http://www.siam.org/journals/sirev/50-1/70271.html

†Department of Computer Sciences, Purdue University, West Lafayette, IN 47907-2066 (wxg@cs.purdue.edu).

2. His Life.

2.1. Basel 1707–1727: Auspicious Beginnings. Leonhard Euler was born on April 15, 1707, the first child of Paulus Euler and Margaretha Brucker. Paulus Euler came from modest folk, mostly artisans, while Margaretha Brucker's ancestors include a number of well-known scholars in the classics. Euler's father at the time was a vicar at the church of St. Jakob, just outside the old city walls of Basel. Although a theologian, Paulus had interests in mathematics and took courses from the famous Jakob Bernoulli during the first two years of his study at the university. About a year and a half after Leonhard's birth, the family moved to Riehen, a suburb of Basel, where Paulus Euler assumed the position of Protestant minister at the local parish. He served in that capacity faithfully and devotedly for the rest of his life.

Fig. 1 *The parish residence and church in Riehen.*

The parish residence, as it looks today (Figure 1), seems comfortable enough, but at the time it had one floor less and only two rooms with heating. The living quarters it provided, therefore, were rather cramped, especially after the family increased by another child, Anna Maria, in 1708. Two more children, Maria Magdalena and Johann Heinrich, were to follow later on.

Leonhard received his first schooling in mathematics at home from his father. Around the age of eight he was sent to the Latin school in Basel and given room and board at his maternal grandmother's house. To compensate for the poor quality then prevailing at the school, Paulus Euler hired a private tutor for his son, a young theologian by the name of Johannes Burckhardt, himself an enthusiastic lover of mathematics. In October of 1720, at the age of thirteen (not unusual at the time), Leonhard enrolled at the University of Basel, first at the philosophical faculty, where he took the freshman courses on elementary mathematics given by Johann Bernoulli, the younger brother of the now deceased Jakob. The young Euler pursued his mathematical studies with such a zeal that he soon caught the attention of Bernoulli, who encouraged him to study more advanced books on his own and even offered him assistance at his house every Saturday afternoon. In 1723, Euler graduated with a master's degree and a public lecture (in Latin) comparing Descartes's system of natural philosophy with that of Newton.

Following the wishes of his parents, he then entered the theological faculty, devoting, however, most of his time to mathematics. Euler's father eventually had to concede, probably at the urging of Johann Bernoulli, that Leonhard was predestined

Fig. 2 *The old university of Basel and Johann* I *Bernoulli.*

to a career in mathematics rather than one in theology. This is how Euler himself recounts this early learning experience at the university in his brief autobiography of 1767 (here freely translated from German; see Fellmann [10, Engl. transl., pp. 1–7]):

> In 1720 I was admitted to the university as a public student, where I soon found the opportunity to become acquainted with the famous professor Johann Bernoulli, who made it a special pleasure for himself to help me along in the mathematical sciences. Private lessons, however, he categorically ruled out because of his busy schedule. However, he gave me a far more beneficial advice, which consisted in myself getting a hold of some of the more difficult mathematical books and working through them with great diligence, and should I encounter some objections or difficulties, he offered me free access to him every Saturday afternoon, and he was gracious enough to comment on the collected difficulties, which was done with such a desired advantage that, when he resolved one of my objections, ten others at once disappeared, which certainly is the best method of making happy progress in the mathematical sciences.

These personal meetings have become known, and famous, as the *privatissima*, and they continued well beyond his graduation. It was during these *privatissima* that Johann Bernoulli more and more began to admire the extraordinary mathematical talents of the young Euler.

Barely nineteen years old, Euler dared to compete with the greatest scientific minds of the time by responding to a prize question of the Paris Academy of Sciences with a memoir on the optimal placing of masts on a ship. He, who at that point in his life had never so much as seen a ship, did not win first prize, but still a respectable second. A year later, when the physics chair at the University of Basel became vacant, the young Euler, dauntlessly again, though with the full support of his mentor, Johann Bernoulli, competed for the position, but failed, undoubtedly because of his youth and lack of an extensive record of publications. In a sense, this was a blessing in disguise, because in this way he was free to accept a call to the Academy of Sciences in St. Petersburg, founded a few years earlier by the czar Peter I (the Great), where he was to find a much more promising arena in which to fully develop himself. The groundwork for this appointment had been laid by Johann Bernoulli and two of his sons, Niklaus II and Daniel I, both of whom were already active at the Academy.

2.2. St. Petersburg 1727–1741: Meteoric Rise to World Fame and Academic Advancement. Euler spent the winter of 1726 in Basel studying anatomy and physiology in preparation for his anticipated duties at the Academy. When he arrived in St. Petersburg and started his position as an adjunct of the Academy, it was soon determined, however, that he should devote himself entirely to the mathematical sciences. In addition, he was to participate in examinations for the cadet corps and act as a consultant to the Russian state in a variety of scientific and technological questions.

Fig. 3 *The Academy in St. Petersburg and Peter* I. *(Photograph of the Academy of Sciences courtesy of Andreas Verdun.)*

Euler adjusted easily and quickly to the new and sometimes harsh life in the northern part of Europe. Contrary to most other foreign members of the Academy he began immediately to study the Russian language and mastered it quickly, both in writing and speaking. For a while he shared a dwelling with Daniel Bernoulli, and he was also on friendly terms with Christian Goldbach, the permanent Secretary of the Academy and best known today for his—still open—conjecture in number theory. The extensive correspondence between Euler and Goldbach that ensued has become an important source for the history of science in the 18th century.

Euler's years at the Academy of St. Petersburg proved to be a period of extraordinary productivity and creativity. Many spectacular results achieved during this time (more on this later) brought him instant world fame and increased status and esteem within the Academy. A portrait of Euler stemming from this period is shown in Figure 4.

In January of 1734 Euler married Katharina Gsell, the daughter of a Swiss painter teaching at the Academy, and they moved into a house of their own. The marriage brought forth thirteen children, of whom, however, only five reached the age of adulthood. The first-born child, Johann Albrecht, was to become a mathematician himself and later in life was to serve Euler as one of his assistants.

Euler was not spared misfortunes. In 1735, he fell seriously ill and almost lost his life. To the great relief of all, he recovered, but suffered a repeat attack three years later of (probably) the same infectious disease. This time it cost him his right eye, which is clearly visible on all portraits of Euler from this time on (for example, the famous one in Figure 6, now hanging in the Basel Museum of Arts).

The political turmoil in Russia that followed the death of the czarina Anna Ivanovna induced Euler to seriously consider, and eventually decide, to leave St. Pe-

Fig. 4 *Euler, ca.* 1737.

tersburg. This all the more as he already had an invitation from the Prussian king Frederick II to come to Berlin and help establish an Academy of Sciences there. This is how Euler put it in his autobiography:

> ... in 1740, when His still gloriously reigning Royal Majesty [Frederick II] came to power in Prussia, I received a most gracious call to Berlin, which, after the illustrious Empress Anne had died and it began to look rather dismal in the regency that followed, I accepted without much hesitation ....

In June of 1741, Euler, together with his wife Katharina, the six-year-old Johann Albrecht, and the one-year-old toddler Karl, set out on the journey from St. Petersburg to Berlin.

2.3. Berlin 1741–1766: The Emergence of Epochal Treatises. Because of his preoccupation with the war campaign in Silesia, Frederick II took his time to set up the Academy. It was not until 1746 that the Academy finally took shape, with Pierre-Louis Moreau de Maupertuis its president and Euler the director of the Mathematics Class. In the interim, Euler did not remain idle; he completed some twenty memoirs, five major treatises (and another five during the remaining twenty years in Berlin), and composed over 200 letters!

Even though Euler was entrusted with manifold duties at the Academy—he had to oversee the Academy's observatory and botanical gardens, deal with personnel matters, attend to financial affairs, notably the sale of almanacs, which constituted the major source of income for the Academy, not to speak of a variety of technological and engineering projects—his mathematical productivity did not slow down. Nor was he overly distracted by an ugly priority dispute that erupted in the early 1750s over Euler's principle of least action, which was also claimed by Maupertuis and which the Swiss fellow mathematician and newly elected academician Johann Samuel König asserted to have already been formulated by Leibniz in a letter to the mathematician Jakob Hermann. König even came close to accusing Maupertuis of plagiarism. When challenged to produce the letter, he was unable to do so, and Euler was asked to investigate. Not sympathetic to Leibniz's philosophy, Euler sided with Maupertuis and in turn accused König of fraud. This all came to a boil when Voltaire, aligned with König, came forth with a scathing satire ridiculing Maupertuis and not sparing

Fig. 5 *The Berlin Academy and Frederick* II. *(Left panel reprinted with permission from the Archiv der Berlin-Brandenburgischen Akademie der Wissenschaften.)*

Fig. 6 *Euler,* 1753.

Euler either. So distraught was Maupertuis that he left Berlin soon thereafter, and Euler had to conduct the affairs of the Academy as *de facto*, if not *de jure*, president of the Academy.

By now, Euler was sufficiently well-off that he could purchase a country estate in Charlottenburg, in the western outskirts of Berlin, which was large enough to provide a comfortable home for his widowed mother (whom he had come to Berlin in 1750), his sister-in-law, and all the children. At just twenty years old, his first-born son, Johann Albrecht, was elected in 1754 to the Berlin Academy on the recommendation of Maupertuis. With a memoir on the perturbation of cometary orbits by planetary attraction he won in 1762 a prize of the Petersburg Academy, but had to share it with Alexis-Claude Clairaut. Euler's second son, Karl, went to study medicine in Halle, whereas the third, Christoph, became an officer in the military. His daughter Charlotte married into Dutch nobility, and her older sister Helene married a Russian officer later in 1777.

Euler's relation with Frederick II was not an easy one. In part, this was due to the marked difference in personality and philosophical inclination: Frederick—

proud, self-assured, worldly, a smooth and witty conversationalist, sympathetic to French enlightenment; Euler—modest, inconspicuous, down-to-earth, and a devout protestant. Another, perhaps more important, reason was Euler's resentment for never having been offered the presidency of the Berlin Academy. This resentment was only reinforced after Maupertuis's departure and Euler's subsequent efforts to keep the Academy afloat, when Frederick tried to interest Jean le Rond d'Alembert in the presidency. The latter indeed came to Berlin, but only to inform the king of his disinterest and to recommend Euler for the position instead. Still, Frederick not only ignored d'Alembert's advice, but ostentatiously declared himself the head of the Academy! This, together with many other royal rebuffs, finally led Euler to leave Berlin in 1766, in defiance of several obstacles put in his way by the king. He indeed already had a most cordial invitation from Empress Catherine II (the Great) to return to the Academy of St. Petersburg, which he accepted, and was given an absolutely triumphant welcome back.

Fig. 7 *The Euler house and Catherine* II. *(Left panel reprinted with permission from the Archiv der Berlin-Brandenburgischen Akademie der Wissenschaften.)*

2.4. St. Petersburg 1766–1783: The Glorious Final Stretch. Highly respected at the Academy and adored at Catherine's court, Euler now held a position of great prestige and influence that had been denied him in Berlin for so long. He in fact was the spiritual if not the appointed leader of the Academy. Unfortunately, however, there were setbacks on a personal level. A cataract in his left (good) eye, which already began to bother him in Berlin, now became increasingly worse, so that in 1771 Euler decided to undergo an operation. The operation, though successful, led to the formation of an abscess, which soon destroyed Euler's vision almost entirely. Later in the same year, his wooden house burned down during the great fire of St. Petersburg, and the almost blind Euler escaped from being burnt alive only by a heroic rescue by Peter Grimm, a workman from Basel. To ease the misfortune, the Empress granted funds to build a new house (the one shown in Figure 7 with the top floor having been added later). Another heavy blow hit Euler in 1773 when his wife Katharina Gsell died. Euler remarried three years later so as not to be dependent on his children.

In spite of all these fateful events, Euler remained mathematically as active as ever, if not more so. Indeed, about half of his scientific output was published, or originated, during this second St. Petersburg period, among which his two "best-sellers," *Letters to a German Princess* and *Algebra*. Naturally, he could not have done it without good secretarial and technical help, which he received from, among

Fig. 8 *Euler,* 1778.

others, Niklaus Fuss, a compatriot from Basel and future grandson-in-law of Euler, and Euler's own son, Johann Albrecht. The latter, by now secretary of the Academy, also acted as the protocolist of the Academy sessions, over which Euler, as the oldest member of the Academy, had to preside.

The high esteem in which Euler was held at the Academy and at court is touchingly revealed by a passage in the memoirs of the Countess Dashkova, a directress of the Academy appointed by the empress. She recounts the first time she accompanied the old Euler to one of the sessions of the Academy, probably Euler's last. Before the session started, a prominent professor and State Councilor as a matter of course claimed the chair of honor, next to the director's chair. The countess then turned to Euler and said: "Please be seated wherever you want; the seat you select will of course become the first of all."

Leonhard Euler died from a stroke on September 18, 1783 while playing with one of his grandchildren. Formulae that he had written down on two of his large slates describing the mathematics underlying the spectacular balloon flight undertaken on June 5, 1783, by the brothers Montgolfier in Paris were found on the day of his death. Worked out and prepared for publication by his son, Johann Albrecht, they became the last article of Euler; it appeared in the 1784 volume of the *Mémoires.* A stream of memoirs, however, all queued up at the presses of the Academy, were still to be published for nearly fifty years after Euler's death.

3. His Works. In the face of the enormous volume of Euler's writings, we content ourselves with briefly identifying his principal works, and then select, and describe in more detail, a few of Euler's prominent results in order to convey a flavor of his work and some of its characteristic features. The papers will be cited by their Eneström-Index numbers (E-numbers).

3.1. The Period in Basel. During the relatively short time of Euler's creative activity in Basel, he published two papers (E1, E3) in the *Acta Eruditorum* (Leipzig), one on isochronous curves, the other on so-called reciprocal curves, both influenced by Johann Bernoulli, and the work on the Paris Academy prize question (E4). The major work of this period is probably his *Dissertatio physica de sono* (E2), which he submitted in support of his application to the physics chair at the University of Basel and had printed in 1727 in Basel. In it, Euler discusses the nature and propagation of


Q. F. F. Q. S.
DISSERTATIO PHYSICA
DE SONO,
QUAM
ANNUENTE NUMINE DIVINO
JUSSU MAGNIFICI ET SAPIENTISSIMI PHILOSOPHORUM ORDINIS
PRO
VACANTE PROFESSIONE PHYSICA
Ad d. 18. Febr. A. MDCCXXVII.
In Auditorio Juridico horis p.
Publico Eruditorum Examini subjicit
LEONHARDUS EULERUS
A. L. M.
Adjuvabunt
Praestantissimo Adolescente
ERNESTO LUDOVICO BURCARDO
Phil. Cand.
BASILEÆ,
Typis E. & J. R. THURNISIORUM, Fratrum.

Fig. 9 *Physical Dissertation on Sound,* 1727. *(Reprinted with permission from Birkhäuser Verlag.)*

sound, in particular the speed of sound, and also the generation of sound by musical instruments. Some of this work is preliminary and has been revisited by Euler in his *Tentamen* (cf. section 3.2.1) and, thirty years later, in several memoirs (E305–E307).

3.2. First St. Petersburg Period. In spite of the serious setbacks in health, Euler's creative output during this period is nothing short of astonishing. Major works on mechanics, music theory, and naval architecture are interspersed with some 70 memoirs on a great variety of topics that run from analysis and number theory to concrete problems in physics, mechanics, and astronomy. An account of the mathematical work during this period is given in Sandifer [22].


MECHANICA
SIVE
MOTVS
SCIENTIA
ANALYTICE
EXPOSITA
AVCTORE
LEONHARDO EVLERO
ACADEMIAE IMPER. SCIENTIARVM MEMBRO ET
MATHESEOS SVBLIMIORIS PROFESSORE.
TOMVS I.
PETROPOLI
EX TYPOGRAPHIA ACADEMIAE SCIENTIARVM.
A. 1736.

MECHANICA
SIVE
MOTVS
SCIENTIA
ANALYTICE
EXPOSITA
AVCTORE
LEONHARDO EVLERO
ACADEMIAE IMPER. SCIENTIARVM MEMBRO ET
MATHESEOS SVBLIMIORIS PROFESSORE.
TOMVS II.
PETROPOLI
EX TYPOGRAPHIA ACADEMIAE SCIENTIARVM.
A. 1736.

Fig. 10 *Mechanics,* 1736. *(Reprinted with permission from Birkhäuser Verlag.)*

3.2.1. Major Works. The two-volume *Mechanica* (E15, E16) is the beginning of a far-reaching program, outlined by Euler in Vol. I, sect. 98, of composing a comprehensive treatment of all aspects of mechanics, including the mechanics of rigid, flexible, and elastic bodies, as well as fluid mechanics and celestial mechanics. The present work is restricted almost entirely to the dynamics of a point mass, to its free

motion in Vol. I and constrained motion in Vol. II. In either case the motion may take place either in a vacuum or in a resisting medium. The novelty of the *Mechanica* consists in the systematic use of (the then new) differential and integral calculus, including differential equations, and in this sense it represents the first treatise on what is now called analytic (or rational) mechanics. It had won the praise of many leading scientists of the time, Johann Bernoulli among them, who said of the work that "it does honor to Euler's genius and acumen." Also Lagrange, who in 1788 wrote his own *Mécanique analytique*, acknowledges Euler's mechanics to be "the first great work where Analysis has been applied to the science of motion." Implementation and systematic treatment of the rest of Euler's program, never entirely completed, occupied him throughout much of his life.

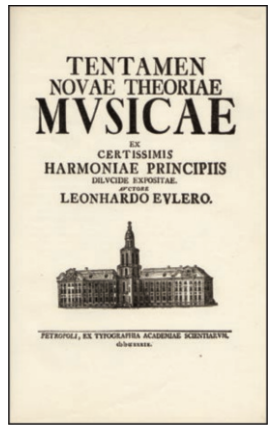

TENTAMEN
NOVAE THEORIAE
MVSICAE
EX
CERTISSIMIS
HARMONIAE PRINCIPIIS
DILVCIDE EXPOSITAE.
AVCTORE
LEONHARDO EVLERO.

Fig. 11 *Tentamen,* 1739 (1731). *(Reprinted with permission from Birkhäuser Verlag.)*

It is evident from Euler's notebooks that he thought a great deal about music and musical composition while still in Basel and had plans to write a book on the subject. These plans matured only later in St. Petersburg and gave rise to the *Tentamen novae theoriae musicae* (E33), usually referred to as the *Tentamen*, published in 1739 but completed already in 1731. (An English translation was made available by Smith [27, pp. 21–347].) The work opens with a discussion of the nature of sound as vibrations of air particles, including the propagation of sound, the physiology of auditory perception, and the generation of sound by string and wind instruments. The core of the work, however, deals with a theory of pleasure that music can evoke, which Euler develops by assigning to a tone interval, a chord, or a succession of such, a numerical value—the "degree"—which is to measure the agreeableness, or pleasure, of the respective musical construct: the lower the degree, the more pleasure. This is done in the context of Euler's favorite diatonic-chromatic temperament, but a complete mathematical theory of temperaments, both antique and contemporary ones, is also given.

In trying to make music an exact science, Euler was not alone: Descartes and Mersenne did the same before him, as did d'Alembert and many others after him (cf. Bailhache [2] and Assayag, Feichtinger, and Rodrigues [1]). In 1766 and 1774, Euler returns to music in three memoirs (E314, E315, and E457).

Euler's two-volume *Scientia navalis* (E110, E111) is a second milestone in his development of rational mechanics. In it, he sets forth the principles of hydrostatics and develops a theory of equilibrium and oscillations about the equilibrium of three-

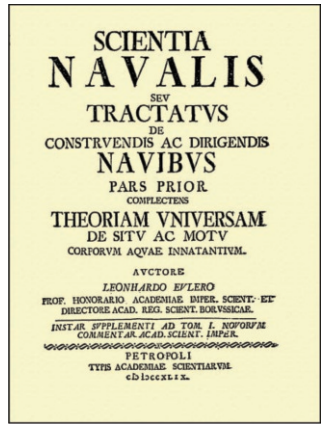

SCIENTIA
NAVALIS
SEV
TRACTATVS
DE
CONSTRVENDIS AC DIRIGENDIS
NAVIBVS
PARS PRIOR
COMPLECTENS
THEORIAM VNIVERSAM
DE SITV AC MOTV
CORPORVM AQVAE INNATANTIVM.
AVCTORE
LEONHARDO EVLERO
PROF. HONORARIO ACADEMIAE IMPER. SCIENT. ET
DIRECTORE ACAD. REG. SCIENT. BORVSSICAE.
INSTAR SVPPLEMENTI AD TOM. I. NOVORVM
COMMENTAR. ACAD. SCIENT. IMPER.
PETROPOLI
TYPIS ACADEMIAE SCIENTIARVM
cIↃIↃCCXLIX.

Fig. 12 *Naval Science,* 1749 (1740–1741).

dimensional bodies submerged in water. This already contains the beginnings of the mechanics of rigid bodies, which much later is to culminate in his *Theoria motus corporum solidorum seu rigidorum*, the third major treatise on mechanics (cf. section 3.3.1). The second volume applies the theory to ships, shipbuilding, and navigation.

3.2.2. Selecta Euleriana.

Selectio 1. The Basel Problem. This is the name that has become attached to the problem of determining the sum of the reciprocal squares,

$$1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots \,. \tag{3.1}$$

In modern terminology, this is called the zeta function of 2, where more generally

$$\zeta(s) = 1 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \cdots \,. \tag{3.2}$$

The problem had stumped the leading mathematicians of the time—Leibniz, Stirling, de Moivre, and all the Bernoullis—until Euler came along. Typically for Euler, he started, using his tremendous dexterity of calculation and his adroitness in speeding up slowly converging series, to calculate $\zeta(2)$ in E20 to seven decimal places (cf. Gautschi [13, sect. 2]). (Stirling, already in 1730, actually calculated the series to nine decimal places, but Euler did not yet know this.) The breakthrough came in 1735 (published as E41 in 1740) when he showed by a brilliant but daring procedure (using Newton's identities for polynomials of infinite degree!) that

$$\zeta(2) = \frac{\pi^2}{6} \,.$$

Spectacular as this achievement was, Euler went on to use the same method, with considerably more labor, to determine $\zeta(s)$ for all even $s = 2n$ up to 12. He found $\zeta(2n)$ to be always a rational number multiplied by the $2n$th power of π. It was in connection with the Basel problem that Euler in 1732 discovered a general summation procedure, found independently by Maclaurin in 1738, and promptly used it to calculate $\zeta(2)$ to twenty decimal places (cf. Gautschi [13, sect. 5.1]). Eventually, Euler

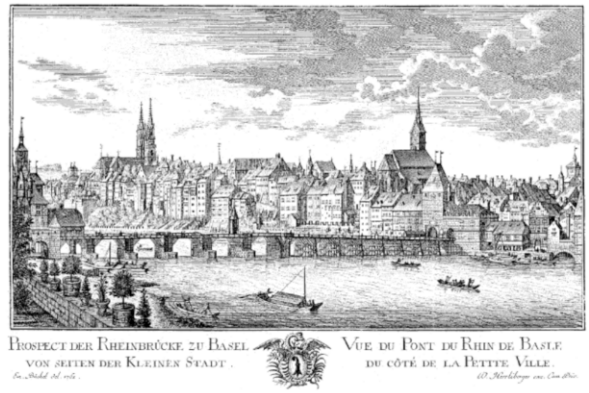


Fig. 13 *Basel, mid* 18*th century. (Reprinted with permission from the University Library of Berne, Central Library, Ryhiner Collection.)*

managed to place his approach on a more solid footing, using his own partial fraction expansion of the cotangent function, and he succeeded, in E130 (see also E212, Part II, Chap. 5, p. 324), to prove the general formula

$$\zeta(2n) = \frac{2^{2n-1}}{(2n)!}\,|B_{2n}|\pi^{2n}. \tag{3.3}$$

Here, B_{2n} are the Bernoulli numbers (introduced by Jakob Bernoulli in his *Ars conjectandi*), which Euler already encountered in his general summation formula.

Euler also tried odd values of s, but wrote in a letter to Johann Bernoulli that "the odd powers I cannot sum, and I don't believe that their sums depend on the quadrature of the circle [that is, on π]" (Fellmann [9, p. 84, footnote 56]). The problem in this case, as a matter of fact, is still open today. The Zürich historian Eduard Fueter once wrote that "where mathematical reason could not go any further, this for Euler was where the kingdom of God began." Could it be that here was an instance where Euler felt a brush with the kingdom of God?

Selectio 2. Prime Numbers and the Zeta Function. Let

$$\mathcal{P} = \{2, 3, 5, 7, 11, 13, 17, \ldots\}$$

be the set of all prime numbers, i.e., the integers > 1 that are divisible only by 1 and themselves. Euler's fascination with prime numbers started quite early and continued throughout his life, even though the rest of the mathematical world at the time (Lagrange excluded!) was rather indifferent to problems of this kind. An example of his profound insight into the theory of numbers is the discovery in 1737 (E72) of the fabulous product formula

$$\prod_{p\in\mathcal{P}} \frac{1}{1-1/p^s} = \zeta(s), \quad s > 1, \tag{3.4}$$

connecting prime numbers with the zeta function (3.2). How did he do it? Simply by starting with the zeta function and peeling away, layer after layer, all the terms whose integers in the denominators are divisible by a prime! Thus, from

$$\zeta(s) = 1 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \cdots,$$

dividing by 2^s and subtracting, one gets

$$\left(1-\frac{1}{2^s}\right)\zeta(s)=1+\frac{1}{3^s}+\frac{1}{5^s}+\frac{1}{7^s}+\cdots\ .$$

All the denominator integers divisible by 2 are gone. Doing the same with the next prime, 3, i.e., dividing the last equation by 3^s and subtracting, one gets

$$\left(1-\frac{1}{2^s}\right)\left(1-\frac{1}{3^s}\right)\zeta(s)=1+\frac{1}{5^s}+\frac{1}{7^s}+\frac{1}{11^s}+\cdots,$$

where all integers divisible by 3 are gone. After continuing in this way ad infinitum, everything will be gone except for the first term, 1,

$$\prod_{p\in\mathcal{P}}\left(1-\frac{1}{p^s}\right)\zeta(s)=1.$$

But this is the same as (3.4)!

The result provides a neat analytic proof of the fact (already known to the Greeks) that the number of primes is infinite. Indeed, since $\zeta(1)$—the harmonic series—diverges to ∞ (cf. Selectio 4), the product on the left of (3.4), if $s=1$, cannot be finite.

The formula—the beginning of "analytic number theory"—in fact paved the way to important later developments in the distribution of primes.

Selectio 3. The Gamma Function. Following a correspondence in 1729 with Goldbach, Euler in E19 considers the problem of interpolating the sequence of factorials

$$n!=1\cdot 2\cdot 3\cdots n,\quad n=1,2,3,\ldots, \tag{3.5}$$

at noninteger values of the argument. Euler quickly realized that this cannot be done algebraically, but requires "transcendentals," that is, calculus. He writes $n!$ as an infinite product,

$$\frac{1\cdot 2^n}{1+n}\cdot\frac{2^{1-n}\cdot 3^n}{2+n}\cdot\frac{3^{1-n}\cdot 4^n}{3+n}\cdot\frac{4^{1-n}\cdot 5^n}{4+n}\ \cdots, \tag{3.6}$$

which formally, by multiplying out the numerators, can be seen to be the ratio of two infinite products, $1\cdot 2\cdot 3\cdot 4\cdot 5\cdots$ and $(n+1)(n+2)(n+3)\cdots$, which indeed reduces to (3.5). Now for $n=\frac{1}{2}$, Euler manages to manipulate the infinite product (3.6) into the square root of an infinite product for $\pi/4$ due to John Wallis; therefore, $\frac{1}{2}!=\frac{1}{2}\sqrt{\pi}$. This is why Euler knew that some kind of integration was necessary to solve the problem.

By a stroke of insight, Euler takes the integral $\int_0^1 x^e(1-x)^n\mathrm{d}x$—up to the factor $1/n!$, the n-times iterated integral of x^e, where e is an arbitrary number (not the basis of the natural logarithms!)—and finds the formula

$$(e+n+1)\int_0^1 x^e(1-x)^n\mathrm{d}x=\frac{n!}{(e+1)(e+2)\cdots(e+n)}\,. \tag{3.7}$$

He now lets $e=f/g$ be a fraction, so that

$$\frac{f+(n+1)g}{g^{n+1}}\int_0^1 x^{f/g}(1-x)^n\mathrm{d}x=\frac{n!}{(f+g)(f+2g)\cdots(f+ng)}\,.$$

If $f=1$, $g=0$, then on the right we have $n!$; on the left, we have to determine the limit as $f\to 1$, $g\to 0$, which Euler takes to be the desired interpolant, since it is

meaningful also for noninteger n. Skillfully, as always, Euler carries out the limit by first changing variables, $x = t^{g/(f+g)}$, to obtain

$$\frac{f+(n+1)g}{f+g}\int_0^1 \left(\frac{1-t^{g/(f+g)}}{g}\right)^n \mathrm{d}t,$$

and then doing the limit as $g \to 0$ with $f = 1$ by the Bernoulli–l'Hôpital rule. The result is $\int_0^1 (-\ln t)^n \mathrm{d}t$. Here we can set $n = x$ to be any positive number, and thus we obtain $x! = \int_0^1 (-\ln t)^x \mathrm{d}t$, which today is written as

$$x! = \int_0^\infty \exp(-t)t^x \mathrm{d}t = \Gamma(x+1) \tag{3.8}$$

in terms of the gamma function Γ. It is easily verified that

$$\Gamma(x+1) = x\Gamma(x), \quad \Gamma(1) = 1, \tag{3.9}$$

so that indeed $\Gamma(n+1) = n!$ if n is an integer ≥ 0.

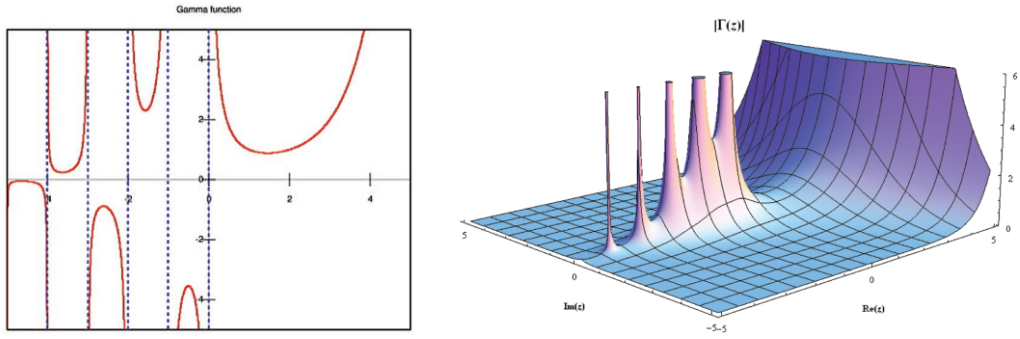


Fig. 14 *The gamma function; graph and contour map. (Per Wikipedia, permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version* 1.2 *or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. Subject to disclaimers.)*

Euler's unfailing intuition in producing the gamma function had been vindicated early in the 20th century when it was shown independently by Harald Bohr and Johannes Mollerup that there is no other function on $(0, \infty)$ interpolating the factorials if, in addition to satisfying the difference equation (3.9), it is also required to be logarithmically convex. The gamma function indeed has become one of the most fundamental functions in analysis—real as well as complex.

The integral in (3.8) is often referred to as the second Eulerian integral, the first being

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\mathrm{d}t,$$

also called the beta function. The latter can be beautifully expressed in terms of the gamma function by

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)},$$

which is nothing but (3.7) for $e = x - 1$, $n = y - 1$.

For a recent historical essay on the gamma function, see Srinivasan [28].

Selectio 4. Euler's Constant. It is generally acknowledged that, aside from the imaginary unit $i = \sqrt{-1}$, the two most important constants in mathematics are $\pi = 3.1415\ldots$, the ratio of the circumference of a circle to its diameter, and $e = 2.7182\ldots$, the basis of the natural logarithms, sometimes named after Euler. They pop up everywhere, often quite unexpectedly. The 19th-century logician Auguste de Morgan said about π that "it comes on many occasions through the window and through the door, sometimes even down the chimney." The third most important constant is undoubtedly Euler's constant γ introduced by him in 1740 in E43. Of the three together—the "holy trinity," as they are sometimes called—the last one, γ, is the most mysterious one, since its arithmetic nature, in contrast to π and e, is still shrouded in obscurity. It is not even known whether γ is rational, even though most likely it is not; if it were, say, equal to p/q in reduced form, then high-precision continued fraction calculations of γ have shown that q would have to be larger than $10^{244,663}$ (Haible and Papanikolaou [14, p. 349]).

Euler's constant arises in connection with the harmonic series $\zeta(1) = 1+\frac{1}{2}+\frac{1}{3}+\cdots$ (so called because each of its terms is the harmonic mean of the two neighboring terms) and is defined as the limit

$$\gamma = \lim_{n\to\infty}\left(1+\frac{1}{2}+\frac{1}{3}\cdots+\frac{1}{n}-\ln n\right) = 0.57721\ldots. \tag{3.10}$$

It has been known as early as the 14th century that the harmonic series diverges, but a rigorous proof of it is usually attributed to Jakob Bernoulli, who also mentioned another proof by his younger brother Johann, which, however, is not entirely satisfactory. At any rate, Euler, in defining his constant and showing it to be finite, puts in evidence not only the divergence of the harmonic series, but also its logarithmic rate of divergence. Beyond this, using his general summation formula (mentioned in Selectio 1), he computes γ to 16 correct decimal places (cf. Gautschi [13, sect. 5.2]), and to equally many decimals the sum of the first million terms of the harmonic series! Since later (in 1790) Lorenzo Mascheroni also considered Euler's constant, gave it the name γ, and computed it to 32 decimal places (of which, curiously, the 19th, 20th, and 21st are incorrect), the term "Euler–Mascheroni constant" is also in use. As of today, it appears that γ has been computed to 108 million decimal places, compared to over 2×10^{11} decimals for π and 50.1 billion for e.

An inspiring tale surrounding Euler's constant can be found in Havil [15], and a rather encyclopedic account in Krämer [18].

After all these spectacular achievements, the numerous other memoirs written on many different topics, and his responsibilities at the Academy, it is incredible that Euler still had the time and stamina to write a 300-page volume on elementary arithmetic for use in the St. Petersburg gymnasia. How fortunate were those St. Petersburg kids for having had Euler as their teacher!

3.3. Berlin. Next to some 280 memoirs, many quite important, and consultation on engineering and technology projects, this period saw the creation of a number of epochal scientific treatises and a highly successful and popular work on the philosophy of science.

3.3.1. Major Works. The brachistochrone problem—finding the path along which a mass point moves under the influence of gravity down from one point of a vertical plane to another (not vertically below) in the shortest amount of time—is an early

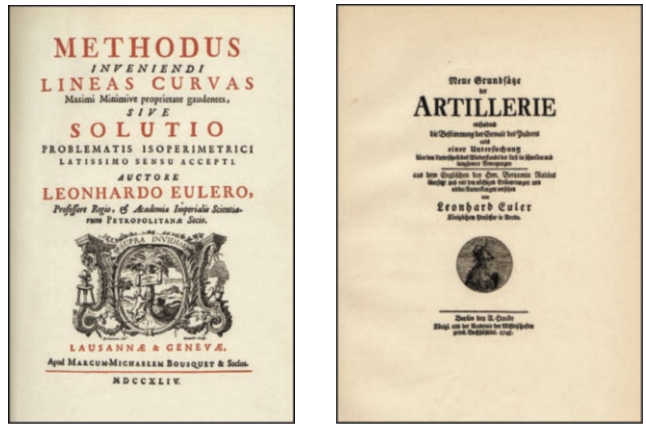


Fig. 15 *Calculus of Variations,* 1744*, and Artillerie,* 1745*. (Reprinted with permission from Birkhäuser Verlag.)*

example of an optimization problem, posed by Johann Bernoulli, which seeks a function (or a curve) that renders optimal an analytic expression that depends on this function. In 1744 (E65), and later in 1766 (E296) adopting an improved approach of Lagrange, Euler vastly generalizes this problem, thereby creating an entirely new branch of mathematics, called (already by Euler) the "calculus of variations." He derives his famous Euler equation: a necessary condition in the form of a differential equation that any solution of the problem must satisfy. Typically for Euler, he illustrates this by many—some hundred!—examples, among them the principle of least action that caused so much turmoil in the mid-1700s (cf. section 2.3).

Two smaller treatises, one on planetary and cometary trajectories (E66) and another on optics (E88), appeared at about the same time (1744, resp., 1746). The latter is of historical interest insofar as it started the debate of Newton's particle versus Euler's own wave theory of light.

In deference to his master, king Frederick II, Euler translated an important work on ballistics by the Englishman Benjamin Robins, even though the latter had been unfairly critical of Euler's *Mechanica* of 1736. He added, however, so many commentaries and explanatory notes (also corrections!) that the resulting book—his *Artillerie* of 1745 (E77)—is about five times the size of the original. Niklaus Fuss in his 1783 *Eulogy* of Euler (cf. *Opera omnia*, Ser. I, Vol. 1, pp. xliii–xcv) remarks: "... the only revenge [Euler] took against his adversary because of the old injustice consists in having made [Robins's] work so famous as, without him, it would never have become."

The two-volume *Introductio in analysin infinitorum* of 1748 (E101, E102) together with the *Institutiones calculi differentialis* of 1755 (E212) and the three-volume *Institutiones calculi integralis* of 1768–1770 (E342, E366, E385)—a "magnificent trilogy" (Fellmann [9, sect. 4])—establishes analysis as an independent, autonomous discipline, and represents an important precursor of analysis as we know it today.

In the first volume of the *Introductio*, after a treatment of elementary functions, Euler summarizes his many discoveries in the areas of infinite series, infinite products, partition of numbers, and continued fractions. On several occasions, he uses the fundamental theorem of algebra, clearly states it, but does not prove it. He develops a clear concept of function—real- as well as complex-valued—and emphasizes the fundamental role played in analysis by the number e and the exponential and logarithm

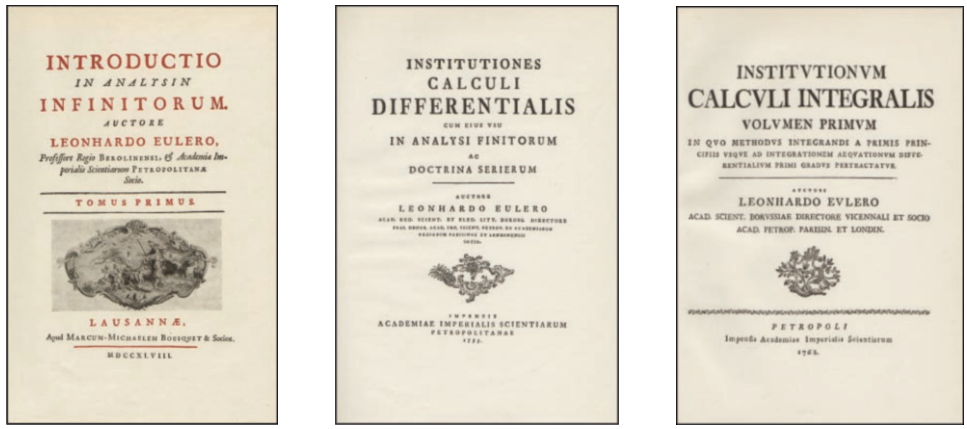


Fig. 16 *Infinitesimal Analysis,* 1748*, and Differential and Integral Calculus,* 1755*,* 1763*,* 1773*. (Reprinted with permission from Birkhäuser Verlag.)*

functions. The second volume is devoted to analytic geometry: the theory of algebraic curves and surfaces.

Differential Calculus also has two parts, the first being devoted to the calculus of differences and differentials, the second to the theory of power series and summation formulae, with many examples given for each. Chapter 4 of the second part, incidentally, contains the first example, in print, of a Fourier series; cf. also p. 297 of the *Opera omnia*, Ser. I, Vol. 10. Another chapter deals with Newton's method, and improvements thereof, for solving nonlinear equations, and still another with criteria for algebraic equations to have only real roots.

The three-volume *Integral Calculus* is a huge foray into the realm of quadrature and differential equations. In the first volume, Euler treats the quadrature (i.e., indefinite integration) of elementary functions and techniques for reducing the solution of linear ordinary differential equations to quadratures. In the second volume, he presents, among other things, a detailed theory of the important linear second-order differential equations, and in the third volume a treatment, to the extent known at the time (mostly through Euler's own work), of linear partial differential equations. A fourth volume, published posthumously in 1794, contains supplements to the preceding volumes. Euler's method—a well-known approximate method for solving arbitrary first-order differential equations, and the more general Taylor series method, are embedded in Chapter 7 of the second section of Volume 1.

Euler's program for mechanics (cf. section 3.2.1) progressed steadily as he tackled the problem of developing a theory of the motion of solids. An important milestone in this effort was the memoir E177 in which was stated for the first time, in full generality, what today is called Newtonian mechanics. The great treatise *Theoria motus corporum solidorum seu rigidorum* (E289) which followed in 1765, also called the "Second *Mechanics*," represents a summary of Euler's mechanical work up to this time. In addition to an improved exposition of his earlier mechanics of mass points (cf. section 3.2.1), it now contains the differential equations (Euler's equations) of motion of a rigid body subject to external forces. Here, Euler introduces the original idea of employing two coordinate systems—one fixed, the other moving, attached to the body—and deriving differential equations for the angles between the respective


THEORIA MOTVS
CORPORVM
SOLIDORVM seu RIGIDORVM
EX
PRIMIS NOSTRAE COGNITIONIS PRINCIPIIS
STABILITA
ET AD OMNES MOTVS,
QVI IN HVIVSMODI CORPORA CADERE POSSVNT,
ACCOMMODATA.
AVCTORE
LEONH. EVLERO
ROSTOCHI ET GRYPHISWALDIAE

Fig. 17 *Theoria motus corporum,* 1765. *(Reprinted with permission from Birkhäuser Verlag.)*

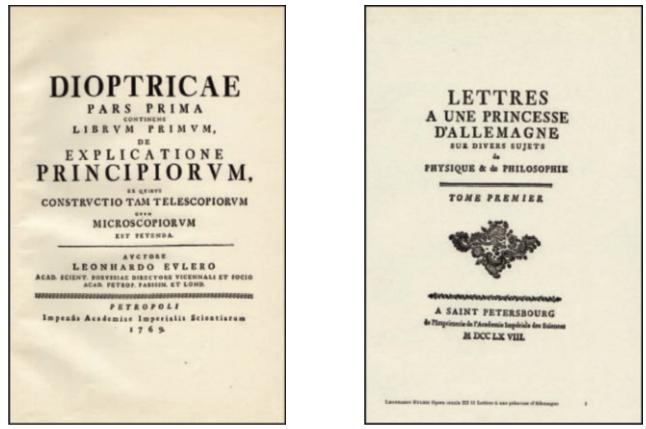

DIOPTRICAE
PARS PRIMA
CONTINENS
LIBRVM PRIMVM,
DE
EXPLICATIONE
PRINCIPIORVM,
CONSTRVCTIO TAM TELESCOPIORVM
QVAM
MICROSCOPIORVM
EST PETENDA.
AVCTORE
LEONHARDO EVLERO
PETROPOLI
Impensis Academiae Imperialis Scientiarum
1769
LETTRES
A UNE PRINCESSE
D'ALLEMAGNE
SUR DIVERS SUJETS
de
PHYSIQUE & de PHILOSOPHIE
TOME PREMIER
A SAINT PETERSBOURG
M. DCC. LX. VIII.

Fig. 18 *Optics,* 1769–1771*, and Letters,* 1768*,* 1772 (1760–1762). *(Reprinted with permission from Birkhäuser Verlag.)*

coordinate axes, now called the Euler angles. The intriguing motion of the spinning top is one of many examples worked out by Euler in detail.

Later, in 1776, Euler returns to mechanics again with his seminal work E479, where one finds the definitive formulation of the principles of linear and angular momentum.

Throughout his years in Berlin and beyond, Euler was deeply occupied with geometric optics. His memoirs and books on this topic, including the monumental three-volume *Dioptrics* (E367, E386, E404), written mostly while still in Berlin, fill no fewer than seven volumes in his *Opera omnia*. A central theme and motivation of this work was the improvement of optical instruments like telescopes and microscopes, notably ways of eliminating chromatic and spherical aberration through intricate systems of lenses and interspaced fluids.

Euler's philosophical views on science, religion, and ethics are expressed in over 200 letters written between 1760 and 1762 (in French) to a German princess and published later in 1768 and 1772 (E343, E344, E417). (For a recent edition of these letters,

see Euler [8].) While Euler's role as a philosopher may be controversial (even his best friend Daniel Bernoulli advised him to better deal with "more sublime matters"), his *Letters*, written with extreme clarity and also accessible to people not trained in the sciences, "even to the gentle sex," as Fuss remarks in his *Eulogy*, became an instant success and were translated into all major languages.

3.3.2. Selecta Euleriana.

Selectio 5. The Königsberg Bridge Problem. The river Pregel, which flows through the Prussian city of Königsberg, divides the city into an island and three distinct land masses, one in the north, one in the east, and one in the south. There are altogether seven bridges, arranged as shown in green on the left of Figure 19, connecting the three land masses with each other and with the island. The problem is this: Can one take a stroll from one point in the city to another by traversing each bridge exactly once? In particular, can one return to the starting point in the same manner?

Evidently, this is a problem that cannot be dealt with by the traditional methods of analysis and algebra. It requires a new kind of analysis that deemphasizes metric properties in favor of positional properties. Euler solved the problem in 1735, published as E53 in 1741, by showing that such paths cannot exist. He does this by an ingenious process of abstraction, associating with the given land and bridge configuration (what today is called) a connected graph, i.e., a network of vertices and connecting edges, each vertex representing a piece of land and each edge a bridge connecting the respective pieces of land. In the problem at hand, there are four distinct pieces of land, hence four vertices, and they are connected with edges as shown on the right of Figure 19. It is obvious what is meant by a path along edges from one vertex to another. A closed path is called a circuit, and paths or circuits are (today) called Eulerian if each edge is traversed exactly once.

Euler recognized that in modern terminology a crucial concept here is the *degree* of a vertex, i.e., the number of edges emanating from it. If, in an arbitrary connected

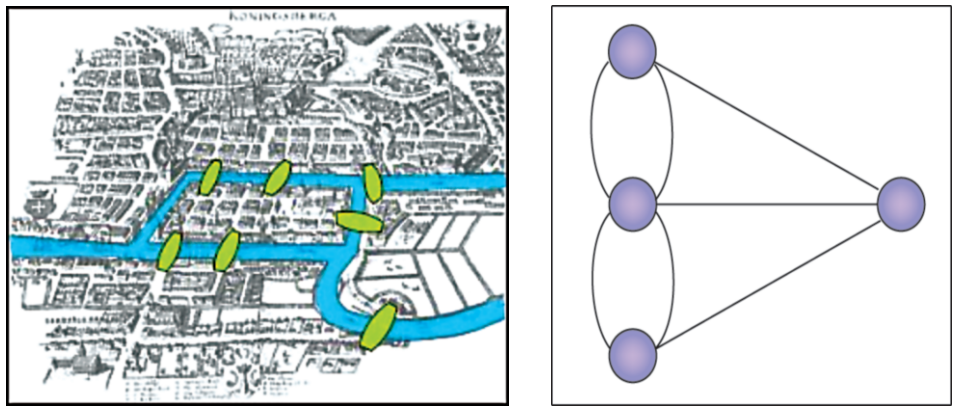


Fig. 19 *The Königsberg bridge problem. (Left image created by Bogdan Giuşcă, as displayed in the Wikipedia article "Leonhard Euler." Per Wikipedia, permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version* 1.2 *or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. Subject to disclaimers.)*

graph, n denotes the number of vertices of odd degree, he in effect proves that (a) if $n = 0$, the graph has at least one Eulerian circuit, and he indicates how to find it; (b) if $n = 2$, it has at least one Eulerian path, but no circuit, and again he shows us how to find it; (c) if $n > 2$, it has neither. (The case $n = 1$ is impossible.) Since the Königsberg bridge graph has $n = 4$, we are in case (c), hence it is impossible to traverse the city in the manner required in the problem.

Here again, like in the calculus of variations, one can admire Euler's powerful drive and capacity of starting with a concrete example and deriving from it, by a process of sweeping generalization, the beginnings of a whole new theory, in the present case, the theory of graphs and topological networks.

Selectio 6. Euler's Buckling Formula (1744). In a first supplement to his *Methodus* (cf. Figure 15, left), Euler applies the calculus of variations to elasticity theory, specifically to the bending of a rod subject to an axial load. He derives the critical load under which the rod buckles. This load depends on the stiffness constant of the material, on the way the rod is supported at either end, and it is inversely proportional to the square of the length of the rod. A particular configuration of two rods loaded on top by a connecting bar (assumed to be of infinite stiffness) is shown in Figure 20, during the initial phase (left), and at the time of buckling (right). Here, the top end of the rods is slidably supported and the bottom end clamped. For a video, see http://epubs.siam.org/sam-bin/getfile/SIREV/articles/70271_02.avi.

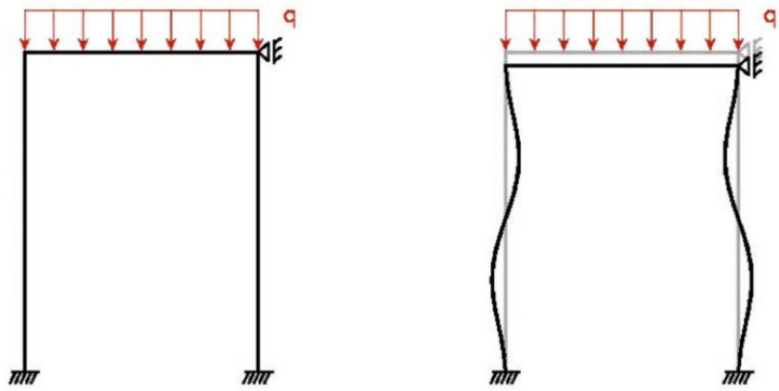


Fig. 20 *The buckling of a rod. (Images and video courtesy of Wolfgang Ehlers.)*

The critical load is the first elasto*static* eigenvalue of the problem. Euler also calculates the elasto*kinetic* eigenvalues, the eigenfrequencies of the rod's transversal oscillations, and the associated eigenfunctions, which determine the shapes of the deformed rod.

Selectio 7. Euler Flow. In a series of three memoirs, E225–E227, all published in 1757, and another three papers (E258, E396, E409), Euler gave his definitive treatment of continuum and fluid mechanics, the culmination of a number of earlier memoirs on the subject. It contains the celebrated Euler equations, expressing the conservation of mass, momentum, and energy. In two (three) dimensions, these constitute a system of four (five) nonlinear hyperbolic partial differential equations, which have to be solved, given appropriate initial and boundary conditions. Naturally, in Euler's time,

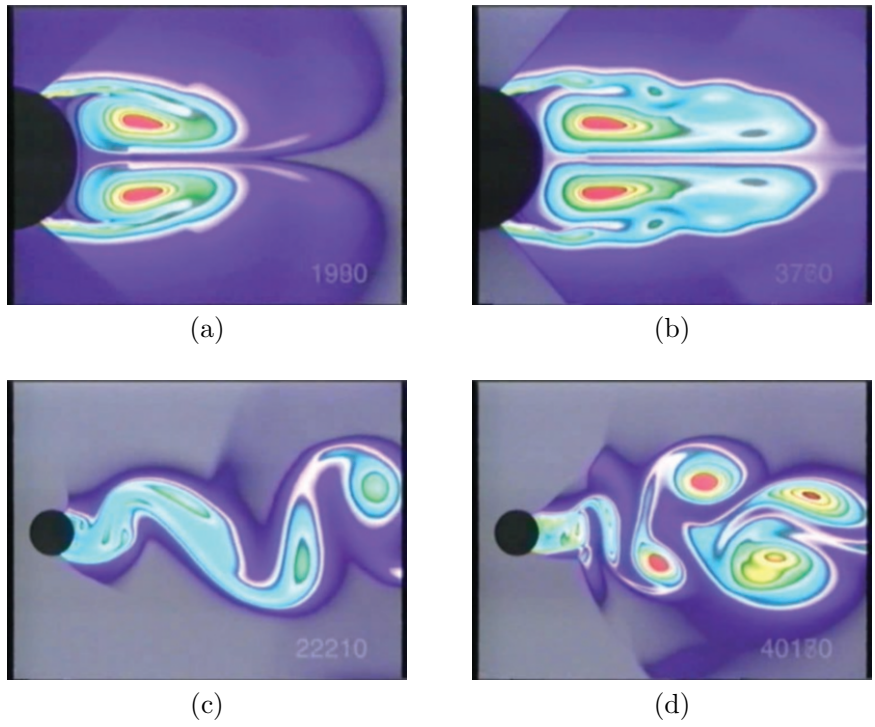


Fig. 21 *Transonic Euler flow at Mach .85 about a cylinder. (Images and video courtesy of Nicola Botta.)*

this was virtually impossible to do, except in very special cases, and indeed Euler in the introduction to E226 had to write that "if there remain any difficulties, they shall not be on the side of mechanics, but solely on the side of analysis: for this science has not yet been carried to the degree of perfection which would be necessary in order to develop analytic formulae which include the principles of the motion of fluids." Nowadays, however, the Euler equations are widely being used in computer simulation of fluids.

An example is the asymmetric flow of a compressible, inviscid fluid about a circular cylinder at transonic speed, calculated in 1995 by Botta [4]. Four color-coded snapshots of the two-dimensional flow (vorticity contour lines), as it develops behind the cylinder, are shown in Figure 21: (a) the onset of the flow, (b) a regimen of Kelvin–Helmholtz instability, (c) the flow after breakdown of symmetry, and (d) the formation of vortex pairs. (The scaling of (c) and (d) differs from that of (a) and (b).) For the complete Euler-flow video, see http://epubs.siam.org/sambin/getfile/SIREV/articles/70271_03.avi.

Selectio 8. Euler's Polyhedral Formula (1758). In a three-dimensional convex polyhedron (not necessarily regular), let V denote the number of vertices, E the number of edges, and F the number of faces. Thus, in the case of an octahedron (cf. Figure 22), one has $V = 6, E = 12$, and $F = 8$. Mentioned in 1750 in a letter to Goldbach, and later published in E231, Euler proves for the first time the extremely simple but stunning formula

$$V - E + F = 2. \tag{3.11}$$

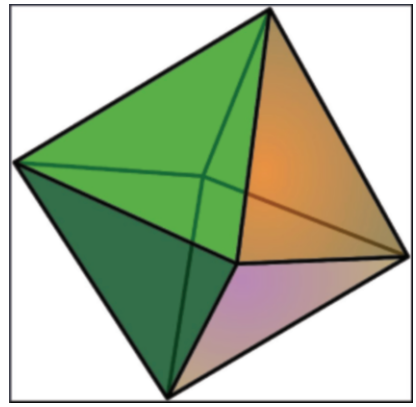


Fig. 22 *Octahedron. (From the Wikipedia article "Octahedron." Per Wikipedia, permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version* 1.2 *or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. Subject to disclaimers.)*

The way he did it is to chop off triangular pyramids from the polyhedron, one after another, in such a manner that the sum on the left of (3.11) remains the same. Once he got it chopped down to a tetrahedron, that sum is easily seen to be 2. (For a critical and historical review of Euler's proof, see Francese and Richeson [11].) Descartes, some 100 years earlier, already knew, but did not prove, something close to the formula (3.11).

The expression on the left-hand side of (3.11) is an example of an Euler characteristics, a topological invariant for polyhedra. Euler characteristics have been defined for many other topological spaces and today still come up often in homological algebra.

The generalization to higher-dimensional polytopes leads to what is called Euler–Poincaré characteristics, where the pattern of alternating signs can be seen to come from the dimensionality of the respective facets, something already noted in 1852 by another Swiss mathematician, Ludwig Schläfli [25, sect. 32].

Selectio 9. Euler and *q*-Theory. The story here begins with a letter Euler wrote in 1734 to Daniel Bernoulli, in which he considered the (somewhat bizarre) problem of interpolating the common logarithm $\log x$ at the points $x_r = 10^r$, $r = 0, 1, 2, \ldots$. He essentially writes down Newton's interpolation series $S(x)$ (without mentioning Newton by name) and remarks that, when $x = 9$, the series converges quickly, but to a wrong value, $S(9) \neq \log 9$ (cf. Gautschi [12]). Rather than losing interest in the problem, Euler must have begun pondering the question about the nature of the limit function $S(x)$: what is it, if not the logarithm?

Almost twenty years later, in 1753, he returned to this problem in E190, now more generally for the logarithm to base $a > 1$, and studied the respective limit function $S(x; a)$ in great detail. Intuitively, he must have perceived its importance. Today we know (Koelink and Van Assche [17]) that it can be thought of as a q-analogue of the logarithm, where $q = 1/a$, and some of the identities derived by Euler (in part already contained in Vol. 1, Chap. 16 of his *Introductio*) are in fact special cases of the q-binomial theorem—a centerpiece of q-theory in combinatorial analysis and physics. Thus, Euler must be counted among the precursors of q-theory, which was only developed about 100 years later by Heinrich Eduard Heine.

Selectio 10. The Euler–Fermat Theorem and Cryptology. Let $\mathbb{N}$ be the set of positive integers, and $\varphi(n)$, $n \in \mathbb{N}$, Euler's totient function, that is, the number of integers $1, 2, 3, \ldots, n$ coprime to n. The Euler–Fermat theorem, published 1763 as E271, states that for any $a \in \mathbb{N}$ coprime to n,

$$a^{\varphi(n)} \equiv 1 \pmod{n}. \tag{3.12}$$

It generalizes the "little Fermat" theorem, which is the case $n = p$ a prime number, and therefore $\varphi(p) = p - 1$.

In cryptography, one is interested in the secure transmission of messages, whereby a message M is transmitted from a sender to the receiver in encrypted form: The sender encodes the message M into E, whereupon the receiver has to decode E back into M. It is convenient to think of M as a number in $\mathbb{N}$, for example, the number obtained by replacing each letter, character, and space in the text by its ASCII code. The encrypted message E is then $E = f(M)$, where $f : \mathbb{N} \to \mathbb{N}$ is some function on $\mathbb{N}$. The problem is to find a function f that can be computed by the general public but is extremely difficult to invert (i.e., to obtain M from E), unless one is in the possession of a secret key associated with the function f.

A solution to this problem is the now widely used RSA encryption scheme (named after its inventors R. Rivet, A. Shamir, and L. Adleman). To encode the message M, one selects two distinct (and very large) prime numbers p, q and defines a "modulus" $n = pq$ assumed to be larger than M. Then an integer e, $1 < e < \varphi(n)$, is chosen with e coprime to $\varphi(n)$. The numbers n, e form the "public key," i.e., they are known to the general public. The encoded message M is $E = f(M)$, where $f(M) \equiv M^e \pmod{n}$. The "private key" is n, d, where d is such that $de \equiv 1 \pmod{\varphi(n)}$. To compute d, one needs to know p and q, since $n = pq$, $\varphi(n) = (p-1)(q-1)$. The general public, however, knows only n, so must factor n into prime numbers to get a hold of p, q. If n is sufficiently large, say $n > 10^{300}$, this, today, is virtually impossible. The person who selected p and q, on the other hand, is in possession of d, and can decode the ciphertext E as follows,

$$E^d \equiv (M^e)^d \pmod{n} \equiv M^{ed} \pmod{n} \equiv M^{N\varphi(n)+1} \pmod{n}, \quad N \in \mathbb{N},$$

by the choice of d. Using now the Euler–Fermat theorem (3.12), with $a = M^N$ (almost certainly coprime to $n = pq$ or can be made so), one gets

$$E^d \equiv Ma^{\varphi(n)} \pmod{n} \equiv M \pmod{n} = M,$$

since $M < n$. (It is true that M, e, n, and d are typically very large numbers so that the computations described may seem formidable. There are, however, efficient schemes to execute them; see, e.g., Silverman [26, Chaps. 16, 17].)

3.4. Second St. Petersburg Period. This may well be Euler's most productive period, with well over 400 published works to his credit, not only on each of the topics already mentioned, but also on geometry, probability theory and statistics, cartography, and even widow's pension funds and agriculture. In this enormous body of work there figure three treatises on algebra, lunar theory, and naval science, and what appear to be fragments of major treatises on number theory (E792), natural philosophy (E842), and dioptrics (E845).

3.4.1. Major Works. Soon into this second St. Petersburg period, another of Euler's "bestsellers" appeared: the *Vollständige Anleitung zur Algebra* (E387, E388),


Vollſtändige
Anleitung
zur
Algebra
von
Hrn. Leonhard Euler.
Erſter Theil.
Von den verſchiedenen Rechnungs-Arten,
Verhältniſſen und Proportionen.
St. Petersburg,
gedruckt bey der Kayſ. Acad. der Wiſſenſchaften 1770.

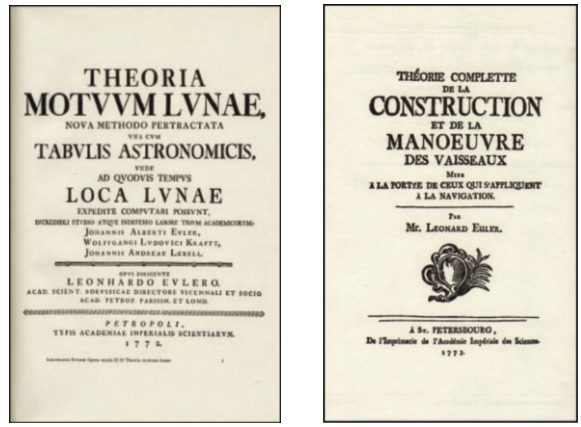
Fig. 23 *Algebra,* 1770. *(Reprinted with permission from Birkhäuser Verlag.)*

or *Algebra* for short. Even before publication of the German original, a translation into Russian came out, and translations into all major languages were soon to follow. (The French translation by Johann III Bernoulli includes a long supplement by Lagrange containing an exposé on the arithmetic theory of continued fractions and many addenda to the last section of the *Algebra* dealing with Diophantine equations.)

Euler wrote this 500-page work to introduce the absolute beginner into the realm of algebra. He dictated the work to a young man—a tailor's apprentice—whom he brought with him from Berlin, and who (according to the preface of the work) "was fairly good at computing, but beyond that did not have the slightest notion about mathematics .... As far as his intellect is concerned, he belonged among the mediocre minds." Nevertheless, it is said that, when the work was completed, he understood everything perfectly well and was able to solve algebraic problems posed to him with great ease.

It is indeed a delight to witness in this work Euler's magnificent didactic skill, to watch him progress in ever so small steps from the basic principles of arithmetic to algebraic (up to quartic) equations, and finally to the beautiful art of Diophantine analysis. Equally delightful is to see how the theory is illustrated by numerous well-chosen examples, many taken from everyday life.

The orbit of the moon, with all its irregularities, had long fascinated mathematicians like Clairaut and d'Alembert, as well as Euler, who already in 1753 published his *Theoria motus lunae* (E187), the "First Lunar Theory." The theory he developed there, while tentative, provided astronomers with formulae needed to prepare lunar tables, which in turn served seafaring nations for over a century with accurate navigational aids. Euler's definitive work on the subject, however, is his "Second Lunar Theory" (E418) of 1772, a monumental work dealing in a more effective way than before with the difficult three-body problem, i.e., the study of the motion of three bodies—in this case the sun, the earth, and the moon, thought of as point masses—moving under the influence of mutual gravitational forces. Already Newton is reputed to have said that "an exact solution of the three-body problem exceeds, if I am not mistaken, the power of any human mind." Today it is known, indeed, that an exact solution is not possible. Euler grapples with the problem by introducing appropriate variables, again choosing two coordinate systems—one fixed, the other moving—applying processes of successive approximation, and making use, when needed, of observational data.


THEORIA
MOTVVM LVNAE,
NOVA METHODO PERTRACTATA
VNA CVM
TABVLIS ASTRONOMICIS,
QVIBVS
AD QVOVIVS TEMPVS
LOCA LVNAE
EXPEDITE COMPVTARI POSSVNT.
INCREDIBILI STVDIO ATQVE INDEFESSO LABORE TRIVM ACADEMICORVM
JOHANNIS ALBERTI EULER,
WOLFGANGI LVDOVICI KRAFFT,
JOHANNIS ANDREAE LEXELL.
OPVS DILIGENTER
LEONHARDO EVLERO,
ACAD. SCIENT. BORVSSICAE DIRECTORE TECHNOLOGIAE ET SOCIO
ACAD. PETROP. PARISIIN. ET LOND.
PETROPOLI,
TYPIS ACADEMIAE IMPERIALIS SCIENTIARVM.
1772.

THÉORIE COMPLETTE
DE LA
CONSTRUCTION
ET DE LA
MANOEUVRE
DES VAISSEAUX
MISE
A LA PORTEE DE CEUX QUI S'APPLIQUENT
A LA NAVIGATION.
PAR
Mr. LEONARD EULER.
A St. PETERSBOURG,
De l'Imprimerie de l'Académie Impériale des Sciences.
1773.

Fig. 24 *Second Lunar Theory,* 1772*, and Second Theory of Ships,* 1773*. (Reprinted with permission from Birkhäuser Verlag.)*

According to L. Courvoisier (cf. *Opera omnia*, Ser. II, Vol. 22, p. xxviii), "all later progress in celestial mechanics is based, more or less, on the ideas contained in the works of Euler, [and the later works of] Laplace and Lagrange."

The *Théorie complete de la construction et de la manœuvre des vaisseaux* (E426), also called the "Second Theory of Ships," is a work that treats the topic indicated in the title for people having no or little mathematical knowledge, in particular for the sailors themselves. Not surprisingly, given the level of presentation and the author's extraordinary didactic skill, the work proved to be very successful. The French maritime and finance minister (and famous economist) Anne Robert Jacques Turgot proposed to King Louis XVI that all students in marine schools (and also those in schools of artillery) be required to study Euler's relevant treatises. Very likely, Napoléon Bonaparte was one of those students. The king even paid Euler 1,000 rubles for the privilege of having the works reprinted, and czarina Catherine II, not wanting to be outdone by the king, doubled the amount and pitched in an additional 2,000 rubles!

3.4.2. Selecta Euleriana.

Selectio 11. Partition of Numbers. Euler's interest in the partition of numbers, i.e., in expressing an integer as a sum of integers from some given set, goes back to 1740 when Philippe Naudé the younger, of the Berlin Academy, in a letter to Euler asked in how many ways the integer 50 can be written as a sum of seven different positive integers. This gave rise to a series of memoirs, spanning a time interval of about 20 years, beginning with E158, published (with a delay of 10 years) in 1751, and ending with E394, published in 1770. In this work, Euler almost single-handedly created the theory of partition. A systematic exposition of part of this work can also be found in Volume 1, Chapter 16, of his *Introductio* (cf. section 3.3.1) and relevant correspondence with Niklaus I Bernoulli in the *Opera omnia*, Ser. IVA, Vol. 2, pp. 481–643, especially pp. 518, 537ff, 555ff.

Euler, as de Moivre before him (cf. Scharlau [24, p. 141f]), attacked problems of this type by a brilliant use of generating functions and formal power series. Thus, in the case of Naudé's inquiry, in Euler's hands this becomes the problem of finding the coefficient of z^7x^{50} in the expansion of $(1+xz)(1+x^2z)(1+x^3z)(1+x^4z)\cdots$, for

which Euler finds the answer 522, "a most perfect solution of Naudé's problem," as he proudly wrote (at the end of section 19 of E158). In the context of "unrestricted partitions," Euler in the penultimate paragraph of E158 surprises us with the marvelous expansion

$$(1-x)(1-x^2)(1-x^3)(1-x^4)\cdots = \sum_{n=-\infty}^{\infty} (-1)^n x^{n(3n-1)/2},$$

which he conjectured as early as 1742 by numerical computation, and then labored on it for almost ten years to find a proof (in E244, a "masterpiece" according to C. G. J. Jacobi). He used (in E175) the expansion to obtain his astonishing recurrence relation for $s(n)$, the sum of divisors of n (including 1 and n), and (in E191) the reciprocal expansion to obtain a similar recurrence for the partition function $p(n)$, the number of ways n can be written as a sum of natural numbers. In E394, Euler considers the problem of how many ways any given number can be thrown by n ordinary dice. He shows that the answer is given by the appropriate coefficient in the expansion of $(x+x^2+x^3+x^4+x^5+x^6)^n$. Of course, Euler also solves the same problem for more general dice having an arbitrary number of sides, which may even differ from die to die.

Euler's magnificent work on partitions has not found much response among his contemporaries; it was only in the 20th century that his work was continued and significantly expanded by such mathematicians as Ramanujan, Hardy, and Rogers.

Selectio 12. Euler's Gear Transmission. In connection with the design of water turbines, Euler developed optimal profiles for teeth in cogwheels that transmit motion with a minimum of resistance and noise (E330, OII.17, pp. 196–219). These profiles involve segments of circular evolvents as shown in Figure 25. For the gear in action, see the video at http://epubs.siam.org/sam-bin/getfile/SIREV/articles/70271_04.avi.

The technical realization of this design took shape only later in what is called the involute gear. Euler not only is the inventor of this kind of gear, but he also anticipated the underlying geometric equations now usually called the Euler–Savary equations.

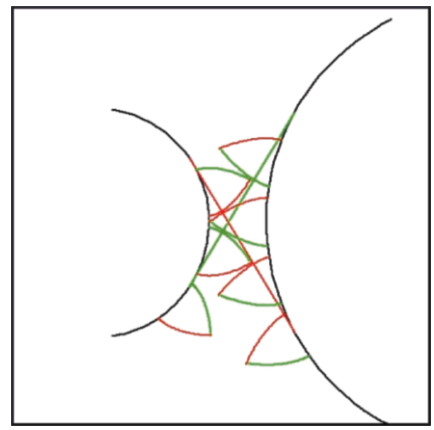


Fig. 25 *Euler gear, 1767. (Image and video courtesy of Bert Jüttler.)*

Selectio 13. Euler's Disk. In a number of memoirs (E257, E292, E336, E585) from the 20-year period 1761–1781, Euler analyzes the motion of a rigid body around a moving axis, including the effects of friction. An interesting example is the Euler disk, a circular (homogeneous) metal disk being spun on a clean smooth surface. At first, it will rotate around its vertical axis, but owing to friction, the axis is beginning to tilt and the disk to roll on a circular path. The more the axis is tilting, the wider the circular path and the higher the pitch of the whirring sound emitted by the point of contact of the disk with the surface. Thus, paradoxically, the speed of the motion seems to increase, judging from the rising pitch of the sound, although energy is being dissipated through friction. The disk, eventually, comes to an abrupt halt, flat on the surface.

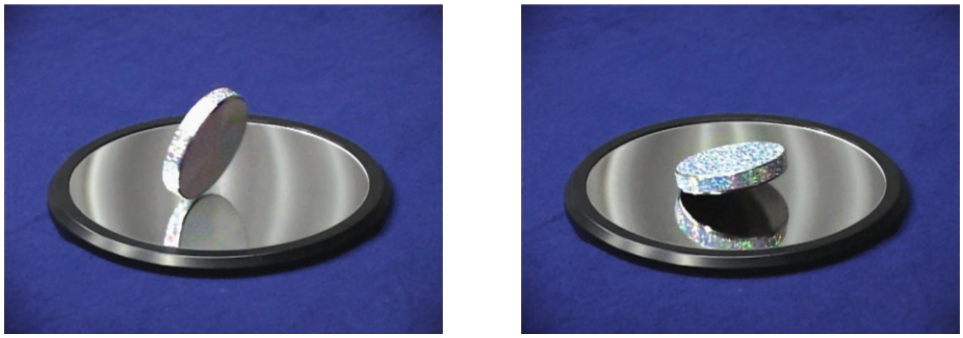


Fig. 26 *Euler disk. (Produced by Multimedia Services, ETH Zürich.)*

Two snapshots, one from the initial phase and the other from a later phase of the motion, are shown in Figure 26 on the left and right, respectively. For the complete Euler-disk video, see http://epubs.siam.org/sam-bin/getfile/SIREV/articles/70271_05.avi.

The key toward explaining the motion are Euler's equations, a set of differential equations involving the Euler angles and other parameters. The technical details of the motion, though, are still being analyzed today (cf., e.g., Le Saux, Leine, and Glocker [19] and the literature cited therein).

4. The Man.

4.1. Personality. From various testimonials of Euler's contemporaries, and also, of course, from Euler's extensive correspondence, one can form a fairly accurate picture of Euler's personality. A valuable source is the eulogy of Niklaus Fuss (*Opera omnia*, Ser. I, Vol. 1, pp. xliii–xcv), who during the last ten years of Euler's life had seen him regularly, almost on a daily basis, as one of his assistants. Also based on personal acquaintance is the eulogy of the marquis Nicolas de Condorcet (*Opera omnia*, Ser. III, Vol. 12, pp. 287–310), which, however, deals more with Euler's work. Euler comes across as a modest, inconspicuous, uncomplicated, yet cheerful and sociable person. He was down-to-earth and upright; "honesty and uncompromising rectitude, acknowledged Swiss national virtues, he possessed to a superior degree," writes Fuss. Euler never disavowed—in fact was proud of—his Swiss heritage. Fuss (who also originated from Basel) recalled that Euler "always retained the Basel dialect with all the peculiarities of its idiom. Often he amused himself to recall for me certain provincialisms and figures of speech, or mix into his parlance Basel expressions whose

use and meaning I had long forgotten." He even made sure that he and his children retained the Basel civic rights.

Feelings of rancor, due to either priority issues or unfair criticism, were totally foreign to Euler. When Maclaurin, for example, discovered the well-known summation formula which Euler obtained six years earlier, Euler did not object, let alone complain, when for some time the formula was generally referred to as the "Maclaurin summation formula." It may even have pleased him that others hit upon the same fortunate idea. In due time, of course, the formula became justly known as the Euler–Maclaurin summation formula. Another example is Maupertuis's claim for the principle of least action (cf. section 2.3), which Euler had already enunciated before, much more clearly and exhaustively; yet Euler remained supportive of Maupertuis. Euler's forgiving way of reacting to Robins's criticism of the *Mechanica* has already been mentioned in section 3.3.1.

Sharing ideas with others and letting others take part in the process of discovery is another noble trait of Euler. A case in point is the way he put on hold his already extensive work on hydrodynamics, so that his friend Daniel Bernoulli, who was working on the same topic, could complete and publish his own *Hydrodynamics* first! It became a classic.

An important aspect of Euler's personality is his religiousness: By his upbringing in the Riehen parish environment, he was a devout protestant and even served as an elder in one of the protestant communities in Berlin. Indeed, he felt increasingly uncomfortable and frustrated in the company of so many "free-spirits"—as he and others called the followers of French enlightenment—that populated and began to dominate the Berlin Academy. He gave vent to his feelings in the (anonymously published) pamphlet *Rettung der göttlichen Offenbarung gegen die Einwürfe der Freygeister* (E92, *Opera omnia*, Ser. III, Vol. 12, pp. 267–286). This frustration may well have had something to do with his atypically harsh treatment of Johann Samuel König in the dispute about the Euler/Maupertuis principle of least action (cf. section 2.3). It may also have been one, and not the least, of the reasons why Euler left Berlin and returned to St. Petersburg.

4.2. Intellect. There are two outstanding qualities in Euler's intellect: a phenomenal memory, coupled with an unusual power of mental calculation, and an ease in concentrating on mental work irrespective of any hustle and bustle going on around him: "A child on the knees, a cat on his back, that's how he wrote his immortal works," recounts Dieudonné Thiébault, the French linguist and confidant of Frederick II. With regard to memory, the story is well known of Euler's ability, even at an advanced age, to recite by heart all the verses of Virgil's *Aeneid*. One of these, Euler says in a memoir, has given him the first ideas in solving a problem in mechanics. Niklaus Fuss also tells us that during a sleepless night, Euler mentally calculated the first six powers of all the numbers less than twenty (less than 100 in Condorcet's account), and several days later was able to recall the answers without hesitation. "Euler calculates as other people breathe," Condorcet wrote.

Equipped with such intellectual gifts, it is not surprising that Euler was extremely well read. In Fuss's words,

> he possessed to a high degree what commonly is called erudition; he had read the best writers of antique Rome; the older mathematical literature was very well known to him; he was well versed in the history of all times and all people. Even about medical and herbal remedies, and chemistry, he knew more than one could expect from a scholar who doesn't make these sciences a special subject of his study.

Many visitors who came to see Euler went away "with a mixture of astonishment and admiration. They could not understand how a man who during half a century seemed to have occupied himself solely with discoveries in the natural sciences and mathematics could retain so many facts that to him were useless and foreign to the subject of his researches."

4.3. Craftsmanship. Euler's writings have the marks of a superb expositor. He always strove for utmost clarity and simplicity, and he often revisited earlier work when he felt they were lacking in these qualities. Characteristically, he will proceed from very simple examples to ever more complicated ones before eventually revealing the underlying theory in its full splendor. Yet, in his quest for discovery, he could be fearless, even reckless, but owing to his secure instinct, he rarely went astray when his argumentation became hasty. He had an eye for what is essential and unifying. In mechanics, Gleb Konstantinovich Mikhailov [20, p. 67] writes, "Euler possessed a rare gift of systematizing and generalizing scientific ideas, which allowed him to present large parts of mechanics in a relatively definitive form." Euler was open and receptive to new ideas. In the words of André Weil [30, pp. 132–133],

> ... what at first is striking about Euler is his extraordinary quickness in catching hold of any suggestion, wherever it came from.... There is not one of these suggestions which in Euler's hands has not become the point of departure of an impressive series of researches.... Another thing, not less striking, is that Euler never abandons a research topic, once it has excited his curiosity; on the contrary, he returns to it, relentlessly, in order to deepen and broaden it on each revisit. Even if all problems related to such a topic seem to be resolved, he never ceases until the end of his life to find proofs that are "more natural," "simpler," "more direct."

4.4. Epilogue. In closing, let me cite the text (translated from German)—concise but to the point—that Otto Spiess had inscribed on a memorial plaque attached near the house in Riehen in which Euler grew up:

LEONHARD EULER
1707–1783
Mathematician, physicist, engineer,
astronomer and philosopher, spent his
youth in Riehen. He was a great scholar
and a kind man.

5. Further Reading. For readers interested in more details, we recommend the authoritative scientific (yet formula-free!) biography by Fellmann [10], the essays in the recent book by Henry [16], and several accounts on Euler and parts of his work that have recently appeared: Bogolyubov, Mikhailov, and Yushkevich [3], Bradley, D'Antonio, and Sandifer [5], Dunham [6], [7], Nahin [21], Sandifer [22], [23], and Varadarajan [29].

The web site of the U.S. Euler Archive,

http://www.math.dartmouth.edu/∼euler,

also provides detailed information about Euler's complete works, arranged by their E-numbers.

Sources and Acknowledgments. The sources for the videos posted here, with permission, are as follows. Video `buckle.avi`: Professor Wolfgang Ehlers, Institute of Applied Mechanics (CE), University of Stuttgart, Germany. Video `eulerflow.avi`: 2-*dimensional compressible inviscid flow about a circular cylinder—a computer simulation by Nicola Botta*, ©1993 Eidgenössische Technische Hochschule Zürich. Video `zahn.avi`: Professor Bert Jüttler, Institute of Applied Geometry, Johannes Kepler Universität, Linz, Austria. Video `eulerdisk.avi`: produced at the author's request by Olaf A. Schulte, Multimedia Services, ETH Zürich, Zürich, Switzerland, ©2007 Walter Gautschi.

The author is grateful to a number of colleagues for having read preliminary versions of this article and for providing useful suggestions or technical help. In particular, he is indebted to R. Askey for suggesting the inclusion of material on partitions, to F. Cerulus for reviewing and commenting on my coverage of mechanics, and to E.A. Fellmann for historical guidance and continuous encouragement. He also wishes to acknowledge Walter Gander for reference [19], H. Hunziker for reference [26], Robert Schaback for reference [18], and Rolf Jeltsch for pointing the author to the `eulerflow.avi` video. He is thankful to Pedro Gonnet for scanning many title pages from Euler's *Opera omnia*.

REFERENCES

[1] G. Assayag, H.-G. Feichtinger, and J. F. Rodrigues, eds., *Mathematics and Music: A Diderot Mathematical Forum*, Springer, Berlin, 2002.

[2] P. Bailhache, *Deux mathématiciens musiciens: Euler et d'Alembert*, Physis Riv. Internaz. Storia Sci. (N.S.), 32 (1995), pp. 1–35.

[3] N. N. Bogolyubov, G. K. Mikhailov, and A. P. Yushkevich, eds., *Euler and Modern Science*, MAA Spectrum, Mathematical Association of America, Washington, D.C., 2007; translated from the Russian by Robert Burns.

[4] N. Botta, *Numerical Investigations of Two-Dimensional Euler Flows: Cylinder at Transonic Speed*, Ph.D. dissertation, Swiss Federal Institute of Technology, Zürich, 1995.

[5] R. E. Bradley, L. A. D'Antonio, and C. E. Sandifer, eds., *Euler at* 300: *An Appreciation*, MAA Spectrum, Mathematical Association of America, Washington, D.C., 2007.

[6] W. Dunham, *Euler: The Master of Us All*, Dolciani Math. Exp. 22, Mathematical Association of America, Washington, D.C., 1999.

[7] W. Dunham, ed., *The Genius of Euler: Reflections on His Life and Work*, MAA Spectrum, Mathematical Association of America, Washington, D.C., 2007.

[8] L. Euler, *Lettres à une princesse d'Allemagne sur divers sujets de physique et de philosophie*, S. D. Chatterji, ed., Presses Polytechniques et Universitaires Romandes, Lausanne, 2003.

[9] E. A. Fellmann, *Leonhard Euler—Ein Essay über Leben und Werk*, in Leonhard Euler 1707–1783: Beiträge zu Leben und Werk, Gedenkband des Kantons Basel-Stadt, Birkhäuser, Basel, 1983, pp. 13–98.

[10] E. A. Fellmann, *Leonhard Euler*, Rowohlt, Reinbek bei Hamburg, 1995 (out of print). English translation by Erika and Walter Gautschi, Birkhäuser, Basel, 2007; Japanese translation by Springer, Tokyo, 2002.

[11] C. Francese and D. Richeson, *The flaw in Euler's proof of his polyhedral formula*, Amer. Math. Monthly, 114 (2007), pp. 286–296.

[12] W. Gautschi, *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., to appear.

[13] W. Gautschi, *Leonhard Eulers Umgang mit langsam konvergenten Reihen*, Elem. Math., 62 (2007), pp. 174–183.

[14] B. Haible and T. Papanikolaou, *Fast multiprecision evaluation of series of rational numbers*, in Algorithmic number theory (Portland, OR, 1998), Lecture Notes in Comput. Sci. 1423, Springer, Berlin, 1998, pp. 338–350.

[15] J. Havil, *Gamma. Exploring Euler's Constant*, Princeton University Press, Princeton, NJ, 2003.

[16] Ph. Henry, *Leonhard Euler: Incomparable géomètre*, Editions Médecine et Hygiène, Chêne-Bourg, 2007

[17] E. Koelink and W. Van Assche, *Leonhard Euler and a q-analogue of the logarithm*, Proc. Amer. Math. Soc., to appear.

[18] S. Krämer, *Die Eulersche Konstante γ und verwandte Zahlen: Eine mathematische und historische Betrachtung*, Diplomarbeit Universität Göttingen, Göttingen, 2005.

[19] C. Le Saux, R. I. Leine, and C. Glocker, *Dynamics of a rolling disk in the presence of dry friction*, J. Nonlinear Sci., 15 (2005), pp. 27–61.

[20] G. K. Mikhailov, *Euler und die Entwicklung der Mechanik*, in Ceremony and Scientific Conference on the Occasion of the 200th Anniversary of the Death of Leonhard Euler (Berlin, 1983), Abh. Akad. Wiss. DDR, Abt. Math. Naturwiss. Tech. 85-1, Akademie-Verlag, Berlin, 1985, pp. 64–82.

[21] P. J. Nahin, *Dr. Euler's Fabulous Formula. Cures Many Mathematical Ills*, Princeton University Press, Princeton, NJ, 2006.

[22] C. E. Sandifer, *The Early Mathematics of Leonhard Euler*, MAA Spectrum, Mathematical Association of America, Washington, D.C., 2007.

[23] C. E. Sandifer, *How Euler Did It*, MAA Spectrum, Mathematical Association of America, Washington, D.C., 2007.

[24] W. Scharlau, *Eulers Beiträge zur partitio numerorum und zur Theorie der erzeugenden Funktionen*, in Leonhard Euler 1707–1783: Beiträge zu Leben und Werk, Gedenkband des Kantons Basel-Stadt, Birkhäuser, Basel, 1983, pp. 135–149.

[25] L. Schläfli, *Theorie der vielfachen Kontinuität*, Zürcher & Furrer, Zürich, 1850–1852; published posthumously in 1901. Also in *Gesammelte Mathematische Abhandlungen*, Bd. 1, Birkhäuser, Basel, 1950, pp. 167–387.

[26] J. H. Silverman, *A Friendly Introduction to Number Theory*, 3rd ed., Prentice–Hall, Upper Saddle River, NJ, 2006.

[27] C. S. Smith, *Leonhard Euler's* Tentamen novae theoriae musicae: *A translation and commentary*, Ph.D. thesis, Indiana University, 1960. Accessible through UMI Dissertation Services, Ann Arbor, MI.

[28] G. K. Srinivasan, *The gamma function: An eclectic tour*, Amer. Math. Monthly, 114 (2007), pp. 297–315.

[29] V. S. Varadarajan, *Euler Through Time: A New Look at Old Themes*, American Mathematical Society, Providence, RI, 2006.

[30] A. Weil, *L'œuvre arithmétique d'Euler*, in Leonhard Euler 1707–1783: Beiträge zu Leben und Werk, Gedenkband des Kantons Basel-Stadt, Birkhäuser, Basel, 1983, pp. 111–133.

29.11. [189] (with C. Giordano) "Luigi Gatteschi's work on asymptotics of special functions and their zeros"

---

[189] (with C. Giordano) "Luigi Gatteschi's work on asymptotics of special functions and their zeros," in *A collection of essays in memory of Luigi Gatteschi* (G. Allasia, C. Brezinski, and M. Redivo-Zaglia, eds.), *Numer. Algorithms* **49**, 11–31 (2008).

Numer Algor (2008) 49:11–31
DOI 10.1007/s11075-008-9208-5

ORIGINAL PAPER

Luigi Gatteschi's work on asymptotics of special functions and their zeros

Walter Gautschi · Carla Giordano

Received: 9 April 2008 / Accepted: 10 April 2008 /
Published online: 20 May 2008

Abstract A good portion of Gatteschi's research publications—about 65%—is devoted to asymptotics of special functions and their zeros. Most prominently among the special functions studied figure classical orthogonal polynomials, notably Jacobi polynomials and their special cases, Laguerre polynomials, and Hermite polynomials by implication. Other important classes of special functions dealt with are Bessel functions of the first and second kind, Airy functions, and confluent hypergeometric functions, both in Tricomi's and Whittaker's form. This work is reviewed here, and organized along methodological lines.

Keywords Luigi Gatteschi's work · Asymptotics · Special functions · Zeros

Mathematics Subject Classifications (2000) 26C10 · 30C15 · 33C10 · 33C15 · 33C45 · 41A60

1 Introduction

In asymptotics there are two kinds of theories: a qualitative theory, and a quantitative theory. They differ in the way the error of an asymptotic approximation is characterized. In the former, the error is estimated by an order-of-magnitude term $O(\omega(x))$, i.e., by a statement that there exists a positive,

W. Gautschi ( )
Department of Computer Sciences, Purdue University, West Lafayette, IN 47907-2066, USA
e-mail: wxg@cs.purdue.edu

C. Giordano
Dipartimento di Matematica, Università di Torino, Turin, 10123, Italy
e-mail: carla.giordano@unito.it

unspecified constant C such that the error is bounded in absolute value by $C\omega(x)$ as the variable (or parameter) x is in the neighborhood of a limit value x_0. Here, ω is a known, computable, positive function of x, for example a reciprocal power of x if $x_0 = +\infty$. A quantitative theory, in contrast, provides a numerical upper bound for the constant C, or better still, concrete numerical lower and upper bounds for the error, $\omega_-(x)$ and $\omega_+(x)$, along with a precise description of the domain of validity (in x). The approximation, in effect, then takes on the form of a two-sided inequality. Much of the older, classical theory of asymptotics is of a qualitative nature, while modern exigencies of computing require a quantitative theory. In the realm of special functions and their zeros, Luigi Gatteschi is without doubt one of the major exponents of, and contributor to, the quantitative theory of asymptotics. His results are not only of a concrete numerical nature, but often attain a degree of sharpness rarely found elsewhere in the literature.

In the following we briefly summarize Gatteschi's relevant work as it pertains to orthogonal polynomials, Bessel and Airy functions, and confluent hypergeometric functions. We arrange the presentation according to the type of methods used, and in each case proceed in more or less chronological order. Even though we can give only a quick and superficial account of finished results, it must be emphasized that, underneath it all, there is a great deal of hard analysis, imaginatively and skillfully executed.

2 The early influence of Szegö, Van der Corput, and Tricomi

Among important individuals who had an influence in shaping Luigi's formation as a research mathematician, one must mention Giovanni Sansone, who guided Luigi's first research efforts, Gabor Szegö and Johannes Van der Corput, with whom Luigi interacted during a visit in 1951 to Stanford University, and above all, from the start of Luigi's career at the University of Turin, Francesco Tricomi, who became his mentor.

2.1 A general method of Tricomi

Already in the very first papers of Luigi, dealing with zeros of Legendre and ultraspherical polynomials of large degrees, and high-order zeros of Bessel functions, an important ingredient is a method of Tricomi for deriving the asymptotics of zeros of functions from the asymptotics of the functions themselves (see [57], or [59, p. 151]). While Tricomi formulated his method in qualitative terms, Luigi in the special cases studied supplies concrete error bounds by tracing and estimating remainder terms in all Taylor expansions employed.

2.1.1 Zeros of ultrashperical polynomials

In the case of Legendre and ultraspherical polynomials, the results obtained in [10–12] are somewhat preliminary inasmuch as they cover only limited ranges

of zeros. This deficiency is overcome later in [33], though at the expense of sharpness, where Tricomi's method is again applied to ultraspherical polynomials $P_n^{(\lambda)} = P_n^{(\lambda-1/2,\lambda-1/2)}$, $0 < \lambda < 1$. For the rth zero $\theta_{n,r}^{(\lambda)}$ of $P_n^{(\lambda)}(\cos\theta)$ it is found that for each $r = 1, 2, \ldots, \lfloor n/2 \rfloor$ (which, by symmetry, is all we need),

$$\theta_{n,r}^{(\lambda)} = \vartheta_{n,r}^{(\lambda)} + \frac{\lambda(1-\lambda)}{2(n+\lambda)(n+\lambda+1)} \cot\vartheta_{n,r}^{(\lambda)} + \rho, \tag{1}$$

where $\vartheta_{n,r}^{(\lambda)} = (r - (1-\lambda)/2)\pi/(n+\lambda)$, and[1]

$$|\rho| < \frac{\lambda(1-\lambda)}{(n+\lambda+1)(2r+\lambda-1)^2}, \quad 0 < \lambda < 1. \tag{2}$$

It can be seen that when r is fixed and $n \to \infty$, the first two terms on the right of (1), as well as the bound in (2), are all $\sim cn^{-1}$, with the respective constants c decreasing [actually, when $r = 1$, and in part also when $r = 2$, the constant c for the bound in (2) is a bit larger than the one for the second term]. On the other hand, when $r = \lfloor \delta n/2 \rfloor$, with $0 < \delta \le 1$ fixed, the two terms and bound are respectively $O(1)$, $O(n^{-2})$, and $O(n^{-3})$. For the zeros $x_{n,r}^{(\lambda)} = \cos\theta_{n,r}^{(\lambda)}$ themselves, one finds

$$x_{n,r}^{(\lambda)} = \xi_{n,r}^{(\lambda)} \left[1 - \frac{\lambda(1-\lambda)}{2(n+\lambda)(n+\lambda+1)}\right] + \varepsilon, \tag{3}$$

where $\xi_{n,r}^{(\lambda)} = \cos((r - (1-\lambda)/2)\pi/(n+\lambda))$, and

$$|\varepsilon| < \frac{1.55\,\lambda(1-\lambda)}{(n+\lambda)(n+\lambda+1)(2r+\lambda-1)}, \quad 0 < \lambda < 1. \tag{4}$$

2.1.2 Zeros of Bessel functions

Similarly complete are the results in [13] for the Bessel function J_ν, $0 \le \nu \le 1$. Thus, for the rth positive zero $j_{\nu,r}$ of J_ν, Luigi shows that

$$j_{\nu,r} = x_r - \frac{4\nu^2 - 1}{8x_r} + \varepsilon(\nu, r), \quad r = 1, 2, 3, \ldots, \tag{5}$$

where $x_r = (r + \nu/2 - 1/4)\pi$, and

$$|\varepsilon(\nu, r)| < \frac{(7.4A^2 + 1.1A)r}{64(6r-5)} (2r + \nu - 1)^{-3}, \quad A = |4\nu^2 - 1|, \tag{6}$$

valid for each $r = 1, 2, 3, \ldots$. The formula (6), in fact, quantifies the $O(r^{-3})$ term in a classical asymptotic formula of McMahon [54]. In another formula of McMahon for the rth zero of $J_0(kx)Y_0(x) - J_0(x)Y_0(kx)$, where J_0 and Y_0 are the zeroth-order Bessel functions of first and second kind, an $O(r^{-7})$ term is similarly quantified in [14] for $r \ge 2$ and values of the parameter k satisfying $1 < k < 3 + 2\sqrt{2}$. The calculations, however, are rather more formidable in this case. For the rth positive zero $j'_{\nu,r}$, $0 \le \nu \le 1$, of the derivative J'_ν of the

[1] The square in the second factor of the denominator is missing in Eq. (2.13_1) of [33].

Bessel function, a formula analogous to (5), (6), also due (without error bound) to McMahon, is derived in [20].

2.1.3 Zeros of Jacobi polynomials

The application of Tricomi's method to more general Jacobi polynomials $P_n^{(\alpha,\beta)}$ had to wait until 1980, when a suitable asymptotic expansion for $P_n^{(\alpha,\beta)}$ became available through the work of Hahn [53]. Using the first three terms of this expansion in conjunction with Tricomi's method (in fact, a slight extension thereof), and assuming $|\alpha| \le 1/2$, $|\beta| \le 1/2$, Luigi jointly with Pittaluga [50] proves that for the zeros $\theta_{n,r}^{(\alpha,\beta)}$ of $P_n^{(\alpha,\beta)}(\cos\theta)$ contained in any compact subinterval of $(-1, 1)$, there holds

$$\theta_{n,r}^{(\alpha,\beta)} = \vartheta_{n,r}^{(\alpha,\beta)} + \frac{1}{(2n+\alpha+\beta+1)^2}\left[\left(\frac{1}{4}-\alpha^2\right)\cot\left(\frac{1}{2}\vartheta_{n,r}^{(\alpha,\beta)}\right) - \left(\frac{1}{4}-\beta^2\right)\tan\left(\frac{1}{2}\vartheta_{n,r}^{(\alpha,\beta)}\right)\right] + O(n^{-4}), \tag{7}$$

where $\vartheta_{n,r}^{(\alpha,\beta)} = (2r+\alpha-1/2)\pi/(2n+\alpha+\beta+1)$. If $\alpha^2 = \beta^2 = 1/4$, not only the expression in brackets, but also the error term in (7) vanish. In the ultraspherical case $\alpha = \beta$, the result (7) is asymptotically in agreement with earlier ones in [11]. There is of course a result analogous to (7) for the zeros $x_{n,r}^{(\alpha,\beta)}$ of $P_n^{(\alpha,\beta)}(x)$ themselves. Numerical tests revealed that already for $n = 16$, these asymptotic approximations (with the error term removed) typically yield $4\frac{1}{2} - 6$ correct significant digits for *all* zeros $x_{n,r}^{(\alpha,\beta)}$. Interestingly, if one of the parameters α, β has the value $\pm 1/2$, the accuracy is several orders higher near the appropriate boundary of $[-1, 1]$, a phenomenon duly explained by Luigi.

2.2 A general method of Gatteschi and Van der Corput's theory of enveloping series

2.2.1 Zeros of Bessel functions by Gatteschi's method

In [15], with the assistance of Van der Corput, Luigi develops a general procedure of his own for generating inequalities for the zeros of a function

$$f(x) = (1+\delta)\sin x + \varepsilon\cos x - \rho, \quad \delta > -1, \tag{8}$$

where δ, ε, and ρ may depend on x but are small in magnitude. This kind of functions is often encountered in asymptotic expansions (for large x) of certain Bessel-type functions. Luigi in [15] applies his new procedure to Bessel functions $J_\nu(x)$, where ν can now be arbitrary nonnegative, and supplements the results in [13] by estimating the zeros $j_{\nu,r}$ that are larger than $(2\nu+1)(2\nu+3)/\pi$. The same procedure is applied in [19] to Airy functions $\mathrm{Ai}(-x)$, $\mathrm{Bi}(-x)$ and their positive zeros.

In two of his late papers, [47] and [48], Luigi, jointly with Giordano, returns to his procedure and makes further applications to Bessel functions. In [47], McMahon's formula for $j_{\nu,r}$ is taken up again and in the case $|\nu| \leq 1/2$ supplied with lower and upper bounds for the $O(r^{-5})$ term, and in the case $\nu > 1/2$ with similar bounds for the $O(r^{-3})$ and $O(r^{-5})$ terms. A two-term asymptotic approximation with explicit error bounds is obtained in [48] for the positive zeros $i_{\nu,r} > (r + \nu/2 - 3/4)\pi$, $r \geq 10$, of $(d/dx)[\sqrt{x}J_\nu(x)]$ in the case that $|\nu| \leq 1/2$.

2.2.2 Bessel functions at and near the transition point

A new methodological element—Van der Corput's theory of "enveloping series"—appears in [21]. Given a series $\sum_{n=0}^{\infty} a_n$ (not necessarily convergent) and a majorizing series $\sum_{n=0}^{\infty} A_n$ thereof, i.e., $|a_n| \leq A_n$ for all n, the series $\sum_{n=0}^{\infty} a_n$ is said to envelope a number (or function) s relative to the majorant $\sum_{n=0}^{\infty} A_n$, if for each $n = 0, 1, 2, \ldots$

$$s = \sum_{k=0}^{n-1} a_k + \vartheta_n A_n, \quad |\vartheta_n| \leq 1.$$

Using two key theorems in Van der Corput's theory of enveloping series, one relating to the formal substitution of a series into another series, and another relating to integration of (functional) enveloping series, both applied to contour integral representations of Hankel functions, Luigi in [21] derives very impressive asymptotic expansions for $J_\nu(\nu)$ and $Y_\nu(\nu)$ as $\nu \to \infty$, both supplied with error estimates. They are not simple, involving as they do incomplete gamma functions and coefficients $A_k^{(m)}$ in the Taylor expansion of $\left(\frac{1}{5!} + \frac{1}{7!}z + \frac{1}{9!}z^2 + \cdots\right)^m$, $m = 2, 3, \ldots$ (which today, however, are easily obtainable by symbolic computation systems such as Maple). As an application, Luigi takes the first two terms of his expansion for $J_\nu(\nu)$ (the second term happening to be zero) and obtains

$$J_\nu(\nu) = \frac{\Gamma(1/3)}{2^{2/3}3^{1/6}\pi}\,\nu^{-1/3} - \theta\eta, \quad 0 < \theta \leq 1,\ \nu \geq 6, \tag{9}$$

where[2]

$$\eta = \frac{1}{\pi\nu}\left(e^{-\nu\pi/\sqrt{3}} + .521e^{-(2\pi/\sqrt{3})^3\nu/6}\right) + \frac{1.4}{\pi}\left(\frac{6}{\nu}\right)^{5/3}.$$

This recovers an asymptotic formula of Cauchy, but endows it with an explicit error bound. The simpler bound $\eta < \nu^{-5/3}$ is given in the lecture [25].

As observed in [22], there is a slight inaccuracy (on p. 275) in [21], but the results obtained there are shown to continue to hold. Also, from the first term

[2]The first term in parentheses is misprinted in [21, Eq. (20′)] as $e^{-2\pi/\sqrt{3}}$.

of the asymptotic expansion for $Y_\nu(\nu)$ in [21], the following companion result to (9) is obtained,

$$Y_\nu(\nu) = -\frac{\Gamma(1/3)}{(4/3)^{1/3}\pi}\,\nu^{-1/3} + \rho, \quad |\rho| \le \frac{.252}{\pi\nu}, \ \nu \ge 1. \tag{10}$$

An interesting consequence of this is

$$\left\{\begin{matrix} |J_\nu(\nu x)| \\ |Y_\nu(\nu x)| \end{matrix}\right\} < \frac{1}{\sqrt{x}}\left[\frac{3.841}{\pi\nu^{1/3}} + \frac{.252}{\pi\nu}\right], \tag{11}$$

valid for $x > 1$, $\nu \ge 1$.

An asymptotic estimate of $J_\nu(x)$ around the transition point $x = \nu$ is developed in [26], by using a Liouville–Steklov-type approach (cf. Section 3.1). It is shown that

$$J_\nu\big(\nu\exp(6^{-1/3}\nu^{-2/3}t)\big) = 3^{2/3}\Gamma(2/3)J_\nu(\nu)\mathrm{Ai}\big(-3^{-1/3}t\big) + \rho, \tag{12}$$

where for $\nu \ge 6$

$$|\rho| < \begin{cases} \dfrac{t^4 + 5.6\,t}{\pi\nu} & \text{if } 0 < t < 6^{1/3}\nu^{2/3}, \\ \dfrac{1}{\nu}\left[.005\,t^4\exp(4(|t|/3)^{3/2}) + 1.77\,|t|\exp(2(|t|/3)^{3/2})\right] & \text{if } t < 0. \end{cases} \tag{13}$$

Sharper estimates are obtained by a reapplication of the Liouville–Steklov method.

Luigi also gives an asymptotic estimate of the derivative $J'_\nu(x)$ at $x = \nu$,

$$J'_\nu(\nu) = \frac{1}{2\sqrt{3}\pi}\left[\Gamma(2/3)\left(\frac{6}{\nu}\right)^{2/3} - \frac{\Gamma(1/3)}{30}\left(\frac{6}{\nu}\right)^{4/3}\right] + \vartheta\,\frac{2}{\nu^2}, \quad |\vartheta| < 1, \ \nu \ge 6, \tag{14}$$

an interesting subsidiary result.

3 Methods based on differential equations

Linear second-order differential equations, which are at the heart of much of special function theory, can be used in many ways to obtain asymptotic approximations and inequalities. There are two techniques, in particular, that Luigi frequently, and early on, availed himself of: One is the method of Liouville–Steklov (sometimes also attributed to Fubini), which is based on transforming the differential equation into a Volterra integral equation; the other is the use of Sturm-type comparison theorems.

3.1 The method of Liouville–Steklov

3.1.1 Hilb's formula and zeros of Legendre polynomials

Already in one of his early papers, [16], Luigi applies the method of Liouville–Steklov, following Szegö's treatment in [55, Section 8.62], to the differential

equation satisfied by $(\sin\theta)^{1/2} P_n(\cos\theta)$. (By symmetry, it suffices to consider the interval $0 \le \theta \le \pi/2$.) This yields immediately Hilb's formula,

$$P_n(\cos\theta) = \left(\frac{\theta}{\sin\theta}\right)^{1/2} J_0((n+1/2)\theta) + \sigma, \tag{15}$$

where for large n, when θ is away from the origin (i.e., $\theta \ge cn^{-1}$ for some positive constant c), the error is $\sigma = \theta^{1/2} O(n^{-3/2})$, otherwise $\sigma = O(n^{-2})$. In his quest for quantification, Luigi derives explicit inequalities for the error σ: In the first case,

$$|\sigma| < .358\,\theta^{-1/2} n^{-5/2} + .394\,\theta^{1/2} n^{-3/2} \quad \text{if } \pi/2n < \theta \le \pi/2 \tag{16}$$

(which may also be written as $|\sigma| < .622\,\theta^{1/2} n^{-3/2}$; cf. [23, Eq. (2)]), and in the second case,

$$|\sigma| < .09\,\theta^2 \quad \text{if } 0 < \theta \le \pi/2n. \tag{17}$$

This is then applied to obtain two-sided inequalities for the zeros $\theta_{n,r}$ (in ascending order) of $P_n(\cos\theta)$, namely[3]

$$0 < \frac{j_{0,r}}{n+1/2} - \theta_{n,r} < (1.6 + 3.7r)n^{-4}, \quad n = 1, 2, \ldots, \lfloor n/2 \rfloor, \tag{18}$$

where $j_{0,r}$ is the rth positive zero of the Bessel function J_0.

A reapplication of the Liouville–Steklov method to the same differential equation, but now with (15) inserted in the integral of the Volterra integral equation, in [23] yields an improved two-term asymptotic approximation for $P_n(\cos\theta)$, and in consequence also two-term approximations for the zeros $\theta_{n,r}$ of $P_n(\cos\theta)$, and likewise for the zeros $x_{n,r}$ of $P_n(x)$. Thus, for example,

$$x_{n,r} = 1 - \frac{j_{0,r}^2}{2(n+1/2)^2} + \frac{j_{0,r}^2 + j_{0,r}^4}{24(n+1/2)^4} + O(n^{-6}), \tag{19}$$

which for $n = 16$, $r = 1$ and $r = 2$ (neglecting the error term), yields approximations for the respective zeros having errors 2.28×10^{-8} resp. 2.15×10^{-6}.

3.1.2 Hilb's formula for ultraspherical polynomials

Hilb's formula for ultraspherical polynomials is supplied with error bounds in [18] and applied to the zeros of $P_n^{(\lambda)}$. A slightly different application of the method of Liouville–Steklov, especially if applied successively as suggested by Szegö [55, Section 8.61(2)] in the case of Legendre polynomials, yields more accurate approximations of ultraspherical polynomials $P_n^{(\lambda)}$, valid in any compact subinterval of $(-1, 1)$, and of their zeros contained therein [34].

[3] There is a misprint in Eq. (17) of [16], where the number 16 in the denominator should be 10. The upper bound given there (and in our Eq. (18)) has been checked by us on the computer and was found to be too small, at least for larger values of n. The reason for this may be inaccuracies in the numerical constants supplied.

3.1.3 Hilb's formula for Jacobi polynomials

There is a Hilb's formula also for Jacobi polynomials $P_n^{(\alpha,\beta)}$, $\alpha > -1$ and β arbitrary real [55, Section 8.63], which in the classical form reads as follows,

$$\theta^{-1/2}\left(\sin\frac{1}{2}\theta\right)^{\alpha+1/2}\left(\cos\frac{1}{2}\theta\right)^{\beta+1/2}P_n^{(\alpha,\beta)}(\cos\theta)$$
$$= 2^{-1/2}N^{-\alpha}\,\frac{\Gamma(n+\alpha+1)}{n!}\,J_\alpha(N\theta)+\sigma_\alpha(n,\theta), \tag{20}$$

where $N = n + (\alpha+\beta+1)/2$ and $\sigma_\alpha = \theta^{1/2}O(n^{-3/2})$ away from the origin, and $\sigma_\alpha = \theta^{\alpha+2}O(n^\alpha)$ otherwise. In [31], this is improved in two ways: First, the method of Liouville–Steklov is refined, with the result that in (20) the number N can be replaced by

$$\nu = \left[\left(n+\frac{\alpha+\beta+1}{2}\right)^2+\frac{1-\alpha^2-3\beta^2}{12}\right]^{1/2} \tag{21}$$

and the error term improved to $\sigma_\alpha = \theta^{5/2}O(n^{-3/2})$ and $\sigma_\alpha = \theta^{\alpha+4}O(n^\alpha)$ away from, and near the origin, respectively.[4] Secondly, the method of Liouville–Steklov is iterated once more, similarly as in [23], producing a two-term approximation,

$$\theta^{-1/2}\left(\sin\frac{1}{2}\theta\right)^{\alpha+1/2}\left(\cos\frac{1}{2}\theta\right)^{\beta+1/2}P_n^{(\alpha,\beta)}(\cos\theta)$$
$$= 2^{-1/2}\nu^{-\alpha}\frac{\Gamma(n+\alpha+1)}{n!}\left[\left(1-\frac{4-\alpha^2-15\beta^2}{1440\,\nu^2}\,\theta^2\right)J_\alpha(\nu\theta)\right.$$
$$\left.+\frac{4-\alpha^2-15\beta^2}{720\,\nu^3}\,\theta\left(\frac{1}{2}\nu^2\theta^2+\alpha^2-1\right)J'_\alpha(\nu\theta)\right]+\rho_\alpha(n,\theta), \tag{22}$$

with the remainder term further improved to respectively $\rho_\alpha = \theta^{9/2}O(n^{-3/2})$ and $\rho_\alpha = \theta^{\alpha+6}O(n^\alpha)$. The result (22) can easily be specialized to ultraspherical polynomials (i.e., to $\alpha = \beta = \lambda - 1/2$) and to Legendre polynomials ($\lambda = 1/2$). In the latter case, by expressing J'_0 in terms of J_0 and J_2, one obtains the rather simple formula

$$\left(\frac{\sin\theta}{\theta}\right)^{1/2}P_n(\cos\theta) = J_0(\nu\theta)-\frac{\theta^3}{360\,\nu}+\frac{\theta^2}{360\,\nu^2}\,J_2(\nu\theta)+\rho(n,\theta), \tag{23}$$

with $\nu = [(n+1/2)^2+1/12]^{1/2}$ and $\rho = \theta^{9/2}O(n^{-3/2})$ resp. $\rho = \theta^6\,O(1)$.

[4] In the second of these formulae, the factor $\theta^{\alpha+4}$ is misprinted as $\theta^{\alpha+1}$ in the original Eq. (19) of [31].

Luigi now once again applies Tricomi's theorem (cf. Section 2.1) to derive from the asymptotic approximations in [31] asymptotic results for zeros of Jacobi polynomials in terms of zeros of Bessel functions, and vice versa. In the ultraspherical case, for example, he finds for the rth positive zero $j_{s,r}$ of J_s, $-1/2 < s < 1/2$, when r is fixed, the following asymptotic approximation,

$$j_{s,r} = \nu\theta_{n,r} + \frac{1-4s^2}{360\,\nu}\,\theta_{n,r}^3 - (1-s^2)\,\frac{1-4s^2}{180\,\nu^3}\,\theta_{n,r} + O(n^{-6}), \tag{24}$$

where $\nu = \left[(n+s+1/2)^2 + (1-4s^2)/12\right]^{1/2}$ and $\theta_{n,r} = \theta_{n,r}^{(s+1/2)}$ is the rth zero of $P_n^{(s+1/2)}(\cos\theta)$.

In [2], the method of Liouville-Steklov is used to derive a new asymptotic approximation of Hilb's type for Jacobi polynomials $P_n^{(\alpha,\beta)}$, $|\alpha| \le 1/2$, $|\beta| \le 1/2$, with realistic and explicit error bounds, and from it an asymptotic estimate of the zeros $\theta_{n,r}^{(\alpha,\beta)}$ of $P_n^{(\alpha,\beta)}(\cos\theta)$ obtained previously in a different manner by Frenzen and Wong [7]. Continuation of this work in [41, 42] led to a number of significant improvements.

The classical Hilb's formula (20) for Jacobi polynomials is applied in [17] to study the relative extrema of $P_n^{(\alpha,\beta)}$. If $y_{n,r}$ are their abscissae, and $y_{n,r} = \cos\varphi_{n,r}$, a short and elegant proof is given of the limit relation

$$\lim_{n\to\infty}\left(\sin\frac{1}{2}\,\varphi_{n,r}\right)^{\alpha}\left(\cos\frac{1}{2}\,\varphi_{n,r}\right)^{\beta} P_n^{(\alpha,\beta)}(y_{n,r}) = J_\alpha(j_{\alpha+1,r}). \tag{25}$$

3.2 Methods based on Sturm comparison theorems

Sturm-type comparison theorems, for example in the form stated by Szegö in [55, Section 1.82], are a natural tool for comparing zeros of one type of special functions with zeros of another type, the types of special functions depending on the choice of differential equations that are being compared. This is a recurring theme in Luigi's work and gives rise to many interesting inequalities.

3.2.1 Zeros of Jacobi polynomials and Bessel functions

In [32], the comparison is between zeros $\theta_{n,r}^{(\alpha,\beta)}$ of Jacobi polynomials $P_n^{(\alpha,\beta)}(\cos\theta)$ and zeros $j_{\alpha,r}$ of Bessel functions J_α, which, under the assumption $|\alpha| \le 1/2$, $\beta \le 1/2$, finds expression in the inequalities

$$j_{\alpha,r}\left[N^2 + \frac{1}{4} - \frac{\alpha^2+\beta^2}{2} - \frac{1-4\alpha^2}{\pi^2}\right]^{-1/2}$$

$$< \theta_{n,r}^{(\alpha,\beta)} < j_{\alpha,r}\left[N^2 + \frac{1-\alpha^2-3\beta^2}{12}\right]^{-1/2}, \tag{26}$$

valid for $r = 1, 2, \ldots, \lfloor n/2\rfloor$, where $N = n + (\alpha+\beta+1)/2$.

The first zero, $j_\nu = j_{\nu,1}$ of the Bessel function J_ν, $\nu > 0$, and also the abscissa j'_ν of its first maximum, are studied in [49], where Sturm's theorem is used in

a form given by Watson in [61, Section 15.83] and is slightly extended and combined, in part, with Tricomi's theorem (cf. Section 2.1). The result can be written in the form

$$\begin{aligned} j_\nu &= \nu \exp\left(2^{-1/3}\nu^{-2/3}a_1 - 1.623\,\vartheta\,\nu^{-4/3}\right), \\ j'_\nu &= \nu \exp\left(2^{-1/3}\nu^{-2/3}a'_1 - 1.06\,\vartheta'\,\nu^{-4/3}\right), \end{aligned} \tag{27}$$

where $0 < \vartheta, \vartheta' < 1$, and $a_1 = 2.33810741$, $a'_1 = 1.01879297$ are the first zero, resp. maximum, of the Airy function $\mathrm{Ai}(-x)$. The bounds implied by (27) compare favorably with earlier estimates by Schafheitlin and Tricomi.

Restricting ν to the "principal" interval $|\nu| < 1/2$, Luigi, together with Giordano, in [46] obtains a very sharp upper bound for j_ν, namely

$$j_\nu < \Theta(\nu)K(\nu), \quad -1/2 < \nu < 1/2, \tag{28}$$

where

$$\Theta(\nu) = \arccos\sqrt{\frac{10\nu + 35 + 2\sqrt{10\nu^2 + 55\nu + 70}}{4\nu^2 + 32\nu + 63}}$$

is the first zero $\theta_{5,1}^{(\nu)}$ of $P_5^{(\nu,\nu)}(\cos\theta)$,

$$K(\nu) = \left[(\nu + 11/2)^2 + \left(\frac{1}{4} - \nu^2\right)\left(\frac{1}{\sin^2\phi(\nu)} - \frac{1}{\phi^2(\nu)}\right)\right]^{1/2},$$

and

$$\phi(\nu) = \frac{\sqrt{\nu+1}\left(\sqrt{\nu+2}+1\right)}{\nu + 11/2}.$$

Outside the principal interval, there holds

$$j_\nu < \frac{1}{3}\,\Theta(\nu)\sqrt{6\nu^2 + 99\nu + 273}, \quad \nu \notin (-1/2, 1/2). \tag{29}$$

These inequalities are generally sharper (often considerably so) than the best inequalities (valid for $\nu > -1$) known in the literature.

3.2.2 Zeros of Laguerre polynomials

The application of Sturm's theorem (again in Szegö's form) to zeros $0 < \lambda_{n,1}^{(\alpha)} < \lambda_{n,2}^{(\alpha)} < \cdots < \lambda_{n,n}^{(\alpha)}$ of Laguerre polynomials $L_n^{(\alpha)}$ is carried out in [37]. Two types of comparison differential equations are used, one giving rise to Bessel functions, the other to Airy functions. In the former case, under the assumption $-1 < \alpha \le 1$, Luigi finds that

$$\lambda_{n,r}^{(\alpha)} < \nu\cos^2\left(\frac{1}{2}\,x_{n,r}^{(\alpha)}\right), \quad r = 1, 2, \ldots, n, \tag{30}$$

where $x_{n,r}^{(\alpha)}$ is the root of the equation

$$x - \sin x = \pi - \frac{4\,j_{\alpha,r}}{\nu}, \tag{31}$$

and $\nu = 4n + 2\alpha + 2$. In the latter case, he shows that

$$\lambda_{n,r}^{(\alpha)} > \nu \cos^2\left(\frac{1}{2}\, x_{n,r}^{*(\alpha)}\right) \quad \text{if } -1/2 \le \alpha \le 1/2, \tag{32}$$

and

$$\lambda_{n,r}^{(\alpha)} < \nu \cos^2\left(\frac{1}{2}\, x_{n,r}^{*(\alpha)}\right) \quad \text{if } -1 < \alpha \le -2/3 \text{ or } \alpha \ge 2/3, \tag{33}$$

where $x_{n,r}^{*(\alpha)}$ is the root of the equation

$$x - \sin x = \frac{8}{3\nu}\, a_{n+1-r}^{3/2} \tag{34}$$

and a_k the kth zero in ascending order of $\mathrm{Ai}(-x)$.

Since Hermite polynomials are related to Laguerre polynomials with parameters $\alpha = \pm 1/2$, and $j_{1/2,r} = r\pi$, $j_{-1/2,r} = (r - 1/2)\pi$, the inequalities (30) and (32) yield upper[5] and lower bounds for the positive zeros $0 < h_{n,\lfloor (n+1)/2\rfloor + r}$, $r = 1, 2, \ldots, \lfloor n/2\rfloor$, of the Hermite polynomial H_n:

$$h_{n,\lfloor (n+1)/2\rfloor + r} < \sqrt{2n+1} \times \begin{cases} \cos\left[\dfrac{1}{2}\, x\left(\dfrac{2n-4r+3}{2n+1}\,\pi\right)\right], & n \text{ even}, \\ \cos\left[\dfrac{1}{2}\, x\left(\dfrac{2n-4r+1}{2n+1}\,\pi\right)\right], & n \text{ odd}, \end{cases} \tag{35}$$

where $x = x(y)$ is the inverse function of $y = \sin x - x$ and

$$h_{n,\lfloor (n+1)/2\rfloor + r} > \sqrt{2n+1}\, \cos\left[\frac{1}{2}\, x\left(\frac{8}{3(2n+1)}\, a_{\lfloor n/2\rfloor + 1 - r}^{3/2}\right)\right], \quad r = 1, 2, \ldots, \lfloor n/2\rfloor. \tag{36}$$

All these inequalities are remarkably sharp.

3.2.3 Zeros of confluent hypergeometric functions

Since Laguerre polynomials $L_n^{(\alpha)}$ are special cases of confluent hypergeometric functions $\Phi(a, c; x)$ and $\Psi(a, c; x)$ (in Tricomi's notation), namely $a = -n$, $c = 1 + \alpha$, it is natural to try extending the inequalities obtained in [37] for Laguerre polynomials to confluent hypergeometric functions. This is done in [39], where Sturm-type comparison theorems are used in both Szegö's and Watson's form. With regard to the first ("regular") confluent hypergeometric function $\Phi(a, c; x)$, it is known that, if $c > 0$, there are no positive zeros of $\Phi(a, c; x)$ if $a \ge 0$, and precisely $-\lfloor a\rfloor$ positive zeros if $a < 0$. Under the assumption $a < 0$, $0 < c \le 2$, Luigi then proves that for the rth positive zero ϕ_r there holds

$$\phi_r < 4k \cos^2\left(\frac{1}{2}\, x_r\right), \quad r = 1, 2, \ldots, s, \tag{37}$$

[5] In the upper bound of (35) for n odd, the numerator $2n - 4r + 1$ in [37] is misprinted as $2n - 4r + 3$.

where $k = \frac{1}{2}c - a$, $s = \lfloor \frac{1}{4} - a \rfloor$, and x_r is the root of the equation

$$x - \sin x = \pi - \frac{j_{c-1,r}}{k}. \tag{38}$$

Note that (38) is identical with (31) in the case $a = -n$, $c = 1 + \alpha$ of Laguerre polynomials, since $k = \frac{1}{2}(1 + \alpha) + n = \nu/4$. Also, $s = \lfloor \frac{1}{4} - a \rfloor$ is either the total number of positive zeros, or one less, depending on whether $a - \lfloor a \rfloor$ is less than, or greater or equal to, 1/4.

As for the ("irregular") confluent hypergeometric function $\Psi(a, c; x)$, it is known that, if $c \geq 1$, it has no positive zeros if $a \geq 0$, and precisely $-\lfloor a \rfloor$ positive zeros ψ_r if $a < 0$. Here, assuming $a < 0$, $1 \leq c \leq 2$, Luigi proves the inequality

$$\psi_r < 4k \cos^2\left(\frac{1}{2}x_r^0\right), \quad r = 1, 2, \ldots, -\lfloor a \rfloor, \tag{39}$$

where $k = \frac{1}{2}c - a$ and x_r^0 is the root of

$$x - \sin x = \pi - \frac{j_{c-1,r}^0}{k}, \quad k = \frac{1}{2}c - a, \tag{40}$$

with $j_{c-1,r}^0$ the rth positive zero of $\cos((a - \lfloor a \rfloor)\pi)J_{c-1}(x) - \sin((a - \lfloor a \rfloor)\pi)Y_{c-1}(x)$. The case $0 < c < 1$ can be reduced to $1 < c < 2$ by applying the identity $\Psi(a - c + 1, 2 - c; x) = x^{c-1}\Psi(a, c; x)$.

Using a different differential equation for comparison in Sturm's theorem, Luigi derives additional inequalities for ϕ_r and ψ_r, where the former reduce to the inequalities (32), (33) in the case of Laguerre polynomials. Another interesting special case is $a = (1 - \nu)/2$, $\nu > 1$ and $c = 3/2$, which leads to parabolic cylinder functions D_ν and upper and lower bounds for their positive zeros $\delta_{\nu,r}$, $r = 1, 2, \ldots, -\lfloor (1 - \nu)/2 \rfloor$.

3.2.4 Inequalities from asymptotic estimates

Applications of Sturm's theorem of a somewhat different character are made in [36] and [43], where known asymptotic estimates containing order-of-magnitude terms are shown to actually become inequalities if the O-term is omitted. Such is the case, e.g., in a result of Frenzen and Wong [7, Corollary 2] concerning the zeros $\theta_{n,r}^{(\alpha,\beta)}$ of $P_n^{(\alpha,\beta)}(\cos\theta)$, which in the hands of Luigi becomes the inequality

$$\theta_{n,r}^{(\alpha,\beta)} \geq \frac{1}{N} j_{\alpha,r} - \frac{1}{4N^2}\left[\left(\frac{1}{4} - \alpha^2\right)\left(\frac{2}{t} - \cot\frac{1}{2}t\right) + \left(\frac{1}{4} - \beta^2\right)\tan\frac{1}{2}t\right],$$

$$N = n + \frac{\alpha + \beta + 1}{2}, \quad t = \frac{1}{N} j_{\alpha,r}, \tag{41}$$

valid for $|\alpha| \leq 1/2$, $|\beta| \leq 1/2$ and $r = 1, 2, \ldots, n$, with equality holding if $\alpha^2 = \beta^2 = 1/4$. In fact, (41) can be improved by replacing N in the definition of t (but not elsewhere) by $\nu = [N^2 + (1 - \alpha^2 - 3\beta^2)/12]^{1/2}$. A similar *upper* bound can be obtained by switching the parameters α and β and using a well-known identity relating $P_n^{(\alpha,\beta)}$ with $P_n^{(\beta,\alpha)}$. These inequalities are quite sharp, especially

near the respective end points π and 0. Sometimes, the upper bound in (26) may be better for the first few values of r than the upper bound obtainable from (41) by switching α and β, and likewise the lower bound obtainable similarly from the upper bound of (26) may be better than (41) for the last few values of r. Thus, in applications, (41) and (26) should be considered conjointly. All these inequalities are easily specialized to the ultraspherical case $\alpha = \beta$.

Similarly, by omitting the O-term in (7), the right-hand side becomes an upper bound in the ultraspherical case $\alpha = \beta$.

For the zeros $j_{\nu,r}$ of Bessel functions J_ν, the removal of the O-terms in some asymptotic (for large ν) estimates of Olver is conjectured in [43] to lead to upper and lower bounds, specifically to

$$\nu x_{\nu,r} < j_{\nu,r} < \nu x_{\nu,r} + g_\nu(x_{\nu,r}), \quad r = 1, 2, 3, \ldots, \tag{42}$$

where $x_{\nu,r}$ is the root of the equation $\sqrt{x^2-1} - \arctan\sqrt{x^2-1} = (2/3\nu)a_r^{3/2}$,

$$g_\nu(x) = \frac{x}{\nu}\,\frac{1}{(x^2-1)^{1/2}}\left[\frac{-5\nu}{48\,a_r^{3/2}} + \frac{5}{24(x^2-1)^{3/2}} + \frac{1}{8(x^2-1)^{1/2}}\right], \tag{43}$$

and a_r is th rth zero of $\mathrm{Ai}(-x)$. The lower bound is actually proved to hold for $\nu > 0$ and all r, and the upper bound for $\nu > 0$ and all r sufficiently large. In fact, if in (43) the right-hand side is multiplied by the factor $1 + 2^{1/3}/(280\,a_r\nu^{4/3})$, then the conjecture is proved to hold for $\nu \geq 1/2$ and all r. Heavy use is made in this work of symbolic computation with Maple V.

4 Uniform expansions

4.1 Zeros of Laguerre polynomials

Asymptotic estimates of the zeros of Laguerre polynomials $L_n^{(\alpha)}$ that resemble the inequalities in (30)–(34) are obtined in [38] from the initial terms of uniform asymptotic expansions for Laguerre polynomials due to Frenzen and Wong [8]. With $x_{n,r}^{(\alpha)}$ again denoting the root of (31), and setting $\tau_{n,r}^{(\alpha)} = \cos^2\left(\frac{1}{2}\,x_{n,r}^{(\alpha)}\right)$, from the expansion [8, Eq. (4.7)] Luigi finds the asymptotic estimate

$$\lambda_{n,r}^{(\alpha)} = \nu\tau_{n,r}^{(\alpha)} - \frac{1}{2\nu}\left[\frac{(1-4\alpha^2)\nu}{2\,j_{\alpha,r}}\left(\frac{\tau_{n,r}^{(\alpha)}}{1-\tau_{n,r}^{(\alpha)}}\right)^{1/2} + \frac{4\alpha^2-1}{2}\right.$$
$$\left. + \frac{\tau_{n,r}^{(\alpha)}}{1-\tau_{n,r}^{(\alpha)}} + \frac{5}{6}\left(\frac{\tau_{n,r}^{(\alpha)}}{1-\tau_{n,r}^{(\alpha)}}\right)^2\right] + O\left(\nu^{-3}\right), \tag{44}$$

where $\nu = 4n + 2\alpha + 2$, and the O-term is uniformly bounded for all $r = 1, 2, \ldots, \lfloor qn \rfloor$, with $0 < q < 1$ fixed.

A companion estimate, valid in the range $r = \lfloor pn \rfloor, \lfloor pn \rfloor + 1, \ldots, n$, $0 < p < 1$, which overlaps with the range for (44) when $p \leq q$, is similarly obtained from the expansion [8, Eq. (5.13)]. With $x_{n,r}^{*(\alpha)}$ again denoting the root

of (34), and setting $\tau_{n,r}^{*(\alpha)} = \cos^2\left(\frac{1}{2}\, x_{n,r}^{*(\alpha)}\right)$, the asymptotic estimate now reads

$$\lambda_{n,r}^{(\alpha)} = \nu\tau_{n,r}^{*(\alpha)} + \frac{1}{\nu}\left[\frac{5\nu}{24\, a_{n+1-r}^{3/2}}\left(\frac{\tau_{n,r}^{*(\alpha)}}{1-\tau_{n,r}^{*(\alpha)}}\right)^{1/2} + \frac{1}{4} - \alpha^2 - \frac{1}{2}\,\frac{\tau_{n,r}^{*(\alpha)}}{1-\tau_{n,r}^{*(\alpha)}} - \frac{5}{12}\left(\frac{\tau_{n,r}^{*(\alpha)}}{1-\tau_{n,r}^{*(\alpha)}}\right)^2\right] + O\big(\nu^{-3}\big), \tag{45}$$

where ν is as above in (44).

In the case where r is fixed and $\nu \to \infty$, the estimate (44) can be sharpened to

$$\lambda_{n,r}^{(\alpha)} = \frac{j_{\alpha,r}^2}{\nu}\left[1 + \frac{j_{\alpha,r}^2 + 2(\alpha^2-1)}{3\,\nu^2}\right] + O\big(\nu^{-5}\big), \quad r \text{ fixed}, \tag{46}$$

which is an old estimate of Tricomi from the 1940s. Likewise, the estimate (45) for $r = n+1-s$ and s fixed can be sharpened to

$$\lambda_{n,n+1-s}^{(\alpha)} = \nu - 2^{2/3} a_s \nu^{1/3} + \frac{1}{5}\, 2^{4/3} a_s^2 \nu^{-1/3} + O\big(\nu^{-1}\big), \quad s \text{ fixed}, \tag{47}$$

which is another of Tricomi's earlier estimate.

4.2 Zeros of confluent hypergeometric functions

Two types of uniform asymptotic expansions for Whittaker's confluent hypergeometric functions $M_{\kappa,\mu}$, $W_{\kappa,\mu}$, given by Dunster [5], are used in [9] to develop asymptotic estimates (for large κ) of the positive zeros $m_r^{(\kappa,\mu)}$, $w_r^{(\kappa,\mu)}$ of $M_{\kappa,\mu}(x)$ and $W_{\kappa,\mu}(x)$, respectively. If specialized to Laguerre polynomials, $\kappa = n + (\alpha+1)/2$, $\mu = \alpha/2$, they yield approximations for the zeros $\lambda_{n,r}^{(\alpha)}$ of $L_n^{\alpha)}$ that are now applicable for unrestrictedly large values of both n and α. Uniformity of the results, of course, comes at a price of increased complexity of the formulae.

In [45], Luigi develops two new uniform asymptotic expansions for Whittaker functions, one involving Bessel functions, the other Airy functions. Using three terms of the former, he then derives asymptotic estimates of the respective zeros, which are simpler than those obtained previously in [9] and valid as $\kappa \to \infty$ for fixed μ. Thus, for the rth positive zero of $M_{\kappa,\mu}(x)$ he finds

$$m_r^{(\kappa,\mu)} = 4\kappa\xi_r + \frac{1}{2\kappa}\left(\frac{\xi_r}{1-\xi_r}\right)^{1/2}\left[\frac{\kappa}{2}\,\frac{16\mu^2-1}{j_{2\mu,r}} - 2\mu^2\left(\frac{1-\xi_r}{\xi_r}\right)^{1/2} + \frac{1}{24}\,\frac{4\xi_r^2 - 12\xi_r + 3}{(1-\xi_r)^{3/2}\xi_r^{1/2}}\right] + O(\kappa^{-3}), \tag{48}$$

where $\xi_r = \xi_r^{(\kappa,\mu)}$ is the root of the equation

$$\arcsin\sqrt{\xi} + \sqrt{\xi - \xi^2} = \frac{j_{2\mu,r}}{2\kappa}\,. \tag{49}$$

The O-term is uniformly bounded for all $r = 1, 2, 3, \ldots$ such that $m_r^{(\kappa,\mu)} \le 4q\kappa$ with q fixed, $0 < q < 1$. Similarly, for the rth positive zero of $W_{\kappa,\mu}(x)$,

$$w_r^{(\kappa,\mu)} = 4\kappa\tau_r + \frac{1}{2\kappa}\left(\frac{\tau_r}{1-\tau_r}\right)^{1/2}\left[\frac{\kappa}{2}\,\frac{16\mu^2-1}{j_{2\mu,r}^0} - 2\mu^2\left(\frac{1-\tau_r}{\tau_r}\right)^{1/2} + \frac{1}{24}\,\frac{4\tau_r^2-12\tau_r+3}{(1-\tau_r)^{3/2}\tau_r^{1/2}}\right] + O\big(\kappa^{-3}\big), \tag{50}$$

where $\tau_r = \tau_r^{(\kappa,\mu)}$ is the root of

$$\arcsin\sqrt{\tau} + \sqrt{\tau-\tau^2} = \frac{j_{2\mu,r}^0}{2\kappa}\,, \tag{51}$$

and $j_{2\mu,r}^0$ the rth positive zero of $\sin((\kappa-\mu)\pi)J_{2\mu}(x) - \cos((\kappa-\mu)\pi)Y_{2\mu}(x)$.

Three terms of Luigi's Airy-type asymptotic expansion yield estimates valid for all zeros $m_r^{(\kappa,\mu)}$, $w_r^{(\kappa,\mu)}$ larger than $4p\kappa$, with p fixed, $0 < p < 1$. Specifically, with $n = \lceil \kappa - \mu - 1/2 \rceil$ denoting the number of positive zeros,

$$m_r^{(\kappa,\mu)} = 4\kappa\xi_r^* + \frac{1}{24\kappa}\left(\frac{\xi_r^*}{1-\xi_r^*}\right)^{1/2} \times\left[\frac{5\kappa}{c_{n+1-r}^{3/2}} - \frac{1}{2}\,\frac{(48\mu^2-4)(\xi_r^*-1)^2+4\xi_r^*+1}{(1-\xi_r^*)^{3/2}\xi_r^{*1/2}}\right] + O\big(\kappa^{-3}\big), \tag{52}$$

where $\xi_r^* = \xi_r^{*(\kappa,\mu)}$ is the root of the equation

$$\arccos\sqrt{\xi} - \sqrt{\xi-\xi^2} = \frac{c_{n+1-r}^{2/3}}{3\kappa} \tag{53}$$

and c_k the kth positive zero in ascending order of $\sin((\kappa-\mu)\pi)\mathrm{Ai}(-x) + \cos((\kappa-\mu)\pi)\mathrm{Bi}(-x)$, and

$$w_r^{(\kappa,\mu)} = 4\kappa\tau_r^* + \frac{1}{24\kappa}\left(\frac{\tau_r^*}{1-\tau_r^*}\right)^{1/2} \times\left[\frac{5\kappa}{a_{n+1-r}^{3/2}} - \frac{1}{2}\,\frac{(48\mu^2-4)(\tau_r^*-1)^2+4\tau_r^*+1}{(1-\tau_r^*)^{3/2}\tau_r^{*1/2}}\right] + O\big(\kappa^{-3}\big), \tag{54}$$

where $\tau_r^* = \tau_{n,r}^*$ is the root of

$$\arccos\sqrt{\tau} - \sqrt{\tau-\tau^2} = \frac{a_{n+1-r}^{2/3}}{3\kappa}\,, \tag{55}$$

and a_k the kth positive zero in ascending order of $\mathrm{Ai}(-x)$.

5 Miscellanea

In this section, a few of Luigi's papers are collected, which do not fit into the classification scheme we have adopted.

5.1 Retouching asymptotic formulae

The idea of "retouching" asymptotic formulae, going back to Tricomi [56], consists in introducing into the asymptotic approximation small correction terms, which can be compactly tabulated or presented graphically so as to enable a quick and relatively accurate determination of the desired quantity. The idea is particularly useful if two or more variables are involved. In [27], Luigi experiments with this idea in connection with asymptotic formulae for Bessel functions $J_\nu(x)$, $Y_\nu(x)$ in the range $x \geq 10$ and arbitrary ν with $-1 < \nu < 1$. He is able, in this way, to produce approximations accurate to about six decimals. He does the same in [28] for Laguerre polynomials $L_n(x)$, $n \geq 7$, in the oscillatory region $0 \leq x \leq 4n + 2$, where retouching is applied to two asymptotic formulae, one appropriate for the left tenth, the other for the remaining part, of the interval.

Retouching of sorts is taking place also in the paper [29], dedicated to the computation of all zeros of the generalized Laguerre polynomial $L_n^{(\alpha)}$, $\alpha > -1$. Classical results need to distinguish between zeros in three zones: a central zone and two lateral zones. Appropriate retouching of the asymptotic formula for the central zone gives rise to a unique procedure for computing *all* zeros. It involves the first $\lfloor n/2 \rfloor$ zeros of the Bessel function $J_\alpha(x)$ and of the Airy function $\mathrm{Ai}(-x)$.

5.2 Reversing asymptotic approximations

Hilb-type formulae such as (15) and their generalizations to ultraspherical and Jacobi polynomials are intended to approximate these polynomials in terms of Bessel functions, and likewise for the respective zeros. There is no intrinsic reason why this process cannot be turned around and thus be used to approximate Bessel functions in terms of, say, ultraspherical polynomials. This in fact is done in [30], where an improved Hilb formula for ultraspherical polynomials $P_n^{(\nu+1/2)}(\cos\theta)$, $\nu > -1/2$, is used to compute Bessel functions $J_\nu(x)$ in terms of them, the variable x being an appropriate multiple (depending on ν and n) of θ. Luigi's intention was to bridge in this way the gap of moderately large x, where neither the power series expansion of $J_\nu(x)$ (for small x) nor its asymptotic expansion (for large x) is numerically satisfactory. Strong competitors, however, are computational algorithms based on three-term recurrence relations satisfied by Bessel functions, which have been developed by one of us (W. G.) and others at just about the same time.

5.3 Bernstein-type inequalities

A well-known inequality for Legendre polynomials is Bernstein's inequality

$$(\sin\theta)^{1/2}|P_n(\cos\theta)| < (2/\pi)^{1/2}n^{-1/2}, \quad 0 \le \theta \le \pi, \tag{56}$$

where the constant $(2/\pi)^{1/2}$ is best possible. This result has been sharpened and generalized to ultraspherical polynomials by various authors. A generalization to Jacobi polynomials is due to Baratella [1]. By improving the constant in Baratella's result, Luigi jointly with Chow and Wong in [4] proves, for $|\alpha| \le 1/2$, $|\beta| \le 1/2$, that

$$\left(\sin\frac{1}{2}\theta\right)^{\alpha+1/2}\left(\cos\frac{1}{2}\theta\right)^{\beta+1/2}|P_n^{(\alpha,\beta)}(\cos\theta)| \le \frac{\Gamma(q+1)}{\Gamma(1/2)}\binom{n+q}{n}N^{-q-1/2},$$
$$N = n + (\alpha+\beta+1)/2, \quad 0 \le \theta \le \pi, \tag{57}$$

where $q = \max(\alpha, \beta)$. The numerical constant in (57) is best possible (cf. [52]).

5.4 Jacobi polynomials in the complex plane

In [6], Elliott obtained an asymptotic expansion for Jacobi polynomials $P_n^{(\alpha,\beta)}(z)$ which is valid uniformly for all z in the complex plane cut along the real axis from $-\infty$ to 1, with a neighborhood of $z = -1$ deleted, and with regard to the parameters α and β holds for arbitrary real β but only for $\alpha \ge 0$. From this expansion, Luigi in [3], together with Baratella, derives one- and two-term asymptotic approximations for $P_n^{(\alpha,\beta)}(z)$ with the same region of validity for z as stated above, and the same assumption on β, but with the restriction $\alpha \ge 0$ relaxed to $\alpha > -1$ by a judicious use of the differential equation and differentiation formulae satisfied by Jacobi polynomials. Analogous approximations that are valid in the z-plane cut along the real axis from -1 to $+\infty$, with a neighborhood of $z = 1$ deleted, can be obtained by switching α and β and using the reflection formula for Jacobi polynomials.

5.5 An expansion of Jacobi polynomials in Laguerre polynomials

In [40], for $\alpha > -1$, $\beta > -1$, the following curious expansion is derived,

$$P_n^{(\alpha,\beta)}(x) = \frac{(2k+t)^{n+\alpha+\beta+1}}{(2k)^{n+\alpha+1}(2k-t)^{\beta}}\,e^{-t}\sum_{m=0}^{\infty}A_m\left(k, \frac{\alpha+1}{2}\right)\left(\frac{t}{2k}\right)^m L_n^{(\alpha+m)}(t), \tag{58}$$

where

$$t = 2k\,\frac{1-x}{3+x}, \quad k = n+\beta+\frac{\alpha+1}{2}, \quad |t| < 2k,$$

and $A_m = A_m(k, \ell)$ satisfies the recurrence relation

$$(m+1)A_{m+1} = (m+2\ell-1)A_{m-1} - 2kA_{m-2}, \quad m = 2, 3, \ldots,$$

with $A_0 = 1$, $A_1 = 0$, $A_2 = \ell$. The condition $|t| < 2k$ translates into $x > -1$. In the special case $\beta = 0$, the expansion is due to Tricomi; see [58, Eq. (26)] as corrected in [60, p. 98].

5.6 Surveys

On a number of occasions, Luigi has taken time out to survey recent progress he and others had made. In an early lecture, [24], beautifully written, he explains the nature of asymptotics, the need for error bounds and techniques to obtain them, for special functions as well as for their zeros, all carefully illustrated on the example of Legendre polynomials.[6]

In [35], work on asymptotic estimates for the zeros of Jacobi polynomials and Bessel functions is reviewed.[7] There are also many original results in this survey, for example a new application of (22) to obtain the following estimate for the zeros of Jacobi polynomials $P_n^{(\alpha,\beta)}(\cos\theta)$, $|\alpha| \le 1/2$, $|\beta| \le 1/2$,

$$\theta_{n,r}^{(\alpha,\beta)} = \frac{j_{\alpha,r}}{\nu}\left[1 - \frac{4-\alpha^2-15\beta^2}{720\,\nu^4}\left(\frac{1}{2}\,j_{\alpha,r}^2 + \alpha^2 - 1\right)\right] + j_{\alpha,r}^5\,O\big(n^{-7}\big) \qquad (59)$$

valid for $r = 1, 2, \ldots, \lfloor \gamma n \rfloor$, with γ fixed in $0 < \gamma < 1$, and ν defined as in (21). Moreover, when r is fixed, (59) with the error term replaced by $O(n^{-7})$, is shown to hold for any $\alpha > -1$ and arbitrary real β. If solved for $j_{\alpha,r}$, it yields a good approximation for the first few zeros of the Bessel function J_α. The simplified $O(n^{-5})$ version of (59), with $r = 1$, has been found useful by one of us (W. G.) to discuss (in [51]) a conjectured inequality involving $\theta_{n,1}^{(\alpha,\beta)}$ and $\theta_{n+1,1}^{(\alpha,\beta)}$.

The final sections of [35] discuss inequalities holding between zeros of Jacobi polynomials and zeros of Bessel functions, some of which sharpening (26), and others extending (26), with the bounds switched, to[8] $|\alpha| > 1/2$, $|\beta| > 1/2$. In particular, many interesting and sharp upper and lower bounds are obtained for the first zero $j_{\alpha,1}$, and first few zeros $j_{\alpha,r}$, of the Bessel function J_α.

Asymptotic estimates and inequalities for the zeros $\lambda_{n,r}^{(\alpha)}$ of Laguerre polynomials $L_n^{(\alpha)}$ are reviewed in [44] and, here too, supplemented by new results.

[6] There are some misprints that may distract the reader: $\varepsilon_1(^*)$ at the bottom of p. 88 should be $\varepsilon_1(x^*)$; on p. 89, second text line, x should read x^*; and in the displayed equation that follows, the first term on the left should be multiplied by δ.

[7] For unexplained reasons, the numbers in Table 1 differ somewhat from those in the corresponding table in [50, p. 85]. In the survey, plots for these numbers are also provided.

[8] In [35, Theorem 5.1 ii)], the inclusion sign $\in$ should be $\notin$.

Thus, e.g., the formula (45) is used to derive a very sharp and interesting estimate for the last few zeros of $L_n^{(\alpha)}$, namely, when s is fixed and $n \to \infty$,

$$\lambda_{n,n+1-s}^{(\alpha)} = \nu - 2^{2/3} a_s \nu^{1/3} + \frac{1}{5} 2^{4/3} a_s^2 \nu^{-1/3} + \left(\frac{11}{35} - \alpha^2 + \frac{12}{175} a_s^3 \right) \nu^{-1}$$
$$+ \left(\frac{92}{7875} a_s^4 - \frac{16}{1575} a_s \right) 2^{2/3} \nu^{-5/3} + \left(\frac{15152}{3031875} a_s^5 - \frac{1088}{121275} a_s^2 \right) 2^{1/3}$$
$$\times \nu^{-7/3} + O\left(\nu^{-3}\right), \tag{60}$$

where $\nu = 4n + 2\alpha + 2$. This, in fact, improves an old $O(\nu^{-1})$ result of Tricomi. In obtaining (60), heavy use is made of Maple V. Luigi also conjectures that in the case $|\alpha| \leq 1/2$, when O-terms are omitted, the right-hand side of (44) becomes a lower bound for all $r = 1, 2, \ldots, n$, whereas the right-hand side of (45) becomes an upper bound for all, except the first few, zeros, and for all zeros if $-.4999 \leq \alpha \leq 1/2$.

The last three sections of [44] review results obtained by Luigi and others in the case where the parameter α is large compared to n, or both parameters α and n are large.

References

1. Baratella, P.: Bounds for the error term in Hilb fromula for Jacobi polynomials. Atti Accad. Sci. Torino, Cl. Sci. Fis. Mat. Nat. **120**(5–6), 207–223 (1986/1987)
2. Baratella, P., Gatteschi, L.: The bounds for the error term of an asymptotic approximation of Jacobi polynomials. Orthogonal Polynomials and their Applications (Segovia, 1986). Lecture Notes in Math., vol. 1329, 203–221. Springer, Berlin (1988)
3. Baratella, P., Gatteschi, L.: Remarks on asymptotics for Jacobi polynomials. Calcolo **28**(1–2), 129–137 (1991/1992)
4. Chow, Y., Gatteschi, L., Wong, R.: A Bernstein-type inequality for the Jacobi polynomial. Proc. Amer. Math. Soc. **121**(3), 703–709 (1994)
5. Dunster, T.M.: Uniform asymptotic expansions for Whittaker's confluent hypergeometric functions. SIAM J. Math. Anal. **20**(3), 744–760 (1989)
6. Elliott, D.: Uniform asymptotic expansions of the Jacobi polynomials and an associated function. Math. Comp. **25**(114), 309–315 (1971)
7. Frenzen, C.L., Wong, R.: A uniform asymptotic expansion of the Jacobi polynomials with error bounds. Can. J. Math. **37**(5), 979–1007 (1985)
8. Frenzen, C.L., Wong, R.: Uniform asymptotic expansions of Laguerre polynomials. SIAM J. Math. Anal. **19**(5), 1232–1248 (1988)
9. Gabutti, B., Gatteschi, L.: New asymptotics for the zeros of Whittaker's functions. In memory of W. Gross. Numer. Algorithms **28**(1–4), 159–170 (2001)
10. Gatteschi, L.: Una formula asintotica per l'approssimazione degli zeri dei polinomi di Legendre. Boll. Unione Mat. Ital. **4**(3), 240–250 (1949)
11. Gatteschi, L.: Approssimazione asintotica degli zeri dei polinomi ultrasferici. Univ. Roma. Ist. Naz. Alta Mat. Rend. Mat. e Appl. **8**(5), 399–411 (1949)
12. Gatteschi, L.: Sull'approssimazione asintotica degli zeri dei polinomi sferici ed ultrasferici. Boll. Unione Mat. Ital. **5**(3), 305–313 (1950)
13. Gatteschi, L.: Valutazione dell'errore nella formula di McMahon per gli zeri della $J_n(x)$ di Bessel nel caso $0 \leq n \leq 1$. Riv. Mat. Univ. Parma **1**, 347–362 (1950)
14. Gatteschi, L.: Valutazione dell'errore nella formula di McMahon per gli zeri della funzione $J_0(kz)Y_0(z) - J_0(z)Y_0(kz)$. Ann. Mat. Pura Appl. **32**(4), 271–279 (1951)

15. Gatteschi, L.: On the zeros of certain functions with application to Bessel functions. Nederl. Akad. Wetensch. Proc. Ser. A **55**; Indagationes Math. **14**, 224–229 (1952)
16. Gatteschi, L.: Limitazione dell'errore nella formula di Hilb e una nuova formula per la valutazione asintotica degli zeri dei polinomi di Legendre. Boll. Unione Mat. Ital. **7**(3), 272–281 (1952)
17. Gatteschi, L.: Una proprietà degli estremi relativi dei polinomi di Jacobi. Boll. Unione Mat. Ital. **8**(3), 398–400 (1953)
18. Gatteschi, L.: Il termine complementare nella formula di Hilb–Szegö ed una nuova valutazione asintotica degli zeri dei polinomi ultrasferici. Ann. Mat. Pura Appl. **36**(4), 143–158 (1954)
19. Gatteschi, L.: Sugli zeri di una classe di funzioni di Bessel. Atti e Relaz. Accad. Pugliese delle Scienze (nuova ser.) **12**, 3–13 (1954)
20. Gatteschi, L.: Sugli zeri della derivata delle funzioni di Bessel di prima specie. Boll. Unione Mat. Ital. **10**(3), 43–47 (1955)
21. Gatteschi, L.: Sulla rappresentazione asintotica delle funzioni di Bessel di uguale ordine ed argomento. Ann. Mat. Pura Appl. **38**(4), 267–280 (1955)
22. Gatteschi, L.: Sulla rappresentazione asintotica delle funzioni di Bessel di uguale ordine ed argomento. Boll. Unione Mat. Ital. **10**(3), 531–536 (1955)
23. Gatteschi, L.: Una nuova rappresentazione asintotica dei polinomi di Legendre mediante funzioni di Bessel. Boll. Unione Mat. Ital. **11**(3), 203–209 (1956)
24. Gatteschi, L.: Limitazione degli errori nelle formule asintotiche per le funzioni speciali. Univ. e Politec. Torino Rend. Sem. Mat. **16**, 83–94 (1956/1957)
25. Gatteschi, L.: Sulle serie inviluppanti e loro applicazioni alla valutazione asintotica delle funzioni di Bessel. Conf. Semin. Mat. Univ. Bari (**1957**)(22), 12pp. (1957)
26. Gatteschi, L.: Sul comportamento asintotico delle funzioni di Bessel di prima specie di ordine ed argomento quasi uguali. Ann. Mat. Pura Appl. **43**(4), 97–117 (1957)
27. Gatteschi, L.: Formule asintotiche "ritoccate" per le funzioni di Bessel. Tabulazione e grafici delle funzioni ausiliarie. Atti Accad. Sci. Torino, Cl. Sci. Fis. Mat. Nat. **93**, 506–514 (1958/1959)
28. Gatteschi, L.: Formule asintotiche "ritoccate" per il calcolo numerico dei polinomi di Laguerre nella zona oscillatoria. Atti Accad. Sci. Torino, Cl. Sci. Fis. Mat. Nat. **96**, 285–306 (1961/1962)
29. Gatteschi, L.: Proprietà asintotiche di una funzione associata ai polinomi di Laguerre e loro utilizzazione al calcolo numerico degli zeri dei polinomi stessi. Atti Accad. Sci. Torino, Cl. Sci. Fis. Mat. Nat. **98**, 113–124 (1963/1964)
30. Gatteschi, L.: Su un metodo di calcolo numerico delle funzioni di Bessel di prima specie. Univ. e Politec. Torino Rend. Sem. Mat. **25**, 109–120 (1965/1966)
31. Gatteschi, L.: Una nuova rappresentazione asintotica dei polinomi di Jacobi. Univ. e Politec. Torino Rend. Sem. Mat. **27**, 165–184 (1967/1968)
32. Gatteschi, L.: Una nuova disuguaglianza per gli zeri dei polinomi di Jacobi. Atti Accad. Sci. Torino, Cl. Sci. Fis. Mat. Nat. **103**, 259–265 (1969)
33. Gatteschi, L.: Sugli zeri dei polinomi ultrasferici. In: Studi in Onore di Fernando Giaccardi Giraud, pp. 111–122. Baccola & Gili, Torino (1972)
34. Gatteschi, L.: Una nuova rappresentazione asintotica dei polinomi ultrasferici, Calcolo **16**(4), 447–458 (1979/1980)
35. 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 vol. 1985, pp. 149–177
36. Gatteschi, L.: New inequalities for the zeros of Jacobi polynomials. SIAM J. Math. Anal. **18**(6), 1549–1562 (1987)
37. Gatteschi, L.: Some new inequalities for the zeros of Laguerre polynomials. Numerical methods and approximation theory, III (Niš, 1987), 23–38, Univ. Niš, Niš (1988)
38. Gatteschi, L.: Uniform approximations for the zeros of Laguerre polynomials. In: Numerical Mathematics. Internat. Schriftenreihe Numer. Math., vol. 86, 137–148. Birkhäuser, Basel (1988)
39. Gatteschi, L.: New inequalities for the zeros of confluent hypergeometric functions. In: Asymptotic and Computational Analysis (Winnipeg, MB, 1989), Lecture Notes in Pure and Appl. Math. vol. 124, pp. 175–192. Dekker, New York (1990)
40. Gatteschi, L.: On a representation of Jacobi polynomials. Atti Accad. Sci. Torino, Cl. Sci. Fis. Mat. Nat. **125**(5–6), 148–153 (1991)

41. Gatteschi, L.: New error bounds for asymptotic approximations of Jacobi polynomials and their zeros. Rend. Mat. Appl. **14**(2)(7), 177–198 (1994)
42. Gatteschi, L.: On some approximations for the zeros of Jacobi polynomials. In: Approximation and Computation (West Lafayette, IN, 1993), Internat. Ser. Numer. Math., vol. 119, pp. 207–218. Birkhäuser Boston, Boston, MA (1994)
43. Gatteschi, L.: Uniform bounds for the zeros of Bessel functions. Mem. Accad. Sci. Torino Cl. Sci. Fis. Mat. Nat. **22**(5), 185–210 (1998)
44. Gatteschi, L.: Asymptotics and bounds for the zeros of Laguerre polynomials: a survey. J. Comput. Appl. Math. **144**(1–2), 7–27 (2002)
45. Gatteschi, L.: Asymptotics for the zeros of Whittaker's functions. Atti Accad. Sci. Torino, Cl. Sci. Fis. Mat. Nat. **136**, 59–71 (2002)
46. Gatteschi, L., Giordano, C.: Upper bounds for the first zero of the Bessel function $J_\alpha(x)$. Atti Accad. Sci. Torino, Cl. Sci. Fis. Mat. Nat. **133**, 177–185 (1999)
47. Gatteschi, L., Giordano, C.: Error bounds for McMahon's asymptotic approximations of the zeros of the Bessel functions. Integral Transform. Spec. Funct. **10**(1), 41–56 (2000)
48. Gatteschi, L., Giordano, C.: On a method for generating inequalities for the zeros of certain functions. J. Comput. Appl. Math. **207**(2), 186–191 (2007)
49. Gatteschi, L., Laforgia, A.: Nuove disuguaglianze per il primo zero ed il primo massimo della funzione di Bessel $J_\nu(x)$. Rend. Sem. Mat. Univ. e Politec. Torino **34**, 411–424 (1975/1976)
50. Gatteschi, L., Pittaluga, G.: An asymptotic expansion for the zeros of Jacobi polynomials. In: Mathematical Analysis. Teubner-Texte Math., vol. 79, pp. 70–86. Teubner, Leipzig (1985)
51. Gautschi, W.: On a conjectured inequality for the largest zero of Jacobi polynomials. Numer. Algorithms **46**, (2008, this issue)
52. Gautschi, W.: How sharp is Bernstein's inequality for Jacobi polynomials? (submitted for publication)
53. Hahn, E.: Asymptotik bei Jacobi-Polynomen und Jacobi-Funktionen. Math. Z. **171**, 201–226 (1980)
54. McMahon, J.: On the roots of the Bessel and certain related functions. Ann. Math. **9**, 23–30 (1894)
55. Szegö, G.: Orthogonal polynomials, 4th edn. In: American Mathematical Society, Colloquium Publications. Amer. Math. Soc., vol. 23. Providence, RI (1975)
56. Tricomi, F.: Generalizzazione di una formula asintotica sui polinomi di Laguerre e sue applicazioni. Atti Accad. Sci. Torino, Cl. Sci. Fis. Mat. Nat. **76**, 288–316 (1941)
57. Tricomi, F.: Sugli zeri delle funzioni di cui si conosce una rappresentazione asintotica. Ann. Mat. Pura Appl. **26**(4), 283–300 (1947)
58. Tricomi, F.G.: Expansion of the hypergeometric function in series of confluent ones and application to the Jacobi polynomials. Comment. Math. Helv. **25**, 196–204 (1951)
59. Tricomi, F.G.: Funzioni Ipergeometriche Confluenti. Edizioni Cremonese, Roma (1954)
60. Tricomi, F.G.: La mia vita di matematico attraverso la cronistoria dei miei lavori. (Bibliografia commentata 1916–1967), CEDAM (Casa Editrice Dott. Antonio Milani), Padova (1967)
61. Watson, G.N.: A treatise on the theory of Bessel functions. Reprint of the second (1944) edition. In: Cambridge Mathematical Library, Cambridge University Press, Cambridge (1995)

29.12. [196] "Alexander M. Ostrowski (1893–1986): His life, work, and students"

---

[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 (2010).

Alexander M. Ostrowski (1893–1986): His life, work, and students*

Walter Gautschi

As a former student of Professor Ostrowski—one of his last—I am delighted to recall here the life and work of one of the great mathematicians of the 20th century. Needless to say that, in view of Ostrowski's immense and vastly diverse mathematical legacy, this can be done only in a most summary fashion. Further literature on Ostrowski can be found in some of the references at the end of this article. We also assemble a complete list of his Ph.D. students and trace the careers of some of them.

1. His life

The mother of Alexander

Alexander Markovich Ostrowski was born in Kiev on September 25, 1893, the son of Mark Ostrowski, a merchant in Kiev, and Vera Rashevskaya. He attended primary school in Kiev and a private school for a year before entering the Kiev School of Commerce. There, his teachers soon became aware of Alexander's extraordinary talents in mathematics and recommended him to Dmitry Aleksandrovich Grave, a professor of mathematics at the University of Kiev. Grave himself had been a student of Chebyshev in St. Petersburg before assuming a position at the University of Kharkov and, in 1902, the chair of mathematics at the University of Kiev. He is considered the founder of the Russian school of algebra, having worked on Galois theory, ideals, and equations of the fifth degree. The seminar on algebra he ran at the University of Kiev was famous at the time.

*Expanded version of a lecture presented at a meeting of the Ostrowski Foundation in Bellinzona, Switzerland, May 24–25, 2002, and published in Italian in [14].

D. A. Grave

After a few personal interviews with Alexander, Grave became convinced of Alexander's exceptional abilities and accepted him — then a boy of 15 years — as a full-fledged member of his seminar. Alexander attended the seminar for three years while, at the same time, completing his studies at the School of Commerce. During this time, with Grave's assistance, he wrote his first mathematical paper, a long memoir on Galois fields, written in Ukrainian, which a few years later (in 1913) appeared in print.

When the time came to enroll at the university, Ostrowski was denied entrance to the University of Kiev on purely bureaucratic grounds: he graduated from the School of Commerce and not from High School! This prompted Grave to write to E. Landau and K. Hensel and to ask for their help. Both responded favorably, inviting Ostrowski to come to Germany. Ostrowski opted for Hensel's offer to study with him at the University of Marburg. Two years into his stay at Marburg, another disruptive event occurred — the outbreak of World War I — which left Ostrowski a civil prisoner. Only thanks to the intervention of Hensel, the restrictions on his movements were eased somewhat, and he was allowed to use the university library. That was all he really needed. During this period of isolation, Ostrowski almost single-handedly developed his now famous theory of valuation on fields.

Alexander, ca. 1915

After the war was over and peace was restored between the Ukraine and Germany, Ostrowski in 1918 moved on to Göttingen, the world center of mathematics at that time. There, he soon stood out among the students by his phenomenal memory and his already vast and broadly based knowledge of the mathematical literature. One student later recalled that the tedious task of literature search, in Göttingen, was extremely simple: all one had to do was to ask the Russian student Alexander Ostrowski and one got the answer — instantly and exhaustively! At one time, he even had to come to the rescue of David Hilbert, when during one of his lectures Hilbert needed, as he put it, a beautiful theorem whose author unfortunately he could not recall. It was Ostrowski who had to whisper to him: "But, Herr Geheimrat, it is one of your own theorems!"

David Hilbert

Not surprisingly, therefore, Felix Klein, always keen in recognizing young talents, became interested in Ostrowski, took him on as one of his assistants, and entrusted him, together with R. Fricke, with editing the first volume of his collected works. In 1920, Ostrowski graduated *summa cum laude* with a thesis written under the guidance of Hilbert and Landau. This, too, caused quite a stir, since it answered, in part, Hilbert's 18th problem. Ostrowski succeeded in proving, among other things, that the Dirichlet zeta series $\zeta(x,s) = 1^{-s}x + 2^{-s}x^2 + 3^{-s}x^3 + \cdots$ does not satisfy an algebraic partial differential equation.

After his graduation, Ostrowski left Göttingen for Hamburg, where as assistant of E. Hecke he worked on his Habilitation Thesis. Dealing with

Ostrowski, the skater

modules over polynomial rings, this work was also inspired by Hilbert. The habilitation took place in 1922, at which time he returned to Göttingen to teach on recent developments in complex function theory and to receive habilitation once again in 1923. The academic year 1925–26 saw him as a

Rockefeller Research Fellow at Oxford, Cambridge, and Edinburgh. Shortly after returning to Göttingen, he received —and accepted— a call to the

Ostrowski, in his 40s and 50s, and at 60

University of Basel. The local newspaper (on the occasion of Ostrowski's 80th birthday) could not help recalling that 200 years earlier, the university lost Euler to St. Petersburg because, according to legend, he found himself at the losing end of a lottery system then in use for choosing candidates (in reality, he was probably considered too young for a professorship at the university). Now, however, the university hit the jackpot by bringing Ostrowski from Russia to Basel!

Ostrowski, Washington, D.C., 1964

Ostrowski remained in Basel for his entire academic career, acquiring the Basel citizenship in 1950. It was here where the bulk of his mathematical work unfolded. Much of it lies in the realm of pure mathematics, but important impulses received from repeated visits to the United States in the late forties and early fifties stirred his interest in more applied problems, particularly numerical methods in conformal mapping and problems, then emerging, relating to the iterative solution of large systems of linear algebraic equations. He went about this work with great enthu-

siasm, even exuberance, having been heard, in the halls of the *National Bureau of Standards*, to exclaim Gottfried Keller's lines "Trinkt, o Augen, was die Wimper hält, von dem goldnen Überfluss der Welt!"[1]. And indeed, exciting problems of pressing significance began to burst forward at this time and demanded nothing less than farsighted and imaginative uses of advanced mathematical techniques.

Margret Ostrowski, 1970

In 1949, Ostrowski married Margret Sachs, a psychoanalyst from the school of Carl Gustav Jung and at one time, as she once revealed to me, a secretary and confidante of Carl Spitteler[2]. Her warm and charming personality greatly helped soften the severe lifestyle of Ostrowski, the scholar, and brought into their lives some measure of joyfulness. This, in fact, is the time the author got to know the Ostrowskis, having become his student and assistant, and, on several occasions, having had the pleasure of being a guest at their house in the old part of the city.

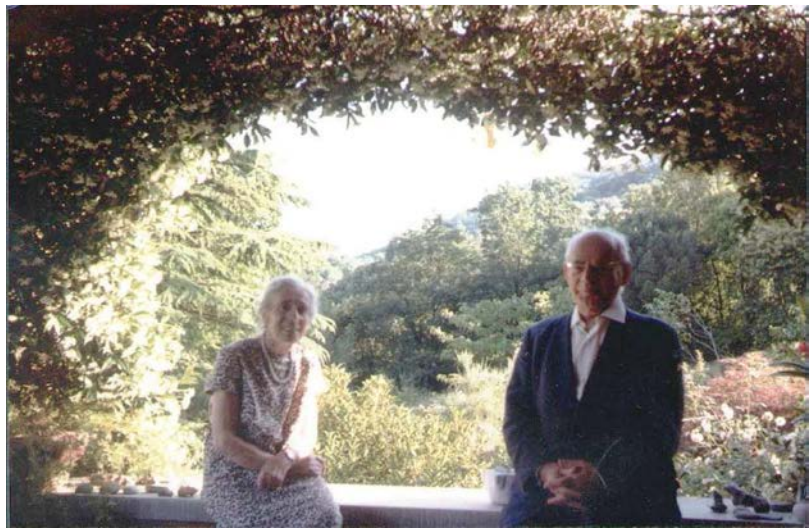
Ostrowski retired from the University in 1958. This did not bring an end to his scientific activities. On the contrary! He continued, perhaps at an even accelerated pace, to produce new and important results until his late eighties. At the age of 90, he was still able to oversee the publication by Birkhäuser of his collected papers, which appeared 1983–85 in six volumes.

75th birthday, Buffalo

After Ostrowski's retirement, he and his wife took up residence in Montagnola, where they earlier had built a beautiful villa — Casa Almarost (ALexander MARgret OSTrowski), as they named it — overlooking the Lake of Lugano. They were always happy to receive visitors at Almarost, and their gracious hospitality was legendary. Mrs. Ostrowski, knowing well the inclinations of mathematicians, always led them down to Ostrowski's library in

[1] As recalled, and kindly related to the author, by Olga Taussky-Todd.

[2] Swiss poet (1845–1924), 1919 Nobel Laureate in Literature.



Margret and Alexander Ostrowski at Almarost

order to leave them alone for a while, so they could catch up on the newest mathematics and mathematical gossip. The walls of the library were filled with books, not all mathematical, but also a good many on science fiction and mystery stories, Ostrowski's favored pastime reading.

Mrs. Ostrowski passed away in 1982, four years before Ostrowski's death in 1986. They are buried in the lovely cemetery of Gentilino, not far from the grave of Hermann Hesse, with whom they were friends.

Ostrowski's merits are not restricted to research alone; they are eminent also on the didactic level, and he exerted a major influence on mathematical publishing. With regard to teaching, his three volumes on the differential and integral calculus [22], which began to appear in the mid-1940s, and in particular the extensive collection of exercises, later published separately with solutions [23], are splendid models of mathematical exposition, which still today serve to educate generations of mathematicians and scientists. His book on the solution of nonlinear equations and systems of equations, published in the United States in 1960 and going through several edi-

Ostrowski at the age of 90

tions [24], [25], continues to be one of the standard works in the field. And last but not least, he had well over a dozen doctoral students, some having attained international stature of their own, and all remaining grateful to him for having opened to them the beauty of mathematics and imparted on them his high standards of intellectual integrity. On the publishing front, Ostrowski was a long-time consultant to the Birkhäuser-Verlag and was instrumental in establishing and supervising their well-known Green Series of textbooks. To a good extent, he can be credited for Birkhäuser having attained the leading position it now occupies in mathematical publishing.

Cemetery of Gentilino

Ostrowski's achievements did not remain unrecognized. He was awarded three honorary doctorates, one from the Federal Institute of Technology (ETH Zurich) in 1958, one from the University of Besançon in 1967, and another in 1968 from the University of Waterloo.

In the early 1980s Professor and Mrs. Ostrowski established an International Prize to be awarded every two years after their deaths [13]. It is to recognize the best achievements made in the preceding five years in Pure Mathematics and the theoretical foundations of Numerical Analysis. So far, eleven prizes have been awarded, the first in 1989 to Louis de Branges for his proof of the Bieberbach conjecture, the fourth in 1995 to Andrew Wiles for his proof of Fermat's last theorem. Characteristically of Ostrowski's view of mathematics as an international and universal science, he expressly stipulated that the award should be made "entirely without regard to politics, race, religion, place of domicile, nationality, or age." This high esteem of scientific merits, regardless of political, personal, or other shortcomings of those attaining them, came across already in 1949, when he had the courage of inviting Bieberbach — then disgraced by his Nazi past and ostracized by the European intelligentsia — to spend a semester as guest of the University of Basel and conduct a seminar on geometric constructions. Undoubtedly, it was Ostrowski who successfully persuaded Birkhäuser to publish the seminar in book form [3].

2. His work

Let us now take a quick look at Ostrowski's mathematical work. A first appreciation of the vast scope of this work can be gained from the headings in the six volumes of his collected papers [27]:

Vol. 1 Determinants, Linear Algebra, Algebraic Equations;

Vol. 2 Multivariate Algebra, Formal Algebra;

Vol. 3 Number Theory, Geometry, Topology, Convergence;

Vol. 4 Real Function Theory, Differential Equations, Differential Transformations;

Vol. 5 Complex Function Theory;

Vol. 6 Conformal Mapping, Numerical Analysis, Miscellany.

Much of this work is at the highest levels of mathematics and can be indicated here only by key words and phrases. The same applies to work that, although more accessible, is difficult to adequately summarize in a few words. From the remaining papers, a few results are selected in chronological order and briefly sketched in "excerpts", hoping in this way to provide a glimpse into Ostrowski's world of mathematics. We go through this work volume by volume and add dates to indicate the period of his life in which the respective papers have been written.

2.1. Volume 1

Key words: Sign rules of Descartes, Budan–Fourier, and Runge (1928–65); critique and correction of Gauss's first and fourth proof of the Fundamental Theorem of Algebra (1933); long memoir on Graeffe's method (1940); linear iterative methods for symmetric matrices (1954); general theory of vector and matrix norms (1955); convergence of the Rayleigh quotient iteration for computing real eigenvalues of a matrix (1958–59); Perron–Frobenius theory of nonnegative matrices (1963–64)

Excerpt 1.1. Matrices with *dominant diagonal* (1937),

$$A = [a_{ij}], \quad d_i := |a_{ii}| - \sum_{j \neq i} |a_{ij}| > 0, \quad \text{all } i.$$

Hadamard in 1899 proved that for such matrices $\det A \neq 0$. Ostrowski sharpens this to $|\det A| \geq \prod_i d_i$.

Excerpt 1.2. *M-matrices* (1937),

$$A = [a_{ij}], \quad a_{ii} > 0, \quad a_{ij} \leq 0 \ \ (i \neq j),$$

$$a_{11} > 0, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0, \quad \ldots, \quad \det A > 0.$$

Theorem. *If A is an M-matrix, then $A^{-1} \geq \mathbf{0}$.*

The theory of M-matrices and the related theory of H-matrices, stemming from Ostrowski's 1937 paper, have proved to be powerful tools in the analysis of iterative methods for solving large systems of linear equations. In addition, this theory forms the basis for the general theory of eigenvalue inclusion regions for matrices, as in the case of the well-known Gershgorin Theorem. See also Excerpt 2.2.

Excerpt 1.3. *Continuity of the roots* of an algebraic equation (1939).

It is well known that the roots of an algebraic equation depend continuously on the coefficients of the equation. Ostrowski gives us a quantitative formulation of this fact.

Theorem. *Let x_ν, y_ν be the zeros of*

$$p(z) = a_0 z^n + a_1 z^{n-1} + \cdots + a_n, \quad a_0 a_n \neq 0,$$

resp.

$$q(z) = b_0 z^n + b_1 z^{n-1} + \cdots + b_n, \quad b_0 b_n \neq 0.$$

If

$$b_\nu - a_\nu = \varepsilon_\nu a_\nu, \quad |\varepsilon_\nu| \leq \varepsilon, \quad 16 n \varepsilon^{1/n} \leq 1,$$

then

$$\left| \frac{x_\nu - y_\nu}{x_\nu} \right| \leq 15 n \varepsilon^{1/n}.$$

Excerpt 1.4. Convergence of the *successive overrelaxation method* (1954).

The iterative solution of large (nonsingular) systems of linear algebraic equations

$$Ax = b, \quad A \in \mathbb{R}^{n \times n}, \ b \in \mathbb{R}^n,$$

was an object of intense study in the 1950s culminating in the "successive overrelaxation method" (SOR)

$$Dx^{k+1} = \omega(b - Lx^{k+1} - Ux^k) - (\omega - 1)Dx^k, \quad k = 0, 1, 2, \ldots,$$

where ω is a real parameter and D, L, U are, respectively, the diagonal, lower triangular, and upper triangular part of A. The method is said to converge if $\lim_{k \to \infty} x^k = A^{-1}b$ for arbitrary b and arbitrary $x^0 \in \mathbb{R}^n$.

Ostrowski–Reich Theorem. *If A is symmetric with positive diagonal elements, and $0 < \omega < 2$, then SOR converges if and only if A is positive definite.*

Reich proved the theorem for $\omega = 1$ in 1949. Ostrowski proved it for general ω in $(0,2)$, even when $\omega = \omega_k$ depends on k but remains in any compact subinterval of $(0,2)$.

Excerpt 1.5. A little mathematical jewel (1979).

Theorem. *Let p and q be polynomials of degrees m and n, respectively. Define*

$$M_f = \max_{|z|=1} |f(z)|.$$

Then

$$\gamma M_p M_q \le M_{pq} \le M_p M_q, \quad \gamma = \sin^m \frac{\pi}{8m} \sin^n \frac{\pi}{8n}.$$

The interest here lies in the lower bound, the upper one being trivial. It is true that this lower bound may be quite small, especially if m and/or n are large. But jewels need not be useful as long as they shine!

2.2. Volume 2

Key words: Algebra of finite fields (1913); theory of valuation on a field (1913–17); necessary and sufficient conditions for the existence of a finite basis for a system of polynomials in several variables (1918–20); various questions of irreducibility (1922, 1975–77); theory of invariants of binary forms (1924); arithmetic theory of fields (1934); structure of polynomial rings (1936); convergence of block iterative methods (1961); Kronecker's elimination theory for polynomial rings (1977).

The fact, proved by Ostrowski in 1917, that the fields of real and complex numbers are the only fields, up to isomorphisms, which are complete (Ostrowski used the older term "perfect" for "complete") with respect to an Archimedean valuation is known today as "Ostrowski's Theorem" in valuation theory (P. Roquette [31]).

Excerpt 2.1. Evaluation of polynomials (1954). If

$$p(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n,$$

then, by *Horner's rule*, $p(x) = p_n$, where

$$p_0 = a_0, \quad p_\nu = x p_{\nu-1} + a_\nu, \quad \nu = 1, 2, \ldots, n.$$

Complexity: n additions, n multiplications.

Theorem. *Horner's rule is optimal for addition and optimal for multiplication when* $n \leq 4$.

It has later been shown by V. Ja. Pan [28] that Horner's scheme indeed is *not* optimal with respect to multiplication when $n > 4$.

Because of this paper, the year 1954 is generally considered "the year of birth of algebraic complexity theory" (P. Bürgisser and M. Clausen [5]).

Excerpt 2.2. Metric properties of *block matrices* (1961),

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & \vdots & & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{bmatrix}, \quad A_{\nu\mu} \in \mathbb{R}^{\nu\times\mu}.$$

Question. *Is Hadamard's theorem still valid if* $|\cdot|$ *is replaced by* $\|\cdot\|$?

Answer: Yes, if

$$\begin{bmatrix} \|A_{11}\|^* & -\|A_{12}\| & \dots & -\|A_{1n}\| \\ -\|A_{21}\| & \|A_{22}\|^* & \dots & -\|A_{2n}\| \\ \vdots & \vdots & & \vdots \\ -\|A_{n1}\| & -\|A_{n2}\| & \dots & \|A_{nn}\|^* \end{bmatrix}$$

is an M-matrix, where

$$\|B\|^* = \min_{\|x\|=1} \|Bx\|, \quad \|B\| = \max_{\|x\|=1} \|Bx\|.$$

2.3. Volume 3

Key words: Existence of a "regular" basis for polynomials with coefficients in a finite arithmetic field that take on integer values for integer arguments (1919); arithmetic theory of algebraic numbers (1919); Diophantine equations and approximations (1921–27, 1964–82); existence criterion for a common zero of two real functions continuous inside and on the boundary of a disk (1933); topology of oriented line elements (1935); evolutes and evolvents of a plane curve (1955) and an oval in particular (1957); differential geometry of plane parallel curves (1955); Ermakov's convergence and divergence criteria for $\int^{\infty} f(x)\,dx$ (1955); necessary and sufficient conditions for two line elements to be connectable by a curve with monotone curvature (1956); behavior of fixed-point iterates in the case of divergence (1956); summation of slowly convergent positive or alternating series (1972).

Excerpt 3.1. Infinite products (1930),

$$x_0 = x, \quad x_{\nu+1} = \varphi(x_\nu), \quad \nu = 0, 1, 2, \ldots,$$
$$\prod_{\nu=0}^{\infty} (1 + x_\nu) = \Phi(x).$$

Example. Euler's product $\varphi(x) = x^2$, $\Phi(x) = (1 - x)^{-1}$.

Problem. Determine *all* products which converge in a neighborhood of $x = 0$, and for which φ is rational and Φ algebraic.

Solution: completely enumerated.

Excerpt 3.2. "Normal" power series (1930),

$$\sum_{\nu=-\infty}^{\infty} a_\nu z^\nu \text{ with } a_\nu \geq 0, \; a_\nu^2 \geq a_{\nu-1} a_{\nu+1},$$

and all coefficients between two positive ones are also positive.

Theorem. *The product of two normal power series, if it exists, is also normal.*

2.4. Volume 4

Key words: Dirichlet series and algebraic differential equations, thesis Göttingen (1919); strengthening, or simplifying, proofs of many known results from real analysis (1919–38); various classes of contact transformations in the sense of S. Lie (1941–42); invertible transformations of line elements (1942); conditions of integrability for partial differential equations (1943); indefinite integrals of "elementary" functions, Liouville Theory (1946); convex functions in the sense of Schur with applications to spectral properties of Hermitian matrices (1952); theory of characteristics for first-order partial differential equations (1956); points of attraction and repulsion for fixed-point iteration in Euclidean space (1957); univalence of nonlinear transformations in Euclidean space (1958); a decomposition of an ordinary second-order matrix differential operator (1961); theory of Fourier transforms (1966); study of the remainder term in the Euler–Maclaurin formula (1969-70); asymptotic expansion of integrals containing a large parameter (1975).

A technique introduced in the 1946 paper on Liouville's Theory is now known in the literature as the "Hermite–Ostrowski method" (J. H. Davenport, Y. Siret, and E. Tournier [7]). This work has attained renewed relevance because of its use in formal integration techniques of computer algebra.

Excerpt 4.1. The (frequently cited) *Ostrowski–Grüss inequality* (1970),

$$\left| \int_0^1 f(x)g(x)\,dx - \int_0^1 f(x)\,dx \int_0^1 g(x)dx \right| \le \frac{1}{8} \operatorname*{osc}_{[0,1]} f \max_{[0,1]} |g'|.$$

Excerpt 4.2. Generalized *Cauchy–Frullani integral* (1976),

$$\int_0^\infty \frac{f(at) - f(bt)}{t}\,dt = [M(f) - m(f)] \ln \frac{a}{b}, \quad a > 0,\ b > 0,$$

where

$$M(f) = \lim_{x\to\infty} \frac{1}{x} \int_1^x f(t)\,dt, \quad m(f) = \lim_{x\downarrow 0} x \int_x^1 \frac{f(t)}{t^2}\,dt.$$

In the original version of the formula, there were point evaluations, $f(\infty)$ and $f(0)$, in place of the mean values $M(f)$ and $m(f)$.

2.5. Volume 5

Key words: Gap theorems for power series and related phenomena of "overconvergence" (1921–30); investigations related to Picard's theorem (1925–33); quasi-analytic functions, the theory of Carleman (1929); analytic continuation of power series and Dirichlet series (1933, 1955).

Excerpt 5.1. Alternative characterization of *normal families of meromorphic functions* (1925).

Theorem. *A family $\mathcal{F}$ of meromorphic functions is normal (i.e., precompact) if and only if it is equicontinuous with respect to the spherical metric.*

Excerpt 5.2. *Carleman's theorem* on quasianalytic functions, as reformulated by Ostrowski (1929).

Given a sequence $m = \{m_\nu\}_{\nu=1}^\infty$ of positive numbers m_ν, an infinitely-differentiable function f on $I = [0, \infty)$ is said to belong to the class $C(m)$ if

$$|f^{(\nu)}(x)| \le m_\nu \text{ on } I, \quad \nu = 0, 1, 2, \ldots.$$

The class $C(m)$ is called quasianalytic if $f \in C(m)$ and $f^{(\nu)}(0) = 0$, $\nu = 0, 1, 2, \ldots$, implies $f(x) \equiv 0$ on I.

Ostrowski reformulates, and gives a simplified proof of, one of the main results of Carleman's theory of quasianalytic functions by introducing the function $T(r) = \sup_\nu r^\nu / m_\nu$ (sometimes named after him).

Theorem. *The class $C(m)$ is quasianalytic if and only if*

$$\int_1^\infty \log T(r) \frac{dr}{r^2} = \infty.$$

Ostrowski's work related to Picard's theorem, though predating R. Nevanlinna's own theory of meromorphic functions, points in the same direction.

2.6. Volume 6

Key words: Constructive proof of the Riemann Mapping Theorem (1929); boundary behavior of conformal maps (1935–36); Newton's method for a single equation and a system of two equations: convergence, error estimates, robustness with respect to rounding (1937–38); convergence of relaxation methods for linear $n \times n$ systems, optimal relaxation parameters for $n = 2$ (1953); iterative solution of a nonlinear integral equation for the boundary function of a conformal map, application to the conformal map of an ellipse onto the disk (1955); "absolute convergence" of iterative methods for solving linear systems (1956); convergence of Steffensen's iteration (1956); approximate solution of homogeneous systems of linear equations (1957); a device of Gauss for speeding up iterative methods (1958); convergence analysis of Muller's method for solving nonlinear equations (1964); convergence of the fixed-point iteration in a metric space in the presence of "rounding errors" (1967); convergence of the method of steepest descent (1967); a descent algorithm for roots of algebraic equations (1969); Newton's method in Banach spaces (1971); a posteriori error bounds in iterative processes (1972–73); probability theory (1946–1980); book reviews, public addresses, obituaries (G. H. Hardy, Wilhelm Süss, Werner Gautschi) (1932–75).

Excerpt 6.1. Matrices close to a triangular matrix (1954),

$$A = [a_{ij}], \quad |a_{ij}| \le m \ (i > j), \quad |a_{ij}| \le M \ (i < j), \quad 0 < m < M.$$

The limit case $m = 0$ corresponds to a triangular matrix with its eigenvalues being the elements on the diagonal. If m is small, one expects the eigenvalues to remain near the diagonal elements. This is expressed by Ostrowski in the following way.

Theorem. *All eigenvalues of A are contained in the union of disks $\bigcup_i D_i$, $D_i = \{z \in \mathbb{C} : |z - a_{ii}| \le \delta(m, M)\}$, where*

$$\delta(m, M) = \frac{M m^{\frac{1}{n}} - m M^{\frac{1}{n}}}{M^{\frac{1}{n}} - m^{\frac{1}{n}}}.$$

The constant $\delta(m, M)$ is best possible.

Excerpt 6.2. The *Moivre–Laplace formula* (1980). If

$$M(n) = \sum_{|\nu - np| \le \eta\sqrt{2npq}} \binom{n}{\nu} p^\nu q^{n-\nu}, \quad 0 < p < 1,\ p + q = 1,\ n > 0,$$

then

$$M(n) = \frac{2}{\sqrt{\pi}} \int_0^\eta e^{-t^2} dt + \rho(\eta, n),$$

where

$$\rho(\eta, n) = \frac{r_n}{\sqrt{2\pi npq}} e^{-\eta^2} + O(1/n), \quad n \to \infty,$$

and, with $R(x) = x - \lfloor x \rfloor$,

$$r_n = 1 - R(nq + \eta\sqrt{2npq}) - R(np + \eta\sqrt{2npq}).$$

The numbers r_n are everywhere dense in $[-1, 1]$. Prior to Ostrowski's work, the formula has been stated (incorrectly) with 1 in place of r_n.

3. His students

Professor Ostrowski has been the primary advisor ("Referent") for the doctoral students listed below. All dissertations, except one, were written at the Faculty of Mathematics and Natural Sciences of the University of Basel. (The exception is the thesis by Willy Richter.)

1932 Stefan Emanuel Warschawski (1904–1989)
"Über das Randverhalten der Ableitung der Abbildungsfunktion bei konformer Abbildung"

1933 Alwin von Rohr (1903–2001)
"Über die Hilbert-Story'schen invariantenerzeugenden Prozesse"

1934 Leo Leib Krüger (1903–?)
"Über eine Klasse von kontinuierlichen Untergruppen der allgemeinen linearen homogenen projektiven Gruppe des $(2N-1)$-dimensionalen Raumes"

1936 Theodor Samuel Motzkin (1908–1970)
"Beiträge zur Theorie der linearen Ungleichungen"

1938 Caleb Gattegno (1911–1988)
"Le cas essentiellement géodésique dans les équations de Hamilton-Jacobi intégrables par séparation des variables"

1938 Fritz Blumer (1904–1988)
"Untersuchungen zur Theorie der halbregelmässigen Kettenbruchentwicklungen, I & II"

1944 Eduard Batschelet (1914–1979)
"Untersuchungen über die absoluten Beträge der Wurzeln algebraischer, insbesondere kubischer Gleichungen"

1945 Gerhard Stohler (1915–1999)
"Über eine Klasse von einparametrigen Differential-Transformationsgruppen"

1948 Rolf Conzelmann (1916–)
"Beiträge zur Theorie der singulären Integrale bei Funktionen von mehreren Variablen, I & II"

1949 Karl-Felix Moppert (1920–1984)
"Über Relationen zwischen m- und p-Funktionen"

1951 Hermann Georg Wundt (1921–?)
"Eine neue Methode der Periodogramm-Analyse und ihre Anwendung auf die Reihe der Sonnenflecken-Relativzahlen"

1952 Willy Richter (1915–1998)
"Estimation de l'erreur commise dans la méthode de M. W. E. Milne pour l'intégration d'un système de n équations différentielles du premier ordre" (Thèse, Faculté des Sciences, Université de Neuchâtel)

1953 Rudolf Thüring (1924–)
"Studien über den Holditchschen Satz"

1954 Werner Gautschi (1927–1959)
"On norms of matrices and some relations between norms and eigenvalues"

1954 Walter Gautschi (1927–)
"Analyse graphischer Integrationsmethoden"

1959 Hans Richard Gutmann (1907–2001)
"Anwendung Tauberscher Sätze und Lambertscher Reihen in der zahlentheoretischen Asymptotik"

Many of these students have had successful careers either in academia or in secondary school education. Like Ostrowski himself, some of the earlier students came to Basel from abroad: Warschawski from Königsberg; Krüger from Riga; Motzkin from Berlin; and Gattegno from Alexandria, Egypt. All the other students, except Wundt, a native of Aalen, Württemberg, were born and grew up in, or near, Basel.

We have no information about the careers of *von Rohr*, *Krüger*, and *Wundt*.

Warschawski became a Ph.D. student of Ostrowski while the latter was still in Göttingen, and moved with him to Basel, where he completed his thesis in 1932. He returned to Göttingen to start his teaching career but was forced to escape from Nazi persecution. He was able, eventually, to reach the United States, where he developed into a highly respected researcher in the area of conformal mapping. He also distinguished himself as a successful academic administrator by building up to prominence two departments of mathematics, one at the University of Minnesota, the other at the University of California at San Diego. For a biography, see [21].

Motzkin, the son of Leo Motzkin, a prominent member of the Zionist movement who participated at the First Zionist Congress (1897) in Basel and in his youth started on a doctoral dissertation under Kronecker, after completion of his thesis moved to the Hebrew University in Jerusalem, where during World War II he worked as a cryptographer for the British government. In 1948 he emigrated to the United States, where in 1950 he became a member of the Institute of Numerical Analysis at the University of California at Los Angeles and a professor ten years later. Motzkin's work as a mathematician is widely recognized to be brilliant and ingenious. Extremely versatile, he contributed significantly to fields such as linear programming, combinatorics, approximation theory, algebraic geometry, number theory, complex function theory, and numerical analysis. Motzkin numbers and Motzkin paths are mathematical objects still studied extensively in today's literature. See [1] for an obituary.

Gattegno turned his attention to the psychology and didactics of teaching in general, and of teaching mathematics, reading and writing, and foreign languages, in particular. He promoted his innovative and unorthodox approaches in more than 50 books and other publications, conducted seminars throughout the world, founded numerous organizations, and produced relevant teaching material. He earned a second doctorate in psychology in 1952 from the University of Lille. In 1965, Gattegno moved to New York, where he established an educational laboratory and continued his pedagogical activities. For more on Gattegno's life and work, see [29].

Batschelet was a teacher at the Humanistischen Gymnasium Basel from 1939 to 1960 and a Privatdozent at the University of Basel from 1952 to 1957. In 1958 he was awarded the title of extraordinary professor and two years later moved to Washington, D.C. to assume a professorship at the Catholic University. He returned to Switzerland in 1971 where he became professor of mathematics at the University of Zurich. His field of research was statistics and biomathematics; he taught and wrote successful textbooks in this area. See [18] for an obituary.

Moppert, after five years of teaching at schools in Basel, emigrated to Australia, where he assumed a lectureship at the University of Tasmania and in 1958 became a senior lecturer in mathematics at the University of Melbourne. In 1967 he joined the Department of Mathematics at Monash University, where he remained until his death. His mathematical work addressed Riemann surfaces — his thesis topic — and miscellaneous other topics including operators in Hilbert space, Diophantine analysis, Brownian motion, and Euclidean and non-Euclidean geometry. He had a knack for scientific instruments, of which a sundial mounted on one of the walls of the Union Building at Monash, "often a better indicator of the correct time than most other clocks on campus" [6], remains a lasting witness.

Werner Gautschi, a twin brother of the author, emigrated in 1953 to the United States, where during postdoctoral years at Princeton University and the University of California at Berkeley he worked himself into the areas of mathematical statistics and probability theory. He started his academic career in 1956 at Ohio State University, moved to Indiana University at Bloomington in 1957 and two years later back to Ohio State University. Soon after he arrived there, a massive heart attack put an abrupt end to his life and to a very promising career. See [4] and [26] for obituaries.

Walter Gautschi, after two years of postdoctoral work in Rome and at Harvard University, took on positions as a research mathematician at (what was then called) the National Bureau of Standards in Washington, D.C. and at Oak Ridge National Laboratory, Oak Ridge, Tennessee. In 1963 he accepted a professorship in mathematics and computer sciences at Purdue University, where he remained until his retirement in 2000. He worked in the areas of special functions, constructive approximation theory, and numerical analysis, as documented in [15].

Among the students who chose a teaching career at schools in Basel are *Blumer*, Humanistisches Gymnasium (HG), 1932–1973; *Stohler*, Mädchengymnasium (MG) (later Holbein-Gymnasium), 1946–1980; *Conzelmann*, HG (later Mathematisch-Naturwissenschaftliches Gymnasium (MNG)), 1949–1982; *Thüring*, Realgymnasium (RG), 1956–1986; *Gutmann*, RG, 1935–1970 (rector thereof from 1962–1970). Both, Blumer and Conzelmann held also academic positions at the University of Basel, the former a lectorship from 1960 to 1974, the latter a Lehrauftrag in 1956/57, a lectorship from 1958 to 1974, and an extraordinary professorship from 1975 until his retirement in 1984. *Richter*, injured in a military accident and battling tuberculosis, absolved his university studies by correspondence in the military sanatorium of Novaggio and the sanatorium in Leysin during World War II and wrote most of his thesis on the sick-bed. He became a teacher in Neuchâtel, for a few years at the École de Commerce and then at the Gymnase Cantonal until his retirement in 1978.

Ostrowski is listed as secondary advisor ("Korreferent") to the following students:

1931 Heinrich Johann Ruch (1895–1960)
"Über eine Klasse besonders einfacher Modulargleichungen zweiten Grades von der Form $y^2 = R(x)$" (Referent: Otto Spiess)

1942 Ernst Fischer (1914–2000)
"Das Zinsfussproblem der Lebensversicherungsrechnung als Interpolationsaufgabe" (Referent: Ernst Zwinggi)

1947 Heinz Hermann Müller (1913–1996)
"Scharfe Fassung des Begriffes faisceau in einer gruppentheoretischen Arbeit Camille Jordans"
(Referent: Andreas Speiser)

1955 Mario Gottfried Howald (1925–2001)
"Die akzessorische Irrationalität der Gleichung fünften Grades" (Referent: Andreas Speiser)

Nothing is known to us about the curricula vitae of these students except for *Howald*, who was teaching at the MNG from 1951 to 1990 (in between for four years at the Gymnasium Bäumlihof). For two years (1962–63) he was working at the Natural Science section of the Goetheanum in Dornach. Besides his teaching activity at the Gymnasien, Howald regularly organized courses in Carona (near Lugano) for amateur astronomers. He is the author of two informative articles [16], [17] on Maupertuis's Lapland expedition to measure the length of a meridional degree that led to the affirmation of the flatness of the earth near the poles. He also edited, and wrote commentaries to, Daniel Bernoulli's work on positional astronomy [2] and from 1997 to his death was a member of the Curatorium of the Otto Spiess foundation which supports the Bernoulli edition.

4. Epilogue

To conclude, let me make a few general remarks about Ostrowski's work. Apart from the kaleidoscopic variety of themes treated by him, a characteristic quality of his work is a strong desire to go to the bottom of things, to unravel the essential features of a problem and the basic concepts needed to deal with it in a satisfactory manner. This is coupled with a relentless drive to be exhaustive. Notable are also his frequent attempts to establish results, even entirely classical ones, under the weakest assumptions possible, and his delight in finding proofs that are short and succinct. A

good part of Ostrowski's work has a definite constructive bent, and all of it exhibits a masterly skill in the use of advanced mathematical techniques, particularly analytic techniques of estimation. His work bears the stamp of scholarly thoroughness, coming from a careful study of the literature, not only the contemporary literature, but also, and perhaps more importantly, the original sources.

Acknowledgments and sources. The author gratefully acknowledges help from Dr. H. Wichers of the Staatsarchiv Basel-Stadt to find all Ph.D. students of Prof. Ostrowski and biographical data for some of them. He is indebted to Professor D. Drasin for help with §2.5, to Mireille Richter for providing information about her father's life, and to Dr. F. Nagel for details about Howald's career. The photographs in Section 1 are, for the most part, from the private property of Prof. Dr. R. Conzelmann. Those of Ostrowski in Washington, D.C., and of Margret Ostrowski, are from the Oberwolfach Photo Collection, and the one of Ostrowski in his 50s from the author's own possession.

References

[1] Anonymous, Obituary: Theodore Samuel Motzkin, Professor of Mathematics, 1908–1970. *J. Combin. Theory Ser. A* **14** (1973), 271–272.

[2] D. Bernoulli, *Die Werke von Daniel Bernoulli.* Bd. 1: Medizin und Physiologie. Mathematische Jugendschriften. Positionsastronomie. Birkhäuser, Basel 1996.

[3] L. Bieberbach, *Theorie der geometrischen Konstruktionen.* Birkhäuser, Basel 1952.

[4] J. R. Blum, Werner Gautschi, 1927–1959. *Ann. Math. Statist.* **31** (1960), 557.

[5] P. Bürgisser und M. Clausen, Algebraische Komplexitätstheorie. I. Eine Einführung. *Sém. Lothar. Combin.* **36** (1996), Art. S36a.

[6] J. N. Crossley and J. B. Miller, Obituary: Carl Felix Moppert, 1920–1984. *J. Austral. Math. Soc. Ser. A* **42** (1987), 1–4.

[7] J. H. Davenport, Y. Siret, and E. Tournier, *Computer algebra.* 2nd ed., Academic Press Ltd., London 1993.

[8] M. Eichler, Alexander Ostrowski. Über sein Leben und Werk. *Acta Arith.* **51** (1988), 295–298.

[9] D. K. Faddeev, On R. Jeltsch-Fricker's paper "In memoriam Alexander M. Ostrowski (1893–1986)". *Algebra i Analiz* **2** (1990), no. 1, 242–243; English transl. *Leningrad Math. J.* **2** (1991), no. 1, 205–206 (see [19]).

[10] Walter Gautschi, To Alexander M. Ostrowski on his ninetieth birthday. *Linear Algebra Appl.* **52/53** (1983), xi–xiv.

[11] Walter Gautschi, Alexander Ostrowski 90jährig. *Neue Zürcher Zeitung*, September 24, 1983.

[12] Walter Gautschi, Obituary: A. M. Ostrowski (1893–1986). *SIAM Newsletter* **20** (January 1987), 2, 19.

[13] Walter Gautschi, Ostrowski and the Ostrowski Prize. *Math. Intelligencer* **20** (1998), 32–34; German edited transl. *Uni Nova* **87** (June 2000), Universität Basel, 60–62.

[14] Walter Gautschi, Alessandro M. Ostrowski (1893–1986): la sua vita e le opere. *Boll. Docenti Matem.* **45** (2002), 9–19.

[15] Walter Gautschi, A guided tour through my bibliography. *Numer. Algorithms* **45** (2007), 11–35.

[16] M. Howald-Haller, Wie hat Maupertuis die Abplattung der Erde gemessen? *Mitteilungen des Heimatmuseums Schwarzbubenland*, Nr. 30/31 (1993/1994), 5–20. [Reissued in 2009.]

[17] M. Howald-Haller, Maupertuis' Messungen in Lappland. In *Pierre Louis Moreau de Maupertuis* (H. Hecht, ed.), Schriftenr. Frankreich-Zentr. TU Berlin 3., Berlin Verlag Arno Spitz, Berlin 1999, 71–88.

[18] R. Ineichen, Professor Dr. Eduard Batschelet (6. April 1914 bis 3. Oktober 1979). *Elem. Math.* **35** (1980), 105–107.

[19] R. Jeltsch-Fricker, In memoriam Alexander M. Ostrowski (1893 bis 1986). *Elem. Math.* **43** (1988), 33–38; Russian transl. *Algebra i Analiz* **2** (1990), no. 1, 235–241; annotated English transl. *Leningrad Math. J.* **2** (1991), no. 1, 199–203.

[20] P. Lancaster, Alexander M. Ostrowski, 1893–1986. *Aequationes Math.* **33** (1987), 121–122.

[21] F. D. Lesley, Biography of S. E. Warschawski. *Complex Variables Theory Appl.* **5** (1986), 95–109.

[22] A. Ostrowski, *Vorlesungen über Differential- und Integralrechnung*. Bd. I: Funktionen einer Variablen. Bd. II: Differentialrechnung auf dem Gebiete mehrerer Variablen; Bd. III: Integralrechnung auf dem Gebiete mehrerer Variablen. Birkhäuser, Basel 1945, 1951, 1954.

[23] A. Ostrowski, *Aufgabensammlung zur Infinitesimalrechnung*. Bd. I: Funktionen einer Variablen; Bd. IIA: Differentialrechnung auf dem Gebiete mehrerer Variablen. Aufgaben und Hinweise; Bd. IIB: Differentialrechnung auf dem Gebiete mehrerer Variablen. Lösungen; Bd. III: Integralrechnung auf dem Gebiete mehrerer Variablen. Birkhäuser, Basel 1964, 1972, 1972, 1977.

[24] A. M. Ostrowski, *Solution of equations and systems of equations*. 1st and 2nd ed., Pure Appl Math. 9, Academic Press, New York 1960, 1966.

[25] A. M. Ostrowski, *Solution of equations in Euclidean and Banach spaces*, 3rd ed. of [24]. Academic Press, New York 1973.

[26] A. Ostrowski, Werner Gautschi, 1927–1959. *Verh. Naturforsch. Ges. Basel* **71** (1960), 314–316.

[27] A. Ostrowski, *Collected mathematical papers.* Vols. 1–6. Birkhäuser, Basel 1983–1985.

[28] V. Ja. Pan, Some schemes for computation of polynomials with real coefficients (Russian). *Dokl. Akad. Nauk SSSR* **127** (1959), 266–269.

[29] A. B. Powell, Caleb Gattegno (1911–1988): a famous mathematics educator from Africa? *Rev. Bras. Hist. Mat.* **2007** (2007), 199–209.

[30] J. M. Rassias, Stefan Banach, Alexander Markowiç Ostrowski, Stanisław Marcin Ulam. In *Functional analysis, approximation theory and numerical analysis*, World Sci. Publ., Singapore 1994, 1–4.

[31] P. Roquette, History of valuation theory. I. In *Valuation theory and its applications*, Vol. I (Saskatoon, SK, 1999), Fields Inst. Commun. 32, Amer. Math. Soc., Providence, RI, 2002, 291–355.

29.13. [201] "My Collaboration with Gradimir V. Milovanović"

---

[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** (2011).

My Collaboration with Gradimir V. Milovanović

Walter Gautschi

1. Collaborative efforts, and joint publications resulting therefrom, are much more prevalent in the physical sciences than they are in mathematics. The reason is that research in the physical sciences usually requires team work involving a number of scientists with specialized skills, whereas research in mathematics is a much more individual and solitary enterprise. Nevertheless, even in mathematics, collaboration between different mathematicians may come about through a variety of circumstances. In my own experience, most of my collaboration originated in my attending mathematical conferences, visiting other institutions, or entertaining guests at my own institution. Another not insignificant group of collaborators comes from Ph.D. or postdoctoral students. In all these cases, an important aspect is interpersonal communication and oral exchange of ideas. Not so in the case of Gradimir! Here, collaboration started anonymously, almost ghostlike, during a process of refereeing, exactly 25 years ago. (I may be permitted to divulge information that normally is held confidential!) That is when I received a manuscript from the editor of *Mathematics of Computation* authored by some Gradimir Milovanović, a name I had never heard of before. I was asked to referee it for the journal, whose editor-in-chief I was to become shortly thereafter.

2. The topic of the manuscript looked interesting enough: It was a matter of computing integrals that frequently occur in solid state physics, e.g. the total energy of thermal vibration of a crystal lattice, which is expressible as an integral

$$\int_0^\infty f(t)\frac{t}{e^t-1}\,dt, \tag{1}$$

Walter Gautschi
Department of Computer Sciences, Purdue University, West Lafayette, IN 47907-2066, USA
e-mail: wxg@cs.purdue.edu

where $f(t)$ is related to the phonon density of states, or the crystal lattice heat capacity at constant volume, which is

$$\int_0^\infty g(t)\left(\frac{t}{e^t-1}\right)^2 dt,$$

with $g(t)=e^t f(t)$. Gradimir's idea was to compute these integrals, and similar ones with $t/(e^t-1)$ replaced by $1/(e^t+1)$, by Gaussian quadrature, treating $t/(e^t-1)$, or its square, as a weight function. This is a neat way of dealing with the poles of this function at $\pm 2k\pi i$, $k=0,1,2,\ldots$, which otherwise would adversely interfere with more standard integration techniques.

Another simple, but interesting observation of Gradimir was this: Integrals of the type (1) can be used to sum infinite series,

$$\sum_{k=1}^{\infty} a_k = \int_0^\infty h(t)\frac{t}{e^t-1}dt, \tag{2}$$

if the general term of the series, $a_k=-F^{(k)}$, is the (negative) derivative of the Laplace transform $F(p)=\int_0^\infty e^{-pt}h(t)\,dt$ evaluated at $p=k$ of some known function h. Since series of this kind are typically slowly convergent, the representation (2) offers a useful summation procedure, the sequence of n-point Gaussian quadrature rules, $n=1,2,3,\ldots$, applied to the integral on the right converging rapidly if h is sufficiently smooth.

This is all very nice, but how do we generate Gaussian quadrature rules with such unusual weight functions? Classically, there is an approach via orthogonal polynomials and the moments of the weight function,

$$\mu_k=\int_0^\infty t^k\frac{t}{e^t-1}dt, \quad k=0,1,2,\ldots .$$

In fact, this is the road Gradimir took in his manuscript, noting that the moments are expressible in terms of the Riemann zeta function,

$$\mu_k=(k+1)!\zeta(k+2), \quad k=0,1,2,\ldots .$$

It was at this point where I felt I had to exercise my prerogatives as a referee: I criticized the highly ill-conditioned nature of this approach and proposed more stable alternative methods that I developed just a year or two earlier. In the process, I rewrote a good portion of the manuscript and informed the editor that the manuscript so revised would be an appropriate and interesting contribution to computational mathematics. I suggested, subject to the author's approval, to publish the work as a joint paper. The approval was forthcoming, and that is how our first joint publication [6] came about.

In retrospect, Gradimir's original approach via moments has regained some viability since software has become available inthe last few years that allows generating

the required orthogonal polynomials in variable-precision arithmetic. One such program is the Matlab symbolic Chebyshev algorithm `chebyshev.m` (downloadable from

http://www.cs.purdue.edu/archives/2002/wxg/codes/SOPQ.html),

which generates the required recurrence coefficients directly from the moments. Table 1 in [6], and similarly Tables 2–4 (cf. 4 [Sects. 4–5]), can thus be produced very simply using the following Matlab script:

```
syms mom ab
digits(65); dig=65;
for k=1:80
   mom(k)=vpa(gamma(vpa(k+1))*zeta(vpa(k+1)));
end
ab=schebyshev(dig,40,mom);
ab=vpa(ab,25)
```

True, it takes 65-decimal-digit arithmetic to overcome the severe ill-conditioning and obtain the first 40 recursion coefficients (in the array `ab`) of the orthogonal polynomials to 25 decimal digits. But this is a one-time shot; once these coefficients are available, one can revert to ordinary arithmetic to compute the desired Gaussian quadratures and the integrals in question.

3. In March of 1984, on a visit to Niš, I had the opportunity to finally meet my collaborator in person. He invited me to dinner at his home (my compliments to Dobrila for her culinary art!), after which Gradimir and I retired to his study, where we engaged in a most lively brainstorming session. I was astonished how well he knew earlier work of mine. He must have read my short 1984 paper on spherically symmetric distributions and their approximation by step functions matching as many moments of the distribution as possible. Because he obviously had thought about extending this type of approximation to more general spline approximations. Another idea that surfaced during this discussion was orthogonality on the semicircle and related (complex-valued) orthogonal polynomials. We agreed to pursue these topics further, which provided enough material to keep us busy for several years to come. It so happened that it was the second of these problems that received our attention first, but soon enough we worked on both problems concurrently.

4. Polynomials that are orthogonal on curves Γ in the complex plane have a long history in the case where the underlying inner product is Hermitian, i.e., of the form $(u,v) = \int_\Gamma u(z)\overline{v(z)}\,d\sigma(z)$, $d\sigma$ being a positive measure on Γ; see, e.g., [14, Chaps. 11 and 16]. The case most studied, by far, is the unit circle, $\Gamma = \{e^{i\theta}, 0 \le \theta < 2\pi\}$, which gives rise to Szegö's theory of orthogonal polynomials on the unit circle. The question we asked ourselves is this: what happens if the second factor in the inner product is not conjugated? We decided to begin our study with a prototype inner product, namely, $d\sigma$ the Lebesgue measure, and Γ the upper half of the unit

circle (the whole unit circle being ruled out by Cauchy's theorem). Thus, we began looking at

$$(u,v) = \int_0^\pi u(e^{i\theta})v(e^{i\theta})\,d\theta, \tag{3}$$

postponing for later the study of more general weight functions.

The moment functional associated with (3) is $\mathcal{L}z^k = (1,z^k)$, $k = 0,1,2,\ldots$; it is well known that a sequence of monic polynomials $\{\pi_n\}$ orthogonal with respect to the inner product (3) exists uniquely if the moment sequence $\{\mu_k\}$, $\mu_k = \mathcal{L}z^k$, is quasi-definite, i.e., $\Delta_n \neq 0$ for all $n \geq 1$, where

$$\Delta_n = \det \begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_{n-1} \\ \mu_1 & \mu_2 & \cdots & \mu_n \\ \cdots & \cdots & \cdots & \cdots \\ \mu_{n-1} & \mu_n & \cdots & \mu_{2n-2} \end{bmatrix}. \tag{4}$$

We were able to prove quasi-definiteness by explicit computation of the moments and the determinant in (4).

Since $(zu,v) = (u,zv)$, there must exist a three-term recurrence relation to the polynomials $\{\pi_n\}$; we found it to be of the form

$$\begin{aligned} \pi_{k+1}(z) &= (z - i\alpha_k)\pi_k(z) - \beta_k\pi_{k-1}(z), \quad k = 0,1,2,\ldots, \\ \pi_{-1}(z) &= 0, \quad \pi_0(z) = 1, \end{aligned} \tag{5}$$

where

$$\alpha_0 = \vartheta_0, \quad \alpha_k = \vartheta_k - \vartheta_{k-1}, \ \beta_k = \vartheta_{k-1}^2, \ k = 1,2,\ldots, \tag{6}$$

and

$$\vartheta_k = \frac{2}{2k+1}\left[\frac{\Gamma((k+2)/2)}{\Gamma((k+1)/2)}\right]^2, \quad k \geq 0. \tag{7}$$

As $k \to \infty$, one finds $\alpha_k \to 0$, $\beta_k \to \frac{1}{4}$, familiar from Szegö's class of polynomials orthogonal on the interval $[-1,1]$.

Interestingly, the polynomials π_n are closely connected to Legendre polynomials,

$$\pi_n(z) = \hat{P}_n(z) - i\vartheta_{n-1}\hat{P}_{n-1}(z), \quad n \geq 1, \tag{8}$$

where $\hat{P}_k$ is the monic Legendre polynomial of degree k. This allowed us to derive a linear second-order differential equation for π_n, which, like the differential equation for Legendre polynomials, has regular singular points at 1, -1, and ∞, but unlike Legendre polynomials, an additional singular point on the negative imaginary axis, which depends on n and approaches the origin monotonically as $n \uparrow \infty$.

All zeros of π_n are contained in the half disk $D_+ = \{z \in \mathbb{C}: |z| < 1, \operatorname{Im} z > 0\}$ and located symmetrically with respect to the imaginary axis. They are all simple and can be computed as the eigenvalues of the real, nonsymmetric, tridiagonal matrix

having the first n of the coefficients α_k on the diagonal, the first $n-1$ of the ϑ_k on the upper side diagonal and their negatives on the lower side diagonal.

There is a Gaussian quadrature formula for integrals over the semicircle,

$$\int_0^\pi g(e^{i\theta})\,d\theta = \sum_{\nu=1}^{n} \sigma_\nu g(\zeta_\nu), \quad g \in \mathbb{P}_{2n-1}, \tag{9}$$

where ζ_ν are the zeros of π_n and σ_ν the (complex) Christoffel numbers. The latter can be computed by an adaptation of the well-known Golub/Welsch procedure.

All these results are briefly announced in [7] and fully developed in [8], where one also finds applications of the Gauss formula (9) to numerical differentiation and the evaluation of Cauchy principal value integrals.

Partial results for Gegenbauer weight functions had already been obtained, when new impulses were received through collaboration with Henry J. Landau, cf. [12]. This resulted in a considerable simplification of the existence and uniqueness theory. Indeed, if the inner product is

$$(u,v) = \int_0^\pi u(e^{i\theta})v(e^{i\theta})w(e^{i\theta})\,d\theta, \tag{10}$$

where w is positive on $(-1,1)$ and holomorphic in D_+, then the (monic) polynomials $\{\pi_n\}$ orthogonal with respect to (10) exist uniquely if

$$\operatorname{Re}(1,1) = \operatorname{Re}\int_0^\pi w(e^{i\theta})\,d\theta \neq 0.$$

This is always true for symmetric weight functions,

$$w(-z) = w(z) \quad \text{and} \quad w(0) > 0, \tag{11}$$

for example, the Gegenbauer weight $w(z) = (1-z^2)^{\lambda-1/2}$, $\lambda > -1/2$, and also for the Jacobi weight function $w(z) = (1-z)^\alpha(1+z)^\beta$, $\alpha > -1$, $\beta > -1$.

There are interesting interrelations between the (monic) complex polynomials $\{\pi_n\}$ orthogonal with respect to the inner product (10), the (monic) real polynomials $\{p_n\}$ orthogonal with respect to the inner product $[u,v] = \int_{-1}^{1} u(x)v(x)w(x)\,dx$, and the associated polynomials of the second kind,

$$q_n(z) = \int_{-1}^{1} \frac{p_n(z)-p_n(x)}{z-x}\,w(x)\,dx, \quad n = 0,1,2,\ldots; \quad q_{-1}(z) = -1.$$

Thus, for example (cf. (8)),

$$\pi_n(z) = p_n(z) - i\vartheta_{n-1}p_{n-1}(z), \quad n = 0,1,2,\ldots,$$

where

$$\vartheta_{n-1} = \frac{\mu_0\,p_n(0) + iq_n(0)}{i\mu_0\,p_{n-1}(0) - q_{n-1}(0)}, \quad \mu_0 = (1,1),$$

or, alternatively,

$$\vartheta_n = ia_n + \frac{b_n}{\vartheta_{n-1}}, \quad n = 0,1,2,\ldots; \quad \vartheta_{-1} = \mu_0, \tag{12}$$

where a_k, b_k are the recursion coefficients for the real orthogonal polynomials $\{p_n\}$. For symmetric weight functions (11), one can prove $\mu_0 = \pi w(0) > 0$, so that by (12), since $a_n = 0$ and $b_n > 0$, all ϑ_n are positive.

Moreover, the three-term recurrence relation for the π_n again has the form (5), where now

$$\alpha_0 = \vartheta_0 - ia_0, \quad \alpha_k = \vartheta_k - \vartheta_{k-1} - ia_k \quad (k \geq 1),$$
$$\beta_0 = \mu_0, \quad \beta_k = \vartheta_{k-1}(\vartheta_{k-1} - ia_{k-1}) \quad (k \geq 1).$$

For symmetric weight functions ($a_k = 0$), this reduces to (6), and for Gegenbauer weight functions, one finds

$$\vartheta_0 = \frac{\Gamma(\lambda + 1/2)}{\sqrt{\pi}\,\Gamma(\lambda + 1)}, \quad \vartheta_k = \frac{1}{\lambda + k} \frac{\Gamma((k+2)/2)\Gamma(\lambda + (k+1)/2)}{\Gamma((k+1)/2)\Gamma(\lambda + (k/2))}, \; k \geq 1,$$

generalizing (7).

With regard to the location of the zeros of π_n, we showed for symmetric weight functions that they are contained in D_+, with the possible exception of a single (simple) zero on the positive imaginary axis outside the unit disk. (For a related result, see also [5]). The exception cannot occur for Gegenbauer weights, at least not when $n \geq 2$, and all zeros in this case can be shown to be simple. For Gegenbauer weights, one can also obtain the linear second-order differential equation for π_n, which has properties analogous to those stated above for Legendre weight functions.

5. Our second joint venture deals with a problem of spline approximation on the half line $\mathbb{R}_+ = \{t : t \geq 0\}$. Given a function f on $\mathbb{R}_+$ having finite moments, we want to approximate f by a spline function s of degree $m \geq 0$ that also has finite moments; in fact, we want f and s to have the same successive moments up to an order as high as possible.

Now any spline function of degree m is the sum of a polynomial of degree m and a linear combination of truncated mth powers. If this is to have finite moments on $\mathbb{R}_+$, then the polynomial part must be identically zero, and the spline s therefore is of the form

$$s_{n,m}(t) = \sum_{\nu=1}^{n} a_\nu (\tau_\nu - t)_+^m, \tag{13}$$

where u_+^m are the truncated powers

$$u_+^m = \begin{cases} u^m & \text{if } u \geq 0, \\ 0 & \text{if } u < 0, \end{cases} \qquad m = 0,1,2,\ldots.$$

The coefficients a_ν are real and the "knots" τ_ν mutually distinct and positive, say $0 < \tau_1 < \tau_2 < \cdots < \tau_n$, but otherwise can be freely chosen. Since there are $2n$ unknowns, we can impose $2n$ moment conditions,

$$\int_{\mathbb{R}_+} t^j s_{n,m}(t)\,dt = \mu_j, \quad j = 0,1,2,\ldots,2n-1, \tag{14}$$

where $\mu_j = \int_{\mathbb{R}_+} t^j f(t)\,dt$ are the (given) moments of f. The problem thus amounts to solving the system (14) of $2n$ nonlinear equations in the $2n$ unknowns a_ν, τ_ν, $\nu = 1,2,\ldots,n$.

The problem is reminiscent of Gaussian quadrature and in fact can be solved by constructing a suitable n-point Gaussian quadrature rule [9]. Indeed, if f is such that

(a) $f \in C^{m+1}(\mathbb{R}_+)$

(b) The moments $\mu_j = \int_{\mathbb{R}_+} t^j f(t)\,dt$, $j = 0,1,2,\ldots,2n-1$ exist

(c) $f^{(\mu)}(t) = o(t^{-2n-\mu})$ as $t \to \infty$, $\mu = 0,1,\ldots,m$

then the equations (14) have a unique solution if and only if the measure

$$d\lambda_m(t) = \frac{(-1)^{m+1}}{m!} t^{m+1} f^{(m+1)}(t)\,dt \quad \text{on } \mathbb{R}_+ \tag{15}$$

admits an n-point Gaussian quadrature formula

$$\int_{\mathbb{R}_+} g(t)\,d\lambda_m(t) = \sum_{\nu=1}^{n} \lambda_\nu^G g(t_\nu^G), \quad g \in \mathbb{P}_{2n-1}, \tag{16}$$

satisfying $0 < t_1^G < t_2^G < \cdots < t_n^G$. If so, then the solution to (14) is

$$\tau_\nu = t_\nu^G; \quad a_\nu = \frac{\lambda_\nu^G}{[t_\nu^G]^{m+1}}, \quad \nu = 1,2,\ldots,n. \tag{17}$$

In general, of course, $d\lambda_m$ is not a positive measure, and therefore the existence of the Gauss formula (16) with positive nodes is by no means guaranteed. However, when f is completely monotone, i.e., $(-1)^k f^{(k)}(t) > 0$ on $\mathbb{R}_+$ for $k = 0,1,2,\ldots$, then the measure (15) is obviously positive and under the assumptions (a)–(b) can be shown to have finite moments of orders up to $2n-1$. In this case, the quadrature formula (16) exists uniquely and has distinct positive nodes t_ν^G. Moreover, by (17), the coefficients a_ν are all positive, so that $s_{n,m}$ is also completely monotone, at least in the weak sense that $(-1)^k s_{n,m}^{(k)}(t) \geq 0$ for all $k \geq 0$ a.e. on $\mathbb{R}_+$.

In case the spline approximation $s_{n,m}(t)$ exists, its error $f(t) - s_{n,m}(t)$ at $t = x$ can be expressed in terms of the error of the Gauss formula (16) for a special spline function $g(t) = t^{-(m+1)}(t-x)_+^m$ (cf. [10, Theorem 2.3]).

Similar problems of moment-preserving spline approximation can be considered on a finite interval, say $[0,1]$. In this case, we can add to (13) a polynomial of degree m, which increases the degree of freedom by $m+1$. We may use this increased degree of freedom either to add $m+1$ more moment conditions, or to impose $m+1$

boundary conditions of the form $s_{n,m}^{(k)}(1) = f^{(k)}(1)$, $k = 0, 1, \ldots, m$. The relevant measure then becomes

$$d\lambda_m(t) = \frac{(-1)^{m+1}}{m!} f^{(m+1)}(t)\,dt \quad \text{on } [0,1], \tag{18}$$

and the solution of the two problems can be given (if it exists) in terms of generalized Gauss–Lobatto formulae for the first problem, and generalized Gauss–Radau formulae for the other, both for integration with respect to the measure (18); cf. [1]. The numerical construction of such formulae, however, is rather more complicated, and has been considered by the author only recently in [2, 3]. Gradimir, together with M.A. Kovačević [13], also studied moment-preserving approximation on $\mathbb{R}_+$ by defective splines, which gives rise to Gauss–Turán quadrature rules for the measure (15). Undoubtedly, this led Gradimir to wonder about how to compute these quadratures effectively.

6. Gauss–Turán quadrature fromulae are of Gaussian type, i.e., have maximum algebraic degree of exactness, and involve not only values of the integrand function, but also values of its successive derivatives up to an even order $2s$, all evaluated at a common set of n nodes. Since there are $(2s+1)n$ coefficients (n for each derivative) and n nodes to be determined, the maximum degree of exactness is expected to be $2(s+1)n - 1$, and the formula thus has the form

$$\int_{\mathbb{R}} f(t)\,d\lambda(t) = \sum_{i=0}^{2s} \sum_{\nu=1}^{n} \lambda_{i,\nu} f^{(i)}(\tau_\nu), \quad f \in \mathbb{P}_{2(s+1)n-1}, \tag{19}$$

where $d\lambda$ is a given positive measure. It is known that the nodes τ_ν must be the zeros of the (monic) polynomial $\pi_n = \pi_{n,s}$ of degree n whose $(2s+1)$st power is orthogonal (relative to the measure $d\lambda$) to all polynomials of degree $<n$. In other words, π_n is the nth-degree polynomial orthogonal with respect to the positive measure

$$d\mu(t) = \pi_n^{2s}(t)\,d\lambda(t). \tag{20}$$

We have here a case of implicit orthogonality–also called s-orthogonality–since the polynomial π_n to be determined appears also in the measure of orthogonality. The problem of computing Gauss–Turán quadrature rules (19), considered in [10], thus will in some way come down to a problem of solving a system of nonlinear equations.

Gradimir's idea was to embed π_n in a sequence of $n+1$ polynomials π_0, $\pi_1, \ldots, \pi_n$, namely, the polynomials orthogonal with respect to the measure (20). As such, they must satisfy a three-term recurrence relation

$$\begin{aligned} &\pi_{k+1}(t) = (t - \alpha_k)\pi_k(t) - \beta_k \pi_{k-1}(t), \quad k = 0, 1, \ldots, n-1, \\ &\pi_{-1}(t) = 0, \quad \pi_0(t) = 1. \end{aligned} \tag{21}$$

Although the coefficients $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$; $\beta_0, \beta_1, \ldots, \beta_{n-1}$ are not known, in fact, need to be determined, we know from the general theory of orthogonal polynomials that they are expressible in terms of inner products involving the polynomials $\pi_0, \pi_1, \ldots, \pi_{n-1}$; specifically,

$$\alpha_k = \frac{\int_{\mathbb{R}} t\pi_k^2(t)\,d\mu(t)}{\int_{\mathbb{R}} \pi_k^2(t)\,d\mu(t)}, \quad k = 0, 1, \ldots, n-1;$$

$$\beta_k = \frac{\int_{\mathbb{R}} \pi_k^2(t)\,d\mu(t)}{\int_{\mathbb{R}} \pi_{k-1}^2(t)\,d\mu(t)}, \quad k = 1, 2, \ldots, n-1,$$

and $\beta_0 = \int_{\mathbb{R}} d\mu(t)$ by convention. If we insert here the definition (20) of the measure $d\mu$ and clear all fractions, we arrive at the following system of $2n$ equations:

$$\begin{aligned}
&\beta_0 - \int_{\mathbb{R}} \pi_n^{2s}(t)\,d\lambda(t) = 0,\\
&\int_{\mathbb{R}} (\alpha_k - t)\pi_k^2(t)\pi_n^{2s}(t)\,d\lambda(t) = 0, \quad k = 0, 1, \ldots, n-1;\\
&\int_{\mathbb{R}} (\beta_k \pi_{k-1}^2(t) - \pi_k^2(t))\pi_n^{2s}(t)\,d\lambda(t) = 0, \quad k = 1, 2, \ldots, n-1.
\end{aligned} \tag{22}$$

We note that (22) represents a system of $2n$ nonlinear equations in the $2n$ unknowns α_k, β_k. Indeed, each of the polynomials π_r appearing in (22) can be thought of as a function of $\alpha_0, \alpha_1, \ldots, \alpha_{r-1}$; $\beta_0, \beta_1, \ldots, \beta_{r-1}$ by virtue of the relations in (21). Furthermore, each integrand in (22) is a polynomial of degree $\leq 2(s+1)n - 2$, and therefore all integrals in (22) can be evaluated exactly by an N-point Gaussian quadrature rule relative to the measure $d\lambda$, where $N = (s+1)n$. The same is true for the integrals appearing in the Jacobian matrix of the system (22). We were able, therefore, to apply to (22) the Newton–Kantorovich method, which could be made to converge by a careful choice of the initial approximations $\alpha_k^{(0)}$, $\beta_k^{(0)}$ to the unknowns α_k, β_k. Once the α_k and β_k have been obtained, the zeros τ_ν of π_n, i.e., the nodes in the quadrature rule (19), can be computed by the well-known Golub/Welsch procedure.

All that remains is to compute the coefficients $\lambda_{i,\nu}$ in (19). By inserting in (19) suitably selected polynomials of degree $\leq 2(s+1)n - 1$, the coefficients $\lambda_{i,\nu}$ for each fixed ν, $1 \leq \nu \leq n$, can be found by solving an upper triangular system of $2s+1$ linear algebraic equations (cf. [10, Theorem 3.3]).

7. Conformal maps in fluid mechanics often require Cauchy principal value integrals of the form

$$(I_{[\alpha,\beta]}\phi)(\xi) = \int_\alpha^\beta \phi(\tau) \coth \frac{\tau - \xi}{2}\,d\tau, \quad \alpha < \xi < \beta, \tag{23}$$

which are notoriously difficult to evaluate numerically. Some possible approaches are discussed in [11]. If the interval $[\alpha,\beta]$ is finite, (23) can be transformed to

$$(I_a f)(x) = \frac{1}{a} \int_{-1}^{1} \frac{f(t)}{t-x} w(t)\,dt, \quad -1 < x < 1, \tag{24}$$

where $x = [2\xi - (\alpha+\beta)]/(\beta-\alpha)$, $a = (\beta-\alpha)/4$, and

$$w(t) = \omega(a(t-x)), \quad \omega(u) = u\coth u.$$

The difficulty here is caused by the poles of $w(t)$ at $t = x \pm ki\pi/a, k = 1,2,\ldots$, which can be close to the interval $[-1,1]$ if a is large. Following standard procedure, (24) is first written in the form

$$(I_a f)(x) = \frac{1}{a}\left\{ f(x) \ln \frac{\sinh a(1-x)}{\sinh a(1+x)} + \int_{-1}^{1} g(t) w(t)\,dt \right\}, \tag{25}$$

where $g(t) = [x,t]f$ is the divided difference of f at x and t. The integral in (25) can then be computed either by Gauss quadrature relative to the (nonstandard) weight function w, which gives excellent results but is somewhat expensive, or by the less expensive Gauss–Legendre quadrature, which, however, works well only for a relatively small. An alternative procedure is interpolatory quadrature of (24) on the zeros of the respective orthogonal polynomials, which cirumvents the need of computing a divided difference. Error analyses are provided, either in real variable form, involving derivatives, or in terms of contour integration in the complex plane.

Cauchy principal value integrals (23) with infinite interval $[\alpha,\beta] = [-\infty,\infty]$ can be rendered accessible to the same approaches after suitable truncation of the interval.

8. In looking back on my collaboration with Gradimir, I can only marvel at the spontaneity and originality of his input, which often reduced my own role to one of implementor and organizer. It has been truly a pleasure to work together with Gradimir, and I am sure I am sharing this feeling with the many other individuals who have had the privilege of collaborating with Gradimir. I wish him many more years of good health and continued excellence and success in his research.

References

1. Frontini, M., Gautschi, W., Milovanović, G.V.: Moment-preserving spline approximation on finite intervals. Numer. Math. **50** (5), 503–518 (1987)
2. Gautschi, W.: Generalized Gauss–Radau and Gauss–Lobatto formulae. BIT **44** (4), 711–720 (2004)
3. Gautschi, W.: High-order generalized Gauss–Radau and Gauss–Lobatto formulae for Jacobi and Laguerre weight functions. Numer. Algorithm **51** (2), 143–149 (2009)
4. Gautschi, W.: Variable-precision recurrence coefficients for nonstandard orthogonal polynomials. Numer. Algorithms **52** (3), 409–418 (2010)

5. Gautschi, W., Milovanović, G.V.: On a class of complex polynomials having all zeros in a half disc. Numerical methods and approximation theory (Niš, 1984), 49–53, Univ. Niš, Niš (1984)
6. Gautschi, W., Milovanović, G.V.: Gaussian quadrature involving Einstein and Fermi functions with an application to summation of series. Math. Comp. **44** (169), 177–190 (1985)
7. Gautschi, W., Milovanović, G.V.: Polynomials orthogonal on the semicircle. Rend. Sem. Mat. Univ. e Politec. Torino. Special issue (1985), 179–185
8. Gautschi, W., Milovanović, G.V.: Polynomials orthogonal on the semicircle. J. Approx. Theory **46** (3), 230–250 (1986)
9. Gautschi, W., Milovanović, G.V.: Spline approximations to spherically symmetric distributions. Numer. Math. **49** (1), 111–121 (1986)
10. Gautschi, W., Milovanović, G.V.: s-orthogonality and construction of Gauss–Turán-type quadrature formulae. J. Comput. Appl. Math. **86** (1), 205–218 (1997)
11. Gautschi, W., Kovačević, M.A., Milovanović, G.V.: The numerical evaluation of singular integrals with coth-kernel. BIT **27** (3), 389–402 (1987)
12. Gautschi, W., Landau, H.J., Milovanović, G.V.: Polynomials orthogonal on the semicircle, II. Constr. Approx. **3** (4), 389–404 (1987)
13. Milovanović, G.V., Kovačević, M.A.: Moment-preserving spline approximation and Turán quadratures. In: Numerical Mathematics, Singapore 1988, 357–365, Internat. Schriftenreihe Numer. Math., vol. 86, Birkhäuser, Basel (1988)
14. Szegö, G.: Orthogonal Polynomials. 4th ed., American Mathematical Society, Colloquium Publications, vol. 23, Amer. Math. Soc., Providence, RI (1975)

30

Papers on Miscellanea

---

71 Families of algebraic test equations, Calcolo 16, 383–398 (1979)

96 (with B. N. 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 (1986)

124 A class of slowly convergent series and their summation by Gaussian quadrature, Math. Comp. 57, 309–324 (1991)

125 On certain slowly convergent series occurring in plate contact problems, Math. Comp. 57, 325–338 (1991)

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, 1997. [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.]

175 The Hardy–Littlewood function: an exercise in slowly convergent series, J. Comput. Appl. Math. 179, 249–254 (2005)

197 The spiral of Theodorus, numerical analysis, and special functions, J. Comput. Appl. Math. 235, 1042–1052 (2010)

30.1. [71] "FAMILIES OF ALGEBRAIC TEST EQUATIONS"

---

[71] "Families of Algebraic Test Equations," *Calcolo* **16**, 383–398 (1979).

FAMILIES OF ALGEBRAIC TEST EQUATIONS

W. GAUTSCHI (1)

1. Introduction.

Test matrices have been in use for some time to scrutinize computer algorithms for solving linear algebraic systems and eigenvalue problems; see, e. g., Gregory and Karney [1969]. For the problem of finding roots of algebraic equations, the construction of appropriate test equations has been given less attention. Here we wish to propose two families of algebraic test equations, the first consisting of equations with predominantly complex roots, the second of equations with only real roots.

To be useful for testing purposes, a test equation (of some fixed degree) should have the following characteristics:

(1) All roots of the equation can be calculated directly (i. e., without recourse to a rootfinding algorithm).

(2) The equation contains a parameter (or parameters) which can be used to control the numerical condition of the roots. By varying the parameter (s), the condition number of the worst-conditioned root can be made to range from relatively small to arbitrarily large values.

(3) All coefficients of the equation are integer-valued.

It may not be easy, in practice, to achieve all these characteristics, particularly the last one, if we are interested in relatively large degrees. Even when this is possible, the integer coefficients may become so large as to make exact representation in floating-point arithmetic impossible. Although equations which do not satisfy (3) are less desirable, they are still useful for testing purposes,

---

— Received June 6, 1979.

(1) Department of Computer Sciences, Purdue University, Lafayette, Indiana 47907, U.S.A..

provided one takes properly into account the influence of rounding errors in the coefficients upon the results.

Isolated examples of test equations, particularly ill-conditioned ones, have been known for a long time. Perhaps the best-known example, due to Wilkinson [1963], is the equation with roots at $1, 2, \ldots, n$. Some of these roots are relatively well-conditioned, while others are quite ill-conditioned, more so the larger the degree n. (The numerical condition of Wilkinson's equation is analyzed in Gautschi [1973]; see also Gautschi [1978, § 4]). The roots of unity lead to another interesting example if one removes half of them and retains only those on the half-circle (Jenkins and Traub [1975]). Our first family of test equations, indeed, is a simple extension of this latter example.

2. A first family of test equations and their numerical condition.

Given an algebraic equation of degree n,

$$p(z)=0, \qquad p(z)=z^n+a_1 z^{n-1}+\ldots+a_{n-1} z+a_n, \tag{2.1}$$

with complex coefficients a_λ, and a simple root ζ of (2.1), we adopt as condition number for ζ the quantity (Wilkinson [1963, p. 38 ff], Gautschi [1973])

$$\operatorname{cond} \zeta=\frac{\sum_{\lambda=1}^{n} |a_\lambda|}{|\zeta|\,|p'(\zeta)|}. \tag{2.2}$$

It measures the sensitivity of ζ, in terms of relative errors, to small (relative) perturbations in all nonzero coefficients a_λ.

Our first family of test equations (with parameter α) is

$$\begin{gathered} p_\alpha(z)=0, \qquad p_\alpha(z)=\prod_{\lambda=1}^{n}[z-\zeta_\lambda(\alpha)]=z^n-\sigma_1 z^{n-1}+\ldots+(-1)^n \sigma_n, \\ \zeta_\lambda(\alpha)=e^{(\lambda-1)i\alpha}, \qquad 0<\alpha\leq\frac{2\pi}{n}. \end{gathered} \tag{2.3}$$

If $\alpha=2\pi/n$, the ζ_λ are the *n-th* roots of unity, thus $p_\alpha(z)=z^n-1$, and we are in the case of a well-conditioned equation, all roots having condition $1/n$. For $\alpha=\pi/(n-1)$, we get the roots of unity on a half-circle, which are all relatively ill-conditioned. As $\alpha \downarrow 0$, the condition deteriorates unboundedly.

From Eq. (2.2) we find for the condition of the roots in (2.3),

$$\text{(2.4)} \qquad \operatorname{cond} \zeta_\nu = \frac{\sum_{\lambda=1}^{n} |\sigma_\lambda|}{2^{n-1} \prod_{\substack{\lambda=1 \\ \lambda \neq \nu}}^{n} \left| \sin (\nu-\lambda) \frac{\alpha}{2} \right|}, \quad \nu = 1, 2, \ldots, n.$$

In view of the symmetry property $\operatorname{cond} \zeta_\nu = \operatorname{cond} \zeta_{n+1-\nu}$, $\nu = 1, 2, \ldots, n$, it suffices to consider the condition numbers for $\nu = 1, 2, \ldots, n'$, where

$$n' = \left\lfloor \frac{n+1}{2} \right\rfloor.$$

We show that

$$\text{(2.5)} \qquad \operatorname{cond} \zeta_1 < \operatorname{cond} \zeta_2 < \ldots < \operatorname{cond} \zeta_{n'} \text{ for } 0 < \alpha < \frac{2\pi}{n}.$$

If $n=2$, there is nothing to prove. If $n>2$, we let π_ν denote the product in the denominator of (2.4) and observe that

$$\text{(2.6)} \qquad \frac{\pi_\nu}{\pi_{\nu-1}} = \left| \frac{\sin (\nu-1) \frac{\alpha}{2}}{\sin (n-\nu+1) \frac{\alpha}{2}} \right|, \qquad \nu = 2, 3, \ldots, n'.$$

Our assumption on α implies that both sine functions in the numerator and denominator are evaluated in the open interval $(0, \pi)$, the former in fact in $(0, \pi/2)$. The absolute value signs in (2.6) can thus be deleted, and an elementary calculation shows that all ratios in (2.6) are less than 1 precisely if

$$\text{(2.7)} \qquad \tan \left(n \frac{\alpha}{4} \right) > \tan (\nu-1) \frac{\alpha}{2}, \quad \nu = 2, 3, \ldots, n'.$$

Since the tangent is monotone increasing on $[0, \pi/2)$, the inequality (2.7) for $\nu = n'$ implies all others, and indeed holds true by virtue of

$$\tan (n'-1) \frac{\alpha}{2} = \begin{cases} \tan \left(n \frac{\alpha}{4} - \frac{\alpha}{2} \right) & \text{if } n \text{ is even,} \\ \tan \left(n \frac{\alpha}{4} - \frac{\alpha}{4} \right) & \text{if } n \text{ is odd.} \end{cases}$$

It follows that $\pi_1 > \pi_2 > \ldots > \pi_{n'}$, hence (2.5).

We recall that the condition number in (2.2) is invariant with respect to scaling of the independent variable. Any transformation $z=\omega z^*$, where $\omega \neq 0$ is arbitrary complex, leaves the condition of the roots unchanged. As a consequence, the assumption $|\zeta_\nu| = 1$ made in (2.3) does not restrict generality, and we are free to subject the roots to a rigid rotation on the unit circle. Taking advantage of this last remark, we may bring the equation $p_\alpha(z)=0$, which generally has complex coefficients, into a form with real coefficients. It suffices to rotate the roots into a position symmetric with respect to the real axis, i. e., to put

$$z=e^{i(n-1)\alpha/2}\, z^*.$$

Then

$$p_\alpha(e^{i(n-1)\alpha/2}\, z^*)=e^{in(n-1)\alpha/2}\, p_\alpha^*(z^*), \tag{2.8}$$

where

$$p_\alpha^*(u)=\prod_{\lambda=1}^{n}[u-\zeta_\lambda^*(\alpha)], \quad \zeta_\lambda^*(\alpha)=e^{i\left(\lambda-1-\frac{n-1}{2}\right)\alpha},$$

and $\zeta_\lambda^*=\overline{\zeta^*_{n+1-\lambda}}$, $\lambda=1, 2, \dots, n$, implying that all coefficients of p_α^* indeed are real.

3. Computation of the coefficients of p_α^* and the condition numbers cond ζ_ν.

In analogy to (2.3), we write

$$p_\alpha^*(u)=\prod_{\lambda=1}^{n}[u-\zeta_\lambda^*(\alpha)]=\sum_{\lambda=0}^{n}(-1)^\lambda\,\sigma_\lambda^*\,u^{n-\lambda}, \qquad \sigma_0^*=1, \tag{3.1}$$

where all σ_λ^* are real. Observing from (2.8) that

$$p_\alpha(z)=e^{in(n-1)\alpha/2}\, p_\alpha^*(e^{-i(n-1)\alpha/2}\, z),$$

and using (3.1), we find

$$p_\alpha(z)=\sum_{\lambda=0}^{n}(-1)^\lambda\, e^{i(n-1)\lambda\alpha/2}\,\sigma_\lambda^*\, z^{n-\lambda}.$$

Hence, by comparison with (2.3),

$$\sigma_\lambda=e^{i(n-1)\lambda\alpha/2}\,\sigma_\lambda^*, \qquad \lambda=0, 1, 2, \dots, n. \tag{3.2}$$

For the following, it will be necessary to introduce the more precise notation $\sigma_\lambda = \sigma_{\lambda,n}$, indeed, to define $\sigma_{\lambda,\mu}$ by

$$\prod_{\lambda=1}^{\mu} [z-\zeta_\lambda(\alpha)] = \sum_{\lambda=0}^{\mu} (-1)^\lambda \sigma_{\lambda,\mu} z^{\mu-\lambda}, \qquad \mu=1,2,3,\dots, \tag{3.3}$$

and similarly $\sigma^*_{\lambda,\mu}$ by

$$\prod_{\lambda=1}^{\mu} [u-\zeta^*_{\lambda,\mu}(\alpha)] = \sum_{\lambda=0}^{\mu} (-1)^\lambda \sigma^*_{\lambda,\mu} u^{\mu-\lambda}, \quad \mu=1,2,3,\dots, \tag{3.4}$$

where

$$\zeta^*_{\lambda,\mu}(\alpha) = e^{i\left(\lambda-1-\frac{\mu-1}{2}\right)\alpha}, \quad \lambda=1,2,\dots,\mu.$$

Then, as in (3.2),

$$\sigma_{\lambda,\mu} = e^{i(\mu-1)\lambda\alpha/2}\,\sigma^*_{\lambda,\mu}. \tag{3.5}$$

Defining

$$\sigma_{0,\mu}=1, \quad \sigma_{\mu+1,\mu}=0 \quad \text{for} \quad \mu=0,1,2,\dots, \tag{3.6}$$

one obtains from (3.3) the well-known recursion

$$\sigma_{\lambda,\mu+1} = \sigma_{\lambda,\mu} + \zeta_{\mu+1}\,\sigma_{\lambda-1,\mu}, \quad \lambda=1,2,\dots,\mu+1, \tag{3.7}$$

where $\zeta_{\mu+1}=e^{i\mu\alpha}$. Substituting (3.5) into (3.6), (3.7), and observing that all $\sigma^*_{\lambda,\mu}$ are real, one finds

$$\left.\begin{aligned} &\sigma^*_{0,\mu}=1, \quad \sigma^*_{\mu+1,\mu}=0, \\ &\sigma^*_{\lambda,\mu+1} = \cos(\lambda\alpha/2)\,\sigma^*_{\lambda,\mu} + \cos((\mu-\lambda+1)\alpha/2)\,\sigma^*_{\lambda-1,\mu} \\ &\qquad\qquad \lambda=1,2,\dots,\mu+1 \end{aligned}\right\} \quad \mu=0,1,2,\dots.$$

A simple induction argument will show that

$$\sigma^*_{\lambda,\mu} = \sigma^*_{\mu-\lambda,\mu}, \quad \lambda=0,1,2,\dots,\mu.$$

It suffices, therefore, to compute

$$\sigma^*_{0,\mu}=1, \quad \mu=0,1,\dots,n-1,$$

$$\left.\begin{aligned} &\sigma^*_{\lambda,\mu+1} = \cos(\lambda\alpha/2)\,\sigma^*_{\lambda,\mu} + \cos((\mu-\lambda+1)\,\alpha/2)\,\sigma^*_{\lambda-1,\mu} \\ &\qquad \lambda = 1, 2, \ldots, \left\lfloor \frac{\mu+1}{2} \right\rfloor, \\ &\sigma^*_{\frac{\mu}{2}+1,\mu+1} = \sigma^*_{\frac{\mu}{2},\mu+1} \quad (\text{if } \mu \text{ is even}) \end{aligned}\right\} \quad \mu = 0,1,\ldots,n-1. \tag{3.8}$$

The numerator in the condition number (2.4), by (3.2), can now be obtained from

$$\sum_{\lambda=1}^{n} |\sigma_\lambda| = \sum_{\lambda=1}^{n} |\sigma^*_{\lambda,n}| = \begin{cases} 1 + |\sigma^*_{\frac{n}{2},n}| + 2\sum_{\lambda=1}^{\frac{n}{2}-1} |\sigma^*_{\lambda,n}|, & n \text{ even}, \\ 1 + 2\sum_{\lambda=1}^{\frac{n-1}{2}} |\sigma^*_{\lambda,n}|, & n \text{ odd}. \end{cases}$$

In Figure 1 we show the largest condition number, cond $\zeta_{n'}$, as function of $\alpha = 2\pi x/n$, $0 \le x \le 1$, for $n = 10, 20, 40$.

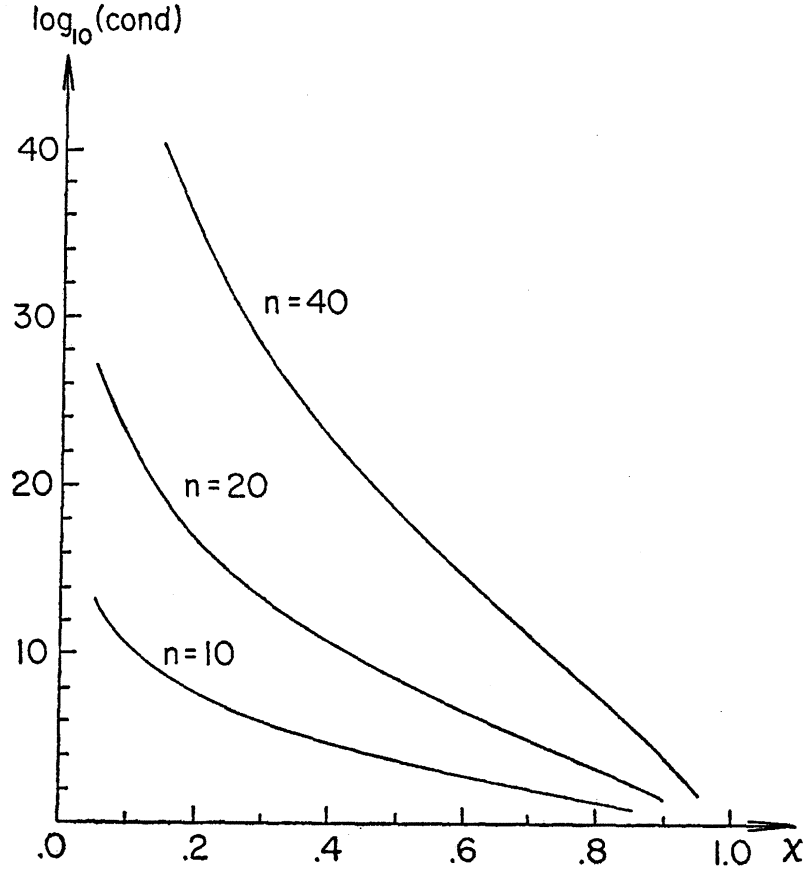


Fig. 1. Condition number of the worst-conditioned root of $p_\alpha(z) = 0$, $\alpha = 2\pi x/n$, for $0 \le x \le 1$, $n = 10, 20, 40$.

4. Equations with integer coefficients.

The polynomials p_α^*, as constructed in Section 3, will not have integer coefficients, in general. It is desirable to find a subset of parameter values α, and a suitable constant multiple p_α^{**} of p_α^* such that all coefficients $\sigma_{\lambda,n}^{**}$ of p_α^{**} become integer-valued when α is restricted to the subset in question. We can achieve this by letting

$$\cos\frac{\alpha}{2}=1-\frac{1}{p},\quad p>0. \tag{4.1}$$

Then all cosine factors in (3.8), hence all $\sigma^*_{\lambda,\mu}$, become polynomials in p^{-1} with integer coefficients, since $\cos\left(m\frac{\alpha}{2}\right)=T_m\left(1-\frac{1}{p}\right)$, $m=1,2,3,\ldots$, where T_m is the Chebyshev polynomial of the first kind. Multiplying through by an appropriate power of p we obtain a polynomial p_α^{**} whose coefficients $\sigma_{\lambda,n}^{**}$ are polynomials in p with integer coefficients, hence integer-valued if p is an integer. Selecting p sufficiently large we can get α arbitrarily small, and thus produce equations which are arbitrarily ill-conditioned. The only serious limitation of this procedure is due to the finiteness of the machine arithmetic in which the integer p and, above all, the resulting integer coefficients, are to be represented. A lower limit on p is imposed by the condition $\alpha\leq 2\pi/n$, which translates into

$$p\geq\frac{1}{2\sin^2\frac{\pi}{2n}}. \tag{4.2}$$

In Table 1 the results are summarized for $2\leq n\leq 8$. Only the coefficients $\sigma_{\lambda,n}^{**}$, $\lambda\leq[n/2]$, are listed, as the others can be obtained by symmetry, $\sigma_{\lambda,n}^{**}=\sigma^{**}_{n-\lambda,n}$.

n	$\sigma_{0,n}^{**}$	$\sigma_{1,n}^{**}$	$\sigma_{2,n}^{**}$	$\sigma_{3,n}^{**}$	$\sigma_{4,n}^{**}$
2	p	$s_2(p)$			
3	p^2	$s_3(p)$			
4	p^4	$2ps_2(p)\,s_4(p)$	$2s_3(p)\,s_4(p)$		
5	p^6	$p^2 s_5(p)$	$2s_4(p)\,s_5(p)$		
6	p^9	$p^4 s_2(p)\,s_6(p)$	$ps_5(p)\,s_6(p)$	$2s_2(p)\,s_4(p)\,s_5(p)\,\bar{s}_6(p)$	
7	p^{12}	$p^5 s_7(p)$	$p^2 s_6(p)\,s_7(p)$	$s_5(p)\,s_7(p)\,\bar{s}_6(p)$	
8	p^{16}	$4p^9 s_2(p)\,s_8(p)$	$4p^4 s_7(p)\,s_8(p)$	$4ps_2(p)\,s_7(p)\,s_8(p)\,\bar{s}_6(p)$	$2s_5(p)\,s_7(p)\,\bar{s}_6(p)\,\bar{s}_8(p)$

TABLE 1. The coefficients of the test polynomial $p_\alpha^{**}(u)=\sum_{\lambda=0}^{n}(-1)^\lambda\sigma_{\lambda,n}^{**}u^{n-\lambda}$.

The notations used in Table 1 are as follows:

$$s_2(p)=2(p-1)$$

$$s_3(p)=(p-2)(3p-2)$$

$$s_4(p)=p^2-4p+2$$

$$s_5(p)=5p^4-40p^3+84p^2-64p+16$$

$$s_6(p)=3p^4-32p^3+80p^2-64p+16$$

$$s_7(p)=7p^6-112p^5+504p^4-960p^3+880p^2-384p+64$$

$$s_8(p)=p^6-20p^5+106p^4-224p^3+216p^2-96p+16$$

$$\bar{s}_6(p)=p^2-8p+4$$

$$\bar{s}_8(p)=p^4-16p^3+40p^2-32p+8.$$

Inequality (4.2) imposes the constraints

$$p\geq 1,\ p\geq 2,\ p\geq 4,\ p\geq 6,\ p\geq 8,\ p\geq 11,\ p\geq 14 \tag{4.3}$$

for $n=2, 3, \ldots, 8$, respectively.

Thus. for example, the family of test equations of degree 5, according to Table 1 ($n=5$), is

$$p^6 z^5-p^2 s_5(p) z^4+2s_4(p) s_5(p) z^3-2s_4(p) s_5(p) z^2+p^2 s_5(p) z-p^6=0, \quad p\geq 6.$$

5. Numerical tests.

Test equations can be put to many uses. One of the more important ones is the comparative analysis of different algorithms (and their computer implementations) for solving equations. Since we can vary the numerical condition of the roots, for each fixed degree, we are able to observe the performance of these algorithms on equations of fixed degree under increasing pressures of ill-conditioning. Our main interest, here, being in the accuracy of the algorithms, it is interesting to observe how closely the relative errors in the results are going to conform with what one ideally expects, namely that all these errors are of

the order of the machine precision multiplied by the respective condition numbers.

If $\zeta \neq 0$ denotes an exact root of the equation, $\tilde{\zeta}$ the approximation to ζ returned by the computer algorithm, and ε denotes the machine precision (i. e., $\varepsilon = 2^{-t}$, where t is the number of binary digits in the mantissa of the floating-point word), then for a good algorithm the quantity

$$\mu(\zeta) = \frac{|\tilde{\zeta} - \zeta|}{\varepsilon |\zeta| \operatorname{cond} \zeta} \tag{5.1}$$

ought to be of the order of magnitude 1, or even smaller (considering that cond ζ is a somewhat conservative measure of the condition of ζ). The larger $\mu(\zeta)$, the poorer the performance of the algorithm (with regard to accuracy). If we are interested in the performance of the algorithm on an equation, rather than an isolated root, we can measure it by the average

$$\mu_{av} = \frac{1}{n} \sum_{k=1}^{n} \mu(\zeta_k), \tag{5.2}$$

where n is the degree of the equation and ζ_k its n (simple) roots.

By way of illustration, we compare two subroutines, ZRPOLY and ZPOLR, cut of the IMSL library (International Mathematical Statistical Libraries, Version 7). The first implements the three-stage algorithm of Jenkins and Traub [1970], while the other is based on Laguerre's method, as implemented by Smith [1967]. We compiled both subroutines on the CDC 6500 computer, using the FTN compiler (CDC Fortran Extended Compiler), and ran them on the test equations $p_\alpha^* = 0$ of degrees $n = 5, 10, 20, 40$, with $\alpha = 2\pi x/n$, $x = .2, .4, \ldots, 1.0$. The coefficients $\sigma^*_{\lambda,n}$ were generated by (3.8) in double precision, and then rounded to single precision. Likewise, we used double precision to compute reference values for the roots $\zeta_\lambda^*(\alpha)$. The results are summarized in Table 2 (2). The last two columns exhibit, for ZRPOLY and ZPOLR, respectively, the values observed for the average measure μ_{av} in (5.2). The two preceding columns contain the smallest and the largest condition number associated with the n roots in question. It is seen that the performance of both algorithms is quite good, in general, even on ill-conditioned equations. ZPOLR yields consistently smaller values of μ_{av} (with one exception, for $n = 5, x = .6$). ZRPOLY seems to have some difficulties maintaining accuracy when x is near 1.0 and n is large (i. e., for relatively well-conditioned equations!).

---

(2) Numbers in parentheses indicate decimal exponents.

TABLE 2. Performance of ZRPOLY and ZPOLR on the test equation $p_\alpha^*=0$, $\alpha=2\pi x/n$, $x=.2\,(.2)\,1.$, $n=5, 10, 20, 40$.

n	x	min cond	max cond	μ_{av} ZRPOLY	μ_{av} ZPOLR
5	.2	.323 (3)	.184 (4)	.252	.748 (—1)
	.4	.199 (2)	.960 (2)	.185	.201 (—1)
	.6	.376 (1)	.135 (2)	.182	.236
	.8	.103 (1)	.229 (1)	.750 (2)	.194
	1.0	.200	.200	.160 (2)	.000
10	.2	.370 (6)	.408 (8)	.399	.151
	.4	.774 (3)	.564 (5)	.346	.541 (—1)
	.6	.222 (2)	.761 (3)	.298	.165
	.8	.178 (1)	.176 (2)	.845 (2)	.288
	1.0	.100	.100	.831 (2)	.200 (1)
20	.2	.637 (12)	.436 (17)	.295	.377 (—1)
	.4	.153 (7)	.412 (11)	.205	.104
	.6	.100 (4)	.489 (7)	.281	.136
	.8	.668 (1)	.188 (4)	.687 (1)	.115
	1.0	.500 (—1)	.500 (—1)	.409 (5)	.100 (1)
40	.2	.265 (25)	.973 (35)	.183 (—11)	.161 (—11)
	.4	.846 (13)	.431 (23)	.432	.688 (—2)
	.6	.291 (7)	.399 (15)	.225 (1)	.707 (—1)
	.8	.136 (3)	.426 (8)	.200 (3)	.804 (—1)
	1.0	.250 (—1)	.250 (—1)	.534 (10)	.200 (1)

Interestingly, both routines do not fall to pieces when confronted with an extremely ill-conditioned equation, such as the one for $n=40$, $x=.2$. All roots are obtained with relative errors ranging between 30 and 120%. This accounts for the very small values of μ_{av} shown in this case.

Similar results ([3]) are observed on equations with integer coefficients (those of Table 1), where $n=2, 3, \ldots, 8$, and where p is varied between the smallest admissible integer (cf. (4.3)) and the largest possible integer subject to exact representation of the coefficients $\sigma_{\lambda,n}$** in CDC 6500 floating-point arithmetic.

([3]) One of our tests runs (for $n=3$, $p=8\times10^7$) uncovered an implementation error in one of the auxiliary routines called by ZRPOLY.

6. A second family of test equations.

We now briefly describe our second family of test equations, viz.,

$$q_\alpha(z)=0, \quad q_\alpha(z)=\prod_{\lambda=1}^{n}[z-\xi_\lambda(\alpha)]=z^n-\tau_1 z^{n-1}+\ldots+(-1)^n\tau_n,$$
$$\xi_\lambda(\alpha)=\alpha^{-\lambda}, \quad 1<\alpha<\infty. \tag{6.1}$$

The equations (6.1) are well-conditioned, when α is large, and become progressively more ill-conditioned as α approaches 1. The condition number of ξ_λ, using Gautschi [1973, Thm. 3.1], is easily found to be

$$\operatorname{cond}\xi_\nu=\frac{2\pi^+_{\nu-1}\pi^+_{n-\nu}-\alpha^{-\nu(\nu-1)/2}}{\pi^-_{\nu-1}\pi^-_{n-\nu}}, \quad \nu=1,2,\ldots,n, \tag{6.2}$$

where

$$\pi_0^\pm=1, \quad \pi_\mu^\pm=\prod_{\lambda=1}^{\mu}(1\pm\alpha^{-\lambda}), \quad \mu=1,2,\ldots,n.$$

It follows that

$$\operatorname{cond}\xi_1\to 1, \ \operatorname{cond}\xi_\nu\to 2\ (\nu>1), \ \text{as} \ \alpha\to\infty,$$

while

$$\operatorname{cond}\xi_\nu\to\infty \ \text{as} \ \alpha\downarrow 1.$$

Furthermore,

$$\operatorname{cond}\xi_\nu<2\frac{\pi^+_{\nu-1}\pi^+_{n-\nu}}{\pi^-_{\nu-1}\pi^-_{n-\nu}}, \quad \nu=1,2,\ldots,n.$$

The bounds on the right are symmetric, i. e., invariant under the substitution $\nu\to n+1-\nu$, and increase monotonically for $\nu<n/2$, attaining their maximum value at $\nu=[n/2]+1$. The true values of the condition number behave similarly. We have, in particular,

$$\max_{1\le\nu\le n}\operatorname{cond}\xi_\nu<\prod_{\lambda=1}^{[n/2]}\left(\frac{1+\alpha^{-\lambda}}{1-\alpha^{-\lambda}}\right)^2. \tag{6.3}$$

The largest condition number is shown in Figure 2 as a function of α, for $n=10, 20, 40$.

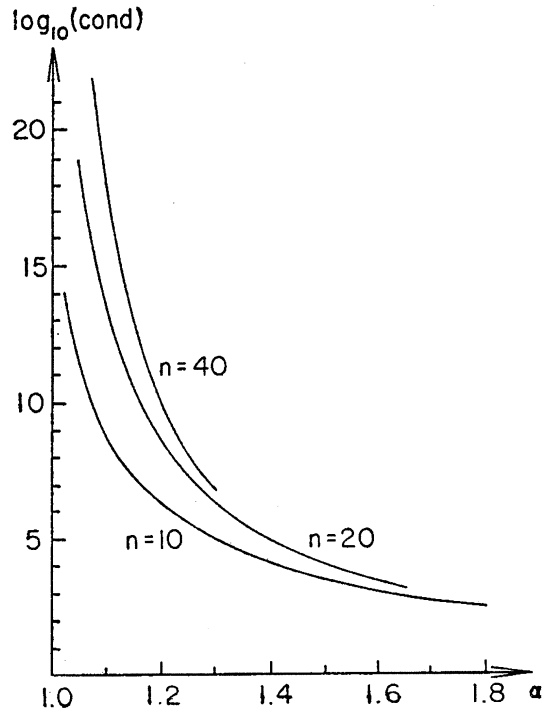


Fig. 2. Condition number of the worst-conditioned root of $q_\alpha(z)=0$, for $\alpha>1$, $n=10, 20, 40$.

The coefficients $\tau_\lambda=\tau_{\lambda,n}$ in (6.1) are easily generated by recursion, as in (3.6), (3.7),

$$\tau_{0,\mu}=1,\ \tau_{\mu+1,\mu}=0,\quad \mu=0,1,2,\dots,$$

(6.4)

$$\tau_{\lambda,\mu+1}=\tau_{\lambda,\mu}+\xi_{\mu+1}\,\tau_{\lambda-1,\mu},\quad \lambda=1,2,\dots,\mu+1,\quad \mu=0,1,2,\dots,n-1.$$

If we let

$$\alpha=1+\frac{1}{p},\quad p>0,$$

and multiply through by appropriate powers of $p+1$, the new coefficients $\tau^*_{\lambda,n}$ become polynomials in p with integer coefficients, hence integer-valued, if p is an integer. More specifically, they assume the form

$$\tau^*_{\lambda,n}=p^{\frac{\lambda(\lambda+1)}{2}}(p+1)^{\frac{(n-\lambda)(n+1-\lambda)}{2}}\tau_{\lambda,n}^{**}(p),\quad \lambda=0,1,2,\dots,n, \tag{6.5}$$

where $\tau_{\lambda,n}^{**}$ are polynomials satisfying

$$\tau_{0,\lambda}^{**}(p)=1,$$

$$\tau_{\lambda,n}^{**}(p)=\tau^{**}_{n-\lambda,n}(p),\quad \lambda=0,1,2,\dots,n.$$

The polynomials $\tau_{\lambda,n}^{**}(p)$, $\lambda=1,2,\dots,[n/2]$, for $2\leq n\leq 8$, are shown in Table 3.

TABLE 3. The coefficients of the test polynomial

$$q_\alpha^*(z)=\sum_{\lambda=0}^{n}(-1)^\lambda p^{\lambda(\lambda+1)/2}(p+1)^{(n-\lambda)(n+1-\lambda)/2}\tau_{\lambda,n}^{**}(p)\,z^{n-\lambda}$$

n	$\tau_{1,n}^{**}$	$\tau_{2,n}^{**}$	$\tau_{3,n}^{**}$	$\tau_{4,n}^{**}$
2	$t_2(p)$			
3	$t_3(p)$			
4	$t_2(p)\,t_4(p)$	$t_3(p)\,t_4(p)$		
5	$t_5(p)$	$t_4(p)\,t_5(p)$		
6	$t_2(p)\,t_3(p)\,t_6(p)$	$t_3(p)\,t_5(p)\,t_6(p)$	$t_2(p)\,t_4(p)\,t_5(p)\,t_6(p)$	
7	$t_7(p)$	$t_3(p)\,t_6(p)\,t_7(p)$	$t_5(p)\,t_6(p)\,t_7(p)$	
8	$t_2(p)\,t_4(p)\,t_8(p)$	$t_4(p)\,t_7(p)\,t_8(p)$	$t_2(p)\,t_4(p)\,t_6(p)\,t_7(p)\,t_8(p)$	$t_5(p)\,t_6(p)\,t_7(p)\,t_8(p)$

The notations used in Table 3 are as follows:

$$t_2(p)=2p+1$$

$$t_3(p)=3p^2+3p+1$$

$$t_4(p)=2p^2+2p+1$$

$$t_5(p)=5p^4+10p^3+10p^2+5p+1$$

$$t_6(p)=p^2+p+1$$

$$t_7(p)=7p^6+21p^5+35p^4+35p^3+21p^2+7p+1$$

$$t_8(p)=2p^4+4p^3+6p^2+4p+1.$$

Thus, for example, the family of test equations of degree 5, according to Table 3 ($n=5$), is

$$(p+1)^{15} z^5 - p\,(p+1)^{10}\, t_5(p)\, z^4 + p^3\,(p+1)^6\, t_4(p)\, t_5(p)\, z^3 -$$

$$- p^6\,(p+1)^3\, t_4(p)\, t_5(p)\, z^2 + p^{10}\,(p+1)\, t_5(p)\, z - p^{15} = 0, \quad p \geq 1.$$

7. Numerical tests.

The same two subroutines, as in Section 5, were tested on the equations $q_\alpha^* = 0$ of degrees $n=5, 10, 20, 40$, for selected values of α between 1 and 5. The results are reported in Table 4, in an outlay similar to the one in Table 2. On the whole, the subroutines perform similarly as on the first test equations, ZPOLR generally yielding smaller values of μ_{av}, but the differences are not as pronounced as before. In the case $n=40$, $\alpha=5$, a series of low-order coefficients underflows, making a meaningful test impossible.

TABLE 4. Performance of ZRPOLY and ZPOLR on the test equation

$q_\alpha^* = 0$, $\alpha = 1.05$ (var.) 5.0, $n = 5, 10, 20, 40$.

n	α	min cond	max cond	μ_{av} ZRPOLY	μ_{av} ZPOLR
5	1.05	.227 (6)	.137 (7)	.256 (2)	.287
	1.1	.157 (5)	.947 (5)	.977 (—1)	.564 (—1)
	1.25	.558 (3)	.326 (4)	.598	.427
	1.5	.619 (2)	.329 (3)	.212	.858 (—1)
	2.0	.113 (2)	.491 (2)	.197	.466 (—1)
	5.0	.200 (1)	.527 (1)	.519	.285
10	1.05	.190 (10)	.231 (12)	.381 (1)	.450 (—1)
	1.1	.533 (7)	.584 (9)	.148 (1)	.142
	1.25	.560 (4)	.363 (6)	.367	.191
	1.5	.129 (3)	.333 (4)	.312	.294
	2.0	.130 (2)	.113 (3)	.328	.000
	5.0	.200 (1)	.549 (1)	.458	.190
20	1.05	.116 (15)	.759 (19)	.231 (—2)	.409 (—2)
	1.1	.115 (10)	.324 (14)	.225	.662 (—1)
	1.25	.148 (5)	.206 (8)	.332	.152
	1.5	.143 (3)	.105 (5)	.279	.182
	2.0	.130 (2)	.136 (3)	.250	.101
	5.0	.200 (1)	.550 (1)	.850	.319
40	1.05	.289 (19)	.148 (29)	.133 (—5)	.758 (—6)
	1.1	.187 (11)	.889 (18)	.269	.149 (—1)
	1.25	.166 (5)	.116 (9)	.235	.129
	1.5	.143 (3)	.124 (5)	.369	.199
	2.0	.130 (2)	.136 (3)	.354	.155
	5.0	.200 (1)	.550 (1)	—	—

REFERENCES

W. GAUTSCHI: *On the condition of algebraic equations,* Numer. Math. **21**, (1973), 405-424.

W. GAUTSCHI: *Questions of numerical condition related to polynomials,* in: « Recent Advances in Numerical Analysis » (1978), C. de Boor and G. H. Golub, eds., Academic Press, New York, 45-72.

R. T. GREGORY and D. L. KARNEY: *A Collection of Matrices for Testing Computational Algorithms* (1969), Wiley-Interscience, New York. [Corrected reprint, R. E. Krieger Publ. Co., Huntington, N. Y., 1978].

M. A. JENKINS and J. F. TRAUB: *A three-stage algorithm for real polynomials using quadratic iteration,* SIAM J. Numer. Anal. **7**, (1970), 545-566.

M. A. JENKINS and J. F. TRAUB: *Principles for testing polynomial zerofinding programs,* ACM Trans. Math. Software **1**, (1975), 26-34.

B. T. SMITH: *ZERPOL, a zero finding algorithm for polynomials using Laguerre's method,* M. S. Thesis, Depertment of Computer Science, University of Toronto, (1967).

J. H. WILKINSON: *Rounding Errors in Algebraic Processes* (1963), Prentice-Hall, Englewood Cliffs, N. J..

30.2. [96] (With B. N. Flury) "AN ALGORITHM FOR SIMULTANEOUS ORTHOGONAL TRANSFORMATION OF SEVERAL POSITIVE DEFINITE SYMMETRIC MATRICES TO NEARLY DIAGONAL FORM"

---

[96] (with B. N. 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 (1986).

SIAM J. SCI. STAT. COMPUT.
Vol. 7, No. 1, January 1986

013

AN ALGORITHM FOR SIMULTANEOUS ORTHOGONAL TRANSFORMATION OF SEVERAL POSITIVE DEFINITE SYMMETRIC MATRICES TO NEARLY DIAGONAL FORM*

BERNHARD N. FLURY† AND WALTER GAUTSCHI‡

Abstract. For $k \geqq 1$ positive definite symmetric matrices $A_1, \cdots, A_k$ of dimension $p \times p$ we define the function $\Phi(A_1, \cdots, A_k; n_1, \cdots, n_k) = \prod_{i=1}^{k} [\det(\operatorname{diag} A_i)]^{n_i} / [\det(A_i)]^{n_i}$, where n_i are positive constants, as a measure of simultaneous deviation of $A_1, \cdots, A_k$ from diagonality. We give an iterative algorithm, called the FG-algorithm, to find an orthogonal $p \times p$-matrix B such that $\Phi(B^T A_1 B, \cdots, B^T A_k B; n_1, \cdots, n_k)$ is minimum. The matrix B is said to transform $A_1, \cdots, A_k$ simultaneously to nearly diagonal form. Conditions for the uniqueness of the solution are given.

The FG-algorithm can be used to find the maximum likelihood estimates of common principal components in k groups (Flury (1984)). For $k = 1$, the FG-algorithm computes the characteristic vectors of the single positive definite symmetric matrix A_1.

Key words. diagonalization, principal components, eigenvectors

1. The problem. It is well known (see, e.g., Basilevsky (1983, § 5.3) that if A is a positive definite symmetric (p.d.s.) matrix of dimension $p \times p$, then there exists a real orthogonal matrix B such that

$$B^T A B = \Lambda = \operatorname{diag}(\lambda_1, \cdots, \lambda_p), \tag{1.1}$$

where the λ_i are all positive. For $k > 1$ p.d.s. matrices $A_1, \cdots, A_k$ the associated orthogonal matrices are in general different. We call $A_1, \cdots, A_k$ simultaneously diagonalizable if there exists an orthogonal matrix B such that

$$B^T A_i B = \Lambda_i \text{ (diagonal)} \quad \text{for } i = 1, \cdots, k. \tag{1.2}$$

Conditions equivalent to (1.2) have been given by Flury (1983).

Now suppose that $A_1, \cdots, A_k$ are not simultaneously diagonalizable, but we wish to find an orthogonal matrix B which makes them simultaneously "as diagonal as possible" in a sense to be defined. As a simple measure of "deviation from diagonality" of a p.d.s. matrix F we can take

$$\varphi(F) = |\operatorname{diag} F| / |F|, \tag{1.3}$$

where $|\cdot|$ is the determinant and diag F is the diagonal matrix having the same diagonal elements as F. The fact that φ is a reasonable measure of deviation from diagonality can be seen from Hadamard's inequality (Noble and Daniel (1977, exercise 11.51)):

$$|F| \leqq |\operatorname{diag} F| \tag{1.4}$$

with equality exactly if F is diagonal. Therefore, $\varphi(F) \geqq 1$ holds, with equality exactly when F is diagonal. Actually, $\varphi(G)$ increases monotonically as G is continuously "inflated" from diag F to F. This can be seen from the following lemma.

LEMMA 1. *If $F = (f_{ij})$ is a p.d.s. matrix of dimension $p \times p$, then*

$$d(\alpha) := \det \begin{pmatrix} f_{11} & \alpha f_{12} & \cdots & \alpha f_{1p} \\ \alpha f_{21} & f_{22} & \cdots & \alpha f_{2p} \\ \vdots & \vdots & & \vdots \\ \alpha f_{p1} & \alpha f_{p2} & \cdots & f_{pp} \end{pmatrix} \tag{1.5}$$

* Received by the editors March 15, 1984. The work of the first author was supported by the Swiss National Science Foundation under contract #82.008.0.82 in the Department of Statistics at Purdue University.

† Department of Statistics, University of Berne, CH-3012 Berne, Switzerland.

‡ Department of Computer Sciences, Purdue University, West Lafayette, Indiana 47907. The work of this author was supported in part by the National Science Foundation under grant DCR 8320561.

is a decreasing function of α for $\alpha \in [0, 1]$. If F is not diagonal, $d(\alpha)$ is strictly decreasing.

Proof. The case $F = \operatorname{diag}(f_{11}, \cdots, f_{pp})$ is trivial; assume therefore that F is not diagonal. Write

$$\begin{aligned} d(\alpha) &= |\alpha F + (1-\alpha)\operatorname{diag} F| \\ &= |\operatorname{diag} F| \cdot |\alpha(\operatorname{diag} F)^{-1/2} F (\operatorname{diag} F)^{-1/2} + (1-\alpha) I_p| \end{aligned} \tag{1.6}$$

and note that $d(\alpha) > 0$ for all $\alpha \in [0, 1]$, since both F and $\operatorname{diag} F$ are p.d.s. Let $R = (\operatorname{diag} F)^{-1/2} F (\operatorname{diag} F)^{-1/2}$. R is p.d.s. with 1's on the main diagonal. Let $d_1(\alpha) = |\alpha R + (1-\alpha) I_p|$. Then $d_1(0) = 1$ and $d_1(1) < 1$ by Hadamard's inequality. It remains to show that $d_1(\alpha)$ is strictly decreasing in $(0, 1)$. Let $\rho_1 \geqq \rho_2 \geqq \cdots \geqq \rho_p > 0$ denote the eigenvalues of R. The eigenvalues of $\alpha R + (1-\alpha) I_p$ are

$$\gamma_j = \alpha \rho_j + 1 - \alpha \qquad (j = 1, \cdots, p), \tag{1.7}$$

and therefore

$$d_1(\alpha) = \prod_{j=1}^{p} \gamma_j = \prod_{j=1}^{p} (1 + \alpha(\rho_j - 1)). \tag{1.8}$$

Taking the first derivative gives

$$\frac{\partial d_1}{\partial \alpha} = \sum_{h=1}^{p} (\rho_h - 1) \prod_{\substack{j=1 \\ j \neq h}}^{p} [1 + \alpha(\rho_j - 1)] = d_1(\alpha) \sum_{j=1}^{p} \frac{\rho_j - 1}{1 + \alpha(\rho_j - 1)}, \tag{1.9}$$

where all denominators are positive because of $\rho_j > 0$ and $\alpha \leqq 1$. Letting

$$d_2(\alpha) = \sum_{j=1}^{p} \frac{\rho_j - 1}{1 + \alpha(\rho_j - 1)}, \tag{1.10}$$

we note that $d_2(0) = \sum_{j=1}^{p} (\rho_j - 1) = \operatorname{tr} R - p = 0$ and

$$\frac{\partial d_2}{\partial \alpha} = -\sum_{j=1}^{p} \frac{(\rho_j - 1)^2}{(1 + \alpha(\rho_j - 1))^2} < 0. \tag{1.11}$$

Therefore, $d_2(\alpha) < 0$ on $(0, 1]$, implying that $\partial d_1 / \partial \alpha < 0$ for $0 < \alpha \leqq 1$. This proves the lemma.

The reader may notice a similarity to ridge regression: Hoerl and Kennard (1970, Thms. 4.1 and 4.2) have given monotonicity properties of some functions related to the trace of the matrix $(F + \alpha I_p)^{-1}$ for $\alpha > 0$.

Let us now consider k p.d.s. matrices $F_1, \cdots, F_k$ and positive weights $n_1, \cdots, n_k$. Then we define the simultaneous deviation from diagonality of the matrices $F_1, \cdots, F_k$ with given weights $n_1, \cdots, n_k$ as

$$\Phi(F_1, \cdots, F_k; n_1, \cdots, n_k) = \prod_{i=1}^{k} [\varphi(F_i)]^{n_i}. \tag{1.12}$$

Let now $F_i = B^T A_i B$ $(i = 1, \cdots, k)$ for a given orthogonal matrix B. Then we can take

$$\Phi_0(A_1, \cdots, A_k; n_1, \cdots, n_k) = \min_{B \in O(p)} \Phi(B^T A_1 B, \cdots, B^T A_k B; n_1, \cdots, n_k), \tag{1.13}$$

where $O(p)$ is the group of orthogonal $p \times p$-matrices, as a measure of simultaneous diagonalizability of $A_1, \cdots, A_k$. Clearly, $\Phi_0 \geqq 1$ holds, with equality if and only if (1.2) is satisfied.

It can be shown (Flury (1984)) that if the minimum Φ_0 is attained for a matrix $B_0=(b_1,\cdots,b_p)\in O(p)$, then the following system of equations holds:

$$b_l^T\left(\sum_{i=1}^{k} n_i\frac{\lambda_{il}-\lambda_{ij}}{\lambda_{il}\lambda_{ij}}A_i\right)b_j=0 \qquad (l,j=1,\cdots,p;\ l\neq j) \tag{1.14}$$

where

$$\lambda_{ih}=b_h^TA_ib_h \qquad (i=1,\cdots,k;\ h=1,\cdots,p). \tag{1.15}$$

In this paper we give an algorithm for finding B_0.

It may be noted that our measure of "deviation from diagonality" (formula (1.3)) is not the only natural one; one could, for instance, also consider the sum of squared off-diagonal elements. Our reason for considering (1.3) was that this criterion arises naturally from a statistical problem of maximizing a likelihood function in principal component analysis of several groups; see Flury (1984) for details.

2. The FG-algorithm. The FG-algorithm consists of two algorithms, called F and G respectively, which minimize Φ by iteration on two levels:

On the outer level (F-level), every pair (b_l, b_j) of column vectors of the current approximation B to the solution B_0 is rotated such that the corresponding equation in (1.14) is satisfied. One iteration step of the F-algorithm consists of rotations of all $p(p-1)/2$ pairs of vectors of B. The F-algorithm is similar to algorithms used in factor analysis to perform varimax and other rotations (see, e.g., Weber (1974)).

On the inner level (G-level), an orthogonal 2×2-matrix Q which solves a two dimensional analog of (1.14) is found by iteration. This matrix defines the rotation of a pair of vectors currently being used on the F-level.

THE F-ALGORITHM. Let

$$\Phi(B)=\Phi(B^TA_1B,\cdots,B^TA_kB;\ n_1,\cdots,n_k) \tag{2.1}$$

denote the simultaneous deviation of $B^TA_1B,\cdots,B^TA_kB$ from diagonality as a function of B, the A_i and n_i being considered as fixed. The F-algorithm yields a converging sequence of orthogonal matrices $B^{(0)}, B^{(1)},\cdots$, such that $\Phi(B^{(f+1)})\leqq\Phi(B^{(f)})$.

The algorithm proceeds as follows:

step F_0: Define $B=(b_1,\cdots,b_p)\in O(p)$ as an initial approximation to the orthogonal matrix minimizing Φ, e.g. $B\leftarrow I_p$. Put $f\leftarrow 0$.

step F_1: Put $B^{(f)}\leftarrow B$ and $f\leftarrow f+1$

step F_2: Repeat steps F_{21} to F_{24} for all pairs (l,j), $1\leqq l<j\leqq p$:

step F_{21}: Put $H(p\times 2)\leftarrow(b_l, b_j)$ and

$$T_i(2\times 2)\leftarrow\begin{pmatrix} b_l^TA_ib_l & b_l^TA_ib_j\\ b_j^TA_ib_l & b_j^TA_ib_j\end{pmatrix} \qquad (i=1,\cdots,k).$$

The T_i are p.d.s.

step F_{22}: Perform the G-algorithm on $(T_1,\cdots,T_k)$ to get an orthogonal

$$2\times 2-\text{matrix } Q=\begin{pmatrix}\cos\alpha & -\sin\alpha\\ \sin\alpha & \cos\alpha\end{pmatrix}.$$

step F_{23}: Put $H^*(p\times 2)=(b_l^*, b_j^*)\leftarrow HQ$. (This corresponds to an orthogonal rotation of the two columns of H by an angle α).

step F_{24}: In the matrix B, replace columns b_l and b_j by b_l^* and b_j^*, respectively, and call the new matrix again B.

step F_3: If, for some small $\varepsilon_F>0$, $\Phi(B^{(f-1)})-\Phi(B)<\varepsilon_F$ holds, stop. Otherwise, start the next iteration step at F_1.

THE G-ALGORITHM. This algorithm solves the equation

$$q_1^T\left(n_1\frac{\delta_{11}-\delta_{12}}{\delta_{11}\delta_{12}}T_1+\cdots+n_k\frac{\delta_{k1}-\delta_{k2}}{\delta_{k1}\delta_{k2}}T_k\right)q_2=0, \tag{2.2}$$

where $T_1,\cdots,T_k$ are fixed p.d.s. 2×2-matrices, $n_i>0$ are fixed constants,

$$\delta_{ij}=q_j^T T_i q_j \qquad (i=1,\cdots,k;\, j=1,2), \tag{2.3}$$

and $Q=(q_1, q_2)$ is an orthogonal 2×2-matrix. The iteration of the algorithm yields a sequence of orthogonal matrices $Q^{(0)}, Q^{(1)},\cdots$, converging to a solution of (2.2).

The algorithm proceeds as follows:

step G_0: Define $Q(2\times 2)$ as an initial approximation to the solution of (2.2), e.g. $Q\leftarrow I_2$. Put $g\leftarrow 0$.

step G_1: Put $Q^{(g)}\leftarrow Q$ and $g\leftarrow g+1$.

step G_2: Compute the δ_{ij} (2.3), using the current Q. Put

$$T(2\times 2)\leftarrow n_1\frac{\delta_{11}-\delta_{12}}{\delta_{11}\delta_{12}}T_1+\cdots+n_k\frac{\delta_{k1}-\delta_{k2}}{\delta_{k1}\delta_{k2}}T_k.$$

step G_3: Compute the (normalized) eigenvectors of T. In $Q=(q_1, q_2)$, put $q_1\leftarrow$ first eigenvector of T, $q_2\leftarrow$ second eigenvector of T.

step G_4: If $\|Q^{(g-1)}-Q\|<\varepsilon_G$ (where $\|\cdot\|$ denotes a matrix norm and $\varepsilon_G>0$ is a small positive constant), stop. Otherwise, start the next iteration step at G_1. Note that, since the order of eigenvectors is arbitrary, as well as their signs, it may be necessary to exchange q_1 and q_2 and/or to multiply one or both columns of Q by -1 before comparing Q with $Q^{(g-1)}$.

The motivation for the two algorithms and their connection with the basic system of equations (1.14) is as follows. Suppose that the (l, j)th equation of (1.14) is to be solved. With $H=(b_l: b_j)$ denoting the current lth and jth columns of B, and λ_{ih} being defined as in (1.15), b_l and b_j are the desired solution if and only if the 2×2-matrix

$$\sum_{i=1}^{k} n_i\frac{\lambda_{il}-\lambda_{ij}}{\lambda_{il}\lambda_{ij}}T_i \tag{2.4}$$

is diagonal, where

$$T_i=H^T A_i H \qquad (i=1,\cdots,k). \tag{2.5}$$

Assume now that b_l and b_j do not solve the (l, j)th equation, but $b_l^*=Hq_1$ and $b_j^*=Hq_2$ do, where $Q=(q_1: q_2)$ is an orthogonal 2×2 matrix. Then

$$b_l^{*T}\left[\sum_{i=1}^{k} n_i\frac{\lambda_{il}^*-\lambda_{ij}^*}{\lambda_{il}^*\lambda_{ij}^*}A_i\right]b_j^*=0, \tag{2.6}$$

where

$$\lambda_{ih}^*=b_h^{*T}A_i B_h^* \qquad (i=1,\cdots,k,\, h=l,j). \tag{2.7}$$

Putting $H^*=(b_l^*: b_j^*)=HQ$, (2.6) holds precisely if

$$\sum_{i=1}^{k} n_i\frac{\lambda_{il}^*-\lambda_{ij}^*}{\lambda_{il}^*\lambda_{ij}^*}H^{*T}A_i H^* \tag{2.8}$$

is diagonal. Now we note that

$$H^{*T}A_i H^*=(HQ)^T A_i(HQ)=Q^T T_i Q, \tag{2.9}$$

$$\lambda_{ii}^* = (Hq_1)^T A_i (Hq_1) = q_1^T T_i q_1 \qquad (i = 1, \cdots, k), \tag{2.10}$$

and

$$\lambda_{ij}^* = q_2^T T_i q_2 \qquad (i = 1, \cdots, k). \tag{2.11}$$

Thus the problem of rotating the lth and jth columns of B so as to satisfy (1.14) can be reduced completely to the problem of finding an orthogonal 2×2-matrix $Q = (q_1 : q_2)$ such that

$$q_1^T \left[\sum_{i=1}^{k} n_i \frac{\delta_{i1} - \delta_{i2}}{\delta_{i1}\delta_{i2}} T_i \right] q_2 = 0, \tag{2.12}$$

where δ_{i1} and δ_{i2} have been written in place of λ_{ii}^* and λ_{ij}^*, respectively.

Since (2.12) is a 2-dimensional analogue of (1.14), and since the group of orthogonal 2×2-matrices is compact, it follows that a solution of (2.12) always exists.

The problem of solving (2.12) is itself nontrivial. Although (2.12) can be written in terms of a rotation angle α, solving for α would involve solving a polynomial equation of degree $4k$ in $\cos \alpha$ and $\sin \alpha$, which seems rather tedious. A more elegant solution is provided by the G-algorithm, which is based on the observation that the vectors q_1, q_2 satisfying (2.12) are eigenvectors of the matrix in brackets. Since the latter, however, depends also on q_1, q_2, an iterative procedure is required.

3. Convergence of the FG-algorithm.

3.1. Convergence of the F-algorithm. We show that the F-algorithm, in theory (i.e. if $\varepsilon_F = \varepsilon_G = 0$), does not stop unless the equations (1.14) are satisfied for the current B, and that, if B does not satisfy (1.14), an iteration step of the F-algorithm will decrease Φ.

Suppose that the current orthogonal matrix $B = (b_1, \cdots, b_p)$ does not satisfy the (l, j)th equation of (1.14). For notational simplicity, we can take $l = 1$ and $j = 2$ without loss of generality. Let us write $B = (B^{(1)} : B^{(2)})$, where $B^{(1)} = (b_1, b_2)$. In step F_{21}, the matrices $T_i = B^{(1)T} A_i B^{(1)}$ are passed to the G-algorithm. The G-algorithm gives back an orthogonal 2×2 matrix Q (step F_{22}). (Note that Q is not necessarily unique, depending upon the conventions used in the G-algorithm. We will consider every matrix $\bar{Q}$ obtained from Q by interchainging the columns of Q and/or multiplying one or both columns by -1 as *equivalent* to Q). Steps F_{23} and F_{24} correspond to the transformation

$$B^* = B \begin{pmatrix} Q & 0 \\ 0 & I_{p-2} \end{pmatrix} = (B^{(1)} Q : B^{(2)}). \tag{3.1}$$

B^* is orthogonal, since it is the product of two orthogonal matrices. Now we have

$$\begin{aligned} \Phi(B^*) &= \prod_{i=1}^{k} [|\operatorname{diag} B^{*T} A_i B^*| / |B^{*T} A_i B^*|]^{n_i} \\ &= \prod_{i=1}^{k} [|\operatorname{diag} Q^T B^{(1)T} A_i B^{(1)} Q| \cdot |\operatorname{diag} B^{(2)T} A_i B^{(2)}| / |A_i|]^{n_i}. \end{aligned} \tag{3.2}$$

It will be shown in § 3.2 that if, as assumed, (1.14) is not satisfied for $l = 1$ and $j = 2$, then

$$\prod_{i=1}^{k} |\operatorname{diag} Q^T B^{(1)T} A_i B^{(1)} Q|^{n_i} < \prod_{i=1}^{k} |\operatorname{diag} B^{(1)T} A_i B^{(1)}|^{n_i}. \tag{3.3}$$

If (1.14) is satisfied, Q will be equivalent to I_2, and hence (3.3) holds with equality.

Therefore, each iteration step of the F-algorithm decreases Φ, and the algorithm will stop only if (1.14) is satisfied.

3.2. Convergence of the G-algorithm. In analogy to (1.13), (1.14), the equation (2.2) is satisfied for the matrix $Q=(q_1, q_2)$ which minimizes

$$\left(\prod_{i=1}^{k} |T_i|^{n_i}\right)\Phi(Q) = \prod_{i=1}^{k} |\text{diag}\,(Q^T T_i Q)|^{n_i}.$$

Let $Q^{(g)}$ denote the orthogonal 2×2 matrix after the gth iteration. We will show that

$$\prod_{i=1}^{k} |\text{diag}\, Q^{(g+1)^T} T_i Q^{(g+1)}|^{n_i} \leqq \prod_{i=1}^{k} |\text{diag}\, Q^{(g)^T} T_i Q^{(g)}|^{n_i}, \tag{3.4}$$

and that the sequence $Q^{(g)}$ converges to an orthogonal matrix which solves (2.2).

Suppose now that the $(g+1)$st iteration of the G-algorithm is being performed. It is somewhat simpler to prove the convergence if we introduce the following notation: Let $Q=(q_1, q_2)$ contain the current approximation to the solution of (2.2) and δ_{ij} be defined as in (2.3). Then we put

$$a_i = \frac{\delta_{i1}-\delta_{i2}}{\delta_{i1}\delta_{i2}} \qquad (i=1, \cdots, k), \tag{3.5}$$

$$T = \sum_{i=1}^{k} n_i a_i T_i, \tag{3.6}$$

and

$$U_i = Q^T T_i Q = \begin{pmatrix} u_{11}^{(i)} & u_{12}^{(i)} \\ u_{21}^{(i)} & u_{22}^{(i)} \end{pmatrix}. \tag{3.7}$$

The U_i are p.d.s., and clearly

$$T = \sum_{i=1}^{k} n_i a_i Q U_i Q^T \tag{3.8}$$

and

$$\delta_{i1} = u_{11}^{(i)}, \qquad \delta_{i2} = u_{22}^{(i)}. \tag{3.9}$$

In step G_3 the characteristic vectors of T are computed. Denote the solution by Q^*, so that

$$Q^{*T} T Q^* = \Lambda \tag{3.10}$$

is diagonal. From (3.8) it follows that

$$\sum_{i=1}^{k} n_i a_i Q^{*T} Q U_i Q^T Q^* = \Lambda. \tag{3.11}$$

The characteristic vectors of the symmetric matrix

$$U = \sum_{i=1}^{k} n_i a_i U_i \tag{3.12}$$

are therefore given by the orthogonal matrix

$$P = Q^T Q^*, \tag{3.13}$$

and Q^* can therefore be obtained by

$$Q^* = QP. \tag{3.14}$$

Note that $U = Q^T T Q$ is diagonal if and only if $Q = (q_1, q_2)$ is a solution of (2.2). From (3.9) and (3.12) it follows that

$$U = \sum_{i=1}^{k} n_i \frac{u_{11}^{(i)} - u_{22}^{(i)}}{u_{11}^{(i)} u_{22}^{(i)}} \begin{pmatrix} u_{11}^{(i)} & u_{12}^{(i)} \\ u_{21}^{(i)} & u_{22}^{(i)} \end{pmatrix}. \tag{3.15}$$

Let

$$\vartheta_i = \begin{cases} 1 & \text{if } u_{11}^{(i)} > u_{22}^{(i)}, \\ -1 & \text{if } u_{11}^{(i)} < u_{22}^{(i)}, \\ 0 & \text{if } u_{11}^{(i)} = u_{22}^{(i)}, \end{cases} \tag{3.16}$$

$$a_i' = \vartheta_i a_i, \tag{3.17}$$

and

$$S_i = \begin{pmatrix} s_{11}^{(i)} & s_{12}^{(i)} \\ s_{21}^{(i)} & s_{22}^{(i)} \end{pmatrix} = a_i' U_i \qquad (i = 1, \cdots, k). \tag{3.18}$$

S_i is p.d.s., unless $\vartheta_i = 0$. With this notation, we have

$$U = \sum_{i=1}^{k} n_i \vartheta_i S_i. \tag{3.19}$$

Now let $k' \leqq k$ denote the number of ϑ_i's which are not zero. Without loss of generality assume that the $k - k'$ matrices S_i which are zero have the indices $k'+1, \cdots, k$. Therefore the sum (3.19) extends only up to k', and we are going to show that

$$\prod_{i=1}^{k'} |\text{diag } P^T S_i P|^{n_i} \leqq \prod_{i=1}^{k'} (s_{11}^{(i)} s_{22}^{(i)})^{n_i}, \tag{3.20}$$

with equality if and only if U is diagonal. Assuming for the moment that (3.20) holds true, the proof of (3.4) can be completed by noting that (3.20) implies

$$\prod_{i=1}^{k'} (a_i^2 \vartheta_i^2 |\text{diag } P^T U_i P|)^{n_i} \leqq \prod_{i=1}^{k'} (a_i^2 \vartheta_i^2 u_{11}^{(i)} u_{22}^{(i)})^{n_i} \tag{3.21}$$

and therefore

$$\prod_{i=1}^{k'} |\text{diag } Q^{*T} T_i Q^*|^{n_i} \leqq \prod_{i=1}^{k'} |\text{diag } Q^T T_i Q|^{n_i}. \tag{3.22}$$

For the remaining $k - k'$ matrices $U_i (i = k'+1, \cdots, k)$ we have $u_{11}^{(i)} = u_{22}^{(i)}$, and therefore, as is easily verified,

$$|\text{diag } B^T U_i B| \leqq |\text{diag } U_i| \tag{3.23}$$

for any $B \in 0(2)$, with equality exactly if B is equivalent to I_2 or $u_{12}^{(i)} = 0$. This holds, in particular, for $B = P$. Putting (3.22) and (3.23) together gives now the desired result (3.4). It remains to show (3.20).

Let $P = (p_1, p_2)$ denote the eigenvectors of $U = \sum_{i=1}^{k'} \vartheta_i n_i S_i$, with p_1 being associated with the algebraically larger root. Since U is symmetric, P is orthogonal (or can be so chosen if the two roots are identical), and both characteristic roots are real. Assume

that U is not diagonal, and let $\varepsilon_i (i=1,\cdots,k')$ be defined by

$$P^T S_i P = \begin{pmatrix} s_{11}^{(i)} + \vartheta_i \varepsilon_i & \cdot \\ \cdot & s_{22}^{(i)} - \vartheta_i \varepsilon_i \end{pmatrix}. \tag{3.24}$$

From (3.5), (3.9) and (3.16) to (3.18) we have

$$s_{11}^{(i)} s_{22}^{(i)} = \vartheta_i (s_{11}^{(i)} - s_{22}^{(i)}), \quad \text{or} \quad 1 = \vartheta_i \left(\frac{1}{s_{22}^{(i)}} - \frac{1}{s_{11}^{(i)}} \right) \qquad (i=1,\cdots,k'), \tag{3.25}$$

which implies that either $s_{11}^{(i)}$ or $s_{22}^{(i)}$ is smaller than 1. It then follows that $\varepsilon_i < 1$ $(i=1,\cdots,k')$. Indeed, if $\vartheta_i = 1$, then $s_{22}^{(i)} < 1$ by (3.25), and the positivity of $s_{22}^{(i)} - \varepsilon_i$ implies $\varepsilon_i < 1$. If $\vartheta_i = -1$, then $s_{11}^{(i)} < 1$, and $s_{11}^{(i)} - \varepsilon_i > 0$ implies again $\varepsilon_i < 1$.

The product of the diagonal elements of $P^T S_i P$ is

$$\begin{aligned} |\text{diag } P^T S_i P| &= (s_{11}^{(i)} + \vartheta_i \varepsilon_i)(s_{22}^{(i)} - \vartheta_i \varepsilon_i) \\ &= s_{11}^{(i)} s_{22}^{(i)} - \varepsilon_i \vartheta_i (s_{11}^{(i)} - s_{22}^{(i)}) - \varepsilon_i^2 \\ &= (1-\varepsilon_i) s_{11}^{(i)} s_{22}^{(i)} - \varepsilon_i^2 \\ &\leqq (1-\varepsilon_i) s_{11}^{(i)} s_{22}^{(i)} \qquad (i=1,\cdots,k'). \end{aligned} \tag{3.26}$$

Thus,

$$\prod_{i=1}^{k'} |\text{diag } P^T S_i P|^{n_i} \leqq \left(\prod_{i=1}^{k'} (1-\varepsilon_i)^{n_i} \right) \left(\prod_{i=1}^{k'} |\text{diag } S_i|^{n_i} \right), \tag{3.27}$$

and (3.20) holds if we can prove that

$$\prod_{i=1}^{k'} (1-\varepsilon_i)^{n_i} < 1. \tag{3.28}$$

To demonstrate this, we note that, since U is assumed not diagonal,

$$p_1^T U p_1 > u_{11}, \tag{3.29}$$

or equivalently,

$$\begin{aligned} &\sum_{i=1}^{k'} \vartheta_i n_i p_1^T S_i p_1 > \sum_{i=1}^{k'} \vartheta_i n_i s_{11}^{(i)}, \\ &\sum_{i=1}^{k'} \vartheta_i n_i (p_1^T S_i p_1 - s_{11}^{(i)}) > 0. \end{aligned} \tag{3.30}$$

Since $p_1^T S_i p_1 - s_{11}^{(i)} = \vartheta_i \varepsilon_i (i=1,\cdots,k')$, this implies

$$\sum_{i=1}^{k'} n_i \varepsilon_i > 0, \tag{3.31}$$

so that not all ε_i can be zero. On the other hand, if U is diagonal, then P is equivalent to I_2, and all ε_i are zero. Therefore the ε_i vanish simultaneously if and only if U is diagonal. Now we need the following lemma.

LEMMA 2. *If* $x_i > 0$, $n_i > 0 (i=1,\cdots,k')$ *and* $\sum_{i=1}^{k'} n_i x_i \leqq \sum_{i=1}^{k'} n_i$, *then* $\prod_{i=1}^{k'} x_i^{n_i} \leqq 1$.

Proof. Maximize the function $\prod_{i=1}^{k'} x_i^{n_i}$ under the restriction $\sum_{i=1}^{k'} n_i x_i = g$ (>0), using a Lagrange multiplier. The maximum has the value $(g/n)^n$ and is attained for $x_1 = \cdots = x_{k'} = g/n$, where $n = \sum_{i=1}^{k'} n_i$. Noting that $g \leqq n$ completes the proof.

Since $\varepsilon_i < 1 (i=1,\cdots,k')$ and $\sum_{i=1}^{k'} n_i \varepsilon_i > 0$, we can use Lemma 2 with $x_i = 1 - \varepsilon_i$ and get (3.28). Note that equality in (3.28) holds exactly if all ε_i are zero. This completes the proof of convergence of the G-algorithm.

4. Conditions for uniqueness of the solution. In § 3 we have shown that the FG-algorithm converges to a minimum of (2.1), unless the initial approximation of the orthogonal matrix B is (badly) chosen as a stationary point of Φ. However, we do not know whether Φ has a unique minimum. We are now going to show that in some "extreme" cases there exist more than one local minimum, and we give approximate conditions when this will happen. Throughout this section (unless otherwise stated) we will only consider the case $k=2$ and $p=2$.

Let the p.d.s. matrix S_1 have the characteristic roots $l_1 > l_2$ (the case $l_1 = l_2$ being trivial), and assume, for simplicity, that

$$S_1 = \begin{pmatrix} l_1 & 0 \\ 0 & l_2 \end{pmatrix}. \tag{4.1}$$

From (3.5) it can be seen that the solutions of (2.2) are unaffected by proportionality, i.e., we can assume (see also (3.25))

$$l_1 - l_2 = l_1 l_2 \tag{4.2}$$

without loss of generality. Consider now an orthogonal matrix

$$B = B(\varphi) = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}. \tag{4.3}$$

The product of the diagonal elements of $B^T S_1 B$ is

$$\begin{aligned} |\mathrm{diag}(B^T S_1 B)| &= [l_2 + (l_1 - l_2)\cos^2\varphi][l_1 - (l_1 - l_2)\cos^2\varphi] \\ &= l_1 l_2 + (l_1 - l_2)^2 \cos^2\varphi \sin^2\varphi \\ &= l_1 l_2 [1 + l_1 l_2 \cos^2\varphi \sin^2\varphi] \\ &= r_1 [1 + r_1 \cos^2\varphi \sin^2\varphi], \end{aligned} \tag{4.4}$$

where

$$r_1 = l_1 l_2 \tag{4.5}$$

denotes the product of the characteristic roots of S_1. Let the eccentricity d_1 of S_1 be defined as the ratio of the larger to the smaller root of S_1,

$$d_1 = l_1 / l_2, \tag{4.6}$$

which is also the Euclidean condition number of S_1. From (4.2) it follows that $l_2 = l_1/(l_1+1)$, and therefore $d_1 = l_1 + 1$. Similarly, $d_1 = 1/(1-l_2)$, and therefore

$$l_1 = d_1 - 1, \qquad l_2 = (d_1 - 1)/d_1. \tag{4.7}$$

Multiplying these two equations gives

$$r_1 = (d_1 - 1)^2 / d_1. \tag{4.8}$$

Note that d_1 does not depend on the absolute size of S_1 (every matrix proportional to S_1 has the same eccentricity), and so the same is true for r_1.

For a second p.d.s. matrix S_2, let d_2 denote its eccentricity, and

$$r_2 = (d_2 - 1)^2 / d_2. \tag{4.9}$$

Let

$$B_0 = \begin{pmatrix} \cos\varphi_0 & -\sin\varphi_0 \\ \sin\varphi_0 & \cos\varphi_0 \end{pmatrix}$$

denote the orthogonal matrix which diagonalizes S_2. Then, in analogy to (4.4), we get

$$|\text{diag}\,(B^T S_2 B)| = r_2[1 + r_2 \cos^2(\varphi - \varphi_0) \sin^2(\varphi - \varphi_0)]. \tag{4.10}$$

The function Φ to be minimized is

$$\Phi(\varphi) = [1 + r_1 \cos^2\varphi \sin^2\varphi]^{n_1}[1 + r_2 \cos^2(\varphi - \varphi_0) \sin^2(\varphi - \varphi_0)]^{n_2}. \tag{4.11}$$

Let us now assume that $n_1 = n_2$, so that it remains to minimize

$$G(\varphi) = [1 + \tfrac{1}{4} r_1 \sin^2(2\varphi)][1 + \tfrac{1}{4} r_2 \sin^2(2(\varphi - \varphi_0))]. \tag{4.12}$$

$G(\varphi)$ is $\pi/2$-periodic, and from (4.12) it becomes clear that for $\varphi_0 \neq 0$, $G(\varphi)$ may have more than one local minimum in one period, depending on r_1, r_2 and φ_0 (and, in the general situation, on n_1 and n_2). Note that φ_0 is the minimum angle between two characteristic vectors of S_1 and S_2.

Let us first look at the extreme situation $\varphi_0 = \pi/4$. From a Taylor expansion it can be seen that in a neighborhood of 0,

$$G(\varphi) = 1 + \tfrac{1}{4} r_2 + (r_1 - r_2 + \tfrac{1}{4} r_1 r_2)\varphi^2 + O(\varphi^4). \tag{4.13}$$

The function $G(\varphi)$ has therefore a stationary point at $\varphi = 0$, which is a

$$\begin{aligned} &\text{minimum,} \quad \text{if } r_1 - r_2 + \tfrac{1}{4} r_1 r_2 > 0, \\ &\text{maximum,} \quad \text{if } r_1 - r_2 + \tfrac{1}{4} r_1 r_2 < 0. \end{aligned} \tag{4.14}$$

Note that for $r_1 \geqq 4$ or $r_1 = r_2$ this is always a minimum.

Similarly, at $\varphi = \pi/4$, we get a

$$\begin{aligned} &\text{minimum,} \quad \text{if } r_2 - r_1 + \tfrac{1}{4} r_1 r_2 > 0, \\ &\text{maximum,} \quad \text{if } r_2 - r_1 + \tfrac{1}{4} r_1 r_2 < 0. \end{aligned} \tag{4.15}$$

For $r_2 \geqq 4$ or $r_1 = r_2$ this is always a minimum.

Since r_1 and r_2 are both positive, there cannot be a maximum at 0 and $\pi/4$ simultaneously. Local minima at both points, however, are obtained e.g. if both r_1 and r_2 are larger than 4, or if $r_1 = r_2$ (even if $r_1 = r_2$ is very close to zero!). Thus the case of equal eccentricity of both matrices seems most "dangerous" in terms of multiple local minima.

Using the relation $r_i = (d_i - 1)^2 / d_i$ (4.8, 4.9), the conditions (4.14) and (4.15) can be transformed to conditions on the eccentricities d_i ($i = 1, 2$). Figure 1 shows a partition of $[1, \infty) \times [1, \infty)$ into three areas in which a minimum is attained at 0 only, at $\pi/4$ only, or at both points, depending on the values of d_1 and d_2. Note that for $d_1 > 5.828427$ ($d_2 > 5.828427$) there is always a minimum at $\varphi = 0$ ($\varphi = \pi/4$), and if $d_1 = d_2$, there are always two minima.

The case $\varphi_0 = \pi/4$ treated so far is of course the "worst possible" case, since the minimum angle between two characteristic vectors of S_1 and S_2 cannot exceed $\pi/4$. For the application in common principal component analysis (Flury (1984)), however, we expect φ_0 rather close to zero, if the null hypothesis of identical principal components in the populations holds. Therefore we look now at the situation where φ_0 is close to zero. Without loss of generality we can assume $\varphi_0 > 0$. Again, for simplicity, we take $n_1 = n_2 = 1$.

Approximating the trigonometric functions in the two factors of G by Taylor series at $\varphi = 0$ (first factor) and at $\varphi = \varphi_0$ (second factor), and taking the first derivative

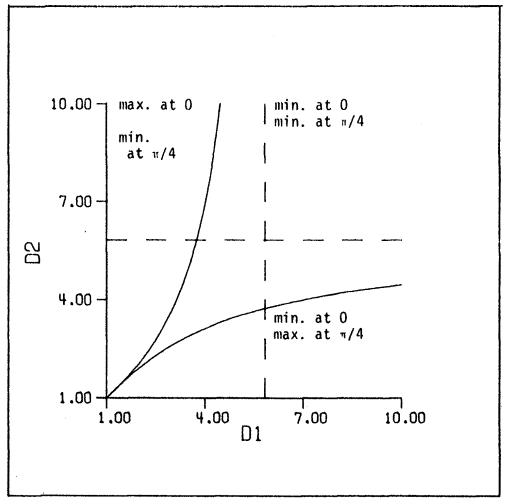


FIG. 1. *Conditions for minima and maxima if* $n_1 = n_2$, $\varphi_0 = \pi/4$.

of G with respect to φ yields

$$G'(\varphi) = 2r_1[\varphi + O(\varphi^3)][1 + r_2((\varphi-\varphi_0)^2 + O(\varphi-\varphi_0)^4)] + 2r_2[(\varphi-\varphi_0) + O(\varphi-\varphi_0)^3][1 + r_1(\varphi^2 + O(\varphi^4))]. \tag{4.16}$$

If φ_0 is close to zero, sufficient accuracy can be had for $0 \leqq \varphi \leqq \varphi_0$ if we ignore all terms of order higher than 2. An approximation to the solution(s) of $G'(\varphi) = 0$ within $[0, \varphi_0]$ is therefore given by the solution(s) of

$$r_1\varphi[1 + r_2(\varphi-\varphi_0)^2] + r_2(\varphi-\varphi_0)[1 + r_1\varphi^2] = 0. \tag{4.17}$$

This equation has either one or three real roots, depending on r_1, r_2 and φ_0. For $r_1 = r_2 = r$, (4.17) can be written as

$$\varphi(1 + r(\varphi-\varphi_0)^2) + (\varphi-\varphi_0)(1 + r\varphi^2) = 2\left(\varphi - \frac{\varphi_0}{2}\right)(r\varphi^2 - r\varphi_0\varphi + 1) = 0. \tag{4.18}$$

Thus $\varphi = \varphi_0/2$ is a solution of (4.18) (and also of (4.16) if $r_1 = r_2$). If, approximately,

$$\frac{4}{r} > \varphi_0^2, \tag{4.19}$$

$G(\varphi)$ takes a minimum at $\varphi_0/2$. Under the same condition (4.19), the polynomial

$$r\varphi^2 - r\varphi_0\varphi + 1 \tag{4.20}$$

has no real root, and the minimum at $\varphi_0/2$ is unique.

If, always approximately for small φ_0, $4/r < \varphi_0^2$, we get a maximum at $\varphi_0/2$, and two minima at

$$\tfrac{1}{2}(\varphi_0 \pm \sqrt{\varphi_0^2 - 4/r}). \tag{4.21}$$

In terms of the eccentricity parameters $d_1 = d_2 = d$, condition (4.19) becomes

$$\frac{4d}{(d-1)^2} > \varphi_0^2, \tag{4.22}$$

which shows that two minima are to be expected only if the eccentricity is high. For large d, (4.22) is approximately the same as

$$d < \left(\frac{2}{\varphi_0}\right)^2. \tag{4.23}$$

For example, if $\varphi_0 = .2 (\approx 11.5$ degrees), a single minimum can be expected approximately if $d < 100$.

Figure 2 shows the typical behavior of the function $G(\varphi)$ for $\varphi_0 = .2$ and $r = 160$ ($d = 161.99$). The two minima are approximately at .039 and .161. If different values are chosen for r_1 and r_2, the two minima are in general not identical, but the shape of the graph is similar, with one "valley" being less deep than the other.

Although these results are only approximate, they give a general idea about the conditions for uniqueness of the minimum. For $k > 2$ matrices, the relations are of course more complicated, but still we can expect a unique minimum unless some of the matrices are highly eccentric.

For dimension $p > 2$, the minimum is certainly unique if all the $p(p-1)/2$ equations (1.14) have a unique minimizing solution. (By a minimizing solution we mean a solution which corresponds to a local minimum of G, or, in the p-dimensional case, of the function Φ.) On the other hand, if some of the equations have more than one minimizing solution, this does not necessarily imply that the whole system (1.14) has more than one minimizing solution.

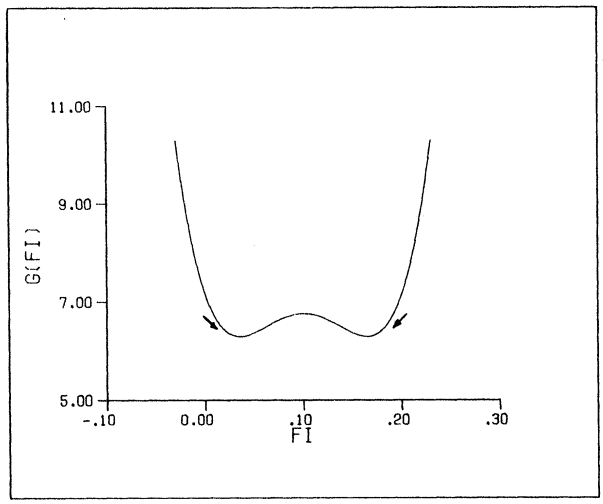


FIG. 2. *Graph of* $G(\varphi)$ *for* $\varphi_0 = .2$, $n_1 = n_2 = 1$ *and* $d_1 = d_2 = 162$.

A solution given by the FG-algorithm does of course not prove its uniqueness. However, Fig. 2 suggests the following: If we start the FG-algorithm with

$$B(0) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

as an initial approximation, it will converge to the left minimum, while

$$B(\varphi_0) = \begin{pmatrix} \cos\varphi_0 & -\sin\varphi_0 \\ \sin\varphi_0 & \cos\varphi_0 \end{pmatrix}$$

as an initial approximation leads to convergence to the right minimum. (This is indicated

by the arrows in Fig. 2.) Since the function Φ is a product of k functions (see 1.12), minimizing solutions can always be expected to be somehow "close" to the characteristic vectors of one of the matrices. Therefore, if there is doubt about the uniqueness of the solution, it is recommended that one run the FG-algorithm k times, using the k sets of characteristic vectors of $A_1, \cdots, A_k$ as initial approximations. If all k solutions found are equal, it is reasonable to assume that there is a unique global minimum.

As a numerical example, let

$$S_1 = \begin{pmatrix} 100 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad S_2 = \begin{pmatrix} 96.0143 & 19.4603 \\ 19.4603 & 4.9857 \end{pmatrix},$$

so that $d_1 = d_2 = 100$ and $\varphi_0 = 202 (\approx 11.57$ degrees), which is a borderline case according to approximation (4.22). $G(\varphi)$ assumes two minima at .08 and .12, approximately. If we reduce the eccentricity to 90 (leaving φ_0 unchanged), we get the matrices

$$S_1 = \begin{pmatrix} 90 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad S_2 = \begin{pmatrix} 86.4168 & 17.4946 \\ 17.4946 & 4.5831 \end{pmatrix}.$$

For these two matrices, there is a unique minimum at $\varphi_0/2$. The bound (4.22) for d is in general too high, but the approximation becomes better when φ_0 gets smaller.

5. Remarks.

1. The proof of convergence of the G-algorithm makes strong use of the assumption that the matrices T_i are positive definite. If one or several of the matrices A_i are close to singularity, this could cause numerical problems, because the a_i (3.5) might become very large.

2. Since the stopping rule given in step F_3 depends on the absolute size of the matrices A_i, it may be better to replace it by a criterion similar to the one used in the G-algorithm:

F_3: If $\|B^{(f-1)} - B\| < \varepsilon_F$ for some small $\varepsilon_F > 0$, stop. Otherwise, start the next iteration step at F_1.

3. If the current version of B in the F-algorithm is a stationary point of Φ, and I_2 is taken as an initial approximation of Q in the G-algorithm, FG will not change B, since (1.14) is satisfied. This occurs, e.g., if the diagonal elements of the A_i-matrices are identical for each A_i, that is, diag $A_i = \text{diag}(c_i, \cdots, c_i)$ for some $c_i > 0 (i = 1, \cdots, k)$, and I_p is taken as an initial approximation of B. An important special case of this are correlation matrices, where the diagonal elements are all 1. If the first iteration of the F-algorithm does not change B, it might therefore be helpful to try FG with another initial approximation.

4. On the F-level, a better initial approximation than I_p might be to take the eigenvectors of one of the A_i (e.g. the one with the largest n_i), or the eigenvectors of $\sum_{i=1}^{k} n_i A_i$. On the G-level, I_2 is a good initial approximation for Q, when the current B on the F-level is already close to the correct solution.

5. In step F_{24}, the lth and jth column of B are adjusted using the matrix Q given by the G-algorithm. Since these two columns will undergo changes in subsequent executions of steps F_{21} to F_{24}, it is not necessary to iterate on the G-level until full convergence is reached. In most cases the first iteration steps of the G-algorithm will decrease $\Phi(B)$ much more than the later iterations. If $k = 1$, only one iteration step is required in each execution of the G-algorithm.

6. In order to avoid permutations of the columns of B and multiplications by -1, it is convenient to order the columns of Q such that

$$Q=\begin{pmatrix}\cos\alpha & -\sin\alpha\\ \sin\alpha & \cos\alpha\end{pmatrix} \tag{5.1}$$

where $-\pi/2<\alpha<\pi/2$.

7. If $k=1$, the FG-algorithm reduces to a Jacobi-method (Parlett (1980, Chap. 9)) for diagonalizing the single p.d.s. matrix $A=A_1$.

8. The listing of a FORTRAN program performing the FG-algorithm (Flury (1985)) can be obtained from the first author upon request.

6. Example. In this section we illustrate the performance of the FG-algorithm by a numerical example of dimension $p=6$ with $k=2$ matrices and weights $n_1=n_2=1$. The matrices are

$$A_1=\begin{pmatrix}45 & 10 & 0 & 5 & 0 & 0\\ 10 & 45 & 5 & 0 & 0 & 0\\ 0 & 5 & 45 & 10 & 0 & 0\\ 5 & 0 & 10 & 45 & 0 & 0\\ 0 & 0 & 0 & 0 & 16.4 & -4.8\\ 0 & 0 & 0 & 0 & -4.8 & 13.6\end{pmatrix},$$

$$A_2=\begin{pmatrix}27.5 & -12.5 & -.5 & -4.5 & -2.04 & 3.72\\ -12.5 & 27.5 & -4.5 & -.5 & 2.04 & -3.72\\ -.5 & -4.5 & 24.5 & -9.5 & -3.72 & -2.04\\ -4.5 & -.5 & -9.5 & 24.5 & 3.72 & 2.04\\ -2.04 & 2.04 & -3.72 & 3.72 & 54.76 & -4.68\\ 3.72 & -3.72 & -2.04 & 2.04 & -4.68 & 51.24\end{pmatrix}.$$

The characteristic vectors of A_1 are the columns of the matrix

$$B_1=\begin{pmatrix}.5 & .5 & .5 & .5 & 0 & 0\\ .5 & .5 & -.5 & -.5 & 0 & 0\\ .5 & -.5 & -.5 & .5 & 0 & 0\\ .5 & -.5 & .5 & -.5 & 0 & 0\\ 0 & 0 & 0 & 0 & .8 & .6\\ 0 & 0 & 0 & 0 & -.6 & .8\end{pmatrix}.$$

The associated roots are 60, 50, 40, 30, 20 and 10. For matrix A_2, the characteristic vectors are

$$B_2=\begin{pmatrix}.5 & .5 & .3 & -.6 & .1 & -.2\\ .5 & .5 & -.3 & .6 & -.1 & .2\\ .5 & -.5 & -.6 & -.3 & -.2 & -.1\\ .5 & -.5 & .6 & .3 & .2 & .1\\ 0 & 0 & -.18 & -.26 & .54 & .78\\ 0 & 0 & -.26 & .18 & .78 & -.54\end{pmatrix},$$

with roots 10, 20, 30, 40, 50 and 60. We used the FG-algorithm as programmed by Flury (1985) with $\varepsilon_F=\varepsilon_G=.0001$. The stopping rule for the F-algorithm was as described in Remark 2 above. As an initial approximation we used $B=I_6$, the identity matrix of dimension 6. The rotation pairs in the F-algorithm were chosen cyclically

(see Golub and Van Loan (1983, p. 299)). A sweep (or iteration step) of the F-algorithm consists therefore of $\binom{6}{2} = 15$ pairwise rotations. For each sweep of the F-algorithm, we give the current orthogonal matrix B, the value of the criterion $\Phi(B) = \prod_{i=1}^{2} |\mathrm{diag}(B^T A_i B)|/|A_i|$ and the average number of iterations of the G-algorithm per call. At the beginning, the value of the criterion is $\Phi(I_6) = 2.24718$.

after sweep 1

$$B^{(1)} = \begin{pmatrix} .8138 & -.0721 & .4584 & .3474 & -.0398 & .0124 \\ .0000 & .7646 & .4627 & -.4451 & .0553 & -.0114 \\ -.1321 & -.6346 & .5378 & -.5358 & -.0456 & -.0370 \\ -.5642 & 0384 & .5353 & .6256 & .0321 & .0358 \\ .0390 & -.0699 & .0002 & -.0357 & .7998 & .5938 \\ -.0224 & .0326 & .0003 & -.0379 & -.5937 & .8028 \end{pmatrix},$$

$\Phi(B^{(1)}) = 1.25461$

average number of iterations of G-algorithm: 2.73

after sweep 2

$$B^{(2)} = \begin{pmatrix} .4983 & -.5648 & .4993 & .4174 & -.0955 & .0072 \\ .5025 & .5613 & .4998 & -.4164 & .0955 & -.0072 \\ -.5003 & -.4124 & .5003 & -.5711 & -.0537 & -.0180 \\ -.4988 & .4150 & .5006 & .5703 & .0538 & .0180 \\ .0003 & -.1276 & -.0001 & -.0010 & .7921 & .5968 \\ .0003 & .0864 & .0000 & -.0323 & -.5903 & .8019 \end{pmatrix}$$

$\Phi(B^{(2)}) = 1.03574$

average number of iterations of G-algorithm: 2.4

after sweep 3

$$B^{(3)} = \begin{pmatrix} .5000 & -.5548 & .5000 & .4278 & -.0956 & .0083 \\ .5000 & .5548 & .5000 & -.4278 & .0956 & -.0083 \\ -.5000 & -.4247 & .5000 & -.5625 & -.0541 & -.0170 \\ -.5000 & .4247 & .5000 & .5625 & .0541 & .0170 \\ .0000 & -.1265 & .0000 & .0013 & .7919 & .5974 \\ .0000 & .0878 & .0000 & -.0336 & -.5905 & .8015 \end{pmatrix}$$

$\Phi(B^{(3)}) = 1.03568$

average number of iterations of G-algorithm: 1.8

after sweep 4

$$B^{(4)} = \begin{pmatrix} .5000 & -.5545 & .5000 & .4281 & -.0956 & .0083 \\ .5000 & .5545 & .5000 & -.4281 & .0956 & -.0083 \\ -.5000 & -.4250 & .5000 & -.5623 & -.0541 & -.0169 \\ -.5000 & .4250 & .5000 & .5623 & .0541 & .0169 \\ .0000 & -.1265 & .0000 & .0014 & .7919 & .5974 \\ .0000 & .0878 & .0000 & -.0337 & -.5906 & .8015 \end{pmatrix}$$

$\Phi(B^{(4)}) = 1.03568$

average number of iterations of G-algorithm: 1.07.

Sweep 5 did not produce any changes in the first four digits of the elements of B and the algorithm stopped. The "nearly diagonal" matrices B^TA_1B and B^TA_2B were given by the program as

$$B^TA_1B = \begin{pmatrix} 50.0000 & .0000 & .0000 & .0000 & .0000 & .0000 \\ .0000 & 29.9305 & .0000 & -1.2531 & 1.5738 & .0753 \\ .0000 & .0000 & 60.0000 & .0000 & .0000 & .0000 \\ .0000 & -1.2531 & .0000 & 39.7904 & -.6200 & .7728 \\ .0000 & 1.5738 & .0000 & -.6200 & 20.2584 & -.0351 \\ .0000 & .0753 & .0000 & .7728 & -.0351 & 10.0207 \end{pmatrix},$$

$$B^TA_2B = \begin{pmatrix} 20.0000 & .0000 & .0000 & .0000 & .0000 & .0000 \\ .0000 & 40.2336 & .0000 & -1.6232 & 3.1472 & 1.1790 \\ .0000 & .0000 & 10.0000 & .0000 & .0000 & .0000 \\ .0000 & -1.6232 & .0000 & 32.0055 & -1.0458 & 5.4272 \\ .0000 & 3.1472 & .0000 & -1.0458 & 59.2485 & .4738 \\ .0000 & 1.1790 & .0000 & 5.4272 & .4738 & 48.5123 \end{pmatrix}.$$

The FG-algorithm has clearly recovered the two common eigenvectors of A_1 and A_2. The four other columns of $B = B^{(4)}$ can be considered as "compromises" between eigenvectors of A_1 and A_2. Of course the order of the columns of B is not relevant; it is simply determined by the initial approximation used in the F-algorithm.

It is worth noting that the convergence is rather fast: after only two sweeps, the coefficients of B are already correct to two digits. This was typically also the case in statistical examples (see Flury (1984)), where the weights n_i are not necessarily equal. In none of these examples, more than five sweeps were needed to reach convergence.

The computation of the above example required .07 seconds of CPU time (not including input/output operations) on the CDC 170/855 computer of Indiana University.

REFERENCES

A. BASILEVSKY, (1983), *Applied Matrix Algebra in the Statistical Sciences*, North-Holland, New York.

B. N. FLURY, (1983), *Some relations between the comparison of covariance matrices and principal component analysis*, Computational Statistics and Data Analysis, 1, pp. 97–109.

——— (1984), *Common principal components in k groups*, J. Amer. Statist. Assoc, 79, pp. 892–898.

——— (1985), *The FG-algorithm*, accepted for publication in the algorithms' section of Applied Statistics.

G. H. GOLUB, AND C. F. VAN LOAN, (1983), *Matrix Computations*, Johns Hopkins Univ. Press, Baltimore.

A. E. HOERL, AND R. W. KENNARD, (1970), *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics, 12, pp. 55–67.

B. NOBLE, AND J. W. DANIEL, (1977), *Applied Linear Algebra*, Prentice-Hall. Englewood Cliffs, N.J.

B. N. PARLETT, (1980), *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ.

E. WEBER, (1974), *Einführung in die Faktorenanalyse*, Verlag Gustav Fischer, Stuttgart & New York.

30.3. [124] "A CLASS OF SLOWLY CONVERGENT SERIES AND THEIR SUMMATION BY GAUSSIAN QUADRATURE"

---

[124] "A Class of Slowly Convergent Series and Their Summation by Gaussian Quadrature," *Math. Comp*. **57**, 309–324 (1991).

MATHEMATICS OF COMPUTATION
VOLUME 57, NUMBER 195
JULY 1991, PAGES 309–325

A CLASS OF SLOWLY CONVERGENT SERIES AND THEIR SUMMATION BY GAUSSIAN QUADRATURE

WALTER GAUTSCHI

ABSTRACT. Series are considered whose general term is a rational function multiplied by a fractional power. The summation of such series is reduced, via Laplace transformation techniques, to a problem of quadrature, which is then solved by Gaussian quadrature relative to Einstein and Fermi weight functions. A number of examples are worked out in detail.

1. INTRODUCTION

We consider series of the type

$$S_0 = \sum_{k=1}^{\infty} k^{\nu-1} r(k) \tag{1.0}$$

or

$$S_1 = \sum_{k=1}^{\infty} (-1)^{k-1} k^{\nu-1} r(k), \tag{1.1}$$

where $0 < \nu \leq 1$ and $r(\cdot)$ is a rational function

$$r(s) = \frac{p(s)}{q(s)} \tag{1.2}$$

with p, q real polynomials of degrees $\deg p \leq \deg q$. Strict inequality is assumed when necessary for convergence. It is further assumed that the zeros of q all have nonpositive real parts:

$$\text{if } q(-a) = 0 \quad \text{then } \operatorname{Re} a \geq 0. \tag{1.3}$$

This condition can always be achieved by a preliminary summation of a few initial terms.

The problem can be simplified by first obtaining the partial fraction decomposition of r,

$$r(s) = \sum_{\rho} \sum_{m=1}^{m_\rho} c_{\rho m} (s + a_\rho)^{-m} + \sum_{\gamma} \sum_{m=1}^{m_\gamma} [c_{\gamma m} (s + a_\gamma)^{-m} + \overline{c}_{\gamma m} (s + \overline{a}_\gamma)^{-m}], \tag{1.4}$$

Received March 7, 1990; revised September 17, 1990.
1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 40A25; Secondary 44A10, 65D30, 33A65.
Key words and phrases. Slowly convergent series, Laplace transformation, summation by quadrature, Gaussian quadrature, orthogonal polynomials.
Work supported, in part, by the National Science Foundation under grant CCR-8704404.

where the first sum is over all real zeros $(-a_\rho)$ of q (with multiplicities m_ρ), and the second sum is over all pairs of conjugate complex zeros $(-a_\gamma, -\overline{a}_\gamma)$ (with multiplicities m_γ). The coefficients $c_{\rho m}$ in the first sum are real, those in the second complex, in general. (We have assumed in (1.4) that $\deg p < \deg q$. If $\deg p = \deg q$, and we are thus dealing with S_1 in (1.1), there will be an additional constant term in (1.4). Its contribution to the series in (1.1) is expressible in terms of the Riemann zeta function; cf. (3.9) below for $m = 0$.) Once the decomposition (1.4) has been obtained (for relevant constructive methods, see, e.g., [8, §7.1]), it clearly suffices to consider

$$r(s) = \frac{1}{(s+a)^m}, \qquad \operatorname{Re} a \geq 0, \quad m \geq 1. \tag{1.5}$$

Without restriction of generality, it may be further assumed that $\operatorname{Im} a \geq 0$.

By interpreting the terms in the series (1.0) and (1.1) as Laplace transforms at integer values, it is possible to express the sum of the series as a weighted integral over $\mathbb{R}_+$ of certain special functions related to the incomplete gamma function. The weighting involves the product of a fractional power and either Einstein's function $t(e^t - 1)^{-1}$ (in the case of (1.0)), or Fermi's function $(e^t + 1)^{-1}$ (in the case of (1.1)), both having infinitely many poles on the imaginary axis of the complex plane. Properties of the required special functions are briefly developed in §2. Section 3 discusses the summation of (1.0), (1.1) via Gaussian quadrature. The case $\nu = 1$ of purely rational series is treated in §4 and complements more traditional techniques (e.g., those in [8, §7.2II]). Numerical examples for the case $\nu = \frac{1}{2}$ are given in §5, where also comparisons are made with direct summation and accelerated summation using the ε-algorithm.

2. Preliminaries

Define, for $t > 0$,

$$g_{-1}(t; \nu) = g_{-1}(t) = \frac{t^{-1}}{\Gamma(1-\nu)}, \qquad 0 < \nu < 1, \tag{2.1$_{-1}$}$$

and for $t > 0$, $n = 0, 1, 2, \ldots$,

$$g_n(t; a, \nu) = g_n(t) = \frac{e^{-at} t^{\nu-1}}{n!\Gamma(1-\nu)} \int_0^t e^{a\tau} (t-\tau)^n \tau^{-\nu} \, d\tau, \tag{2.1$_n$}$$
$$\operatorname{Re} a \geq 0, \quad \operatorname{Im} a \geq 0, \quad a \neq 0, \quad 0 < \nu < 1.$$

Lemma 2.1. *We have*

$$g_0(t; a, \nu) = e^{-at} \gamma^*(1-\nu, -at), \tag{2.2}$$

where γ^* is Tricomi's form of the incomplete gamma function (cf. [3, eq. 6.5.4]*). Furthermore,*

$$g_{n+1}(t) = \frac{1}{n+1} \left\{ \left(t + \frac{n+1-\nu}{a} \right) g_n(t) - \frac{t}{a} g_{n-1}(t) \right\}, \tag{2.3}$$
$$n = 0, 1, 2, \ldots.$$

Proof. By definition,

$$g_0(t) = \frac{e^{-at}t^{\nu-1}}{\Gamma(1-\nu)}\int_0^t e^{a\tau}\tau^{-\nu}\,d\tau,$$

which, upon the change of variables $u = -a\tau$, gives

$$\begin{aligned} g_0(t) &= \frac{e^{-at}(-at)^{\nu-1}}{\Gamma(1-\nu)}\int_0^{-at} e^{-u}u^{-\nu}\,du \\ &= \frac{e^{-at}(-at)^{-(1-\nu)}}{\Gamma(1-\nu)}\gamma(1-\nu, -at) = e^{-at}\gamma^*(1-\nu, -at) \end{aligned}$$

(cf. [3, eqs. 6.5.2, 6.5.4]). Next,

$$\begin{aligned} g_{n+1}(t) &= \frac{e^{-at}t^{\nu-1}}{(n+1)!\Gamma(1-\nu)}\int_0^t e^{a\tau}(t-\tau)^n(t-\tau)\tau^{-\nu}\,d\tau \\ &= \frac{t}{n+1}g_n(t) - \frac{e^{-at}t^{\nu-1}}{(n+1)!\Gamma(1-\nu)}\int_0^t e^{a\tau}(t-\tau)^n\tau^{1-\nu}\,d\tau. \end{aligned}$$

To the last integral we apply integration by parts, letting $u = \tau^{1-\nu}(t-\tau)^n$, $v' = e^{a\tau}$, hence

$$\begin{aligned} u' &= (1-\nu)\tau^{-\nu}(t-\tau)^n - n\tau^{1-\nu}(t-\tau)^{n-1} \\ &= (n+1-\nu)(t-\tau)^n\tau^{-\nu} - nt(t-\tau)^{n-1}\tau^{-\nu}, \\ v &= \frac{1}{a}e^{a\tau}. \end{aligned}$$

This yields (2.3) for $n \geq 1$. A similar calculation gives

$$g_1(t) = \left(t + \frac{1-\nu}{a}\right)g_0(t) - \frac{1}{a\Gamma(1-\nu)},$$

which, in view of (2.1_{-1}), shows that (2.3) holds also for $n = 0$. □

To the author's knowledge, no software seems to be readily available for computing the incomplete gamma function in the domain of interest here (left half plane), and one thus has to rely on standard techniques such as power series and asymptotic expansions. An effort of developing good software for Tricomi's function $\gamma^*(1-\nu, z)$, $0 < \nu < 1$, applicable for arbitrary complex z, would certainly be worthwhile. For the case $\nu = \frac{1}{2}$, however, see §5.

3. Summation of S_0 and S_1

We employ a technique already used in [7], namely, to interpret the general term of the series as a Laplace transform and thereby converting the series into a suitably weighted integral. We first treat the series S_0 in (1.0).

Assume $r(\cdot)$ given as in (1.5), and consider first the case $a \neq 0$. Then

$$(3.1) \qquad s^{\nu-1}\cdot\frac{1}{(s+a)^m} = \mathscr{L}\left\{\frac{t^{-\nu}}{\Gamma(1-\nu)} * \frac{t^{m-1}e^{-at}}{(m-1)!}\right\}, \qquad 0 < \nu < 1,$$

where $\mathscr{L}\{\cdot\}$ is the Laplace transform and $*$ means convolution. We thus have

$$f(t) := \mathscr{L}^{-1}\left\{\frac{s^{\nu-1}}{(s+a)^m}\right\} = \frac{1}{(m-1)!\Gamma(1-\nu)}\int_0^t e^{-a(t-\tau)}(t-\tau)^{m-1}\tau^{-\nu}\,d\tau$$

$$= \frac{e^{-at}}{(m-1)!\Gamma(1-\nu)}\int_0^t e^{a\tau}(t-\tau)^{m-1}\tau^{-\nu}\,d\tau \tag{3.2}$$

$$= t^{1-\nu}g_{m-1}(t;a,\nu), \qquad m \geq 1,$$

where $g_n(t;a,\nu)$ is defined in (2.1_n). There follows

$$\sum_{k=1}^{\infty}\frac{k^{\nu-1}}{(k+a)^m} = \sum_{k=1}^{\infty}(\mathscr{L}f)(k) = \sum_{k=1}^{\infty}\int_0^{\infty}e^{-kt}f(t)\,dt$$

$$= \sum_{k=1}^{\infty}\int_0^{\infty}e^{-t}t^{1-\nu}e^{-(k-1)t}g_{m-1}(t;a,\nu)\,dt$$

$$= \int_0^{\infty}e^{-t}t^{1-\nu}\sum_{k=1}^{\infty}e^{-(k-1)t}g_{m-1}(t;a,\nu)\,dt$$

$$= \int_0^{\infty}t^{1-\nu}\frac{1}{e^t-1}g_{m-1}(t;a,\nu)\,dt.$$

Thus,

$$\sum_{k=1}^{\infty}\frac{k^{\nu-1}}{(k+a)^m} = \int_0^{\infty}t^{-\nu}\varepsilon(t)g_{m-1}(t;a,\nu)\,dt, \qquad m \geq 1,\ 0<\nu<1, \tag{3.3}$$

where $\varepsilon(\cdot)$ is "Einstein's function" (cf. [7]),

$$\varepsilon(t) = \frac{t}{e^t-1}, \qquad 0 \leq t < \infty. \tag{3.4}$$

Since $g_{m-1}(\cdot;a,\nu)$, by Lemma 2.1, is an entire function, formula (3.3) suggests to apply Gaussian quadrature to the integral on the right, using $w(t) = t^{-\nu}\varepsilon(t)$ as a weight function on $[0,\infty]$. The required orthogonal polynomials can be computed by techniques already discussed in [7] (see, in particular, (2.3) and Example 4.4 of that paper). Gauss quadrature will converge quite rapidly, unless $\operatorname{Re} a$ and/or $\operatorname{Im} a$ is large, in which case "stratified" summation can be employed to regain rapid convergence (see Examples 5.1 and 5.4). For $\nu = \frac{1}{2}$, the first 80 recursion coefficients for the required orthogonal polynomials are given to 25 significant digits in Table 1 of the Appendix.

It is easily seen that (3.3) holds also for $a = 0$ if we define

$$g_0(t) = \frac{1}{\Gamma(2-\nu)}, \qquad g_{n+1}(t) = \frac{t}{n+2-\nu}g_n(t), \tag{3.5}$$

$$n = 0, 1, 2, \ldots \ (a = 0).$$

The series can then be expressed in terms of Riemann's zeta function,

$$\sum_{k=1}^{\infty}k^{-(m+1-\nu)} = \zeta(m+1-\nu), \tag{3.6}$$

and since g_{m-1} is a monomial of degree $m-1$, the n-point Gauss formula for the integral in (3.3) gives exact answers (modulo rounding) if $n = \lfloor (m+1)/2 \rfloor$.

For the sum S_1 in (1.1), a calculation similar to the one which led to (3.3) now yields, for $a \neq 0$,

$$\sum_{k=1}^{\infty} (-1)^{k-1} \frac{k^{\nu-1}}{(k+a)^m} = \int_0^{\infty} t^{-\nu} \varphi(t) \cdot t g_{m-1}(t; a, \nu)\, dt, \tag{3.7}$$

$$m \geq 0, \quad 0 < \nu < 1,$$

where $\varphi(\cdot)$ is the "Fermi weight function" (cf. [7]),

$$\varphi(t) = \frac{1}{e^t + 1}, \qquad 0 \leq t < \infty. \tag{3.8}$$

The result (3.7), as noted, holds also for $m = 0$, if g_{-1} is defined as in (2.1_{-1}). If $a = 0$, then (3.7) holds with g_{m-1}, $m \geq 0$, defined in (2.1_{-1}) and (3.5), and represents the series

$$\sum_{k=1}^{\infty} (-1)^{k-1} k^{-(m+1-\nu)} = (1 - 2^{-m+\nu}) \zeta(m+1-\nu). \tag{3.9}$$

Again, Gauss quadrature for the integral in (3.7), in this case, is exact if we take $n = \lfloor (m+2)/2 \rfloor$ points.

The first 80 recursion coefficients for the orthogonal polynomials with respect to the weight function $t^{-\nu}\varphi(t)$, $\nu = \frac{1}{2}$, are listed in Table 2 of the Appendix.

Remarks. 1. The fractional power $k^{\nu-1}$ in (1.0), (1.1) can easily be generalized to $(k+b)^{\nu-1}$, $\operatorname{Re} b \geq 0$, since this only introduces a factor e^{-bt} in (3.1) and gives $f(t) = t^{1-\nu} e^{-bt} g_{m-1}(t; a-b, \nu)$ in place of (3.2), hence

$$\sum_{k=1}^{\infty} \frac{(k+b)^{\nu-1}}{(k+a)^m} = \int_0^{\infty} t^{-\nu} \varepsilon(t) e^{-bt} g_{m-1}(t; a-b, \nu)\, dt, \tag{3.10}$$

and the similarly generalized version of (3.7),

$$\sum_{k=1}^{\infty} (-1)^{k-1} \frac{(k+b)^{\nu-1}}{(k+a)^m} = \int_0^{\infty} t^{-\nu} \varphi(t) \cdot t e^{-bt} g_{m-1}(t; a-b, \nu)\, dt. \tag{3.11}$$

2. As an alternative to Gauss quadrature with weight function $t^{-\nu}\varepsilon(t)$, one could write the integral in (3.3) as

$$\int_0^{\infty} t^{-\nu} \varepsilon(t) g_{m-1}(t; a, \nu)\, dt = \int_0^{\infty} t^{-\nu} e^{-t} \cdot \frac{t}{1 - e^{-t}} g_{m-1}(t; a, \nu)\, dt \tag{3.12}$$

and apply Gauss-Laguerre quadrature to the integral on the right. The poles that were previously incorporated in the weight function $t^{-\nu}\varepsilon(t)$ then, however, become part of the integrand, which may adversely affect the speed of convergence of the quadrature scheme. This was indeed found to be the case when a is relatively small, but for larger values of a, in particular when used

in conjunction with stratified summation (cf. Examples 5.1 and 5.4), the Gauss-Laguerre method is competitive and, in Example 5.1, even more efficient.

The same alternative approach is possible for the integral in (3.7), if written as

$$\int_0^\infty t^{-\nu}\varphi(t)\cdot t g_{m-1}(t;a,\nu)\,dt = \int_0^\infty t^{-\nu}e^{-t}\cdot\frac{t}{1+e^{-t}}g_{m-1}(t;a,\nu)\,dt, \tag{3.13}$$

although this time the poles affect convergence more severely, since they are half as close to the real axis than before. Still, for sufficiently large values of a, Gauss-Laguerre quadrature here, too, gains the upper hand.

Finally, if many different values of ν were contemplated, then Gauss-Laguerre would also be preferable, since the recursion coefficients for the respective orthogonal polynomials—the generalized Laguerre polynomials—are then explicitly known and need not be tabulated. In the case of the weight function $w(t) = t^{-\nu}e^{-t}$ on $(0,\infty)$, the corresponding (monic) orthogonal polynomials $\pi_k(\cdot) = \pi_k(\cdot; w)$ indeed satisfy

$$\begin{aligned} \pi_{k+1}(t) &= (t-\alpha_k)\pi_k(t) - \beta_k\pi_{k-1}(t), \qquad k = 0,1,2,\dots, \\ \pi_{-1}(t) &= 0, \qquad \pi_0(t) = 1, \end{aligned} \tag{3.14}$$

with the coefficients $\alpha_k = \alpha_k(w)$, $\beta_k = \beta_k(w)$ $[\beta_0(w) = \int_0^\infty w(t)\,dt]$ having the particularly simple form

$$\begin{aligned} &\alpha_k(w) = 2k+1-\nu, \qquad k \geq 0; \\ &\beta_0(w) = \Gamma(1-\nu), \qquad \beta_k(w) = k(k-\nu), \quad k \geq 1 \end{aligned} \qquad (w = t^{-\nu}e^{-t}). \tag{3.15}$$

4. Series of purely rational terms

So far, we assumed that $0 < \nu < 1$ in (1.0) and (1.1). The same techniques, however, are applicable when $\nu = 1$, i.e., for series

$$S_0' = \sum_{k=1}^\infty r(k), \qquad S_1' = \sum_{k=1}^\infty (-1)^{k-1}r(k). \tag{4.1}$$

One finds, when

$$r(s) = \frac{1}{(s+a)^m}, \qquad \operatorname{Re} a \geq 0, \tag{4.2}$$

for $m \geq 2$ that

$$\sum_{k=1}^\infty \frac{1}{(k+a)^m} = \frac{1}{(m-1)!}\int_0^\infty \varepsilon(t)t^{m-2}e^{-at}\,dt, \tag{4.3}$$

and for $m \geq 1$ that

$$\sum_{k=1}^\infty \frac{(-1)^{k-1}}{(k+a)^m} = \frac{1}{(m-1)!}\int_0^\infty \varphi(t)t^{m-1}e^{-at}\,dt, \tag{4.4}$$

with ε and φ as given in (3.4) and (3.8), respectively.

For more general rational functions

$$r(s) = \sum_i \sum_{m=1}^{m_i} c_{im}(s + a_i)^{-m}, \tag{4.5}$$

we get from (4.3)

$$\sum_{k=1}^{\infty} r(k) = \int_0^{\infty} \varepsilon(t) g(t)\, dt, \tag{4.6}$$

where

$$g(t) = \sum_i \sum_{m=1}^{m_i} \frac{c_{im}}{(m-1)!} t^{m-2} e^{-a_i t}. \tag{4.7}$$

Evidently, g is again an entire function, provided

$$\sum_i c_{i1} = 0, \tag{4.8}$$

which is required for the series in (4.6) to converge. Similarly,

$$\sum_{k=1}^{\infty} (-1)^{k-1} r(k) = \int_0^{\infty} \varphi(t) h(t)\, dt, \tag{4.9}$$

where

$$h(t) = \sum_i \sum_{m=1}^{m_i} \frac{c_{im}}{(m-1)!} t^{m-1} e^{-a_i t}, \tag{4.10}$$

an entire function for any choice of the coefficients c_{im}.

The first 40 recursion coefficients for the polynomials orthogonal with respect to ε and φ can be found to 25 significant digits in [7, Appendices A1 and A2].

The method described in this section provides an alternative to other summation/integration methods, such as those discussed in [8, §7.2II]. An advantage of the present method is that it leads to Gaussian quadrature of *entire* functions, a possible complication, that the interval of integration is infinite.

5. Examples

In all of our examples we take $\nu = \frac{1}{2}$. In this case, the function g_0 in (2.2) is given by (cf. [3, eq. 6.5.18])

$$g_0\left(t; a, \frac{1}{2}\right) = e^{-at} \gamma^*\left(\frac{1}{2}, -at\right) = \frac{2}{\sqrt{\pi}} \frac{F(\sqrt{at})}{\sqrt{at}}, \tag{5.1}$$

where F is Dawson's integral,

$$F(z) = e^{-z^2} \int_0^z e^{t^2}\, dt. \tag{5.2}$$

For real z, this can be evaluated with an accuracy of up to about 20 significant digits, using the rational Chebyshev approximations given in [2].

All computations reported below were done in double precision on the Cyber 205 computer (the equivalent of about 29 decimal places).

Example 5.1. $S_0 = \sum_{k=1}^{\infty} k^{-1/2}/(k+a)^m$.

This series with $a = m = 1$ was communicated to the author by Professor P. J. Davis, who encountered it in his study of spirals [4].

We computed S_0 from (3.3) (with $\nu = \frac{1}{2}$), (5.1), and (2.1_{-1}), (2.3), for $a = .5, 1., 2., 4., 8.$ and $m = 1(1)5$. The integral in (3.3) was evaluated by n-point Gaussian quadrature rules, which were generated with the help of the recursion coefficients in Table 1 of the Appendix and well-known eigenvalue techniques (see, e.g., [6, §1.3(iv)]). In Table 5.1 we show only the results for $m = 1$; those for $m > 1$ are similar, but exhibit somewhat slower convergence. It is evident that, as a increases, convergence of the Gauss quadrature formula slows down considerably. The reason for this is the behavior of the function g_0 on the right of (5.1), which for increasing a approaches a discontinuous function (see Figure 5.1).

TABLE 5.1

n-point quadrature approximations to the integral in (3.3) *with* $\nu = \frac{1}{2}$, $m = 1$, $a = .5, 1., 2., 4., 8.$

n	$a = .5$	$a = 1.$	$a = 2.$
5	2.1344163	1.8599	1.537
10	2.1344166429861	1.860025078	1.53967
15	2.1344166429862372611	1.86002507922117	1.539680509
20	2.1344166429862372611	1.860025079221190306	1.539680512350
25		1.860025079221903071	1.53968051235329
30		1.860025079221903072	1.539680512353302010
35			1.5396805123533020128
40			1.5396805123533020128

n	$a = 4.$	$a = 8.$
5	1.19	.8
10	1.217	.91
15	1.21826	.930
20	1.218273	.9312
25	1.218274011	.93135
30	1.21827401461	.931371
35	1.218274014668	.9313727
40	1.2182740146698	.93137291

To achieve better accuracy, when a is large, we proceed as follows. With $a_0 = \lfloor a \rfloor$ denoting the largest integer $\leq a$, and $a = a_0 + a_1$, where $a_0 \geq 1$, $0 \leq a_1 < 1$, the summation over all $k \geq 1$ may be "stratified" by letting $k = \lambda + \kappa a_0$ and summing over all $\kappa \geq 0$ for $\lambda = 1, 2, \ldots, a_0$. Thus,

$$S_0 = \sum_{k=1}^{\infty} \frac{k^{-1/2}}{(k + a_0 + a_1)^m} = \sum_{\lambda=1}^{a_0} \sum_{\kappa=0}^{\infty} \frac{(\lambda + \kappa a_0)^{-1/2}}{(\lambda + \kappa a_0 + a_0 + a_1)^m}$$

$$= a_0^{-(m+1/2)} \sum_{\lambda=1}^{a_0} \sum_{\kappa=0}^{\infty} \frac{(\kappa + \lambda/a_0)^{-1/2}}{(\kappa + 1 + (\lambda + a_1)/a_0)^m} \tag{5.3}$$

$$= a_0^{-(m+1/2)} \sum_{\lambda=1}^{a_0} \left\{ \sum_{\kappa=1}^{\infty} \frac{(\kappa + \lambda/a_0)^{-1/2}}{(\kappa + 1 + (\lambda + a_1)/a_0)^m} + \frac{(\lambda/a_0)^{-1/2}}{(1 + (\lambda + a_1)/a_0)^m} \right\}.$$

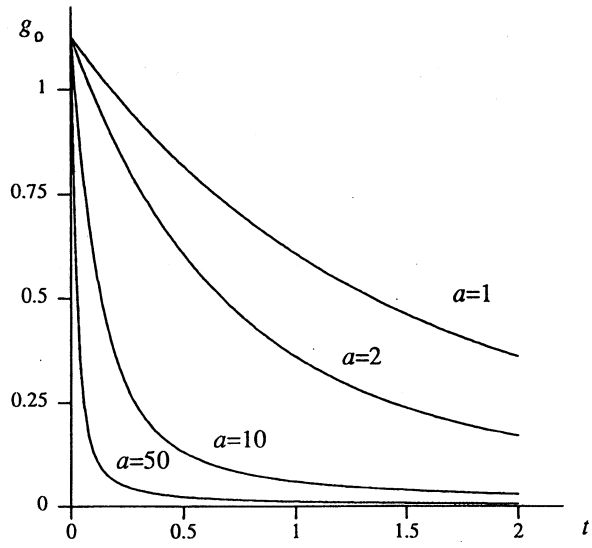


FIGURE 5.1
The function g_0 *in* (5.1)

To the inner sum we now apply (3.10) to obtain

$$\sum_{\kappa=1}^{\infty} \frac{(\kappa+\lambda/a_0)^{-1/2}}{(\kappa+1+(\lambda+a_1)/a_0)^m} = \int_0^{\infty} t^{-1/2}\varepsilon(t)e^{-(\lambda/a_0)t} g_{m-1}\left(t;\, 1+\frac{a_1}{a_0},\, \frac{1}{2}\right) dt. \tag{5.4}$$

The "effective" parameter in g_{m-1} is now $1 + a_1/a_0$, a number close to 1, and the coefficient λ/a_0 in the exponential is bounded by 1. Gauss quadrature applied to the integral in (5.4) should therefore converge quite rapidly, indeed, more so the larger a_0! This is borne out by the results displayed for $m = 1$ and $a = a_0 = 8., 16., 32.$ in Table 5.2.

The correct number of significant decimal digits produced by direct summation of the series, using 1000 terms, is shown in Table 5.3. The numbers in parentheses are the correct digits obtained by applying the ε-algorithm with the same number of terms. For the entries marked by an asterisk, we used (5.3),

TABLE 5.2
Approximations to S_0 *in* (5.3) *for* $m = 1$, $a_1 = 0$, $a_0 = 8., 16., 32.$, *using n-point quadrature in* (5.4)

n	$a = 8.$	$a = 16.$	$a = 32.$
5	.93098	.6946	.5097
10	.9313726	.6949315	.5099264
15	.931372933	.6949317145	.5099265169
20	.9313729340028	.6949317146409	.5099265170271
25	.9313729340031036	.6949317146410454	.5099265170272112
30	.9313729340031038714	.6949317146410455900	.5099265170272113479
35	.9313729340031038716 9	.6949317146410455901 6	.5099265170272113480 4
40	.9313729340031038716 9	.6949317146410455901 6	.5099265170272113480 4

TABLE 5.3
Number of correct significant decimal digits in direct (and accelerated) summation of S_0 *using 1000 terms*

m	$a = .5$	$a = 1.$	$a = 2.$	$a = 4.$	$a = 8.$
1	2(2)	1(2)	2(2)	2(2)	1(1)
2	4(5)	4(5)	4(5)	4(5)	3(4)
3	7(9)	7(9)	6(8)	6(8)	5(7)*
4	10(13)	9(12)	9(12)	8(11)*	8(10)*
5	14(17)	13(15)	11(14)	11(13)*	10(12)*

(5.4) to verify the number of correct digits. As is evident from Table 5.3, the ε-algorithm is only marginally effective on this particular series.

Example 5.2. $S_1 = \sum_{k=1}^{\infty}(-1)^{k-1}k^{-1/2}/(k+a)^m$.

We now apply Gauss quadrature (obtained from the recursion coefficients in Table 2 of the Appendix) relative to the weight function $t^{-1/2}\varphi(t)$ to the integral in (3.7) (with $\nu = \frac{1}{2}$), using the same values of a and m as in Example 5.1. The results are similar to those in Table 5.1 of Example 5.1, except that convergence is slightly slower. Stratified summation similar to (5.3), (5.4), on the other hand, gives

$$S_1 = a_0^{-(m+1/2)} \sum_{\lambda=1}^{a_0} (-1)^{\lambda} \left\{ s_\lambda - \frac{(\lambda/a_0)^{-1/2}}{(1+(\lambda+a_1)/a_0)^m} \right\}, \tag{5.5}$$

where

$$s_\lambda = \begin{cases} -\sum_{\kappa=1}^{\infty} \dfrac{(\kappa+\lambda/a_0)^{-1/2}}{(\kappa+1+(\lambda+a_1)/a_0)^m} & \text{if } a_0 \text{ is even}, \\ \sum_{\kappa=1}^{\infty} \dfrac{(-1)^{\kappa-1}(\kappa+\lambda/a_0)^{-1/2}}{(\kappa+1+(\lambda+a_1)/a_0)^m} & \text{if } a_0 \text{ is odd}, \end{cases} \tag{5.6}$$

that is, by (3.10) and (3.11),

$$s_\lambda = \begin{cases} -\int_0^{\infty} t^{-1/2}\varepsilon(t)e^{-(\lambda/a_0)t} g_{m-1}(t;\, 1+a_1/a_0,\, 1/2)\,dt, & a_0 \text{ even}, \\ \int_0^{\infty} t^{-1/2}\varphi(t)\cdot te^{-(\lambda/a_0)t} g_{m-1}(t;\, 1+a_1/a_0,\, 1/2)\,dt, & a_0 \text{ odd}. \end{cases} \tag{5.7}$$

The use of (5.5), with n-point quadrature applied to (5.7), yields results converging at the same speed as those in Table 5.2.

Direct summation using 1000 terms yields 2–3 more correct digits than in Table 5.3, but the ε-algorithm is now surprisingly effective, giving full accuracy (20 decimals) with only 23–27 terms!

Example 5.3. $S_0(b) = \sum_{k=1}^{\infty}(k+b)^{-1/2}/(k+b+1)$, $0 \le b < 1$.

This is another series of interest in P. J. Davis's work on spirals [4]. It can be readily evaluated with the help of Remark 1 in §3, taking $m = 1$, $a = b+1$ in (3.10), and noting (5.1); one finds

$$S_0(b) = \int_0^{\infty} t^{-1/2}\varepsilon(t)e^{-bt}g_0\left(t; 1, \frac{1}{2}\right)dt$$
$$= \frac{2}{\sqrt{\pi}}\int_0^{\infty} e^{-bt}\frac{F(\sqrt{t})}{\sqrt{t}}\cdot t^{-1/2}\varepsilon(t)\,dt. \tag{5.8}$$

Thus, $S_0(b)$ is the Laplace transform of the function $(F(\sqrt{t})/\sqrt{t})t^{-1/2}\varepsilon(t)$, which has a square root singularity at the origin and poles at integer multiples of $2\pi i$. The last integral in (5.8) is easily computed by Gaussian quadrature relative to the weight function $t^{-1/2}\varepsilon(t)$. No more than 24 quadrature points are needed to get 14 correct significant decimal digits in the range $0 < b < 1$.

Example 5.4. $S_0 = \sum_{k=1}^{\infty} k^{-1/2}/(k+i\alpha)$, $\alpha > 0$.

Here, $a = i\alpha$, $m = 1$, hence

$$S_0 = \int_0^{\infty} t^{-1/2}\varepsilon(t)g_0\left(t; i\alpha, \frac{1}{2}\right)dt. \tag{5.9}$$

Use of (5.1) and a simple change of variables gives

$$g_0\left(t; i\alpha, \frac{1}{2}\right) = \frac{ie^{-z^2}}{z}\operatorname{erf}(-iz), \qquad z = \sqrt{i\alpha t},$$

where erf is the error function. Letting $-iz = \frac{1}{2}\sqrt{\pi}(1-i)x$, i.e., $x = \sqrt{2\alpha t/\pi}$, one finds (cf. [5, eq. 7.3.22])

$$g_0(t; i\alpha, \tfrac{1}{2}) = \sqrt{2/(\alpha t)}e^{-i\alpha^2 t^2}[C(\sqrt{2\alpha t/\pi}) + iS(\sqrt{2\alpha t/\pi})]. \tag{5.10}$$

Here, $C(x)$, $S(x)$ denote the Fresnel integrals [5, eqs. 7.3.1, 7.3.2]. They can be computed with an accuracy of up to 18 significant digits from the rational Chebyshev approximations provided (on microfiche cards) in [1]. Gaussian quadrature of (5.9), with g_0 given by (5.10), yields the results shown in Table 5.4. For each n, the first entry is the real part, the second the imaginary part of the Gauss approximation to S_0. Convergence is seen to deteriorate rapidly with increasing α, which is to be expected in view of the highly oscillatory behavior of g_0 in (5.10) when α is large. The device of stratified summation, used successfully in Examples 5.1 and 5.2, however, can also be applied here,

TABLE 5.4

n-point quadrature approximations to the integral in (3.3) *with* $\nu = \frac{1}{2}$, $m = 1$, $a = i\alpha$, $\alpha = .5, 1., 2., 4., 8.$

n	$\alpha = .5$	$\alpha = 1.$
10	2.382181322854	2.006153
	−.564259325223	−.7964883
20	2.38218132285517164	2.0061526552273
	−.56425932522086830	−.79648812356982
40	2.38218132285517164	2.00615265522741423
	−.56425932522086830	−.79648812356984801
80		2.00615265522741424
		−.79648812356984802

n	$\alpha = 2.$	$\alpha = 4.$	$\alpha = 8.$
10	1.52	1.14	1.0
	−.846	−.91	−.3
20	1.51822	1.10	.9
	−.84397	−.72	−.7
40	1.518231590364	1.0978	.73
	−.843981047692	−.745	−.54
80	1.51823159036615356	1.0976938	.77
	−.84398104769701668	−.7460348	−.599

and gives, with $\alpha_0 = \lfloor \alpha \rfloor$, $\alpha = \alpha_0 + \alpha_1$, $\alpha_0 \geq 1$, $0 \leq \alpha_1 < 1$,

$$S_0 = \alpha_0^{-3/2} \sum_{\lambda=1}^{\alpha_0} \left\{ \int_0^\infty t^{-1/2} \varepsilon(t) e^{-(\lambda/\alpha_0)t} g_0\left(t; i\left(1 + \frac{\alpha_1}{\alpha_0}\right), \frac{1}{2}\right) dt + \frac{(\alpha_0/\lambda)^{3/2}}{1 + i(\alpha_0 + \alpha_1)/\lambda} \right\}. \tag{5.11}$$

Using (5.11) instead of (5.9) yields the results shown in Table 5.5.

Direct summation using 1000 terms gives 1–2 correct decimal digits in the real part, and 4–5 in the imaginary part. The epsilon algorithm produces no more than one additional correct digit.

Example 5.5. $S_1 = \sum_{k=1}^\infty (-1)^{k-1} k^{-1/2}/(k + i\alpha)$, $\alpha > 0$.

Applying Gauss quadrature to

$$S_1 = \int_0^\infty t^{-1/2} \varphi(t) \cdot t g_0(t; i\alpha, \tfrac{1}{2})\, dt, \tag{5.12}$$

with g_0 as in (5.10), or to formulae analogous to those in (5.5)–(5.7), yields results similar in quality to those in Table 5.4 (but converging at a slightly slower rate) and to those in Table 5.5. As in Example 5.2, here too, the ε-algorithm produces full accuracy (20 decimals) with as few as 25–28 terms.

TABLE 5.5

Approximations to S_0 *for* $\alpha_1 = 0$, $\alpha_0 = 8., 16., 32.$, *using n-point quadrature for the integral in* (5.11)

n	$\alpha = 8.$	$\alpha = 16.$	$\alpha = 32.$
5	.78217	.55456	.39250
	−.6028	−.46405	−.34703
10	.78214786	.554548189	.392496065
	−.60290377	−.464094441	−.347063784
15	.782147849839	.554548181558	.392496059680
	−.602903762412	−.464094436689	−.347063781179
20	.7821478498420740	.5545481815605363	.3924960596818862
	−.6029037624091237	−.46409443668759260	−.3470637811774942
25	.78214784984207490	.55454818156053686	.39249605968188663
	−.60290376240912468	−.46409443668759260	−.34706378117749456
30	.78214784984207491	.55454818156053686	.39249605968188663
	−.60290376240912469	−.46409443668759260	−.34706378117749456

ACKNOWLEDGMENT

The stimulus to write this paper came from Professor P. J. Davis who showed the author the two series in Example 5.1 (with $a = m = 1$) and Example 5.3 and their interesting connections with analytic properties of certain spirals. The author thanks Professor Davis for bringing these matters to his attention.

APPENDIX

Coefficients α_k, β_k in the recurrence relation (3.14) for the (monic) polynomials $\pi_k(\cdot; w_1)$ and $\pi_k(\cdot; w_2)$ orthogonal on $[0, \infty]$ with respect to the weight functions $w_1(t) = t^{-1/2}\varepsilon(t)$ and $w_2(t) = t^{-1/2}\varphi(t)$, where ε and φ are the Einstein and Fermi functions, respectively.

TABLE 1

Recursion coefficients for the polynomials $\{\pi_k(\cdot; w_1)\}$

k	alpha(k)	beta(k)
0	0.77026867019278179736191 58D+00	0.23151573733941170004258 19D+01
1	0.31875985567615246793664 14D+01	0.10240846879834073034233 87D+01
2	0.52637469230456075097135 90D+01	0.44142904264700503516592 16D+01
3	0.73017185139793212284724 01D+01	0.98113025695540054901760 24D+01
4	0.93256497425572464039917 17D+01	0.17220198465973390696553 35D+02
5	0.11342533803387059107008 59D+02	0.26638169277777889946706 44D+02
6	0.13355278620003509127761 66D+02	0.38063081451713061889973 79D+02
7	0.15365344638234987638798 41D+02	0.51493472261239039805402 11D+02
8	0.17373558079167667562843 31D+02	0.66928314185979485177735 95D+02
9	0.19380426713403465987817 99D+02	0.84366859492647833064115 61D+02
10	0.21386282307577038402018 35D+02	0.10380854621690826434949 63D+03
11	0.23391352121618863872832 04D+02	0.12525294023979745947121 27D+03
12	0.25395797800772431866967 02D+02	0.14869969834382843365066 72D+03
13	0.27399737908126324383360 57D+02	0.17414854380937822111560 81D+03
14	0.29403261660321889569536 19D+02	0.20159924977173251906126 71D+03
15	0.31406437659759026983137 27D+02	0.23105162754565671608347 20D+03
16	0.33409319645249837542931 90D+02	0.26250551822904687433989 86D+03
17	0.35411950394815987595886 94D+02	0.29596078653156353235649 55D+03
18	0.37414364444106038877489 66D+02	0.33141731615047276900823 55D+03

TABLE 1 (*continued*)

k	alpha(k)	beta(k)
19	0.39416590023263585452615 33D+02	0.36887500624612839234891 00D+03
20	0.41418650464761856741814 97D+02	0.40833376871438369164044 52D+03
21	0.43420565244991268776951 73D+02	0.44979352604673523203340 15D+03
22	0.45422350767174583074598 72D+02	0.49325420963079124902694 02D+03
23	0.47424020958286094251402 12D+02	0.53871575838535563785044 31D+03
24	0.49425587730056977442132 63D+02	0.58617811765310286795840 24D+03
25	0.51427061339201650743365 92D+02	0.63564123829389873772350 91D+03
26	0.53428450671916934060193 81D+02	0.68710507593610413709676 09D+03
27	0.55429763470782432459859 49D+02	0.74056959035350743740703 77D+03
28	0.57431006517359307228941 78D+02	0.79603474494307220406826 76D+03
29	0.59432185780362916511967 56D+02	0.85350050628427242712060 02D+03
30	0.61433306536828190182706 84D+02	0.91296684376497226030864 84D+03
31	0.63434373471900528796986 44D+02	0.97443372926197615538750 38D+03
32	0.65435390761571238355914 52D+02	0.10379011368667987515100 73D+04
33	0.67436362141699625774572 29D+02	0.11033690426490744434031 61D+04
34	0.69437290965930137820205 92D+02	0.11708374244514828389943 80D+04
35	0.71438180254556600517079 70D+02	0.12403062617112092545027 93D+04
36	0.73439032735960077220210 81D+02	0.13117755353038632147856 56D+04
37	0.75439850881918665662853 80D+02	0.13852452274064977526901 27D+04
38	0.77440636937832458610121 71D+02	0.14607153213769494381217 14D+04
39	0.79441392948707162373152 67D+02	0.15381858016471847294911 38D+04
40	0.81442120781582404323611 67D+02	0.16176566536287162050918 56D+04
41	0.83442822144965808849797 74D+02	0.16991278636284607506928 31D+04
42	0.85443498605734155233472 28D+02	0.17825994187736649614361 73D+04
43	0.87444151603882807452353 53D+02	0.18680713069447318425658 80D+04
44	0.89444782465439898908580 56D+02	0.19555435167149559671002 50D+04
45	0.91445392413809230825583 84D+02	0.20450160372963183493430 18D+04
46	0.93445982579762955930272 48D+02	0.21364888584906127960666 75D+04
47	0.95446554010270029651380 26D+02	0.22299619706452766869494 99D+04
48	0.97447107676317407560870 72D+02	0.23254353646133844460390 56D+04
49	0.99447644479857079327406 54D+02	0.24229090317173341621904 48D+04
50	0.10144816525999212006277 57D+03	0.25223829637158191367866 09D+04
51	0.10344867079849836496548 26D+03	0.26238571527737284013690 17D+04
52	0.10544916182476443092111 05D+03	0.27273315914346649422114 47D+04
53	0.10744963902022114574067 75D+03	0.28328062725958087121389 72D+04
54	0.10945010302232161215028 07D+03	0.29402811894848845056255 06D+04
55	0.11145055442812481455001 72D+03	0.30497563356390232510218 68D+04
56	0.11345099379752861586368 01D+03	0.31612317048853299220559 71D+04
57	0.11545142165619198045292 14D+03	0.32747072913229926627512 63D+04
58	0.11745183849818112574823 17D+03	0.33901830893067863346470 40D+04
59	0.11945224478836990902782 43D+03	0.35076590934318399342904 77D+04
60	0.12145264096462098031507 55D+03	0.36271352985195515302087 66D+04
61	0.12345302743977098129155 52D+03	0.37486116996045468169718 76D+04
62	0.12545340460344026333460 34D+03	0.38720882919225883207108 26D+04
63	0.12745377282368516849534 21D+03	0.39975650708993519202567 22D+04
64	0.12945413244850880957019 36D+03	0.41250420321399958451169 65D+04
65	0.13145448380724445251977 88D+03	0.42545191714194548244673 66D+04
66	0.13345482721182400692152 74D+03	0.43859964846733987173941 94D+04
67	0.13545516295794273463653 27D+03	0.45194739679898008628725 56D+04
68	0.13745549132613006528000 63D+03	0.46549516176010666422057 89D+04
69	0.13945581258273533553232 21D+03	0.47924294298766774277297 11D+04
70	0.14145612698083632750973 70D+03	0.49319074013163092693359 63D+04
71	0.14345643476107765202062 29D+03	0.50733855285433894052545 18D+04
72	0.14545673615244529077211 67D+03	0.52168638082990570280145 56D+04
73	0.14745703137298296479311 84D+03	0.53623422374364977361851 35D+04
74	0.14945732063045542362519 09D+03	0.55098208129156237971183 66D+04
75	0.15145760412296324184459 29D+03	0.56592995317980747701647 11D+04
76	0.15345788203951325814744 97D+03	0.58107783912425152240401 07D+04
77	0.15545815456054839058441 50D+03	0.59642573885002082527804 61D+04
78	0.15745842185844020353565 27D+03	0.61197365209108452753560 23D+04
79	0.15945868409794728243625 73D+03	0.62772157858986142150565 71D+04

TABLE 2

Recursion coefficients for the polynomials $\{\pi_k(\cdot; w_2)\}$

k	alpha(k)	beta(k)
0	0.63245886971850936236610460+00	0.10721549299401913395308970+01
1	0.26184924841473600282012220+01	0.67521709631759430156397120+00
2	0.45795649636410436101717030+01	0.31675835602583982388211780+01
3	0.65640307792994767460871910+01	0.76961592305189008195692770+01
4	0.85550699231017914919898420+01	0.14222840658306459618714580+02
5	0.10549055512086048984290300+02	0.22747103956778693637439300+02
6	0.12544660675830604084014320+02	0.33269636867785143113505170+02
7	0.14541268416263327765338090+02	0.45790003399176020819747280+02
8	0.16538547584107779369781120+02	0.60309314352732408490887330+02
9	0.18536302389020313856558920+02	0.76827515995496473448402300+02
10	0.20534408661457238081223300+02	0.95344775569597250176840180+02
11	0.22532783340432013142933420+02	0.11586122321188056567075160+03
12	0.24531368481173921194445440+02	0.13837696218079208224295570+03
13	0.26530122264191848592115500+02	0.16289207586793601379874050+03
14	0.28529013649267782240601800+02	0.18940663268191962647617660+03
15	0.30528019047512582930054610+02	0.21792068951820909101532070+03
16	0.32527120168812919681497600+02	0.24843429427418868760992560+03
17	0.34526302583572362649125600+02	0.28094748770694697765938790+03
18	0.36525554734518330181429800+02	0.31546030482966644598659590+03
19	0.38524867241069635231210970+02	0.35197277597800563345288450+03
20	0.40524232399118860328547840+02	0.39048492763629561596800600+03
21	0.42523643814499269297692570+02	0.43099678308609004051163120+03
22	0.44523096129865617562915080+02	0.47350836292136877393545460+03
23	0.46522584818096776986896490+02	0.51801968546228273911583010+03
24	0.48522106023882696096284210+02	0.56453076709073644293490430+03
25	0.50521656440753942160941500+02	0.61304162252506360690444090+03
26	0.52521233214548101958414000+02	0.66355226504674006934938520+03
27	0.54520833866848474656957190+02	0.71606270668895878214613530+03
28	0.56520456233688419808260950+02	0.77057295839460569446735430+03
29	0.58520098416049804004403410+02	0.82708303014947991732583530+03
30	0.60519758739564156557171860+02	0.88559293109533017411340290+03
31	0.62519435721460662218508360+02	0.94610266962631620820926770+03
32	0.64519128043269586392564240+02	0.10086122534717668122061500+04
33	0.66518834528133022804534780+02	0.10731216897675371712987610+04
34	0.68518554121831255702095160+02	0.11396309851178252288884830+04
35	0.70518285876826409469872370+02	0.12081401456489591355932390+04
36	0.72518028938772244433083800+02	0.12786491770563929689326330+04
37	0.74517782535051945126398730+02	0.13511580846459290417564380+04
38	0.76517545964993181094464700+02	0.14256668733700096819064690+04
39	0.78517318591477892365063300+02	0.15021755478597798603524910+04
40	0.80517099833717178497355304D+02	0.15806841124535072291924620+04
41	0.82516889161008843089517750+02	0.16611925712218524460328120+04
42	0.84516686087311834874113400+02	0.17437009279904058102436780+04
43	0.86516490166532739042895270+02	0.18282091863598428750301830+04
44	0.88516300988398713143783990+02	0.19147173497239992076459300+04
45	0.90516118174842782692559300+02	0.20032254212861207841468410+04
46	0.92515941376824714480485610+02	0.20937334040735099887420940+04
47	0.94515770271527215239623500+02	0.21862413009507565371329960+04
48	0.96515604559876188752741020+02	0.22807491146317168135981690+04
49	0.98515443964341698304127520+02	0.23772568476903832618469010+04
50	0.10051528822698283859030950+03	0.24757645025707669192869960+04
51	0.10251513710770517661502500+03	0.25762720815959003802064700+04
52	0.10451499038270397759220850+03	0.26787795869760549638867580+04
53	0.10651484784307025186877580+03	0.27832870208162542786450680+04
54	0.10851470929353987292901980+03	0.28897943851231564075111880+04
55	0.11051457455136873011069300+03	0.29983016818113683438596890+04
56	0.11251444344531917828506350+03	0.31088089127092488669694630+04
57	0.11451431581474500009306840+03	0.32213160795642495942948660+04
58	0.11651419150876376134854740+03	0.33358231840478383339654740+04
59	0.11851407038550686366992580+03	0.34523302277600439659559070+04

TABLE 2 (*continued*)

k	alpha(k)	beta(k)
60	0.12051395231143881879755 63D+03	0.35708372122336576801133153D+04
61	0.12251383716073831823441 34D+03	0.36913441389381226176721 31D+04
62	0.12451372481473457622376 03D+03	0.38138510092831372811057 33D+04
63	0.12651361516139320562866 78D+03	0.39383578246220019555894 68D+04
64	0.12851350809484656335520 84D+03	0.40648645862547263571056 46D+04
65	0.13051340351496408993084 10D+03	0.41933712954309212393451 42D+04
66	0.13251330132695867961882 34D+03	0.43238779533524913014306 89D+04
67	0.13451320144102556389004 82D+03	0.44563845611761459305106 39D+04
68	0.13651310377201058136811 02D+03	0.45908911200157426132141 27D+04
69	0.13851300823910504928852 39D+03	0.47273976309444764435983 10D+04
70	0.14051291476556475167220 25D+03	0.48659040949969279004484 75D+04
71	0.14251282327845092342769 92D+03	0.50064105131709799453614 52D+04
72	0.14451273370839054226291 63D+03	0.51489168864296144891219 02D+04
73	0.14651264598935624570779 58D+03	0.52934232157025973737105 12D+04
74	0.14851256005846077224225 90D+03	0.54399295018880602088796 17D+04
75	0.15051247585576798651057 43D+03	0.55884357458539866751374 62D+04
76	0.15251239332411709148675 63D+03	0.57389419484396102500536 94D+04
77	0.15451231240895955748115 05D+03	0.58914481104567297240260 14D+04
78	0.15651223305820761098769 96D+03	0.60459542326909483380020 06D+04
79	0.15851215522209332724680 49D+03	0.62024603159028418929514 91D+04

BIBLIOGRAPHY

1. W. J. Cody, *Chebyshev approximations for the Fresnel integrals*, Math. Comp. **22** (1968), 450–453. Loose microfiche suppl. A1–B4.
2. W. J. Cody, K. A. Paciorek, and H. C. Thacher, Jr., *Chebyshev approximations for Dawson's integral*, Math. Comp. **24** (1970), 171–178.
3. P. J. Davis, *Gamma function and related functions*, Handbook of Mathematical Functions, Chapter 6 (M. Abramowitz and I. A. Stegun, eds.), NBS Appl. Math. Series, vol. 55, U.S. Government Printing Office, Washington, D.C., 1964.
4. ____, *Spirals: From Theodorus of Cyrene to Meta-Chaos*, The Hedrick Lectures, Math. Assoc. Amer., August, 1990.
5. W. Gautschi, *Error function and Fresnel integrals*, Handbook of Mathematical Functions, Chapter 7 (M. Abramowitz and I. A. Stegun, eds.), NBS Appl. Math. Series, vol. 55, U.S. Government Printing Office, Washington, D.C., 1964.
6. ____, *Computational aspects of orthogonal polynomials*, Orthogonal Polynomials—Theory and Practice (P. Nevai, ed.), NATO ASI Series, Series C: Mathematical and Physical Sciences, vol. 294, Kluwer, Dordrecht, 1990, pp. 181–216.
7. W. Gautschi and G. V. Milovanović, *Gaussian quadrature involving Einstein and Fermi functions with an application to summation of series*, Math. Comp. **44** (1985), 177–190.
8. P. Henrici, *Applied and computational complex analysis*, vol. 1, Wiley, New York, 1984.

DEPARTMENT OF COMPUTER SCIENCES, PURDUE UNIVERSITY, WEST LAFAYETTE, INDIANA 47907
E-mail address: wxg@cs.purdue.edu

30.4. [125] "ON CERTAIN SLOWLY CONVERGENT SERIES OCCURRING IN PLATE CONTACT PROBLEMS"

---

[125] "On Certain Slowly Convergent Series Occurring in Plate Contact Problems," *Math. Comp*. **57**, 325–338 (1991).

MATHEMATICS OF COMPUTATION
VOLUME 57, NUMBER 195
JULY 1991, PAGES 325–338

ON CERTAIN SLOWLY CONVERGENT SERIES OCCURRING IN PLATE CONTACT PROBLEMS

WALTER GAUTSCHI

ABSTRACT. A simple computational procedure is developed for accurately summing series of the form $\sum_{k=0}^{\infty}(2k+1)^{-p}z^{2k+1}$, where z is complex with $|z| \le 1$ and $p = 2$ or 3, as well as series of the type

$$\sum_{k=0}^{\infty}(2k+1)^{-p}\cosh(2k+1)x/\cosh(2k+1)b$$

and

$$\sum_{k=0}^{\infty}(2k+1)^{-p}\sinh(2k+1)x/\cosh(2k+1)b,$$

where $0 \le x \le b$, $p = 2$ or 3. The procedures are particularly useful in cases where the series converge slowly. Numerical experiments illustrate the effectiveness of the procedures.

1. INTRODUCTION

Our concern, in §§2–4, is with series of the type

$$R_p(z) = \sum_{k=0}^{\infty}\frac{z^{2k+1}}{(2k+1)^p} \tag{1.1$_p$}$$

or the type

$$S_p(z) = \sum_{k=0}^{\infty}(-1)^k\frac{z^{2k+1}}{(2k+1)^p}, \tag{1.2$_p$}$$

where

$$z \in \mathbb{C}, \quad |z| \le 1, \quad \text{and} \quad p = 2 \text{ or } 3. \tag{1.3}$$

Of particular interest to us is the numerical evaluation of these series in cases of slow convergence, i.e., when $|z|$ is close or equal to 1. It clearly suffices to concentrate on the first of the two series, R_p, since

$$S_p(z) = iR_p(-iz). \tag{1.4}$$

Received July 23, 1990.

1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 40A25; Secondary 44A10, 33A65.

Key words and phrases. Slowly convergent series, Laplace transformation, Stieltjes transform, orthogonal polynomials.

Work supported, in part, by the National Science Foundation under grant CCR-8704404.

Furthermore, $R_p(-z) = -R_p(z)$ and $R_p(\overline{z}) = \overline{R_p(z)}$, so that attention can be restricted to the first quadrant of the complex plane.

Series of the type (1.1_p), with

$$(1.5) \qquad z = A, \quad 0 < A \leq 1, \qquad \text{and} \qquad z = e^{i\alpha}, \quad \alpha \in \mathbb{R},$$

occur in the mathematical treatment of unilateral plate contact problems, and their numerical evaluation, in this context, has recently been discussed by K. M. Dempsey, D. Liu, and J. P. Dempsey [1]. The method proposed by these authors consists in applying Plana's summation formula, which in turn requires the numerical evaluation of several definite integrals—for example by Romberg integration.

Here we develop a technique which appears to be considerably simpler. All it requires is the application (in the backward direction) of a three-term recurrence relation, once a set of numerical constants has been precomputed. Results of high accuracy are easily achieved, even for $|z|$ near or equal to 1.

Some of the series (1.1_p), (1.2_p) with $p = 2$ or $p = 3$ can be summed explicitly as Fourier series when z is given by the second expression in (1.5). We thus have

$$(1.6_2) \qquad \sum_{k=0}^{\infty} \frac{\cos(2k+1)\alpha}{(2k+1)^2} = \pi(\pi - 2|\alpha|)/8, \qquad -\pi \leq \alpha \leq \pi \quad [8, (17.2.16)],$$

$$(1.6_3) \qquad \sum_{k=0}^{\infty} \frac{\sin(2k+1)\alpha}{(2k+1)^3} = \pi\alpha(\pi - |\alpha|)/8, \qquad -\pi \leq \alpha \leq \pi \quad [8, (14.2.21)],$$

and analogous formulae for

$$\sum_{k=0}^{\infty}(-1)^k(2k+1)^{-2}\sin(2k+1)\alpha, \qquad \sum_{k=0}^{\infty}(-1)^k(2k+1)^{-3}\cos(2k+1)\alpha,$$

which can be obtained from (1.6) by applying (1.4). When $z = 1$, the sum of (1.1_p) is expressible in terms of the Riemann zeta function,

$$(1.7) \qquad R_p(1) = (1 - 2^{-p})\zeta(p),$$

whereas $S_2(1)$ is known as Catalan's constant, and $S_3(1) = \pi^3/32$. All these explicit formulae will be useful for testing purposes.

In §5 we combine our techniques of §2 with series expansion to deal with the more difficult series

$$(1.8_p) \qquad T_p(x, b) = \sum_{k=0}^{\infty} \frac{1}{(2k+1)^p} \frac{\cosh(2k+1)x}{\cosh(2k+1)b}$$

and

$$(1.9_p) \qquad U_p(x, b) = \sum_{k=0}^{\infty} \frac{1}{(2k+1)^p} \frac{\sinh(2k+1)x}{\cosh(2k+1)b},$$

where

$$0 \le x \le b, \quad b > 0, \quad \text{and} \quad p = 2, 3. \tag{1.10}$$

Both are also of interest in plate contact problems [1]. Here again, we are able to sum these series effectively and to high precision, the major (as yet unresolved) difficulty occurring when b is very small.

2. Summation of R_p and S_p, $p = 2$ and $p = 3$

We begin with an idea used previously in [7, 6], namely to express part of each term of the series (not the whole term, as in [7, 6]) as a Laplace transform with integer argument. Specifically,

$$\frac{1}{(k+1/2)^p} = (\mathscr{L}f)(k), \qquad \mathscr{L} = \text{Laplace transform}, \tag{2.1}$$

where

$$f(t) = \frac{1}{(p-1)!} t^{p-1} e^{-t/2}. \tag{2.2}$$

Then

$$\begin{aligned} R_p(z) &= \frac{z}{2^p} \sum_{k=0}^{\infty} \frac{z^{2k}}{(k+1/2)^p} = \frac{z}{2^p} \sum_{k=0}^{\infty} z^{2k} \int_0^{\infty} e^{-kt} \cdot \frac{t^{p-1} e^{-t/2}}{(p-1)!}\, dt \\ &= \frac{z}{2^p (p-1)!} \int_0^{\infty} \sum_{k=0}^{\infty} (z^2 e^{-t})^k \cdot t^{p-1} e^{-t/2}\, dt \\ &= \frac{z}{2^p (p-1)!} \int_0^{\infty} \frac{1}{1 - z^2 e^{-t}} t^{p-1} e^{-t/2}\, dt, \end{aligned}$$

that is,

$$R_p(z) = \frac{z}{2^p (p-1)!} \int_0^{\infty} \frac{t^{p-1} e^{t/2}}{e^t - z^2}\, dt. \tag{2.3}$$

We distinguish two cases.

Case 1: $z = 1$. In this case, (2.3) takes the form

$$R_p(1) = \frac{1}{2^p (p-1)!} \int_0^{\infty} \frac{t}{e^t - 1} \cdot t^{p-2} e^{t/2}\, dt \tag{2.4}$$

and can be evaluated by Gaussian quadrature relative to the weight function (cf. [7])

$$\varepsilon(t) = \frac{t}{e^t - 1} \quad \text{on } [0, \infty] \qquad (\text{"Einstein function"}). \tag{2.5}$$

However, there is no real need for this, since by (1.7) the sum is expressible in terms of the well-tabulated Riemann zeta function [9]. In particular, $R_2(1) = \pi^2/8$.

The more difficult case is

Case 2: $z \neq 1$. Here we could proceed similarly as in (2.4) and write

$$R_p(z) = \frac{z}{2^p(p-1)!}\int_0^\infty \varepsilon(t)\cdot\frac{e^t-1}{e^t-z^2}\cdot t^{p-2}e^{t/2}\,dt. \tag{2.6}$$

Unfortunately, the second factor in the integrand is quite ill-behaved when $|z|$ is close to 1, exhibiting a steep boundary layer near $t = 0$. (Consider, e.g., $z^2 = 1-\eta$, $0 < \eta \ll 1$.) Gaussian quadrature, therefore, will no longer be effective.

Instead, we make the change of variable $e^{-t} = \tau$ in (2.3) (and then replace τ again by t) to obtain

$$R_p(z) = \frac{1}{2^p(p-1)!z}\int_0^1 \frac{[\ln(1/t)]^{p-1}}{\sqrt{t}}\frac{dt}{z^{-2}-t}. \tag{2.7}$$

This expresses $R_p(z)$ as a Stieltjes transform of the weight function

$$w_p(t) = \frac{[\ln(1/t)]^{p-1}}{\sqrt{t}} \quad \text{on } [0,1]. \tag{2.8}$$

Our assumptions on z are such that the point z^{-2} at which the transform is evaluated lies *outside* of the interval $[0,1]$,

$$z^{-2} \in \mathbb{C}\backslash[0,1]. \tag{2.9}$$

The integral in (2.7), therefore, can be evaluated by backward recursion, as is discussed in [3, §5].

Indeed, if

$$y_{n+1} = (z^{-2}-\alpha_n)y_n - \beta_n y_{n-1}, \qquad n = 0,1,2,\ldots, \tag{2.10}$$

is the recurrence relation for the orthogonal polynomials $\{\pi_n(z^{-2}; w_p)\}$ relative to the weight function (2.8), thus,

$$\alpha_n = \alpha_n(w_p), \quad \beta_n = \beta_n(w_p) \qquad \left[\beta_0(w_p) = \int_0^1 w_p(t)\,dt\right], \tag{2.11}$$

and if we define the sequence $\{r_{n-1}^{[\nu]}(z)\}_{n=0}^{\nu+1}$ for any integer $\nu > 0$ by

$$r_\nu^{[\nu]}(z) = 0, \qquad r_{n-1}^{[\nu]}(z) = \frac{\beta_n}{z^{-2}-\alpha_n - r_n^{[\nu]}(z)}, \quad n = \nu, \nu-1, \ldots, 1, 0, \tag{2.12}$$

then (cf. [3, equation (5.2) for $N = 0$])

$$\int_0^1 \frac{w_p(t)\,dt}{z^{-2}-t} = \lim_{\nu\to\infty} r_{-1}^{[\nu]}(z). \tag{2.13}$$

Thus, by (2.7),

$$R_p(z) = \frac{r_{-1}^{[\infty]}(z)}{2^p(p-1)!z}. \tag{2.14}$$

Convergence in (2.13) is faster the further away z^{-2} is from the interval $[0, 1]$.

The evaluation of $r_{-1}^{[\infty]}(z)$ is quite cheap, once the coefficients α_n, β_n in (2.11) have been precomputed for sufficiently many n. One simply lets ν increase through a sequence $\{\nu_i\}$ of integers $0 < \nu_1 < \nu_2 < \cdots$ and stops at the smallest i, say $i = i_{\min}$, for which $|r_{-1}^{[\nu_i]}(z) - r_{-1}^{[\nu_{i-1}]}(z)| \le \varepsilon |r_{-1}^{[\nu_i]}(z)|$, where ε is a preset error tolerance. One then accepts $r_{-1}^{[\nu_i]}(z)$ with $i = i_{\min}$ as the desired approximation of $r_{-1}^{[\infty]}(z)$ in (2.14). For the two choices of z in (1.5), practical guidelines for determining an acceptable value of ν (i.e., one for which $r_{-1}^{[\nu]}(z)$ sufficiently approximates $r_{-1}^{[\infty]}(z)$) will be given in §4.

The coefficients α_n, β_n can be computed by known methods, as will be further discussed in §3. The first 100 coefficients are tabulated in Tables 1 and 2 of the Appendix for $p = 2$ and $p = 3$ to an accuracy of 25 and 20 significant decimal digits, respectively.

The procedure (2.12)–(2.14), in view of (1.4), is readily adapted to the series S_p in (1.2_p). Indeed, letting $s_n^{[\nu]}(z) = -r_n^{[\nu]}(-iz)$, one finds

$$S_p(z) = \frac{s_{-1}^{[\infty]}(z)}{2^p (p-1)! z}, \tag{2.15}$$

where

$$s_\nu^{[\nu]}(z) = 0, \quad s_{n-1}^{[\nu]}(z) = \frac{\beta_n}{z^{-2} + \alpha_n - s_n^{[\nu]}(z)}, \quad n = \nu, \nu-1, \ldots, 1, 0. \tag{2.16}$$

Since $S_p(z)$ is effectively the Stieltjes transform of $w_p(\cdot)$ evaluated at $-z^{-2}$, the process (2.15), (2.16) converges more rapidly (as $\nu \to \infty$) the further away $-z^{-2}$ is from the interval $[0, 1]$. In particular, it converges rapidly for $z = 1$, yielding a fast way of computing Catalan's constant when $p = 2$. Indeed, taking $\nu = 16$ in (2.16) produces $s_{-1}^{[\infty]}(1)$, hence $S_2(1)$, accurately to 25 decimal digits!

3. Generating the coefficients $\alpha_n(w_p)$, $\beta_n(w_p)$ for $p = 2$ and $p = 3$

Consider the weight function

$$w_p(t; \alpha) = t^\alpha [\ln(1/t)]^{p-1}, \qquad 0 < t \le 1, \ \alpha > -1, \ p \ge 2, \tag{3.1}$$

and let

$$m_n(\alpha; p) = \int_0^1 P_n^*(t) w_p(t; \alpha)\, dt, \qquad n = 0, 1, 2, \ldots, \tag{3.2}$$

denote the "modified moments" of $w_p(\cdot; \alpha)$ relative to the shifted Legendre polynomials $P_n^*(t) = P_n(2t - 1)$. In the case $p = 2$ these modified moments

are explicitly known (cf. [2]):

$$m_n(\alpha;2) = \frac{1}{\alpha+1}\left\{\frac{1}{\alpha+1} + 2\sum_{k=1}^{n}\frac{k}{(k+\alpha+1)(k-\alpha-1)}\right\} \cdot \prod_{k=1}^{n}\frac{\alpha+1-k}{\alpha+1+k}. \tag{3.3}$$

It is also well known how the modified moments of a weight function w can be used to generate the recursion coefficients $\alpha_n(w)$, $\beta_n(w)$ of the respective orthogonal polynomials $\{\pi_k(\cdot;w)\}$ by means of the so-called modified Chebyshev algorithm [4, §2.4]. This algorithm indeed works particularly well in the case of the weight function (3.1) with $p=2$, $\alpha=-\frac{1}{2}$, i.e., for $w(t)=w_2(t)$ (cf. (2.8)), as was demonstrated in [5, Example 5.3]. This, then, is the way we computed the quantities $\alpha_n(w_p)$, $\beta_n(w_p)$ for $p=2$. Compensating for a loss of about four decimal digits, when n runs from 0 to 99, we tabulate the results in Table 1 of the Appendix to only 25 decimals (having done the computation in 29-decimal arithmetic).

In order to get the same quantities for $p=3$, it suffices to observe that

$$\frac{\partial w_p}{\partial \alpha}(t;\alpha) = -w_{p+1}(t;\alpha),$$

and therefore

$$m_n(\alpha;p+1) = -\int_0^1 P_n^*(t)\frac{\partial w_p}{\partial \alpha}(t;\alpha)\,dt = -\frac{\partial m_n}{\partial \alpha}(\alpha;p).$$

Thus, the required modified moments $m_n(\alpha;3)$ can be obtained by differentiating both sides of (3.3) with respect to α (after multiplication by -1). The result is

$$m_n(\alpha;3) = \frac{2}{(\alpha+1)^3}\left\{1 + 2(\alpha+1)\sum\nolimits_1 + 2(\alpha+1)^2\sum\nolimits_1^2 - 2(\alpha+1)^3\sum\nolimits_2\right\}\Pi, \tag{3.4}$$

where

$$\sum\nolimits_1 = \sum_{k=1}^{n}\frac{k}{(k+\alpha+1)(k-\alpha-1)},$$

$$\sum\nolimits_2 = \sum_{k=1}^{n}\frac{k}{(k+\alpha+1)^2(k-\alpha-1)^2}, \tag{3.5}$$

$$\Pi = \prod_{k=1}^{n}\frac{\alpha+1-k}{\alpha+1+k}.$$

Putting $\alpha=-\frac{1}{2}$ in (3.4), and using the resulting quantities as input to the modified Chebyshev algorithm, produces the coefficients $\alpha_n(w_3)$, $\beta_n(w_3)$. The procedure is somewhat less stable than in the case $p=2$, suffering a loss of

about eight to nine decimal digits when applied up to $n = 99$. For this reason we tabulate $\alpha_n(w_3)$, $\beta_n(w_3)$ in Table 2 of the Appendix to only 20 decimals.

4. Implementation and numerical examples

It would clearly be desirable in our procedure (2.12) to know a priori what value to choose for the starting index ν, given any z in the first quadrant of $\mathbb{C}$ and given the required accuracy. The recursion in (2.12) then would need to be run through only once, and the iterative procedure suggested in §2 could be dispensed with.

To deal with this problem, we consider only the two cases of practical interest stated in (1.5). More precisely, we address the following related problem: Given ν and the desired relative accuracy ε, determine the set of values A in $[0, 1]$, resp. α in $[0, \pi/2]$, for which $r_{-1}^{[\nu]}$ in (2.13) approximates $r_{-1}^{[\infty]}$ within a relative error of ε.

As to the values of A, we note that the speed of convergence in (2.13) decreases as A increases in $[0, 1]$. The desired set of A-values must thus have the form $0 \le A \le A(\nu, \varepsilon) \le 1$, and the problem is to determine $A(\nu, \varepsilon)$. We solve this empirically by a bisection procedure: Start with two numbers A_0^-, A_0^+ such that $A_0^- \le A(\nu, \varepsilon) \le A_0^+$, for example, $A_0^- = 0$, $A_0^+ = 1$. Having already obtained A_{k-1}^- and A_{k-1}^+ with $A_{k-1}^- < A_{k-1}^+$, test the midpoint $M = \frac{1}{2}(A_{k-1}^- + A_{k-1}^+)$ to see whether at M the procedure (2.12) yields an approximation $r_{-1}^{[\nu]}$ with relative error larger or smaller than ε. In the former case we set $A_k^- = A_{k-1}^-$, $A_k^+ = M$, in the latter case $A_k^- = M$, $A_k^+ = A_{k-1}^+$. We quit this iteration as soon as, say, $A_k^+ - A_k^- \le \frac{1}{2}10^{-6}$ and take $\frac{1}{2}(A_k^- + A_k^+)$ to approximate $A(\nu, \varepsilon)$. In order to determine the relative errors of $r_{-1}^{[\nu]}$, as required in this procedure, we approximate $r_{-1}^{[\infty]}$ by $r_{-1}^{[99]}$ and, at the same time, check to see that $r_{-1}^{[99]}$ and $r_{-1}^{[98]}$ agree to within a relative accuracy $\varepsilon/100$. If they do, it is safe to assume that $r_{-1}^{[99]}$ can reliably substitute $r_{-1}^{[\infty]}$ in determining whether $r_{-1}^{[\nu]}$ has relative error $> \varepsilon$ or $< \varepsilon$. If they do not, we print a cautionary message, and take A_k^- as a (conservative) estimate from below of $A(\nu, \varepsilon)$.

The results of this procedure are summarized in Table 4.1 for both $p = 2$ and $p = 3$. An asterisk indicates a conservative lower estimate of $A(\nu, \varepsilon)$ for reasons explained above.

We can see from Table 4.1, for example, that if we are interested in 12-digit accuracy and only in positive values of A satisfying $A \le .99$, then we can safely use $\nu = 50$ in (2.12) when $p = 2$, and $\nu = 40$ when $p = 3$. On the other hand, the choice $\nu = 10$ for the same range of A-values, always gives at least four correct decimal digits.

Interestingly, the procedure (2.12), (2.13) seems to converge even for $A = 1$, albeit slowly, but there is no theoretical justification for it (to our knowledge).

TABLE 4.1

Values of $A(\nu, \varepsilon)$, $\varepsilon = \frac{1}{2}10^{-acc}$, *such that* $r_{-1}^{[\nu]}$ *approximates* $r_{-1}^{[\infty]}$ *to acc digits for all* A *with* $0 \le A \le A(\nu, \varepsilon)$

ν	*acc*	$p=2$	$p=3$	ν	*acc*	$p=2$	$p=3$	ν	*acc*	$p=2$	$p=3$
10	4	.9902	1.0000	40	4	.9999	1.0000	70	4	1.0000	1.0000
	8	.9313	.9592		8	.9962	.9993		8	.9989	1.0000
	12	.8422	.8732		12	.9889	.9936		12	.9961*	.9980*
	16	.7384	.7688		16	.9786	.9842		16	.9922*	.9955
	20	.6325	.6603		20	.9653	.9717		20	.9887	.9914
20	4	.9985	1.0000	50	4	1.0000	1.0000	80	4	1.0000	1.0000
	8	.9827	.9931		8	.9977	.9998		8	.9990*	1.0000
	12	.9551	.9685		12	.9931	.9963		12	.9961*	.9980*
	16	.9178	.9330		16	.9864	.9903		16	.9922*	.9961*
	20	.8729	.8891		20	.9777	.9823		20	.9914	.9922*
30	4	.9996	1.0000	60	4	1.0000	1.0000	90	4	1.0000	1.0000
	8	.9927	.9980		8	.9985	1.0000		8	.9990*	1.0000
	12	.9800	.9873		12	.9953	.9976		12	.9961*	.9980*
	16	.9620	.9708		16	.9906	.9936		16	.9922*	.9961*
	20	.9395	.9492		20	.9845	.9880		20	.9922*	.9922*

For the second choice $z = e^{i\alpha}$, $0 \le \alpha \le \pi/2$, in (1.5), it was observed empirically that the speed of convergence in (2.13) decreases—slowly at first, and then faster—as α decreases from $\pi/2$ to 0. Therefore, a similar procedure as above for A-values can be applied to determine the number $\omega(\nu, \varepsilon)$ with the property that for all α satisfying $0 \le \omega(\nu, \varepsilon)\pi/2 \le \alpha \le \pi/2$, the procedure (2.12) produces $r_{-1}^{[\nu]}$ with (at least approximately) $|(r_{-1}^{[\nu]} - r_{-1}^{[\infty]})/r_{-1}^{[\infty]}| \le \varepsilon$. The results are displayed in Table 4.2.

TABLE 4.2

Values of $\omega(\nu, \varepsilon)$, $\varepsilon = \frac{1}{2}10^{-acc}$, *such that* $r_{-1}^{[\nu]}$ *approximates* $r_{-1}^{[\infty]}$ *to acc digits for all* α *with* $\omega(\nu, \varepsilon)\pi/2 \le \alpha \le \pi/2$

ν	*acc*	$p=2$	$p=3$	ν	*acc*	$p=2$	$p=3$	ν	*acc*	$p=2$	$p=3$
10	4	.0159	0.0000	40	4	.0002	0.0000	70	4	0.0000	0.0000
	8	.1066	.0690		8	.0054	.0013		8	.0020*	0.0000
	12	.2864	.2313		12	.0152	.0096		12	.0046	.0025
	16	.6789	.5576		16	.0294	.0226		16	.0092	.0078*
	20	1.0000*	1.0000*		20	.0482	.0404		20	.0156*	.0120
20	4	.0026	0.0000	50	4	.0001	0.0000	80	4	0.0000	0.0000
	8	.0247	.0118		8	.0033	.0005		8	.0020*	0.0000
	12	.0649	.0481		12	.0095	.0056		12	.0039*	.0020*
	16	.1251	.1045		16	.0185	.0138		16	.0078*	.0078*
	20	.2095	.1844		20	.0304	.0249		20	.0156*	.0089*
30	4	.0007	0.0000	60	4	0.0000	0.0000	90	4	0.0000	0.0000
	8	.0103	.0036		8	.0022	.0002		8	.0020*	0.0000
	12	.0278	.0190		12	.0064	.0036		12	.0039*	.0020*
	16	.0535	.0428		16	.0126	.0091		16	.0078*	.0078*
	20	.0878	.0754		20	.0208	.0168		20	.0156*	.0078*

TABLE 4.3
Results for Example 1

A	$p=2$	$p=3$	$R_2(A)$	$R_3(A)$
.8	21	16	.87728809392146472530085 18	.82248858052014232615
.9	30	23	1.02593895111111017277187 7	.93414857586540185586
.95	43	31	1.11409957792905248150121 3	.99191543992242877550
.99	95	65	1.20207566477685753806290 1	1.0395722318736413458
.999	–	–	1.2293981974	1.0505677498304
1.000	–	–	1.2336	1.051799789

Example 1. $R_p(A)$ for $A = .8, .9, .95, .99, .999, 1.000$, and $p = 2, 3$.

We applied the procedure (2.12) with $\nu = 1, 2, 3, \ldots$, terminating it for the first value of ν, $\nu = \nu_{\min}$, for which $|(r_{-1}^{[\nu]} - r_{-1}^{[\nu-1]})/r_{-1}^{[\nu]}| \le \varepsilon$, where $\varepsilon = \frac{1}{2}10^{-25}$ for $p = 2$, and $\varepsilon = \frac{1}{2}10^{-20}$ for $p = 3$. Table 4.3 shows the values of $\nu_{\min}$ along with 25-, resp. 20-digit results for $R_p(A)$, $p = 2, 3$.

For $A \ge .999$, full accuracy could not be achieved with $\nu \le 99$, only the partially accurate results shown in Table 4.3.

Example 2. $R_p(e^{i\alpha})$ for $\alpha = \omega\pi/2$, $\omega = .2, .1, .05, .01, .001, 0.000$, and $p = 2, 3$.

The same experiment as in Example 1 was run in this case, with the results being shown in Table 4.4. The first entry under each heading $R_p(e^{i\omega\pi/2})$ represents the real part, the second the imaginary part. The results for $\operatorname{Re} R_2$, $\operatorname{Im} R_3$ were checked against formulas (1.6_2) and (1.6_3), respectively, and revealed agreement to all digits shown.

TABLE 4.4
Results for Example 2

ω	$p=2$	$p=3$	$R_2(e^{i\omega\pi/2})$	$R_3(e^{i\omega\pi/2})$
.2	27	21	.98696044010893586188344 91	.96915102126251836837
			.44740227008596319725325 77	.34882061265337297697
.1	37	28	1.11033049512255284461888 0	1.0268555576593748316
			.27830297928558039181589 69	.18409976778928018229
.05	51	38	1.17201552262936133598659 6	1.0444944153967221625
			.16639152396897369411952 21	.09447224926028851460
.01	–	76	1.22136354463480812908 08	1.0514082919738793229
			.04592009281744058404956	.01928202831056145067
.001	–	–	1.232466849	1.051794454929
			.006400460	.001936923346
0.000	–	–	1.2337	1.051799789
			0.	0.

5. Summation of T_p and U_p, $p=2$ and $p=3$

We first take up the series (1.8_p). We expand the ratio of hyperbolic cosines as follows:

$$\frac{\cosh(2k+1)x}{\cosh(2k+1)b} = \sum_{n=0}^{\infty}(-1)^n\{e^{-(2k+1)[(2n+1)b-x]} + e^{-(2k+1)[(2n+1)b+x]}\}.$$

Then, upon using again the Laplace transform technique (2.1), (2.2), and interchanging the summations over k and n, one obtains after an elementary calculation

$$T_p(x,b) = \frac{1}{2^p(p-1)!}\sum_{n=0}^{\infty}(-1)^n e^{(2n+1)b}[\varphi_n(-x)+\varphi_n(x)], \tag{5.1}$$

where

$$\varphi_n(s) = e^s\int_0^1 \frac{w_p(t)\,dt}{e^{2[(2n+1)b+s]}-t}, \qquad -b\le s\le b. \tag{5.2}$$

The integral in (5.2) again is a Stieltjes transform of the weight function (2.8), this time evaluated at $u=\exp(2[(2n+1)b+s])$. Clearly, $u>1$, unless $n=0$ and $s=-b$, in which case, by (2.7) and (1.7),

$$\varphi_0(-b) = e^{-b}\int_0^1 \frac{w_p(t)}{1-t}\,dt = (2^p-1)(p-1)!\zeta(p)e^{-b}. \tag{5.3}$$

The integral in (5.2), hence both $\varphi_n(x)$ and $\varphi_n(-x)$ in (5.1) (the latter if $n>0$ or $x<b$), can be computed, as before, by the recursive procedure (2.12), (2.13) (where z^{-2} is to be replaced by u). For large n, this procedure converges almost instantaneously.

The series in (5.1), on the other hand, converges geometrically, with ratio $\exp(-2b)$. This is easily seen by noting that its general term (including the factor in front of the series) behaves like $2(-1)^n\cosh x\cdot e^{-2nb}$ as $n\to\infty$. Thus, convergence is quite satisfactory, unless b is small, the speed of convergence being independent of x. Table 5.1 shows the number of terms, N, required

TABLE 5.1

Number of terms required in the series of (5.1) *to achieve an accuracy of acc significant decimal digits*

b	*acc*	$p=2$	$p=3$	b	*acc*	$p=2$	$p=3$	b	*acc*	$p=2$	$p=3$
.05	4	104	105	.20	4	26	26	.80	4	7	7
	8	196	198		8	49	49		8	13	13
	12	288	290		12	72	72		12	18	18
	16	380	382		16	95	96		16	24	24
	20	473	474		20	118	119		20	30	30
.10	4	52	53	.40	4	13	13	1.60	4	4	4
	8	98	99		8	25	25		8	6	6
	12	144	145		12	36	36		12	9	9
	16	190	191		16	48	48		16	12	12
	20	236	237		20	59	59		20	15	15

in (5.1) to achieve various accuracies. As mentioned, N does not depend on x. It can be seen that the convergence characteristics of the series are virtually the same for $p = 2$ and $p = 3$. (When x is very close to b, the backward recursion (2.12) with $\nu \leq 99$ for evaluating $\varphi_0(-x)$ in (5.1) may provide only limited accuracy; cf. Example 1.)

For the series (1.9_p) one finds similarly

$$U_p(x, b) = \frac{1}{2^p(p-1)!}\sum_{n=0}^{\infty}(-1)^n e^{(2n+1)b}[\varphi_n(-x) - \varphi_n(x)], \tag{5.4}$$

with $\varphi_n(\cdot)$ defined in (5.2); the convergence behavior, when $x > 0$, is similar to the one shown in Table 5.1 for the series (5.1).

Series of the types (1.8_p), (1.9_p), which include alternating sign factors, can be treated similarly.

APPENDIX

Recursion coefficients α_n, β_n for the (monic) polynomials $\{\pi_k(\cdot; w_2)\}$ and $\{\pi_k(\cdot; w_3)\}$ orthogonal on $[0, 1]$ with respect to the weight functions $w_2(t) = t^{-1/2}\ln(1/t)$ and $w_3(t) = t^{-1/2}[\ln(1/t)]^2$.

TABLE 1

Recursion coefficients for the polynomials $\{\pi_k(\cdot; w_2)\}$

n	alpha(n)	beta(n)
0	0.11111111111111111111111111D+00	0.40000000000000000000000000D+01
1	0.46614836410754778101716 88D+00	0.27654320987654320987654320D-01
2	0.48806905819764265617396 54D+00	0.55342926841707111832654 76D-01
3	0.49387434192080573312748 22D+00	0.59405262984888651830670 45D-01
4	0.49626395786134592637002 77D+00	0.60777146066747328938272 87D-01
5	0.49748051363454704994043 27D+00	0.61403711431264107469512 99D-01
6	0.49818464245393947128190 88D+00	0.61741676592022707968813 79D-01
7	0.49862903362598437705294 48D+00	0.61944536279147177113286 88D-01
8	0.49892760828492354155461 95D+00	0.62075765809336261443409 40D-01
9	0.49913795646648500479808 02D+00	0.62165502445884118610013 40D-01
10	0.49929176974499659259765 74D+00	0.62229551936309398534370 94D-01
11	0.49940767088590894835203 75D+00	0.62276853260785448991122 81D-01
12	0.49949719160946385662422 02D+00	0.62312770828774884775638 86D-01
13	0.49956778517513642661565 99D+00	0.62340681999294532513398 64D-01
14	0.49962444394626209576455 66D+00	0.62362798995205712980762 83D-01
15	0.49967061459798408085488 42D+00	0.62380619889958642133257 18D-01
16	0.49970873924645614917289 15D+00	0.62395188347252497533381 38D-01
17	0.49974058765317366261908 72D+00	0.62407249483560017106891 11D-01
18	0.49976746790109466064172 45D+00	0.62417346753844348994681 91D-01
19	0.49979036384406940474666 12D+00	0.62425884050278824859136 15D-01
20	0.49981002705096913813591 06D+00	0.62433166588468583056150 25D-01
21	0.49982703969019018110531 16D+00	0.62439428474877549860531 15D-01
22	0.49984185840112695470616 28D+00	0.62444851692092106636980 73D-01
23	0.49985484545257844247060 20D+00	0.62449579424524223534376 94D-01
24	0.49986629123242189433015 92D+00	0.62453725573422426004572 26D-01
25	0.49987643072044243974059 48D+00	0.62457381657371861246587 78D-01
26	0.49988545571692466694490 19D+00	0.62460621888084789791669 57D-01
27	0.49989352403282268865915 72D+00	0.62463506952697074534795 68D-01
28	0.49990076647503651079180 77D+00	0.62466086865952047985422 40D-01
29	0.49990729221153824306286 71D+00	0.62468403144728664686728 16D-01
30	0.49991319293218225649394 69D+00	0.62470490482828369678795 67D-01
31	0.49991854610466230691806 92D+00	0.62472378053067142820290 63D-01
32	0.49992341754380909657649 87D+00	0.62474090528510611614299 77D-01
33	0.49992786345494608768660 86D+00	0.62475648889997358365542 1D-01
34	0.49993193207089328386338 69D+00	0.62477071069564418235950 53D-01

TABLE 1 (*continued*)

n	alpha(n)	beta(n)
35	0.49993566497245406232743870+00	0.62478372466799400949739840-01
36	0.49993909816047174225911360+00	0.62479566366005707088442480-01
37	0.49994226293149237040864770+00	0.62480664275367942846832460-01
38	0.49994518659711757542975720+00	0.62481676204346723780222450-01
39	0.49994789307815401140731280+00	0.62482610891830340060054430-01
40	0.49995040339786883924638830+00	0.62483475994783822493212880-01
41	0.49995273609347511826150160+00	0.62484278245021161187513160-01
42	0.49995490756098630051579170+00	0.62485023580109695368168720-01
43	0.49995693234549617386198330+00	0.62485717253170974603860490-01
44	0.49995882338654000220856400+00	0.62486363925377658340444820-01
45	0.49996059222631174261703180+00	0.62486967744193529050156970-01
46	0.49996224918702990583070700+00	0.62487532409813195044691080-01
47	0.49996380352256988086167250+00	0.62488061231792005965147220-01
48	0.49996526354854458006619690+00	0.62488557177486847424336190-01
49	0.49996663675426574349517600+00	0.62489022913633424970980430-01
50	0.49996792989941510586974580+00	0.62489460842149083321263850-01
51	0.49996914909776702528955350+00	0.62489873131059629430074730-01
52	0.49997029988990820961917610+00	0.62490261741294392844848990-01
53	0.49997138730657725491594920+00	0.62490628449968384339076960-01
54	0.49997241592398227492928210+00	0.62490974870668072753981020-01
55	0.49997338991223750466566490+00	0.62491302471173420879341720-01
56	0.49997431307788035905095070+00	0.62491612588979806010242960-01
57	0.49997518890128183973639030+00	0.62491906444926453024891310-01
58	0.49997602056963968446411640+00	0.62492185155190765363169240-01
59	0.49997681100614066062366410+00	0.62492449741868646704979040-01
60	0.49997756289579223287645170+00	0.62492701142328116447748200-01
61	0.49997827870835151001620050+00	0.62492940217496070868440320-01
62	0.49997896071871848931336890+00	0.62493167759214988509543610-01
63	0.49997961102510920793750520+00	0.62493384496786960159997890-01
64	0.49998023156528088329313350+00	0.62493591102806020155188000-01
65	0.49998082413104416628476110+00	0.62493788198365859759996310-01
66	0.49998139038126617097765000+00	0.62493976357718199841148000-01
67	0.49998193185354109268926700+00	0.62494156112447047498505120-01
68	0.49998244997468225298481290+00	0.62494327955215478293573300-01
69	0.49998294607016970662732910+00	0.62494492343134239328565330-01
70	0.49998342137267060738312020+00	0.62494649700795165509036310-01
71	0.49998387702973493562844620+00	0.62494800423006981141695780-01
72	0.49998431411075658897628900+00	0.62494944877266387481900900-01
73	0.49998473361327893149889470+00	0.62495083405993301774307460-01
74	0.49998513646871444389858050+00	0.62495216328555620635385420-01
75	0.49998552354753989555587000+00	0.62495343943105851189052470-01
76	0.49998589566402131313076270+00	0.62495466528249320491956040-01
77	0.49998625358051677657882820+00	0.62495584344561381116599000-01
78	0.49998659801139962346524640+00	0.62495697635969030566714540-01
79	0.49998692926663987058905350+00	0.62495806631010613961025700-01
80	0.49998724905507747365108150+00	0.62495911543985748666603890-01
81	0.49998755688741737239537490+00	0.62496012576006266900662000-01
82	0.49998785367897303053181260+00	0.62496109915957792638830890-01
83	0.49998813995218232968472900+00	0.62496203741380530970683220-01
84	0.49998841619891715908902880+00	0.62496294219276932887863060-01
85	0.49998866828826058173914092D+00	0.62496381506853090526296560-01
86	0.49998894044018537244231550+00	0.62496465752200003462669330-01
87	0.49998918928389937766567310+00	0.62496547094920224011749820-01
88	0.49998942980295479197062460+00	0.62496625666704828408756540-01
89	0.49998966236505057038048590+00	0.62496701591865162479387640-01
90	0.49998988731778916389181310+00	0.62496774987823367258530750-01
91	0.49999010498998107152682920+00	0.62496845965565295379479030-01
92	0.49999031569285160940919280+00	0.62496914630059077143233140-01
93	0.49999051972115818727787720+00	0.62496981080642280958818350-01
94	0.49999071735422560018384380+00	0.62497045411380331927055380-01
95	0.49999090885690614170861970+00	0.62497107711398600881171530-01
96	0.49999109448047071571510210+00	0.62497168065190350835475360-01
97	0.49999127446343655832302730+00	0.62497226552902525579844230-01
98	0.49999144903233667340357460+00	0.62497283250601183504392530-01
99	0.49999161840243562716707900+00	0.62497338[illegible]51821636937157D-01

TABLE 2

Recursion coefficients for the polynomials $\{\pi_k(\cdot; w_3)\}$

n	alpha(n)	beta(n)
0	0.37037037037037037037D-01	0.16000000000000000000D+02
1	0.35811288669783060917D+00	0.66282578875171467764D-02
2	0.44293596764346020311D+00	0.41154551017361395415D-01
3	0.46925333322630284096D+00	0.51741782003936091009D-01
4	0.48077976287000283670D+00	0.56064845754865753829D-01
5	0.48684740638240343111D+00	0.58228578157740470460D-01
6	0.49043291987640869651D+00	0.59461384764548385798D-01
7	0.49272768237210454609D+00	0.60229080854911299451D-01
8	0.49428489903347684509D+00	0.60739066803493998597D-01
9	0.49539008804668943172D+00	0.61094906789182697675D-01
10	0.49620279412350964003D+00	0.61352955641236678499D-01
11	0.49681787613674196493D+00	0.61545998375888748886D-01
12	0.49729461823511554334D+00	0.61694155907229941430D-01
13	0.49767162298766581034D+00	0.61810329736885846614D-01
14	0.49797490489096250254D+00	0.61903100105428254212D-01
15	0.49822251490262493942D+00	0.61978352734467430133D-01
16	0.49842729720309510051D+00	0.62040233674523126765D-01
17	0.49859859328823678290D+00	0.62091731596138376834D-01
18	0.49874332901963438146D+00	0.62135044810738308068D-01
19	0.49886672742402863250D+00	0.62171819377072260690D-01
20	0.49897278759666604684D+00	0.62203307551568758377D-01
21	0.49906461349768728317D+00	0.62230475638549453317D-01
22	0.49914464410949686541D+00	0.62254078896209909882D-01
23	0.49921481738564071689D+00	0.62274714515063386935D-01
24	0.49927668889965310293D+00	0.62292859708354606101D-01
25	0.49933151895604900098D+00	0.62308899510226347107D-01
26	0.49938033739401414223D+00	0.62323147341070148965D-01
27	0.49942399238220700438D+00	0.62335860413249654858D-01
28	0.49946318757058857679D+00	0.62347251405111280344D-01
29	0.49949851066980882148D+00	0.62357497401578225918D-01
30	0.49953045564672500675D+00	0.62366746808956619876D-01
31	0.49955944011544773239D+00	0.62375124751988228320D-01
32	0.49958581907688664129D+00	0.62382737322232212745D-01
33	0.49960989585754508391D+00	0.62389674948882016821D-01
34	0.49963193088162418308D+00	0.62396015093213554691D-01
35	0.49965214875344591167D+00	0.62401824417420930673D-01
36	0.49967074401221995461D+00	0.62407160541823668615D-01
37	0.49968788583618920278D+00	0.62412073477358958151D-01
38	0.49970372190980540385D+00	0.62416606800160811616D-01
39	0.49971838161991717404D+00	0.62420798619957045131D-01
40	0.49973197871081541506D+00	0.62424682382629097021D-01
41	0.49974461350037973769D+00	0.62428287538611475284D-01
42	0.49975637473833709910D+00	0.62431640102160245772D-01
43	0.49976734117120030262D+00	0.62434763121387056841D-01
44	0.49977758286563735392D+00	0.62437677074965693541D-01
45	0.49978716233197190813D+00	0.62440400208298018187D-01
46	0.49979613548158744430D+00	0.62442948819471456645D-01
47	0.49980455244572020271D+00	0.62445337503398093115D-01
48	0.49981245827811267119D+00	0.62447579360980611775D-01
49	0.49981989355998203428D+00	0.62449686178915178922D-01
50	0.49982689492252312177D+00	0.62451668584748955026D-01
51	0.49983349549954855405D+00	0.62453536181008802223D-01
52	0.49983972532074252462D+00	0.62455297661568124452D-01
53	0.49984561165426967645D+00	0.62456960912889718816D-01
54	0.49985117930605904348D+00	0.62458533102349863062D-01
55	0.49985645088191383563D+00	0.62460020755493639085D-01
56	0.49986144701763254412D+00	0.62461429823778722643D-01
57	0.49986618658152699341D+00	0.62462765744122685831D-01
58	0.49987068685305789539D+00	0.62464033491367810709D-01
59	0.49987496368075360253D+00	0.62465237624609943064D-01
60	0.49987903162211334570D+00	0.62466382328197957392D-01
61	0.49988290406780632029D+00	0.62467471448093066886D-01
62	0.49988659335214961618D+00	0.62468508524178538288D-01
63	0.49989011085157064972D+00	0.62469496819027145677D-01
64	0.49989346707252485814D+00	0.62470439343563310277D-01

TABLE 2 (*continued*)

n	alpha(n)	beta(n)
65	0.4998966717301399239 4D+00	0.6247133887999713757 4D-01
66	0.4998997338186879236 4D+00	0.6247219800235684068 2D-01
67	0.4999026616748417765 1D+00	0.6247301909490267255 2D-01
68	0.4999054630345482611 2D+00	0.6247380436866853652 0D-01
69	0.4999081450842434011 5D+00	0.6247455587634570473 2D-01
70	0.4999107145070444767 5D+00	0.6247527552569584521 5D-01
71	0.4999131775244740261 4D+00	0.6247596509165711061 7D-01
72	0.4999155399342030648 2D+00	0.6247662622728681292 0D-01
73	0.4999178071442417772 1D+00	0.6247726047366671844 5D-01
74	0.4999199842039547889 7D+00	0.6247786926888184382 3D-01
75	0.4999220758322336801 8D+00	0.6247845395617047758 6D-01
76	0.4999240864431206975 0D+00	0.6247901579133170769 0D-01
77	0.4999260201691438612 5D+00	0.6247955594946676088 7D-01
78	0.4999278808825941548 6D+00	0.6248007553112175029 3D-01
79	0.4999296722149496485 3D+00	0.6248057556789180862 5D-01
80	0.4999313975746287488 5D+00	0.6248105702753990795 9D-01
81	0.4999330601632348571 9D+00	0.6248152081867780591 2D-01
82	0.4999346629904371993 6D+00	0.6248196779505140462 2D-01
83	0.4999362088876171455 8D+00	0.6248239875946826988 3D-01
84	0.4999377005203957056 6D+00	0.6248281446740105379 3D-01
85	0.4999391404001458286 3D+00	0.6248321563029702660 1D-01
86	0.4999405308945824607 3D+00	0.6248360291862079310 1D-01
87	0.4999418742375138484 3D+00	0.6248397696465449494 1D-01
88	0.4999431725378291651 5D+00	0.6248433836507733840 9D-01
89	0.4999444277877900630 8D+00	0.6248468768334409946 0D-01
90	0.4999456418706870924 9D+00	0.6248502545188031079 0D-01
91	0.4999468165679159930 1D+00	0.6248535217411010073 2D-01
92	0.4999479535655235592 5D+00	0.6248566832633110565 0D-01
93	0.4999490544602680432 0D+00	0.6248597435944949432 3D-01
94	0.4999501207652348126 2D+00	0.6248627070058690547 9D-01
95	0.4999511539150441819 8D+00	0.6248655775456999141 2D-01
96	0.4999521552706849203 1D+00	0.6248683590531226681 7D-01
97	0.4999531261240038750 1D+00	0.6248710551709706973 2D-01
98	0.4999540677018793938 9D+00	0.6248736693576963954 0D-01
99	0.4999549811701037455 6D+00	0.6248762048984559532 6D-01

BIBLIOGRAPHY

1. K. M. Dempsey, D. Liu, and J. P. Dempsey, *Plana's summation formula for* $\sum_{m=1,3,\ldots}^{\infty} m^{-2}\sin(m\alpha)$, $m^{-3}\cos(m\alpha)$, $m^{-2}A^m$, $m^{-3}A^m$, Math. Comp. **55** (1990), 693–703.
2. W. Gautschi, *On the preceding paper "A Legendre polynomial integral" by James L. Blue*, Math. Comp. **33** (1979), 742–743.
3. ——, *Minimal solutions of three-term recurrence relations and orthogonal polynomials*, Math. Comp. **36** (1981), 547–554.
4. ——, *On generating orthogonal polynomials*, SIAM J. Sci. Statist. Comput. **3** (1982), 289–317.
5. ——, *Questions of numerical condition related to polynomials*, Studies in Numerical Analysis (G. H. Golub, ed.), Math. Assoc. Amer., 1984, pp. 140–177.
6. ——, *A class of slowly convergent series and their summation by Gaussian quadrature*, Math. Comp. **57** (1991), 309–324.
7. W. Gautschi and G. V. Milovanović, *Gaussian quadrature involving Einstein and Fermi functions with an application to summation of series*, Math. Comp. **44** (1985), 177–190.
8. E. R. Hansen, *A table of series and products*, Prentice-Hall, Englewood Cliffs, NJ, 1975.
9. A. McLellan IV, *Tables of the Riemann zeta function and related functions*, Math. Comp. **22** (1968), Review **69**, 687–688.

DEPARTMENT OF COMPUTER SCIENCES, PURDUE UNIVERSITY, WEST LAFAYETTE, INDIANA 47907
E-mail address: wxg@cs.purdue.edu

30.5. [149] (with J. Waldvogel) "Contour Plots of Analytic Functions"

---

[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, 1997. [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.]

Chapter 25. Contour Plots of Analytic Functions

W. Gautschi and J. Waldvogel

25.1 Introduction

There are two easy ways in MATLAB to construct contour plots of analytic functions, i.e., lines of constant modulus and constant phase. One is to use the MATLAB `contour` command for functions of two variables, another to solve the differential equations satisfied by the contour lines. This is illustrated here for the function $f(z) = e_n(z)$, where

$$e_n(z) = 1 + z + \frac{z^2}{2!} + \cdots + \frac{z^n}{n!} \tag{25.1}$$

is the nth partial sum of the exponential series. The lines of constant modulus 1 of e_n are of interest in the numerical solution of ordinary differential equations, where they delineate regions of absolute stability for the Taylor expansion method of order n and also for any n-stage explicit Runge-Kutta method of order n, $1 \le n \le 4$ (cf. [4, §9.3.2]).

25.2 Contour Plots by the contour Command

Let f be analytic and $f(z) = re^{i\varphi}$. We may consider the modulus r as a function of two variables x, y, where $z = x + iy$; similarly for the phase φ, $-\pi < \varphi \le \pi$. Hence, we can apply the MATLAB command `contour` to r and φ to obtain the lines of constant modulus and phase.

In the MATLAB program below, the set of all x- and y-values is collected (in true MATLAB spirit) in a matrix `a`, which is operated upon to compute the desired values of r and φ for $f = e_n$ as input matrices to the routine `contour`.

The program begins with the definitions of the mesh `h` and the number `nmax` of contour plots to be generated. The vector `bounds` contains common lower and upper bounds for the x- and y-coordinates applicable for all plots. The bounds used here have been chosen to accommodate contour plots of the first four exponential sums. Then the contour levels `vabs0` and `vang0` for the modulus and phase of $f(z)$ are defined. The last preparatory step is generating the vectors `x` and `y` containing the discrete x- and y-values to be used in the matrix `a` of grid points. In the loop over `n` the values `f` of e_n on the entire grid are

generated by almost the same statements that would evaluate e_n at a single point, where `t` stands for an individual term of the series (25.1). The only difference is the statement `t=t.*a/n`, in which the operation symbol `.*` invokes the element-by-element product of the matrices `t` and `a`. The last line of the program (here turned off by the comment sign %) generates the encapsulated postscript file `fign.eps` of `figure(n)`, ready to be printed or incorporated into a text file.

```
% Contour plots of the first nmax exponential sums (Figure 1)
%
>> h = 1/64;  nmax = 4;  bounds = [-3.25  .75  -3.375  3.375];
>> vabs0 = [0:.1:1];  vang0 = [-.875:.125:1]*pi;
>> x = bounds(1):h:bounds(2);  y=bounds(3):h:bounds(4);
>> a = ones(size(y'))*x + i*y'*ones(size(x));
% Next line: a shorter way of generating a (more memory!)
>> %  [xx,yy]=meshgrid(x,y); a=xx+i*yy;
>> t = ones(size(a));  f = t;
>> for n = 1: nmax
>>   if n <= 2, vabs = vabs0; vang = vang0;
>>     elseif n == 3, vabs = [vabs0 .47140452]; vang = vang0;
>>     else vabs = [vabs0 .58882535 .27039477];
>>          vang = [vang0 1.48185376 -1.48185376];
>>   end;
>>   t = t.*a/n;  f = f + t;
>>   figure(n);  clf;  hold on;
>>   axis(bounds);  axis image;
>>   contour(x, y, abs(f), vabs);
>>   contour(x, y, angle(f), vang);
>> end;
>> % figure(n); print -deps fign;
```

The results for $n = 1:1:4$[1] are shown in the plots below. Clearly visible are the n zeros of e_n from which emanate the lines of constant phase. Near these zeros, the lines of constant modulus become circle-like with radii tending to 0 as the zeros are approached. The contour lines are for $r = .1:.1:1$ and $\varphi = -\frac{7}{8}\pi : \frac{1}{8}\pi : \pi$.

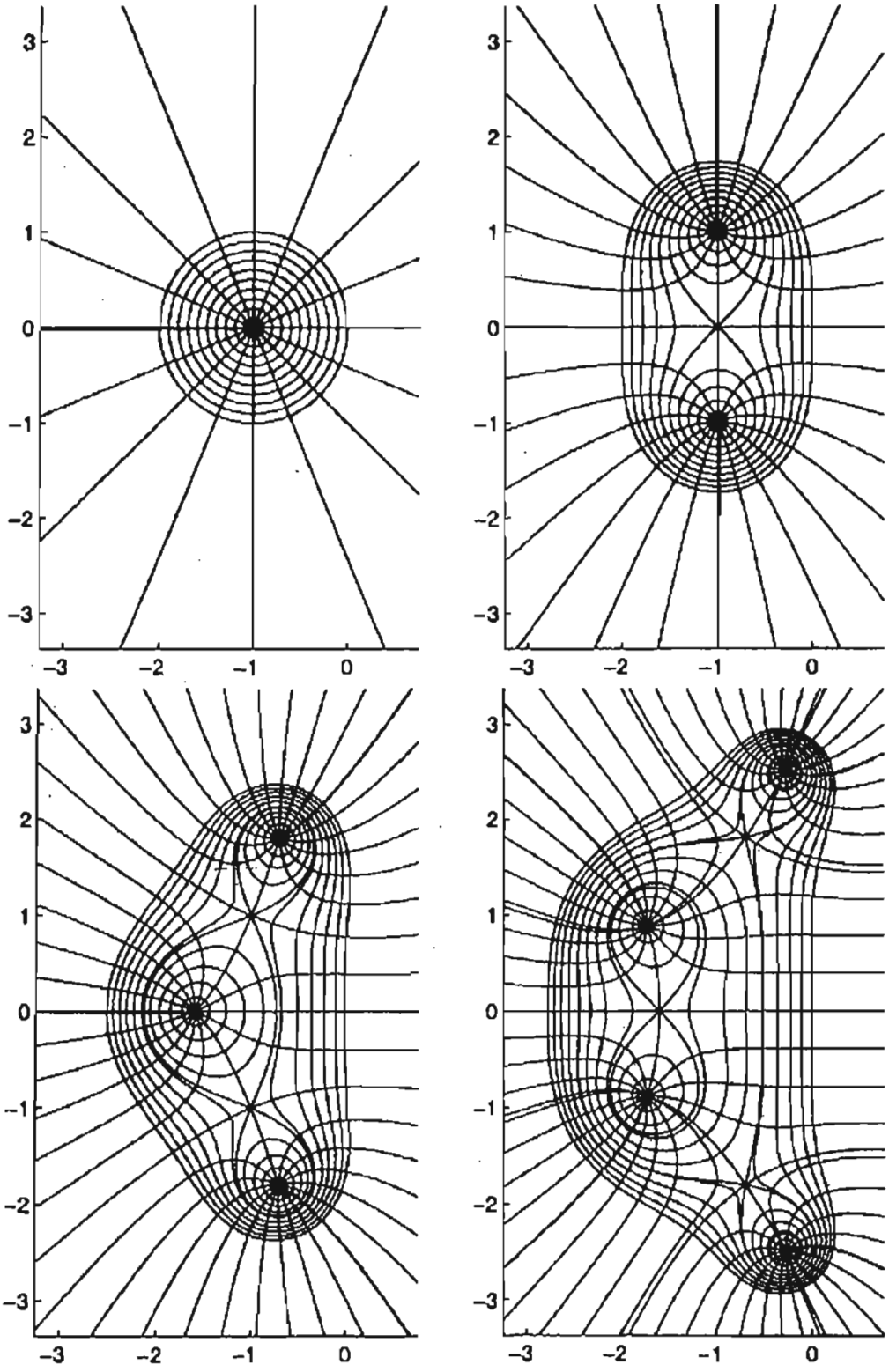
At points z_0 where $e_n'(z_0) = e_{n-1}(z_0) = 0$, $n \geq 2$, two lines of constant modulus intersect (cf. §3.1 below). The respective r-values are $r = |e_n(z_0)|$, or $r = |z_0|^n/n!$, since

$$e_n(z) = e_{n-1}(z) + \frac{z^n}{n!}. \tag{25.2}$$

These critical lines are also included in the plots (see the `if` statement of the program). When $n = 2$, they go through $z_0 = -1$, where $r = \frac{1}{2}$, while for $n = 3$ and $n = 4$, one has to 8 decimal digits: $z_0 = -1 \pm i$, $r = \sqrt{2}/3 = .47140452$ and $z_0 = -.70196418 \pm 1.80733949i$, $r = (1.93887332)^4/24 = .58882535$, $z_0 = -1.59607164$, $r = .27039477$, respectively.

[1] This MATLAB notation stands for $n = 1, 2, 3, 4$.

FIGURE 25.1.
Contour Plots of the First 4 Exponential Sums



What's good for the r-lines is good for the φ-lines! The singular points for them are also the zeros z_0 of e_n' (cf. §3.2), to which there correspond φ-values defined by $e_n(z_0)/|e_n(z_0)| = (z_0/|z_0|)^n = e^{i\varphi}$, i.e., $\varphi = n \arg z_0$. Thus, for $n = 2$, we have $\varphi = 0 \pmod{2\pi}$, whereas for $n = 3$ we get $\varphi = \pm\frac{\pi}{4}$ corresponding to $z_0 = -1 \pm i$, respectively. All three of these φ-values are included among the values already listed above. For $n = 4$, the two complex values of z_0 shown in the previous paragraph yield $\varphi = 4 \arg z_0 = \pm 1.48185376 \pmod{2\pi}$, and the real value of z_0 yields $\varphi = 0$. These critical φ-lines are also shown in the plots in Figure 25.1

The figure was generated by means of the step size `h=1/64` in order to obtain a good resolution, even for the "branch cuts" corresponding to `|angle(f)|=pi`. The choice `h=1/32` is a good compromise, whereas `h=1/16` is very fast while still producing satisfactory plots.

25.3 Differential Equations

For an analytic function f, let

$$w = f(z), \quad w = re^{i\varphi}, \quad z = x + iy. \tag{25.3}$$

25.3.1 Contour Lines $r = \text{const.}$

To describe the lines $r = \text{const}$, it is natural to take φ as independent variable. Differentiating

$$f(z(\varphi)) = re^{i\varphi}, \quad r = \text{const}, \tag{25.4}$$

with respect to φ then gives $f'(z)\frac{dz}{d\varphi} = ire^{i\varphi} = if(z)$, that is,

$$\frac{dz}{d\varphi} = i\, q(z)\,, \quad \text{where} \quad q(z) = \frac{f(z)}{f'(z)}. \tag{25.5}$$

With s the arc length, one has

$$\frac{ds}{d\varphi} = \sqrt{\left(\frac{dx}{d\varphi}\right)^2 + \left(\frac{dy}{d\varphi}\right)^2} = \left|\frac{dz}{d\varphi}\right| = |q(z)|,$$

so that

$$\frac{dz}{ds} = \frac{dz}{d\varphi}\frac{d\varphi}{ds} = i\,\frac{q(z)}{|q(z)|}. \tag{25.6}$$

Written as a system of differential equations, this is

$$\begin{aligned} \frac{dx}{ds} &= -\operatorname{Im}\left\{\frac{q(z)}{|q(z)|}\right\}, \\ \frac{dy}{ds} &= \operatorname{Re}\left\{\frac{q(z)}{|q(z)|}\right\}, \end{aligned} \qquad z = x + iy. \tag{25.7}$$

If we are interested in a contour line crossing the real axis, we must find an initial point $x(0) = x_r$, $y(0) = 0$ for (25.7) with real x_r such that $f(x_r) = r$ (assuming $f(x)$ real for real x). In the case $f(x) = e_n(x)$, this is easy if $r \geq 1$, since $e_n(0) = 1$ and $e_n(x)$ monotonically increases for $x \geq 0$. There is thus a unique $x_r \geq 0$ such that $e_n(x_r) = r$. If $0 < r < 1$, this is still possible when n is odd. Then, $e_n'(x) = e_{n-1}(x) > 0$, since all zeros of e_m, when m is even, are known to be complex [3] (cf. also [1]). Thus, e_n monotonically increases from $-\infty$ to $+\infty$ as x increases from $-\infty$ to $+\infty$, and there is a unique $x_r < 0$ such that $e_n(x_r) = r$. When n is even, we have $e_n(x) > 0$ for all real x, and $e_n' = e_{n-1}$ vanishes at exactly one point $x_0 < 0$, where e_n has a minimum (cf. [2]). Owing to (25.2) and $e_{n-1}(x_0) = 0$, we have $e_n(x_0) = x_0^n/n!$, and there is a solution $x_r < 0$ of $e_n(x_r) = r$ if and only if $r \geq x_0^n/n!$. For smaller positive values of r, one must find a complex initial point $x(0)$, $y(0) > 0$ near one of the complex zeros of e_n.

The point z_0 where $f'(z_0) = 0$ is a singular point of (25.7), a point where two r-lines intersect at a right angle. This requires special care to get the integration of (25.7) started in all four directions. The initial point, of course, is z_0, that is, $x(0) = \operatorname{Re} z_0$, $y(0) = \operatorname{Im} z_0$. What needs some analysis is the value of the right-hand side of (25.7) at z_0. Let $h(z) = (z - z_0)\,q(z)$; then h is smooth near z_0 and has the Taylor expansion

$$h(z) = \frac{f_0 + \frac{1}{2}(z - z_0)^2 f_0'' + \cdots}{f_0'' + \frac{1}{2}(z - z_0) f_0''' + \cdots}, \quad h(z_0) = \frac{f_0}{f_0''},$$

where $f_0 = f(z_0)$, etc. (we assume $f_0 \neq 0$ and $f_0'' \neq 0$). Letting

$$\frac{z - z_0}{|z - z_0|} = e^{i\theta}, \quad -\frac{1}{2}\pi < \theta \leq \frac{1}{2}\pi, \quad h(z_0)/|h(z_0)| = e^{i\omega}, \quad -\pi < \omega \leq \pi,$$

(being mindful that to each θ there is a $\theta + \pi$ corresponding to the backward continuation of the line), we then have

$$\frac{q(z)}{|q(z)|} = \frac{|z - z_0|}{z - z_0} \frac{h(z)}{|h(z)|} \to e^{i(\omega - \theta_0)} \quad \text{as } z \to z_0,$$

where $\theta_0 = \lim_{z \to z_0} \theta$. It remains to determine θ_0.
Along an r-line through z_0, we have

$$\begin{aligned} r^2 = |f(z)|^2 &= |f_0 + \tfrac{1}{2}(z - z_0)^2 f_0'' + \cdots|^2 \\ &= |f_0|^2 + \operatorname{Re}[(z - z_0)^2 f_0'' \overline{f_0}] + O(|z - z_0|^3). \end{aligned}$$

Since $|f_0|^2 = r^2$, this gives

$$\operatorname{Re}\left\{\left(\frac{z - z_0}{|z - z_0|}\right)^2 f_0'' \overline{f_0}\right\} = O(|z - z_0|),$$

hence, as $z \to z_0$,

$$\operatorname{Re}(e^{2i\theta_0} f_0'' \overline{f_0}) = 0.$$

Therefore,

$$\tan 2\theta_0 = \frac{\operatorname{Re}(f_0'' \overline{f_0})}{\operatorname{Im}(f_0'' \overline{f_0})}. \tag{25.8}$$

There are exactly two solutions in $-\frac{1}{2}\pi < \theta_0 \le \frac{1}{2}\pi$, which differ by $\frac{1}{2}\pi$, confirming the orthogonality of the two r-lines through z_0.

Note that in the case $f(z) = e_n(z)$, we have $f'(z) = e_{n-1}(z)$, so that z_0 is a zero of e_{n-1}. This is clearly visible in the plots of §2. Furthermore, $f_0 = e_n(z_0)$, $f_0'' = e_{n-2}(z_0)$ if $n \ge 2$, so that (25.2) with $z = z_0$, once applied as is, and once with n replaced by $n-1$, gives $f_0 = z_0^n/n!$, $f_0'' = -z_0^{n-1}/(n-1)!$, and the equation for θ_0 reduces to

$$\tan 2\theta_0 = -\frac{\operatorname{Re} z_0}{\operatorname{Im} z_0} \quad (f = e_n).$$

25.3.2 Contour Lines $\varphi = \text{const.}$

For the lines $\varphi = \text{const}$, we take r as the independent variable and, by differentiating

$$f(z(r)) = re^{i\varphi}, \quad \varphi = \text{const},$$

with respect to r, obtain

$$\frac{dz}{dr} = \frac{e^{i\varphi}}{f'(z)}.$$

In terms of the arc length s, we now have

$$\frac{ds}{dr} = \left|\frac{dz}{dr}\right| = \frac{1}{|f'(z)|},$$

so that

$$\frac{dz}{ds} = \frac{dz}{dr}\frac{dr}{ds} = e^{i\varphi}\,\frac{|f'(z)|}{f'(z)},$$

or, written as a system of differential equations,

$$\begin{aligned} \frac{dx}{ds} &= \operatorname{Re}\left\{e^{i\varphi}\,\frac{|f'(z)|}{f'(z)}\right\}, \\ \frac{dy}{ds} &= \operatorname{Im}\left\{e^{i\varphi}\,\frac{|f'(z)|}{f'(z)}\right\}, \end{aligned} \qquad z = x + iy. \tag{25.9}$$

The singular point of (25.9) is again z_0, a zero of f'. At this point,

$$\frac{f_0}{|f_0|} = e^{i\varphi}, \quad -\pi < \varphi \le \pi,$$

which determines φ. The limit of $|f'(z)|/f'(z)$ as $z \to z_0$ may be determined by a procedure similar to the one in §3.1. Instead, we directly use the Taylor series of f in z_0 in order to study the φ-lines (and the r-lines as well) near z_0 with $f'(z_0) = 0$. Let $z = z_0 + \zeta$, where ζ is a complex increment, and let

$$f_k := f^{(k)}(z_0), \quad k \geq 0, \qquad f_0 \neq 0, \quad f_1 = 0, \quad f_2 \neq 0, \tag{25.10}$$

be the derivatives of f at z_0. Then the Taylor series is

$$f(z_0 + \zeta) = f_0 + f_2 \frac{\zeta^2}{2!} + f_3 \frac{\zeta^3}{3!} + \cdots . \tag{25.11}$$

Next, we observe that by defining $w = f_0\, e^u$ in (25.3), i.e., by putting

$$f(z_0 + \zeta) = f(z_0)\, e^u, \tag{25.12}$$

the r-lines through z_0 are given by the values of ζ corresponding to purely imaginary values $u = it$, whereas the φ-lines through z_0 are given by $u \in R$. The point z_0 itself corresponds to $\zeta = u = 0$. We therefore need to solve Equ. (25.12), with $f(z_0 + \zeta)$ substituted from (25.11), for ζ, which is a typical task for MAPLE.

In the program below[2] the series (25.11) and the equation (25.12) are denoted by `s` and `eq`, respectively. The `solve` command automatically expands e^u in a Taylor series and solves the equation by means of a series progressing in appropriate powers of u (here half-integer powers). As expected, two solutions corresponding to the two possible values of the square root are found. Only the first solution `zet0[1]` is processed further: first by substituting the abbreviations `fk` defined in Equ. (25.10), then by introducing the variable `v` according to

$$u = \frac{v^2 f_2}{2 f_0} \quad \text{or} \quad v = \left(\frac{2u f_0}{f_2}\right)^{\frac{1}{2}} . \tag{25.13}$$

The symbols `D(f)` and `(D@@k)(f)` stand for the derivative of f and the kth derivative of f, respectively. The call to the function `map` causes the operation defined by its first argument, here the simplification of the radicals, to be applied to each term of the expression defined by the second argument. Finally, the call to `series` causes the O-term to be simplified.

```
> N := 5:  Order := N:
> s := series(f(z0 + dz), dz):
> s0 := subs(D(f)(z0) = 0, s);
```

$$s0 := \mathrm{f}(z0) + \frac{1}{2}(D^{(2)})(f)(z0)\, dz^2 + \frac{1}{6}(D^{(3)})(f)(z0)\, dz^3 + \frac{1}{24}(D^{(4)})(f)(z0)\, dz^4 + \mathrm{O}(dz^5)$$

[2]The authors are indebted to Dominik Gruntz for this program.

```
> eq := s0 = f(z0)*exp(u):
> zet0 := solve(eq, dz);
```

$$zet0 := \frac{f(z0)\sqrt{2}\sqrt{u}}{\sqrt{\%1}} - \frac{1}{3}\frac{(D^{(3)})(f)(z0)\,f(z0)\,u}{(D^{(2)})(f)(z0)^2} + \frac{1}{288}($$
$$40\,f(z0)^3\,(D^{(3)})(f)(z0)^2 - 24\,(D^{(2)})(f)(z0)\,f(z0)^3\,(D^{(4)})(f)(z0)$$
$$+ 72\,(D^{(2)})(f)(z0)^3\,f(z0)^2)\sqrt{2}\,u^{3/2}\Big/((D^{(2)})(f)(z0)^2\,\%1^{3/2}) + O(u^2),$$

$$-\frac{f(z0)\sqrt{2}\sqrt{u}}{\sqrt{\%1}} - \frac{1}{3}\frac{(D^{(3)})(f)(z0)\,f(z0)\,u}{(D^{(2)})(f)(z0)^2} - \frac{1}{288}(40\,f(z0)^3\,(D^{(3)})(f)(z0)^2$$
$$- 24\,(D^{(2)})(f)(z0)\,f(z0)^3\,(D^{(4)})(f)(z0) + 72\,(D^{(2)})(f)(z0)^3\,f(z0)^2)\sqrt{2}\,u^{3/2}$$
$$\Big/((D^{(2)})(f)(z0)^2\,\%1^{3/2}) + O(u^2)$$
$$\%1 := f(z0)\,(D^{(2)})(f)(z0)$$

```
> zet1 := subs(seq( (D@@k)(f)(z0) = f.k, k=0..N-1), zet0[1]):
> zet2 := map(radsimp, subs(u = v^2*f2/2/f0, zet1)):
> zeta := series(zet2, v);
```

$$\zeta := v - \frac{1}{6}\frac{f3}{f2}v^2 - \frac{1}{72}\frac{-5\,f0\,f3^2 + 3\,f2\,f0\,f4 - 9\,f2^3}{f0\,f2^2}v^3 + O(v^4)$$

The MAPLE program works for any $N \geq 3$, producing $N-2$ terms of the above series. However, it is fairly slow, since no "intelligence", such as information on the form of the resulting series, is built in. To be able to find this series, nevertheless, is a good accomplishment of a general-purpose symbolic manipulator. It can be seen that ζ may be written as a formal power series in the variable v defined in (25.13). If the original series (25.11) converges in a neighborhood of z_0, the resulting series converges in a neighborhood of $v = 0$.

The directions θ_0 of the r-lines at z_0 are now given by the values of ζ corresponding to $u = it$ in the limit $t \to 0$. The above series and Equ. (25.13) immediately yield

$$\theta_0 = \arg v = \frac{1}{2}(\arg f_0 - \arg f_2 \pm \frac{\pi}{2})\,.$$

Hence there are two r-lines through z_0 intersecting at a right angle, in perfect agreement with Equ. (25.8).

The directions of the φ-lines through z_0, on the other hand, are given by (25.13) for real values of u. We obtain the two directions $\theta_0 \pm \frac{\pi}{4}$, i.e., the tangents of the two φ-lines through z_0 are the bisectors of the tangents of the r-lines.

25.4 The Contour Lines $r = 1$ of $f = e_n$

As indicated in §3.1, we need to solve (25.7) with initial values $x = y = 0$. Let S_n be the point of intersection of the 1-line of e_n with the negative real

axis. By symmetry, only the portion of each 1-line lying in the upper half of the complex plane needs to be computed.

We will discuss two implementations of this process: first a naive approach only using termination of the numerical integration at a precomputed value s_f of the independent variable (as available in the previous version MATLAB 4). At the end of this section we will present a simplified algorithm taking advantage of the Events capability of the current version MATLAB 5. For both implementations a good upper bound s_f for the arc length on the 1-line between the origin and S_n is needed.

Such an upper bound s_f may be obtained as follows. We first observe that the region $|e_n(z)| \leq 1$ approaches a semidisk of radius $\rho(n)$ as $n \to \infty$. An asymptotic analysis shows that

$$\rho(n) = \exp(-1) \cdot (n + \log\sqrt{2\pi n} + O(1)).$$

A good empirical choice of $O(1)$ is 3; then we obtain

$$s_f = (1 + \frac{\pi}{2}) \exp(-1) \cdot (n + \log\sqrt{2\pi n} + 3) \tag{25.14}$$

as a close upper bound for the arc length up to the point S_n for $n \geq 1$.

If the Events capability is not used, we first integrate the differential equations up to the final value s_f. Then, only the points satisfying the condition $y \geq 0$ need to be plotted. The point S_n can be approximated by linear interpolation between the two points on the 1-line closest to S_n.

In the MATLAB program below this is done by using the `find` command with the parameter $y \geq 0$ in order to find the subset of all points satisfying the condition $y \geq 0$. Their indices are collected in the vector `indices`. The indices of the points used in linear interpolation are then `l=length(indices)` and `l1=l+1`. Finally, `w` is the normalized row vector containing the two interpolation weights, and the actual interpolation is carried out by the product `w*z(l:l1,:)`.

```
% Level curves r = 1 for the first 21 exponential sums (Figure 2)
>> global n
>> nmin = 1;  nmax = 21;  tol = 3.e-8;
>> axis equal;  hold on; % arc = [];
>> options = odeset('RelTol',tol,'AbsTol',tol);
>> for n = nmin:nmax,
>>    sf = 0.94574*(n +.5*log(2*pi*n) + 3);
% 0.94574=(1+pi/2)*exp(-1)
>>    [s,z] = ode45('level4', [0 sf], [0; 0], options);
>>    indices = find(z(:, 2) >= 0);
>>    l = length(indices);  l1 = l + 1;
>>    w = [-z(l1, 2), z(l, 2)]; w = w/sum(w);
>>    z(l1,:) = w*z(l:l1,:);
>>    plot(z(1:l1, 1), z(1:l1, 2));
% approximate arclengths and bounds sf
>> %  arc = [arc; w*s(l:l1), sf];
>> end;
```

ALGORITHM 25.1. *level4.m*

```
function zdot = level4(s, z)
%LEVEL4 generates the right-hand side of
%       the differential equation for the 1-lines of
%       f(z) = 1 + z + z^2/2! + ... + z^n/n!

global n

zc = z(1) + i*z(2);  t = 1;  f0 = 0; f = t;
for k = 1: n, t = t*zc/k; f0 = f;  f = f + t; end;
q = f/f0; zdot = [-imag(q); real(q)]/abs(q);
```

The MATLAB program begins with the definitions of the parameters `nmin`, `nmax` and the error tolerance `tol` to be used in the definition of the integrator options structure, `options`, by means of `odeset`. The 1-lines in the range `nmin` $\le n \le$ `nmax` are generated and plotted in the accuracy given by `tol`. Rerunning the program with new values of `nmin` and `nmax` adds new curves to the figure. The statements turned off by the comment marks % generate a table `arc` containing the actual arc lengths and the upper bounds s_f computed from (25.14).

The actual integration is done in the call to the integrator `ode45`. The input parameters of this procedure are: the name `'level4'` (in string quotes) of the M-file defined in Alg. 25.1 according to Equ. (25.7), a vector containing the initial value 0 and the final value `sf` of the independent variable, the column vector `[0;0]` of the initial values, and the options structure, `options`. The choice `tol = 3.0e-8` yields a high-resolution plot, whereas the default `RelTol = 1.0e-3`, `AbsTol = 1.0e-6` (when the parameter `options` is omitted in the call) still yields a satisfactory plot. The values of the independent and dependent variables generated by the integrator are stored as the vectors `s` and `z`, respectively, ready to be plotted.

The results for $n = 1 : 1 : 21$ are shown in Figure 25.2 below. The features near the imaginary axis at the transition to the circular part seem to show a periodicity in n of a little over 5. For example, the curves corresponding to $n = 5, 10, 15, 21$ all show a particularly large protrusion into the right half-plane.

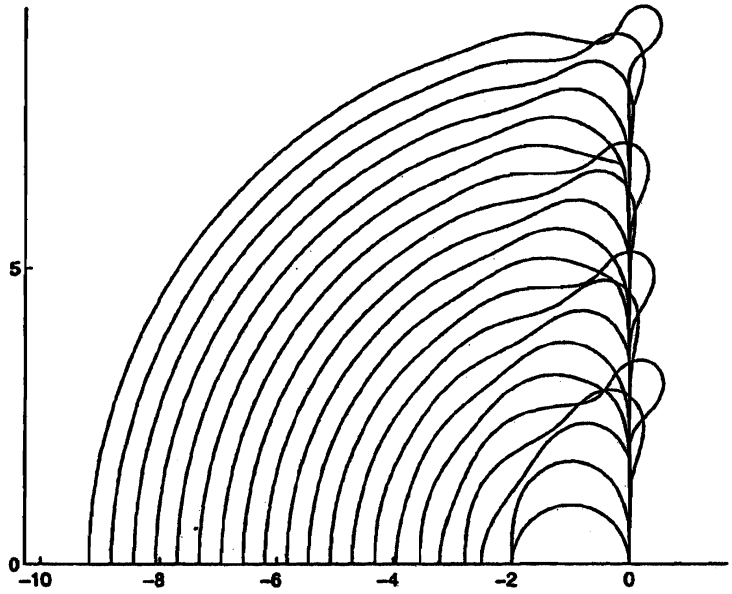
An investigation of this phenomenon is interesting, but exceeds the scope of this article. We limit ourselves to reporting that as $n \to \infty$, the period tends to

$$\frac{2\pi}{\frac{\pi}{2} - \exp(-1)} = 5.22329130.$$

This result was obtained by considering the function $e_\nu(z)$ for real values of ν (which leads to the incomplete gamma function) and requiring the 1-line of $e_\nu(z)$ to contain a saddle point with $e'_\nu(z) = 0$.

The `Events` capability of MATLAB 5 allows to stop a numerical integration at an "event", i.e. if a so-called "event function" passes through zero. The

FIGURE 25.2.
Level Curves $r = 1$ for the First 21 Exponential Sums



event function, in our case the imaginary part `imag(z)` of the complex dependent variable, has to be defined in the function `level` (Alg. 25.2), which also defines the right-hand sides `zdot` of the differential equations. This function needs to be coded with the additional input parameter `flag` and the additional output parameters `isterminal` and `direction` in such a way that `zdot` is the vector of the right-hand sides of the differential equations if `flag` is missing or undefined. If `flag` has the value `'events'`, however, the vector `zdot` must be defined as the event function, and the parameters `isterminal` (the indices of the relevant components of `zdot`) and `direction` (the direction of the zero passage) must be appropriately defined. Since `ode45` of MATLAB 5 allows to integrate complex-valued dependent variables, this simplification is taken advantage of in the program below.

```
% Level curves r = 1 for the first 21 exponential sums (Fig. 2)
% using the 'Events' capability and integration of complex
% dependent variables
>> global n
>> nmin = 1;  nmax = 21;  tol = 1e-6;
>> axis equal;  hold on; % arc = [];
>> options= odeset('RelTol',tol,'AbsTol',tol,'Events','on');
>> for n = nmin:nmax,
>>    sf = 0.94574*(n +.5*log(2*pi*n) + 3);
>>    [s,z] = ode45('level', [0 sf],0,options);
>>    plot(z);
% arclengths and bounds sf
>>    % arc = [arc; s(length(s)), sf];
>> end;
```

ALGORITHM 25.2. *level.m*

```
function [zdot, isterminal, direction] = level(s, z, flag)
%LEVEL generates the right-hand side of
%      the differential equation for the 1-lines of
%      f(z) = 1 + z + z^2/2! + ... + z^n/n!

global n

if nargin<3 | isempty(flag),
   t = 1;  f0 = 0; f = t;
   for k = 1:n, t = t*z/k; f0 = f;  f = f + t; end;
   q = f/f0; zdot = i*q/abs(q);
else
   switch(flag)
   case 'events'
      zdot= imag(z);
      isterminal= 1;
      direction= -1;
   otherwise
      error(['Unknown flag: ', flag]);
   end;
end
```

25.5 The Contour Lines $\varphi = \text{const}$ of $f = e_n$

Below is a MATLAB program that implements the method of §3.2 for any fixed $n > 0$, where the differential equations (25.9) must be implemented in the function phase and stored in the M-file phase.m (Alg. 25.3).

```
% Lines of constant phase for the 10th exponential sum (Fig 3)
%
>> global n phi
>> n = 10; tol = 1.e-5; sf = 1.5;
>> clf; axis([-6 6 -1 8]); hold on;
>> r = roots(1./gamma(n + 1:-1:1));
>> indices = find(imag(r) >= 0); zero = r(indices)
>> options= odeset('RelTol',tol,'AbsTol',tol);
>> for k = 1: length(zero),
>>    z0 = zero(k);
>>    for phi = -7/8*pi:pi/8:pi,
>>       [s,z] = ode45('phase', [0 sf], z0, options);
>>       plot(z);
>>    end;
>> end;
```

The program begins with the definitions of n, the error tolerance tol, and the desired arc length sf of the curve segments emanating from the zeros. Then, the vector r of the zeros of e_n is computed by means of the function roots, where the coefficients of e_n are generated by means of the gamma function. On the next line the subset of the zeros in the upper half-plane is formed by means

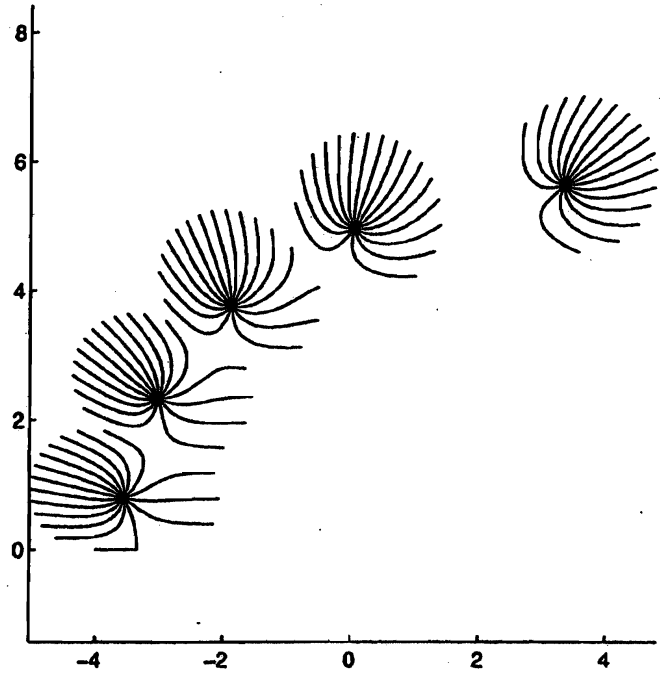
ALGORITHM 25.3. *phase.m*

```
function zdot = phase(s,z)

global n phi

eiphi = exp(i*phi);
t = 1; f = t;
for k=1:n-1, t = t*z/k; f = f+t; end;
zdot = eiphi*abs(f)/f;
```

of the `find` command with the argument `imag(r)>=0`. The statement used in the program stores all the indices defining the subset in the vector `indices`; then `r(indices)` is the vector of the zeros of e_n in the upper half-plane (which is printed for convenience).

The input parameters in the call to the integrator `ode45` are: the name `'phase'` (in string quotes) of the differential equations defined in Alg. 25.3 according to Equ. (25.9), a vector containing the initial value 0 and the final value `sf` of the independent variable, the complex initial value `z0`, and the options structure, `options` (defaults 10^{-3},10^{-6} when omitted). The values of the independent and dependent variables generated by the integrator are stored as the vectors `s` and `z`, respectively, ready to be plotted. The result for $n = 10$ is shown in Figure 25.3[3].

FIGURE 25.3.
Lines of Constant Phase for the 10th Exponential Sum



[3]We wrote and ran the script on August 1, 1996, while fireworks went off in celebration of the Swiss national holiday.

None of the complex singular points is in evidence in Figure 25.3 since the φ-values chosen do not correspond to level lines passing through a complex singular point. The real singular point at $z_0 = -3.333551485$, however, is clearly visible by an abrupt right-angled turn of the line $\varphi = 0$ (near the bottom of the figure). It is curious to note how the ode45 integrator was able to integrate right through the singularity, or so it seems.

Acknowledgments

The first author was supported, in part, by the US National Science Foundation under grant DMS-9305430.

References

[1] A.J. CARPENTER, R.S. VARGA, J. WALDVOGEL, *Asymptotics for the zeros of the partial sums of e^z. I*, Rocky Mountain J. Math., 21, 1991, pp. 99–120.

[2] K.E. IVERSON, *The zeros of the partial sums of e^z*, Math. Tables and Other Aids to Computation, 7, 1953, pp. 165–168.

[3] G. PÓLYA, G. SZEGÖ, *Problems and Theorems in Analysis*, Vol. II, Part V, Exercise 74, Springer-Verlag, New York, 1976.

[4] H.R. SCHWARZ, *Numerical Analysis. A Comprehensive Introduction*, John Wiley & Sons, Chichester, 1989.

30.6. [175] "The Hardy–Littlewood function: an exercise in slowly convergent series"

---

[175] "The Hardy–Littlewood function: an exercise in slowly convergent series," *J. Comput. Appl. Math*. **179**, 249–254 (2005).

Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER
Journal of Computational and Applied Mathematics 179 (2005) 249–254
JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS
www.elsevier.com/locate/cam

The Hardy–Littlewood function: an exercise in slowly convergent series

Walter Gautschi*

Department of Computer Sciences, Purdue University, West Lafayette, IN 47907-2066, USA

Received 8 September 2003

Dedicated to Olav Njåstad on the occasion of his 70th birthday

Abstract

The function in question is $H(x) = \sum_{k=1}^{\infty} \sin(x/k)/k$. In deference to the general theme of this conference, a summation procedure is first described using orthogonal polynomials and polynomial/rational Gauss quadrature. Its effectiveness is limited to relatively small (positive) values of x. Direct summation with acceleration is shown to be more powerful for very large values of x. Such values are required to explore a (in the meantime disproved) conjecture of Alzer and Berg, according to which $H(x)$ is bounded from below by $-\pi/2$.

Keywords: Hardy–Littlewood function; Slowly convergent series; Summation by polynomial/rational Gauss quadrature; Direct summation with acceleration

1. Introduction

Work on the complete monotonicity of certain functions involving the polygamma functions led Alzer et al. [2] to consider the function

$$H(x) = \sum_{k=1}^{\infty} \frac{1}{k} \sin \frac{x}{k} \tag{1}$$

* Tel.: +1 765 494 1995; fax: +1 765 494 0739.
E-mail address: wxg@cs.purdue.edu (W. Gautschi).
URL: http://www.cs.purdue.edu/people/wxg.

0377-0427/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2004.09.043

already studied by Hardy and Littlewood [7, Section 7] in connection with a summation procedure of Lambert. Hardy and Littlewood prove that the function is unbounded, there being infinitely many (though rare) positive values of x with $x \to \infty$ for which $H(x) > C(\log\log x)^{1/2}$. The complete monotonicity property alluded to above was shown by Alzer et al. [2] to be equivalent to the inequality $H(x) > -\pi/2$ for all $x > 0$. Although this inequality was eventually disproved by these authors, there may be some interest in studying the behavior of the function $H(x)$ numerically. Given the slow convergence of the series in (1), this is a challenging task in its own right.

We describe two procedures for computing $H(x)$. The first is one that has been used previously with some success (cf. [6, Section 4], and for further references [4, Section 3.2]) and employs Gaussian quadrature. In the present context, its effectiveness is somewhat limited, and does not allow us to go much beyond $x = 100$. We therefore develop another more direct method which can deal with values of x that are considerably larger.

2. Summation by quadrature

Consider an infinite series

$$S = \sum_{k=1}^{\infty} a_k, \quad a_k = (\mathscr{L}f)(k), \tag{2}$$

whose general term is the Laplace transform

$$(\mathscr{L}f)(s) = \int_0^{\infty} \mathrm{e}^{-st} f(t)\,\mathrm{d}t$$

evaluated at $s = k$ of some known function f. Then we have

$$S = \sum_{k=1}^{\infty} (\mathscr{L}f)(k) = \sum_{k=1}^{\infty} \int_0^{\infty} \mathrm{e}^{-kt} f(t)\,\mathrm{d}t = \int_0^{\infty} \sum_{k=1}^{\infty} \mathrm{e}^{-(k-1)t}\mathrm{e}^{-t} f(t)\,\mathrm{d}t$$
$$= \int_0^{\infty} \frac{1}{1-\mathrm{e}^{-t}} \mathrm{e}^{-t} f(t)\,\mathrm{d}t,$$

that is

$$S = \int_0^{\infty} \frac{t}{1-\mathrm{e}^{-t}} \frac{f(t)}{t} \mathrm{e}^{-t}\,\mathrm{d}t. \tag{3}$$

In general, if $a_k \sim k^{-p}$ as $k \to \infty$, $p > 1$, then $f(t) \sim t^{p-1}$ as $t \to 0$.

To determine the function f in the case of series (1) we note that [1, Eq. (29.3.81)]

$$\frac{1}{s}\mathrm{e}^{x/s} = (\mathscr{L}_{(t)} I_0(2\sqrt{xt}))(s),$$

where I_0 is the modified Bessel function of order zero. There follows, by Euler's formula,

$$\frac{1}{s}\sin(x/s) = \frac{1}{s}\frac{1}{2\mathrm{i}}(\mathrm{e}^{\mathrm{i}x/s} - \mathrm{e}^{-\mathrm{i}x/s}) = \frac{1}{2\mathrm{i}}(\mathscr{L}_{(t)}[I_0(2\sqrt{\mathrm{i}xt}) - I_0(2\sqrt{-\mathrm{i}xt})](s)),$$

Table 1
Number of Gauss points required in (3) for 6-digit accuracy, and severity of cancellation

x	10	25	50	75	100
$n_{\max}$	20	35	55	75	95
d	1	4	10	15	20

that is,

$$f(t) = f(t;x) = \frac{1}{2\mathrm{i}}[I_0(2\sqrt{\mathrm{i}xt}) - I_0(2\sqrt{-\mathrm{i}xt})]. \tag{4}$$

From the known power series expansion of I_0 one finds that

$$f(t;x) = \sum_{k=0}^{\infty}(-1)^k \frac{u^{2k+1}}{(2k+1)!^2}, \quad u = xt. \tag{5}$$

In particular, $\lim_{t\to 0} f(t;x)/t = x$. Series (5) is useful for computation as long as u is not too large, but is subject to severe cancellation errors otherwise. The number of decimal digits lost, owing to cancellation, is approximately 2, 6, 8, 17, and 25 for u respectively equal to 100, 500, 1000, 5000, and 10,000.

Alternatively, we may use the integral representation (cf. [1, Eq. (9.6.16)])

$$I_0(z) = \frac{1}{\pi}\int_0^{\pi} \mathrm{e}^{z\cos\theta}\,\mathrm{d}\theta,$$

and write (4) in the form

$$f(t;x) = \frac{1}{\pi}\int_0^{\pi} \mathrm{e}^{\sqrt{2u}\cos\theta}\sin(\sqrt{2u}\cos\theta)\,\mathrm{d}\theta, \quad u = xt. \tag{6}$$

Here, the integrand is a 2π-periodic even function of θ, so that integration, in effect, is over the full period. The fact that it is also an entire function makes the composite trapezoidal rule the method of choice for evaluating the integral. For the u-values considered above, there is practically no cancellation in calculating the trapezoidal sums, in stark contrast to the series in (5).

With regard to the integral in (3), Gauss quadrature relative to the Laguerre weight function e^{-t} on $[0,\infty)$ would seem to be an option. One can do better, however, by noting that the integrand has poles $\pm 2\nu\mathrm{i}\pi$, $\nu = 1, 2, 3, \ldots$. This suggests using Gauss-type formulae that are exact not only for polynomials, but also for rational functions having the same poles, or at least a few of those closest to the real axis. Such formulae have been developed in [3] and are implemented in [5]. Motivated by experience gained in [5], we choose, for $n = 5, 10, 15, \ldots$, an n-point quadrature rule that is exact for elementary rational functions corresponding to the first $m = 2\lfloor (n+1)/2 \rfloor$ poles (taken in conjugate complex pairs) and for polynomials of degree $2n - 1 - m$. If $n_{\max}$ denotes the smallest n for which two consecutive quadratures agree within a tolerance of $\frac{1}{2}\cdot 10^{-6}$, then, as a function of x, the value of $n_{\max}$ observed has the behavior shown in Table 1.

Table 1 also shows the approximate number d of decimal digits lost, owing to cancellation errors in the quadrature sum for the integral in (3). (For $x = 100$, the error tolerance had to be lowered to $\frac{1}{2}\cdot 10^{-3}$ to be able to achieve it.) All computations were done in quadruple precision. It is seen that values of x

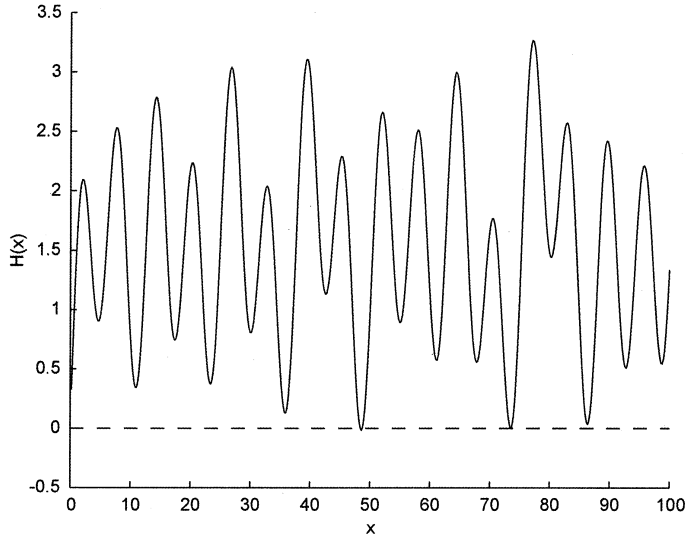


Fig. 1. The function $H(x)$ for $0 \leqslant x \leqslant 100$.

much beyond $x = 100$ are beginning to strain even quadruple-precision calculations. Results produced by these calculations in the range $0 \leqslant x \leqslant 100$ are shown in Fig. 1.

3. Direct summation

Summing the series in (1) directly, as is, would be too time consuming if a reasonably high accuracy is desired. However, we may sum the first n terms directly, where $n \approx x$, and then observe that in the remaining terms $0 < x/k < 1$, so that a few terms in the Taylor expansion of $\sin(x/k)$ may be subtracted to speed up convergence and then added back for compensation. Thus, with $n = \lfloor x \rfloor$,

$$\begin{aligned} H(x) &= \sum_{k=1}^{n} \frac{1}{k} \sin \frac{x}{k} + \sum_{k=n+1}^{\infty} \frac{1}{k} \sin \frac{x}{k} \\ &= \sum_{k=1}^{n} \frac{1}{k} \sin \frac{x}{k} + \sum_{k=n+1}^{\infty} \frac{1}{k} \left(\sin \frac{x}{k} - \frac{x}{k} + \frac{1}{6} \left(\frac{x}{k} \right)^3 \right) \\ &\quad + x \left(\frac{\pi^2}{6} - \sum_{k=1}^{n} \frac{1}{k^2} \right) - \frac{x^3}{6} \left(\frac{\pi^4}{90} - \sum_{k=1}^{n} \frac{1}{k^4} \right), \end{aligned} \tag{7}$$

where the well-known formulae $\zeta(2) = \pi^2/6$, $\zeta(4) = \pi^4/90$ for the zeta function $\zeta(s) = \sum_{k=1}^{\infty} k^{-s}$ have been used (cf. [1, Eqs. (23.2.24–25)]). Since x will be very large, and $H(x)$ of the order of magnitude 1, the two remainder terms at the end of (7) must be calculated very accurately. As written, too much accuracy may

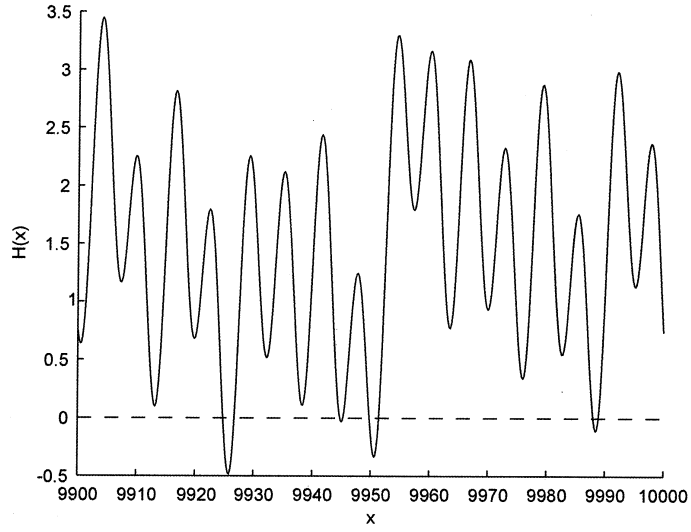


Fig. 2. The function $H(x)$ for $9900 \leqslant x \leqslant 10,000$.

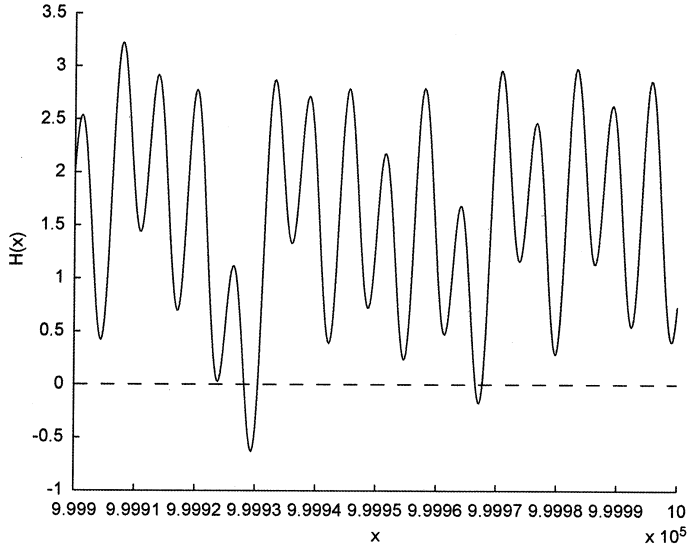


Fig. 3. The function $H(x)$ for $999,900 \leqslant x \leqslant 1,000,000$.

be lost owing to cancellation. A better way to compute these terms is via the Euler–Maclaurin summation formula (cf. [1, Eq. (3.6.28)]). Thus,

$$\frac{\pi^2}{6} - \sum_{k=1}^{n} \frac{1}{k^2} \sim \frac{B_0}{n+1} - \frac{B_1}{(n+1)^2} + \frac{B_2}{(n+1)^3} + \frac{B_4}{(n+1)^5} + \cdots + \frac{B_{10}}{(n+1)^{11}} \tag{8}$$

and

$$\frac{\pi^4}{90} - \sum_{k=1}^{n} \frac{1}{k^4} \sim \frac{1}{3}\frac{B_0}{(n+1)^3} - \frac{B_1}{(n+1)^4} + \frac{2B_2}{(n+1)^5} + \frac{5B_4}{(n+1)^7} + \frac{28}{3}\frac{B_6}{(n+1)^9} + \frac{3B_8}{(n+1)^{11}} + \frac{22B_{10}}{(n+1)^{13}}, \tag{9}$$

where B_i are the Bernoulli numbers

$$B_0 = 1, \quad B_1 = -\tfrac{1}{2}, \quad B_2 = \tfrac{1}{6}, \quad B_4 = -\tfrac{1}{30}, \quad B_6 = \tfrac{1}{42}, \quad B_8 = -\tfrac{1}{30}, \quad B_{10} = \tfrac{5}{66}.$$

The first seven terms in (8) and (9) will be ample to provide sufficient accuracy. Results thus produced are shown in Figs. 2 and 3.

Evidently, neither the unboundedness of H from above nor the one from below can be as much as suggested by these calculations. To do so, in view of the $(\log\log x)^{1/2}$ behavior of $|H(x)|$, would require values of x so large as to not even be machine representable, let alone be such that the summation procedure of this subsection would be feasible.

References

[1] M. Abramowitz, I. Stegun (Eds.), Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, National Bureau of Standards, Applied Mathematics Series, vol. 55, US Government Printing Office, Washington, DC, 1964.
[2] H. Alzer, C. Berg, S. Koumandos, On a conjecture of Clark and Ismail, J. Approx. Theory, to appear.
[3] W. Gautschi, Gauss-type quadrature rules for rational functions, in: H. Brass, G. Hämmerlin (Eds.), Numerical Integration IV, International Series of Numerical Mathematics, vol. 112, Birkhäuser, Basel, 1993, pp. 111–130.
[4] W. Gautschi, Orthogonal polynomials: applications and computation, in: A. Iserles (Ed.), Acta Numerica 1996, Cambridge University Press, Cambridge, 1996, pp. 45–119.
[5] W. Gautschi, Algorithm 793: GQRAT—Gauss quadrature for rational functions, ACM Trans. Math. Software 25 (1999) 213–239.
[6] W. Gautschi, G. Milovanović, Gaussian quadrature involving Einstein and Fermi functions with an application to summation of series, Math. Comp. 44 (1985) 177–190.
[7] G.H. Hardy, J.E. Littlewood, Notes on the theory of series (xx): on Lambert series, Proc. London Math. Soc. 41 (2) (1936) 257–270.

30.7. [197] "The spiral of Theodorus, numerical analysis, and special functions"

---

[197] "The spiral of Theodorus, numerical analysis, and special functions," *J. Comput. Appl. Math*. **235**, 1042–1052 (2010).

Journal of Computational and Applied Mathematics 235 (2010) 1042–1052

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

The spiral of Theodorus, numerical analysis, and special functions☆

Walter Gautschi

Department of Computer Sciences, Purdue University, West Lafayette, IN 47907-2066, United States

ARTICLE INFO

Article history:
Received 8 September 2009
Received in revised form 24 November 2009

Dedicated to Adhemar Bultheel on his 60th birthday

MSC:
01A20
30E05
33C47 51-03
65B10
65D30
65D32

Keywords:
Spiral of Theodorus
Slowly convergent series
Gaussian quadrature
Bose–Einstein distribution

ABSTRACT

Theodorus of Cyrene (ca. 460–399 B.C.), teacher of Plato und Theaetetus, is known for his proof of the irrationality of $\sqrt{n}$, $n = 2, 3, 5, \ldots, 17$. He may have known also of a discrete spiral, today named after him, whose construction is based on the square roots of the numbers $n = 1, 2, 3, \ldots$. The subject of this lecture is the problem of interpolating this discrete, angular spiral by a smooth, if possible analytic, spiral. An interesting solution was proposed in 1993 by P.J. Davis, which is based on an infinite product. The computation of this product gives rise to problems of numerical analysis, in particular the summation of slowly convergent series, and the identification of the product raises questions regarding special functions. The former are solved by a method of integration, in particular Gaussian integration, the latter by means of Dawson's integral und the Bose–Einstein distribution. Number-theoretic questions also loom behind this work.

1. Introduction

The topic of this lecture may be somewhat peripheral to the core of today's mathematical activities, yet it has a certain aesthetic appeal that may well compensate for its borderline status. The principal ideas go back to Greek antiquity, specifically to the 5th century B.C. mathematician and philosopher Theodorus of Cyrene (ca. 460–399 B.C.). He was born and grew up in Cyrene, then a sprawling Greek colony at the Northern coast of Africa (in what today is Libya), directly south of Greece. He also traveled to Athens, where he encountered Socrates. Not much, however, is known about his life and work. From the writings of Plato, who had been a student of Theodorus, in particular from his *Theaetetus*, we know about Theodorus's great fascination with questions of incommensurability. He was to have proved, for example, the irrationality of the square roots of the integers $n = 2, 3, 5, 6, 7, \ldots$, and, so Plato writes, for some reason he stopped at $n = 17$. This cryptic remark has given rise to all sorts of speculation as to what the reasons might have been. One of these, probably the least credible, will be mentioned later.

But let me first introduce the three topics mentioned in the title. First, the *spiral of Theodorus*, depicted in Fig. 1—a harmonious, very pleasing, and elegant spiral. The name "spiral of Theodorus", though, may be misleading, since Theodorus most certainly did not know of this spiral; it is a product of the late 20th century! Very likely, however, he was aware of, or even invented, a more primitive, angular precursor of this spiral, which we will call the "discrete spiral of Theodorus" (cf. Section 2) to distinguish it from the spiral in Fig. 1, which may be called the "analytic spiral of Theodorus".

☆ Lecture presented February 9, 2009 at Purdue University, and February 26, 2009 at the University of Basel.

E-mail address: wxg@cs.purdue.edu.

0377-0427/$ – see front matter
doi:10.1016/j.cam.2009.11.054

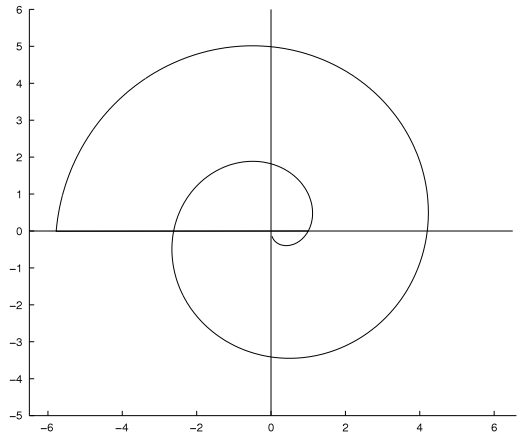


Fig. 1. Spiral of Theodorus.

The second topic – *numerical analysis* – has to do with the summation of slowly convergent series, in particular the series

$$\sum_{k=1}^{\infty} \frac{1}{k^{3/2}+k^{1/2}}. \tag{1}$$

It was in fact this series that gave the impetus to my interest in this area. I was visiting Brown University, early in the 1990s, where I was to give a colloquium lecture. Before the talk, I dropped by Prof. Philip Davis's office to chat a little about the newest mathematical gossip. I knew Prof. Davis well from our days at the (what was then called) National Bureau of Standards in Washington, DC. At one point during our conversation, he pulled out a crumpled envelope from his waste basket, scribbled the series (1) on the back of the envelope, and handed it to me with the words "compute it!". I responded that I couldn't do it on the spot, but promised to look at it once I was back home. (I already had an idea of how to go about it.) A few days later, I sent him back my answer,

$$\sum_{k=1}^{\infty} \frac{1}{k^{3/2}+k^{1/2}} = 1.86002507922119030718069591571714332466652412152345 1493049199503\ldots \tag{2}$$

if not to sixty-four digits, then at least to fifteen (or maybe twenty). This must have impressed Prof. Davis enough to let me in on what was behind this series, and what he was working on at the time: preparing for the Hedrick Lectures he was to give at the 75th anniversary meeting of the Mathematical Association of America. The theme of these lectures was spirals, not only those in mathematics, but also spirals as they occur in nature, in celestial mechanics, and elsewhere. An expanded version of these lectures later appeared in book form [1].

The third topic – *special functions* – finally involves Dawson's integral

$$F(x) = \mathrm{e}^{-x^2} \int_0^x \mathrm{e}^{t^2}\,\mathrm{d}t, \tag{3}$$

probably better known with the opposite signs in the exponents of the exponentials, which then becomes the familiar Gaussian error function.

The theme of this lecture is to show how these three seemingly disparate topics hang together.

2. The discrete spiral of Theodorus

As is well known, in the mathematics of Greek antiquity, numbers and algebraic expressions were thought of differently than they are today. A number like 3 was viewed not so much as a numerical value but as a geometric object: a straight line that has three units in length. Likewise, $\sqrt{2}$ was viewed as the length of the diagonal of a unit square. Since Theodorus was concerned with square roots of successive numbers, he must have viewed them also in geometric terms. Almost inevitably, then, he must have arrived at the construction indicated in Fig. 2. Here, the points $T_0, T_1, T_2, \ldots$ ("T" for "Theodorus") are constructed as follows: T_0 is the origin, and T_1 on the real axis a distance of 1 away from T_0. Thus, the distance $|T_1T_0|$ is $1 = \sqrt{1}$. From T_1 one proceeds in a perpendicular upward direction a distance of 1 to T_2, so $|T_2T_0| = \sqrt{2}$. Then again, perpendicularly, one proceeds a distance of 1 to T_3 and has $|T_3T_0| = \sqrt{2+1} = \sqrt{3}$. Continuing in this manner, the points $T_4, T_5, T_6, \ldots$ so obtained have distances from the origin that are $|T_nT_0| = \sqrt{n}$, $n = 4, 5, 6, \ldots$. One can therefore interpret the successive square roots $\sqrt{n}$ geometrically as being the radial distances of the vertices T_n of the spiral-like construct of

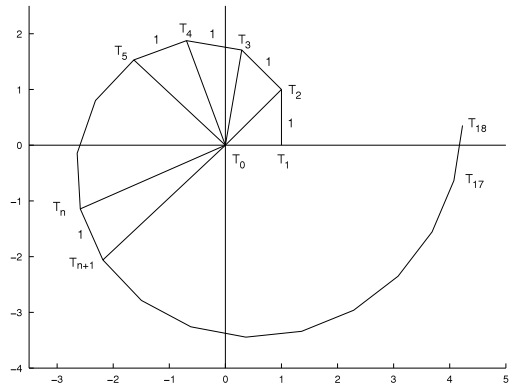


Fig. 2. Discrete spiral of Theodorus.

Fig. 2. It is natural to call it the *discrete spiral of Theodorus*. (It is also known as the "Quadratwurzelschnecke", a term given it by Hlawka in [2].)

It is convenient to view this spiral as a curve in the complex plane, represented parametrically by a complex-valued function

$$T(\alpha) \in \mathbb{C}, \quad \alpha \geq 0.$$

We want this function for integer values of the parameter to produce the vertices of the spiral, $T(n) = T_n$, $n = 0, 1, 2, \ldots$. These are uniquely defined by the relations

$$\left.\begin{aligned} |T_n| &= \sqrt{n} \\ |T_{n+1} - T_n| &= 1 \end{aligned}\right\} \quad n = 0, 1, 2, \ldots \tag{4}$$

with $T_1 = 1$. Linear interpolation between integer-valued parameters then defines $T(\alpha)$ for all $\alpha \geq 0$.

Why did Theodorus stop at $n = 17$? The graph in Fig. 2 gives a clue: The line from T_{17} to T_0, whose length is $\sqrt{17}$, can be drawn without any obstruction. Not so for the line from T_{18} to T_0, and all subsequent lines, which intersect part of the figure already drawn. Since legend has it that geometers in antiquity drew their lines in sand, such intersections become messy, and that's why Theodorus stopped at 17. As I indicated before, *se non é vero, é ben trovato*! ["If it's not true, it's a good story!"]

3. The analytic spiral of Theodorus

3.1. Definition and properties

Davis in [1] posed the problem of interpolating the discrete Theodorus spiral by a smooth, if possible analytic, curve. This is an interpolation problem involving an infinite number of data points, a problem of the type Euler already faced in 1729 when he tried to interpolate the successive factorials on the real line. Ingeniously, Euler discovered the gamma function (now also called the second Eulerian integral) as a valid analytic interpolant. In addition, he derived a number of properties of the gamma function involving product representations, including an infinite product formula for the reciprocal of the gamma function. Davis, who knew Euler's work very well (cf. [3]), used it as a source of inspiration and came up with an interpolant, also expressed as an infinite product,

$$T(\alpha) = \prod_{k=1}^{\infty} \frac{1 + \mathrm{i}/\sqrt{k}}{1 + \mathrm{i}/\sqrt{k + \alpha - 1}}, \quad \alpha \geq 0. \tag{5}$$

Since the general term of the product is $\sim 1 + k^{-3/2}$ as $k \to \infty$, and the series $\sum_{k=1}^{\infty} k^{-3/2}$ converges (absolutely, though slowly), the same is true for the infinite product, as follows from well-known theorems.

Simple calculations will show that the function in (5) satisfies (cf. also (12))

$$|T(\alpha)| = \sqrt{\alpha} \tag{6}$$

and the first-order difference equation

$$T(\alpha + 1) = \left(1 + \frac{\mathrm{i}}{\sqrt{\alpha}}\right) T(\alpha). \tag{7}$$

As a consequence of (6) and (7) one also has

$$|T(\alpha+1)-T(\alpha)| = \left|\frac{\mathrm{i}}{\sqrt{\alpha}}T(\alpha)\right| = |\mathrm{i}| = 1. \tag{8}$$

The relations (6) and (8), for integer values $\alpha = n$, coincide exactly with the analogous relations (4) for the discrete spiral of Theodorus, and since $T(1) = 1$, the function $T(\alpha)$ does indeed interpolate the discrete spiral of Theodorus.

The arc $T(\alpha)$, $1 \le \alpha < 2$, may be considered the "heart" of the spiral; it completely determines the entire spiral, the infinite outer part corresponding to $2 \le \alpha < \infty$ by repeated forward application of (7), and the inner part corresponding to $0 < \alpha < 1$ by a backward application of (7). In the limit as $\alpha \downarrow 0$, one gets $T(0) = 0$.

Recall that Euler's gamma function also satisfies a first-order difference equation, the much simpler $y(\alpha+1) = \alpha y(\alpha)$. Harold Bohr and Johannes Mollerup in 1921 proved the beautiful result that this difference equation has no other solution, with $y(1) = 1$, than the gamma function, if one requires it to be logarithmically convex; cf. [4]. Davis posed the question of whether his own function $T(\alpha)$ in (5), as a solution of the difference equation (7), has a similar uniqueness property. This was answered in 2004 by Gronau [5], who proved, among other things, that $T(\alpha)$ is the only solution of the difference equation (7) with $T(1) = 1$, if one requires $|T(\alpha)|$ to be monotonic and $\arg T(\alpha)$ monotonic and continuous. In the same way as the Bohr–Mollerup result reinforces the legitimacy and importance of the gamma function, the Gronau result does the same for Davis's function.

3.2. Some number theory

An interesting number-theoretic question regards the distribution of the angles $\varphi_n = \angle T_1T_0T_{n+1}$ in the discrete spiral of Theodorus. From the geometry of Fig. 2, it is easily seen that

$$\varphi_n = \sum_{k=1}^{n} \sin^{-1}\frac{1}{\sqrt{k+1}}, \quad n = 1, 2, 3, \ldots. \tag{9}$$

Considering $\varphi_n \bmod 2\pi$, Hlawka in [2] proved that the sequence $\{\varphi_n\}_{n=1}^{\infty}$ is equidistributed mod 2π. In his book [6], Hlawka gives a very elegant proof based on an analytic equidistribution criterion of Fejér (cf. [7, Part II, Probl. 174, p. 281] and [8, pp. 843–844]).

The author, when preparing this lecture, wondered whether a similar equidistribution result holds for the angles $\varphi_n(\alpha) = \angle T(\alpha)T_0T(\alpha+n)$, $1 < \alpha < 2$, in Davis's analytic spiral of Theodorus. These are, from (13), $\varphi_n(\alpha) = \varphi(\alpha+n)-\varphi(\alpha)$, and by analogy with the discrete spiral one suspects that

$$\varphi_n(\alpha) = \sum_{k=1}^{n} \sin^{-1}\frac{1}{\sqrt{k+\alpha}}, \quad n = 1, 2, 3, \ldots, \tag{10}$$

which for $\alpha = 1$ in fact reduces to (9). We shall prove (10) in Section 3.3. The answer to the question of equidistribution was provided by Harald Niederreiter, a former Ph.D. student of Hlawka, and communicated to the author by email on February 3, 2009: The sequence $\{\varphi_n(\alpha)\}_{n=1}^{\infty}$ is indeed also equidistributed mod 2π for any fixed α with $1 < \alpha < 2$ (in fact, for any $\alpha > 0$), and the proof is a simple extension of the proof given by Hlawka in [6].

3.3. Polar representation

When dealing with spirals, it is useful to have a polar representation thereof. For the spiral in Fig. 1, this can be nicely obtained by logarithmic differentiation of $T(\alpha)$. Since $T(\alpha)$ is a product, its logarithmic derivative is the sum of the logarithmic derivatives of the factors,

$$\begin{aligned}
\frac{T'(\alpha)}{T(\alpha)} &= \sum_{k=1}^{\infty} \frac{1+\mathrm{i}/\sqrt{k+\alpha-1}}{1+\mathrm{i}/\sqrt{k}} \frac{\mathrm{d}}{\mathrm{d}\alpha}\left(\frac{1+\mathrm{i}/\sqrt{k}}{1+\mathrm{i}/\sqrt{k+\alpha-1}}\right) \\
&= \sum_{k=1}^{\infty} (1+\mathrm{i}/\sqrt{k+\alpha-1})\frac{\mathrm{i}}{2}\frac{(k+\alpha-1)^{-3/2}}{(1+\mathrm{i}/\sqrt{k+\alpha-1})^2} \\
&= \frac{\mathrm{i}}{2}\sum_{k=1}^{\infty} \frac{1}{(k+\alpha-1)(\sqrt{k+\alpha-1}+\mathrm{i})} \\
&= \frac{\mathrm{i}}{2}\sum_{k=1}^{\infty} \frac{\sqrt{k+\alpha-1}-\mathrm{i}}{(k+\alpha-1)(k+\alpha)}.
\end{aligned}$$

Decomposing the last series into its real and imaginary parts yields

$$\begin{aligned}\frac{T'(\alpha)}{T(\alpha)} &= \frac{1}{2}\sum_{k=1}^{\infty}\frac{1}{(k+\alpha-1)(k+\alpha)} + \frac{\mathrm{i}}{2}\sum_{k=1}^{\infty}\frac{1}{(k+\alpha-1)^{3/2}+(k+\alpha-1)^{1/2}}\\ &= \frac{1}{2}\sum_{k=1}^{\infty}\left(\frac{1}{k+\alpha-1}-\frac{1}{k+\alpha}\right)+\frac{\mathrm{i}}{2}U(\alpha)\\ &= \frac{1}{2\alpha}+\frac{\mathrm{i}}{2}U(\alpha),\end{aligned}$$

where

$$U(\alpha)=\sum_{k=1}^{\infty}\frac{1}{(k+\alpha-1)^{3/2}+(k+\alpha-1)^{1/2}}. \tag{11}$$

Now integrating from 1 to α gives

$$\ln T(\alpha)=\ln(\alpha^{1/2})+\frac{\mathrm{i}}{2}\int_1^{\alpha}U(\alpha)\mathrm{d}\alpha,$$

which by exponentiation yields the desired representation,

$$T(\alpha)=\sqrt{\alpha}\exp\left(\frac{\mathrm{i}}{2}\int_1^{\alpha}U(\alpha)\mathrm{d}\alpha\right),\quad \alpha>0. \tag{12}$$

Thus, in polar coordinates (r,φ), the analytic spiral of Theodorus has the parametric representation

$$r=r(\alpha),\qquad \varphi=\varphi(\alpha)\quad \text{where } r(\alpha)=\sqrt{\alpha},\qquad \varphi(\alpha)=\frac{1}{2}\int_1^{\alpha}U(\alpha)\mathrm{d}\alpha. \tag{13}$$

In terms of this representation, we can rewrite (10) (multiplied by 2) as follows:

$$\int_1^{\alpha+n}U(\alpha)\mathrm{d}\alpha-\int_1^{\alpha}U(\alpha)\mathrm{d}\alpha=2\sum_{k=1}^{n}\sin^{-1}\frac{1}{\sqrt{k+\alpha}}.$$

We know this to be true for $\alpha=1$. To prove it for general α, it suffices to prove that the derivatives with respect to α of the two sides are equal,

$$U(\alpha+n)-U(\alpha)=-\sum_{k=1}^{n}\frac{1}{(k+\alpha)\sqrt{k+\alpha-1}}.$$

This, however, follows readily from the definition of U in (11).

We note that the tangent vector to the spiral at $\alpha=1$ is $T'(1)=\frac{1}{2}+\frac{\mathrm{i}}{2}U(1)$, so

$$U(1)=\sum_{k=1}^{\infty}\frac{1}{k^{3/2}+k^{1/2}}$$

is precisely the slope of the tangent vector to the spiral at $\alpha=1$ where it crosses the real axis for the first time.

We have come halfway to the mysterious series introduced at the beginning of this lecture. As a universal constant, like π, with a solid geometric meaning, it deserves to be given a name, and to be calculated to high precision; we name it, as Davis already did in [1], the "Theodorus constant", and denote it by

$$\theta=\sum_{k=1}^{\infty}\frac{1}{k^{3/2}+k^{1/2}} \tag{14}$$

("θ" for "$\theta\epsilon\omega\delta o\rho o\sigma$").

There is, of course, another number-theoretic problem awaiting attention: the arithmetic nature of the number θ. A solution, however, seems far beyond sight at this time.

We now proceed to the next topic on our agenda, the computation and identification of the function $U(\alpha)$ in (11) and its integral $\int_1^{\alpha}U(\alpha)\mathrm{d}\alpha$ for $1<\alpha<2$. This requires two digressions, one on an appropriate summation procedure, the other on Gaussian quadrature.

4. Two digressions

4.1. Summation by integration

There are several ways to convert a problem of summation, especially the summation of slowly convergent series, to a problem of integration. Here we consider a procedure proposed in 1985 in a joint paper with Milovanović [9] that applies to a special class of series in which the generic term is the Laplace transform of some known function f,

$$s = \sum_{k=1}^{\infty} a_k, \quad a_k = (\mathcal{L}f)(k). \tag{15}$$

Then

$$s = \sum_{k=1}^{\infty} (\mathcal{L}f)(k) = \sum_{k=1}^{\infty} \int_0^{\infty} \mathrm{e}^{-kt} f(t)\mathrm{d}t,$$

and interchanging summation and integration yields

$$s = \int_0^{\infty} \left(\sum_{k=1}^{\infty} \mathrm{e}^{-kt} \right) f(t)\mathrm{d}t = \int_0^{\infty} \frac{t}{\mathrm{e}^t - 1} \frac{f(t)}{t}\mathrm{d}t.$$

Thus

$$\sum_{k=1}^{\infty} a_k = \int_0^{\infty} \frac{f(t)}{t} \varepsilon(t)\mathrm{d}t, \quad f = \mathcal{L}^{-1}a, \tag{16}$$

where

$$\varepsilon(t) = \frac{t}{\mathrm{e}^t - 1}, \quad t \in \mathbb{R}_+. \tag{17}$$

In statistical mechanics, (17) is known as the Bose–Einstein distribution; it is also the generating function of the Bernoulli numbers.

Changing the minus sign in the denominator of (17) to a plus sign and replacing t in the numerator by 1 gives another distribution important in statistical mechanics: the Fermi–Dirac distribution. In our context it arises when the series in (15) contains alternating sign factors.

How does this apply to the Theodorus constant? Here,

$$a_k = \frac{1}{k^{3/2} + k^{1/2}} = \frac{k^{-1/2}}{k+1}.$$

Since

$$k^{-1/2} = \left(\mathcal{L} \frac{t^{-1/2}}{\sqrt{\pi}} \right)(k), \quad \frac{1}{k+1} = \left(\mathcal{L}\mathrm{e}^{-t} \right)(k),$$

the convolution theorem for Laplace transforms yields

$$a_k = \left(\mathcal{L} \frac{t^{-1/2}}{\sqrt{\pi}} \right)(k) \cdot \left(\mathcal{L}\mathrm{e}^{-t} \right)(k) = \left(\mathcal{L} \frac{1}{\sqrt{\pi}} \int_0^t \tau^{-1/2} \mathrm{e}^{-(t-\tau)}\mathrm{d}\tau \right)(k), \tag{18}$$

where the integral on the right is the convolution of $t^{-1/2}$ and e^{-t}. Thus,

$$f(t) = \frac{1}{\sqrt{\pi}} \mathrm{e}^{-t} \int_0^t \tau^{-1/2} \mathrm{e}^{\tau}\mathrm{d}\tau = \frac{2}{\sqrt{\pi}} \mathrm{e}^{-t} \int_0^{\sqrt{t}} \mathrm{e}^{x^2}\mathrm{d}x = \frac{2}{\sqrt{\pi}} F(\sqrt{t}),$$

where $F(x)$ is Dawson's integral (3). There follows, from (16),

$$\sum_{k=1}^{\infty} \frac{1}{k^{3/2} + k^{1/2}} = \int_0^{\infty} \frac{f(t)}{t} \varepsilon(t)\mathrm{d}t.$$

By writing $t = \sqrt{t} \cdot \sqrt{t}$ in the denominator of the integrand and associating one square root with f and the other with ε, we obtain

$$\sum_{k=1}^{\infty} \frac{1}{k^{3/2} + k^{1/2}} = \frac{2}{\sqrt{\pi}} \int_0^{\infty} \frac{F(\sqrt{t})}{\sqrt{t}} w(t)\mathrm{d}t, \tag{19}$$

where

$$w(t) = t^{-1/2}\varepsilon(t) = \frac{t^{1/2}}{\mathrm{e}^t - 1}. \tag{20}$$

We recall that $F(x)$ is an entire, odd function of x; hence $F(\sqrt{t})/\sqrt{t}$ in (19) is a power series in t that converges in the whole complex plane, and hence in turn an entire function. For the purpose of numerical integration, entire functions are usually conducive to rapid convergence; hence the first factor, $F(\sqrt{t})/\sqrt{t}$, in the integrand of (19) is a nice, benign function. The second factor, $w(t)$, though positive on $\mathbb{R}_+$, is difficult: for one thing, it blows up like $t^{-1/2}$ at $t = 0$, and for another, it has an infinite string of poles on the imaginary axis at the integer multiples of $2\pi\mathrm{i}$. Both are troublesome for numerical integration. But in numerical analysis one knows of an effective approach for integrating such a product: one treats the difficult factor as a weight function and applies weighted numerical integration, for example, Gaussian quadrature.

4.2. Gaussian quadrature

An n-point Gaussian quadrature formula for an integral as in (19) is a relation

$$\int_0^\infty g(t)w(t)\mathrm{d}t = \sum_{k=1}^{n} \lambda_k^{(n)} g(\tau_k^{(n)}), \quad g \in \mathbb{P}_{2n-1}, \tag{21}$$

which expresses the integral exactly as a linear combination of n function values provided the function is a polynomial of degree $\leq 2n-1$. It is known that such a representation exists uniquely, and that the "weights" $\lambda_k^{(n)}$ are positive (if w is positive) and the "nodes" $\tau_k^{(n)}$ are mutually distinct and contained in the open interval $(0, \infty)$. If g is not a polynomial, but is polynomial-like, for example an entire function as in (19), then (21) will no longer be an exact equality but very likely a good approximation, especially if n is large.

But how do we find the weights $\lambda_k^{(n)}$ and nodes $\tau_k^{(n)}$ for any given n? The answer is well known in principle: we need the orthogonal polynomials with respect to the weight function w, that is, the (monic) polynomials π_k of degree k, $k = 0, 1, 2, \ldots$, satisfying

$$(\pi_k, \pi_\ell) = 0, \quad k \neq \ell, \quad \text{where } (u, v) = \int_0^\infty u(t)v(t)w(t)\mathrm{d}t.$$

It is known that they exist uniquely and satisfy a three-term recurrence relation

$$\begin{aligned} &\pi_{k+1}(t) = (t - \alpha_k)\pi_k(t) - \beta_k\pi_{k-1}(t), \quad k = 0, 1, 2, \ldots, \\ &\pi_{-1}(t) = 0, \qquad \pi_0(t) = 1, \end{aligned}$$

where the coefficients $\alpha_k = \alpha_k(w)$ and $\beta_k = \beta_k(w)$ are respectively real and positive numbers depending on w. Although β_0 is arbitrary, it is convenient to define $\beta_0 = \int_0^\infty w(t)\mathrm{d}t$. The nth-order Jacobi matrix

$$\boldsymbol{J}_n(w) = \begin{bmatrix} \alpha_0 & \beta_1 & & & \mathbf{0} \\ \beta_1 & \alpha_1 & \beta_2 & & \\ & \beta_2 & \alpha_2 & \ddots & \\ & & \ddots & \ddots & \beta_{n-1} \\ \mathbf{0} & & & \beta_{n-1} & \alpha_{n-1} \end{bmatrix} \tag{22}$$

is formed by placing the first n coefficients $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$ on the diagonal, the $n-1$ coefficients $\beta_1, \beta_2, \ldots, \beta_{n-1}$ on the two side diagonals, and filling the rest of the matrix with zeros. It is the eigenvalues and eigenvectors of this symmetric, tridiagonal matrix that yield the Gaussian nodes and weights: the nodes $\tau_k^{(n)}$ are the eigenvalues of $\boldsymbol{J}_n$, and the weights $\lambda_k^{(n)}$ expressible as $\lambda_k^{(n)} = \beta_0 \boldsymbol{v}_{k,1}^2$ in terms of the first components $\boldsymbol{v}_{k,1}$ of the corresponding (normalized) eigenvectors $\boldsymbol{v}_k$ [10].

We are done, once we are in possession of the recurrence coefficients α_k, β_k. There are various numerical techniques for computing them (cf., for example, [11, Sections 2.1, 2.2]). For our purposes here, the classical approach based on moments

$$\mu_k = \int_0^\infty t^k w(t)\mathrm{d}t, \quad k = 0, 1, 2, \ldots, \tag{23}$$

suffices. An algorithm due to Chebyshev takes the first $2n$ moments (23), and from them generates the first n coefficients $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$ and $\beta_0, \beta_1, \ldots, \beta_{n-1}$ by a simple nonlinear recursion. The algorithm is elegant but highly unstable, the more so the larger n. This drawback, however, can be overcome by running the algorithm in sufficiently high precision. Relevant software is available; see, e.g., [12].

To show how this works for the Theodorus constant, we first note that the moments of the weight function w in (20) are

$$\mu_k = \int_0^\infty \frac{t^{k+1/2}}{\mathrm{e}^t - 1}\mathrm{d}t = \Gamma(k + 3/2)\zeta(k + 3/2).$$

Table 1
Gaussian quadrature approximations to the Theodorus constant.

n	s_n
5	1.85997…
15	1.86002507922117…
25	1.86002507922119030718068 9…
35	1.86002507922119030718069591571714 1…
45	1.86002507922119030718069591571714332466652 35…
55	1.86002507922119030718069591571714332466652412152345 13…
65	1.86002507922119030718069591571714332466652412152345149304919944…
75	1.86002507922119030718069591571714332466652412152345149304919950 3…

Both the gamma function Γ and the Riemann zeta function ζ are computable by variable-precision calculation. Applying the Chebyshev algorithm in sufficiently high precision to get the Jacobi matrix (22), and then well-known eigenvalue/eigenvector techniques to get the Gaussian quadrature formula, we can now approximate

$$\sum_{k=1}^{\infty} \frac{1}{k^{3/2}+k^{1/2}} = \frac{2}{\sqrt{\pi}} \int_0^{\infty} [F(\sqrt{t})/\sqrt{t}]w(t)\mathrm{d}t \tag{24}$$

by

$$s_n = \frac{2}{\sqrt{\pi}} \sum_{k=1}^{n} \lambda_k^{(n)} F\left(\sqrt{\tau_k^{(n)}}\right) \Big/ \sqrt{\tau_k^{(n)}}.$$

Numerical results for $n = 5 : 10 : 75$ are shown in Table 1. We see now how the answer given in (2) comes about. Allowing for a sufficient amount of computer time, we could obtain it to an arbitrary number of decimal digits. Faster high-precision computational techniques, however, can be found in [13].

5. Computation and identification

We are now in a position to deal with the computation of $U(\alpha)$ (cf. (11)) and $\int_1^{\alpha} U(\alpha)\mathrm{d}\alpha$ for $1 < \alpha < 2$. The series in (11) is the same as the series (14) for the Theodorus constant except that k in the latter has to be replaced by $k + \alpha - 1$. From the computation in (18), we thus find that

$$\frac{(k+\alpha-1)^{-1/2}}{(k+\alpha-1)+1} = \frac{1}{\sqrt{\pi}} \left(\mathscr{L} \int_0^t \tau^{-1/2} \mathrm{e}^{-(t-\tau)} \mathrm{d}\tau \right) (k+\alpha-1).$$

It is now a matter of applying the shift property of the Laplace transform to obtain

$$\frac{(k+\alpha-1)^{-1/2}}{(k+\alpha-1)+1} = \frac{1}{\sqrt{\pi}} \mathscr{L} \left(\mathrm{e}^{-\alpha t} \int_0^t \tau^{-1/2} \mathrm{e}^{\tau} \mathrm{d}\tau \right) (k);$$

hence, the function f in (15) is

$$f(t) = \frac{1}{\sqrt{\pi}} \mathrm{e}^{-\alpha t} \int_0^t \tau^{-1/2} \mathrm{e}^{\tau} \mathrm{d}\tau = \frac{2}{\sqrt{\pi}} \mathrm{e}^{-(\alpha-1)t} F(\sqrt{t}).$$

We find, analogously to (19),

$$U(\alpha) = \frac{2}{\sqrt{\pi}} \int_0^{\infty} \mathrm{e}^{-(\alpha-1)t} \frac{F(\sqrt{t})}{\sqrt{t}} w(t)\mathrm{d}t, \quad 1 < \alpha < 2. \tag{25}$$

This again can be readily computed by Gauss quadrature, just like (24). We note, incidentally, that $U(\alpha)$ can be identified as a Laplace transform itself, namely

$$U(\alpha) = (\mathscr{L}u)(\alpha - 1),$$

where

$$u(t) = \frac{2}{\sqrt{\pi}} \frac{F(\sqrt{t})}{\sqrt{t}} w(t) = \frac{2}{\sqrt{\pi}} \frac{F(\sqrt{t})}{\mathrm{e}^t - 1}.$$

As far as the integral of $U(\alpha)$ is concerned, we only need to integrate under the integral sign in (25) to obtain

$$\int_1^{\alpha} U(\alpha)\mathrm{d}\alpha = \frac{2(\alpha-1)}{\sqrt{\pi}} \int_0^{\infty} \frac{1-\mathrm{e}^{-(\alpha-1)t}}{(\alpha-1)t} \frac{F(\sqrt{t})}{\sqrt{t}} w(t)\mathrm{d}t. \tag{26}$$

This, too, is amenable to Gauss quadrature but requires a little extra care in the evaluation near $t = 0$ of the first factor on the right.

6. Epilogue

6.1. The Theodorus constant to very high precision

Waldvogel [13] has calculated the Theodorus constant θ to over a thousand decimal places, using a line integral representation in the complex plane, the trapezoidal rule, and the computer algebra system PARI. With the same package, at the suggestion of N. A'Campo, he computed the continued fraction

$$\theta = 1 + \frac{1}{1+}\,\frac{1}{6+}\,\frac{1}{6+}\,\frac{1}{1+}\,\frac{1}{15+}\,\frac{1}{11+}\,\frac{1}{5+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{5+}\cdots$$

to some 300 partial denominators to look for patterns. None were found.

6.2. Summation by integration; extensions

The summation process of Section 4.1 can be generalized to series

$$s_+ = \sum_{k=1}^{\infty} k^{\nu-1} R(k), \qquad s_- = \sum_{k=1}^{\infty} (-1)^k k^{\nu-1} R(k), \quad 0 < \nu < 1,$$

where R is a rational function having all its poles in the left half of the complex plane (cf. [14]). The integration measures that arise are, as in Section 4.1, the Bose–Einstein distribution for the series s_+ and the Fermi–Dirac distribution for the series s_-. The special functions involved, however, are more elaborate, being based on Tricomi's form of the incomplete gamma function. Also, there are serious complications that arise when the poles of R are large in magnitude, in which case Gaussian quadrature converges very slowly. Satisfactory convergence can be restored by a process called "stratified summation" in [14].

6.3. The analytic spiral of Theodorus; an alternative approach

Heuvers et al. [15], apparently unaware of Davis's work, gave the following analytic interpolant of the discrete Theodorus spiral, expressed in polar coordinates:

$$\varphi = g(r), \qquad g(r) = \sum_{j=0}^{\infty} \left(\tan^{-1} \frac{1}{\sqrt{j+1}} - \tan^{-1} \frac{1}{\sqrt{j+r^2}} \right), \quad r \geq 1. \tag{27}$$

They proved that $g(r)$ in (27) is the unique monotonically increasing solution, satisfying $g(1) = 0$, of the functional equation

$$g\left(\sqrt{1+r^2}\right) - g(r) = \tan^{-1} \frac{1}{r}, \quad r \geq 1, \tag{28}$$

thus anticipating Gronau's uniqueness result.

The connection of (27) and (28) with Davis's spiral is as follows. The angle φ in Davis's spiral, as a function of r, can be seen from (13), since $\alpha = r^2$, to be

$$\varphi = \frac{1}{2} \int_1^{r^2} U(\alpha) \mathrm{d}\alpha, \tag{29}$$

which is identical to (27). In fact, when $r = 1$, this is obvious, and differentiating with respect to r we get $rU(r^2)$ from (29) and

$$-\sum_{j=0}^{\infty} \frac{1}{1 + (j+r^2)^{-1}} \left(-\frac{1}{2}\right) (j+r^2)^{-3/2} 2r = r \sum_{j=0}^{\infty} \frac{1}{(j+r^2)^{3/2} + (j+r^2)^{1/2}},$$

from (27), which by (11) is indeed $rU(r^2)$. On writing

$$1 + \frac{\mathrm{i}}{\sqrt{\alpha}} = \sqrt{\frac{\alpha+1}{\alpha}}\, \mathrm{e}^{\mathrm{i}\theta(\alpha)}, \quad \theta(\alpha) = \tan^{-1} \frac{1}{\sqrt{\alpha}},$$

the difference equation (7) splits into

$$\frac{r(\alpha+1)}{r(\alpha)} = \sqrt{\frac{\alpha+1}{\alpha}}, \qquad \varphi(\alpha+1) = \varphi(\alpha) + \tan^{-1} \frac{1}{\sqrt{\alpha}},$$

the latter, on setting $\varphi(r^2) = g(r)$, becoming (28).

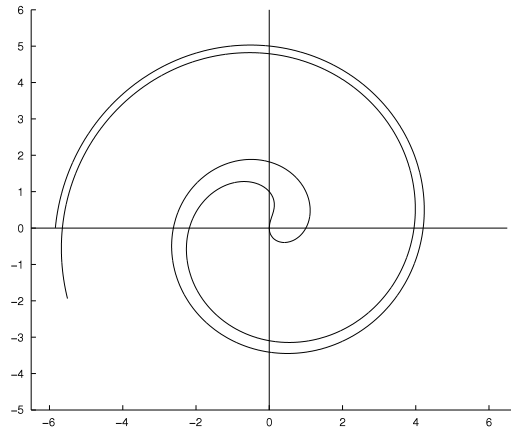


Fig. 3. Twin-spiral of Theodorus.

6.4. Analytic continuation of the spiral of Theodorus

For complex α, Davis's function $T(\alpha)$ in (5) is multivalued, owing to the square roots in the denominator. A useful "regularizing" transformation,

$$\alpha = r^2, \quad r \in \mathbb{R}, \tag{30}$$

is suggested by Waldvogel in [13]. It has the effect of transforming $T(\alpha)$ into a function

$$T(r^2) = \frac{1+\mathrm{i}}{1+\mathrm{i}/r} \prod_{k=2}^{\infty} \frac{1+\mathrm{i}/\sqrt{k}}{1+\mathrm{i}/\sqrt{r^2+k-1}} \tag{31}$$

that is regular analytic in the complex r-plane cut along the lines from i to i∞ and $-$i to $-$i∞ on the imaginary axis. The part of (31) corresponding to positive values of r coincides with the spiral shown in Fig. 1, whereas the part corresponding to negative values of r may be considered the analytic continuation of the spiral into the second sheet of the Riemann surface for the square root. Both parts together, shown in Fig. 3, constitute what may be called the "twin-spiral of Theodorus".

If $T(\alpha)$, $\alpha > 0$, is on the original spiral (5), then

$$S(\alpha) = \frac{1+\mathrm{i}/\sqrt{\alpha}}{1-\mathrm{i}/\sqrt{\alpha}} T(\alpha)$$

is the corresponding point on the twin branch of the spiral. Therefore, by (7),

$$S(\alpha) = \frac{1}{1-\mathrm{i}/\sqrt{\alpha}} T(\alpha+1), \tag{32}$$

whereas

$$T(\alpha) = \frac{1}{1+\mathrm{i}/\sqrt{\alpha}} T(\alpha+1), \tag{33}$$

showing that the two points in (32) and (33) are mirror images with respect to the line $T_0T(\alpha+1)$. In the special case $\alpha = n^2$, $n > 0$ an integer, i.e., in the case of the discrete Theodorus spiral, this was observed in [13].

Acknowledgements

The author is indebted to Jörg Waldvogel for useful discussions and grateful to him for providing a copy of [15] and access to the manuscript for [13] as it progressed.

References

[1] Philip J. Davis, Spirals: From Theodorus to Chaos, AK Peters, Wellesley, MA, 1993, With contributions by Walter Gautschi and Arieh Iserles.
[2] Edmund Hlawka, Gleichverteilung und Quadratwurzelschnecke, Monatsh. Math. 89 (1) (1980) 19–44.
[3] Philip J. Davis, Leonhard Euler's integral: A historical profile of the gamma function, Amer. Math. Monthly 66 (1959) 849–869.

[4] Emil Artin, The Gamma Function (Michael Butler, Trans.), in: Athena Series: Selected Topics in Mathematics, Holt, Rinehart and Winston, New York, 1964.
[5] Detlef Gronau, The spiral of Theodorus, Amer. Math. Monthly 111 (3) (2004) 230–237.
[6] Edmund Hlawka, The Theory of Uniform Distribution (Henry Orde, Trans.), AB Academic Publishers, Berkhamsted, 1984 (in German).
[7] G. Pólya, G. Szegö, Problems and Theorems in Analysis I (D. Aeppli, Trans.), Springer, Berlin, 1978.
[8] Turán Pál (Ed.), Leopold Fejér: Gesammelte Arbeiten, vol. II, Akadémiai Kiadó, Budapest, 1970.
[9] Walter Gautschi, Gradimir V. Milovanović, Gaussian quadrature involving Einstein and Fermi functions with an application to summation of series, Math. Comp. 44 (169) (1985) 177–190.
[10] Gene H. Golub, John H. Welsch, Calculation of Gauss quadrature rules, Math. Comp. 23 (1969) 221–230;
Gene H. Golub, John H. Welsch, Calculation of Gauss quadrature rules, Math. Comp. 23 (106) (1969) loose microfiche suppl. A1–A10 (addendum).
[11] Walter Gautschi, Orthogonal Polynomials: Computation and Approximation, in: Numerical Mathematics and Scientific Computation, Oxford Science Publications, Oxford University Press, New York, 2004.
[12] Walter Gautschi, Variable-precision recurrence coefficients for nonstandard orthogonal polynomials, Numer. Algorithms 52 (3) (2009) 409–441.
[13] Jörg Waldvogel, Analytic continuation of the Theodorus spiral (in preparation); see http://www.sam.math.ethz.ch/~waldvoge.
[14] Walter Gautschi, A class of slowly convergent series and their summation by Gaussian quadrature, Math. Comp. 57 (1991) 309–324.
[15] Konrad J. Heuvers, Daniel S. Moak, Blake Boursaw, The functional equation of the square root spiral, in: T.M. Rassias (Ed.), Functional Equations and Inequalities, in: Math. Appl., vol. 518, Kluwer Acad. Publ., Dordrecht, 2000, pp. 111–117.

Part III

Werner Gautschi

31

Publications

Werner Gautschi

1 The asymptotic behaviour of powers of matrices, Duke Math. J. 20, 127–140 (1953)
2 The asymptotic behaviour of powers of matrices. II, Duke Math. J. 20, 375–379 (1953)
3 Bounds of matrices with regard to an Hermitian metric, Compositio Math. 12, 1–16 (1954)
4 Some remarks on systematic sampling, Ann. Math. Statist. 28, 385–394 (1957)
5 Some remarks on Herbach's paper "Optimum nature of the F-test for model II in the balanced case", Ann. Math. Statist. 30, 960–963 (1959)

THE ASYMPTOTIC BEHAVIOUR OF POWERS OF MATRICES

"The asymptotic behaviour of powers of matrices", *Duke Math. J.* **20,** 127–140 (1953).

THE ASYMPTOTIC BEHAVIOUR OF POWERS OF MATRICES

By Werner Gautschi

I. The Main Results

1. **Introduction.** Let $A = (a_{\nu\mu})(\nu, \mu = 1, \cdots, n)$ be an $n \times n$ matrix with real or complex numbers as elements. By $A^* = \bar{A}'$ we denote the conjugate-transpose of A and by tr A the trace $\sum_{\nu=1}^{n} a_{\nu\nu}$ of A. The *norm* (or *absolute value*) $N(A)$ of A is defined by

$$N(A) = (\text{tr } AA^*)^{\frac{1}{2}} = \left(\sum_{\nu,\mu=1}^{n} | a_{\nu\mu} |^2 \right)^{\frac{1}{2}}. \tag{1}$$

It is well known (J. H. M. Wedderburn [11] or [12; 125]; see also Hardy-Littlewood-Pólya [5; 36]) that

$$N(A + B) \leq N(A) + N(B), \quad N(AB) \leq N(A)N(B), \quad N(\lambda A) = | \lambda | N(A) \tag{2}$$

λ being a scalar and A, B two $n \times n$ matrices.

In (2) of §1 we shall give bounds for the norms of powers $A^p (p = 1, 2, \cdots)$. A *lower* bound can readily be found: by a well-known theorem of I. Schur [7] there exists a unitary matrix U which transforms A to a triangular matrix D. The principal diagonal of D consists of the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$ of A, not necessarily all distinct, arranged in any desired order. From

$$U^*AU = D, \qquad U^*U = UU^* = I \tag{3}$$

it follows that

$$U^*A^pU = D^p, \qquad U^*(A^p)^*U = (D^p)^*$$

and

$$\text{tr } A^p(A^p)^* = \text{tr } U^*A^p(A^p)^*U = \text{tr } D^p(D^p)^*.$$

Hence by (1) we have

$$N(A^p) = N(D^p) \geq \left(\sum_{\nu=1}^{n} | \lambda_\nu |^{2p} \right)^{\frac{1}{2}} \qquad (p = 1, 2, \cdots) \tag{4}$$

with equality for all $p = 1, 2, \cdots$, if and only if A is normal, $A^*A = AA^*$.

Suppose now that A can be transformed to the *diagonal form*

$$X^{-1}AX = \Lambda = \text{diag } (\lambda_1, \lambda_2, \cdots, \lambda_n)$$

Received February 9, 1952; in revised form, October 18, 1952. This paper is part of the thesis for the doctor of philosophy at the University of Basle, Switzerland.

by a nonsingular matrix X. A necessary and sufficient condition for this is that all elementary divisors of A are simple, *i.e.*, that the minimal polynomial $\phi(\lambda)$ of A has no multiple roots. We then get $\Lambda^p = X^{-1}A^pX$,

$$A^p = X\Lambda^pX^{-1}$$

and, taking norms on both sides and using (2),

$$N(A^p) \leq N(X)N(X^{-1})N(\Lambda^p),$$

that is,

$$N(A^p) \leq c\left(\sum_{\nu=1}^{n} |\lambda_\nu|^{2p}\right)^{\frac{1}{2}} \qquad (p = 1, 2, \cdots), \tag{5}$$

where $c > 0$ and only depends on A. (5) can be generalized to cover the case in which $\phi(\lambda)$ has multiple roots. We will show that the right-hand side has to be multiplied by p^{k-1}, where k is, roughly speaking, the highest multiplicity of the roots of $\phi(\lambda)$. The proof is based on a simple application of Sylvester's interpolation formula in the general form.

So far we have only considered *norms* of matrices. We will deal now with the asymptotic behavior of the *individual elements* in the powers of matrices. Necessary and sufficient conditions for the existence of the infinite power $A^\infty = \lim_{p\to\infty} A^p$ and for A^∞ to be zero have been given by R. Oldenburger [6]. From the point of view of topological algebra O. Taussky [8], O. Taussky and J. Todd [9] have further investigated properties of the class of matrices whose infinite powers are zero. In this paper we shall be concerned in finding upper bounds for the elements of A^p. Matrices B_p which can serve as "majorants" for A^p have been given by Frazer-Duncan-Collar [4; 145–147]. The authors assign to all elements of B_p the same order of magnitude. For certain *triangular* matrices however the order can be graduated as will be shown in 3-4 of §1 by elementary considerations. The result may roughly be characterized as follows: if $D = (d_{\nu\mu})$, $d_{\nu\mu} = 0$ $(\nu > \mu)$ is a triangular matrix whose elements along the principal diagonal are arranged in descending order of moduli, the elements in the i-th row of D^p are, if not zero, in modulus at most equal to a constant $\times |d_{ii}|^p$, eventually multiplied by a certain fixed power of p.

In §2 immediate consequences are deduced from Theorem 1 yielding sequences converging towards $\max_\nu |\lambda_\nu|$.

Finally I should like to express my indebtedness to Prof. A. Ostrowski for his valuable criticism and for many helpful suggestions.

2. THEOREM 1. *Let A be a (real or complex) $n \times n$ matrix and suppose that not all eigenvalues λ_ν of A are zero. Denote by m_ν the multiplicity of λ_ν in the minimal polynomial of A and put*

$$k = \max_{\substack{\nu=1,\cdots,n \\ \lambda_\nu \neq 0}} m_\nu .$$

Then we have for a constant $c > 0$ depending only on A

$$1 \leq \frac{N(A^p)}{\left(\sum_{\nu=1}^{n} |\lambda_\nu|^{2p}\right)^{\frac{1}{2}}} \leq cp^{k-1} \qquad (p = 1, 2, \cdots). \tag{6}$$

If all λ_ν are zero we have $N(A^p) = 0$ $(p \geq l, l = m_\nu)$.

Proof. Because of (4) it remains only to prove the right-hand side of (6). If all λ_ν are zero, the minimal polynomial of A is λ^l and $A^p = 0$ $(p \geq l)$. (By 0 we also denote *matrices* with zero elements.) We may therefore assume not all λ_ν to be zero. Let

$$\phi(\lambda) = (\lambda - \lambda_1)^{m_1}(\lambda - \lambda_2)^{m_2} \cdots (\lambda - \lambda_s)^{m_s}$$

be the minimal polynomial of A *with* $\lambda_1, \lambda_2, \cdots, \lambda_s$ *all distinct* and put $m = \sum_{\sigma=1}^{s} m_\sigma$. For any polynomial $f(\lambda)$ we form the $s(m+1) \times m_\sigma$-matrices Λ_σ and an $(m+1) \times 1$-matrix Λ_{s+1}:

$$\Lambda_\sigma = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \lambda_\sigma & 1 & \cdots & 0 \\ \lambda_\sigma^2 & \frac{2}{1!}\lambda_\sigma & & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \lambda_\sigma^{m-1} & \frac{m-1}{1!}\lambda_\sigma^{m-2} & \cdots & \vdots \\ f(\lambda_\sigma) & \frac{f'(\lambda_\sigma)}{1!} & \cdots & \frac{f^{(m_\sigma-1)}(\lambda_\sigma)}{(m_\sigma-1)!} \end{bmatrix} \quad (\sigma = 1, \cdots, s), \quad \Lambda_{s+1} = \begin{bmatrix} 1 \\ \lambda \\ \lambda^2 \\ \vdots \\ \vdots \\ \lambda^{m-1} \\ f(\lambda) \end{bmatrix}.$$

The elements in the $(\mu+1)$-th column of $\Lambda_\sigma (\sigma = 1, \cdots, s)$ are derived from those of the first column by applying the operation $1/\mu!(d/d\lambda_\sigma)^\mu$. Putting together the blocks $\Lambda_1, \Lambda_2, \cdots, \Lambda_{s+1}$ into one row we get a square matrix

$$\Lambda = [\Lambda_1, \Lambda_2, \cdots, \Lambda_{s+1}].$$

Let us for a moment consider $\det \Lambda$ as a polynomial of the $(s+1)$ independent variables $\lambda_1, \lambda_2, \cdots, \lambda_s, \lambda$; then, as is well known (Turnbull and Aitken [10; 63]), $\det \Lambda$ is divisible by the corresponding "confluent" difference-product Δ,

$$\det \Lambda \equiv 0 \pmod{\Delta}, \qquad \Delta = \prod_{\sigma=1}^{s} (\lambda - \lambda_\sigma)^{m_\sigma} \prod_{\tau>\sigma}^{1,s} (\lambda_\tau - \lambda_\sigma)^{m_\tau m_\sigma}$$

$$= \phi(\lambda) \prod_{\tau>\sigma}^{1,s} (\lambda_\tau - \lambda_\sigma)^{m_\tau m_\sigma},$$

that is,

$$\det \Lambda \equiv 0 \pmod{\phi(\lambda)}. \tag{7}$$

We now expand $\det \Lambda$ in terms of its last row; this gives a relation of the type

$$\det \Lambda = cf(\lambda) + P(\lambda), \tag{8}$$

where

$$c = \prod_{\tau>\sigma}^{1,s} (\lambda_\tau - \lambda_\sigma)^{m_\tau m_\sigma} \neq 0$$

(Turnbull and Aitken [10; 63]; see also Aitken [1; 119–121]) and where $P(\lambda)$ is a polynomial with coefficients in which $f(\lambda_1), f'(\lambda_1), \cdots, f(\lambda_2), f'(\lambda_2), \cdots$ appear linearly. On the right-hand side of (8) we replace λ by the matrix A; since $\phi(A) = 0$, it follows from (7) that

$$cf(A) + P(A) = 0.$$

Thus we may write for any polynomial $f(\lambda)$

$$f(A) = \sum_{\sigma=1}^{s} f_\sigma, \qquad f_\sigma = \sum_{\tau=0}^{m_\sigma-1} \frac{f^{(\tau)}(\lambda_\sigma)}{\tau!} C_\tau^{(\sigma)} \tag{9}$$

where $C_\tau^{(\sigma)}$ are matrices *independent of the choice of* $f(\lambda)$ and completely determined by the matrix A. (Compare for this argument Turnbull and Aitken [10; 76–78]. Another proof and more details on formula (9) are given in Wedderburn [12; 27–30].)

We now specialize (9) choosing $f(\lambda) = \lambda^p$ so that

$$A^p = \sum_{\sigma=1}^{s} f_\sigma, \qquad f_\sigma = \sum_{\tau=0}^{m_\sigma-1} \binom{p}{\tau} \lambda_\sigma^{p-\tau} C_\tau^{(\sigma)}. \tag{10}$$

If $\lambda_\sigma \neq 0$ we can write

$$f_\sigma = \lambda_\sigma^p \sum_{\tau=0}^{m_\sigma-1} \binom{p}{\tau} \lambda_\sigma^{-\tau} C_\tau^{(\sigma)},$$

and thus by (2), majorizing the polynomial $\sum \binom{p}{\tau}$ by means of its highest power,

$$N(f_\sigma) \leq c_\sigma \mid \lambda_\sigma \mid^p \sum_{\tau=0}^{m_\sigma-1} \binom{p}{\tau} \leq d_\sigma \mid \lambda_\sigma \mid^p p^{m_\sigma-1} \qquad (p \geq k), \tag{11}$$

c_σ, d_σ being positive constants. Since $f_\sigma = 0$ for $\lambda_\sigma = 0$ and sufficiently large

p, say $p \geq p_0$, it follows from (10), (11) and (2) that for a certain constant $d > 0$

$$N(A^p) \leq \sum_{\sigma=1}^{s} N(f_\sigma) \leq dp^{k-1} \sum_{\sigma=1}^{s} |\lambda_\sigma|^p \qquad (p \geq p_0). \tag{11a}$$

Since not all $\lambda_\sigma (\sigma = 1, \cdots, s)$ are zero, this upper bound is positive for all integers $p \geq 1$. Hence the constant d can be chosen such that the last inequality also holds for $p = 1, 2, \cdots$.

If we change our notations and suppose $\lambda_1, \lambda_2, \cdots, \lambda_n$ to be the n (no longer necessarily distinct) eigenvalues of A, we have *a fortiori*

$$N(A^p) \leq dp^{k-1} \sum_{\nu=1}^{n} |\lambda_\nu|^p \qquad (p = 1, 2, \cdots),$$

which, by Schwarz's inequality, is

$$\leq n^{\frac{1}{2}} dp^{k-1} \left(\sum_{\nu=1}^{n} |\lambda_\nu|^{2p} \right)^{\frac{1}{2}}.$$

This completes the proof of Theorem 1.

We may indicate an alternative, though less elementary proof of (6): adapting the method by which we proved (5) we can transform A to Jordan's canonical form, whose powers are readily calculated. The estimation of their norms can then be carried through similarly as before.

Finally we may remark that the lower bound in (6) can also be attained by nonnormal matrices for certain values of p. Take *e.g.*

$$D = \begin{pmatrix} 1 & 1 \\ 0 & e^{i\psi} \end{pmatrix}$$

and write

$$D^p = \begin{pmatrix} 1 & \epsilon_p \\ 0 & e^{ip\psi} \end{pmatrix};$$

it is easily seen that (compare (16) of the lemma in 4)

$$\epsilon_p = 1 + e^{i\psi} + \cdots + e^{i(p-1)\psi} = \frac{e^{ip\psi} - 1}{e^{i\psi} - 1}. \tag{12}$$

Hence if ψ is a rational multiple of 2π, for infinitely many p we have $N(D^p) = 2^{\frac{1}{2}}$ while otherwise $N(D^p)$ may come arbitrarily near to $2^{\frac{1}{2}}$.

On the other hand there exist matrices for which $N(A^p)/(\sum_{\nu=1}^{n} |\lambda_\nu|^{2p})^{\frac{1}{2}}$ is of the same order of magnitude as the upper bound in (6). An example is $D = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$, where $k = 2$,

$$D^p = \begin{pmatrix} 1 & p \\ 0 & 1 \end{pmatrix}, \qquad N(D^p) = (2 + p^2)^{\frac{1}{2}}. \tag{13}$$

3. **Notations and definitions.** Let $A = (a_{\nu\mu})(\nu = 1, \cdots, n; \mu = 1, \cdots, m)$ be an $n \times m$ matrix with real or complex elements. If $\Delta = \operatorname{diag}(\delta_1, \cdots, \delta_n)$ $(\delta_\nu \neq 0)$ is a diagonal matrix of order n, we write A/Δ for $\Delta^{-1}A$. A/Δ is obtained from A by dividing the ν-th row of A by $\delta_\nu(\nu = 1, \cdots, n)$. $\langle A \rangle$ will denote the matrix whose elements are $| a_{\nu\mu} |$; if A_1 and A_2 are any $n \times m$ matrices we write $A_1 \ll A_2$, if every element of A_1 is in modulus less than or equal to the corresponding element in A_2 and $\langle A_2 \rangle = A_2$. Clearly $\langle A_1 + A_2 \rangle \ll \langle A_1 \rangle + \langle A_2 \rangle$ and $\langle A_1 A_2 \rangle \ll \langle A_1 \rangle \langle A_2 \rangle$ if both products exist. By I we will denote the unity matrix and by E, E_1 or E_2 (square or rectangular) matrices with elements 1. We call a sequence of $n \times m$ matrices $A_\nu(\nu = 1, 2, \cdots)$ *bounded*, if for a certain constant $c > 0$

$$\langle A_\nu \rangle \ll cE \qquad (\nu = 1, 2, \cdots).$$

A *triangular* matrix is a square matrix with zeros *below* the principal diagonal.

In order to state our second theorem we consider the following seven types of triangular matrices D to each of which we define a "p-th adjoint" diagonal matrix $D^{(p)}(p = 1, 2, \cdots)$:

I. $D = (d_{\nu\mu}) \quad (\nu, \mu = 1, \cdots, m), \quad | d_{11} | > | d_{22} | > \cdots > | d_{mm} | > 0.$

$$D^{(p)} = \operatorname{diag}(d_{11}^p, \cdots, d_{mm}^p).$$

II. $D = (d_{\nu\mu}) \quad (\nu, \mu = 1, \cdots, m), \quad | d_{11} | = | d_{22} | = \cdots = | d_{mm} | > 0,$

$$d_{\nu\nu} \neq d_{\mu\mu} \quad (\nu \neq \mu).$$

$$D^{(p)} = \operatorname{diag}(d_{11}^p, \cdots, d_{mm}^p).$$

III. $D = (D_{\lambda\kappa}) \quad (\lambda, \kappa = 1, \cdots, k),$

where $D_{\kappa\kappa}(\kappa = 1, \cdots, k)$ are triangular matrices of order $n_\kappa > 1$, all diagonal elements of which have the same value d_κ, $D_{\lambda\kappa}(\lambda > \kappa)$ are zero $n_\lambda \times n_\kappa$ matrices and $D_{\lambda\kappa}(\lambda < \kappa)$ arbitrary $n_\lambda \times n_\kappa$ matrices. We further assume

$$| d_1 | = | d_2 | = \cdots = | d_k | > 0, \qquad d_\kappa \neq d_\lambda \qquad \text{if} \qquad \kappa \neq \lambda.$$

In terms of its (scalar) elements we may write $D = (d_{\nu\mu})$ $(\nu, \mu = 1, \cdots, m)$ and define $D^{(p)}$ to be

$$D^{(p)} = \operatorname{diag}(p^{m-1}d_{11}^p, p^{m-2}d_{22}^p, \cdots, d_{mm}^p).$$

III′. $D = (d_{\nu\mu}) \quad (\nu, \mu = 1, \cdots, m), \quad | d_{11} | = \cdots = | d_{mm} | > 0.$

$$D^{(p)} = \operatorname{diag}(p^{m-1}d_{11}^p, p^{m-2}d_{22}^p, \cdots, d_{mm}^p).$$

IV.
$$D = \begin{pmatrix} D_1 & B \\ 0 & D_2 \end{pmatrix},$$

where the block D_1 is of type III, the modulus of all diagonal elements being d, $D_2 = (d_{\nu\mu})$ $(\nu, \mu = 1, \cdots, m)$ is a matrix of type II such that $|d_{\mu\mu}| = d$ $(\mu = 1, \cdots, m)$ and no $d_{\mu\mu}$ is equal to a diagonal element in D_1. Then

$$D^{(p)} = \operatorname{diag}(pD_1^{(p)}, D_2^{(p)}).$$

V. $D = (D_{\lambda\kappa})$ $(\lambda, \kappa = 1, \cdots, k)$,

where the diagonal blocks $D_{\kappa\kappa}(\kappa = 1, \cdots, k)$ are of one of the types II, III, or IV and d_κ are the moduli of the diagonal elements in $D_{\kappa\kappa}$ such that

$$d_1 > d_2 > \cdots > d_k > 0.$$

As in III we suppose $D_{\lambda\kappa} = 0$ $(\lambda > \kappa)$ and $D_{\lambda\kappa}$ $(\lambda < \kappa)$ to be arbitrary. We define

$$D^{(p)} = \operatorname{diag}(D_{11}^{(p)}, D_{22}^{(p)}, \cdots, D_{kk}^{(p)}).$$

VI.
$$D = \begin{pmatrix} D_1 & B \\ 0 & D_2 \end{pmatrix},$$

where D_1 can be of type I-V and $D_2 = (d_{\nu\mu})$ $(\nu, \mu = 1, \cdots, m)$ is a triangular matrix with zeros along the principal diagonal. Then

$$D^{(p)} = \operatorname{diag}(D_1^{(p)}, 0).$$

4. THEOREM 2. *For a matrix of type* I *there exists a triangular matrix* G *with elements* 1 *along the principal diagonal such that* $D^p/D^{(p)} \to G$ *as* $p \to \infty$. *For the remaining types* II-VI, III′ *we have*

$$\langle D^p \rangle \ll c\langle D^{(p)} \rangle E \qquad (p \geq m), \tag{14}$$

where c *is a positive constant,* $m = 1$, *if* D *is of type* II-V *or* III′, *and* m *is equal to the number of zero diagonal elements, if* D *is of type* VI.

In the cases II-V, III′ (14) clearly means that $D^p/D^{(p)}$ $(p = 1, 2, \cdots)$ is a *bounded* sequence.

LEMMA. *Let* $A = \left(\begin{smallmatrix} R & T \\ 0 & S \end{smallmatrix}\right)$, *where* R, S *are square matrices. Then* $A^p = \left(\begin{smallmatrix} R^p & T_p \\ 0 & S^p \end{smallmatrix}\right)$, *where*

$$T_p = R^{p-1}T + R^{p-2}TS + \cdots + RTS^{p-2} + TS^{p-1}, \tag{15}$$

i.e. if R^{-1} *exists,*

$$T_p = R^{p-1}[T + R^{-1}TS + \cdots + (R^{-1})^{p-1}TS^{p-1}]. \tag{16}$$

Proof. For $p = 2$ we have, using block-multiplication,

$$T_2 = RT + TS.$$

Hence we may assume our lemma to be true for powers less than p. Then

$$T_p = RT_{p-1} + TS^{p-1} = R[R^{p-2}T + \cdots + TS^{p-2}] + TS^{p-1}$$

which is (15).

Proof of Theorem 2. We first prove the Theorem 2 for matrices of one of the types I, II or III′. Since any matrix of type III is also a matrix of type III′ this will include the proof for matrices of type III.

The assertion concerning I, II and III′ are clearly true for $m = 1$. We may therefore assume Theorem 2 to be true for matrices of order $< m$ and of type I, II and III′. Let D be of order m; we divide D in four submatrices, in the following way:

$$D = \begin{pmatrix} d_{11} & \vdots & B \\ \cdots & \cdots & \cdots \\ 0 & \vdots & D_{m-1} \end{pmatrix}.$$

Then we have

$$D^p = \begin{pmatrix} d_{11}^p & \vdots & B_p \\ \cdots & \cdots & \cdots \\ 0 & \vdots & D_{m-1}^p \end{pmatrix}, \qquad \frac{D^p}{D^{(p)}} = \begin{pmatrix} d_{11}^p/\delta_{11}^{(p)} & \vdots & B_p/\delta_{11}^{(p)} \\ \cdots & \cdots & \cdots \\ 0 & \vdots & D_{m-1}^p/D_{m-1}^{(p)} \end{pmatrix},$$

where by (16)

$$B_p = d_{11}^{p-1}B\left[I + \frac{D_{m-1}}{d_{11}} + \cdots + \left(\frac{D_{m-1}}{d_{11}}\right)^{p-1}\right], \tag{17}$$

and for $\delta_{11}^{(p)}$ we put d_{11}^p or $p^{m-1}d_{11}^p$ according to the types I, II or III′. It is sufficient to prove the boundedness (or, when D is of type I, the convergence) of $B_p/\delta_{11}^{(p)}$. We consider now two cases:

(a) *Types* I, II. From (17) it follows that

$$\frac{B_p}{d_{11}^p} = B(D_{m-1} - d_{11}I)^{-1}\left[\left(\frac{D_{m-1}}{d_{11}}\right)^p - I\right]. \tag{18}$$

For the type I,

$$\left(\frac{D_{m-1}}{d_{11}}\right)^p \to 0 \quad \text{as} \quad p \to \infty.$$

(For if all eigenvalues of a matrix A lie inside the unity circle, we have the convergent development $(I - A)^{-1} = \sum_{\nu=0}^{\infty} A^\nu$, so that $A^\nu \to 0$ $(\nu \to \infty)$.) Therefore

$$\lim_{p\to\infty} \frac{B_p}{d_{11}^p} = B(d_{11}I - D_{m-1})^{-1}, \tag{19}$$

and since, if D is of Type II,

$$\left\langle \frac{D_{m-1}^{p}}{d_{11}^{p}} \right\rangle = \left\langle \frac{D_{m-1}^{p}}{D_{m-1}^{(p)}} \right\rangle \qquad (p = 1, 2, \cdots),$$

the right-hand side of (18) is bounded by hypothesis.

(b) *Type* III′. Since $\delta_{11}^{(p)} = p^{m-1} d_{11}^{p}$, it suffices to prove the boundedness of

$$\phi_p = \frac{1}{p^{m-1}} \sum_{\pi=1}^{p-1} \left(\frac{D_{m-1}}{d_{11}} \right)^{\pi} \qquad \text{with} \qquad p = 1, 2, \cdots .$$

By hypothesis for a constant $c > 0$ we have

$$\langle D_{m-1}^{\pi} \rangle \ll c \langle D_{m-1}^{(\pi)} \rangle E, \qquad \sum_{\pi=1}^{p-1} \langle D_{m-1}^{\pi} \rangle \ll c \sum_{\pi=1}^{p-1} \langle D_{m-1}^{(\pi)} \rangle E. \tag{20}$$

Put

$$\left\langle \frac{D_{m-1}^{(\pi)}}{d_{11}^{\pi}} \right\rangle = \operatorname{diag} (\pi^{m-2}, \pi^{m-3}, \cdots, 1) = \Delta_{m-1}^{(\pi)} ;$$

then it follows from (20) that

$$\sum_{\pi=1}^{p-1} \left\langle \frac{D_{m-1}^{\pi}}{d_{11}^{\pi}} \right\rangle \ll c \sum_{\pi=1}^{p-1} \Delta_{m-1}^{(\pi)} E \ll cp \Delta_{m-1}^{(p)} E.$$

Hence ϕ_p is bounded.

Type IV. Without loss of generality we may assume $d = 1$; for if $d \neq 1$, write $D = d\overline{D}$; then $D^{(p)} = d^p \overline{D}^{(p)}$, $D^p / D^{(p)} = \overline{D}^p / \overline{D}^{(p)}$.

We put $D^p = \begin{pmatrix} D_1^p & B_p \\ 0 & D_2^p \end{pmatrix}$ and have to show that $B_p / p D_1^{(p)}$ is bounded. By (15) we have

$$B_p = D_1^{p-1} B_1 + D_1^{p-2} B_1 D_2 + \cdots + B_1 D_2^{p-1}.$$

Since D_2^{π} by II is bounded with $\pi = 1, 2, \cdots$, we have for a constant $c > 0$

$$\langle B_p \rangle \ll c[\langle D_1^{p-1} \rangle + \cdots + \langle D_1 \rangle + I]E.$$

It is therefore sufficient to prove that

$$\frac{1}{p D_1^{(p)}} \sum_{\pi=1}^{p-1} \langle D_1^{\pi} \rangle$$

is bounded, and this follows at once from

$$\langle D_1^{\pi} \rangle \ll c_1 \langle D_1^{(\pi)} \rangle E_1 , \qquad \sum_{\pi=1}^{p-1} \langle D_1^{\pi} \rangle \ll c_1 \sum_{\pi=1}^{p-1} \langle D_1^{(\pi)} \rangle E_1 \ll c_1 p \langle D_1^{(p)} \rangle E_1 ,$$

c_1 being a positive constant.

Type V. Our assertion for the type V is now true in the case $k = 1$. For

$k > 1$ we assume it to be true for smaller values of k. Let $D = (D_{\lambda\kappa})$ $(\lambda,\kappa = 1, \cdots, k)$ and write

$$D = \begin{pmatrix} D_{11} & B \\ 0 & D_{k-1} \end{pmatrix} \qquad \frac{D^p}{D^{(p)}} = \begin{pmatrix} D_{11}^p/D_{11}^{(p)} & \vdots & B_p/D_{11}^{(p)} \\ \cdots & \vdots & \cdots \\ 0 & \vdots & D_{k-1}^p/D_{k-1}^{(p)} \end{pmatrix}.$$

$$D^p = \begin{pmatrix} D_{11}^p & B_p \\ 0 & D_{k-1}^p \end{pmatrix}$$

By (16) we have

$$\frac{B_p}{D_{11}^{(p)}} = \frac{D_{11}^{p-1}}{D_{11}^{(p)}} [B + D_{11}^{-1}BD_{k-1} + \cdots + (D_{11}^{-1})^{p-1}BD_{k-1}^{p-1}].$$

Since $D_{11}^{p-1}/D_{11}^{(p)}$ is bounded we have only to show that

$$\phi_p = \sum_{\pi=1}^{p-1} (D_{11}^{-1})^\pi BD_{k-1}^\pi$$

is bounded. Put $D_{11}^{-1} = 1/d_1\ \overline{D}_{11}$; there exists a constant $c_1 > 0$ such that

$$\langle (D_{11}^{-1})^\pi \rangle \ll \frac{c_1}{d_1^\pi} \langle \overline{D}_{11}^{(\pi)} \rangle E_1 , \qquad \langle D_{k-1}^\pi \rangle \ll c_1 \langle D_{k-1}^{(\pi)} \rangle E_2 .$$

The diagonal of $\langle D_{k-1}^{(\pi)} \rangle$ consists of $d_2^\pi , \cdots , d_k^\pi$ multiplied by certain powers of π. We enlarge $\langle D_{k-1}^{(\pi)} \rangle$ replacing $d_i^\pi (i = 2, \cdots , k)$ by d_2^π and we write $\langle D_{k-1}^{(\pi)} \rangle \ll d_2^\pi\ \Delta_{k-1}^{(\pi)}$. Hence for a constant $c_2 > 0$ we have

$$\frac{1}{c_2} \langle (D_{11}^{-1})^\pi BD_{k-1}^\pi \rangle \ll \frac{1}{d_1^\pi} \langle \overline{D}_{11}^{(\pi)} \rangle E \langle D_{k-1}^{(\pi)} \rangle E_2 \ll \epsilon^\pi \langle \overline{D}_{11}^{(\pi)} \rangle E \Delta_{k-1}^{(\pi)} E_2 ,$$

where $0 < \epsilon < 1$, E is rectangular and $\Delta_{k-1}^{(\pi)}$, as well as $\langle \overline{D}_{11}^{(\pi)} \rangle$, is a diagonal matrix containing only powers of π along the diagonal. Thus for a constant $c_3 > 0$ and an integer $s \geq 0$ we can write

$$\epsilon^\pi \langle \overline{D}_{11}^{(\pi)} \rangle E \Delta_{k-1}^{(\pi)} E_2 \ll c_3 \epsilon^\pi \pi^s E,$$

and therefore with $c_4 > 0$

$$\langle \phi_p \rangle \ll c_4 \left[\sum_{\pi=1}^{p-1} \epsilon^\pi \pi^s \right] E.$$

Since the sum in brackets converges as $p \to \infty$, ϕ_p is bounded with $p = 1, 2, \cdots$.

Type VI. Write

$$D^p = \begin{pmatrix} D_1^p & B_p \\ 0 & D_2^p \end{pmatrix};$$

from (16) it follows that

$$\frac{B_p}{D_1^{(p)}} = \frac{D_1^{p-1}}{D_1^{(p)}}[B + D_1^{-1}BD_2 + \cdots + (D_1^{-1})^{p-1}BD_2^{p-1}], \tag{21}$$

and by the Hamilton-Cayley theorem $D_2^p = 0$ $(p \geq m)$. Hence the expression in brackets on the right-hand side of (21) is bounded with $p = 1, 2, \cdots$ and for a constant $c > 0$

$$\langle D^p \rangle \ll c\langle D^{(p)} \rangle E \qquad (p \geq m).$$

This completes the proof of Theorem 2.

The same example as given at the end of §2 shows that for matrices of type II $D^p/D^{(p)}$ is not convergent in general. In fact, (12) converges only for $\psi \equiv 0 \pmod{2\pi}$. Moreover from (13) it follows that the exponents of p in III cannot be reduced in general.

We should also note that by means of Schur's transformation (3) and (4) an inequality of type (6) could easily be deduced from Theorem 2. The exponent of p however would turn out to be less favorable than in (6).

II. Corollaries

1. Put $A^p = (a_{\nu\mu}^{(p)})$ $(\nu,\mu = 1, \cdots, n)$,

$$F_p(A) = N^{1/p}(A^p) = [\operatorname{tr} A^p(A^p)^*]^{1/2p} = \left(\sum_{\nu,\mu=1}^{n} |\, a_{\nu\mu}^{(p)} \,|^2\right)^{1/2p} \quad (p = 1, 2, \cdots) \tag{22}$$

and $\omega_A = \max_\nu |\, \lambda_\nu \,|$. Then from Theorem 1 we deduce

Theorem 3. *For every integer $p \geq 1$ we have $\omega_A \leq F_p(A)$ and $F_p(A) \to \omega_A$ as $p \to \infty$.*

Proof. We may assume $\omega_A > 0$; then by Theorem 1

$$\omega_A \leq \left(\sum_{\nu=1}^{n} |\, \lambda_\nu \,|^{2p}\right)^{1/2p} \leq F_p(A) \leq (cp^{k-1})^{1/p}\left(\sum_{\nu=1}^{n} |\, \lambda_\nu \,|^{2p}\right)^{1/2p}, \tag{23}$$

and our second assertion follows by making $p \to \infty$.

This theorem has been stated by A. B. Farnell [3] in the case that p runs through powers of 2. For the convergence of $F_p(A)$ he has indicated a proof for matrices of second order, valid also when p runs through all integers. In a way the theorem is analogous to the fact that for an essentially bounded function, integrable in the sense of Lebesgue

$$N_p(f) = \left[\int |\, f \,|^p \, dx\right]^{1/p} \to \max |\, f \,| \qquad (p \to \infty).$$

The analogy however is not perfect, since the numbers $F_p(A)$ cannot be considered as norms for matrices the triangular inequality being not satisfied in general. For from

$$F_p(A+B) \leq F_p(A) + F_p(B) \qquad (p = 1, 2, \cdots),$$

making $p \to \infty$, it would follow $\omega_{A+B} \leq \omega_A + \omega_B$ which is false in the case

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad A+B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

where $\omega_A = \omega_B = 0$, $\omega_{A+B} = 1$.

The following corollary comes out of Theorem 3 particularly easily:

COROLLARY. *For any square matrix A let $\rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_n^2$ $(\rho_\nu \geq 0)$ be the eigenvalues of AA^* and put $M(A) = \rho_1$. Then*

$$\omega_A \leq M^{1/p}(A^p) \to \omega_A \qquad (p \to \infty).$$

In fact, we have

$$\omega_A \leq M(A) \leq \left(\sum_{\nu=1}^{n} \rho_\nu^2\right)^{\frac{1}{2}} = (\operatorname{tr} AA^*)^{\frac{1}{2}},$$

and therefore by (22)

$$\omega_A \leq M^{1/p}(A^p) \leq [\operatorname{tr} A^p(A^p)^*]^{1/2p} = F_p(A).$$

2. As the values $F_p(A)$ are of rather intricate structure we give a sequence $G_p(A)$, equally convergent to ω_A, whose elements are somewhat simpler to calculate. Put

$$R_p = \max_\nu \sum_{\mu=1}^{n} |a_{\nu\mu}^{(p)}|, \qquad T_p = \max_\mu \sum_{\nu=1}^{n} |a_{\nu\mu}^{(p)}|,$$

$$G_p(A) = R_p^{1/p} \qquad (p = 1, 2, \cdots).$$

THEOREM 4. *For every integer $p \geq 1$ we have $G_p(A) \geq \omega_A$ and $G_p(A) \to \omega_A$ as $p \to \infty$.*

Proof. The inequality follows immediately from the well-known inequality $\omega_A \leq R_1$, applied to the matrix A^p. For any matrix $C = (c_{\nu\mu})(\nu,\mu = 1, 2, \cdots, n)$ we have using Schwarz's inequality

$$\sum_{\mu=1}^{n} |c_{\nu\mu}| \leq n^{\frac{1}{2}}\left(\sum_{\mu=1}^{n} |c_{\nu\mu}|^2\right)^{\frac{1}{2}} \leq n^{\frac{1}{2}}\left(\sum_{\nu,\mu=1}^{n} |c_{\nu\mu}|^2\right)^{\frac{1}{2}} = n^{\frac{1}{2}}N(C). \tag{24}$$

We apply (24) to the matrix $C = A^p$ and take ν such that $\sum_{\mu=1}^{n} |a_{\nu\mu}^{(p)}| = R_p$. This gives

$$\omega_A \leq R_p^{1/p} \leq n^{1/2p}F_p(A)$$

and, by making $p \to \infty$, our assertion follows from Theorem 3.

Since A and its transpose A' have the same eigenvalues, Theorem 4 holds for the sequence $T_p^{1/p}$ as well; in particular

$$H_p(A) = \min^{1/p}(R_p, T_p)$$

tends to ω_A as $p \to \infty$.

3. It is natural to ask under which conditions the sequences $F_p(A)$ and $G_p(A)$ will *monotonically* decrease towards ω_A. In this direction we prove

THEOREM 5. *If $p_\nu > 0$ $(\nu = 1, 2, \cdots)$ is a strictly increasing sequence of integers, such that every p_ν is divisible by $p_{\nu-1}$ $(\nu = 2, 3, \cdots)$, then $F_{p_\nu}(A)$, $G_{p_\nu}(A)$ are monotonically decreasing towards ω_A as $\nu \to \infty$.* The monotony of $F_{p_\nu}(A)$ and $G_{p_\nu}(A)$ has been proved by A. B. Farnell [3], A. Brauer [2] respectively for the special case $p_\nu = 2^{\nu-1}$.

Proof. We first prove that for any integer $k \geq 1$

$$F_{k+1}(A) \leq F_1(A)\left[\frac{F_k(A)}{F_1(A)}\right]^{k/(k+1)} \qquad (k = 1, 2, \cdots). \tag{25}$$

In fact, applying Schwarz's inequality we have

$$|a_{\nu\mu}^{(k+1)}|^2 = \left|\sum_{\lambda=1}^{n} a_{\nu\lambda}^{(k)} a_{\lambda\mu}\right|^2 \leq \left(\sum_{\lambda=1}^{n} |a_{\nu\lambda}^{(k)}|\,|a_{\lambda\mu}|\right)^2 \leq \sum_{\lambda,\kappa=1}^{n} |a_{\nu\lambda}^{(k)}|^2 |a_{\kappa\mu}|^2.$$

Summing over ν and μ we get

$$\sum_{\nu,\mu=1}^{n} |a_{\nu\mu}^{(k+1)}|^2 \leq \left(\sum_{\kappa,\mu=1}^{n} |a_{\kappa\mu}|^2\right)\left(\sum_{\nu,\lambda=1}^{n} |a_{\nu\lambda}^{(k)}|^2\right),$$

that is,

$$F_{k+1}(A) \leq F_1(A)^{1/(k+1)} F_k(A)^{k/(k+1)},$$

which is (25).

We now have to show that for any two integers q, p $(q > p \geq 1)$ related by $q = kp$ (k an integer) $F_q(A) \leq F_p(A)$. It is sufficient to prove

$$F_k(A) \leq F_1(A) \qquad (k = 1, 2, \cdots). \tag{26}$$

For since $F_{kp}^p(A) = F_k(A^p)$, (26) implies

$$F_q^p(A) = F_{kp}^p(A) = F_k(A^p) \leq F_1(A^p) = F_p^p(A).$$

(26) is trivial for $k = 1$, therefore using the inequality (25), (26) is seen to be true for $k = 1, 2, \cdots$.

The proof for $G_p(A)$ goes in a similar way. For $\nu = 1, 2, \cdots$ we have

$$\sum_{\mu=1}^{n} |a_{\nu\mu}^{(k+1)}| = \sum_{\mu=1}^{n} \left|\sum_{\lambda=1}^{n} a_{\nu\lambda}^{(k)} a_{\lambda\mu}\right| \leq \sum_{\lambda=1}^{n} \left(\sum_{\mu=1}^{n} |a_{\nu\lambda}^{(k)}|\,|a_{\lambda\mu}|\right) \leq R_1 R_k .$$

Choosing ν such that $\sum_{\mu=1}^{n} | a_{\nu\mu}^{(k+1)} | = R_{k+1}$, we get

$$G_{k+1}^{k+1}(A) \leq G_1(A) G_k^k(A),$$

that is,

$$G_{k+1}(A) \leq G_1(A) \left[\frac{G_k(A)}{G_1(A)} \right]^{k/(k+1)} \qquad (k = 1, 2, \cdots). \tag{27}$$

Again, since $G_k(A^p) = G_{kp}^p(A)$ it is sufficient to prove $G_k(A) \leq G_1(A)$ ($k = 1, 2, \cdots$) which can be done in the same way as before using (27) instead of (25). This completes the proof of Theorem 5.

In a second note we will extend the results obtained in this section to more general norms.

References

1. A. C. Aitken, *Determinants and matrices*, University Mathematical Texts, vol. 1(1939).
2. Alfred Brauer, *Limits for the characteristic roots of a matrix*, this Journal, vol. 13(1946), pp. 387–395.
3. A. B. Farnell, *Limits for the field of values of a matrix*, American Mathematical Monthly, vol. 52(1945), pp. 488–493.
4. R. A. Frazer, W. J. Duncan, A. R. Collar, *Elementary matrices*, Cambridge, 1938.
5. G. H. Hardy, J. E. Littlewood, G. Pólya, *Inequalities*, Cambridge, 1934.
6. Rufus Oldenburger, *Infinite powers of matrices and characteristic roots*, this Journal, vol. 6(1940), pp. 357–361.
7. I. Schur, *Über die charakteristischen Wurzeln einer linearen Substitution mit einer Anwendung auf die Theorie der Integralgleichungen*, Mathematische Annalen, vol. 66(1909), pp. 488–510.
8. Olga Taussky, *Analytical methods in hypercomplex systems*, Compositio Mathematica, vol. 3(1936), pp. 399–407.
9. Olga Taussky and John Todd, *Infinite powers of matrices*, Journal of the London Mathematical Society, vol. 17(1942), pp. 146–151.
10. H. W. Turnbull, A. C. Aitken, *An introduction to the theory of canonical matrices*, second edition, London and Glasgow, 1945.
11. J. H. M. Wedderburn, *The absolute value of the product of two matrices*, Bulletin American Mathematical Society, vol. 31(1925), pp. 304–308.
12. J. H. M. Wedderburn, *Lectures on matrices*, American Mathematical Society Colloquium Publications, vol. 17(1934).

University of Basle,
Switzerland.

THE ASYMPTOTIC BEHAVIOUR OF POWERS OF MATRICES. II.

"The asymptotic behaviour of powers of matrices. II", *Duke Math. J.* **20**, 375–379 (1953).

THE ASYMPTOTIC BEHAVIOUR OF POWERS OF MATRICES. II.

By Werner Gautschi

This note is an addendum to our paper [3]. We will extend the results obtained in II of [3] by introducing more general norms, and from this we derive further sequences converging towards ω_A. The enumeration of the equations, theorems and sections will be continued from [3].

III. Generalizations

1. To any column-vector $x = (x_1, \cdots, x_n)$ of the n-dimensional complex Euclidean space let a number $\phi(x)$, called the *norm* of x, be assigned, satisfying the following three conditions:

(i) $\phi(x) > 0$ except for the null vector $x = 0$, for which $\phi(0) = 0$,
(ii) $\phi(\lambda x) = |\lambda| \phi(x)$ for any complex scalar λ,
(iii) $\phi(x + y) \leq \phi(x) + \phi(y)$.

Furthermore, suppose that $\phi(x)$ is bounded over the set of vectors with Euclidean length $|x| = 1$,

(iv) $$\phi(x) \leq C \qquad (|x| = 1).$$

Let a function of the vector x, $\psi(x)$, satisfy (i) and (ii) and be bounded from below by a *positive* constant for all vectors x with $|x| = 1$,

(v) $$\psi(x) \geq c > 0 \qquad (|x| = 1).$$

Then for an $n \times n$ matrix $A = (a_{\nu\mu})$ the ratio $\phi(Ax)/\psi(x)$ remains bounded over the set of all vectors $x \neq 0$; we may therefore define its least upper bound

(28) $$\Omega_{\phi,\psi}(A) \equiv \sup_{x \neq 0} \frac{\phi(Ax)}{\psi(x)}$$

as the (upper) *norm of A induced by ϕ and ψ*. (Compare for this definition A. Ostrowski [5].) $\Omega_{\phi,\psi}(A)$ is a special case of the most general norm $\Omega(A)$ defined by the three properties:

(vi) $\Omega(A) > 0$ except when $A = 0$, in which case $\Omega(0) = 0$,
(vii) $\Omega(\lambda A) = |\lambda| \Omega(A)$ for any complex scalar λ,
(viii) $\Omega(A + B) \leq \Omega(A) + \Omega(B)$, A, B being $n \times n$ matrices.

If in particular we take $\psi(x) = \phi(x)$ assuming of course that ϕ satisfies (v), $\Omega_\phi \equiv \Omega_{\phi,\phi}$ also satisfies

(ix) $$\Omega_\phi(AB) \leq \Omega_\phi(A)\Omega_\phi(B) \qquad (\Omega_\phi \equiv \Omega_{\phi,\phi}),$$

Received November 5, 1952. This note is part of the author's doctoral dissertation presented to the University of Basle, Switzerland.

since

$$\frac{\phi(ABx)}{\phi(x)} = \frac{\phi(A(Bx))}{\phi(Bx)} \frac{\phi(Bx)}{\phi(x)}.$$

We mention the following three examples, where $\Omega_{\phi,\psi}$ can be given explicitly in terms of characteristic values of A (compare *e.g.* [1; Chapter 1], [2] or [6]):

(a) $\phi(x) = \psi(x) = \operatorname{Max}_{\nu=1,\cdots,n} |x_\nu|, \qquad \Omega_{\phi,\psi} = \operatorname{Max}_{\nu=1,\cdots,n} \sum_{\mu=1}^{n} |a_{\nu\mu}|.$

(b) $\phi(x) = \psi(x) = \sum_{\nu=1}^{n} |x_\nu|, \qquad \Omega_{\phi,\psi} = \operatorname{Max}_{\mu=1,\cdots,n} \sum_{\nu=1}^{n} |a_{\nu\mu}|.$

(c) $\phi(x) = \psi(x) = |x| = \left(\sum_{\nu=1}^{n} |x_\nu|^2\right)^{\frac{1}{2}}, \qquad \Omega_{\phi,\psi} = \rho_{\max},$

where $\rho_{\max}^2$ denotes the dominant eigenvalue of the matrix A^*A. Since $\operatorname{Inf}_{x\neq 0} |Ax|/|x|$ is equal to $\rho_{\min}$, $\rho_{\min}^2$ being the smallest eigenvalue of A^*A, we can estimate Ω_ϕ in terms of $\rho_{\max}$, $\rho_{\min}$:

$$\frac{c}{C}\rho_{\min} \leq \Omega_\phi(A) \leq \frac{C}{c}\rho_{\max}. \tag{29}$$

Indeed, from (iv) and (v) it follows that

$$\frac{c}{C}\rho_{\min} \leq \frac{c}{C}\frac{|Ax|}{|x|} \leq \frac{\phi(Ax)}{\phi(x)} \leq \frac{C}{c}\frac{|Ax|}{|x|} \leq \frac{C}{c}\rho_{\max}.$$

Another class of examples is given by

(d) $$\phi(x) = (\bar{x}'Hx)^{\frac{1}{2}}, \qquad \psi(x) = (\bar{x}'Kx)^{\frac{1}{2}},$$

where H, K denote two positive definite Hermitian matrices. A detailed discussion of the norm $\Omega_{H,K}(A)$ induced by this choice of ϕ, ψ will be given elsewhere [4].

(e) $\phi(x) = \left(\sum_{\nu=1}^{n} |x_\nu|^r\right)^{1/r}$, $\psi(x) = \left(\sum_{\nu=1}^{n} |x_\nu|^{r'}\right)^{1/r'}$, $\frac{1}{r} + \frac{1}{r'} = 1$ $(1 \leq r \leq \infty)$.

We will denote the corresponding norm of A by $\Omega_{r,r'}(A)$. This norm can easily be estimated from above as follows: put $y = Ax$, $y = (y_1, \cdots, y_n)$; then by Hölder's inequality

$$|y_\nu|^r \leq \left(\sum_{\mu=1}^{n} |a_{\nu\mu}|^r\right)\left(\sum_{\mu=1}^{n} |x_\mu|^{r'}\right)^{r/r'} \qquad (\nu = 1, \cdots, n).$$

Hence summing over $\nu = 1, \cdots, n$ and taking the r-th root on both sides we get

$$\phi(y) \leq \left(\sum_{\nu,\mu=1}^{n} |a_{\nu\mu}|^r\right)^{1/r} \psi(x), \tag{30}$$

that is,

$$\Omega_{r,r'}(A) \leq \left(\sum_{\nu,\mu=1}^{n} |a_{\nu\mu}|^r\right)^{1/r}. \tag{31}$$

The limiting cases $r = \infty$ or $r' = \infty$ of (31) follow from (30) by letting $r \to \infty$, $r' \to \infty$ respectively in (30) (compare [5], in particular p. 788).

2. The arguments by which we proved the right-hand side of (6) remain unchanged if we use instead of $N(A)$ any norm $\Omega(A)$ satisfying (vi)-(viii). Indeed, the steps leading from (10) to (11) and (11a) can still be applied with Ω instead of N, since they only require (vi)-(viii). We thus obtain the following extension of (6):

THEOREM 7. *Let Ω be a norm for $n \times n$ matrices with the properties* (vi)-(viii). *Then under the assumptions and with the notations of Theorem* 1 *we have*

$$\Omega(A^p) \leq cp^{k-1}\Big(\sum_{\nu=1}^{n} |\lambda_\nu|^{2p}\Big)^{\frac{1}{2}} \qquad (p = 1, 2, \cdots), \tag{32}$$

where c is a certain positive constant depending only on A and Ω.

Obviously the left-hand side of (6) with Ω instead of N is not true in general, since (vi)-(viii) still hold if Ω is replaced by $\epsilon\Omega$ for any $\epsilon > 0$.

3. In analogy to (22) we put

$$\Phi_p(A) \equiv \Omega_{\phi,\psi}^{1/p}(A^p) \qquad (p = 1, 2, \cdots), \tag{33}$$

where, for the sake of simplicity, we have dropped the indices ϕ, ψ on the left-hand side. Then from Theorem 7 we deduce the following theorem which contains Theorem 4 and the corollary of Theorem 3 as special cases:

THEOREM 8. *For any ϕ, ψ as defined in* 1 *we have*

$$\Phi_p(A) \to \omega_A \qquad (p \to \infty). \tag{34}$$

Moreover, if $\phi(x) = \psi(x)$ then all $\Phi_p(A)$ are bounded from below by ω_A :

$$\Phi_p(A) \geq \omega_A \qquad (p = 1, 2, \cdots). \tag{35}$$

Proof. Denote by λ an eigenvalue of A with $|\lambda| = \omega_A$ and by x an eigenvector corresponding to λ. Then obviously $\lambda^p x = A^p x$ $(p = 1, 2, \cdots)$ and from (ii) and (28) it follows that

$$\omega_A^p \phi(x) = \phi(A^p x) \leq \Omega_{\phi,\psi}(A^p)\psi(x),$$

$$\omega_A\Big(\frac{\phi(x)}{\psi(x)}\Big)^{1/p} \leq \Phi_p(A) \qquad (p = 1, 2, \cdots). \tag{36}$$

Since (vi)-(viii) hold for $\Omega = \Omega_{\phi,\psi}$ we can combine (36) with (32) and get

$$\omega_A\Big(\frac{\phi(x)}{\psi(x)}\Big)^{1/p} \leq \Phi_p(A) \leq (cp^{k-1})^{1/p}\Big(\sum_{\nu=1}^{n} |\lambda_\nu|^{2p}\Big)^{1/2p}$$

whence (34) follows by making $p \to \infty$.

If $\phi(x) = \psi(x)$, then (35) follows directly from (36).

Again, from Theorem 8 it follows that $\Phi_p(A)$ cannot be considered as a norm for matrices, since the triangular inequality is not satisfied in general, the reason being the same as that given in [3; II].

It is interesting to notice that Theorem 5 can also be extended in the same way:

THEOREM 9. *If $p_\nu > 0$ ($\nu = 1, 2, \cdots$) is a strictly increasing sequence of integers such that every p_ν is divisible by $p_{\nu-1}$ ($\nu = 2, 3, \cdots$), and if we take $\phi(x) = \psi(x)$, then the sequence $\Phi_{p_\nu}(A)$ as defined by (33) is monotonically decreasing towards ω_A as $\nu \to \infty$.*

Proof. The proof is essentially the same as for Theorem 5. Since $\phi(x) = \psi(x)$ we can apply (ix) to $A^{k+1} = A \cdot A^k$, which gives

$$\Omega_\phi(A^{k+1}) \leq \Omega_\phi(A)\Omega_\phi(A^k),$$

whence

$$\Phi_{k+1}^{k+1}(A) \leq \Phi_1(A)\Phi_k^k(A),$$

$$\Phi_{k+1}(A) \leq \Phi_1(A)\left(\frac{\Phi_k(A)}{\Phi_1(A)}\right)^{k/k+1} \qquad (k = 1, 2, \cdots). \tag{37}$$

Again we have to show that for any two integers q, p ($q > p \geq 1$) related by $q = kp$ (k an integer) $\Phi_q(A) \leq \Phi_p(A)$. Since $\Phi_k(A^p) = \Phi_{kp}^p(A)$, it is sufficient to prove

$$\Phi_k(A) \leq \Phi_1(A) \qquad (k = 1, 2, \cdots),$$

which follows from (37) by the induction argument.

4. By means of Theorems 7 and 8 we now extend Theorem 3. For any real number $r \geq 1$ let the norm $N_r(A)$ be defined by

$$N_r(A) = \left(\sum_{\nu,\mu=1}^{n} |\, a_{\nu\mu} \,|^r\right)^{1/r} \qquad (1 \leq r \leq \infty). \tag{38}$$

For $\Omega = N_r$ obviously (vi) and (vii) are satisfied while (viii) follows from Minkowski's inequality. Hence Theorem 7 is applicable and by (31), applied to the matrix A^p, we get

$$\Omega_{r,r'}(A^p) \leq N_r(A^p) \leq cp^{k-1}\left(\sum_{\nu=1}^{n} |\, \lambda_\nu \,|^{2p}\right)^{\frac{1}{2}} \qquad (p = 1, 2, \cdots). \tag{39}$$

On the other hand $\Omega_{r,r'}^{1/p}(A^p) \to \omega_A$ ($p \to \infty$) by Theorem 8. Thus by taking the p-th root on both sides of (39) and letting $p \to \infty$ the following theorem is obtained:

THEOREM 10. *For any $r \geq 1$ we have*

$$N_r^{1/p}(A^p) \to \omega_A \qquad (p \to \infty).$$

The limiting cases $r = 1, \infty$ of Theorem 10 are of particular interest, so that we may state them explicitly as

COROLLARY. *Let A be an $n \times n$ (real or complex) matrix and put $A^p = (a_{\nu\mu}^{(p)})$,*

$$L_p = \left(\sum_{\nu,\mu=1}^{n} | a_{\nu\mu}^{(p)} | \right)^{1/p}, \qquad M_p = \left(\operatorname*{Max}_{\nu,\mu=1,\cdots,n} | a_{\nu\mu}^{(p)} | \right)^{1/p} \qquad (p = 1, 2, \cdots).$$

Then the sequences L_p and M_p tend to ω_A as $p \to \infty$.

REFERENCES

1. V. N. FADDEEVA, *Computational methods of linear algebra* (Russian), Moscow-Leningrad, 1950.
2. V. N. FADDEEVA, *Basic material from linear algebra* (Translation of Chapter 1 of [1]), National Bureau of Standards, Washington, D. C.—Report 1644, 1952.
3. WERNER GAUTSCHI, *The asymptotic behaviour of powers of matrices*, this Journal, vol. 20(1953), pp. 127–140.
4. WERNER GAUTSCHI, *Bounds of matrices with regard to an Hermitian metric*, to appear in Compositio Mathematica.
5. ALEXANDRE OSTROWSKI, *Un nouveau théorème d'existence pour les systèmes d'équations*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), vol. 232(1951), pp. 786–788.
6. ALEXANDRE OSTROWSKI, *Lectures on solutions of equations and systems of equations*, Working paper CL-52-2, Washington, D. C., 1952.

UNIVERSITY OF BASLE,
SWITZERLAND.

Bounds of matrices with regard to an Hermitian metric

"Bounds of matrices with regard to an Hermitian metric", *Compositio Math*. **12**, 1–16 (1954).

Bounds of matrices with regard to an Hermitian metric [1])

by

Werner Gautschi.

§ 1. The bounds $\Omega_{H,K}$, $\omega_{H,K}$.

1. *Introduction.* In various questions concerning the solutions of systems of equations and the errors made by rounding off, the following definition of upper and lower bounds $\Omega(A)$, $\omega(A)$ of a matrix A has frequently been used:

$$\Omega(A) = \operatorname*{Max}_{\nu} \sqrt{\varrho_\nu}, \quad \omega(A) = \operatorname*{Min}_{\nu} \sqrt{\varrho_\nu}, \tag{1}$$

where ϱ_ν denote the eigenvalues of $\bar{A}'A$ (cf. e.g. [8] p. 1042 ff., or [9], p. 787, for the special case in which φ, χ are Euclidean lengths). In this paper we will discuss a generalization of this definition introducing as "parameters" two positive definite Hermitian matrices H, K. If H, K vary independently, the generalized bounds $\Omega_{H,K}(A)$, $\omega_{H,K}(A)$ can in general take values in the whole range $(0, \infty)$ (cf. § 1, section 3(vi)); to obtain appropriate values one has to couple H, K in some way. This can be done very naturally when A is an $n \times n$ matrix, by taking $K = H$. The bounds $\Omega_{H,H} \equiv \Omega_H$, $\omega_{H,H} \equiv \omega_H$ are in fact often more favourable for A than (1), but at the same time their actual calculation is considerably more difficult, as is shown by the examples given in § 2. If, however, A contains only a few non-vanishing elements, $\Omega_{H,K}(A)$ can fairly well be estimated from above by means of our theorem 2 in § 3, section 1, which generalizes a theorem due to W. Ledermann [7]. We will also make use of the theorem 2 in § 3, section 2, to determine both $\operatorname*{Inf}_H \Omega_H(A)$ and $\operatorname*{Sup}_H \omega_H(A)$, where H runs through all positive definite matrices. In sections 2 and 3 of § 1 we give the exact definitions and a few elementary properties of $\Omega_{H,K}(A)$ and $\omega_{H,K}(A)$, while in section 4 of § 1 a property not quite so trivial is proved.

[1]) This paper is part of the thesis for the Dr. phil.-degree at the University of Basle, Switzerland.

The idea of relating lengths of vectors to a positive definite Hermitian matrix H has recently been applied to the solution of linear equations $Ax = b$ by M. R. Hestenes and M. L. Stein [6]. Their main problem is to minimize the "H-length" of the residual vectors $r(x) = b - Ax$. Our definition of $\Omega_{H,K}(A)$, $\omega_{H,K}(A)$ involves a similar extremum problem, but (in contrast to [6]) with a *side condition.*

In defining the H-length of a vector we make use of the "scalar product" (x, y) with regard to H of two vectors x, y, as given e.g. in H. L. Hamburger and M. E. Grimshaw [4], p. 153. Such products (x, y) have also recently been used by W. Givens [2] to obtain theorems on the fields of values of a square matrix, which considerably extend the well known results due to O. Toeplitz [12] and F. Hausdorff [5].

I am very much indebted to Prof. Dr. A. Ostrowski for having most kindly allowed me to see through the manuscript of the yet unpublished book [10] from which I received many suggestions. In particular a chapter of [10] on the bounds (1) was the starting point of our investigations, which rather closely follow the disposition of this chapter.

2. *Notations and definitions.* Let $A = (a_{\mu\nu})$ $(\mu = 1, \ldots, m;$ $\nu = 1, \ldots, n)$ be an $m \times n$ matrix with real or complex elements $a_{\mu\nu}$. By $A^* = \bar{A}'$ we denote its conjugate-transpose and by $A^{(p)}$ $(p = 1, \ldots, \text{Min}\,(m, n))$ its p^{th} compound matrix, i.e. the $\binom{m}{p} \times \binom{n}{p}$ matrix consisting of all minors of A of order p. The groups of p rows and columns which form the minors are supposed to be arranged in lexicographical order. We have to use the following rules concerning $A^{(p)}$:

$$(AB)^{(p)} = A^{(p)}B^{(p)}, \quad (A^*)^{(p)} = (A^{(p)})^*, \tag{2}$$

if the product AB exists (cf. e.g. [1], p. 90ff). The first relation in (2) (the so called *Binet-Cauchy* theorem) is readily extended to more than two factors. Further if $m = n$ and A^{-1} exists, from (2) (with $B = A^{-1}$) it follows that

$$(A^{-1})^{(p)} = (A^{(p)})^{-1}. \tag{3}$$

tr A will denote the trace $\sum_\nu a_{\nu\nu}$ of a square matrix A, λ_A an eigenvalue of A and $|\lambda_A|^{\max}$, $|\lambda_A|^{\min}$ respectively the maximal, minimal modulus of the eigenvalues of A.

By x, y etc. we denote *column*-vectors of a k-dimensional complex Euclidean space, by x^* the conjugate-transposed row-vector

$\bar{x}'$ and by $|x|$ the Euclidean length of x. In order to introduce an Hermitian metric we define the scalar product (x, y) of two vectors x, y by

$$(x, y) \equiv y^* H x \quad (H > 0), \tag{4}$$

where H is an Hermitian matrix of order k; the meaning of the relation $H > 0$ is that H is positive definite. In particular (x, x) is real and $\geqq 0$ with $(x, x) = 0$ only when $x = 0$. We therefore define

$$\| x \| \equiv \sqrt{(x, x)} \tag{5}$$

as the *norm of x with regard to H*. Sometimes we add the subscript H and write $\| x \|_H$ instead of $\| x \|$. By routine arguments (cf. e.g. [11], p. 5, [3], p. 90—92, or [4], p. 4—5) the following three properties of $\| x \|$ are obtained:

$$\left.\begin{aligned} &\| x \| \geqq 0 \text{ with equality if and only if } x = 0 \\ &\| \gamma x \| = |\gamma| \, \| x \| \ (\gamma \text{ any complex scalar}) \\ &\| x + y \| \leqq \| x \| + \| y \|. \end{aligned}\right\} \tag{6}$$

Now let A be an $m \times n$ matrix and $H > 0, K > 0$ be Hermitian matrices of orders m, n respectively; we then define the upper and lower bounds $\Omega_{H,K}(A)$, $\omega_{H,K}(A)$ of A by

$$\left.\begin{aligned} \Omega_{H,K}(A) &= \operatorname*{Max}_{\|x\|_K = 1} \| Ax \|_H = \left(\operatorname*{Max}_{\|x\|_K = 1} x^* A^* H A x \right)^{\frac{1}{2}}, \\ \omega_{H,K}(A) &= \operatorname*{Min}_{\|x\|_K = 1} \| Ax \|_H = \left(\operatorname*{Min}_{\|x\|_K = 1} x^* A^* H A x \right)^{\frac{1}{2}}. \end{aligned}\right\} \tag{7}$$

If in particular $m = n$ and $H = K$ we write $\Omega_{H,H} \equiv \Omega_H$, $\omega_{H,H} \equiv \omega_H$. The definition (7) can also (partly) be expressed in terms of Euclidean lengths: Let K be transformed to a diagonal matrix by the unitary matrix U:

$$D = U^* K U = \operatorname{Diag}(k_1, \ldots, k_n), \ U^* U = I_n, \tag{8}$$

where I_n is the $n \times n$ unity matrix. Since $k_\nu > 0$ $(\nu = 1, \ldots, n)$ D can further be reduced to I_n by multiplying on the right and left by $\Delta = \operatorname{Diag}\left(\frac{1}{\sqrt{k_1}}, \ldots, \frac{1}{\sqrt{k_n}}\right)$:

$$\Delta U^* K U \Delta = I_n, \ \Delta = \operatorname{Diag}\left(\frac{1}{\sqrt{k_1}}, \ldots, \frac{1}{\sqrt{k_n}}\right). \tag{9}$$

If we now apply to x the substitution $x = U\Delta y$ we get

$$\| Ax \|_H^2 = x^* A^* H A x = y^* \Delta U^* A^* H A U \Delta y \ (x = U \Delta y)$$

and by (9)

$$x^*Kx = y^*\Delta U^*KU\Delta y = y^*y.$$

Hence $\| x \|_K = 1$ implies $| y | = 1$ and viceversa; we therefore have

$$\mathfrak{F}_{\|x\|_K=1} \| Ax \|_H^2 = \mathfrak{F}_{|y|=1} y^*By, \quad B = \Delta U^*A^*HAU\Delta, \tag{10}$$

where $\mathfrak{F}_{\|x\|_K=1}$, $\mathfrak{F}_{|y|=1}$ denote the fields of values over the sets of vectors x, y with $\| x \|_K = 1$, $| y | = 1$ respectively. Since B is non-negative definite, from (10) we see that both Max, Min in (7) actually exist and

$$\Omega_{H,K}^2(A) = \lambda_B^{\max}, \quad \omega_{H,K}^2(A) = \lambda_B^{\min}, \quad B = \Delta U^*A^*HAU\Delta. \tag{11}$$

If in particular we take $H = I_m$, $K = I_n$, so that clearly $U = \Delta = I_n$, we obtain the bounds defined in (1).

Throughout this paper we denote respectively by $h_1, \ldots, h_m > 0$, $k_1, \ldots, k_n > 0$ the eigenvalues (not necessarily distinct and arranged in any order) of H, K and we put $h' = \operatorname{Max}_{\mu=1,\ldots,m} h_\mu$, $h'' = \operatorname{Min}_{\mu=1,\ldots,m} h_\mu$; $k' = \operatorname{Max}_{\nu=1,\ldots,n} k_\nu$, $k'' = \operatorname{Min}_{\nu=1,\ldots,n} k_\nu$.

3. *Elementary properties of $\Omega_{H,K}$, $\omega_{H,K}$.* If not otherwise stated in this section A, $H > 0$, $K > 0$ are respectively $m \times n$, $m \times m$, $n \times n$ matrices.

(i) The following properties of $\Omega_{H,K}$, $\omega_{H,K}$ are immediate consequences of (6) and (7):

$$\begin{gathered}\Omega_{H,K}(\gamma A) = |\gamma| \Omega_{H,K}(A), \omega_{H,K}(\gamma A) = |\gamma| \omega_{H,K}(A) \ (\gamma \text{ any complex scalar}) \\ \Omega_{H,K}(A+B) \leqq \Omega_{H,K}(A) + \Omega_{H,K}(B), \quad \omega_{H,K}(A+B) \geqq \omega_{H,K}(A) - \Omega_{H,K}(B),\end{gathered} \tag{12}$$

$$\begin{aligned}&\Omega_{H,K}(A) = 0 \text{ if and only if } A = 0 \\ &\omega_{H,K}(A) = 0 \text{ if and only if the rank of } A \text{ is } < n.\end{aligned} \tag{13}$$

(ii) Obviously we can also write

$$\Omega_{H,K}(A) = \operatorname*{Max}_{x \neq 0} \frac{\| Ax \|_H}{\| x \|_K}, \quad \omega_{H,K}(A) = \operatorname*{Min}_{x \neq 0} \frac{\| Ax \|_H}{\| x \|_K}, \tag{14}$$

so that for any $m \times n$ matrix C:

$\| Cx \|_H \leqq \Omega_{H,K}(C) \| x \|_K$, $\| Cx \|_H \geqq \omega_{H,K}(C) \| x \|_K$. Hence, if A, B, $L > 0$ respectively are $m \times l$, $l \times n$, $l \times l$ matrices, we have

$$\Omega_{H,K}(AB) \leqq \Omega_{H,L}(A)\Omega_{L,K}(B), \omega_{H,K}(AB) \geqq \omega_{H,L}(A)\omega_{L,K}(B). \tag{15}$$

On the other hand, for any vector x with $Bx \neq 0$

$$\frac{\| ABx \|_H}{\| x \|_K} = \frac{\| A(Bx) \|_H}{\| Bx \|_L} \frac{\| Bx \|_L}{\| x \|_K} \quad (Bx \neq 0). \tag{16}$$

Suppose now that for the vector x: $\frac{\| Bx \|_L}{\| x \|_K} = \Omega_{L,K}(B)$. Then by (16) and (14)

$$\Omega_{H,K}(AB) \geqq \frac{\| ABx \|_H}{\| x \|_K} = \frac{\| A(Bx) \|_H}{\| Bx \|_L} \Omega_{L,K}(B) \geqq \omega_{H,L}(A) \Omega_{L,K}(B).$$

Similarly, if B is of rank n, from (16) we deduce $\omega_{H,K}(AB) \leqq \Omega_{H,L}(A) \omega_{L,K}(B)$. If B is of rank $< n$, then the same holds for AB (cf. e.g. [1], p. 96—97) and therefore $\omega_{H,K}(AB) = \omega_{L,K}(B) = 0$. Thus we can extend (15) as follows:

$$\Omega_{H,K}(AB) \geqq \omega_{H,L}(A) \Omega_{L,K}(B), \omega_{H,K}(AB) \leqq \Omega_{H,L}(A) \omega_{L,K}(B). \tag{17}$$

(iii) Suppose that $m = n$ and A^{-1} exists; then putting $x = A^{-1}y$ we see that

$$\mathop{\mathfrak{F}}_{x \neq 0} \frac{\| Ax \|_H}{\| x \|_K} = \mathop{\mathfrak{F}}_{y \neq 0} \frac{\| y \|_H}{\| A^{-1}y \|_K} = \mathop{\mathfrak{F}}_{y \neq 0} \left(\frac{\| A^{-1}y \|_K}{\| y \|_H} \right)^{-1} \quad (x = A^{-1}y).$$

Hence in using (14) we get

$$\Omega_{H,K}(A) = \frac{1}{\omega_{K,H}(A^{-1})}, \quad \omega_{H,K}(A) = \frac{1}{\Omega_{K,H}(A^{-1})}. \tag{18}$$

(iv) *Let S, T be two nonsingular matrices of orders m, n respectively; then we have*

$$\Omega_{H,K}(A) = \Omega_{S^*HS,\, T^*KT}(S^{-1}AT), \omega_{H,K}(A) = \omega_{S^*HS,\, T^*KT}(S^{-1}AT). \tag{19}$$

If in particular $m = n$, $\Omega_{H,K}$, $\omega_{H,K}$ do not change, if a unitary transformation S is applied both to A, H and K.

Indeed, putting $x = Ty$ we see that the field of values x^*A^*HAx over the set of vectors x with $x^*Kx = 1$ coincides with the field of values

$$y^*T^*A^*HATy = y^*(T^*A^*(S^*)^{-1})(S^*HS)(S^{-1}AT)y$$

taken for all vectors y with $y^*T^*KTy = 1$. Hence (19) follows at once from the definition (7).

(v) *Suppose that both A, H and K are respectively the "direct sums" of $A_1, \ldots, A_s$, $H_1, \ldots, H_s$ and $K_1, \ldots, K_s$, i.e. that in an obvious notation*

$$A = \mathrm{Diag}(A_1, \ldots, A_s), \; H = \mathrm{Diag}(H_1, \ldots, H_s), \; K = \mathrm{Diag}(K_1, \ldots, K_s),$$

where A_σ is an $m_\sigma \times n_\sigma$, $H_\sigma > 0$ an $m_\sigma \times m_\sigma$ and $K_\sigma > 0$ an $n_\sigma \times n_\sigma$ matrix $(\sigma = 1, \ldots, s)$. Then we have

$$\Omega_{H,K}(A) = \operatorname*{Max}_{\sigma=1,\ldots,s} \Omega_{H_\sigma,K_\sigma}(A_\sigma),\ \omega_{H,K}(A) = \operatorname*{Min}_{\sigma=1,\ldots,s} \omega_{H_\sigma,K_\sigma}(A_\sigma). \quad (20)$$

In fact, let K_σ be transformed to a diagonal matrix by the unitary matrix U_σ $(\sigma = 1, \ldots, s)$ and put $U = \operatorname{Diag}(U_1, \ldots, U_s)$, so that clearly (8) holds. Put $\Delta = \operatorname{Diag}\left(\frac{1}{\sqrt{k_1}}, \ldots, \frac{1}{\sqrt{k_n}}\right) = \operatorname{Diag}(\Delta_1, \ldots, \Delta_s)$, Δ_σ being of the same order as K_σ $(\sigma = 1, \ldots, s)$. Then obviously the matrix B in (10) is the direct sum of $\Delta_\sigma U_\sigma^* A_\sigma^* H_\sigma A_\sigma U_\sigma \Delta_\sigma$ $(\sigma = 1, \ldots, s)$, whence (20) follows from (11).

(vi) *For every $H > 0$, $K > 0$ we have*

$$\sqrt{\frac{h''}{k'}}\,\omega(A) \leqq \omega_{H,K}(A) \leqq \Omega_{H,K}(A) \leqq \sqrt{\frac{h'}{k''}}\,\Omega(A), \quad (21)$$

where $\omega(A)$, $\Omega(A)$ are the bounds defined in (1).

Indeed, (21) follows from (14) by putting $y = Ax$ in

$$h''\,|y|^2 \leqq \|y\|_H^2 \leqq h'\,|y|^2,\ k''\,|x|^2 \leqq \|x\|_K^2 \leqq k'\,|x|^2.$$

(vii) *We have for any eigenvalue λ_A of a square matrix A:*

$$\omega_H(A) \leqq |\lambda_A| \leqq \Omega_H(A). \quad (22)$$

In fact, let x be an eigenvector corresponding to λ_A with $\|x\|_H = 1$. Then $Ax = \lambda_A x$, $\|Ax\|_H = |\lambda_A|$, whence (22) follows directly from (7).

4. For the proof of our first theorem we need the following

LEMMA 1. *Let S be an $n \times m$ matrix and T an $m \times n$ matrix. Then, if $m < n$, we have*

$$(ST)^{(p)} = 0 \quad (p > m). \quad (23)$$

PROOF. Put $S_0 = (SO_1)$, $T_0 = \begin{pmatrix} T \\ O_2 \end{pmatrix}$, where O_1, O_2 are $n \times (n-m)$, $(n-m) \times n$ zero-matrices respectively. Obviously both S_0 and T_0 are $n \times n$ matrices and $ST = S_0 T_0$. Hence by (2) $(ST)^{(p)} = S_0^{(p)} T_0^{(p)}$, and (23) follows from $S_0^{(p)} = T_0^{(p)} = 0$ $(p > m)$.

LEMMA 2. *Suppose that $D = \operatorname{Diag}(k_1, \ldots, k_n)$ $(k_\nu > 0)$, $\Delta = \operatorname{Diag}\left(\frac{1}{\sqrt{k_1}}, \ldots, \frac{1}{\sqrt{k_n}}\right)$ and $G = \operatorname{Diag}(h_1, \ldots, h_m)$ $(h_\mu > 0)$, $\Gamma = \operatorname{Diag}\left(\frac{1}{\sqrt{h_1}}, \ldots, \frac{1}{\sqrt{h_m}}\right)$. Further let $R = (r_{\mu\nu})$ be an $m \times n$*

matrix and put $B = \Gamma RDR^*\Gamma$, $C = \Delta^{-1}R^*G^{-1}R\Delta^{-1}$. *Then, if* $\varphi(\lambda) = |\lambda I_m - B|$, $\psi(\lambda) = |\lambda I_n - C|$ *are the characteristic polynomials of* B, C, *we have*

$$\psi(\lambda) = \lambda^{n-m}\varphi(\lambda).$$

PROOF. Without loss of generality we may assume $m \leqq n$. Put $B = (b_{\mu\nu})$ $(\mu, \nu = 1, \ldots, m)$, $C = (c_{\mu\nu})$ $(\mu, \nu = 1, \ldots, n)$; by direct multiplication we get

$$b_{\mu\nu} = \frac{1}{\sqrt{h_\mu}} \frac{1}{\sqrt{h_\nu}} \sum_{\sigma=1}^{n} k_\sigma r_{\mu\sigma} \bar{r}_{\nu\sigma}$$

$$c_{\mu\nu} = \sqrt{k_\mu}\sqrt{k_\nu} \sum_{\tau=1}^{m} \frac{1}{h_\tau} \bar{r}_{\tau\mu} r_{\tau\nu}.$$

Hence

$$\operatorname{tr} B = \sum_{\mu=1}^{m} b_{\mu\mu} = \sum_{\mu=1}^{m}\sum_{\sigma=1}^{n} \frac{k_\sigma}{h_\mu} |r_{\mu\sigma}|^2 = \sum_{\nu=1}^{n}\sum_{\tau=1}^{m} \frac{k_\nu}{h_\tau} |r_{\tau\nu}|^2 = \sum_{\nu=1}^{n} c_{\nu\nu} = \operatorname{tr} C.$$

We now form the p^{th} compound matrices $B^{(p)}$, $C^{(p)}$ of B, C; from (2), (3) it follows that

$$B^{(p)} = \Gamma^{(p)}R^{(p)}D^{(p)}(R^{(p)})^*\Gamma^{(p)}$$
$$C^{(p)} = (\Delta^{(p)})^{-1}(R^{(p)})^*(G^{(p)})^{-1}R^{(p)}(\Delta^{(p)})^{-1} \quad (p = 1, \ldots, m).$$

Evidently $B^{(p)}$, $C^{(p)}$ are built analogously to B, C. Therefore our first conclusion again is applicable and we get

$$\operatorname{tr} B^{(p)} = \operatorname{tr} C^{(p)} \quad (p = 1, \ldots, m). \tag{24}$$

If $m < n$ by the lemma 1 with $S = \Delta^{-1}R^*$, $T = G^{-1}R\Delta^{-1}$ we have

$$C^{(p)} = 0 \quad (p > m). \tag{25}$$

Since generally $(-1)^p \operatorname{tr} A^{(p)}$ is the coefficient of λ^{n-p} in the characteristic polynomial $|\lambda I_n - A|$ of an $n \times n$ matrix A, (cf. e.g. [1], p. 88), our assertion now follows immediately from (24) and (25).

THEOREM 1. *Let* A *be an* $m \times n$ *matrix and* $H > 0$, $K > 0$ *be respectively of orders* m, n; *then we have*

$$\Omega_{K,H}(A^*) = \Omega_{H^{-1},K^{-1}}(A), \tag{26}$$

and, if $m = n$,

$$\omega_{K,H}(A^*) = \omega_{H^{-1},K^{-1}}(A). \tag{27}$$

PROOF. Let

$$G = V^*HV = \operatorname{Diag}(h_1, \ldots, h_m), \quad V^*V = I_m, \tag{28}$$

$$\Gamma V^*HV\Gamma = I_m, \quad \Gamma = \operatorname{Diag}\left(\frac{1}{\sqrt{h_1}}, \ldots, \frac{1}{\sqrt{h_m}}\right) \tag{29}$$

be the equations corresponding to (8), (9), applied to the matrix H. Then in using (8), (28) we have

$$D = U^*KU,\ D^{-1} = U^*K^{-1}U;\ G = V^*HV,\ G^{-1} = V^*H^{-1}V, \quad (30)$$
$$K = UDU^*,\ K^{-1} = UD^{-1}U^*;\ H = VGV^*,\ H^{-1} = VG^{-1}V^*. \quad (31)$$

According to (11) and (28)—(30) we have to examine the eigenvalues of

$$B = \Gamma V^*AKA^*V\Gamma,\ C = \Delta^{-1}U^*A^*H^{-1}AU\Delta^{-1}.$$

By means of (31) we can write

$$B = \Gamma(V^*AU)D(U^*A^*V)\Gamma = \Gamma RDR^*\Gamma$$
$$C = \Delta^{-1}(U^*A^*V)G^{-1}(V^*AU)\Delta^{-1} = \Delta^{-1}R^*G^{-1}R\Delta^{-1},$$

putting $R = V^*AU$. If we now apply the lemma 2, our assertion follows at once.

COROLLARY 1. *For any square matrix A and $H > 0$ we have*

$$\Omega_H(A^*) = \Omega_{H^{-1}}(A),\ \omega_H(A^*) = \omega_{H^{-1}}(A). \quad (32)$$

COROLLARY 2. *If A is an Hermitian matrix, then for any $H > 0$*

$$\Omega_H(A) = \Omega_{H^{-1}}(A),\ \omega_H(A) = \omega_{H^{-1}}(A). \quad (33)$$

§ 2. Examples.

For the sake of simplicity in this section we only consider square $n \times n$ matrices A and we take $H = K$. As to the selection of examples we follow very closely the arrangement given by A. Ostrowski in [10].

(i) Let $A = (a_{\mu\nu})$ be a matrix the only non-vanishing element of which is $a_{ik} \equiv a$. Put $H = (h_{\mu\nu})$, $B = (b_{\mu\nu})$, $U = (u_{\mu\nu})$, where U satisfies (8) and B is the matrix defined in (10). By direct multiplication we get

$$b_{\mu\nu} = |a|^2 h_{ii}\bar{u}_{k\mu}u_{k\nu}\frac{1}{\sqrt{h_\mu}}\frac{1}{\sqrt{h_\nu}}.$$

If by v we denote the row-vector $\left(\frac{u_{k1}}{\sqrt{h_1}}, \ldots, \frac{u_{kn}}{\sqrt{h_n}}\right)$, B can be considered as the product $|a|^2 h_{ii} v^* v$ and is therefore of rank 1. Hence by (11) we have

$$\Omega_H^2(A) = \max \lambda_B = \operatorname{tr} B = |a|^2 h_{ii} \sum_{\nu=1}^{n} \frac{|u_{k\nu}|^2}{h_\nu}.$$

On the other hand, by (31), $h_{ii} = \sum_{\nu=1}^{n} h_\nu \mid u_{i\nu} \mid^2$ and so

$$\Omega_H^2(A) = \mid a \mid^2 \sum_{\nu=1}^{n} h_\nu \mid u_{i\nu} \mid^2 \sum_{\nu=1}^{n} \frac{\mid u_{k\nu} \mid^2}{h_\nu}. \tag{34}$$

If in particular H is a diagonal matrix, and therefore $U = I_n$, we get

$$\Omega_H(A) = \mid a \mid \sqrt{\frac{h_i}{h_k}}, \quad H = \text{Diag}\,(h_1, \ldots, h_n). \tag{35}$$

Let us in this example discuss, to what extent $\Omega_H(A)$ is determined by the eigenvalues of H. Clearly all Hermitian matrices having the *fixed* eigenvalues $h_1, \ldots, h_n > 0$ are obtained by letting U in $H = UDU^*$, $D = \text{Diag}\,(h_1, \ldots, h_n)$, run through *all unitary* $n \times n$ matrices. In the case $i \neq k$, from (34) we can derive the following bounds, between which $\Omega_H^2(A)$ varies:

$$\frac{h''}{h'} \leqq \frac{1}{\mid a \mid^2} \Omega_H^2(A) \leqq \frac{h'}{h''},$$

where the upper and lower bounds are attained by taking in (34) for $(u_{i1}, \ldots, u_{in})$, $(u_{k1}, \ldots, u_{kn})$ suitable unit vectors. Similarly, if $i = k$, from (34) we see that $\frac{1}{\mid a \mid^2} \Omega_H^2(A)$ takes values in a certain closed interval, the left-hand end point of which by (22) is equal to 1.

If on the other hand we let H run through all diagonal matrices, (35) shows that in the case $i \neq k$ the range of $\Omega_H(A)$ is the whole interval $(0, \infty)$, while $\Omega_H(A)$ for $i = k$ is always equal to $\mid a \mid$.

Evidently in this example $\omega_H(A) = 0$ by (13).

(ii) Let $A = (a_{\mu\nu})$ be a matrix all elements of which are zero except those lying in the i^{th} row, and put $(a_{i1}, \ldots, a_{in}) = (a_1, \ldots, a_n) = \alpha$. We suppose that $H = \text{Diag}\,(h_1, \ldots, h_n)$, i.e. $U = I_n$. Then for the matrix B in (10) we have $B = \Delta A^* H A \Delta$, $b_{\mu\nu} = h_i \bar{a}_\mu a_\nu \frac{1}{\sqrt{h_\mu}} \frac{1}{\sqrt{h_\nu}}$ and as in our example (i) the rank of B is equal to 1, so that

$$\Omega_H^2(A) = \text{tr}\, B = h_i \sum_{\nu=1}^{n} \frac{\mid a_\nu \mid^2}{h_\nu} = \mid a_i \mid^2 + h_i \sum_{\substack{\nu=1 \\ \nu \neq i}}^{n} \frac{\mid a_\nu \mid^2}{h_\nu}. \tag{36}$$

Clearly we have always $\omega_H(A) = 0$, and, by a suitable choice of H, $\Omega_H(A)$ can take values arbitrarily near to $\mid a_i \mid = \mid \lambda_A \mid^{\max}$.

(iii) If all elements of the matrix A are equal to $a \neq 0$ and if we take $H = \text{Diag}\,(h_1, \ldots, h_n)$, we have $b_{\mu\nu} = \dfrac{1}{\sqrt{h_\mu}} \dfrac{1}{\sqrt{h_\nu}} |a|^2 \sum_{\varkappa=1}^{n} h_\varkappa$ and therefore by the Cauchy-Schwarz inequality

$$\Omega_H^2(A) = \text{tr}\, B = |a|^2 \left(\sum_{\nu=1}^{n} h_\nu\right)\left(\sum_{\nu=1}^{n} \frac{1}{h_\nu}\right) \geqq |a|^2 n^2. \tag{37}$$

The lower bound for $\Omega_H(A)$, $|a|n = |\lambda_A|^{\max}$, is attained for $H = I_n$, while $\Omega_H(A)$ is not bounded at all from above. On the other hand $\omega_H(A) = 0$.

(iv) Let $A = \text{Diag}\,(a_1, \ldots, a_n)$, $H = \text{Diag}\,(h_1, \ldots, h_n)$. Then $B = \text{Diag}\,(|a_1|^2, \ldots, |a_n|^2)$, so that

$$\Omega_H(A) = \max_\nu |a_\nu| = |\lambda_A|^{\max}, \quad \omega_H(A) = \min_\nu |a_\nu| = |\lambda_A|^{\min}. \tag{38}$$

(v) Let $A = (a_{\mu\nu})$ be a matrix all elements of which are zero except those lying in the i^{th} row and k^{th} column, while we have also $a_{ik} = 0$. We further assume H to be a diagonal matrix. In applying to both A and H the same permutation to the rows and columns, whereby in virtue of § 1, section 3(iv), $\Omega_H(A)$, $\omega_H(A)$ are not changed, we can make $k = 1$. Having carried through this transformation we denote by $\alpha = (a_2, \ldots, a_n)$ the i^{th} $(n-1)$-dimensional row-vector of A (without its first element), by $\beta = (b_1, \ldots, b_n)$ the first (n-dimensional) column-vector of A (where $b_i = 0$) and we put $H = \text{Diag}\,(h_1, \ldots, h_n)$, $\Delta_1 = \text{Diag}\left(\dfrac{1}{\sqrt{h_2}}, \ldots, \dfrac{1}{\sqrt{h_n}}\right)$. For the matrix B of (10) we then obtain by direct multiplication (observing that $U = I_n$)

$$B = \begin{pmatrix} \dfrac{1}{h_1}\beta^* H\beta & 0 \\ 0 & h_i \Delta_1 \alpha^* \alpha \Delta_1 \end{pmatrix}.$$

Since again the $(n-1) \times (n-1)$ matrix in the lower right-hand corner of B is of rank 1, it follows from (11) that

$$\Omega_H(A) = \text{Max}\,(\Omega_1, \Omega_2), \text{ where } \begin{cases} \Omega_1 = \left(\dfrac{1}{h_1} \sum\limits_{\substack{\nu=1 \\ \nu \neq i}}^{n} h_\nu |b_\nu^2|\right)^{\frac{1}{2}} \\ \Omega_2 = \left(h_i \sum\limits_{\nu=2}^{n} \dfrac{1}{h_\nu} |a_\nu|^2\right)^{\frac{1}{2}} \end{cases}. \tag{39}$$

$$\omega_H(A) = 0 \quad (n > 2).$$

§ 3. A generalization of Ledermann's theorem and the determination of $\operatorname{Inf}_{H>0} \Omega_H$, $\operatorname{Sup}_{H>0} \omega_H$.

1. The reason we succeeded to calculate directly Ω_H, ω_H in the examples given in § 2 was that the matrices A contained a sufficiently large number of zeros. We now prove a general theorem which in similar cases always yields an upper bound for $\Omega_{H,K}(A)$ and which is a generalization of a theorem due to W. Ledermann [7]. More precisely:

THEOREM 2. *Let $A = (a_{\mu\nu})$ be an $m \times n$ matrix and denote by α_μ its μ^{th} row-vector; then, if $H = \operatorname{Diag}(h_1, \ldots, h_m)$ $(h_\mu > 0)$, $K = \operatorname{Diag}(k_1, \ldots, k_n)$ $(k_\nu > 0)$ and if every column-vector of A contains at most s non-vanishing elements, we have*

$$\Omega^2_{H,K}(A) \leqq \sum_{\sigma=1}^{s} h_{\mu_\sigma} \| \alpha_{\mu_\sigma} \|^2_{K^{-1}}, \tag{40}$$

where the sum on the right-hand side has to be taken over the s largest numbers $h_{\mu_\sigma} \| \alpha_{\mu_\sigma} \|^2_{K^{-1}}$ $(\sigma = 1, \ldots, s)$ among $h_\mu \| \alpha_\mu \|^2_{K^{-1}}$ $(\mu = 1, \ldots, m)$.

PROOF. Our proof is essentially the same as that given for the case $H = I_m$, $K = I_n$ by A. Ostrowski in [10].

Without loss of generality we may assume that

$$h_1 \| \alpha_1 \|^2_{K^{-1}} \geqq h_2 \| \alpha_2 \|^2_{K^{-1}} \geqq \ldots \geqq h_m \| \alpha_m \|^2_{K^{-1}}. \tag{41}$$

Indeed, let a permutation P be applied to the rows of A; if we further permute the rows *and the columns* of H according to P, by (19) (with $T = I_n$) $\Omega_{H,K}(A)$ does not change and the numbers $h_\mu \| \alpha_\mu \|^2_{K^{-1}}$ $(\mu = 1, \ldots, m)$ are arranged as required.

Let $\operatorname{Max}_{\| x \|_{K^{-1}}} \| Ax \|^2_H$ be attained for the vector $x = (x_1, \ldots, x_n)$ and put $y = Ax$, $y = (y_1, \ldots, y_m)$. For every μ $(\mu = 1, \ldots, m)$ replace the coordinates x_ν of x for which $a_{\mu\nu} = 0$ by zeros and denote the vector so obtained by $x^{(\mu)}$. Then we have

$$\Omega^2_{H,K}(A) = \| y \|^2_H = \sum_{\mu=1}^{m} h_\mu | y_\mu |^2 = \sum_{\mu=1}^{m} h_\mu | \alpha_\mu x^{(\mu)} |^2. \tag{42}$$

We further put $x^{(\mu)} = (x^{(\mu)}_1, \ldots, x^{(\mu)}_n)$ $(\mu = 1, \ldots, m)$. Since by the Cauchy-Schwarz inequality

$$| \alpha_\mu x^{(\mu)} |^2 = \left| \sum_{\nu=1}^{n} a_{\mu\nu} x^{(\mu)}_\nu \right|^2 \leqq \left(\sum_{\nu=1}^{n} \frac{1}{\sqrt{k_\nu}} | a_{\mu\nu} | \sqrt{k_\nu} \, | x^{(\mu)}_\nu | \right)^2 \leqq$$

$$\leqq \left(\sum_{\nu=1}^{n} \frac{1}{k_\nu} | a_{\mu\nu} |^2 \right) \left(\sum_{\nu=1}^{n} k_\nu | x^{(\mu)}_\nu |^2 \right) = \| \alpha_\mu \|^2_{K^{-1}} \sum_{\nu=1}^{n} k_\nu | x^{(\mu)}_\nu |^2,$$

from (42) we get

$$\Omega^2_{H,K}(A) \leqq \sum_{\mu=1}^{m} h_\mu \,||\, \alpha_\mu \,||^2_{K^{-1}} \sum_{\nu=1}^{n} k_\nu \,|\, x_\nu^{(\mu)} \,|^2 =$$

$$= \sum_{\nu=1}^{n} k_\nu \left[\sum_{\mu=1}^{m} |\, x_\nu^{(\mu)} \,|^2 h_\mu \,||\, \alpha_\mu \,||^2_{K^{-1}} \right]. \quad (43)$$

If $x_\nu^{(\mu)} \neq 0$ then

$$x_\nu^{(\mu)} = x_\nu \quad (x_\nu^{(\mu)} \neq 0) \quad (44)$$

and $a_{\mu\nu} \neq 0$. From this and the hypothesis it follows that for any fixed ν at most s of the $x_\nu^{(\mu)}$ are $\neq 0$. Therefore taking the sum in brackets on the right-hand side of (43) only over the terms with $x_\nu^{(\mu)} \neq 0$ and using (44), (41) we see that

$$\sum_{\mu=1}^{m} |\, x_\nu^{(\mu)} \,|^2 h_\mu \,||\, \alpha_\mu \,||^2_{K^{-1}} \leqq |\, x_\nu \,|^2 \sum_{\sigma=1}^{s} h_\sigma \,||\, \alpha_\sigma \,||^2_{K^{-1}} \quad (\nu = 1, \ldots, n),$$

whence by (43)

$$\Omega^2_{H,K}(A) \leqq \left(\sum_{\nu=1}^{n} k_\nu \,|\, x_\nu \,|^2 \right) \left(\sum_{\sigma=1}^{s} h_\sigma \,||\, \alpha_\sigma \,||^2_{K^{-1}} \right).$$

This proves our assertion, since $\sum_{\nu=1}^{n} k_\nu \,|\, x_\nu \,|^2 = ||\, x \,||^2_K = 1$.

REMARKS. The theorem of Ledermann is obtained by taking $H = I_m$, $K = I_n$. If in particular we apply (40) with $H = K$ to our examples (i), (ii) and (iv) we obtain respectively as upper bounds $\frac{h_i}{h_k} \,|\, a \,|^2$, $h_i \,||\, \alpha \,||^2_{H^{-1}}$, $\operatorname{Max}_\nu |\, a_\nu \,|^2$, which all coincide with the corresponding Ω^2_H.

Even if $s = m$ the theorem 2 is often useful. Take e.g.

$$A = \begin{pmatrix} 2 & 1 & 8 \\ 7 & 0 & 5 \\ 0 & 1 & 3 \end{pmatrix},$$

where the elements of the second column are comparatively small. In order to get favourable bounds for $\Omega_H(A)$ in applying (40), we choose h_2 relatively small. With $H = \operatorname{Diag}(4, 1, 20)$ we obtain $\Omega^2_H(A) \leqq 63{,}3$, while $H = I$ gives $\Omega^2(A) \leqq 153$.

2. We now use our theorem 2 to give a refinement of (22):

THEOREM 3. *For any $n \times n$ matrix A we have*

$$\operatorname*{Inf}_{H>0} \Omega_H(A) = |\, \lambda_A \,|^{\max}, \quad \operatorname*{Sup}_{H>0} \omega_H(A) = |\, \lambda_A \,|^{\min}. \quad (45)$$

If in particular A has only simple elementary divisors both Inf and Sup in (45) are attained for suitable matrices $H > 0$.

PROOF. Since for a nonsingular matrix $|\lambda_{A^{-1}}|^{\max} = 1/|\lambda_A|^{\min}$ and by (18) $\omega_H(A) = \dfrac{1}{\Omega_H(A^{-1})}$, it is sufficient to prove the relation concerning $\operatorname{Inf} \Omega_H$. Let A be transformed to Jordan's canonical form by the nonsingular matrix S:

$$S^{-1}AS = \Lambda + C,$$

where Λ is a diagonal matrix the elements of which are the eigenvalues of A, and C denotes a matrix consisting of zeros except possibly some elements $c_{\mu\nu} = 1$ with $\nu = \mu + 1$. If in (19) we take $T = S$ we have by (22), (12)

$$|\lambda_A|^{\max} \leqq \Omega_H(A) = \Omega_K(\Lambda + C) \leqq \Omega_K(\Lambda) + \Omega_K(C), \quad (46)$$

where $K = S^*HS$. It suffices to show that for a suitable choice of K the sum on the right-hand side of (46) is arbitrarily near to $|\lambda_A|^{\max}$. Take $K = \operatorname{Diag}(k_1, \ldots, k_n)$; then by (38) $\Omega_K(\Lambda) = |\lambda_A|^{\max}$; if on the other hand γ_ν is the ν^{th} row-vector of C, then $\|\gamma_\nu\|^2_{K^{-1}} = \begin{cases} 0 (\gamma_\nu = 0) \\ 1/k_{\nu+1} (\gamma_\nu \neq 0) \end{cases}$. Hence by the theorem 2

$$\Omega^2_K(C) \leqq \operatorname*{Max}_{\nu=1,\ldots,n-1} \frac{k_\nu}{k_{\nu+1}},$$

which obviously can be made as small as we please.

If all elementary divisors of A are simple we have in (46) $C = 0$, $\Omega_K(C) = 0$, so that $\Omega_H(A)$ attains the value $|\lambda_A|^{\max}$ for a suitable matrix $H > 0$.

It is natural to ask whether we could in (45) take Inf, Sup only over the set of all *diagonal* matrices $H > 0$. This is however not true as the following example shows: Take

$$A = \begin{pmatrix} 0 & i & 1 \\ i & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad (i^2 = 1),$$

where $|\lambda_A|^{\max} = 0$. If $H = \operatorname{Diag}(h_1, h_2, h_3)$, it follows from § 2, Ex. (v), that

$$\Omega_H(A) = \operatorname{Max}(\Omega_1, \Omega_2), \quad \text{where} \quad \begin{cases} \Omega_1 = \sqrt{\dfrac{h_2 + h_3}{h_1}} \\ \Omega_2 = \sqrt{h_1\left(\dfrac{1}{h_2} + \dfrac{1}{h_3}\right)} = \sqrt{\dfrac{h_1(h_2+h_3)}{h_2 h_3}}. \end{cases}$$

But by the inequality of the arithmetic and geometric mean

$$\Omega_1\Omega_2 = \frac{h_2 + h_3}{\sqrt{h_2 h_3}} \geqq 2,$$

so that certainly $\Omega_H(A) \geqq \sqrt{2}$ for all diagonal matrices $H > 0$.

The second statement in theorem 3 can be made more precise by the following

THEOREM 4. *In order that for some matrix* $H > 0$

$$\Omega_H(A) = |\lambda_A|^{\max} \tag{47}$$

it is necessary and sufficient that the elementary divisors corresponding to the eigenvalues of A *with maximal modulus are simple. Similarly, if* A *is a non-singular matrix, we have*

$$\omega_H(A) = |\lambda_A|^{\min} \quad (|\lambda_A|^{\min} > 0) \tag{48}$$

for some $H > 0$, *if and only if all elementary divisors associated with the eigenvalues of* A *of minimal modulus are simple.*

PROOF. *Necessity*: let λ be an eigenvalue of A of either maximal or minimal modulus having multiple elementary divisors. It then suffices to show that, given a matrix $H > 0$, there always exists a vector x for which

$$\frac{\|Ax\|_H}{\|x\|_H} \begin{cases} < |\lambda_A|^{\min}, & \text{if } |\lambda| = |\lambda_A|^{\min} \\ > |\lambda_A|^{\max}, & \text{if } |\lambda| = |\lambda_A|^{\max}. \end{cases} \tag{49}$$

From Jordan's canonical form of A it is easily seen that under our hypothesis on λ there exist two linearly independent vectors u_1, u_2 such that $Au_1 = \lambda u_1$, $Au_2 = \lambda u_2 + u_1$. Put $v_1 = u_1$, $v_2 = \alpha u_1 + u_2$; in order to make v_1, v_2 orthogonal with respect to H, using the notation (4) we must have

$$(v_2, v_1) = (\alpha u_1 + u_2, u_1) = \alpha(u_1, u_1) + (u_2, u_1) = 0,$$
$$\alpha = -(u_2, u_1)/(u_1, u_1).$$

Clearly $Av_1 = \lambda v_1$, $Av_2 = \alpha\lambda u_1 + \lambda u_2 + u_1 = \lambda v_2 + v_1$, and so the vectors $w_1 = v_1/\|v_1\|_H$, $w_2 = v_2/\|v_2\|_H$ satisfy

$$\begin{aligned} Aw_1 &= \lambda w_1 \\ Aw_2 &= \lambda w_2 + \beta w_1 \end{aligned} \quad (\|w_1\|_H = \|w_2\|_H = 1,\ (w_2, w_1) = 0), \tag{50}$$

where $\beta = \|v_1\|_H/\|v_2\|_H > 0$. We now take

$$x = \gamma w_1 + w_2 \tag{51}$$

and determine the scalar γ in such a way that (49) holds. In fact, by (50)

$$Ax = \gamma\lambda w_1 + \lambda w_2 + \beta w_1 = \lambda x + \beta w_1,$$
$$\|Ax\|_H^2 = x^*A^*HAx = (\bar{\lambda}x^* + \beta w_1^*)(\lambda Hx + \beta Hw_1) =$$
$$= |\lambda|^2 \|x\|_H^2 + 2\Re(\beta\lambda w_1^* Hx) + \beta^2.$$

Substituting the expression (51) for x in w_1^*Hx we obtain

$$\frac{\|Ax\|_H^2}{\|x\|_H^2} = |\lambda|^2 + \frac{\beta}{\|x\|_H^2}\left\{2\Re(\gamma\lambda) + \beta\right\} \qquad (\beta > 0).$$

Now (49) certainly holds, if in the case $|\lambda| = |\lambda_A|^{\min}$ we choose γ such that $\Re(\gamma\lambda) < \frac{-\beta}{2}$ and $\gamma = 0$ if $|\lambda| = |\lambda_A|^{\max}$.

Sufficiency: suppose that all eigenvalues with maximal modulus have simple elementary divisors. Let A be transformed to Jordan's canonical form $S^{-1} A S = J = \mathrm{Diag}(J_1, J_2)$, where J_1 is a *diagonal* matrix containing tha eigenvalues λ with $|\lambda| = |\lambda_A|^{\max}$. By (19) we have $\Omega_H(A) = \Omega_K(J)$ $(K = S^*HS)$. To show that for a suitable matrix $K > 0$: $\Omega_K(J) = |\lambda_A|^{\max}$, take $K = \mathrm{Diag}(K_1, K_2)$, where $K_1, K_2 > 0$ are matrices of the same order as J_1, J_2 respectively and K_1 is a unity matrix. Since $\Omega_{K_1}(J_1) = |\lambda_A|^{\max}$, from (20) we get

$$\Omega_K(J) = \mathrm{Max}\{|\lambda_A|^{\max}, \Omega_{K_2}(J_2)\}.$$

On the other hand $|\lambda_{J_2}|^{\max} < |\lambda_A|^{\max}$, whence, by theorem 3, K_2 can be chosen such that $\Omega_{K_2}(J_2) < |\lambda_A|^{\max}$.

A similar argument shows that (48) holds for some $H > 0$, if all eigenvalues of A with minimal modulus are simple.

REFERENCES.

A. C. Aitken.

[1] Determinants and matrices, University Math. Texts, vol. 1, 7th edition. Edinburgh & London 1951.

[2] W. Givens.
Fields of values of a matrix, Bull. Amer. Math. Soc., vol. 58 (1952), p. 53.

P. R. Halmos.

[3] Finite dimensional vector spaces, Annals of Mathematics Studies, nr. 7, Princeton 1942.
(P. 91, line 10: instead of $|y|^2$ read $\|y\|^2$
p. 91, line 15: instead of $R[\alpha\beta(x, y)]$ read $R[\alpha\bar{\beta}(x, y)]$.)

H. L. Hamburger and M. E. Grimshaw.

[4] Linear Transformations in n-dimensional vector space, Cambridge University Press, Cambridge 1951.

F. Hausdorff.

[5] Der Wertevorrat einer Bilinearform, Math. Z., Bd. 3 (1919), pp. 314—316.

M. R. Hestenes and M. L. Stein.

[6] The Solution of Linear Equations by Minimization, NBS-NAML Report 52—45, Washington, D.C., 1951.

W. Ledermann.

[7] On an upper limit for the latent roots of a certain class of matrices, J. Lond. Math. Soc., vol. 12 (1937), pp. 12—18.

J. VON NEUMANN and H. H. GOLDSTINE.

[8] Numerical inverting of matrices of high order, Bull. Amer. Math. Soc., vol. 53 (1947), pp. 1021—1099.

A. OSTROWSKI.

[9] Un nouveau théorème d'existence pour les systèmes d'équations, C. R. Acad. Sci. Paris, t. 232 (1951), pp. 786—788.

[10] Leçons sur la résolution des systèmes d'équations, to appear in the series of the cahiers scientifiques af Gauthier-Villars, Paris.

M. H. STONE.

[11] Linear Transformations in Hilbert Space and their applications to Analysis, Amer. Math. Soc. Coll. Publ., vol. XV (1932).

O. TOEPLITZ.

[12] Das algebraische Analogon zu einem Satze von Fejér, Math. Z., Bd. 2 (1918), pp. 187—197.

(Oblatum 10-2-53).
University of Basle, Switzerland.

SOME REMARKS ON SYSTEMATIC SAMPLING

"Some remarks on systematic sampling", *Ann. Math. Statist.* **28**, 385–394 (1957).

SOME REMARKS ON SYSTEMATIC SAMPLING[1]

By Werner Gautschi[2]

University of California, Berkeley

1. Introduction and summary. Consider a finite population consisting of N elements $y_1, y_2, \cdots, y_N$. Throughout the paper we will assume that $N = nk$. A systematic sample of n elements is drawn by choosing one element at random from the first k elements $y_1, \cdots, y_k$, and then selecting every kth element thereafter. Let $y_{ij} = y_{i+(j-1)k} (i = 1, \cdots, k; j = 1, \cdots, n)$; obviously systematic sampling is equivalent to selecting *one* of the k "clusters"

$$C_i = \{y_{ij}; j = 1, \cdots, n\}$$

at random. From this it follows that the sample mean $\bar{y}_i = 1/n \sum_{j=1}^{n} y_{ij}$ is an unbiased estimate for the population mean $\bar{y} = 1/N \sum_{i=1}^{k} \sum_{j=1}^{n} y_{ij}$ and that $\operatorname{Var} \bar{y}_i = 1/k \sum_{i=1}^{k} (\bar{y}_i - \bar{y})^2$. We will denote this variance by $V_{sy}^{(1)}$ indicating by the superscript that only one cluster is selected at random. $V_{sy}^{(1)}$ can be written as

$$V_{sy}^{(1)} = S^2 - \frac{1}{k} \sum_{i=1}^{k} S_i^2, \quad \text{where} \quad S^2 = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y})^2,$$

(1)

$$S_i^2 = \frac{1}{n} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2.$$

It is natural to compare systematic sampling with stratified random sampling, where *one* element is chosen independently in each of the n strata $\{y_1, \cdots, y_k\}$, $\{y_{k+1}, \cdots, y_{2k}\}, \cdots$, and with simple random sampling using sample size n. The corresponding variances of the sample mean will be denoted by $V_{st}^{(1)}$ $V_{ran}^{(n)}$ respectively.

We consider now the following generalization of systematic sampling which appears to have been suggested by J. Tukey (see [3], p. 96, [4], [5]). Instead of choosing at first only one element at random we select a simple random sample of size s (without replacement) from the first k elements and then every kth element following those selected. In this way we obtain a sample of ns elements and, if $i_1, i_2, \cdots, i_s$ are the serial numbers of the elements first chosen, the sample mean $1/s(\bar{y}_{i_1} + \cdots + \bar{y}_{i_s})$ can be used as an estimate for the population mean. This sampling procedure is clearly equivalent to drawing a simple random sample of size s from the k clusters $C_i (i = 1, \cdots, k)$. It therefore follows (see, for example, [2], Chapter 2.3 to 2.4) that the sample mean is an unbiased estimate for the population mean and that its variance, which we denote by $V_{sy}^{(s)}$, is given by[3]

Received June 12, 1956.

[1] This investigation was supported (in part) by a research grant (RG-3666) from the Institutes of Health, Public Health Service.

[2] Now at Ohio State University

[3] This formula is not new, but appeared already in [6] and, more recently, in [5].

$$V_{sy}^{(s)} = \frac{k-s}{ks}\frac{1}{k-1}\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2 = \frac{1}{s}\frac{k-s}{k-1}V_{sy}^{(1)}. \tag{2}$$

Again, it is natural to compare this sampling procedure with stratified random sampling, where a simple random sample of size s is drawn independently in each of the n strata $\{y_1, \cdots, y_k\}$, $\{y_{k+1}, \cdots, y_{2k}\}$, $\cdots$ or with simple random sampling employing sample size ns. We denote the corresponding variances of the sample mean (which in both cases is an unbiased estimate for the population mean) by $V_{st}^{(s)}$, $V_{ran}^{(ns)}$ respectively. From well-known variance formulae (see, for example, [2], Chapters 2.4 and 5.3) it follows that

$$\begin{aligned} V_{st}^{(s)} &= \frac{1}{s}\frac{k-s}{k-1}V_{st}^{(1)}, \\ V_{ran}^{(ns)} &= \frac{N-ns}{s(N-n)}V_{ran}^{(n)} = \frac{1}{s}\frac{k-s}{k-1}V_{ran}^{(n)}. \end{aligned} \tag{3}$$

Thus the relative magnitudes of the three variances $V_{sy}^{(s)}$, $V_{st}^{(s)}$, $V_{ran}^{(ns)}$ are the same as for $V_{sy}^{(1)}$, $V_{st}^{(1)}$, $V_{ran}^{(n)}$, of which comparisons were made for several types of populations by W. G. Madow and L. H. Madow [6] and W. G. Cochran [1]. Some of the results will be reviewed in Section 3.

The object of this note is to compare systematic sampling with s random starts, as described above, with systematic sampling employing only one random start but using a sample of the same size ns. To make this comparison we obviously have to assume that k is an integral multiple of s, say $k = ls$. The latter procedure then consists in choosing one element at random from the first l elements $\{y_1, \cdots, y_l\}$ and selecting every lth consecutive element. We denote the variances of the sample mean of the two procedures by $V_k^{(s)}$, $V_l^{(1)}$ respectively, indicating by the subscript the size of the initial "counting interval." (In our notation $V_{sy}^{(s)} \equiv V_k^{(s)}$.) We shall show in Section 4 that $V_l^{(1)} = V_k^{(s)}$ in the case of a population "in random order," but $V_l^{(1)} < V_k^{(s)}$ for a population with a linear trend or with a positive correlation between the elements which is a decreasing convex function of their distance apart. Some numerical results on the relative precision of the two procedures will be given in Section 5 for the case of a large population with an exponential correlogram.

2. Acknowledgment. I wish to express my debt to Professor W. Kruskal for having brought the question treated in this note to my attention.

3. Cochran's approach. Extension of Cochran's results to systematic sampling with multiple random starts. Instead of considering a particular single population $\{y_1, y_2, \cdots, y_N\}$ we assume, following Cochran [1], [2], Chapter 8, that the y_i's are drawn from an infinite population having some specified properties. We are then interested in comparing the expected variance $E(V \mid y_1, \cdots, y_N)$ rather than $(V \mid y_1, \cdots, y_N)$ for the sampling procedures under consideration. More specifically, we consider the following three types of populations.

(i) *Population in random order.* The variates y_i are assumed to be uncor-

related and to have the same expectations. The variances may change with i

$$(4)\qquad \begin{aligned} Ey_i &= \mu, \qquad E(y_i - \mu)^2 = \sigma_i^2 \qquad (i = 1, \cdots, N); \\ E(y_i - \mu)(y_j - \mu) &= 0 \qquad (i \neq j). \end{aligned}$$

It is not difficult to show ([2], Chapter 8.5) that in this case

$$(5)\qquad EV_{sy}^{(1)} = EV_{st}^{(1)} = EV_{ran}^{(n)} = \frac{N-n}{N}\frac{\sigma^2}{n} = \frac{k-1}{k}\frac{\sigma^2}{n},$$

where $\sigma^2 = \sum_{i=1}^{N} \sigma_i^2/N$.

(ii) *Population with a linear trend.* We assume that the y_i's are uncorrelated variates whose expectations change linearly in i, more precisely

$$(6)\qquad \begin{aligned} Ey_i &= \alpha + \beta i, \qquad \operatorname{Var} y_i = \sigma^2 \qquad (i = 1, 2, \cdots, N), \\ \operatorname{Cov}(y_i, y_j) &= 0 \qquad (i \neq j). \end{aligned}$$

Applying standard linear regression theory (see, for instance, [7], Chapter 14.2) to the sum of squares in (1), it is easily found that

$$(8)\qquad EV_{sy}^{(1)} = \frac{N-n}{Nn}\sigma^2 + \beta^2\frac{k^2-1}{12} = \frac{k-1}{nk}\sigma^2 + \beta^2\frac{k^2-1}{12}.$$

In a similar way we obtain

$$(9)\qquad \begin{aligned} EV_{st}^{(1)} &= \frac{k-1}{nk}\sigma^2 + \beta^2\frac{k^2-1}{12n}, \\ EV_{ran}^{(n)} &= \frac{k-1}{nk}\sigma^2 + \beta^2\frac{(k-1)(nk+1)}{12}. \end{aligned}$$

Thus

$$(10)\qquad EV_{st}^{(1)} \leqq EV_{sy}^{(1)} \leqq EV_{ran}^{(n)},$$

with equality only if $n = 1$.

(iii) *Population with serial correlation.* It is assumed that two elements y_i, y_j are positively correlated with a correlation which depends only on the "distance" $z = |j - i|$ and which decreases as z increases. The mean and variance of all the y_i are supposed to be constant

$$(11)\qquad \begin{aligned} Ey_i &= \mu, \qquad E(y_i - \mu)^2 = \sigma^2 \qquad (i = 1, 2, \cdots, N), \\ E(y_i - \mu)(y_{i+z} - \mu) &= \rho_z\sigma^2, \end{aligned}$$

where $\rho_{z_1} \geqq \rho_{z_2} \geqq 0$ for $z_1 < z_2$. For this type of population Cochran [1] obtained the following results relevant to our purpose:

$$(12)\qquad EV_{sy}^{(1)} = \frac{k-1}{N}\sigma^2\left\{1 - \frac{2}{N(k-1)}\sum_{z=1}^{N-1}(N-z)\rho_z + \frac{2k}{n(k-1)}\sum_{z=1}^{n-1}(n-z)\rho_{kz}\right\},$$

(13) $EV_{st}^{(1)} \leqq EV_{ran}^{(n)}$,

(14) $EV_{sy}^{(1)} \leqq EV_{st}^{(1)}$,

(14) applying if, in addition, ρ_z is convex downwards.

In virtue of (2) and (3) all the results (5), (10), (13) and (14) carry over immediately to the more general sampling procedure discussed in Section 1 and, moreover, the relative sizes of the variances $V_{sy}^{(s)}$, $V_{st}^{(s)}$, $V_{ran}^{(ns)}$ remain the same as those of $V_{sy}^{(1)}$, $V_{st}^{(1)}$, $V_{ran}^{(n)}$. Numerical results of the relative precision

$$EV_{st}^{(1)}/EV_{sy}^{(1)}$$

were given by Cochran [1] for populations with a linear and exponential correlogram.

4. Comparison of systematic sampling and systematic sampling with multiple random starts.

(i) *Population in random order.* From (5), replacing k by l and n by ns, we obtain

$$EV_l^{(1)} = \frac{l-1}{lns}\sigma^2 = \frac{l-1}{N}\sigma^2.$$

On the other hand, by (2) and (5), remembering that $k = sl$,

$$EV_k^{(s)} = \frac{1}{s}\frac{k-s}{k-1}\frac{k-1}{k}\frac{\sigma^2}{n} = \frac{l-1}{N}\sigma^2.$$

Thus

(15) $$EV_l^{(1)} = EV_k^{(s)}.$$

(ii) *Population with linear trend.* By (2) and (8)

$$EV_l^{(1)} = \frac{l-1}{N}\sigma^2 + \beta^2\frac{(l-1)(l+1)}{12},$$

$$EV_k^{(s)} = \frac{1}{s}\frac{k-s}{k-1}\left[\frac{k-1}{nk}\sigma^2 + \beta^2\frac{k^2-1}{12}\right] = \frac{l-1}{N}\sigma^2 + \beta^2\frac{(l-1)(ls+1)}{12}.$$

Hence

(16) $$EV_l^{(1)} \leqq EV_k^{(s)}$$

with equality only if $s = 1$.

Both these results are, of course, to be expected intuitively. The comparison of $V_l^{(1)}$ and $V_k^{(s)}$ is, perhaps, mostly relevant for a population with a convex decreasing correlogram, since in this case $EV_l^{(1)}$ turns out to be the smallest among all the variances $EV_l^{(1)}, EV_k^{(s)}, EV_{st}^{(s)}, EV_{ran}^{(ns)}$.

(iii) *Population with serial correlation.* From (12) and (2),

$$EV_l^{(1)} = \frac{l-1}{N}\sigma^2\left\{1 - \left[\frac{2}{N(l-1)}\sum_{z=1}^{N-1}(N-z)\rho_z\right.\right.$$

$$- \frac{2l}{ns(l-1)} \sum_{z=1}^{ns-1} (ns - z)\rho_{lz} \Big] \Big\} \tag{17}$$

$$= \frac{l-1}{N} \sigma^2 \{1 - L_1\},$$

$$EV_k^{(s)} = \frac{l-1}{N} \sigma^2 \Big\{ 1 - \Big[\frac{2}{N(k-1)} \sum_{z=1}^{N-1} (N - z)\rho_z$$

$$- \frac{2k}{n(k-1)} \sum_{z=1}^{n-1} (n - z)\rho_{kz} \Big] \Big\} \tag{18}$$

$$= \frac{l-1}{N} \sigma^2 \{1 - L_2\}.$$

It is easy to check that both L_1 and L_2 are linear forms in the ρ_z's in each of which the sum of coefficients is equal to 1. Hence, in order to show that $EV_l^{(1)} \leqq EV_k^{(s)}$, it is enough to prove that

$$L = L_1 - L_2 \geqq 0, \tag{19}$$

L being a linear form of the ρ_z's whose sum of coefficients is zero. If in addition to the monotonicity the ρ_z are assumed to be convex, the following lemma, which is analogous to the lemma proved in [1], is applicable to forms of this type.

LEMMA. *Let S be the set of $\rho = \{\rho_1, \rho_2, \cdots, \rho_m\}$ for which*

$$\rho_1 \geqq \rho_2 \geqq \cdots \geqq \rho_m \geqq 0 \tag{20}$$

and

$$\Delta^2 \rho_{\mu-1} = \rho_{\mu+1} - 2\rho_\mu + \rho_{\mu-1} \geqq 0 \qquad (\mu = 2, 3, \cdots, m-1). \tag{21}$$

Let $\alpha_1, \cdots, \alpha_m$ be constants such that $\sum_{\mu=1}^{m} \alpha_\mu = 0$ and put $A_i = \sum_{\mu=1}^{i} \alpha_\mu$. Then

$$L = \sum_{\mu=1}^{m} \alpha_\mu \rho_\mu \geqq 0 \qquad \text{for all } \rho \varepsilon S$$

if and only if

$$B_j = \sum_{i=1}^{j} A_i \geqq 0 \qquad \text{for } j = 1, 2, \cdots, m-1. \tag{22}$$

Moreover, if in addition to (20) *and* (21) *strict inequality holds in* (22), *then $L > 0$ unless $\rho_1 = \cdots = \rho_m$.*

PROOF. Writing $\alpha_\mu = A_\mu - A_{\mu-1} (\mu = 1, \cdots, m; A_0 = 0)$ and using the fact that $A_m = 0$, we find

$$L = \sum_{\mu=1}^{m} A_\mu \rho_\mu - \sum_{\mu=1}^{m} A_{\mu-1} \rho_\mu = - \sum_{\mu=1}^{m-1} A_\mu \Delta\rho_\mu .$$

Similarly,

$$\sum_{\mu=1}^{m-1} A_\mu \Delta\rho_\mu = - \sum_{\mu=1}^{m-2} B_\mu \Delta^2 \rho_\mu + B_{m-1}(\rho_m - \rho_{m-1}).$$

Thus

$$L = \sum_{\mu=1}^{m-2} B_\mu \Delta^2 \rho_\mu + B_{m-1}(\rho_{m-1} - \rho_m). \tag{23}$$

Since, by hypothesis, the coefficients of all the B_μ are nonnegative, the sufficiency of (22) is clear. On the other hand, if $B_{m-1} < 0$, we could choose the ρ_μ linearly decreasing and obtain $L < 0$. If $B_j < 0$, $1 \leqq j \leqq m-2$, L could be made negative by taking, for example,

$$\rho_\mu = \begin{cases} j + 2 - \mu, & 1 \leqq \mu < j+1, \\ 1, & j+1 \leqq \mu \leqq m. \end{cases}$$

Thus (22) is also a necessary condition. If all the B_j are positive, then $L = 0$ implies $\Delta^2\rho_\mu = 0 (\mu = 1, \cdots, m-2)$, $\rho_{m-1} = \rho_m$. This in turn implies that $\rho_{m-2} = \rho_{m-1}, \rho_{m-3} = \rho_{m-2}, \cdots, \rho_1 = \rho_2$.

THEOREM. *For any population in which*

$$\rho_1 \geqq \rho_2 \geqq \cdots \geqq \rho_{N-1} \geqq 0,$$
$$\Delta^2\rho_{z-1} = \rho_{z+1} - 2\rho_z + \rho_{z-1} \geqq 0 \quad (z = 2, \cdots, N-2)$$

we have

$$EV_l^{(1)} \leqq EV_k^{(s)} \tag{24}$$

with equality only if $s = 1$ *or* $\rho_1 = \cdots = \rho_{N-1}$.

PROOF. There is nothing to prove if $s = 1$. If $s > 1$ we apply the above lemma (with $m = N-1$ and L given by (17), (18) and (19)) and show that

$$B_j > 0 \qquad j = 1, 2, \cdots, N-2. \tag{25}$$

We notice that

$$\frac{N}{2} L_1 = \frac{1}{l-1}\left[\sum_{z=1}^{N-1} (N-z)\rho_z - l^2 \sum_{z=1}^{ns-1} (ns-z)\rho_{lz}\right]$$

$$\frac{N}{2} L_2 = \frac{1}{ls-1}\left[\sum_{z=1}^{N-1} (N-z)\rho_z - (ls)^2 \sum_{z=1}^{n-1} (n-z)\rho_{(ls)z}\right].$$

To prove (25) it is enough to show that the sums B_j are positive for the form $NL/2 = NL_1/2 - NL_2/2$. We compute these sums separately for $NL_1/2$, $NL_2/2$ and then take their differences. Put[4]

$$\begin{aligned} &j = \nu k + \sigma l + \lambda = (\nu s + \sigma) l + \lambda, \qquad \text{where } \nu = 0, 1, \cdots, n-1; \\ &\sigma = 0, 1, \cdots, s-1; \qquad \lambda = 0, 1, \cdots, l-1. \end{aligned} \tag{26}$$

[4] We use the Greek letters ν, σ, λ to indicate their range $n-1$, $s-1$, $l-1$, respectively; σ, λ should not be confused with the variance symbol and the parameter to be introduced in Section 5.

By elementary computations the sums $B_j^{(1)}$ for $NL_1/2$ are found to be

$$B_j^{(1)} = \frac{1}{l-1}\{\mathrm{I} - \mathrm{II}\},$$

where

$$\mathrm{I} = \sum_{i=1}^{(\nu s+\sigma)l+\lambda} \frac{i(2N-i-1)}{2}$$

$$= \tfrac{1}{6}[(\nu s+\sigma)l+\lambda][(\nu s+\sigma)l+\lambda+1][3N-(\nu s+\sigma)l-\lambda-2]$$

$$\mathrm{II} = l^2\left[l\sum_{i=1}^{\nu s+\sigma-1}\frac{i(2ns-i-1)}{2} + (\lambda+1)\frac{(\nu s+\sigma)(2ns-\nu s-\sigma-1)}{2}\right]$$

$$= \frac{l^2(\nu s+\sigma)}{6}[l(\nu s+\sigma-1)(3ns-\nu s-\sigma-1)$$

$$+ 3(\lambda+1)(2ns-\nu s-\sigma-1)].$$

Similarly the sums $B_j^{(2)}$ for $NL_2/2$ are obtained as

$$B_j^{(2)} = \frac{1}{ls-1}\{\mathrm{I} - \mathrm{III}\},$$

where

$$\mathrm{III} = (ls)^2\left[ls\sum_{i=1}^{\nu-1}\frac{i(2n-i-1)}{2} + (\sigma l+\lambda+1)\frac{\nu(2n-\nu-1)}{2}\right]$$

$$= \frac{\nu(ls)^2}{6}[ls(\nu-1)(3n-\nu-1) + 3(\sigma l+\lambda+1)(2n-\nu-1)].$$

We have to show that

$$B_j = B_j^{(1)} - B_j^{(2)} = \frac{1}{6}\frac{l}{(l-1)(ls-1)}$$

$$\cdot\left[(s-1)6\mathrm{I} - (ls-1)\frac{6\mathrm{II}}{l} + (l-1)\frac{6\mathrm{III}}{l}\right] > 0.$$

After some elementary algebra the expression in brackets is found to be a polynomial $f(\sigma)$ in σ of third degree with the following coefficients

$$\sigma^3: l^2(l-1)$$

$$\sigma^2: -3l(l-1)[(n-\nu)ls - (\lambda+1)]$$

$$\sigma^1: l\{(l-1)[3s(n-\nu)(sl-2(\lambda+1)) - sl + 3s\lambda + 2s + 1]$$

$$- 3\lambda(\lambda+1)(s-1)\} \tag{27}$$

$$\sigma^0: (s-1)\{\nu ls(l-1)(ls-1) + \lambda(\lambda+1)[3ls(n-\nu) - (\lambda+2)]\}.$$

We notice that the second derivative $f''(\sigma)$ vanishes at

$$\sigma^* = (n - \nu)s - \frac{\lambda + 1}{l}$$

which is $\geqq s - 1$ whatever be the values of ν, λ specified by (26). For any of those values $f(\sigma)$ is therefore concave between $\sigma = 0$ and $\sigma = s - 1$ so that it is enough to show $f(0) > 0$, $f(s - 1) > 0$. Now, if $\sigma = 0$ then not both ν, λ can vanish. Hence, $f(0) > 0$ follows immediately from (27). On the other hand, $f(s - 1)/(s - 1)$, after some slight rearranging, can be written as

$$\begin{aligned}\frac{f(s-1)}{s-1} &= 3(n-\nu)sl[(l-1)(l-2(\lambda+1)) + \lambda(\lambda+1)] \\ &+ l\{l(l-1)((s-1)^2 - s) + 3(s-1)(\lambda+1)(l-1-\lambda)\} \\ &+ \lambda\{3sl(l-1) - (\lambda+1)(\lambda+2)\} + l(l-1)\{2s + \nu s(ls-1) + 1\}.\end{aligned} \tag{28}$$

The expression in brackets is a polynomial of second degree in λ with a positive leading coefficient and with roots $\lambda = l - 2$, $\lambda = l - 1$. It is therefore nonnegative for $\lambda = 0, 1, \cdots, l - 1$. It is easily verified that the quantities in the three braces are nonnegative for $l > 1$, $s > 2$ and λ, ν satisfying (26). Furthermore, the last term is positive. It remains to consider (28) for the particular case $s = 2$. We have

$$\begin{aligned}f(1) \geqq\ & 6l[(l-1)(l-2(\lambda+1)) + \lambda(\lambda+1)] \\ &+ l\{3(\lambda+1)(l-1-\lambda) - l(l-1)\} \\ &+ \lambda\{6l(l-1) - (\lambda+1)(\lambda+2)\} + 5l(l-1).\end{aligned}$$

The right-hand side is a polynomial $\varphi(\lambda)$ of third degree,

$$\varphi(\lambda) = -\lambda^3 + 3(l-1)\lambda^2 - (3l^2 - 6l + 2)\lambda + l(l-1)(5l-4),$$

whose second derivative $\varphi''(\lambda)$ vanishes at $\lambda = l - 1$. It is easy to verify that $\varphi(\lambda)$ has its relative minimum at $\lambda = l - 1 - \sqrt{3}/3$. Hence $\varphi(\lambda) > 0$ for $\lambda = 0, 1, \cdots, l - 1$ follows from

$$\varphi(l-2) = \varphi(l-1) = 2l(2l-1)(l-1) > 0.$$

This completes the proof of our theorem.

For populations with serial correlation the result (24) is to be expected also on intuitive grounds; in fact, the systematic sample is spread more evenly through the population than the sample with multiple random starts which may contain elements very close together, giving about the same information. Our proof, however, does not make clear why (24) only holds for populations with a *convex* correlogram. That (24) does not generally hold for any monotone decreasing correlogram can readily be seen by trying to apply Cochran's lemma [1] to the linear form (19). It turns out that, for example, the sum of the first l coefficients of $NL/2$ is equal to

$$\frac{-l^2}{2(ls-1)}[(2n-1)s-1] < 0,$$

One might suspect that $EV_l^{(1)} \geqq EV_k^{(s)}$ for all populations with a *concave* decreasing correlogram. However, according to our theorem $EV_l^{(1)} < EV_k^{(s)}$ for the example of a linear correlogram, so that the conjecture is not generally true.

5. Asymptotic results in the case of an exponential correlogram. We assume that $\rho_z = e^{-\lambda z}(z = 1, \cdots, N-1)$ and that both l and n are large. For n, k large Cochran [1] showed that the expression in braces of (12) is approximately equal to $1 - 2/\lambda k + 2/(e^{\lambda k} - 1)$. Since the corresponding expression $1 - L_1$ in (17) is obtained by replacing k by l and n by ns, we find

$$1 - L_1 \sim 1 - \frac{2}{\lambda l} + \frac{2}{e^{\lambda l} - 1}. \tag{29}$$

On the other hand, replacing l by $k = ls$, s by 1 in the brace of (17), we obtain $1 - L_2$ of (18). Thus

$$1 - L_2 \sim 1 - \frac{2}{\lambda ls} + \frac{2}{e^{\lambda ls} - 1}.$$

Introducing $\rho = e^{-\lambda l}$, we see that the relative precision of systematic sampling over systematic sampling with multiple random starts

$$RP = \frac{EV_k^{(s)}}{EV_l^{(1)}} \sim \frac{1 + \dfrac{2}{s \log \rho} + \dfrac{2\rho^s}{1-\rho^s}}{1 + \dfrac{2}{\log \rho} + \dfrac{2\rho}{1-\rho}}$$

depends, apart from s, only on the correlation ρ of elements of a distance l apart. Clearly $\lim_{\rho \downarrow 0} RP = 1$; also, expanding numerator and denominator in power series, it is readily seen that $\lim_{\rho \uparrow 1} RP = s$. The numerical values in Table 1 show that the limit as $\rho \downarrow 0$ is approached rather slowly.

6. Concluding remark. When the statistician has a choice between systematic sampling and systematic sampling with multiple random starts, he is more

TABLE 1

Relative precision RP of systematic sampling over systematic sampling with multiple random starts for an exponential correlogram

s	ρ										
	.01	.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2	1.34	1.53	1.66	1.80	1.87	1.92	1.95	1.97	1.99	2.00	2.00
5	1.56	1.98	2.34	2.92	3.43	3.88	4.25	4.55	4.76	4.92	4.99
10	1.63	2.13	2.58	3.40	4.26	5.19	6.23	7.32	8.39	9.31	9.85

likely to use the latter procedure because its variance can be estimated from the sample and the estimate is unbiased whatever be the form of the population. On the other hand, as we have seen in Section 5, systematic sampling is considerably more precise in the case of a population with an exponential correlogram. Thus, it may be worth while to try to find an estimate for the variance of systematic sampling which is at least consistent in some sense if the underlying assumption of an exponential correlogram is realized. In view of (17) or (29) this would involve estimating the correlation between the elements as well as σ^2.

REFERENCES

[1] W. G. COCHRAN, "Relative accuracy of systematic and stratified random samples for a certain class of populations," *Ann. Math. Stat.*, Vol. 17 (1946), pp. 164–177.
[2] W. G. COCHRAN, *Sampling Techniques*, John Wiley and Sons, 1953.
[3] W. E. DEMING, *Some Theory of Sampling*, John Wiley and Sons, 1950.
[4] H. L. JONES, "The application of sampling procedures to business operations," *J. Amer. Stat. Assoc.*, Vol. 50 (1955), pp. 763–774.
[5] H. L. JONES, "Investigating the properties of a sample mean by employing random subsample means," *J. Amer. Stat. Assoc.*, Vol. 51 (1956), pp. 54–83.
[6] W. G. MADOW AND L. H. MADOW, "On the theory of systematic sampling," *Ann. Math. Stat.*, Vol. 15 (1944), pp. 1–24.
[7] A. MOOD, *Introduction to the Theory of Statistics*, McGraw-Hill Co., 1950.

SOME REMARKS ON HERBACH'S PAPER, "OPTIMUM NATURE OF THE F-TEST FOR MODEL II IN THE BALANCED CASE"

Some remarks on Herbach's paper, "Optimum nature of the F-test for model II in the balanced case", *Ann. Math. Statist.* **30**, 960–963 (1959).

SOME REMARKS ON HERBACH'S PAPER, "OPTIMUM NATURE OF THE F-TEST FOR MODEL II IN THE BALANCED CASE"[1]

By Werner Gautschi †

Indiana University

1. Summary. The purpose of this note is to present a lemma which will settle a question of completeness left open in Section 6 of the above mentioned paper [5]. We give two applications of the lemma,

(i) by proving that, in addition to Herbach's results, also the standard F-test for $\sigma^2_{ab} = 0$ is a uniformly most powerful similar test,

(ii) by pointing out that the standard form introduced in [5] together with our lemma provide convenient tools to prove that in a balanced model II design (with the usual normality assumptions) *the standard estimates of variance components are minimum variance unbiased.* This result is well known ([2], [3]) and it has in fact been pointed out by Graybill and Wortham [3] that a completeness argument may be used to demonstrate the minimum variance property of the usual estimators for the variance components. The present lemma shows that the estimators do indeed have the necessary completeness property. We will follow Herbach's notation throughout.

2. A completeness lemma. The following lemma guarantees completeness for a certain class of probability densities to which the results of Lehmann and Scheffé do not apply directly. It takes care of a difficulty mentioned in [5], Section 6, which is caused when $g(\theta)$ does not equal one of the θ_i $(i = 2, \cdots, r)$. If $g(\theta)$ does, the product-densities could immediately be reduced to the exponential form considered by Lehmann and Scheffé in [7], Theorem 7.3. Our lemma is more general than the Lehmann and Scheffé Theorem 7.1 [7] in the sense that we allow instead of their $g''_{\theta''}(x'')$ to have $g''_{\theta',\theta''}(x'')$ which, however, we assume to factor into $h'_{\theta'}(x'')h''_{\theta''}(x'')$ with $h'_{\theta'}(x'') > 0$ and $\{h_{\theta''}(x'')\, d\mu^{x''}\}$ strongly complete. It is of course more special in that we take both $\mu^{x''}$ and $\mu^{x'|x''}$ as Lebesgue measure and for $g'_{\theta'}(x')$, $g''_{\theta',\theta''}(x'')$ specific functions. Our proof is modelled along the same lines as the one given by Lehmann and Scheffé in [7] p. 221.

Lemma: *Let*

$$\mathfrak{P}^t = \{P^t_\theta\,;\ \theta \in \mathfrak{D}\}, \qquad t = (t_2, \cdots, t_r),\ \theta = (\theta_2, \cdots, \theta_r)$$

$$\mathfrak{P}^{t_1} = \{P^{t_1}_{\theta_1,\theta}\,;\ (\theta_1, \theta) \in \mathfrak{D}_1 \times \mathfrak{D}\}, \qquad \theta_1 \text{ real}$$

Received October 8, 1957; revised January 27, 1959.

[1] This is a cut-down version of a paper in which the author independently considered standard forms for model II designs. He acknowledges, however, the priority of Dr. Herbach's approach (see [4] as compared to [1]) and restricts himself to giving some results supplementing those of Herbach.

† Werner Gautschi died on October 3, 1959. *Editor.*

be two families of probability measures on the Borel sets of the Euclidean space E_{r-1} *and the real line* E_1 *respectively, having the densities*

$$p_\theta(t) = c(\theta)h(t_2, \cdots, t_r)e^{\theta_2 t_2 + \cdots + \theta_r t_r} \tag{1}$$

$$p_{\theta_1,\theta}(t_1) = c(\theta_1, \theta)e^{g(\theta)t_1^2 + \theta_1 t_1} \tag{2}$$

with respect to Lebesgue measure. If $\mathfrak{D}_1$ *is the real line and* $\mathfrak{D}$ *a Borel set in* E_{r-1} *containing a non-degenerate* $(r-1)$*-dimensional interval then the family of product measures* $\mathfrak{P} = \{P^{t_1}_{\theta_1,\theta} \times P^{t}_{\theta} ; (\theta_1, \theta) \in \mathfrak{D}_1 \times \mathfrak{D}\}$ *is strongly complete (in the sense of Lehmann and Scheffé* [7]).

PROOF: Suppose

$$I = \iint f(t_1, t)p_{\theta_1,\theta}(t_1)p_\theta(t)\, dt_1\, dt = 0 \qquad \text{(a.e. } L^{\theta_1 \times \theta}\text{)}.^2 \tag{3}$$

Let N be the set of parameter points (θ_1, θ) for which $I \neq 0$. If N_θ denotes the θ-section of N, i.e. $N_\theta = \{\theta_1 ; (\theta_1, \theta) \in N\}$, then $L^{\theta_1}(N_\theta) = 0$ except possibly for $\theta \in N_0$, where $L^{\theta}(N_0) = 0$.

According to Fubini's theorem we may write

$$I = \int p_{\theta_1,\theta}(t_1)\Phi(t_1, \theta)\, dt_1,$$

where $\Phi(t_1, \theta) = \int f(t_1, t)p_\theta(t)\, dt$. Since $p_{\theta_1,\theta}(t_1) > 0$, for fixed $\theta \notin N_0$ the exceptional set of points t_1 for which the integral defining $\Phi(t_1, \theta)$ does not exist has L^{t_1}-measure zero. Furthermore, if $\theta \notin N_0$, we can, in virtue of (2), rewrite (3) as

$$\int e^{\theta_1 t_1}\left[e^{g(\theta)t_1^2}\Phi(t_1, \theta)\right] dt_1 = 0 \qquad \text{(a.e. } L^{\theta_1}\text{)}, \theta \notin N_0.$$

From the unicity property of the bilateral Laplace transform (see, for instance, [8], Ch. VI, Theorem 6b) it follows that

$$\Phi(t_1, \theta) = 0 \qquad \text{(a.e. } L^{t_1}\text{)}, \theta \notin N_0.$$

Thus, if S denotes the (measurable) set of points (t_1, θ) for which Φ is either not defined or $\neq 0$, almost every θ-section of S has L^{t_1}-measure zero, hence $L^{t_1 \times \theta}(S) = 0$.

This in turn implies that almost all t_1-sections of S have L^{θ}-measure zero, i.e.

$$\Phi(t_1, \theta) = \int f(t_1, t)p_\theta(t)\, dt = 0 \qquad \text{(a.e. } L^{\theta}\text{)} \quad \text{if} \quad t_1 \notin N_1,$$

where $L^{t_1}(N_1) = 0$. Since the family of probability densities $p_\theta(t)$ is strongly complete (Lehmann and Scheffé [7], Theorem 7.3) we conclude

---

[2] L with a superscript denotes Lebesgue measure. The superscript indicates the space on which the measure is taken.

$$f(t_1, t) = 0 \qquad \text{(a.e. } \mathfrak{P}^t\text{)}, t_1 \notin N_1,$$

from which $f(t_1, t) = 0$ (a.e. $\mathfrak{P}$) follows immediately.

3. Applications. (a) *Tests of hypotheses in balanced model* II *designs.* Consider the balanced two-way classification ([5], Section 6) and the hypothesis $\omega: \sigma_{ab}^2 = 0$. The statistic

$$T_1 = Z_{111}, \qquad T_2 = S_2, \qquad T_3 = S_3, \qquad T_4 = S_4 + S_5$$

is not only sufficient under ω but also complete on ω. In fact, if we let

$$\theta_1 = \frac{\sqrt{N}\mu}{\lambda_2 + \lambda_3 - \lambda_4}, \qquad \theta_2 = -\frac{1}{2\lambda_2}, \qquad \theta_3 = -\frac{1}{2\lambda_3}, \qquad \theta_4 = -\frac{1}{2\lambda_4},$$

the densities of T_1 and $T = (T_2, T_3, T_4)$ are easily recognized to have the form given in our lemma. Proceeding therefore in the same fashion as in [5], Section 6, we would find that *also the standard F-test of the hypothesis* $\omega: \sigma_{ab}^2 = 0$ *is a uniformly most powerful similar test.* The same situation prevails in higher order classifications. As is well known, in a complete n-way classification F-tests exist for the non-existence of anyone of the $(n-1)$st or $(n-2)$nd order interactions. All these tests are uniformly most powerful similar tests.

(b) *Point estimation in balanced model* II *designs.* To fix the ideas consider the standard form for the balanced two-way classification. A sufficient statistic for the parameters involved is

$$T_1 = Z_{111}, \qquad T_2 = S_2, \cdots, \qquad T_5 = S_5. \tag{4}$$

If we let

$$\theta_1 = \frac{\sqrt{N}\mu}{\lambda_2 + \lambda_3 - \lambda_4}, \qquad \theta_2 = -\frac{1}{2\lambda_2}, \cdots, \qquad \theta_5 = -\frac{1}{2\lambda_5},$$

the densities of T_1 and $T = (T_2, \cdots, T_5)$ are again of the form given in our lemma and thus the statistic (4) is complete on Ω. Unbiased estimates for the variance components, in terms of (4), are

$$\hat{\sigma}_e^2 = \frac{T_5}{\nu_e}, \quad \hat{\sigma}_{ab}^2 = \frac{1}{K}\left[\frac{T_4}{\nu_{ab}} - \frac{T_5}{\nu_e}\right], \quad \hat{\sigma}_b^2 = \frac{1}{IK}\left[\frac{T_3}{\nu_b} - \frac{T_4}{\nu_{ab}}\right], \tag{5}$$

$$\hat{\sigma}_a^2 = \frac{1}{JK}\left[\frac{T_2}{\nu_a} - \frac{T_4}{\nu_{ab}}\right],$$

where $\nu_a = I - 1$, $\nu_b = J - 1$, $\nu_{ab} = (I-1)(J-1)$, $\nu_e = IJ(K-1)$ and are therefore minimum variance unbiased estimates ([6], Theorem 5.1). On the other hand the standard estimates in terms of the various mean squares have the same distribution as those in (5) and must consequently be of minimum variance among all unbiased estimates based on the original observation vector X.

Higher order layouts could be treated in a similar manner.

REFERENCES

[1] W. GAUTSCHI, "On an optimal property of variance-components estimates" (Abstract), *Ann. Math. Stat.* Vol. 28 (1957), p. 1058.

[2] F. A. GRAYBILL, "On quadratic estimates of variance components", *Ann. Math. Stat.* Vol. 25 (1954), pp. 367–372.

[3] F. A. GRAYBILL AND A. W. WORTHAM, "A note on uniformly best unbiased estimators for variance components", *J. Amer. Stat. Assn.*, Vol. 51 (1956), pp. 266–268.

[4] L. H. HERBACH, "Topics in analysis of variance: A. Optimum properties of tests for model II, B. Generalizations of model II" (Abstract), *Ann. Math. Stat.* Vol. 24 (1953), p. 137.

[5] L. H. HERBACH, "Properties of model II—Type analysis of variance tests, A: Optimum nature of the F-test for model II in the Balanced Case", *Ann. Math. Stat.* Vol. 30 (1959), pp. 939–959.

[6] E. L. LEHMANN AND H. SCHEFFÉ, "Completeness, similar regions and unbiased estimation, Part I" *Sankhya*, Vol. 10 (1950), pp. 305–340.

[7] E. L. LEHMANN AND H. SCHEFFÉ, "Completeness, similar regions and unbiased estimation, Part II", *Sankhya*, Vol. 15 (1955), pp. 219–236.

[8] D. V. WIDDER, *The Laplace Transform*, Princeton University Press, 1941.

32

Obituaries

1 A. Ostrowski, "Werner Gautschi 1927–1959", *Verh. Naturf. Ges. Basel* **71**, Nr. 2, 314–316 (1960). (English translation by Walter Gautschi)
2 J. R. Blum, "Werner Gautschi 1927–1959", *Ann. Math. Statist.* **31**, 557 (1960).

Werner Gautschi 1927–1959

A. Ostrowski, “Werner Gautschi 1927–1959”, *Verh. Naturf. Ges. Basel* **71**, Nr. 2, 314–316 (1960). (English translation by Walter Gautschi)

Werner Gautschi*

1927–1959

On October 3, 1959, mathematics in Basel suffered a great loss. Dr. Werner Gautschi, associate professor[1] at Ohio State University, died unexpectedly of a heart attack at the age of only 32 years. A promising academic career has thus come to a bitter untimely end.

Werner Gautschi

Werner Gautschi was born in Basel on December 11, 1927, together with his twin brother Walter, with whom later in life he would intimately share both a vocation to mathematics and a love of music. After attending the

*English translation by Walter Gautschi of A. Ostrowski, "Werner Gautschi, 1927–1959", *Verh. Naturf. Ges. Basel* 71, Nr. 2 (1960), 314–316.

[1]The original erroneously has "assistant professor". (Translator's note)

primary schools and classical "Gymnasium" in Basel, he enrolled 1946 at the University of Basel. During the years 1948–1950, the undersigned had the pleasure of having him as an assistant, and to the devoted collaboration of the young student, who already then proved to have a rare acumen in the assessment of mathematical reasoning, I owe valuable furtherance of my own work.

Since in 1950 one could think again of going abroad, as used to be the custom, he went for a year to Cambridge, England. In 1952 he graduated from our university *summa cum laude* with a fine piece of work on the so-called matrix theory. Since he already expected to first enter an academic career in America, the thesis, from the start, was written in English and eventually appeared in two parts in America, in the *Duke Mathematical Journal*, and another part in Holland, in the *Compositio Mathematica*.

In 1953, Werner, with a fellowship of the Swiss National Foundation, went to the USA where, first in Princeton, he began to familiarize himself with the workings of electronic computers. But soon, he became attracted to mathematical statistics, which at the time was flourishing in Princeton. In the following years he in fact occupied himself almost exclusively with this discipline. He had to begin here almost from scratch. Indeed, during his student days there were no courses offered in actual mathematical statistics, neither in Basel nor in other Swiss universities, and the only Swiss instructor in this field had to concern himself, in Zurich and Geneva, with the more elementary parts of the subject.

In 1954, Gautschi moved to Berkeley, where around Jerzy Neyman a circle of researchers had been formed who were passionately interested in mathematical statistics and where the probably most intense mathematical-statistical activity, world-wide, had been developed. There, he had the fortune to be given the opportunity of collaborating with a group of researchers led by the famous statistician Blackwell, today probably the most important "black" mathematician.

In 1956, he began to set his sights on starting an academic teaching career. After a year as an instructor at Ohio State University, in 1958 he went to Bloomington, Indiana as an assistant professor, only to return again to Ohio in 1959.

Two publications authored by Gautschi in the *Annals of Mathematical Statistics* give only a most fragmentary picture of the struggle with problems of statistics the last years of his life were devoted to. As late as June of 1959, he reported to the undersigned about a new turn he has given to the

investigations on the classical problem of the so-called Bernoulli distribution. He must have worked on the elaboration of this result when his life came to such an abrupt end.

While scientific work has given Werner Gautschi's life the decisive direction, he has found in music an other dimension of his life. He often used to play four-hand piano with his brother Walter. It was also through mutual musical interests that he first got to know his future wife Erika, b. Wüst. The marriage brought forth a son Thomas, born only after his death.

To all who have known him, the memory of a man of unassuming demeanor and open to all that is humane, and of a devoted scholar and teacher, will remain alive.

A. Ostrowski

Werner Gautschi, 1927–1959

J. R. Blum, "Werner Gautschi, 1927–1959", *Ann. Math. Statist.* **31**, 557 (1960).

Reprinted from THE ANNALS OF MATHEMATICAL STATISTICS
Vol. 31, No. 3, September, 1960
Printed in U.S.A.

Werner Gautschi, 1927–1959

BY J. R. BLUM

Sandia Corporation

Werner Gautschi was born on December 11, 1927, in Basel. A serious heart ailment suffered as a young boy prevented him from participating in many of the usual childhood activities and led to an early devotion to mathematics and music. In 1946 he entered the University of Basel and remained there until 1952, with the exception of three terms at Cambridge University during 1950–51. He graduated *summa cum laude* from the University of Basel in 1952, with a dissertation written under the direction of Professor A. Ostrowski.

An early interest in Statistics and Computing brought him to the United States in 1953 in order to study these fields. He spent his first year here at the Institute for Advanced Studies, where he did computational work on eigenvalues and norms of matrices. In 1954 he joined the Statistical Laboratory at Berkeley for a two year period. Aside from his studies, research, and teaching, he made many valuable suggestions to Erich Lehmann who was writing *Testing Statistical Hypotheses* and to Henry Scheffé who was writing *The Analysis of Variance*.

In the fall of 1956 he joined the faculty of Ohio State University and in the fall of 1957 he came to Indiana University for a two year period. During the summer of 1958 he returned to Switzerland where he married Erika Wüst and brought her back to the United States. In the summer of 1959 he rejoined Ohio State University where he remained until his death on October 3, 1959. A son, Thomas, was born on January 25, 1960.

The death of a good man is a loss to all of us. Werner Gautschi was a good man, a fine scientist, and a sensitive pianist. His many friends and colleagues mourn him and remember him.

Bibliography of Werner Gautschi

[1] "The asymptotic behaviour of powers of matrices," *Duke Math. J.* Vol. 20 (1953), pp. 127–140.

[2] "The asymptotic behaviour of powers of matrics II," *Duke Math. J.* Vol. 20 (1953), pp. 375–379.

[3] "Bounds of matrices with regard to an Hermitian metric," *Compositio Math.* Vol. 12 (1954), pp. 1–16.

[4] "Some remarks on systematic sampling," *Ann. Math. Stat.* Vol. 28 (1957), pp. 385–394.

[5] "Some remarks on Herbach's paper, 'Optimum nature of the F-test for model II in the balanced case,'" *Ann. Math. Stat.*, Vol. 30 (1959), pp. 960–963.

Received April 7, 1960.

33

Recording

Music, as Professor Ostrowski wrote in his obituary, was "an other dimension" of Werner's life. A tribute is paid here to this dimension of Werner's life by providing a link,

`springer.com`

(type in the ISBN of Vol. 3, 978-1-4614-7131-8, and click on `Trout.wma`) to a recording of Schubert's Trout Quintet. The recording was made in Champaign, Illinois in May of 1959, five months prior to Werner's death. The performers, Werner himself, a fellow mathematician, and members of the University of Illinois Music faculty, are

Uni Sprengling Thomas – Violin
Howard Osborn – Viola
Peter Farrell – Cello
Thomas Frederickson – Double Bass
Werner Gautschi – Piano