Journal of Computational and Applied Mathematics

Volume 127, Issues 1-2, Pages 1-369 (15 January 2001)

North-Holland

Display Checked Docs | E-mail Articles | Export Citations

View: Citations Go

JOURNAL OF
COMPUTATIONAL AND
APPLIED MATHEMATICS

Volume 127, Numbers 1–2, 15 January 2001

Contents

PII: S0377-0427(00)00692-0

ELSEVIER

Journal of Computational and Applied Mathematics 127 (2001) ix–xi

JOURNAL OF
COMPUTATIONAL AND
APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Preface

Orthogonal polynomials play a prominent role in pure, applied, and computational mathematics, as well as in the applied sciences. It is the aim of the present volume in the series "Numerical Analysis in the 20th Century" to review, and sometimes extend, some of the many known results and properties of orthogonal polynomials and related quadrature rules. In addition, this volume discusses techniques available for the analysis of orthogonal polynomials and associated quadrature rules. Indeed, the design and computation of numerical integration methods is an important area in numerical analysis, and orthogonal polynomials play a fundamental role in the analysis of many integration methods.

The 20th century has witnessed a rapid development of orthogonal polynomials and related quadrature rules, and we therefore cannot even attempt to review all significant developments within this volume. We primarily have sought to emphasize results and techniques that have been of significance in computational or applied mathematics, or which we believe may lead to significant progress in these areas in the near future. Unfortunately, we cannot claim completeness even within this limited scope. Nevertheless, we hope that the readers of this volume will find the papers of interest and many references to related work of help.

We outline the contributions in the present volume. Properties of orthogonal polynomials are the focus of the papers by Marcellán and Álvarez-Nodarse and by Freund. The former contribution discusses "Favard's theorem", i.e., the question under which conditions the recurrence coefficients of a family of polynomials determine a measure with respect to which the polynomials in this family are orthogonal. Polynomials that satisfy a three-term recurrence relation as well as Szegő polynomials are considered. The measure is allowed to be signed, i.e., the moment matrix is allowed to be indefinite. Freund discusses matrix-valued polynomials that are orthogonal with respect to a measure that defines a bilinear form. This contribution focuses on breakdowns of the recurrence relations and discusses techniques for overcoming this difficulty. Matrix-valued orthogonal polynomials form the basis for algorithms for reduced-order modeling. Freund's contribution to this volume provides references to such algorithms and their application to circuit simulation.

The contribution by Peherstorfer and Steinbauer analyzes inverse images of polynomial mappings in the complex plane and their relevance to extremal properties of polynomials orthogonal with respect to measures supported on a variety of sets, such as several intervals, lemniscates, or equipotential lines. Applications include fractal theory and Julia sets.

Orthogonality with respect to Sobolev inner products has attracted the interest of many researchers during the last decade. The paper by Martinez discusses some of the recent developments

PII: S0377-0427(00)00490-8

in this area. The contribution by López Lagomasino, Pijeira, and Perez Izquierdo deals with orthogonal polynomials associated with measures supported on compact subsets of the complex plane. The location and asymptotic distribution of the zeros of the orthogonal polynomials, as well as the *n*th-root asymptotic behavior of these polynomials is analyzed, using methods of potential theory.

Investigations based on spectral theory for symmetric operators can provide insight into the analytic properties of both orthogonal polynomials and the associated Padé approximants. The contribution by Beckermann surveys these results.

Van Assche and Coussement study multiple orthogonal polynomials. These polynomials arise in simultaneous rational approximation; in particular, they form the foundation for simultaneous Hermite–Padé approximation of a system of several functions. The paper compares multiple orthogonal polynomials with the classical families of orthogonal polynomials, such as Hermite, Laguerre, Jacobi, and Bessel polynomials, using characterization theorems.

Bultheel, González-Vera, Hendriksen, and Njåstad consider orthogonal rational functions with prescribed poles, and discuss quadrature rules for their exact integration. These quadrature rules may be viewed as extensions of quadrature rules for Szegő polynomials. The latter rules are exact for rational functions with poles at the origin and at infinity.

Many of the papers of this volume are concerned with quadrature or cubature rules related to orthogonal polynomials. The analysis of multivariable orthogonal polynomials forms the foundation of many cubature formulas. The contribution by Cools, Mysovskikh, and Schmid discusses the connection between cubature formulas and orthogonal polynomials. The paper reviews the development initiated by Radon's seminal contribution from 1948 and discusses open questions. The work by Xu deals with multivariate orthogonal polynomials and cubature formulas for several regions in $\mathbb{R}^d$. Xu shows that orthogonal structures and cubature formulas for these regions are closely related.

The paper by Milovanović deals with the properties of quadrature rules with multiple nodes. These rules generalize the Gauss–Turán rules. Moment-preserving approximation by defective splines is considered as an application.

Computational issues related to Gauss quadrature rules are the topic of the contributions by Ehrich and Laurie. The latter paper discusses numerical methods for the computation of the nodes and weights of Gauss-type quadrature rules, when moments, modified moments, or the recursion coefficients of the orthogonal polynomials associated with a nonnegative measure are known. Ehrich is concerned with how to estimate the error of quadrature rules of Gauss type. This question is important, e.g., for the design of adaptive quadrature routines based on rules of Gauss type.

The contribution by Mori and Sugihara reviews the double exponential transformation in numerical integration and in a variety of Sinc methods. This transformation enables efficient evaluation of the integrals of analytic functions with endpoint singularities.

Many algorithms for the solution of large-scale problems in science and engineering are based on orthogonal polynomials and Gauss-type quadrature rules. Calvetti, Morigi, Reichel, and Sgallari describe an application of Gauss quadrature to the computation of bounds or estimates of the Euclidean norm of the error in iterates (approximate solutions) generated by an iterative method for the solution of large linear systems of equations with a symmetric matrix. The matrix may be positive definite or indefinite.

The computation of zeros of polynomials is a classical problem in numerical analysis. The contribution by Ammar, Calvetti, Gragg, and Reichel describes algorithms based on Szegő polynomials.

In particular, knowledge of the location of zeros of Szegő polynomials is important for the analysis and implementation of filters for time series.

Walter Gautschi[a]
Francisco Marcellán[b]
Lothar Reichel[c]
[a] *Department of Computer Sciences,*
Purdue University,
West Lafayette, IN 47907-1398, USA
[b] *Departamento de Matemáticas,*
Universidad Carlos III de Madrid,
Avenida Universidad 30,
E-28911 Leganés, Madrid, Spain
[c] *Department of Mathematics and Computer Science,*
Kent State University,
Kent OH 44242-0001, USA

ELSEVIER

Journal of Computational and Applied Mathematics 127 (2001) 1–16

JOURNAL OF
COMPUTATIONAL AND
APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Polynomial zerofinders based on Szegő polynomials

G.S. Ammar[a,*], D. Calvetti[b,1], W.B. Gragg[c], L. Reichel[d,2]

[a]*Department of Mathematical Sciences, Northern Illinois University, DeKalb, IL 60115, USA*
[b]*Department of Mathematics, Case Western Reserve University, Cleveland, OH 44106, USA*
[c]*Department of Mathematics, Naval Postgraduate School, Monterey, CA 93943, USA*
[d]*Department of Mathematics and Computer Science, Kent State University, Kent, OH 44242, USA*

Received 28 December 1999; received in revised form 27 March 2000

Abstract

The computation of zeros of polynomials is a classical computational problem. This paper presents two new zerofinders that are based on the observation that, after a suitable change of variable, any polynomial can be considered a member of a family of Szegő polynomials. Numerical experiments indicate that these methods generally give higher accuracy than computing the eigenvalues of the companion matrix associated with the polynomial. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Szegő–Hessenberg matrix; Companion matrix; Eigenvalue problem; Continuation method; Parallel computation

1. Introduction

The computation of the zeros of a polynomial

$$\psi_n(z) = z^n + \alpha_{n-1}z^{n-1} + \cdots + \alpha_1 z + \alpha_0, \quad \alpha_j \in \mathbb{C}, \tag{1}$$

is a fundamental problem in scientific computation that arises in many diverse applications. The conditioning of this problem has been investigated by Gautschi [8,9]. Several classical methods for determining zeros of polynomials are described by Henrici [17, Chapter 6] and Stoer and Bulirsch [26, Chapter 5]. A recent extensive bibliography of zerofinders is provided by McNamee [21].

[*] Corresponding author.

E-mail addresses: ammar@math.niu.edu (G.S. Ammar), dxc57@po.cwru.edu (D. Calvetti), gragg@math.nps.navy.mil (W.B. Gragg), reichel@mcs.kent.edu (L. Reichel).

[1] Supported in part by NSF Grant DMS 9806702.

[2] Supported in part by NSF Grant DMS 9806413.

PII: S0377-0427(00)00491-X

Among the most popular numerical methods for computing zeros of polynomials are the Jenkins–Traub algorithm [18], and the computation of the zeros as eigenvalues of the companion matrix

$$C_n = \begin{bmatrix} 0 & & & \cdots & 0 & -\alpha_0 \\ 1 & 0 & & \cdots & 0 & -\alpha_1 \\ & 1 & 0 & \cdots & 0 & -\alpha_2 \\ & & & & \vdots & \vdots \\ & & \ddots & \ddots & & \\ & & & 1 & 0 & -\alpha_{n-2} \\ 0 & & & & 1 & -\alpha_{n-1} \end{bmatrix} \in \mathbb{C}^{n\times n} \tag{2}$$

associated with the polynomial (1) by the QR algorithm after balancing; see Edelman and Murakami [7] and Moler [22]. Recently, Goedecker [10] compared these methods and found the latter approach to be competitive with several available implementations of the Jenkins–Traub algorithm with regard to both accuracy and execution time for polynomials of small to moderate degree.

This paper describes two new methods for computing zeros of polynomials. The methods are based on the observation that, after a change of variable, any polynomial can be considered a member of a family of Szegő polynomials. The new zerofinders use the recursion relation for the Szegő polynomials, which are defined as follows. Let ω be a nondecreasing distribution function with infinitely many points of increase on the unit circle in the complex plane and define the inner product

$$(f,g) := \frac{1}{2\pi}\int_{-\pi}^{\pi} f(z)\overline{g(z)}\,\mathrm{d}\omega(t), \quad z := \exp(\mathrm{i}t), \;\; \mathrm{i} := \sqrt{-1}, \tag{3}$$

for polynomials f and g, where the bar denotes complex conjugation. We assume for notational convenience that $\mathrm{d}\omega(t)$ is scaled so that $(1,1)=1$. Introduce orthonormal polynomials with respect to this inner product, $\phi_0, \phi_1, \phi_2, \ldots$, where ϕ_j is of degree j with positive leading coefficient. These polynomials are known as Szegő polynomials and many of their properties are discussed by Grenander and Szegő [16]. In particular, they satisfy the recursion relation

$$\begin{aligned} &\phi_0(z) = \phi_0^*(z) = 1, \\ &\sigma_{j+1}\phi_{j+1}(z) = z\phi_j(z) + \gamma_{j+1}\phi_j^*(z), \quad j = 0,1,2.,\ldots,n-1, \\ &\sigma_{j+1}\phi_{j+1}^*(z) = \bar{\gamma}_{j+1} z\phi_j(z) + \phi_j^*(z), \end{aligned} \tag{4}$$

where the recursion coefficients γ_{j+1} and the auxiliary coefficients σ_{j+1} are defined by

$$\begin{aligned} &\gamma_{j+1} = -\frac{(z\phi_j, 1)}{\delta_j}, \\ &\sigma_{j+1} = \sigma_j(1 - |\gamma_{j+1}|^2), \quad j = 0,1,2,\ldots, \\ &\delta_{j+1} = \delta_j\sigma_{j+1}, \quad \delta_0 = \sigma_0 = 1. \end{aligned} \tag{5}$$

It follows from (4) that the auxiliary polynomials ϕ_j^* satisfy

$$\phi_j^*(z):=z^j\bar{\phi}_j(1/z). \tag{6}$$

The zeros of the Szegő polynomials are strictly inside the unit circle and all recursion coefficients γ_j are of magnitude smaller than one; see, e.g., [1,16]. The leading coefficient of ϕ_j is $1/\delta_j$.

The first step in the new zerofinders of this paper is to determine recursion coefficients $\{\gamma_j\}_{j=1}^n$, such that the Szegő polynomial ϕ_n satisfies

$$\delta_n\phi_n(\zeta)=\eta_1^n\psi_n(z), \tag{7}$$

where

$$\zeta=\eta_1 z+\eta_2, \tag{8}$$

and the constants η_1 and η_2 are chosen so that the zeros z_j of ψ_n are mapped to zeros ζ_j of ϕ_n inside the unit circle. We refer to this change of variable as a *rescaling* of the monic polynomial $\psi_n(z)$. Its construction is discussed in Section 2. Thus, the problem of determining the zeros of ψ_n is reduced to the problem of computing the zeros of a Szegő polynomial of degree n. Section 3 considers two methods for this purpose, based on a matrix formulation of the recursion relation (4). This gives an $n \times n$ upper Hessenberg matrix whose eigenvalues are the zeros of ϕ_n. We refer to this matrix, which is described in [11], as the *Szegő–Hessenberg matrix* associated with ϕ_n. Having computed the eigenvalues ζ_j of this matrix, we use the relation (8) to compute the zeros z_j of ψ_n.

A third method for computing the zeros of $\psi_n(z)$ is to use the power-basis coefficients of the monic Szegő polynomial $\Phi_n(\zeta):=\delta_n\phi_n(\zeta)$ of (7) to form the companion matrix associated with Φ_n, compute its eigenvalues, and transform these back to the z-variable using (8). In other words, to use the companion matrix of the rescaled monic polynomial Φ_n instead of that of ψ_n. This method is included in the numerical results we report in Section 4.

Section 4 compares the use of the QR algorithm with balancing for computing the eigenvalues of the Szegő–Hessenberg, the companion matrix (2) of ψ_n, and the companion matrix of the rescaled polynomial Φ_n. We note in passing that these are all upper Hessenberg matrices. Balancing is commonly used for improving the accuracy of the computed eigenvalues; see [7] for a discussion on balancing of the companion matrix. In our experiments we found that when the parameters η_1 and η_2 for the rescaling are chosen so that all zeros of ϕ_n are inside the unit circle and one zero is close to the unit circle, the computed eigenvalues of the Szegő–Hessenberg matrix and of the companion matrix of the rescaled polynomial (7) generally provide more accurate zeros of ψ_n than those of the companion matrix of ψ_n. This rescaling is achieved by application of the Schur–Cohn test as described in Section 3. Numerous computed examples, some of which are reported in Section 4, indicate that computing eigenvalues of the Szegő–Hessenberg matrix after balancing often gives the zeros of ψ_n with higher accuracy than computing eigenvalues of the companion matrix of the scaled polynomial (7) after balancing. Both methods, in general, give higher accuracy in the computed zeros than computing the zeros of ψ_n as eigenvalues of the balanced companion matrix.

The other zerofinder for Szegő polynomials discussed in Section 3 is the continuation method previously introduced in [2]. For many polynomials ψ_n, this method yields higher accuracy than the computation of the eigenvalues of the associated companion or Szegő–Hessenberg matrices. Section 4 presents numerical examples and Section 5 contains concluding remarks.

2. Computation of Szegő polynomials

Given a polynomial $\psi_n(z)$ in power-basis form (1), we compute the recursion coefficients $\{\gamma_j\}_{j=1}^n$ of the family of Szegő polynomials $\{\phi_j\}_{j=0}^n$, chosen so that ϕ_n satisfies (7), by first transforming the polynomial ψ_n so that the average of its zeros vanishes. Then we determine a disk centered at the origin that contains all zeros of the transformed polynomial. The complex plane is then scaled so that this disk becomes the unit disk. In this fashion, the problem of determining the zeros of the polynomial ψ_n has been transformed into an equivalent problem of determining the zeros of a polynomial with all zeros in the unit disk. We may assume that the latter polynomial has leading coefficient one, and identify it with the monic Szegő polynomial $\Phi_n = \delta_n\phi_n$. Given the power-basis coefficients of Φ_n, the recursion coefficients of the family of Szegő polynomials $\{\phi_j\}_{j=0}^n$ can be computed by the Schur–Cohn algorithm. The remainder of this section describes details of the computations outlined.

Let $\{z_j\}_{j=1}^n$ denote the zeros of ψ_n and introduce the average of the zeros

$$\rho := \frac{1}{n}\sum_{j=1}^{n} z_j. \tag{9}$$

We evaluate this quantity as $\rho = -\alpha_{n-1}/n$, and define the new variable $\hat{z} = z - \rho$. The polynomial $\hat{\psi}_n(\hat{z}) := \psi_n(z)$ can be written as

$$\hat{\psi}_n(\hat{z}) = \hat{z}^n + \hat{\alpha}_{n-2}\hat{z}^{n-2} + \cdots + \hat{\alpha}_1\hat{z} + \hat{\alpha}_0. \tag{10}$$

The coefficients $\{\hat{\alpha}_j\}_{j=0}^{n-2}$ can be computed from the coefficients $\{\alpha_j\}_{j=0}^{n-1}$ in $\mathcal{O}(n^2)$ arithmetic operations.

We now scale the $\hat{z}$-plane in two steps in order to move the zeros of $\hat{\psi}_n$ inside the unit circle. Our choice of scaling is motivated by the following result mentioned by Ostrowski [23].

Proposition 2.1. *Let χ_n be a polynomial of degree n of the form*

$$\chi_n(z) = z^n + \beta_{n-2}z^{n-2} + \cdots + \beta_1 z + \beta_0, \tag{11}$$

and assume that

$$\max_{0\leqslant j\leqslant n-2} |\beta_j| = 1.$$

Then all zeros of χ_n are contained in the open disk $\{z: |z| < \frac{1}{2}(1+\sqrt{5})\}$ in the complex plane.

Proof. Let z be a zero of χ_n and assume that $|z| > 1$. Then

$$z^n = -\beta_{n-2}z^{n-2} - \cdots - \beta_1 z - \beta_0,$$

and it follows that

$$|z|^n \leqslant \sum_{j=0}^{n-2} |z|^j = \frac{|z|^{n-1}-1}{|z|-1}.$$

This inequality can be written as

$$|z|^{n-1}(|z|^2 - |z| - 1) \leqslant -1. \tag{12}$$

Since $|z|^2-|z|-1=(|z|-\frac{1}{2}(1-\sqrt{5}))(|z|-\frac{1}{2}(1+\sqrt{5}))$, inequality (12) can only hold for $|z| < \frac{1}{2}(1+\sqrt{5})$. □

After the change of variable $\tilde{z}:=\sigma\hat{z}$, where $\sigma > 0$ is chosen so that

$$\max_{2\leqslant j\leqslant n} \sigma^j|\hat{\alpha}_{n-j}| = 1,$$

the polynomial $\tilde{\psi}_n(\tilde{z}):=\sigma^n\hat{\psi}_n(\hat{z})$ satisfies the conditions of the proposition.

Define the scaling factor

$$\tau:=\frac{2}{1+\sqrt{5}}. \tag{13}$$

By Proposition 2.1 the change of variables

$$\zeta:=\tau\tilde{z} \tag{14}$$

yields a monic polynomial

$$\Phi_n^{(\tau)}(\zeta):=\tau^n\tilde{\psi}_n(\tilde{z}) \tag{15}$$

with all zeros inside the unit circle.

We identify $\Phi_n^{(\tau)}$ with the monic Szegő polynomial $\delta_n\phi_n$, and wish to compute the recursion coefficients $\{\gamma_j\}_{j=1}^n$ that determine polynomials of lower degree $\{\phi_j\}_{j=0}^{n-1}$ in the same family of Szegő polynomials; see (4). This can be done by using the relationship between the coefficients of ϕ_j in power form and the coefficients of the associated auxiliary polynomial. Specifically, it follows from (6) that if

$$\phi_j(z) = \sum_{k=0}^{j} \beta_{j,k}z^k, \tag{16}$$

then

$$\phi_j^*(z) = \sum_{k=0}^{j} \bar{\beta}_{j,k-j}z^k.$$

Thus, given the Szegő polynomial ϕ_n in power form, we can determine the coefficients of the associated auxiliary polynomial ϕ_n^* in power form and apply the recursion formula (4) "backwards" in order to determine the recursion coefficient γ_n and the coefficients of the polynomials ϕ_{n-1} and ϕ_{n-1}^* in power form. In this manner we can determine the recursion coefficients γ_j for decreasing values of the index j.

The Schur–Cohn algorithm, see, e.g., Henrici [17, Chapter 6], is an implementation of these computations. The algorithm requires $\mathcal{O}(n^2)$ arithmetic operations to determine the recursion coefficients $\{\gamma_j\}_{j=1}^n$ from the representation of ϕ_n in power form (16).

We remark that the Schur–Cohn algorithm is known for its use in determining whether a given polynomial, in power form, has all zeros inside the unit circle. In this context it is known as the Schur–Cohn test; see [17, Chapter 6]. All zeros being strictly inside the unit circle is equivalent with all recursion coefficients $\{\gamma_j\}_{j=1}^n$ being of magnitude strictly smaller than one. We will return to this property of the recursion coefficients in Section 3.

Perhaps the first application of the Schur–Cohn algorithm to the computation of zeros of polynomials was described by Lehmer [19], who covered the complex plane by disks and used the Schur–Cohn test to determine which disks contain zeros of the polynomial. Lehmer's method can be viewed as a generalization of the bisection method to the complex plane. It is discussed in [17, Chapter 6].

3. The zerofinders

We present two zerofinders for ϕ_n and assume that the recursion coefficients $\{\gamma_j\}_{j=1}^n$ as well as the auxiliary coefficients $\{\sigma_j\}_{j=1}^n$ are available.

3.1. An eigenvalue method

Eliminating the auxiliary polynomials ϕ_j^* in the recursion formula (4) yields an expression for ϕ_{j+1} in terms of Szegő polynomials of lower degree. Writing the expressions for the first $n+1$ Szegő polynomials in matrix form yields

$$[\phi_0(z), \phi_1(z), \ldots, \phi_{n-1}(z)]H_n = z[\phi_0(z), \phi_1(z), \ldots, \phi_{n-1}(z)] - [0, \ldots, 0, \phi_n(z)], \tag{17}$$

where

$$H_n = \begin{bmatrix} -\gamma_1 & -\sigma_1\gamma_2 & -\sigma_1\sigma_2\gamma_3 & \cdots & -\sigma_1\cdots\sigma_{n-1}\gamma_n \\ \sigma_1 & -\bar{\gamma}_1\gamma_2 & -\bar{\gamma}_1\sigma_2\gamma_3 & \cdots & -\bar{\gamma}_1\sigma_2\cdots\sigma_{n-1}\gamma_n \\ & \sigma_2 & -\bar{\gamma}_2\gamma_3 & \cdots & -\bar{\gamma}_2\sigma_3\cdots\sigma_{n-1}\gamma_n \\ & & \ddots & & \vdots \\ & & \sigma_{n-2} & -\bar{\gamma}_{n-2}\gamma_{n-1} & -\bar{\gamma}_{n-2}\sigma_{n-1}\gamma_n \\ 0 & & & \sigma_{n-1} & -\bar{\gamma}_{n-1}\gamma_n \end{bmatrix} \in \mathbb{C}^{n\times n} \tag{18}$$

is the Szegő–Hessenberg matrix associated with the Szegő polynomials $\{\phi_j\}_{j=0}^n$; see [11]. Eq. (17) shows that the eigenvalues of the upper Hessenberg matrix H_n are the zeros of ϕ_n. Thus, we can compute the zeros of ϕ_n by determining the eigenvalues of H_n.

Let ζ_j, $1 \leqslant j \leqslant n$, denote the zeros of ϕ_n. The scaling parameters η_1 and η_2 in (8) are chosen so that all zeros of ϕ_n are inside the unit circle. However, for some polynomials ψ_n, the scaling may be such that

$$\kappa_n := \max_{1\leqslant j\leqslant n} |\zeta_j| \ll 1.$$

We have noticed that we can determine the zeros of ψ_n with higher accuracy when the disk is rescaled to make κ_n close to one. Such a rescaling is easy to achieve by repeated application of the Schur–Cohn test as follows. Instead of scaling $\tilde{z}$ by the factor (13) in (14), we scale $\tilde{z}$ by $\tau := \sqrt{2}/(1+\sqrt{5})$ and apply the Schur–Cohn test to determine whether all zeros of the scaled polynomial (15) so obtained are inside the unit circle. If they are not, then we increase the scaling factor τ in (14) by a factor $\Delta\tau{:=}(2/(1+\sqrt{5}))^{1/10}$ and check whether the (re)scaled polynomial (15) obtained has all zeros inside the unit circle. The scaling factor τ is increased repeatedly by the factor $\Delta\tau$ until the polynomial (15) has all its zeros inside the unit circle. On the other hand, if the polynomial (15) associated with the scaling factor $\tau=\sqrt{2}/(1+\sqrt{5})$ has all zeros inside the unit circle, we repeatedly decrease τ by a factor $(\Delta\tau)^{-1}$ until a scaling factor τ has been determined, such that all zeros of the polynomial $\Phi_n^{(\tau)}$ are inside the unit disk, but at least one zero of $\Phi_n^{(\tau/\Delta\tau)}$ is not. Our choice of scaling factor τ in (14) assures that the monic polynomial (15) has all its zeros inside the unit circle and (at least) one zero close to the unit circle.

The scaling factors τ in (14) for the computed examples reported in Section 4 have been determined as described above. In our experience, the time spent rescaling the disk is negligible compared to the time required to compute the eigenvalues of H_n, because each rescaling only requires $\mathcal{O}(n^2)$ arithmetic operations.

After determining the scaling factor τ as described above and computing the recursion coefficients $\{\gamma_j\}_{j=1}^n$ via the Schur–Cohn test, we form the Szegő–Hessenberg matrix (18), balance it, and compute its eigenvalues using the QR algorithm.

3.2. A continuation method

Similarly as in the method described in Section 3.1, we first determine the recursion coefficients of the Szegő polynomials $\{\phi_j\}_{j=0}^n$ such that Eq. (7) holds, as described above. We then apply the continuation method for computing zeros of Szegő polynomials developed in [2]. In this method the Szegő–Hessenberg matrix (18) is considered a function of the last recursion parameter γ_n. Denote this parameter by $t \in \mathbb{C}$ and the associated Szegő–Hessenberg matrix by $H_n(t)$. Thus, we write the matrix (18) as $H_n(\gamma_n)$. When $|t|=1$, the Szegő–Hessenberg matrix $H_n(t)$ is unitary. Assume that $\gamma_n \neq 0$. Then $H_n(\gamma_n/|\gamma_n|)$ is the closest unitary matrix to $H_n(\gamma_n)$; see [2] for details. The continuation method for computing zeros of Szegő polynomials consists of the following steps:

(i) Compute the eigenvalues of the unitary upper Hessenberg matrix $H_n(\gamma_n/|\gamma_n|)$.

(ii) Apply a continuation method for tracking the path of each eigenvalue of the matrix $H_n(t)$ as t is moved from $\gamma_n/|\gamma_n|$ to γ_n.

Several algorithms that require only $\mathcal{O}(n^2)$ arithmetic operations for the computations of Step (i) are available; see, e.g. [4–6,12–15]. If the coefficients α_j in (1) are real, then the method discussed in [3] can also be applied. These methods compute the eigenvalues of $H_n(\gamma_n/|\gamma_n|)$ without explicitly forming the matrix elements. In the numerical experiments reported in Section 4, we used the implementation [4,5] of the divide-and-conquer method described in [14,15]. The computations required for this method can readily be implemented on a parallel computer. This may be of importance in the application of the zerofinder in real-time filter design; see, e.g., Parks and Burrus [24] and references therein for more on this application of polynomial zerofinders.

We have found that for many polynomials ψ_n, the continuation method determines the zeros with higher accuracy than the method discussed in Section 3.1. The continuation method determines the zeros of the Szegő polynomial ϕ_n close to the unit circle particularly rapidly. However, our present implementation of the continuation method may fail to determine all zeros for some polynomials ψ_n when the pathfollowing is complicated by (numerous) bifurcation points. These cases are easy to identify; see [2] for a discussion and remedies.

Table 1
Ten polynomials of degree $n = 15$ with zeros in D_1

Differences:	CB	SHB	CM	CBS
	6.67E−05	4.89E−06	4.57E−06	6.82E−06
	1.66E−03	7.57E−05	5.49E−05	2.11E−04
	1.20E−01	3.06E−03	—	1.83E−02
	8.41E−04	2.45E−05	3.91E−05	6.22E−04
	9.66E−04	5.88E−05	5.82E−05	1.51E−04
	2.75E−05	5.20E−07	1.79E−07	2.40E−06
	3.34E−05	5.75E−06	2.71E−07	2.05E−05
	1.67E−05	2.85E−06	2.25E−06	5.52E−05
	2.72E−04	6.60E−06	7.48E−07	3.77E−05
	7.60E−05	1.16E−06	7.40E−07	3.30E−06
Averages:	1.24E−02	3.24E−04	1.79E−05	1.94E−03

Residuals:	CB	SHB	CM	CBS	ψ_n
	3.85E−06	9.06E−07	4.89E−07	1.10E−06	6.94E−07
	3.31E−07	9.68E−08	2.05E−08	1.15E−07	1.47E−08
	3.16E−05	1.30E−05	—	2.41E−05	5.80E−07
	2.48E−06	9.15E−07	3.16E−07	1.47E−06	6.62E−08
	5.24E−06	6.74E−07	1.18E−06	1.50E−06	3.58E−07
	8.64E−08	2.13E−08	1.47E−08	4.12E−08	2.18E−09
	1.87E−06	6.88E−07	5.66E−07	8.80E−07	2.92E−08
	2.93E−06	2.48E−06	2.76E−07	2.71E−06	4.34E−08
	2.14E−07	7.87E−08	6.35E−08	3.23E−08	6.32E−09
	1.07E−06	4.44E−07	9.72E−08	9.11E−07	2.11E−08
Averages:	4.97E−06	1.93E−06	3.36E−07	3.28E−06	1.82E−07

	Differences				Residuals			
CB	0				0			
SHB	10	2			10	2		
CM	9	8	8		9	8	7	
CBS	9	0	1	0	10	1	2	1

We remark that other continuation methods also are available, such as the method proposed by Li and Zeng [20] for computing the eigenvalues of a general Hessenberg matrix. This method does not use the structure of the Hessenberg matrices (18), i.e., the fact that the last recursion coefficient γ_n is a natural continuation parameter. However, it may be possible to apply some techniques developed in [20] to improve the performance of the continuation method of this paper; see [2] for a discussion and references to other continuation methods.

Table 2
Ten polynomials of degree $n = 15$ with zeros in D_2

Differences:	CB	SHB	CM	CBS
	3.06E−04	4.98E−05	4.25E−05	9.49E−05
	1.47E−04	4.30E−05	4.22E−05	8.30E−05
	9.99E−06	2.40E−06	2.67E−07	8.38E−06
	5.97E−06	3.04E−05	2.09E−06	1.59E−05
	2.72E−04	3.44E−05	3.05E−05	3.37E−05
	1.10E−06	1.77E−06	5.06E−07	1.53E−06
	4.77E−04	1.56E−05	1.78E−05	5.08E−05
	7.30E−04	1.02E−03	8.53E−04	8.76E−04
	7.92E−06	2.82E−06	1.53E−06	6.90E−06
	4.88E−04	8.80E−05	1.33E−05	1.55E−04
Averages:	2.44E−04	1.29E−04	1.00E−04	1.33E−04

Residuals:	CB	SHB	CM	CBS	ψ_n
	5.85E−02	6.82E−03	1.11E−02	6.22E−03	1.06E−03
	1.50E−01	3.04E−02	1.96E−02	4.09E−02	1.95E−02
	8.29E−02	1.90E−02	4.67E−03	1.26E−02	2.27E−03
	4.56E−01	4.67E−01	2.94E−02	2.13E−01	7.14E−03
	1.98E−03	2.93E−03	8.92E−04	8.11E−04	1.00E−03
	1.77E−02	1.92E−02	7.89E−03	7.24E−03	1.30E−03
	7.42E−01	3.88E−01	4.22E−01	5.35E−01	1.84E−02
	9.64E−03	7.14E−03	3.95E−03	1.23E−02	4.08E−03
	7.70E−02	2.89E−02	2.19E−02	1.21E−01	4.53E−03
	3.02E−02	3.05E−03	6.00E−04	4.11E−04	2.43E−03
Averages:	1.62E−01	9.73E−02	5.22E−02	9.50E−02	6.17E−03

	Differences				Residuals			
CB	1				0			
SHB	7	1			7	1		
CM	9	9	8		10	8	5	
CBS	7	4	0	0	8	6	4	4

4. Computed examples

We present the results of several computed examples which illustrate the performance of the zerofinders discussed in Section 3. The computer programs used were all written in FORTRAN 77, and the numerical experiments were carried out on a SUN SparcStation 5 in single-precision arithmetic, i.e., with approximately 7 significant decimal digits of accuracy, except where explicitly stated otherwise. The eigenvalues of the companion and Szegő–Hessenberg matrices were computed by single-precision subroutines from EISPACK [25].

Table 3
Comparison of methods for 100 polynomials of each degree n with zeros in D_1

	Average differences					
n	CB	SHB	CM	CBS		N
10	1.20E−03	1.78E−05	1.75E−05	2.08E−05		99
15	3.12E−03	1.34E−04	1.14E−02	3.22E−04		94
20	3.48E−02	6.59E−03	7.27E−03	9.89E−03		86
30	1.75E−01	5.28E−02	1.67E−03	1.04E−01		47
40	3.95E−01	1.60E−01	1.07E−03	3.20E−01		12
	Average residuals					
n	CB	SHB	CM	CBS	ψ_n	N
10	1.70E−06	1.09E−06	3.58E−07	9.85E−07	1.20E−07	99
15	6.99E−06	2.89E−06	7.89E−07	3.39E−06	2.95E−07	94
20	4.06E−03	9.95E−06	1.90E−06	2.35E−05	8.01E−07	86
30	3.08E+01	7.52E−03	1.36E−05	1.03E−03	4.83E−06	47
40	1.05E+04	3.92E−02	6.31E−06	4.30E−02	4.64E−05	12

In our experiments, we input a set of n real or complex conjugate zeros of the polynomial ψ_n, see (1), and compute the coefficients α_j of the power-basis representation by a recursion formula. These computations are carried out in double-precision arithmetic, i.e., with about 15 significant digits, in order to avoid loss of accuracy. After their computation, the α_j are stored as single-precision real numbers. We now seek to determine the zeros of ψ_n, given the coefficients α_j, with one of several methods:

CB: The QR algorithm applied to the companion matrix (2) of ψ_n after balancing, using the EISPACK routines `balanc` and `hqr`.

CBS: The QR algorithm applied to the companion matrix of the monic Szegő polynomial Φ_n, after balancing, using the EISPACK routines `balanc` and `hqr`.

SHB: The QR algorithm applied to the Szegő–Hessenberg matrix after balancing, using the EISPACK routines `balanc` and `hqr`.

CM: The continuation method for real Szegő–Hessenberg matrices, described in [2].

We compare the following computed quantities:

Residuals: The maximum modulus of the values of the initial monic polynomial ψ_n in power form (1) at the computed roots.

Differences: The computed zeros are put into correspondence with the initial zeros, which were used to generate ψ_n as described above, and the maximum difference after this pairing is computed. Note that this is not exactly the error in the computed zeros; the error is the maximum difference of the computed roots and the exact roots of the monic polynomial ψ_n. However, since the coefficients of ψ_n were computed from the given zeros in floating-point arithmetic, the exact zeros of the ψ_n need not be close to the input zeros. Nevertheless, the computed differences provide a way to compare the various methods.

In the tables we also display in the column labeled ψ_n the residuals computed at the input zeros; i.e., at the zeros that were used to compute the power-basis coefficients of ψ_n. This provides some

Table 4
Comparative counts for 100 polynomials for each degree n with zeros in D_1

	Differences				Residuals			
$n=10$								
CB	0				2			
SHB	97	12			79	13		
CM	99	83	71		95	81	74	
CBS	90	28	20	17	78	39	15	11
$n=15$								
CB	0				2			
SHB	100	17			77	8		
CM	97	80	78		95	85	81	
CBS	88	16	9	5	71	48	12	9
$n=20$								
CB	0				4			
SHB	97	15			79	16		
CM	88	78	77		88	79	73	
CBS	85	19	17	8	70	32	18	7
$n=30$								
CB	1				8			
SHB	97	42			84	29		
CM	61	55	53		58	50	46	
CBS	73	6	40	4	73	29	48	17
$n=40$								
CB	4				8			
SHB	94	74			88	55		
CM	26	17	16		19	13	12	
CBS	61	8	78	6	76	29	83	25

indication of how ill-conditioned the roots of ψ_n and the computation of its power-basis coefficients are, as well as an indication of the significance of the differences and the other computed residuals that are displayed.

The polynomials ψ_n in all computed examples except those for Tables 7–8 have real or complex conjugate zeros uniformly distributed in a disk

$$D_R := \{z : \ |z| \leqslant R\} \subset \mathbb{C}. \tag{19}$$

In particular, the coefficients α_j in the representation (1) are real. We generate zeros of ψ_n in D_R as follows. Two random numbers are determined according to a uniform distribution on the interval $[-R, R]$ and used as the real and imaginary parts of a candidate zero z. If $z \in D_R$ and $\mathrm{Im}(z) > 1 \times 10^{-6}$, then both z and $\bar{z}$ are accepted as zeros of ψ_n. If $z \in D_R$ and $\mathrm{Im}(z) \leqslant 1 \times 10^{-6}$ then $\mathrm{Re}(z)$ is accepted as a real zero of ψ_n. The purpose of the condition on the imaginary part of z is to avoid that ψ_n has very close zeros. We generate candidate points until n zeros of ψ_n have been determined. When n is odd, then at least one of the zeros of ψ_n is in the real interval $[-R, R]$.

Table 5
Comparison of methods for 100 polynomials of degree $n = 20$ for each radius R

	Average differences					
R	CB	SHB	CM	CBS		N
0.2	2.96E−01	1.54E−03	1.25E−03	2.01E−03		86
0.7	9.76E−02	4.45E−03	5.59E−04	6.54E−03		84
1.0	3.48E−02	6.59E−03	7.27E−03	9.89E−03		86
1.5	2.20E−02	1.16E−02	8.55E−04	1.84E−02		83
3.0	6.81E−02	2.32E−02	1.71E−03	3.38E−02		83
	Average residuals					
R	CB	SHB	CM	CBS	ψ_n	N
0.2	6.79E−10	1.32E−19	2.27E−20	1.15E−19	7.20E−21	86
0.7	4.69E−07	7.86E−09	1.73E−09	7.43E−09	5.79E−10	84
1.0	4.06E−03	9.95E−06	1.90E−06	2.35E−05	8.01E−07	86
1.5	6.91E+01	4.03E−02	5.13E−03	7.10E−02	3.15E−03	83
3.0	1.17E+08	4.06E+04	5.50E+03	6.79E+04	3.30E+03	83

Table 1 shows results for 10 polynomials ψ_{15} generated in this manner with zeros in the disk D_1. We display the maximum modulus of the residuals and the maximum difference of the computed zeros with the input zeros for the methods CB, SHB, CM, and CBS. The results for CM for one of these 10 polynomials are marked with a "—" to indicate that the continuation method did not yield all n zeros. The averages for CM ignore the entries marked by —. In Table 1 the standard companion matrix approach (CB) consistently yields the least accuracy as measured both by the residuals and by the differences with the input zeros.

The integer arrays at the bottom of Table 1 display the relative performance of the algorithms. The (j,k) entry for $j > k$ is the number of times the jth algorithm gave smaller maximal differences or residuals than the kth algorithm, and the (j,j) entry indicates the number of times the jth algorithm gave the smallest maximal differences or residuals among the four methods compared. For example, the arrays for Table 1 show that CM produces the smallest residuals for 7 of the 10 polynomials generated. This count includes the polynomial for which CM failed to determine all zeros. The maximum residual for CM was smaller than for CB, SHB, and CBS for 9, 8, and 8 polynomials, respectively. CB produced larger residuals than any of the other three methods for all polynomials, except for the polynomial for which CM failed to determine all zeros.

Table 2 gives the results for 10 polynomials of degree 15 with uniformly distributed real and complex conjugate zeros in the disk D_2. In this experiment, CM successfully determined all zeros of all polynomials.

Tables 3 and 4 show summary data for 100 polynomials of each of several degrees n with uniformly distributed real and complex conjugate zeros in the disk D_1. We display in Tables 3 the average of the maximum differences and the average of the maximum residuals for the methods CB, SHB and CBS over all polynomials. For CM we compute these averages only over those polynomials for which the method successfully determined all zeros. The number of those polynomials of each degree n, out of 100, is denoted by N and is displayed in the last column of Table 3.

Table 6
Comparative counts for 100 polynomials of degree $n = 20$ for each radius R

	Differences				Residuals			
$R = 0.2$								
CB	0				0			
SHB	100	22			100	12		
CM	100	73	72		100	79	73	
CBS	100	23	15	6	100	48	20	15
$R = 0.7$								
CB	0				0			
SHB	100	19			100	10		
CM	91	81	76		90	80	75	
CBS	100	15	17	5	100	35	20	15
$R = 1.0$								
CB	0				4			
SHB	97	15			79	16		
CM	88	78	77		88	79	73	
CBS	85	19	17	8	70	32	18	7
$R = 1.5$								
CB	7				4			
SHB	58	18			77	21		
CM	87	76	72		85	72	67	
CBS	36	16	17	3	72	37	21	8
$R = 3.0$								
CB	4				0			
SHB	66	16			95	22		
CM	89	78	76		89	69	66	
CBS	44	17	17	4	90	37	22	12

In the experiments in Tables 5 and 6, we generated 100 polynomials of degree 20 with uniformly distributed real or complex conjugate zeros in disks (19) of radius R for several different values of R. The entries in the columns "Average differences" and "Average residuals" of Table 5 are computed as for Table 3. We display results obtained for disks with radii between 0.2 and 3.

Finally, Tables 7 and 8 illustrate the performance of the zerofinders for polynomials ψ_{20} with real zeros only. The zeros are uniformly distributed in the interval $[-1,1]$. Tables 7 and 8 are analogous to Tables 3 and 4. We see that CBS often gives significantly higher accuracy than CB, and SHB usually yields slightly higher accuracy than CBS. Our present implementation of CM is able to accurately determine all or most zeros for the polynomials in this experiment of fairly low degree, $n \leqslant 10$, only, due to numerous bifurcation points encountered during pathfollowing. The performance of CM might be improved by using a more sophisticated pathfollowing method; see [2] for a discussion.

Table 7
Comparison of methods for 100 polynomials of each degree n zeros in $[-1,1]$

	Average differences		
n	CB	SHB	CBS
10	8.73E−03	1.53E−03	3.16E−03
15	5.83E−02	1.43E−02	3.47E−02
20	2.07E−01	8.64E−02	1.67E−01
30	4.97E−01	2.93E−01	5.62E−01
40	7.18E−01	5.62E−01	7.94E−01

	Average residuals			
n	CB	SHB	CBS	ψ_n
10	7.90E−07	4.64E−07	4.23E−07	6.92E−08
15	1.59E−06	8.51E−07	1.48E−06	9.62E−08
20	1.03E−05	4.05E−06	9.74E−06	2.69E−07
30	3.07E−04	5.24E−05	8.11E−05	7.90E−07
40	3.70E+01	5.01E−02	6.71E−02	3.34E−06

Table 8
Comparative counts for 100 polynomials of each degree n with zeros in $[-1,1]$

	Differences			Residuals		
$n=10$						
CB	2			6		
SHB	96	31		71	23	
CBS	74	9	7	74	59	28
$n=15$						
CB	2			17		
SHB	98	63		75	50	
CBS	77	5	4	60	31	26
$n=20$						
CB	2			17		
SHB	98	95		77	68	
CBS	71	4	3	59	17	15
$n=30$						
CB	10			7		
SHB	89	86		93	85	
CBS	39	6	4	64	11	8
$n=40$						
CB	19			10		
SHB	78	67		88	80	
CBS	37	20	13	53	11	10

In addition to the examples reported above, we carried out numerous numerical experiments with the zerofinders applied to polynomials whose zeros were uniformly distributed in squares and wedges in the complex plane. The performance of the zerofinders for these problems is similar to the performance reported in the Tables 1–6, and we therefore omit the details. We noted that for some classes of problems CBS performed comparatively better than in the Tables 1–6, and gave about the same accuracy as SHB. In all examples considered, CB gave the poorest overall accuracy.

5. Conclusions

Numerous numerical experiments, some of which have been presented in Section 4, indicate that the polynomial zerofinders CBS, CM and SHB presented in this paper, in general, yield higher accuracy than computing eigenvalues of the associated balanced companion matrix, the CB method. When CM finds all zeros, this method typically yields the highest accuracy. Presently, we are using a fairly simple path-following scheme described in [2], and this implementation of CM may occasionally occasionally fail to find all zeros. Our numerical experiments suggest that CM with an improved pathfollowing scheme would be an attractive zerofinder. Alternatively, one can use CM as presently implemented and switch to a different zerofinding method when CM fails to determine all zeros. This approach has the advantage of allowing us to keep the pathfollowing scheme simple. The numerical examples of Section 4, as well as other examples not reported, indicate that the SHB method may be a good method to switch to. It is simple to implement and often gives higher accuracy than the CB and CBS methods.

References

[1] N.I. Akhiezer, The Classical Moment Problem, Oliver & Boyd, London, 1965.

[2] G.S. Ammar, D. Calvetti, L. Reichel, Continuation methods for the computation of zeros of Szegő polynomials, Linear Algebra Appl. 249 (1996) 125–155.

[3] G.S. Ammar, W.B. Gragg, L. Reichel, On the eigenproblem for orthogonal matrices, in: Proceedings of the 25th IEEE Conference on Decision and Control, Athens, 1986, Institute for Electrical and Electronic Engineers, New York, 1986, pp. 1963–1966.

[4] G.S. Ammar, L. Reichel, D.C. Sorensen, An implementation of a divide and conquer algorithm for the unitary eigenproblem, ACM Trans. Math. Software 18 (1992) 292–307.

[5] G.S. Ammar, L. Reichel, D.C. Sorensen, Algorithm 730, ACM Trans. Math. Software 20 (1994) 161.

[6] A. Bunse-Gerstner, L. Elsner, Schur parameter pencils for the solution of the unitary eigenproblem, Linear Algebra Appl. 154–156 (1991) 741–778.

[7] A. Edelman, H. Murakami, Polynomial roots from companion matrix eigenvalues, Math. Comp. 64 (1995) 763–776.

[8] W. Gautschi, On the condition of algebraic equations, Numer. Math. 21 (1973) 405–424.

[9] W. Gautschi, Question of numerical condition related to polynomials, in: G.H. Golub (Ed.), Studies in Numerical Analysis, The Mathematical Association of America, Washington, D.C., 1984, pp. 140–177.

[10] S. Goedecker, Remark on algorithms to find roots of polynomials, SIAM J. Sci. Comput. 15 (1994) 1059–1063.

[11] W.B. Gragg, Positive definite Toeplitz matrices, the Arnoldi process for isometric operators, and Gaussian quadrature on the unit circle, J. Comput. Appl. Math. 46 (1993) 183–198. (This is a slight revision of a paper originally published (in Russian) E.S. Nikolaev (Ed.), Numerical Methods in Linear Algebra, Moscow University Press, Moscow, 1982, pp. 16–32.

[12] W.B. Gragg, The QR algorithm for unitary Hessenberg matrices, J. Comput. Appl. Math. 16 (1986) 1–8.

[13] W.B. Gragg, Stabilization of the UHQR algorithm, in: Z. Chen, Y. Li, C. Micchelli, Y. Xu (Eds.), Advances in Computational Mathematics, Lecture Notes in Pure and Applied Mathematics, Vol. 202, Marcel Dekker, Hong Kong, 1999, pp. 139–154.

[14] W.B. Gragg, L. Reichel, A divide and conquer method for the unitary eigenproblem, in: M.T. Heath (Ed.), Hypercube Multiprocessors 1987, SIAM, Philadelphia, PA, 1987, pp. 639–647.

[15] W.B. Gragg, L. Reichel, A divide and conquer method for unitary and orthogonal eigenproblems, Numer. Math. 57 (1990) 695–718.

[16] U. Grenander, G. Szegő, Toeplitz Forms and Applications, Chelsea, New York, 1984.

[17] P. Henrici, Applied and Computational Complex Analysis, Vol. 1, Wiley, New York, 1974.

[18] M.A. Jenkins, J.F. Traub, A three-stage variable shift iteration for polynomial zeros and its relation to generalized Ritz iteration, Numer. Math. 14 (1970) 252–263.

[19] D.H. Lehmer, A machine method for solving polynomial equations, J. Assoc. Comput. Mach. 8 (1961) 151–162.

[20] T.-Y. Li, Z. Zeng, Homotopy-determinant algorithm for solving nonsymmetric eigenvalue problems, Math. Comp. 59 (1992) 483–502.

[21] J.M. McNamee, An updated supplementary bibliography on roots of polynomials, J. Comput. Appl. Math. 110 (1999) 305–306. (This reference discusses the bibliography. The actual bibliography is available on the world wide web at `http://www.elsevier.nl/locate/cam`.)

[22] C. Moler, Cleve's corner: roots – of polynomials, that is, The MathWorks Newsletter 5 (1) (1991) 8–9.

[23] A.M. Ostrowski, A method for automatic solution of algebraic equations, in: B. Dejon, P. Henrici (Eds.), Constructive Aspects of the Fundamental Theorem of Algebra, WileyInterscience, London, 1969, pp. 209–224.

[24] T.W. Parks, C.S. Burrus, Digital Filter Design, Wiley, New York, 1987.

[25] B. Smith, J. Boyle, Y. Ikebe, V. Klema, C. Moler, Matrix Eigensystem Routines: EISPACK Guide, 2nd Edition, Springer, Berlin, 1976.

[26] J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, 2nd Edition, Springer, New York, 1993.

ELSEVIER

Journal of Computational and Applied Mathematics 127 (2001) 17–65

JOURNAL OF
COMPUTATIONAL AND
APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Complex Jacobi matrices

Bernhard Beckermann

*Laboratoire d'Analyse Numérique et d'Optimisation, UFR IEEA – M3, UST Lille,
F-59655 Villeneuve d'Ascq Cedex, France*

Received 11 October 1999; received in revised form 22 December 1999

Abstract

Complex Jacobi matrices play an important role in the study of asymptotics and zero distribution of formal orthogonal polynomials (FOPs). The latter are essential tools in several fields of numerical analysis, for instance in the context of iterative methods for solving large systems of linear equations, or in the study of Padé approximation and Jacobi continued fractions. In this paper we present some known and some new results on FOPs in terms of spectral properties of the underlying (infinite) Jacobi matrix, with a special emphasis to unbounded recurrence coefficients. Here we recover several classical results for real Jacobi matrices. The inverse problem of characterizing properties of the Jacobi operator in terms of FOPs and other solutions of a given three-term recurrence is also investigated. This enables us to give results on the approximation of the resolvent by inverses of finite sections, with applications to the convergence of Padé approximants.

MSC: 39A70; 47B39; 41A21; 30B70

Keywords: Difference operator; Complex Jacobi matrix; Formal orthogonal polynomials; Resolvent convergence; Convergence of J-fractions; Padé approximation

1. Introduction

We denote by ℓ^2 the Hilbert space of complex square-summable sequences, with the usual scalar product $(u,v)=\sum \overline{u_j}v_j$, and by $(e_n)_{n\geqslant 0}$ its usual orthonormal basis. Furthermore, for a linear operator T in ℓ^2, we denote by $\mathscr{D}(T)$, $\mathscr{R}(T)$, $\mathscr{N}(T)$, and $\sigma(T)$, its domain of definition, its range, its kernel, and its spectrum, respectively.

E-mail address: bbecker@ano.univ-lille1.fr (B. Beckermann).

0377-0427/01/\$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0377-0427(00)00492-1

Given complex numbers a_n, b_n, $n \geqslant 0$, with $a_n \neq 0$ for all n, we associate the infinite tridiagonal *complex Jacobi matrix*

$$\mathscr{A} = \begin{pmatrix} b_0 & a_0 & 0 & \cdots & \cdots \\ a_0 & b_1 & a_1 & 0 & \\ 0 & a_1 & b_2 & a_2 & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{1.1}$$

In the *symmetric case* $b_n, a_n \in \mathbb{R}$ for all n one recovers the classical Jacobi matrix. Denoting by $\mathscr{C}_0 \subset \ell^2$ the linear space of finite linear combinations of the basis elements $e_0, e_1, \ldots$, we may identify via the usual matrix product a complex Jacobi matrix $\mathscr{A}$ with an operator acting on $\mathscr{C}_0$. Its closure A is called the corresponding second-order difference operator or Jacobi operator (see Section 2.1 for a more detailed discussion).

Second-order (or higher-order) difference operators have received much attention in the last years, partly motivated by applications to nonlinear discrete dynamical systems (see [7,20,21,29,38] and the references therein). Also, Jacobi matrices are known to be a very useful tool in the study of (formal) orthogonal polynomials ((F)OPs), which again have applications in numerous fields of numerical analysis. To give an example, (formal) orthogonal polynomials have been used very successfully in numerical linear algebra for describing both algorithmic aspects and convergence behavior of iterative methods like conjugate gradients, GMRES, Lanczos, QMR, and many others. Another example is given by the study of convergence of continued fractions and Padé approximants. Indeed, also the study of higher-order difference operators is of interest in all these applications; let us mention the Bogoyavlenskii discrete dynamical system [8], Ruhe's block version of the Lanczos method in numerical linear algebra, or the problem of Hermite–Padé and matrix Padé approximation (for the latter see, e.g., the surveys [5,6]). In the present paper we will restrict ourselves to the less involved case of three diagonals.

To start with, a linear functional c acting on the space of polynomials with complex coefficients is called regular if and only if $\det(c(x^{j+k}))_{j,k=0,\ldots,n} \neq 0$ for all $n \geqslant 0$. Given a regular c (with $c(1) = 1$), there exists a sequence $(q_n)_{n\geqslant 0}$ of FOPs, i.e., q_n is a polynomial of degree n (unique up to a sign), and $c(q_j \cdot q_k)$ vanishes if $j \neq k$ and is equal to 1 otherwise. These polynomials are known to verify a three-term recurrence of the form

$$a_n q_{n+1}(z) = (z - b_n) q_n(z) - a_{n-1} q_{n-1}(z), \quad n \geqslant 0, \quad q_0(z) = 1, \quad q_{-1}(z) = 0,$$

where $a_n = c(z q_{n+1} q_n) \in \mathbb{C} \backslash \{0\}$, and $b_n = c(z q_n q_n) \in \mathbb{C}$. Here a_n, b_n are known to be real if and only if c is positive, i.e., $c(P) > 0$ for each nontrivial polynomial P taking nonnegative values on the real axis, or, equivalently, $\det(c(x^{j+k}))_{j,k=0,\ldots,n} > 0$ for all $n \geqslant 0$. Conversely, the Shohat–Favard Theorem says that any $(q_n(z))_{n\geqslant 0}$ verifying a three-term recurrence relation of the above form is a sequence of formal orthogonal polynomials with respect to some regular linear functional c. As shown in Remark 2.3 below, this linear functional can be given in terms of the Jacobi operator A defined above, namely $c(P) = (e_0, P(A) e_0)$ for each polynomial P. In the real case one also knows that there is orthogonality with respect to some positive Borel measure μ supported on the real axis, i.e., $c(P) = \int P(x)\, \mathrm{d}\mu(x)$.

Notice that q_n is (up to normalization) the characteristic polynomial of the finite submatrix $\mathscr{A}_n$ of order n of $\mathscr{A}$. Also, the second-order difference equation

$$z \cdot y_n = a_n y_{n+1} + b_n y_n + a_{n-1} y_{n-1}, \quad n \geqslant 0 \tag{1.2}$$

(a_{-1}:=1) together with the initialization $y_{-1} = 0$ may be formally rewritten as the spectral equation $(z\mathscr{I} - \mathscr{A}) \cdot y = 0$. This gives somehow the idea that spectral properties of the Jacobi operator should be determined by the spectral or asymptotic properties of FOPs, and vice versa. Indeed, in the real case the link is very much known (see, for instance, [38] or [37]): if A is self-adjoint, then there is just one measure of orthogonality (obtained by the spectral theorem applied to A), with support being equal to the spectrum $\sigma(A)$ of A. Also, the zeros of OPs lie all in the convex hull of $\sigma(A)$, are interlacing, and every point in $\sigma(A)$ attracts zeros. Furthermore, in case of bounded A one may describe the asymptotic behavior of OPs on and outside $\sigma(A)$. Surprisingly, for formal orthogonal polynomials these questions have been investigated only recently in terms of the operator A, probably owing to the fact that here things may change quite a bit (see, for instance, Example 3.2 below).

To our knowledge, the first detailed account on (a class of) complex Jacobi matrices was given by Wall in his treatise [59] on continued fractions. He dealt with the problem of convergence of Jacobi continued fractions (J-fractions)

$$\frac{1 \rfloor}{\lfloor z - b_0} + \frac{-a_0^2 \rfloor}{\lfloor z - b_1} + \frac{-a_1^2 \rfloor}{\lfloor z - b_2} + \frac{-a_2^2 \rfloor}{\lfloor z - b_3} + \cdots \tag{1.3}$$

having at infinity the (possibly formal) expansion $f(z) = \sum_j c(x^j) z^{-j-1} = \sum_j (e_0, A^j e_0) z^{-j-1}$. Their nth convergent may be rewritten as $p_n(z)/q_n(z)$, where $(p_n(z))_{n \geqslant -1}, (q_n(z))_{n \geqslant -1}$ are particular solutions of (1.2) with initializations

$$q_0(z) = 1, \quad q_{-1}(z) = 0, \quad p_0(z) = 0, \quad p_{-1}(z) = -1, \tag{1.4}$$

i.e., q_n are the FOPs mentioned above. Also, p_n/q_n is just the nth Padé approximant (at infinity) of the perfect power series f. Notice that, in case of a bounded operator A, f is the Laurent expansion at infinity of the so-called *Weyl function* [21]

$$\phi(z) := (e_0, (zI - A)^{-1} e_0), \quad z \in \Omega(A),$$

where here and in the sequel $\Omega(A) = \mathbb{C} \backslash \sigma(A)$ denotes the resolvent set, i.e., the set of all $z \in \mathbb{C}$ such that $\mathscr{N}(zI - A) = \{0\}$ and $\mathscr{R}(zI - A) = \ell^2$ (and thus the resolvent $(zI - A)^{-1}$ is bounded).

The aim of the present paper is threefold: we try to give a somehow complete account on connections between FOPs, complex J-fractions and complex Jacobi matrices presented in the last five years. In this context we report about recent work by Aptekarev, Kaliaguine, Van Assche, Barrios, López Lagomasino, Martínez-Finkelshtein, Torrano, Castro Smirnova, Simon, Magnus, Stahl, Baratchart, Ambroladze, Almendral Vázquez, and the present author. Special attention in our study is given to unbounded complex Jacobi matrices, where similar uniqueness problems occur as for the classical moment problem. Secondly, we present some new results concerning ratio-normality of FOPs and compact perturbations of complex Jacobi matrices. In addition, we show that many recent results on convergence of complex J-fractions in terms of Jacobi operators [13–16,18,20] are, in fact, results on the approximation of the resolvent of complex Jacobi operators. Finally, we mention several open problems in this field of research. A main (at least partially) open question is however omitted: do these results have a counterpart for higher-order difference operators?

The paper is organized as follows: Some preliminaries and spectral properties of Jacobi operators in terms of solutions of (1.2) are presented in Section 2. In Section 2.1 we report about the problem of associating a unique operator to (1.1), and introduce the notion of *proper* Jacobi matrices. Next to some preliminary observations, we recall in Section 2.2 Wall's definition of determinate Jacobi matrices and relate it to proper ones. Also, some sufficient conditions for determinacy are discussed [59,13,22]. Known characterizations [7,19] of elements z of the resolvent set in terms of the asymptotic behavior of solutions of (1.2) are described in Section 2.3. Here we also show in Theorem 2.11 that for indeterminate complex Jacobi operators we have a similar behavior as for nonself-adjoint Jacobi operators (where the corresponding moment problem does not have a unique solution). In Section 2.4 we highlight the significance of the Weyl function and of functions of the second kind. Their representation as Cauchy transforms is investigated in Section 2.5, where we also study the case of totally positive moment sequences leading to nonreal compact Jacobi matrices.

In Section 3 we describe results on the asymptotic behavior of FOPs in the resolvent set. nth-root asymptotics for bounded complex Jacobi matrices obtained in [7,18,20] are presented in Section 3.1. In Section 3.2 we deal with the problem of localizing zeros of FOPs, thereby generalizing some results from [18]. We show that, roughly, under some additional hypotheses, there are only "few" zeros in compact subsets of the resolvent set. An important tool is the study of ratios of two successive monic FOPs. An inverse open problem concerning zero-free regions is presented in Section 3.3. In Section 3.4 we characterize compact perturbations of complex Jacobi matrices in terms of the ratios mentioned above. Strong asymptotics for trace class perturbations are the subject of Section 3.5. In the final Section 4, we investigate the problem of convergence of Padé approximants (or J-fractions) and more generally of (weak, strong or norm) resolvent convergence. A version of the Kantorovich Theorem for complex Jacobi matrices is given in Section 4.1, together with a discussion of its assumptions. We describe in Section 4.2 consequences for the approximation of the Weyl function, and finally illustrate in Section 4.3 some of our findings by discussing (asymptotically) periodic complex Jacobi matrices.

2. The Jacobi operator

2.1. Infinite matrices and operators

Given an infinite matrix $\mathscr{A} = (a_{j,k})_{j,k \geqslant 0}$ of complex numbers, can we define correctly a (closed and perhaps densely defined) operator via matrix calculus by identifying elements of ℓ^2 with infinite column vectors? Of course, owing to Hilbert and his collaborators, an answer to this question is known, see, e.g., [2]. In this section we briefly summarize the most important facts. Here we will restrict ourselves to matrices $\mathscr{A}$ whose rows and columns are elements of ℓ^2, an assumption which is obviously true for banded matrices such as our complex Jacobi matrices.

By assumption, the formal product $\mathscr{A} \cdot y$ is defined for any $y \in \ell^2$. Thus, as a natural candidate of an operator associated with $\mathscr{A}$, we could consider the so-called *maximal operator* (see [30, Example III.2.3]) $[\mathscr{A}]_{\max}$ with

$$\mathscr{D}([\mathscr{A}]_{\max}) = \{y \in \ell^2\colon \mathscr{A} \cdot y \in \ell^2\} \tag{2.1}$$

and $[\mathscr{A}]_{\max} y := \mathscr{A} \cdot y \in \ell^2$. However, there are other operators having interesting properties which may be associated with $\mathscr{A}$. For instance, since the columns of $\mathscr{A}$ are elements of ℓ^2, we may define a linear operator on $\mathscr{C}_0$ (also denoted by $\mathscr{A}$) by setting

$$\mathscr{A} e_k = (a_{j,k})_{j \geqslant 0}, \quad k = 0, 1, 2, \ldots .$$

Notice that $[\mathscr{A}]_{\max}$ is an extension[1] of $\mathscr{A}$.

A minimum requirement in the spectral theory of linear operators is that the operator in question be closed [30, Section III.5.2]. In general, our operator $\mathscr{A}$ is not closed, but it is closable [30, Section III.5.3], i.e., for any sequence $(y^{(n)})_{n \geqslant 0} \subset \mathscr{D}(\mathscr{A})$ with $y^{(n)} \to 0$ and $\mathscr{A} y^{(n)} \to v$ we have $v = 0$. To see this, notice that

$$(e_j, v) = \lim_{n \to \infty} \sum_{k=0}^{\infty} a_{j,k} y_k^{(n)} = \lim_{n \to \infty} (v_j, y^{(n)}) = 0,$$

where $v_j = (\overline{a_{j,k}})_{k \geqslant 0} \in \ell^2$ by assumption on the rows of $\mathscr{A}$. Thus, we may consider the *closure* $[\mathscr{A}]_{\min}$ of $\mathscr{A}$, i.e., the smallest closed extension of $\mathscr{A}$. Notice that

$$\mathscr{D}([\mathscr{A}]_{\min}) = \{ y \in \ell^2 \colon \exists (y^{(n)})_{n \geqslant 0} \subset \mathscr{C}_0 \text{ converging to } y, \text{ and } (\mathscr{A} y^{(n)})_{n \geqslant 0} \subset \ell^2 \text{ converging (to } [\mathscr{A}]_{\min} y) \}. \tag{2.2}$$

We have the following links between the operators $[\mathscr{A}]_{\min}$, $[\mathscr{A}]_{\max}$, and their adjoints.

Lemma 2.1. *Let the infinite matrix $\mathscr{A}^{\mathrm{H}}$ be obtained from $\mathscr{A}$ by transposing and by taking complex conjugates of the elements. Then*

$$([\mathscr{A}]_{\min})^* = [\mathscr{A}^{\mathrm{H}}]_{\max}, \quad ([\mathscr{A}]_{\max})^* = [\mathscr{A}^{\mathrm{H}}]_{\min}.$$

In particular, the maximal operator $[\mathscr{A}]_{\max}$ is a closed extension of $[\mathscr{A}]_{\min}$.

Proof. In order to show the first equality, for short we write $A = [\mathscr{A}]_{\min}$. By definition of the adjoint [30, Section III.5.5]), $\mathscr{D}(A^*)$ equals the set of all $y \in \ell^2$ such that there exists a $y^* \in \ell^2$ with

$$(y, Ax) = (y^*, x) \quad \text{for all } x \in \mathscr{D}(A)$$

and $y^* = A^* y$. According to the characterization of $\mathscr{D}(A)$ given above and the continuity of the scalar product, it is sufficient to require that $(y, Ax) = (y^*, x)$ holds for all $x \in \mathscr{C}_0$, or, equivalently,

$$y_j^* = (e_j, y^*) = (Ae_j, y) \quad \text{for all } j \geqslant 0.$$

Since (Ae_j, y) coincides with the jth component of the formal product $\mathscr{A}^{\mathrm{H}} y$, we obtain $A^* = [\mathscr{A}^{\mathrm{H}}]_{\max}$ by definition (2.1) of $\mathscr{D}([\mathscr{A}^{\mathrm{H}}]_{\max})$.

The second identity of Lemma 2.1 follows from the fact that $A^{**} = A$ by [30, Theorem III.5.29]. Finally, we obtain the last claim by observing that $[\mathscr{A}]_{\max}$ is an extension of $\mathscr{A}$, and $[\mathscr{A}]_{\max}$ is closed (since an adjoint of a densely defined operator is closed [30, Theorem III.5.29]). □

[1] Given two operators T, S in ℓ^2, we say that S is an extension of T (and write $T \subset S$) if $\mathscr{D}(T) \subset \mathscr{D}(S)$, and $Ty = Sy$ for all $y \in \mathscr{D}(T)$.

Definition 2.2. The infinite matrix $\mathscr{A}$ with rows and columns in ℓ^2 is called *proper* if the operators $[\mathscr{A}]_{\max}$ and $[\mathscr{A}]_{\min}$ coincide.

Notice that any operator B defined by matrix product (i.e., $By = \mathscr{A} \cdot y$ for $y \in \mathscr{D}(B)$) necessarily is a restriction of $[\mathscr{A}]_{\max}$ by (2.1). From Lemma 2.1 we obtain the equivalent description $\mathscr{A}^{\mathrm{H}} \subset B^*$. If in addition $\mathscr{C}_0 \subset \mathscr{D}(B)$, then $\mathscr{A} \subset B \subset [\mathscr{A}]_{\max}$. We may conclude that any closed operator B defined by matrix product and $\mathscr{C}_0 \subset \mathscr{D}(B)$ satisfies $[\mathscr{A}]_{\min} \subset B \subset [\mathscr{A}]_{\max}$, and such an operator B is unique if and only if $\mathscr{A}$ is proper.[2]

Let us have a look at the special case of Hermitian matrices $\mathscr{A}$, i.e., $\mathscr{A} = \mathscr{A}^{\mathrm{H}}$. Here Lemma 2.1 tells us that $A := [\mathscr{A}]_{\min}$ has the adjoint $A^* = [\mathscr{A}]_{\max}$ (see also [52, p. 90] or [30, Example V.3.13]), which is an extension of A. Hence A is symmetric, and we obtain the following equivalencies: A is self-adjoint (i.e., $A = A^*$) if and only if $\mathscr{A}$ is proper if and only if there exists a unique symmetric closed extension of $\mathscr{A}$ (cf. with [30, Problem III.5.25]).

Remark 2.3. The notion of proper Jacobi matrices may be motivated by considering the following problem: given a regular functional c acting on the space of polynomials, can we describe its action by a densely defined closed operator B, namely $c(P) = (g, P(B)f)$ and, more generally,

$$c(P \cdot Q) = (Q(B)^* g, P(B)f) \quad \text{for all polynomials } P, Q \tag{2.3}$$

with suitable $f, g \in \ell^2$?

We first show that any closed operator B with $\mathscr{A}_{\min} \subset B \subset \mathscr{A}_{\max}$ satisfies (2.3) with $f = g = e_0$. Obviously, it is sufficient to show the relation

$$e_j = q_j(B)e_0 = q_j(B)^* e_0, \quad j \geqslant 0.$$

Indeed, $e_0 = q_0(B)e_0$ by (1.4), and by induction, using (1.2), we obtain

$$a_j q_{j+1}(B)e_0 = Bq_j(B)e_0 - b_j q_j(B)e_0 - a_{j-1}q_{j-1}(B)e_0 = Be_j - b_j e_j - a_{j-1}e_{j-1} = a_j e_{j+1},$$

the last equality following from $\mathscr{A} \subset B$. Since $a_j \neq 0$, the relation $e_{j+1} = q_{j+1}(B)e_0$ follows. In a similar way the other identity is shown using the relation $\mathscr{A}^{\mathrm{H}} \subset B^*$.

We now show that these are essentially all the operators satisfying (2.3). Notice first that B is only properly characterized by (2.3) if f is a cyclic element of B (i.e., $f \in \mathscr{D}(B^k)$ for all k, and $\operatorname{span}\{B^j f \colon j \geqslant 0\}$ is dense in ℓ^2), and g is a cyclic element of B^*. In this case, using the orthogonality relations of the FOPs q_j, we may conclude that $(f_n)_{n \geqslant 0}$ and $(g_n)_{n \geqslant 0}$, defined by $f_n = q_n(B)f$ and $g_n = q_n(B)^* g$, is a complete normalized biorthogonal system. The expansion coefficients of Bf_k (and $B^* g_j$, resp.) with respect to the system $(f_n)_{n \geqslant 0}$ (and $(g_n)_{n \geqslant 0}$, resp.) are given by

$$(g_j, Bf_k) = \overline{(f_k, B^* g_j)} = c(zq_jq_k) = (e_j, \mathscr{A}e_k) = \begin{cases} a_{\min(j,k)} & \text{if } j = k+1 \text{ or } k = j+1, \\ b_j & \text{if } j = k, \\ 0 & \text{else.} \end{cases}$$

[2] Some authors consider other extensions of $\mathscr{A}$ which are not defined by matrix product, or which are defined by Hilbert space extensions, see, for instance, [45, Section 6] or [42].

In other words, up to the representation of ℓ^2 and its dual with the help of these different bases, we have $\mathscr{A} \subset B$ and $\mathscr{A}^{\mathrm{H}} \subset B^*$, and thus $\mathscr{A}_{\min} \subset B \subset \mathscr{A}_{\max}$. We may conclude in particular that an operator as in (2.3) is unique (up to basis transformations) if and only if $\mathscr{A}$ is proper. □

For an infinite matrix $\mathscr{A}$, define the quantity

$$||\mathscr{A}|| := \sup_{u,v \in \mathscr{C}_0} \left| \frac{(u, \mathscr{A} v)}{(u,v)} \right| = \sup_{n \geqslant 0} ||\mathscr{A}_n||,$$

where on the right-hand side $\mathscr{A}_n$ denotes the principal submatrix of order n of $\mathscr{A}$. Clearly, $||\mathscr{A}||$ is the operator norm of $\mathscr{A}$, and $[\mathscr{A}]_{\min}$ is bounded (with $||[\mathscr{A}]_{\min}|| = ||\mathscr{A}||$) if and only if $||\mathscr{A}|| < \infty$. One easily checks that in this case $\mathscr{D}([\mathscr{A}]_{\min}) = \ell^2$. Conversely, if $\mathscr{D}([\mathscr{A}]_{\min}) = \ell^2$, then $\mathscr{A}$ is bounded by the closed graph theorem [30, Theorem III.5.20]. Finally, we recall the well-known estimate [30, Example III.2.3]

$$||\mathscr{A}||^2 \leqslant \left[\sup_j \sum_{k=0}^{\infty} |a_{j,k}| \right] \left[\sup_k \sum_{j=0}^{\infty} |a_{j,k}| \right]. \tag{2.4}$$

Using this formula, one easily verifies the well-known fact that banded matrices $\mathscr{A}$ are bounded if and only if their entries are uniformly bounded.

We may conclude that a bounded matrix $\mathscr{A}$ is proper, and thus we may associate a unique closed operator $A = [\mathscr{A}]_{\min}$ whose action is described via matrix calculus. However, these properties do not remain necessarily true for unbounded matrices.

2.2. Spectral properties of Jacobi operators

In what follows, $\mathscr{A}$ will be the complex Jacobi matrix of (1.1) with entries $a_n, b_n \in \mathbb{C}$, $a_n \neq 0$, We refer to its closure $A = [\mathscr{A}]_{\min}$ as the corresponding *difference operator* or *Jacobi operator*, and denote by $A^\# = [\mathscr{A}]_{\max}$ the maximal closed extension of A defined by matrix product. Since $\mathscr{A}^{\mathrm{H}}$ is obtained from $\mathscr{A}$ by taking the complex conjugate of each entry, we may conclude from Lemma 2.1 that $A^\# = \Pi A^* \Pi$, where Π denotes the complex conjugation operator defined by $\Pi(y_j)_{j \geqslant 0} = (\overline{y_j})_{j \geqslant 0}$.

The aim of this section is to summarize some basic properties of the operators A, $A^\#$ in terms of solutions $q(z) := (q_n(z))_{n \geqslant 0}$ and $p(z) := (p_n(z))_{n \geqslant 0}$ of recurrence (1.2), (1.4).

We will make use of the projection operators Π_j defined by

$$\Pi_j(y_0, y_1, y_2, \ldots) = (y_0, y_1, \ldots, y_{j-1}, 0, 0, \ldots) \in \mathscr{C}_0, \quad j \geqslant 1.$$

Clearly, $\Pi_j y \to y$ for $j \to \infty$ for any $y \in \ell^2$. Also, one easily checks that, for any sequence $y = (y_n)_{n \geqslant 0}$, one has $\Pi_j y \in \mathscr{D}(A)$, with

$$A(\Pi_j y) = \Pi_j(\mathscr{A} \cdot y) + a_{j-1} \cdot (\underbrace{0, \ldots, 0}_{j-1}, -y_j, y_{j-1}, 0, 0, \ldots.). \tag{2.5}$$

Using (2.4), one easily verifies that A is bounded if and only if the entries of $\mathscr{A}$ are uniformly bounded; more precisely,

$$\sup_{n \geqslant 0} \sqrt{|a_{n-1}|^2 + |b_n|^2 + |a_n|^2} \leqslant ||A|| \leqslant \sup_{n \geqslant 0} (|a_{n-1}| + |b_n| + |a_n|) \tag{2.6}$$

(where we tacitly put $a_{-1} = 0$). Also, notice that $\mathscr{A}$ is Hermitian if and only if it is real.

Some further elementary observations are summarized in

Lemma 2.4. (a) *For $z \in \mathbb{C}$ there holds*

$$0 \leqslant \dim \mathscr{N}(zI - A) \leqslant \dim \mathscr{N}(zI - A^{\#}) \leqslant 1$$

with $\mathscr{N}(zI - A^{\#}) = \ell^2 \cap \{\lambda q(z)\colon \lambda \in \mathbb{C}\}$.

(b) *For $n \geqslant 0$ and $z \in \mathbb{C}$ we have $e_n - q_n(z)e_0 \in \mathscr{R}(zI - A)$.*

(c) *For $z \in \mathbb{C}$ there holds $\{y \in \mathscr{D}(A^{\#})\colon (zI - A^{\#})y = e_0\} = \ell^2 \cap \{\gamma q(z) - p(z)\colon \gamma \in \mathbb{C}\}$.*

(d) *For all $z \in \mathbb{C}$ we have $\mathscr{D}(zI - A^{\#}) = \mathscr{D}(A^{\#}) = \Pi \mathscr{D}(A^*)$, $\mathscr{N}(zI - A^{\#}) = \Pi \mathscr{N}((zI - A)^*)$, and $\mathscr{R}(zI - A^{\#}) = \Pi \mathscr{R}((zI - A)^*)$.*

Proof. (a) Since $A \subset A^{\#}$, we only have to show the last assertion. By (2.1), $y = (y_n)_{n \geqslant 0} \in \mathscr{N}(zI - A^{\#})$ if and only if $y \in \ell^2$, and we have $(z\mathscr{I} - \mathscr{A}) \cdot y = 0$. The latter identity may be rewritten as

$$-a_n y_{n+1} + (z - b_n) y_n - a_{n-1} y_{n-1} = 0, \quad n \geqslant 0, \ y_{-1} = 0.$$

Comparing with (1.2), (1.4), we see that $(z\mathscr{I} - \mathscr{A}) \cdot y = 0$ if and only if $y = y_0 \cdot q(z)$, leading to the above description of $\mathscr{N}(zI - A^{\#})$.

(b) Notice that (1.2), (1.4) may be rewritten as

$$(z\mathscr{I} - \mathscr{A}) \cdot q(z) = 0, \quad (z\mathscr{I} - \mathscr{A}) \cdot p(z) = -e_0.$$

Combining this with (2.5), we obtain

$$(zI - A)\Pi_{n+1} p(z) = -e_0 + a_n \cdot (\underbrace{0, \ldots, 0}_{n}, p_{n+1}(z), -p_n(z), 0, 0, \ldots), \tag{2.7}$$

$$(zI - A)\Pi_{n+1} q(z) = a_n \cdot (\underbrace{0, \ldots, 0}_{n}, q_{n+1}(z), -q_n(z), 0, 0, \ldots). \tag{2.8}$$

Also, one easily verifies by induction, using (1.2), that

$$a_n \cdot (q_n(z) \cdot p_{n+1}(z) - q_{n+1}(z) \cdot p_n(z)) = 1, \quad n \geqslant -1, \ z \in \mathbb{C}. \tag{2.9}$$

Thus, we have found an element of $\mathscr{C}_0 \subset \mathscr{D}(A)$ satisfying

$$(zI - A)\Pi_{n+1}[q_n(z)p(z) - p_n(z)q(z)]$$

$$= -q_n(z)e_0 + a_n \cdot (\underbrace{0, \ldots, 0}_{n}, q_n(z)p_{n+1}(z) - p_n(z)q_{n+1}(z), 0, 0, \ldots) = e_n - q_n(z)e_0.$$

(c) Since $(zI - \mathscr{A}) \cdot (\gamma \cdot q(z) - p(z)) = e_0$ for all γ, a proof for this assertion follows the same lines as the one of part (a). We omit the details.

(d) This part is an immediate consequence of the fact that

$$(zI - A)^* = \bar{z}I - A^* = \bar{z}I - \Pi A^{\#} \Pi = \Pi(zI - A^{\#})\Pi. \qquad \square$$

For a closed densely defined linear operator T in ℓ^2, the integer $\dim \mathscr{N}(T)$ is usually referred to as the *nullity* of T, and $\dim \mathscr{N}(zI - T)$ coincides with the geometric multiplicity of the "eigenvalue" z (if larger than zero). One also defines the *deficiency* of T as the codimension in ℓ^2 of $\mathscr{R}(T)$. Provided that $\mathscr{R}(T)$ is closed, it follows from [30, Theorem IV.5.13, Lemma III.1.40] that the deficiency of T

coincides with $\dim \mathscr{N}(T^*)$, and that also $\mathscr{R}(T^*)$ is closed. Taking into account Lemma 2.4(a), (d), we may conclude that both deficiency and nullity are bounded by one for our operators $zI - A$ and $zI - A^\#$ provided one of them has closed range. Consequently, we obtain for the essential spectrum [30, Chapter IV.5.6]

$$\sigma_{\text{ess}}(A) = \sigma_{\text{ess}}(A^\#) = \{z \in \mathbb{C}\colon \mathscr{R}(zI - A) \text{ is not closed}\}. \tag{2.10}$$

Recall that this (closed) part of the spectrum of A (or $A^\#$) remains invariant under compact perturbations [30, Theorem IV.5.35]. Let us relate Definition 2.2 to the notion of determinacy as introduced by Wall.

Definition 2.5 (See Wall [59, Definition 22.1]). The complex Jacobi matrix $\mathscr{A}$ is called *determinate* if at least one of the sequences $p(0)$ or $q(0)$ is not an element of ℓ^2.

According to [59, Theorem 22.1], $\mathscr{A}$ is indeterminate if $p(z)$ and $q(z)$ are elements of ℓ^2 for one $z \in \mathbb{C}$, and in this case they are elements of ℓ^2 for all $z \in \mathbb{C}$. It is also known (see [1, pp. 138–141] or [38, p. 76]) that a real Jacobi matrix is proper (i.e., self-adjoint) if and only if it is determinate. In the general case we have the following

Theorem 2.6 (Cf. with Beckermann [19, Proposition 3.2]). (a) *A determinate complex Jacobi matrix $\mathscr{A}$ with $\sigma_{\text{ess}}(A) \neq \mathbb{C}$ is proper.*

(b) *A proper complex Jacobi matrix $\mathscr{A}$ is determinate.*

(c) *If $\Omega(A) \neq \emptyset$ then $\mathscr{A}$ is proper.*

Proof. Part (a) has been established in [19, Proposition 3.2], the proof is mainly based on Lemma 2.4(d) and the fact that in the case $z \notin \sigma_{\text{ess}}(A)$ the set $\mathscr{R}(zI - A)$ (and $\mathscr{R}((zI - A)^*)$, resp.) coincides with the orthogonal complement of $\mathscr{N}((zI - A)^*)$ (and of $\mathscr{N}(zI - A)$, resp.).

In order to show part (b), suppose that $\mathscr{A}$ is not determinate. Then $\dim \mathscr{N}(zI - A^\#) = 1$ and $\mathscr{C}_0 \subset \mathscr{R}(zI - A^\#)$ for all $z \in \mathbb{C}$ according to Lemma 2.4(a)–(c), and thus $\mathscr{C}_0 \subset \mathscr{R}((zI - A)^*)$ by Lemma 2.4(d). Taking into account that $\mathscr{N}(zI - A)$ is just the orthogonal complement $\mathscr{R}((zI - A)^*)^\perp$ of $\mathscr{R}((zI - A)^*)$, we may conclude that $\dim \mathscr{N}(zI - A) = 0$, showing that $A \neq A^\#$.

Part (c) was also mentioned in [19, Proposition 3.2]: Let $z \in \Omega(A)$. Then $\mathscr{R}(zI - A) = \ell^2$ by definition of the resolvent set. Hence $\mathscr{N}((zI - A)^*) = \{0\}$, implying that $\mathscr{N}(zI - A^\#) = \{0\}$ by Lemma 2.4(d). From Lemma 2.4(a) we may conclude that $(q_n(z))_{n \geqslant 0} \notin \ell^2$, and hence $\mathscr{A}$ is determinate. Finally, since $\mathbb{C} \backslash \sigma_{\text{ess}}(A) \supset \Omega(A)$ is nonempty, it follows from Theorem 2.6(a) that $\mathscr{A}$ is proper. □

In order to complete the statement of Theorem 2.6, we should mention the following characterization in terms of operators of indeterminate complex Jacobi matrices which will be shown in Theorem 2.11 below: if $\mathscr{A}$ is indeterminate, then $\sigma_{\text{ess}}(A)$ is empty and $\sigma(A) = \mathbb{C}$; more precisely, for all $z \in \mathbb{C}$, the kernel of $zI - A$ is empty, $\mathscr{R}(zI - A)$ is closed and has codimension 1, the kernel of $zI - A^\#$ has dimension 1, and $\mathscr{R}(zI - A^\#) = \ell^2$.

The *numerical range* (or *field of values*) [30, Section V.3.2] of a linear operator T in ℓ^2 is defined by

$$\Theta(T) = \{(y, Ty)\colon y \in \mathscr{D}(T), ||y|| = 1\}.$$

By a theorem of Hausdorff, $\Theta(T)$ and its closure $\Gamma(T)$ are convex. Also, $\sigma_{\rm ess}(A) \subset \Gamma(A)$ by [30, Theorem V.3.2]. Hence, for complex Jacobi matrices with $\sigma_{\rm ess}(A) \neq \mathbb{C}$ or $\Gamma(A) \neq \mathbb{C}$, the notions of determinacy and properness are equivalent. This includes the case of real Jacobi matrices since here $\Gamma(A) \subset \mathbb{R}$. Notice that $\Gamma(A) \neq \mathbb{C}$ implies that $\Gamma(A)$ is included in some half-plane of $\mathbb{C}$. The case of the lower half-plane $\{\mathrm{Im}(z) \leqslant 0\}$ was considered by Wall [59, Definition 16.1], who called the corresponding J-fraction *positive definite* and gave characterizations of such complex Jacobi matrices in terms of chain sequences [59, Theorem 16.2]. In this context we should also mention that $\Gamma(A)$ is compact if and only if A is bounded; indeed, one knows [27, Eq. (1.6)] that $\sup\{|z|\colon z \in \Gamma(A)\} \in [||A||/2, ||A||]$.

It is not known whether there exists a determinate complex Jacobi matrix which is not proper. Since many of the results presented below are valid either for proper or for indeterminate Jacobi matrices, a clarification of this problem seems to be desirable.

Results related to Theorem 2.6 have been discussed by several authors: Barrios et al. [13, Lemma 3] showed that a complex Jacobi matrix $\mathscr{A} = \mathscr{A}' + \mathscr{A}''$ with $\mathscr{A}'$ self-adjoint and $\mathscr{A}''$ bounded is determinate. More generally, Castro Smirnova [22, Theorem 2] proved that a bounded perturbation of a real Jacobi matrix $\mathscr{A}$ is determinate[3] if and only if $\mathscr{A}$ is determinate. It is an interesting open problem to characterize determinacy or properness in terms of the real and the imaginary part of a Jacobi matrix.

Let us here have a look at a sufficient condition which will be used later.

Example 2.7. It is known [59, Theorem 25.1] that $\mathscr{A}$ is determinate provided that

$$\sum_{n=0}^{\infty} \frac{1}{|a_n|} = +\infty.$$

We claim that then $\mathscr{A}$ is also proper. To see this, let $y \in \mathscr{D}(A^{\#})$. Choose integers $n_0 < n_1 < \cdots$ with

$$\alpha_\ell := \sum_{j=n_\ell}^{n_{\ell+1}-1} \frac{1}{|a_{j-1}|} \geqslant 1, \quad \ell \geqslant 0,$$

and put

$$y^{(\ell)} = \frac{1}{\alpha_\ell} \sum_{j=n_\ell}^{n_{\ell+1}-1} \frac{1}{|a_{j-1}|} \Pi_j y \in \mathscr{C}_0.$$

Since $n_\ell \to \infty$ and $\Pi_j y \to y$, one obtains $y^{(\ell)} \to y$. Furthermore, according to (2.5),

$$\begin{aligned} ||Ay^{(\ell)} - A^{\#}y|| &\leqslant \frac{1}{\alpha_\ell} \sum_{j=n_\ell}^{n_{\ell+1}-1} \frac{||\Pi_j A^{\#} y - A^{\#} y||}{|a_{j-1}|} + \frac{1}{\alpha_\ell} \left\| \sum_{j=n_\ell}^{n_{\ell+1}-1} \frac{A\Pi_j y - \Pi_j A^{\#} y}{|a_{j-1}|} \right\| \\ &\leqslant ||\Pi_{n_\ell} A^{\#} y - A^{\#} y|| + \frac{1}{\alpha_\ell} \left\| \sum_{j=n_\ell}^{n_{\ell+1}-1} (\underbrace{0,\ldots,0}_{j-1}, |y_j|, |y_{j-1}|, 0, 0, \ldots) \right\| \end{aligned}$$

[3] Of course, by (2.1), (2.2), a proper Jacobi matrix $\mathscr{A}$ remains proper after adding some bounded perturbation.

$$\leqslant ||\Pi_{n_\ell} A^{\#} y - A^{\#} y|| + \frac{2}{\alpha_\ell} ||\Pi_{n_\ell} y - y||$$

and the right-hand side clearly tends to 0 for $\ell \to \infty$. Thus $y \in \mathscr{D}(A)$ by (2.2).

Combining Example 2.7 with the techniques of [18, Example 5.2], one may construct for any closed set $E \subset \mathbb{C}$ a proper difference operator A satisfying $\sigma_{\text{ess}}(A) = E$. In particular [18, Example 5.2], notice that (in contrast to the real case) the resolvent set may consist of several connected components.

To conclude this part, recall from [19] a further characterization of the essential spectrum in terms of *associated* Jacobi matrices. We denote by $\mathscr{A}^{(k)}$ the "shifted" (complex) Jacobi matrix obtained by replacing (a_j, b_j) in $\mathscr{A}$ by (a_{j+k}, b_{j+k}), $j \geqslant 0$. As in [30, Chapter IV.6.1] one shows that the operators corresponding to $\mathscr{A} = \mathscr{A}^{(0)}$, and to $\mathscr{A}^{(k)}$, respectively, have the same essential spectrum for any $k \geqslant 0$. In our case we have the following stronger assertion.

Theorem 2.8 (See Beckermann [19, Proposition 3.4]). *Suppose that $\mathscr{A}$ is determinate. Then $\sigma_{\text{ess}}(A) = \sigma(A^{(k)}) \cap \sigma(A^{(k+1)})$ for any $k \geqslant 0$. More precisely, for any $z \in \mathbb{C} \backslash \sigma_{\text{ess}}(A)$ there exists a nontrivial ℓ^2-solution $(s_n(z))_{n \geqslant -1}$ of* (1.2), *with*

$$\Omega(A^{(k)}) = \{z \in \mathbb{C} \backslash \sigma_{\text{ess}}(A) \colon s_{k-1}(z) \neq 0\}, \quad k \geqslant 0. \tag{2.11}$$

If the entries of the difference of two (complex) Jacobi matrices tend to zero along diagonals, then the difference of the corresponding difference operators is known to be *compact* [2]. We can now give a different characterization of the essential spectrum, namely

$$\sigma_{\text{ess}}(A) = \bigcap \{\sigma(A') \colon A' \text{ is a difference operator and } A - A' \text{ is compact}\}. \tag{2.12}$$

Here the inclusion $\subset$ is true even in a more general setting [30, Theorem IV.5.35]. In order to see the other inclusion, notice that for the particular solution of Theorem 2.8 there necessarily holds $|s_{-1}(z)| + |s_0(z)| \neq 0$. Therefore, by Theorem 2.8, the essential spectrum is already obtained by taking the intersection with respect to all difference operators found by varying the entry a_0 of $\mathscr{A}$, i.e., by rank 1 perturbations.

2.3. Characterization of the spectrum

In this subsection we are concerned with the problem of characterizing the spectrum of a difference operator in terms of solutions of the recurrence relation (1.2). This connection can be exploited in several ways. On the one hand, one sometimes knows the asymptotic behavior of solutions of (1.2) (as for instance in the case of (asymptotically) periodic recurrence coefficients, cf. [15,16,20,26,34]), and it is possible to determine the shape of the spectrum. On the other hand, we will see in Section 3 that we obtain nth-root asymptotics for FOPs and functions of the second kind on the resolvent set.

A description of the resolvent operator (or more precisely of a (formal) "right reciprocal") in terms of the solutions $p(z)$, $q(z)$ of recurrence (1.2) has been given already by Wall [59, Sections 59–61]. Starting with a paper of Aptekarev et al. [7], the problem of characterizing the spectrum has received much attention in the last years, see [15,16,19,20] for Jacobi matrices and the survey papers [5,6] and the references therein for higher-order difference operators. A typical example of characterizing the spectrum in terms of only one solution of (1.2) is the following.

Theorem 2.9 (See Beckermann [19, Theorem 2.3]). *Let A be bounded. Then $z \in \Omega(A)$ if and only if*

$$\sup_{n\geqslant 0} \frac{\sum_{j=0}^{n} |q_j(z)|^2}{|a_n|^2[|q_n(z)|^2 + |q_{n+1}(z)|^2]} < \infty. \tag{2.13}$$

Indeed, using (2.8), we obtain for $z \in \Omega(A)$ and $n \geqslant 0$

$$\frac{1}{||zI-A||^2} \leqslant \frac{\sum_{j=0}^{n} |q_j(z)|^2}{|a_n|^2[|q_n(z)|^2 + |q_{n+1}(z)|^2]} = \frac{||\Pi_{n+1}q(z)||^2}{||(zI-A)\Pi_{n+1}q(z)||^2} \leqslant ||(zI-A)^{-1}||^2, \tag{2.14}$$

showing that (2.13) holds. The other implication is more involved; here one applies the characterization of Theorem 2.12 below. Notice that we may reformulate Theorem 2.9 as follows: we have $z \in \sigma(A)$ if and only if the sequence $(\Pi_n q(z)/||\Pi_n q(z)||)_{n\geqslant 0}$ contains a subsequence of approximate eigenvectors (i.e., a sequence of elements of $\mathscr{D}(A)$ of norm one so that their images under $zI - A$ tend to zero).

In view of (2.13), (2.14), we can give another formulation of Theorem 2.9: we have $z \in \Omega(A)$ if and only if the sequence of numerators, and denominators in (2.13), respectively, have the same asymptotic behavior. It becomes clear from the following considerations (and can also be checked directly) that then both sequences will grow exponentially. It seems that, even for the classical case of real bounded Jacobi matrices, this result has only been found recently [19]. As mentioned before, here the spectrum of A coincides with the support of the measure of orthogonality μ of $(q_n)_{n\geqslant 0}$.

Some further consequences of relation (2.13) concerning the distribution of zeros of FOPs will be discussed in Section 3.

In order to describe other characterizations of the spectrum, we will fix $z \in \mathbb{C}$, and denote by $\mathscr{R}(\gamma)$, $\gamma \in \mathbb{C}$, the infinite matrix with elements

$$\mathscr{R}(\gamma)_{j,k} = \begin{cases} q_j(z)\cdot\{\gamma\cdot q_k(z) - p_k(z)\} & \text{if } 0\leqslant j\leqslant k, \\ \{q_j(z)\cdot\gamma - p_j(z)\}\cdot q_k(z) & \text{if } 0\leqslant k\leqslant j, \end{cases}$$

$j,k = 0,1,2,\ldots$. These matrices are just the (formal) "right reciprocals" mentioned by Wall [59, Theorem 60.2]. In the next statement we characterize the resolvent set of possibly unbounded difference operators in terms of two solutions of (1.2). A special case of this assertion may be found in [59, Theorem 61.2].

Theorem 2.10. *We have $z \in \Omega(A)$ if and only if $\mathscr{A}$ is proper, and there exists a $\gamma \in \mathbb{C}$ such that $\mathscr{R}(\gamma)$ is bounded. In this case, γ is unique, and the resolvent is given by $(zI-A)^{-1} = [\mathscr{R}(\gamma)]_{\min}$, in particular $\gamma = (e_0, (zI-A)^{-1}e_0)$.*

Proof. Let $z \in \Omega(A)$, and denote by $\mathscr{R} = (\mathscr{R}_{j,k})_{j,k=0,1,\ldots}$ the (bounded) infinite matrix corresponding to the resolvent operator $(zI-A)^{-1}$. It follows from Theorem 2.6(c) that $\mathscr{A}$ is proper. Thus the first implication follows by showing that $\mathscr{R} = \mathscr{R}(\gamma)$ for $\gamma = (e_0,(zI-A)^{-1}e_0)$. Since $\mathscr{R}((zI-A)^{-1}) =$

$\mathscr{D}(zI - A)$, we obtain

$$(zI - A)[\mathscr{R} \cdot e_0] = (zI - A)[(zI - A)^{-1}e_0] = e_0.$$

From Lemma 2.4(c) we get the form of the first column of $\mathscr{R}$, namely $\mathscr{R}_{j,0} = \gamma q_j(z) - p_j(z) = \mathscr{R}(\gamma)_{j,0}$ for some $\gamma \in \mathbb{C}$. Here the identity $\gamma = (e_0, (zI - A)^{-1}e_0)$ is obtained by comparing the values for the index $j = 0$. The form of the other columns of $\mathscr{R}$ is obtained from Lemma 2.4(b) and its proof. Indeed, we have for $j \geqslant 1$, $k \geqslant 0$

$$\mathscr{R}_{j,k} - q_k(z)\mathscr{R}_{j,0} = (e_j, (zI - A)^{-1}[e_k - q_k(z)e_0]) = (e_j, \Pi_{k+1}[q_k(z)p(z) - p_k(z)q(z)])$$

and thus $\mathscr{R}_{j,k} = \mathscr{R}(\gamma)_{j,k}$.

Conversely, suppose that $\mathscr{R}(\gamma)$ is bounded, and denote by R its closure. By some elementary calculations using (1.2) and (2.9) one verifies that

$$\mathscr{R}(\gamma) \cdot (ze_j - \mathscr{A}e_j) = e_j, \quad j \geqslant 0$$

and thus $R(zI - A)y = y$ for all $y \in \mathscr{C}_0$ by linearity. Recalling (2.2), we may conclude that

$$\inf_{y \in \mathscr{D}(A)} \frac{||(zI - A)y||}{||y||} = \inf_{y \in \mathscr{C}_0} \frac{||(zI - A)y||}{||y||} = \inf_{y \in \mathscr{C}_0} \frac{||(zI - A)y||}{||R(zI - A)y||} \geqslant \frac{1}{||R||} > 0.$$

Consequently, $\mathscr{N}(zI - A) = \{0\}$, and from [30, Theorem IV.5.2] if follows that $\mathscr{R}(zI - A)$ is closed. In order to establish our claim $z \in \Omega(A)$, it remains to show that $\mathscr{R}(zI - A)$ is dense in ℓ^2. Since $\mathscr{R}(\gamma)$ is bounded, its first column $y(\gamma)$ is an element of ℓ^2. Using Lemma 2.4(c), we may conclude that $e_0 \in \mathscr{R}(zI - A^{\#})$, and thus $e_0 \in \mathscr{R}(zI - A)$ since $\mathscr{A}$ is proper. Combining this with Lemma 2.4(b), we find that $\mathscr{C}_0 \subset \mathscr{R}(zI - A)$, and hence $\mathscr{R}(zI - A) = \ell^2$.

For establishing the second sentence of Theorem 2.10, we still need to show that the γ of the preceding part of the proof necessarily coincides with $(e_0, (zI - A)^{-1}e_0)$. By construction of $y(\gamma)$ we have $(zI - A^{\#})y(\gamma) = (zI - A)y(\gamma) = e_0$, and thus $\gamma = (e_0, y(\gamma)) = (e_0, (zI - A)^{-1}e_0)$. □

For the sake of completeness, let us also describe the case of operators A which are not proper. Here we have either the trivial case $\sigma_{\text{ess}}(A) = \mathbb{C}$, or otherwise $\mathscr{A}$ is indeterminate by Theorem 2.6(a). In the latter case, we find exactly the same phenomena as for real Jacobi matrices (see, e.g., [1,38] or [45, Theorem 2.6]).

Theorem 2.11. *Suppose that $\mathscr{A}$ is indeterminate. Then the following assertions hold*:

(a) $\sigma_{\text{ess}}(A) = \emptyset$ *and* $\sigma(A) = \sigma(A^{\#}) = \mathbb{C}$.

(b) $A^{\#}$ *is a two-dimensional extension of A. Furthermore, all other operators $A_{[\eta]}$ with $A \subset A_{[\eta]} \subset A^{\#}$ are one-dimensional extensions of A; they may be parametrized by $\eta \in \mathbb{C} \cup \{\infty\}$ via*

$$\mathscr{D}(A_{[\eta]}) = \left\{ y + \lambda \frac{\eta q(0) - p(0)}{1 + |\eta|} \colon y \in \mathscr{D}(A), \lambda \in \mathbb{C} \right\}.$$

(c) *We have $\sigma_{\text{ess}}(A_{[\eta]}) = \emptyset$, and $A^*_{[\eta]} = \Pi A_{[\eta]} \Pi$ for all η. Furthermore, there exist entire functions $a_1, a_2, a_3, a_4 : \mathbb{C} \to \mathbb{C}$ with $a_1 a_4 - a_2 a_3 = 1$ such that*

$$\sigma(A_{[\eta]}) = \{z \in \mathbb{C} \colon \phi_{[\eta]}(z) = \infty\}, \quad \textit{where } \phi_{[\eta]}(z) := \frac{a_1(z) - a_2(z)\eta}{a_3(z) - a_4(z)\eta}.$$

Finally, the resolvent of $A_{[\eta]}$ at $z \in \Omega(A_{[\eta]})$ is given by the closure of $\mathscr{R}(\phi_{[\eta]}(z))$, which is a compact operator of Schmidt class.

Proof. For $z \in \mathbb{C}$, consider the infinite matrix $\mathscr{S}(z)$ with entries

$$\mathscr{S}(z)_{j,k} = \begin{cases} q_k(z)p_j(z) - q_j(z)p_k(z) & \text{if } 0 \leqslant j \leqslant k, \\ 0 & \text{if } 0 \leqslant k \leqslant j. \end{cases}$$

Since $\mathscr{A}$ is indeterminate, we get $\sum_{j,k} |\mathscr{S}(z)_{j,k}|^2 < \infty$. In particular, the closure $S(z)$ of $\mathscr{S}(z)$ is bounded, and more precisely a compact operator of Schmidt class [30, Section V.2.4]. By some elementary calculations using (1.2) and (2.9) one verifies that $\mathscr{S}(z) \cdot (ze_j - \mathscr{A}e_j) = e_j$ for $j \geqslant 0$. As in the last part of the proof of Theorem 2.10 it follows that $S(z)$ is a left-inverse of $zI - A$, and it follows that $\mathscr{R}(zI - A)$ is closed and $\mathscr{N}(zI - A) = \{0\}$. Since by assumption $\mathscr{N}(zI - A^{\#}) = \mathrm{span}(q(z))$, we obtain

$$\mathscr{R}(zI - A^{\#}) = \mathscr{N}(zI - A)^{\perp} = \ell^2, \quad \mathscr{R}(zI - A) = \Pi \mathscr{N}(zI - A^{\#})^{\perp} = \mathrm{span}(\Pi q(z))^{\perp}, \quad z \in \mathbb{C}.$$

Using (2.10), we may conclude that part (a) holds.

For a proof of (b), let $y \in \mathscr{D}(A^{\#})$. Since $A^{\#} p(0) = -e_0$ by Lemma 2.4(c), we have

$$A^{\#}(y + (\Pi q(0), A^{\#} y) \cdot p(0)) \in \mathrm{span}(\Pi q(0))^{\perp} = \mathscr{R}(A).$$

Consequently, there exists a $y' \in \mathscr{D}(A)$ with $0 = A^{\#}(y + (\Pi q(0), A^{\#} y) \cdot p(0)) - Ay' = A^{\#}(y + (\Pi q(0), A^{\#} y) \cdot p(0) - y')$, showing that $y + (\Pi q(0), A^{\#} y) \cdot p(0) - y'$ is a multiple of $q(0)$. Hence $A^{\#}$ is a two-dimensional extension of A. Since any other extension either has a nontrivial kernel ($\eta = \infty$) or otherwise the image ℓ^2 ($\eta \neq \infty$), the second part of the assertion follows.

It remains to show part (c). Following [59, Section 23], we define the entire functions

$$a_1(z) = z \sum_{j=0}^{\infty} p_j(0) p_j(z), \quad a_2(z) = 1 + z \sum_{j=0}^{\infty} q_j(0) p_j(z),$$

$$a_3(z) = -1 + z \sum_{j=0}^{\infty} p_j(0) q_j(z), \quad a_4(z) = z \sum_{j=0}^{\infty} q_j(0) q_j(z).$$

It is shown in [59, Theorem 23.1] that indeed $a_1(z)a_4(z) - a_2(z)a_3(z) = 1$ for all $z \in \mathbb{C}$. Let $z \in \mathbb{C}$. We claim that, for a suitable unique $\gamma \in \mathbb{C} \cup \{\infty\}$ (depending on η, z),

$$\Pi \frac{\gamma q(z) - p(z)}{1 + |\gamma|} \in \mathscr{D}((zI - A_{[\eta]})^*), \quad \text{with } (zI - A_{[\eta]})^* \Pi \frac{\gamma q(z) - p(z)}{1 + |\gamma|} = \frac{e_0}{1 + |\gamma|}. \tag{2.15}$$

Indeed, for any $y \in \mathscr{D}(A)$ and $\lambda \in \mathbb{C}$

$$\begin{aligned} &\left(\frac{e_0}{1 + |\gamma|}, y + \lambda \frac{\eta q(0) - p(0)}{1 + |\eta|} \right) - \left(\Pi \frac{\gamma q(z) - p(z)}{1 + |\gamma|}, (zI - A_{[\eta]}) \left(y + \lambda \frac{\eta q(0) - p(0)}{1 + |\eta|} \right) \right) \\ &= \frac{(e_0, y) + \lambda \eta / (1 + |\eta|)}{1 + |\gamma|} - \left((zI - A)^* \Pi \frac{\gamma q(z) - p(z)}{1 + |\gamma|}, y \right) \\ &\quad - \left(\Pi \frac{\gamma q(z) - p(z)}{1 + |\gamma|}, \frac{\lambda(\eta z q(0) - z p(0) + e_0)}{1 + |\eta|} \right) \end{aligned}$$

$$=\frac{\lambda}{(1+|\eta|)(1+|\gamma|)}[\eta-\gamma-z(\Pi(\gamma q(z)-p(z)),\eta q(0)-p(0))]$$

$$=\frac{\lambda}{(1+|\eta|)(1+|\gamma|)}[-[a_1(z)-a_2(z)\eta]+\gamma[a_3(z)-a_4(z)\eta]]$$

and the term on the right-hand side equals zero for $\gamma=\phi_{[\eta]}(z)$. Thus (2.15) holds.

We are now prepared to show part (c). First, notice that also $zI-A_{[\eta]}$ is a one-dimensional extension of $zI-A$ for all $z\in\mathbb{C}$. Therefore, $\mathscr{R}(zI-A_{[\eta]})$ equals either ℓ^2 or $\mathscr{R}(zI-A)$, and hence is closed for all $z\in\mathbb{C}$. Consequently, $\sigma_{\text{ess}}(A_{[\eta]})=\emptyset$. Secondly, $A\subset A_{[\eta]}\subset A^{\#}$ implies that $\Pi A\Pi=(A^{\#})^*\subset A^*_{[\eta]}\subset\Pi A^{\#}\Pi=A^*$, and hence $A^*_{[\eta]}$ is a one-dimensional extension of $\Pi A\Pi$. Noticing that $\phi_{[\eta]}(0)=\eta$, we may conclude from (2.15) for $z=0$ that $\Pi(\eta q(0)-p(0))/(1+|\eta|)\in\mathscr{D}(A^*_{[\eta]})$. Since the one-dimensional extensions of $\Pi A\Pi$ have been parametrized in part (b), it follows that $A^*_{[\eta]}=\Pi A_{[\eta]}\Pi$ for all η. Taking into account that $\mathscr{R}(zI-A_{[\eta]})$ is closed, we may conclude that $\mathscr{N}(zI-A_{[\eta]})=\emptyset$ if and only if $\mathscr{R}(zI-A_{[\eta]})=\ell^2$, which by (2.15) is equivalent to $\phi_{[\eta]}(z)\neq\infty$. In the latter case, applying [30, Theorem IV.5.2], we find that $z\in\Omega(A_{[\eta]})$, and thus $\sigma(A_{[\eta]})$ has the form claimed in the assertion.

Finally, in the case $\gamma=\phi_{[\eta]}(z)\neq\infty$, it follows again from (2.15) that $(e_j,(zI-A_{[\eta]})^{-1}e_0)=\phi_{[\eta]}(z)q_j(z)-p_j(z)$ for $j\geqslant 0$, and the characterization $(zI-A)^{-1}=[\mathscr{R}(\phi_{[\eta]}(z))]_{\min}$ is proved as in the first part of the proof of Theorem 2.10. Since

$$\mathscr{R}(\phi_{[\eta]}(z))=\mathscr{S}(z)+((\phi_{[\eta]}(z)q_j(z)-p_j(z))q_k(z))_{j,k=0,1,\dots}$$

and $\mathscr{S}(z)$ is of Schmidt class, the same is true for $\mathscr{R}(\phi_{[\eta]}(z))$. This terminates the proof of Theorem 2.11. □

Under the assumptions of Theorem 2.11, suppose in addition that $\mathscr{A}$ is real. Then the extension $A_{[\eta]}$ of A is symmetric if and only if $\eta\in\mathbb{R}\cup\{\infty\}$. It follows from part (b) that $A_{[\eta]}$ is self-adjoint, i.e., we obtain all self-adjoint extensions of the difference operator A in ℓ^2 (cf. with [1; 45, Theorem 2.6]). Notice also that then the corresponding functions $\phi_{[\eta]}(z)$ are just the Cauchy transforms of the *extremal* [44, Theorem 2.13] or *Neumann* solutions [45] of the moment problem (which according to part (c) are discrete).

Suppose that A is bounded (and thus $\mathscr{A}$ is proper and determinate). In this case it is known [23] that there is an exponential decay rate for the entries of the resolvent of the form (2.16). Conversely, any infinite matrix with entries verifying (2.16) is bounded. We thus obtain the following result of Aptekarev, Kaliaguine and Van Assche mentioned already in the introduction.

Theorem 2.12 (cf. with Aptekarev et al. [7, Theorem 1]). *Let A be bounded. Then $z\in\Omega(A)$ if and only if there exists a $\gamma(z)\in\mathbb{C}$ and positive constants $\beta(z)$ and $\delta(z)$ such that for all $j,k\geqslant 0$*

$$|\mathscr{R}(\gamma(z))_{j,k}|\leqslant\beta(z)\cdot\delta(z)^{|k-j|},\quad \delta(z)<1. \tag{2.16}$$

The equivalence of Theorem 2.12 remains true for unbounded difference operators where the sequence of offdiagonal entries $(a_n)_{n\geqslant 0}$ is bounded (see, e.g., [25, Proposition 2.2]) or contains a "sufficiently dense" bounded subsequence (namely, there exists an increasing sequence $(n_k)_{k\geqslant 0}$ of indices so that both sequences $(a_{n_k})_{k\geqslant 0}$ and $(n_{k+1}-n_k)_{k\geqslant 0}$ are bounded, see [18, Theorem 2.1]). In these two cases, the matrix $\mathscr{A}$ is proper according to Example 2.7.

Notice that in the original statement of [7, Theorem 1] the authors impose some additional conditions on z which are not necessary. Also, the authors treat general tridiagonal matrices $\mathscr{A}$ where the entries of the superdiagonal may differ from those on the subdiagonal. Such operators can be obtained by multiplying a complex Jacobi matrix on the left by some suitable diagonal matrix and on the right by its inverse, i.e., we rescale our recurrence relation (1.2). Such recurrence relations occur for instance in the context of monic (F)OPs, whereas we have chosen the normalization of orthonormal FOPs. The following result of Kaliaguine and Beckermann shows that our normalization gives the smallest spectrum.

Theorem 2.13 (Beckermann and Kaliaguine [20, Theorem 2.3]). *Let $\mathscr{A}$ be a bounded complex Jacobi matrix, and consider a bounded tridiagonal matrix $\mathscr{A}' = \mathscr{D}\mathscr{A}\mathscr{D}^{-1}$ with diagonal $\mathscr{D}$. Then for the corresponding difference operators A and A' we have $\Omega(A') \subset \Omega(A)$.*

As an example, take the tridiagonal Toeplitz matrix with diagonal entries $a/2, 0, 1/(2a)$. Here it is known that the spectrum is the interior and the boundary of an ellipse with foci ± 1 and half axes $|a \pm 1/a|/2$, and it is minimal (namely the interval $[-1, 1]$) for $a = 1$. Notice also that for monic FOPs one chooses the normalization $a = \frac{1}{2}$.

It would be interesting to generalize Theorem 2.13 to unbounded Jacobi matrices.

2.4. The Weyl function and functions of the second kind

Following Berezanskii (see [21]), we call

$$\phi(z) := (e_0, (zI - A)^{-1} e_0), \quad z \in \Omega(A), \tag{2.17}$$

the *Weyl function* of A. Since the resolvent is analytic on $\Omega(A)$, the same is true for the Weyl function. If the operator A is bounded (or, equivalently, if the entries of $\mathscr{A}$ are uniformly bounded), then ϕ is analytic for $|z| > ||A||$. Then its Laurent series at infinity is given by

$$\phi(z) \sim \sum_{j=0}^{\infty} \frac{(e_0, A^j e_0)}{z^{j+1}}, \tag{2.18}$$

i.e., its coefficients are the moments of the linear functional c of formal orthogonality (some authors refer to the series on the right-hand side of (2.18) as the symbol of c). In the case where the numerical range of A is not the whole plane (as for instance for real Jacobi matrices), one may show (see, e.g., [59, Theorem 84.3]) that (2.18) can be interpreted as an asymptotic expansion of ϕ in some sector.

The associated *functions of the second kind* are given by

$$r_n(z) = (e_n, (zI - A)^{-1} e_0) = q_n(z)\phi(z) - p_n(z), \quad n \geqslant 0, \; z \in \Omega(A),$$

where the last representation follows from Theorem 2.10 and the construction of $\mathscr{R}(\phi(z))$. Similarly, we may express the other entries as

$$(e_j, (zI - A)^{-1} e_k) = (e_k, (zI - A)^{-1} e_j) = r_k(z) q_j(z), \quad 0 \leqslant j \leqslant k, \; z \in \Omega(A). \tag{2.19}$$

In case of a bounded operator A, we know from [18, Theorem 5.3] that the Weyl function contains already all information about isolated points of the spectrum. The proof given for this assertion only uses the representation (2.19), and thus also applies for unbounded operators.

Theorem 2.14 (Beckermann [18, Theorem 5.3 and Corollary 5.6]). *Let ζ be an isolated point of $\sigma(A)$. Then $\zeta \in \sigma_{\rm ess}(A)$ if and only if ϕ has an essential singularity in ζ, and ζ is an eigenvalue of algebraic multiplicity $m<\infty$ if and only if ϕ has a pole of multiplicity m. In particular, if $\sigma(A)$ is countable, then the set of singularities of ϕ coincides with $\sigma(A)$.*

A proof of the second sentence of Theorem 2.14 is based on the observation that any element of a closed and countable set $\Sigma \subset \mathbb{C}$ is either an isolated point or a limit of isolated points of Σ. Notice that the spectrum is in particular countable and has the only accumulation point $b \in \mathbb{C}$ if $A - bI$ is compact, i.e., $a_n \to 0$ and $b_n \to b$ (see for instance Corollary 2.17 below). Here the Weyl function is analytic in $\Omega(A)$ (and in no larger set), meromorphic in $\mathbb{C}\setminus\{b\}$, and has an essential singularity at b. For a nice survey on compact Jacobi matrices we refer the reader to [57].

Relation (2.19) allows us also to compare the growth of FOPs and of functions of the second kind. Indeed, according to (2.9) we have

$$a_n(q_{n+1}(z)r_n(z) - r_{n+1}(z)q_n(z)) = 1, \quad n \geqslant 0,\ z \in \Omega(A). \tag{2.20}$$

This implies that

$$\begin{aligned} 1 &\leqslant |a_n|\sqrt{|q_n(z)|^2 + |q_{n+1}(z)|^2}\sqrt{|r_n(z)|^2 + |r_{n+1}(z)|^2} \\ &\leqslant 1 + 2|a_n| \cdot ||(zI - A)^{-1}||, \end{aligned} \tag{2.21}$$

$$\begin{aligned} 1 &\leqslant \sqrt{|q_n(z)|^2 + |a_n q_{n+1}(z)|^2}\sqrt{|r_n(z)|^2 + |a_n r_{n+1}(z)|^2} \\ &\leqslant 1 + (1 + |a_n|^2)||(zI - A)^{-1}|| \end{aligned} \tag{2.22}$$

for all $z \in \Omega(A)$ and $n \geqslant 0$. Indeed, the left-hand inequalities of (2.21), (2.22) follow by applying the Cauchy–Schwarz inequality on (2.20). In order to verify the right-hand estimate in, e.g., (2.21), we notice that, by (2.20),

$$\begin{aligned} &|a_n|^2(|q_n(z)|^2 + |q_{n+1}(z)|^2)(|r_n(z)|^2 + |r_{n+1}(z)|^2) \\ &\quad = |a_n|^2[|q_n(z)r_n(z)|^2 + |q_n(z)r_{n+1}(z)|^2 + |q_{n+1}(z)r_{n+1}(z)|^2] + |1 + a_n r_{n+1}(z)q_n(z)|^2. \end{aligned}$$

Each term of the form $r_j(z)q_k(z)$ occurring on the right-hand side may be bounded by $||(zI - A)^{-1}||$, leading to (2.21).

If additional information on the sequence $(a_n)_{n\geqslant 0}$ is available, we may be even much more precise.

Corollary 2.15. *Let $(a_n)_{n\geqslant 0}$ be bounded. Then there exist continuous functions $\beta : \Omega(A) \to (0, +\infty)$ and $\delta : \Omega(A) \to (0, 1)$ such that for all $0 \leqslant j \leqslant k$ and for all $z \in \Omega(A)$*

$$|r_k(z) \cdot q_j(z)| \leqslant \beta(z) \cdot \delta(z)^{k-j}. \tag{2.23}$$

If in addition A is bounded, then the functions $\beta(z)$ and $|z| \cdot \delta(z)$ are continuous at infinity.

Here (2.23) follows from Theorem 2.12. The continuity of the functions β, δ has been discussed in [20, Lemma 3.3; 18, Lemma 2.3] for bounded A, and implicitly in [19, proof of Theorem 2.1] for bounded $(a_n)_{n\geqslant 0}$.

2.5. Some special cases

It is well known (see, e.g., [11]) that a linear functional c having real moments is positive (i.e., $\det(c(x^{j+k}))_{j,k=0,\dots,n} > 0$ for all $n \geqslant 0$) if and only if c has the representation

$$c(P) = \int P(x)\,\mathrm{d}\mu(x) \quad \text{for any polynomial } P, \tag{2.24}$$

where μ is some positive Borel measure with real infinite support $\operatorname{supp}(\mu)$. Under these assumptions, the support is a part of the positive real axis if and only if in addition $\det(c(x^{j+k+1}))_{j,k=0,\dots,n} > 0$ for all $n \geqslant 0$. Furthermore, the sequence of moments is totally monotone (i.e., $(-1)^k \Delta^k c(x^n) > 0$ for all $n, k \geqslant 0$, see [11, Section 5.4.1]) if and only if (2.24) holds with μ some positive Borel measure with infinite support $\operatorname{supp}(\mu) \subset [0,1]$.

In all these cases, the corresponding Jacobi matrix is real, and the corresponding measure is unique (uniqueness of the moment problem) if and only if $\mathscr{A}$ is proper (in other words, A is self-adjoint). In this case, μ can be obtained by the Spectral Theorem, with $\operatorname{supp}(\mu) = \sigma(A) \subset \mathbb{R}$, and

$$\phi(z) = \int \frac{\mathrm{d}\mu(x)}{z - x} \tag{2.25}$$

holds for all $z \notin \sigma(A)$.

In case of complex bounded Jacobi matrices (or more general proper operators with $\Omega(A) \not\subset \Gamma(A)$), we may also obtain a complex-valued measure μ satisfying (2.24) and (2.25) via the Cauchy integral formula; however, in general $\sigma(A) \neq \operatorname{supp}(\mu)$. In all these cases, we recover the following well-known representation of functions of the second kind as Cauchy transforms.

Lemma 2.16. *Suppose that there exists some* (*real- or complex-valued*) *Borel measure μ such that* (2.24) *holds*, *and some set $U \subset \Omega(A)$ such that* (2.25) *is true for all $z \in U$. Then*

$$r_k(z) q_j(z) = \int \frac{q_j(x) q_k(x)}{z - x}\,\mathrm{d}\mu(x), \quad 0 \leqslant j \leqslant k,\ z \in U.$$

Proof. Consider the sequence of Cauchy transforms

$$\tilde{r}_n(z) := \int \frac{q_n(x)}{z - x}\,\mathrm{d}\mu(x), \quad n \geqslant 0$$

and $\tilde{r}_{-1} = 0$. One easily checks, using (2.24) and (1.2), that $-a_n \tilde{r}_{n+1}(z) + (z - b_n)\tilde{r}_n(z) - a_{n-1}\tilde{r}_{n-1}(z) = \int q_n(x)\,\mathrm{d}\mu(x) = c(q_n) = \delta_{n,0}$ for $n \geqslant 0$. Moreover, $\tilde{r}_0(z) = \phi(z) = r_0(z)$ for $z \in U$ by (2.25) and (2.19). Consequently, for $z \in U$, $(\tilde{r}_n(z))_{n \geqslant 0}$ satisfies the same recurrence and initialization as the sequence $(r_n(z))_{n \geqslant 0}$, implying that $\tilde{r}_n(z) = r_n(z)$. Furthermore, for $j \leqslant k$ there holds

$$r_k(z) q_j(z) - \int \frac{q_j(x) q_k(x)}{z - x}\,\mathrm{d}\mu(x) = \int \frac{q_j(z) - q_j(x)}{z - x} q_k(x)\,\mathrm{d}\mu(x).$$

Since the fraction on the right-hand side is a polynomial of degree $< j \leqslant k$ in x, the right-hand integral vanishes by orthogonality and (2.24). □

If A is bounded, then any measure with compact support satisfying (2.24) will fulfill (2.25) with U being equal to the unbounded component of the complement of $\sigma(A) \cup \operatorname{supp}(\mu)$, since the functions on both sides of (2.25) have the same Laurent expansion at infinity. It would be very interesting to prove for general nonreal (unbounded but proper) Jacobi matrices that if (2.24) holds for some measure with compact support, then also (2.25) is true for $z \in U$, where U is the intersection of $\Omega(A)$ with the unbounded connected component of $\mathbb{C}\backslash\mathrm{supp}(\mu)$.

To conclude this section let us have a look at a different class of functionals which to our knowledge has not yet been studied in the context of complex Jacobi matrices: It is known from the work of Schoenberg and Edrei that the sequence of (real) moments $(c_n)_{n\geqslant 0}$, $c_n = c(x^n)$, $c_0 = 1$, $c_n = 0$ for $n < 0$, is totally positive (i.e., $\det(c_{m+j-k})_{j,k=0,\dots,n-1} \geqslant 0$ for all $n, m \geqslant 0$) if and only if $\sum c_j z^j$ is the expansion at zero of a meromorphic function ψ having the representation

$$\psi(z) = \mathrm{e}^{\gamma z} \prod_{j=1}^{\infty} \frac{1 + \alpha_j z}{1 - \beta_j z}, \quad \alpha_j, \beta_j, \gamma \geqslant 0, \quad \sum_{j=1}^{\infty} (\alpha_j + \beta_j) < \infty$$

(including, for instance, the exponential function). Following [9], we exclude the case that ψ is rational. Many results about convergence of Padé approximants (at zero) of these functions have been obtained in [9], see also [11]. Let us consider the linear functionals $c^{[k]}$ defined by

$$c^{[k]}(x^n) = c_{n+k}, \quad n, k \geqslant 0, \quad \text{with symbol } \phi^{[k]}(z) = z^{k-1} \left(\psi(1/z) - \sum_{j=0}^{k-1} \frac{c_j}{z^j} \right)$$

($c^{[0]} = c$), having symbols which are meromorphic in $\mathbb{C}\backslash\{0\}$, and analytic around infinity. We have the following

Corollary 2.17. *The functionals $c^{[k]}$ as described above are regular for all $k \geqslant 0$. The associated complex Jacobi matrices $\mathscr{A}^{[k]}$ are compact, with Weyl function given by $\phi^{[k]}$, and spectrum $\{0\} \cup \{\beta_j: j \geqslant 1\}$. Finally, $(a_n^{[k]})^2 < 0$ for all $n, k \geqslant 0$.*

Proof. In [9, Theorem 1.I], the authors show that the Padé table of ψ (at zero) is normal. Denote by $Q_{m,n}$ the denominator of the Padé approximant of type $[m|n]$ at zero, normalized so that $Q_{m,n}(0) = 1$, and define

$$Q_n^{[k]}(z) := z^n Q_{n+k,n}\left(\frac{1}{z}\right) = z^n + Q_{n,1}^{[k]} z^{n-1} + Q_{n,2}^{[k]} z^{n-2} + \cdots.$$

It is well known and easily verified that $Q_n^{[k]}$ is an nth monic FOP of the linear functional $c^{[k]}$, and thus $c^{[k]}$ is regular. The sign of the recurrence coefficient $a_n^{[k]}$ follows from well-known determinantal representations; we omit the details. Precise asymptotics for $(Q_{n+k,n})_{n\geqslant 0}$ are given in [9, Theorem 1.II] (see also [11]), implying that

$$\lim_{n\to\infty} \frac{Q_n^{[k]}(z)}{z^n} = \exp\left(\frac{-\gamma}{2z}\right) \prod_{j=1}^{\infty} \left(1 - \frac{\beta_j}{z}\right) \tag{2.26}$$

for all $k \geqslant 0$ uniformly on closed subsets of $(\mathbb{C} \cup \{\infty\})\backslash\{0\}$. In particular, the sequences $(Q_{n,1}^{[k]})_{n\geqslant 0}$ and $(Q_{n,2}^{[k]})_{n\geqslant 0}$ converge. On the other hand, we know from (1.2) that $Q_{n+1}^{[k]}(z) = (z - b_n^{[k]}) Q_n^{[k]}(z) - (a_{n-1}^{[k]})^2 Q_{n-1}^{[k]}(z)$. Thus

$$Q_{n+1,1}^{[k]} = Q_{n,1}^{[k]} - b_n^{[k]}, \quad Q_{n+1,2}^{[k]} = Q_{n,2}^{[k]} - b_n^{[k]} Q_{n,1}^{[k]} - (a_{n-1}^{[k]})^2, \tag{2.27}$$

implying that $b_n^{[k]} \to 0$ and $a_n^{[k]} \to 0$ for $n \to \infty$. Hence $\mathscr{A}^{[k]}$ is compact. Since its Weyl function ϕ has the same (convergent) Laurent expansion around infinity as $\phi^{[k]}$, we have $\phi = \phi^{[k]}$, and the rest of the assertion follows from Theorem 2.14 and the explicit knowledge of the singularities of $\phi^{[k]}$. □

3. Asymptotics of FOPs

3.1. *n*th-root asymptotics of FOPs

In this subsection we will restrict ourselves to bounded Jacobi matrices A. We present some recent results of [7,18,20,53,54].

In their work on tridiagonal infinite matrices, Aptekarev, Kaliaguine and Van Assche also observed [7, Corollary 3] that

$$\limsup_{n\to\infty} |q_n(z)|^{1/n} > 1, \quad z \in \Omega(A).$$

Indeed, a combination of (2.23) for $j = 0$ and (2.21) yields the stronger relation

$$\liminf_{n\to\infty} [|q_n(z)|^2 + |q_{n+1}(z)|^2|]^{1/(2n)} > 1, \quad z \in \Omega(A). \tag{3.1}$$

For real bounded Jacobi matrices, this relation was already established by Szwarc [53, Corollary 1], who showed by examples [54] that there may be also exponential growth inside the spectrum.

Kaliaguine and Beckermann [20, Theorem 3.6] applied the maximum principle to the sequence of functions of the second kind and showed that, in the unbounded connected component $\Omega_0(A)$ of the resolvent set $\Omega(A)$, one may replace 1 on the right-hand side of (3.1) by $\exp(g_{\sigma(A)}(z))$. Here $g_{\sigma(A)}$ denotes the (generalized) Green function with pole at ∞ of the compact set $\sigma(A)$, being characterized by the three properties (see, e.g., [41, Section II.4]):

(i) $g_{\sigma(A)}$ is nonnegative and harmonic in $\Omega_0(A)\backslash\{\infty\}$,
(ii) $g_{\sigma(A)}(z) - \log|z|$ has a limit for $|z| \to \infty$,
(iii) $\lim_{z\to\zeta, z\in\Omega_0(A)} g_{\sigma(A)}(z) = 0$ for quasi-every $\zeta \in \partial\Omega_0(A)$.

We also recall that the limit in (ii) equals $-\log \mathrm{cap}(\sigma(A))$, where $\mathrm{cap}(\cdot)$ is the logarithmic capacity. A detailed study of nth-root asymptotics of formal orthogonal polynomials with bounded recurrence coefficients has been given in [18]. We denote by k_n the leading coefficient of q_n, i.e.,

$$k_n = \frac{1}{a_0 \cdot a_1 \cdots a_{n-1}},$$

and define the quantities

$$\kappa_{\sup} := \limsup_{n\to\infty} |k_n|^{-1/n}, \quad \kappa_{\inf} := \liminf_{n\to\infty} |k_n|^{-1/n}.$$

Notice that $|a_n| \leqslant ||A||$, and thus $|k_n|^{1/n} \geqslant 1/||A||$, implying that $0 \leqslant \kappa_{\inf} \leqslant \kappa_{\sup} \leqslant ||A||$.

Theorem 3.1 (See Beckermann [18, Theorems 2.5 and 2.10]). *Let A be bounded. Then there exist functions $g_{\inf}, g_{\sup}$ such that*

$$\liminf_{n\to\infty} (|q_n(z)|^2 + |a_n q_{n+1}(z)|^2)^{-1/(2n)} = \exp(-g_{\sup}(z)), \tag{3.2}$$

$$\limsup_{n\to\infty}(|q_n(z)|^2 + |a_n q_{n+1}(z)|^2)^{-1/(2n)} = \exp(-g_{\inf}(z)), \tag{3.3}$$

holds uniformly on closed subsets of $\Omega(A)$.

Here $g_{\inf} = +\infty$ (*and* $g_{\sup} = +\infty$, *resp.*) *if and only if* $\kappa_{\sup} = 0$ (*and* $\kappa_{\inf} = 0$, *resp.*). *Otherwise*, $g_{\inf}$ *is superharmonic*, *strictly positive*, *and continuous in* $\Omega(A)\setminus\{\infty\}$, *with*

$$\lim_{|z|\to\infty} g_{\inf}(z) - \log|z| = \log\frac{1}{\kappa_{\sup}}.$$

Also, $g_{\sup}$ *is subharmonic*, *strictly positive*, *and continuous in* $\Omega(A)\setminus\{\infty\}$, *with*

$$\lim_{|z|\to\infty} g_{\sup}(z) - \log|z| = \log\frac{1}{\kappa_{\inf}}.$$

In addition,

$$0 \leqslant \kappa_{\inf} \leqslant \kappa_{\sup} \leqslant \operatorname{cap}(\sigma(A)), \qquad g_{\sigma(A)}(z) \leqslant g_{\inf}(z) \leqslant g_{\sup}(z), \quad z \in \Omega_0(A). \tag{3.4}$$

Various further properties and relations between $g_{\sigma(A)}$, $g_{\inf}$ and $g_{\sup}$ may be found in [18, Sections 2.2, 2.3]. The proof of Theorem 3.1 is based on (2.22), Corollary 2.15, and applies tools from logarithmic potential theory. Instead of giving details, let us discuss some consequences and special cases. First, since $(a_n)_{n\geqslant 0}$ is bounded, we obtain from (3.2) that

$$\limsup_{n\to\infty} |q_n(z)|^{1/n} = \exp(g_{\sup}(z)) > 1, \quad z \in \Omega(A). \tag{3.5}$$

Furthermore, we will show below that (3.3) implies the relation

$$\liminf_{n\to\infty} |q_n(z)|^{1/n} = \exp(g_{\inf}(z)) > 1, \quad z \in F, \tag{3.6}$$

provided that the set $F \subset \Omega_0(A)$ does not contain any of the zeros of q_n for sufficiently large n. In addition, by combining (3.3) with (2.22) we get

$$\limsup_{n\to\infty} |r_n(z)|^{1/n} = \exp(-g_{\inf}(z)) < 1, \quad z \in \Omega(A). \tag{3.7}$$

Indeed, relation (2.22) allows us to restate Theorem 3.1 in terms of functions of the second kind.

The simplest case which may illustrate these findings is the Toeplitz operator $\mathscr{A}$ with $a_n = \frac{1}{2}$, $b_n = 0$, $n \geqslant 0$, see [38, Section II.9.2].[4] Here one may write down explicitly q_n and r_n in terms of the Joukowski function; in particular one finds that $\sigma(A) = [-1, 1]$, and $g_{\inf} = g_{\sup} = g_{[-1,1]}$. Of course, in the generic case there will be no particular relation between $g_{\sup}$, $g_{\inf}$, and $g_{\sigma(A)}$. Some extremal cases of Theorem 3.1 have been discussed in [18, Example 2.9] (see also [25, Examples 4.1, 4.2]). For instance, there are operators with $\sigma(A) = [-1, 3]$, and $g_{\inf} = g_{\sup} = g_{[-1,1]}$. Also, the case $\sigma(A) = [-2, 2]$, $\kappa_{\inf} = \kappa_{\sup} = \frac{1}{2} < \operatorname{cap}(\sigma(A)) = 1$, and $g_{\inf} \neq g_{\sup}$ may occur. In addition there is an example where $g_{\sup}(z) - g_{\inf}(z) = \log(\kappa_{\sup}/\kappa_{\inf}) \neq 0$ for all $z \in \Omega(A)$.

The nth-root asymptotics of general orthogonal polynomials are investigated by Stahl and Totik [51]. Of course, in case of orthogonality on the real line (i.e., real Jacobi matrices) results such as (3.4)–(3.6) have been known before, see, e.g., [51, Theorem 1.1.4, Corollary 1.1.7].

In this context we should mention the deep work of Stahl concerning the convergence of Padé approximants and asymptotics of the related formal orthogonal polynomials. He considers linear

[4] See also the case of periodic complex Jacobi matrices discussed in Section 4.3 below.

functionals as in (2.24), where μ is some (real or complex valued but not positive) Borel measure with compact support. Of course, such functionals are not necessarily regular, but we can always consider the asymptotics of the subsequence of (unique) FOPs corresponding to normal points. In [46, Corollary of Theorem 1] Stahl constructs a measure μ supported on $[-1,1]$ such that the sequence of normalized zero counting measures is weakly dense in the set of positive Borel measures of total mass $\leqslant 1$ supported on $\mathbb{C}$. In contrast, in the case of regular functionals and bounded recurrence coefficients, it is shown in [18, Theorem 2.5] that the support of any partial weak limit of the sequence of normalized zero counting measures is a subset of $\mathbb{C}\backslash\Omega_0(A)$.

Another very interesting class has been considered by Stahl in a number of papers (see, for instance, [49]), here the symbols (the Cauchy transform of μ) are multivalued functions having, e.g., a countable number of branchpoints. Here it follows from [49, Theorems 1.7, 1.8] that (3.5) and (3.7) hold quasi-everywhere outside of supp(μ), with $g_{\inf} = g_{\text{supp}}$ being the Green function of supp(μ). Again, it is not clear whether the functional is regular and the corresponding recurrence coefficients are bounded.

Linear functionals of the form

$$c_w(P) = \int_{-1}^{1} \frac{w(x)P(x)}{\sqrt{1-x^2}}\,\mathrm{d}x \tag{3.8}$$

with some possibly complex-valued weight function w have been discussed by a number of authors, see, e.g., the introduction of [46]. Nuttall [39], Nuttall and Wherry [40], Baxter [17], Magnus [33], and Baratchart and Totik [12] suggested conditions on w insuring that all (at least sufficiently large) indices n are normal, and that there are only "few" zeros outside of $[-1,1]$. In particular, nth-root asymptotics for the sequence of FOPs are derived.

3.2. Ratio asymptotics and zeros of FOP

It is well known that the monic polynomial q_n/k_n is the characteristic polynomial of the finite section $\mathscr{A}_n$ obtained by taking the first n rows and columns of $\mathscr{A}$. In this section we will be concerned with the location of zeros of FOPs, i.e., of eigenvalues of $\mathscr{A}_n$. In numerical linear algebra, one often refers to these zeros as Ritz values. The motivation for our work is the idea that the sequence of matrices $\mathscr{A}_n$ approximates in some sense the infinite matrix $\mathscr{A}$ and thus the corresponding difference operator A; therefore the corresponding spectra should be related. In the sequel, we will try to make this statement more precise.

An important tool in our investigations is the rational function[5]

$$u_n(z) := \frac{q_n(z)}{a_n q_{n+1}(z)} = \frac{q_n(z)/k_n}{q_{n+1}(z)/k_{n+1}}$$

[5] Most of the results presented in this paper for the sequence (u_n) are equally valid for the ratio

$$\phi^{(n+1)}(z) := \frac{r_{n+1}(z)}{a_n r_n(z)},$$

which can be shown to have a meromorphic continuation in $\mathbb{C}\backslash\sigma_{\text{ess}}(A)$, and coincides with the Weyl function of the associated Jacobi matrix $\mathscr{A}^{(n+1)}$. Some additional interesting properties are presented in a future publication.

$$= \frac{\det(z\mathscr{I}_n - \mathscr{A}_n)}{\det(z\mathscr{I}_{n+1} - \mathscr{A}_{n+1})} = (e_n, (z\mathscr{I}_{n+1} - \mathscr{A}_{n+1})^{-1} e_n).$$

Here and in the sequel we denote by $e_0, \ldots, e_n$ also the canonical basis of $\mathbb{C}^{n+1}$, the length of the vectors being clear from the context. In the theory of continued fractions, the sequence $(1/u_n)_{n \geqslant 0}$ of meromorphic functions is referred to as a tail sequence of the J-fraction (1.3) [32, Section II.1.2, Eq. (1.2.7)]. Using (1.2), one easily verifies that

$$\frac{1}{zu_n(z)} = \frac{a_n q_{n+1}(z)}{z \cdot q_n(z)} = 1 - \frac{b_n}{z} - \frac{a_{n-1}^2}{z^2} + \mathscr{O}\left(\frac{1}{z^3}\right)_{z \to \infty}. \tag{3.9}$$

In order to motivate our results presented below, we briefly recall some properties of orthogonal polynomials, i.e., real Jacobi matrices. It is well known that here the zeros of q_n are simple, and lie in the convex hull $\mathscr{S}$ of $\sigma(A)$. Also, they interlace with the zeros of q_{n+1}, and thus u_n has positive residuals. These two facts allow us to conclude that $(u_n)_{n \geqslant 0}$ is bounded uniformly in closed subsets of $\mathbb{C} \backslash \mathscr{S}$. Finally, q_n can have at most one zero in a *gap* of the form $(a,b) \subset \mathscr{S} \backslash \sigma(A)$.

We should mention first that none of these properties remains valid for FOPs. Classical counterexamples known from Padé approximation (such as the examples of Perron and of Gammel-Wallin, see [11]) use linear functionals c which are highly nonregular. But there also exist other ones.

Example 3.2. (a) The linear functional (3.8) with weight $w(x) = (x - \cos(\alpha_1 \pi))(x - \cos(\alpha_2 \pi))$ has been studied in detail by Stahl [47]. Provided that 1, α_1, α_2 are rationally independent, Stahl showed that c is regular, but (two) zeros of the sequence of FOPs cluster everywhere in $\mathbb{C}$.

Not very much is known about the associated (nonreal) Jacobi matrix. Theorem 3.4(a) below shows that $\Gamma(A) = \mathbb{C}$; in particular, A is unbounded. Also, it follows from [47, Section 5] that $(a_n)_{n \geqslant 0}$ contains a bounded subsequence, and hence $\mathscr{A}$ is proper (and determinate) by Example 2.7. On the other hand, it is unknown whether $\sigma(A)$ (or $\sigma_{\text{ess}}(A)$) equals the whole plane.

(b) Beckermann [18, Example 5.7] investigated the linear functional with generating function

$$\phi_d(z) = (z - d)\left[\exp\left(\frac{1}{z^2 - 1}\right) - 1\right].$$

Here the coefficients of the recurrence relation are given by $a_0^2 = \frac{3}{2} - d^2$, and

$$b_{2n} = -d, \quad b_{2n+1} = d, \quad -a_{2n}^2 a_{2n+1}^2 = \frac{1}{4(2n+1)(2n+3)}, \quad a_{2n+2}^2 + a_{2n+1}^2 = 1 - d^2$$

for $n \geqslant 0$ (provided that there is no division by zero, which can be insured for instance if $d \in (-\infty, -\sqrt{3/2}) \cup [-1, 1] \cup (\sqrt{3/2}, +\infty)$). One may show that $a_{2n-1} \to 0$, and thus $\mathscr{A}$ is bounded but not real. Also, $\sigma(A) = \sigma_{\text{ess}}(A) = \{\pm 1\}$. Furthermore, $q_{2n-1}(-d) = 0$ for all $n \geqslant 0$, and $-d$ may be far from the convex hull of $\sigma(A)$.

Below we will see, however, that many of the properties for OPs remain valid for FOPs outside[6] the numerical range $\mathscr{S} = \Gamma(A)$. An important tool in these investigations is the notion of normal families as introduced by Montel: a sequence of functions analytic in some domain D is called a

[6] Notice that, for real $\mathscr{A}$, the numerical range $\Gamma(A)$ coincides with the convex hull of the spectrum. It is known from examples [20] that this property is no longer true for general complex Jacobi matrices.

normal family if from each subsequence we may extract a subsequence which converges locally uniformly in D (i.e., uniformly on closed subsets[7] of D), with the limit being different from the constant ∞. By a Theorem of Montel [43, Section 2.2, Theorem 2.2.2], a family of functions analytic in D is normal in D if and only if it is uniformly bounded on any closed subset of D. More generally, we will also consider sequences of functions being meromorphic in D. Such a sequence is called normal in D if, given a subsequence, we may extract a subsequence converging locally uniformly in D with respect to the chordal metric $\chi(\cdot)$ on the Riemann sphere [43, Definition 3.1.1]. Notice that normal families of analytic functions are also normal families of meromorphic functions, but the converse is clearly not true.

Theorem 3.3. (a) *The sequence $(u_n)_{n\geqslant 0}$ is bounded above uniformly on compact subsets of $\mathbb{C}\backslash\Gamma(A)$.*

(b) *The sequence $(u_n)_{n\geqslant 0}$ of meromorphic functions is normal around infinity if and only if A is bounded.*

(c) (*Cf. with Beckermann* [18, *Proposition* 2.2]). *Let Λ be some infinite set of integers such that $(a_n)_{n\in\Lambda}$ is bounded. Then the sequence $(u_n)_{n\in\Lambda}$ of meromorphic functions is normal in $\Omega(A)$.*

Proof. (a) We first observe that there is a connection between the numerical range of the difference operator and the numerical range of the finite sections $\mathscr{A}_n$, namely[8]

$$\Gamma(\mathscr{A}_n)=\Theta(\mathscr{A}_n)=\left\{\frac{(y,Ay)}{(y,y)}\colon\ y\in\mathscr{C}_0,\ \Pi_n y=y\right\}\subset\Theta(A)\subset\Gamma(A).$$

Since

$$\frac{1}{||(z\mathscr{I}_n-\mathscr{A}_n)^{-1}||}=\min_{y\in\mathbb{C}^n}\frac{||(z\mathscr{I}_n-\mathscr{A}_n)y||}{||y||}\geqslant\min_{y\in\mathbb{C}^n}\left|\frac{(y,(z\mathscr{I}_n-\mathscr{A}_n)y)}{(y,y)}\right|=\operatorname{dist}(z,\Theta(\mathscr{A}_n)),$$

we may conclude that

$$|u_n(z)|=|(e_n,(z\mathscr{I}_{n+1}-\mathscr{A}_{n+1})^{-1}e_n)|\leqslant||(z\mathscr{I}_{n+1}-\mathscr{A}_{n+1})^{-1}||\leqslant\frac{1}{\operatorname{dist}(z,\Gamma(A))},$$

leading to the claim of part (a).

(b) If A is bounded then $\Gamma(A)$ is compact. Hence its complement contains a neighborhood D of infinity (for instance the set $|z|>||A||$), and $(u_n)_{n\geqslant 0}$ is a normal family of analytic functions in U according to part (a) and the Theorem of Montel. Conversely, suppose that $(u_n)_{n\geqslant 0}$ is a normal family of meromorphic functions in a neighborhood D of infinity. Then $(u_n)_{n\geqslant 0}$ is equicontinuous in D (with respect to the chordal metric). Since $u_n(\infty)=0$ for all $n\geqslant 0$, there exists some $R>0$

[7] All the subsequent considerations are in the extended complex plane $\overline{\mathbb{C}}=\mathbb{C}\cup\{\infty\}$, equipped with the chordal metric $\chi(\cdot)$.

[8] Indeed, using (2.2) one immediately obtains the more precise relation

$$\Gamma(A)=\operatorname{Clos}\left(\bigcup_{n\geqslant 0}\Gamma(\mathscr{A}_n)\right).$$

such that

$$\chi(u_n(z),0)=\frac{|u_n(z)|}{\sqrt{1+|u_n(z)|^2}}\leqslant\frac{1}{2},\quad n\geqslant 0,\ |z|\geqslant R.$$

It follows that $|u_n(z)|\leqslant 1$ for all $n\geqslant 0$ and $|z|\geqslant R$. Applying the maximum principle for analytic functions, we obtain $|\tilde{u}_n(z)|\leqslant R$ for all $n\geqslant 0$ and $|z|\geqslant R$, where $\tilde{u}_n(z)=z\cdot u_n(z)$. Consequently, both $(\tilde{u}_n)_{n\geqslant 0}$ and $(1/\tilde{u}_n)_{n\geqslant 0}$ are normal families of meromorphic functions in $|z|>R$. Since $\tilde{u}_n(\infty)=1$, it follows again from equicontinuity that $(1/\tilde{u}_n)_{n\geqslant 0}$ is bounded above by some constant M for $|z|>R'$ with some suitable $R'>R$. Using the Cauchy formula, we obtain

$$|(1/\tilde{u}_n)'(\infty)|\leqslant M\cdot R',\qquad |(1/\tilde{u}_n)''(\infty)|\leqslant\frac{M\cdot(R')^2}{2},\quad n\geqslant 0.$$

Taking into account (3.9), we may conclude that both sequences $(b_n)_{n\geqslant 0}$, $(a_n)_{n\geqslant 0}$ are bounded, and thus the operator A is bounded.

(c) Here we closely follow arguments from [18, Proof of Proposition 2.2]. By the Marty Theorem [43, Section 3], the sequence $(u_n)_{n\in\Lambda}$ is a normal family of meromorphic functions in some domain $D\subset\mathbb{C}$ if and only if the spherical derivative

$$\rho(u_n):=\frac{|u_n'|}{1+|u_n|^2}$$

is bounded uniformly with respect to $n\in\Lambda$ on compact subsets of D. Using the confluent limit of the Christoffel–Darboux formula

$$a_n\cdot\frac{q_n(x)q_{n+1}(z)-q_n(z)q_{n+1}(x)}{z-x}=\sum_{j=0}^{n}q_j(x)\cdot q_j(z),$$

one obtains

$$|\rho(u_n)(z)|=\frac{|\sum_{j=0}^{n}q_j(z)^2|}{|q_n(z)|^2+|a_nq_{n+1}(z)|^2}.$$

According to (2.14), the right-hand side is bounded above by $\max(1,|a_n|^2)||(zI-A)^{-1}||$, and this quantity is bounded on closed subsets of $\Omega(A)$ uniformly for $n\in\Lambda$ by assumption on (a_n). □

Let us briefly comment on Theorem 3.3. Part (c) has been stated in [18, Proposition 2.2] for bounded difference operators. Then, of course, the whole sequence (u_n) is normal in $\Omega(A)$, and from the proof of part (b) we see that any partial limit of (u_n) is different from the constants 0, ∞ in the unbounded connected component $\Omega_0(A)$ of $\Omega(A)$. If A is no longer bounded, then things become much more involved. However, for real Jacobi matrices we still obtain from part (a) the normality in $\mathbb{C}\backslash\mathbb{R}$. On the other hand, we see from part (b) that expansion (3.9) can only be exploited for bounded difference operators.

A different proof of part (b) can be based on the following observation. For unbounded operators, it is interesting to consider the so-called contracted zero distribution (for real Jacobi matrices see, e.g., [25,56]): Since the eigenvalues of $\mathscr{A}_n/||\mathscr{A}_n||$ are all in the unit disk, one easily verifies that $\tilde{q}_n(z)=q_n(||\mathscr{A}_n||z)$ has its zeros in the unit disk. As a consequence, one may derive nth-root asymptotics for $(\tilde{q}_n)$. Indeed, for particular families of recurrence coefficients (Hermite, Laguerre, or

Freud polynomials) even stronger asymptotics have been derived in the last years, see, e.g., [31]. In our context, one may verify that the rational functions

$$\tilde{u}_n(z) = ||\mathscr{A}_{n+1}|| \cdot u_n(||\mathscr{A}_{n+1}|| \cdot z) = \frac{||\mathscr{A}_{n+1}|| \cdot q_n(||\mathscr{A}_{n+1}|| \cdot z)}{a_n \cdot q_{n+1}(||\mathscr{A}_{n+1}|| \cdot z)}$$

form a normal family in $|z| > 1$, which has at least one partial limit being different from the constant 0. Then the assertion of Theorem 3.3(b) follows by applying a criterion of Zalcman [60]. Indeed, the contracted zero distribution has proved to be very useful in describing properties of OPs for unbounded supports, and it seems to be interesting to explore the implications for complex Jacobi matrices and FOPs.

In the following statement we summarize some implications for the zeros of FOPs.

Theorem 3.4. (a) *There are no zeros of FOPs outside $\Gamma(A)$.*

(b) *Let Λ be some infinite set of integers such that $(a_n)_{n\in\Lambda}$ is bounded. Then for each closed $F \subset \Omega(A)$ there exists a $\delta = \delta(F)$ such that, for all $n \in \Lambda$, the zeros of q_n in F are at least at a distance δ from the zeros of q_{n+1} in F. If $\mathscr{A}$ is real, then $\Omega(A)$ is the largest open set with this property.*

(c) (*Cf. with Beckermann* [18, *Proposition* 2.1]). *Let Λ be some infinite set of integers such that $(a_n)_{n\in\Lambda}$ is bounded, and denote by Ω a connected component of $\Omega(A)$ which is not a subset of $\Gamma(A)$. Then for each closed $F \subset \Omega$ there exists a constant $v = v(F)$ such that, for all $n \in \Lambda$, the number of zeros of q_{n+1} in F (counting multiplicities) is bounded by $v(F)$. If $\mathscr{A}$ is real, then Ω is a maximal open connected set with this property.*

Proof. Part (a) follows immediately from Theorem 3.3(a) by observing that zeros of q_{n+1} are poles of u_n. In order to show part (b), recall from Theorem 3.3(c) that $(u_n)_{n\in\Lambda}$ is normal and thus equicontinuous in closed subsets of $\Omega(A)$. Given F as above, we can find a $\delta > 0$ such that $\chi(u_n(z'), u_n(z'')) \leqslant \frac{1}{2}$ for all $n \in \Lambda$ and for all $z', z'' \in F$ satisfying $|z' - z''| < \delta$. If now $z', z'' \in F$ with $q_n(z') = 0 = q_{n+1}(z'')$, then

$$\chi(u_n(z'), u_n(z'')) = \chi(0, \infty) = 1$$

and thus $|z' - z''| \geqslant \delta$, showing that the zeros in F of q_n and of q_{n+1} are separated.

Suppose now that $\mathscr{A}$ is real. Then, according to, e.g., Example 2.7, the corresponding difference operator A is self-adjoint, and the corresponding moment problem has a unique solution μ, with $\mathrm{supp}(\mu) = \sigma(A)$. It follows that, for any function f continuous on $\mathbb{R}$ with compact support, we have $I_n(f) \to \int f(x)\,\mathrm{d}\mu(x)$, where $I_n(\cdot)$ denotes the nth Gaussian quadrature rule. Given any $z_0 \notin \Omega(A)$ (i.e., $z_0 \in \mathrm{supp}(\mu)$) and $\delta > 0$, there exists a continuous function f with support in $U = (z_0 - \delta, z_0 + \delta)$ such that $\int f(x)\,\mathrm{d}\mu(x) > 0$. In particular, there exists some N such that $I_n(f) > 0$, $n \geqslant N$, showing that all polynomials q_n must have at least one zero in U. This terminates the proof of part (b).

If the assertion of part (c) is not true, then using Theorem 3.3(c) we may construct a closed set $F \subset \Omega$ and a subsequence $(v_n)_{n\geqslant 0}$ of $(u_n)_{n\in\Lambda}$, v_n having at least n poles in F, with $(v_n)_{n\geqslant 0}$ converging to some function v locally uniformly in Ω. Notice that v is meromorphic in Ω. From Theorem 3.3(a) we know that v is different from the constant ∞ in $\Omega \backslash \Gamma(A)$, and thus in Ω. Clearly, poles of (v_n) only accumulate in the set $F' := \{z \in F\colon |v(z)| \geqslant 2\}$, and thus we may suppose, without loss of generality, that there exists an open set $U \supset F$ with its closure U' contained in Ω such that $|v(z)| \geqslant 1$ for $z \in U'$, and $v(z) \neq \infty$ on the boundary $\partial U'$ of U'. As a consequence, for a sufficiently large N, the sequence $(1/v_n)_{n \geqslant N}$ consists of functions being analytic in U', and tends to $1/v$ uniformly in U' with respect to the Euclidean metric. Applying the principle of argument to the connected components of U, we may conclude that, for sufficiently large n, the number of poles of v_n in U' coincides with the number of poles of v in U'. Since the latter number is finite, we have a contradiction to the construction of v_n. A proof for the final sentence of part (c) follows the same lines as the second part of the proof of (b); we omit the details. □

Of course, for real Jacobi matrices, assertions related to Theorem 3.4 have been known before, see [4, Corollary 2] for part (b), and [55, Theorem 6.1.1] for part (c). Part (a) for complex Jacobi matrices has already been mentioned in [20, Theorem 3.10]. For complex bounded Jacobi matrices, $\Gamma(A)$ is bounded and contains $\sigma(A)$, and thus Ω necessarily coincides with the unbounded connected component $\Omega_0(A)$ of $\Omega(A)$. Consequently, for bounded A, part (c) gives a bound for the number of zeros of (all) FOPs in closed subsets of $\Omega_0(A)$, and this statement has already been established in [18, Proposition 2.1].

We terminate this section with a discussion of the closed convex set

$$\Gamma_{\mathrm{ess}}(A) = \bigcap_{k \geqslant 0} \Gamma(A^{(k)}),$$

where $A^{(k)}$ denotes the difference operator of the associated Jacobi matrix $\mathscr{A}^{(k)}$ introduced before Theorem 2.8, $A^{(0)} = A$. This set has been considered before in [13,14]. In the next statement we collect some properties of this set. Our main purpose is to generalize Theorems 3.3(a) and 3.4(a).

Theorem 3.5. (a) *There holds* $\Gamma_{\mathrm{ess}}(A) \subset \Gamma(A^{(k+1)}) \subset \Gamma(A^{(k)})$ *for all* $k \geqslant 0$, *and* $\Gamma_{\mathrm{ess}}(A) \neq \mathbb{C}$ *if and only if* $\Gamma(A) \neq \mathbb{C}$.

(b) *For any compact difference operator* B *we have* $\Gamma_{\mathrm{ess}}(A) = \Gamma_{\mathrm{ess}}(A + B)$.

(c) *Let* $\mathscr{A}$ *be proper. Then* $\sigma(A) \subset \Gamma(A)$ *and* $\sigma_{\mathrm{ess}}(A) \subset \Gamma_{\mathrm{ess}}(A)$. *Furthermore,* $\sigma(A) \backslash \Gamma_{\mathrm{ess}}(A)$ *consists of isolated points which accumulate only on* $\Gamma_{\mathrm{ess}}(A)$.

(d) *The sequence* $(u_n)_{n \geqslant 0}$ *of meromorphic functions is normal in* $\Omega(A) \backslash \Gamma_{\mathrm{ess}}(A)$, *and any partial limit is different from the constant* ∞.

(e) *For any compact subset* F *of* $\Omega(A) \backslash \Gamma_{\mathrm{ess}}(A)$ *there exists a constant* $N = N(F)$ *such that none of the FOPs* q_n *for* $n \geqslant N$ *has a zero in* F.

Proof. (a) The first inclusions follow immediately from the definition of the numerical range. It remains to discuss the case $\Gamma_{\mathrm{ess}}(A) \neq \mathbb{C}$. Then at least for one $k \geqslant 0$ we must have $\Gamma(A^{(k)}) \neq \mathbb{C}$. Since $\Gamma(A^{(k)})$ is convex, it must be contained in some half-plane. Furthermore, one easily checks that any $z \in \Theta(A)$ may be written as $z = z_1 + z_2$, with $z_2 \in \Gamma(A^{(k)})$ and $|z_1| \leqslant 2||\mathscr{A}_k||$. Thus $\Theta(A)$ and $\Gamma(A)$ are contained in some half-plane, and $\Gamma(A) \neq \mathbb{C}$.

(b) This assertion follows from the fact that any $z \in \Gamma(A^{(k)} + B^{(k)})$ may be written as $z = z_A + z_B$, with $z_A \in \Gamma(A^{(k)})$, $|z_B| \leqslant ||B^{(k)}||$, and $||B^{(k)}|| \to 0$.

(c) It is known [30, Theorem V.3.2] that, in connected components of $\mathbb{C} \backslash \Gamma(A)$, $\mathscr{R}(zI - A)$ is closed and $\dim \mathscr{N}(zI - A) = 0$. Since $\mathscr{A}$ is proper, it follows from Lemma 2.4(d) that $\mathscr{R}(zI - A)^{\perp} = \mathscr{N}((zI - A)^*) = \{0\}$, and thus $\mathscr{R}(zI - A) = \ell^2$. Consequently, $\mathbb{C} \backslash \Gamma(A) \subset \Omega(A)$. Also, it follows (implicitly) from [30, Theorem IV.5.35] that $\sigma_{\mathrm{ess}}(A) = \sigma_{\mathrm{ess}}(A^{(k)})$ for all $k \geqslant 0$, and $\sigma_{\mathrm{ess}}(A^{(k)}) \subset \Gamma(A^{(k)})$ by [30, Problem V.3.6]. Thus, we have also established the second inclusion $\sigma_{\mathrm{ess}}(A) \subset \Gamma_{\mathrm{ess}}(A)$.

In order to see the last sentence of part (c), denote by D a connected component of $\mathbb{C}\backslash\sigma_{\text{ess}}(A)$. From [30, Section IV.5.6] we know that either $D\subset\sigma(A)$, or the elements of $\sigma(A)$ in D are isolated and accumulate only in $\sigma_{\text{ess}}(A)\subset\Gamma_{\text{ess}}(A)$. If now $\sigma(A)\backslash\Gamma_{\text{ess}}(A)\subset\mathbb{C}\backslash\sigma_{\text{ess}}(A)$, there is nothing to show. Otherwise, suppose that D contains a point $z\in\sigma(A)\backslash\Gamma_{\text{ess}}(A)$. Then the assertion follows by showing that $D\not\subset\sigma(A)$. Indeed, we know from part (a) and the preceding paragraph that there exists a $\zeta\in\mathbb{C}\backslash\Gamma(A)\subset\Omega(A)$. By convexity of $\Gamma_{\text{ess}}(A)$, it follows that the segment $[z,\zeta]$ is a subset of $\mathbb{C}\backslash\Gamma_{\text{ess}}(A)\subset\mathbb{C}\backslash\sigma_{\text{ess}}(A)$. Hence $[z,\zeta]\subset D$, which implies that $D\not\subset\sigma(A)$.

(d), (e) We show in Corollary 4.4(a) below that for any compact subset F of $\Omega(A)\backslash\Gamma_{\text{ess}}(A)$ there exists a constant $N=N(F)$ such that

$$\sup_{n\geqslant N}\max_{z\in F}||(z\mathscr{I}_n-\mathscr{A}_n)^{-1}||<\infty.$$

Since $u_n(x)=(e_n,(z\mathscr{I}_{n+1}-\mathscr{A}_{n+1})^{-1}e_n)$, it follows that

$$\sup_{n\geqslant N}\max_{z\in F}|u_n(z)|<\infty.$$

Then assertions (d), (e) follow immediately. □

A particularly interesting case contained in Theorem 3.5 has been discussed by Barrios, López, Martínez and Torrano, see [13–16]: here $A=G+C$, where G is a self-adjoint difference operator (resulting from a real proper Jacobi matrix) and C is a compact complex difference operator. Then $\mathscr{A}$ is proper (and determinate), and

$$\sigma_{\text{ess}}(A)=\sigma_{\text{ess}}(G)\subset\Gamma_{\text{ess}}(A)=\Gamma_{\text{ess}}(G)\subset\Gamma(G)=\text{conv}(\sigma(G))\subset\mathbb{R}$$

by (2.12) and parts (a)–(c). Several of the results given in the present paper for general complex Jacobi matrices have been shown for the above class already earlier, see, e.g., [13, Lemmas 3,4; 14]. In particular, Theorem 3.5(e) for this class was established in [13, Corollary 1].

3.3. An open problem concerning zero-free regions

We have seen above that, for bounded operators A, the zeros of all FOPs are contained in the convex compact set $\Gamma(A)$, and most of them are "close" to the polynomial convex hull[9] of the spectrum $\sigma(A)$.

Let us have a closer look at an inverse problem: Suppose that c is some regular linear functional and Γ is some compact[10] convex set containing *all* zeros of *all* FOPs. Can we give some (spectral) properties of the underlying difference operator, or the sequence (u_n)?

Zero-free regions can be obtained from the recurrence relation, e.g., by applying techniques from continued fractions. There are, for instance, Cassini ovals [58, Corollary 4.1], or the Worpitski set (see [59, Theorem V.26.2; 20, Section 3.1]).

[9] Indeed, it is also unclear whether there is an example of a (complex) operator A where the number of zeros of FOPs in some compact subset of a bounded component of $\Omega(A)$ is unbounded.

[10] Example 3.2(a) of Stahl shows that there exist regular functionals induced by some measure on $[-1,1]$ where all but two zeros stay in $[-1,1]$, but the sequence of exceptional zeros is not bounded (and thus the underlying operator also is unbounded). Thus the restriction to bounded Γ seems to be natural.

A related question in Padé approximation has been discussed by Gonchar [28], who showed that the sequence of rational functions $(p_n/q_n)_{n\geqslant 0}$ converges locally uniformly in $\mathbb{C}\backslash\Gamma$ to some function f with a geometric rate. In other words, the absence of poles in sets (with a particular shape) is already sufficient to insure convergence of Padé approximants. Let us recall here some of his intermediate findings: writing a_n, b_n in terms of the coefficients of q_n/k_n (cf. with (2.27)), and taking into account that the zeros of q_n/k_n are bounded, one finds the relation (see [28, Proof of Proposition 4])

$$\sup_{n\geqslant 0}\frac{|a_n|}{n+1}<\infty, \quad \sup_{n\geqslant 0}\frac{|b_n|}{n+1}<\infty. \tag{3.10}$$

Notice that combining this result with Example 2.7 shows that the underlying complex Jacobi matrix is proper. A combination of [28, Propositions 2, 4] leads to the relations

$$\liminf_{n\to\infty}|q_n(z)|^{1/n}=\liminf_{n\to\infty}(|q_n(z)|^2+|a_nq_{n+1}(z)|^2)^{1/(2n)}\geqslant\exp(g_\Gamma(z)), \quad z\in\mathbb{C}\backslash\Gamma,$$

$$\kappa_{\sup}=\limsup_{n\to\infty}|a_0\cdot a_1\cdots a_{n-1}|^{1/n}\leqslant\mathrm{cap}(\Gamma),$$

$$\limsup_{n\to\infty}|\tilde{r}_n(z)|^{1/n}=\liminf_{n\to\infty}(|\tilde{r}_n(z)|^2+|a_n\tilde{r}_{n+1}(z)|^2)^{1/(2n)}\leqslant\exp(-g_\Gamma(z)), \quad z\in\mathbb{C}\backslash\Gamma,$$

where $\tilde{r}_n(z)=q_n(z)f(z)-p_n(z)$. Of course, in the case $\sigma(A)\subset\Gamma$, these relations (with $f(z)=\phi(z)$ and $\tilde{r}_n(z)=r_n(z)$) would follow from our Theorem 3.1. But this is exactly our problem: does it follow only from the knowledge about zeros of FOPs that $\sigma(A)\subset\Gamma$? Clearly, for real Jacobi matrices the answer is yes, but for complex Jacobi matrices?

Since an operator A with compact spectrum is necessarily bounded, a first step in this direction would be to sharpen (3.10) and to show that A is bounded. According to Theorem 3.3(b), this is equivalent to the fact that $(u_n)_{n\geqslant 0}$ (or $(z\cdot u_n)_{n\geqslant 0}$) is normal in some neighborhood of infinity.

Notice that $(z\cdot u_n)_{n\geqslant 0}$ does not take the values $0,\infty$ in $\bar{\mathbb{C}}\backslash\Gamma$. Moreover, by a theorem of Montel [43], any sequence of meromorphic functions which does not take three different values in some region D is normal. It would be interesting to know whether, for our particular sequence of (rational) functions, the information on the zeros of FOPs is already sufficient for normality.

Another interesting approach to our problem would be to impose in addition that A is bounded. If this implies $\sigma(A)\subset\Gamma$, then we would have at least a partial answer to the following problem raised by Aptekarev et al. [7]: does the convergence of the whole sequence of Padé approximants with a geometric rate at a fixed point z implies that $z\in\Omega(A)$?

3.4. Compact perturbations of Jacobi matrices and ratio asymptotics

An important element in the study of FOPs is the detection of so-called *spurious zeros* (or spurious poles in Padé approximation). We have seen in the preceding section that the absence of zeros in some region has some important consequences concerning, e.g., the convergence of Padé approximants. Roughly speaking, we call *spurious* the zeros of FOPs which are not related to the spectrum of the underlying difference operator. To give an example, consider a real Jacobi matrix induced by a measure supported on $[-2,-1]\cup[1,2]$ which is symmetric with respect to the origin. Then the zeros of the OPs q_{2n} lie all in the spectrum of A, and also $2n$ of the zeros of the OPs q_{2n+1} lie in the spectrum of A, but $q_{2n+1}(0)=0$ by symmetry.

We will not give a proper definition of spurious zero in the general case; see [50, Section 4] for a more detailed discussion. Here we will restrict ourselves to bounded complex Jacobi matrices: a sequence $(z_n)_{n\in\Lambda}$ is said to consist of spurious zeros if $q_n(z_n)=0$, $n\in\Lambda$, and $(z_n)_{n\in\Lambda}$ lies in some closed subset of the unbounded connected component $\Omega_0(A)$ of the resolvent set. Notice that $|z_n|\leqslant||A||$ by Theorem 3.4(a), implying that $(z_n)_{n\in\Lambda}$ remains in some compact subset of $\Omega_0(A)$. Therefore, we may (and will) assume that $(z_n)_{n\in\Lambda}$ converges to some $\zeta\in\Omega_0(A)$.

From Theorem 3.4(c) and the remarks after Theorem 3.4, we see that there are only “few” such spurious zeros, and that the set of their limits ζ coincides with the set of zeros (or poles) in $\Omega_0(A)$ of partial limits of the normal family (u_n). Also, $\zeta\in\sigma(A)\cup\Gamma_{\mathrm{ess}}(A)$ by Theorem 3.5(e).

One motivation for the considerations of this section is to show that the set of limits of spurious zeros remains invariant with respect to compact perturbations. This follows as a corollary to the following

Theorem 3.6. *Let $\mathscr{A}$, $\tilde{\mathscr{A}}$ be two complex Jacobi matrices with entries a_n, b_n, and $\tilde{a}_n, \tilde{b}_n$, respectively. Suppose that $\mathscr{A}, \tilde{\mathscr{A}}$ are bounded, and*[11] *that* $\arg(\tilde{a}_n/a_n)\in(-\pi/2,\pi/2]$ *for* $n\geqslant 0$.

Then the difference $A-\tilde{A}$ of the corresponding difference operators is compact if and only if

$$\lim_{n\to\infty}\chi(u_n,\tilde{u}_n)=0 \tag{3.11}$$

uniformly in closed subsets of $\Omega_0(A)\cap\Omega_0(\tilde{A})$.

Theorem 3.6 has been known before (at least partially) for real Jacobi matrices. Take as reference system the entries $\tilde{a}_n=a\neq 0$, $\tilde{b}=b$, $n\geqslant 0$. Then

$$\tilde{u}_n(z)=\frac{\tilde{q}_n(z)}{\tilde{a}_n\tilde{q}_{n+1}(z)}\to\frac{2}{z-b+\sqrt{(z-b)^2-4a^2}}$$

uniformly on closed subsets of $\mathbb{C}\backslash[b-2a,b+2a]=\mathbb{C}\backslash\sigma(\tilde{A})$ (we choose a branch of the square root such that the right-hand side vanishes at infinity). Thus Theorem 3.6 includes as a special case the well-known description of the Nevai–Blumenthal class $\mathscr{M}(a;b)$, see, e.g., [35]. This description is usually shown by applying the Poincaré Theorem, and a similar description is known for compact perturbations of (real) periodic Jacobi matrices (being considered more in detail in Section 4.3 below). Finally, Nevai and Van Assche [36] showed that a relation similar to (3.11) holds provided that $\tilde{\mathscr{A}}$ is a real compact perturbation of a real $\mathscr{A}$.

Before proving Theorem 3.6, let us motivate and state a related more general result. Given any (not necessarily regular) linear functional c acting on the space of polynomials, the (unique) *monic* FOPs Q_{n_j} corresponding to normal indices n_j together with some auxiliary monic polynomials Q_n, $n\neq n_j$ are known to satisfy a recurrence of the form

$$z\cdot Q_n(z)=Q_{n+1}(z)+\sum_{j=n-\gamma_n}^{n}b_{n,j}Q_j(z),\quad n\geqslant 0,\quad Q_0(z)=1, \tag{3.12}$$

[11] Such a normalization is known from orthogonal polynomials where usually $a_n,\tilde{a}_n>0$. It can be insured by possibly multiplying $\tilde{q}_n$ by -1.

where $b_{n,j}$ are some complex numbers, and the integer $\gamma_n \geqslant 0$ is bounded above by some multiple of the maximal distance of two successive normal indices. We may rewrite the recurrence formally as

$$0 = (z\mathscr{I} - \mathscr{B}) \cdot \begin{bmatrix} Q_0(z) \\ Q_1(z) \\ Q_2(z) \\ \vdots \end{bmatrix}, \quad \mathscr{B} = \begin{bmatrix} b_{0,0} & 1 & 0 & 0 & \cdots\cdots & \\ b_{1,0} & b_{1,1} & 1 & 0 & & \\ b_{2,0} & b_{2,1} & b_{2,2} & 1 & 0 & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}, \tag{3.13}$$

i.e., $\mathscr{B}$ is a lower Hessenberg matrix. If in addition the distance of two successive normal indices is uniformly bounded, then $\mathscr{B}$ is banded. This occurs, for instance, for symbols like $\sin(1/z)$, or for functionals c which are *asymptotically regular*, i.e., all sufficiently large indices are normal. Notice that, by (3.13), Q_n is the characteristic polynomial of the finite principal submatrix $\mathscr{B}_n$ of order n.

A class of asymptotically regular functionals was studied by Magnus [33], who considered c_w of (3.8) with a complex and continuous w which is different from 0 in $[-1, 1]$ (in fact, his class is larger). By, e.g., the Theorem of Rakhmanov, the real Jacobi matrix associated to $c_{|w|}$ is a compact perturbation of the Toeplitz operator having $\frac{1}{2}$ on the super- and the subdiagonal and 0 elsewhere. The functional c_w may not be regular, but is asymptotically regular by [33, Theorem 6.1(i)]. Therefore, the corresponding matrix $\mathscr{B}$ will in general not be tridiagonal, but is a compact perturbation of the Toeplitz operator having 1 on the super-, $\frac{1}{4}$ on the subdiagonal and 0 elsewhere (see [33, Theorem 6.1(iii)] and Theorem 3.7 below).

For regular functionals, recurrence (3.13) holds with

$$b_{n,n} = b_n, \qquad b_{n+1,n} = a_n^2, \qquad b_{k,n} = 0, \quad k - 1 > n \geqslant 0 \tag{3.14}$$

showing that $\mathscr{B}$ is bounded if and only if the corresponding Jacobi matrix is bounded. Recurrences of the above form are also valid for more general sequences of polynomials. For instance, for monic OPs with respect to the Hermitian scalar product

$$(f, g)_\mu = \int \overline{f(z)} g(z) \, \mathrm{d}\mu(z),$$

μ being some positive measure with compact infinite support, we always have a recurrence (3.12) with $b_{n,k} = (Q_k, zQ_n)_\mu / (Q_k, Q_k)_\mu$. We have the following complement to Theorem 3.6.

Theorem 3.7. *Let* $\mathscr{B}$ *be a tridiagonal matrix as in* (3.14), *with coefficients* $b_{n,k}$ *and associated monic FOPs* Q_n, *and let* $\tilde{\mathscr{B}}$ *be a lower Hessenberg matrix as in* (3.13) *with coefficients* $\tilde{b}_{n,k}$ *and associated polynomials* $\tilde{Q}_n$, *Provided that* $\mathscr{B}$ *and* $\tilde{\mathscr{B}}$ *are bounded, we have*

$$\lim_{n\to\infty} \left(\frac{Q_n(z)}{Q_{n+1}(z)} - \frac{\tilde{Q}_n(z)}{\tilde{Q}_{n+1}(z)} \right) = 0 \tag{3.15}$$

uniformly for $|z| \geqslant R$ *for sufficiently large* R *if and only if* [12]

$$\lim_{n\to\infty} (b_{n+j,n} - \tilde{b}_{n+j,n}) = 0, \quad j = 0, 1, 2, \ldots\ . \tag{3.16}$$

[12] If $\mathscr{B}$ is in addition banded, then this second condition is equivalent to the fact that $\mathscr{B} - \tilde{\mathscr{B}}$ is compact.

Proof. Suppose that (3.15) holds. Since $Q_n, \tilde{Q}_n$ are monic and of degree exactly n, we have the expansions

$$\frac{Q_n(z)}{Q_{n+1}(z)} = \sum_{j=0}^{\infty} \frac{u_{n,j}}{z^{j+1}}, \quad \frac{\tilde{Q}_n(z)}{\tilde{Q}_{n+1}(z)} = \sum_{j=0}^{\infty} \frac{\tilde{u}_{n,j}}{z^{j+1}},$$

where $u_{n,0} = \tilde{u}_{n,0} = 1$, and $u_{n,j} - \tilde{u}_{n,j}$ tends to zero for $n \to \infty$ for all fixed $j \geqslant 1$ by (3.15). From (3.13) we obtain

$$z\frac{Q_n(z)}{Q_{n+1}(z)} = 1 + b_{n,n}\frac{Q_n(z)}{Q_{n+1}(z)} + b_{n,n-1}\frac{Q_{n-1}(z)}{Q_n(z)}\frac{Q_n(z)}{Q_{n+1}(z)} + \cdots + b_{n,n-j}\prod_{\ell=0}^{j}\frac{Q_{n-\ell}(z)}{Q_{n+1-\ell}(z)} + \mathcal{O}\left(\frac{1}{z^{j+2}}\right)_{z\to\infty}$$

for any $n \geqslant j \geqslant 0$, and a similar equation for the quantities related to $\tilde{\mathscr{B}}$. Inserting the expansions at infinity and comparing coefficients leads to

$$u_{n,1} - \tilde{u}_{n,1} = b_{n,n} - \tilde{b}_{n,n} \to 0, \quad u_{n,2} - \tilde{u}_{n,2} = (b_{n,n-1} + b_{n,n}^2) - (\tilde{b}_{n,n-1} + \tilde{b}_{n,n}^2) \to 0$$

and similarly $u_{n,j+1} - \tilde{u}_{n,j+1} = b_{n,n-j} - \tilde{b}_{n,n-j} + C_{n,j} - \tilde{C}_{n,j}$ for $j \geqslant 2$, where $C_{n,j}$ is a polynomial expression of the quantities $b_{n-\ell,n-i}$ for $0 \leqslant \ell \leqslant i < j$, and $\tilde{C}_{n,j}$ is obtained from $C_{n,j}$ by replacing the quantities $b_{n-\ell,n-i}$ by $\tilde{b}_{n-\ell,n-i}$. One concludes by recurrence on j that the claimed limit relation (3.16) for the recurrence coefficients holds.

The other implication of Theorem 3.7 is slightly more involved. We choose

$$|z| \geqslant R := 2\max\{||\mathscr{B}||, ||\tilde{\mathscr{B}}||\}.$$

Then $|z| \geqslant 2\max\{||\mathscr{B}_n||, ||\tilde{\mathscr{B}}_n||\}$ for all n, implying that

$$||(z\mathscr{I}_n - \mathscr{B}_n)^{-1}|| \leqslant \frac{2}{|z|}, \quad ||(z\mathscr{I}_n - \tilde{\mathscr{B}}_n)^{-1}|| \leqslant \frac{2}{|z|}, \quad n \geqslant 0.$$

It follows from (3.13) that

$$(z\mathscr{I}_n - \tilde{\mathscr{B}}_n) \cdot (\tilde{Q}_0(z), \ldots, \tilde{Q}_{n-1}(z))^{\mathrm{T}} = (0, \ldots, 0, \tilde{Q}_n(z))^{\mathrm{T}},$$

and thus $\tilde{Q}_n(z)/\tilde{Q}_{n+1}(z) = (e_n, (z\mathscr{I}_{n+1} - \tilde{\mathscr{B}}_{n+1})^{-1}e_n)$, as well as

$$\sum_{j=0}^{n} |\tilde{Q}_j(z)|^2 \leqslant ||(z\mathscr{I}_{n+1} - \tilde{\mathscr{B}}_{n+1})^{-1}||^2 |\tilde{Q}_{n+1}(z)|^2 \leqslant \frac{4|\tilde{Q}_{n+1}(z)|^2}{|z|^2}.$$

From the latter relation one deduces by recurrence on $n - j$ that

$$|(e_j, (z\mathscr{I}_{n+1} - \tilde{\mathscr{B}}_{n+1})^{-1}e_n)|^2 = \left|\frac{\tilde{Q}_j(z)}{\tilde{Q}_{n+1}(z)}\right|^2 \leqslant \frac{4}{|z|^2(1 + |z|^2/4)^{n-j}}, \quad 0 \leqslant j \leqslant n. \tag{3.17}$$

We claim that also

$$|(e_n, (z\mathscr{I}_{n+1} - \mathscr{B}_{n+1})^{-1}e_j)|^2 \leqslant \frac{4}{|z|^2(1 + |z|^2/(4a^2))^{n-j}}, \quad 0 \leqslant j \leqslant n, \tag{3.18}$$

where $a = \max\{1, \sup|b_{n+1,n}|\} \leqslant ||\mathscr{B}|| < \infty$. This inequality is based on the observation that the polynomials $Q_n^L(z) := k_n q_n(z) = k_n^2 Q_n(z)$ satisfy

$$(Q_0^L(z), \ldots, Q_{n-1}^L(z)) \cdot (z\mathscr{I}_n - \mathscr{B}_n) = b_{n,n-1} \cdot (0, \ldots, 0, Q_n^L(z)).$$

Thus a proof for (3.18) follows the same lines as the proof of (3.17); we omit the details.

Given an $\varepsilon > 0$, by assumption (3.16) on the recurrence coefficients we may find an $L > 0$ and an $N > 0$ such

$$\lambda := (1 + R^2/(4a^2))^{-1/2} < \varepsilon^{1/L}, \quad \text{and} \quad |b_{n+\ell,n} - \tilde{b}_{n+\ell,n}| < \varepsilon, \quad n \geqslant N,\ \ell = 0,\dots,L.$$

For all other indices we have the trivial upper bound $|b_{n+\ell,n} - \tilde{b}_{n+\ell,n}| \leqslant (\|\mathscr{B}\| + \|\tilde{\mathscr{B}}\|) =: b$. Using (3.17), (3.18), we obtain for $|z| \geqslant R$, $n \geqslant N + L$,

$$\begin{aligned}\left|\frac{Q_n(z)}{Q_{n+1}(z)} - \frac{\tilde{Q}_n(z)}{\tilde{Q}_{n+1}(z)}\right| &= |(e_n, [(z\mathscr{I}_{n+1} - \mathscr{B}_{n+1})^{-1} - (z\mathscr{I}_{n+1} - \tilde{\mathscr{B}}_{n+1})^{-1}]e_n)| \\ &= |(e_n, (z\mathscr{I}_{n+1} - \mathscr{B}_{n+1})^{-1}(\tilde{\mathscr{B}}_{n+1} - \mathscr{B}_{n+1})(z\mathscr{I}_{n+1} - \tilde{\mathscr{B}}_{n+1})^{-1}e_n)| \\ &\leqslant \frac{2}{R}\sum_{j=0}^{n}\sum_{k=0}^{j}\lambda^{n-j+n-k}|b_{j,k} - \tilde{b}_{j,k}| \\ &\leqslant \frac{2b}{R}\sum_{j=0}^{n}\sum_{k=0}^{\min\{j,n-L\}}\lambda^{n-j+n-k} + \frac{2\varepsilon}{R}\sum_{j=n-L}^{n}\sum_{k=n-L}^{j}\lambda^{n-j+n-k} \leqslant \frac{2(b+1)}{R(1-\lambda)^2}\varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, we have established (3.15). Hence the second implication of Theorem 3.7 is shown. □

Proof of Theorem 3.6. We apply Theorem 3.7 with

$$\tilde{b}_{n,n} = \tilde{b}_n, \qquad \tilde{b}_{n+1,n} = \tilde{a}_n^2, \qquad \tilde{b}_{k,n} = 0, \quad k - 1 \geqslant n \geqslant 0.$$

Since $q_n/(a_n q_{n+1}) = Q_n/Q_{n+1}$ is bounded around infinity by Theorem 3.3(a), and similarly for the quantities with tildes, we see that (3.11) implies (3.15). In order to show that also the converse is true, suppose that (3.15) holds but not (3.11). Then there is some infinite set Λ and some $z_n \in \Omega_0(A) \cap \Omega_0(\tilde{A})$, $(z_n)_{n\in\Lambda}$ tending to some $\zeta \in \Omega_0(A) \cap \Omega_0(\tilde{A})$, such that $(\chi(u_n(z_n), \tilde{u}_n(z_n)))_{n\in\Lambda}$ does not converge to zero. Using the normality established in Theorem 3.3(c), we find a subset also denoted by Λ such that $(u_n)_{n\in\Lambda}$ (and $(\tilde{u}_n)_{n\in\Lambda}$ resp.) tends to some meromorphic function u (and $\tilde{u}$ resp.) locally uniformly in $\Omega_0(A)$ (and in $\Omega_0(\tilde{A})$ resp.). Notice that $u(\zeta) \neq \tilde{u}(\zeta)$ by construction of ζ, and $u = \tilde{u}$ in some neighborhood of infinity by (3.15), which is impossible for meromorphic functions. Hence (3.11) and (3.15) are equivalent.

Notice that (3.16) may be rewritten in our setting as $\tilde{b}_n - b_n \to 0$, and $\tilde{a}_n^2 - a_n^2 \to 0$. The normalization $\arg(\tilde{a}_n/a_n) \in (-\pi/2, \pi/2]$ of Theorem 3.6 allows us to conclude that $|a_n - \tilde{a}_n| \leqslant |a_n + \tilde{a}_n|$, showing that $(a_n^2 - \tilde{a}_n^2)_{n\geqslant 0}$ tends to zero if and only if $(a_n - \tilde{a}_n)_{n\geqslant 0}$ does. Thus $A - \tilde{A}$ is compact if and only if (3.16) holds, and Theorem 3.6 follows from Theorem 3.7. □

It is known for many examples (see, e.g., [50, Proposition 4.2]) that spurious poles of Padé approximants p_n/q_n are accompanied by a "close" zero. As a further consequence of Theorem 3.7, we can be more precise. In fact, consider $\tilde{\mathscr{B}}$ obtained from $\mathscr{B}$ by changing the values $\tilde{b}_{1,0} = 0$ and $\tilde{b}_{0,0} \in \sigma(A)$. Comparing with (1.2) one easily sees that $\tilde{Q}_n(z) = (z - \tilde{b}_{0,0})p_n(z)/k_n$, and as in the above proof it follows that

$$\chi\left(\frac{p_n}{a_n p_{n+1}}, \frac{q_n}{a_n q_{n+1}}\right) \to 0$$

locally uniformly in $\Omega_0(A)\cap\Omega_0([\tilde{\mathscr{B}}]_{\min})=\Omega_0(A)\cap\Omega_0(A^{(1)})$, which according to Theorem 2.8 coincides with $\Omega:=\{z\in\Omega_0(A)\colon\ \phi(z)\neq 0\}$. In particular, applying the argument principle, we may conclude that, for every sequence $(z_n)_{n\in\Lambda}$ tending to $\zeta\in\Omega_0(A)$ with $q_n(z_n)=0$, there exists a sequence $(z_n')_{n\in\Lambda}$ tending to ζ with $p_n(z_n')=0$.

3.5. Trace class perturbations and strong asymptotics

It is known for real Jacobi matrices [36] that if $A-\tilde{A}$ is not only compact but of trace class, then we may have a stronger form of convergence: A similar assertion is true for complex Jacobi matrices.

Theorem 3.8. *Let $\mathscr{A}$, $\tilde{\mathscr{A}}$ be two bounded complex Jacobi matrices. Provided that the difference $A-\tilde{A}$ of the corresponding difference operators is of trace class, i.e.,*

$$\sum_{n=0}^{\infty}(|a_n-\tilde{a}_n|+|b_n-\tilde{b}_n|)<\infty,$$

the corresponding monic FOPs satisfy

$$\lim_{n\to\infty}\frac{\tilde{Q}_n(z)}{Q_n(z)}=\det(I+(A-\tilde{A})(zI-A)^{-1})$$

uniformly on closed subsets of subdomains D of $\Omega_0(A)\cap\Omega_0(\tilde{A})$ which are (asymptotically) free of zeros of the FOPs q_n and $\tilde{q}_n$, $n\geqslant 0$.

Proof. Define the projections $E_n:\ell^2\to\mathbb{C}^n$ by $E_n(y_j)_{j\geqslant 0}=(y_j)_{0\leqslant j<n}$. We start by establishing for $z\in\Omega(A)$ the formula

$$E_n(zI-A)^{-1}E_n^*-(z\mathscr{I}_n-\mathscr{A}_n)^{-1}=(q_0(z),\ldots,q_{n-1}(z))^{\mathrm{T}}\frac{r_n(z)}{q_n(z)}(q_0(z),\ldots,q_{n-1}(z)).\tag{3.19}$$

Indeed, by (2.19),

$$\begin{aligned}I_n&=E_n(zI-A)(zI-A)^{-1}E_n^*\\&=E_n(zI-A)E_n^*E_n(zI-A)^{-1}E_n^*-(0,\ldots,0,a_{n-1})^{\mathrm{T}}r_n(z)(q_0(z),\ldots,q_{n-1}(z)).\end{aligned}$$

With $E_n(zI-A)E_n^*=z\mathscr{I}_n-\mathscr{A}_n$ and

$$(0,\ldots,0,a_{n-1})^{\mathrm{T}}=\frac{1}{q_n(z)}(z\mathscr{I}_n-\mathscr{A}_n)(q_0(z),\ldots,q_{n-1}(z))^{\mathrm{T}}$$

taken into account, identity (3.19) follows. In a similar way one obtains for $z\in\Omega(A)$, using (3.19),

$$\begin{aligned}&E_n(zI-\tilde{A})(zI-A)^{-1}E_n^*-(z\mathscr{I}_n-\tilde{\mathscr{A}}_n)(z\mathscr{I}_n-\mathscr{A}_n)^{-1}\\&\quad=(z\mathscr{I}_n-\tilde{\mathscr{A}}_n)[E_n(zI-A)^{-1}E_n^*-(z\mathscr{I}_n-\mathscr{A}_n)^{-1}]+E_n(zI-\tilde{A})(I-E_n^*E_n)(zI-A)^{-1}E_n^*\\&\quad=((z\mathscr{I}_n-\tilde{\mathscr{A}}_n)(q_0(z),\ldots,q_{n-1}(z))^{\mathrm{T}}\frac{r_n(z)}{q_n(z)}-(0,\ldots,0,\tilde{a}_{n-1})^{\mathrm{T}}r_n(z))(q_0(z),\ldots,q_{n-1}(z))\end{aligned}$$

$$=a_{n-1}r_n(z)q_n(z)(z\mathscr{I}_n-\tilde{\mathscr{A}}_n)\left(\frac{q_0(z)}{q_n(z)}-\frac{\tilde{q}_0(z)}{\tilde{q}_n(z)},\ldots,\frac{q_{n-1}(z)}{q_n(z)}-\frac{\tilde{q}_{n-1}(z)}{\tilde{q}_n(z)}\right)^{\mathrm{T}}$$

$$\times(0,\ldots,0,1)(z\mathscr{I}_n-\mathscr{A}_n)^{-1}.$$

Consequently,

$$\det(E_n(zI-\tilde{A})(zI-A)^{-1}E_n^*)$$

$$=\det((z\mathscr{I}_n-\tilde{\mathscr{A}}_n)(z\mathscr{I}_n-\mathscr{A}_n)^{-1})$$

$$\det(I_n-a_{n-1}r_n(z)q_n(z)\left(\frac{q_0(z)}{q_n(z)}-\frac{\tilde{q}_0(z)}{\tilde{q}_n(z)},\ldots,\frac{q_{n-1}(z)}{q_n(z)}-\frac{\tilde{q}_{n-1}(z)}{\tilde{q}_n(z)}\right)^{\mathrm{T}}(0,\ldots,0,1))$$

$$=\frac{\tilde{Q}_n(z)}{Q_n(z)}\left[1-a_{n-1}r_n(z)q_n(z)\left(\frac{q_{n-1}(z)}{q_n(z)}-\frac{\tilde{q}_{n-1}(z)}{\tilde{q}_n(z)}\right)\right]. \tag{3.20}$$

Using the projector $\Pi_n=E_n^*E_n$ introduced in Section 2, we may write the term on the left-hand side as

$$\det(E_n(zI-\tilde{A})(zI-A)^{-1}E_n^*)=\det(I_n+E_n(A-\tilde{A})(zI-A)^{-1}E_n^*)$$
$$=\det(I+\Pi_n(A-\tilde{A})(zI-A)^{-1}),$$

where the term of the right is the determinant of a finite-rank perturbation of the identity; see, e.g., [30, Section III.4.3]. Since $A-\tilde{A}$ is a trace class operator, the same is true for $(A-\tilde{A})(zI-A)^{-1}$ and thus

$$\lim_{n\to\infty}\det(I+\Pi y_n(A-\tilde{A})(zI-A)^{-1})=\det(I+(A-\tilde{A})(zI-A)^{-1})$$

uniformly in closed subsets of $\Omega(A)$. It remains to see whether the term in brackets on the right-hand side of (3.20) tends to 1. Let F be some closed subset of the zero-free region $D\subset\Omega:=\Omega_0(A)\cap\Omega_0(\tilde{A})$. According to Theorem 3.3(c), both (u_n) and $(\tilde{u}_n)$ are normal families of meromorphic functions in Ω, and the functions are analytic in the subdomain D. Furthermore, we know from Theorem 3.3 that any partial limit is different from the constant infinity. It is known (see, e.g., [18, Lemma 2.4(d)]) that then (u_n) and $(\tilde{u}_n)$ are bounded on F uniformly in n. Combining this with Theorem 3.6, we find that $|u_n-\tilde{u}_n|\to 0$ uniformly in F, and

$$\max_{z\in F}\left|\frac{q_{n-1}(z)}{q_n(z)}-\frac{\tilde{q}_{n-1}(z)}{\tilde{q}_n(z)}\right|\leqslant|a_{n-1}|\max_{z\in F}|u_{n-1}(z)-\tilde{u}_{n-1}(z)|+|a_{n-1}-\tilde{a}_{n-1}|\max_{z\in F}|\tilde{u}_{n-1}(z)|$$

tends to zero for $n\to\infty$. Moreover, the remaining term $a_{n-1}r_n(z)q_n(z)$ is bounded uniformly for $z\in F$ and $n\geqslant 0$ according to (2.23). This terminates the proof of Theorem 3.8. □

We conclude this section with some general remarks concerning the strong asymptotics

$$\max_{z\in U}\left|\frac{\tilde{Q}_n(z)}{Q_n(z)}-g(z)\right|=0 \quad \text{where } U \text{ is some closed disk around } \infty.$$

Indeed, by examining the proof, we see that this assertion is true also for the more general matrices $\mathscr{B},\tilde{\mathscr{B}}$ of Theorem 3.7 provided that $\mathscr{B}-\tilde{\mathscr{B}}$ is of trace class. Finally, already from the real case it is known that for this form of strong asymptotics it is necessary that $\mathscr{A}-\tilde{\mathscr{A}}$ is compact, but it does not need to be of trace class. Indeed, a necessary and sufficient condition seems to be that $\mathscr{A}-\tilde{\mathscr{A}}$ is compact, and that

$$\sum_{j=0}^{n-1}[r_j(z)q_j(z)-\tilde{r}_j(z)\tilde{q}_j(z)]=\operatorname{trace}\Pi_n[(zI-A)^{-1}-(zI-\tilde{A})^{-1}]\Pi_n$$

converges for $n\to\infty$ uniformly in U (to g'/g). It would be very interesting to explore the connection to some complex counterpart of the Szegő condition.

4. Approximation of the resolvent and the Weyl function

The goal of this section is to investigate the question whether we can approximate the resolvent $(zI-A)^{-1}$ by means of inverses $(z\mathscr{I}_n-\mathscr{A}_n)^{-1}$ of finite sections of $z\mathscr{I}-\mathscr{A}$. This question is of interest, e.g., for discrete Sturm–Liouville problems on the semiaxis: for solving in ℓ^2 the equation $(zI-A)y=f$ for given $f\in\ell^2$ via a projection method, one considers instead the finite-dimensional problems $(z\mathscr{I}_n-\mathscr{A}_n)y^{(n)}=E_nf$.

Another motivation comes from convergence questions for Padé approximation and continued fractions: With p_n,q_n as in (1.2), (1.4) we define the rational function

$$\pi_n(z)=\frac{p_n(z)}{q_n(z)}=(e_0,(z\mathscr{I}_n-\mathscr{A}_n)^{-1}e_0).$$

It is known [59] that $\pi_n(z)$ has the J-fraction expansion

$$\pi_n(z)=\frac{1|}{|z-b_0}+\frac{-a_0^2|}{|z-b_1}+\frac{-a_1^2|}{|z-b_2}+\frac{-a_2^2|}{|z-b_3}+\cdots+\frac{-a_{n-2}^2|}{|z-b_{n-1}}$$

being the nth convergent of the J-fraction (1.3). In addition, the (formal) expansion at infinity of this J-fraction is known to coincide with (2.18), and one also shows that π_n is its nth Padé approximant (at infinity). The question is whether we can expect the convergence of $\pi_n(z)=(e_0,(z\mathscr{I}_n-\mathscr{A}_n)^{-1}e_0)$ to the Weyl function $\phi(z)=(e_0,(zI-A)^{-1}e_0)$.

This question has been studied by means of operators by many authors, see [59, Section 26; 7, 15, 16, 18, 20] for bounded A and [13,14] for bounded perturbations of possibly unbounded self-adjoint A. Our aim is to show that most of these results about the approximation of the Weyl function are in fact results about the approximation of the resolvent $(zI-A)^{-1}$ by $(z\mathscr{I}_n-\mathscr{A}_n)^{-1}$.

4.1. Approximation of the resolvent

Different kinds of resolvent convergence may be considered for $z\in\Omega(A)$, for instance *norm convergence*

$$\lim_{\substack{n\to\infty\\ n\in\Lambda}}||(z\mathscr{I}_n-\mathscr{A}_n)^{-1}-E_n(zI-A)^{-1}E_n^*||=0, \tag{4.1}$$

strong resolvent convergence

$$\lim_{\substack{n\to\infty\\ n\in\Lambda}} E_n^*(z\mathscr{I}_n - \mathscr{A}_n)^{-1}E_n y = (zI - A)^{-1}y \quad \forall y \in \ell^2, \tag{4.2}$$

or *weak resolvent convergence*

$$\lim_{\substack{n\to\infty\\ n\in\Lambda}} (E_n y', (z\mathscr{I}_n - \mathscr{A}_n)^{-1}E_n y) = (y', (zI - A)^{-1}y) \quad \forall y, y' \in \ell^2. \tag{4.3}$$

The interested reader may easily check that (4.1) implies (4.2), and the latter implies (4.3). Notice also that (pointwise) convergence results for Padé approximation of the Weyl function are obtained by choosing in (4.3) the vectors $y = y' = e_0$. In all these forms of convergence we assume implicitly that $z\mathscr{I}_n - \mathscr{A}_n$ is invertible for (sufficiently large) $n \in \Lambda$. We also mention the related condition

$$\limsup_{\substack{n\to\infty\\ n\in\Lambda}} ||(z\mathscr{I}_n - \mathscr{A}_n)^{-1}|| =: C < \infty. \tag{4.4}$$

A Kantorovich-type theorem gives connections between properties (4.2) and (4.4). For complex (possibly unbounded) Jacobi matrices we have the following result.

Theorem 4.1. *Let A be a difference operator resulting from a complex Jacobi matrix, Λ some infinite set of integers, and $z \in \mathbb{C}$. The following assertions are equivalent*:
(a) *$z \in \Omega(A)$, and* (4.2) *holds.*
(b) *$z \in \Omega(A)$, and* (4.3) *holds.*
(c) *$\mathscr{A}$ is proper, and* (4.4) *holds.*
In addition, if property (*c*) *holds for some $z = z_0$, then the limit relations* (4.2), (4.3) *take place uniformly for $|z - z_0| \leqslant 1/(2C)$.*

Proof. Trivially, (b) follows from (a). Also, $\Omega(A) \neq \emptyset$ implies that $\mathscr{A}$ is proper by Theorem 2.6. In addition, (b) only makes sense if $z\mathscr{I}_n - \mathscr{A}_n$ is invertible for sufficiently large $n \in \Lambda$. Furthermore, a sequence of weakly converging bounded linear operators is necessarily uniformly bounded, see, e.g., [30, Section III.3.1]. Thus (b) implies (c).

Suppose now that (c) holds. By possibly dropping some elements from Λ we may replace condition (4.4) by

$$\sup_{n\in\Lambda} ||(z\mathscr{I}_n - \mathscr{A}_n)^{-1}|| \leqslant C' := 3C/2 < \infty. \tag{4.5}$$

For any $y \in \mathscr{C}_0$, say, $y = \Pi_k y$, we find an index $n \in \Lambda$, $n > k$, with

$$||(zI - A)y|| = ||E_{k+1}(zI - A)\Pi_k y|| = ||(z\mathscr{I}_n - \mathscr{A}_n)E_n y|| \geqslant \frac{||E_n y||}{||(z\mathscr{I}_n - \mathscr{A}_n)^{-1}||} \geqslant \frac{||y||}{C'}.$$

As in the second part of the proof of Theorem 2.10 we obtain

$$\inf_{y\in\mathscr{D}(A)} \frac{||(zI - A)y||}{||y||} = \inf_{y\in\mathscr{C}_0} \frac{||(zI - A)y||}{||y||} \geqslant \frac{1}{C'} > 0.$$

Consequently, $\mathscr{N}(zI-A)=\{0\}$. Since $\mathscr{A}$ is proper, it follows from Lemma 2.4(a),(d) that $\mathscr{N}((zI-A)^*)=\{0\}$. Furthermore, by [30, Theorem IV.5.2], $\mathscr{R}(zI-A)$ is closed. Since its orthogonal complement is given by $\mathscr{N}((zI-A)^*)$, we may conclude that $\mathscr{R}(zI-A)=\ell^2$, and thus $z\in\Omega(A)$.

In order to show the second part of (a), let $y\in\ell^2=\mathscr{R}(zI-A)$, and $x\in\mathscr{D}(A)$ with $(zI-A)x=y$. Let $\varepsilon>0$. By (2.2), we find $\tilde{x}\in\mathscr{C}_0$, $\tilde{y}=(zI-A)\tilde{x}$, such that

$$C'\|y-\tilde{y}\|\leqslant\varepsilon/3, \quad \text{and} \quad \|x-\tilde{x}\|\leqslant\varepsilon/3.$$

Also, since $y\in\ell^2$ and $\tilde{x}\in\mathscr{C}_0$, we find an $N\geqslant 0$ such that

$$\Pi_n\tilde{x}=\tilde{x}, \quad \text{and} \quad C'\|(I-\Pi_n)y\|\leqslant\varepsilon/3, \quad n\geqslant N.$$

Recalling that $E_nE_n^*=\mathscr{I}_n$ and $E_n^*E_n=\Pi_n$, we obtain

$$\begin{aligned}\|E_n^*(z\mathscr{I}_n-\mathscr{A}_n)^{-1}E_ny-(zI-A)^{-1}y\| &\leqslant \|E_n^*(z\mathscr{I}_n-\mathscr{A}_n)^{-1}E_n[y-E_n^*(z\mathscr{I}_n-\mathscr{A}_n)E_n\tilde{x}]\|\\ &\quad+\|E_n^*(z\mathscr{I}_n-\mathscr{A}_n)^{-1}E_nE_n^*(z\mathscr{I}_n-\mathscr{A}_n)E_n\tilde{x}-x\|\\ &\leqslant C'\|y-\Pi_n(zI-A)\Pi_n\tilde{x}\|+\|\Pi_n\tilde{x}-x\|\\ &= C'\|y-\Pi_n(zI-A)\tilde{x}\|+\|\tilde{x}-x\|\\ &\leqslant C'(\|(I-\Pi_n)y\|+\|\Pi_n(y-\tilde{y})\|)+\|\tilde{x}-x\|\leqslant\varepsilon\end{aligned}$$

for all $n\geqslant N$, $n\in\Lambda$, and thus (4.2) holds.

It remains to show the last sentence. We first mention that if $z_0\in\mathbb{C}$ satisfies (4.5), then for any z with $|z-z_0|\leqslant\varepsilon\leqslant 1/(2C)$ and for any $n\in\Lambda$ there holds

$$\|(z\mathscr{I}_n-\mathscr{A}_n)^{-1}\|\leqslant\|(z_0\mathscr{I}_n-\mathscr{A}_n)^{-1}\|\cdot\|(\mathscr{I}_n+(z-z_0)(z_0\mathscr{I}_n-\mathscr{A}_n)^{-1})^{-1}\|\leqslant 4C'=6C \tag{4.6}$$

and

$$\|(z\mathscr{I}_n-\mathscr{A}_n)^{-1}-(z_0\mathscr{I}_n-\mathscr{A}_n)^{-1}\|=|z-z_0|\|(z\mathscr{I}_n-\mathscr{A}_n)^{-1}(z_0\mathscr{I}_n-\mathscr{A}_n)^{-1}\|\leqslant\varepsilon\cdot 9C^2.$$

The same estimates are obtained for the resolvent. Thus, given $\varepsilon>0$ and $y\in\ell^2$, we may cover $U:=\{z\in\mathbb{C}\colon |z-z_0|<1/(2C)\}$ by a finite number of closed disks of radius $\varepsilon'\leqslant\varepsilon/(9C^2\cdot\|y\|)$ centred at $z_1,\ldots,z_K\in U$, and find an N such that

$$\|E_n^*(z_k\mathscr{I}_n-\mathscr{A}_n)^{-1}E_ny-(z_kI-A)^{-1}y\|<\varepsilon, \quad n\in\Lambda,\ n\geqslant N,\ k=1,\ldots,K.$$

Then for each $z\in U$ we find a k with $|z-z_k|\leqslant\varepsilon'$, and

$$\begin{aligned}&\|E_n^*(z\mathscr{I}_n-\mathscr{A}_n)^{-1}E_ny-(zI-A)^{-1}y\|\\ &\quad\leqslant 2\varepsilon'\cdot(9C^2)\cdot\|y\|+\|E_n^*(z_k\mathscr{I}_n-\mathscr{A}_n)^{-1}E_ny-(z_kI-A)^{-1}y\|\leqslant 3\varepsilon\end{aligned}$$

for all $n\geqslant N$, $n\in\Lambda$, showing that the convergence in (4.2) (and thus in (4.3)) takes place uniformly in U. □

Different variants of the Kantorovich Theorem have been discussed before in the context of FOPs and Padé approximation, see [33, Theorems 4.1, 4.2] or [14, Lemmas 4, 5]. Usually, the condition $z\in\Omega(A)$ is imposed for all equivalences; then the proof simplifies considerably, and also applies to general proper matrices.

We see from Theorem 4.1 that the notion of weak and strong resolvent convergence are equivalent for proper complex Jacobi matrices. On the other hand, by (3.19),

$$||(z\mathscr{I}_n-\mathscr{A}_n)^{-1}-E_n(zI-A)^{-1}E_n^*||=\left|\frac{r_n(z)}{q_n(z)}\right|\sum_{j=0}^{n-1}|q_j(z)|^2=|\phi(z)-\pi_n(z)|\sum_{j=0}^{n-1}|q_j(z)|^2, \tag{4.7}$$

and at least for particular examples it is known that the right-hand side of (4.7) does not tend to zero. Thus we may not expect to have norm convergence.

If $\mathscr{A}$ is not proper, then Theorem 4.1 does not give any information (notice that $\Omega(A)=\emptyset$ by Theorem 2.6(c)). However, at least in the indeterminate case we clearly understand what happens.

Theorem 4.2. *Let $\mathscr{A}$ be indeterminate. If Λ is some infinite set of integers and $\zeta\in\mathbb{C}$ is such that*

$$\lim_{\substack{n\to\infty\\ n\in\Lambda}}(e_0,(\zeta\mathscr{I}_n-\mathscr{A}_n)^{-1}e_0)=:\pi, \tag{4.8}$$

then with the unique $\eta\in\mathbb{C}\cup\{\infty\}$ satisfying $\phi_{[\eta]}(\zeta)=\pi$ (see Theorem 2.11) there holds

$$\lim_{\substack{n\to\infty\\ n\in\Lambda}}||(z\mathscr{I}_n-\mathscr{A}_n)^{-1}-E_n(zI-A_{[\eta]})^{-1}E_n^*||=0$$

uniformly on compact subsets of $\Omega(A_{[\eta]})$.

Proof. We will only show pointwise norm convergence for $z\in\Omega(A_{[\eta]})$, the extension to uniform convergence follows as in the proof of Theorem 4.1. First one shows as in (3.19) and (4.7) that

$$||(z\mathscr{I}_n-\mathscr{A}_n)^{-1}-E_n(zI-A_{[\eta]})^{-1}E_n^*||=|\phi_{[\eta]}(z)-\pi_n(z)|\sum_{j=0}^{n-1}|q_j(z)|^2,\quad z\in\Omega(A_{[\eta]}).$$

Since $\mathscr{A}$ is indeterminate, the sum is bounded uniformly in n for all $z\in\mathbb{C}$, and $\phi_{[\eta]}(z)\neq\infty$. Therefore, it remains only to show that $\pi_n(\zeta)\to\pi$ for $n\to\infty$, $n\in\Lambda$ implies $\pi_n(z)\to\phi_{[\eta]}(z)$ for $n\to\infty$, $n\in\Lambda$ and $z\in\mathbb{C}$. Here we follow [59, Proof of Theorem 23.2]: According to [59, Theorem 23.1, Eqs. (23.2), (23.6)], there exist polynomials $a_{j,n}$, $j=1,2,3,4$ with

$$\lim_{n\to\infty}a_{j,n}(z)=a_j(z),\quad j=1,2,3,4,\ z\in\mathbb{C}, \tag{4.9}$$

$$a_{1,n}(z)a_{4,n}(z)-a_{2,n}(z)a_{3,n}(z)=1,\quad n\geqslant 0,\ z\in\mathbb{C}, \tag{4.10}$$

$$p_n(z)=p_n(0)a_{2,n}(z)-q_n(0)a_{1,n}(z),\quad q_n(z)=p_n(0)a_{4,n}(z)-q_n(0)a_{3,n}(z) \tag{4.11}$$

with $a_1,\dots,a_4$ as in Theorem 2.11. Combining (4.10) and (4.11), we get

$$p_n(0)=-p_n(\zeta)a_{3,n}(\zeta)+q_n(\zeta)a_{1,n}(\zeta),\quad q_n(0)=-p_n(\zeta)a_{4,n}(\zeta)+q_n(\zeta)a_{2,n}(\zeta),$$

and, by assumption on $\pi_n(\zeta)=p_n(\zeta)/q_n(\zeta)$, we may conclude from (4.9) that

$$\lim_{\substack{n\to\infty\\ n\in\Lambda}}\pi_n(0)=\frac{a_1(\zeta)-\pi a_3(\zeta)}{a_2(\zeta)-\pi a_4(\zeta)}.$$

Here the right-hand side equals η by definition. Applying again (4.9) and (4.11), we obtain for $z \in \mathbb{C}$ the desired relation

$$\lim_{\substack{n\to\infty \\ n\in\Lambda}} \pi_n(z) = \frac{a_1(z) - \eta a_2(z)}{a_3(z) - \eta a_4(z)} = \phi_{[\eta]}(z). \qquad \square$$

Theorem 4.2 implies that in the indeterminate case we obtain weak and strong resolvent convergence to $(zI - A_{[\eta]})^{-1}$. Since (4.8) follows from weak convergence, we may conclude that here all three notions of convergence are equivalent. Notice that (4.8) is equivalent to the convergence of a subsequence of Padé approximants at one point.

Let us return to the more interesting case of proper complex Jacobi matrices. In order to be able to exploit Theorem 4.1, we need to know whether there exist infinite sets Λ (possibly depending on z) satisfying (4.4). In the following theorem we show that, under some additional assumptions, the existence can be insured.

Theorem 4.3. (a) *Suppose that the infinite sequence $(a_n)_{n\in\Lambda'}$ is bounded. Then $z \in \Omega(A)$ if and only if there exists an infinite set of integers Λ satisfying* (4.4).

(b) *Suppose that $(a_{n-1})_{n\in\Lambda}$ is bounded, and let $z \in \Omega$, where Ω is a connected component of $\Omega(A)$ which is not a subset of $\Gamma(A)$. Then* (4.4) *holds if and only if z is not an accumulation point of $\{zeros\ of\ q_n\colon n \in \Lambda\}$.*

(c) *Suppose that $(a_{n-1})_{n\in\Lambda}$ tends to zero. Then $z \in \Omega(A)$ if and only if* (4.4) *holds.*

(d) *Let $A,\tilde{A}$ be two difference operators with compact $A - \tilde{A}$, and $z \in \Omega(A) \cap \Omega(\tilde{A})$. Then* (4.4) *for A implies* (4.4) *for $\tilde{A}$.*

(e) *Relation* (4.4) *with $\Lambda = \{0, 1, 2, \ldots\}$ holds for $z \in \Omega(A) \backslash \Gamma_{\text{ess}}(A)$.*

It seems that the assertions of Theorem 4.3 have gone unnoticed so far for general possibly unbounded complex Jacobi matrices. For bounded or compact perturbations of self-adjoint Jacobi matrices, results related to Theorem 4.3(e) may be found in [14, Sections 1, 2].

Combining Theorem 4.3 with Theorem 4.1 (specially the last sentence) and using classical compactness arguments, we may get uniform counterparts of (4.1) and (4.4). Since these results play an important role for the convergence of Padé approximants, we state them explicitly in

Corollary 4.4. *We have*

$$\limsup_{\substack{n\to\infty \\ n\in\Lambda}} \max_{z\in F} ||(z\mathscr{I}_n - \mathscr{A}_n)^{-1}|| < \infty$$

and

$$\limsup_{\substack{n\to\infty \\ n\in\Lambda}} \max_{z\in F} ||E_n^*(z\mathscr{I}_n - \mathscr{A}_n)^{-1}E_n y - (zI - A)^{-1}y|| = 0$$

for a compact set F and $y \in \ell^2$ provided that one of the following conditions is satisfied:

(a) $\Lambda = \{0, 1, 2, \ldots\}$ *and* $F \subset \Omega(A) \backslash \Gamma_{\text{ess}}(A)$.

(b) $(a_{n-1})_{n\in\Lambda}$ *tends to zero and* $F\subset\Omega(A)$.
(c) $(a_{n-1})_{n\in\Lambda}$ *is bounded*, $F\subset\Omega$, *with* $\Omega\not\subset\Gamma(A)$ *being some subdomain of* $\Omega(A)$, *and* F *does not contain accumulation points of zeros of* q_n, $n\in\Lambda$.

For the proof of Theorem 4.3(d),(e) we will need the following lemma, which for bounded operators was already stated before by Magnus [33].

Lemma 4.5 (Cf. with Magnus [33, Theorem 4.4]). *Let* $\mathscr{B}$ *be some infinite proper matrix, and write* $B=[\mathscr{B}]_{\min}$. *Furthermore, let* $\tilde{B}$ *be an operator in* ℓ^2, *with* $\mathscr{C}_0\subset\mathscr{D}(\tilde{B})$, $0\in\Omega(B)\cap\Omega(\tilde{B})$, *and* $B-\tilde{B}$ *being compact. Then for any infinite set of integers* Λ *we have the implication*

$$\sup_{n\in\Lambda}||(E_nBE_n^*)^{-1}||=C'<\infty\Rightarrow\limsup_{\substack{n\to\infty\\ n\in\Lambda}}||(E_n\tilde{B}E_n^*)^{-1}||<\infty.$$

Proof. We claim that there exist N,C such that, for all $n\geqslant N$, $n\in\Lambda$, the system

$$(E_n\tilde{B}E_n^*)x_n=y_n$$

admits a unique solution x_n for all $y_n\in\mathbb{C}^n$, with $||x_n||\leqslant C\cdot||y_n||$. Then the assertion follows. For proving this claim, we rewrite the system as

$$[\mathscr{I}_n+(E_nBE_n^*)^{-1}E_n(\tilde{B}-B)E_n^*]x_n=(E_nBE_n^*)^{-1}y_n.$$

Since $E_nE_n^*=\mathscr{I}_n$, $E_n^*\mathscr{I}_n=E_n^*$, $E_n^*E_n=\Pi_n$, the system takes the form

$$[I+B^{-1}(\tilde{B}-B)+\Delta_n](E_n^*x_n)=E_n^*(E_nBE_n^*)^{-1}y_n,\tag{4.12}$$

where

$$\begin{aligned}\Delta_n&=E_n^*(E_nBE_n^*)^{-1}E_n(\tilde{B}-B)-B^{-1}(\tilde{B}-B)\\&=[E_n^*(E_nBE_n^*)^{-1}E_n-B^{-1}](I-\Pi_m)(\tilde{B}-B)+[E_n^*(E_nBE_n^*)^{-1}E_n-B^{-1}]\Pi_m(\tilde{B}-B)\end{aligned}$$

for any integer m. Here the expression in brackets is bounded in norm by $C'+||B^{-1}||$. Since $(B-\tilde{B})$ is compact, it is known that $||(I-\Pi_m)(\tilde{B}-B)||\to 0$ for $m\to\infty$. Hence we may find an m such that

$$||(I-\Pi_m)(\tilde{B}-B)||\leqslant\frac{1}{4||\tilde{B}^{-1}B||(C'+||B^{-1}||)}.$$

Since $\mathscr{B}$ is proper, with $0\in\Omega(B)$, one shows as in the proof of Theorem 4.1 that

$$\lim_{\substack{n\to\infty\\ n\in\Lambda}}E_n^*(E_nBE_n^*)^{-1}E_ny=B^{-1}y,\quad y\in\ell^2.$$

In particular, we may find for $\varepsilon:=1/(4\sqrt{m}||\tilde{B}^{-1}B||\cdot||E_m(\tilde{B}-B)||)$ an $N\geqslant m$ such that

$$||[E_n^*(E_nBE_n^*)^{-1}E_n-B^{-1}]e_j||\leqslant\varepsilon,\quad j=0,\ldots,m-1,\ n\geqslant N,\ n\in\Lambda,$$

implying that $||[E_n^*(E_nBE_n^*)^{-1}E_n - B^{-1}]E_m^*|| \leqslant \sqrt{m}\varepsilon$. Collecting the individual terms, we may conclude that $||\Delta_n|| \leqslant 1/(2||\tilde{B}^{-1}B||)$. In particular, $B^{-1}\tilde{B} + \Delta_n$ is invertible, with its inverse having a norm bounded by $2||\tilde{B}^{-1}B||$. Thus, system (4.12) has a unique solution for $n \geqslant N$, $n \in \Lambda$, with

$$||x_n|| \leqslant 2||\tilde{B}^{-1}B||\,||E_n^*(E_nBE_n^*)^{-1}||\,||y_n|| \leqslant 2||\tilde{B}^{-1}B||C'||y_n||,$$

as claimed above. □

Proof of Theorem 4.3. (a) By Example 2.7, $\mathscr{A}$ is proper. Thus (4.4) implies that $z \in \Omega$. In order to show the converse, let $z \in \Omega(A)$. According to (4.7), it will be sufficient to give a suitable error estimate for the error of Padé approximation. For $n \in \Lambda'$, define $\varepsilon_n = 1$ if $|u_n(z)| \leqslant 1$, and $\varepsilon_n = 0$ otherwise. Furthermore, let $\Lambda = \{n + \varepsilon_n : n \in \Lambda'\}$. Then we get for $n \in \Lambda'$

$$|a_n|^{2\varepsilon_n}|q_{n+\varepsilon_n}(z)|^2 \geqslant \tfrac{1}{2}(|q_n(z)|^2 + |a_nq_{n+1}(z)|^2)$$

by construction of ε_n, and trivially

$$|a_n|^{2\varepsilon_n}|r_{n+\varepsilon_n}(z)|^2 \leqslant |r_n(z)|^2 + |a_nr_{n+1}(z)|^2.$$

Using the left-hand estimate of (2.22), we may conclude that

$$\begin{aligned}\left|\frac{r_{n+\varepsilon_n}(z)}{q_{n+\varepsilon_n}(z)}\right| \sum_{j=0}^{n} |q_j(z)|^2 &\leqslant \sqrt{2\frac{|r_n(z)|^2 + |a_nr_{n+1}(z)|^2}{|q_n(z)|^2 + |a_nq_{n+1}(z)|^2}} \sum_{j=0}^{n} |q_j(z)|^2 \\ &\leqslant \sqrt{2}(|r_n(z)|^2 + |a_nr_{n+1}(z)|^2) \sum_{j=0}^{n} |q_j(z)|^2 \\ &= \sqrt{2}(||\Pi_{n+1}(zI - A)^{-1}e_n||^2 + |a_n|^2||\Pi_{n+1}(zI - A)^{-1}e_{n+1}||^2),\end{aligned}$$

where in the last equality we have applied (2.19). Notice that the term on the right-hand side is bounded by $\sqrt{2}(1 + |a_n|^2)||(zI - A)^{-1}||^2$. Hence, using (4.7), we obtain

$$\begin{aligned}||(z\mathscr{I}_{n+\varepsilon_n} - \mathscr{A}_{n+\varepsilon_n})^{-1}|| &\leqslant ||E_n(zI - A)^{-1}E_n^*|| + ||(z\mathscr{I}_{n+\varepsilon_n} - \mathscr{A}_{n+\varepsilon_n})^{-1} - E_n(zI - A)^{-1}E_n^*|| \\ &\leqslant ||(zI - A)^{-1}|| + \sqrt{2}(1 + |a_n|^2)||(zI - A)^{-1}||^2,\end{aligned}$$

which is bounded in n by assumption on (a_n). Thus (4.4) holds.

(b) We first show that (4.4) implies that z may not be an accumulation point of zeros of q_n, $n \in \Lambda$. In fact, as in (4.6) we may find some $N > 0$ such that

$$||(\zeta\mathscr{I}_n - \mathscr{A}_n)^{-1}|| \leqslant 6C, \quad n \in \Lambda,\ n \geqslant N, \quad |z - \zeta| < \frac{1}{2C},$$

showing that eigenvalues of $\mathscr{A}_n$ (i.e., zeros of q_n) have to stay away from z for sufficiently large $n \in \Lambda$. Suppose now that Λ is as described in part (b). Then there exists an open neighborhood $U \subset \Omega$ of z such that u_{n-1} is analytic in U for $n \in \Lambda$ (at least after dropping a finite number of elements of Λ). Also, from Theorem 3.3(c) we know that $(u_{n-1})_{n\in\Lambda}$ is a normal family of meromorphic functions in Ω, with any partial limit being different from the constant ∞ by Theorem 3.3(a).

It follows from [18, Lemma 2.4(d)] that then $(u_{n-1})_{n\in\Lambda}$ is bounded uniformly on compact subsets of U, in particular,

$$d := \sup_{n\in\Lambda} |u_{n-1}(z)| < \infty.$$

Consequently,

$$|a_{n-1}q_n(z)|^2 = \frac{|q_{n-1}(z)|^2 + |a_{n-1}q_n(z)|^2}{|u_{n-1}(z)|^2 + 1} \geqslant \frac{|q_{n-1}(z)|^2 + |a_{n-1}q_n(z)|^2}{d^2 + 1}.$$

As in the proof of part (a) (with $\varepsilon_n = 1$ and n replaced by $n-1$) we may conclude that

$$||(z\mathscr{I}_n - \mathscr{A}_n)^{-1}|| \leqslant ||(zI - A)^{-1}|| + \sqrt{1 + d^2}(1 + |a_{n-1}|^2)||(zI - A)^{-1}||^2$$

and thus (4.4) is true.

(c) As in part (a), it is sufficient to show that $z \in \Omega(A)$ implies (4.4). Denote by $\mathscr{B}^{[n]}$ the infinite matrix obtained from $\mathscr{A}$ by replacing a_{n-1} by 0, and write $B^{[n]} := [\mathscr{B}^{[n]}]_{\min}$. Let $z \in \Omega(A)$. By assumption on $(a_{n-1})_{n\in\Lambda}$, we find an $N > 0$ such that

$$||A - B^{[n]}|| \leqslant \frac{1}{2||(zI - A)^{-1}||}, \quad n \in \Lambda,\ n \geqslant N.$$

Thus $z \in \Omega(B^{[n]})$ and $||(zI - B^{[n]})^{-1}|| \leqslant 2||(zI - A)^{-1}||$. On the other hand, $\mathscr{B}^{[n]}$ is block diagonal, with $E_n(zI - B^{[n]})^{-1}E_n^* = (z\mathscr{I}_n - \mathscr{A}_n)^{-1}$. Thus $||(z\mathscr{I}_n - \mathscr{A}_n)^{-1}|| \leqslant 2||(zI - A)^{-1}||$ for all $n \in \Lambda$, $n \geqslant N$, implying (4.4).

(d) This part follows immediately from Lemma 4.5.

(e) The complex Jacobi matrix $\mathscr{A}$ is proper by Theorem 2.6(c), and one easily deduces that the same is true for all associated Jacobi matrices $\mathscr{A}^{(k)}$. By the definition of $\Gamma_{\text{ess}}(A)$, there exists a $k > 0$ with $z \in \mathbb{C}\backslash\Gamma(A^{(k)})$, the latter being a subset of $\Omega(A^{(k)})$ by Theorem 3.5(c). From the proof of Theorem 3.3(a) we know that

$$||(z\mathscr{I}_n - \mathscr{A}_n^{(k)})^{-1}|| \leqslant \frac{1}{\text{dist}(z, \Gamma(A^{(k)}))} < \infty, \quad n \geqslant 0.$$

Let $\mathscr{B}$ be obtained from $\mathscr{A}$ by keeping the elements from $\mathscr{A}^{(k)}$, putting $\zeta \neq z$ on the first k diagonal positions, and zero elsewhere. One easily verifies that then

$$||(z\mathscr{I}_n - \mathscr{B}_n)^{-1}|| \leqslant \frac{1}{\text{dist}(z, \{\zeta\} \cup \Gamma(A^{(k)}))} < \infty, \quad n \geqslant 0.$$

Writing $B = [\mathscr{B}]_{\min}$, we trivially have $z \in \Omega(B)$, and $A - B$ is compact (and even of finite rank). Thus the assertion follows from Lemma 4.5. □

4.2. Some consequences for the approximation of the Weyl function

We summarize some consequences of the preceding section for the convergence of Padé approximants $\pi_n(z) = (e_0, (z\mathscr{I}_n - \mathscr{A}_n)^{-1}e_0)$ (i.e., Weyl functions of the finite sections $\mathscr{A}_n$) to the Weyl function $\phi(z) = (e_0, (zI - A)^{-1}e_0)$ in the following statement, which is an immediate consequence of Corollary 4.4.

Corollary 4.6. *The subsequence $(\pi_n)_{n\in\Lambda}$ converges to the Weyl function ϕ uniformly in the compact set F provided that one of the following conditions is satisfied*:
(a) $\Lambda=\{0,1,2,\ldots\}$ *and* $F\subset\Omega(A)\backslash\Gamma_{\text{ess}}(A)$.
(b) $(a_{n-1})_{n\in\Lambda}$ *tends to zero and* $F\subset\Omega(A)$.
(c) $(a_{n-1})_{n\in\Lambda}$ *is bounded*, $F\subset\Omega$, *with* $\Omega\not\subset\Gamma(A)$ *being some subdomain of* $\Omega(A)$, *and F does not contain accumulation points of zeros of* q_n, $n\in\Lambda$.

For convergence outside $\Gamma(A)$ (which is included in Corollary 4.6(a)) we refer the reader to [59, Theorems 26.2, 26.3; 20, Theorem 3.10] in the case of bounded $\mathscr{A}$, and [59, Theorem 25.4] in the case of determinate $\mathscr{A}$. For the special case of a compact perturbation of a self-adjoint Jacobi operator, Corollary 4.6(a) may be found in [14, Corollary 6; 13, Theorem 2]. The latter assertion applies a different technique of proof, and contains additional information about the number of poles at isolated points of $\Omega(A)\backslash\Gamma_{\text{ess}}(A)$. Assertion [14, Theorem 2] on bounded perturbations of a self-adjoint Jacobi operator is contained in Corollary 4.6(a).

For bounded complex Jacobi matrices, Corollary 4.6(b) may be found in [18, Corollary 4.2]. As shown in [18, Corollary 5.6], this statement can be used to prove the Baker–Gammel–Wills conjecture for Weyl functions of operators with countable compact spectrum. Corollary 4.6(c) for bounded complex Jacobi matrices was established in [18, Theorem 4.1] (containing additional results on the rate of convergence in terms of the functions g_{inf} and g_{sup} of Theorem 3.1). Here as set Ω we may choose the unbounded connected component $\Omega_0(A)$ of $\Omega(A)$. Notice that a connected component Ω of $\Omega(A)$ with $\Omega\not\subset\Gamma(A)$ is unbounded also for unbounded $\mathscr{A}$. Thus Corollary 4.6(c) has to be compared with the result of Gonchar [28] mentioned in Section 3.3.

In their work on bounded tridiagonal infinite matrices, Aptekarev, Kaliaguine and Van Assche observed [7, Theorem 2] that

$$\liminf_{n\to\infty}|\pi_n(z)-\phi(z)|=0,\quad z\in\Omega(A). \tag{4.13}$$

Notice that this relation also holds for unbounded $\mathscr{A}$ since otherwise a nontrivial multiple of the sequence $(|q_n(z)|)_{n\geqslant 0}\notin\ell^2$ would minorize the sequence $(|r_n(z)|)_{n\geqslant 0}=(|\phi(z)-\pi_n(z)|\cdot|q_n(z)|)_{n\geqslant 0}\in\ell^2$. If a subsequence of (a_n) is bounded, then by combining Theorem 4.3(a) with Theorem 4.1 we see that relation (4.13) holds even uniformly in some neighborhood of any $z\in\Omega(A)$. This was observed before in [4, Corollaries 3, 4] for bounded real, and in [18, Theorem 4.4] for bounded complex Jacobi matrices.

In this context, let us discuss the related question whether (pointwise) convergence of (a subsequence of) Padé approximants at some z implies that $z\in\Omega(A)$. Clearly, the answer is no; see for instance the counterexamples presented in the last paragraph of [7]. If, however, we replace Padé convergence by weak (or strong) resolvent convergence, and we limit ourselves to sequences (a_n) containing a bounded subsequence, then the answer is yes: we have $z\in\Omega(A)$ if and only if there exists an infinite set Λ of indices such that

$$\lim_{\substack{n\to\infty\\ n\in\Lambda}}(E_ny',(z\mathscr{I}_n-\mathscr{A}_n)^{-1}E_ny)\quad\text{exists }\forall y,y'\in\ell^2. \tag{4.14}$$

Indeed, if $z\in\Omega(A)$, then we may use Theorems 4.3(a) and 4.1 to establish (4.14). Conversely, (4.14) implies (4.4) by [30, Problem V.1.6], and thus $z\in\Omega(A)$ by Theorem 4.3(a).

We terminate this section with a generalization of [18, Theorem 3.1], where convergence in (logarithmic) capacity of (π_n) is established for bounded $\mathscr{A}$ on compact subsets of the unbounded connected component of the resolvent set.

Theorem 4.7. *Let $(a_n)_{n\in\Lambda}$ be bounded, and denote by Ω a connected component of $\Omega(A)$. Then there exist $\varepsilon_n \in \{0,1\}$ such that, for each compact $F\subset\Omega$ and for each $\varepsilon>0$, we have*

$$\lim_{\substack{n\to\infty\\ n\in\Lambda}} \operatorname{cap}\{z\in F\colon |\phi(z)-\pi_{n+\varepsilon_n}(z)|\geqslant\varepsilon\}=0.$$

If in addition $\Omega\not\subset\Gamma(A)$, then we may choose $\varepsilon_n=1$ for all $n\in\Lambda$.

Proof. From Theorem 3.3(c) we know that $(u_n)_{n\in\Lambda}$ is a normal family of meromorphic functions in Ω. If in addition $\Omega\not\subset\Gamma(A)$, then any partial limit is different from the constant ∞ by Theorem 3.3(a), and we put $v_n=u_n$, $\varepsilon_n=1$. Otherwise, let $\zeta\in\Omega$. If $|u_n(\zeta)|\leqslant 1$, then again $v_n=u_n$, $\varepsilon_n=1$, and otherwise $v_n=1/u_n$, $\varepsilon_n=0$. In this way we have constructed a normal family $(v_n)_{n\in\Lambda}$ of meromorphic functions in Ω with any partial limit being different from the constant ∞.

Let $F,F'\subset\Omega$ be compact, the interior of F' containing F. Let ω_n, $n\in\Lambda$, be a monic polynomial of minimal degree such that $\omega_n v_n$ is analytic in F'. From the proof of Theorem 3.4(c) we know that the degree v_n of ω_n is bounded by some $v(F')$ uniformly for $n\in\Lambda$. We claim that

$$\sup_{n\in\Lambda} C_n =: C(F)<\infty,\qquad C_n := \max_{z\in F}|\omega_n(z)\cdot v_n(z)|. \tag{4.15}$$

Otherwise, there would be integers $n_k\in\Lambda$ such that $C_{n_k}>k$. By normality, we may assume, without loss of generality, that $(v_{n_k})_k$ converges to some meromorphic v uniformly in F'. Since $v\neq\infty$, we find some open set D, $F\subset D\subset F'$, having a finite number of open components, and $v(z)\neq\infty$ for $z\in\partial D$. By uniform convergence on ∂D it follows that

$$\limsup_{k\to\infty}\max_{z\in\partial D}|v_{n_k}(z)|<\infty.$$

Since D is bounded and the degrees of the ω_n are uniformly bounded, we may conclude that the above relation remains true after multiplication of v_{n_k} with ω_{n_k}. Using the maximum principle for analytic functions, we obtain a bound for $\omega_{n_k}\cdot v_{n_k}$ on F uniformly in k, in contradiction to the construction of n_k. Thus (4.5) holds.

From (4.15) we conclude that, for any $d>\max\{2,2C(F)\}$ and $n\in\Lambda$,

$$\operatorname{cap}\{z\in F\colon \sqrt{1+|v_n(z)|^2}>d\}\leqslant\operatorname{cap}\{z\in F\colon |v_n(z)|>d/2\}$$

$$\leqslant\operatorname{cap}\left\{z\in F\colon |\omega_n(z)|\leqslant\frac{2C(F)}{d}\right\}=\left(\frac{2C(F)}{d}\right)^{1/v_n}\leqslant\left(\frac{2C(F)}{d}\right)^{1/v(F')}.$$

Notice that by construction (compare with the proof of Theorem 4.3(a))

$$\phi(z)-\pi_{n+\varepsilon_n}(z)=\frac{a_n^{\varepsilon_n}r_{n+\varepsilon_n}(z)\sqrt{1+|v_n(z)|^2}}{\sqrt{|q_n(z)|^2+|a_nq_{n+1}(z)|^2}}\leqslant\sqrt{1+|v_n(z)|^2[|r_n(z)|^2+|a_nr_{n+1}(z)|^2|]}.$$

Since the term in brackets tends to zero uniformly in F by (2.19), we obtain the claimed convergence by combining the last two formulas. □

Combining the reasoning of the proofs of Theorems 4.3(a) and 4.7, we may also show that with $\Omega, \Lambda, \varepsilon_n$ as in Theorem 4.7 there holds for any compact $F \subset \Omega$

$$\lim_{\varepsilon\to 0} \operatorname{cap}\left(\left\{z \in F:\ \limsup_{\substack{n\to\infty\\ n\in\Lambda}} \|(z\mathscr{I}_{n+\varepsilon_n} - \mathscr{A}_{n+\varepsilon_n})^{-1}\| > \frac{1}{\varepsilon}\right\}\right) = 0.$$

Thus, our result on convergence in capacity of Padé approximants is again connected to a type of strong resolvent convergence in capacity.

In [48], Stahl suggested to replace in the Baker–Gammel–Wills conjecture [11] locally uniform convergence of a subsequence by convergence in capacity of a subsequence. Theorem 4.7 confirms (under assumptions on the regularity of the underlying function and assumptions on some of the coefficients of its J-fraction expansion) that this is true in the resolvent set. Of course, this open set does not need to contain the maximal disk of analyticity of the Weyl function, but it might be helpful in investigating the above conjecture for special classes of functions. We refer the reader to Baker's survey [10] for further recent developments in convergence questions for Padé approximation.

4.3. An application to asymptotically periodic Jacobi matrices

A complex Jacobi-matrix $\mathscr{A}$ is called m-periodic if $a_{jm+k} = a_k,\ b_{jm+k} = b_k,\ k = 0, 1, \ldots, m-1,\ j \geqslant 0$, and $\tilde{\mathscr{A}}$ is called *asymptotically periodic* if it is a compact perturbation of an m-periodic matrix, i.e.,

$$\lim_{j\to\infty} \tilde{a}_{jm+k} = a_k, \quad \lim_{j\to\infty} \tilde{b}_{jm+k} = b_k, \quad k = 0, 1, \ldots, m-1.$$

Real periodic and asymptotically periodic Jacobi matrices have been studied by a number of authors, see, e.g., [26,24,34]. Complex perturbations of real periodic Jacobi matrices are investigated in [15,16], and complex (asymptotically) periodic Jacobi matrices in [20, Sections 2.2, 2.3; 19, Example 3.6,3].

It is well known (see, e.g., [20, Section 2.2]) that, for m-periodic $\mathscr{A}$, the sequences $(p_n(z))_{n\geqslant -1}$ and $(q_n(z))_{n\geqslant -1}$ satisfy the recurrence relation [13]

$$y_{(j+1)m+k} = h(z) \cdot y_{jm+k} - y_{(j-1)m+k}, \quad j \geqslant 0,\ k \geqslant -1 \tag{4.16}$$

with some polynomial h for which we have several representations:

$$h(z) = \frac{q_{2m-1}(z)}{q_{m-1}(z)} = \frac{p_{2m}(z)}{p_m(z)} = q_m(z) - a_{m-1} p_{m-1}(z).$$

In [20, Section 2.3], the authors show (see also [19, Example 3.6] or [3]) that $\sigma_{\text{ess}}(A) = \{z \in \mathbb{C}\colon h(z) \in [-2, 2]\}$, which by [20, Lemma 2.5] has empty interior and connected complement. The Weyl function of $\mathscr{A}$ is an algebraic function, meromorphic (and single valued) in $\mathbb{C}\backslash\sigma_{\text{ess}}(A)$, with possible poles at the zeros of q_{m-1} [20, Section 2.2], and $\sigma(A)$ is just the extremal set of Stahl [49], i.e., the set of minimal capacity outside of which the Weyl function has a single-valued analytic continuation from infinity [20, Remark 2.9].

[13] Here we need to put $a_{-1} = a_{m-1}$, and thus $1 = -a_{m-1} p_{-1}(z) = a_{m-1} r_{-1}(z)$. This slight modification does not change the other elements of the sequences $(p_n(z))_{n\geqslant -1}$ or $(r_n(z))_{n\geqslant -1}$.

Let us show here that we may localize the spurious zeros of the FOPs associated with $\mathscr{A}$ (and with $\tilde{\mathscr{A}}$). First, using (1.2) and (2.20), one easily verifies the well-known fact that

$$q_n^{(k+1)}(z) := a_k(q_{n+k+1}(z)r_k(z) - r_{n+k+1}(z)q_k(z)) \tag{4.17}$$

is the nth FOP of the associated Jacobi matrix $\mathscr{A}^{(k+1)}$. For $z \notin \sigma_{\mathrm{ess}}(A)$, the equation $y^2 = h(z)y - 1$ has one solution $w(z)$ of modulus $|w(z)| < 1$, and the second solution $1/w(z)$. From (4.16), (3.7) we may conclude that there exist (algebraic) functions $\alpha_k, \beta_k, \gamma_k$ such that

$$r_{jm+k}(z) = \alpha_k(z) \cdot w(z)^j, \quad q_{jm+k}(z) = \beta_k(z) \cdot w(z)^j + \gamma_k(z) \cdot w(z)^{-j} \tag{4.18}$$

for all $k \geqslant -1$, $j \geqslant 0$ and $z \in \Omega(A)$. Injecting this information in (2.20) we obtain

$$a_k(\gamma_{k+1}(z)\alpha_k(z) - \gamma_k(z)\alpha_{k+1}(z)) = 1, \tag{4.19}$$

showing that $|\gamma_k(z)| + |\gamma_{k+1}(z)| \neq 0$ for all $z \in \Omega(A)$. We may deduce that

$$\lim_{j\to\infty} \chi\left(\frac{q_{jm+k}(z)}{a_{jm+k}q_{jm+k+1}(z)}, \frac{\gamma_k(z)}{a_k\gamma_{k+1}(z)}\right) = 0, \quad k = 0, 1, \ldots, m-1, \ z \in \Omega(A). \tag{4.20}$$

Also, by periodicity, $q_n^{(k)}(z) = q_n^{(k+m)}(z)$, and by combining (4.18) with (4.17) we may conclude that

$$q_{m-1}^{(k+1)}(z) = a_k[w(z)^{-1} - w(z)]\alpha_k(z)\gamma_k(z). \tag{4.21}$$

From (4.20) and (4.21) we see that spurious zeros of $(q_{jm+k})_{j\geqslant 0}$ accumulating in $\zeta \in \Omega(A)$ satisfy $\gamma_k(\zeta) = 0$ and thus $q_{m-1}^{(k+1)}(\zeta) = 0$.

Combining this finding with Theorem 3.6 and Corollary 4.6(c), we obtain the following statement:

Corollary 4.8. *Let $\tilde{\mathscr{A}}$ be an asymptotically periodic complex Jacobi matrix, denote by $\mathscr{A}$ the corresponding m-periodic Jacobi matrix, and let $k \in \{0, \ldots, m-1\}$. Then for each compact subset F of $\Omega(A) \cap \Omega(\tilde{A})$ which does not contain zeros of $q_{m-1}^{(k+1)}$ there exists a $J = J(F)$ such that $\tilde{q}_{mj+k}$ has no zeros in F for $j \geqslant J$, and*

$$\lim_{j\to\infty} \max_{z\in F} |\tilde{\phi}(z) - \tilde{\pi}_{mj+k}(z)| = 0.$$

Notice that pointwise convergence for asymptotically periodic complex Jacobi matrices was already obtained in [20, Theorem 2.11]. If $\mathscr{A}$ is real, then clearly $\sigma_{\mathrm{ess}}(A)$ consists of at most m real intervals. Barrios et al. [15,16] showed that then the zeros of all $q_{m-1}^{(k+1)}$ lie in the convex hull $\mathscr{S}$ of $\sigma_{\mathrm{ess}}(A)$ and obtained uniform convergence of $(\tilde{\pi}_n)_{n\geqslant 0}$ on compact subsets of $\mathbb{C}\backslash\mathscr{S}$.

Acknowledgements

Parts of this manuscript originated from very fruitful discussions during a stay at the University of Almeria (Spain). The author gratefully acknowledges the hospitality of the Grupo de Investigación Teoría de Aproximación y Polinomios Ortogonales.

I also would like to thank V. Sorokin and G. López Lagomasino for reading an earlier version of this manuscript and making valuable comments. Finally, I am grateful to the anonymous referees for their careful reading and their remarks which helped to improve the manuscript.

References

[1] N.I. Akhiezer, Classical Moment Problems and some Related Questions in Analysis, Oliver&Boyd, Edinburgh, 1965.

[2] N.I. Akhiezer, I.M. Glazman, Theory of Linear Operators in a Hilbert Space, Vols. I,II, Pitman, Boston, 1981.

[3] A. Almendral Vázquez, The spectrum of a periodic complex Jacobi matrix revisited, J. Approx. Theory, to appear.

[4] A. Ambroladze, On exceptional sets of asymptotic relations for general orthogonal polynomials, J. Approx. Theory 82 (1995) 257–273.

[5] A.I. Aptekarev, Multiple orthogonal polynomials, J. Comput. Appl. Math. 99 (1998) 423–447.

[6] A.I. Aptekarev, V.A. Kaliaguine, Complex rational approximation and difference equations, Suppl. Rend. Circ. Mat. Palermo 52 (1998) 3–21.

[7] A.I. Aptekarev, V.A. Kaliaguine, W. Van Assche, Criterion for the resolvent set of nonsymmetric tridiagonal operators, Proc. Amer. Math. Soc. 123 (1995) 2423–2430.

[8] A.I. Aptekarev, V.A. Kaliaguine, J. Van Iseghem, Genetic sum representation for the moments of a system of Stieltjes functions and its application, Constr. Approx., to appear.

[9] R.J. Arms, A. Edrei, The Padé tables and continued fraction representations generated by totally positive sequences in: Mathematical Essays, Ohio University Press, Athens, OH, 1970, pp. 1–21.

[10] G.A. Baker, Defects and the convergence of Padé Approximants, LA-UR-99-1570, Los Alamos Nat. Lab., 1999.

[11] G.A. Baker, P.R. Graves-Morris, in: Padé Approximants, 2nd Edition, Encyclopedia of Mathematics, Cambridge University Press, New York, 1995.

[12] L. Baratchart, personal communication, June 1999.

[13] D. Barrios, G. López, A. Martínez, E. Torrano, On the domain of convergence and poles of complex J-fractions, J. Approx. Theory 93 (1998) 177–200.

[14] D. Barrios, G. López, A. Martínez, E. Torrano, Finite-dimensional approximations of the resolvent of an infinite banded matrix and continued fractions, Sbornik: Mathematics 190 (1999) 501–519.

[15] D. Barrios, G. López, E. Torrano, Location of zeros and asymptotics of polynomials defined by three-term recurrence relation with complex coefficients, Russian Acad. Sci. Sb. Math. 80 (1995) 309–333.

[16] D. Barrios, G. López, E. Torrano, Polynomials generated by asymptotically periodic complex recurrence coefficients, Sbornik: Mathematics 186 (1995) 629–660.

[17] G. Baxter, A norm inequality for a finite-section Wiener–Hopf equation, Illinois Math. 7 (1963) 97–103.

[18] B. Beckermann, On the convergence of bounded J-fractions on the resolvent set of the corresponding second order difference operator, J. Approx. Theory 99 (1999) 369–408.

[19] B. Beckermann, On the classification of the spectrum of second order difference operators, Publication ANO 379, Université de Lille, 1997 Math. Nachr., to appear.

[20] B. Beckermann, V.A. Kaliaguine, The diagonal of the Padé table and the approximation of the Weyl function of second order difference operators, Constr. Approx. 13 (1997) 481–510.

[21] Yu.M. Berezanskii, Integration of nonlinear difference equation by the inverse spectral problem method, Soviet Mathem. Dokl. 31 (2) (1985) 264–267.

[22] M. Castro Smirnova, Determinacy of bounded complex perturbations of Jacobi matrices, J. Approx. Theory, to appear.

[23] S. Demko, W.F. Moss, P.W. Smith, Decay rates for inverses of band matrices, Math. Comp. 43 (1984) 491–499.

[24] J. Geronimo, W. Van Assche, Orthogonal polynomials with asymptotically periodic recurrence coefficients, J. Approx. Theory 46 (1986) 251–283.

[25] J.S. Geronimo, E.M. Harrell II, W. Van Assche, On the asymptotic distribution of eigenvalues of banded matrices, Constr. Approx. 4 (1988) 403–417.

[26] J.L. Geronimus, On some difference equations and associated systems of orthogonal polynomials, Zapiski Math. Otdel. Phys.-Math. Facul. Kharkov Universiteta and Kharkov Mathem. Obchestva XXV (1957) 87–100 (in Russian).

[27] M. Goldberg, E. Tadmor, On the numerical radius and its applications, Linear Algebra Appl. 42 (1982) 263–284.

[28] A.A. Gonchar, On uniform convergence of diagonal Padé approximants, Math. USSR Sb. 46 (1983) 539–559.

[29] V.A. Kaliaguine, On rational approximation of the resolvent function of second order difference operator, Russian Math. Surveys 49 (3) (1994) 187–189.

[30] T. Kato, Perturbation Theory for Linear Operators, Springer, Berlin, 1966.

[31] A.B.J. Kuijlaars, W. Van Assche, The asymptotic zero distribution of orthogonal polynomials with varying recurrence coefficients, J. Approx. Theory 99 (1999) 167–197.

[32] L. Lorentzen, H. Waadeland, Continued Fractions with Applications, North-Holland, Amsterdam, 1992.

[33] A.P. Magnus, Toeplitz matrix techniques and convergence of complex weight Padé approximants, J. Comput. Appl. Math. 19 (1987) 23–38.

[34] A. Máté, P. Nevai, W. Van Assche, The supports of measures associated with orthogonal polynomials and the spectra of the related self adjoint operators, Rocky Mountain J. Math. 21 (1991) 501–527.

[35] P. Nevai, in: Orthogonal Polynomials, Memoirs Amer. Math. Soc., Vol. 213, AMS, Providence, RI, 1979.

[36] P. Nevai, W. Van Assche, Compact perturbations of orthogonal polynomials, Parcific J. Math. 153 (1992) 163–184.

[37] E.M. Nikishin, Discrete Sturm–Liouville operators and some problems of function theory, J. Soviet Math. 35 (1986) 2679–2744.

[38] E.M. Nikishin, V.N. Sorokin, Rational Approximations and Orthogonality, Translations of Mathematical Monographs, Vol. 92, American Mathematical Society, Providence, RI, 1991.

[39] J. Nuttall, Padé polynomial asymptotics from a singular integral equation, Constr. Approx. 6 (1990) 157–166.

[40] J. Nuttall, C.J. Wherry, Gaussian integration for complex weight functions, J. Inst. Math. Appl. 21 (1978) 165–170.

[41] E.B. Saff, V. Totik, Logarithmic Potentials with External Fields, Springer, Berlin, 1997.

[42] D. Sarason, Moment problems and operators in Hilbert spaces. Proc. Symp. Appl. Math. 37, 1987, pp. 54–70.

[43] J.L. Schiff, Normal Families, Universitext, Springer, New York, 1993.

[44] J. Shohat, J. Tamarkin, in: The Problem of Moments, Mathematical Surveys, Vol. 1, AMS, Providence, RI, 1950.

[45] B. Simon, The classical moment problem as a self-adjoint finite difference operator, Adv. Math. 137 (1998) 82–203.

[46] H. Stahl, Divergence of diagonal Padé approximants and the asymptotic behavior of orthogonal polynomials associated with nonpositive measures, Constr. Approx. 1 (1985) 249–270.

[47] H. Stahl, On the Divergence of certain Padé Approximants and the Behavior of the Associated Orthogonal Polynomials, Lecture Notes in Mathematics, Vol. 1771, Springer, Berlin, 1985, pp. 321–330.

[48] H. Stahl, Conjectures around the Baker–Gammel–Wills conjecture, Constr. Approx. 13 (1997) 287–292.

[49] H. Stahl, The convergence of Padé approximants to functions with branch points, J. Approx. Theory 91 (1997) 139–204.

[50] H. Stahl, Spurious poles in Padé approximation, J. Comput. Appl. Math. 99 (1998) 511–527.

[51] H. Stahl, V. Totik, General Orthogonal Polynomials, Encyclopedia of Mathematics, Cambridge University Press, New York, 1992.

[52] M.H. Stone, Linear Transformations in Hilbert Spaces and their Applications to Analysis, American Mathematical Society, Providence, RI, 1932.

[53] R. Szwarc, A lower bound for orthogonal polynomials with an application to polynomial hypergroups, J. Approx. Theory 81 (1995) 145–150.

[54] R. Szwarc, A counterexample to subexponential growth of orthogonal polynomials, Constr. Approx. 11 (1995) 381–389.

[55] G. Szegő, Orthogonal Polynomials, AMS, Providence, RI, 1975.

[56] W. Van Assche, Orthogonal polynomials, associated polynomials and functions of the second kind, J. Comput. Appl. Math. 37 (1991) 237–249.

[57] W. Van Assche, Compact Jacobi matrices: from Stieltjes to Krein and $M(a;b)$, Ann. Fac. Sci. Toulouse, Numero Spécial Stieltjes (1996) 195–215.

[58] E.A. van Doorn, Representations and bounds for zeros of orthogonal polynomials and eigenvalues of sign-symmetric tri-diagonal matrices, J. Approx. Theory 51 (1987) 254–266.

[59] H.S. Wall, Analytic Theory of Continued Fractions, Chelsea, Bronx, NY, 1973.

[60] L. Zalcman, Normal families: new perspectives, Bull. Amer. Math. Soc. 35 (1998) 215–230.

ELSEVIER

Journal of Computational and Applied Mathematics 127 (2001) 67–91

JOURNAL OF
COMPUTATIONAL AND
APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Quadrature and orthogonal rational functions

A. Bultheel[a,*,1], P. González-Vera[b,2], E. Hendriksen[c], Olav Njåstad[d]

[a] *Department of Computer Science, Celestijnenlaan 200 A, 3001 K.U. Leuven, Belgium*
[b] *Department Análisis Math., Univ. La Laguna, Tenerife, Spain*
[c] *Department of Mathematics, University of Amsterdam, Netherlands*
[d] *Department of Math. Sc., Norwegian University of Science and Technology, Trondheim, Norway*

Received 5 April 1999; received in revised form 18 January 2000

Abstract

Classical interpolatory or Gaussian quadrature formulas are exact on sets of polynomials. The Szegő quadrature formulas are the analogs for quadrature on the complex unit circle. Here the formulas are exact on sets of Laurent polynomials. In this paper we consider generalizations of these ideas, where the (Laurent) polynomials are replaced by rational functions that have prescribed poles. These quadrature formulas are closely related to certain multipoint rational approximants of Cauchy or Riesz–Herglotz transforms of a (positive or general complex) measure. We consider the construction and properties of these approximants and the corresponding quadrature formulas as well as the convergence and rate of convergence. © 2001 Elsevier Science B.V. All rights reserved.

MSC: 65D30; 33D45; 41A21

Keywords: Numerical quadrature; Orthogonal rational functions; Multipoint Padé approximation

1. Introduction

It is an important problem in numerical analysis to compute integrals of the form $\int_a^b f(x)\,\mathrm{d}\mu(x)$ where μ is in general a complex measure on the interval $[a,b]$ with $-\infty \leqslant a < b \leqslant +\infty$. Most quadrature formulas approximate this integral by a weighted combination of function values: $\sum_{i=1}^{n} A_{ni} f(\xi_{ni})$.

[*] Corresponding author. Tel.: +32-16-327540; fax: +32-16-327996.
E-mail address: adhemar.bultheel@cs.kuleuven.ac.be (A. Bultheel).

[1] The work of this author is partially supported by the Fund for Scientific Research (FWO), project “Orthogonal systems and their applications”, grant #G.0278.97 and the Belgian Programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister’s Office for Science, Technology and Culture. The scientific responsibility rests with the author.

[2] The work of this author was partially supported by the scientific research project PB96-1029 of the Spanish D.G.E.S.

PII: S0377-0427(00)00493-3

The quadrature parameters are the abscissas or knots $\{\xi_{ni}\}_{i=1}^{n}$ and the coefficients or weights $\{A_{ni}\}_{i=1}^{n}$. One objective in constructing quadrature formulas could be to find the quadrature parameters such that the formulas are exact for all functions in a class that is as large as possible.

The most familiar quadrature formulas based on this principle are the Gauss–Christoffel formulas. These formulas choose for a positive measure μ as abscissas the zeros of φ_n, which is the polynomial of degree n orthogonal with respect to the inner product $\langle f,g\rangle = \int_a^b f(x)\overline{g(x)}\,\mathrm{d}\mu(x)$. These zeros are simple and all in the interval (a,b). The weights are the so-called Christoffel numbers. They are positive and are constructed in such a way that the quadrature formula is exact for all $f\in\Pi_{2n-1}$, i.e., for any polynomial of degree at most $2n-1$. These Gauss–Christoffel formulas are optimal in the sense that it is impossible to construct an n-point formula that is exact in Π_k with $k\geqslant 2n$. For a survey, see for example [25]. The study of such quadrature formulas was partly motivated by the role they played in the solution of the corresponding Stieltjes–Markov moment problem. That is, given a sequence of complex numbers c_k, find a positive or complex measure μ such that $c_k = \int_a^b x^k\,\mathrm{d}\mu(x)$, $k=0,1,\ldots$. For complex measures see [41,31].

In this survey, it is our intention to concentrate on the computation of integrals of the form $I_\mu\{f\} := \int_{-\pi}^{\pi} f(\mathrm{e}^{\mathrm{i}\theta})\,\mathrm{d}\mu(\theta)$ where f is a complex function defined on the unit circle and μ is in general a complex measure on $[-\pi,\pi]$.

The motivation for this problem is that, just as the previous case is related to a Stieltjes moment problem for an interval, this integral can be related to the solution of a trigonometric moment problem when μ is a positive measure. The Stieltjes–Markov moment problems suggested the construction of quadrature formulas in the largest possible subset of polynomials. However, in the case of the trigonometric moment problem, it is very natural to consider Laurent polynomials (L-polynomials) instead. This is motivated by the fact that a function continuous on the unit circle can be uniformly approximated by L-polynomials. Since L-polynomials are rational functions with poles at the origin and at infinity, the step towards a more general situation where the poles are at several other (fixed) positions in the complex plane seems natural. This gives rise to a discussion of orthogonal L-polynomials and orthogonal rational functions (with arbitrary but fixed poles).

The outline of the paper is as follows. First we introduce the basic ideas and techniques by considering the case of Szegő quadrature formulas that are exact in the largest possible sets of Laurent polynomials for integrals with a positive measure on the unit circle. We introduce the rational (two-point Padé or Padé type) approximants, based on orthogonal polynomials, and the error estimates for the rational approximants and for the quadrature. Section 3 introduces the rational variants of these formulas and approximants. Their convergence is established in Section 5. The necessary properties of orthogonal rational functions needed are discussed briefly in Section 4. Next, we discuss the corresponding problems for a complex measure on the unit circle in Section 6. In Section 7 we also discuss the case where the poles of the rational functions are chosen inside the support of the measure. For ideas related to integrals on an interval (compact or not) of the real line we refer to Section 8. Finally, in Section 9 we state some open problems for further research.

Some notation before we start: $\mathbb{C}$ is the complex plane and $\hat{\mathbb{C}} = \mathbb{C}\cup\{\infty\}$. We denote the unit circle by $\mathbb{T} = \{z\in\mathbb{C}\colon |z|=1\}$, the open unit disk by $\mathbb{D} = \{z\in\mathbb{C}\colon |z|<1\}$, and the external of the closed disk by $\mathbb{E} = \{z\in\mathbb{C}\colon |z|>1\}$. For any function f, the para-hermitian conjugate is defined as $f_*(z) := \overline{f(1/\bar{z})}$. The set of polynomials of degree at most $n\geqslant 0$ is denoted by Π_n, and Π is the set of all polynomials. By $\Lambda_{p,q} = \{\sum_{k=p}^{q} a_k z^k\colon a_k\in\mathbb{C}\}$ we denote subsets of L-polynomials, and Λ is the set of all L-polynomials. Note that $\Lambda_{0,n}=\Pi_n$. If $P\in\Pi_n\backslash\Pi_{n-1}$ (where $\Pi_{-1}=\emptyset$), then $P^*(z)=z^nP_*(z)$.

2. The unit circle and L-polynomials

A systematic study of quadrature problems for integrals of the form

$$I_\mu\{f\} := \int_{-\pi}^{\pi} f(\mathrm{e}^{\mathrm{i}\theta})\,\mathrm{d}\mu(\theta) \tag{2.1}$$

with μ a positive measure was initiated by Jones et al. [36]. We shall introduce the main ingredients here as an introduction to our more general discussion in the following sections. The quadrature formula has the form (the measure μ_n is discrete with mass A_{nk} at the points ξ_{nk}, $k = 1,\dots,n$)

$$I_{\mu_n}\{f\} = \sum_{k=1}^{n} A_{nk} f(\xi_{nk}), \tag{2.2}$$

where the abscissas are all simple and on $\mathbb{T}$. The objective is an analogue of the Gauss–Christoffel formula. That is, find knots and weights such that these formulas are exact in the largest possible set of L-polynomials. So we consider here the polynomial space $\mathscr{L}_n = \Pi_n$ and the space of Laurent polynomials $\mathscr{R}_{p,q} = \Lambda_{-p,q}$, where p and q are always assumed to be nonnegative integers. Note that the dimension of $\mathscr{R}_{p,q}$ is $p+q+1$. It can be shown that for n different points $\xi_{ni} \in \mathbb{T}$, there is no quadrature formula of form (2.2) that is exact in some $\mathscr{R}_{p,q}$ of dimension $p+q+1 > 2n-1$. But there is a quadrature formula that is exact in $\mathscr{R}_{n-1,n-1}$, and this has the maximal possible dimension. All n-point quadrature formulas with this maximal domain of validity can be described with one free parameter $\tau_n \in \mathbb{T}$. The formulas are called Szegő formulas. They can be described as follows. First we need the orthonormal polynomials φ_k, obtained by orthogonalizing $1, z, z^2, \dots$ with respect to the inner product

$$\langle f, g\rangle = \int_{-\pi}^{\pi} f(\mathrm{e}^{\mathrm{i}\theta})\overline{g(\mathrm{e}^{\mathrm{i}\theta})}\,\mathrm{d}\mu(\theta).$$

The para-orthogonal polynomials are then defined by $Q_n(z;\tau_n) := \varphi_n(z) + \tau_n\varphi_n^*(z)$. Para-orthogonal means that $Q_n \perp \operatorname{span}\{z,\dots,z^{n-1}\}$ while $\langle Q_n, 1\rangle \neq 0 \neq \langle Q_n, z^n\rangle$. If $\tau_n \in \mathbb{T}$, then it can be shown that $Q_n(z;\tau_n)$ has n simple zeros $\{\xi_{nk}\}_{k=1}^n \subset \mathbb{T}$. These depend on the parameter τ_n. We can use this parameter to place one zero, e.g., ξ_{n1}, in some arbitrary $w \in \mathbb{T}$. The other knots $\{\xi_{nk}\}_{k=2}^n$ are then the zeros of $k_{n-1}(z,w)$, where k_{n-1} represents the reproducing kernel for $\mathscr{L}_{n-1}$, that is $k_{n-1}(z,w) = \sum_{i=0}^{n-1}\varphi_i(z)\overline{\varphi_i(w)}$. It reproduces in the sense that $\langle f, k_{n-1}(\cdot,w)\rangle = f(w)$ for every $f \in \mathscr{L}_{n-1}$. This implies, for example, that we only need to know the first n polynomials $\varphi_0,\dots,\varphi_{n-1}$ to find the n knots $\{\xi_{nk}\}_{k=1}^n$.

Theorem 2.1 (González-Vera et al. [33]). *If* (2.2) *is a Szegő formula, then the distinct knots* $\{\xi_{ni}\}_{i=1}^n \subset \mathbb{T}$ *are given by the zeros of the para-orthogonal functions* $Q_n(z;\tau_n) = \varphi_n(z) + \tau_n\varphi_n^*(z)$ *with* $\tau_n \in \mathbb{T}$ *arbitrary, or equivalently by some arbitrary point* $\xi_{n1} \in \mathbb{T}$ *and the zeros of* $k_{n-1}(z,\xi_{n1})$. *The* (*positive*) *weights* A_{ni} *are given by the reciprocals*

$$A_{ni}^{-1} = \sum_{k=0}^{n-1} |\varphi_k(\xi_{ni})|^2 = k_{n-1}(\xi_{ni},\xi_{ni}).$$

To study error formulas and convergence properties, we use the link with moment problems and certain rational approximations.

Therefore, we introduce some rational approximants to the Riesz–Herglotz transform of the measure μ. Let us start by defining the *Riesz–Herglotz transform* as

$$F_\mu(z) = I_\mu\{D(\cdot,z)\}, \quad D(t,z) = \frac{t+z}{t-z},$$

which is a function analytic in $\hat{\mathbb{C}}\backslash\mathbb{T}$ having a radial limit a.e. to the unit circle whose real part is the absolutely continuous part of μ. Moreover, it has expansions in $\mathbb{D}$ and $\mathbb{E}$ that can be described in terms of the moments $c_k = I_\mu\{z^{-k}\}$, $k \in \mathbb{Z}$. We have

$$F_\mu(z) \sim L_0(z) = c_0 + 2\sum_{j=1}^{\infty} c_k z^k, \quad z \in \mathbb{D},$$

$$F_\mu(z) \sim L_\infty(z) = -c_0 - 2\sum_{j=1}^{\infty} c_{-k} z^{-k}, \quad z \in \mathbb{E}.$$

Here we have a motivation for finding approximants that converge to F_μ, since this could help solving the moment problem. Some rational approximants with fixed denominators are constructed as follows.

Consider a triangular table $\mathbb{X} = \{\xi_{ni} \in \mathbb{T}\colon i = 1,\ldots,n;\ n \in \mathbb{N}\}$, where $\xi_{ni} \neq \xi_{nj}$ for $i \neq j$. We shall use the rows of this array as knots for quadrature formulas. Therefore, we shall call such an array a *node array*. Let $Q_n \in \Pi_n$ be a node polynomial, that is, a polynomial whose zeros are $\{\xi_{ni},\ i = 1,\ldots,n\}$. For any such polynomial Q_n, and for any pair of nonnegative integers (p,q) such that $p+q = n-1 \geqslant 0$, we can find a unique polynomial $P_n \in \Pi_n$ such that for $F_{\mu_n} = P_n/Q_n$ we have

$$F_\mu(z) - F_{\mu_n}(z) = \mathrm{O}[z^{p+1}], \quad z \to 0,$$

$$F_\mu(z) - F_{\mu_n}(z) = \mathrm{O}[(1/z)^{q+1}], \quad z \to \infty.$$

The rational function F_{μ_n} is called a *two-point Padé-type approximation* (2PTA). The relation with quadrature formulas is that if the zeros of Q_n are the abscissas of the n-point Szegő quadrature formula, then the 2PTA $F_{\mu_n} = P_n/Q_n$ has the partial fraction expansion

$$F_{\mu_n}(z) = \sum_{i=1}^{n} A_{ni} D(\xi_{ni}, z),$$

where the A_{ni} are the weights of the quadrature formula.

Now let us consider a function f that is analytic in a neighborhood of $\mathbb{T}$. More precisely, let $\mathbb{G}$ be an open and bounded annulus such that $0 \notin \mathbb{G}$ and $\mathbb{T} \subset \mathbb{G}$, and assume that f is analytic in (a neighborhood of) the closure of $\mathbb{G}$. Let Γ_1 be the inner and Γ_2 the outer boundary of $\mathbb{G}$ so that $\partial\mathbb{G} = \Gamma_1 \cup \Gamma_2$. Then, by Cauchy's theorem we have for $t \in \mathbb{G}$

$$f(t) = \frac{1}{2\pi i}\int_\Gamma D(t,z) g(z)\,\mathrm{d}z, \quad g(z) = -f(z)/(2z),$$

where the integral over Γ runs clockwise over Γ_1 and counter-clockwise over Γ_2. Now applying the operator I_μ and using Fubini's theorem, we get

$$I_\mu\{f\} = \frac{1}{2\pi i}\int_\Gamma F_\mu(z) g(z)\,\mathrm{d}z \quad \text{and} \quad I_{\mu_n}\{f\} = \frac{1}{2\pi i}\int_\Gamma F_{\mu_n}(z) g(z)\,\mathrm{d}z.$$

Thus, for the error, one has

Theorem 2.2 (Bultheel et al. [11]). *Let f be analytic in the closure of $\mathbb{G}$, where $\mathbb{G}$ is an open and bounded annulus such that $0 \notin \mathbb{G}$ and $\mathbb{T} \subset \mathbb{G}$. Assume $\partial\mathbb{G} = \Gamma = \Gamma_1 \cup \Gamma_2$ with Γ_1 the inner and Γ_2 the outer boundary of $\mathbb{G}$. Then, if the triangular table $\mathbb{X}$, the 2PTA $F_{\mu_n} = P_n/Q_n$, and the n-point Szegő formula are as above, we have*

$$E_{\mu_n}\{f\} = I_\mu\{f\} - I_{\mu_n}\{f\} = \frac{1}{2\pi i}\int_\Gamma [F_\mu(t) - F_{\mu_n}(t)]\frac{-f(t)\,\mathrm{d}t}{2t},$$

where the integral over Γ is clockwise for Γ_1 and counter-clockwise for Γ_2, and

$$R_{\mu_n}(z) := F_\mu(z) - F_{\mu_n}(z) = \frac{2z^{p+1}}{Q_n(z)}\int_{-\pi}^{\pi}\frac{Q_n(\mathrm{e}^{\mathrm{i}\theta})\mathrm{e}^{-\mathrm{i}p\theta}}{\mathrm{e}^{\mathrm{i}\theta} - z}\,\mathrm{d}\mu(\theta).$$

This shows that there is an intimate relationship between the convergents of 2PTA for F_μ and the convergence of quadrature formulas.

We shall not develop our treatment of the polynomial case any further, but use the thread of this section as a motivation for the more general case of rational functions replacing the polynomials. We shall do this in the next sections. The polynomial situation is there just a special case.

3. Rational Szegő formulas and modified approximants

Let $\mathbb{A} = \{\alpha_n: n = 1,2,\ldots\}$ be an arbitrary sequence of points in $\mathbb{D}$. Sometimes we abuse this notation to mean also the point set of the α_k. It should be clear from the context what it is meant to be. Define the Blaschke factors ζ_k by

$$\zeta_k(z) := \frac{\bar{\alpha}_k}{|\alpha_k|}\frac{\alpha_k - z}{1 - \bar{\alpha}_k z}, \quad k = 1,2,\ldots,$$

where if $\alpha_k = 0$, then $\bar{\alpha}_k/|\alpha_k|$ is replaced by -1, and the Blaschke products $B_0 = 1$ and $B_k = \zeta_1 \cdots \zeta_k$, $k \geqslant 1$. The spaces of polynomials of the previous section are replaced by the spaces of rational functions $\mathscr{L}_n = \operatorname{span}\{B_0,\ldots,B_n\}$. Note that if we set $\pi_0 = 1$ and $\pi_n(z) = \prod_{k=1}^n (1 - \bar{\alpha}_k z)$, $n \geqslant 1$, then $f \in \mathscr{L}_n$ is of the form p/π_n with $p \in \Pi_n$. The spaces of negative powers of z are replaced by $\mathscr{L}_{n*} = \operatorname{span}\{B_0, B_{1*},\ldots,B_{n*}\}$. Thus, setting $\omega_0 = 1$ and $\omega_n(z) = \prod_{k=1}^n (z - \alpha_k)$, $n \geqslant 1$, then $f \in \mathscr{L}_{n*}$ is of the form q/ω_n with $q \in \Pi_n$. The space of L-polynomials is replaced by $\mathscr{R}_{p,q} = \mathscr{L}_{p*} + \mathscr{L}_q = \{N/(\pi_q\omega_p): N \in \Pi_{p+q}\}$. Note that if all $\alpha_k = 0$, then we are back in the situation of the polynomials and the L-polynomials as in the previous section.

Let $\hat{\mathbb{A}} = \{1/\bar{\alpha}_k: \alpha_k \in \mathbb{A}\}$. Since $\mathscr{R}_{p,q}$ is a Chebyshev system on any set $\mathbb{X} \subset \mathbb{C}\setminus(\mathbb{A} \cup \hat{\mathbb{A}})$, it follows that for any $i = 1,\ldots,n$, there is a unique rational function $L_{ni} \in \mathscr{R}_{p,q}$, $p + q = n - 1 \geqslant 0$, that satisfies $L_{ni}(\xi_{nj}) = \delta_{ij}$, where as before $\mathbb{X} = \{\xi_{ni}, i = 1,\ldots,n; n = 1,2,\ldots\}$ is a triangular table of points on $\mathbb{T}$ such that $\xi_{ni} \neq \xi_{nj}$ for $i \neq j$. Hence $f_n(z) = \sum_{i=1}^n L_{ni}(z)f(\xi_{ni})$ is the unique function in $\mathscr{R}_{p,q}$ interpolating a given function f in the points ξ_{ni}, $i = 1,\ldots,n$, and $I_{\mu_n}\{f\} := I_\mu\{f_n\} = \sum_{i=1}^n A_{ni}f(\xi_{ni})$ with $A_{ni} = I_\mu\{L_{ni}\}$ is a *quadrature formula of interpolatory type* having *domain of validity* $\mathscr{R}_{p,q}$.

Again, by an appropriate choice of the knots ξ_{ni}, we want to extend the domain of validity to make it as large as possible. As in the L-polynomial case, it can be shown that

Theorem 3.1 (Bultheel et al. [9]). *There does not exist an n-point quadrature formula of the form* (2.2) *with distinct knots on the unit circle that is exact in* $\mathscr{R}_{n-1,n}$ *or* $\mathscr{R}_{n,n-1}$.

This means that $\mathscr{R}_{n-1,n-1}$ is a candidate for a *maximal domain of validity*. As in the polynomial case, we can obtain this maximal domain of validity by choosing the abscissas as the zeros of the para-orthogonal functions. Such an optimal formula is called a *rational Szegő formula* (or R-Szegő formula for short).

Therefore we have to extend our previous notion of para-orthogonality. First we define the operation indicated by a superstar. For any $f_n \in \mathscr{L}_n \backslash \mathscr{L}_{n-1}$ ($\mathscr{L}_{-1} = \emptyset$), we set $f_n^* := B_n f_{n*}$. This generalizes the superstar conjugate for polynomials, since indeed if all α_k are zero these notions coincide. Suppose that by Gram–Schmidt orthogonalization of the Blaschke products $\{B_n\}$, we generate an orthogonal sequence $\{\varphi_n\}$. Then $0 = \langle \mathscr{L}_{n-1}, \varphi_n \rangle = \langle \phi_{n*}, \mathscr{L}_{(n-1)*} \rangle = \langle \varphi_n^*, B_n \mathscr{L}_{(n-1)*} \rangle$. Now note that $B_n \mathscr{L}_{(n-1)*} = \{f \in \mathscr{L}_n\colon f(\alpha_n) = 0\}$. Thus, if we set $\mathscr{L}_n(w) = \{f \in \mathscr{L}_n\colon f(w) = 0\}$, then $B_n \mathscr{L}_{(n-1)*} = \mathscr{L}_n(\alpha_n)$. Thus, $\varphi_n \perp \mathscr{L}_{n-1} \Leftrightarrow \varphi_n^* \perp \mathscr{L}_n(\alpha_n)$. Moreover, note that $\langle \varphi_n, B_n \rangle = \langle 1, \varphi_n^* \rangle \neq 0$. This motivates the following definition.

Definition 3.2. We say that a sequence of functions $Q_n \in \mathscr{L}_n$ is para-orthogonal if $Q_n \perp \mathscr{L}_{n-1} \cap \mathscr{L}_n(\alpha_n)$ for $n \geqslant 1$ while $\langle Q_n, 1 \rangle \neq 0$ and $\langle Q_n, B_n \rangle \neq 0$.

Definition 3.3. A function $Q_n \in \mathscr{L}_n$ is called c-invariant if $Q_n^* = cQ_n$ for some nonzero constant $c \in \mathbb{C}$.

It can be shown that any para-orthogonal and c-invariant function $Q_n \in \mathscr{L}_n$ has to be some constant multiple of $Q_n(z; \tau_n) = \varphi_n(z) + \tau_n \varphi_n^*(z)$ with $\tau_n \in \mathbb{T}$. The most important property is stated in the next theorem.

Theorem 3.4 (Bultheel et al. [9]). *Any para-orthogonal and c-invariant function from* $\mathscr{L}_n$ *has precisely n zeros*; *they are all simple and lie on* $\mathbb{T}$.

Thus, the functions $Q_n(z; \tau_n) = \varphi_n(z) + \tau_n \varphi_n^*(z)$ with $\tau_n \in \mathbb{T}$ can provide the knots for an R-Szegő formula and indeed they do, and what is more: this is the only possibility.

Theorem 3.5 (Bultheel et al. [9]). *The quadrature formula* (2.2) *with distinct knots on* $\mathbb{T}$ *is an R-Szegő formula* (*with maximal domain of validity* $\mathscr{R}_{n-1,n-1}$) *if and only if*
(a) *it is of interpolatory type in* $\mathscr{R}_{p,q}$ *with* p, q *nonnegative integers with* $p + q = n - 1$;
(b) *the abscissas are the zeros of a para-orthogonal c-invariant function from* $\mathscr{L}_n$.

Note that for each n, we have a one-parameter family of R-Szegő quadrature formulas, since the parameter $\tau_n \in \mathbb{T}$ is free.

We now introduce the reproducing kernels, since both the abscissas and the weights can be expressed in terms of these kernels. If $\{\varphi_k\}$ are the orthonormal functions, then the kernel function $k_n(z, w) = \sum_{k=0}^{n} \varphi_k(z)\overline{\varphi_k(w)}$ is reproducing for $\mathscr{L}_n$, meaning that $\langle f, k_n(\cdot, w) \rangle = f(w)$ for all $f \in \mathscr{L}_n$. As for the Szegő polynomials, these kernels appear in a Christoffel–Darboux formula.

Theorem 3.6 (Bultheel et al. [8]). *Let $\{\varphi_k\}$ be the orthonormal polynomials for the spaces $\mathscr{L}_n$. Then the reproducing kernel satisfies*

$$k_{n-1}(z,w)=\frac{\varphi_n^*(z)\overline{\varphi_n^*(w)}-\varphi_n(z)\overline{\varphi_n(w)}}{1-\zeta_n(z)\overline{\zeta_n(w)}}.$$

So far we have characterized the abscissas of the R-Szegő formulas as the zeros of the para-orthogonal functions $Q_n(z;\tau_n)$. They can also be written as the zeros of a reproducing kernel. Indeed, by varying $\tau_n\in\mathbb{T}$, we can place for example ξ_{n1} at any position $w\in\mathbb{T}$. The other $n-1$ zeros ξ_{ni} are then the zeros of $k_{n-1}(z,w)$. Conversely, if $\xi_{n1}\in\mathbb{T}$ arbitrary and $\{\xi_{ni}\}_{i=2}^n$ are the zeros of $k_{n-1}(z,\xi_{n1})$, then there is some $\tau_n\in\mathbb{T}$ such that these $\{\xi_{ni}\}_{i=1}^n$ are the zeros of the para-orthogonal function $Q_n(z;\tau_n)$ (see [7]).

Also the weights can be expressed in terms of the kernels exactly as in the polynomial case. This results in the fact that Theorem 2.1 is still true if we replace Szegő formula by R-Szegő formula (see [13]).

The 2PTA of the previous section are generalized to *multipoint rational approximants* (MRA) in the following sense. Let $Q_n=\varphi_n+\tau_n\varphi_n^*$, $\tau_n\in\mathbb{T}$, be the para-orthogonal function as above. Now define the so-called functions of the second kind $\psi_n\in\mathscr{L}_n$ as

$$\psi_n(z):=I_\mu\{E(t,z)\varphi_n(t)-D(t,z)\varphi_n(z)\},\quad E(t,z)=D(t,z)+1 \tag{3.1}$$

and set $P_n=\psi_n-\tau_n\psi_n^*$. Then it turns out that the rational function $F_{\mu_n}(\cdot;\tau_n):=F_{\mu_n}=-P_n/Q_n$ (depending on z and τ_n) is a MRA for the Riesz–Herglotz transform F_μ, since

$$zB_{n-1}(z)[F_\mu(z)-F_{\mu_n}(z)] \quad\text{and}\quad [zB_{n-1}(z)]_*[F_\mu(z)-F_{\mu_n}(z)]$$

are both analytic in $\hat{\mathbb{C}}\backslash\mathbb{T}$. This means that F_{μ_n} interpolates F_μ in the points $\{0,\alpha_1,\dots,\alpha_{n-1}\}$ and in $\{\infty,1/\bar\alpha_1\dots,1/\bar\alpha_{n-1}\}$. Note that there are $2n+1$ degrees of freedom while there are $2n$ interpolation conditions. So there is one condition short for F_{μ_n} to be a multipoint Padé approximant. These approximants are called *modified approximants* (MA) since they are modifications of the true multipoint Padé approximants (MPA) $F_{\mu_n}(\cdot;0)=\varphi_n/\psi_n$ (interpolates in all the MA points and in the extra point $1/\bar\alpha_n$) and $F_{\mu_n}(\cdot;\infty)=-\psi_n^*/\varphi_n^*$ (interpolates in all the MA points and in the extra point α_n). Furthermore, it follows from the partial fraction expansion $F_{\mu_n}(z)=\sum_{i=1}^n A_{ni}D(\xi_{ni},z)$ that $F_{\mu_n}(z)=I_{\mu_n}\{D(\cdot,z)\}$, and this relates it to the quadrature formula.

If more generally, we have a rational function F_{μ_n}, whose denominator is a node polynomial of degree n for some node array, and suppose it interpolates F_μ in the points $\{0,\alpha_1,\dots,\alpha_p\}$ and in $\{\infty,1/\bar\alpha_1\dots,1/\bar\alpha_q\}$, then we say that it is an MRA of order $(p+1,q+1)$. Thus, our MA is an MRA of order (n,n). Let $F_{\mu_n}(z)=\sum_{i=1}^n A_{ni}D(\xi_{ni},z)$ be the partial fraction decomposition of F_{μ_n} which defines the weights A_{nj} as

$$A_{nj}:=\frac{\omega_p(\xi_{nj})\pi_q(\xi_{nj})}{X_n'(\xi_{nj})}I_\mu\left\{\frac{X_n(t)}{\omega_p(t)\pi_q(t)(t-\xi_{nj})}\right\},\qquad X_n(t)=\prod_{i=1}^n(t-\xi_{ni}),$$

then it can be shown that for given $(\mathbb{A},\mathbb{X})$, the quadrature formula $I_{\mu_n}\{f\}=\sum_{i=1}^n A_{ni}f(\xi_{ni})$ is exact in $\mathscr{R}_{p,q}$ if and only if F_{μ_n} is an MRA of type $(p+1,q+1)$ with respect to the point sets $(\mathbb{A},\mathbb{X})$ for the Riesz–Herglotz transform F_μ (see [15]).

We can now work toward an expression for the error of the quadrature formula. Assume f is analytic in the closure of $\mathbb{G}$ where $\mathbb{G}$ is an open and bounded annulus such that $0 \notin \mathbb{G}$ and $\mathbb{T} \subset \mathbb{G} \subset \mathbb{C}\backslash(\mathbb{A} \cup \hat{\mathbb{A}})$. This is only possible if $\mathbb{A}$ is in a compact subset of $\mathbb{D}$, i.e., if the α_k do not tend to $\mathbb{T}$. Exactly the same type of proof as in the polynomial case can be given for the following theorem.

Theorem 3.7 (Bultheel et al. [12]). *Let f be analytic in the closure of the open and bounded annulus $\mathbb{G}$ which is such that $0 \notin \mathbb{G}$ and $\mathbb{T} \subset \mathbb{G} \subset \mathbb{C}\backslash(\mathbb{A} \cup \hat{\mathbb{A}})$. Let $\mathbb{X}$ be a given node array. Assume that $I_{\mu_n}\{f\} = \sum_{i=1}^n A_{ni} f(\xi_{ni})$ is exact in $\mathscr{R}_{p,q}$ and hence $F_{\mu_n} = \sum_{i=1}^n A_{ni} D(\xi_{ni}, \cdot)$ is an MRA of order $(p+1, q+1)$ for F_μ in the strong sense; then*

$$E_{\mu_n}\{f\} := I_\mu\{f\} - I_{\mu_n}\{f\} = \frac{1}{2\pi i} \int_\Gamma [F_\mu(t) - F_{\mu_n}(t)] \frac{-f(t)\,\mathrm{d}t}{2t}. \tag{3.2}$$

Define the node polynomial $X_n(z) := \prod_{i=1}^n (z - \xi_{ni})$; then

$$R_{\mu_n}(z) := F_\mu(z) - F_{\mu_n}(z) = \frac{2z\omega_p(z)\pi_q(z)}{X_n(z)} \int_{-\pi}^{\pi} \frac{X_n(\mathrm{e}^{\mathrm{i}\theta})\,\mathrm{d}\mu(\theta)}{(\mathrm{e}^{\mathrm{i}\theta} - z)\omega_p(\mathrm{e}^{\mathrm{i}\theta})\pi_q(\mathrm{e}^{\mathrm{i}\theta})}. \tag{3.3}$$

There is no MRA of degree n with simple poles in $\mathbb{T}$ and of order $(n+1, n)$ or order $(n, n+1)$, hence there is no quadrature formula with knots on $\mathbb{T}$ with domain of validity $\mathscr{R}_{n,n-1}$ or $\mathscr{R}_{n-1,n}$.

The only quadrature formulas exact in $\mathscr{R}_{n-1,n-1}$ with knots on $\mathbb{T}$ are the R-Szegő formulas, and hence the MA are the only MRA of order (n, n), i.e., the ones whose poles are zeros of the para-orthogonal function of degree n.

Here too, it is seen that the convergence of the quadrature formulas is closely related to the convergence of the MAs or MRAs.

By means of orthogonality properties it can be shown that for the MAs the previous error formula can be transformed into

$$R_{\mu_n}(z) = \frac{2z\omega_{n-1}(z)\pi_{n-1}(z)}{[X_n(z)]^2} \left[\int_{-\pi}^{\pi} \frac{[X_n(\mathrm{e}^{\mathrm{i}\theta})]^2\,\mathrm{d}\mu(\theta)}{(\mathrm{e}^{\mathrm{i}\theta} - z)\omega_{n-1}(\mathrm{e}^{\mathrm{i}\theta})\pi_{n-1}(\mathrm{e}^{\mathrm{i}\theta})} + D_n \right], \tag{3.4}$$

where $D_n = I_\mu\{Q_n(z)(1 - \bar{\alpha}_n z)\}$. Note that this term D_n is caused by the fact that para-orthogonality is a deficient orthogonality. When in classical formulas, zeros of orthogonal polynomials are used, then such a term does not appear.

4. Orthogonal rational functions

The quadrature formulas we consider in this paper are closely related to the properties of orthogonal and quasi-orthogonal rational functions. We collect some properties of these functions in this section. A fairly complete account of what is known about these orthogonal rational functions can be found in the monograph [19].

The orthonormal rational functions $\{\varphi_0, \varphi_1, \ldots\}$ are obtained by orthonormalization of the sequence of Blaschke products $\{B_0, B_1, \ldots, \}$. They were first considered by Djrbashian (see the references in [24]). Later on, the reproducing kernels were considered in the context of linear prediction and the

Nevanlinna–Pick interpolation problem [4]. We assume that they are normalized by the condition that in the expansion $\varphi_n(z)=\sum_{k=0}^n a_{nk}B_k(z)$, the leading coefficient with respect to this basis is positive: $\kappa_n=a_{nn}>0$. If $k_n(z,w)=\sum_{k=0}^n \varphi_k(z)\overline{\varphi_k(w)}$ is the reproducing kernel for $\mathscr{L}_n=\operatorname{span}\{B_0,\dots,B_n\}$, then it can be verified that $k_n(z,\alpha_n)=\kappa_n\varphi_n^*(z)$, where $\varphi_n^*(z)=B_n(z)\varphi_{n*}(z)$, hence also $k_n(\alpha_n,\alpha_n)=\kappa_n^2$. Although it is not essential for the present application, we mention for completeness that the φ_n satisfy a recurrence relation, i.e., there exist specific complex constants ε_n and δ_n such that $|\varepsilon_n|^2-|\delta_n|^2=(\kappa_{n-1}^2(1-|\alpha_n|^2)/\kappa_n^2(1-|\alpha_{n-1}|^2))>0$ and

$$\varphi_n(z)=\frac{\kappa_n}{\kappa_{n-1}}\left[\varepsilon_n\frac{z-\alpha_{n-1}}{1-\bar{\alpha}_n z}\varphi_{n-1}(z)+\delta_n\frac{1-\bar{\alpha}_{n-1}z}{1-\bar{\alpha}_n z}\varphi_{n-1}^*(z)\right].$$

The initial condition is $\varphi_0=1/\sqrt{c_0}$ with $c_0=\int_{\mathbb{T}}\mathrm{d}\mu$. There is also a Favard-type theorem: if some φ_n satisfy a recurrence relation of this form, then they will form an orthonormal sequence with respect to some positive measure on $\mathbb{T}$. In this respect see also the contribution by Marcellán and Alvarez in this volume.

The functions of the second kind ψ_n associated with φ_n are another independent solution of the same recurrence relation. They can also be derived from the φ_n by relation (3.1). In fact, this means that $\psi_0=\sqrt{c_0}=1/\varphi_0$ and $\psi_n=I_\mu\{D(\cdot,z)[\varphi_n(\cdot)-\varphi_n(z)]\}$ for $n\geqslant 1$. The para-orthogonal functions $Q_n(z;\tau_n)=\varphi_n(z)+\tau_n\varphi_n^*(z)$ and the associated functions of the second kind $P_n(z;\tau_n)=\psi_n(z)-\tau_n\psi_n^*(z)$ were introduced before.

Example 4.1 (Malmquist basis). Assume we take the normalized Lebesgue measure $\mathrm{d}\mu(\theta)=\mathrm{d}\theta/(2\pi)$. Then it is known that the orthonormal functions are given by

$$\varphi_n(z)=\kappa_n\frac{zB_n(z)}{z-\alpha_n},\qquad \kappa_n=\sqrt{1-|\alpha_n|^2}.$$

This basis is known as the Malmquist basis. Then $\varphi_n^*(z)=\kappa_n/(1-\bar{\alpha}_n z)$ and therefore $Q_n(z;\tau_n)=\varphi_n(z)+\tau_n\varphi_n^*(z)=\kappa_n[(zB_n(z)/z-\alpha_n)+(\tau_n/1-\bar{\alpha}_n z)]$. Noting that $B_n(z)=\eta_n\omega_n(z)/\pi_n(z)$, with $\eta_n\in\mathbb{T}$, and $\tau_n\in\mathbb{T}$ is arbitrary, we can choose $\tau_n=\eta_n$, so that the expression for $Q_n(z;\tau_n)$ becomes $Q_n(z;\tau_n)=\kappa_n\tau_n[z\omega_{n-1}(z)+\pi_{n-1}(z)]/\pi_n(z)$. If all $\alpha_k=0$, then $\omega_{n-1}(z)=z^{n-1}$ and $\pi_{n-1}(z)=1$, so that the zeros of $Q_n(z;\tau_n)$ are (a rotated version of) the nth roots of unity.

We also have to introduce the spaces $\mathscr{R}_{p,q}=\operatorname{span}\{B_{-p},\dots,B_q\}$, where p,q are nonnegative integers and $B_{-p}=1/B_p=B_{p*}$. We set $\mathscr{L}=\bigcup_{n=0}^\infty\mathscr{L}_n$ and $\mathscr{R}=\bigcup_{n=0}^\infty\mathscr{R}_{n,n}$.

Theorem 4.2 (Bultheel et al. [12]). *The space $\mathscr{L}$ is dense in $H^p(\mathbb{D})$, $1\leqslant p<\infty$, if and only if $\sum_k(1-|\alpha_k|)=\infty$.*

The space $\mathscr{R}$ is dense in $L^p(\mathbb{T})$, $1\leqslant p<\infty$, and in $C(\mathbb{T})$ if and only if $\sum_k(1-|\alpha_k|)=\infty$.

We note that the condition $\sum_k(1-|\alpha_k|)=\infty$ means that the Blaschke product B_n converges uniformly to zero in $\mathbb{D}$. Also we should have $p\neq\infty$ in this theorem, and not $1\leqslant p\leqslant\infty$ as erroneously stated in [12].

Finally, we mention the rational variant of the trigonometric moment problem. Suppose a linear functional M is defined on $\mathscr{L}$ by the moments

$$c_0 = M\{1\}, \quad c_k = M\{1/\pi_k\}, \quad \pi_k(z) = \prod_{i=1}^{k}(1 - \bar{\alpha}_k z), \ k = 1, 2, \ldots$$

and by $M\{\omega_{k*}\} = \bar{c}_k$; we can then define M on the whole of $\mathscr{R} = \mathscr{L} + \mathscr{L}_* = \mathscr{L} \cdot \mathscr{L}_*$, where $\mathscr{L}_* = \{f_*: f \in \mathscr{L}\}$. The moment problem is to find under what conditions there exists a positive measure μ on $\mathbb{T}$ such that $M\{\cdot\} = I_\mu\{\cdot\}$, and if the problem is solvable, to find conditions under which the solution is unique and possibly, if there are more solutions, to describe all of them. For a proof of the following results about moment problems, we refer to [14,18].

If it is assumed that M satisfies $M\{f_*\} = \overline{M\{f\}}$ for all $f \in \mathscr{R}$ and $M\{ff_*\} > 0$ for all nonzero $f \in \mathscr{L}$, then this functional defines an inner product $\langle f, g\rangle_M := M\{fg_*\}$. Under these conditions one can guarantee that a solution for the moment problem exists. Denote by $\mathscr{M}$ the set of all solutions. For solving the uniqueness question, the MRAs that we considered in the previous section play a central role in the solution of the problem. Recall that the MRA is given by $F_n(z;\tau_n) = -P_n(z;\tau_n)/Q_n(z;\tau_n)$, $\tau_n \in \mathbb{T}$, where $Q_n(z;\tau_n)$ are the para-orthogonal functions and $P_n(z;\tau_n)$ the associated functions of the second kind. It turns out that the set $K_n(z) = \{F_n(z;\tau): \tau \in \mathbb{T}\}$ is a circle. Moreover, the circular disks: $\Delta_n(z)$ with boundary $K_n(z)$ are nested: $\Delta_{n+1}(z) \subset \Delta_n(z)$ and their boundaries touch. The limiting set $\Delta(z) = \bigcap_n \Delta_n(z)$ will be either one point or a circular disk, and this fact is independent of the value chosen for the complex number $z \in \mathbb{C} \setminus (\mathbb{A} \cup \hat{\mathbb{A}})$. If the Blaschke product diverges, i.e., $\sum(1-|\alpha_k|) = \infty$, then the limiting set is a point and the moment problem has a unique solution (is determinate). If the limiting set $\Delta(z)$ is a disk with positive radius, then the Blaschke product converges, i.e., $\sum(1 - |\alpha_k|) < \infty$, and the moment problem has infinitely many solutions. The set $\Delta(z)$ can be characterized as $\Delta(z) = \{F_\mu(z): \mu \in \mathscr{M}\}$, where $F_\mu(z)$ denotes the Riesz–Herglotz transform of μ. A solution $\mu \in \mathscr{M}$ is called N-extremal if its Riesz–Herglotz transform $F_\mu(z)$ belongs to the boundary of $\Delta(z)$. It can be proved that $\mu \in \mathscr{M}$ is N-extremal if and only if $\mathscr{L}$ is dense in L^2_μ.

The last density result is interesting because N-extremal solutions exist if the Blaschke product converges. Thus, $\mathscr{L}$ may be dense in L^2_μ even if $\sum(1 - |\alpha_k|) < \infty$. However, if the Blaschke product diverges, then $\mathscr{L}$ will be dense in L^2_μ.

5. Convergence of MA and R-Szegő quadrature

We are still considering the case of a positive measure μ and MA's $F_{\mu_n} = P_n/Q_n$, where $Q_n = \varphi_n + \tau_n\varphi_n^*$, $\tau_n \in \mathbb{T}$, is the para-orthogonal function in $\mathscr{L}_n$ and $P_n = \psi_n - \tau_n\psi_n^*$ is its associated function. Also $I_{\mu_n}\{f\}$ is the nth R-Szegő formula. We have seen that the convergence of $I_{\mu_n}\{f\}$ is related to the convergence of F_{μ_n}. That F_{μ_n} does converge is essentially a consequence of the Stieltjes–Vitali theorem.

Theorem 5.1 (Bultheel et al. [8,15]). *If* $\sum_{n=1}^{\infty}(1 - |\alpha_n|) = \infty$, *then the MA's* $F_{\mu_n}(z;\tau_n)$ *converge to* $F_\mu(z)$ *uniformly on compact subsets of* $\hat{\mathbb{C}} \setminus \mathbb{T}$.

Remark. Because the Stieltjes–Vitali theorem is used, the proof of the above theorem is not constructive. It is an interesting open problem to give a constructive proof using continued fraction methods [34].

To estimate the rate of convergence, we recall that $Q_n = X_n/\pi_n$ and $|B_n| = |\omega_n/\pi_n|$, so that from (3.4)

$$|F_{\mu_n}(z)| \leqslant \left|\frac{\omega_{n-1}(z)}{\pi_n(z)}\right| \left|\frac{\pi_{n-1}(z)}{\pi_n(z)}\right| \frac{2|z|}{|Q_n(z)|^2} \left[I_\mu \left\{ \left|\frac{\pi_n}{\pi_{n-1}}\right| \left|\frac{\pi_n}{\omega_{n-1}}\right| \frac{|Q_n|^2}{|\cdot - z|} \right\} + |D_n| \right]$$

and hence there is a constant M not depending on n such that

$$|F_{\mu_n}(z)|^{1/n} \leqslant M^{1/n}[|B_{n-1}(z)||1-\bar{\alpha}_n z|^2 |Q_n(z)|^2]^{-1/n}(S_n + |D_n|)^{1/n}$$

with $S_n = ||Q_n^2||_\infty \max_{t\in\mathbb{T}}|1-\bar{\alpha}_n t|^2$. This explains why we need the root asymptotics of Q_n and B_n to estimate the rate of convergence for F_{μ_n}. Therefore we need some assumptions on μ and the point set $\mathbb{A} = \{\alpha_1, \alpha_2, \ldots\}$. For the given set $\mathbb{A}$, let $v_n^{\mathbb{A}} = 1/n \sum_{j=1}^n \delta(\alpha_j)$ be the counting measure, which assigns a mass at α_j proportional to its multiplicity. Assume that $v_n^{\mathbb{A}}$ converges to some $v^{\mathbb{A}}$ in the weak star sense, in the dual of the Banach space $C(\hat{\mathbb{C}})$, where $\hat{\mathbb{C}}$ is the Riemann sphere. Thus $\lim_{n\to\infty}\int f\,\mathrm{d}v_n^{\mathbb{A}} = \int f\,\mathrm{d}v^{\mathbb{A}}$ for all $f \in C(\hat{\mathbb{C}})$. We denote this as $v_n^{\mathbb{A}} \overset{*}{\to} v^{\mathbb{A}}$. Then the root asymptotics for the Blaschke products are given by:

Theorem 5.2 (Bultheel et al. [15]). *If B_n is the Blaschke product with zeros $\{\alpha_k\}_{k=1}^n$, and $v_n^{\mathbb{A}} \overset{*}{\to} v^{\mathbb{A}}$, then*

$$\lim_{n\to\infty} |B_n(z)|^{1/n} = \exp\{\lambda(z)\} \quad \textit{and} \quad \lim_{n\to\infty} |B_n(z)|^{-1/n} = \exp\{\lambda(\hat{z})\}$$

locally uniformly for $z \in \hat{\mathbb{C}} \backslash (\{0\} \cup \operatorname{supp}(v^{\mathbb{A}}) \cup \operatorname{supp}(v^{\hat{\mathbb{A}}}))$, where

$$\lambda(z) = \int \log|\zeta_z(x)|\,\mathrm{d}v^{\mathbb{A}}(x), \quad \zeta_z(x) = \frac{x-z}{1-\bar{z}x} \tag{5.1}$$

and where $\hat{z} = 1/\bar{z}$ and $\hat{\mathbb{A}} = \{\hat{\alpha} = 1/\bar{\alpha}\colon \alpha \in \mathbb{A}\}$. For $z \in \mathbb{C}\backslash\{0\}$ we have the inequalities

$$\limsup_{n\to\infty} |B_n(z)|^{1/n} \leqslant \exp\{\lambda(z)\} \quad \textit{and} \quad \limsup_{n\to\infty} |B_n(z)|^{-1/n} \leqslant \exp\{\lambda(\hat{z})\}.$$

As for the root asymptotics of the para-orthogonal functions, one has

Theorem 5.3 (Bultheel et al. [15]). *Let μ be a positive measure satisfying the Szegő condition $\int_{-\pi}^{\pi} \log \mu'(\theta)\,\mathrm{d}\theta > -\infty$ and assume that the point set $\mathbb{A}$ is compactly included in $\mathbb{D}$ and that $v_n^{\mathbb{A}} \overset{*}{\to} v^{\mathbb{A}}$. Then, for the para-orthogonal functions Q_n, it holds locally uniformly in the indicated regions that*

$$\lim_{n\to\infty} |Q_n(z)|^{1/n} = 1, \quad z \in \mathbb{D},$$

$$\lim_{n\to\infty} |Q_n(z)|^{1/n} = \exp\{\lambda(z)\}, \quad z \in \mathbb{E}\backslash\operatorname{supp}(v^{\hat{\mathbb{A}}}),$$

$$\limsup_{n\to\infty} |Q_n(z)|^{1/n} \leqslant \exp\{\lambda(z)\}, \quad z \in \mathbb{E},$$

$$\lim_{n\to\infty} ||Q_n(z)||_\infty^{1/n} = 1.$$

A combination of the previous results now leads to the rate of convergence for the MA's.

Theorem 5.4 (Bultheel et al. [15]). *Under the same conditions as in the previous theorem, the following estimates hold for the convergence of the MAs F_{μ_n} to the Riesz–Herglotz transform F_μ. Setting $R_{\mu_n} = F_\mu - F_{\mu_n}$, we have*

$$\limsup_{n\to\infty} |R_{\mu_n}(z)|^{1/n} \leqslant \exp\{\lambda(z)\} < 1, \quad \forall z \in \mathbb{D},$$

$$\limsup_{n\to\infty} |R_{\mu_n}(z)|^{1/n} \leqslant \exp\{\lambda(\hat{z})\} < 1, \quad \forall z \in \mathbb{E}, \ \textit{where } \hat{z} = 1/\bar{z},$$

and $\lambda(z)$ as in (5.1).

Example 5.5. Consider the simple case where $\lim_{k\to\infty} \alpha_k = a \in \mathbb{D}$. Then $v^{\mathbb{A}}(z) = \delta_a$ and $\lambda(z) = \log|\zeta_z(a)|$, $\zeta_z(a) = (a - z)/(1 - \bar{z}a)$. Therefore, $\limsup_{n\to\infty} |R_{\mu_n}(z)|^{1/n} \leqslant |\zeta_z(a)| < 1$ for $z \in \mathbb{D}$ and $\limsup_{n\to\infty} |R_{\mu_n}(z)|^{1/n} \leqslant 1/|\zeta_z(a)| < 1$ for $z \in \mathbb{E}$. The best rates of convergence are obtained for z near a and $\hat{a} = 1/\bar{a}$, as one could obviously expect.

Similar results hold for the true MPAs:

Theorem 5.6 (Bultheel et al. [15]). *Under the same conditions as in the previous theorem, the following estimates hold for the convergence of the MPAs $F_n = \psi_n/\varphi_n$ and $F_n^\times = -\psi_n^*/\varphi_n^*$ to the Riesz–Herglotz transform F_μ. Set $R_n = F_\mu - F_n$ and $R_n^\times = F_\mu - F_n^\times$; then locally uniformly in the indicated regions*:

$$\limsup_{n\to\infty} |R_n^\times(z)|^{1/n} \leqslant \exp\{\lambda(z)\} < 1, \quad \forall z \in \mathbb{D},$$

$$\limsup_{n\to\infty} |R_{\mu_n}(z)|^{1/n} \leqslant \exp\{\lambda(\hat{z})\} < 1, \quad \forall z \in \mathbb{E}, \ \textit{where } \hat{z} = 1/\bar{z},$$

and $\lambda(z)$ as in (5.1).

Now we can move on to the convergence of the R-Szegő formulas. This is a direct consequence of the previous analysis. For example, we get from (3.2) that

$$|E_{\mu_n}\{f\}| \leqslant \frac{1}{4\pi} \max_{t\in\Gamma} \left|\frac{f(t)}{t}\right| \int_\Gamma |F_\mu(t) - F_{\mu_n}(t)||\mathrm{d}t|.$$

Therefore, it follows under the conditions of Theorems 3.7 and 5.4 that the R-Szegő quadrature formula converges to the integral for all functions analytic in the closure of $\mathbb{G}$ with the annulus $\mathbb{G}$ as above in Theorem 3.7. For this situation, we can even obtain an estimate for the rate of convergence that relies on the previous estimates.

Theorem 5.7 (Bultheel et al. [15]). *Let $I_{\mu_n}\{f\}$ be the R-Szegő formula for a function f that is analytic in the closure of the open bounded annulus $\mathbb{G}$ such that $0 \notin \mathbb{G}$ and $\mathbb{T} \subset \mathbb{G} \subset \mathbb{C}\backslash(\mathbb{A} \cup \hat{\mathbb{A}})$. Then under the conditions of Theorem* 5.4,

$$\limsup_{n\to\infty} |I_\mu\{f\} - I_{\mu_n}\{f\}|^{1/n} \leqslant \gamma < 1,$$

where $\gamma = \max\{\gamma_1, \gamma_2\}$, with $\gamma_1 = \max_{z\in\Gamma\cap\mathbb{D}} \exp\{\lambda(z)\}$ and $\gamma_2 = \max_{z\in\Gamma\cap\mathbb{E}} \exp\{\lambda(\hat{z})\}$, where $\hat{z} = 1/\bar{z}$ and $\lambda(z)$ as in (5.1).

To prove convergence for the broader class of continuous functions $f \in C(\mathbb{T})$, we define $\gamma_n(f) := \inf_{f_n \in \mathscr{R}_{n,n}} \|f - f_n\|_\infty$. By Theorem 4.2, $\lim_{n\to\infty} \gamma_n(f) = 0$ if $\sum_k (1 - |\alpha_k|) = \infty$. Assume that $r_{n-1} \in \mathscr{R}_{n-1,n-1}$ is such that $\|f - r_{n-1}\|_\infty = \gamma_{n-1}(f)$. If we take into account that $I_{\mu_n}\{r_{n-1}\} = I_\mu\{r_{n-1}\}$, then it follows that

$$|E_{\mu_n}\{f\}| = |I_\mu\{f - r_n\} + I_{\mu_n}\{r_{n-1} - f\}| \leqslant \gamma_{n-1}(f)[I_\mu\{1\} + I_{\mu_n}\{1\}].$$

Thus, $|E_{\mu_n}\{f\}| \leqslant C_1 \gamma_{n-1}(f)$, with C_1 a constant. So it follows from the convergence of γ_n that also the R-Szegő formula converges for continuous functions. If we then also take into account the standard argument proving that, if a quadrature formula with positive coefficients converges for continuous functions, then it also converges for bounded integrable functions, we arrive at

Theorem 5.8. *The R-Szegő formulas $I_{\mu_n}\{f\}$ converge for any bounded integrable function f and positive μ if $\sum_n (1 - |\alpha_n|) = \infty$.*

Thus, we have obtained convergence in the largest possible class.

With the help of [12, Theorem 4.7], it can be shown that γ_n can be bounded in terms of the modulus of continuity $\omega(f, \delta) = \sup\{|f(t) - f(\tau)|: t, \tau \in \mathbb{T}, |\mathrm{Arg}(t/\tau)| < \delta\}$. So, there exists a constant C_2 such that $|E_{\mu_n}\{f\}| \leqslant C_2 \omega(f, \pi/(n+1))$ for n large enough.

6. The case of a complex measure

If the measure μ is complex (not real), then we cannot guarantee the existence of a sequence of orthogonal rational functions. In that case we can choose an arbitrary auxiliary positive measure v on $\mathbb{T}$ and compute the knots of the quadrature formula as the zeros of a para-orthogonal function for this measure. The obvious question is what would be a good choice for this auxiliary measure. Choosing the Lebesgue measure as v would lead to equidistant nodes on $\mathbb{T}$. There are few other examples of measures that lead to explicit expressions for the knots. In general, they must be computed numerically. If we are prepared to do this, then we could choose the measure v as a function of the convergence behavior of the quadrature formulas.

In this case we shall consider absolutely continuous measures. So, let $\mathrm{d}\mu(\theta) = \rho(\theta)\,\mathrm{d}\theta$ and $\mathrm{d}v(\theta) = \omega(\theta)\,\mathrm{d}\theta$. We assume $\omega(\theta) > 0$, $\int_{-\pi}^{\pi} |\rho(\theta)|\mathrm{d}\theta < \infty$, and $\rho/\omega \in L^2_\omega$, i.e.,

$$\int_{-\pi}^{\pi} \frac{|\rho(\theta)|^2}{\omega(\theta)}\,\mathrm{d}\theta = K^2 < \infty. \tag{6.1}$$

We are concerned with the computation of integrals of the form $I_\rho\{f\} = \int_{-\pi}^{\pi} f(\mathrm{e}^{\mathrm{i}\theta})\rho(\theta)\,\mathrm{d}\theta$ approximated by $I_{\rho_n}\{f\} = \sum_{i=1}^{n} W_{ni} f(\xi_{ni})$, where $\mathbb{X} = \{\xi_{ni}\} \subset \mathbb{T}$ is a node array.

Inspired by the results of the previous sections, our first guess is to choose the knots as the zeros of the para-orthogonal functions for the positive weight ω and construct interpolatory quadrature formulas in a subspace $\mathscr{R}_{p,q}$ of dimension n. For this kind of quadrature formulas, we can show that the coefficients do not grow too fast: There is an absolute constant (i.e., not depending on n) C_3 such that $\sum_{j=1}^{n} |W_{nj}| \leqslant C_3\sqrt{n}$. Then, by using a rational generalization of the Jackson III theorem (see [12, Theorem 4.7]), it can be shown that the following convergence result holds.

Theorem 6.1 (Bultheel et al. [12]). *Let ω and ρ be as in* (6.1). *Let $f \in C(\mathbb{T})$ with modulus of continuity $\omega(f,\delta) = O(\delta^p)$, $p > \frac{1}{2}$. Let $p(n)$ and $q(n)$ be nonnegative integers with $p(n)+q(n)=n-1$ and $\lim_{n\to\infty} p(n)/n = \frac{1}{2}$. Then the interpolatory quadrature formulas $I_{\rho_n}\{f\}$ whose knots are the zeros of the para-orthogonal function for ω and which are exact in $\mathscr{R}_{p(n),q(n)}$ converge to $I_\rho\{f\}$.*

Note that we need $p > \frac{1}{2}$ so that convergence in $C(\mathbb{T})$ is not proved.

To get convergence for a larger class, we consider N-point quadrature formulas of interpolatory type in $\mathscr{R}_{n,n}$ with $N = 2n + 1$. The basic idea for constructing such quadrature formulas is the following. Define $g(\mathrm{e}^{\mathrm{i}\theta}) := \rho(\theta)/\omega(\theta)$; then it is clear that

$$I_\rho\{f\} = \int_{-\pi}^{\pi} f(\mathrm{e}^{\mathrm{i}\theta})\rho(\theta)\,\mathrm{d}\theta = \int_{-\pi}^{\pi} f(\mathrm{e}^{\mathrm{i}\theta})g(\mathrm{e}^{\mathrm{i}\theta})\omega(\theta)\,\mathrm{d}\theta = I_\omega\{fg\}.$$

Now ω is positive and we can apply our previous theory of R-Szegő formulas. However, the integrand is now a product fg. If we want equality $I_\rho\{f\} = I_\rho^N\{f\}$, for $f \in \mathscr{R}_{n,n}$, then we must be able to integrate fg exactly for $f \in \mathscr{R}_{n,n}$. It can be shown that $I_\omega\{fg\} = I_\omega\{fg_{2n}\}$ for all $f \in \mathscr{R}_{n,n}$ if g_{2n} is the orthogonal projection of g onto $\mathscr{R}_{n,n}$ in L^2_ω [17]. Thus, it is sufficient to construct quadrature formulas exact in $\mathscr{R}_{n,n} \cdot \mathscr{R}_{n,n}$ so that $I_\omega\{fg\}$ and hence also $I_\rho\{f\}$ can be computed exactly for all $f \in \mathscr{R}_{n,n}$. Note that by the product $\mathscr{R}_{n,n} \cdot \mathscr{R}_{n,n}$, we double each pole.

Therefore, we associate with the sequence $\mathbb{A} = \{\alpha_k\}$ the doubling sequence $\tilde{\mathbb{A}} = \{0, \alpha_1, \alpha_1, \alpha_2, \alpha_2, \ldots\}$, denoted as $\{\tilde{\alpha}_0, \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4, \ldots\}$. This doubling sequence can be used in exactly the same way as before to define Blaschke products $\tilde{B}_n$ and spaces $\tilde{\mathscr{L}}_n$, and orthogonal rational functions $\tilde{\varphi}_n$. The para-orthogonal rational functions $\tilde{Q}_n(z;\tau_n) = \tilde{\varphi}_n + \tau_n\tilde{\varphi}_n^*$ with $\tau_n \in \mathbb{T}$ and $\tilde{\varphi}_n^* = \tilde{B}_n\tilde{\varphi}_{n*}$ have n simple zeros ξ_{ni}, $i = 1, \ldots, n$, that can be used to construct R-Szegő quadrature formulas. Now set $N = 2n + 1$ and let $\tilde{I}_\omega^N\{f\}$ be the R-Szegő formula that is exact in $\tilde{R}_{N-1,N-1} = \tilde{\mathscr{L}}_{N-1} \cdot \tilde{\mathscr{L}}_{(N-1)*}$. Since $F \in \tilde{\mathscr{R}}_{N-1,N-1} \Leftrightarrow F = fg$ with $f, g \in \mathscr{R}_{n,n}$ we have reached our objective. We have

Theorem 6.2 (Bultheel et al. [17]). *As in* (6.1), *let ρ be complex, ω positive, and $g(\mathrm{e}^{\mathrm{i}\theta}) = \rho(\theta)/\omega(\theta) \in L^2_\omega(\mathbb{T})$. For $N = 2n + 1$, let $\{\xi_{Nj}\}_{j=1}^N$ be the zeros of the para-orthogonal function from $\tilde{\mathscr{L}}_N$ associated with the doubling sequence $\tilde{\mathbb{A}}$ and the weight ω. Moreover, let $\tilde{A}_{Nj}$ be the weights of the corresponding N-point R-Szegő formula, exact in $\tilde{\mathscr{R}}_{N-1,N-1}$. Then the quadrature formula $I_\rho^N\{f\} = \sum_{i=1}^N W_{Nj} f(\xi_{Nj})$ computes the integral $I_\rho\{f\}$ exactly for all $f \in \mathscr{R}_{n,n}$ if the weights are given by $W_{Nj} = \tilde{A}_{Nj} g_{2n}(\xi_{Nj})$, where g_{2n} is the projection of g onto $\mathscr{R}_{n,n}$ in $L^2_\omega(\mathbb{T})$.*

This defines the quadrature formulas $I_{\rho_n}\{f\}$. Now to prove convergence, we use a rational extension of the classical Erdős–Turán theorem: if $f_N \in \mathscr{R}_{n,n}$ interpolates f in the points $\{\xi_{Nk}\}_{k=1}^N$, then, under the conditions given in Theorem 6.2, f_N converges to f in $L^2_\omega(\mathbb{T})$. Using the bound $\sum_{k=1}^N |W_{Nk}| < C_3\sqrt{n}$, the Cauchy–Schwarz inequality, and a rational generalization of [43, Theorem 1.5.4], we get the following convergence result.

Theorem 6.3 (Bultheel et al. [17]). *Assume the same conditions and the same interpolatory quadrature formulas as in Theorem* 6.2. *If, moreover, $\sum_{j=1}^\infty (1 - |\alpha_j|) = \infty$, then the following convergence results hold.*

For any bounded f for which $I_\rho\{f\} < \infty$ exists as a Riemann integral, $I_\rho^N\{f\}$ converges to $I_\rho\{f\}$.

For any bounded Riemann integrable f, $\sum_{j=1}^{N} |W_{Nj}| f(\xi_{Nj})$ converges to $\int_{-\pi}^{\pi} f(\mathrm{e}^{\mathrm{i}\theta})|\rho(\theta)|\mathrm{d}\theta$.

The MRA $F_N(z) = I_\rho^N\{D(\cdot,z)\}$ interpolates the Riesz–Herglotz transform $F_\rho(z) = I_\rho\{D(\cdot,z)\}$ at the points $\{0,\alpha_1,\ldots,\alpha_n\}$ and $\{\infty,1/\bar{\alpha}_1,\ldots,1/\bar{\alpha}_n\}$, and it converges to F_ρ uniformly on compact subsets of $\hat{\mathbb{C}}\backslash\mathbb{T}$.

The previous results are related to interpolatory quadrature formulas in $\mathscr{R}_{n,n}$. We can, however, generalize to the asymmetric case and consider more general spaces $\mathscr{R}_n = \mathscr{R}_{p(n),q(n)}$, where $p(n)$ and $q(n)$ are nondecreasing sequences of nonnegative integers such that $p(n) + q(n) = n - 1$ and $\lim_{n\to\infty} p(n)/n = r \in (0,1)$. Note that the spaces $\mathscr{R}_n$ have dimension n and they are nested. As before, we need to introduce an asymmetric doubling sequence as follows. Set $r(n) = \max\{p(n),q(n)\}$, $s(n) = \min\{p(n),q(n)\}$, $\alpha_0 = 0$, $\tilde{\mathbb{A}}_n = \{\alpha_0,\alpha_1,\alpha_1,\ldots,\alpha_{s(n)},\alpha_{s(n)},\alpha_{s(n)+1},\ldots,\alpha_{r(n)}\} = \{\tilde{\alpha}_0,\tilde{\alpha}_1,\ldots,\tilde{\alpha}_{n-1}\}$. Since increasing n to $n+1$ increases either $p(n)$ or $q(n)$ by one, this increases either $r(n)$ or $s(n)$ by one. The numbering of the $\tilde{\alpha}_k$ is such that $\tilde{\alpha}_n$ is either a repeated point $\alpha_{s(n)+1}$ or a new point $\alpha_{r(n)+1}$. This defines the sequence $\tilde{\mathbb{A}} = \{\tilde{\alpha}_1,\tilde{\alpha}_1,\ldots\}$ uniquely. The quantities related to the $\tilde{\mathbb{A}}$ are as before denoted with a tilde. We construct quadrature formulas whose nodes are the zeros ξ_{ni} of the para-orthogonal function $\tilde{Q}_n(z;\tau_n) = \tilde{\varphi}_n(z) + \tau_n\tilde{\varphi}_n^*(z)$. The $\tilde{\varphi}_n \in \tilde{\mathscr{L}}_n\backslash\tilde{\mathscr{L}}_{n-1}$ are the orthogonal functions with respect to the positive measure ω with the properties introduced before. The weights W_{nk} of these quadrature formulas $I_{\rho_n}\{f\} = \sum_{k=1}^{n} W_{nk} f(\xi_{nk})$ are constructed such that $I_{\rho_n}\{f\}$ is exact in $\mathscr{R}_n = \mathscr{R}_{p(n),q(n)}$ of dimension n. With this setting, one can follow the same approach as in Theorem 5.4, but now the MAs are replaced by MRAs of order $(p(n)+1,q(n)+1)$.

Theorem 6.4 (Bultheel et al. [20]). *Assume $\int_{-\pi}^{\pi} \log\omega(\theta)\,\mathrm{d}\theta > -\infty$ and let the sequence $\mathbb{A}$, hence also $\tilde{\mathbb{A}}$, be included in a compact subset of $\mathbb{D}$. Denote by F_{ρ_n} the MRA of order $(p(n)+1,q(n)+1)$ to the Riesz–Herglotz transform F_ρ. The denominator of F_{ρ_n} is $\prod_{k=1}^{n}(z-\xi_{nk})$, where the ξ_{nk} are the zeros of the para-orthogonal functions in $\tilde{\mathscr{L}}_n$ with respect to the sequence $\tilde{\mathbb{A}}$ and the positive function ω. This sequence $\tilde{\mathbb{A}}$ is defined as above in terms of the sequence $\mathbb{A}$ and the sequences of integers $(p(n),q(n))$, $p(n)+1(n) = n-1$. The functions ρ and ω satisfy* (6.1). *Then the following convergence results hold*:

$$\limsup_{n\to\infty} |F_\rho(z) - F_{\rho_n}(z)|^{1/n} \leqslant \exp\{r\lambda(z)\} < 1, \quad \forall z \in \mathbb{D},$$

$$\limsup_{n\to\infty} |F_\rho(z) - F_{\rho_n}(z)|^{1/n} \leqslant \exp\{s\lambda(\hat{z})\} < 1, \quad \forall z \in \mathbb{E} \text{ where } \hat{z} = 1/\bar{z},$$

where $r = \lim_{n\to\infty} p(n)/n$, $s = \lim_{n\to\infty} q(n)/n = 1 - r$, and $\lambda(z)$ as in (5.1).

From this theorem, the following theorem follows directly by using Theorem 2.2.

Theorem 6.5 (Bultheel et al. [20]). *Under the same conditions as in the previous theorem, assume the quadrature formulas $I_{\rho_n}\{f\}$ have nodes $\{\xi_{nk}\}_{k=1}^{n}$ and their weights are defined such that the formulas are exact in $\mathscr{R}_{p(n),q(n)}$ of dimension n. Then it holds that*

$$\limsup_{n\to\infty} |I_\rho\{f\} - I_{\rho_n}\{f\}|^{1/n} \leqslant \gamma < 1$$

for any function f analytic in a closed region $\mathbb{G}$ *such that* $\mathbb{T} \subset \mathbb{G} \subset \mathbb{C}\backslash(\mathbb{A}\cup\hat{\mathbb{A}})$, *where* $\gamma = \max\{\gamma_1, \gamma_2\}$ *with*

$$\gamma_1 = \max_{z\in\Gamma\cap\mathbb{D}} \exp\{r\lambda(z)\} \quad \textit{and} \quad \gamma_2 = \max_{z\in\Gamma\cap\mathbb{E}} \exp\{s\lambda(\hat{z})\},$$

$\lambda(z)$ *as in* (5.1), $\Gamma = \partial\mathbb{G}$ *is the boundary of* $\mathbb{G}$ *consisting of finitely many rectifiable curves*, $r = \lim_{n\to\infty} p(n)/n$, $s = \lim_{n\to\infty} q(n)/n = 1 - r$, $\hat{z} = 1/\bar{z}$, $\hat{\mathbb{A}} = \{\hat{\alpha} = 1/\bar{\alpha}\colon \alpha \in \mathbb{A}\}$.

Note that if we take the balanced situation, i.e., when $r = s = \frac{1}{2}$, then from this theorem it follows that $\gamma = \sqrt{\tilde{\gamma}}$ with $\tilde{\gamma} = \max\{\tilde{\gamma}_1, \tilde{\gamma}_2\}$, where $\tilde{\gamma}_1 = \max_{z\in\Gamma\cap\mathbb{D}} \exp\{\lambda(z)\}$ and $\tilde{\gamma}_2 = \max_{z\in\Gamma\cap\mathbb{E}} \exp\{\lambda(\hat{z})\}$. If we assume that ρ is positive, then we can take $\omega = \rho$ and $\hat{\mathbb{A}} = \mathbb{A}$, so that the quadrature formulas then considered in this theorem are precisely the R-Szegő formulas. This result is confirmed by Theorem 5.6, where indeed the bound $\tilde{\gamma}$ is given. The squaring $\gamma = \tilde{\gamma}^2$ is of course to be expected.

7. Poles in the support of the measure

So far, we have assumed that the poles of the rational functions were outside the support of the measure. If the poles are selected in the support, then we can refer to the theory of orthogonal rational functions with poles on $\mathbb{T}$ when we want to compute integrals over $\mathbb{T}$. This theory is analogous and yet different from what was explained in Sections 3–6. It generalizes the differences that also exist between polynomials orthogonal on the real line and polynomials orthogonal on the unit circle.

So instead of choosing points α_k inside $\mathbb{D}$, we choose them all on the boundary $\mathbb{T}$. We need to define one exceptional point on $\mathbb{T}$ that is different from all α_k. We assume without loss of generality that it is -1. So $\mathbb{A} = \{\alpha_1, \alpha_2, \ldots\} \subset \mathbb{T}\backslash\{-1\}$. We consider the spaces $\mathscr{L}_n = \operatorname{span}\{1/\omega_0, 1/\omega_1, \ldots, 1/\omega_n\}$, where $\omega_k(z) = \prod_{i=1}^{k}(z - \alpha_i)$ as before. The theory can be developed along the same lines, but it is a bit more involved. We use the same notation where possible. Now it is important that if $\varphi_n(z) = p_n(z)/\pi_n(z)$, then $p_n(\alpha_{n-1}) \neq 0$. If this holds, then φ_n is called regular, and the system $\{\varphi_n\}$ is regular if every function in the system is regular. It is for such a regular system that one can prove that the orthogonal functions satisfy a recurrence relation of the following form [10]. For $n = 2, 3, \ldots$ and with $\alpha_0 = 0$,

$$\varphi_n(z) = \frac{A_n}{z - \alpha_n}\varphi_{n-1}(z) + B_n\frac{z - \alpha_{n-2}}{z - \alpha_n}\varphi_{n-1}(z) + C_n\frac{z - \alpha_{n-2}}{z - \alpha_n}\varphi_{n-2}(z).$$

These constants satisfy $A_n + B_n(\alpha_{n-1} - \alpha_{n-2}) \neq 0$ and $C_n \neq 0$ for $n = 2, 3, \ldots$.

The para-orthogonal functions are in this case replaced by quasi-orthogonal functions. These are defined as

$$Q_n(z; \tau_n) := \varphi_n(z) + \tau_n\frac{(1 + \alpha_n)(z - \alpha_{n-1})}{(1 + \alpha_{n-1})(z - \alpha_n)}\varphi_{n-1}(z), \quad \tau_n \in \mathbb{R}.$$

We have:

Theorem 7.1 (Bultheel et al. [10]). *If the system* $\{\varphi_n\}$ *is regular, then it is always possible to find* (*infinitely many so-called regular values*) $\tau_n \in \mathbb{R}$ *such that the quasi-orthogonal functions* $Q_n(z; \tau_n)$ *have precisely n zeros, all simple and on* $\mathbb{T}\backslash\{\alpha_1, \ldots, \alpha_n\}$.

Let $\{\xi_{nk}\}_{k=1}^{n}$ be these zeros. If we take them as knots for an interpolating quadrature formula for $\mathscr{L}_{n-1}$, then this quadrature formula will have positive weights and it will be exact in $\mathscr{L}_{n-1}\cdot\mathscr{L}_{n-1}$. If $\tau_n=0$ is a regular value, then the corresponding quadrature formula is exact in $\mathscr{L}_{n-1}\cdot\mathscr{L}_{n}$.

These quadrature formulas are the analogs of the R-Szegő formulas. We shall denote them again by $I_{\mu_n}\{f\}=\sum_{k=1}^{n}A_{nk}f(\xi_{nk})$. In fact, one can, exactly as for the R-Szegő formulas, express A_{nk} and ξ_{nk} in terms of the reproducing kernels. Note that A_{nk} and ξ_{nk} depend as before on the choice of τ_n. Then it can be shown that $I_{\mu_n}\{D(\cdot,z)\}=F_{\mu_n}(z)=-P_n(z;\tau_n)/Q_n(z;\tau_n)$. This is a rational approximant for the Riesz–Herglotz transform $F_\mu(z)=I_\mu\{D(\cdot,z)\}$ in a particular sense. Indeed, $F_{\mu_n}(z)$ and $F_\mu(z)$ are defined for $z\in\mathbb{D}$, but by extending them by a nontangential limit to the boundary $\mathbb{T}$, we have interpolation in $\{0,\infty,\alpha_1,\alpha_1,\ldots,\alpha_{n-1},\alpha_{n-1}\}$ (repeated points imply interpolation in the Hermite sense). In case $\tau_n=0$ is a regular value, then $F_{\mu_n}=\psi_n/\varphi_n$, with ψ_n as before the functions of the second kind associated with φ_n. Then this F_{μ_n} will also interpolate in the extra point α_n. It can be shown that if $\{\varphi_n\}$ is a regular system, then there is a subsequence $F_{\mu_{n(s)}}$ that converges to F_μ locally uniformly in $\mathbb{C}\backslash\mathbb{T}$. However, convergence has been explored only partially, and here is a wide-open domain for future research.

8. Integrals over an interval

By conformally mapping the unit circle to the real line, we can obtain analogous results on the real line. The results have different formulations, but they are essentially the same as the ones we gave in the previous sections. We consider instead some other quadrature formulas that were derived, making use of rational functions. We restrict ourselves in the first place to a compact interval $\varDelta$ on the real line, which we can always renormalize to be $[-1,1]$. It will be assumed everywhere in this section that $\varDelta$ denotes this interval.

So we now consider measures that are supported on an interval of the real line and we assume that this interval is $\operatorname{supp}(\mu)\subset\varDelta=[-1,1]$. The problem is to approximate the integral $\int_\varDelta f(x)\,\mathrm{d}x$. Several quadrature formulas of the form $\sum_{k=1}^{n}A_kf(x_k)$, exact for other functions than polynomials have been considered in the literature before. We shall discuss formulas exact for spaces of rational functions with prescribed poles outside $\varDelta$. For more general cases, see the classical book [23, p. 122] and references therein.

Consider a positive measure μ. In [26], Gautschi considers the following problem. Let α_k, $k=1,\ldots,M$ be distinct numbers in $\mathbb{C}\backslash\varDelta$. For given integers m and n, with $1\leqslant m\leqslant 2n$, find an n-point quadrature formula exact for all monomials x^j, $j=0,\ldots,2n-m-1$, as well as for the rational functions $(x-\alpha_k)^{-s}$, $k=1,\ldots,M$, $s=1,\ldots,s_k$, with $s_k\geqslant 1$ and $\sum_{k=1}^{M}s_k=m$. The solution is given in the following theorem (which is also valid for an unbounded support $\varDelta$).

Theorem 8.1 (Gautschi [26]). *Given a positive measure μ with $\operatorname{supp}(\mu)\subset\varDelta\subset\mathbb{R},\{\alpha_k\}_{k=1}^{M}\subset\mathbb{C}\backslash\varDelta$, and positive integers s_k, $\sum_{k=1}^{M}s_k=m$, and define $\omega_m(x):=\prod_{k=1}^{M}(x-\alpha_k)^{s_k}$, a polynomial of degree m. Assume that the measure $\mathrm{d}\mu/\omega_m$ admits an (polynomial) n-point Gaussian quadrature formula, i.e., there are $\xi_j^G\in\varDelta$ and $A_j^G>0$ such that*

$$\int_\varDelta f(x)\frac{\mathrm{d}\mu(x)}{\omega_m(x)}=\sum_{j=1}^{n}A_j^Gf(\xi_j^G)+E_n^G\{f\}\quad \text{with } E_n^G\{f\}=0,\quad \forall f\in\Pi_{2n-1}.$$

Define $\xi_j{:=}\xi_j^G$ *and* $A_j{:=}A_j^G\omega_m(\xi_j^G)$, $j=1,\dots,n$. *Then* $\int_\Delta f(x)\,\mathrm{d}\mu(x)=\sum_{j=1}^n A_j f(\xi_j)+E_n\{f\}$, *where* $E_n\{f\}=0$ *for all* $f\in\Pi_{2n-m-1}$ *and for all* $f\in\{(x-\alpha_k)^{-s}:k=1,\dots,M;s=1,\dots,s_k\}$.

Depending on the application, several special choices of $\{\alpha_k\}$ are proposed: they may contain real numbers and/or complex conjugate pairs, and they may be of order 1 or 2. Independently, Van Assche and Vanherwegen [44] discuss two special cases of Theorem 8.1: the α_k are real and either all $s_k=1$ and $m=2n$ (a polynomial of degree -1 is understood as identically zero), or all but one $s_k=2$ and $m=2n-1$. The first case is called "Gaussian quadrature", the second "orthogonal quadrature".

The main observation to be made with these quadrature formulas is that the nodes and weights are closely related to the zeros and Christoffel numbers for polynomials orthogonal on Δ with respect to a varying measure. This interaction, also observed by López and Illan [34,35], makes it possible to use results from orthogonal polynomials to get useful properties for the nodes and weights for quadrature based on rational interpolation. This is the main contribution of [44] along with the convergence in the class of continuous functions. It should be pointed out that unlike [26,44], in [34,35] non-Newtonian tables of prescribed poles are used, so that when considering convergence results, some additional conditions on the poles are necessary in order to assure the density of the rational functions that are considered in the space $C(\Delta)$ of functions continuous in Δ. For instance, when all the α_k are different

$$\sum_{k=1}^{\infty}(1-|c_k|)=\infty \quad \text{where } c_k=\alpha_k-\sqrt{\alpha_k^2-1} \tag{8.1}$$

(see [1, p. 254]).

We also mention here the work of Gautschi and coauthors [27–29]. In [27] the quadrature method of Theorem 8.1 for the interval $\Delta=[0,\infty]$ is applied with $m=2n$ to Fermi–Dirac integrals, and with $m=2n-1$ to Bose–Einstein integrals, the poles selected being those of the integrand closest to, or on, the real line. The paper [28] describes software implementing Theorem 8.1 and pays special attention to the treatment of poles very close to the support of μ. In [29] results analogous to Theorem 8.1 are developed for other quadrature rules, e.g., Gauss–Kronrod and Gauss–Turán rules, and for other integrals, e.g., Cauchy principal value integrals. In the case of the Lebesgue measure, similar interpolatory formulas are also considered in [45].

Finally, we should mention the works of Min [39,40], where also quadrature formulas based on rational functions are considered when taking $\mathrm{d}\mu(x)=\mathrm{d}x/\sqrt{1-x^2}$, $x\in\Delta$. The author makes use of the properties of the generalized Chebyshev "polynomials" associated with the rational system

$$\left\{1,\frac{1}{x-\alpha_1},\frac{1}{x-\alpha_2},\dots,\frac{1}{x-\alpha_n}\right\},\quad n=1,2,\dots,\quad x\in\Delta. \tag{8.2}$$

This generalized notion is introduced in [2]. The term *polynomial* is misleading because they are in fact *rational* functions in the span of the functions (8.2). The qualification *Chebyshev* is justified by the fact that they have properties similar to classical Chebyshev polynomials. Using these Chebyshev functions and assuming that $\{\alpha_k\}_{k=1}^n\subset\mathbb{R}\backslash\Delta$, $n=1,2,\dots,$ Min constructs the n-point interpolatory quadrature formula

$$Q_n\{f\}=\sum_{k=1}^n A_k f(\xi_k)\approx\int_{-1}^1\frac{f(x)}{\sqrt{1-x^2}}\,\mathrm{d}x,$$

in the zeros of the generalized Chebyshev “polynomials”, and it turns out that this formula is exact for all functions $f\in\operatorname{span}\{(x-\alpha_1)^{-1},(x-\alpha_1)^{-2},\ldots,(x-\alpha_n)^{-1},(x-\alpha_n)^{-2}\}=\mathscr{R}_{2n-1}(\alpha_1,\ldots,\alpha_n)$.

Theorem 8.2 (Min [39]). *Let $\{\alpha_k\}$ and $Q_n\{f\}$ be defined as above. Let $\{\xi_k\}_{k=1}^n$ be the zeros of $T_n(x)$, the generalized Chebyshev “polynomial” associated with* (8.2). *Then* (*a*) *$Q_n\{f\}$ is a positive quadrature formula* (*i.e. $A_k>0$ for $k=1,\ldots,n$*) *and* (*b*) *$\int_{-1}^1 f(x)/\sqrt{1-x^2}\,\mathrm{d}x=Q_n\{f\}$ for any $f\in\mathscr{R}_{2n-1}(\alpha_1,\ldots,\alpha_n)$.*

Since the converse of Theorem 8.1 is also true, it follows that the zeros of the nth Chebyshev polynomial for (8.2) coincide with the zeros of the orthonormal polynomial of degree n with respect to the varying measure

$$\mathrm{d}\mu(x)=\frac{1}{\sqrt{1-x^2}(x-\alpha_1)^2\cdots(x-\alpha_n)^2},\quad \{\alpha_j\}_{j=1}^n\subset\mathbb{R}\backslash\Delta.$$

On the other hand, let U_n be the Chebyshev polynomial of the second kind associated with (8.2). It is known [3, Theorem 1.2] that (a) $T_n^2(x)+(\sqrt{1-x^2}U_n(x))^2=1$, (b) there are $n+1$ points $\{\tilde{\xi}_k\}$ with $-1=\tilde{\xi}_n<\tilde{\xi}_{n-1}<\cdots<\tilde{\xi}_1<\tilde{\xi}_0=1$ such that $T(\tilde{\xi}_k)=(-1)^k$, $k=0,\ldots,n$. Since $||T_n||_{[-1,1]}=1$, $\{\tilde{\xi}_k\}_0^n$ are the extreme points of T_n and also $\{\tilde{\xi}_k\}_1^{n-1}$ are the zeros of U_n.

Theorem 8.3 (Min [39]). *Let the elements $\{\alpha_k\}_1^n\subset\mathbb{C}\backslash\mathbb{R}$ be paired by complex conjugation and let $\{\tilde{\xi}_k\}_1^{n-1}$ be the zeros of $U_n(x)$ as defined above. Then there exist positive parameters $\tilde{A}_0,\ldots,\tilde{A}_n$ such that for all $f\in\mathscr{R}_{2n-1}(\alpha_1,\ldots,\alpha_n)$*

$$\tilde{Q}_n\{f\}=\tilde{A}_0f(1)+\sum_{k=1}^{n-1}\tilde{A}_kf(\tilde{\xi}_k)+\tilde{A}_nf(-1)=\int_{-1}^1\frac{f(x)}{\sqrt{1-x^2}}\,\mathrm{d}x.$$

This is a Lobatto-type quadrature formula.

Let us next assume that μ is a complex measure in $\Delta=[-1,1]$. Theorem 8.1 is still valid, however some difficulties have to be addressed. We need to guarantee the existence of n-point Gaussian quadrature formulas for a measure of the type $\mathrm{d}\mu(x)/\omega_m(x)$ as defined in Theorem 8.1. This requires orthogonal polynomials with respect to a complex measure, and these need not be of degree n, and if they are, their zeros can be anywhere in the complex plane. In [31] the authors could rely on known results about the asymptotic behavior of polynomials orthogonal with respect to fixed complex measures and their zeros to overcome these difficulties. For a general rational setting, a treatment similar to the one in Section 6 is given in [30,22,21]. The idea is as follows. Assume $\mathrm{d}\mu(x)=\rho(x)\,\mathrm{d}x$ with $\rho(x)\in L^1(\Delta)$, possibly complex. Let $\mathbb{A}_n=\{\alpha_{jn}\colon j=1,\ldots,n\}$ and $\mathbb{A}=\bigcup_{n\in\mathbb{N}}\mathbb{A}_n$ with $\mathbb{A}\subset\hat{\mathbb{C}}\backslash\Delta$ be given, and set $\omega_n(x)=(x-\alpha_{1n})\cdots(x-\alpha_{nn})$. For each n, define the space $\mathscr{R}_n:=\{P(x)/\omega_n(x)\colon P\in\Pi_{n-1}\}$. Given n distinct points $\{\xi_{1n},\ldots,\xi_{nn}\}\subset\Delta$, there exist parameters $A_{1n},\ldots,A_{nn}$ such that

$$I_\rho\{f\}:=\int_{-1}^1 f(x)\rho(x)\,\mathrm{d}x=I_{\rho_n}\{f\}:=\sum_{j=1}^n A_{jn}f(\xi_{jn}),\quad\forall f\in\mathscr{R}_n.$$

We call $I_{\rho_n}\{f\}$ an n-interpolatory quadrature formula for $\mathscr{R}_n$. By introducing an auxiliary positive weight function $\beta(x)$ on Δ and taking $\{\xi_{jn}\}_{j=1}^n$ as the zeros of the nth orthogonal polynomial with respect to $\beta(x)/|\omega_n(x)|^2$, several results on the convergence for these quadrature formulas have been proved. González-Vera et al. prove in [30] the convergence of this type of quadrature formulas in the class of continuous functions satisfying a certain Lipschitz condition. Cala–Rodriguez and López–Lagomasino in [22] derive exact rates of convergence when approximating Markov-type analytic functions. In both of these papers, the intimate connection between multipoint Padé-type approximants and interpolatory quadrature formulas is explicitly exploited. The same kind of problem is considered in [21] from a purely "numerical integration" point of view. The most relevant result is:

Theorem 8.4 (Cala-Rodriguez et al. [21]). *Set* $\Delta=[-1,1]$, $\mathbb{A}=\bigcup_{n\in\mathbb{N}}\mathbb{A}_n\subset\hat{\mathbb{C}}\backslash\Delta$ *with* $\mathbb{A}_n=\{\alpha_{jn}\colon j=1,\ldots,n\}$. *Assume that* $\operatorname{dist}(\mathbb{A},\Delta)=\delta>0$ *and that for each* $n\in\mathbb{N}$ *there exists an integer* $m=m(n)$, $1\leqslant m\leqslant n$, *such that* $\alpha_m\in\mathbb{A}_n$ *satisfies* $|\operatorname{Re}(\alpha_m)|>1$. *Let* $\rho(x)\in L^1(\Delta)$ *and* $\beta(x)\geqslant 0$ *on* Δ *such that* $\int_\Delta|\rho(x)|^2/\beta(x)\,\mathrm{d}x=K_1^2<\infty$. *Let* $I_{\omega_n}\{f\}=\sum_{j=1}^n A_{jn}f(\xi_{jn})$ *be the n-point interpolatory quadrature formula in* $\mathscr{R}_n$ *for the nodes* $\{\xi_{jn}\}$ *that are zeros of* $Q_n(x)$, *the* nth *orthogonal polynomial with respect to* $\beta(x)/|\omega_n(x)|^2$, $x\in\Delta$. *Then*, $\lim_{n\to\infty}I_{\rho_n}\{f\}=I_\rho\{f\}$ *for all bounded complex valued functions on* Δ *such that the integral* $I_\rho\{f\}$ *exists.*

As we have seen, when dealing with convergence of sequences of quadrature formulas based on rational functions with prescribed poles, one makes use of some kind of condition about the separation of the poles and the support of the measure (positive or complex) as for example in Theorem 8.4 or some other condition like (8.1). Thus, it is a natural question to ask what happens when sequences of points in the Table $\mathbb{A}$ tend to $\operatorname{supp}(\mu)\subset\Delta$ or when some points are just chosen in Δ. Consider for example the situation where the points in $\mathbb{A}$ are just a repetition of the boundary points -1 or $+1$ of the interval $\Delta=[-1,1]$. According to the approach given in [38], let us consider the transformation $\varphi:[-1,1]\to[0,\infty]$ given by $t=\varphi(x)=(1+x)/(1-x)$. Thus, after this change of variables, we can pass from an integral $\int_{-1}^1 f(x)\,\mathrm{d}\mu(x)$ to an integral $\int_0^\infty g(t)\,\mathrm{d}\lambda(t)$. The poles at $x=1$ are moved to poles at $t=\infty$ and the poles at $x=-1$ are moved to poles at $t=0$. This means that our rational functions are reduced to Laurent polynomials. This special situation is closely related to the so-called strong Stieltjes moment problem (see Section 9.3). The L-polynomials appear in two-point Padé approximants in a situation similar to what was discussed in Section 2. However, the difference is that now 0 and ∞ are points in the support of the measure. The generalization is that we consider a sequence of poles α_k that are in the support of the measure. Then we are back in the situation similar to the one discussed in Section 7.

9. Open problems

Several open problems have been explicitly mentioned or at least been hinted at in the text, and others may have jumped naturally to the mind of attentive readers. We add a few more in this section.

9.1. Error bounds

In this paper we have considered the convergence of modified approximants (MA) or multipoint rational approximants (MRA) and the corresponding convergence of R-Szegő quadrature formulas or interpolatory quadrature formulas. We have used some error bounds both for the approximants and the quadrature formulas (and these are closely related by Theorem 3.7). However, we did not give sharp bounds, and this is of course essential for computing exact rates of convergence.

Using continued fraction theory, Jones and Waadeland [37] recently gave computable sharp error bounds for MAs, in the polynomial case, i.e., $\alpha_k = 0$ for all k. A similar treatment could be done in the rational case. As an illustration, we consider the case of the normalized Lebesgue measure $\mathrm{d}\mu(\theta) = \mathrm{d}\theta/(2\pi)$. Let $R_{\mu_n}(z;\tau_n) = F_\mu(z) - F_{\mu_n}(z)$ be the error for the MA. Recall (3.3), where $p+q = n-1$, $X_n(z) = \prod_{k=1}^{n}(z-\xi_{nk})$, with ξ_{nk} the zeros of the para-orthogonal function $Q_n(z;\tau_n)$. Let us take $p = 0$ and $q = n-1$; then for the normalized Lebesgue measure we have

$$R_{\mu_n}(z) = \frac{2z\pi_{n-1}(z)}{X_n(z)}\,\frac{1}{2\pi i}\int_{-\pi}^{\pi}\frac{X_n(t)\,\mathrm{d}\theta}{\pi_{n-1}(t)(t-z)}, \quad t = \mathrm{e}^{\mathrm{i}\theta}.$$

If $z \in \mathbb{D}$, then by the residue theorem, one gets

$$R_{\mu_n}(z) = 2\left[1 - \frac{X_n(0)\pi_{n-1}(z)}{X_n(z)}\right].$$

Now using the choice $\tau_n = \eta_n$ as in Example 4.1, we have

$$Q_n(z;\tau_n) = \frac{X_n(z)}{\pi_n(z)}, \quad X_n(z) = c\prod_{k=1}^{n}(z-\xi_{nk}) = \kappa_n\tau_n[z\omega_{n-1}(z) + \pi_{n-1}(z)].$$

Therefore,

$$R_{\mu_n}(z) = 2\left[1 - \frac{\pi_{n-1}(z)}{z\omega_{n-1}(z)+\pi_{n-1}(z)}\right] = \frac{2z\omega_{n-1}(z)}{z\omega_{n-1}(z)+\pi_{n-1}(z)}, \quad z \in \mathbb{D}.$$

This is an example of an explicit expression for the error of approximation. It is an open problem to extend this to the general case.

9.2. *Exact rates of convergence*

In Theorem 5.6 we obtained estimates for the rate of convergence. It is however not clear under what conditions on $\mathbb{A}$ and μ we obtain equality, i.e., when a formula of the form $\limsup_{n\to\infty}|R_{\mu_n}(z)|^{1/n} = \exp\{\lambda(z)\}$ holds.

If μ is the normalized Lebesgue measure and all $\alpha_k = 0$, then $R_{\mu_n}(z) = 2z^n/(z^n+1)$, and therefore one gets

$$\lim_{n\to\infty}|R_{\mu_n}(z)|^{1/n} = |z| = \exp\{\lambda(z)\}, \quad z \in \mathbb{D}.$$

Is it true for the normalized Lebesgue measure and $\mathbb{A}$ included in a compact subset of $\mathbb{D}$ that

$$\lim_{n\to\infty}\left|\frac{2z\omega_{n-1}(z)}{z\omega_{n-1}(z)+\pi_{n-1}(z)}\right|^{1/n} = \exp\{\lambda(z)\},$$

where $\lambda(z)$ is as in (5.1)?

9.3. *Stieltjes problems*

Concerning Stieltjes and strong Stieltjes moment problems, several situations are considered that correspond to special choices of the poles like a finite number of poles that are cyclically repeated.

Several results were obtained concerning the moment problem and the multipoint Padé approximants. See for example [16] for the rational moment problem where poles are allowed on the unit circle. The existence proof given there is closely related to the convergence of the quadrature formulas. Several other papers exist about convergence of multipoint Padé approximants with or without a cyclic repetition of the poles. The quadrature part is still largely unexplored. In [38], this situation is briefly mentioned.

For the case of L-polynomials where one considers only the poles 0 and ∞, the convergence of two-point Padé approximants to Stieltjes transforms was studied in [32,5,6]. The latter papers also give error estimates and consider the corresponding convergence and rates of convergence of the quadrature formulas. To illustrate this, we formulate some of the theorems. Note that in this section we are dealing with intervals Δ that may be finite or infinite, so it is not always possible to renormalize it to the standard interval $[-1,1]$ as in Section 8. Thus Δ has another meaning here.

Theorem 9.1 (Bultheel et al. [6]). *Let μ be a positive measure on $\Delta=[a,b]$ with $0\leqslant a<b\leqslant\infty$. Let $Q_n^\mu(x)=\kappa_n\prod_{k=1}^n(x-\xi_{jn})$, $\kappa_n>0$, be the nth orthonormal polynomial with respect to $x^{-p}\,\mathrm{d}\mu(x)$, $p\geqslant 0$. Let $F_\mu(z)=I_\mu\{(x-z)^{-1}\}=\int_\Delta 1/(x-z)\,\mathrm{d}\mu(x)$ be the Cauchy transform of μ. Let $I_{\mu_n}\{f\}=\sum_{j=1}^n A_{jn}f(\xi_{jn})$ be the n-point Gaussian formula in $\Lambda_{-p,q}$ where $0\leqslant p\leqslant 2n$, $q\geqslant -1$ and $p+q=2n-1$, and set $F_{\mu_n}(z)=I_{\mu_n}\{(x-z)^{-1}\}$. Then F_{μ_n} is a rational function of type $(n-1,n)$ that is a two-point Padé approximant (2PA) for F_μ (order p at the origin and order $q+2$ at infinity).*

López-Lagomasino et al. prove in several papers (see for example [38]) the uniform convergence of the 2PA in compact subsets of $\mathbb{C}\backslash\Delta$ and give estimates for the rate of convergence. They assume some Carleman-type conditions, namely either $\lim_{n\to\infty}p(n)=\infty$ and $\sum_{n=1}^\infty c_{-n}^{-1/2n}=\infty$ or $\lim_{n\to\infty}[2n-1-p(n)]=\infty$ and $\sum_{n=1}^\infty c_n^{-1/2n}=\infty$, where the moments are defined as $c_n:=\int x^n\,\mathrm{d}\mu(x)$, $n\in\mathbb{Z}$.

When the measure $\mathrm{d}\mu(x)=\rho(x)\,\mathrm{d}x$ is complex, with $\int_\Delta|\rho(x)|\,\mathrm{d}x<\infty$, then an auxiliary positive measure $\omega(x)\,\mathrm{d}x$ with $\omega(x)>0$, $x\in\Delta$, is introduced such that $\int_\Delta|\rho(x)|^2/\omega(x)\,\mathrm{d}x=K^2<\infty$.

Theorem 9.2 (Bultheel et al. [6]). *Let Q_n^ω be the nth orthogonal polynomial with respect to $x^{-2p}\omega(x)$, whose zeros are $\xi_{jn}\in\Delta$, and let $I_{\mu_n}\{f\}=\sum_{j=1}^n A_{jn}f(\xi_{jn})$ be the interpolatory quadrature formula exact in $\Lambda_{-p,q}$, $p+q=n-1$. Then $F_{\mu_n}(z)=I_{\mu_n}\{(x-z)^{-1}\}$ is a rational function of type $(n-1,n)$ and it is a two-point Padé-type (2PTA) approximant for F_μ.*

Let $d_k=\int_\Delta x^k\omega(x)\,\mathrm{d}x$, $k\in\mathbb{N}$, be the moments of ω and assume that $p=p(n)$ and $q=q(n)=n-1-p(n)$, such that either $\lim_{n\to\infty}p(n)=\infty$ and $\sum_{n=1}^\infty d_{-n}^{-1/2n}=\infty$ or $\lim_{n\to\infty}q(n)=\infty$ and $\sum_{n=1}^\infty d_n^{-1/2n}=\infty$. Then the 2PTA $F_{\mu_n}(z)$ converge to $F_\mu(z)$ uniformly in compact subsets of $\mathbb{C}\backslash\Delta$. The quadrature formula converges to the integral for all $f\in C^{\mathrm{B}}[0,\infty)=\{f\in C[0,\infty)\colon\ \lim_{x\to\infty}f(x)=L\in\mathbb{C}\}$ if and only if $\sum_{k=1}^n|A_{kn}|\leqslant M$ for $n\in\mathbb{N}$.

Note that if in this theorem $\mathrm{d}\mu$ is a positive Borel measure, we can set $\omega=\rho$, so that the quadrature formula becomes the n-point Gaussian formula, and then the Carleman conditions on its moments imply the convergence of the quadrature formulas in the class $C^{\mathrm{B}}[0,\infty)$.

As for the rate of convergence, we assume that $\lim_{n\to\infty}p(n)/n=r\in[0,1]$, and we assume that μ is of the form $\mathrm{d}\mu(x)=x^v\exp(-\tau(x))\,\mathrm{d}x$, $v\in\mathbb{R}$, $\tau(x)$ continuous on $(0,\infty)$ and for $\gamma>\frac{1}{2}$ and

$s > 0$: $\lim_{x\to 0^+}(sx)^\gamma \tau(x) = \lim_{x\to\infty}(sx)^{-\gamma}\tau(x) = 1$. Set

$$D(\gamma) = \frac{2\gamma}{2\gamma - 1}\left[\frac{\Gamma(\gamma + 1/2)}{\sqrt{\pi}\Gamma(\gamma)}\right]^{1/2\gamma},$$

where Γ is the Euler Gamma function, and with $\theta = 1 - 1/(2\gamma) < 1$, $\delta(z) = (1-r)^\theta \operatorname{Im}((sz)^{1/2}) + r^\theta \operatorname{Im}((sz)^{-1/2})$, where the branch is taken such that $(-1)^{1/2} = \mathrm{i}$. Furthermore, for f analytic in the domain $\mathbb{G}$ such that $[0,\infty) \subset \mathbb{G} \subset \hat{\mathbb{C}}$ and some compact K, define $\lambda(K) := \exp(-R)$ with $R = 2D(\gamma)\inf_{z\in K}\{\delta(z)\} > 0$ for some compact K. It is then proved, using results from [38], that

Theorem 9.3 (Bultheel et al. [6]). *With the notation just introduced, let $I_{\mu_n}\{f\}$ be the n-point Gaussian quadrature formula exact in $\Lambda_{-p(n),2n-1-p(n)}$ with error $E_{\mu_n} = I_\mu - I_{\mu_n}$. Let F_{μ_n} be the corresponding 2PA for the Cauchy transform F_μ and $R_{\mu_n} = F_\mu - F_{\mu_n}$ the associated error. Then we have that* $\lim_{n\to\infty}||R_{\mu_n}||_K^{1/(2n)^\theta} = \lambda(K)$, *where K is a compact subset of* $\mathbb{C}\backslash[0,\infty)$ *and* $||\cdot||_K$ *is the supremum norm in K. Also* $\lim_{n\to\infty}E_{\mu_n}\{f\} = 0$ *for all f analytic in the domain* $\mathbb{G}$, *and* $\limsup_{n\to\infty}|E_{\mu_n}\{f\}|^{1/(2n)^\theta} \leqslant \lambda(\mathbb{J}) < 1$, *where* $\mathbb{J} \subset \mathbb{G}$ *is a Jordan curve.*

If I_{μ_n} is the interpolatory quadrature formula of Theorem 9.2 *and F_{μ_n} the corresponding 2PTA, then $R_{\mu_n} \to 0$ uniformly in compact subsets K of* $\mathbb{C}\backslash[0,\infty)$ *and* $\lim_{n\to\infty}||R_{\mu_n}||_K^{1/(2n)^\theta} = \sqrt{\lambda(K)} < 1$. *For the quadrature formula it holds that* $\limsup_{n\to\infty}|E_{\mu_n}\{f\}|^{1/n^\theta} \leqslant \sqrt{\lambda(\mathbb{J})} < 1$ *with* $\mathbb{J} \subset \mathbb{G}$ *a Jordan curve and f analytic in* $\mathbb{G}$.

It is still an open problem to generalize this kind of results to the multipoint case where we select a number of poles $\alpha_k \in [0,\infty]$. Also the multipoint problem corresponding to the Hamburger moment problem (the measure is supported on the whole of $\mathbb{R}$ as explained in Section 7) needs generalization. There is almost nothing published about error estimates, convergence or rate of convergence for the rational approximants or for the quadrature formulas.

9.4. Miscellaneous problems

(1) In the convergence results where the poles of the rational functions are outside the support of the measure, it was assumed that they stayed away (they were in a compact subset of $\mathbb{D}$). What if the latter is not true?

(2) In [42], Santos-León considers integrals of the form $\int_{-\pi}^{\pi} f(\mathrm{e}^{\mathrm{i}\theta})K(\theta)\,\mathrm{d}\theta$ with K such that $\int_{-\pi}^{\pi}|K(\theta)|\,\mathrm{d}\theta < \infty$. He proposes quadrature formulas of interpolatory type with nodes uniformly distributed on $\mathbb{T}$. Properties for the weights and estimates for the error of the quadrature formulas are given. A similar treatment can be given when the nodes are the zeros of the para-orthogonal rational functions with respect to the Lebesgue measure, which means the zeros of $z\omega_{n-1}(z) + \pi_{n-1}(z)$.

References

[1] N.I. Achieser, Theory of Approximation, Frederick Ungar, New York, 1956.

[2] P. Borwein, T. Erdélyi, Polynomials and Polynomial Inequalities, Springer, New York, 1995.

[3] P. Borwein, T. Erdélyi, J. Zhang, Chebyshev polynomials and Markov-Bernstein type inequalities for the rational spaces, J. London Math. Soc. (2) 50 (1994) 501–519.

[4] A. Bultheel, P. Dewilde, Orthogonal functions related to the Nevanlinna-Pick problem, in: P. Dewilde (Ed.), Proceedings of the Fourth International Conference on Mathematical Theory of Networks and Systems, Delft, Western Periodicals, North-Hollywood, 1979, pp. 207–212.

[5] A. Bultheel, C. Díaz-Mendoza, P. González-Vera, R. Orive, Quadrature on the half line and two-point Padé approximants to Stieltjes functions, Part II: convergence, J. Comput. Appl. Math. 77 (1997) 53–76.

[6] A. Bultheel, C. Díaz-Mendoza, P. González-Vera, R. Orive, Quadrature on the half line and two-point Padé approximants to Stieltjes functions, Part III: the unbounded case, J. Comput. Appl. Math. 87 (1997) 95–117.

[7] A. Bultheel, P. González-Vera, Wavelets by orthogonal rational kernels, in: B. Berndt, F. Gesztesy (Eds.), Continued Fractions: From Number Theory to Constructive Approximation, Contemporary Mathematics, Vol. 236, American Mathematical Society, New York, 1999, pp. 101–126.

[8] A. Bultheel, P. González-Vera, E. Hendriksen, O. Njåstad, The computation of orthogonal rational functions and their interpolating properties, Numer. Algorithms 2 (1) (1992) 85–114.

[9] A. Bultheel, P. González-Vera, E. Hendriksen, O. Njåstad, Orthogonal rational functions and quadrature on the unit circle, Numer. Algorithms 3 (1992) 105–116.

[10] A. Bultheel, P. González-Vera, E. Hendriksen, O. Njåstad, Orthogonal rational functions with poles on the unit circle, J. Math. Anal. Appl. 182 (1994) 221–243.

[11] A. Bultheel, P. González-Vera, E. Hendriksen, O. Njåstad, Quadrature formulas on the unit circle and two-point Padé approximation, in: A.M. Cuyt (Ed.), Nonlinear Numerical Methods and Rational Approximation II, Kluwer, Dordrecht, 1994, pp. 303–318.

[12] A. Bultheel, P. González-Vera, E. Hendriksen, O. Njåstad, On the convergence of multipoint Padé-type approximants and quadrature formulas associated with the unit circle, Numer. Algorithms 13 (1996) 321–344.

[13] A. Bultheel, P. González-Vera, E. Hendriksen, O. Njåstad, Orthogonal rational functions and modified approximants, Numer. Algorithms 11 (1996) 57–69.

[14] A. Bultheel, P. González-Vera, E. Hendriksen, O. Njåstad, Orthogonal rational functions and nested disks, J. Approx. Theory 89 (1997) 344–371.

[15] A. Bultheel, P. González-Vera, E. Hendriksen, O. Njåstad, Rates of convergence of multipoint rational approximants and quadrature formulas on the unit circle, J. Comput. Appl. Math. 77 (1997) 77–102.

[16] A. Bultheel, P. González-Vera, E. Hendriksen, O. Njåstad, A rational moment problem on the unit circle, Methods Appl. Anal. 4 (3) (1997) 283–310.

[17] A. Bultheel, P. González-Vera, E. Hendriksen, O. Njåstad, Orthogonal rational functions and interpolatory product rules on the unit circle, II, Quadrature and convergence, Analysis 18 (1998) 185–200.

[18] A. Bultheel, P. González-Vera, E. Hendriksen, O. Njåstad, A density problem for orthogonal rational functions, J. Comput. Appl. Math. 105 (1999) 199–212.

[19] A. Bultheel, P. González-Vera, E. Hendriksen, O. Njåstad, Orthogonal Rational Functions, Cambridge Monographs on Applied and Computational Mathematics, Vol. 5, Cambridge University Press, Cambridge, 1999.

[20] A. Bultheel, P. González-Vera, E. Hendriksen, O. Njåstad, Orthogonal rational functions and interpolatory product rules on the unit circle, III, Convergence of general sequences, Analysis 20 (2000) 99–120.

[21] F. Cala-Rodriguez, P. González-Vera, M. Jiménez-Páiz, Quadrature formulas for rational functions, Electron. Trans. Numer. Anal. 9 (1999) 39–52.

[22] F. Cala-Rodriguez, G. López-Lagomasino, Multipoint Padé-type approximants, Exact rate of convergence, Constr. Approx. 14 (1998) 259–272.

[23] P.J. Davis, P. Rabinowitz, Methods of Numerical Integration, 2nd Edition, Academic Press, New York, 1984.

[24] M.M. Djrbashian, A survey on the theory of orthogonal systems and some open problems, in: P. Nevai (Ed.), Orthogonal Polynomials: Theory and Practice, NATO-ASI, Series C: Mathematical and Physical Sciences, Vol. 294, Kluwer Academic Publishers, Boston, 1990, pp. 135–146.

[25] W. Gautschi, A survey of Gauss-Christoffel quadrature formulae, in: P.L. Butzer, F. Fehér (Eds.), E.B. Christoffel, The Influence of his Work on Mathematical and Physical Sciences, Birkhäuser, Basel, 1981, pp. 72–147.

[26] 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, 1993, pp. 111–130.

[27] W. Gautschi, On the computation of generalized Fermi-Dirac and Bose-Einstein integrals, Comput. Phys. Comm. 74 (1993) 233–238.

[28] W. Gautschi, Algorithm 793: GQRAT — Gauss quadrature for rational functions, ACM Trans. Math. Software 25 (1999) 213–239.

[29] W. Gautschi, L. Gori, M.L. Lo Cascio, Quadrature rules for rational functions, Numer. Math. (2000), to appear.

[30] P. González-Vera, M. Jiménez-Páiz, G. López-Lagomasino, R. Orive, On the convergence of quadrature formulas connected with multipoint Padé-type approximants, J. Math. Anal. Appl. 202 (1996) 747–775.

[31] P. González-Vera, G. López-Lagomasino, R. Orive, J.C. Santos-León, On the convergence of quadrature formulas for complex weight functions, J. Math. Anal. Appl. 189 (1995) 514–532.

[32] P. González-Vera, O. Njåstad, Convergence of two-point Padé approximants to series of Stieltjes, J. Comput. Appl. Math. 32 (1990) 97–105.

[33] P. González-Vera, J.C. Santos-León, O. Njåstad, Some results about numerical quadrature on the unit circle, Adv. Comput. Math. 5 (1996) 297–328.

[34] J. Illan, G. López-Lagomasino, A note on generalized quadrature formulas of Gauss-Jacobi type, Constructive Theory of Functions '84, Sofia, 1994, pp. 513–518.

[35] J. Illan, G. López-Lagomasino, Sobre los metodos interpolatorios de integración numérica y su conexion con la approximación racional, Rev. Cienc. Mat. 8 (1987) 31–44.

[36] W.B. Jones, O. Njåstad, W.J. Thron, Moment theory, orthogonal polynomials, quadrature and continued fractions associated with the unit circle, Bull. London Math. Soc. 21 (1989) 113–152.

[37] W.B. Jones, H. Waadeland, Bounds for remainder terms in Szegő quadrature on the unit circle, in: Approximation and Computation, International Series of Numerical Mathematics, Vol. 119, Birkhäuser, Basel, 1995, pp. 325–342. A Festschrift in honor of Walter Gautschi.

[38] G. López-Lagomasino, A. Martínez-Finkelshtein, Rate of convergence of two-point Padé approximants and logarithmic asymptotics of Laurent-type orthogonal polynomials, Constr. Approx. 11 (1995) 255–286.

[39] G. Min, Lagrange interpolation and quadrature formula in rational systems, J. Approx. Theory 95 (1998) 123–145.

[40] G. Min, Lobatto type quadrature formulas in rational spaces, J. Comput. Appl. Math. 94 (1998) 1–12.

[41] J. Nuttal, C.J. Wherry, Gauss integration for complex weight functions, J. Inst. Math. Appl. 21 (1987) 165–170.

[42] J.C. Santos-León, Product rules on the unit circle with uniformly distributed nodes, Error bounds for analytic functions, J. Comput. Appl. Math. 108 (1999) 195–208.

[43] G. Szegő, Orthogonal Polynomials, 3rd Edition, American Mathematical Society Colloquial Publication, Vol. 33, American Mathematical Society Providence, RI, 1967, 1st Edition, 1939.

[44] W. Van Assche, I. Vanherwegen, Quadrature formulas based on rational interpolation, Math. Comp. 16 (1993) 765–783.

[45] J.A.C. Weideman, D.P. Laurie, Quadrature rules based on partial fraction expansions, Numer. Algorithms (1999) submitted for publication.

ELSEVIER

Journal of Computational and Applied Mathematics 127 (2001) 93–119

JOURNAL OF
COMPUTATIONAL AND
APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

An iterative method with error estimators

D. Calvetti[a,*,1], S. Morigi[b], L. Reichel[c,2], F. Sgallari[b]

[a] *Department of Mathematics, Case Western Reserve University, Cleveland, OH 44106, USA*
[b] *Dipartimento di Matematica, Università di Bologna, Bologna, Italy*
[c] *Department of Mathematics and Computer Science, Kent State University, Kent, OH 44242, USA*

Received 23 November 1999; received in revised form 4 April 2000

Abstract

Iterative methods for the solution of linear systems of equations produce a sequence of approximate solutions. In many applications it is desirable to be able to compute estimates of the norm of the error in the approximate solutions generated and terminate the iterations when the estimates are sufficiently small. This paper presents a new iterative method based on the Lanczos process for the solution of linear systems of equations with a symmetric matrix. The method is designed to allow the computation of estimates of the Euclidean norm of the error in the computed approximate solutions. These estimates are determined by evaluating certain Gauss, anti-Gauss, or Gauss–Radau quadrature rules. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Lanczos process; Conjugate gradient method; Symmetric linear system; Gauss quadrature

1. Introduction

Large linear systems of equations

$$A\boldsymbol{x}=\boldsymbol{b}, \quad A\in\mathbb{R}^{n\times n}, \quad \boldsymbol{x}\in\mathbb{R}^n, \quad \boldsymbol{b}\in\mathbb{R}^n \tag{1}$$

with a nonsingular symmetric matrix are frequently solved by iterative methods, such as the conjugate gradient method and variations thereof; see, e.g., [12, Chapter 10] or [17, Chapter 6]. It is the purpose of the present paper to describe a modification of the conjugate gradient method that allows the computation of bounds or estimates of the norm of the error in the computed approximate solutions.

[*] Corresponding author.

E-mail addresses: dxc57@po.cwru.edu (D. Calvetti), morigi@dm.unibo.it (S. Morigi), reichel@mcs.kent.edu (L. Reichel), sgallari@dm.unibo.it (F. Sgallari).

[1] Research supported in part by NSF grant DMS-9806702.

[2] Research supported in part by NSF grant DMS-9806413.

PII: S0377-0427(00)00494-5

Assume for notational simplicity that the initial approximate solution of (1) is given by $\boldsymbol{x}_0 = \boldsymbol{0}$, and let Π_{k-1} denote the set of polynomials of degree at most $k-1$. The iterative method of this paper yields approximate solutions of (1) of the form

$$\boldsymbol{x}_k = q_{k-1}(A)\boldsymbol{b}, \quad k = 1, 2, \ldots, \tag{2}$$

where the iteration polynomials $q_{k-1} \in \Pi_{k-1}$ are determined by the method.

The residual error associated with $\boldsymbol{x}_k$ is defined by

$$\boldsymbol{r}_k := \boldsymbol{b} - A\boldsymbol{x}_k \tag{3}$$

and the error in $\boldsymbol{x}_k$ is given by

$$\boldsymbol{e}_k := A^{-1}\boldsymbol{r}_k. \tag{4}$$

Using (3) and (4), we obtain

$$\boldsymbol{e}_k^{\mathrm{T}}\boldsymbol{e}_k = \boldsymbol{r}_k^{\mathrm{T}}A^{-2}\boldsymbol{r}_k = \boldsymbol{b}^{\mathrm{T}}A^{-2}\boldsymbol{b} - 2\boldsymbol{b}^{\mathrm{T}}A^{-1}\boldsymbol{x}_k + \boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{x}_k. \tag{5}$$

Thus, the Euclidean norm of $\boldsymbol{e}_k$ can be evaluated by computing the terms on the right-hand side of (5). The evaluation of the term $\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{x}_k$ is straightforward. This paper discusses how to evaluate bounds or estimates of the other terms on the right-hand side of (5). The evaluation is made possible by requiring that the iteration polynomials satisfy

$$q_{k-1}(0) = 0, \quad k = 1, 2, \ldots. \tag{6}$$

Then $\boldsymbol{b}^{\mathrm{T}}A^{-1}\boldsymbol{x}_k = \boldsymbol{b}^{\mathrm{T}}A^{-1}q_{k-1}(A)\boldsymbol{b}$ can be computed for every k without using A^{-1}, and this makes easy evaluation of the middle term on the right-hand side of (5) possible. The iterative method obtained is closely related to the SYMMLQ method, see, e.g., [16] or [8, Section 6.5], and can be applied to solve linear systems of equations (1) with a positive definite or indefinite symmetric matrix. Details of the method are presented in Section 2.

Section 3 discusses how bounds or estimates of the first term on the right-hand side of (5) can be computed by evaluating certain quadrature rules of Gauss-type. Specifically, when the matrix A is positive definite and we have evaluated $\boldsymbol{x}_k$, a lower bound of $\boldsymbol{b}^{\mathrm{T}}A^{-2}\boldsymbol{b}$ can be computed inexpensively by evaluating a k-point Gauss quadrature rule. An estimate of an upper bound is obtained by evaluating an associated k-point anti-Gauss rule. When A is indefinite, an estimate of the Euclidean norm of the error $\boldsymbol{e}_k$ is obtained by evaluating a $(k+1)$-point Gauss–Radau quadrature rule with a fixed node at the origin. We also describe how the quadrature rules can be updated inexpensively when k is increased. Section 4 presents a few computed examples, and Section 5 contains concluding remarks.

The application of quadrature rules of Gauss-type to the computation of error bounds for approximate solutions generated by an iterative method was first described by Dahlquist et al. [6], who discussed the Jacobi iteration method. When the matrix A is symmetric and positive definite, the linear system (1) can conveniently be solved by the conjugate gradient method. Dahlquist et al. [7], and subsequently Golub and Meurant [10,14], describe methods for computing bounds in the A-norm of approximate solutions determined by the conjugate gradient method. A new approach, based on extrapolation, for computing estimates of the norm of the error in approximate solutions determined by iterative methods has recently been proposed in [1].

Assume for the moment that the matrix A in (1) is symmetric and positive definite, and approximate solutions $\boldsymbol{x}_k$ of the linear system (1) are computed by the conjugate gradient method. The method of Golub and Meurant [10] for computing upper bounds for the A-norm of the error in the approximate solutions requires that a lower positive bound for the smallest eigenvalue of the matrix A is available, and so does the scheme in [14], based on two-point Gauss quadrature rules, for computing upper bounds of the Euclidean norm of the error in the iterates. Estimates of the smallest eigenvalue can be computed by using the connection between the conjugate gradient method and the Lanczos method, see, e.g., [12, Chapter 10]; however, it is generally difficult to determine positive lower bounds. The methods of the present paper for computing error estimates do not require knowledge of any of the eigenvalues of the matrix A.

The performance of iterative methods is often enhanced by the use of preconditioners; see, e.g., [12, Chapter 10, 17, Chapters 9–10]. In the present paper, we assume that the linear system of equations (1) represents the preconditioned system. Alternatively, one can let (1) represent the unpreconditioned linear system and modify the iterative method to incorporate the preconditioner. Meurant [15] shows how the computation of upper and lower bounds of the A-norm of the error in approximate solutions determined by the conjugate gradient method can be carried out when this approach is used. Analogous formulas can be derived for the iterative method of the present paper.

2. The iterative method

This section presents an iterative method for the solution of linear systems of equations (1) with a nonsingular symmetric matrix A. The description is divided into two subsections, the first of which discusses basic properties of the method. The second subsection derives updating formulas for the approximate solutions $\boldsymbol{x}_k$ computed. The method may be considered a modification of the conjugate gradient method or of the SYMMLQ method, described, e.g., in [8,16].

Our description uses the spectral factorization

$$A = U_n \Lambda_n U_n^{\mathrm{T}}, \quad U_n \in \mathbb{R}^{n\times n}, \ U_n^{\mathrm{T}} U_n = I_n,$$

$$\Lambda_n = \operatorname{diag}[\lambda_1, \lambda_2, \ldots, \lambda_n], \quad \lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n. \tag{7}$$

Here and throughout this paper, I_j denotes the identity matrix of order j. Let $\hat{\boldsymbol{b}}=[\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_n]^{\mathrm{T}} := U_n^{\mathrm{T}} \boldsymbol{b}$ and express the matrix functional

$$F(A) := \boldsymbol{b}^{\mathrm{T}} f(A) \boldsymbol{b}, \quad f(t) := 1/t^2, \tag{8}$$

as a Stieltjes integral

$$F(A) = \hat{\boldsymbol{b}}^{\mathrm{T}} f(\Lambda_n) \hat{\boldsymbol{b}} = \sum_{k=1}^{n} f(\lambda_k) \hat{b}_k^2 = \int_{-\infty}^{\infty} f(t)\, \mathrm{d}\omega(t). \tag{9}$$

The measure ω is a nondecreasing step function with jump discontinuities at the eigenvalues λ_k of A. We will use the notation

$$\mathscr{I}(f) := \int_{-\infty}^{\infty} f(t)\, \mathrm{d}\omega(t). \tag{10}$$

2.1. Basic properties

Our method is based on the Lanczos process. Given the right-hand side vector $\boldsymbol{b}$, k steps of the Lanczos process yield the Lanczos decomposition

$$AV_k = V_kT_k + \boldsymbol{f}_k\tilde{\boldsymbol{e}}_k^{\mathrm{T}}, \tag{11}$$

where $V_k = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k] \in \mathbb{R}^{n\times k}$ and $\boldsymbol{f}_k \in \mathbb{R}^n$ satisfy $V_k^{\mathrm{T}}V_k = I_k$, $V_k^{\mathrm{T}}\boldsymbol{f}_k = \boldsymbol{0}$ and

$$\boldsymbol{v}_1 = \boldsymbol{b}/||\boldsymbol{b}||. \tag{12}$$

Moreover, $T_k \in \mathbb{R}^{k\times k}$ is symmetric and tridiagonal. Throughout this paper $\tilde{\boldsymbol{e}}_j$ denotes the jth axis vector and $||\cdot||$ the Euclidean vector norm. We may assume that T_k has nonvanishing subdiagonal entries; otherwise the Lanczos process breaks down and the solution of (1) can be computed as a linear combination of the columns $\boldsymbol{v}_j$ generated before break down.

Eq. (11) defines a recursion relation for the columns of V_k. This relation, combined with (12), shows that

$$\boldsymbol{v}_j = s_{j-1}(A)\boldsymbol{b}, \quad 1\leqslant j\leqslant k \tag{13}$$

for certain polynomials s_{j-1} of degree $j-1$. These polynomials are orthogonal with respect to the following inner product induced by (10) for functions g and h defined on the spectrum of A,

$$(g,h) := \mathscr{I}(gh). \tag{14}$$

We have

$$\begin{aligned}(s_{j-1}, s_{\ell-1}) &= \int_{-\infty}^{\infty} s_{j-1}(t)s_{\ell-1}(t)\,\mathrm{d}\omega(t) = \boldsymbol{b}^{\mathrm{T}}U_n s_{j-1}(\Lambda_n)s_{\ell-1}(\Lambda_n)U_n^{\mathrm{T}}\boldsymbol{b}\\ &= \boldsymbol{b}^{\mathrm{T}}s_{j-1}(A)s_{\ell-1}(A)\boldsymbol{b}\\ &= \boldsymbol{v}_j^{\mathrm{T}}\boldsymbol{v}_\ell = \begin{cases}0, & j\neq \ell,\\ 1, & j=\ell,\end{cases}\end{aligned} \tag{15}$$

where we have applied manipulations analogous to those used in Eq. (9). The last equality of (15) follows from the orthogonality of the columns $\boldsymbol{v}_j$ of V_k. Since the polynomial s_ℓ is of degree ℓ, the columns of V_k span the Krylov subspace

$$\mathbb{K}_k(A,\boldsymbol{b}) := \mathrm{span}\{\boldsymbol{b}, A\boldsymbol{b}, \ldots, A^{k-1}\boldsymbol{b}\},$$

i.e.,

$$\mathrm{range}\,(V_k) = \mathbb{K}_k(A,\boldsymbol{b}). \tag{16}$$

We also will use the following form of the Lanczos decomposition:

$$AV_{k-1} = V_kT_{k,k-1}, \tag{17}$$

where $T_{k,k-1}$ is the leading principal $k\times(k-1)$ submatrix of T_k.

Introduce the QR-factorization of T_k, i.e., let

$$T_k = Q_kR_k, \quad Q_k, R_k \in \mathbb{R}^{k\times k}, \; Q_k^{\mathrm{T}}Q_k = I_k, \tag{18}$$

where $R_k = [r_{j\ell}^{(k)}]_{j,\ell=1}^{k}$ is upper triangular. Also, define

$$T_{k,k-1} = Q_k \begin{bmatrix} \bar{R}_{k-1} \\ 0 \end{bmatrix} = Q_{k,k-1}\bar{R}_{k-1}, \tag{19}$$

where $\bar{R}_{k-1}$ is the leading principal submatrix of order $k-1$ of R_k, and $Q_{k,k-1} \in \mathbb{R}^{k\times(k-1)}$ consists of the first $k-1$ columns of Q_k. For definiteness, we assume that the diagonal entries in the triangular factors in all QR-factorizations of this paper are nonnegative.

The following manipulations of the Lanczos decomposition (11) give an iterative method, whose associated iteration polynomials satisfy (6). The manipulations are closely related to those required in the derivation of the implicitly restarted Lanczos method; see, e.g., [5]. Substituting the QR-factorization (18) into the Lanczos decomposition (11) yields

$$AV_k = V_kQ_kR_k + \boldsymbol{f}_k\tilde{\boldsymbol{e}}_k^{\mathrm{T}}, \tag{20}$$

which after multiplication by Q_k from the right gives

$$A\tilde{V}_k = \tilde{V}_k\tilde{T}_k + \boldsymbol{f}_k\tilde{\boldsymbol{e}}_k^{\mathrm{T}}Q_k, \quad \tilde{V}_k{:=}V_kQ_k, \quad \tilde{T}_k{:=}R_kQ_k. \tag{21}$$

The matrix $\tilde{V}_k = [\tilde{\boldsymbol{v}}_1^{(k)}, \tilde{\boldsymbol{v}}_2^{(k)}, \ldots, \tilde{\boldsymbol{v}}_k^{(k)}]$ has orthonormal columns and $\tilde{T}_k$ is the symmetric tridiagonal matrix obtained from T_k by applying one step of the QR-algorithm with shift zero.

A relation between the first columns $\boldsymbol{v}_1$ and $\tilde{\boldsymbol{v}}_1^{(k)}$ of V_k and $\tilde{V}_k$, respectively, is easily shown. Assume that $k > 1$ and multiply (20) by $\tilde{\boldsymbol{e}}_1$ from the right. We obtain

$$AV_k\tilde{\boldsymbol{e}}_1 = \tilde{V}_kR_k\tilde{\boldsymbol{e}}_1 + \boldsymbol{f}_k\tilde{\boldsymbol{e}}_k^{\mathrm{T}}\tilde{\boldsymbol{e}}_1,$$

which simplifies to

$$A\boldsymbol{v}_1 = r_{11}^{(k)}\tilde{\boldsymbol{v}}_1^{(k)},$$

where we have used that $R_k\tilde{\boldsymbol{e}}_1 = r_{11}^{(k)}\tilde{\boldsymbol{e}}_1$. Thus,

$$\tilde{\boldsymbol{v}}_1^{(k)} = A\boldsymbol{b}/\|A\boldsymbol{b}\|.$$

Since T_k is tridiagonal, the orthogonal matrix Q_k in the QR-factorization (18) is of upper Hessenberg form. It follows that all but the last two components of the vector $\tilde{\boldsymbol{e}}_k^{\mathrm{T}}Q_k$ are guaranteed to vanish. Therefore, decomposition (21) differs from a Lanczos decomposition in that the last two columns of the matrix $\boldsymbol{f}_k\tilde{\boldsymbol{e}}_k^{\mathrm{T}}Q_k$ may be nonvanishing.

Let $\bar{V}_{k-1}$ be the matrix made up by the first $k-1$ columns of $\tilde{V}_k$. Note that

$$\bar{V}_{k-1} = V_kQ_{k,k-1}, \tag{22}$$

where $Q_{k,k-1}$ is defined by (19). Generally, $\bar{V}_{k-1} \neq \tilde{V}_{k-1}$; see Section 2.2 for details. Removing the last column from each term in Eq. (21) yields the decomposition

$$A\bar{V}_{k-1} = \bar{V}_{k-1}\tilde{T}_{k-1} + \bar{\boldsymbol{f}}_{k-1}\tilde{\boldsymbol{e}}_{k-1}^{\mathrm{T}}, \tag{23}$$

where $\bar{V}_{k-1}^{\mathrm{T}}\bar{\boldsymbol{f}}_{k-1} = \boldsymbol{0}$, $\bar{V}_{k-1}^{\mathrm{T}}\bar{V}_{k-1} = I_{k-1}$ and $\tilde{T}_{k-1}$ is the leading principal submatrix of order $k-1$ of the matrix $\tilde{T}_k$. Thus, decomposition (23) is a Lanczos decomposition with initial vector $\tilde{\boldsymbol{v}}_1^{(k)}$ of $\bar{V}_{k-1}$ proportional to $A\boldsymbol{b}$. Analogously to (16), we have

$$\mathrm{range}(\bar{V}_{k-1}) = \mathbb{K}_{k-1}(A, A\boldsymbol{b}). \tag{24}$$

We determine the iteration polynomials (2), and thereby the approximate solutions $\boldsymbol{x}_k$ of (1), by requiring that

$$\boldsymbol{x}_k = q_{k-1}(A)\boldsymbol{b} = \bar{V}_{k-1}\boldsymbol{z}_{k-1} \tag{25}$$

for some vector $\boldsymbol{z}_{k-1} \in \mathbb{R}^{k-1}$. It follows from (24) that any polynomial q_{k-1} determined by (25) satisfies (6). We choose $\boldsymbol{z}_{k-1}$, and thereby $q_{k-1} \in \Pi_{k-1}$, so that the residual error (3) associated with the approximate solution $\boldsymbol{x}_k$ of (1) satisfies the Petrov–Galerkin equation

$$\boldsymbol{0} = V_{k-1}^{\mathrm{T}}\boldsymbol{r}_k = V_{k-1}^{\mathrm{T}}\boldsymbol{b} - V_{k-1}^{\mathrm{T}}A\bar{V}_{k-1}\boldsymbol{z}_{k-1}, \tag{26}$$

which, by using (12) and factorization (22), simplifies to

$$||\boldsymbol{b}||\tilde{\boldsymbol{e}}_1 = (AV_{k-1})^{\mathrm{T}}V_kQ_{k,k-1}\boldsymbol{z}_{k-1}. \tag{27}$$

We remark that if the matrix $\bar{V}_{k-1}$ in (26) is replaced by V_{k-1}, then the standard SYMMLQ method [16] is obtained. The iteration polynomial q_{k-1} associated with the standard SYMMLQ method, in general, does not satisfy condition (6). The implementation of our method uses the QR-factorization of the matrix T_k, similarly as the implementation of the SYMMLQ method described in [8, Section 6.5]. In contrast, the implementation of the SYMMLQ method presented in [16] is based on the LQ-factorization of T_k.

It follows from (17) and (19) that

$$(AV_{k-1})^{\mathrm{T}}V_kQ_{k,k-1} = T_{k,k-1}^{\mathrm{T}}Q_{k,k-1} = \bar{R}_{k-1}^{\mathrm{T}}. \tag{28}$$

Substituting (28) into (27) yields

$$\bar{R}_{k-1}^{\mathrm{T}}\boldsymbol{z}_{k-1} = ||\boldsymbol{b}||\tilde{\boldsymbol{e}}_1. \tag{29}$$

This defines the iterative method.

Recursion formulas for updating the approximate solutions $\boldsymbol{x}_k$ inexpensively are derived in Section 2.2. In the remainder of this subsection, we discuss how to evaluate the right-hand side of (5). Eqs. (24) and (25) show that $\boldsymbol{x}_k \in \mathbb{K}_{k-1}(A, A\boldsymbol{b})$, and therefore there is a vector $\boldsymbol{y}_{k-1} \in \mathbb{R}^{k-1}$, such that

$$A^{-1}\boldsymbol{x}_k = V_{k-1}\boldsymbol{y}_{k-1}. \tag{30}$$

Thus, by (17),

$$\boldsymbol{x}_k = AV_{k-1}\boldsymbol{y}_{k-1} = V_kT_{k,k-1}\boldsymbol{y}_{k-1}, \tag{31}$$

and, by (25) and (22), we have

$$\boldsymbol{x}_k = V_kQ_{k,k-1}\boldsymbol{z}_{k-1}.$$

It follows that

$$Q_{k,k-1}\boldsymbol{z}_{k-1} = T_{k,k-1}\boldsymbol{y}_{k-1}. \tag{32}$$

Multiplying this equation by $Q_{k,k-1}^{\mathrm{T}}$ yields, in view of (19), that

$$\boldsymbol{z}_{k-1} = Q_{k,k-1}^{\mathrm{T}}T_{k,k-1}\boldsymbol{y}_{k-1} = \bar{R}_{k-1}\boldsymbol{y}_{k-1}. \tag{33}$$

Application of (30), (12), (33) and (29), in order, yields

$$\boldsymbol{b}^{\mathrm{T}}A^{-1}\boldsymbol{x}_k = \boldsymbol{b}^{\mathrm{T}}V_{k-1}\boldsymbol{y}_{k-1} = ||\boldsymbol{b}||\tilde{\boldsymbol{e}}_1^{\mathrm{T}}\boldsymbol{y}_{k-1} = ||\boldsymbol{b}||\tilde{\boldsymbol{e}}_1^{\mathrm{T}}\bar{R}_{k-1}^{-1}\boldsymbol{z}_{k-1} = \boldsymbol{z}_{k-1}^{\mathrm{T}}\boldsymbol{z}_{k-1}. \tag{34}$$

It follows from (25) that $\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{x}_k = \boldsymbol{z}_{k-1}^{\mathrm{T}}\boldsymbol{z}_{k-1}$. This observation and (34) show that Eq. (5) can be written in the form

$$\boldsymbol{e}_k^{\mathrm{T}}\boldsymbol{e}_k = \boldsymbol{r}_k^{\mathrm{T}}A^{-2}\boldsymbol{r}_k = \boldsymbol{b}^{\mathrm{T}}A^{-2}\boldsymbol{b} - \boldsymbol{z}_{k-1}^{\mathrm{T}}\boldsymbol{z}_{k-1}. \tag{35}$$

The term $\boldsymbol{z}_{k-1}^{\mathrm{T}}\boldsymbol{z}_{k-1}$ is straightforward to evaluate from (29). Section 3 describes how easily computable upper and lower bounds, or estimates, of $\boldsymbol{b}^{\mathrm{T}}A^{-2}\boldsymbol{b}$ can be derived by using Gauss-type quadrature rules. In this manner, we obtain easily computable upper and lower bounds, or estimates, of the norm of $\boldsymbol{e}_k$. Details are described in Section 3.

Assume for the moment that n steps of the Lanczos process have been carried out to yield the Lanczos decomposition $AV_n = V_nT_n$, analogous to (11). Using the QR-factorization (18) of T_n and the property (12) yields

$$\begin{aligned}\boldsymbol{b}^{\mathrm{T}}A^{-2}\boldsymbol{b} &= ||\boldsymbol{b}||^2\tilde{\boldsymbol{e}}_1^{\mathrm{T}}V_n^{\mathrm{T}}A^{-2}V_n\tilde{\boldsymbol{e}}_1 = ||\boldsymbol{b}||^2\tilde{\boldsymbol{e}}_1^{\mathrm{T}}T_n^{-2}\tilde{\boldsymbol{e}}_1\\ &= ||\boldsymbol{b}||^2\tilde{\boldsymbol{e}}_1^{\mathrm{T}}R_n^{-1}R_n^{-\mathrm{T}}\tilde{\boldsymbol{e}}_1.\end{aligned}$$

Substituting this expression into (35) and using (29) shows that

$$\boldsymbol{e}_k^{\mathrm{T}}\boldsymbol{e}_k = ||\boldsymbol{b}||^2\tilde{\boldsymbol{e}}_1^{\mathrm{T}}R_n^{-1}R_n^{-\mathrm{T}}\tilde{\boldsymbol{e}}_1 - ||\boldsymbol{b}||^2\tilde{\boldsymbol{e}}_1^{\mathrm{T}}\bar{R}_{k-1}^{-1}\bar{R}_{k-1}^{-\mathrm{T}}\tilde{\boldsymbol{e}}_1. \tag{36}$$

The right-hand side of (36) is analogous to expressions for the A-norm of the error $\boldsymbol{e}_k$ discussed in [10,11,14].

2.2. Updating formulas for the iterative method

We describe how the computation of the iterates $\boldsymbol{x}_k$ defined by (25) can be organized so that storage of only a few n-vectors is required.

Let the matrix T_k in (11) have the entries

$$T_k = \begin{bmatrix} \alpha_1 & \beta_1 & & & & 0 \\ \beta_1 & \alpha_2 & \beta_2 & & & \\ & \beta_2 & \alpha_3 & & & \\ & & & \ddots & & \\ & & \ddots & \ddots & \beta_{k-2} & \\ & & & \beta_{k-2} & \alpha_{k-1} & \beta_{k-1} \\ 0 & & & & \beta_{k-1} & \alpha_k \end{bmatrix} \in \mathbb{R}^{k\times k}, \tag{37}$$

where according to the discussion following equation (12) we may assume that the β_j are nonvanishing. This property of the β_j secures that the eigenvalues of T_k are distinct. Introduce the spectral factorization

$$T_k = W_k\Theta_kW_k^{\mathrm{T}}, \quad W_k \in \mathbb{R}^{k\times k}, \quad W_k^{\mathrm{T}}W_k = I_k,$$

$$\Theta_k = \mathrm{diag}[\theta_1^{(k)}, \theta_2^{(k)}, \ldots, \theta_k^{(k)}], \quad \theta_1^{(k)} < \theta_2^{(k)} < \cdots < \theta_k^{(k)}. \tag{38}$$

The QR-factorization (18) of T_k is computed by applying $k-1$ Givens rotations

$$G_k^{(j)} := \begin{bmatrix} I_{j-1} & & & \\ & c_j & s_j & \\ & -s_j & c_j & \\ & & & I_{k-j-1} \end{bmatrix} \in \mathbb{R}^{k\times k}, \quad c_j^2+s_j^2=1, \; s_j \geqslant 0, \tag{39}$$

to T_k, i.e.,

$$R_k := G_k^{(k-1)} G_k^{(k-2)} \cdots G_k^{(1)} T_k, \qquad Q_k := G_k^{(1)\mathrm{T}} G_k^{(2)\mathrm{T}} \cdots G_k^{(k-1)\mathrm{T}}, \tag{40}$$

see, e.g., [12, Chapter 5] for a discussion on Givens rotations. In our iterative method the matrix Q_k is not explicitly formed; instead we use representation (40). Since T_k is tridiagonal, the upper triangular matrix R_k has nonvanishing entries on the diagonal and the two adjacent superdiagonals only.

The matrix T_k in (37) is determined by k steps of the Lanczos process. After an additional step, we obtain the Lanczos decomposition

$$AV_{k+1} = V_{k+1}T_{k+1} + \boldsymbol{f}_{k+1}\tilde{\boldsymbol{e}}_{k+1}^{\mathrm{T}}, \tag{41}$$

analogous to (11). For future reference, we remark that the last subdiagonal entry of the symmetric tridiagonal matrix T_{k+1} may be computed by

$$\beta_k := \|\boldsymbol{f}_k\| \tag{42}$$

already after completion of k Lanczos steps.

The matrix T_{k+1} has the QR-factorization

$$T_{k+1} = Q_{k+1}R_{k+1}, \tag{43}$$

whose factors can be computed from Q_k and R_k in a straightforward manner. We have

$$Q_{k+1} = \begin{bmatrix} Q_k & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{bmatrix} G_{k+1}^{(k)\mathrm{T}} \in \mathbb{R}^{(k+1)\times(k+1)},$$

$$Q_{k+1,k} = \begin{bmatrix} Q_k & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{bmatrix} G_{k+1,k}^{(k)\mathrm{T}} \in \mathbb{R}^{(k+1)\times k}, \tag{44}$$

where $G_{k+1}^{(k)}$ is defined by (39) and $G_{k+1,k}^{(k)} \in \mathbb{R}^{(k+1)\times k}$ is made up of the first k columns of $G_{k+1}^{(k)}$.

We obtain updating formulas for computing the triangular matrix R_{k+1} in (43) from the matrix R_k in (40) by expressing these matrices in terms of their columns

$$R_k = [\boldsymbol{r}_1^{(k)}, \boldsymbol{r}_2^{(k)}, \ldots, \boldsymbol{r}_k^{(k)}], \qquad R_{k+1} = [\boldsymbol{r}_1^{(k+1)}, \boldsymbol{r}_2^{(k+1)}, \ldots, \boldsymbol{r}_k^{(k+1)}, \boldsymbol{r}_{k+1}^{(k+1)}].$$

Comparing (18) and (43) yields

$$\boldsymbol{r}_j^{(k+1)} = \begin{bmatrix} \boldsymbol{r}_j^{(k)} \\ 0 \end{bmatrix}, \quad 1 \leqslant j < k \tag{45}$$

and

$$\boldsymbol{r}_k^{(k+1)} = G_{k+1}^{(k)} \begin{bmatrix} \boldsymbol{r}_k^{(k)} \\ \beta_k \end{bmatrix},$$

$$\boldsymbol{r}_{k+1}^{(k+1)} = G_{k+1}^{(k)} G_{k+1}^{(k-1)} T_{k+1} \tilde{\boldsymbol{e}}_{k+1}. \tag{46}$$

Thus, the entries of all the matrices $R_1, R_2, \ldots, R_{k+1}$ can be computed in only $\mathrm{O}(k)$ arithmetic floating-point operations.

The matrix $\bar{R}_k = [\bar{r}_{j\ell}^{(k)}]_{j,\ell=1}^k$ defined by (19) is the leading principal submatrix of R_{k+1} of order k and agrees with $R_k = [r_{j\ell}^{(k)}]_{j,\ell=1}^k$ except for the last diagonal entry. Eq. (46) and the fact that β_k is nonvanishing yield

$$\bar{r}_{kk}^{(k)} > r_{kk}^{(k)} \geqslant 0, \tag{47}$$

and when T_k is nonsingular, we have $r_{kk}^{(k)} > 0$.

We turn to the computation of the columns of

$$\tilde{V}_{k+1} = [\tilde{\boldsymbol{v}}_1^{(k+1)}, \tilde{\boldsymbol{v}}_2^{(k+1)}, \ldots, \tilde{\boldsymbol{v}}_{k+1}^{(k+1)}] := V_{k+1} Q_{k+1} \tag{48}$$

from those of the matrix $\tilde{V}_k$, where V_{k+1} is determined by the Lanczos decomposition (41) and Q_{k+1} is given by (44). Substituting (44) into the right-hand side of (48) yields

$$\begin{aligned} \tilde{V}_{k+1} &= [V_k, \boldsymbol{v}_{k+1}] Q_{k+1} = [\tilde{V}_k, \boldsymbol{v}_{k+1}] G_{k+1}^{(k)\mathrm{T}} \\ &= [\bar{V}_{k-1}, c_k \tilde{\boldsymbol{v}}_k^{(k)} + s_k \boldsymbol{v}_{k+1}, -s_k \tilde{\boldsymbol{v}}_k^{(k)} + c_k \boldsymbol{v}_{k+1}]. \end{aligned} \tag{49}$$

Thus, the first $k-1$ columns of the matrix $\tilde{V}_{k+1}$ are the columns of $\bar{V}_{k-1}$. The columns $\tilde{\boldsymbol{v}}_k^{(k+1)}$ and $\tilde{\boldsymbol{v}}_{k+1}^{(k+1)}$ of $\tilde{V}_{k+1}$ are linear combinations of the last columns of $\tilde{V}_k$ and V_{k+1}.

Assume that the solution $\boldsymbol{z}_{k-1}$ of the linear system (29) is available. Since the matrix $\bar{R}_k$ is upper triangular and $\bar{R}_{k-1}$ is the leading principal submatrix of order $k-1$ of $\bar{R}_k$, the computation of the solution $\boldsymbol{z}_k = [\zeta_1, \zeta_2, \ldots, \zeta_k]^{\mathrm{T}}$ of

$$\bar{R}_k^{\mathrm{T}} \boldsymbol{z}_k = ||\boldsymbol{b}|| \tilde{\boldsymbol{e}}_1 \tag{50}$$

is easy. We have

$$\boldsymbol{z}_k = \begin{bmatrix} \boldsymbol{z}_{k-1} \\ \zeta_k \end{bmatrix}, \qquad \zeta_k = -(\bar{r}_{k-2,k}^{(k)} \zeta_{k-2} + \bar{r}_{k-1,k}^{(k)} \zeta_{k-1}) / \bar{r}_{kk}^{(k)}. \tag{51}$$

Hence, only the last column of the matrix $\bar{R}_k$ is required.

We are now in a position to compute $\boldsymbol{x}_{k+1}$ from $\boldsymbol{x}_k$. Eqs. (25) and (49) yield

$$\boldsymbol{x}_{k+1} = \bar{V}_k \boldsymbol{z}_k = \bar{V}_{k-1} \boldsymbol{z}_{k-1} + \zeta_k \tilde{\boldsymbol{v}}_k^{(k+1)} = \boldsymbol{x}_k + \zeta_k \tilde{\boldsymbol{v}}_k^{(k+1)},$$

where we have used that $\tilde{\boldsymbol{v}}_k^{(k+1)}$ is the last column of $\bar{V}_k$. Note that only the last few columns of V_k and $\tilde{V}_k$ have to be stored in order to update the approximate solution $\boldsymbol{x}_k$.

3. Quadrature rules of Gauss-type for error estimation

This section describes how to bound or compute estimates of the matrix functional (8) by approximating the Stieltjes integral representation (9) by quadrature rules of Gauss-type. A nice discussion on the application of Gauss quadrature rules to the evaluation of upper and lower bounds of certain matrix functionals is presented in [9]. Related discussions can also be found in [2,4,11].

3.1. Gauss quadrature rules

Let f be a $2k$ times continuously differentiable function defined on the interval $[\lambda_1, \lambda_n]$, which contains the support of the measure ω. The k-point Gauss quadrature rule associated with ω for the computation of an approximation of the integral (10) is given by

$$\mathscr{G}_k(f) := \sum_{j=1}^{k} f(\theta_j^{(k)})\omega_j^{(k)}, \qquad \omega_j^{(k)} := ||\boldsymbol{b}||^2(\tilde{\boldsymbol{e}}_1^{\mathrm{T}} W_k \tilde{\boldsymbol{e}}_j)^2, \tag{52}$$

where the $\theta_j^{(k)}$ and W_k are defined by (38). The nodes and weights of the Gauss rule are uniquely determined by the requirement

$$\mathscr{G}_k(p) = \mathscr{I}(p), \quad \forall p \in \Pi_{2k-1}, \tag{53}$$

where $\mathscr{I}$ is defined by (10). We also will use the representation

$$\mathscr{G}_k(f) = ||\boldsymbol{b}||^2 \tilde{\boldsymbol{e}}_1^{\mathrm{T}} f(T_k) \tilde{\boldsymbol{e}}_1. \tag{54}$$

The equivalence of (52) and (54) is shown in [9] and follows by substituting the spectral factorization (38) into (54). The integration error

$$\mathscr{E}_k(f) := \mathscr{I}(f) - \mathscr{G}_k(f)$$

can be expressed as

$$\mathscr{E}_k(f) = \frac{f^{(2k)}(\tilde{\theta}^{(k)})}{(2k)!} \int_{-\infty}^{\infty} \prod_{\ell=1}^{k} (t - \theta_\ell^{(k)})^2 \, \mathrm{d}\omega(t) \tag{55}$$

for some $\tilde{\theta}^{(k)}$ in the interval $[\lambda_1, \lambda_n]$, where $f^{(2k)}$ denotes the derivative of order $2k$ of the function f; see, e.g., [9] or [18, Section 3.6] for details.

In the remainder of this section, we will assume that f is given by (8) and that the matrix A is positive definite. Then $f^{(2k)}(t) > 0$ for $t > 0$, and the constant $\tilde{\theta}^{(k)}$ in (55) is positive. It follows from (55) that $\mathscr{E}_k(f) > 0$, and therefore

$$\mathscr{G}_k(f) < \mathscr{I}(f) = F(A) = \boldsymbol{b}^{\mathrm{T}} A^{-2} \boldsymbol{b}, \tag{56}$$

where $F(A)$ is defined by (8).

Representation (54) of the Gauss quadrature rule can be simplified by using the QR-factorization (18) of T_k when f is given by (8),

$$\mathscr{G}_k(f) = ||\boldsymbol{b}||^2 \tilde{\boldsymbol{e}}_1^{\mathrm{T}} T_k^{-2} \tilde{\boldsymbol{e}}_1 = ||\boldsymbol{b}||^2 \tilde{\boldsymbol{e}}_1^{\mathrm{T}} R_k^{-1} R_k^{-\mathrm{T}} \tilde{\boldsymbol{e}}_1 = ||\boldsymbol{b}||^2 ||R_k^{-\mathrm{T}} \tilde{\boldsymbol{e}}_1||^2. \tag{57}$$

It is easy to evaluate the right-hand side of (57) when the solution $\boldsymbol{z}_{k-1}$ of (29) is available. Let $\tilde{\boldsymbol{z}}_k \in \mathbb{R}^k$ satisfy

$$R_k^{\mathrm{T}} \tilde{\boldsymbol{z}}_k = ||\boldsymbol{b}|| \tilde{\boldsymbol{e}}_1. \tag{58}$$

Then

$$\mathscr{G}_k(f) = \tilde{\boldsymbol{z}}_k^{\mathrm{T}} \tilde{\boldsymbol{z}}_k. \tag{59}$$

Since all entries $r_{j\ell}^{(k)}$ of R_k and $\bar{r}_{j\ell}^{(k)}$ of $\bar{R}_k$ are the same, except for $r_{kk}^{(k)} \neq \bar{r}_{kk}^{(k)}$, the solution of (58) is given by

$$\tilde{\boldsymbol{z}}_k = \begin{bmatrix} \boldsymbol{z}_{k-1} \\ \tilde{\zeta}_k \end{bmatrix}, \qquad \tilde{\zeta}_k = -(\bar{r}_{k-2,k}^{(k)}\zeta_{k-2} + \bar{r}_{k-1,k}^{(k)}\zeta_{k-1})/r_{kk}^{(k)}. \tag{60}$$

Substituting inequality (56) into (35) (with k replaced by $k+1$) and using representation (59) yields

$$\boldsymbol{e}_{k+1}^{\mathrm{T}}\boldsymbol{e}_{k+1} > \tilde{\boldsymbol{z}}_k^{\mathrm{T}}\tilde{\boldsymbol{z}}_k - \boldsymbol{z}_k^{\mathrm{T}}\boldsymbol{z}_k = \tilde{\zeta}_k^2 - \zeta_k^2, \tag{61}$$

where the equality follows from (51) and (60). A comparison of (51) and (60) yields, in view of inequality (47), that $|\tilde{\zeta}_k| \geqslant |\zeta_k|$, and therefore the right-hand side of (61) is nonnegative. Moreover, if $\tilde{\zeta}_k \neq 0$, then $|\tilde{\zeta}_k| > |\zeta_k|$, and we obtain

$$||\boldsymbol{e}_{k+1}|| > \sqrt{\tilde{\zeta}_k^2 - \zeta_k^2} > 0. \tag{62}$$

Thus, Gauss quadrature rules give easily computable lower bounds for the error in the approximate solutions generated by the iterative method when applied to linear systems of equations with a symmetric positive-definite matrix.

3.2. Anti-Gauss quadrature rules

Let the matrix A be symmetric and positive definite. If the smallest eigenvalue λ_1 of A were explicitly known, then an upper bound of (56) could be computed by a $(k+1)$-point Gauss–Radau quadrature rule with a fixed node between λ_1 and the origin; see [9,10] for details. The computed bound typically improves the further away from the origin we can allocate the fixed node. However, accurate lower bounds for λ_1 are, in general, not available. We therefore propose to use anti-Gauss quadrature rules to compute estimates of the error that generally are of opposite sign as $\mathscr{E}_k(f)$.

Anti-Gauss rules were introduced in [13], and their application to the evaluation of matrix functionals was explored in [4]. Let f be a smooth function. Analogously to representation (54) of the k-point Gauss rule, the $(k+1)$-point anti-Gauss quadrature rule associated with ω for the computation of an approximation of integral (10) is given by

$$\breve{\mathscr{G}}_{k+1}(f) := ||\boldsymbol{b}||^2\tilde{\boldsymbol{e}}_1^{\mathrm{T}} f(\breve{T}_{k+1})\tilde{\boldsymbol{e}}_1, \tag{63}$$

where

$$\breve{T}_{k+1} = \begin{bmatrix} \alpha_1 & \beta_1 & & & & & 0 \\ \beta_1 & \alpha_2 & \beta_2 & & & & \\ & \beta_2 & \alpha_3 & & & & \\ & & & \ddots & & & \\ & & & \ddots & \ddots & \beta_{k-1} & \\ & & & & \beta_{k-1} & \alpha_k & \sqrt{2}\beta_k \\ 0 & & & & & \sqrt{2}\beta_k & \alpha_{k+1} \end{bmatrix} \in \mathbb{R}^{(k+1)\times(k+1)}. \tag{64}$$

Thus, $\breve{T}_{k+1}$ is obtained from T_{k+1} by multiplying the last off-diagonal entries by $\sqrt{2}$. We note that the determination of $\breve{T}_{k+1}$ requires application of $k+1$ steps of the Lanczos process; cf. (11).

The $(k+1)$-point anti-Gauss rule is characterized by the requirement that the integration error

$$\breve{\mathscr{E}}_{k+1}(f) := \mathscr{I}(f) - \breve{\mathscr{G}}_{k+1}(f)$$

satisfies

$$\breve{\mathscr{E}}_{k+1}(p) = -\mathscr{E}_k(p), \quad \forall p \in \Pi_{2k+1},$$

which can be written in the equivalent form

$$\breve{\mathscr{G}}_{k+1}(p) = (2\mathscr{I} - \mathscr{G}_k)(p), \quad \forall p \in \Pi_{2k+1}. \tag{65}$$

Assume for the moment that we can carry out n steps of the Lanczos process without break down. This yields an orthonormal basis $\{\boldsymbol{v}_j\}_{j=1}^n$ of $\mathbb{R}^n$ and an associated sequence of polynomials $\{s_j\}_{j=0}^{n-1}$ defined by (13) that satisfy (15). Expanding the function f on the spectrum of A, denoted by $\lambda(A)$, in terms of the polynomials s_j yields

$$f(t) = \sum\nolimits_{j=0}^{n-1} \eta_j s_j(t), \quad t \in \lambda(A), \tag{66}$$

where $\eta_j = (f, s_j)$, with the inner product defined by (14).

In view of $\mathscr{I}(s_j) = 0$ for $j > 0$ and (53), it follows from (66) that

$$\mathscr{I}(f) = \eta_0 \mathscr{I}(s_0) = \eta_0 \mathscr{G}_k(s_0). \tag{67}$$

Therefore, applying the Gauss rule $\mathscr{G}_k$ and anti-Gauss rule $\breve{\mathscr{G}}_{k+1}$ to (66), using (53), (65) and (67), yields for $n \geqslant 2k+2$ that

$$\mathscr{G}_k(f) = \mathscr{I}(f) + \sum_{j=2k}^{n-1} \eta_j \mathscr{G}_k(s_j), \tag{68}$$

$$\begin{aligned}
\breve{\mathscr{G}}_{k+1}(f) &= \sum_{j=0}^{n-1} \eta_j \breve{\mathscr{G}}_{k+1}(s_j) = \sum_{j=0}^{2k+1} \eta_j (2\mathscr{I} - \mathscr{G}_k)(s_j) + \sum_{j=2k+2}^{n-1} \eta_j \breve{\mathscr{G}}_{k+1}(s_j) \\
&= \sum_{j=0}^{2k+1} \eta_j 2\mathscr{I}(s_j) - \sum_{j=0}^{2k+1} \eta_j \mathscr{G}_k(s_j) + \sum_{j=2k+2}^{n-1} \eta_j \breve{\mathscr{G}}_{k+1}(s_j) \\
&= \mathscr{I}(f) - \eta_{2k} \mathscr{G}_k(s_{2k}) - \eta_{2k+1} \mathscr{G}_k(s_{2k+1}) + \sum_{j=2k+2}^{n-1} \eta_j \breve{\mathscr{G}}_{k+1}(s_j).
\end{aligned} \tag{69}$$

Assume that the coefficients η_j converge rapidly to zero with increasing index. Then the leading terms in expansions (68) and (69) dominate the error, i.e.,

$$\begin{aligned}
\mathscr{E}_k(f) &= (\mathscr{I} - \mathscr{G}_k)(f) \approx -\eta_{2k} \mathscr{G}_k(s_{2k}) - \eta_{2k+1} \mathscr{G}_k(s_{2k+1}), \\
\breve{\mathscr{E}}_{k+1}(f) &= (\mathscr{I} - \breve{\mathscr{G}}_{k+1})(f) \approx \eta_{2k} \mathscr{G}_k(s_{2k}) + \eta_{2k+1} \mathscr{G}_k(s_{2k+1}),
\end{aligned} \tag{70}$$

where $\approx$ stands for "approximately equal to". This leads us to expect that, in general, the errors $\mathscr{E}_k(f)$ and $\breve{\mathscr{E}}_{k+1}(f)$ are of opposite sign and of roughly the same magnitude.

In the remainder of this subsection, we let f be defined by (8) and discuss the evaluation of anti-Gauss rules for this particular integrand. Introduce the QR-factorization

$$\breve{T}_{k+1} = \breve{Q}_{k+1}\breve{R}_{k+1}, \quad \breve{Q}_{k+1}, \breve{R}_{k+1} \in \mathbb{R}^{(k+1)\times(k+1)}, \quad \breve{Q}_{k+1}^{\mathrm{T}}\breve{Q}_{k+1} = I_{k+1}, \tag{71}$$

where $\breve{R}_{k+1} = [\breve{r}_{j\ell}^{(k+1)}]_{j,\ell=1}^{k+1}$ is upper triangular. Using representation (63), we obtain, analogously to (57), that

$$\breve{\mathscr{G}}_{k+1}(f) = ||\boldsymbol{b}||^2 \tilde{\boldsymbol{e}}_1^{\mathrm{T}} \breve{T}_{k+1}^{-2} \tilde{\boldsymbol{e}}_1 = ||\boldsymbol{b}||^2 \tilde{\boldsymbol{e}}_1^{\mathrm{T}} \breve{R}_{k+1}^{-1} \breve{R}_{k+1}^{-\mathrm{T}} \tilde{\boldsymbol{e}}_1 = ||\boldsymbol{b}||^2 ||\breve{R}_{k+1}^{-\mathrm{T}} \tilde{\boldsymbol{e}}_1||^2. \tag{72}$$

Since by (56) we have $\mathscr{E}_k(f) > 0$, Eq. (70) suggests that, typically, $\breve{\mathscr{E}}_{k+1}(f) < 0$. Thus, we expect that for many symmetric positive-definite matrices A, right-hand side vectors $\boldsymbol{b}$ and values of k, the inequality

$$\breve{\mathscr{G}}_{k+1}(f) > \mathscr{I}(f) = F(A) = \boldsymbol{b}^{\mathrm{T}} A^{-2} \boldsymbol{b} \tag{73}$$

holds, where f and F are given by (8).

Let $\breve{\boldsymbol{z}}_{k+1}$ satisfy

$$\breve{R}_{k+1}^{\mathrm{T}} \breve{\boldsymbol{z}}_{k+1} = ||\boldsymbol{b}|| \tilde{\boldsymbol{e}}_1. \tag{74}$$

Then it follows from (72) that

$$\breve{\mathscr{G}}_{k+1}(f) = \breve{\boldsymbol{z}}_{k+1}^{\mathrm{T}} \breve{\boldsymbol{z}}_{k+1}. \tag{75}$$

The matrix $\breve{R}_{k+1}$ can be determined when $k+1$ Lanczos steps have been completed, and so can the approximate solution $\boldsymbol{x}_{k+1}$ of (1). Substituting (73) into (35) (with k replaced by $k+1$) and using representation (75) suggests that the inequality

$$\boldsymbol{e}_{k+1}^{\mathrm{T}} \boldsymbol{e}_{k+1} < \breve{\boldsymbol{z}}_{k+1}^{\mathrm{T}} \breve{\boldsymbol{z}}_{k+1} - \boldsymbol{z}_k^{\mathrm{T}} \boldsymbol{z}_k \tag{76}$$

holds for many symmetric positive-definite matrices A, right-hand side vectors $\boldsymbol{b}$ and values of k.

We evaluate the right-hand side of (76) by using the close relation between the upper triangular matrices $\breve{R}_{k+1}$ and $\bar{R}_k$. Assume that $\bar{R}_k$ is nonsingular and that $\beta_{k+1} \neq 0$. It is easy to see that the $k \times k$ leading principal submatrix of $\breve{R}_{k+1}$ agrees with $\bar{R}_k$ except for its last diagonal entry. A comparison of (74) with (29) (with $k-1$ replaced by k) shows that

$$\breve{\boldsymbol{z}}_{k+1} = \begin{bmatrix} \boldsymbol{z}_{k-1} \\ \breve{\zeta}_k^{(k+1)} \\ \breve{\zeta}_{k+1}^{(k+1)} \end{bmatrix},$$

where

$$\begin{aligned} \breve{\zeta}_k^{(k+1)} &= -(\bar{r}_{k-2,k}^{(k)} \zeta_{k-2} + \bar{r}_{k-1,k}^{(k)} \zeta_{k-1}) / \breve{r}_{kk}^{(k+1)}, \\ \breve{\zeta}_{k+1}^{(k+1)} &= -(\breve{r}_{k-1,k+1}^{(k+1)} \zeta_{k-1} + \breve{r}_{k,k+1}^{(k+1)} \breve{\zeta}_k^{(k+1)}) / \breve{r}_{k+1,k+1}^{(k+1)} \end{aligned}$$

and the ζ_j are entries of $\boldsymbol{z}_{k-1}$. Thus,

$$\breve{\boldsymbol{z}}_{k+1}^{\mathrm{T}} \breve{\boldsymbol{z}}_{k+1} - \boldsymbol{z}_k^{\mathrm{T}} \boldsymbol{z}_k = (\breve{\zeta}_{k+1}^{(k+1)})^2 + (\breve{\zeta}_k^{(k+1)})^2 - \zeta_k^2.$$

Substitution of this identity into (76) yields

$$||\boldsymbol{e}_{k+1}|| < \sqrt{(\breve{\zeta}_{k+1}^{(k+1)})^2 + (\breve{\zeta}_{k}^{(k+1)})^2 - \zeta_k^2}. \tag{77}$$

According to the above discussion, we expect the argument of the square root to be positive and the inequality to hold for many symmetric positive-definite matrices A, right-hand side vectors $\boldsymbol{b}$ and values of k. We refer to the right-hand side of (77) as an upper estimate of the norm of the error $\boldsymbol{e}_{k+1}$. However, we point out that inequality (77) might be violated for some values of k. This is illustrated in the numerical examples of Section 4.

3.3. Gauss–Radau quadrature rules

Throughout this section we assume that the matrix A is nonsingular and indefinite. Thus, there is an index m such that eigenvalues (7) of A satisfy

$$\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_m < 0 < \lambda_{m+1} \leqslant \cdots \leqslant \lambda_n. \tag{78}$$

The application of Gauss quadrature rules (52) to estimate the norm of the error in approximate solutions $\boldsymbol{x}_k$ might not be possible for all values of k when A is indefinite, because for some $k > 0$ one of the nodes $\theta_j^{(k)}$ of the Gauss rule (52) may be at the origin, and the integrand f given by (8) is not defined there. In fact, numerical difficulties may arise also when one of the nodes $\theta_j^{(k)}$ is very close to the origin. We circumvent this problem by modifying the integrand and estimating the norm of the error in the computed approximate solutions by Gauss–Radau quadrature rules associated with the measure ω and with a fixed node at the origin. Note that since the matrix A is indefinite, the origin is inside the smallest interval containing the spectrum of A. Some of the desired Gauss–Radau rules therefore might not exist. We will return to this issue below.

Let f be a smooth function on a sufficiently large interval that contains $\lambda(A)$ in its interior. We may, for instance, think of f as analytic. The $(k+1)$-point Gauss–Radau quadrature rule associated with the measure ω and with a fixed node $\hat{\theta}_1$ at the origin for the integration of f is of the form

$$\hat{\mathscr{G}}_{k+1}(f) := \sum_{j=1}^{k+1} f(\hat{\theta}_j^{(k+1)})\hat{\omega}_j^{(k+1)}. \tag{79}$$

It is characterized by the requirements that

$$\hat{\mathscr{G}}_{k+1}(p) = \mathscr{I}(p), \quad \forall p \in \Pi_{2k} \quad \text{and} \quad \hat{\theta}_1^{(k+1)} = 0.$$

The nodes and weights in (79) are given by formulas analogous to those for the nodes and weights of standard Gauss rules (52). Introduce the symmetric tridiagonal matrix

$$\hat{T}_{k+1} = \begin{bmatrix} \alpha_1 & \beta_1 & & & & & 0 \\ \beta_1 & \alpha_2 & \beta_2 & & & & \\ & \beta_2 & \alpha_3 & & & & \\ & & & \ddots & & & \\ & & & \ddots & \ddots & \beta_{k-1} & \\ & & & & \beta_{k-1} & \alpha_k & \beta_k \\ 0 & & & & & \beta_k & \hat{\alpha}_{k+1} \end{bmatrix} \in \mathbb{R}^{(k+1)\times(k+1)}, \tag{80}$$

where

$$\hat{\alpha}_{k+1} := \beta_k^2 \tilde{\boldsymbol{e}}_k^{\mathrm{T}} T_k^{-1} \tilde{\boldsymbol{e}}_k$$

and T_k is given by (37). In view of the discussion on the computation of β_k, see (42), all entries of the matrix $\hat{T}_{k+1}$ can be computed after k Lanczos steps have been completed, provided that the matrix T_k is invertible. Since A is indefinite, we cannot exclude that T_k is singular. However, because of the interlacing property of the eigenvalues of the matrices T_k and T_{k+1}, it follows that if T_k is singular, then T_{k+1} is not. Thus, the desired $(k+1)$-point Gauss–Radau rules can be determined for at least every other value of k.

Define the spectral factorization

$$\hat{T}_{k+1} = \hat{W}_{k+1} \hat{\Theta}_{k+1} \hat{W}_{k+1}^{\mathrm{T}}, \quad \hat{W}_{k+1} \in \mathbb{R}^{(k+1)\times(k+1)}, \quad \hat{W}_{k+1}^{\mathrm{T}} \hat{W}_{k+1} = I_{k+1},$$
$$\hat{\Theta}_{k+1} = \mathrm{diag}\,[\hat{\theta}_1^{(k+1)}, \hat{\theta}_2^{(k+1)}, \ldots, \hat{\theta}_{k+1}^{(k+1)}], \quad 0 = \hat{\theta}_1^{(k+1)} < |\hat{\theta}_2^{(k+1)}| \leqslant \cdots \leqslant |\hat{\theta}_{k+1}^{(k+1)}|.$$

The eigenvalues $\hat{\theta}_j^{(k+1)}$ are distinct and may be positive or negative. The nodes in the Gauss–Radau quadrature rule (79) are the eigenvalues $\hat{\theta}_j^{(k+1)}$ and the weights are given by

$$\hat{\omega}_j^{(k+1)} := \|\boldsymbol{b}\|^2 (\tilde{\boldsymbol{e}}_1^{\mathrm{T}} \hat{W}_{k+1} \tilde{\boldsymbol{e}}_j)^2,$$

see [9] for details. Analogously to (54), the quadrature rule (79) also can be represented by

$$\hat{\mathscr{G}}_{k+1}(f) = \|\boldsymbol{b}\|^2 \tilde{\boldsymbol{e}}_1^{\mathrm{T}} f(\hat{T}_{k+1}) \tilde{\boldsymbol{e}}_1. \tag{81}$$

Let for the moment f be a function that is analytic on an interval that contains all eigenvalues of A and all Gauss–Radau nodes $\hat{\theta}_j^{(k+1)}$, and satisfies

$$f(t) := \begin{cases} 1/t^2, & t \in \lambda(A) \cup \{\hat{\theta}_j^{(k+1)}\}_{j=2}^{k+1}, \\ 0, & t = 0. \end{cases} \tag{82}$$

Then

$$\mathscr{I}(f) = \boldsymbol{b}^{\mathrm{T}} A^{-2} \boldsymbol{b} = \boldsymbol{b}^{\mathrm{T}} (A^{\dagger})^2 \boldsymbol{b}$$

and representations (79) and (81) yield

$$\hat{\mathscr{G}}_{k+1}(f) = \sum_{j=2}^{k+1} (\hat{\theta}_j^{(k+1)})^{-2} \hat{\omega}_j^{(k+1)} = \|\boldsymbol{b}\|^2 \tilde{\boldsymbol{e}}_1^{\mathrm{T}} (\hat{T}_{k+1}^{\dagger})^2 \tilde{\boldsymbol{e}}_1 = \|\boldsymbol{b}\|^2 \|\hat{T}_{k+1}^{\dagger} \tilde{\boldsymbol{e}}_1\|^2, \tag{83}$$

where $M^{\dagger}$ denotes the Moore–Penrose pseudoinverse of the matrix M.

Proposition 3.1. *Let the index m be determined by* (78). *Then the nonvanishing eigenvalues $\hat{\theta}_j^{(k+1)}$ of the Gauss–Radau matrix $\hat{T}_{k+1}$ satisfy*

$$\hat{\theta}_j^{(k+1)} \leqslant \lambda_m \quad \textit{or} \; \hat{\theta}_j^{(k+1)} \geqslant \lambda_{m+1}, \quad 2 \leqslant j \leqslant k+1.$$

Proof. The result follows by combining Lemmas 5.2 and 5.3 of [3]. □

The proposition secures that none of the nonvanishing Gauss–Radau nodes is closer to the origin than the eigenvalue of A of smallest magnitude. This property does not hold for nodes in Gauss rules (52). Therefore, the symmetric tridiagonal matrices (37) associated with Gauss rules may be nearly singular, even when A is well conditioned. Near singularity of the tridiagonal matrices (37) makes the computed error estimates sensitive to propagated round-off errors, and may cause the computed estimates to be of poor quality. This is illustrated in Examples 3 and 4 of Section 4.

The error $(\mathscr{I}-\hat{\mathscr{G}}_{k+1})(f)$ can be expressed by a formula similar to (55). However, the derivatives of the integrand f change sign on the interval $[\lambda_1,\lambda_n]$ and the sign of the error cannot be determined from this formula. The Gauss–Radau rule only provides estimates of the error in the computed approximate solutions. The computed examples of Section 4 show these estimates to be close to the norm of the error in the computed approximate solutions. This is typical for our experience from a large number of computed examples.

We turn to the evaluation of Gauss–Radau rules (83). Define the QR-factorization

$$\hat{T}_{k+1}=Q_{k+1}\hat{R}_{k+1},\quad Q_{k+1},\hat{R}_{k+1}\in\mathbb{R}^{(k+1)\times(k+1)},\quad Q_{k+1}^{\mathrm{T}}Q_{k+1}=I_{k+1},\tag{84}$$

where $\hat{R}_{k+1}=[\hat{r}_{j\ell}^{(k+1)}]_{j,\ell=1}^{k+1}$ is upper triangular. Since $\hat{T}_{k+1}$ is singular, the entry $\hat{r}_{k+1,k+1}^{(k+1)}$ vanishes. Note that the matrix Q_{k+1} in (84) is the same as in (43). Moreover, the leading $k\times k$ principal submatrix of $\hat{R}_{k+1}$ is given by the matrix $\bar{R}_k$ in (50).

Let $\boldsymbol{q}_{k+1}^{(k+1)}$ denote the last column of Q_{k+1}. Then

$$\boldsymbol{q}_{k+1}^{(k+1)\mathrm{T}}\hat{T}_{k+1}=\boldsymbol{q}_{k+1}^{(k+1)\mathrm{T}}Q_{k+1}\hat{R}_{k+1}=\tilde{\boldsymbol{e}}_{k+1}^{\mathrm{T}}\hat{R}_{k+1}=\boldsymbol{0}^{\mathrm{T}}.$$

By symmetry of $\hat{T}_{k+1}$ it follows that

$$\hat{T}_{k+1}\boldsymbol{q}_{k+1}^{(k+1)}=\boldsymbol{0},$$

i.e., $\boldsymbol{q}_{k+1}^{(k+1)}$ spans the null space of $\hat{T}_{k+1}$ and is orthogonal to the range of $\hat{T}_{k+1}$. In particular, $I_{k+1}-\boldsymbol{q}_{k+1}^{(k+1)}\boldsymbol{q}_{k+1}^{(k+1)\mathrm{T}}$ is the orthogonal projector onto the range of $\hat{T}_{k+1}$.

We evaluate the right-hand side of (83) by using the QR-factorization (84) as follows. The vector $||\boldsymbol{b}||\hat{T}_{k+1}^{\dagger}\tilde{\boldsymbol{e}}_1$ is the solution of minimal norm of the least-squares problem

$$\min_{\boldsymbol{y}_{k+1}\in\mathbb{R}^{k+1}}||\hat{T}_{k+1}\boldsymbol{y}_{k+1}-||\boldsymbol{b}||\tilde{\boldsymbol{e}}_1||.\tag{85}$$

We may replace the vector $||\boldsymbol{b}||\tilde{\boldsymbol{e}}_1$ in (85) by its orthogonal projection onto the range of $\hat{T}_{k+1}$ without changing the solution of the least-squares problem. Thus, $||\boldsymbol{b}||\hat{T}_{k+1}^{\dagger}\tilde{\boldsymbol{e}}_1$ also is the solution of minimal norm of the least-squares problem

$$\min_{\boldsymbol{y}_{k+1}\in\mathbb{R}^{k+1}}||\hat{T}_{k+1}\boldsymbol{y}_{k+1}-(I_{k+1}-\boldsymbol{q}_{k+1}^{(k+1)}\boldsymbol{q}_{k+1}^{(k+1)\mathrm{T}})||\boldsymbol{b}||\tilde{\boldsymbol{e}}_1||.\tag{86}$$

Substituting $\hat{T}_{k+1}=\hat{T}_{k+1}^{\mathrm{T}}=\hat{R}_{k+1}^{\mathrm{T}}Q_{k+1}^{\mathrm{T}}$ into (86) and letting $\hat{\boldsymbol{y}}_{k+1}=Q_{k+1}^{\mathrm{T}}\boldsymbol{y}_{k+1}$ yields the consistent linear system of equations

$$\hat{R}_{k+1}^{\mathrm{T}}\hat{\boldsymbol{y}}_{k+1}=(I_{k+1}-\boldsymbol{q}_{k+1}^{(k+1)}\boldsymbol{q}_{k+1}^{(k+1)\mathrm{T}})||\boldsymbol{b}||\tilde{\boldsymbol{e}}_1.\tag{87}$$

Let $\hat{\boldsymbol{y}}_{k+1}$ denote the minimal norm solution of (87). Then $\hat{\boldsymbol{y}}_{k+1}=||\boldsymbol{b}||Q_{k+1}^{\mathrm{T}}\hat{T}_{k+1}^{\dagger}\tilde{\boldsymbol{e}}_1$ and therefore

$$||\hat{\boldsymbol{y}}_{k+1}||=||\boldsymbol{b}||||\hat{T}_{k+1}^{\dagger}\tilde{\boldsymbol{e}}_1||.\tag{88}$$

Since $\hat{r}_{k+1,k+1}^{(k+1)} = 0$ and $\hat{r}_{jj}^{(k+1)} > 0$ for $1 \leqslant j \leqslant k$, the minimal norm solution of (87) is of the form

$$\hat{\boldsymbol{y}}_{k+1} = \begin{bmatrix} \bar{\boldsymbol{y}}_k \\ 0 \end{bmatrix}, \quad \bar{\boldsymbol{y}}_k \in \mathbb{R}^k.$$

The vector $\bar{\boldsymbol{y}}_k$ satisfies the linear system of equations obtained by removing the last row and column of the matrix and the last entry of the right-hand side in (87), i.e.,

$$\bar{R}_k^{\mathrm{T}} \bar{\boldsymbol{y}}_k = ||\boldsymbol{b}||\tilde{\boldsymbol{e}}_1 - ||\boldsymbol{b}||\bar{\boldsymbol{q}}_k \boldsymbol{q}_{k+1}^{(k+1)\mathrm{T}} \tilde{\boldsymbol{e}}_1,$$

where $\bar{\boldsymbol{q}}_k \in \mathbb{R}^k$ consists of the first k entries of $\boldsymbol{q}_{k+1}^{(k+1)}$. Thus,

$$\bar{\boldsymbol{y}}_k = \boldsymbol{z}_k + \bar{\boldsymbol{z}}_k, \tag{89}$$

where $\boldsymbol{z}_k$ solves (50) and $\bar{\boldsymbol{z}}_k$ satisfies

$$\bar{R}_k^{\mathrm{T}} \bar{\boldsymbol{z}}_k = -||\boldsymbol{b}||\bar{\boldsymbol{q}}_k \boldsymbol{q}_{k+1}^{(k+1)\mathrm{T}} \tilde{\boldsymbol{e}}_1. \tag{90}$$

A recursion formula for the vector $\boldsymbol{q}_{k+1}^{(k+1)}$ can be derived easily. It follows from representation (44) of the matrix Q_{k+1} that

$$\boldsymbol{q}_{k+1}^{(k+1)} = Q_{k+1}\tilde{\boldsymbol{e}}_{k+1} = \begin{bmatrix} Q_k & \boldsymbol{0} \\ \boldsymbol{0}^{\mathrm{T}} & 1 \end{bmatrix} G_{k+1}^{(k)\mathrm{T}} \tilde{\boldsymbol{e}}_{k+1} = \begin{bmatrix} -s_k \boldsymbol{q}_k^{(k)} \\ c_k \end{bmatrix}, \tag{91}$$

where $\boldsymbol{q}_k^{(k)}$ denotes the last column of Q_k. Repeated application of Eq. (91) for increasing values of k makes it possible to compute the vectors $\boldsymbol{q}_2^{(2)}, \boldsymbol{q}_3^{(3)}, \ldots, \boldsymbol{q}_{k+1}^{(k+1)}$ in about $k^2/2$ arithmetic floating-point operations.

The solutions of the linear systems (90) can be evaluated by a recursion formula based on (91) for increasing values of k as follows. Eq. (91) yields that

$$\boldsymbol{q}_{k+1}^{(k+1)\mathrm{T}} \tilde{\boldsymbol{e}}_1 = -s_k \boldsymbol{q}_k^{(k)\mathrm{T}} \tilde{\boldsymbol{e}}_1,$$

$$\bar{\boldsymbol{q}}_k = -s_k \boldsymbol{q}_k^{(k)} \tag{92}$$

and

$$\bar{\boldsymbol{q}}_{k+1} = -s_{k+1} \begin{bmatrix} \bar{\boldsymbol{q}}_k \\ c_k \end{bmatrix}, \tag{93}$$

where the vector $\bar{\boldsymbol{q}}_{k+1}$ consists of the $k+1$ first entries of $\boldsymbol{q}_{k+2}^{(k+2)}$, the last column of Q_{k+2}. Assume that the solution $\bar{\boldsymbol{z}}_k$ of (90) is available. We would like to compute the vector $\bar{\boldsymbol{z}}_{k+1} = [\bar{\zeta}_1^{(k+1)}, \bar{\zeta}_2^{(k+1)}, \ldots, \bar{\zeta}_{k+1}^{(k+1)}]^{\mathrm{T}}$ that satisfies

$$\bar{R}_{k+1}^{\mathrm{T}} \bar{\boldsymbol{z}}_{k+1} = -||\boldsymbol{b}||\bar{\boldsymbol{q}}_{k+1} \boldsymbol{q}_{k+2}^{(k+2)\mathrm{T}} \tilde{\boldsymbol{e}}_1. \tag{94}$$

Substituting (92) and (93) into (94) yields

$$\bar{R}_{k+1}^{\mathrm{T}} \bar{\boldsymbol{z}}_{k+1} = -||\boldsymbol{b}||s_{k+1}^2 \begin{bmatrix} \bar{\boldsymbol{q}}_k \\ c_k \end{bmatrix} \boldsymbol{q}_{k+1}^{(k+1)\mathrm{T}} \tilde{\boldsymbol{e}}_1,$$

which shows that

$$\bar{\boldsymbol{z}}_{k+1} = \begin{bmatrix} s_{k+1}^2 \bar{\boldsymbol{z}}_k \\ \bar{\zeta}_{k+1}^{(k+1)} \end{bmatrix},$$

$$\bar{\zeta}_{k+1}^{(k+1)} = -(||\boldsymbol{b}||s_{k+1}^2 c_k \boldsymbol{q}_{k+1}^{(k+1)\mathrm{T}} \tilde{\boldsymbol{e}}_1 + \bar{r}_{k-1,k+1}^{(k+1)} \bar{\zeta}_{k-1}^{(k+1)} + \bar{r}_{k,k+1}^{(k+1)} \bar{\zeta}_k^{(k+1)}) / \bar{r}_{k+1,k+1}^{(k+1)}.$$

Thus, assuming that the matrix $\bar{R}_{k+1}$ is available, all the vectors $\bar{\boldsymbol{z}}_1, \bar{\boldsymbol{z}}_2, \ldots, \bar{\boldsymbol{z}}_{k+1}$ can be computed in $\mathrm{O}(k^2)$ arithmetic floating-point operations.

Having computed the solutions of (50) and (90), the above development, and in particular Eqs. (88) and (89), show that we can evaluate the $(k+1)$-point Gauss–Radau rule (83) with integrand (82) according to

$$\hat{\mathscr{G}}_{k+1}(f) = ||\boldsymbol{z}_k + \bar{\boldsymbol{z}}_k||^2.$$

Substituting this approximation of $\boldsymbol{b}^{\mathrm{T}} A^{-2} \boldsymbol{b}$ into (35) yields

$$\begin{aligned} \boldsymbol{e}_k^{\mathrm{T}} \boldsymbol{e}_k &= |\boldsymbol{b}^{\mathrm{T}} A^{-2} \boldsymbol{b} - \boldsymbol{z}_{k-1}^{\mathrm{T}} \boldsymbol{z}_{k-1}| \\ &\approx |||\boldsymbol{z}_k + \bar{\boldsymbol{z}}_k||^2 - \boldsymbol{z}_{k-1}^{\mathrm{T}} \boldsymbol{z}_{k-1}| \\ &= |\bar{\boldsymbol{z}}_k^{\mathrm{T}} (2\boldsymbol{z}_k + \bar{\boldsymbol{z}}_k) + \zeta_k^2|, \end{aligned}$$

where the last equality follows from (51). This suggests the approximation

$$||\boldsymbol{e}_k|| \approx |\bar{\boldsymbol{z}}_k^{\mathrm{T}} (2\boldsymbol{z}_k + \bar{\boldsymbol{z}}_k) + \zeta_k^2|^{1/2}. \tag{95}$$

We note that the approximate solution $\boldsymbol{x}_k$ of (1) and the right-hand side of (95) can be evaluated after k Lanczos steps have been carried out and the last subdiagonal entry of the Gauss–Radau matrix (80) has been determined by (42). Computed examples in the following section indicate that approximation (95) typically gives accurate estimates of the norm of the error.

4. Computed examples

We describe four examples that illustrate the performance of the iterative method, the error bound and the error estimates. All computations were carried out on an XP1000 Alpha workstation in Matlab with about 16 significant digits. In all examples we chose the initial approximate solution $\boldsymbol{x}_0 = \boldsymbol{0}$ and terminated the iterations as soon as

$$||\boldsymbol{e}_k|| < \varepsilon \tag{96}$$

with $\varepsilon := 1 \cdot 10^{-10}$ or $1 \cdot 10^{-11}$. These values of ε are likely to be smaller than values of interest in many application. Our choices of ε demonstrates the possibility of computing accurate solutions and error estimates. In fact, the error bounds and estimates perform well also for values of ε smaller than $1 \cdot 10^{-11}$.

We determined the matrices in the linear systems in Examples 1–3 in the following fashion. Let

$$A := U_n \Lambda_n U_n^{\mathrm{T}}, \quad \Lambda_n = \mathrm{diag}[\lambda_1, \lambda_2, \ldots, \lambda_n], \quad U_n \in \mathbb{R}^{n \times n}, \quad U_n^{\mathrm{T}} U_n = I_n, \tag{97}$$

where the eigenvector matrix U_n either is the $n \times n$ identity matrix I_n or a random orthogonal matrix determined by orthogonalizing the columns of an $n \times n$ real matrix with random entries. The matrix

A is diagonal when $U_n = I_n$ and dense when U_n is a random orthogonal matrix. We remark that the matrices T_k and V_k in the Lanczos decomposition (11) depend on the choice of U_n. Moreover, propagated round-off errors, due to round-offs introduced during matrix–vector product evaluations with the matrix A, may depend on the matrix U_n.

Example 1. Let $n{:=}1000$ and assume that the diagonal entries of the matrix Λ_n in (97) are given by $\lambda_j = 5j$. We first let U_n be a random orthogonal matrix. Then the matrix A defined by (97) is symmetric positive definite and dense. The right-hand side vector $\boldsymbol{b}$ is chosen so that $\boldsymbol{x} = \frac{1}{10}[1, 1, \ldots, 1]^{\mathrm{T}}$ solves (1). We terminate the iterations as soon as (96) is satisfied with $\varepsilon = 1 \cdot 10^{-11}$.

Fig. 1 (a) shows the 10-logarithm of $\|\boldsymbol{e}_k\|$ (solid curve), the 10-logarithm of the lower bound of $\|\boldsymbol{e}_k\|$ computed by Gauss quadrature (62) (dash–dotted curve), and the 10-logarithm of the upper estimate of $\|\boldsymbol{e}_k\|$ computed by anti-Gauss quadrature (77) (dashed curve) as functions of the number of iterations k. After the first 50 iterations, the computed lower bounds and upper estimates can be seen to be quite close to the norm of the error in the computed approximate solutions.

The closeness between the lower bound (62), upper estimate (77), and the norm of the error of the computed approximate solutions is also illustrated in Figs. 1(b) and (c). The former figure displays $(\tilde{\zeta}_k^2 - \zeta_k^2)^{1/2} - \|\boldsymbol{e}_k\|$ (solid curve) and $((\breve{\zeta}_k^{(k)})^2 + (\breve{\zeta}_{k-1}^{(k)})^2 - \zeta_{k-1}^2)^{1/2} - \|\boldsymbol{e}_k\|$ (dash–dotted curve) as functions of k. These quantities are seen to converge to zero as k increases. To shed some light on the rate of convergence, Fig. 1(c) shows the relative differences $((\tilde{\zeta}_k^2 - \zeta_k^2)^{1/2} - \|\boldsymbol{e}_k\|)/\|\boldsymbol{e}_k\|$ and $(((\breve{\zeta}_k^{(k)})^2 + (\breve{\zeta}_{k-1}^{(k)})^2 - \zeta_{k-1}^2)^{1/2} - \|\boldsymbol{e}_k\|)/\|\boldsymbol{e}_k\|$, both of which converge to zero as k increases.

Fig. 1(a) also shows the 10-logarithm of the norm of the residual error (3) as a function of k (dotted curve). The norm of the residual error is about a factor $1 \cdot 10^3$ larger than the norm of the error in the corresponding approximate solution. If we would like to stop the iterations when the error in the computed approximate solution is below a certain tolerance, then we can terminate the computations much sooner if we base the stopping criterion on formulas (62) and (77) than on the norm of the residual error.

We now replace the random orthogonal matrix U_n in definition (97) of the matrix A by I_n. The matrix A obtained is diagonal and has the same spectrum as the matrix used for the computations shown in Fig. 1. The right-hand side vector $\boldsymbol{b}$ is chosen so that $\boldsymbol{x} = \frac{1}{10}[1, 1, \ldots, 1]^{\mathrm{T}}$ solves (1). The performance of the iterative method applied to this linear system is displayed by Fig. 2, which is analogous to Fig. 1.

Figs. 1 and 2 show the Gauss and anti-Gauss rules to give good lower bounds and upper estimates of the norm of the error in the computed approximate solutions, with the lower bounds and upper estimates being closer to the norm of the error when $U_n = I_n$ than when U_n was chosen to be a random orthogonal matrix. This example illustrates that the quality of the computed error bounds and estimates may depend on the eigenvector matrix of A.

Example 2. Let the matrix $A \in \mathbb{R}^{48\times 48}$ in the linear system (1) be of the form (97) with U_{48} a random orthogonal matrix and Λ_{48} defined by

$$\lambda_i{:=}c + \frac{i-1}{47}(d-c)\rho^{48-i}, \quad i = 1, 2, \ldots, 48.$$

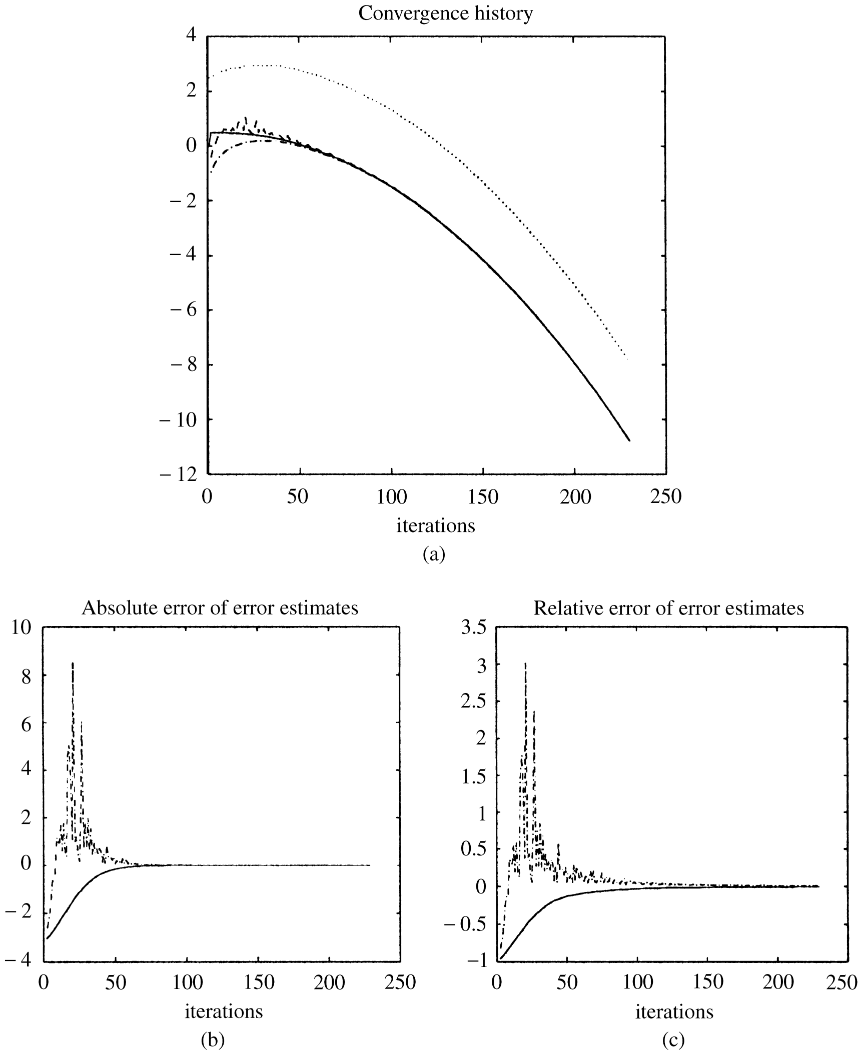


Fig. 1. Example 1: Symmetric positive-definite dense matrix. (a) shows the 10-logarithm of the norm of the error (solid curve), of the Gauss bound (62) (dash–dotted curve), of the anti-Gauss upper estimate (77) (dashed curve) and of the norm of the residual error (dotted curve). (b) displays the error in the Gauss bound (solid curve) and anti-Gauss upper estimate (dash–dotted curve). (c) shows the relative error in the Gauss bound (solid curve) and anti-Gauss upper estimate (dash–dotted curve).

Here $c{:=}0.1$, $d{:=}100$ and $\rho{:=}0.875$. Thus, A is symmetric, positive definite and dense. The right-hand side vector $\boldsymbol{b}$ is chosen so that $\boldsymbol{x} = [1, 1, \ldots, 1]^{\mathrm{T}}$ solves the linear system (1). We terminate the iterations as soon as (96) is satisfied with $\varepsilon = 1 \cdot 10^{-10}$.

Fig. 3 is analogous to Fig. 1 and shows the performance of the iterative method, of the lower error bound (62) and of the upper error estimate (77). The error bound (62) and error estimate (77) are seen to be close to the norm of the error in the computed approximate solutions. The “spikes” in Figs. 3(b) and (c) correspond to anti-Gauss rules associated with ill-conditioned tridiagonal matrices

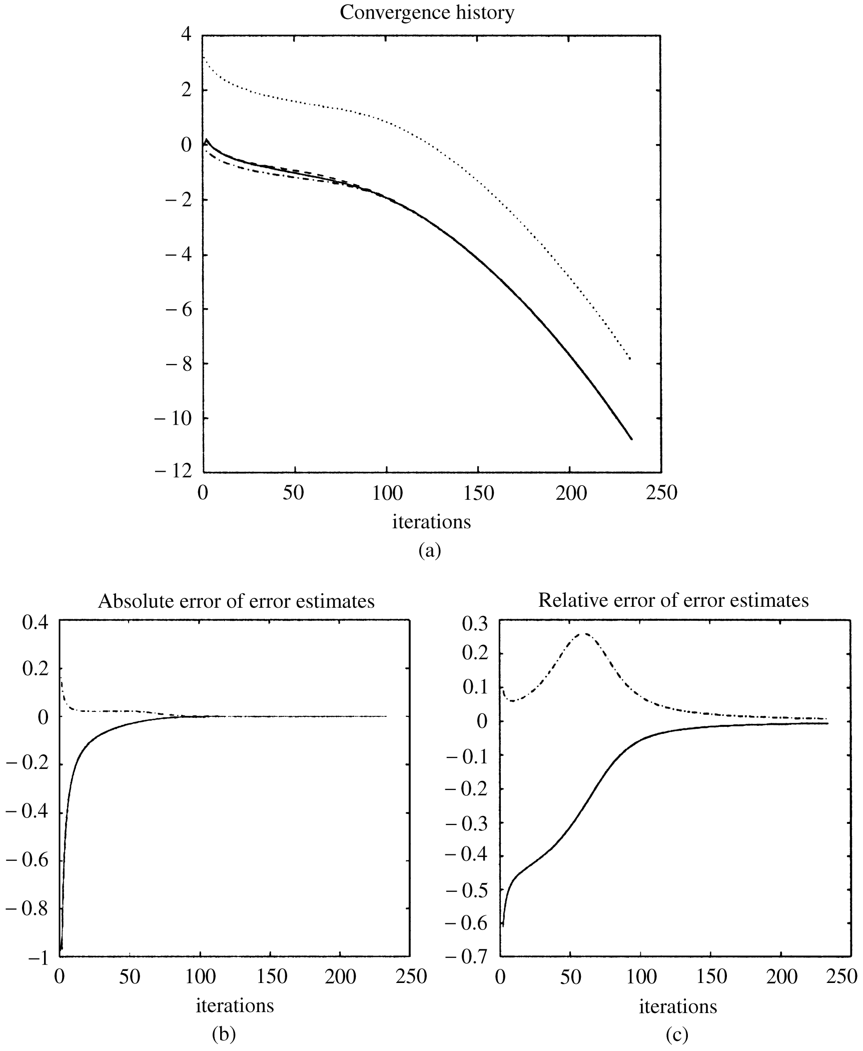


Fig. 2. Example 1: Symmetric positive-definite dense matrix. (a) shows the 10-logarithm of the norm of the error (solid curve), of the Gauss bound (62) (dash–dotted curve), of the anti-Gauss upper estimate (77) (dashed curve) and of the norm of the residual error (dotted curve). (b) displays the error in the Gauss bound (solid curve) and anti-Gauss upper estimate (dash–dotted curve). (c) shows the relative error in the Gauss bound (solid curve) and anti-Gauss upper estimate (dash–dotted curve).

(64). Ill-conditioning of the tridiagonal matrices (64) can cause loss of accuracy in the computed error estimates.

Now replace the random orthogonal matrix U_{48} in definition (97) of the matrix A by the identity matrix I_{48}. The matrix A so defined is diagonal and has the same spectrum as the matrix used for the computations shown in Fig. 3. The right-hand side vector $\boldsymbol{b}$ is chosen so that $\boldsymbol{x} = [1, 1, \ldots, 1]^{\mathrm{T}}$ solves (1). This linear system has previously been used in computed examples in [10,11,14] with a stopping criterion, based on the A-norm instead of the Euclidean norm, with $\varepsilon = 1 \cdot 10^{-10}$. We

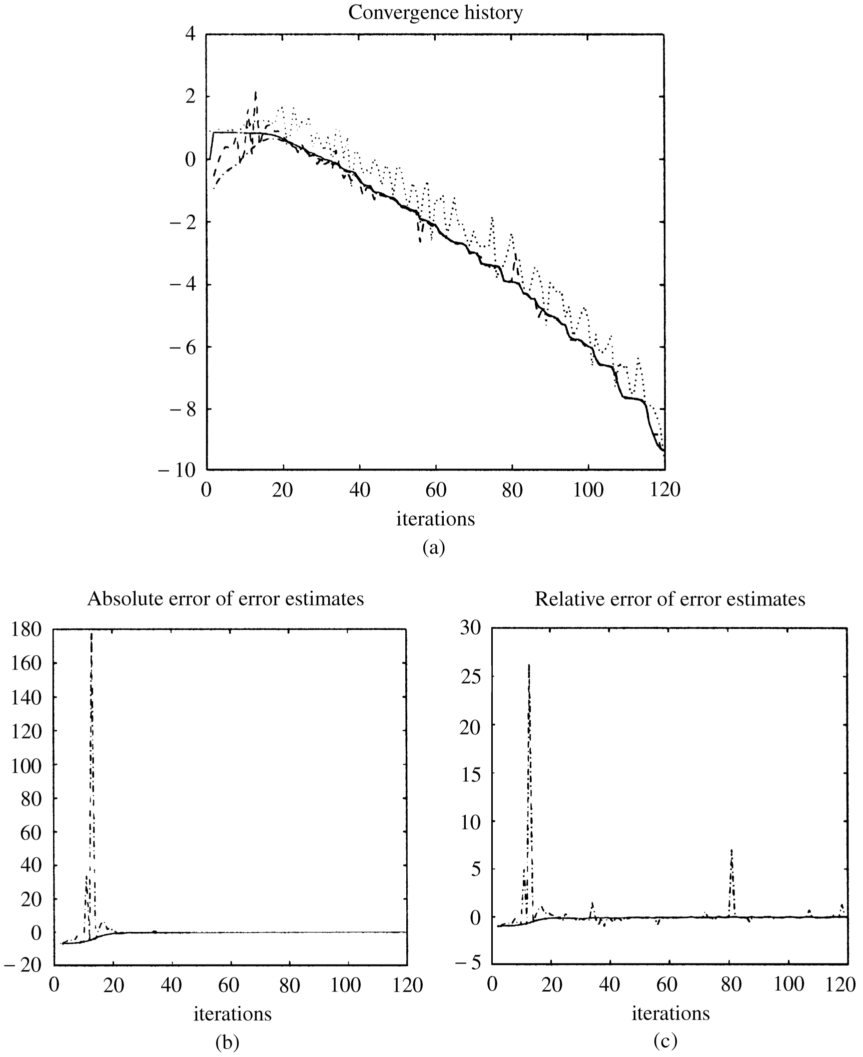
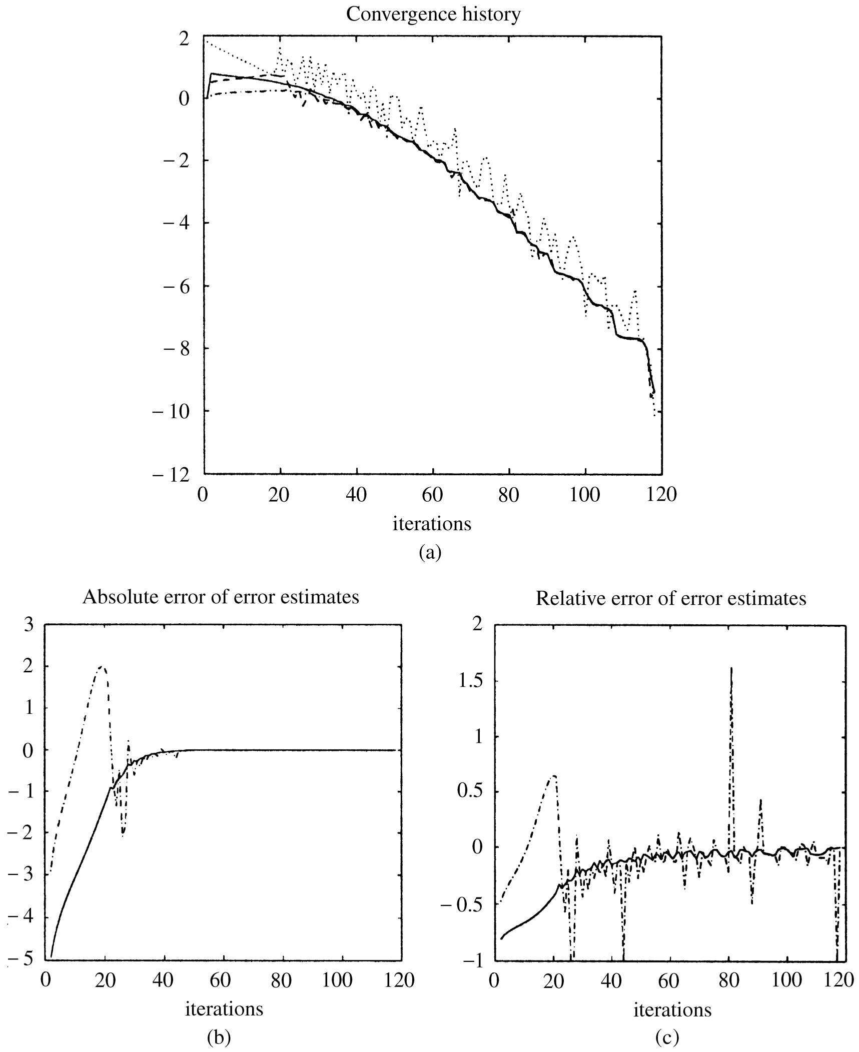


Fig. 3. Example 2: Symmetric positive-definite dense matrix. (a) shows the 10-logarithm of the norm of the error (solid curve), of the Gauss bound (62) (dash–dotted curve), of the anti-Gauss upper estimate (77) (dashed curve) and of the norm of the residual error (dotted curve). (b) displays the error in the Gauss bound (solid curve) and anti-Gauss upper estimate (dash–dotted curve). (c) shows the relative error in the Gauss bound (solid curve) and anti-Gauss upper estimate (dash–dotted curve).

therefore use the same value of ε in the present example. The performance of the iterative method, as well as of the error bounds and estimates, are shown in Fig. 4.

Figs. 3 and 4 display that the lower bounds and upper estimates of the norm of the error in the computed approximate solutions are closer to the norm of the error when $U_{48} = I_{48}$ than when U_{48} was chosen to be a random orthogonal matrix. Thus, similarly as in Example 1, the quality of the error bounds and estimates depends on the eigenvector matrix of A.



(a)

(b)

(c)

Fig. 4. Example 2: Symmetric positive-definite diagonal matrix. (a) shows the 10-logarithm of the norm of the error (solid curve), of the Gauss bound (62) (dash–dotted curve), of the anti-Gauss upper estimate (77) (dashed curve) and of the norm of the residual error (dotted curve). (b) displays the error in the Gauss bound (solid curve) and anti-Gauss upper estimate (dash–dotted curve). (c) shows the relative error in the Gauss bound (solid curve) and anti-Gauss upper estimate (dash–dotted curve).

The following two examples are concerned with linear systems of equations with symmetric indefinite matrices. For such matrices, the convex hull of the spectrum contains the origin, and some Gauss rules (52) may have a node in the interval between the largest negative and the smallest positive eigenvalues, where the matrix has no eigenvalues. The presence of a node close to the origin can give inaccurate estimates of the norm of the error in the computed approximate solution. This is illustrated by Figs. 5 and 6. This difficulty is circumvented by Gauss–Radau quadrature rules, cf. Proposition 3.1.

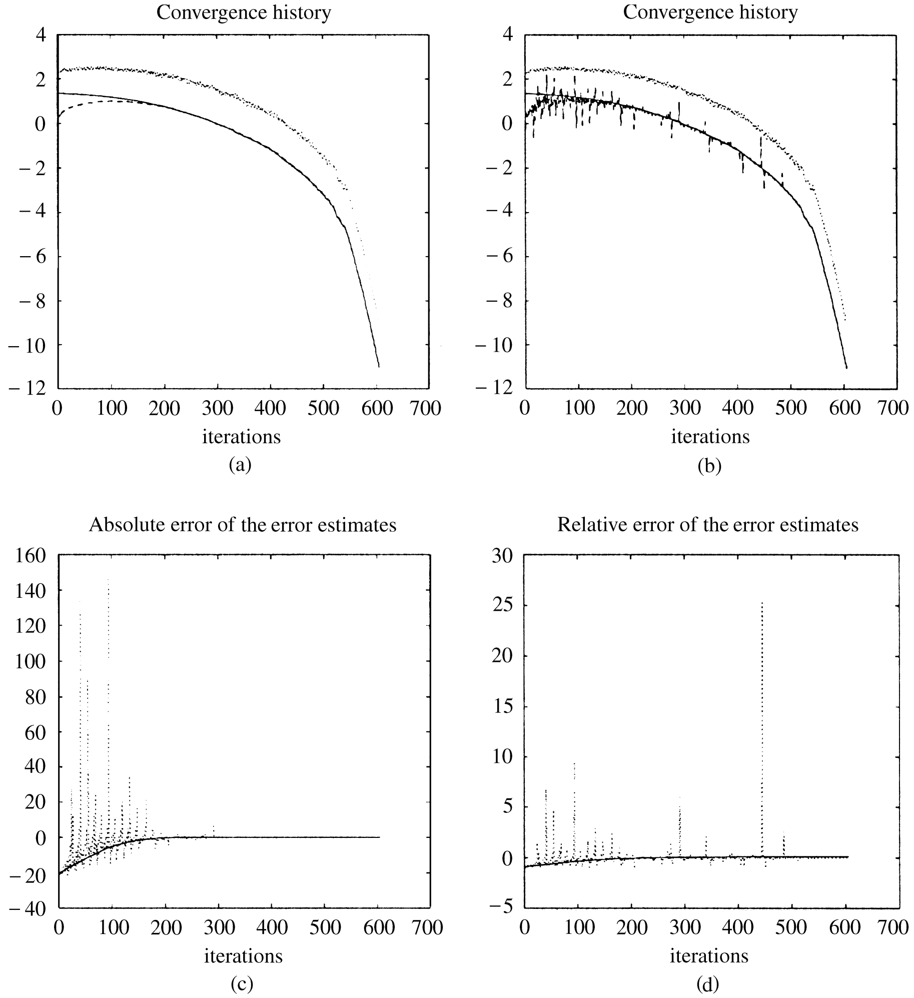


Fig. 5. Example 3: Symmetric indefinite dense matrix. (a) shows the 10-logarithm of the norm of the error (solid curve), of the Gauss–Radau estimate (95) (dashed curve) and of the norm of the residual error (dotted curve). (b) displays the 10-logarithm of the norm of the error (solid curve), of the Gauss estimate (62) (dashed curve) and of the norm of the residual error (dotted curve). (c) shows the error in the Gauss–Radau estimate (solid curve) and Gauss estimate (dotted curve). (d) displays the relative error in the Gauss–Radau estimate (solid curve) and Gauss estimate (dotted curve).

Example 3. Let the matrix A in (1) be of order 491 and of the form (97), where U_{491} is a random orthogonal matrix and the entries of the diagonal matrix Λ_{491} are given by

$$\lambda_i = \begin{cases} -150 + (i-1), & i = 1,\dots,141, \\ i - 141, & i = 142,\dots,491. \end{cases}$$

Then A is a dense matrix with eigenvalues in the interval $[-150, 350]$. We determine the right-hand side vector $\boldsymbol{b}$ so that $\boldsymbol{x} = [1, 1, \dots, 1]^{\mathrm{T}}$ solves the linear system (1). The iterations are terminated as soon as (96) is satisfied with $\varepsilon = 1 \cdot 10^{-11}$.

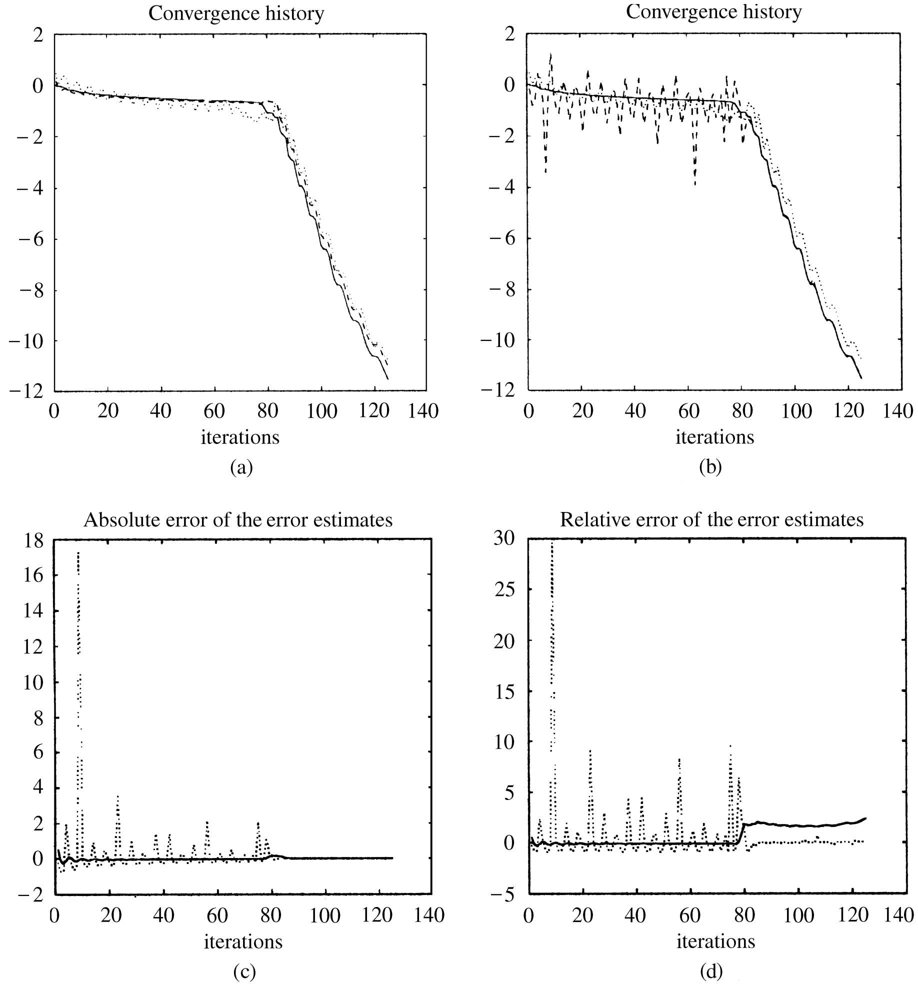


Fig. 6. Example 4: Symmetric indefinite banded matrix. (a) shows the 10-logarithm of the norm of the error (solid curve), of the Gauss–Radau estimate (95) (dashed curve) and of the norm of the residual error (dotted curve). (b) displays the 10-logarithm of the norm of the error (solid curve), of the Gauss estimate (62) (dashed curve) and of the norm of the residual error (dotted curve). (c) shows the error in the Gauss–Radau estimate (solid curve) and Gauss estimate (dotted curve). (d) displays the relative error in the Gauss–Radau estimate (solid curve) and Gauss estimate (dotted curve).

Fig. 5(a) shows the 10-logarithm of the error in the computed approximate solutions (solid curve), the 10-logarithm of the error estimate determined by Gauss–Radau quadrature (95) (dashed curve), and the 10-logarithm of the norm of the residual error (dotted curve). The error estimates computed by Gauss–Radau quadrature can be seen to be quite close to the norm of the error in the computed approximate solutions.

Fig. 5(b) is obtained from Fig. 5(a) by replacing the curve for the Gauss–Radau estimates (95) with a curve that displays error estimates computed by Gauss quadrature (62). Thus, the dashed curve of Fig. 5(b) displays the 10-logarithm of the right-hand side of (62). Note that since A is indefinite, formula (55) for the integration error does not reveal the sign of the error and inequality

(56) is not guaranteed to hold. The Gauss rules only give estimates of the norm of the error in the computed approximate solutions. The "spikes" of the dashed curve are caused by nodes of Gauss rules being very close to the origin.

Fig. 5(c) is analogous to Fig. 3(b). The solid curve displays the error in the Gauss–Radau estimates $|\bar{\boldsymbol{z}}_k^{\mathrm{T}}(2\boldsymbol{z}_k+\bar{\boldsymbol{z}}_k)+\zeta_k^2|^{1/2}-||\boldsymbol{e}_k||$, cf. (95), and the dashed curve shows the error in the Gauss estimates $(\tilde{\zeta}_k^2-\zeta_k^2)^{1/2}-||\boldsymbol{e}_k||$. Fig. 5(d) displays the corresponding relative errors, i.e., $(|\bar{\boldsymbol{z}}_k^{\mathrm{T}}(2\boldsymbol{z}_k+\bar{\boldsymbol{z}}_k)+\zeta_k^2|^{1/2}-||\boldsymbol{e}_k||)/||\boldsymbol{e}_k||$ (solid curve) and $((\tilde{\zeta}_k^2-\zeta_k^2)^{1/2}-||\boldsymbol{e}_k||)/||\boldsymbol{e}_k||$ (dashed curve). The Gauss–Radau estimates are seen to be more reliable than the Gauss estimates.

Example 4. Let $A\in\mathbb{R}^{200\times 200}$ be defined by $A:=B^2-\mu I_{200}$, where B is the standard 3-point discretization of the one-dimensional Laplacian and $\mu:=\sqrt{3}$. Thus, B^2 is a pentadiagonal matrix; a typical row has the nonvanishing entries $\{1,-4,6,-4,1\}$. Then A has 77 negative eigenvalues and condition number $3.9\cdot 10^3$. The right-hand side vector $\boldsymbol{b}$ is chosen so that $\boldsymbol{x}=[1,1,\ldots,1]^{\mathrm{T}}$ solves the linear system (1). We terminate the iterations as soon as the stopping criterion (96) is satisfied with $\varepsilon=1\cdot 10^{-11}$.

Figs. 6(a)–(d) are analogous to Figs. 5(a)–(d). The error estimates obtained by Gauss–Radau rules are quite accurate, while the estimates determined by Gauss rules oscillate widely during the first 77 iterations. After these initial iterations both Gauss–Radau and Gauss rules provide accurate error estimates.

5. Conclusion

This paper describes an iterative method for the solution of linear systems of equations with a symmetric nonsingular matrix. The iterative method is designed to allow the computation of bounds or estimates of the error in the computed approximate solutions. Computed examples show that the computed bounds and estimates are close to the norm of the actual errors in the computed approximate solutions.

Acknowledgements

We would like to thank Walter Gautschi for comments. This work was concluded while D.C. and L.R. were visiting the University of Bologna. They would like to thank Fiorella Sgallari for making this visit possible and enjoyable.

References

[1] C. Brezinski, Error estimates for the solution of linear systems, SIAM J. Sci. Comput. 21 (1999) 764–781.

[2] D. Calvetti, G.H. Golub, L. Reichel, A computable error bound for matrix functionals, J. Comput. Appl. Math. 103 (1999) 301–306.

[3] D. Calvetti, L. Reichel, An adaptive Richardson iteration method for indefinite linear systems, Numer. Algorithms 12 (1996) 125–149.

[4] D. Calvetti, L. Reichel, F. Sgallari, Application of anti-Gauss quadrature rules in linear algebra, in: W. Gautschi, G.H. Golub, G. Opfer (Eds.), Applications and Computation of Orthogonal Polynomials, Birkhäuser, Basel (1999) 41–56.

[5] D. Calvetti, L. Reichel, D.C. Sorensen, An implicitly restarted Lanczos method for large symmetric eigenvalue problems, Electr. Trans. Numer. Anal. 2 (1994) 1–21.

[6] G. Dahlquist, S.C. Eisenstat, G.H. Golub, Bounds for the error of linear systems of equations using the theory of moments, J. Math. Anal. Appl. 37 (1972) 151–166.

[7] G. Dahlquist, G.H. Golub, S.G. Nash, Bounds for the error in linear systems, in: R. Hettich (Ed.), Semi-Infinite Programming, Lecture Notes in Control and Computer Science, Vol. 15, Springer, Berlin (1979) 154–172.

[8] B. Fischer, Polynomial Based Iteration Methods for Symmetric Linear Systems, Teubner-Wiley, New York, 1996.

[9] G.H. Golub, G. Meurant, Matrices, moments and quadrature, in: D.F. Griffiths, G.A. Watson (Eds.), Numerical Analysis 1993, Longman, Essex, England (1994) 105–156.

[10] G.H. Golub, G. Meurant, Matrices, moments and quadrature II: how to compute the norm of the error in iterative methods, BIT 37 (1997) 687–705.

[11] G.H. Golub, Z. Strakos, Estimates in quadratic formulas, Numer. Algorithms 8 (1994) 241–268.

[12] G.H. Golub, C.F. Van Loan, Matrix Computations, 3rd edition, Johns Hopkins University Press, Baltimore, 1996.

[13] D.P. Laurie, Anti-Gaussian quadrature formulas, Math. Comp. 65 (1996) 739–747.

[14] G. Meurant, The computation of bounds for the norm of the error in the conjugate gradient algorithm, Numer. Algorithms 16 (1997) 77–87.

[15] G. Meurant, Numerical experiments in computing bounds for the norm of the error in the preconditioned conjugate gradient algorithm, Numer. Algorithms 22 (1999) 353–365.

[16] C.C. Paige, M.A. Saunders, Solution of sparse indefinite systems of linear equations, SIAM J. Numer. Anal. 12 (1975) 617–629.

[17] Y. Saad, Iterative Methods for Sparse Linear Systems, PWS, Boston, 1996.

[18] J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, 2nd edition, Springer, New York, 1993.

Journal of Computational and Applied Mathematics 127 (2001) 121–152

JOURNAL OF
COMPUTATIONAL AND
APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Cubature formulae and orthogonal polynomials

R. Cools[a,*], I.P. Mysovskikh[b], H.J. Schmid[c]

[a] *Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, B-3001 Heverlee, Belgium*
[b] *Mathematical and Mechanical Faculty, Saint-Petersburg State University, Stary Peterhof, Bibliotechnaya pl. 2,*
198904 Saint Petersburg, Russia
[c] *Math. Institut, Universität Erlangen-Nürnberg, Bismarckstr. 1 1/2, D-91054 Erlangen, Germany*

Received 8 September 1999; received in revised form 1 December 1999

Abstract

The connection between orthogonal polynomials and cubature formulae for the approximation of multivariate integrals has been studied for about 100 yr. The article J. Radon published about 50 yr ago (J. Radon, Zur mechanischen Kubatur, Monatsh. Math. 52 (1948) 286–300) has been very influential. In this text we describe some of the results that were obtained during the search for answers to questions raised by his article. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Cubature; Multivariate integrals; Orthogonal polynomials

1. Introduction

The connection between orthogonal polynomials and algebraic integration formulae in higher dimension was already studied about 100 yr ago (early papers are, e.g., [1,6]). The problem became widely noticed after the second edition of Krylov's book "On the approximate calculation of integrals" [43], published in 1967, wherein Mysovskikh introduced Radon's construction of a formula of degree 5 published in 1948 [72].

Though no final solution – similar to the one-dimensional case – has been found up to now, the work in this field has been tremendous. In the textbooks by Krylov [43], Stroud [90], Sobolev [84], Engels [20], Mysovskikh [66], Davis and Rabinowitz [17], Xu [94] and Sobolev and Vaskevich [85], and in the survey article of Cools [10], the growth of knowledge in the field is documented. In this text we will only try to describe some relevant results – following Radon's ideas – that have been found in the meantime.

[*] Corresponding author.
E-mail address: ronald.cools@cs.kuleuven.ac.be (R. Cools).

PII: S0377-0427(00)00495-7

The one-dimensional algebraic case, i.e., interpolatory quadrature formulae and their relation to orthogonal polynomials, is well known. In dimension 2 and beyond, things look worse: there are more questions than answers. Nevertheless, some progress has been made. Though several essential problems – important in applications – are still open, e.g., how minimal formulae of an arbitrary degree of exactness look like for the integral over the square with constant weight function, several results of some generality have been found. They make transparent why answers to important questions must be quite complex. We leave aside many particular results, in spite of their importance for applications. For these we refer to the surveys in [90,14,11].

2. Basic concepts and notations

We would have liked to preserve the flair of the old papers; however, we finally decided to use modern notations in order to achieve an easy and consistent way of presenting the results.

We denote by $\mathbb{N}$ the nonnegative integers. The monomials of degree m in n variables are written in the short notation

$$\boldsymbol{x}^{\boldsymbol{m}} = x_1^{m_1} x_2^{m_2} \cdots x_n^{m_n} \quad \text{with } |\boldsymbol{m}| = m,$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, $\boldsymbol{m} = (m_1, m_2, \ldots, m_n) \in \mathbb{N}^n$, and

$$|\boldsymbol{m}| = \sum_{i=1}^{n} m_i \text{ is the length of the multi-index } \boldsymbol{m}.$$

A polynomial $f(\boldsymbol{x}) = f(x_1, x_2, \ldots, x_n)$ of (total) degree m can be represented as

$$f(\boldsymbol{x}) = \sum_{s=0}^{m} \sum_{|\boldsymbol{k}|=s} c_{\boldsymbol{k}} \boldsymbol{x}^{\boldsymbol{k}}, \quad c_{\boldsymbol{k}} \in \mathbb{C},$$

and the summation in

$$g_s(\boldsymbol{x}) = \sum_{|\boldsymbol{k}|=s} c_{\boldsymbol{k}} \boldsymbol{x}^{\boldsymbol{k}}$$

is done over all multi-indices $\boldsymbol{k}$ of length s. The polynomial g_s is called a *homogeneous component* of degree s. Hence f is an element of the ring of polynomials with complex coefficients, which will be denoted by $\mathbb{C}[\boldsymbol{x}] = \mathbb{C}[x_1, x_2, \ldots, x_n]$. The degree of a polynomial f will be denoted by $\deg(f)$.

The number of linearly independent polynomials of degree $\leqslant m$ is

$$M(n,m) = \binom{m+n}{n},$$

the number of pairwise distinct monomials of degree m is $M(n-1, m)$. When the linearly independent monomials are needed as an ordered sequence, we will represent them by

$$\{\varphi_j(\boldsymbol{x})\}_{j=1}^{\infty},$$

where $j < k$ whenever $\deg(\varphi_j(\boldsymbol{x})) < \deg(\varphi_k(\boldsymbol{x}))$. Hence

$$\{\varphi_j(\boldsymbol{x})\}_{j=1}^{\mu}, \quad \mu = M(n,m),$$

contains all monomials of degree $\leqslant m$. Most of the results in the sequel will be stated in the ring of polynomials with real coefficients, $\mathbb{R}[x_1, x_2, \ldots, x_n] = \mathbb{R}[\boldsymbol{x}]$, which will be denoted by $\mathbb{P}^n$. The polynomials in $\mathbb{P}^n$ of degree $\leqslant m$ will be denoted by $\mathbb{P}^n_m$. The elements of $\mathbb{P}^n_m$ form a vector space over $\mathbb{R}$ with dimension $\dim \mathbb{P}^n_m = M(n,m)$.

We consider integrals of the type

$$\mathscr{I}^n[f] = \int_\Omega f(\boldsymbol{x})\omega(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \quad f \in \mathscr{C}(\Omega), \tag{1}$$

where $\Omega \subseteq \mathbb{R}^n$ is a region with inner points and the real weight function ω is chosen such that the moments

$$\mathscr{I}^n[\boldsymbol{x}^{\boldsymbol{m}}], \quad \boldsymbol{m} \in \mathbb{N}^n,$$

exist. In many applications $\omega(\boldsymbol{x})$ is nonnegative. Hence $\mathscr{I}^n$ is a positive linear functional. In most of the results presented, $\mathscr{I}^n$ will be even strictly positive, i.e.,

$$\mathscr{I}^n[f] > 0 \text{ whenever } 0 \neq f \geqslant 0 \text{ on } \Omega,$$

so orthogonal polynomials with respect to $\mathscr{I}^n$ are defined.

The type of integrals we consider includes integrals over the so-called *standard regions*, for which we follow Stroud's notation [90].

C_n: the n-dimensional cube

$$\Omega = \{(x_1,\dots,x_n):\ -1 \leqslant x_i \leqslant 1, i = 1,\dots,n\}$$

with weight function $\omega(\boldsymbol{x}) = 1$,

S_n: the n-dimensional ball

$$\Omega = \left\{(x_1,\dots,x_n):\ \sum_{i=1}^n x_i^2 \leqslant 1\right\}$$

with weight function $\omega(\boldsymbol{x}) = 1$,

T_n: the n-dimensional simplex

$$\Omega = \left\{(x_1,\dots,x_n):\ \sum_{i=1}^n x_i \leqslant 1 \text{ and } x_i \geqslant 0, i = 1,\dots,n\right\}$$

with weight function $\omega(\boldsymbol{x}) = 1$,

$E_n^{r^2}$: the entire n-dimensional space $\Omega = \mathbb{R}^n$ with weight function

$$\omega(\boldsymbol{x}) = \mathrm{e}^{-r^2}, \quad r^2 = \sum_{i=1}^n x_i^2,$$

E_n^r: the entire n-dimensional space $\Omega = \mathbb{R}^n$ with weight function

$$\omega(\boldsymbol{x}) = \mathrm{e}^{-r},$$

H_2: the region bounded by the regular hexagon with vertices $(\pm 1, 0)$, $(\pm\frac{1}{2}, \pm\frac{1}{2}\sqrt{3})$ and weight function $\omega(\boldsymbol{x}) = 1$.

A *cubature formula* for (1) is of the form

$$\mathscr{I}^n[f] = \mathscr{Q}[f] + R[f], \tag{2}$$

where

$$\mathscr{Q}[f]=\sum_{j=1}^{N} w_j f(\boldsymbol{x}^{(j)}) \tag{3}$$

is called a *cubature sum*. The $\boldsymbol{x}^{(j)}$'s are called *nodes*, the w_j's *weights* or *coefficients*, and $R[f]$ is the error. The shorthand notation

$$\mathscr{I}^n[f]\cong\sum_{j=1}^{N} w_j f(\boldsymbol{x}^{(j)})$$

is often used.

A nonnegative integer d is called *degree of exactness* or *degree of precision* or simply *degree* of formula (2), if $R[f]=0$ for all polynomials f with $\deg(f)\leqslant d$ and if a polynomial g with $\deg(g)=d+1$ exists such that $R[g]\neq 0$.

Let $f\in\mathbb{C}[\boldsymbol{x}]$ be given and $\deg(f)=m$; the algebraic manifold of degree m generated by f will be denoted by

$$\mathscr{H}_m=\{\boldsymbol{x}\in\mathbb{C}^n\colon\ f(\boldsymbol{x})=0\}.$$

A cubature formula (2) with $N=M(n,m)$ which is exact for all polynomials of degree $\leqslant m$ is called *interpolatory* if the nodes do not lie on an algebraic manifold of degree m and the coefficients are uniquely determined by the nodes.

If $n=1$, then $N=m+1$ and the converse is true, too. If the degree of exactness of the quadrature formula is m, then it is interpolatory. For $n\geqslant 2$ this does not hold in general. The number of nodes might be lower than $M(n,m)$ since some of the coefficients may vanish.

Theorem 1. *Let* (2) *be given such that* $R[f]=0$ *for all polynomials of degree* $\leqslant m$ *and* $N\leqslant\mu=M(n,m)$. *The formula is interpolatory if and only if*

$$\operatorname{rank}([\varphi_1(\boldsymbol{x}^{(j)}),\varphi_2(\boldsymbol{x}^{(j)}),\ldots,\varphi_\mu(\boldsymbol{x}^{(j)})]_{j=1}^{N})=N.$$

We are specifying $\mathscr{Q}$ by $\mathscr{Q}(n,m,N)$ if we refer to a cubature sum in n dimensions with a degree of exactness m and N nodes. We only consider interpolatory cubature formulae. A noninterpolatory formula can be transformed to an interpolatory formula by deleting nodes. In particular, minimal formulae (N is minimal for fixed m) are interpolatory.

A polynomial P with $\deg(P)=m$ is called *orthogonal* with respect to the underlying $\mathscr{I}^n$ if $\mathscr{I}^n[PQ]=0$ for all Q, $\deg(Q)\leqslant m-1$. It is called *quasi-orthogonal* if $\mathscr{I}^n[PQ]=0$ for all $Q,\deg(Q)\leqslant m-2$.

A set of polynomials in $\mathbb{R}[x_1,\ldots,x_n]$ is called a *fundamental system of degree m* whenever it consists of $M(n-1,m)$ linearly independent polynomials of the form

$$\boldsymbol{x}^{\boldsymbol{m}}+Q_{\boldsymbol{m}},\quad \boldsymbol{m}\in\mathbb{N}^n,\quad \deg(Q_{\boldsymbol{m}})\leqslant|\boldsymbol{m}|-1.$$

A set $\mathscr{M}$ of polynomials is called a *fundamental set of degree m* if a fundamental system of degree m is contained in $\operatorname{span}\{\mathscr{M}\}$.

In the two-dimensional case we drop the superscript n and use x and y as variables, i.e., $\mathbb{P}, \mathbb{P}_m$, and,

$$\mathscr{I}[f] = \int_{\Omega} f(x,y)\omega(x,y)\,\mathrm{d}x\,\mathrm{d}y, \quad f \in \mathscr{C}(\Omega), \;\; \Omega \subseteq \mathbb{R}^2.$$

Whenever possible, one tries to find cubature sums (3) such that the following constraints are satisfied:

(i) $w_j > 0$,

(ii) $\boldsymbol{x}^{(j)} \in \Omega$,

(iii) N is minimal for fixed degree m.

If $n = 1$, Gaussian quadrature formulae satisfy all constraints. These formulae are closely connected to orthogonal polynomials. The zeros of a quasi-orthogonal polynomial of degree k, $P_k^1 + \gamma P_{k-1}^1$ with free parameter γ, are the nodes of a minimal quadrature formula of degree $2k-2$ with all weights positive. The parameter γ can be chosen such that all nodes are inside the domain of integration, and, if $\gamma = 0$ one obtains a uniquely determined formula of degree $2k-1$ satisfying (i), (ii), (iii). Nonminimal interpolatory quadrature formulae have been characterised by Sottas and Wanner, Peherstorfer, and many others, most recently by Xu [86,68,69,96].

3. Radon's formulae of degree 5

The paper by Johann Radon [72], which appeared in 1948, is not the oldest studying the application of orthogonal polynomials to cubature formulae (earlier papers are, e.g., [1] to which Radon refers, and [6]). However, Radon's contribution became fundamental for all research in that field. Although the word "cubature" appeared in the written English language already in the 17th century, this paper is probably the first that used the term "Kubaturformel" (i.e., German for "cubature formula") for a weighted sum of function values to approximate a multiple integral (in contrast to quadrature formula to approximate one-dimensional integrals). As an introduction to the survey which follows, we will briefly sketch its main ideas.

Radon discusses the construction of cubature formulae of degree 5 with 7 nodes for integrals over T_2, C_2, S_2. We are sure Radon knew the estimate (22) and knew that this bound will not be attained for classical (standard) regions in the case of degree 5. In order to construct a cubature formula of degree m he counted the number of monomials of degree $\leqslant m$ and used this divided by 3 as number of necessary nodes. He was aware that for degrees 2, 3 and 4 this will not lead to a solution and thus degree 5 is the first nontrivial case he could consider.

Assuming a formula of degree 5 with 7 nodes for an integral $\mathscr{I}$, a geometric consideration leads to polynomials of degree 3 vanishing at the nodes. These polynomials have to be orthogonal with respect to $\mathscr{I}$ to all polynomials of degree 2, and exactly three such polynomials, P_1, P_2, P_3, can vanish at the nodes. In the next step further necessary conditions are derived for the P_i. They must satisfy

$$\sum_{i=1}^{3} L_i P_i = 0 \tag{4}$$

for some linear polynomials $L_i \neq 0$. This can be reduced to

$$xK_1 + yK_2 = K_3, \tag{5}$$

where K_i are orthogonal polynomials of degree 3. Assuming the orthogonal basis of degree 3 in a form

$$P_i^3 = x^{3-i}y^i + Q_i, \quad Q_i \in \mathbb{P}_2, \ i = 0,1,2,3,$$

one obtains

$$K_1 = \alpha P_1^3 + \beta P_2^3 + \gamma P_3^3 \quad \text{and} \quad K_2 = -\alpha P_0^3 - \beta P_1^3 - \gamma P_2^3.$$

The parameters α, β, γ have to be chosen such that K_3 is also an orthogonal polynomial of degree 3. Thus, by setting

$$A = \mathscr{I}[P_0^3 P_2^3 - P_1^3 P_1^3], \quad B = \mathscr{I}[P_0^3 P_3^3 - P_1^3 P_2^3], \quad C = \mathscr{I}[P_1^3 P_3^3 - P_2^3 P_2^3], \tag{6}$$

one obtains the linear system

$$\begin{bmatrix} 0 & A & B \\ -A & 0 & C \\ -B & -C & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = 0. \tag{7}$$

Two cases may occur. The parameters α, β, γ can be determined up to a common factor, if

$$A^2 + B^2 + C^2 > 0, \tag{8}$$

otherwise they can be chosen arbitrarily. Radon did not further pursue the last case. He just remarked that he did not succeed in proving that this case never occurs.

In case (8) the polynomials K_1, K_2, K_3 can be computed. If they are linearly independent, the desired equation (4) is given by $xK_1 + yK_2 = K_3$. If these polynomials vanish at 7 pairwise distinct nodes, the degree of exactness follows from the orthogonality property of the K_i's.

If the K_i are linearly dependent, it follows that

$$K_1 = yQ \quad \text{and} \quad K_2 = -xQ$$

for some $Q \in \mathbb{P}_2$. In this case it can be shown that there is a K_3 such that all K_i vanish at 7 pairwise distinct nodes. This construction again is based on geometric considerations and finally allows the conclusion that such K_3's can be computed.

Radon's article continues with the construction of formulae of degree 5 with 7 nodes for integrals over the standard regions with constant weight function T_2, C_2 and S_2. The amount of computational work – in a pre-computer time – is tremendous. The article finishes with an examination of the cubature error.

Though the results are limited to a special case, Radon's approach is the basis for fundamental questions that were studied in the years following the publication of his result:

(i) Can this constructive method be generalised to a higher degree of exactness?
(ii) Can this constructive method be generalised to more than two dimensions?
(iii) Are there integrals for which the second case occurs, i.e., $A = B = C = 0$?
(iv) Is 7 a lower bound for the number of nodes of cubature formulae of degree 5 if (8) holds?
(v) Are there lower bounds of some generality for the number of nodes?
(vi) What intrinsic tools were applied for the solution?

Fifty years after the publication of Radon's paper, it is still not possible to answer these questions completely. We will outline what is known in the sequel.

4. Multivariate orthogonal polynomials

We assume that $\mathscr{I}^n$ is strictly positive. Different methods of generating an orthogonal polynomial basis were discussed by Hirsch [34]. If the moments $\mathscr{I}^n[\boldsymbol{x}^{\boldsymbol{m}}]$, $\boldsymbol{m} \in \mathbb{N}^n$, are known, one can either orthogonalise the monomial basis by the Gram–Schmidt procedure, or compute step by step fundamental systems of orthogonal polynomials of degree m.

A third way is to find a partial differential equation with boundary conditions the polynomial solutions of which lead to orthogonal systems. This permits finding formulae for the coefficients of the polynomials and deriving recursion formulae. However, one has to find out if there is an integral for which the polynomials form an orthogonal system.

4.1. Simple properties

The strictly positive integral (1) defines a scalar product in $\mathbb{C}[\boldsymbol{x}]$ by

$$(\phi,\psi) = \mathscr{I}^n[\phi,\bar{\psi}] = \int_\Omega \phi(\boldsymbol{x})\overline{\psi(\boldsymbol{x})}\omega(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \quad \phi,\psi \in \mathbb{C}[\boldsymbol{x}]. \tag{9}$$

Consider the polynomial

$$P_{k+1}(\boldsymbol{x}) = g_{k+1}(\boldsymbol{x}) + \sum_{i=1}^{\kappa} a_i\varphi_i(\boldsymbol{x}), \quad \kappa = M(n,k), \tag{10}$$

where $g_{k+1}(\boldsymbol{x})$ is a given homogeneous component of degree $k+1$ and the a_i's are unknown coefficients. Assuming (10) to be orthogonal to $\varphi_j(\boldsymbol{x})$, $j = 1,2,\ldots,\kappa$, with respect to (9), we obtain

$$\sum_{i=1}^{\kappa} a_i(\varphi_i,\varphi_j) = -(g_{k+1}(\boldsymbol{x}),\varphi_j), \quad j = 1,2,\ldots,\kappa. \tag{11}$$

The matrix of this system is the Gram matrix of the linearly independent polynomials $\varphi_1(\boldsymbol{x})$, $\varphi_2(\boldsymbol{x}),\ldots,\varphi_\kappa(\boldsymbol{x})$. Hence, the a_i are uniquely determined. The polynomial (10) is uniquely determined by its homogeneous component of degree $k+1$ and by orthogonality to all polynomials of degree $\leqslant k$.

We state some simple properties of orthogonal polynomials, which will be of use later.

Theorem 2. *The following properties hold for an orthogonal polynomial* P_{k+1}.

(1) *If the homogeneous component of degree* $k+1$ *has real coefficients, then all coefficients are real. This follows from* (11).

(2) *A real polynomial* P_{k+1} *changes sign in* Ω. *In particular*,

$$\{\boldsymbol{x} \in \Omega\colon P_{k+1} > 0\} \quad \textit{and} \quad \{\boldsymbol{x} \in \Omega\colon P_{k+1} < 0\}$$

are of positive measure, which follows from

$$\int_\Omega P_{k+1}(\boldsymbol{x})\omega(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0.$$

(3) *Whenever* $P_{k+1} = UV$ *with polynomials* U *and* V *of degree at least* 1, *then* U *is orthogonal with respect to* Ω *and the weight function* $\omega(\boldsymbol{x})|V(\boldsymbol{x})|^2$. *This implies Properties* (1) *and* (2) *for the factors as well.*

(4) *If the coefficients belonging to the highest-degree terms in the homogeneous component of a factor are real, then the remaining coefficients are real, too.*

(5) *A real factor* U *of an orthogonal polynomial changes sign in* Ω. *In particular,*

$$\{\boldsymbol{x} \in \Omega\colon U > 0\} \quad \textit{and} \quad \{\boldsymbol{x} \in \Omega\colon U < 0\}$$

are of positive measure. From this we obtain

(6) *An orthogonal polynomial has no real multiple factors.*

We normalise the orthogonal polynomials of degree k to

$$P_{\boldsymbol{k}} = \boldsymbol{x}^{\boldsymbol{k}} + Q_{\boldsymbol{k}}, \quad \boldsymbol{k} \in \mathbb{N}^n, \ |\boldsymbol{k}| = k, \ \deg(Q_{\boldsymbol{k}}) \leqslant k - 1.$$

This fundamental system of degree k will be gathered in a polynomial vector of dimension $M(n-1,k)$ and be written as $\boldsymbol{P}_k$. We refer to the common zeros of all $P_{\boldsymbol{k}}$ as zeros of $\boldsymbol{P}_k$. The known explicit expressions for these normalised orthogonal polynomials are collected in [10].

4.2. Recursion formulae

For $n = 2$ the following results were found. Jackson [37] discusses a three-term recursion formula for a given orthogonal system, Gröbner [27] generates orthogonal systems by solving a variational problem under constraints; Krall and Sheffer [42] study in a class of second-order differential equations special cases the polynomial solutions of which generate classical orthogonal systems. Since their approach is closely related to recursion formulae and leads to concrete results we will outline the main ideas.

Let

$$P_j^k = x^{k-j} y^j + Q_j, \quad Q_j \in \mathbb{P}_{k-1}, \ j = 0, 1, \ldots, k, \ k \in \mathbb{N},$$

be a basis of $\mathbb{P}$. We can collect these fundamental systems of degree k in vectors

$$\boldsymbol{P}_k = (P_0^k, P_1^k, \ldots, P_k^k)^{\mathrm{T}}.$$

The basis $\boldsymbol{P}_k$, $k \in \mathbb{N}$, is said to be a *weak orthogonal system* if there exist matrices

$$C_k, \bar{C}_k \in \mathbb{R}^{k+1 \times k+1} \quad \text{and} \quad D_k, \bar{D}_k \in \mathbb{R}^{k+1 \times k},$$

such that

$$\begin{aligned} x\boldsymbol{P}_k &= L_{k+1}\boldsymbol{P}_{k+1} + C_k\boldsymbol{P}_k + D_k\boldsymbol{P}_{k-1}, \\ y\boldsymbol{P}_k &= F_{k+1}\boldsymbol{P}_{k+1} + \bar{C}_k\boldsymbol{P}_k + \bar{D}_k\boldsymbol{P}_{k-1}, \end{aligned} \tag{12}$$

with shift matrices L_{k+1} and F_{k+1} defined by $[E_k \ 0]$ and $[0 \ E_k]$, where E_k is the identity in $\mathbb{R}^{k+1 \times k+1}$ and $\boldsymbol{P}_{-1} = 0$.

A polynomial basis is said to be orthogonal with respect to a linear functional $\mathscr{L} : \mathbb{P} \to \mathbb{R}$, if, for each $k \in \mathbb{N}$, $\mathscr{L}[\boldsymbol{P}_k\boldsymbol{P}_l^{\mathrm{T}}] = 0$, $l = 0, 1, \ldots, k-1$, and if $\operatorname{rank}(\mathscr{L}[\boldsymbol{P}_k\boldsymbol{P}_k^{\mathrm{T}}]) = k+1$. Here, $\boldsymbol{P}_k\boldsymbol{P}_l^{\mathrm{T}}$ is the tensor product of the vectors $\boldsymbol{P}_k$ and $\boldsymbol{P}_l$, and $\mathscr{L}[\boldsymbol{P}_k\boldsymbol{P}_l^{\mathrm{T}}]$ is the matrix whose elements are determined by the functional acting on the polynomial coefficients of the tensor product. The matrix $M_k = \mathscr{L}[\boldsymbol{P}_k\boldsymbol{P}_k^{\mathrm{T}}]$ is known as the kth moment matrix. The basis $\{\boldsymbol{P}_k\}_{k\in\mathbb{N}}$ is said to be a (*positive*) definite orthogonal system in case the matrices M_k, $k \in \mathbb{N}$, are (positive) definite.

A definite system $\{\boldsymbol{P}_k\}_{k\in\mathbb{N}}$ is a weak orthogonal system, i.e., it satisfies the recurrence relations (12). Conversely, it follows from work by Xu [93] that a weak orthogonal system is an orthogonal system with respect to some $\mathscr{L}$ if and only if

$$\operatorname{rank}(S_k) = k+1 \quad \text{where } S_k = [D_k \quad \bar{D}_k] \in \mathbb{R}^{k+1\times 2k}.$$

The associated moment problem consists in assigning a measure to the functional $\mathscr{L}$ defined by a definite system. In particular, assigning a positive measure in case the system is positive definite (Favard's theorem) is quite complicated. We refer to Fuglede [24] and Xu [95].

Krall and Sheffer [42] studied the orthogonal polynomial systems which are generated by the following second-order differential equation:

$$\mathscr{D}\omega = -\lambda_k\omega, \quad \lambda_k \in \mathbb{R},\ k \in \mathbb{N}, \tag{13}$$

where

$$\begin{aligned}\mathscr{D}\omega = {}&(ax^2 + d_1x + e_1y + f_1)\frac{\partial^2\omega}{\partial x^2} + (2axy + d_2x + e_2y + f_2)\frac{\partial^2\omega}{\partial x\partial y}\\ &+ (ay^2 + d_3x + e_3y + f_3)\frac{\partial^2\omega}{\partial y^2} + (gx + h_1)\frac{\partial\omega}{\partial x} + (gy + h_2)\frac{\partial\omega}{\partial y}\end{aligned}$$

for some real constants $a \neq 0, g, d_i, e_i, f_i, h_i$, and for

$$\lambda_k = -k((k-1)a + g), \quad g + ka \neq 0,\ k \in \mathbb{N}.$$

They determined all weak orthogonal systems which are generated from (13) and proved that they are definite or positive definite, finding the classical orthogonal systems which had been derived in [2] and some new definite systems.

In [4] the recursion formulae for all positive-definite systems have been computed in the following way. Let $\{\boldsymbol{P}_k\}_{k=0,1,\ldots}$ be a definite orthogonal system with respect to $\mathscr{I}$. Multiplying (12) by $\boldsymbol{P}_{k-1}^{\mathrm{T}}$, $\boldsymbol{P}_k^{\mathrm{T}}$, and $\boldsymbol{P}_{k+1}^{\mathrm{T}}$, respectively, and applying $\mathscr{I}$, we obtain

$$\begin{aligned}&C_kM_k = \mathscr{I}[x\boldsymbol{P}_k\boldsymbol{P}_k^{\mathrm{T}}], \quad D_kM_{k-1} = \mathscr{I}[x\boldsymbol{P}_k\boldsymbol{P}_{k-1}^{\mathrm{T}}] = M_kL_k^{\mathrm{T}},\\ &\bar{C}_kM_k = \mathscr{I}[y\boldsymbol{P}_k\boldsymbol{P}_k^{\mathrm{T}}], \quad \bar{D}_kM_{k-1} = \mathscr{I}[y\boldsymbol{P}_k\boldsymbol{P}_{k-1}^{\mathrm{T}}] = M_kF_k^{\mathrm{T}}.\end{aligned}$$

By means of these identities the moment matrices can be computed by induction. Indeed, let $G_k = \operatorname{diag}\{[2, E_{k-2}]\}$ and $\bar{G}_k = \operatorname{diag}\{[E_{k-2}, 2]\}$; then

$$2\,E_k = L_k^{\mathrm{T}}G_kL_k + F_k^{\mathrm{T}}\bar{G}_kF_k,$$

and consequently,

$$2\,M_k = M_kL_k^{\mathrm{T}}G_kL_k + M_kF_k^{\mathrm{T}}\bar{G}_kF_k = D_kM_{k-1}G_kL_k + \bar{D}_kM_{k-1}\bar{G}_kF_k.$$

If one sets $M_0 = 1$, the last equation allows us to compute M_k from M_{k-1}, $k \in \mathbb{N}$. Based on [2], Verlinden [91] has computed explicit recursion formulae for classical two-dimensional integrals, too. So we refer to [91,4], if explicit recursion formulae are needed for standard integrals.

Not all two-dimensional orthogonal systems of interest can be obtained from (13). For further systems we refer to Koornwinder [38] and the references given there.

Kowalski in [39] presented a n-dimensional recursion formula and characterised it in [40,41]; Xu [93] refined this characterisation by dropping one condition. We will briefly outline these results. For a more complete insight into this development of a general theory of orthogonal polynomials in n dimensions we refer to the excellent survey by Xu [98].

Let $\mathscr{I}^n$ be given, and let

$$M_k = \mathscr{I}^n[\boldsymbol{P}_k\boldsymbol{P}_k^{\mathrm{T}}] \in \mathbb{R}^{M(n-1,k)\times M(n-1,k)}$$

be the moment matrix for $\mathscr{I}^n$. Then the recursion formula can be stated as follows.

Theorem 3. *For $k=0,1,\ldots$ there are matrices*

$$A_{k,i} \in \mathbb{R}^{M(n-1,k)\times M(n-1,k+1)}, \quad B_{k,i} \in \mathbb{R}^{M(n-1,k)\times M(n-1,k)},$$

and

$$C_{k,i} \in \mathbb{R}^{M(n-1,k)\times M(n-1,k-1)},$$

such that

$$x_i\boldsymbol{P}_k = A_{k,i}\boldsymbol{P}_{k+1} + B_{k,i}\boldsymbol{P}_k + C_{k,i}\boldsymbol{P}_{k-1}, \quad i=1,2,\ldots,n, \;\; k=0,1,\ldots,$$

where $\boldsymbol{P}_{-1}=0$ and for all $i=1,2,\ldots,n$ and all k

$$A_{k,i}M_{k+1} = \mathscr{I}^n[x_i\boldsymbol{P}_k\boldsymbol{P}_{k+1}^{\mathrm{T}}],$$

$$B_{k,i}M_k = \mathscr{I}^n[x_i\boldsymbol{P}_k\boldsymbol{P}_k^{\mathrm{T}}],$$

$$A_{k,i}M_{k+1} = M_kC_{k+1,i}^{\mathrm{T}}.$$

Furthermore, there are matrices

$$D_{k,i}, G_k \in \mathbb{R}^{M(n-1,k+1)\times M(n-1,k)}, \quad H_k \in \mathbb{R}^{M(n-1,k+1)\times M(n-1,k)}$$

such that

$$\boldsymbol{P}_{k+1} = \sum_{i=1}^{n} x_iD_{k,i}\boldsymbol{P}_k + G_k\boldsymbol{P}_k + H_k\boldsymbol{P}_{k-1},$$

where

$$\sum_{i=1}^{n} D_{k,i}A_{k,i} = E_{M(n-1,k+1)\times M(n-1,k+1)}$$

and

$$\sum_{i=1}^{n} D_{k,i}B_{k,i} = -G_k, \quad \sum_{i=1}^{n} D_{k,i}C_{k,i} = -H_k.$$

We will denote the fundamental set of orthonormal polynomials (with respect to $\mathscr{I}^n$) of degree k by $\boldsymbol{p}_k$. The recursion for orthonormal polynomials is given by Xu [95]. We reuse the notations $A_{k,i}$, $B_{k,i}$. In the following, these matrices will refer to the recursion for orthonormal matrices.

Theorem 4. *For $k = 0, 1, \ldots$ there are matrices*

$$A_{k,i} \in \mathbb{R}^{M(n-1,k)\times M(n-1,k+1)}, \quad B_{k,i} \in \mathbb{R}^{M(n-1,k)\times M(n-1,k)}$$

such that

$$x_i \boldsymbol{p}_k = A_{k,i}\boldsymbol{p}_{k+1} + B_{k,i}\boldsymbol{p}_k + A_{k-1,i}^{\mathrm{T}}\boldsymbol{p}_{k-1}, \quad i = 1, 2, \ldots, n, \; k = 0, 1, \ldots,$$

where $\boldsymbol{p}_{-1} = 0$, $A_{-1,i} = 0$ and

$$\operatorname{rank}(A_k) = \operatorname{rank}([A_{k,1}^{\mathrm{T}}|A_{k,2}^{\mathrm{T}}|\cdots|A_{k,n}^{\mathrm{T}}]^{\mathrm{T}}) = M(n-1, k+1).$$

For $i, j = 1, 2, \ldots, n$, $i \neq j$, and $k \geqslant 0$, the following matrix equations hold for the coefficient matrices:

(i) $A_{k,i}A_{k+1,j} = A_{k,j}A_{k+1,i}$,

(ii) $A_{k,i}B_{k+1,j} + B_{k,i}A_{k,j} = B_{k,j}A_{k,i} + A_{k,j}B_{k+1,i}$,

(iii) $A_{k-1,i}^{\mathrm{T}}A_{k-1,j} + B_{k,i}B_{k,j} + A_{k,i}A_{k,j}^{\mathrm{T}} = A_{k-1,j}^{\mathrm{T}}A_{k-1,i} + B_{k,j}B_{k,i} + A_{k,j}A_{k,i}^{\mathrm{T}}$.

In order to characterise Gaussian cubature formulae, see Section 7.1.4; the use of orthonormal systems gives more insight and often is easier to apply.

4.3. Common zeros

A direct analog of the Gaussian approach for $n \geqslant 2$ suggests considering the common zeros of all orthogonal polynomials of degree k as nodes of a formula of degree $2k - 1$. So the behaviour of common zeros of all orthogonal polynomials of degree k is of interest.

The following theorem, due to Mysovskikh [60,66], holds for (not necessarily orthogonal or real) fundamental systems of polynomials; it turned out to be essential.

Theorem 5. *Let*

$$R_{\boldsymbol{m}} = \boldsymbol{x}^{\boldsymbol{m}} + Q_{\boldsymbol{m}}, \quad \deg(Q_{\boldsymbol{m}}) \leqslant m - 1, \; |\boldsymbol{m}| = m,$$

be a fundamental system of degree m. Then the following is true.

(i) *The polynomials $R_{\boldsymbol{m}}$ have at most* $\dim \mathbb{P}_{m-1}^n$ *common zeros.*

(ii) *No polynomial of degree $m - 1$ vanishes at the common zeros of the $R_{\boldsymbol{m}}$, if and only if the $R_{\boldsymbol{m}}$ have exactly* $\dim \mathbb{P}_{m-1}^n$ *common pairwise distinct zeros.*

We will briefly derive the main properties of the zeros of fundamental systems of orthogonal polynomials. Orthonormalising the monomials $\{\varphi_j(\boldsymbol{x})\}_{j=1}^{\infty}$ with respect to $\mathscr{I}^n$, e.g., by the Gram–Schmidt procedure, we obtain

$$\{F_j(\boldsymbol{x})\}_{j=1}^{\infty} \quad \text{where } \mathscr{I}^n[F_iF_j] = \delta_{ij}.$$

The reproducing kernel in $\mathbb{P}_m^n$ is a polynomial in $2n$ variables,

$$K_m(\boldsymbol{u}, \boldsymbol{x}) = \sum_{j=1}^{\mu} F_j(\boldsymbol{u})F_j(\boldsymbol{x}), \quad \mu = M(n, m), \tag{14}$$

having the property

$$\mathscr{I}^n[K_m(\boldsymbol{u}, \boldsymbol{x})f(\boldsymbol{x})] = f(\boldsymbol{u}) \quad \text{for all } f \in \mathbb{P}_m^n. \tag{15}$$

Lemma 6. *For an $\boldsymbol{a} \in \mathbb{C}^n$ let l be a linear polynomial such that $l(\boldsymbol{a}) = 0$. Then $R = l(\boldsymbol{x})K_m(\boldsymbol{a}, \boldsymbol{x})$ is quasi-orthogonal. Whenever $\boldsymbol{a}$ is a common zero of $\boldsymbol{P}_{m+1}$, then R is orthogonal.*

Proof. If $Q \in \mathbb{P}^n_{m-1}$ we obtain by (15)

$$\mathscr{I}^n[l(\boldsymbol{x})K_m(\boldsymbol{a}, \boldsymbol{x})Q(\boldsymbol{x})] = l(\boldsymbol{a})Q(\boldsymbol{a}) = 0.$$

If $\boldsymbol{a}$ is a common zero of $\boldsymbol{P}_{m+1}$, then

$$F_{M(n-1,m)+i}(\boldsymbol{a}) = 0, \quad i = 1, 2, \ldots, M(n-1, m+1),$$

and thus $K_m(\boldsymbol{a}, \boldsymbol{x}) = K_{m+1}(\boldsymbol{a}, \boldsymbol{x})$. Hence

$$R(\boldsymbol{x}) = l(\boldsymbol{x})K_{m+1}(\boldsymbol{a}, \boldsymbol{x}) = l(\boldsymbol{x})K_m(\boldsymbol{a}, \boldsymbol{x})$$

is orthogonal to $\mathbb{P}^n_m$. Assuming $\deg(l(\boldsymbol{x})K_m(\boldsymbol{a}, \boldsymbol{x})) \leqslant m$, we obtain that R is zero, in contradiction to

$$K_m(\boldsymbol{a}, \bar{\boldsymbol{a}}) = \sum_{i=1}^{\mu} |F_j(\boldsymbol{a})^2| > 0, \quad \mu = M(n, m). \qquad \square$$

The following theorem was proved in [61,65].

Theorem 7. *The zeros of $\boldsymbol{P}_{m+1}$ are real and simple, and they belong to the interior of the convex hull of Ω. Furthermore, $\boldsymbol{P}_{m+1}$ and $\boldsymbol{P}_m$ have no zeros in common.*

Proof. Let $\boldsymbol{a} \in \mathbb{C}^n$ be a common zero of $\boldsymbol{P}_{m+1}$. By Lemma 6 the polynomials

$$(x_i - a_i)K_m(\boldsymbol{a}, \boldsymbol{x}), \quad i = 1, 2, \ldots, n, \tag{16}$$

are orthogonal to all polynomials of degree m. Because of property (3) in Theorem 2, the linear factor $x_i - a_i$ is real, hence $\boldsymbol{a} \in \mathbb{R}^n$. The Jacobian matrix of (16) in $\boldsymbol{a}$ is diagonal with elements $K_m(\boldsymbol{a}, \boldsymbol{a}) > 0$. This implies that $\boldsymbol{a}$ is simple. If $\boldsymbol{a}$ is not an interior point of the convex hull of Ω, there is a separating hyperplane $l(\boldsymbol{x})$ through $\boldsymbol{a}$, e.g., $l(\boldsymbol{x}) \geqslant 0$ for all $\boldsymbol{x}$ in the interior of the convex hull. Since $l(\boldsymbol{x})$ is a real factor of $l(\boldsymbol{x})K_m(\boldsymbol{a}, \boldsymbol{x})$, this is a contradiction to property (5) in Theorem 2. Finally, $\boldsymbol{a}$ is no common zero of $\boldsymbol{P}_m$ since by Lemma 6 the degree of (16) is $m + 1$, hence $\deg(K_m(\boldsymbol{a}, \boldsymbol{x})) = m$. $\square$

Using the matrices presented in Theorem 4, Xu [97] defines infinite tridiagonal block Jacobi matrices of the form

$$T_i = \begin{bmatrix} B_{0,i} & A_{0,i} & 0 & 0 & \ldots & 0 \\ A^{\mathrm{T}}_{0,i} & B_{1,i} & A_{1,i} & 0 & \ldots & 0 \\ 0 & A^{\mathrm{T}}_{1,i} & B_{1,i} & A_{2,i} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}, \quad i = 1, 2, \ldots, n,$$

and truncated versions of these. He found a relation between an eigenvalue problem for these matrices and the zeros of all orthogonal polynomials of a fixed degree. These results and their relation to cubature formulae are further elaborated in [94].

5. Lower bounds

5.1. Numerical characteristics

Mysovskikh [56] proved that Radon's formulae are minimal whenever (8) holds. In [67] an integral is constructed where the matrix in (7) is zero and a formula of degree 5 with 6 nodes can be constructed. Applying Radon's method to formulae of degree 3, Mysovskikh [59] found that such formulae with 4 nodes exist if and only if

$$\mathscr{I}[P_0^2P_2^2 - P_1^2P_1^2] \neq 0; \tag{17}$$

otherwise the formula has only 3 nodes. This has been further studied by Günther [28,29]. Fritsch [23] gave an example of an integral for which $\mathcal{Q}(n,3,n+1)$ exists. Cernicina [9] constructed a region in $\mathbb{R}^n$, $3 \leqslant n \leqslant 8$, admitting minimal formulae of type $\mathcal{Q}(n,4,(n+1)(n+2)/2)$; for $n=2$ a formula $\mathcal{Q}(2,5,6)$ is obtained. Stroud [90] extended (6) in the following way:

$$B = [\mathscr{I}[P_i^kP_j^k - P_\nu^kP_\mu^k]]_{i+j=\nu+\mu,\ \mu\neq i\neq\nu,\ 0\leqslant i,j,\mu,\nu\leqslant k}, \tag{18}$$

in order to obtain the lower bound in Theorem 10.

Mysovskikh [60] generalised (5) and (6) (in order to study the case which Radon did not further pursue) by defining for given k and $\mathscr{I}$ the following matrices:

$$M_{k-1}^{\star} = [\mathscr{I}[P_{j+1}^kP_i^k - P_j^kP_{i+1}^k]]_{i,j=0,1,\ldots,k-1}, \tag{19}$$

– note that $M_k^{\star}$ is skew-symmetric – and

$$A = \tfrac{1}{2}[\mathscr{I}[P_{i+1}^kP_{j-1}^k - 2P_i^kP_j^k - P_{i-1}^kP_{j+1}^k]]_{i,j=1,2,\ldots,k-1}.$$

The elements of these matrices characterise the behaviour of the orthogonal polynomials with respect to $\mathscr{I}$; so they were called *numerical characteristics*.

Theorem 8. *The following are equivalent*:

(i) *the matrix A vanishes*,

(ii) *the matrix $M_{k-1}^{\star}$ vanishes*,

(iii) *the orthogonal basis of degree k has $k(k+1)/2$ common pairwise distinct real zeros*,

(iv) *a cubature formula of degree $2k-1$ with the lowest possible number of nodes exists. Its nodes are the common zeros of the orthogonal basis of degree k.*

The proof in [66, p. 189], is based on Theorem 5 and the following considerations. The polynomials

$$Q_i = yP_i^k - xP_{i+1}^k, \quad i = 0,1,\ldots,k-1, \tag{20}$$

are of degree k; this is Radon's equation (5). Hence if the common zeros of all P_i^k are the nodes of a formula of degree $2k-1$, then the Q_i are orthogonal polynomials of degree k. This implies $M_{k-1}^{\star}=0$. Evidently, $M_{k-1}^{\star}=0$ implies $A=0$. On the other hand, if $A=0$, then $M_{k-1}^{\star}$ is of Hankel type, and, since $M_k^{\star}$ is skew-symmetric, this implies $M_{k-1}^{\star}=0$. The existence of integrals admitting the conditions of Theorem 8 was studied by Kuzmenkov in [44–46].

The articles based on Mysovskikh's results prefer to work with $M_{k-1}^{\star}$, and it turns out that this matrix in many ways characterises the behaviour of the associated orthogonal polynomials.

In order to generalise Theorem 8 to n dimensions, Eq. (20) has to be studied for all possible variables. By means of the recursion formulae for orthonormal systems, n-dimensional numerical characteristics can be defined. By condition (ii) in Theorem 4,

$$A_{k-1,j}x_i\boldsymbol{p}_k - A_{k-1,i}x_j\boldsymbol{p}_k$$

can be computed under the condition that these polynomials are orthogonal to $\mathbb{P}^n_{k-1}$. This leads to the matrices

$$M^{\star}_{k-1}(i,j) = A_{k-1,i}A^{\mathrm{T}}_{k-1,j} - A_{k-1,j}A^{\mathrm{T}}_{k-1,i}, \quad i,j = 1,2,\ldots,n, \ i \neq j, \tag{21}$$

which are representing the numerical characteristics of orthogonal polynomials in n dimensions.

5.2. Lower bounds

To settle the question of minimal formulae, lower bounds for the number of nodes are needed. The one-dimensional result can be generalised directly to find the following result, which seems to be folklore.

Theorem 9. *If $\mathscr{Q}(n,m,N)$ is a cubature sum for an integral $\mathscr{I}^n$, then*

$$N \geqslant \dim \mathbb{P}^n_{\lfloor m/2 \rfloor} = M(n, \lfloor m/2 \rfloor). \tag{22}$$

As we have seen in Section 3, this lower bound is not sharp for $n = 2$ and $m = 5$. A simple consequence of Theorem 8 is

Theorem 10. *If $\mathscr{Q}(2,2k-1,N)$ is a cubature sum for an integral $\mathscr{I}$ for which* $\operatorname{rank}(M^{\star}_{k-1}) > 0$, *then*

$$N \geqslant \dim \mathbb{P}_{k-1} + 1.$$

Stroud [90] showed this under the condition that B in (18) does not vanish.

Considerable progress was made by Möller [49], who improved the lower bound for $n = 2$.

Theorem 11. *If $\mathscr{Q}(2,2k-1,N)$ is a cubature sum for an integral $\mathscr{I}$, then*

$$N \geqslant \dim \mathbb{P}_{k-1} + \tfrac{1}{2}\operatorname{rank}(M^{\star}_{k-1}). \tag{23}$$

Proof. If a cubature sum $\mathscr{Q}(2,2k-1,N)$ is given, then no polynomial in $\mathbb{P}_{k-1}$ vanishes at all nodes. If no polynomial of degree k vanishes at all nodes, then $N \geqslant \dim \mathbb{P}_k > \dim \mathbb{P}_{k-1} + \frac{1}{2}\operatorname{rank}(M^{\star}_{k-1})$, since $M^{\star}_{k-1} \in \mathbb{R}^{k\times k}$. So Möller assumed the existence of s orthogonal polynomials Q_i of degree k which vanish at the nodes and first searched for a bound on s. Note that Q_i, xQ_i, yQ_i belong to an ideal which does not contain any polynomial of $\mathbb{P}_{k-1}$. Let

$$\mathscr{W} = \operatorname{span}\{Q_i, xQ_i, yQ_i, \ i = 1,2,\ldots,s\};$$

then

$$3s - \eta = \dim \mathscr{W} \leqslant k + 2 + s,$$

where η is the number of those xQ_i, yQ_i which can be dropped without diminishing the dimension of $\mathscr{W}$. These dependencies in $\mathscr{W}$ are of the form

$$x\sum_{i=1}^{k} a_i P_i^k - y\sum_{i=0}^{k-1} a_{i+1} P_i^k = \sum_{i=0}^{k} b_i P_i^k.$$

By orthogonality this leads to

$$\sum_{i=0}^{k-1} a_{i+1}\mathscr{I}[P_j^k P_{i+1}^k - P_{j+1}^k P_i^k] = 0, \quad j = 0,1,\ldots,k-1,$$

i.e.,

$$M_{k-1}^{\star}(a_1, a_2, \ldots, a_k)^{\mathrm{T}} = 0.$$

Thus we get $\eta \leqslant k - \operatorname{rank}(M_{k-1}^{\star})$, from which we finally obtain

$$s \leqslant k + 1 - \tfrac{1}{2}\operatorname{rank}(M_{k-1}^{\star}). \qquad \square$$

For all classes of integrals for which the rank of $M_{k-1}^{\star}$ has been computed, it turned out that either the rank is zero (cf. [82]) or

$$\operatorname{rank}(M_{k-1}^{\star}) = 2\lfloor k/2 \rfloor.$$

Classes of integrals for which the second rank condition holds have been already given by Möller [49]. He showed this for product integrals and integrals enjoying central symmetry. This includes the standard regions $C_2, S_2, H_2, E_2^{r^2}$ and E_2^r. Further classes with the same rank were detected by Rasputin [73], Berens and Schmid [3]. These include the standard region T_2.

Another important fact was observed by Möller. If (23) is attained, then the polynomials xQ_i, yQ_i form a fundamental set of degree $k+1$.

The improved lower bound, in general, is not sharp. Based on a characterisation of cubature sums $\mathscr{Q}(2, 4k+1, 2(k+1)^2 - 1)$ for circularly symmetric integrals in [92], it was shown in [16] that for all $k \in \mathbb{N} \setminus \{1\}$ the integrals

$$\int_{\mathbb{R}^2} f(x,y)(x^2+y^2)^{\alpha-1} \mathrm{e}^{-x^2-y^2}\,\mathrm{d}x\,\mathrm{d}y, \quad \alpha > 0,$$

and for $\alpha, \beta > -1$ the integrals

$$\int_{S_2} f(x,y)(x^2+y^2)^{\alpha}(1-x^2-y^2)^{\beta}\,\mathrm{d}x\,\mathrm{d}y$$

admit cubature sums $\mathscr{Q}(2, 4k+1, N)$ where at least $N \geqslant 2(k+1)^2$. Note that this includes the standard regions S_2 and $E_2^{r^2}$. This result can however not be generalised to all circularly symmetric integrals. In [92] the existence of a circularly symmetric integral admitting a cubature sum $\mathscr{Q}(2,9,17)$ has been proven.

The n-dimensional version of Theorem 11 was stated in [51]. An explicit form of the matrices involved was given in [97], using (21), which allows us to formulate (23) as follows.

Theorem 12. *If $\mathscr{Q}(n, 2k-1, N)$ is a cubature sum for an integral $\mathscr{I}^n$, then*

$$N \geqslant \dim \mathbb{P}_{k-1}^n + \tfrac{1}{2}\max\{\operatorname{rank}(M_{k-1}^{\star}(i,j)) \colon i,j = 1,2,\ldots,n\}. \tag{24}$$

Let us denote by $\mathscr{G}_{2m}$ the linear space of all even polynomials in $\mathbb{P}^n_{2m}$ and by $\mathscr{G}_{2m-1}$ the linear space of all odd polynomials in $\mathbb{P}^n_{2m-1}$. For integrals $\mathscr{I}^n$ which are centrally symmetric, i.e., for which

$$\mathscr{I}^n[Q] = 0 \quad \text{if } Q \in \mathscr{G}_{2m-1}, \ m \in \mathbb{N},$$

holds, another lower bound is known, which is not based on orthogonality.

Theorem 13. *If $\mathscr{Q}(n, 2k-1, N)$ is a cubature sum for a centrally symmetric integral $\mathscr{I}^n$, then*

$$N \geqslant 2 \dim \mathscr{G}_{k-1} - \begin{cases} 1, & if\ 0 \text{ is a node and } k \text{ is even}, \\ 0, & else. \end{cases}$$

This bound was given for degree 3 by Mysovskikh [55]; the general case is due to Möller [47,51] and Mysovskikh [64]. Möller proved that cubature formulae attaining the bound of Theorem 13 (having the node 0, if k is even) are centrally symmetric, too. For $n \geqslant 3$ and $\mathscr{I}^n$ centrally symmetric, the bound of Theorem 13 is better than the one of Theorem 12. For $n=2$ and $\mathscr{I}^n$ centrally symmetric, they coincide.

To conclude this section, we remark that cubature formulae with all nodes real and attaining the bounds of Theorem 9 or Theorem 13 are known to have all weights positive [58,90,47,12].

6. Methods of construction

6.1. Interpolation

Let $\Omega \subseteq \mathbb{R}^n$ be given and assume that Ω contains interior points. By virtue of the linear independence of $\{\varphi_j(\boldsymbol{x})\}_{j=1}^{\infty}$ we can find for each m exactly $\mu = M(n,m)$ points from Ω such that they generate a regular Vandermonde matrix. We remark that v points, $v < \mu$, are always contained in an algebraic manifold of degree m, hence μ is the minimal number of points which do not belong to such a manifold. We denote by

$$V_m = [\varphi_1(\boldsymbol{x}^{(j)}), \varphi_2(\boldsymbol{x}^{(j)}), \ldots, \varphi_\mu(\boldsymbol{x}^{(j)})]_{j=1}^{\mu}, \quad \mu = M(n,m), \tag{25}$$

the Vandermonde matrix defined by $\boldsymbol{x}^{(j)}$, $j = 1, 2, \ldots, \mu$.

Theorem 14. *The points $\boldsymbol{x}^{(j)}$, $j = 1, 2, \ldots, \mu$, do not lie on an algebraic manifold of degree m if and only if* $\det V_m \neq 0$.

A natural way to construct a cubature formula is interpolation. Choose μ points $\boldsymbol{x}^{(j)} \in \mathbb{R}^n$ which do not lie on a manifold of degree m. Because of the nonsingularity of the corresponding Vandermonde matrix we can construct the interpolating polynomial of f:

$$P_m(\boldsymbol{x}) = \sum_{j=1}^{\mu} L_j^{(m)}(\boldsymbol{x}) f(\boldsymbol{x}^{(j)}),$$

where

$$L_j^{(m)}(\boldsymbol{x}^{(i)}) = \delta_{ij}, \quad i,j = 1,2,\ldots,\mu.$$

Substituting P_m for f in (1), we obtain

$$\mathscr{I}[f] = \sum_{j=1}^{\mu} w_j f(\boldsymbol{x}^{(j)}) + R[f], \tag{26}$$

where

$$w_j = \int_{\Omega} L_j^{(m)}(\boldsymbol{x})\omega(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{27}$$

A cubature formula obtained in this way is obviously interpolatory, see Theorem 1.

6.2. Reproducing kernels

The method of reproducing kernel was introduced in [58] in order to construct cubature formulae of degree $2k$ with a minimal number of nodes $N = M(n,k)$. Most often the method will produce cubature formulae with more nodes. By means of the orthonormal basis such cubature formulae may be constructed by inserting $f = F_l F_m$ in (2) and studying

$$\sum_{j=1}^{N} w_j F_l(\boldsymbol{x}^{(j)}) F_m(\boldsymbol{x}^{(j)}) = \delta_{lm}, \quad l,m = 1,2,\ldots,N. \tag{28}$$

Introducing the $N \times N$ matrices

$$F = [F_1(\boldsymbol{x}^{(j)}), F_2(\boldsymbol{x}^{(j)}), \ldots, F_N(\boldsymbol{x}^{(j)})]_{j=1}^{N}$$

and $C = \operatorname{diag}\{w_1, w_2, \ldots, w_N\}$, we can write Eq. (28) as

$$F^{\mathrm{T}} C F = E.$$

This can be written as $FF^{\mathrm{T}} = C^{-1}$, i.e.,

$$\sum_{i=1}^{N} F_i(\boldsymbol{x}^{(r)}) F_i(\boldsymbol{x}^{(s)}) = w_r^{-1}\delta_{rs}, \quad r,s = 1,2,\ldots,N.$$

If we are using (14), this can be rewritten as

$$K_k(x^{(r)}, x^{(s)}) = w_r^{-1}\delta_{rs}, \quad r,s,=1,2,\ldots,N. \tag{29}$$

If we assume that (28) will lead to a cubature formula, then the nodes and coefficients can be determined by (29).

Let $\boldsymbol{a}^{(i)}$, $i=1,2,\ldots,n$, be pairwise distinct nodes of such a formula. We denote by $\mathscr{H}_i$ the algebraic manifold defined by the polynomial $K_k(\boldsymbol{a}^{(i)}, \boldsymbol{x})$. From (29) we obtain

$$K_k(\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)}) = b_i \delta_{ij}, \quad b_i = w_i^{-1}, \ \ i,j = 1,2,\ldots,n. \tag{30}$$

The remaining nodes of the formula belong to $\bigcap_{i=1}^{n} \mathscr{H}_i$ and can be computed by solving for the unknown variables $\boldsymbol{x}$ from

$$K_k(\boldsymbol{a}^{(i)}, \boldsymbol{x}) = 0, \quad i = 1,2,\ldots,n.$$

Since the nodes of the cubature sum $\mathscr{Q}(n, 2k, M(n,k))$ are not known, we proceed in the following way. The nodes $\boldsymbol{a}^{(i)}$ are chosen (whenever possible) in Ω but different from any of the common zeros of the fundamental system of orthogonal polynomials of degree k. So the order of the manifold $\mathscr{H}_i$ generated by $K_k(\boldsymbol{a}^{(i)}, \boldsymbol{x})$ is k.

If $\boldsymbol{a}^{(1)}$ is fixed, then $\boldsymbol{a}^{(2)}$ will be chosen on $\mathscr{H}_1$, and, if possible in Ω. If $\boldsymbol{a}^{(i)}$, $i = 1, 2, \ldots, t-1$, are fixed, the next node $\boldsymbol{a}^{(t)}$ is chosen on $\bigcap_{i=1}^{t-1} \mathscr{H}_i$, if possible in Ω. The $\boldsymbol{a}^{(i)}$, $i = 1, 2, \ldots, n$, constructed in this way satisfy (30). If all nodes are chosen in $\mathbb{R}^n$, then $b_i > 0$; in fact,

$$b_i = \sum_{j=1}^{N} F_j^2(\boldsymbol{a}^{(i)}) > 0,$$

since the $\boldsymbol{a}^{(i)}$ are no zeros of the fundamental system of orthogonal polynomials of degree k.

If there are no points at infinity on $\mathscr{H} = \bigcap_{i=1}^{n} \mathscr{H}_i$, then $\mathscr{H}$ consists of r points $\boldsymbol{x}^{(j)}$. Thus we obtain

$$\int_\Omega f(\boldsymbol{x})\omega(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \cong \mathscr{Q}(n, 2k, n+r) = \sum_{j=1}^{n} A_j f(\boldsymbol{a}^{(j)}) + \sum_{j=1}^{r} B_j f(\boldsymbol{x}^{(j)}). \tag{31}$$

The coefficients A_i can be computed from (31) since the formula is exact for $K_k(\boldsymbol{a}^{(i)}, \boldsymbol{x})$, i.e.,

$$\int_\Omega K_k(\boldsymbol{a}^{(i)}, \boldsymbol{x})\omega(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \sum_{j=1}^{n} A_j K_k(\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)}) = A_i K_k(\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(i)}),$$

or, by using (15) with $f \equiv 1$,

$$A_i = \frac{1}{b_i} = \frac{1}{K_k(\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(i)})}.$$

If $n + r = N = M(n,k)$, the coefficients B_j can be computed in an analogous way; if $n + r > N$, the B_j are determined by the condition for (31) to be of degree $2k$.

The method of reproducing kernels can be applied to regions in $\mathbb{R}^n$ without inner points, see [36,52]. The method was applied in [58,7,8,25] to construct cubature formulae of degree 4 for a variety of regions and in [47] to construct a cubature formula of degree 9 for the region S_2.

Möller [47] and Gegel' [26] proved

Theorem 15. *If* $\boldsymbol{a}^{(i)}$, $i = 1, 2, \ldots, n$, *satisfy* (30) *where* $b_i \neq 0$, *and if* $\bigcap_{i=1}^{n} \mathscr{H}_i$ *consists of pairwise distinct nodes* $\boldsymbol{x}^{(j)}$, $j = 1, 2, \ldots, k^n$, *then*

$$\int_\Omega f(\boldsymbol{x})\omega(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \cong \mathscr{Q}(m, 2k, n+k^n) = \sum_{i=1}^{n} \frac{1}{b_i} f(\boldsymbol{a}^{(i)}) + \sum_{j=1}^{k^n} w_j f(\boldsymbol{x}^{(j)}),$$

where $b_i = K_k(\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(i)})$.

Möller modified this for centrally symmetric integrals by using the following important observation. For these integrals, the orthogonal polynomials of degree m are even (odd) polynomials, if m is even (odd). For the linear space $\tilde{\mathbb{P}}_k^n$ of all even (odd) polynomials of degree $\leqslant k$ if k is even (odd) the reproducing kernel

$$\tilde{K}_k(\boldsymbol{u},\boldsymbol{x}) = {\sum_{j=t}^{N}}' F_j(\boldsymbol{u})F_j(\boldsymbol{x}), \quad t = k - 2\lfloor k/2\rfloor + 1, \;\; N = M(n,k),$$

is considered. Here $\sum'$ denotes summation over all even (odd) polynomials F_j if j is even (odd).

Again, nodes $\boldsymbol{a}^{(i)}$, $i = 1,2,\dots,n$, are chosen (whenever possible) in Ω but different from any of the common zeros of the fundamental system of orthogonal polynomials of degree k.

The manifolds corresponding to $\tilde{K}_k(\boldsymbol{a}^{(i)},\boldsymbol{x})$ will be denoted by $\tilde{\mathscr{H}}_i$. If $\boldsymbol{a}^{(i)}$, $i = 1,2,\dots,t-1$, are already selected, the node $\boldsymbol{a}^{(t)}$ is chosen on $\bigcap_{i=1}^{t-1}\tilde{\mathscr{H}}_i$, if possible in Ω. These nodes satisfy

$$\tilde{K}_k(\boldsymbol{a}^{(i)},\boldsymbol{a}^{(j)}) = b_i\delta_{ij}, \quad i,j = 1,2,\dots,n, \tag{32}$$

where $b_i > 0$ since $\boldsymbol{a}^{(i)} \in \mathbb{R}^n$.

We remark that the nodes in the modified method are chosen as $\boldsymbol{a}^{(i)}, -\boldsymbol{a}^{(i)}$, $i=1,2,\dots,n$. By central symmetry it follows that

$$\tilde{K}_k(\boldsymbol{a}^{(i)},-\boldsymbol{a}^{(j)}) = (-1)^k\tilde{K}_k(\boldsymbol{a}^{(i)},\boldsymbol{a}^{(j)}),$$

hence by (32), if $b_i \neq 0$, we get $\boldsymbol{a}^{(i)} \neq -\boldsymbol{a}^{(j)}$ if $i \neq j$. So the $\boldsymbol{a}^{(i)}, -\boldsymbol{a}^{(i)}$ are pairwise distinct, if $\boldsymbol{a}^{(i)} \neq 0$, $i = 1,2,\dots,n$. If k is odd, this is satisfied; if k is even, the number of pairwise distinct nodes $\boldsymbol{a}^{(i)}$ and $-\boldsymbol{a}^{(i)}$ may be $2n$ or $2n-1$. In [47] the following is derived.

Theorem 16. *Let the integral be centrally symmetric. If the nodes $\boldsymbol{a}^{(i)}$, $i = 1,2,\dots,n$, satisfy* (32) *where $b_i \neq 0$, and $\bigcap_{i=1}^{n}\tilde{H}_i$ consists of pairwise distinct points $\boldsymbol{x}^{(j)}$, $j = 1,2,\dots,k^n$, then*

$$\int_\Omega f(\boldsymbol{x})\omega(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \cong \mathscr{Q}(n,2k+1,2n+k^n)$$
$$= \sum_{i=1}^{n}\frac{1}{2b_i}[f(\boldsymbol{a}^{(i)}) + f(-\boldsymbol{a}^{(i)})] + \sum_{j=1}^{k^n} w_j f(\boldsymbol{x}^{(j)}),$$

where $b_i = \tilde{K}_k(\boldsymbol{a}^{(i)},\boldsymbol{a}^{(i)})$.

6.3. Ideal theory

Let

$$X = \{\boldsymbol{x}^{(j)},\; j = 1,2,\dots,N\} \subset \mathbb{R}^n$$

be a finite set of points, and define the subspace

$$\mathscr{W} = \{P \in \mathbb{P}^n_m\colon P(\boldsymbol{x}) = 0 \text{ for all } \boldsymbol{x} \in X\} \subset \mathbb{P}^n.$$

Sobolev [83] proved

Theorem 17. *The points X are the nodes of $\mathscr{Q}(n,m,N)$ for $\mathscr{I}^n$ if and only if*

$$P \in \mathscr{W} \quad \textit{implies } \mathscr{I}^n[P] = 0.$$

An English rendering of the proof can be found in [10].

The relationship of orthogonal polynomials and cubature formulae was studied since 1967 by Mysovskikh [57,59,61–63] and Stroud [87–90]; in particular they introduced elements from algebraic geometry.

Möller [47] recognised that this connection can be represented more transparently by using ideal theory, and that this theory will help in determining common zeros of orthogonal polynomials. E.g., Theorem 5 follows easily from this theory. Let $\mathcal{Q}(n,m,N)$ be given such that

$$X=\{\boldsymbol{x}^{(j)},\ j=1,2,\dots,N\}\subset\mathbb{R}^n$$

is the finite set of nodes, and define the polynomial ideal

$$\mathfrak{A}=\{P\in\mathbb{P}^n\colon P(\boldsymbol{x})=0 \text{ for all } \boldsymbol{x}\in X\}\subset\mathbb{P}^n.$$

Then we obtain for each $P\in\mathfrak{A}\cap\mathbb{P}^n_m$ the orthogonality condition $\mathscr{I}^n[PQ]=0$, whenever $PQ\in\mathbb{P}^n_m$. Möller introduced the notion of m-orthogonality. A set of polynomials is said to be m-orthogonal if for every element P we have $\mathscr{I}^n[PQ]=0$, when $PQ\in\mathbb{P}^n_m$. Hence, orthogonal polynomials of degree m are $(2m-1)$-orthogonal, while quasi-orthogonal polynomials of degree m are $(2m-2)$-orthogonal.

The main problem is the selection of a suitable basis. It turns out that an H-basis suits best. $\{P_1,P_2,\dots,P_s\}$ is such a basis, if every $Q\in\mathfrak{A}$ can be written as

$$Q=\sum_{i=1}^{s}Q_iP_i \quad \text{where } \deg(Q_iP_i)\leqslant\deg(Q).$$

The ideal then is written as $\mathfrak{A}=(P_1,P_2,\dots,P_s)$.

Theorem 18. *Let Q_i, $i=1,2,\dots,s$, be an H-basis of a zero-dimensional ideal $\mathfrak{A}$. Then the following are equivalent*:

(i) *The Q_i are m-orthogonal with respect to $\mathscr{I}^n$.*

(ii) *There is a $\mathcal{Q}(n,m,N)$ using the N common zeros of $\mathfrak{A}$ as nodes if no multiple nodes appear.*

For multiple nodes, function derivatives can be used; this was proposed in [62,63]. Möller called formulae of this type *generalised cubature formulae* of algebraic degree. This was further developed in [48–50].

In this ideal-theoretic setting, condition (5) can be interpreted as syzygy. If an H-basis of $\mathfrak{A}$ is fixed, then syzygies of higher order will occur; e.g., if $P_1,P_2\in\mathfrak{A}$ are of degree m, then it is possible that

$$Q_1P_1-Q_2P_2\in\mathbb{P}^n_{m-1} \quad \text{for } Q_i\in\mathbb{P}^n.$$

Möller found that such syzygies will occur in an H-basis and that they impose restrictions on the polynomials which can be used constructively to compute a suitable ideal. Furthermore, the Hilbert function can be used to study the number of common zeros. For the connection to Gröbner bases we refer to [53,13].

However, Theorem 18 allows nodes to be in $\mathbb{C}^n$. Schmid [78] proposed to avoid this by considering real ideals.

If the common zeros of an ideal $\mathfrak{A}$ are denoted by $\mathscr{V}(\mathfrak{A})$ and the ideal of all polynomials which vanish at a finite set $X \subset \mathbb{R}^n$ by $\mathfrak{A}_X$, then

$$X \subseteq \mathscr{V}(\mathfrak{A}_X).$$

An ideal $\mathfrak{A}$ is called real if

$$X = \mathscr{V}(\mathfrak{A}_X).$$

So m-orthogonal real ideals characterise cubature formulae, and real ideals are characterised by the following theorem due to Dubois et al. (cf. [18,19,75,76]).

Theorem 19. *The following are equivalent*:
(i) $\mathfrak{A}$ *is a real ideal,*
(ii) *the common zeros of* $\mathfrak{A}$ *are pairwise distinct and real,*
(iii) P *vanishes on* $\mathscr{V}(\mathfrak{A})$ *if and only if* $P \in \mathfrak{A}$,
(iv) *for all* $M \in \mathbb{N}$ *and all* $Q_i \in \mathbb{P}^n$, $i = 1, 2, \dots, M$,

$$\sum_{i=1}^{M} Q_i^2 \in \mathfrak{A} \textit{ implies } Q_i \in \mathfrak{A}, \quad i = 1, 2, \dots, M.$$

By combining Möller's results and the conditions which can be derived from Theorem 19 it is possible to give a complete characterisation of cubature formulae. However, if the degree of the formula m is fixed, the conditions which have to be satisfied strongly depend on the number of nodes. Indeed, the number of nodes influences the number of polynomials in the ideal basis and their degree. The conditions derived from Theorem 19 depend on the structure of the ideal basis, and their complexity therefore increases with m.

In [79] Theorem 5 is extended by using Theorem 19 and applying it to ideals containing a fundamental set of an arbitrary degree. It was then applied to construct cubature formulae for the regions C_2, S_2, T_2.

Theorem 20. *Let* R_i, $i = 1, 2, \dots, t$, *be linearly independent polynomials in* $\mathbb{P}^n_m$ *containing a fundamental system of degree m. If* $\mathfrak{A} = (R_1, R_2, \dots, R_t)$, *then*
(i) $\mathscr{V}(\mathfrak{A}) \leqslant N = \dim \mathbb{P}^n_m - t$,
(ii) $\mathscr{V}(\mathfrak{A}) = N$ *if and only if* $\mathfrak{A}$ *is a real ideal.*

6.3.1. Even-degree formulae

By applying the syzygies of first order to quasi-orthogonal polynomials it is possible to characterise all even-degree formulae attaining the lower bound in (22).

Theorem 21. *Let*

$$R_i = P_i^k + \sum_{j=0}^{k-1} \gamma_{ij} P_j^{k-1}, \quad i = 0, 1, \dots, k,$$

be quasi-orthogonal polynomials generating the ideal $\mathfrak{A}$. A cubature formula $\mathscr{Q}(2,2k-2,\dim\mathbb{P}_{k-1})$ with all weights positive exists if and only if the parameters γ_{ij} can be chosen such that

$$yR_i - xR_{i+1} \in \operatorname{span}\{R_j,\ j=0,1,\ldots,k\},\quad i=0,1,\ldots,k-1, \tag{33}$$

holds. If (33) *holds, then the nodes of the formula are given by $\mathscr{V}(\mathfrak{A})$, and $\mathfrak{A}$ is real.*

Morrow and Patterson [54] proved this by applying Möller's Theorem 18 and using the Hilbert function to count the common zeros, counting multiplicities.

Schmid [77] applied Theorem 19 to prove that (33) is necessary and sufficient for $\mathfrak{A}$ to be a real ideal. From the work in [54] a classical integral is known for which all even-degree minimal formulae can be computed, see Section 7.1.

The complexity of (33) can be realised by considering the equivalent matrix equation given in [81] for integrals having central symmetry. The quadratic matrix equation

$$0 = M_{k-1}^{\star} + \Gamma_k M_k^{-1} M_k^{\star} M_k^{-1} \Gamma_k^{\mathrm{T}}$$

has to be solved. Here $M_k = [\mathscr{I}[P_i^k P_j^k]]_{i,j=0,1,\ldots,k}$ is the moment matrix and $M_{k-1}^{\star}$ the matrix of the numerical characteristics. Γ_k is a $k \times k+1$ Hankel matrix, which has to be determined; from its coefficients the γ_{ij}'s can be computed.

The straightforward generalisation to the n-dimensional case has been studied in [79,74]; however, only moderate-degree formulae could be constructed for C_n, $n=2,3,4,5$.

6.3.2. Odd-degree formulae

Stroud and Mysovskikh [88,59] proved that $\mathscr{Q}(2,2k-1,k^2)$ can be constructed if two orthogonal polynomials of degree k can be found having exactly k^2 common pairwise distinct real zeros. Franke [21] derived sufficient conditions implying the existence of $\mathscr{Q}(2,2k-1,N)$, where $N<k^2$, for special integrals over planar regions. Further generalisations were obtained in [62,63] by admitting point evaluations of derivatives and preassigning nodes.

We recall Theorem 8 in the following form.

Theorem 22. *$\mathscr{Q}(n,2k-1,\dim\mathbb{P}_{k-1})$ exists if and only if the nodes are the zeros of $\boldsymbol{P}_k$.*

For the standard regions, such formulae exist for $n=1$ and $k=1,2,\ldots$ or $k=1$ and $n=1,2,\ldots$; for $n\geqslant 2$, $k\geqslant 2$, such formulae do not exist. The existence of $M(n,k-1)$ common roots of the polynomials gathered in $\boldsymbol{P}_k$ can be reduced to the solution of a nonlinear system in n unknowns; however, the number of equations is larger than n, since for $n,k\geqslant 2$, we have

$$M(n-1,k) \geqslant M(n-1,2) = \frac{n(n+1)}{2} \geqslant n+1.$$

The existence of special regions for which Theorem 22 holds for moderate k have been discussed in Section 5.2. A class of integrals for which Theorem 22 holds for arbitrary k was presented in [82] for $n=2$, and in [5] for n arbitrary.

In order to find cubature sums $\mathcal{Q}(2,2k-1,N)$ where N attains the improved lower bound (23), Möller derived the following necessary conditions:

Theorem 23. *If $\mathcal{Q}(2,2k-1,N)$ exists where $N=\dim \mathbb{P}_{k-1}+\frac{1}{2}\operatorname{rank}(M_{k-1}^{\star})$, then there are $s=k+1-\frac{1}{2}\operatorname{rank}(M_{k-1}^{\star})$ orthogonal polynomials P_i of degree k vanishing at the nodes of the formula and satisfying the following conditions.*

(i) *Whenever orthogonal polynomials of degree k satisfy $xQ_1-yQ_2=Q_3$, then $Q_i\in \operatorname{span}\{P_i,\ i=1,2,\ldots,s\}$.*

(ii) *xP_i, yP_i form a fundamental set of degree $k+1$.*

(iii) *There are $2k-\frac{3}{2}\operatorname{rank} M_{k-1}^{\star}$ linearly independent vectors $a\in\mathbb{R}^{3(k+1)}$ such that*

$$x^2\sum_{i=0}^{k} a_iP_i + xy\sum_{i=0}^{k} a_{k+1+i}P_i + y^2\sum_{i=0}^{k} a_{2k+2+i}P_i = \sum_{i=1}^{s} L_iP_i,$$

where L_i are linear polynomials.

These conditions are almost sufficient, too.

Theorem 24. *If there are $s=k+1-\frac{1}{2}\operatorname{rank}(M_{k-1}^{\star})$ orthogonal polynomials P_i of degree k satisfying the conditions* (i), (ii), *and* (iii) *in Theorem* 23, *then these polynomials have $N=\dim \mathbb{P}_{k-1}+\frac{1}{2}\operatorname{rank}(M_{k-1}^{\star})$ affine common zeros. If they are pairwise distinct and real, then a cubature sum $\mathcal{Q}(2,2k-1,N)$ exists.*

The surprising result from this theorem was the construction of $\mathcal{Q}(2,9,17)$ for C_2, the square with constant weight function. Franke [22] expected that 20 would be the lowest possible number of nodes for such a formula. Haegemans and Piessens [70,33] conjectured that 18 would be lowest possible.

Again, by applying Theorem 19 one can determine further conditions which guarantee that the polynomials P_i generate a real ideal, i.e., have pairwise distinct real zeros. To check this, choose U_i, $i=1,2,\ldots,k+1-s$, such that P_i, U_i are a fundamental system of degree k. By virtue of condition (ii) of Theorem 23 there are polynomials R_{ij} and $P\in\operatorname{span}\{P_i,\ i=1,2,\ldots,s\}$ such that $U_iU_j-R_{ij}P\in\mathbb{P}_{k+1}$. If, in addition, the P_i are chosen such that

$$\mathcal{I}[U^2-RP]>0$$

for all $U\in\operatorname{span}\{U_i,\ i=1,2,\ldots,k+1-s\}$, $P\in\operatorname{span}\{P_i,\ i=1,2,\ldots,s\}$, and R such that $U^2-RP\in\mathbb{P}_{k+1}$, then the ideal $(P_1,P_2,\ldots,P_s)$ is real.

This holds in the n-dimensional case, too, even if we admit ideals with a fundamental system of maximal degree $m+1$ [79].

Theorem 25. *Let R_i, $i=1,2,\ldots,t$, be an m-orthogonal fundamental set of degree $m+1$ of linearly independent polynomials in $\mathbb{P}^n$, and let $\mathfrak{A}=(R_1,R_2,\ldots,R_t)$ and $\mathcal{W}=\operatorname{span}\{R_1,R_2,\ldots,R_t\}$. Let $\mathcal{U}$, $\dim\mathcal{U}=N$, be an arbitrary, but fixed, complement of $\mathcal{W}$ such that $\mathbb{P}_{m+1}^n=\mathcal{W}\oplus\mathcal{U}$. Then the following are equivalent*:

(i) *A positive $\mathcal{Q}(n,m,N)$ for $\mathcal{I}^n$ exists with nodes in $\mathcal{V}(\mathfrak{A})$.*

(ii) $\mathfrak{A}$ *and* $\mathscr{U}$ *are characterised by*
  (a) $\mathfrak{A} \cap \mathscr{U} = (0)$,
  (b) $\mathscr{I}^n[Q^2 - R^+] > 0$ *for all* $Q \in \mathscr{U}$, *where* R^+ *is chosen such that* $Q^2 - R^+ \in \mathbb{P}^n_m$.

(iii) $\mathfrak{A}$ *is a real ideal with a zero-set of N pairwise distinct real points, which are the nodes of the cubature formula of degree m.*

6.4. Formulae characterised by three orthogonal polynomials

The nodes of a Gauss quadrature formula are the zeros of one particular polynomial. The nodes of a Gauss product cubature formula in n dimensions are the common zeros of n polynomials in n variables. Franke [21] derived conditions for planar product regions implying the existence of cubature sums $\mathscr{Q}(2, 2k-1, N)$ where $N < k^2$, see Section 6.3.2. This is based on the common zeros of two orthogonal polynomials.

Huelsman [35] proved that for fully symmetric regions, $\mathscr{Q}(2, 7, 10)$ cannot exist. Franke [22] proved that for these regions, and also for symmetric product regions, a cubature sum $\mathscr{Q}(2, 7, 11)$ cannot exist. He observed that from Stroud's characterisation [90] there follows that a cubature sum $\mathscr{Q}(2, 7, 12)$ is characterised by three orthogonal polynomials of degree 4, and he exploited this to construct some formulae.

In [70,71], Piessens and Haegemans observed that there are actually three orthogonal polynomials of degree k that vanish in the nodes of their cubature formulae of degree $2k-1$ for $k = 5, 6$. Following this observation, and using earlier results of Radon and Franke, in a series of articles [30,32,33] they constructed cubature formulae for a variety of planar regions whose nodes are the common zeros of three orthogonal polynomials in two variables. They restricted their work to regions that are symmetric with respect to both coordinate axes and noticed that Radon's cubature formulae for these regions have the same symmetry.

At first sight, it may look strange that Radon, Franke, Haegemans and Piessens characterised cubature formulae in two dimensions as the common zeros of three polynomial equations in two unknowns, i.e., as an overdetermined system of nonlinear equations. We now know, see Section 5.2, that for centrally symmetric regions there are $\lfloor \frac{k}{2} \rfloor + 1$ linearly independent orthogonal polynomials of degree k that vanish in the nodes of a cubature formula of degree $2k-1$ that attains the lower bound of Theorem 11. We thus know that formulae of degree 5 and 7 that attain this bound are fully characterised by three such polynomials. Formulas of higher degrees $2k-1$ that attain this bound will have even more than three linearly independent orthogonal polynomials of degree k that vanish in their nodes.

Franke, Haegemans and Piessens proceeded as follows. They assumed the existence of three linearly independent orthogonal polynomials of the form

$$\phi_i = \sum_{j=0}^{k} a_{ij} P_j^k, \quad i = 1, 2, 3.$$

The first set of conditions on the unknowns a_{ij} is obtained by demanding that whenever a node (α_i, β_i) is a common zero of ϕ_1, ϕ_2, and ϕ_3, then also $(\pm\alpha_i, \pm\beta_i)$ is. A second set of conditions is obtained by demanding that these three polynomials have sufficiently many common zeros. Obtaining these conditions requires much labour, and a computer algebra system was used to derive some of these. For higher degrees, the result contains some free parameters, and consequently a continuum

Table 1
Number of nodes in known cubature formulae [22,30,32,33][a]

Degree	C_2	S_2	$E_2^{r^2}$	E_2^r	H_2
7	12(∞)	12(∞)	12(∞)	12(∞)	12(∞)
9	19(∞)	19(∞)	19(∞)	19(∞)	19(∞)
	18(2)	18(1)	18(1)		18(1)
11	28(∞)	28(∞)	28(∞)	28(∞)	28(∞)
	26(∞)	26(∞)	26(∞)	26(∞)	26(∞)
	25(2)	25(1)	25(1)		25(1)

[a] In parentheses the number of such cubature formulae is given.

of cubature formulae was obtained. In such a continuum Haegemans and Piessens searched for the formula with the lowest number of nodes, e.g., by searching for a formula with a weight equal to zero. An overview of the cubature formulae for the symmetric standard regions C_2, S_2, $E_2^{r^2}$, E_2^r, and H_2 obtained in this way is presented in Table 1.

This approach was also used to construct cubature formulae of degree 5 for the four standard symmetric regions in three dimensions [31]. A continuum of formulae with 21 nodes is obtained. It is mentioned that this continuum contains formulae with 17, 15, 14 and 13 nodes, the last being the lowest possible.

7. Cubature formulae of arbitrary degree

For an overview of all known minimal formulae, we refer to [10]. In this final section we present those integrals for which minimal cubature formulae for an arbitrary degree of exactness were constructed by using orthogonal polynomials. Though these examples are limited, they illustrate that all lower bounds which have been discussed will be attained for special integrals and that the construction methods based on orthogonal polynomials can be applied. Indirectly this shows that improving these bounds will require more information about the given integral to be taken into account. The symmetry of the region Ω and the weight function ω is not enough!

7.1. Minimal formulae for the square with special weight functions

Two-dimensional integrals with an infinite number of minimal cubature formulae have been presented by Morrow and Patterson [54]. They studied

$$\mathscr{I}_{1/2}[f] = \int_{-1}^{1}\int_{-1}^{1} f(x,y)(1-x^2)^{1/2}(1-y^2)^{1/2}\,\mathrm{d}x\,\mathrm{d}y.$$

The associated fundamental orthogonal system of degree k, U_i^k, $i=0,1,\ldots,k$, is gathered in

$$\boldsymbol{U}_k = (U_0^k, U_1^k, \ldots, U_k^k)^{\mathrm{T}}.$$

Similarly, they studied

$$\mathscr{I}_{-1/2}[f] = \int_{-1}^{1}\int_{-1}^{1} f(x,y)(1-x^2)^{-1/2}(1-y^2)^{-1/2}\,\mathrm{d}x\,\mathrm{d}y,$$

where the associated fundamental orthogonal system of degree k, T_i^k, $i=0,1,\dots,k$, is gathered in

$$\boldsymbol{T}_k = (T_0^k, T_1^k, \dots, T_k^k)^{\mathrm{T}}.$$

7.1.1. Even-degree formulae for $\mathscr{I}_{1/2}$

A minimal cubature sum $\mathscr{Q}(2, 2k-2, \dim \mathbb{P}_{k-1})$ has been derived in [54]; the nodes are the common zeros of

$$\boldsymbol{U}_k + 1/2F_k^{\mathrm{T}}\boldsymbol{U}_{k-1},$$

where $F_k = [0\ E_k]$. This is the special case $\sigma = 1$ from the following result [80,81]:

For $k \geqslant 6$, up to symmetries, all minimal cubature sums $\mathscr{Q}(2, 2k-2, \dim \mathbb{P}_{k-1})$ are generated by the common zeros of

$$\boldsymbol{U}_k + 1/2\Gamma_k^{\mathrm{T}}\boldsymbol{U}_{k-1},$$

where Γ_k is a Hankel matrix of the form

$$\Gamma_k = \begin{bmatrix} \gamma_0 & \sigma\gamma_0 & \sigma^2\gamma_0 & \cdots & \sigma^{k-1}\gamma_0 & 1/\sigma \\ \sigma\gamma_0 & \sigma^2\gamma_0 & \sigma^3\gamma_0 & \cdots & 1/\sigma & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \sigma^{k-1}\gamma_0 & 1/\sigma & 0 & \cdots & 0 & 0 \end{bmatrix}, \quad \gamma_0 = \frac{1-\sigma^2}{\sigma^{k+1}}, \ 0 \neq \sigma \in \mathbb{R}.$$

7.1.2. Odd-degree formulae for $\mathscr{I}_{1/2}$

Up to symmetries, all minimal cubature sums $\mathscr{Q}(2, 2k-1, \dim \mathbb{P}_{k-1} + \lfloor k/2 \rfloor)$, k odd, for $\mathscr{I}_{1/2}$ are generated by the common zeros of

$$(E_k + \Gamma_k)\boldsymbol{U}_k,$$

where Γ_k is an orthogonal Hankel matrix of the form

$$\Gamma_k = \begin{bmatrix} \gamma_0 & \sigma\gamma_0 & \dots & \sigma^{k-1}\gamma_0 & \sigma^k\gamma_0 - \sigma \\ \sigma\gamma_0 & \sigma^2\gamma_0 & \dots & \sigma^k\gamma_0 - \sigma & \gamma_0 \\ \vdots & \vdots & & \vdots & \vdots \\ \sigma^{k-1}\gamma_0 & \sigma^k\gamma_0 - \sigma & \dots & \sigma^{k-3}\gamma_0 & \sigma^{k-2}\gamma_0 \\ \sigma^k\gamma_0 - \sigma & \gamma_0 & \dots & \sigma^{k-2}\gamma_0 & \sigma^{k-1}\gamma_0 \end{bmatrix}, \tag{34}$$

where

$$\gamma_0 = 2/(k+1), \quad \sigma^2 = 1, \quad \text{or } \gamma_0 = \frac{\sigma^2 - 1}{\sigma^{k+1} - 1}, \quad \sigma^2 \neq 1, \ \sigma \in \mathbb{R}.$$

Note that there are redundancies in (34), $\operatorname{rank}(E_k + \Gamma_k) = \lfloor k/2 \rfloor + 2$. The general form is obtained in [81], special cases having been known long before: for $\sigma = 0$, $\gamma_0 = 1$ see [78], for $\sigma = 1$ and $\sigma = -1$ see [15].

For odd-degree formulae, k even, no general formula is known. However, there are minimal cubature sums $\mathscr{Q}(2, 2k-1, \dim \mathbb{P}_{k-1} + k/2)$, k even, for $\mathscr{I}_{1/2}$, generated by the common zeros of

$$(E_k + \Gamma_k)\boldsymbol{U}_k,$$

where Γ_k is an orthogonal Hankel matrix of the form (34), where

$$\gamma_0 = 2/(k+1), \quad \sigma = 1, \quad \text{or} \quad \gamma_0 = \frac{\sigma^2 - 1}{\sigma^{k+1} - 1}, \quad \sigma \neq 1, \ \sigma \in \mathbb{R}.$$

The case $\sigma = -1$, $\gamma_0 = 0$ was stated in [54]. The result for $\sigma = 1$ is obtained in [15], for $\sigma \neq 1$ it is obtained in [81].

7.1.3. Odd-degree formulae for $\mathscr{I}_{-1/2}$

If k is even, a minimal formula of degree $2k-1$ exists, the nodes being the common zeros of

$$T_i^k + T_{k-i}^k, \quad i = 0, 1, \ldots, k/2,$$

this result is due to [54]. Minimal formulae of degree $2k-1$, k odd or even, were derived in [15], the nodes are the common zeros of

$$T_i^k - T_{k-i}^k, \quad i = 0, 1, \ldots, \lfloor k/2 \rfloor, \ \ T_0^k + T_1^k + \cdots + T_{k-1}^k + T_k^k.$$

A third formula of degree $2k-1$ for k even is given in [15], the nodes are the common zeros of

$$T_i^k - T_{k-i}^k, \quad i = 0, 1, \ldots, k/2 - 1, \ \ T_0^k + T_2^k + \cdots + T_{k-2}^k + T_k^k.$$

7.1.4. Gaussian formulae

Cubature formulae attaining the lower bound (22) for even and odd degree are often called formulae of *Gaussian type* or *Gaussian formulae*. They exist for a class of (nonstandard) integrals, which will be shown in this section. This result is due to [82].

Let $\omega(x)$ be a nonnegative function on $I \subseteq \mathbb{R}$ and let $\{p_s\}$ be the orthonormal polynomials with respect to ω. Koornwinder [38] introduced bivariate orthogonal polynomials as follows.

For given $s \in \mathbb{N}$ let $u = x + y$ and $v = xy$ and define

$$P_i^{s,(-1/2)}(u,v) = \begin{cases} p_s(x)p_i(y) + p_s(y)p_i(x) & \text{if } i < s, \\ \sqrt{2}\, p_s(x)p_s(y) & \text{if } i = s, \end{cases}$$

and

$$P_i^{s,(1/2)}(u,v) = \frac{p_{s+1}(x)p_i(y) - p_{s+1}(y)p_i(x)}{x - y}.$$

Then $P_i^{s,(\pm 1/2)}$ are polynomials of total degree s. Koornwinder showed that they form a bivariate orthogonal system with respect to the weight function

$$(u^2 - 4v)^{\pm 1/2} W(u,v).$$

Let $x_{i,s}$ be the zeros of the quasi-orthogonal polynomial $p_s + \rho p_{s-1}$ where $\rho \in \mathbb{R}$ is arbitrary but fixed. The roots are ordered by $x_{1,s} < \cdots < x_{s,s}$. Let $u = x + y$ and $v = xy$, and define $W(u,v) = \omega(x)\omega(y)$.

Then we have the following Gaussian cubature formula of degree $2k-2$:

$$\iint_\Omega f(u,v)W(u,v)(u^2 - 4v)^{-1/2}\,\mathrm{d}u\,\mathrm{d}v \cong \sum_{i=1}^{k} \sum_{j=1}^{i} \omega_{i,j} f(x_{i,k} + x_{j,k}, x_{i,k}x_{j,k}),$$

and

$$\iint_{\Omega} f(u,v)W(u,v)(u^2-4v)^{1/2}\,\mathrm{d}u\,\mathrm{d}v \cong \sum_{i=1}^{k+1}\sum_{j=1}^{i-1}\omega_{i,j} f(x_{i,k+1}+x_{j,k+1}, x_{i,k+1}x_{j,k+1}),$$

where

$$\Omega = \{(u,v)\colon (x,y)\in I\times I \text{ and } x<y\}.$$

If $\rho = 0$, then a uniquely determined formula of degree $2k-1$ will be obtained.

So there are classes of two-dimensional integrals for which the one-dimensional result of Gaussian quadrature formulae can be regained. The lower bound (22) will be attained for odd and even degree, the common zeros of

$$\boldsymbol{P}_k + \rho\Gamma_k\boldsymbol{P}_{k-1}, \quad \Gamma_k \in \mathbb{R}^{k+1\times k}, \;\; \rho\in\mathbb{R},$$

are the nodes of the formula, where Γ_k is determined from commuting properties in the orthonormal recursion formula in Theorem 4 and a matrix equation which follows from (33) in Theorem 21.

These examples have been extended to the n-dimensional case in [5].

Acknowledgements

The authors would like to thank Frau Jutta Zintchenko for her help in translating Russian texts. They are also indebted to Professor Walter Gautschi for removing the influences of Dutch and German from our version of English and for his suggestions to improve the presentation.

References

[1] P. Appell, Sur une classe de polynômes a deux variables et le calcul approché des intégrales doubles, Ann. Fac. Sci. Univ. Toulouse 4 (1890) H1–H20.

[2] P. Appell, J.K. de Fériet, Fonctions hypergéometriques et hypersphériques – Polynomes d'Hermite, Gauthiers-Villars et Cie., Paris, 1926.

[3] H. Berens, H.J. Schmid, On the number of nodes of odd degree cubature formulae for integrals with Jacobi weights on a simplex, in: T.O. Espelid, A. Genz (Eds.), Numerical Integration, Kluwer Academic Publishers, Dordrecht, 1992, pp. 37–44.

[4] H. Berens, H.J. Schmid, Y. Xu, On twodimensional definite orthogonal systems and on a lower bound for the number of nodes of associated cubature formulae, SIAM J. Math. Anal. 26 (1995) 468–487.

[5] H. Berens, H.J. Schmid, Y. Xu, Multivariate Gaussian cubature formulae, Arch. Math. 64 (1995) 26–32.

[6] H. Bourget, Sur une extension de la méthode de quadrature de Gauss, Acad. Sci. Paris 126 (1898) 634–636.

[7] T.M. Bykova, Cubature formulae with a minimal number of nodes in a plane region which are exact for polynomials of degree 4, Vestnik Leningrad. Univ. Math. 7 (1969) 145–147 (in Russian).

[8] T.M. Bykova, Cubature formulae for three-dimensional integrals of degree four with eleven nodes, Isw. AN BSSR. Ser. Phys.-Math. 10 (1) (1970) 51–54 (in Russian).

[9] V.Ja. Cernicina, Construction of domains for which there exist interpolation cubature formulae with the least number of nodes, Vestnik Leningrad. Univ. No. 1, Mat. Meh. Astronom. Vyp. 28 (1973) 144–147 (in Russian).

[10] R. Cools, Constructing Cubature Formulae: The Science Behind the Art, Acta Numerica, Vol. 6, Cambridge University Press, Cambridge, 1997, pp. 1–54.

[11] R. Cools, Monomial cubature rules since "Stroud": a compilation – Part 2, J. Comput. Appl. Math. 112 (1999) 21–27.

[12] R. Cools, A. Haegemans, Why do so many cubature formulae have so many positive weights?, BIT 28 (1988) 792–802.

[13] R. Cools, A. Haegemans, Construction of symmetric cubature formulae with the number of knots (almost) equal to Möller's lower bound, in: H. Braß, G. Hämmerlin (Eds.), Numerical Integration, ISNM, Vol. 85, Birkhäuser, Basel, 1988, pp. 25–36.

[14] R. Cools, P. Rabinowitz, Monomial cubature rules since "Stroud": a compilation, J. Comput. Appl. Math. 48 (1993) 309–326.

[15] R. Cools, H.J. Schmid, Minimal cubature formulae of degree $2k-1$ for two classical functionals, Computing 43 (1989) 141–157.

[16] R. Cools, H.J. Schmid, A new lower bound for the number of nodes in cubature formulae of degree $4n+1$ for some circularly symmetric integrals, in: H. Brass, G. Hämmerlin (Eds.), Integration IV, International Series of Numerical Mathematics, Vol. 112, Birkhäuser, Basel, 1993, pp. 57–66.

[17] P.J. Davis, P. Rabinowitz, Methods of Numerical Integration, 2nd Edition, Academic Press, New York, 1984.

[18] D.W. Dubois, G. Efroymson, Algebraic theory of real varieties I, in: Studies and Essays Presented to Yu-Why Chen on his Sixtieth Birthday, Academica Sinica, Tapei, 1970, pp. 107–135.

[19] G. Efroymson, Local reality on algebraic varieties, J. Algebra 29 (1974) 133–142.

[20] H. Engels, Numerical Quadrature and Cubature, Academic Press, London, 1980.

[21] R. Franke, Obtaining cubatures for rectangles and other planar regions by using orthogonal polynomials, Math. Comp. 25 (1971) 813–817.

[22] R. Franke, Minimal point cubatures of precision seven for symmetric planar regions, SIAM J. Numer. Anal. 10 (1973) 849–862.

[23] N.N. Fritsch, On the existence of regions with minimal third degree integration formulas, Math. Comp. 24 (1970) 855–861.

[24] B. Fuglede, The multidimensional moment problem, Exposition. Math. 1 (1983) 47–65.

[25] G.N. Gegel', On a cubature formula for the four-dimensional ball, Ž. Vyčisl. Mat. i Mat. Fiz. 15 (1) (1975) 234–236 (in Russian), USSR Math. Comp. 15 (1) (1975) 226–228.

[26] G.N. Gegel', Construction of cubature formulae exact for polynomials of degree $2m$, Voprosy Vychisl. i Prikl. Mat., Tashkent 32 (1975) 5–9 (in Russian).

[27] W. Gröbner, Über die Konstruktion von Systemen orthogonaler Polynome in ein- und zweidimensionalen Bereichen, Monatsh. Math. 52 (1948) 38–54.

[28] C. Günther, Über die Anzahl von Stützstellen in mehrdimensionalen Integrationsformeln mit positiven Gewichten, Ph.D., Universität Karlsruhe, 1973.

[29] C. Günther, Third degree integration formulas with four real points and positive weights in two dimensions, SIAM J. Numer. Anal. 11 (1974) 480–493.

[30] A. Haegemans, Tables of symmetrical cubature formulas for the two-dimensional hexagon, Department of Computer Science, K.U. Leuven, Report TW, 1976.

[31] A. Haegemans, Construction of known and new cubature formulas of degree five for three-dimensional symmetric regions, using orthogonal polynomials, in: G. Hämmerlin (Ed.), Numerical Integration, International Series of Numerical Mathematics, Vol. 57, Birkhäuser, Basel, 1982, pp. 119–127.

[32] A. Haegemans, R. Piessens, Construction of cubature formulas of degree eleven for symmetric planar regions, using orthogonal polynomials, Numer. Math. 25 (1976) 139–148.

[33] A. Haegemans, R. Piessens, Construction of cubature formulas of degree seven and nine symmetric planar regions, using orthogonal polynomials, SIAM J. Numer. Anal. 14 (1977) 492–508.

[34] P.M. Hirsch, Evaluation of orthogonal polynomials and relationship to evaluating multiple integrals, Math. Comp. 22 (1968) 280–285.

[35] C.B. Huelsman III, Quadrature formulas over fully symmetric planar regions, SIAM J. Numer. Anal. 10 (1973) 539–552.

[36] G.P. Ismatullaev, On the construction of formulae for hypercube and sphere, Voprosy Vycisl. i Prikl. Mat., Tashkent 51 (1978) 191–203 (in Russian).

[37] D. Jackson, Formal properties of orthogonal polynomials in two variables, Duke Math. J. 2 (1936) 423–434.

[38] T. Koornwinder, Two-variable analogues for the classical orthogonal polynomials, in: R.A. Askey (Ed.), Theory and Applications of Special Functions, Academic Press, New York, 1975.

[39] M.A. Kowalski, The recursion formulas for orthogonal polynomials in *n* variables, SIAM J. Math. Anal. 13 (1982) 309–315.

[40] M.A. Kowalski, Orthogonality and recursion formulas for polynomials in *n* variables, SIAM J. Math. Anal. 13 (1982) 316–323.

[41] M.A. Kowalski, Algebraic Characterization of Orthogonality in the Space of Polynomials, Lecture Notes in Mathematics, Vol. 1171, Springer, Berlin, 1985, pp. 101–110.

[42] H.L. Krall, I.M. Sheffer, Orthogonal polynomials in two variables, Ann. Mat. Pura Appl. (4) 76 (1967) 325–376.

[43] V.I. Krylov, Approximate Calculation of Integrals, 2nd Edition, Nauka, Moscow, 1967 (in Russian).

[44] V.A. Kuzmenkov, The finite moment problem in several variables, Vestnik Leningr. Univ. 19 (1975) 26–31 (in Russian).

[45] V.A. Kuzmenkov, On the existence of cubature formulae with a minimal number of nodes, Ž h. Vyčisl. Mat. i Mat. Fiz. 16(5) (1976) 1337–1334 (in Russian). USSR Math. Comp. 16(5) (1975) 242–245.

[46] V.A. Kuzmenkov, Gaussian type cubature formulae and the finite moment problem, Metody Vychisl. 11 (1978) 21–42 (in Russian).

[47] H.M. Möller, Polynomideale und Kubaturformeln, Ph.D. Thesis, Univ. Dortmund, 1973.

[48] H.M. Möller, Mehrdimensionale Hermite-Interpolation und numerische Integration, Math. Z. 148 (1976) 107–118.

[49] H.M. Möller, Kubaturformeln mit minimaler Knotenzahl, Numerische Mathematik 25 (1976) 185–200.

[50] H.M. Möller, Hermite Interpolation in Several Variables Using Ideal-Theoretic Methods, Lecture Notes in Mathematics, Vol. 571, Springer, Berlin, 1977, pp. 155–163.

[51] H.M. Möller, Lower bounds for the number of nodes in cubature formulae, in: G. Hämmerlin (Ed.), Numerische Integration, International Series of Numerical Mathematics, Vol. 45, Birkhäuser, Basel, 1979, pp. 221–230.

[52] H.M. Möller, The construction of cubature formulae and ideals of principal classes, in: G. Hämmerlin (Ed.), Numerische Integration, International Series of Numerical Mathematics, Vol. 45, Birkhäuser, Basel, 1979, pp. 249–263.

[53] H.M. Möller, On the construction of cubature formulae with few nodes using Gröbner bases, in: P. Keast, G. Fairweather (Eds.), Numerical Integration – Recent Developments, Software and Applications, Reidel, Dordrecht, 1987, pp. 177–192.

[54] C.R. Morrow, T.N.L. Patterson, Construction of algebraic cubature rules using polynomial ideal theory, SIAM J. Numer. Anal. 15 (1978) 953–976.

[55] I.P. Mysovskikh, Proof of the minimality of the number of nodes in the cubature formula for a hypersphere, Ž h. Vyčisl. Mat. i Mat. Fiz. 6 (4) (1966) 621–630 (in Russian), USSR Math. Comp. 6 (4) (1966) 15–27.

[56] I.P. Mysovskikh, Radon's paper on the cubature formula, Ž h. Vyčisl. Mat. i Mat. Fiz. 7(4) (1967) 889–891 (in Russian), USSR Comp. Math. 7(4) (1967) 232–236.

[57] I.P. Mysovskikh, Construction of cubature formulae and orthogonal polynomials, Ž h. Vyčisl. Mat. i Mat. Fiz. 7(1) (1967) 185–189 (in Russian), USSR Comp. Math. 7(1) (1967) 252–257.

[58] I.P. Mysovskikh, On the construction of cubature formulas with fewest nodes, Dokl. Akad. Nauk SSSR 178 (1968) 1252–1254 (in Russian), Soviet Math. Dokl. 9 (1968) 277–280 (in English).

[59] I.P. Mysovskikh, Cubature formulae and orthogonal polynomials, Ž h. Vyčisl. Mat. i Mat. Fiz. 9(2) (1969) 419–425 (in Russian), USSR Comp. Math. 9(2) (1969) 217–228.

[60] I.P. Mysovskikh, Numerical characteristics of orthogonal polynomials in two variables, Vestnik Leningrad. Univ. Mat. 25 (1970) 46–53 (in Russian).

[61] I.P. Mysovskih, A multidimensional analogon of quadrature formulae of Gaussian type and the generalised problem of Radon, Voprosy Vychisl. i Prikl. Mat., Tashkent 38 (1970) 55–69 (in Russian).

[62] I.P. Mysovskikh, On the paper: Cubature formulae and orthogonal polynomials, Ž h. Vyčisl. Mat. i Mat. Fiz. 10(2) (1970) 444–447 (in Russian), USSR Comp. Math. 10(2) (1970) 219–223.

[63] I.P. Mysovskikh, Application of orthogonal polynomials to construct cubature formulae, Ž. Vyčisl. Mat. i Mat. Fiz. 12(2) (1972) 467–475 (in Russian), USSR Comp. Math. 12(2) (1972) 228–239.

[64] I.P. Mysovskikh, Construction of cubature formulae, Voprosy Vycisl. i Prikl. Mat., Tashkent 32 (1975) 85–98 (in Russian).

[65] I.P. Mysovskikh, Orthogonal polynomials in several variables, Metody Vyčisl. 10 (1976) 26–35 (in Russian).

[66] I.P. Mysovskikh, Interpolatory Cubature Formulas, Nauka, Moscow, 1981 (in Russian), Interpolatorische Kubaturformeln, Institut für Geometrie und Praktische Mathematik der RWTH Aachen, Aachen, 1992, Bericht No. 74 (in German).

[67] I.P. Mysovskikh, V. Ja. Cernicina, The answer to a question of Radon, Dokl. Akad. Nauk SSSR 198(3) (1971) (in Russian), Soviet Math. Dokl. 12 (1971) 852–854 (in English).

[68] F. Peherstorfer, Characterization of positive quadrature formulas, SIAM J. Math. Anal. 12 (1981) 935–942.

[69] F. Peherstorfer, Characterization of quadrature formulas II, SIAM J. Math. Anal. 15 (1984) 1021–1030.

[70] R. Piessens, A. Haegemans, Cubature formulas of degree nine for symmetric planar regions, Math. Comp. 29 (1975) 810–815.

[71] R. Piessens, A. Haegemans, Cubature formulas of degree eleven for symmetric planar regions, J. Comput. Appl. Math. 1 (1975) 79–83.

[72] J. Radon, Zur mechanischen Kubatur, Monatsh. Math. 52 (1948) 286–300.

[73] G.G. Rasputin, On the question of numerical characteristics for orthogonal polynomials of two variables, Metody Vyčisl. 13 (1983) 145–154 (in Russian).

[74] G.G. Rasputin, Zur Konstruktion der Kubaturformel mit geradem agebraischen Grad und minimaler Knotenzahl, Wiss. Z. d. Päd. Hochsch. in Erfurt-Mühlhausen, Math.-Nat. Reihe 23 (1987) 158–165.

[75] J.J. Risler, Une charactérisation des idéaux des varietés algébriques réelles, Note aux CRAS, Paris 272 (1970) 522, 531.

[76] J.J. Risler, Un Théorème de Zéroes en Géometrie Algébrique et Analytique Réelles, in: Lecture Notes in Mathematics, Vol. 409, Springer, Berlin, 1974, pp. 522–531.

[77] H.J. Schmid, On cubature formulae with a minimal number of knots, Numer. Math. 31 (1978) 282–297.

[78] H.J. Schmid, Interpolatorische Kubaturformeln und reelle Ideale, Math. Z. 170 (1980) 267–280.

[79] H.J. Schmid, Interpolatorische Kubaturformeln, Diss. Math. CCXX (1983) 1–122.

[80] H.J. Schmid, On minimal cubature formulae of even degree, in: H. Brass, G. Hämmerlin (Eds.), Integration III, International Series of Numerical Mathematics, Vol. 85, Birkhäuser, Basel, 1988, pp. 216–225.

[81] H.J. Schmid, Two-dimensional minimal cubature formulas and matrix equations, SIAM J. Matrix Anal. Appl. 16 (1995) 898–921.

[82] H.J. Schmid, Y. Xu, On bivariate Gaussian cubature formulae, Proc. Amer. Math. Soc. 122 (1994) 833–841.

[83] S.L. Sobolev, Formulas of mechanical cubature on the surface of the sphere, Sibirsk. Math. 3 (1962) 769–796 (in Russian).

[84] S.L. Sobolev, Introduction to Cubature Formulae, Nauka Publishers, Moscow, 1974 (in Russian), Cubature Formulas and Modern Analysis, Gordon and Breach Science Publishers, Philadelphia, PA, 1992, xvi + 379 pp. (in English).

[85] S.L. Sobolev, V.L. Vaskevich, The Theory of Cubature Formulas, Moskau, Nauka, 1996 (in Russian), Kluwer Academic Publishers, Dordrecht, 1997 (in English).

[86] G. Sottas, G. Wanner, The number of weights of a quadrature formula, BIT 22 (1982) 339–352.

[87] A.H. Stroud, Integration formulas and orthogonal polynomials, SIAM J. Numer. Anal. 4 (1967) 381–389.

[88] A.H. Stroud, Integration formulas and orthogonal polynomials for two variables, SIAM J. Numer. Anal. 6 (1969) 222–229.

[89] A.H. Stroud, Integration formulas and orthogonal polynomials II, SIAM J. Numer. Anal. 7 (1967) 271–276.

[90] A.H. Stroud, Approximate Calculation of Multiple Integrals, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[91] P. Verlinden, Expliciete uitdrukkingen en rekursie-betrekkingen voor meerdimensionale invariante orthogonale veeltermen, Master's Thesis, Katholieke Universiteit Leuven, 1986.

[92] P. Verlinden, R. Cools, On cubature formulae of degree $4k + 1$ attaining Möller's lower bound for integrals with circular symmetry, Numer. Math. 61 (1992) 395–407.

[93] Y. Xu, On multivariate orthogonal polynomials, SIAM J. Math. Anal. 24 (1993) 783–794.

[94] Y. Xu, Common zeros of polynomials in several variables and higher dimensional quadrature, Pitman Research Notes in Mathematics Series, Vol. 312, Longman Scientific & Technical, Harlow, 1994.

[95] Y. Xu, Recurrence formulas for multivariate orthogonal polynomials, Math. Comp. 62 (1994) 687–702.

[96] Y. Xu, A characterization of positive quadrature formulae, Math. Comp. 62 (1994) 703–718.

[97] Y. Xu, Block Jacobi matrices and zeros of multivariate orthogonal polynomials, Trans. Amer. Math. Soc. 342 (1994) 855–866.

[98] Y. Xu, On orthogonal polynomials in several variables, in: Special functions, *q*-series and related topics, The Fields Institute for Research in Mathematical Sciences, Communications Series, Vol. 14, American Mathematical Society, Providence, RI, 1997, pp. 247–270.

ELSEVIER

Journal of Computational and Applied Mathematics 127 (2001) 153–171

JOURNAL OF
COMPUTATIONAL AND
APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Stopping functionals for Gaussian quadrature formulas

Sven Ehrich

*GSF-National Research Center for Environment and Health, Institute of Biomathematics and Biometrics,
Ingolstadter Landstr. 1, D-85764 Neuherberg, Germany*

Received 10 January 2000; received in revised form 22 March 2000

Abstract

Gaussian formulas are among the most often used quadrature formulas in practice. In this survey, an overview is given on stopping functionals for Gaussian formulas which are of the same type as quadrature formulas, i.e., linear combinations of function evaluations. In particular, methods based on extended formulas like the important Gauss–Kronrod and Patterson schemes, and methods which are based on Gaussian nodes, are presented and compared. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Gaussian quadrature formulas; Error estimates; Gauss–Kronrod formulas; Stopping functionals

1. Introduction

1.1. Motivation

The problem of approximating definite integrals is of central importance in many applications of mathematics. In practice, a mere approximation of an integral very often will not be satisfactory unless it is accompanied by an estimate of the error. For most quadrature formulas of practical interest, error bounds are available in the literature which use, e.g., norms of higher-order derivatives or bounds for the integrand in the complex plane. However, in many practical situations such information about the integrand is not available. In particular, automatic quadrature routines are designed such that the user only has to insert the limits of integration, a routine for computing the integrand, a tolerance for the error and an upper bound for the number of function evaluations (cf. [51,52,18 p. 418]). Functionals based on function evaluations that provide estimates for the quadrature error are called stopping functionals.

Most quadrature methods used in modern numerical software packages like those of NAG [71] and IMSL [47] are based on Gaussian (Gauss–Kronrod, Patterson) formulas. Furthermore, both numerical experience and theoretical results show the superiority of Gaussian formulas over many

PII: S0377-0427(00)00496-9

other quadrature formulas in many function classes (cf. in particular [11] and the literature cited therein). For these reasons, the problem of practical error estimates in particular for Gaussian formulas is very important, and many papers are devoted to this subject. Several stopping functionals for Gaussian quadrature formulas have been proposed in the literature and as computer algorithms. One may roughly divide these methods into two categories: (1) those based on extensions, i.e., on the addition of nodes, and (2) those based (essentially) on the nodes of the Gaussian formulas. To make this distinction more strict, in the following we say that an error estimate for a quadrature formula Q_n with n nodes is *based on extension* if the number of additional nodes is unbounded when $n \to \infty$; otherwise, we call it *essentially based on the nodes of* Q_n. Important prototypes for the first category are the Gauss–Kronrod and the Patterson methods (see Section 2). The extension strategy is also used by many practical implementations of the important Clenshaw–Curtis formulas. Very often, the extended formula is used for approximating the integral, and the difference between the two quadrature values is used to approximate the error. Such extended formulas are the state of the art, e.g., in the above-mentioned software libraries and in the QUADPACK package [85]. The second category includes the null rules and the recent Peano stopping functionals (see Section 3). Such methods deserve special attention, since function evaluations are generally considered the computationally most expensive part of quadrature algorithms.

Methods from both the categories have been used in automatic integration algorithms. In particular, many automatic integration algorithms use interval subdivision techniques where a fixed pair of a quadrature formula and a stopping functional are used to compute both a local approximation and an error estimate. Based on this information, a decision is made about further subdivision. Presently, the most important univariate general-purpose integrators over finite intervals, like the NAG routine D01AJF, use bisection strategies with pairs of Gaussian and Gauss–Kronrod formulas. The Gauss–Kronrod scheme was introduced by Kronrod in 1964 [53,54]. Kronrod's approach was, for the estimation of the error of an n-point Gaussian formula, to choose $n+1$ additional nodes for the construction of a "better" formula, i.e., a formula that has the highest possible algebraic degree of exactness using $2n+1$ function evaluations, among them the n function values that were computed for the Gaussian formula. There exist exhaustive survey papers on the Gauss–Kronrod scheme [39,67,69,74] and related quadrature formulas.

Presently, however, a general survey on stopping functionals that have been proposed for the practical (computational) estimation of the error of Gaussian quadrature formulas does not seem to exist. From a practical point of view, the most important problems seem to be the availability of the stopping functional, its computational complexity, and its quality for error estimation. The aim of this survey is to present the known methods and results with a focus on these practical aspects, and with a certain emphasis on recent developments and results. For space limitations, we restrict ourselves to the practically most important (linear) stopping functionals that are of the same type as quadrature formulas, i.e., linear combinations of function evaluations. We do not include methods which are based on error bounds from the literature using, e.g., norms of derivatives in conjunction with automatic differentiation techniques and interval analysis (cf. [17,32] for more details on this topic). Furthermore, we do not discuss stopping functionals based on other than Gauss-type formulas in this survey (cf. [34]). The results on extended Gaussian formulas are summarized in Section 2, and the stopping functionals based on Gaussian nodes are presented in Section 3. In the following three subsections, we summarize basic facts on numerical integration and Peano kernels which are necessary for the presentation in Sections 2 and 3.

1.2. Basic facts and notation

For a given nonnegative and integrable weight function ω on (a,b), $-\infty < a < b < \infty$, a quadrature formula Q_n and the corresponding remainder R_n of (algebraic) degree of exactness s are linear functionals on $C[a,b]$ defined by

$$Q_n[f] = \sum_{v=1}^{n} a_{v,n} f(x_{v,n}), \quad R_n[f] = \int_a^b \omega(x) f(x)\,\mathrm{d}x - Q_n[f],$$

$$\deg(R_n) = s \Leftrightarrow R_n[m_k] \begin{cases} = 0, & k = 0,1,\ldots,s, \\ \neq 0, & k = s+1, \end{cases} \qquad m_k(x) = x^k, \tag{1}$$

with nodes $-\infty < x_{1,n} < \cdots < x_{n,n} < \infty$ and weights $a_{v,n} \in \mathbb{R}$. Without restriction, using affine transformations, in the following we set $[a,b] = [-1,1]$ wherever not explicitly states otherwise. Furthermore, we omit the second index in $x_{v,n}$, $a_{v,n}$ whenever the meaning is clear from the context. A quadrature formula is called interpolatory if $\deg(R_n) \geqslant n-1$. The unique quadrature formula with n nodes and highest possible degree of exactness $2n-1$ is the Gaussian formula (with respect to the weight ω)

$$Q_n^{\mathrm{G}}[f] = \sum_{v=1}^{n} a_v^{\mathrm{G}} f(x_v^{\mathrm{G}}).$$

For an overview on Gaussian formulas, cf. [10,11,95] and, in particular, [37].

1.3. Principles of verified numerical integration

Numerical integration problems in practice are often of the following type. Given the limits of integration a and b, a routine for computing $f(x)$ at any $x \in (a,b)$, a tolerance ε and an upper bound N on the number of function evaluations, compute a number Q such that

$$\left| Q - \int_{-1}^{1} f(x)\,\mathrm{d}x \right| \leqslant \varepsilon \quad \text{or} \quad \frac{|Q - \int_{-1}^{1} f(x)\,\mathrm{d}x|}{\int_{-1}^{1} |f(x)|\,\mathrm{d}x} \leqslant \varepsilon, \tag{2}$$

or give an approximation based on N function values and an estimate for the absolute error which does not meet the requirement (2). Any software routine for this purpose is called *automatic integration routine* (cf. [18, Chapter 6]). In order to decide whether a particular quadrature approximation $Q = Q_n[f]$ fulfills (2), most often linear stopping functionals S_m, $m \geqslant n$, of the same type as Q_n are used (cf. [34]),

$$S_m[f] = \sum_{v=1}^{m} b_v f(y_v), \quad b_v \in \mathbb{R}, \quad -1 \leqslant y_1 < \cdots < y_m \leqslant 1,$$
$$\{x_1,\ldots,x_n\} \subseteq \{y_1,\ldots,y_m\}. \tag{3}$$

Such linear stopping functionals have a low computational complexity, in particular if $n \approx m$. Natural requirements for an error estimate are its *efficiency*, i.e., an accurate approximation should be accompanied by a small error estimate, and its *reliability*, i.e., the error estimate should not be smaller than the actual approximation error. However, it is obvious that without knowing more about f than a finite number of function values it cannot be guaranteed that

$$|R_n[f]| \leqslant |S_m[f]|. \tag{4}$$

For nonlinear stopping functionals cf. Section 3.3 as well as [57,85] and the literature cited therein. A standard method for the construction of a (linear) stopping functional S^* for Q_n is by choosing a second (reference) quadrature formula Q_l^* and then computing

$$S^* = \gamma(Q_l^* - Q_n), \tag{5}$$

with some heuristically determined constant $\gamma \in \mathbb{R}$.

1.4. Peano kernels and applications

Let L be a bounded linear functional on $C[-1,1]$ with $\deg(L) \geqslant s-1$, where

$$\deg(L) = \sup\{r \mid L[\mathbb{P}_r] = 0, \quad \mathbb{P}_r\text{: space of polynomials of degree} \leqslant r\}.$$

If $L = R_n$ is a quadrature remainder, this definition coincides with (1). For

$$f \in A_s[-1,1] := \{f \mid f^{(s-1)} \text{ is absolutely continuous in } [-1,1],\ ||f^{(s)}||_\infty < \infty\},$$

the following representation of L due to Peano is well known,

$$L[f] = \int_{-1}^{1} f^{(s)}(x) K_s(L,x)\,\mathrm{d}x,$$

where $K_s(L,\cdot)$ is the so-called Peano kernel of L of order s,

$$K_s(L,x) = \frac{1}{(s-1)!} L[(\cdot - x)_+^{s-1}], \quad u_+^{s-1} = \begin{cases} 0 & \text{for } u < 0, \\ u^{s-1} & \text{for } u \geqslant 0 \end{cases}$$

(cf. [12]). The constants

$$c_s = c_s(L) = \int_{-1}^{1} |K_s(L,x)|\,\mathrm{d}x$$

are the best possible constants in estimates of the type

$$|L[f]| \leqslant c_s ||f^{(s)}||_\infty,$$

i.e.,

$$c_s(L) = \sup\{|L[f]| \mid ||f^{(s)}||_\infty \leqslant 1\}. \tag{6}$$

The functional L is said to be positive (negative) definite of order s if the Peano kernel $K_s(L,\cdot)$ is nonnegative (nonpositive) in the interval (a,b). An important example of a (positive) definite functional is the divided difference

$$\operatorname{dvd}(t_1,\ldots,t_{s+1})[f] = \sum_{v=1}^{s+1} d_v f(t_v), \quad d_v = \prod_{\substack{\mu=1 \\ \mu\neq v}}^{s+1} (t_v - t_\mu)^{-1},$$

$$1 \leqslant t_1 < t_2 < \cdots < t_{s+1} \leqslant 1,$$

which is characterized by

$$\operatorname{dvd}(t_1,\ldots,t_{s+1})[m_k] = \begin{cases} 0, & k=0,1,\ldots,s-1, \\ 1, & k=s. \end{cases}$$

We have

$$K_s(\operatorname{dvd}(t_1,\ldots,t_{s+1}),\cdot) = \frac{1}{s!}B[t_1,\ldots,t_{s+1}](\cdot),$$

where $B[t_1,\ldots,t_{s+1}](\cdot)$ is the B-spline with respect to the knots $t_1,\ldots,t_{s+1}$, normalized by $\|B[t_1,\ldots,t_{s+1}]\|_1 = 1$ (cf. [91, Section 4.3]).

Applying Peano kernel theory to quadrature remainders $L = R_n$ is a systematic and standard way for obtaining error bounds for quadrature formulas (cf. [37, p.115]). In view of (6), the constants $c_s(L)$ can be considered as a measure for the quality of quadrature formulas for the function class $A_s[-1,1]$. Explicit or asymptotic values for these constants are known for many quadrature formulas of practical interest and many function classes (cf. in particular [10,11,83]).

2. Extended Gaussian formulas

2.1. Gauss–Kronrod formulas and Stieltjes polynomials

Let ω be nonnegative and integrable in the open interval $(-1,1)$. It is well known that the nodes of the Gaussian formula Q_n^{G} with respect to the weight ω are precisely the zeros of the nth orthonormal polynomial p_n^{ω} with respect to the weight ω (see, e.g., [97]). The following fundamental theorem gives a more general statement.

Theorem 1. *Let ω be nonnegative and integrable in $(-1,1)$. Let $-1 \leqslant x_1 < \cdots < x_n \leqslant 1$ be fixed numbers. A necessary and sufficient condition that for*

$$Q_{n,m}[f] = \sum_{v=1}^{n} A_v f(x_v) + \sum_{\mu=1}^{m} B_\mu f(\xi_\mu)$$

we have

$$\deg(R_{n,m}) \geqslant 2m+n-1 \tag{7}$$

is that simultaneously (i) *the polynomial $\prod_{\mu=1}^{m}(x-\xi_\mu)$ is orthogonal to all polynomials of degree $\leqslant m-1$ with respect to the sign changing weight $\omega\prod_{v=1}^{m}(\cdot - x_v)$, i.e.,*

$$\int_{-1}^{1} \omega(x)x^k \prod_{v=1}^{n}(x-x_v)\prod_{\mu=1}^{m}(x-\xi_\mu)\,\mathrm{d}x = 0, \quad k=0,1,\ldots,m-1, \tag{8}$$

and (ii) *that $Q_{n,m}$ is interpolatory.*

The orthogonality conditions (8) are a nonlinear system of equations for the unknown nodes $\xi_1,\ldots,\xi_m$. The weight $A_1,\ldots,A_n$, $B_1,\ldots,B_m$ are determined by the interpolation condition. Elementary examples show that the system (8) is not always uniquely solvable. In fact, if $x_1,\ldots,x_n$ are the roots of the

nth orthonormal polynomial of degree n, then for $m < n/2$ every choice of nodes $\xi_1,\dots,\xi_m$ satisfies (8), while for $n = m$ there is no choice of $\xi_1,\dots,\xi_m$ (even complex ones) such that (8) is satisfied for $k = 0$. Well-known special instances of Theorem 1 are the Gaussian formula ($n = 0$), the Radau formulas ($n = 1, x_1 = \pm 1$), and the Lobatto formula ($n = 2, x_1 = -1, x_2 = 1$). For these examples, the function $\omega\prod_{v=1}^{n}(\cdot - x_v)$ has no change of sign in $(-1,1)$. For the Gauss–Kronrod formulas, the fixed nodes are the nodes of the n-point Gaussian formula, i.e., the zeros of p_n^{ω},

$$Q_{2n+1}^{\mathrm{GK}}[f] = \sum_{v=1}^{n} A_v^{\mathrm{GK}} f(x_v^{\mathrm{G}}) + \sum_{\mu=1}^{n+1} B_\mu^{\mathrm{GK}} f(\xi_\mu^{\mathrm{K}}),$$

where $\xi_1^{\mathrm{K}},\dots,\xi_{n+1}^{\mathrm{K}}$ and $A_1^{\mathrm{GK}},\dots,A_n^{\mathrm{GK}}$, $B_1^{\mathrm{GK}},\dots,B_{n+1}^{\mathrm{GK}}$ are chosen such that (7) is satisfied with $m=n+1$. The polynomials

$$E_{n+1}(x) = c_n \prod_{\mu=1}^{n+1}(x - \xi_\mu^{\mathrm{K}}), \quad c_n \in \mathbb{R},$$

are called *Stieltjes polynomials*. These polynomials seem to appear first in a letter of T.J. Stieltjes to C. Hermite in 1894 [2]. Stieltjes conjectured that the zeros of E_{n+1}, for the Legendre weight $\omega \equiv 1$, are all real and in $(-1,1)$ for each $n \in \mathbb{N}$, and that they interlace with the zeros of the Legendre polynomial P_n, i.e., the Gaussian nodes, for all $n \in \mathbb{N}$,

$$-1 < \xi_1^{\mathrm{K}} < x_1^{\mathrm{G}} < \cdots < \xi_n^{\mathrm{K}} < x_n^{\mathrm{G}} < \xi_{n+1}^{\mathrm{K}} < 1.$$

These conjectures were proved by Szegő [96] for the wider class of weights $\omega_\lambda(x) = (1 - x^2)^{\lambda-1/2}$, $\lambda \in (0,2]$. Recent results on the location of the zeros of Stieltjes polynomials can be found in [22,30]. The Gauss–Kronrod formulas have been introduced in 1964 by Kronrod, but there are no hints that Kronrod was aware of Stieltjes' and Szegő's work at that time. The connection has been observed later by Barrucand [4] and by Mysovskih [70].

The most important weight function ω for the application of Gauss–Kronrod formulas in automatic integration is the Legendre weight. In this case, the positivity of the Gauss–Kronrod formulas was proved by Monegato in [66]. In [86] (see also [88] for a correction), Rabinowitz proved that the exact degree of precision of Q_{2n+1}^{GK} is $3n + 1$ if n is even and $3n + 2$ if n is odd. The nondefiniteness of Gauss–Kronrod formulas was proved by Rabinowitz in [88]. Results on the convergence of the interpolation processes based on the nodes of Gauss–Kronrod formulas can be found in [30,31]. For more general weight functions and other constructions of extended positive quadrature formulas, see Sections 2.4 and 2.7; cf. also the survey papers of Gautschi [39], Monegato [67,69] and Notaris [74]; for tables of nodes and weights see Piessens et al. [85] and the original work of Kronrod [53,54].

2.2. The Gauss–Kronrod stopping functional

The standard stopping functional for the Gauss–Kronrod method,

$$S_{2n+1}^{\mathrm{GK}} = Q_{2n+1}^{\mathrm{GK}} - Q_n^{\mathrm{G}} = \sum_{v=1}^{n} \tilde{d}_v f(x_v^{\mathrm{G}}) + \sum_{\mu=1}^{n+1} d_\mu^* f(\xi_\mu^{\mathrm{K}})$$

is a linear combination of point evaluation functionals which satisfies

$$S_{2n+1}^{\mathrm{GK}}[m_k] = 0, \quad k = 0,1,\dots,2n-1,$$

$$S^{\mathrm{GK}}_{2n+1}[m_{2n}] = R^{\mathrm{G}}_n[m_{2n}] = \frac{2^{2n+1}n!^4}{(2n+1)(2n)!^2},$$

where $m_k(x) = x^k$, hence

$$S^{\mathrm{GK}}_{2n+1} = \frac{2^{2n+1}n!^4}{(2n+1)(2n)!^2}\mathrm{dvd}(\xi^{\mathrm{K}}_1, x^{\mathrm{G}}_1, \xi^{\mathrm{K}}_2, \ldots, x^{\mathrm{G}}_n, \xi^{\mathrm{K}}_{n+1})$$

and

$$K_{2n}(S^{\mathrm{GK}}_{2n+1}, \cdot) = \frac{2^{2n+1}n!^4}{(2n+1)(2n)!^3}B[\xi^{\mathrm{K}}_1, x^{\mathrm{G}}_1, \ldots, x^{\mathrm{G}}_n, \xi^{\mathrm{K}}_{n+1}](\cdot).$$

Furthermore,

$$R^{\mathrm{GK}}_{2n+1} = I - Q^{\mathrm{GK}}_{2n+1} = I - Q^{\mathrm{G}}_n - S^{\mathrm{GK}}_{2n+1} = R^{\mathrm{G}}_n - S^{\mathrm{GK}}_{2n+1},$$

where $I[f] = \int_{-1}^{1} f(x)\,\mathrm{d}x$, hence

$$K_{2n}(R^{\mathrm{GK}}_{2n+1}, x) = K_{2n}(R^{\mathrm{G}}_n, x) - K_{2n}(S^{\mathrm{GK}}_{2n+1}, x).$$

Since $\deg(R^{\mathrm{GK}}_{2n+1}) \geqslant 3n+1$, we have

$$\int_{-1}^{1} K_{2n}(R^{\mathrm{GK}}_{2n+1}, x)x^k\,\mathrm{d}x = 0, \quad k = 0, 1, \ldots, n+1,$$

and therefore,

$$\int_{-1}^{1} K_{2n}(R^{\mathrm{G}}_n, x)x^k\,\mathrm{d}x = \int_{-1}^{1} K_{2n}(S^{\mathrm{GK}}_{2n+1}, x)x^k\,\mathrm{d}x, \quad k = 0, 1, \ldots, n+1, \tag{9}$$

i.e., the Peano kernel of the stopping functional reproduces the first $n+2$ moments of the $(2n)$th Peano kernel of Q^{G}_n. Moreover, $\xi^{\mathrm{K}}_1, \ldots, \xi^{\mathrm{K}}_{n+1}$ are characterized by (9). Hence, the construction of the Gauss–Kronrod stopping functional S^{GK}_{2n+1} is essentially the construction of a suitable spline function with partially free knots which approximates the $(2n)$th Peano kernel of Q^{G}_n "best" in the sense of the maximum number of reproduced moments. The connection of Gaussian quadrature formulas and moment-preserving spline approximation problems has been investigated in many papers, cf. [26,35,38,40,46,64]. Other types of approximations lead to other stopping functionals (see Section 3). Peano kernel theory provides a general and very useful framework for the construction and comparison of stopping functionals.

2.3. Gauss–Kronrod vs. Gaussian formulas

A result from [21, Corollary] states that for all $n \geqslant 1$ we have

$$c_{3n+2+\kappa}(R^{\mathrm{G}}_{2n+1}) < c_{3n+2+\kappa}(R^{\mathrm{GK}}_{2n+1}),$$

where $\kappa = 0$ if n is even and $\kappa = 1$ if n is odd, and for $n \geqslant 15$ we have

$$\frac{c_{3n+2+\kappa}(R^{\mathrm{G}}_{2n+1})}{c_{3n+2+\kappa}(R^{\mathrm{GK}}_{2n+1})} < 3^{-n+1}.$$

Asymptotically, we have

$$\lim_{n\to\infty}\left(\frac{c_{3n+2+\kappa}(R^{\mathrm{G}}_{2n+1})}{c_{3n+2+\kappa}(R^{\mathrm{GK}}_{2n+1})}\right)^{1/n} = \sqrt{\frac{6^6}{7^7}} = \frac{1}{4.2013\ldots}.$$

This relation shows that for "smooth", i.e., infinitely often differentiable functions with all derivatives uniformly bounded, Q_{2n+1}^{G} can be expected to give much better results than Q_{2n+1}^{GK}. This has also been observed in many numerical examples. A similar relation as above holds true for the error constants $c_{3n+2+\kappa-s}(R_{2n+1}^{\mathrm{GK}})$, $s \in \mathbb{N}$ independent of n. For the case $s = s(n)$, $\lim_{n\to\infty} s/n = A$, $0 \leqslant A < 1$, it has been proved in [20] that

$$\limsup_{n\to\infty} \left(\frac{c_{3n+2+\kappa-s}(R_{2n+1}^{\mathrm{G}})}{c_{3n+2+\kappa-s}(R_{2n+1}^{\mathrm{GK}})} \right)^{1/n} < 1.$$

Concerning the case $A = 1$, we have

$$\limsup_{n\to\infty} \left(\frac{c_{2n+1}(R_{2n+1}^{\mathrm{G}})}{c_{2n+1}(R_{2n+1}^{\mathrm{GK}})} \right)^{1/n} \leqslant 1,$$

but the precise value of the lim sup is presently an open problem.

A natural question is how Gauss–Kronrod and Gaussian formulas compare with respect to larger classes of "nonsmooth" functions. Let $s \in \mathbb{N}$ be independent of n. In this case we have [24]

$$\lim_{n\to\infty} \frac{c_s(R_{2n+1}^{\mathrm{G}})}{c_s(R_{2n+1}^{\mathrm{GK}})} = 1.$$

For more results on the error of Gauss–Kronrod formulas, cf. the survey papers [27,74] and the literature cited therein.

2.4. Existence of Gauss–Kronrod formulas

Simple counterexamples show that the existence of Gauss–Kronrod formulas with real nodes $\xi_1^{\mathrm{GK}}, \ldots, \xi_{n+1}^{\mathrm{GK}}$ inside the interval $[-1, 1]$ cannot be guaranteed for general ω under the assumptions of Theorem 1. In this section, an overview will be given on Gauss–Kronrod formulas for weight functions which attracted most interest in the literature. Some remarks will also be made on the existence of Kronrod extensions of Lobatto and Radau formulas. The use of Lobatto formulas and Kronrod extensions of Lobatto formulas for automatic integration has recently been suggested by Gander and Gautschi [36] in order to improve the existing automatic quadrature routines of the Matlab software package [63]. Note that for Lobatto–Kronrod formulas, n Kronrod nodes have to be chosen for $n + 1$ Lobatto nodes (including ± 1), while for the Radau–Kronrod formulas n Kronrod nodes have to be chosen for n given Radau nodes.

Following Gautschi and Notaris [41], the following properties have to be included in a systematic study of Gauss–Kronrod formulas:

(a) The nodes $x_1^{\mathrm{G}}, \ldots, x_n^{\mathrm{G}}$ and $\xi_1^{\mathrm{K}}, \ldots, \xi_{n+1}^{\mathrm{K}}$ interlace.
(b) In addition to property (a), all nodes are contained in $(-1, 1)$.
(c) In addition to property (a), all weights $A_1^{\mathrm{GK}}, \ldots, A_n^{\mathrm{GK}}$ and $B_1^{\mathrm{GK}}, \ldots, B_{n+1}^{\mathrm{GK}}$ are positive.
(d) All nodes, without necessarily satisfying (a) or (b), are real.

Monegato showed in [65] that the interlacing property of the nodes is equivalent to the positivity of the weights $B_1^{\mathrm{GK}}, \ldots, B_{n+1}^{\mathrm{GK}}$ at the additional nodes $\xi_1^{\mathrm{GK}}, \ldots, \xi_{n+1}^{\mathrm{GK}}$. This property holds for general weights ω.

As mentioned in Section 2.1, for the ultraspherical or Gegenbauer weight function

$$\omega_\lambda(x) = (1-x^2)^{\lambda-1/2}, \quad x \in (-1,1), \quad \lambda > -\tfrac{1}{2},$$

Szegő [96] has shown that properties (a) and (b) are valid for $\lambda \in (0,2]$. For $\lambda = 0$, two nodes are in ± 1. For $\lambda < 0$, Szegő gives the counterexample $n = 3$, where two nodes are outside of $[-1,1]$. Monegato [67] pointed out that for sufficiently large λ extended Gaussian formulas $Q_{2n+1}^{\mathrm{ext\,G}}$ with respect to w_λ, with $\deg(Q_{2n+1}^{\mathrm{ext\,G}}) \geqslant [2rn+l]$, $r > 1$ and l integer, and with only real nodes and positive weights, cannot exist for all $n \in \mathbb{N}$ (as Gautschi [39] and Notaris [74] mention, the proof is not correct, but can be repaired). Peherstorfer and Petras [82] recently proved that for every $\lambda > 3$, Gauss–Kronrod formulas do not exist with real nodes for all $n \in \mathbb{N}$. Gautschi and Notaris [41] investigated Gauss–Kronrod formulas for ω_λ numerically for $n = 1, 2, \ldots, 20, 24, 28, \ldots, 40$ and computed feasible regions for the parameter λ for each of the four properties. Existence results for Lobatto–Kronrod formulas for w_λ, $\lambda \in (-\frac{1}{2}, 1]$, with real nodes in $(-1,1)$ that have the interlacing property with respect to the Lobatto nodes follow from the results about the Gauss–Kronrod formulas. Monegato [69] conjectures the positivity of all quadrature weights for the Legendre weight function. A partial (asymptotic) positive answer was given in [23] for weights which are associated with nodes inside fixed subintervals of $(-1,1)$.

Stieltjes polynomials and Gauss–Kronrod formulas have been considered for the more general Jacobi weight function

$$\omega_{\alpha,\beta}(x) = (1-x)^\alpha (1+x)^\beta, \quad x \in (-1,1), \quad \alpha, \beta > -1.$$

Rabinowitz [87] showed that (b) is not valid for $\alpha = -\frac{1}{2}, -\frac{1}{2} < \beta \leqslant \frac{3}{2}$ $(\beta \neq \frac{1}{2})$ and $-\frac{1}{2} < \alpha \leqslant \frac{3}{2}$, $\beta = -\frac{1}{2}$ $(\alpha \neq \frac{1}{2})$ for even n and for $\alpha = -\frac{1}{2}, \frac{3}{2} \leqslant \beta \leqslant \frac{5}{2}$ and $\frac{3}{2} < \alpha \leqslant \frac{5}{2}$, $\beta = -\frac{1}{2}$ for odd n. Monegato [69] derived the relations

$$E_{n+1}^{\alpha,-1/2}(2t^2-1) = tE_{2n+1}^{\alpha,\alpha}(t) - d_n,$$

$$E_{n+1}^{\alpha,-1/2}(2t^2-1) = (-1)^{n+1} E_{n+1}^{1/2,\alpha}(1-2t^2) E_{2n+2}^{\alpha,\alpha}(t),$$

where d_n is an explicitly given constant and $E_{n+1}^{\alpha\beta}$ is the (suitable normalized) Stieltjes polynomial associated with the weight $\omega_{\alpha,\beta}$. Hence, for $\alpha = \frac{1}{2}$ and $-\frac{1}{2} < \beta \leqslant \frac{3}{2}$ as well as $\beta = \frac{1}{2}$ and $-\frac{1}{2} < \alpha \leqslant \frac{3}{2}$, results can be carried over from the ultraspherical case. Gautschi and Notaris [41] extended their numerical investigations to the Jacobi weight function and determined feasible regions in the (α, β)-plane for the validity of the four properties. It is well known that the left (right) Radau formula for the weight $\omega_{\alpha,\beta}$ is connected with the Gaussian formula for the weight $\omega_{\alpha+1,\beta}$ $(\omega_{\alpha,\beta+1})$. Numerical results in [3] indicate that Radau–Kronrod formulas for the Legendre weight have positive weights and hence the interlacing property (see [69]).

The most elementary cases of the Jacobi weight function are those with $|\alpha| = |\beta| = \frac{1}{2}$. For the Chebyshev weight functions of the first kind with $\alpha = \beta = -\frac{1}{2}$ and of the second kind with $\alpha = \beta = \frac{1}{2}$, we have the well-known identity

$$2T_{n+1}U_n(x) = U_{2n+1}(x).$$

Therefore, the Stieltjes polynomials are identical to $(1-x^2)U_{n-1}(x)$ in the first case and to T_{n+1} in the second case, and the degree of exactness is $4n-1$ in the first case and $4n+1$ in the second. The Gauss–Kronrod formula for the Chebyshev weight of the first kind is therefore the Lobatto formula with $2n+1$ nodes for the same weight function, and the Gauss–Kronrod formula Q_{2n+1}^{GK} for the Chebyshev weight of the second kind is identical to the Gaussian formula Q_{2n+1}^{G} for this weight. For $\alpha = \frac{1}{2}$, $\beta = -\frac{1}{2}$ ($\alpha = -\frac{1}{2}$, $\beta = \frac{1}{2}$), the Gauss–Kronrod formula is the $(2n+1)$-point left (right) Radau formula (see [69]).

In case of the Laguerre weight $\omega(x)=x^{\alpha}\mathrm{e}^{-x}$, $-1 < \alpha \leqslant 1$, $x \in [0,\infty)$, Kahaner and Monegato [48] showed that no Kronrod extension with real nodes and positive weights exists for $n > 23$, in the case $\alpha = 0$ even for $n > 1$. Furthermore, Monegato [67] proved that extended Gaussian formulas $Q_{2n+1}^{\mathrm{ext\,G}}$ with respect to this weight, with $\deg(Q_{2n+1}^{\mathrm{ext\,G}}) \geqslant [2rn+l]$, $r > 1$ and l integer, and with only real nodes and positive weights do not exits for n sufficiently large. In case of the Hermite weight function $\omega(x)=\mathrm{e}^{-x^2}$, real positive Kronrod extensions do not exist for all $n \in \mathbb{N}\backslash\{1,2\}$ (cf. [48]). For $n=4$, all nodes are real but two weights are negative. Numerical examples suggest that for all $n \in \mathbb{N}\backslash\{1,2,4\}$ complex nodes occur, but this problem is still open. Notaris [73] uses the nonexistence results for Gauss–Laguerre and Gauss–Hermite quadrature formulas and limit relations for the ultraspherical and Hermite resp. Jacobi and Laguerre polynomials in order to deduce nonexistence results for ultraspherical and Jacobi weight functions. Monegato [69] showed the existence of real Gauss–Kronrod formulas with nodes in $(-1,1)$ for the weight function

$$\omega^{(\mu)}(x) = \frac{(1-x^2)^{1/2}}{1-\mu x^2}, \quad -\infty < \mu \leqslant 1.$$

Furthermore, the Kronrod nodes interlace with the Gaussian nodes. We have $\deg(Q_{2n+1}^{\mathrm{GK}}) = 4n-1$, if $n > 1$ and $\mu \neq 0$, respectively $\deg(Q_{2n+1}^{\mathrm{GK}}) = 4n+1$ if $n > 1$, $\mu = 0$, and $\deg(Q_3^{\mathrm{GK}}) = 5$. The weights are always positive, see Gautschi and Rivlin [44]. Gautschi and Notaris [42], Notaris [72] and Peherstorfer [79] considered the Bernstein–Szegő weight $\omega(x)=\sqrt{1-x^2}/s_m(x)$, where s_m is a positive polynomial on $[-1,1]$ of degree m and prove that for all $n \in \mathbb{N}$ the Gauss–Kronrod formulas exists with nodes in $(-1,1)$ that interlace with the Gaussian nodes and with positive weights. Gautschi and Notaris [43] generalized these results to weights for which the corresponding orthogonal polynomials satisfy a three-term recurrence relation whose coefficients $a_n \in \mathbb{R}$ and $b_n > 0$, $n \in \mathbb{N}$, are constant above a fixed index $l \in \mathbb{N}$ $a_n = \alpha$ and $b_n = \beta$ for $n \geqslant l$. More precisely, for such weights and all $n \geqslant 2l-1$, the Gauss–Kronrod formula Q_{2n+1}^{GK} has the interlacing property, and all its weights are positive. Moreover, $\deg(R_{2n+1}^{\mathrm{GK}}) \geqslant 4n-2l+2$ for $n \geqslant 2l-1$. If additionally the support of ω is contained in $[a,b]$, where $a=\alpha-2\sqrt{\beta}$ and $b=\alpha+2\sqrt{\beta}$, then all Kronrod nodes are contained in $[a,b]$ for $n \geqslant 2l-1$.

Peherstorfer [79–81] investigated properties (a)–(d) for more general classes of weight functions. In particular, for sufficiently large n, Peherstorfer proved these properties for all weight functions ω which can be represented by

$$\omega(x) = \sqrt{1-x^2}D(\mathrm{e}^{\mathrm{i}\phi})^2, \quad x = \cos\phi \;\; \phi \in [0,\pi],$$

where D is a real and analytic function with $D(z) \neq 0$ for $|z| \leqslant 1$ (cf. [80]; see also [81] for corrections).

Gautschi and Notaris pointed out the relation of Gauss–Kronrod formulas for the weight function

$$\omega_{\gamma}^{(\alpha)} = |x|^{\gamma}(1-x^2)^{\alpha}, \qquad \alpha > -1, \;\; \gamma > -1, \;\; x \in (-1,1),$$

to those for the Jacobi weight $\omega_{\alpha,(\gamma+1)/2}$. Numerical results support the conjecture of Caliò et al. [14] that Gauss–Kronrod formulas for the weight $\omega(x) = -\ln x$, $x \in (0,1)$, exist for all $n \geqslant 1$ and satisfy (a)–(d). Gautschi [39, p. 40] conjectures similar results for the more general weight

$$\omega(x,\alpha) = -x^{\alpha}\ln x, \quad x \in (0,1), \;\; \alpha = \pm\tfrac{1}{2} \; (\alpha \neq \tfrac{1}{2} \text{ if } n \text{ is even}).$$

Li [60] investigates Kronrod extensions of generalized Radau and Lobatto formulas, which use function values and, in the first case, first derivatives at one boundary and in the second case first derivatives at both boundaries. Explicit expressions are proved for the Stieltjes polynomials with respect to Jacobi weight functions with $|\alpha| = |\beta| = \frac{1}{2}$ and for the weights at the interval boundaries.

A three-point Gauss–Kronrod formula for the discrete weight

$$\omega(x) = \sum_{j=0}^{\infty} \frac{\mathrm{e}^{-x} x^j}{j!} \delta(x - j), \quad x \in (0, \infty),$$

has been constructed a long time before Kronrod's work by Ramanujan in his second notebook [90].[1] As Askey reports [1], Ramanujan computed several Gaussian formulas in this notebook, and his motivation for the Gauss–Kronrod formula for this weight was that the nodes of the three-point Gaussian formula could not be found as simple expressions.

2.5. Patterson extensions

Patterson [75] computed sequences of embedded quadrature formulas by iterating Kronrod's method. The resulting formulas, now called Patterson extensions, are used, e.g., in the NAG routing D01AHF [71]. More precisely, Patterson extensions are quadrature formulas of the type

$$\int_{-1}^{1} p(x) f(x)\, \mathrm{d}x \approx Q_{2^i(n+1)-1}[f] = \sum_{v=1}^{n} \alpha_{iv} f(x_v^{\mathrm{G}}) + \sum_{\rho=1}^{i} \sum_{v=1}^{2^{\rho-1}(n+1)} \beta_{i\rho v} f(\xi_v^{\rho}),$$

$i \geqslant 1$, where $x_1^{\mathrm{G}}, \ldots, x_n^{\mathrm{G}}$ are the nodes of a Gaussian formula, the nodes of $Q_{2^{i-1}(n+1)-1}$ are used by $Q_{2^i(n+1)-1}$, and the free nodes are chosen according to Theorem 1. Hence, the algebraic accuracy of $Q_{2^i(n+1)-1}$ is at least $3 \cdot 2^{i-1}(n+1) - 2$. Very little is known about the existence and positivity of Patterson extensions for $p \equiv 1$ and beyond Kronrod's extension. Numerical examples in [77] show that nodes outside the integration interval can occur. The only two weights for which general existence results are available are the Chebyshev weight of the second kind, for which Patterson formulas are identical to Gaussian formulas, and weight functions and Bernstein–Szegő type [79]. Tables of sequences of Kronrod–Patterson formulas have been given in [75,85]. Computational investigations on the existence of the first Patterson extension are discussed in [89]. Patterson extensions recently received some attention in the context of sparse grid methods of multivariate numerical integration [45].

Let $A_{i,j,k}$ be the weight associated with the ith node (for nodes ordered in increasing magnitude) which is added in the jth Patterson extension in a (interpolatory) formula which results from a total of $k \geqslant j$ extensions. Krogh and Van Snyder [50] observed that $A_{i,j,k} \approx 0.5\, A_{i,j,k-1}$, and used this property for representing Patterson extensions with fewer function values. Laurie [56] constructed sequences of stratified nested quadrature formulas of the type

$$Q_{(k)}[f] = \theta Q_{(k-1)}[f] + \sum_{i=0}^{n_{k-1}} A_{i,k} f(x_{i,k}), \quad 0 < \theta < 1.$$

[1] I thank Prof. Askey for pointing out this reference.

Here, only the value $Q_{(k-1)}$ has to be stored from step $k-1$ to step k. Laurie computed sequences of embedded quadrature formulas, for $\theta=\frac{1}{2}$, and with interlacing nodes in $(-1,1)$ and positive weights. Hybrid methods are discussed in [78].

2.6. Anti-Gaussian formulas

Laurie [58] introduced the stratified pair of quadrature formulas

$$L_{2n+1}=\tfrac{1}{2}Q_n^{\mathrm{G}}[f]+\sum_{i=1}^{n+1}\frac{a_i}{2}f(\xi_i), \tag{10}$$

where the ξ_i, a_i are the nodes of the "anti-Gaussian" formula

$$Q_{n+1}^{\mathrm{AG}}[f]=\sum_{i=1}^{n+1}a_i f(\xi_i),$$

defined by the conditions

$$R_{n+1}^{\mathrm{AG}}[m_k]=-R_n^{\mathrm{G}}[m_k],\quad k=0,1,\ldots,2n+1.$$

An equivalent condition for symmetric formulas is

$$R_{n+1}^{\mathrm{AG}}[m_{2n}]=-R_n^{\mathrm{G}}[m_{2n}]. \tag{11}$$

The nodes of Q_{n+1}^{AG} are real for every integrable ω, even if (a,b) is unbounded, and they interlace with the Gaussian nodes. For the ultraspherical weight ω_λ, also ξ_1 and ξ_{n+1} are inside $(-1,1)$. There are Jacobi weights for which ξ_1 or ξ_{n+1} are outside $(-1,1)$. Laurie [58] proposes the stopping functional

$$S_{2n+1}^{\mathrm{AG}}[f]=\tfrac{1}{2}(Q_{n+1}^{\mathrm{AG}}[f]-Q_n^{\mathrm{G}}[f]) \tag{12}$$

for the estimation of the error of L_{2n+1}. This is a multiple of a divided difference of order $2n$. In particular, for the Legendre weight we have

$$S_{2n+1}^{\mathrm{AG}}=\frac{2^{2n+1}}{2n+1}\frac{n!^4}{(2n)!^2}\operatorname{dvd}(\xi_1,x_1,\xi_2,x_2,\ldots,x_n,\xi_{n+1}).$$

The Lobatto formula, for the Legendre weight "almost" satisfies (11) (cf. [10, p. 149]),

$$R_{n+1}^{\mathrm{L}}[m_{2n}]=-\left(1+\frac{1}{n}\right)R_n^{\mathrm{G}}[m_{2n}].$$

2.7. Other extensions of Gaussian formulas

"Suboptimal" Kronrod extensions Q_{2n+1}^r have been considered for weight functions where Gauss–Kronrod formulas do not exist with real nodes and positive weights, in particular, for the Laguerre and Hermite weight functions (cf. Begumisa and Robinson [6]). Here, using Theorem 1, given the n Gaussian nodes, one chooses $n+1$ additional real nodes such that the degree of the formula is $\deg(R_{2n+1}^r)\geqslant 3n+1-r$, with r as small as possible and such that all weights are positive. Another strategy is the extension by more than $n+1$ nodes ("Kronrod-heavy"); (see [39,49,68]). In terms of moment-preserving spline approximation (see Section 2.2), the first method is based on reproducing less moments, while for the second method more spline knots are introduced.

Kronrod extensions of Turàn type are considered in [7,59,92]. Smith [93,94] considers Kronrod extensions that use high-order derivatives at ±1 but only function values in the interior of the integration interval. Stieltjes polynomials and Gauss–Kronrod formulas on the semicircle have been considered in [15,16]. Kronrod extensions of Wilf-type formulas have been considered by Engels et al. [33]. Rabinowitz [87] considers Gauss–Kronrod-type formulas for the computation of Cauchy principal value integrals. For $\omega \equiv 1$, the Stieltjes polynomial has a double zero in the center of the interval of integration, and hence a derivative value is needed for computing the Gauss–Kronrod formulas in this case.

3. Stopping functionals based on Gaussian nodes

As already mentioned in Section 1, in practice, most often the Kronrod scheme is used "backwards", i.e., the $(2n+1)$-point Gauss–Kronrod formula gives the quadrature value, and the error estimate is based on a comparison with the n-point Gaussian formula. As Laurie points out in [55, p. 427], "viewed from this angle, it becomes somewhat mystifying why the Kronrod rule should have been singled out as a candidate for the parenthood of subset rules. Could the $(2n+1)$-point Gaussian rule not equally well (or even better) have been used?" As discussed in detail in Section 2.3, the $(2n+1)$-point Gaussian formula often gives better results than the $(2n+1)$-point Gauss–Kronrod formula and is a promising candidate, in particular, for automatic integration, if suitable stopping functionals are available. Unlike in Kronrod's approach, several authors considered methods for estimating the error of quadrature formulas essentially without extra function evaluations, i.e., on the basis of the function values that have been computed for the quadrature formula. Most of these stopping functionals can be represented by linear combinations of divided differences (see Section 1.4; cf. also [34,55,61]). As for the Gauss–Kronrod formulas (see Section 2.2), a natural construction principle is the approximation of a Peano Kernel of Q_n^{G} by a Peano kernel of a suitable divided difference (see Section 3.4).

3.1. Successive deletion of alternate nodes

Patterson [76] considered sets of quadrature formulas which are derived from a fixed Gaussian or Lobatto formula with 2^r+1 nodes, $r \in \mathbb{N}$, by successively deleting alternate points from the preceding subset. The interpolatory formulas on these sets of nodes are hence nested by definition. Furthermore, numerical results show that all formulas based on the Gaussian formulas Q_{33}^{G} and Q_{65}^{G} and on the Lobatto formula Q_{65}^{L} are positive (see [76]).

3.2. Dropping the midpoint

Berntsen and Espelid [8] constructed a reference formula Q_{2n}^{BE} for the Gaussian formula Q_{2n+1}^{G} by dropping the node $x_{n+1,2n+1}^{\mathrm{G}} = 0$. Hence, we have

$$S_{2n+1}^{\mathrm{BE}} = Q_{2n+1}^{\mathrm{G}} - Q_{2n}^{\mathrm{BE}} = (-1)^n \frac{2^{4n+1} n!^2 (2n)!}{(4n+1)!} \,\mathrm{dvd}(x_1^{\mathrm{G}}, \ldots, x_{2n+1}^{\mathrm{G}}).$$

Numerical results in favor of this stopping functional are given in [55]. In Section 3.4, a stopping functional will be given which gives smaller estimates (by $O(1/n)$) but which are still guaranteed error bounds for functions whose $(2n)$th derivative does not change sign.

3.3. Null rules

Linear combinations of kth divided differences are often called null rules (of degree $k-1$) [9,57,62]. For $n+1$ nodes, the linear space of null rules of degree $\geqslant n-m$ has dimension m. For any inner product on $\mathbb{R}^{n+1}$, a unique orthonormal basis of null rules can be constructed for this space. In [9], the standard inner product $(\boldsymbol{a},\boldsymbol{b}) = \sum_{i=1}^{n+1} a_i b_i$ is used for null rules $N_{n+1}[f] = \sum_{i=1}^{n+1} a_i f(x_i)$ and $\tilde{N}_{n+1}[f] = \sum_{i=1}^{n+1} b_i f(x_i)$, with $\boldsymbol{a} = (a_1,\dots,a_{n+1})$ and $\boldsymbol{b} = (b_1,\dots,b_{n+1})$. Laurie [57] considers the inner product

$$(\boldsymbol{a},\boldsymbol{b})_W = \boldsymbol{a}^{\mathrm{T}} \boldsymbol{W} \boldsymbol{b}, \quad \boldsymbol{W} = \operatorname{diag}[w_i],$$

using the (nonzero) weights w_i of a positive quadrature formula $Q_{n+1}[f] = \sum_{i=1}^{n+1} w_i f(x_i)$. Denoting the monic orthogonal polynomials with respect to the discrete inner product $(f,g) = \sum_{i=1}^{n+1} w_i f(x_i) g(x_i)$ by p_j, $j=0,1,\dots,n$, the null rules $(Q_{n+1}[p_k^2])^{-1/2} Q[p_k f]$, $k=0,1,\dots,n$, are mutually orthonormal with respect to the inner product $(\cdot,\cdot)_{W^{-1}}$. These null rules are used in [57] to construct actual approximating polynomials f_d and in turn for nonlinear error estimates which are based on approximating $\int_{-1}^{1} |f(x) - f_d(x)|\,\mathrm{d}x$.

3.4. Peano stopping functionals

Using the notation from Section 1.4, most stopping functionals S_m used in practice satisfy $c_s(R_n) < c_s(S_m)$ for special values of s. This inequality implies $\deg(S_m) \geqslant s-1$. A stronger condition can be given using Peano kernels,

$$|K_s(R_n,x)| \leqslant K_s(S_m,x) \quad \text{for every } x \in (a,b). \tag{13}$$

For stopping functionals based on Peano kernel theory, cf. [34] and the literature cited therein. Every functional S_m of the type (3) which satisfies (13) is called a (s,m) *Peano stopping functional for the quadrature formula* Q_n. A restriction for Peano stopping functionals is that the endpoints ± 1 have to be among the nodes of S_m (see [34,61]). An (s,m) Peano stopping functional S_m^{opt} is called optimal for Q_n if

$$c_s(S_m^{\mathrm{opt}}) = \min\{c_s(S_m) \mid S_m \text{ is an } (s,m) \text{ Peano stopping functional for } Q_n\}.$$

In view of (13), the construction of optimal Peano stopping functionals is a problem of best one-sided approximation by spline functions. For every Q_n with $\deg(R_n) \geqslant s-1$ and fixed nodes $y_1,\dots,y_m$ there exists a unique optimal (s,m) Peano stopping functional (cf. [28]). Condition (13) implies that $K_s(S_m,\cdot)$, $K_s(S_m+R_n,\cdot)$ and $K_s(S_m-R_n,\cdot)$ are nonnegative in $(-1,1)$, i.e., S_m, S_m-R_n and S_m+R_n are positive definite of order s. Hence, definiteness criteria are important for the construction of Peano stopping functionals (see [34] and in particular [13]); several algorithms are compared in [5].

A characteristic property of (s,m) Peano stopping functionals is that the inequality (4) is guaranteed for all

$$f \in A_s^+[-1,1] = \{f \in A_s[-1,1], f^{(s)} \text{ has no change of sign in } [-1,1]\}.$$

This feature seems particularly attractive for automatic integration routines. Assuming that a given function f is s times differentiable, its sth derivative may often have only a finite number of changes of sign. Hence, after sufficiently many recursion steps of an interactive automatic quadrature routine, “most” of the resulting subintervals $[a_i, b_i]$ will not contain a change of sign of $f^{(s)}$, such that $f \in A_s^+[a_i, b_i]$. Hence, in such subintervals, reliable error bounds can be obtained without explicit knowledge of $f^{(s)}$.

In view of the well-known definiteness of the Gaussian and Lobatto formulas,

$$S_{2n+1} = \tfrac{1}{2}(Q_{n+1}^{\mathrm{L}} - Q_n^{\mathrm{G}})$$

is a $(2n, 2n+1)$ stopping functional for the quadrature formula

$$Q_{2n+1} = \tfrac{1}{2}(Q_{n+1}^{\mathrm{L}} + Q_n^{\mathrm{G}}).$$

In general, pairs of positive and negative definite formulas lead to analogous constructions of Peano stopping functionals (see also [19] for examples). However, from the point of view of practical calculations, most interesting are stopping functionals for fixed quadrature formulas. In [29], the following optimal $(n, n+1)$ Peano stopping formula has been constructed for the Lobatto formula Q_{n+1}^{L},

$$S_{n+1}^{\mathrm{L}} = \frac{\sqrt{\pi}}{2^{n-1}} \frac{\Gamma(n)}{\Gamma(n+1/2)} \operatorname{dvd}(x_1^{\mathrm{L}}, \ldots, x_{n+1}^{\mathrm{L}}).$$

A $(n+1, n+2)$ Peano stopping functional for the Gaussian formula Q_n^{G} is given by [29]

$$S_{n+2}^{\mathrm{G}} = \frac{1}{2^{n-1}} \frac{\Gamma(n)\Gamma(n/2+1)}{\Gamma(n+1/2)\Gamma(n/2+3/2)} \operatorname{dvd}(-1, x_1^{\mathrm{G}}, \ldots, x_n^{\mathrm{G}}, 1). \tag{14}$$

This stopping functional gives both guaranteed inclusions for $f \in A_s^+[-1, 1]$ and tight bounds, in particular, tighter than (12) and tighter than S_{2n+1}^{BE} in Section 3.2,

$$\frac{c_{2n}(S_{2n+1}^{\mathrm{G}})}{c_{2n}(S_{2n+1}^{\mathrm{AG}})} \leqslant \frac{C}{n}, \quad \frac{c_{2n}(S_{2n+1}^{\mathrm{G}})}{c_{2n}(S_{2n+1}^{\mathrm{BE}})} \leqslant \frac{C}{n}, \quad C \neq C(n).$$

A $(2, n)$ Peano stopping functional for Q_n^{G} which is based only on the nodes $\{-1, 0, 1\}$ can be found in [84],

$$S^P = \frac{2\pi^2}{3(2n+1)^2} \operatorname{dvd}(-1, 0, 1).$$

Since $\xi_1^{\mathrm{K}} > -1$ and $\xi_{n+1}^{\mathrm{K}} < 1$, the Gauss–Kronrod stopping functional S_{2n+1}^{GK} is no $(s, 2n+1)$ Peano stopping functional for any $s \in \mathbb{N}$. In [25], a $(2n+2, 2n+3)$ Peano stopping functional has been constructed for the Gauss–Kronrod formula Q_{2n+1}^{GK},

$$S_{2n+3} = \frac{1}{2^{2n}} \frac{c_n}{n+1} \operatorname{dvd}(-1, \xi_1^{\mathrm{K}}, x_1^{\mathrm{G}}, \xi_2^{\mathrm{K}}, \ldots, x_n^{\mathrm{G}}, \xi_{n+1}^{\mathrm{K}}, 1),$$

where $c_n = (\sqrt{\pi}\sqrt{(6n+3)/(2n+5)})(1 + \sqrt{3\pi}(2\sqrt{n+2} - 2)^{-1})$.

4. Conclusion

In this survey, we gave an overview on practical error estimates for Gaussian formulas that are of the same type as quadrature formulas, i.e., linear combinations of function values. Stopping functionals based both on extended Gaussian formulas and on Gaussian nodes are linear combinations of divided differences. Peano kernels are a very useful tool for the construction and for the comparison of both quadrature formulas and stopping functionals. In many cases, the construction of a stopping functional is essentially the construction of a suitable spline function that approximates best, in a given sense, a Peano kernel of the Gaussian formula (typically the highest-order Peano kernel). The sense of "best" governs the type of the stopping functional: moment-preserving approximation leads to Gauss–Kronrod formulas, while one-sided approximation leads to Peano stopping functionals. Other types of approximations may be applied in many situations, e.g., on infinite intervals, where Gauss–Kronrod formulas are not available. We shall discuss such methods elsewhere.

Acknowledgements

I thank Prof. Walter Gautschi and an unknown referee for the careful reading of the manuscript.

References

[1] R. Askey, Gaussian quadrature in Ramanujan's second notebook, Proc. Indian Acad. Sci. (Math. Sci.) 104 (1994) 237–243.

[2] B. Baillaud, H. Bourget, Correspondence d'Hermite et de Stieltjes I,II, Gauthier–Villars, Paris, 1905.

[3] P. Baratella, Un' estensione ottimale della formula di quadratura di Radau, Rend. Sem. Mat. Univ. Politecn. Torino 37 (1979) 147–158.

[4] P. Barrucand, Intégration numérique, abscisse de Kronrod–Patterson et polynômes de Szegő, C.R. Acad. Sci. Paris, Ser. A 270 (1970) 147–158.

[5] C. Becker, Peano–Stoppregeln zur numerischen Approximation linearer Funktionale: Theoretische Entwicklung und praktische Implementierung universeller Algorithmen, Diploma Thesis, Univ. Hildesheim, 1996.

[6] A. Begumisa, I. Robinson, Suboptimal Kronrod extension formulas for numerical quadrature, Numer. Math. 58 (8) (1991) 807–818.

[7] A. Bellen, S. Guerra, Su alcune possibili estensioni delle formule di quadratura gaussiane, Calcolo 19 (1982) 87–97.

[8] J. Berntsen, T.O. Espelid, On the use of Gauss quadrature in adaptive automatic integration schemes, BIT 24 (1984) 239–242.

[9] J. Berntsen, T.O. Espelid, Error estimation in automatic quadrature rules, ACM Trans. Math. Software 7 (1991) 233–252.

[10] H. Brass, Quadraturverfahren, Vandenhoeck und Ruprecht, Göttingen, 1977.

[11] H. Brass, J.-W. Fischer, K. Petras, The Gaussian quadrature method, Abh. Braunschweig. Wiss. Ges. 47 (1997) 115–150.

[12] H. Brass, K.-J. Förster, On the application of the Peano representation of linear functionals in numerical analysis, in: G.V. Milovanović (Ed.), Recent Progress in Inequalities, Kluwer Academic Publishers, Dordrecht, 1998, pp. 175–202.

[13] H. Brass, G. Schmeisser, Error estimates for interpolatory quadrature formulae, Numer. Math. 37 (1981) 371–386.

[14] F. Calió, W. Gautschi, E. Marchetti, On computing Gauss–Kronrod formulae, Math. Comp. 47 (1986) 639–650.

[15] F. Calió, E. Marchetti, On zeros of complex polynomials related to particular Gauss–Kronrod quadrature formulas, Facta Univ. Ser. Math. Inform. 7 (1992) 49–57.

[16] F. Calió, E. Marchetti, Complex Gauss–Kronrod integration rules for certain Cauchy principal value integrals, Computing 50 (1993) 165–173.

[17] G.F. Corliss, L.B. Rall, Adaptive, self-validating numerical quadrature, SIAM J. Sci. Statist. Comput. 8 (1987) 831–847.

[18] P. Davis, P. Rabinowitz, Methods of Numerical Integration, 2nd Edition, Academic Press, New York, 1984.

[19] L. Derr, C. Outlaw, D. Sarafyan, Upgraded Gauss and Lobatto formulas with error estimation, J. Math. Anal. Appl. 106 (1985) 120–131.

[20] S. Ehrich, Einige neue Ergebnisse zu den Fehlerkonstanten der Gauss–Kronrod–Quadraturformel, Z. Angew. Math. Mech. 73 (1993) T882–T886.

[21] S. Ehrich, Error bounds for Gauss–Kronrod quadrature formulae, Math. Comp. 62 (1994) 295–304.

[22] S. Ehrich, Asymptotic properties of Stieltjes polynomials and Gauss–Kronrod quadrature formulae, J. Approx. Theory 82 (1995) 287–303.

[23] S. Ehrich, Asymptotic behaviour of Stieltjes polynomials for ultraspherical weight functions, J. Comput. Appl. Math. 65 (1995) 135–144.

[24] S. Ehrich, A note on Peano constants of Gauss–Kronrod quadrature schemes, J. Comput. Appl. Math. 66 (1996) 177–183.

[25] S. Ehrich, Practical error estimates for the Gauss–Kronrod quadrature formula, in: G. Alefeld, J. Herzberger (Eds.), Numerical Methods and Error Bounds, Akademie-Verlag, Berlin, 1996, pp. 74–79.

[26] S. Ehrich, On orthogonal polynomials for certain non-definite linear functionals, J. Comput. Appl. Math. 99 (1998) 119–128.

[27] S. Ehrich, Stieltjes polynomials and the error of Gauss–Kronrod quadrature formulas, in: W. Gautschi, G. Golub, G. Opfer (Eds.), Applications and Computation of Orthogonal Polynomials, Proc. Conf. Oberwolfach, International Series Numerical Mathematics, Vol. 131, Birkhäuser, Basel, 1999, pp. 57–77.

[28] S. Ehrich, K.-J. Förster, On exit criteria in quadrature using Peano kernel inclusions, Part I: Introduction and basic results, Z. Angew. Math. Mech. 75(SII) (1995) S625–S626.

[29] S. Ehrich, K.-J. Förster, On exit criteria in quadrature using Peano kernel inclusions, Part II: Exit criteria for Gauss type formulas, Z. Angew. Math. Mech. 75(SII) (1995) S627–S628.

[30] S. Ehrich, G. Mastroianni, Stieltjes polynomials and Langrange interpolation, Math. Comp. 66 (1997) 311–331.

[31] S. Ehrich, G. Mastroianni, Weighted convergence of Langrange interpolation based on Gauss–Kronrod nodes, J. Comput. Anal. Appl. 2 (2000) 129–158.

[32] M.C. Eiermann, Automatic, guaranteed integration of analytic functions, BIT 29 (1989) 270–282.

[33] H. Engels, B. Ley-Knieper, R. Schwelm, Kronrod–Erweiterungen Wilf'scher Quadraturformeln, Mitt. Math. Sem. Giessen 203 (1991) 1–15.

[34] K.-J. Förster, A survey of stopping rules in quadrature based on Peano kernel methods, Suppl. Rend. Circ. Mat. Palermo, Ser. II 33 (1993) 311–330.

[35] M. Frontini, W. Gautschi, G.V. Milovanović, Moment-preserving spline approximation on finite intervals, Numer. Math. 50 (1987) 503–518.

[36] W. Gander, W. Gautschi, Adaptive quadrature – revisited, BIT 40 (2000) 84–101.

[37] W. Gautschi, A survey of Gauss–Christoffel quadrature formulae, in: P.L. Butzer, F. Fehér (Eds.), E.B. Christoffel, Birkhäuser, Basel, 1981, pp. 72–147.

[38] W. Gautschi, Discrete approximations to spherically symmetric distributions, Numer. Math. 44 (1984) 53–60.

[39] W. Gautschi, Gauss–Kronrod quadrature – a survey, in: G.V. Milovanović (Ed.), Numerical Methods of Approximation Theory III, Univ, Niš, 1988, pp. 39–66.

[40] W. Gautschi, G.V. Milovanović, Spline approximations to spherically symmetric distributions, Numer. Math. 49 (1986) 111–121.

[41] W. Gautschi, S.E. Notaris, An algebraic study of Gauss–Kronrod quadrature formulae for Jacobi weight functions, Math. Comp. 51 (1988) 231–248.

[42] W. Gautschi, S.E. Notaris, Gauss–Kronrod quadrature formulae for weight functions of Bernstein–Szegő type, J. Comput. Appl. Math. 25 (1989) 199–224 (Errata: ibid. 27 (1989) 429).

[43] W. Gautschi, S.E. Notaris, Stieltjes polynomials and related quadrature formulas for a class of weight functions, Math. Comp. 65 (1996) 1257–1268.

[44] W. Gautschi, T.J. Rivlin, A family of Gauss–Kronrod quadrature formulae, Math. Comp. 51 (1988) 749–754.

[45] T. Gerstner, M. Griebel, Numerical integration using sparse grids, Numer. Algorithms 18 (3,4) (1998) 209–232.

[46] L. Gori, E. Santi, Moment-preserving approximations: a monospline approach, Rend. Mat. Appl. 12 (7) (1992) 1031–1044.

[47] IMSL Fortran Numerical Libraries, Version 3.0 (Visual Numerics, 1998).

[48] D.K. Kahaner, G. Monegato, Nonexistence of extended Gauss–Laguerre and Gauss–Hermite quadrature rules with positive weights, Z. Angew. Math. Phys. 29 (1978) 983–986.

[49] D.K. Kahaner, J. Waldvogel, L.W. Fullerton, Addition of points to Gauss–Laguerre quadrature formulas, SIAM J. Sci. Statist. Comput. 5 (1984) 42–55.

[50] F. Krogh, W. Van Snyder, Algorithm 699: A new representation of Patterson's quadrature formulae, ACM Trans. Math. Software 17 (4) (1991) 457–461.

[51] A.R. Krommer, C.W. Ueberhuber, Numerical integration on advanced computer systems, Lecture notes on Computer Science, Vol. 848, Springer, Berlin, 1994.

[52] A.R. Krommer, C.W. Ueberhuber, Computational Integration, SIAM, Philadelphia, PA, 1998.

[53] A.S. Kronrod, Nodes and Weights for Quadrature Formulae. Sixteen Place Tables, Nauka, Moscow, 1964 (Translation by Consultants Bureau, New York, 1965).

[54] A.S. Kronrod, Integration with control of accuracy, Soviet Phys. Dokl. 9 (1) (1964) 17–19.

[55] D.P. Laurie, Practical error estimation in numerical integration, J. Comput. Appl. Math. 12 & 13 (1985) 425–431.

[56] D.P. Laurie, Stratified sequences of nested quadrature formulas, Quaestiones Math. 15 (1992) 365–384.

[57] D.P. Laurie, Null rules and orthogonal expansions, in: R.V.M. Zahar (Ed.), Approximation and Computation, a Festschrift in Honor of Walter Gautschi, International Series Numerical Mathematics, Vol. 119, Birkhäuser, Basel, 1994, pp. 359–370.

[58] D.P. Laurie, Anti-Gaussian quadrature formulas, Math. Comp. 65 (1996) 739–747.

[59] S. Li, Kronrod extension of Turán formula, Studia Sci. Math. Hungar. 29 (1994) 71–83.

[60] S. Li, Kronrod extension of generalized Gauss–Radau and Gauss–Lobatto formulae, Rocky Mountain J. Math. 26 (1996) 1455–1472.

[61] F. Locher, Fehlerkontrolle bei der numerischen Quadratur, in: G. Hämmerlin (Ed.), Numerische Integration, Proceedings of Conference on Oberwolfach, International Series Numerical Mathematics, Vol. 45, Birkhäuser, Basel, 1979, pp. 198–210.

[62] J. Lyness, Symmetric integration rules for hypercubes III. Construction of integration rules using null rules, Math. Comp. 19 (1965) 625–637.

[63] Matlab 5.3.1, The MathWorks, Inc., 1999.

[64] C. Micchelli, Monosplines and moment preserving spline approximation, in: H. Brass, G. Hämmerlin (Eds.), Numerical integration III, International Series Numerical Mathematics, Vol. 85, Birkhäuser, Basel, 1998, pp. 130–139.

[65] G. Monegato, A note on extended Gaussian quadrature rules, Math. Comp. 30 (1976) 812–817.

[66] G. Monegato, Positivity of weights of extended Gauss–Legendre quadrature rules, Math. Comp. 32 (1978) 243–245.

[67] G. Monegato, An overview of results and questions related to Kronrod schemes, in: G. Hämmerlin (Ed.), Numerische Integration, Proceedings of Conference on Oberwolfach, International Series Numerical Mathematics, Vol. 45, Birkhäuser, Basel, 1979, pp. 231–240.

[68] G. Monegato, On polynomials orthogonal with respect to particular variable signed weight functions, Z. Angew. Math. Phys. 31 (1980) 549–555.

[69] G. Monegato, Stieltjes polynomials and related quadrature rules, SIAM Rev. 24 (1982) 137–158.

[70] I.P. Mysovskih, A special case of quadrature formulae containing preassigned nodes, Vesci Akad. Navuk BSSR Ser. Fiz.-Tehn. Navuk 4 (1964) 125–127 (in Russian).

[71] NAG Fortran Library, Mark 18, NAG, Inc., 1998.

[72] S.E. Notaris, Gauss–Kronrod quadrature formulae for weight functions of Bernstein–Szegő type, II, J. Comput. Appl. Math. 29 (1990) 161–169.

[73] S.E. Notaris, Some new formulae for Stieltjes polynomials relative to classical weight functions, SIAM J. Numer. Anal. 28 (1991) 1196–1206.

[74] S.E. Notaris, An overview of results on the existence or nonexistence and the error term of Gauss–Kronrod quadrature formulae, in: R.V.M. Zahar (Ed.), Approximation and Computation, a Festschrift in Honor of Walter Gautschi, International Series Numerical Mathematics, Vol. 119, Birkhäuser, Basel, 1994, pp. 485–496.

[75] T.N.L. Patterson, The optimum addition of points to quadrature formulae, Math. Comp. 22 (1968) 847–856 (Microfiche CI–CII; Errata; ibid. 23 (1969) 892).

[76] T.N.L. Patterson, On some Gauss and Lobatto based integration formulae, Math. Comp. 22 (1968) 877–881.

[77] T.N.L. Patterson, Modified optimal quadrature extensions, Numer. Math. 64 (1993) 511–520.

[78] T.N.L. Patterson, Stratified nested and related quadrature rules, J. Comput. Appl. Math. 112 (1999) 243–251.

[79] F. Peherstorfer, Weight functions admitting repeated positive Kronrod quadrature, BIT 30 (1990) 145–151.

[80] F. Peherstorfer, On the asymptotic behaviour of functions of the second kind and Stieltjes polynomials and on Gauss–Kronrod quadrature formulas, J. Approx. Theory 70 (1992) 156–190.

[81] F. Peherstorfer, Stieltjes polynomials and functions of the second kind, J. Comput. Appl. Math. 65 (1–3) (1995) 319–338.

[82] F. Peherstorfer, K. Petras, Ultraspherical Gauss–Kronrod quadrature is not possible for $\lambda > 3$, SIAM J. Numer. Anal. 37 (2000) 927–948.

[83] K. Petras, Asymptotic behaviour of Peano kernels of fixed order, in: H. Brass, G. Hämmerlin (Eds.), Numerical Integration III, International Series Numerical Mathematics, Vol. 85, Birkhäuser, Basel, 1998, pp. 186–198.

[84] K. Petras, Quadrature theory of convex functions, a survey and additions, in: H. Brass, G. Hämmerlin (Eds.), Numerical Integration IV, International Series Numerical Mathematics, Vol. 112, Birkhäuser, Basel, 1993, pp. 315–329.

[85] R. Piessens, E. de Doncker, C. Überhuber, D.K. Kahaner, QUADPACK–a Subroutine Package for Automatic Integration, Springer Series in Computational Mathematics, Vol. 1, Springer, Berlin, 1983.

[86] P. Rabinowitz, The exact degree of precision of generalized Gauss–Kronrod integration rules, Math. Comp. 35 (1980) 1275–1283.

[87] P. Rabinowitz, Gauss–Kronrod integration rules for Cauchy principal value integrals, Math. Comp. 41 (1983) 63–78 (Corrigenda: ibid. 45 (1985) 277).

[88] P. Rabinowitz, On the definiteness of Gauss–Kronrod integration rules, Math. Comp. 46 (1986) 225–227.

[89] P. Rabinowitz, J. Kautsky, S. Elhay, Empirical mathematics: the first Patterson extension of Gauss–Kronrod rules, Internat. J. Computer Math. 36 (1990) 119–129.

[90] S. Ramanujan, Notebooks, Vol. 2, Tata Institute of Fundamental Research, Bombay, 1957.

[91] L.L. Schumaker, Spline Functions: Basic Theory, Wiley-Interscience, New York, 1981.

[92] Y.G. Shi, Generalised Gaussian Kronrod–Turán quadrature formulas, Acta. Sci. Math. (Szeged) 62 (1996) 175–185.

[93] H.V. Smith, An algebraic study of an extension of a class of quadrature formulae, J. Comput. Appl. Math. 37 (1991) 27–33.

[94] H.V. Smith, An extension of a class of quadrature formulae, J. Inst. Math. Comput. Sci. Math. Ser. 4 (1991) 37–40.

[95] A. Stroud, D. Secrest, Gaussian Quadrature Formulas, Prentice-Hall, Englewood Cliffs, NJ, 1996.

[96] G. Szegő, Über gewisse orthogonale Polynome, die zu einer oszillierenden Belegungsfunktion gehören, Math. Ann 110 (1934) 501–513 [cf. also R. Askey (Ed.), Collected Works, Vol. 2, pp. 545–558].

[97] G. Szegő, Orthogonal polynomials, American Mathematical Society Colloquim Publ., Vol. 23, AMS, Providence, RI, 1975.

ELSEVIER

Journal of Computational and Applied Mathematics 127 (2001) 173–199

JOURNAL OF
COMPUTATIONAL AND
APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Computation of matrix-valued formally orthogonal polynomials and applications

Roland W. Freund

Bell Laboratories, Lucent Technologies, Room 2C-525, 700 Mountain Avenue, Murray Hill, NJ 07974-0636, USA

Received 10 December 1999; received in revised form 10 March 2000

Abstract

We present a computational procedure for generating formally orthogonal polynomials associated with a given bilinear Hankel form with rectangular matrix-valued moments. Our approach covers the most general case of moments of any size and is not restricted to square moments. Moreover, our algorithm has a built-in deflation procedure to handle linearly dependent or almost linearly dependent columns and rows of the block Hankel matrix associated with the bilinear form. Possible singular or close-to-singular leading principal submatrices of the deflated block Hankel matrix are avoided by means of look-ahead techniques. Applications of the computational procedure to eigenvalue computations, reduced-order modeling, the solution of multiple linear systems, and the fast solution of block Hankel systems are also briefly described. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Bilinear form; Vector-valued polynomials; Matrix-valued moments; Block Hankel matrix; Deflation; Look-ahead; Realization; Block Krylov matrix; Lanczos-type algorithm; Matrix-Padé approximation; Fast block Hankel solver

1. Introduction

It has been known for a long time that many of the algebraic properties of scalar orthogonal polynomials on the real line carry over to the more general case of formally orthogonal polynomials induced by a given sequence of scalar moments; see, e.g., [3,4,8,15–17] and the references given there. For example, such formally orthogonal polynomials still satisfy three-term recurrences, as long as the scalar Hankel matrix $\boldsymbol{H}$ associated with the moment sequence is *strongly regular*, i.e., all leading principal submatrices of $\boldsymbol{H}$ are nonsingular. If $\boldsymbol{H}$ has some singular or in some sense nearly singular leading principal submatrices, then so-called look-ahead techniques can be used to jump over these submatrices, resulting in recurrence relations that connect the formally orthogonal polynomials corresponding to three consecutive look-ahead steps. In particular, these recurrences reduce to the standard three-term recurrences whenever three consecutive leading principal submatrices of $\boldsymbol{H}$ are nonsingular.

PII: S0377-0427(00)00505-7

Scalar formally orthogonal polynomials are intimately connected with a number of algorithms for matrix computations. For example, the classical Lanczos process [19] for nonsymmetric matrices, fast solvers for linear systems with Hankel structure [14], and the computation of Padé approximants of transfer functions of single-input single-output linear dynamical systems [6,10] are all closely related to formally orthogonal polynomials. Furthermore, the theory of formally orthogonal polynomials has proven to be useful for developing more robust versions of these algorithms. For instance, the look-ahead variants [12,17,22] of the Lanczos process, which remedy possible breakdowns in the classical algorithm, are direct translations of the extended recurrences for formally orthogonal polynomials in the general case of Hankel matrices $\boldsymbol{H}$ with singular or nearly singular leading principal submatrices.

The concept of formally orthogonal polynomials can be extended to the case of arbitrary, in general rectangular, matrix-valued moments. However, except for special cases such as square orthogonal matrix polynomials on the line [25], the theory of the associated matrix-valued polynomials is a lot less developed than in the case of scalar moments. For example, a suitable extension of the classical Lanczos process, which is only applicable to single right and left starting vectors, to multiple, say m right and p left, starting vectors is intimately related to formally orthogonal polynomials associated with sequences of $(p \times m)$-matrix-valued moments that are given by a so-called realization. Such a Lanczos-type method for multiple starting vectors was developed only recently [1,7], motivated mainly by the need for such an algorithm for the computation of matrix-Padé approximants of transfer functions of m-input p-output linear dynamical systems [7,9,10].

There are two intrinsic difficulties that arise in the case of $(p \times m)$-matrix-valued moments, but not in the case of scalar moments. First, in the important case of matrix moments given by a realization, the block Hankel matrix $\boldsymbol{H}$ associated with these moments necessarily exhibits systematic singularities or ill-conditioning due to linearly dependent or nearly linearly dependent columns and rows. These linear or nearly linear dependencies imply that from a certain point on all leading principal submatrices of $\boldsymbol{H}$ are singular or nearly singular, although the moment information contained in $\boldsymbol{H}$ has not been fully exhausted yet. In particular, block Hankel matrices induced by a realization are not strongly regular, and singular or nearly singular submatrices caused by linearly dependent or nearly linearly dependent columns and rows cannot be avoided by means of look-ahead techniques. Instead, so-called deflation is needed in order to remove systematic singularities or ill-conditioning due to linearly dependent or nearly linearly dependent columns and rows of $\boldsymbol{H}$. Second, the fact that $m \neq p$ in general excludes the possibility of constructing the formally orthogonal polynomials directly as right $(m \times m)$-matrix-valued and left $(p \times p)$-matrix-valued polynomials. Moreover, each deflation of a column of $\boldsymbol{H}$ effectively reduces m by one and each deflation of a row of $\boldsymbol{H}$ effectively reduces p by one. Since deflations of columns and rows occur independently in general, this means that the "current" values of m and p in the course of the construction of formally orthogonal polynomials will be different in general, even if $m = p$ initially. The difficulties due to different m and p can be avoided by constructing the polynomials associated with $(p \times m)$-matrix-valued moments vector-wise, rather than matrix-wise.

In this paper, we present a computational procedure for generating right and left formally orthogonal polynomials associated with a given bilinear form induced by a sequence of general rectangular $(p \times m)$-matrix-valued moments. Our approach covers the most general case of arbitrary integers $m, p \geqslant 1$, and we need not assume that the block Hankel matrix $\boldsymbol{H}$ associated with the given bilinear form is strongly regular. Our algorithm has a built-in deflation procedure to handle linearly dependent or almost linearly dependent columns and rows of $\boldsymbol{H}$. Possible singular or close-to-singular leading principal submatrices of the deflated block Hankel matrix are avoided by means of look-ahead techniques. Applications of the computational procedure to eigenvalue computations, reduced-order modeling, the solution of multiple linear systems, and the fast solution of block Hankel systems are also briefly described.

We remark that our approach of constructing formally orthogonal polynomials induced by matrix moments in a vector-wise fashion is related to earlier work, such as [2,24,27]. However, in these papers, the assumption that $\boldsymbol{H}$ is strongly regular is made, and this excludes the intrinsic difficulties described above.

The remainder of this paper is organized as follows. In Section 2, we introduce some notation. Section 3 describes our general setting of bilinear Hankel forms and discusses the need for deflation and look-ahead. In Section 4, we present our notion of formally orthogonal polynomials associated with a given bilinear Hankel form. In Section 5, we explain the structure of the recurrence relations used in our construction of formally orthogonal polynomials. A complete statement of our algorithm for generating formally orthogonal polynomials is given in Section 6, and some properties of this algorithm are stated in Section 7. Applications of the algorithm are sketched in Section 8. Finally, in Section 9, we make some concluding remarks.

2. Notation and some preliminaries

In this section, we introduce some notation used throughout this paper.

2.1. Notation

All vectors and matrices are allowed to have real or complex entries. We use boldface letters to denote vectors and matrices. As usual, $\overline{\boldsymbol{M}} = [\overline{m_{jk}}]$, $\boldsymbol{M}^{\mathrm{T}} = [m_{kj}]$, and $\boldsymbol{M}^{\mathrm{H}} = \overline{\boldsymbol{M}}^{\mathrm{T}} = [\overline{m_{kj}}]$ denote the complex conjugate, transpose, and the conjugate transpose, respectively, of the matrix $\boldsymbol{M}=[m_{jk}]$. We use the notation $[x_j]_{j\in\mathscr{J}}$ for the subvector of $\boldsymbol{x} = [x_j]$ induced by the index set $\mathscr{J}$, and analogously, $[m_{jk}]_{j\in\mathscr{J},k\in\mathscr{K}}$ for the submatrix of $\boldsymbol{M} = [m_{jk}]$ induced by the row indices $\mathscr{J}$ and column indices $\mathscr{K}$. The vector norm $\|\boldsymbol{x}\| := \sqrt{\boldsymbol{x}^{\mathrm{H}}\boldsymbol{x}}$ is always the Euclidean norm, and $\|\boldsymbol{M}\| := \max_{\|\boldsymbol{x}\|=1}\|\boldsymbol{M}\boldsymbol{x}\|$ is the corresponding induced matrix norm.

The ith unit vector of dimension j is denoted by $\boldsymbol{e}_i^{(j)}$. We use $\boldsymbol{I}_n$ to denote the $n\times n$ identity matrix, and we will simply write $\boldsymbol{I}$ if the actual dimension is apparent from the context.

The sets of real and complex numbers are denoted by $\mathbb{R}$ and $\mathbb{C}$, respectively. We use the symbols $\mathbb{N}$ for the set of positive integers and $\mathbb{N}_0 = \mathbb{N}\cup\{0\}$ for the set of non-negative integers.

We denote by $\mathscr{P}^{(j)}$ the set of all vector-valued polynomials

$$\boldsymbol{\phi}(\lambda) \equiv \boldsymbol{a}_0 + \boldsymbol{a}_1\lambda + \cdots + \boldsymbol{a}_i\lambda^i, \quad \text{where } \boldsymbol{a}_0, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_i \in \mathbb{C}^j,\ i \in \mathbb{N}_0, \tag{1}$$

with coefficient vectors of dimension j, and by

$$\mathscr{P}^{(j\times k)} := \{\boldsymbol{\Phi} = [\boldsymbol{\phi}_1 \quad \boldsymbol{\phi}_2 \quad \cdots \quad \boldsymbol{\phi}_k] \mid \boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_k \in \mathscr{P}^{(j)}\}$$

the set of all matrix-valued polynomials with coefficient matrices of size $j\times k$.

We use the symbol 0 both for the number zero and for the scalar zero polynomial, and similarly, the symbol $\mathbf{0}$ for the $m \times n$ zero matrix and the zero polynomials in $\mathscr{P}^{(j)}$ and $\mathscr{P}^{(j\times k)}$. The actual dimension of $\mathbf{0}$ will always be apparent from the context.

2.2. The degree of a vector polynomial

Following [2,5,27], we can associate with any given vector polynomial (1), $\boldsymbol{\phi}$, the scalar polynomial

$$\varphi(\lambda) \equiv [1 \quad \lambda \quad \cdots \quad \lambda^{j-1}] \cdot \boldsymbol{\phi}(\lambda^j) \equiv \sum_{k=0}^{(i+1)j-1} \alpha_k \lambda^k. \tag{2}$$

Here, the α_k's are just the coefficients of the *stacked* coefficient vector

$$\boldsymbol{a} = \begin{bmatrix} \boldsymbol{a}_0 \\ \boldsymbol{a}_1 \\ \vdots \\ \boldsymbol{a}_i \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{(i+1)j-1} \end{bmatrix} \in \mathbb{C}^{(i+1)j} \tag{3}$$

of (1). The (diagonal) *degree*, $\deg \boldsymbol{\phi}$, of $\boldsymbol{\phi}$ is then defined as the degree of the scalar polynomial (2), i.e.,

$$\deg \boldsymbol{\phi} := \begin{cases} \max\{k \mid \alpha_k \neq 0 \text{ in } (2)\} & \text{if } \boldsymbol{\phi} \neq \mathbf{0}, \\ -\infty & \text{if } \boldsymbol{\phi} = \mathbf{0}. \end{cases}$$

In the sequel, we will also use the notation

$$\operatorname{vec} \boldsymbol{\phi} := \begin{cases} [\alpha_k]_{0 \leqslant k \leqslant \deg \phi} & \text{if } \boldsymbol{\phi} \neq \mathbf{0}, \\ 0 & \text{if } \boldsymbol{\phi} = \mathbf{0} \end{cases}$$

for the vector that results from (3) by deleting any trailing zeros.

3. Bilinear Hankel forms and block Hankel matrices

In this section, we describe our general setting of bilinear Hankel forms and their associated infinite block Hankel matrices. We also discuss the need for deflation and look-ahead to avoid possible singular or ill-conditioned submatrices of the block Hankel matrices.

3.1. Bilinear Hankel forms

Let m, $p \geqslant 1$ be given integers. A complex-valued functional

$$\langle \cdot, \cdot \rangle : \mathscr{P}^{(p)} \times \mathscr{P}^{(m)} \mapsto \mathbb{C} \tag{4}$$

is called a *bilinear form* if

$$\langle \boldsymbol{\psi}, \sigma_1 \boldsymbol{\phi}_1 + \sigma_2 \boldsymbol{\phi}_2 \rangle = \sigma_1 \langle \boldsymbol{\psi}, \boldsymbol{\phi}_1 \rangle + \sigma_2 \langle \boldsymbol{\psi}, \boldsymbol{\phi}_2 \rangle \tag{5a}$$

and

$$\langle \sigma_1\boldsymbol{\psi}_1 + \sigma_2\boldsymbol{\psi}_2, \boldsymbol{\phi}\rangle = \sigma_1\langle \boldsymbol{\psi}_1, \boldsymbol{\phi}\rangle + \sigma_2\langle \boldsymbol{\psi}_2, \boldsymbol{\phi}\rangle \tag{5b}$$

for all $\boldsymbol{\phi}, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2 \in \mathscr{P}^{(m)}$, $\boldsymbol{\psi}, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2 \in \mathscr{P}^{(p)}$, $\sigma_1, \sigma_2 \in \mathbb{C}$. We say that (4) is a *Hankel form* if the following shift property is satisfied:

$$\langle \boldsymbol{\psi}, \lambda\boldsymbol{\phi}\rangle = \langle \lambda\boldsymbol{\psi}, \boldsymbol{\phi}\rangle \quad \text{for all } \boldsymbol{\phi} \in \mathscr{P}^{(m)},\ \boldsymbol{\psi} \in \mathscr{P}^{(p)}. \tag{6}$$

In the sequel, it is always assumed that (4) is a given bilinear Hankel form. Furthermore, we extend (4) to matrix-valued polynomials by setting

$$\langle \boldsymbol{\Psi}, \boldsymbol{\Phi}\rangle := \begin{bmatrix} \langle \boldsymbol{\psi}_1, \boldsymbol{\phi}_1\rangle & \langle \boldsymbol{\psi}_1, \boldsymbol{\phi}_2\rangle & \cdots & \langle \boldsymbol{\psi}_1, \boldsymbol{\phi}_k\rangle \\ \langle \boldsymbol{\psi}_2, \boldsymbol{\phi}_1\rangle & \langle \boldsymbol{\psi}_2, \boldsymbol{\phi}_2\rangle & \cdots & \langle \boldsymbol{\psi}_2, \boldsymbol{\phi}_k\rangle \\ \vdots & \vdots & & \vdots \\ \langle \boldsymbol{\psi}_j, \boldsymbol{\phi}_1\rangle & \langle \boldsymbol{\psi}_j, \boldsymbol{\phi}_2\rangle & \cdots & \langle \boldsymbol{\psi}_j, \boldsymbol{\phi}_k\rangle \end{bmatrix} \in \mathbb{C}^{j\times k} \tag{7}$$

for all

$$\boldsymbol{\Phi} = [\boldsymbol{\phi}_1 \quad \boldsymbol{\phi}_2 \quad \cdots \quad \boldsymbol{\phi}_k] \in \mathscr{P}^{(m\times k)}, \qquad \boldsymbol{\Psi} = [\boldsymbol{\psi}_1 \quad \boldsymbol{\psi}_2 \quad \cdots \quad \boldsymbol{\psi}_j] \in \mathscr{P}^{(p\times j)}.$$

In particular, using the notation (7), we define $p \times m$ (matrix-valued) *moments*

$$\boldsymbol{M}_{j,k} := \langle \boldsymbol{I}_p\lambda^j, \boldsymbol{I}_m\lambda^k\rangle \in \mathbb{C}^{p\times m} \quad \text{for all } j,k \in \mathbb{N}_0. \tag{8}$$

In view of the bilinearity (5a) and (5b), any bilinear form (4) is completely determined by its moments (8). Furthermore, the shift property (6) means that the moments $\boldsymbol{M}_{j,k}$ only depend on $j+k$, and we set $\boldsymbol{M}_{j+k} := \boldsymbol{M}_{j,k}$. Therefore, any bilinear Hankel form (4) is completely determined by the sequence of moments

$$\boldsymbol{M}_i := \langle \boldsymbol{I}_p, \boldsymbol{I}_m\lambda^i\rangle = \langle \boldsymbol{I}_p\lambda^i, \boldsymbol{I}_m\rangle, \quad i \in \mathbb{N}_0.$$

The associated infinite block Hankel matrix

$$\boldsymbol{H} := [\boldsymbol{M}_{j+k}]_{j,k\geqslant 0} = \begin{bmatrix} \boldsymbol{M}_0 & \boldsymbol{M}_1 & \boldsymbol{M}_2 & \cdots \\ \boldsymbol{M}_1 & \boldsymbol{M}_2 & & \\ \boldsymbol{M}_2 & & & \\ \vdots & & & \end{bmatrix} \tag{9}$$

is called the *moment matrix* of the bilinear Hankel form (4). Although $\boldsymbol{H}$ has a block Hankel structure, we will also consider $\boldsymbol{H}$ as a scalar matrix with entries $h_{v,\mu}$, i.e.,

$$\boldsymbol{H} = [h_{v,\mu}]_{v,\mu\geqslant 0}, \quad \text{where } h_{v,\mu} \in \mathbb{C} \text{ for all } v,\mu \in \mathbb{N}_0.$$

Furthermore, for each $n,k \in \mathbb{N}_0$, we set

$$\boldsymbol{H}_{n,k} := [h_{v,\mu}]_{0\leqslant v\leqslant n, 0\leqslant\mu\leqslant k} \quad \text{and} \quad \boldsymbol{H}_n := \boldsymbol{H}_{n,n} = [h_{v,\mu}]_{0\leqslant v,\mu\leqslant n}.$$

Note that $\boldsymbol{H}_n$ is the nth scalar leading principal submatrix of $\boldsymbol{H}$.

With the notation just introduced, for any pair of vector-valued polynomials $\boldsymbol{\phi} \in \mathscr{P}^{(m)}$ and $\boldsymbol{\psi} \in \mathscr{P}^{(p)}$, we have

$$\langle \boldsymbol{\psi}, \boldsymbol{\phi}\rangle := \boldsymbol{b}^{\mathrm{T}}\boldsymbol{H}_{\deg\boldsymbol{\psi},\deg\boldsymbol{\phi}}\boldsymbol{a}, \quad \text{where } \boldsymbol{a} = \operatorname{vec}\boldsymbol{\phi},\ \boldsymbol{b} = \operatorname{vec}\boldsymbol{\psi}. \tag{10}$$

3.2. Bilinear Hankel forms associated with realizations

An important special case is bilinear Hankel forms that are associated with so-called realizations of time-invariant linear dynamical systems. We refer the reader to [18, Chapter 10.11], [23], or [26, Chapter 5.5] for a discussion of the concepts and results from realization theory that we will use in this subsection.

Let

$$\boldsymbol{M}_i \in \mathbb{C}^{p\times m}, \quad i=0,1,\ldots, \tag{11}$$

be a given sequence of moments. A triple of matrices

$$\boldsymbol{A} \in \mathbb{C}^{N\times N}, \quad \boldsymbol{R} \in \mathbb{C}^{N\times m}, \quad \boldsymbol{L} \in \mathbb{C}^{N\times p} \tag{12}$$

is called a *realization* of the sequence (11) if

$$\boldsymbol{M}_i = \boldsymbol{L}^{\mathrm{T}}\boldsymbol{A}^i\boldsymbol{R} \quad \text{for all } i \in \mathbb{N}_0.$$

The integer N is called the *dimension* of the realization (12). A realization (12) of a given sequence (11) is said to be *minimal* if its dimension N is as small as possible.

Not every given sequence (11) has a realization. The following well-known result (see, e.g., [26, Theorem 21]) gives a necessary and sufficient condition for the existence of a realization in terms of the infinite block Hankel matrix (9) with block entries (11).

Theorem A. *A sequence* (11) *admits a realization if, and only if, the associated infinite block Hankel matrix* (9), $\boldsymbol{H}$, *has finite rank. Furthermore, if* (11) *has a realization, then* $N=\operatorname{rank}\boldsymbol{H}$ *is the dimension of a minimal realization.*

For the remainder of this subsection, we now assume that $\boldsymbol{H}$ has finite rank and that (12) is a given, not necessarily minimal, realization of the block entries (11) of $\boldsymbol{H}$.

Note that, in view of (9) and (12), the block Hankel matrix $\boldsymbol{H}$ can be factored into block Krylov matrices as follows:

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{L}^{\mathrm{T}} \\ \boldsymbol{L}^{\mathrm{T}}\boldsymbol{A} \\ \boldsymbol{L}^{\mathrm{T}}\boldsymbol{A}^2 \\ \vdots \end{bmatrix} \cdot [\boldsymbol{R} \quad \boldsymbol{A}\boldsymbol{R} \quad \boldsymbol{A}^2\boldsymbol{R} \quad \cdots]. \tag{13}$$

As a first application of (13), we have the following connection of vector-valued polynomials with vectors in $\mathbb{C}^N$.

Remark 1. Let $\boldsymbol{\phi} \in \mathscr{P}^{(m)}$ and $\boldsymbol{\psi} \in \mathscr{P}^{(p)}$ be any pair of polynomials. Then, using the representations

$$\boldsymbol{\phi}(\lambda) \equiv \boldsymbol{a}_0 + \boldsymbol{a}_1\lambda + \cdots + \boldsymbol{a}_j\lambda^j,$$
$$\boldsymbol{\psi}(\lambda) \equiv \boldsymbol{b}_0 + \boldsymbol{b}_1\lambda + \cdots + \boldsymbol{b}_k\lambda^k,$$

we can associate with $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ the pair of vectors

$$\begin{aligned} \boldsymbol{v} &= \boldsymbol{\phi}(\boldsymbol{A}) \circ \boldsymbol{R} := \boldsymbol{R}\boldsymbol{a}_0 + \boldsymbol{A}\boldsymbol{R}\boldsymbol{a}_1 + \cdots + \boldsymbol{A}^j\boldsymbol{R}\boldsymbol{a}_j \in \mathbb{C}^N, \\ \boldsymbol{w} &= \boldsymbol{\psi}(\boldsymbol{A}^{\mathrm{T}}) \circ \boldsymbol{L} := \boldsymbol{L}\boldsymbol{b}_0 + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{b}_1 + \cdots + (\boldsymbol{A}^{\mathrm{T}})^k\boldsymbol{L}\boldsymbol{b}_k \in \mathbb{C}^N. \end{aligned} \tag{14}$$

By (10), (13), and (14), it follows that $\langle \boldsymbol{\psi}, \boldsymbol{\phi} \rangle = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{v}$.

The factorization (13) of $\boldsymbol{H}$ necessarily implies that from a certain n on, all leading principal submatrices $\boldsymbol{H}_n$ are singular. Indeed, consider the *right block Krylov matrix*

$$[\boldsymbol{R} \quad \boldsymbol{A}\boldsymbol{R} \quad \boldsymbol{A}^2\boldsymbol{R} \quad \cdots] \tag{15}$$

in (13). The columns of the matrix (15) are vectors in $\mathbb{C}^N$, and hence at most N of them are linearly independent. By scanning the columns of (15) from left to right and deleting each column that is linearly dependent on earlier columns, we obtain the *deflated* right block Krylov matrix

$$[\boldsymbol{R}_0 \quad \boldsymbol{A}\boldsymbol{R}_1 \quad \boldsymbol{A}^2\boldsymbol{R}_2 \quad \cdots \quad \boldsymbol{A}^{j_{\max}-1}\boldsymbol{R}_{j_{\max}}]. \tag{16}$$

By the structure of (15), a column $\boldsymbol{A}^{j-1}\boldsymbol{r}$ being linearly dependent on earlier columns implies that all columns $\boldsymbol{A}^i\boldsymbol{r}$, $i \geqslant j$, are also linearly dependent on earlier columns. This implies that, for each $j = 0, 1, \ldots, j_{\max}$, $\boldsymbol{R}_j$ is a submatrix of $\boldsymbol{R}_{j-1}$, where, for $j = 0$, we set $\boldsymbol{R}_{-1} := \boldsymbol{R}$. Similarly, by scanning the columns of the *left block Krylov matrix*

$$[\boldsymbol{L} \quad \boldsymbol{A}^{\mathrm{T}}\boldsymbol{L} \quad (\boldsymbol{A}^{\mathrm{T}})^2\boldsymbol{L} \quad \cdots]$$

from left to right and deleting each column that is linearly dependent on earlier columns, we obtain the *deflated* left block Krylov matrix

$$[\boldsymbol{L}_0 \quad \boldsymbol{A}^{\mathrm{T}}\boldsymbol{L}_1 \quad (\boldsymbol{A}^{\mathrm{T}})^2\boldsymbol{L}_2 \quad \cdots \quad (\boldsymbol{A}^{\mathrm{T}})^{k_{\max}-1}\boldsymbol{L}_{k_{\max}}]. \tag{17}$$

Here, for each $k = 0, 1, \ldots, k_{\max}$, $\boldsymbol{L}_k$ is a submatrix of $\boldsymbol{L}_{k-1}$, where, for $k = 0$, we set $\boldsymbol{L}_{-1} := \boldsymbol{L}$. Now let j_0 be the smallest integer such that $\boldsymbol{R}_{j_0} \neq \boldsymbol{R}$ and let k_0 be the smallest integer such that $\boldsymbol{L}_{k_0} \neq \boldsymbol{L}$. Then, by construction, it follows that the nth leading principal submatrix

$$\boldsymbol{H}_n \quad \text{is singular for all} \quad n \geqslant \min\{(j_0 + 1)m, (k_0 + 1)p\} - 1.$$

Finally, we note that this systematic singularity of the submatrices $\boldsymbol{H}_n$ can be avoided by replacing the moment matrix $\boldsymbol{H}$ by the *deflated moment matrix*

$$\boldsymbol{H}^{\mathrm{defl}} := \begin{bmatrix} \boldsymbol{L}_0^{\mathrm{T}} \\ \boldsymbol{L}_1^{\mathrm{T}}\boldsymbol{A} \\ \vdots \\ \boldsymbol{L}_{k_{\max}}^{\mathrm{T}}\boldsymbol{A}^{k_{\max}-1} \end{bmatrix} \cdot [\boldsymbol{R}_0 \quad \boldsymbol{A}\boldsymbol{R}_1 \quad \cdots \quad \boldsymbol{A}^{j_{\max}-1}\boldsymbol{R}_{j_{\max}}].$$

3.3. The need for deflation

We now return to general moment matrices $\boldsymbol{H}$, and we extend the procedure for avoiding systematic singularities of submatrices $\boldsymbol{H}_n$ to this general case.

Generalizing the procedure described in Section 3.2, we scan the columns, $[h_{i,\mu}]_{i\geqslant 0}$, $\mu = 0, 1, \ldots,$ of $\boldsymbol{H}$, from left to right and delete each column that is linearly dependent, or in some sense "almost" linearly dependent, on earlier columns. Similarly, we scan the rows, $[h_{v,i}]_{i\geqslant 0}$, $v=0,1,\dots,$ of $\boldsymbol{H}$, from top to bottom and delete each row that is linearly dependent, or in some sense "almost" linearly dependent, on earlier rows. We refer to this process of deleting linearly dependent and "almost" linearly dependent columns and rows as *deflation*. Moreover, we say that *exact deflation* is performed when only the linearly dependent columns and rows are removed. Clearly, exact deflation is only possible in exact arithmetic, and deflation of "almost" linearly dependent columns and rows has to be included when actual computations are done in finite-precision arithmetic.

In the sequel, we denote by

$$0\leqslant\mu_0<\mu_1<\cdots<\mu_n<\cdots \quad\text{and}\quad 0\leqslant v_0<v_1<\cdots<v_n<\cdots \tag{18}$$

the sequences of indices of those columns and rows of $\boldsymbol{H}$, respectively, that are left after deflation has been performed. Moreover, we denote by

$$\mathscr{M}=\{\mu_n\}_{n=0}^{n_{\mathrm{r}}} \quad\text{and}\quad \mathscr{N}=\{v_n\}_{n=0}^{n_\ell} \tag{19}$$

the sets of all the indices (18), and by

$$\boldsymbol{H}^{\mathrm{defl}}:=[h_{v,\mu}]_{v\in\mathscr{N},\,\mu\in\mathscr{M}} \tag{20}$$

the corresponding *deflated* moment matrix. Note that, in general, each of the two sequences (18) may be finite of infinite, i.e., $n_{\mathrm{r}},n_\ell\in\mathbb{N}_0\cup\{\infty\}$. Hence $\boldsymbol{H}^{\mathrm{defl}}$ can have finitely or infinitely many rows or columns. However, in the special case of Hankel matrices $\boldsymbol{H}$ of finite rank discussed in Section 3.2, $n_{\mathrm{r}},n_\ell<\infty$ and thus $\boldsymbol{H}^{\mathrm{defl}}$ is a finite matrix.

In view of the block Hankel structure (9) of $\boldsymbol{H}$, an exact deflation of a μth column of $\boldsymbol{H}$ implies that also all $(\mu+jm)$th columns, where $j=1,2,\dots,$ need to be deflated. Similarly, an exact deflation of a vth row implies that also all $(v+jp)$th rows, where $j=1,2,\dots,$ need to be deflated. We assume that the same rule is also applied in the case of general deflation, and so whenever a μth column or vth row is deflated, we also deflate all $(\mu+jm)$th columns, $j\in\mathbb{N}$, respectively all $(v+jp)$th rows, $j\in\mathbb{N}$. This implies that the sets (19) always satisfy the following conditions:

$$\begin{aligned}&\mu\notin\mathscr{M}\Rightarrow\mu+jm\notin\mathscr{M}\quad\text{for all } j\in\mathbb{N}_0,\\ &v\notin\mathscr{N}\Rightarrow v+jp\notin\mathscr{N}\quad\text{for all } j\in\mathbb{N}_0.\end{aligned}$$

Finally, we note that, by (18), the mappings $n\mapsto\mu_n$ and $n\mapsto v_n$ are both invertible, and we will use μ^{-1} and v^{-1} to denote the inverse mappings defined by

$$n=\mu^{-1}(\mu_n),\quad \mu_n\in\mathscr{M},\quad\text{and}\quad n=v^{-1}(v_n),\quad v_n\in\mathscr{N},$$

respectively.

3.4. The need for look-ahead

By replacing $\boldsymbol{H}$ with the deflated moment matrix $\boldsymbol{H}^{\mathrm{defl}}$, we have removed any rank deficiencies due to linearly dependent columns and rows. Next, we discuss potential singularities of the leading principal submatrices of $\boldsymbol{H}^{\mathrm{defl}}$.

By (19) and (20), all the leading principal submatrices of $\boldsymbol{H}^{\mathrm{defl}}$ are given by

$$\boldsymbol{H}_{n-1}^{\mathrm{defl}}:=[h_{v_i,\mu_j}]_{0\leqslant i,j<n}\quad\text{for all } n\in\mathbb{N} \text{ with } n\leqslant n_{\max}, \tag{21}$$

where

$$n_{\max} := 1 + \min\{n_{\mathrm{r}}, n_\ell\}. \tag{22}$$

Note that either $n_{\max} \in \mathbb{N}_0$ or $n_{\max} = \infty$. In the special case of Hankel matrices $\boldsymbol{H}$ of finite rank discussed in Section 3.2, $n_{\max}$ is a finite integer.

By construction, the necessary singularities implied by linearly dependent columns or rows of $\boldsymbol{H}$ have been removed from the submatrices (21). Obviously, as in the case of scalar Hankel matrices (see, e.g., [14] and the references given there), this construction alone is not sufficient to always guarantee that

$$\det \boldsymbol{H}_{n-1}^{\mathrm{defl}} \neq 0 \quad \text{for all } n \in \mathbb{N} \text{ with } n \leqslant n_{\max}. \tag{23}$$

However, the situation that (23) is satisfied is the *generic case*.

In the *general case*, some of the submatrices (21) may be singular or in some sense "close" to singular, and we employ so-called *look-ahead* techniques [22,28] to avoid these submatrices. We use the indices

$$n_0 := 0 < n_1 < n_2 < \cdots < n_k < \cdots \tag{24}$$

to mark those submatrices $\boldsymbol{H}_{n_k-1}$ that remain after any singular or close-to-singular submatrix has been removed from (21). In particular, by construction, we have

$$\det \boldsymbol{H}_{n_k-1}^{\mathrm{defl}} \neq 0 \quad \text{for all } k \in \mathbb{N} \text{ with } n_k \leqslant n_{\max}.$$

Finally, note that in the generic case when no look-ahead is necessary, the indices (24) are simply given by

$$n_k = k \quad \text{for all } k \in \mathbb{N}_0 \text{ with } k \leqslant n_{\max}. \tag{25}$$

4. Formally orthogonal polynomials

In this section, we present our notion of formally orthogonal polynomials associated with a given bilinear Hankel form $\langle\cdot,\cdot\rangle : \mathscr{P}^{(p)} \times \mathscr{P}^{(m)} \mapsto \mathbb{C}$.

4.1. Two sequences of polynomials

Let $\mathscr{M} = \{\mu_n\}_{n=0}^{n_{\mathrm{r}}}$ and $\mathscr{N} = \{\nu_n\}_{n=0}^{n_\ell}$ be the column and row indices introduced in (18) and (19). Of course, in practice, these indices are not given beforehand, and instead, they have to be determined within our computational procedure for constructing vector-valued orthogonal polynomials. In Section 4.4 below, we will show how this can be done, but for now, we assume that $\mathscr{M}$ and $\mathscr{N}$ are given.

Our computational procedure generates two sequences of *right* and *left* polynomials,

$$\boldsymbol{\phi}_0, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n, \ldots \in \mathscr{P}^{(m)} \quad \text{and} \quad \boldsymbol{\psi}_0, \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n, \ldots \in \mathscr{P}^{(p)}, \tag{26}$$

respectively. Here, $n \in \mathbb{N}_0$ and $n \leqslant n_{\max}$, where $n_{\max}$ is given by (22). Furthermore, the polynomials (26) are constructed such that their degrees are just the indices $\mathscr{M}$ and $\mathscr{N}$, i.e.,

$$\deg \boldsymbol{\phi}_n = \mu_n \quad \text{and} \quad \deg \boldsymbol{\psi}_n = \nu_n \quad \text{for all } n, \tag{27}$$

and their coefficient vectors,

$$\boldsymbol{a}^{(n)} := \operatorname{vec} \boldsymbol{\phi}_n = \begin{bmatrix} \alpha_0^{(n)} \\ \alpha_1^{(n)} \\ \vdots \\ \alpha_{\mu_n}^{(n)} \end{bmatrix} \quad \text{and} \quad \boldsymbol{b}^{(n)} := \operatorname{vec} \boldsymbol{\psi}_n = \begin{bmatrix} \beta_0^{(n)} \\ \beta_1^{(n)} \\ \vdots \\ \beta_{\nu_n}^{(n)} \end{bmatrix}, \tag{28}$$

have no nonzero entries outside $\mathscr{M}$ and $\mathscr{N}$, i.e., for all n,

$$\alpha_\mu^{(n)} = 0 \quad \text{if } \mu \neq \mathscr{M} \quad \text{and} \quad \beta_\nu^{(n)} = 0 \quad \text{if } \nu \neq \mathscr{N}. \tag{29}$$

4.2. Orthogonality in the generic case

The goal is to construct the polynomials (26) such that they are regular formally orthogonal polynomials in the sense of the following definition.

Definition 2. The polynomial $\boldsymbol{\phi}_n \in \mathscr{P}^{(m)}$ is said to be an nth right formally orthogonal polynomial (RFOP) if $\deg \boldsymbol{\phi}_n = \mu_n$ and

$$\langle \boldsymbol{\psi}, \boldsymbol{\phi}_n \rangle = 0 \quad \text{for all } \boldsymbol{\psi} \in \mathscr{P}^{(p)} \text{ with } \deg \boldsymbol{\psi} = \nu_i, \quad i < n. \tag{30}$$

The polynomial $\boldsymbol{\psi}_n \in \mathscr{P}^{(p)}$ is said to be an nth left formally orthogonal polynomial (LFOP) if $\deg \boldsymbol{\psi}_n = \nu_n$ and

$$\langle \boldsymbol{\psi}_n, \boldsymbol{\phi} \rangle = 0 \quad \text{for all } \boldsymbol{\phi} \in \mathscr{P}^{(m)} \text{ with } \deg \boldsymbol{\phi} = \mu_i, \quad i < n. \tag{31}$$

Moreover, the RFOP $\boldsymbol{\phi}_n$ and the LFOP $\boldsymbol{\psi}_n$ are said to be regular if they are uniquely determined by (30) and (31), respectively, up to a nonzero scalar factor.

Using (10), (21), and (27)–(29), one readily verifies that the condition (30) is equivalent to the system of linear equations,

$$\boldsymbol{H}_{n-1}^{\text{defl}} \begin{bmatrix} \alpha_{\mu_0}^{(n)} \\ \alpha_{\mu_1}^{(n)} \\ \vdots \\ \alpha_{\mu_{n-1}}^{(n)} \end{bmatrix} = -\alpha_{\mu_n}^{(n)} \begin{bmatrix} h_{\nu_0,\mu_n} \\ h_{\nu_1,\mu_n} \\ \vdots \\ h_{\nu_{n-1},\mu_n} \end{bmatrix}, \quad \alpha_{\mu_n}^{(n)} \neq 0, \tag{32}$$

for the potentially nonzero coefficients of $\boldsymbol{\phi}_n$. Similarly, (31) is equivalent to the system of linear equations,

$$(\boldsymbol{H}_{n-1}^{\text{defl}})^{\mathrm{T}} \begin{bmatrix} \beta_{\nu_0}^{(n)} \\ \beta_{\nu_1}^{(n)} \\ \vdots \\ \beta_{\nu_{n-1}}^{(n)} \end{bmatrix} = -\beta_{\nu_n}^{(n)} \begin{bmatrix} h_{\nu_n,\mu_0} \\ h_{\nu_n,\mu_1} \\ \vdots \\ h_{\nu_n,\mu_{n-1}} \end{bmatrix}, \quad \beta_{\nu_n}^{(n)} \neq 0, \tag{33}$$

for the potentially nonzero coefficients of $\boldsymbol{\psi}_n$. In view of (32) and (33), $\boldsymbol{\phi}_n$ and $\boldsymbol{\psi}_n$ can be constructed as a regular RFOP and a regular LFOP, respectively, if, and only if, $\boldsymbol{H}_{n-1}^{\text{defl}}$ is nonsingular. We thus have the following result on the existence of regular RFOPs and LFOPs in the generic case (23).

Theorem 3. *The polynomials* (26) *can all be constructed as regular RFOPs and LFOPs only in the generic case*, *i.e.*, *if the matrices* $\boldsymbol{H}_{n-1}^{\mathrm{defl}}$ *are nonsingular for all* $n \in \mathbb{N}$ *with* $n \leqslant n_{\max}$.

In the generic case, we thus construct the polynomials (26) such that

$$\langle \boldsymbol{\psi}_i, \boldsymbol{\phi}_n \rangle = 0 \quad \text{for all } i \neq n,\ i,n \in \mathbb{N}_0,\ i,n \leqslant n_{\max}. \tag{34}$$

Using the notation

$$\boldsymbol{\Phi}_n := [\boldsymbol{\phi}_0 \quad \boldsymbol{\phi}_1 \quad \cdots \quad \boldsymbol{\phi}_n] \quad \text{and} \quad \boldsymbol{\Psi}_n := [\boldsymbol{\psi}_0 \quad \boldsymbol{\psi}_1 \quad \cdots \quad \boldsymbol{\psi}_n] \tag{35}$$

for the matrix polynomials whose columns are the first $n+1$ right and left polynomials (26), respectively, the orthogonality condition (34) can be stated as follows:

$$\langle \boldsymbol{\Psi}_n, \boldsymbol{\Phi}_n \rangle = \boldsymbol{\Delta}_n \quad \text{for all } n \in \mathbb{N}_0,\ n \leqslant n_{\max}. \tag{36}$$

Here,

$$\boldsymbol{\Delta}_n := \operatorname{diag}(\delta_0, \delta_1, \ldots, \delta_n), \quad \text{where } \delta_i := \langle \boldsymbol{\psi}_i, \boldsymbol{\phi}_i \rangle \text{ for all } i. \tag{37}$$

Finally, let

$$\boldsymbol{A}_n := [\alpha_{\mu_i}^{(j)}]_{0 \leqslant i,j \leqslant n} \quad \text{and} \quad \boldsymbol{B}_n := [\beta_{v_i}^{(j)}]_{0 \leqslant i,j \leqslant n} \tag{38}$$

denote the matrices of the potentially nonzero coefficients of $\boldsymbol{\Phi}_n$ and $\boldsymbol{\Psi}_n$, respectively. Then, by (10), (21), (28), (29), and (38), the condition (36) is equivalent to the matrix factorization

$$\boldsymbol{B}_n^{\mathrm{T}} \boldsymbol{H}_n^{\mathrm{defl}} \boldsymbol{A}_n = \boldsymbol{\Delta}_n, \tag{39}$$

where $\boldsymbol{A}_n$ and $\boldsymbol{B}_n$ are nonsingular upper triangular matrices. Thus, it follows from (37) and (39) that the condition (23) for the generic case is equivalent to

$$\delta_i \neq 0 \quad \text{for all } 0 \leqslant i < n_{\max}. \tag{40}$$

4.3. Orthogonality in the case of look-ahead

We now turn to the general case where look-ahead is used to avoid singular or close-to-singular submatrices $\boldsymbol{H}_{n-1}^{\mathrm{defl}}$. In view of Theorem 3, the polynomials $\boldsymbol{\phi}_n$ and $\boldsymbol{\psi}_n$ cannot be constructed as a regular RFOP and LFOP is $\boldsymbol{H}_{n-1}^{\mathrm{defl}}$ if exactly singular. If $\boldsymbol{H}_{n-1}^{\mathrm{defl}}$ is nonsingular, but in some sense close to singular, then building $\boldsymbol{\phi}_n$ and $\boldsymbol{\psi}_n$ as a regular RFOP and LFOP will result in numerical instabilities in general. Therefore, we only construct the polynomials $\boldsymbol{\phi}_{n_k}$ and $\boldsymbol{\psi}_{n_k}$ corresponding to the index sequence (24) as regular RFOPs and LFOPs, while the remaining polynomials satisfy only a relaxed version of the orthogonality condition (34).

More precisely, based on the indices (24), we partition the right and left polynomials (26) into *clusters*

$$\boldsymbol{\Phi}^{(k)} := [\boldsymbol{\phi}_{n_k} \quad \boldsymbol{\phi}_{n_k+1} \quad \cdots \quad \boldsymbol{\phi}_{n_{k+1}-1}] \tag{41a}$$

and

$$\boldsymbol{\Psi}^{(k)} := [\boldsymbol{\psi}_{n_k} \quad \boldsymbol{\psi}_{n_k+1} \quad \cdots \quad \boldsymbol{\psi}_{n_{k+1}-1}], \tag{41b}$$

respectively. The polynomials (26) are then constructed such that we have the cluster-wise orthogonality

$$\langle \boldsymbol{\Psi}^{(j)}, \boldsymbol{\Phi}^{(k)} \rangle = \boldsymbol{0} \quad \text{for all } j \neq k,\ j,k \in \mathbb{N}_0,\ n_j, n_k \leqslant n_{\max}. \tag{42}$$

We remark that, by (42), the leading polynomials $\boldsymbol{\phi}_{n_k}$ and $\boldsymbol{\psi}_{n_k}$ of each kth cluster (41a) and (41b) are regular RFOPs and LFOPs, respectively. Next, we set

$$\boldsymbol{\Delta}^{(k)} := \langle \boldsymbol{\Psi}^{(k)}, \boldsymbol{\Phi}^{(k)} \rangle \quad \text{for all } k.$$

Then, using essentially the same argument that lead to condition (40) in the generic case, it follows that, in the general case,

$$\boldsymbol{\Delta}^{(k)} \quad \text{is nonsingular for all } k \in \mathbb{N}_0 \text{ with } n_k < n_{\max}.$$

Finally, note that each nth pair of polynomials $\boldsymbol{\phi}_n$ and $\boldsymbol{\psi}_n$ in (26) is part of exactly one pair of clusters (41a) and (41b), namely those with index $k = \gamma(n)$. Here and in the sequel, we use the notation

$$\gamma(n) := \max\{ j \in \mathbb{N}_0 \mid n_j \leqslant n \} \tag{43}$$

for the function that determines the cluster index for the nth pair of polynomials. Recall from (25) that in the case of no look-ahead, $n_k = k$ for all k. Thus, in the generic case, (43) reduces to

$$\gamma(n) = n \quad \text{for all } n. \tag{44}$$

4.4. How deflation is done

In practice, the indices (18) and (19), which describe deflations, are not given beforehand, and instead, they have to be determined as part of the algorithm for constructing the polynomials (26). In this subsection, we describe how this is done.

In the algorithm, we keep track of the *current* block sizes m_c and p_c. Initially, $m_c = m$ and $p_c = p$. Every time a deflation of a right polynomial is performed, we set $m_c = m_c - 1$, and every time a deflation of a left polynomial is performed, we set $p_c = p_c - 1$. Thus, at any stage of the algorithm, $m - m_c$, respectively $p - p_c$, is just the number of deflations of right, respectively left, polynomials that have occurred so far.

Now assume that we already have constructed the right and left polynomials (26) up to index $n-1$. In addition to these polynomials, our algorithm has built m_c right and p_c left auxiliary polynomials,

$$\hat{\boldsymbol{\phi}}_n, \hat{\boldsymbol{\phi}}_{n+1}, \ldots, \hat{\boldsymbol{\phi}}_{n+m_c-1} \quad \text{and} \quad \hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\psi}}_{n+1}, \ldots, \hat{\boldsymbol{\psi}}_{n+p_c-1}, \tag{45}$$

that satisfy the following "partial" orthogonality conditions:

$$\begin{aligned} \langle \boldsymbol{\Psi}^{(k)}, \hat{\boldsymbol{\phi}}_i \rangle &= \mathbf{0} \quad \text{for all } 0 \leqslant k < \gamma(n),\ n \leqslant i < n + m_c, \\ \langle \hat{\boldsymbol{\psi}}_i, \boldsymbol{\Phi}^{(k)} \rangle &= \mathbf{0} \quad \text{for all } 0 \leqslant k < \gamma(n),\ n \leqslant i < n + p_c. \end{aligned} \tag{46}$$

The polynomials (45) are the candidates for the next m_c right, respectively p_c left, polynomials in (26). In particular, in view of (46), the polynomials $\hat{\boldsymbol{\phi}}_n$ and $\hat{\boldsymbol{\psi}}_n$ already satisfy the necessary orthogonality conditions of $\boldsymbol{\phi}_n$ and $\boldsymbol{\psi}_n$. It remains to decide if $\hat{\boldsymbol{\phi}}_n$ or $\hat{\boldsymbol{\psi}}_n$ should be deflated. If $\hat{\boldsymbol{\phi}}_n$ is deflated, it is deleted from (45), the indices of the remaining right polynomials in (45) are reduced by one, and m_c is reduced by one. If $\hat{\boldsymbol{\psi}}_n$ is deflated, one proceeds analogously.

The decision if $\hat{\boldsymbol{\phi}}_n$ or $\hat{\boldsymbol{\psi}}_n$ needs to be deflated is relatively simple for the special case of bilinear Hankel forms given in terms of a realization (12). Indeed, it is easy to see that an exact deflation of $\hat{\boldsymbol{\phi}}_n$, respectively $\hat{\boldsymbol{\psi}}_n$, needs to be performed if, and only if, the associated vectors (see (14)) satisfy

$$\hat{\boldsymbol{\phi}}_n(\boldsymbol{A}) \circ \boldsymbol{R} = \mathbf{0}, \quad \text{respectively } \hat{\boldsymbol{\psi}}_n(\boldsymbol{A}^{\mathrm{T}}) \circ \boldsymbol{L} = \mathbf{0}.$$

In practice, one thus deflates $\hat{\boldsymbol{\phi}}_n$, respectively $\hat{\boldsymbol{\psi}}_n$, if

$$\|\hat{\boldsymbol{\phi}}_n(\boldsymbol{A})\circ\boldsymbol{R}\|\leqslant \texttt{dtol}_n^{\texttt{r}}, \quad \text{respectively } \|\hat{\boldsymbol{\psi}}_n(\boldsymbol{A}^{\mathrm{T}})\circ\boldsymbol{L}\|\leqslant\texttt{dtol}_n^{\texttt{l}}, \tag{47}$$

where $\texttt{dtol}_n^{\texttt{r}}\geqslant 0$ and $\texttt{dtol}_n^{\texttt{l}}\geqslant 0$ are suitably chosen small deflation tolerances. Note that the check (47) reduces to exact deflation only if $\texttt{dtol}_n^{\texttt{r}}=\texttt{dtol}_n^{\texttt{l}}=0$.

In the general case of Hankel matrices $\boldsymbol{H}$ of not necessarily finite rank, one can show that exact deflation of $\hat{\boldsymbol{\phi}}_n$, respectively $\hat{\boldsymbol{\psi}}_n$, needs to be performed if, and only if,

$$\langle \boldsymbol{I}_p\lambda^k, \hat{\boldsymbol{\phi}}_n\rangle=\boldsymbol{0}, \quad \text{respectively } \langle\hat{\boldsymbol{\psi}}_n, \boldsymbol{I}_m\lambda^k\rangle=\boldsymbol{0}, \text{ for all } k\in\mathbb{N}_0. \tag{48}$$

Since (48) represents infinitely many conditions (representing the fact that $\boldsymbol{H}$ is an infinite matrix), in practice, one needs to replace (48) by some appropriate finite version. For example, imitating (47), one can check if, for all $0\leqslant k\leqslant k(n)$,

$$\|\langle \boldsymbol{I}_p\lambda^k, \hat{\boldsymbol{\phi}}_n\rangle\|\leqslant\texttt{dtol}_n^{\texttt{r}}, \quad \text{respectively } \|\langle\hat{\boldsymbol{\psi}}_n, \boldsymbol{I}_m\lambda^k\rangle\|\leqslant\texttt{dtol}_n^{\texttt{l}}, \tag{49}$$

where $k(n)$ is a sufficiently large, but finite integer.

We conclude this section with some comments on the choice of the deflation tolerances $\texttt{dtol}_n^{\texttt{r}}$ and $\texttt{dtol}_n^{\texttt{l}}$ in (47) and (49). Clearly, deflation of $\hat{\boldsymbol{\phi}}_n$ or $\hat{\boldsymbol{\psi}}_n$ should occur independent of the actual scaling of the problem. First, consider the special case of bilinear Hankel forms given in terms of a realization (12). To make the deflation check independent of the actual scaling of the columns $\boldsymbol{r}_j$ of $\boldsymbol{R}$, of the columns $\boldsymbol{l}_j$ of $\boldsymbol{L}$, and of the matrix $\boldsymbol{A}$, we use the tolerances

$$\begin{aligned}\texttt{dtol}_n^{\texttt{r}}&=\begin{cases}\texttt{dtol}\cdot\|\boldsymbol{r}_{\deg\hat{\boldsymbol{\phi}}_n}\| & \text{if } \deg\hat{\boldsymbol{\phi}}_n\leqslant m,\\ \texttt{dtol}\cdot\operatorname{nest}(\boldsymbol{A}) & \text{if } \deg\hat{\boldsymbol{\phi}}_n> m,\end{cases}\\ \texttt{dtol}_n^{\texttt{l}}&=\begin{cases}\texttt{dtol}\cdot\|\boldsymbol{l}_{\deg\hat{\boldsymbol{\psi}}_n}\| & \text{if } \deg\hat{\boldsymbol{\psi}}_n\leqslant p,\\ \texttt{dtol}\cdot\operatorname{nest}(\boldsymbol{A}) & \text{if } \deg\hat{\boldsymbol{\psi}}_n> p.\end{cases}\end{aligned} \tag{50}$$

Here, $\operatorname{nest}(\boldsymbol{A})$ is either $\|\boldsymbol{A}\|$ or an estimate of $\|\boldsymbol{A}\|$, and $\texttt{dtol}$ is an absolute deflation tolerance. Based on our extensive numerical experiences for the applications outlined in Sections 8.1–8.4 below, we recommend $\texttt{dtol}=\sqrt{\texttt{eps}}$, where $\texttt{eps}$, is the machine precision. In practical applications, the matrix $\boldsymbol{A}$ is often not available directly, and then the ideal choice $\operatorname{nest}(\boldsymbol{A})=\|\boldsymbol{A}\|$ in (50) is not feasible in general. However, matrix–vector products with $\boldsymbol{A}$ and $\boldsymbol{A}^{\mathrm{T}}$ can usually be computed efficiently. In this case, one evaluates quotients of the form

$$\frac{\|\boldsymbol{A}\boldsymbol{v}\|}{\|\boldsymbol{v}\|} \quad\text{or}\quad \frac{\|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{w}\|}{\|\boldsymbol{w}\|}$$

for a small number of vectors $\boldsymbol{v}\in\mathbb{C}^N$, $\boldsymbol{v}\neq\boldsymbol{0}$, or $\boldsymbol{w}\in\mathbb{C}^N, \boldsymbol{w}\neq\boldsymbol{0}$, and then takes the largest of these quotients as an estimate $\operatorname{nest}(\boldsymbol{A})$ of $\|\boldsymbol{A}\|$.

In the case of Hankel matrices $\boldsymbol{H}$ not given in terms of a realization (12), the deflation tolerances $\texttt{dtol}_n^{\texttt{r}}$, respectively $\texttt{dtol}_n^{\texttt{l}}$, in (49) are chosen similar to (50) as products of an absolute deflation tolerance $\texttt{dtol}$ and factors that take the actual scaling of the columns, respectively rows, of $\boldsymbol{H}$ into account.

5. Recurrence relations

In this section, we describe the recurrence relations that are used in our computational procedure for generating the polynomials (26). From now on, we always consider the general case where look-ahead may be needed, and we will point out simplifications that occur in the generic case.

5.1. Recurrences in matrix form

Recall that the matrix polynomials $\boldsymbol{\Phi}_n$ and $\boldsymbol{\Psi}_n$ introduced in (35) contain the first $n+1$ pairs,

$$\phi_0, \phi_1, \ldots, \phi_n \quad \text{and} \quad \psi_0, \psi_1, \ldots, \psi_n, \tag{51}$$

of the polynomials (26). Using the notation $\boldsymbol{\Phi}_n$ and $\boldsymbol{\Psi}_n$, the recurrences for generating all the polynomials (51) can be summarized compactly in matrix form as follows:

$$\begin{aligned} [\boldsymbol{I}_m \quad \lambda\boldsymbol{\Phi}_n] &= \boldsymbol{\Phi}_n \boldsymbol{T}_{n,m+n} + [\underbrace{\boldsymbol{0} \quad \cdots \quad \boldsymbol{0}}_{m+n+1-m_c} \quad \underbrace{\hat{\phi}_{n+1} \quad \cdots \quad \hat{\phi}_{n+m_c}}_{m_c}] + \hat{\boldsymbol{\Phi}}^{\text{defl}}_{m+n}, \\ [\boldsymbol{I}_p \quad \lambda\boldsymbol{\Psi}_n] &= \boldsymbol{\Psi}_n \tilde{\boldsymbol{T}}_{n,p+n} + [\underbrace{\boldsymbol{0} \quad \cdots \quad \boldsymbol{0}}_{p+n+1-p_c} \quad \underbrace{\hat{\psi}_{n+1} \quad \cdots \quad \hat{\psi}_{n+p_c}}_{p_c}] + \hat{\boldsymbol{\Psi}}^{\text{defl}}_{p+n}. \end{aligned} \tag{52}$$

Relations (52) hold true for all $n=-1,0,1,2,\ldots$, where $n \leqslant n_{\max}$. Here, we use the convention that $n=-1$ corresponds to the initialization of the first m right and p left auxiliary polynomials,

$$\hat{\phi}_0, \hat{\phi}_1, \ldots, \hat{\phi}_{m-1} \quad \text{and} \quad \hat{\psi}_0, \hat{\psi}_1, \ldots, \hat{\psi}_{p-1},$$

respectively, and we set $\boldsymbol{\Phi}_{-1} := \boldsymbol{\Psi}_{-1} := \emptyset$ and $\boldsymbol{T}_{-1,m-1} := \tilde{\boldsymbol{T}}_{-1,p-1} := \emptyset$ in (52). For $n \geqslant 0$, the matrices

$$\boldsymbol{T}_{n,m+n} = [t_{j,k}]_{0 \leqslant j \leqslant n, -m \leqslant k \leqslant n} \in \mathbb{C}^{(n+1)\times(m+n+1)} \tag{53a}$$

and

$$\tilde{\boldsymbol{T}}_{n,p+n} = [\tilde{t}_{j,k}]_{0 \leqslant j \leqslant n, -p \leqslant k \leqslant n} \in \mathbb{C}^{(n+1)\times(p+n+1)} \tag{53b}$$

contain the recurrence coefficients used for the right and left polynomials, respectively. Corresponding to the partitioning of the matrices on the left-hand sides of (52), the matrices (53a) and (53b) can be written in the form

$$\boldsymbol{T}_{n,m+n} = [\boldsymbol{\rho}_n \quad \boldsymbol{T}_n] \quad \text{and} \quad \tilde{\boldsymbol{T}}_{n,p+n} = [\boldsymbol{\eta}_n \quad \tilde{\boldsymbol{T}}_n], \tag{54}$$

where

$$\begin{aligned} \boldsymbol{\rho}_n &:= [t_{j,k}]_{0 \leqslant j \leqslant n, -m \leqslant k < 0} \in \mathbb{C}^{(n+1)\times m}, \quad \boldsymbol{T}_n = [t_{j,k}]_{0 \leqslant j,k \leqslant n} \in \mathbb{C}^{(n+1)\times(n+1)}, \\ \boldsymbol{\eta}_n &:= [\tilde{t}_{j,k}]_{0 \leqslant j \leqslant n, -p \leqslant k < 0} \in \mathbb{C}^{(n+1)\times p}, \quad \tilde{\boldsymbol{T}}_n = [\tilde{t}_{j,k}]_{0 \leqslant j,k \leqslant n} \in \mathbb{C}^{(n+1)\times(n+1)}. \end{aligned}$$

Finally, the matrix polynomials $\hat{\boldsymbol{\Phi}}^{\text{defl}}_{m+n}$ and $\hat{\boldsymbol{\Psi}}^{\text{defl}}_{p+n}$ in (52) contain mostly zero columns, together with the polynomials that have been deflated. Recall from Section 4.4 that in our algorithm the polynomials $\hat{\phi}_n$ and $\hat{\psi}_n$ are used to check for necessary deflation. If it is decided that $\hat{\phi}_n$ needs to be deflated, then $\hat{\boldsymbol{\phi}}_n$ is moved into $\hat{\boldsymbol{\Phi}}^{\mathrm{defl}}_{m+n}$ and becomes its column with index $m+n-m_{\mathrm{c}}$. Otherwise, $\hat{\boldsymbol{\phi}}_n$ is accepted as the next right polynomials $\boldsymbol{\phi}_n$. Similarly, each deflated left polynomials $\hat{\boldsymbol{\psi}}_n$ becomes the $(p+n-p_{\mathrm{c}})$th column of $\hat{\boldsymbol{\Psi}}^{\mathrm{defl}}_{p+n}$.

5.2. Structure of the recurrence matrices

Recall that the recurrence matrices for generating scalar formally orthogonal polynomials in the special case $m=p=1$ are tridiagonal in the generic case and block-tridiagonal in the general case; see, e.g., [14,16,17]. Similarly, the recurrence matrices (53a) and (53b), $\boldsymbol{T}_{n,m+n}$ and $\tilde{\boldsymbol{T}}_{n,p+n}$, exhibit certain structures, although, in the case of deflation, these structures are somewhat more complicated than those for $m=p=1$. In this subsection, we describe the structures of the matrices (53a) and (53b).

First, consider the simplest case that neither deflation nor look-ahead occur during the construction of the polynomials (51). In this case, $\boldsymbol{T}_{n,m+n}$ and $\tilde{\boldsymbol{T}}_{n,p+n}$ have a banded structure. More precisely, the entries of (53a) and (53b) satisfy

$$t_{j,k}=0 \quad \text{if } j>k+m \text{ or } k>j+p,$$

$$\tilde{t}_{j,k}=0 \quad \text{if } j>k+p \text{ or } k>j+m.$$

In terms of the partitionings (54), this means that $\boldsymbol{\rho}_n$ and $\boldsymbol{\eta}_n$ are upper triangular matrices, $\boldsymbol{T}_n$ is a banded matrix with lower bandwidth m and upper bandwidth p, and $\tilde{\boldsymbol{T}}_n$ is a banded matrix with lower bandwidth p and upper bandwidth m.

Next, consider the case that deflation occurs. Recall that we use the integers m_{c} and p_{c} to count deflations. More precisely, initially $m_{\mathrm{c}}=m$ and $p_{\mathrm{c}}=p$, and then m_{c}, respectively p_{c}, is reduced by one every time a right polynomial $\hat{\boldsymbol{\phi}}_n$, respectively a left polynomial $\hat{\boldsymbol{\psi}}_n$, is deflated. It turns out that m_{c} and p_{c} are also the "current" bandwidths of $\boldsymbol{T}_{n,m+n}$ and $\tilde{\boldsymbol{T}}_{n,p+n}$. This means that the deflation of a right polynomial $\hat{\boldsymbol{\phi}}_n$ reduces m_{c} and thus both the lower bandwidth of $\boldsymbol{T}_{n,m+n}$ and the upper bandwidth of $\tilde{\boldsymbol{T}}_{n,p+n}$ by one. Similarly, the deflation of a left polynomial $\hat{\boldsymbol{\psi}}_n$ reduces p_{c} and thus both the upper bandwidth of $\boldsymbol{T}_{n,m+n}$ and the lower bandwidth of $\tilde{\boldsymbol{T}}_{n,p+n}$ by one. In addition, deflation has a second effect. A deflation of $\hat{\boldsymbol{\phi}}_n$ implies that from now on, the matrix $\tilde{\boldsymbol{T}}_{n,p+n}$ will have additional potentially nonzero entries $\tilde{t}_{n-m_{\mathrm{c}},k}$ in row $n-m_{\mathrm{c}}$ and to the right of its banded part. These additional entries mean that from now on, all left polynomials need to be explicitly orthogonalized against the right polynomial $\boldsymbol{\phi}_{n-m_{\mathrm{c}}}$. Similarly, a deflation of $\hat{\boldsymbol{\psi}}_n$ implies that from now on, the matrix $\boldsymbol{T}_{n,m+n}$ will have additional potentially nonzero entries $t_{n-p_{\mathrm{c}},k}$ in row $n-p_{\mathrm{c}}$ and to the right of its banded part. These additional entries mean that from now on, all right polynomials need to be explicitly orthogonalized against the left polynomial $\boldsymbol{\psi}_{n-p_{\mathrm{c}}}$. The size of these additional entries in $\tilde{\boldsymbol{T}}_{n,p+n}$ and $\boldsymbol{T}_{n,m+n}$ can be shown to be bounded by

$$\max_{j=0,1,\dots,n} \mathtt{dtol}^{\mathtt{r}}_j \quad \text{and} \quad \max_{j=0,1,\dots,n} \mathtt{dtol}^{\mathtt{l}}_j,$$

respectively, where $\mathtt{dtol}^{\mathtt{r}}_j$ and $\mathtt{dtol}^{\mathtt{l}}_j$ are the tolerances used to check for deflation; see Section 4.4. In particular, these additional entries in $\tilde{\boldsymbol{T}}_{n,p+n}$ and $\boldsymbol{T}_{n,m+n}$ all reduce to zero if only exact deflation is performed, i.e., if $\mathtt{dtol}^{\mathtt{r}}_j=\mathtt{dtol}^{\mathtt{l}}_j=0$ for all j.

Finally, in the general case where both deflation and look-ahead occur, non-trivial look-ahead clusters, i.e., those with $n_{k+1} - n_k > 1$, result in "bulges" just above the banded parts of $\boldsymbol{T}_{n,m+n}$ and $\tilde{\boldsymbol{T}}_{n,p+n}$. In Algorithm 1 below, the cluster indices ℓ_ψ and ℓ_ϕ defined in (58) and (57) mark the first potentially nonzero elements of the banded parts, including the bulges due to look-ahead, in the nth column of $\boldsymbol{T}_{n,m+n}$ and $\tilde{\boldsymbol{T}}_{n,p+n}$, respectively. More precisely, $t_{n_{\ell_\psi},n}$ is the first potentially nonzero element in the nth column of $\boldsymbol{T}_{n,m+n}$, and $\tilde{t}_{n_{\ell_\phi},n}$ is the first potentially nonzero element in the nth column of $\tilde{\boldsymbol{T}}_{n,p+n}$. Furthermore, if the row indices $n - m_\mathrm{c}$, respectively $n - p_\mathrm{c}$, of the additional potentially nonzero elements due to deflation are part of a nontrivial look-ahead cluster, then these additional potentially nonzero elements are spread out over all rows corresponding to that look-ahead cluster. In Algorithm 1 below, the sets $\mathscr{D}_\psi$ and $\mathscr{D}_\phi$ of cluster indices are used to record these additional nonzero rows above the banded parts of $\boldsymbol{T}_{n,m+n}$ and $\tilde{\boldsymbol{T}}_{n,p+n}$, respectively.

At each nth pass through the main loop of Algorithm 1, based on ℓ_ψ and $\mathscr{D}_\psi$, we form the set $\mathscr{I}_\psi$ in Step (6b) and, based on ℓ_ϕ and $\mathscr{D}_\phi$, the set $\mathscr{I}_\phi$ in Step (7b). The set $\mathscr{I}_\psi$ in (59) contains the indices k of those clusters for which the associated entries $t_{j,n},\ n_k \leqslant j < n_{k+1}$ of column n of the matrix $\boldsymbol{T}_{n,m+n}$ are potentially nonzero. These are just the indices of the clusters $\boldsymbol{\Psi}^{(k)}$ of left polynomials against which the next right auxiliary polynomial $\hat{\boldsymbol{\phi}}_{n+m_\mathrm{c}}$ has to be explicitly orthogonalized. Note that the set $\mathscr{I}_\psi$ in (59) has two parts. The first part in (59) contains the cluster indices corresponding to spread-out rows due to deflation of earlier left polynomials, while the second part in (59) contains the cluster indices corresponding to the banded part of $\boldsymbol{T}_{n,m+n}$. Similarly, the set $\mathscr{I}_\phi$ in (61) contains the indices k of those clusters for which the associated entries $\tilde{t}_{j,n},\ n_k \leqslant j < n_{k+1}$ of column n of the matrix $\tilde{\boldsymbol{T}}_{n,p+n}$ are potentially nonzero. These are just the indices of the clusters $\boldsymbol{\Phi}^{(k)}$ of right polynomials against which the next left auxiliary polynomial $\hat{\boldsymbol{\psi}}_{n+p_\mathrm{c}}$ has to be explicitly orthogonalized. The first parts of the set $\mathscr{I}_\phi$ in (61) contains the cluster indices corresponding to spread-out rows due to deflation of earlier right polynomials, while the second part in (61) contains the cluster indices corresponding to the banded part of $\tilde{\boldsymbol{T}}_{n,m+n}$.

5.3. An example

In this subsection, we illustrate the structure of the recurrence matrices with an example.

Consider the case that $m = 4$ and $p = 5$, and assume that in the associated block Hankel matrix, the columns with index $k = 2+4i,\ 8+4i,\ 19+4i,\ i \in \mathbb{N}_0$, and the rows with index $j = 7+5i,\ 11+5i,\ 15+5i,\ i \in \mathbb{N}_0$, need to be deflated. The associated indices (18) of the columns and rows left after deflation are as follows:

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	$\cdots$
μ_n	0	1	3	4	5	7	9	11	13	15	17	21	25	29	33	$\cdots$
v_n	0	1	2	3	4	5	6	8	9	10	13	14	18	19	23	$\cdots$

In terms of the auxiliary polynomials (45), these deflations of columns and rows translate into deflations of the right auxiliary polynomials $\hat{\boldsymbol{\phi}}_2$ (when $m_\mathrm{c} = 4$), $\hat{\boldsymbol{\phi}}_6$ (when $m_\mathrm{c} = 3$), and $\hat{\boldsymbol{\phi}}_{11}$ (when $m_\mathrm{c} = 2$), and the left auxiliary polynomials $\hat{\boldsymbol{\psi}}_2$ (when $p_\mathrm{c} = 4$), $\hat{\boldsymbol{\psi}}_6$ (when $p_\mathrm{c} = 3$), and $\hat{\boldsymbol{\psi}}_{11}$ (when $p_\mathrm{c} = 2$).

$$
\mathrm{T}_{14,18} = \left[\begin{array}{cccc|ccccccccccccccc}
\bullet & * & * & * & * & * & * & * & * & * & 1 \\
 & \bullet & * & * & * & * & * & * & * & * & * \\
 & & & \bullet & * & * & * & * & * & * & * & \diamond & \diamond & \diamond & \diamond & \diamond & \diamond & \diamond & \diamond \\
 & & & & \bullet & * & * & * & * & * & * & * & 1 \\
 & & & & & \bullet & * & * & * & * & * & * & * & 1 & 1 & 1 & 1 & 1 & 1 \\
 & & & & & & \bullet & * & * & * & * & * & * & * & 1 & 1 & 1 & 1 & 1 \\
 & & & & & & & & \bullet & * & * & * & * & * & \diamond & \diamond & \diamond & \diamond & \diamond \\
 & & & & & & & & & \bullet & * & * & * & * & * & 1 \\
 & & & & & & & & & & \bullet & * & * & * & * & * \\
 & & & & & & & & & & & \bullet & * & * & * & * & \diamond & \diamond & \diamond \\
 & & & & & & & & & & & & \bullet & * & * & * & * \\
 & & & & & & & & & & & & & & \bullet & * & * & * \\
 & & & & & & & & & & & & & & & \bullet & * & * & * \\
 & & & & & & & & & & & & & & & & \bullet & * & * \\
 & & & & & & & & & & & & & & & & & \bullet & *
\end{array}\right]
$$

Fig. 1. Structure of $\boldsymbol{T}_{14,18}$.

$$
\tilde{\mathrm{T}}_{14,19} = \left[\begin{array}{ccccc|ccccccccccccccc}
\bullet & * & * & * & * & * & * & * & * \\
 & \bullet & * & * & * & * & * & * & * & * & 1 & 1 \\
 & & \bullet & * & * & * & * & * & * & * & * & 1 \\
 & & & \bullet & * & * & * & * & * & * & * & \diamond & \diamond & \diamond & \diamond & \diamond & \diamond & \diamond & \diamond & \diamond \\
 & & & & \bullet & * & * & * & * & * & * & * & 1 & 1 \\
 & & & & & \bullet & * & * & * & * & * & * & * & 1 \\
 & & & & & & \bullet & * & * & * & * & * & * & * \\
 & & & & & & & & \bullet & * & * & * & * & * & * & 1 \\
 & & & & & & & & & \bullet & * & * & * & * & * & * \\
 & & & & & & & & & & \bullet & * & * & * & * & * & \diamond & \diamond & \diamond & \diamond \\
 & & & & & & & & & & & & \bullet & * & * & * & * \\
 & & & & & & & & & & & & & \bullet & * & * & * & * \\
 & & & & & & & & & & & & & & & \bullet & * & * & * \\
 & & & & & & & & & & & & & & & & \bullet & * & * & * \\
 & & & & & & & & & & & & & & & & & \bullet & * & *
\end{array}\right]
$$

Fig. 2. Structure of $\tilde{\boldsymbol{T}}_{14,19}$.

We now determine the structure of the recurrence matrices $\boldsymbol{T}_{14,18}$ and $\tilde{\boldsymbol{T}}_{14,19}$ at $n = 14$. First, we assume that no look-ahead occurs. Recall from (44) that then $\gamma(i) = i$ for all i. Algorithm 1 produces the sets $\mathscr{D}_\phi = \{3, 9\}$ and $\mathscr{D}_\psi = \{6, 9\}$. Moreover, the values of the row indices ℓ_ϕ and ℓ_ψ given by (57) and (58) are as follows:

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
ℓ_ϕ	0	0	0	0	0	1	2	4	5	6	7	8	10	11	12	13
ℓ_ψ	0	0	0	0	0	0	0	1	3	4	5	7	8	10	11	12

The resulting structure of $\boldsymbol{T}_{14,18}$ and $\tilde{\boldsymbol{T}}_{14,19}$ is shown in Figs. 1 and 2, respectively. Here, the following convention is used: guaranteed positive elements are marked by "•", other potentially nonzero elements within the banded parts are marked by "$*$", and potentially nonzero elements outside the banded parts due to deflation are marked by "$\diamond$". Moreover, the vertical lines in Figs. 1 and 2 indicate the partitioning (54) of $\boldsymbol{T}_{14,18}$ into $\boldsymbol{\rho}_{14}$ and of $\boldsymbol{T}_{14}$, and $\tilde{\boldsymbol{T}}_{14,19}$ into $\boldsymbol{\eta}_{14}$ and $\tilde{\boldsymbol{T}}_{14}$, respectively.

Next, we assume that two non-trivial look-head clusters occur: one cluster of length $n_5 - n_4 = 3$ starting at $n_4 = 4$, followed by a cluster of length $n_6 - n_5 = 2$ starting at $n_5 = 7$. Algorithm 1 now produces the sets of cluster indices $\mathscr{D}_\phi = \{3, 6\}$ and $\mathscr{D}_\psi = \{2, 4, 6\}$, and the values of the cluster indices ℓ_ϕ and ℓ_ψ given by (57) and (58) are as follows:

k	0	1	2	3	4	5	6	7	8	9	10	11
n_k	0	1	2	3	4	7	9	10	11	12	13	14
ℓ_ϕ	0	0	0	0	1	4	5	5	7	8	9	10
ℓ_ψ	0	0	0	0	0	3	4	5	5	7	8	9

The resulting structure of $\boldsymbol{T}_{14,18}$ and $\tilde{\boldsymbol{T}}_{14,19}$ is again shown in Figs. 1 and 2, respectively, where we have used "1" to mark the additional potentially nonzero entries caused by the two non-trivial look-ahead clusters.

6. The algorithm

In this section, we present a detailed statement of the complete computational procedure for constructing formally orthogonal polynomials associated with a given bilinear Hankel form.

At pass n through the main loop of Algorithm 1 below, we construct the nth pair of polynomials, $\boldsymbol{\phi}_n$ and $\boldsymbol{\psi}_n$. We use the counter ℓ to denote the index of the clusters to which $\boldsymbol{\phi}_n$ and $\boldsymbol{\psi}_n$ are added. This means that, at pass n, the currently constructed look-ahead clusters are

$$\boldsymbol{\Phi}^{(\ell)} := [\boldsymbol{\phi}_{n_\ell} \quad \cdots \quad \boldsymbol{\phi}_n] \quad \text{and} \quad \boldsymbol{\Psi}^{(\ell)} := [\boldsymbol{\psi}_{n_\ell} \quad \cdots \quad \boldsymbol{\psi}_n].$$

We also check if these look-ahead clusters are complete. If they are, then the polynomials $\boldsymbol{\phi}_{n+1}$ and $\boldsymbol{\psi}_{n+1}$ constructed during the next, $(n+1)$st, pass start new clusters. All the other notation used in the following statement of Algorithm 1 has already been introduced.

Algorithm 1 (Construction of polynomials associated with $\langle\cdot,\cdot\rangle$.).

```
INPUT: A bilinear Hankel form $\langle\cdot,\cdot\rangle: \mathscr{P}^{(p)} \times \mathscr{P}^{(m)} \mapsto \mathbb{C}$.
(0) Set $\hat{\boldsymbol{\phi}}_i = \boldsymbol{e}_{i+1}^{(m)}$ and $\hat{\mu}_i = i$   for $i = 0, 1, \ldots, m-1$.
    Set $\hat{\boldsymbol{\psi}}_i = \boldsymbol{e}_{i+1}^{(p)}$ and $\hat{\nu}_i = i$   for $i = 0, 1, \ldots, p-1$.
    Set $m_c = m$ and $p_c = p$.
    Set $\mathscr{D}_\phi = \mathscr{D}_\psi = \emptyset$ and $\ell_\phi = \ell_\psi = 0$.
    Set $\ell = 0$, $n_0 = 0$, and $\boldsymbol{\Phi}^{(0)} = \boldsymbol{\Psi}^{(0)} = \emptyset$.
For $n = 0, 1, 2, \ldots,$ do:
(1) (If necessary, deflate $\hat{\boldsymbol{\phi}}_n$.)
    Decide if $\hat{\boldsymbol{\phi}}_n$ needs to be deflated.
    If no, continue with Step (2).
    If yes, deflate $\hat{\boldsymbol{\phi}}_n$ by doing the following:
    (a) If $m_c = 1$, then stop.
        (There are no more right orthogonal polynomials.)
```

(b) *If* $n \geqslant m_{\mathrm{c}}$, *set* $\mathscr{D}_\phi = \mathscr{D}_\phi \cup \{\gamma(n-m_{\mathrm{c}})\}$.
($\mathscr{D}_\phi$ records the cluster indices of right polynomials against which we have to explicitly orthogonalize from now on.)

(c) *Set* $\hat{\boldsymbol{\phi}}_i = \hat{\boldsymbol{\phi}}_{i+1}$ *and* $\hat{\mu}_i = \hat{\mu}_{i+1}$ *for* $i = n, n+1, \ldots, n+m_{\mathrm{c}}-2$.
(The polynomial $\hat{\boldsymbol{\phi}}_n$ is deflated and becomes the $(m+n-m_{\mathrm{c}})$th column of the matrix $\hat{\boldsymbol{\Phi}}^{\mathrm{defl}}_{m+n}$ in (52); the indices of the remaining right auxiliary polynomials are reduced by one.)

(d) *Set* $m_{\mathrm{c}} = m_{\mathrm{c}} - 1$.
(The current right block size is reduced by one.)

(e) *Repeat all of Step* (1).

(2) (If necessary, deflate $\hat{\boldsymbol{\psi}}_n$.)
Decide if $\hat{\boldsymbol{\psi}}_n$ *needs to be deflated.*
If no, continue with Step (3).
If yes, deflate $\hat{\boldsymbol{\psi}}_n$ *by doing the following*:

(a) *If* $p_{\mathrm{c}} = 1$, *then stop.*
(There are no more left orthogonal polynomials).

(b) *If* $n \geqslant p_{\mathrm{c}}$, *set* $\mathscr{D}_\psi = \mathscr{D}_\psi \cup \{\gamma(n-p_{\mathrm{c}})\}$.
($\mathscr{D}_\psi$ records the cluster indices of left polynomials against which we have to explicitly orthogonalize from now on.)

(c) *Set* $\hat{\boldsymbol{\psi}}_i = \hat{\boldsymbol{\psi}}_{i+1}$ *and* $\hat{v}_i = \hat{v}_{i+1}$ *for* $i = n, n+1, \ldots, n+p_{\mathrm{c}}-2$.
(The polynomial $\hat{\boldsymbol{\psi}}_n$ is deflated and becomes the $(p+n-p_{\mathrm{c}})$th column of the matrix $\boldsymbol{\Psi}^{\mathrm{defl}}_{p+n}$ in (52); the indices of the remaining left auxiliary polynomials are reduced by one.)

(d) *Set* $p_{\mathrm{c}} = p_{\mathrm{c}} - 1$.
(The current right block size is reduced by one.)

(e) *Repeat all of Step* (2).

(3) (Normalize $\hat{\boldsymbol{\phi}}_n$ and $\hat{\boldsymbol{\psi}}_n$ to obtain $\boldsymbol{\phi}_n$ and $\boldsymbol{\psi}_n$, and add them to the current clusters $\boldsymbol{\phi}^{(\ell)}$ and $\boldsymbol{\psi}^{(\ell)}$.)
Set

$$\boldsymbol{\phi}_n = \frac{\hat{\boldsymbol{\phi}}_n}{t_{n,n-m_{\mathrm{c}}}}, \quad \mu_n = \hat{\mu}_n \quad \textit{and} \quad \boldsymbol{\psi}_n = \frac{\hat{\boldsymbol{\psi}}_n}{\tilde{t}_{n,n-p_{\mathrm{c}}}}, \quad v_n = \hat{v}_n,$$

where $t_{n,n-m_{\mathrm{c}}} > 0$ *and* $\tilde{t}_{n,n-p_{\mathrm{c}}} > 0$ *are suitable scaling factors.*
Set $\boldsymbol{\Phi}^{(\ell)} = [\boldsymbol{\Phi}^{(\ell)} \;\; \boldsymbol{\phi}_n]$ *and* $\boldsymbol{\Psi}^{(\ell)} = [\boldsymbol{\Psi}^{(\ell)} \;\; \boldsymbol{\psi}_n]$.

(4) (Compute $\Delta^{(\ell)}$ and check for end of look-ahead cluster.)
Form the matrix $\Delta^{(\ell)} = \langle \boldsymbol{\Psi}^{(\ell)}, \boldsymbol{\Phi}^{(\ell)} \rangle$.
If the matrix $\Delta^{(\ell)}$ *is singular or in some sense* "*close*" *to singular, continue with Step* (6).

(5) (The ℓth look-ahead clusters $\boldsymbol{\Phi}^{(\ell)}$ and $\boldsymbol{\Psi}^{(\ell)}$ are complete and the following "end-of-clusters" updates are performed.)

(a) (Orthogonalize the polynomials $\hat{\boldsymbol{\phi}}_{n+1}, \hat{\boldsymbol{\phi}}_{n+2}, \ldots, \hat{\boldsymbol{\phi}}_{n+m_{\mathrm{c}}-1}$ against $\boldsymbol{\Psi}^{(\ell)}$.)
For $i = n+1, n+2, \ldots, n+m_{\mathrm{c}}-1$, *set*

$$\begin{aligned} [t_{j,i-m_{\mathrm{c}}}]_{n_\ell \leqslant j \leqslant n} &= (\Delta^{(\ell)})^{-1} \langle \boldsymbol{\Psi}^{(\ell)}, \hat{\boldsymbol{\phi}}_i \rangle, \\ \hat{\boldsymbol{\phi}}_i &= \hat{\boldsymbol{\phi}}_i - \boldsymbol{\Phi}^{(\ell)} [t_{j,i-m_{\mathrm{c}}}]_{n_\ell \leqslant j \leqslant n}. \end{aligned} \tag{55}$$

(b) (Orthogonalize the polynomials $\hat{\boldsymbol{\psi}}_{n+1}, \hat{\boldsymbol{\psi}}_{n+2}, \ldots, \hat{\boldsymbol{\psi}}_{n+p_\mathrm{c}-1}$ against $\boldsymbol{\Phi}^{(\ell)}$.)
For $i = n+1, n+2, \ldots, n+p_\mathrm{c}-1$, *set*

$$\begin{aligned}
&[\tilde{t}_{j,i-p_\mathrm{c}}]_{n_\ell \leqslant j \leqslant n} = (\boldsymbol{\Delta}^{(\ell)})^{-\mathrm{T}} \langle \hat{\boldsymbol{\psi}}_i, \boldsymbol{\Phi}^{(\ell)} \rangle^{\mathrm{T}},\\
&\hat{\boldsymbol{\psi}}_i = \hat{\boldsymbol{\psi}}_i - \boldsymbol{\Psi}^{(\ell)} [\tilde{t}_{j,i-p_\mathrm{c}}]_{n_\ell \leqslant j \leqslant n}.
\end{aligned} \tag{56}$$

(c) *If* $\mu_{n_\ell} \geqslant m$, *set*

$$\ell_{\boldsymbol{\phi}} = \gamma(\mu^{-1}(\mu_{n_\ell} - m)). \tag{57}$$

If $\nu_{n_\ell} \geqslant p$, *set*

$$\ell_{\boldsymbol{\psi}} = \gamma(\nu^{-1}(\nu_{n_\ell} - p)). \tag{58}$$

(d) *Set* $\ell = \ell + 1$, $n_\ell = n+1$, *and* $\boldsymbol{\Phi}^{(\ell)} = \boldsymbol{\Psi}^{(\ell)} = \emptyset$.
(The polynomials $\boldsymbol{\phi}_{n+1}$ and $\boldsymbol{\psi}_{n+1}$ constructed in the next iteration start new look-ahead clusters.)

(6) Obtain new right polynomial $\hat{\boldsymbol{\phi}}_{n+m_\mathrm{c}}$ and make it orthogonal to complete right clusters.)

(a) *Set* $\hat{\boldsymbol{\phi}}_{n+m_\mathrm{c}} = \lambda \boldsymbol{\phi}_n$ *and* $\hat{\mu}_{n+m_\mathrm{c}} = \mu_n + m$.

(b) (Determine the indices of the left clusters $\boldsymbol{\Psi}^{(k)}$ against which $\hat{\boldsymbol{\phi}}_{n+m_\mathrm{c}}$ needs to be orthogonalized.)
Set

$$\mathscr{I}_{\boldsymbol{\psi}} = \{k \mid k \in \mathscr{D}_{\boldsymbol{\psi}} \text{ and } k < \ell_{\boldsymbol{\psi}}\} \cup \{\ell_{\boldsymbol{\psi}}, \ell_{\boldsymbol{\psi}}+1, \ldots, \ell-1\}. \tag{59}$$

(c) (Orthogonalize $\hat{\boldsymbol{\phi}}_{n+m_\mathrm{c}}$ against these clusters.)
For all $k \in \mathscr{I}_{\boldsymbol{\psi}}$ (*in ascending order*), *set*

$$\begin{aligned}
&[t_{j,n}]_{n_k \leqslant j < n_{k+1}} = (\boldsymbol{\Delta}^{(k)})^{-1} \langle \boldsymbol{\Psi}^{(k)}, \hat{\boldsymbol{\phi}}_{n+m_\mathrm{c}} \rangle,\\
&\hat{\boldsymbol{\phi}}_{n+m_\mathrm{c}} = \hat{\boldsymbol{\phi}}_{n+m_\mathrm{c}} - \boldsymbol{\Phi}^{(k)} [t_{j,n}]_{n_k \leqslant j < n_{k+1}}.
\end{aligned} \tag{60}$$

(7) (Obtain new left polynomial $\hat{\boldsymbol{\psi}}_{n+p_\mathrm{c}}$ and make it orthogonal to complete right clusters.)

(a) *Set* $\hat{\boldsymbol{\psi}}_{n+p_\mathrm{c}} = \lambda \boldsymbol{\psi}_n$ *and* $\hat{\nu}_{n+p_\mathrm{c}} = \nu_n + p$.

(b) (Determine the indices of the right clusters $\boldsymbol{\Phi}^{(k)}$ against which $\hat{\boldsymbol{\psi}}_{n+p_\mathrm{c}}$ needs to be orthogonalized.)
Set

$$\mathscr{I}_{\boldsymbol{\phi}} = \{k \mid k \in \mathscr{D}_{\boldsymbol{\phi}} \text{ and } k < \ell_{\boldsymbol{\phi}}\} \cup \{\ell_{\boldsymbol{\phi}}, \ell_{\boldsymbol{\phi}}+1, \ldots, \ell-1\}. \tag{61}$$

(c) (Orthogonalize $\hat{\boldsymbol{\psi}}_{n+p_\mathrm{c}}$ against these clusters.)
For all $k \in \mathscr{I}_{\boldsymbol{\phi}}$ (*in ascending order*), *set*

$$\begin{aligned}
&[\tilde{t}_{j,n}]_{n_k \leqslant j < n_{k+1}} = (\boldsymbol{\Delta}^{(k)})^{-\mathrm{T}} \langle \hat{\boldsymbol{\psi}}_{n+p_\mathrm{c}}, \boldsymbol{\Phi}^{(k)} \rangle^{\mathrm{T}},\\
&\hat{\boldsymbol{\psi}}_{n+p_\mathrm{c}} = \hat{\boldsymbol{\psi}}_{n+p_\mathrm{c}} - \boldsymbol{\Psi}^{(k)} [\tilde{t}_{j,n}]_{n_k \leqslant j < n_{k+1}}.
\end{aligned} \tag{62}$$

In the following, let $\boldsymbol{T}_{n,m+n}$ and $\tilde{\boldsymbol{T}}_{n,p+n}$ be the matrices (53a) and (53b) with entries $t_{j,k}$ and $\tilde{t}_{j,k}$ generated by Algorithm 1. Here we use the convention that entries that are not explicitly generated in Algorithm 1 are set to be zero.

7. Properties

In the following theorem, we summarize the key properties of Algorithm 1.

Theorem 4 (Properties of Algorithm 1).

(a) *For each $n = -1, 0, 1, 2, \ldots,$ the polynomials and matrices that have been generated after the nth pass through the main loop of Algorithm* 1, *satisfy the recurrence relations* (52).

(b) *For each $n = 0, 1, 2, \ldots,$ the polynomials that have been generated after the nth pass through the main loop of Algorithm* 1, *satisfy*

$$\begin{aligned} \boldsymbol{\Phi}_n &:= [\boldsymbol{\phi}_0 \quad \boldsymbol{\phi}_1 \quad \cdots \quad \boldsymbol{\phi}_n] = [\boldsymbol{\Phi}^{(0)} \quad \boldsymbol{\Phi}^{(1)} \quad \cdots \quad \boldsymbol{\Phi}^{(\ell)}], \\ \boldsymbol{\Psi}_n &:= [\boldsymbol{\psi}_0 \quad \boldsymbol{\psi}_1 \quad \cdots \quad \boldsymbol{\psi}_n] = [\boldsymbol{\Psi}^{(0)} \quad \boldsymbol{\Psi}^{(1)} \quad \cdots \quad \boldsymbol{\Psi}^{(\ell)}], \end{aligned} \tag{63}$$

and the cluster-wise the orthogonality conditions

$$\begin{aligned} &\langle \boldsymbol{\Psi}_n, \boldsymbol{\Phi}_n \rangle = \boldsymbol{\Delta}_n := \operatorname{diag}(\boldsymbol{\Delta}^{(0)}, \boldsymbol{\Delta}^{(1)}, \ldots, \boldsymbol{\Delta}^{(\ell)}), \\ &\langle \boldsymbol{\Psi}_{n_\ell - 1}, \hat{\boldsymbol{\phi}}_{n+i} \rangle = \mathbf{0}, \quad i = 1, 2, \ldots, m_{\mathrm{c}}, \\ &\langle \hat{\boldsymbol{\psi}}_{n+i}, \boldsymbol{\Phi}_{n_\ell - 1} \rangle = \mathbf{0}, \quad i = 1, 2, \ldots, p_{\mathrm{c}}, \end{aligned} \tag{64}$$

where $\ell = \gamma(n+1)$.

Proof (Sketch). Part (a), as well as the partitioning property (63), can be directly verified.

The cluster-wise orthogonality conditions (64) are proved using induction on n. By the induction hypothesis, before Step (5a) in Algorithm 1 is performed, the right auxiliary polynomials $\hat{\boldsymbol{\phi}}_{n+1}, \hat{\boldsymbol{\phi}}_{n+2}, \ldots, \hat{\boldsymbol{\phi}}_{n+m_{\mathrm{c}}-1}$ are already orthogonal to all left clusters $\boldsymbol{\Psi}^{(k)}$ with $0 \leqslant k < \ell$. Thus it only remains to orthogonalize these polynomials against $\boldsymbol{\Psi}^{(\ell)}$, and this is obviously achieved by the update (55). Similarly, the update (56) is sufficient to ensure that the left auxiliary polynomials $\hat{\boldsymbol{\psi}}_{n+1}, \hat{\boldsymbol{\psi}}_{n+2}, \ldots, \hat{\boldsymbol{\psi}}_{n+p_{\mathrm{c}}-1}$ are orthogonal against all right clusters $\boldsymbol{\Phi}^{(k)}$ with $0 \leqslant k \leqslant \ell$. Next, consider the update (60) of $\hat{\boldsymbol{\phi}}_{n+m_{\mathrm{c}}}$. Here, we need to show that for the 'omitted' clusters $\boldsymbol{\Psi}^{(k)}$, $k \notin \mathscr{I}_\psi$, we have

$$\langle \boldsymbol{\Psi}^{(k)}, \hat{\boldsymbol{\phi}}_{n+m_{\mathrm{c}}} \rangle = [\langle \boldsymbol{\psi}_j, \hat{\boldsymbol{\phi}}_{n+m_{\mathrm{c}}} \rangle]_{n_k \leqslant j < n_{k+1}} = \mathbf{0}.$$

To this end, assume that $k \notin \mathscr{I}_\psi$ and let $n_k \leqslant j < n_{k+1}$. Since $\hat{\boldsymbol{\phi}}_{n+m_{\mathrm{c}}} = \lambda \boldsymbol{\phi}_n$ and using the shift property (6) of the bilinear Hankel form, it follows that

$$\langle \boldsymbol{\psi}_j, \hat{\boldsymbol{\phi}}_{n+m_{\mathrm{c}}} \rangle = \langle \lambda \boldsymbol{\psi}_j, \boldsymbol{\phi}_n \rangle. \tag{65}$$

The polynomial $\lambda\psi_j$ itself was used at the earlier pass j through the main loop of Algorithm 1 to obtain the auxiliary polynomial $\hat{\psi}_{j+p_c(j)}$. Now, there are two cases. The first one is that $\hat{\psi}_{j+p_c(j)}$ was deflated later on; in this case, however, $k \in \mathscr{D}_\psi$, and thus, by (59), $k \in \mathscr{I}_\psi$, which contradicts our assumption. This leaves the second case that $\hat{\psi}_{j+p_c(j)}$ was not deflated. In this case, $\hat{\psi}_{j+p_c(j)}$ was orthogonalized to become $\psi_{j'}$ and was added to the cluster with index $k' = \gamma(j')$ at pass j'. One can show that ℓ_ψ in (58) is just chosen such that $k < \ell_\psi$ implies $k' < \ell$. Thus the polynomials $\psi_{j'}$ and ϕ_n are part of clusters with different indices, and together with (65), it follows that

$$\langle \psi_j, \hat{\phi}_{n+m_c} \rangle = \langle \lambda\psi_j, \phi_n \rangle = \langle \psi_{j'}, \phi_n \rangle = 0$$

for all $n_k \leqslant j < n_{k+1}$ with $k \notin \mathscr{I}_\psi$.

Similarly, one shows that the clusters with $k \notin \mathscr{I}_\phi$ can indeed be omitted in (62). □.

8. Applications

In this section, we sketch some applications of Algorithm 1. In the following, we assume that $\boldsymbol{A}, \boldsymbol{R}, \boldsymbol{L}$ is a given triplet of matrices of the form (12).

8.1. A Lanczos-type algorithm for multiple starting vectors

The first application is a Lanczos-type method that extends the classical Lanczos process [19] to multiple right and left starting vectors. We denote by $\mathscr{K}_n(\boldsymbol{A}, \boldsymbol{R})$ the *nth right block Krylov subspace* spanned by the first $n+1$ columns of the deflated right block Krylov matrix (16), and by $\mathscr{K}_n(\boldsymbol{A}^{\mathrm{T}}, \boldsymbol{L})$ the *nth left block Krylov subspace* spanned by the first $n+1$ columns of the deflated left block Krylov matrix (17). The goal of the Lanczos-type method is to generate bi-orthogonal basis vectors for $\mathscr{K}_n(\boldsymbol{A}, \boldsymbol{R})$ and $\mathscr{K}_n(\boldsymbol{A}^{\mathrm{T}}, \boldsymbol{L})$. To this end, using (14), we associate with the polynomials generated by Algorithm 1 the so-called right and left Lanczos vectors,

$$\boldsymbol{v}_n = \phi_n(\boldsymbol{A}) \circ \boldsymbol{R} \quad \text{and} \quad \boldsymbol{w}_n = \psi_n(\boldsymbol{A}^{\mathrm{T}}) \circ \boldsymbol{L}, \quad n = 0, 1, \ldots, \tag{66}$$

respectively. Then, by re-stating Algorithm 1 in terms of the vectors (66), instead of polynomials, we obtain the desired Lanczos-type method for multiple right and left starting vectors. In fact, the resulting algorithm is a look-ahead version of the Lanczos-type method stated in [10, Algorithm 9.2], and it can also be viewed as a variant of the Lanczos-type method proposed in [1].

It is easy to verify that the right and left Lanczos vectors (66) indeed span the right and left block Krylov subspaces, i.e.,

$$\operatorname{span}\{\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\} = \mathscr{K}_n(\boldsymbol{A}, \boldsymbol{R}) \tag{67a}$$

and

$$\operatorname{span}\{\boldsymbol{w}_0, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_n\} = \mathscr{K}_n(\boldsymbol{A}^{\mathrm{T}}, \boldsymbol{L}). \tag{67b}$$

Now let $\boldsymbol{V}_n := [\boldsymbol{v}_0\ \boldsymbol{v}_1\ \cdots\ \boldsymbol{v}_n]$ and $\boldsymbol{W}_n := [\boldsymbol{w}_0\ \boldsymbol{w}_1\ \cdots\ \boldsymbol{w}_n]$ denote the matrices whose columns are the first $n+1$ right and left Lanczos vectors. The polynomial recurrences (52) then translate into the following compact formulation of the recurrences used to generate the Lanczos vectors:

$$\begin{aligned}
[\boldsymbol{R}\quad \boldsymbol{A}\boldsymbol{V}_n] &= \boldsymbol{V}_n\boldsymbol{T}_{n,m+n} + [\underbrace{\boldsymbol{0}\ \ \cdots\ \ \boldsymbol{0}}_{m+n+1-m_c}\ \ \underbrace{\hat{\boldsymbol{v}}_n\ \ \cdots\ \ \hat{\boldsymbol{v}}_{n+m_c}}_{m_c}] + \hat{\boldsymbol{V}}^{\text{defl}}_{m+n},\\
[\boldsymbol{L}\quad \boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}_n] &= \boldsymbol{W}_n\tilde{\boldsymbol{T}}_{n,p+n} + [\underbrace{\boldsymbol{0}\ \ \cdots\ \ \boldsymbol{0}}_{p+n+1-p_c}\ \ \underbrace{\hat{\boldsymbol{w}}_{n+1}\ \ \cdots\ \ \hat{\boldsymbol{w}}_{n+p_c}}_{p_c}] + \hat{\boldsymbol{W}}^{\text{defl}}_{p+n}.
\end{aligned} \tag{68}$$

Here, the only possible nonzero columns of $\hat{\boldsymbol{V}}^{\text{defl}}_{m+n}$ and $\hat{\boldsymbol{W}}^{\text{defl}}_{p+n}$ are the columns containing deflated vectors. Moreover, in view of (47), even these columns are zero and thus

$$\hat{\boldsymbol{V}}^{\text{defl}}_{m+n} = \boldsymbol{0} \quad \text{and} \quad \hat{\boldsymbol{W}}^{\text{defl}}_{p+n} = \boldsymbol{0} \tag{69}$$

if only exact deflation is performed. Finally, the orthogonality conditions (64) translate into the following cluster-wise bi-orthogonality relations of the Lanczos vectors:

$$\begin{aligned}
&\boldsymbol{W}_n^{\mathrm{T}}\boldsymbol{V}_n = \boldsymbol{\Delta}_n := \operatorname{diag}(\boldsymbol{\Delta}^{(0)}, \boldsymbol{\Delta}^{(1)}, \ldots, \boldsymbol{\Delta}^{(\ell)}),\\
&\boldsymbol{W}_{n_\ell-1}^{\mathrm{T}}\hat{\boldsymbol{v}}_{n+i} = 0, \quad i = 1, 2, \ldots, m_c,\\
&\boldsymbol{V}_{n_\ell-1}^{\mathrm{T}}\hat{\boldsymbol{w}}_{n+i} = 0, \quad i = 1, 2, \ldots, p_c.
\end{aligned} \tag{70}$$

Here, $\ell = \gamma(n+1)$.

In the remainder of this paper, we always assume that n corresponds to the end of the ℓth look-ahead cluster, i.e., $n = n_\ell - 1$. This condition guarantees that, in (70), all blocks $\boldsymbol{\Delta}^{(k)}$ of the block-diagonal matrix $\boldsymbol{\Delta}_n$ and thus $\boldsymbol{\Delta}_n$ itself are nonsingular.

By multiplying the first relation in (68) from the left by $\boldsymbol{\Delta}_n^{-1}\boldsymbol{W}_n^{\mathrm{T}}$ and using the bi-orthogonality relations (70), as well as the partitioning of $\boldsymbol{T}_{n,m+n}$ in (54), we obtain

$$\begin{aligned}
[\boldsymbol{\Delta}_n^{-1}\boldsymbol{W}_n^{\mathrm{T}}\boldsymbol{R}\quad \boldsymbol{\Delta}_n^{-1}\boldsymbol{W}_n^{\mathrm{T}}\boldsymbol{A}\ \boldsymbol{V}_n] &= [\boldsymbol{\rho}_n\quad \boldsymbol{T}_n] + \boldsymbol{\Delta}_n^{-1}\boldsymbol{W}_n^{\mathrm{T}}\hat{\boldsymbol{V}}^{\text{defl}}_{m+n}\\
&=: [\boldsymbol{\rho}_n^{\text{proj}}\quad \boldsymbol{T}_n^{\text{proj}}].
\end{aligned} \tag{71}$$

Similarly, by multiplying the second relation in (68) from the left by $\boldsymbol{\Delta}_n^{-\mathrm{T}}\boldsymbol{V}_n^{\mathrm{T}}$, we get

$$\begin{aligned}
[\boldsymbol{\Delta}_n^{-\mathrm{T}}\boldsymbol{V}_n^{\mathrm{T}}\boldsymbol{L}\quad \boldsymbol{\Delta}_n^{-\mathrm{T}}\boldsymbol{V}_n^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\ \boldsymbol{W}_n] &= [\boldsymbol{\eta}_n\quad \tilde{\boldsymbol{T}}_n] + \boldsymbol{\Delta}_n^{-\mathrm{T}}\boldsymbol{V}_n^{\mathrm{T}}\hat{\boldsymbol{W}}^{\text{defl}}_{p+n}\\
&=: [\boldsymbol{\eta}_n^{\text{proj}}\quad \tilde{\boldsymbol{T}}_n^{\text{proj}}].
\end{aligned}$$

Recall that $\hat{\boldsymbol{V}}^{\text{defl}}_{m+n}$ and $\hat{\boldsymbol{W}}^{\text{defl}}_{p+n}$ contain mostly zero columns, together with the deflated vectors. This can be used to show that the matrices $\boldsymbol{T}_n$ and $\tilde{\boldsymbol{T}}_n$ and the *projected* versions $\boldsymbol{T}_n^{\text{proj}}$ and $\tilde{\boldsymbol{T}}_n^{\text{proj}}$ differ only in a few entries in their lower triangular parts, respectively. Furthermore, from (71) and (72), it follows that

$$\boldsymbol{W}_n^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}_n = \boldsymbol{\Delta}_n\boldsymbol{T}_n^{\text{proj}} = (\tilde{\boldsymbol{T}}_n^{\text{proj}})^{\mathrm{T}}\boldsymbol{\Delta}_n. \tag{73}$$

This relation implies that $\boldsymbol{T}_n^{\text{proj}}$ can be generated directly from only $\boldsymbol{T}_n$, $\tilde{\boldsymbol{T}}_n$, and $\boldsymbol{\Delta}_n$, without using the term $\boldsymbol{\Delta}_n^{-1}\boldsymbol{W}_n^{\mathrm{T}}\boldsymbol{V}^{\text{defl}}_{m+n}$ in (71).

For later use, we note that in the case of exact deflation, by (69), (71), and (72),

$$\boldsymbol{\rho}_n = \boldsymbol{\Delta}_n^{-1} \boldsymbol{W}_n^{\mathrm{T}} \boldsymbol{R}, \quad \boldsymbol{\eta}_n = \boldsymbol{\Delta}_n^{-\mathrm{T}} \boldsymbol{V}_n^{\mathrm{T}} \boldsymbol{L}, \quad \boldsymbol{T}_n = \boldsymbol{\Delta}_n^{-1} \boldsymbol{W}_n^{\mathrm{T}} \boldsymbol{A} \boldsymbol{V}_n. \tag{74}$$

Next, we describe three applications of the Lanczos-type method sketched in this subsection.

8.2. Padé approximation of matrix-valued transfer functions

The matrix triplet (12) induces the $(p \times m)$-matrix-valued transfer function

$$\boldsymbol{Z}(s) \equiv \boldsymbol{L}^{\mathrm{T}}(\boldsymbol{I} - s\boldsymbol{A})^{-1}\boldsymbol{R}. \tag{75}$$

The matrix size N in (12) is called the *state-space dimension* of (75). Let $n < N$, and consider $(p \times m)$-matrix-valued transfer functions

$$\boldsymbol{Z}_{n+1}(s) \equiv \boldsymbol{L}_n^{\mathrm{T}}(\boldsymbol{I} - s\boldsymbol{G}_n)^{-1}\boldsymbol{R}_n, \tag{76}$$

where $\boldsymbol{G}_n \in \mathbb{C}^{(n+1)\times(n+1)}, \boldsymbol{R}_n \in \mathbb{C}^{(n+1)\times m}$, and $\boldsymbol{L}_n \in \mathbb{C}^{(n+1)\times p}$. Note that (76) is a transfer function of the same form as (75), but with smaller state-space dimension $n+1$, instead of N. A function of the form (76) is said to be an $(n+1)$st *matrix-Padé approximant* of $\boldsymbol{Z}$ (about the expansion point $s_0 = 0$) if the matrices $\boldsymbol{G}_n, \boldsymbol{R}_n$ and $\boldsymbol{L}_n$ are such that

$$\boldsymbol{Z}_{n+1}(s) = \boldsymbol{Z}(s) + \mathcal{O}(s^{q(n)}),$$

where $q(n)$ is as large as possible.

It turns out that, for the case of exact deflation, an $(n+1)$st matrix-Padé approximant can be obtained by a suitable two-sided projection of $\boldsymbol{Z}$ onto the nth block Krylov subspaces $\mathscr{K}_n(\boldsymbol{A}, \boldsymbol{R})$ and $\mathscr{K}_n(\boldsymbol{A}^{\mathrm{T}}, \boldsymbol{L})$. Recall from (67) that these subspaces are spanned by the columns of the matrices $\boldsymbol{V}_n$ and $\boldsymbol{W}_n$. In terms of $\boldsymbol{V}_n$ and $\boldsymbol{W}_n$, the two-sided projection of (75) is as follows:

$$\boldsymbol{Z}_{n+1}(s) \equiv (\boldsymbol{V}_n^{\mathrm{T}}\boldsymbol{L})^{\mathrm{T}}(\boldsymbol{W}_n^{\mathrm{T}}\boldsymbol{V}_n - s\boldsymbol{W}_n^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}_n)^{-1}(\boldsymbol{W}_n^{\mathrm{T}}\boldsymbol{R}). \tag{77}$$

Using the first relation in (70), as well as (74), we can re-write (77) in the following form:

$$\boldsymbol{Z}_{n+1}(s) \equiv (\boldsymbol{\Delta}_n^{\mathrm{T}}\boldsymbol{\eta}_n)^{\mathrm{T}}(\boldsymbol{I} - s\boldsymbol{T}_n)^{-1}\boldsymbol{\rho}_n. \tag{78}$$

Hence $\boldsymbol{Z}_{n+1}$ is a function of the type (76). Furthermore, in [9, Theorem 1], it is shown that (78) is indeed an $(n+1)$st matrix-Padé approximant of $\boldsymbol{Z}$.

8.3. Approximate eigenvalues

In this subsection, we consider the eigenvalue problem,

$$\boldsymbol{A}\boldsymbol{x} = \tau\boldsymbol{x}, \tag{79}$$

for $\boldsymbol{A}$. Using the same two-sided projection as in Section 8.2, we can derive from (79) a smaller eigenvalue problem whose eigenvalues can then be used as approximate eigenvalues of $\boldsymbol{A}$. More precisely, setting $\boldsymbol{x} = \boldsymbol{V}_n\boldsymbol{z}$ and multiplying (79) from the left by $\boldsymbol{W}_n^{\mathrm{T}}$, we get

$$\boldsymbol{W}_n^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}_n\boldsymbol{z} = \tau\boldsymbol{W}_n^{\mathrm{T}}\boldsymbol{V}_n\boldsymbol{z}. \tag{80}$$

By (73), the generalized eigenvalue problem (80) is equivalent to the standard eigenvalue problem,

$$\boldsymbol{T}_n^{\mathrm{proj}}\boldsymbol{z} = \tau\boldsymbol{z},$$

for the matrix $\boldsymbol{T}_n^{\text{proj}}$ generated from the Lanczos-type method sketched in Section 8.1. The eigenvalues of $\boldsymbol{T}_n^{\text{proj}}$ are then used as approximate eigenvalues of $\boldsymbol{A}$. For further details of this approach, we refer the reader to [11].

8.4. Linear systems with multiple right-hand sides

Next, consider systems of linear equations with coefficient matrix $\boldsymbol{A}$ and multiple, say m, given right-hand sides. Such linear systems can be written in compact matrix form,

$$\boldsymbol{A}\boldsymbol{X} = \boldsymbol{B}, \tag{81}$$

where $\boldsymbol{B} \in \mathbb{C}^{N \times m}$. Let $\boldsymbol{X}_0 \in \mathbb{C}^{N \times m}$ be any guess for the solution of (81), and let $\boldsymbol{R}_0 := \boldsymbol{B} - \boldsymbol{A}\boldsymbol{X}_0$ be the associated residual matrix. By running the Lanczos-type method sketched in Section 8.1 applied to $\boldsymbol{A}, \boldsymbol{R} = \boldsymbol{R}_0$, and any (for example, random) matrix $\boldsymbol{L} \in \mathbb{C}^{N \times p}$ until pass n, we obtain the matrices $\boldsymbol{\rho}_n^{\text{proj}}$ and $\boldsymbol{T}_n^{\text{proj}}$, which can be used to generate a Galerkin-type iterate for (81). More precisely, we set

$$\boldsymbol{X}_{n+1} = \boldsymbol{X}_0 + \boldsymbol{V}_n \boldsymbol{Z}_n, \quad \text{where } \boldsymbol{Z}_n \in \mathbb{C}^{(n+1) \times m}, \tag{82}$$

and require that the free parameter matrix $\boldsymbol{Z}_n$ in (82) is chosen such that the Galerkin condition

$$\boldsymbol{W}_n^{\mathrm{T}}(\boldsymbol{B} - \boldsymbol{A}\boldsymbol{X}_{n+1}) = \boldsymbol{0} \tag{83}$$

is satisfied. By inserting (82) into (83) and using the definitions of $\boldsymbol{\rho}_n^{\text{proj}}$ and $\boldsymbol{T}_n^{\text{proj}}$ in (71), it follows that (83) is equivalent to the linear system

$$\boldsymbol{T}_n^{\text{proj}} \boldsymbol{Z}_n = \boldsymbol{\rho}_n^{\text{proj}}. \tag{84}$$

Provided that $\boldsymbol{T}_n^{\text{proj}}$ is nonsingular, the solution of (84) defines a unique iterate (82). In the special case that $m = p$ and that no deflation occurs, the resulting iterative method for solving (81) is mathematically equivalent to block-biconjugate gradients [21].

We remark that the condition on the nonsingularity of $\boldsymbol{T}_n^{\text{proj}}$ can be avoided by replacing (84) by a least-squares problem with a rectangular extension of $\boldsymbol{T}_n^{\text{proj}}$, which always has full column rank. The resulting iterative method for solving (81) is the block-QMR algorithm [13,20].

8.5. A fast block Hankel solver

The last application is an extension of the fast solver for scalar Hankel matrices in [14] to general block Hankel matrices.

Let $\boldsymbol{A}_n$ and $\boldsymbol{B}_n$ the matrices given by (28) and (38). Recall that these matrices contain the potentially nonzero coefficients of the first $n + 1$ pairs of polynomials (51) produced by Algorithm 1. By construction, these matrices are upper triangular and they satisfy $\langle \boldsymbol{\Psi}_n, \boldsymbol{\Phi}_n \rangle = \boldsymbol{B}_n^{\mathrm{T}} \boldsymbol{H}_n^{\text{defl}} \boldsymbol{A}_n$. Hence, the first relation in (64) is equivalent to the following matrix factorization:

$$\boldsymbol{B}_n^{\mathrm{T}} \boldsymbol{H}_n^{\text{defl}} \boldsymbol{A}_n = \boldsymbol{\Delta}_n = \operatorname{diag}(\boldsymbol{\Delta}^{(0)}, \boldsymbol{\Delta}^{(1)}, \ldots, \boldsymbol{\Delta}^{(\ell)}). \tag{85}$$

Note that (85) represents an inverse triangular factorization of $\boldsymbol{H}_n^{\text{defl}}$. By rewriting the recurrences used to generate the polynomials in Algorithm 1 in terms of the columns of $\boldsymbol{A}_n$ and $\boldsymbol{B}_n$, one obtains a fast, i.e., $\mathcal{O}(n^2)$, algorithm for computing the factorization (85) of the deflated block Hankel matrix $\boldsymbol{H}_n^{\text{defl}}$. Details of this fast block Hankel solver will be given in a future publication.

9. Concluding remarks

We have presented a computational procedure for generating vector-valued polynomials that are formally orthogonal with respect to a matrix-valued bilinear form induced by a general block Hankel matrix $\boldsymbol{H}$ with arbitrary, not necessarily square blocks. Existing algorithms for this problem require the assumption that $\boldsymbol{H}$ is strongly regular; unfortunately, this assumption is not satisfied in one of the most important special cases, namely bilinear forms given by a realization. In contrast, our approach can handle the most general case and does not require any assumptions on the block Hankel matrix $\boldsymbol{H}$.

We have briefly discussed some applications of the proposed computational procedure to problems in linear algebra. There are other potential applications, for example Gauss quadrature for matrix-valued bilinear forms, and these will be reported elsewhere.

References

[1] J.I. Aliaga, D.L. Boley, R.W. Freund, V. Hernańdez, A Lanczos-type method for multiple starting vectors, Math. Comp. (2000), to appear. Also available online from `http://cm.bell-labs.com/cs/doc/98`.

[2] B. Beckermann, Nonsymmetric difference operators and polynomials being orthogonal with respect to rectangular matrix valued measures, Technical Report ANO-335, Laboratoire d'Analyse Numérique et d'Optimisation, Université des Sciences et Technologies de Lille, 1995.

[3] C. Brezinski, Padé-Type Approximation and General Orthogonal Polynomials, Birkhäuser, Basel, 1980.

[4] A. Bultheel, M. Van Barel, Linear Algebra, Rational Approximation and Orthogonal Polynomials, North-Holland, Amsterdam, 1997.

[5] A.J. Durán, W. Van Assche, Orthogonal matrix polynomials and higher-order recurrence relations, Linear Algebra Appl. 219 (1995) 261–280.

[6] P. Feldmann, R.W. Freund, Efficient linear circuit analysis by Padé approximation via the Lanczos process, IEEE Trans. Comput.-Aided Des. 14 (1995) 639–649.

[7] P. Feldmann, R.W. Freund, Reduced-order modeling of large linear subcircuits via a block Lanczos algorithm, Proceedings of 32nd Design Automation Conference, ACM, New York, 1995, pp. 474−479.

[8] R.W. Freund, The look-ahead Lanczos process for nonsymmetric matrices and its applications, in: J.D. Brown, M.T. Chu, D.C. Ellision, R.J. Plemmons (Eds.), Proceedings of the Cornelius Lanczos International Centenary Conference, SIAM, Philadelphia, 1994, pp. 33–47.

[9] R.W. Freund, Computation of matrix Padé approximations of transfer functions via a Lanczos-type process, in: C.K. Chui, L.L. Schumaker (Eds.), Approximation Theory VIII, Vol. 1: Approximation and Interpolation, World Scientific, Singapore, 1995, pp. 215–222.

[10] R.W. Freund, Reduced-order modeling techniques based on Krylov subspaces and their use in circuit simulation, in: B.N. Datta (Ed.), Applied and Computational Control, Signals, and Circuits, Vol. 1, Birkhäuser, Boston, 1999, pp. 435–498.

[11] R.W. Freund, Band Lanczos method, in: Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, H. van der Vorst (Eds.), Templates for the Solution of Algebraic Eigenvalue Problems: a Practical Guide, SIAM, Philadelphia, 2000, to appear. Also available online from `http://cm.bell-labs.com/cs/dos/99`.

[12] R.W. Freund, M.H. Gutknecht, N.M. Nachtigal, An implementation of the look-ahead Lanczos algorithm for non-Hermitian matrices, SIAM J. Sci. Comput. 14 (1993) 137–158.

[13] R.W. Freund, M. Malhotra, A block QMR algorithm for non-Hermitian linear systems with multiple right-hand sides, Linear Algebra Appl. 254 (1997) 119–157.

[14] R.W. Freund, H. Zha, A look-ahead algorithm for the solution of general Hankel systems, Numer. Math. 64 (1993) 295–321.

[15] W.B. Gragg, Matrix interpretations and applications of the continued fraction algorithm, Rocky Mountain J. Math. 4 (1974) 213–225.

[16] W.B. Gragg, A. Lindquist, On the partial realization problem, Linear Algebra Appl. 50 (1983) 277–319.

[17] M.H. Gutknecht, A completed theory of the unsymmetric Lanczos process and related algorithms, part I, SIAM J. Matrix Anal. Appl. 13 (1992) 594–639.

[18] R.E. Kalman, P.L. Falb, M.A. Arbib, Topics in Mathematical System Theory, McGraw-Hill, New York, 1969.

[19] C. Lanczos, An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, J. Res. Natl. Bur. Standards 45 (1950) 255–282.

[20] M. Malhotra, R.W. Freund, P.M. Pinsky, Iterative solution of multiple radiation and scattering problems in structural acoustics using a block quasi-minimal residual algorithm, Comput. Methods. Appl. Mech. Eng. 146 (1997) 173–196.

[21] D.P. O'Leary, The block conjugate gradient algorithm and related methods, Linear Algebra Appl. 29 (1980) 293–322.

[22] B.N. Parlett, D.R. Taylor, Z.A. Liu, A look-ahead Lanczos algorithm for unsymmetric matrices, Math. Comp. 44 (1985) 105–124.

[23] J. Rissanen, Basis of invariants and canonical forms for linear dynamical systems, Automat. J. IFAC 10 (1974) 175–182.

[24] A. Salam, Formal vector orthogonal polynomials, Adv. Comput. Math. 8 (1998) 267–289.

[25] A. Sinap, W. Van Assche, Orthogonal matrix polynomials and applications, J. Comput. Appl. Math. 66 (1996) 27–52.

[26] E.D. Sontag, Mathematical Control Theory, Springer, New York, 1990.

[27] V.N. Sorokin, J. Van Iseghem, Algebraic aspects of matrix orthogonality for vector polynomials, J. Approx. Theory 90 (1997) 97–116.

[28] D.R. Taylor, Analysis of the look ahead Lanczos algorithm, Ph.D. Thesis, Department of Mathematics, University of California, Berkeley, 1982.

Journal of Computational and Applied Mathematics 127 (2001) 201–217

JOURNAL OF
COMPUTATIONAL AND
APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Computation of Gauss-type quadrature formulas

Dirk P. Laurie

School for Modelling Sciences, Potchefstroomse Universiteit vir CHO, Vaal Triangle Campus,
P.O. Box 1174 Vanderbijlpark, South Africa

Received 5 May 2000

Abstract

Gaussian formulas for a linear functional L (such as a weighted integral) are best computed from the recursion coefficients relating the monic polynomials orthogonal with respect to L. In Gauss-type formulas, one or more extraneous conditions (such as pre-assigning certain nodes) replace some of the equations expressing exactness when applied to high-order polynomials. These extraneous conditions may be applied by modifying the same number of recursion coefficients. We survey the methods of computing formulas from recursion coefficients, methods of obtaining recursion coefficients and modifying them for Gauss-type formulas, and questions of existence and numerical accuracy associated with those computations. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Gaussian quadrature; Numerical integration; Lobatto; Radau; Kronrod

1. Introduction

By n-point *Gaussian quadrature* we mean the approximation of a given linear functional L by a discrete linear functional G_n, the n-point *Gaussian quadrature formula* given by

$$G_n[f] = \sum_{j=1}^{n} w_j f(x_j),$$

such that $G_n[f] = L[f]$ whenever $f \in \mathscr{P}_{2n-1}$, where $\mathscr{P}_m$ is the space of polynomials of degree not exceeding m. We say that the formula has *degree* $2n-1$. It is convenient to think of L as defined on the space of all functions $f : \mathbb{R} \to \mathbb{R}$, because when f is defined on a smaller interval, one can simply define it to be zero elsewhere.

E-mail address: dlaurie@na-net.ornl.gov (D.P. Laurie).

PII: S0377-0427(00)00506-9

The definition suggests an obvious brute force calculation method, namely:

(1) Select a basis for $\mathscr{P}_{2n-1}$ for which L can reliably be evaluated.
(2) Solve the system of $2n$ non-linear equations for the $2n$ unknowns $\{w_j\}$ and $\{x_j\}$ without taking any advantage of its structure, apart from possibly using a solver that can treat implicitly the "linear" variables $\{w_j\}$.

The brute force method is not totally devoid of theoretical merit, since the Jacobian of the system gives information about the condition of the calculation problem, but is intractable for large n. In some cases the brute force method, with ad hoc simplifying techniques that reflect great ingenuity, may be the only available method to find certain formulas of *Gauss-type*. This nonstandard term is here used to mean an n-point formula of degree at least n (i.e., of higher degree than an interpolatory formula) and defined by $2n$ equations, not all of which express exactness when applied to a polynomial. In general there can be at most $n-1$ such equations, the other $n+1$ equations being those that define the degree.

The best-known formulas of Gauss-type are the Radau, Lobatto and Kronrod formulas, respectively, of degrees $2n-2$, $2n-3$ and approximately $\frac{3}{2}n$. They are discussed together with some others in Section 4.

The problem is easiest when we know the coefficients of the three-term recursion

$$p_{l+1}(x) = (x - a_l)p_l(x) - b_l p_{l-1}(x) \tag{1}$$

satisfied by the monic orthogonal polynomials $\{p_l\}$ (i.e., monic polynomials for which $L[p_l q] = 0$ when $q \in \mathscr{P}_{l-1}$; by pq we mean the pointwise product of p and q, i.e., $(pq)(x) = p(x)q(x)$.) In order that (1) be valid for $l = 0, 1, 2, \ldots$, it is customary to define $p_{-1}(x) = 0$; of course, $p_0(x) = 1$. This allows freedom in the choice of b_0; a convenient value suggested by Gautschi is $b_0 = L[p_0]$, which leads to the useful relation

$$L[p_l^2] = b_0 b_1 b_2 \ldots b_l. \tag{2}$$

Unless stated otherwise, we will assume that all b_l are positive: this is for example the case when L denotes integration with a positive measure. In Section 6 we discuss what could happen if some b_l are not positive.

The purpose of this paper is to survey the calculation methods currently regarded as efficient and accurate for Gaussian and some Gauss-type formulas. To keep the paper concise and devoted to a single theme, any method that does not explicitly use the three-term coefficients for G_n is only mentioned when no method based on them is available. The state of the art in software for computing with three-term coefficients is Gautschi's Fortran package `ORTHPOL` [20].

There are at least two ways of using the recursion coefficients.

(1) The nodes $\{x_j\}$ are the zeros of p_n. The recursion (1) is a numerically stable way of computing p_n, and can easily be differentiated to give a recursion formula for the derivatives. One can therefore calculate the nodes by a method such as Newton's or Weierstrass'. Once a node has been found, any of several classical formulas, such as

$$w_j = \frac{L[p_{n-1}^2]}{p_{n-1}(x_j)p_n'(x_j)}, \tag{3}$$

Table 1
Two- and three-term recursion coefficients for the classical orthogonal polynomials. In the Jacobi case, $\kappa = 2k + \alpha + \beta$

	Jacobi over (0,2)	Laguerre over $(0, \infty)$
a_k	$1 + \dfrac{\beta^2 - \alpha^2}{\kappa(\kappa+2)}$	$2k + 1 + \alpha$
b_k	$\dfrac{4k(k+\alpha)(k+\beta)(k+\alpha+\beta)}{\kappa^2(\kappa+1)(\kappa-1)}$	$k(k+\alpha)$
e_k	$\dfrac{2k(k+\alpha)}{\kappa(\kappa+1)}$	k
q_k	$\dfrac{2(k+\beta)(k+\alpha+\beta)}{\kappa(\kappa-1)}$	$k + \alpha$

can be applied to calculate its weight. This method is used in the monumental work by Stroud and Secrest [47], which was the first to suggest that the use of tabulated formulas (of which the book has many) might be superseded by subroutines (also given there).

(2) The symmetric tridiagonal matrix (known as the *Jacobi matrix*)

$$T = \mathsf{tridiag}\begin{pmatrix} & \sqrt{b_1} & & \sqrt{b_1} & \bullet & \sqrt{b_{n-1}} & \\ a_0 & & a_1 & \bullet & \bullet & & a_{n-1} \\ & \sqrt{b_1} & & \sqrt{b_1} & \bullet & \sqrt{b_{n-1}} & \end{pmatrix}$$

has the nodes as eigenvalues, and when its eigenvectors are normalized to have length 1, the weights are just $w_j = b_0 u_{1,j}^2$ where $u_{1,j}$ is the first component of the eigenvector corresponding to the eigenvalue x_j. Golub and Welsch [23] show how to modify the QR algorithm so that only the required components are found.

In view of the Golub–Welsch result, the calculation of a Gaussian formula is considered to be essentially a solved problem in the case where the recursion coefficients are known, and all $b_j > 0$. Still, even a fully satisfactory algorithm might be improved a little; some alternatives to the original algorithm are discussed in Section 2. In particular, when some b_l are negative, the algebraic identities behind the Golub–Welsch algorithm still hold, but the algorithm itself would require complex arithmetic and its numerical stability and freedom from breakdown would become uncertain.

Analytical expressions for the recursion coefficients are only available in exceptional cases, including the classical orthogonal polynomials (see Table 1). In practice one has to calculate them from other information about the functional. Here two main general approaches have emerged.

(1) If L arises from integration with a weight function (by far the most common application), one can approximate L by a quadrature formula with $N \gg n$ points. This approach was first suggested by Gautschi [12,14]. The problem of recovering the recursion coefficients from a discrete linear functional is an important one in its own right, discussed in Section 3.1, although for our purpose it is not necessary to obtain great accuracy right up to a_{N-1} and b_{N-1}.

(2) In some cases it is possible to obtain *modified moments* $\mu_k = L[\pi_k]$, where the polynomials $\{\pi_k\}$ themselves are orthogonal with respect to another linear functional Λ. This approach is only useful when the recursion coefficients $\{\alpha_k\}$ and $\{\beta_k\}$ for the polynomials $\{\pi_k\}$ are known: the most common cases are the monic Chebyshev and Legendre polynomials (see Section 3.2).

> It is essential to remember that any algorithm for calculating $\{a_l; b_l\}$ from $\{\alpha_k; \beta_k\}$ can in principle deliver no better accuracy than that implied by the condition of the map between the two polynomial bases, a question thoroughly discussed by Gautschi [16]. In particular, it is well known (to numerical analysts, at least) that the case $\alpha_k = \beta_k = 0$, when the polynomials $\{\pi_k\}$ reduce to the monomials and the modified moments to ordinary moments, is catastrophically ill-conditioned.

Finally, it is sometimes possible to view certain Gauss-type quadratures for L as Gaussian quadratures for a related linear functional L', so that the recursion coefficients for L' can be derived with more or less ease from those of L. These include Radau, Lobatto, anti-Gaussian and Kronrod formulas, and are discussed in Section 4.

For much of the paper, the questions of when the methods are applicable and how well they work is at best mentioned in passing. These issues are collected in Section 6. That section also contains a new idea on how to check the accuracy of a Gaussian formula.

Recent developments in eigenvalue methods indicate that when it is known that not only the weights but also the nodes must be positive, it is better to work with two-term recursion coefficients. Laurie [33] discusses this question fully, and therefore only a brief summary is given in Section 5.

2. What to do when the recursion coefficients are known

The timeless fact about the Golub–Welsch algorithm is that any advance in our ability to solve the symmetric tridiagonal eigenproblem immediately implies a corresponding advance in our ability to compute Gaussian formulas. Therefore, the fragment of Matlab code

```
% Given the number of nodes n —
a=zeros(1,n);
b=sqrt(1 ./(4-1 ./(1:n-1).^2));
b0=2;
[S,D]=eig(diag(b,1)+diag(a)+diag(b,-1));
x=diag(D); w= b0*S(1,:).^2;
% — we now have nodes x and weights w
```

that computes the n-point Gaussian quadrature formula for $Lf = \int_{-1}^{1} f(x)\,\mathrm{d}x$ may be inefficient because the eigenvectors are found in their entirety, but as far as accuracy is concerned, it will remain as state-of-the-art as the Linpack or LAPACK or whatever subroutines that underlie one's available Matlab implementation. Only the lines defining a, b and $b0$ need be changed to make it work for other linear functionals with known recursion coefficients.

The above code fragment requires $\mathrm{O}(n^3)$ operations, whereas the Golub–Welsch algorithm only requires $\mathrm{O}(n^2)$, as follows:

(1) The inner loop of the algorithm consists of an implicit QR sweep, implemented as a sequence of plane rotations [44] in the $(i, i+1)$ plane, with $i = 1, 2, \ldots, n-1$.

(2) The same rotations are made to the first row of the eigenvector matrix, which starts at $[1, 0, 0, \ldots, 0]$ and ends as the square roots of the normalized weights.

(3) All decisions as to shift strategy and deflation are made as usual, uninfluenced by the weight calculation.

Since 1968, several improvements to the QR algorithm have been published. When eigenvalues only are needed, there exist ways to avoid taking square roots, e.g., Pal–Walker–Kahan [35] and Gates–Gragg [11] rational QR algorithms. The latter is slightly faster than the former, using one multiplication less per inner loop. Parlett has suggested [35] that it may be better to calculate the eigenvectors after the eigenvalues have already been found.

These ideas could lead to improvements to the Golub–Welsch algorithm if one could also calculate the weights in a square-root free way. Laurie [32] has presented one way of so doing. The idea is to calculate all the eigenvalues first, and then to do a QL instead of a QR sweep to find each weight. In that case the sequence of plane rotations runs $i = n-1, n-2, \dots, 1$, so that the first row of the eigenvector matrix is unchanged until the very last rotation. The first element changes to $\cos\theta$. The required weight is $b_0 \cos^2\theta$. But the quantity $C = \cos^2\theta$ is available directly from the rational QL inner loop.

This idea should not be taken to its logical conclusion, as in algorithm sqlGauss of [32], where the matrix is deflated immediately, and the other weights are computed using the deflated matrix. It is well known that when ‘ultimate’ shifts are used in an attempt to converge to each eigenvalue in a single iteration, the accumulation of roundoff is such that deflation is not always justified. This may happen even when each shift is the correctly rounded machine representation of the corresponding eigenvalue. Instead, each weight should be computed on its own, starting from $[1, 0, 0, \dots, 0]$.

The main recent alternative to the QR algorithm for the symmetric tridiagonal eigenproblem is the divide-and-conquer algorithm. This algorithm is faster than QR on large problems when all eigenvectors are required, and has been modified by Cuppen [4] to yield only those components necessary to the quadrature formula. A numerical study that compares the original Golub–Welsch algorithm to the square-root free QL-based version outlined above, and to the divide-and-conquer method, would be an interesting topic for a master’s thesis. At the moment of writing the Golub–Welsch algorithm is the only one of the three that has been published in the form of actual Fortran code [20], but divide-and-conquer (in the numerically robust version of Gu and Eisenstat [25]) is in LAPACK (called via dsyevd in double precision) and it would be easy to make the abbreviated version that would be required for such a comparison.

3. Obtaining recursion coefficients

The subject of obtaining recursion coefficients is of importance in all applications of orthogonal polynomials, and in this paper we concentrate on methods that are necessary in order to compute Gauss-type quadratures.

3.1. Recursion coefficients via quadrature

Very often the functional L is given in the form

$$L[f] = \int_A^B f(x)w(x)\,\mathrm{d}x$$

(or even more generally as a Stieltjes integral). When the integral can be evaluated analytically, a procedure commonly called the *Stieltjes algorithm* gives the recursion coefficients. The formulas

$$b_l = \frac{L[p_l^2]}{L[p_{l-1}^2]},$$

$$a_l = \frac{L[xp_l^2]}{L[p_l^2]},$$

are used in bootstrap fashion, alternating with the recursion (1): each coefficient is computed just before it is needed to generate the next p_l. (The notation x above stands for the identity polynomial: $x(t) = t$.)

The luxury of analytical evaluation is rarely possible, and therefore Gautschi [12,14] suggests replacing L by a suitable N-point quadrature formula Q with $N \gg n$, of sufficient accuracy that the first n polynomials orthogonal to Q will coincide to working accuracy with those orthogonal to L. The *discrete Stieltjes algorithm* is then obtained by running the above computation with Q in the place of L, stopping when enough coefficients have been found.

One might equally well say that the desired Gaussian quadrature is the n-point formula corresponding to the functional Q. The problem of calculating the coefficients $\{a_l\}$ and $\{b_l\}$ from a given Gaussian quadrature formula can be viewed as an inverse eigenvalue problem. This approach is taken by De Boor and Golub [5] and by Gragg and Harrod [24]. The latter paper gives examples, involving either very nearly coincident nodes or very nearly vanishing weights, where the discrete Stieltjes produces results with no significant figures. Reichel has given some other illustrative examples [41] showing that the Stieltjes algorithm may give inaccurate recursion coefficients a_l and b_l for $l < n$ even when $N \gg n$. So, although there are cases where the Stieltjes procedure is adequate, the proverbial 'cautious man' will prefer to use the Gragg–Harrod algorithm [24].

An algorithm with a further small theoretical edge is discussed in Section 5.

3.2. Recursion coefficients from modified moments

Suppose that it is possible to obtain in a numerically stable way the *modified moments*

$$\mu_k = L[\pi_k],$$

where the monic polynomials $\{\pi_k\}$ satisfy the recursion

$$\pi_{k+1}(x) = (x - \alpha_k)\pi_k(x) - \beta_k \pi_{k-1}(x). \tag{4}$$

To do so is seldom a trivial task, even in simple cases such as when the polynomials $\{\pi_k\}$ are monic Chebyshev or Legendre polynomials. For example, it is likely to be disastrous to try direct quadrature of $\{\pi_k\}$. When the map from the basis $\{\pi_k\}$ to the basis $\{p_l\}$ is well conditioned, π_k typically has k sign changes in the integration interval, thereby leading to severe the loss of significant digits by cancellation. The most promising approach is to look for a recursion formula connecting the $\{\mu_k\}$. Early attempts to do so needed inspiration, but Lewanowicz ([34] and many later papers) has shown how to obtain such recursions when the functional L is an integral involving a weight function that satisfies a differential equation with polynomial coefficients.

We leave the question as to the stable computation of modified moments and ask what can be done with them when they are available.

Sack and Donovan [45] point out that the *mixed moments*

$$\sigma_{k,l} = L[\pi_k p_l]$$

satisfy the five-term two-dimensional recursion relation

$$\begin{pmatrix} & -\beta_k & \\ b_l & a_l - \alpha_k & 1 \\ & -1 & \end{pmatrix} \cdot \begin{pmatrix} & \sigma_{k-1,l} & \\ \sigma_{k,l-1} & \sigma_{k,l} & \sigma_{k,l+1} \\ & \sigma_{k+1,l} & \end{pmatrix} = 0.$$

Here the dot is taken to mean, as in the case of vectors, that the quantities to its left and right are multiplied component by component and summed.

The matrix with entries $[\sigma_{k,l}]$ is lower triangular, and this fact, together with the known first column, allows us to generate the other entries along anti-diagonals running southwest to northeast. When the main diagonal is reached, the stencil is centred over known entries to give the values of b_l and a_l that are required to complete the next anti-diagonal.

The five-term stencil makes it easy to see that when two functionals are equal on the space P_m, then all mixed moments $\sigma_{k,l}$ with $k+l \leqslant m$ and $k \neq l$ are zero, and therefore the first $m+1$ members of the recursion coefficient sequences $b_0, a_0, b_1, a_1, \ldots$ and $\beta_0, \alpha_0, \beta_1, \alpha_1, \ldots$ are equal. For want of a better term, we call this the *degree-revealing* property of the recursion coefficients.

4. Gauss-type quadratures

Our main tool for computing Gauss-type quadratures is to obtain the recursion coefficients for the system of orthogonal polynomials that are generated by the quadrature formula itself. All the machinery of Section 2 can then be brought to bear.

When the formula has degree close to the maximum, only a few of these coefficients differ from those of the Gaussian formula itself, and the problem is easy. When the formula is of comparatively low degree, many coefficients differ and the problem is difficult.

In this section, we use notations such as a_l', b_l'', etc., to indicate recursion coefficients for Gauss-type formulas that do not necessarily equal the a_l and b_l of the Gaussian formula G_n.

4.1. Radau, Lobatto, anti-Gaussian and other related formulas

The best-known Gauss-type quadratures are those with one or two preassigned nodes. We shall call them, respectively, Radau and Lobatto formulas, although strictly speaking these names apply to the case where the preassigned nodes are endpoints of the integration interval. A nice discussion on Radau and Lobatto formulas is provided by Golub [21].

It is possible to calculate Radau and Lobatto formulas just like Gaussian formulas, by specifying the recursion coefficients that define their orthogonal polynomials. By the degree-revealing property, only a_{n-1} in the case of a Radau formula, and both a_{n-1} and b_{n-1} in the case of a Lobatto formula, will differ from that for the Gaussian formula. We obtain

$$a_{n-1}' = a_{n-1} + p_n(\xi)/p_{n-1}(\xi)$$

for a formula with one preassigned node ξ, and

$$\begin{bmatrix} a''_{n-1} \\ b''_{n-1} \end{bmatrix} = \begin{bmatrix} a_{n-1} \\ b_{n-1} \end{bmatrix} + \begin{bmatrix} p_{n-1}(\xi) & p_{n-2}(\xi) \\ p_{n-1}(\eta) & p_{n-2}(\eta) \end{bmatrix}^{-1} \begin{bmatrix} p_n(\xi) \\ p_n(\eta) \end{bmatrix}$$

for a formula with two preassigned nodes ξ and η.

The anti-Gaussian formulas introduced by Laurie [30] are likewise easily calculated. The idea here is that G'_n should have error equal in magnitude but opposite in sign to G_{n-1}. The formula is therefore also of degree $2n-1$. By the degree-revealing property, only a_{n-1} and b_{n-1} need to be modified: in fact, it turns out that only the latter changes. We obtain

$$b'_{n-1} = 2b_{n-1}.$$

4.2. General modification algorithms

If the recursion coefficients for the Gaussian formula corresponding to a weight function $w(x)$ are known, then it is possible to derive the coefficients for the following weights in a reasonably straightforward manner:

(1) $(x-\xi)w(x)$;
(2) $((x-\xi)^2+\eta^2)w(x)$;
(3) $(x-\xi)^{-1}w(x)$;
(4) $((x-\xi)^2+\eta^2)^{-1}w(x)$.

Software for this calculation is included in ORTHPOL [20].

The $(n-1)$-point Gaussian rule G^{ξ}_{n-1} for the weight function $(x-\xi)w(x)$ is related to the n-point Radau formula R_n for $w(x)$ with prescribed node ξ as follows if:

$$R_n[f] = \rho f(\xi) + \sum_{j=1}^{n} r_j f(x_j),$$

then

$$G^{\xi}_{n-1}[f] = \sum_{j=1}^{n} (x_j-\xi) r_j f(x_j).$$

To go from R_n to G^{ξ}_{n-1} we keep the nodes and multiply the weights by $(x_j-\xi)$; to go the other way we divide the weights by $(x_j-\xi)$ and calculate the weight ρ for the fixed node from the condition that $f(x)=1$ is integrated exactly.

This approach to calculating the Radau formula differs from the one in Section 4.1 in that the fixed node is treated in different way to the others. By repeating the calculation, one can obtain a Gaussian formula for functions with any number of preassigned poles or zeros, as is done by Gautschi [19].

Similarly, the modification algorithms can be used to obtain the formulas with up to $n-1$ preassigned nodes. The details of the algorithms involved, including their extension to multiple nodes, are discussed by Elhay, Golub and Kautsky (variously collaborating) in [22,26,27].

4.3. Kronrod formulas

Formulas with m preassigned nodes, where $n = 2m + 1$, have been well studied. The best-known case is the formula K_{2m+1} whose preassigned nodes are those of G_m. Although theoretical results go back to Stieltjes and later Szegő, the quadrature formulas themselves are named after Kronrod [28], who computed them for $Lf = \int_{-1}^{1} f(x)\,\mathrm{d}x$ by a brute force method.

Many methods for computing Kronrod formulas are known (see Gautschi's survey [18]) but most of them separate the calculation of the new nodes and weights from the calculation of new weights for the old nodes. In the spirit of our theme we mention only methods based on the recurrence relation for the polynomials generated by K_{2m+1}.

By the degree-revealing property, the last m entries in the recursion coefficient sequence differ from those of G_{2m+1}. An algorithm to calculate them, based on the five-term recurrence of mixed moments, is given by Laurie [31]. The theoretical base for that algorithm is a theorem, proved in [31], that the leading and trailing $n \times n$ submatrices of the Jacobi matrix corresponding to K_{2m+1} have the same characteristic polynomial. This theorem is used by Calvetti et al. [3] to derive a method for calculating the Kronrod formula by a divide-and-conquer method applied to that Jacobi matrix, in which the new coefficients are never explicitly found.

The divide-and-conquer method seems on the basis of numerical experiments to be a little more accurate than the method based on finding the new recursion coefficients, but the latter also has some theoretical and practical advantages. From the theoretical point of view, it allows us to tell whether the formula is real with positive weights before calculating it: the question is equivalent to asking whether all $b'_l > 0$. For example, in [31], an analytic criterion is obtained for the existence of a 5-point Kronrod formula for the Jacobi weight $(1-x)^{\alpha}(1+x)^{\beta}$ in terms of the positivity of certain sixth-degree polynomials in α and β. From a practical point of view, Kronrod formulas with complex nodes and/or negative weights can also be calculated.

4.4. Other formulas with $m = (n-1)/2$ preassigned nodes

An easy alternative to the Kronrod formula is to require that not only should the nodes of G_m be preassigned, but their weights in the new formula should be the old weights multiplied by $\frac{1}{2}$. In that case, the required formula is simply the mean of the Gaussian formula G_m and the anti-Gaussian formula G'_{m+1}. For lack of a better term, we call the formula $\frac{1}{2}(G_m + G'_{m+1})$ an *averaged Gaussian formula*. This formula exists in many more cases than does the Kronrod formula (see [31]).

The general case with $m = (n-1)/2$ preassigned nodes is numerically very difficult. A software package has been published by Patterson [40], but one must be prepared to compute in much higher precision than that of the required formula.

The philosophy behind the Kronrod and anti-Gaussian formulas can be iterated to form sequences of embedded quadrature formulas in which each member is related to its predecessor in the same way as a Kronrod or averaged Gaussian formula is related to a Gaussian formula. In the first case, the Patterson formulas [38,39] are obtained: each Patterson formula is of Gauss-type under the restriction that all the nodes of the preceding formula are pre-assigned. In the second case, Laurie's stratified formulas [29] are obtained: each formula is of Gauss-type under the restriction that all the nodes of the preceding formula are pre-assigned, with their weights multiplied by $\frac{1}{2}$.

For both of these families of formulas, the best available calculation algorithms (too complicated to describe here) require very high precision. It is an open question whether the difficulty is inherent (which seems likely) or an ingenious algorithm that can calculate the formulas in a numerically stable way might one day be discovered.

5. Two-term recursion coefficients

The symmetric tridiagonal matrix T defined in Section 1 is positive definite if and only if all the nodes are positive. In that case, T can be factorized as $T = R^{\mathrm{T}}R$ where

$$R = \mathsf{bidiag}\begin{pmatrix} & \sqrt{e_1} & & \sqrt{e_2} & & \bullet & & \sqrt{e_{n-1}} & \\ \sqrt{q_1} & & \sqrt{q_1} & & \bullet & & \bullet & & \sqrt{q_n} \end{pmatrix}.$$

The coefficients are related by the equations

$$a_l = q_{l+1} + e_l, \tag{5}$$

$$b_l = q_l e_l. \tag{6}$$

Although the matrix R is useful from a theoretical point of view, practical algorithms involving two-term coefficients all work with two bidiagonal matrices

$$L = \mathsf{bidiag}\begin{pmatrix} 1 & & 1 & & \bullet & & \bullet & & 1 \\ & e_1 & & e_2 & & \bullet & & e_{n-1} & \end{pmatrix}, \tag{7}$$

$$U = \mathsf{bidiag}\begin{pmatrix} & 1 & & 1 & & \bullet & & 1 & \\ q_1 & & q_2 & & \bullet & & \bullet & & q_n \end{pmatrix}, \tag{8}$$

which have the property that LU is similar to T.

The main computational tool is Rutishauser's qd algorithm [43], which requires only rational operations. It is slightly older, and computationally more efficient, than the QR algorithm, but was neglected for many years because the standard form of the algorithm is numerically suspect.

Recent work by Dhillon, Fernando and Parlett (variously collaborating) [6,7,36,37] has catapulted the qd algorithm back into the limelight. The point is that the 'differential' form of the qd algorithm (which Rutishauser knew but did not love, because it uses one multiplication more per inner loop than the standard form) has a remarkable stability property: when the two-term coefficients are known to full floating-point accuracy, the eigenvalues of T can also be calculated to high floating-point accuracy.

Laurie [33] surveys the theoretical and computational aspects of two-term versus three-term recursions. It is argued there that two-term recursions might become the method of choice, but at this stage of software development that is not yet the case.

5.1. Computing formulas from two-term recursion coefficients

Since the eigenvalues of R^TR are the squares of the singular values of R, it is possible to give a 'timeless' algorithm also in the two-term case:

```
% Given the number of nodes n —
q=(1:n)./(1:2:2*n);
e=(1:n-1)./(3:2:2*n);
b0=2;
[U,D,V]=svd(diag(sqrt(q))+diag(sqrt(e),1);
x=diag(D). ^2; w=b0*V(1,:).^2;
% — we now have nodes x and weights w.
```

In practice, this algorithm might not be all that is desired, because high-floating-point accuracy of the smallest singular value is not relevant in the traditional applications of the SVD and the implementation might not deliver such accuracy. It is also annoying that one must take square roots before entering the svd routine, and square the result afterwards, but such is the price one pays for using general-purpose software.

Software for the two-term formulas is not as highly developed yet as in the case of three-term formulas, but the area is one of intense current reaserch, and it is already planned to add such methods to the standard linear algebra package LAPACK which is at present the engine driving higher-level languages like Matlab and Octave.

As for Gaussian quadrature, there exists an algorithm given by Rutishauser himself [43] that computes the weights in a single qd array, but the algorithm relies on 'ultimate' shifts and therefore the later weights may be badly contaminated. At this stage, there is no generally accepted 'best' way to calculate the weights.

Much work remains to be done: in particular, a two-term version of ORTHPOL would be very welcome.

5.2. Obtaining two-term recursion coefficients

The one thing that one should not do is to start from known three-term coefficients and solve Eqs. (5) and (6), because that requires subtraction of numbers with the same sign. The reverse procedure, to obtain a_l and b_l from q_l and e_l, is on the other hand quite a reasonable thing to do.

One should therefore obtain the required two-term coefficients directly. In the case of the classical orthogonal polynomials (if necessary shifted to move the left endpoint of their interval of orthogonality to the origin), this is easily done, and in fact the formulas for the two-term coefficients are simpler than those for the three-term ones (see Table 1). In the case of formulas such as those for the Jacobi weight, the shift back to $[-1,1]$ will of course lead to some cancellation if a node happens to lie close to 0. Such a node cannot be obtained in this way to high-floating-point accuracy.

In the case where the coefficients themselves are to be obtained by quadrature, it is easy to modify the discrete Stieltjes procedure. Alternatively, one may use an algorithm given by Laurie [32], which requires no subtractions and is therefore numerically stable in the sense of componentwise relative error.

An analogue of the Sack–Donovan algorithm is given in by Laurie [33], which allows us to calculate the two-term recursion coefficients directly from modified moments.

6. Existence and numerical considerations

We use the following terminology [30] when discussing questions about the existence of a Gaussian or Gauss-type formula:

- A quadrature formula *exists* if its defining equations have a (possibly complex) solution.
- The formula is *real* if the points and weights are all real.
- A real formula is *internal* if all the points belong to the (closed) interval of integration. A node not belonging to the interval is called an *exterior node.*
- The formula is *positive* if all the weights are positive.

The classical results are:

- The Gaussian formula G_n is real and positive if and only if all $b_l > 0,\ l = 0, 1, \ldots, n-1$.
- When $L[f] = \int_A^B f(x)\,\mathrm{d}\omega(x)$, and $\omega(x)$ is nondecreasing with at least n points of increase in the interval (A, B), then G_n is real, positive and internal. This condition covers the classical integrals with weight functions and also discrete linear functionals with positive coefficients.
- The nodes and weights are well-conditioned functions of the coefficients $\{b_l\}$ and $\{a_l\}$ when all $b_l > 0,\ l = 0, 1, \ldots, n-1$.

The main positive results about conditioning and numerical stability are

- When the Gaussian formula is real and positive, the computation of nodes and weights from recursion coefficients is well conditioned in a vector norm.
- When the Gaussian formula is symmetric around zero, the computation of nodes is well conditioned in the sense of componentwise relative error.
- The computation of recursion coefficients from positive weights and inter-node gaps $x_{j+1} - x_j$ is well conditioned in the sense of componentwise relative error [32].

There are also some negative results

- The computation of recursion coefficients from positive weights and distinct nodes is ill-conditioned in a vector norm [24].

The last two results seem to give contradictory answers to the question: is the map from a positive quadrature formula to its recursion coefficients well conditioned? The answer is generally thought to be ‘yes’, e.g., Gautschi [17] bluntly states: ‘The map H_n is always well conditioned’. The negative result of Gragg and Harrod [24] arises because they measure the distance between two weight vectors $\mathbf{w}$ and $\hat{\mathbf{w}}$ as $\max|w_j - \hat{w}_j|$ and not as $\max|w_j - \hat{w}_j|/|w_j|$. The matter is fully discussed by Laurie [32].

To summarize: when starting from accurate three-term coefficients, with all b_l positive, a formula with positive weights is obtained and its absolute error will be small. When starting from accurate two-term coefficients, all positive, a formula with positive nodes and weights is obtained and its relative error will be small.

When starting from accurate modified moments, all depends on the condition of the map from the one polynomial base to the other. The classic paper of Gautschi [16] analyzes many cases, and further work has been done by Fischer [8–10].

In the case of a finite interval, it has been shown by Beckermann and Bourreau [2] that the auxiliary base must be orthogonal with respect to a measure that has the same minimal support as that of the polynomials we are trying to find, otherwise the condition number rises exponentially. This is only a sufficient condition for instability, and it is therefore still necessary to exhibit caution even when the condition is satisfied. Section 6.1 shows how to obtain an *a posteriori* numerical estimate for the error in such a case.

The temptation is great to use the methods even when the hypotheses required for the theorems do not hold. For example, some of the b_k values obtained by Laurie's algorithm [31] for computing a Kronrod formula via its the three-term coefficients may be negative. This does not mean that the Kronrod formula does not exist, only that it is not positive or not real (or both). In the case where the formula is still real, any method that only uses b_k and not its square root, such as those in [32], should work. Some methods that particularly cater for such a possibility have also been developed by Ammar et al. [1].

The case of a general indefinite measure can be written in terms of a symmetric (not Hermitian) tridiagonal matrix with complex entries. The inverse eigenvalue problem for such a matrix is discussed by Reichel [42]. In general, algorithms with look-ahead have to be used, but these tend to be fairly complicated.

As a general rule, the algebraic properties of the methods remain valid when the hypotheses fail, but the numerical stability might be adversely affected. Therefore, a method should still calculate the correct numbers in exact arithmetic as long as everything stays real, and may well deliver satisfactory results also in finite precision. Of course, as soon as quantities that should be real turn out to be complex, software that requires real numbers will fail, and software in a language that silently goes complex (like Matlab) may give wrong answers, since the original version may have tacitly relied on identities such as $z = \bar{z}$ that are no longer true.

6.1. Estimating accuracy of the computed formula

Gautschi's paper *How and how not to check the accuracy of Gaussian formulae* [15] is one of those little gems of numerical analysis that should be required reading for all postgraduate students. The main 'how not to' dictum is that it is meaningless to demonstrate that the formula integrates the required monomials to machine precision: some very bad formulas also do that. The 'how to' suggestions involve higher precision than was used to compute the formula itself.

If the only available data are the moments with respect to an ill-conditioned base (such as the monomials) then higher precision is indeed unavoidable, even when the moments are given as exact rational numbers. In the cases where we start from modified moments or recursion coefficients, it is relatively easy to calculate the sensitivity of the formula.

We give the argument for three-term recursions but it is equally well applicable to two-term ones. The Sack–Donovan algorithm for computing recursion coefficients from modified moments is noniterative and uses $\mathcal{O}(n^2)$ operations. One can vary each modified moment in turn by approximately $\sqrt{\varepsilon}$, where ε is the machine epsilon, to obtain in $\mathcal{O}(n^3)$ operations a numerical Jacobian J_1 for the dependence of recursion coefficients on modified moments.

The computation of the formula itself is iterative, and one would be squeamish to compute a numerical Jacobian by varying the recursion coefficients. But the Gragg–Harrod algorithm, which uses $\mathcal{O}(n^2)$ operations, is noniterative, and one could therefore obtain a numerical inverse Jacobian J_2^{-1} (where J_2 is the Jacobian for the dependence of the nodes and weights on recursion coefficients) by varying the nodes and weights.

Under the assumption (which may be a highly optimistic one) that the original modified moments are correct to machine precision one could compute an iterative correction to the formula as follows:

(1) Calculate the recursion coefficients $\hat{a}_k, \hat{b}_k$ for the computed formula by the Gragg–Harrod algorithm.
(2) Calculate the modified moments $\hat{\mu}_i$ for these recursion coefficients by the Salzer algorithm [46].
(3) Calculate the residual $\mu_i - \hat{\mu}_i$.
(4) Calculate the correction to the formula by applying $(J_2^{-1})^{-1}J_1$ to the residual.

Of course, without the aid of higher precision in the first three steps of the above procedure, one cannot hope to improve the formula, but the correction can be expected to be of approximately the same order of magnitude as the error in the formula.

6.2. Underflow and overflow

On computers with double-precision IEEE arithmetic, underflow and overflow is less of a problem than on some older machines. Nevertheless, in the case of calculating Gaussian quadrature formulas, it could be a difficulty if care is not taken. For example, the examples in [24] showing ill-conditioning in a vector norm arise because some weights underflow and are set to zero.

In the case of integration over a finite interval, it is easy to avoid underflow and overflow. The critical quantities of interest are the numbers $L[p_l^2]$. These appear in, e.g., the Sack–Donovan algorithm and in certain formulas for the weights. By the formula (2), overflow or underflow will arise if $b_l \to b$ for any positive value of b except 1. When the length of the interval is 4, it is fairly well known that $b_l \to 1$ for a very large family of weight functions. Any software with claims to robustness against overflow or underflow should therefore scale the interval to have length 4 before embarking on any computation.

On an infinite interval it is not typical for $\lim_{l\to\infty} b_l$ to exist. The best that one could achieve by scaling is to have some b_l less than and some greater than 1, a technique that has its limitations. Care must therefore be taken to detect the overflow and underflow in the computation, since their avoidance cannot be guaranteed.

6.3. Floating-point computation and relative accuracy

It is a common practice to perform the calculations for finite intervals on $[-1,1]$. The underflow and overflow argument above suggests that $[-2,2]$ is a much safer interval to work on. This is still not good enough. In practice, points of a Gaussian formula cluster near the endpoints of the interval and have very small weights there. When they are used to integrate functions with singularities at the endpoints (there are better ways to do that, we know, but a numerical analyst has no control over the abuses to which software will be put) the important quantity is the distance between the node and the endpoint, which is subject to cancellation when the endpoint is nonzero.

For careful work, one should store as floating-point numbers not the nodes, but the gaps between them. These are precisely the quantities for which the floating-point stability result in [32] holds. A challenge for future research is the development of software that can reliably compute those gaps. One approach may be to shift each endpoint in turn to the origin, and keep half the gaps and weights from each computation; the difficulty is then transferred to obtaining accurate floating-point values for the shifted coefficients, which is possible for the Jacobi polynomials but may require methods not yet known in the general case.

7. Conclusion

The calculation of Gauss-type quadrature formulas is a well-understood problem for which fully satisfactory methods are available when the recursion coefficients are known. Questions on which work remains to be done include:

- Do rational QR and divide-and-conquer methods really improve on the Golub–Welsch algorithm in practice?
- Does the QL algorithm give higher accuracy in practice than the QR algorithm?
- Does the qd algorithm really improve on the Golub–Welsch algorithm in practice?
- To what extent do the answers to the previous three questions depend on implementation?
- Is the use of very high precision for the Patterson and stratified sequences of embedded integration formulas inherently unavoidable?
- Two-term analogues for current three-term methods need to be developed, in particular in the form of reliable software, culminating in a package like ORTHPOL but based on two-term recursions.
- Can the inter-node gaps of a quadrature formula be accurately computed without using higher precision?

Acknowledgements

When starting work on this article, I asked NA-Net[1] members for references to the post-1980 literature, the pre-1980 literature being very well covered in Gautschi's comprehensive survey [13]. The following members were very helpful, in some cases e-mailing me extensive BiBTEX bibliographies: Bradley Alpert, John Crow, Kai Diethelm, Alex Fedoseyev, Daniele Funaro, Walter Gautschi, Gene Golub, Bill Gragg, Martin Gutknecht, Jaroslav Kautsky, Fred Krogh, Tom Patterson, Lothar Reichel, Van Snyder, Garry Tee. If I have missed anyone, please blame it on a hard disk crash I had in October 1999. The comments of an anonymous referee have led to many improvements in the paper. Lothar Reichel and Walter Gautschi, as editors of this special volume, have saved me from some embarrassing solecisms.

[1] mailto://help@na-net.ornl.gov

References

[1] G.S. Ammar, D. Calvetti, L. Reichel, Computation of Gauss–Kronrod quadrature rules with non-positive weights, Elec. Trans. Numer. Anal. 9 (1999) 26–38.

[2] B. Beckermann, E. Bourreau, How to choose modified moments?, J. Comput. Appl. Math. 98 (1998) 81–98.

[3] D. Calvetti, G.H. Golub, W.B. Gragg, L. Reichel, Computation of Gauss-Kronrod quadrature rules, Math. Comput. 69 (2000) 1035–1052.

[4] J.J.M. Cuppen, A divide and conquer method for the symmetric tridiagonal eigenproblem, Numer. Math. 36 (1981) 177–195.

[5] C. de Boor, G.H. Golub, The numerically stable reconstruction of a Jacobi matrix from spectral data, Linear Algebra Appl. 21 (1978) 245–260.

[6] I.S. Dhillon, A new $O(n^2)$ algorithm for the symmetric tridiagonal eigenvalue/eigenvector problem, Ph.D. Thesis, University of California, Berkeley, 1997.

[7] K.V. Fernando, B.N. Parlett, Accurate singular values and differential quotient-difference algorithms, Numer. Math. 67 (1994) 191–229.

[8] H.-J. Fischer, On the condition of orthogonal polynomials via modified moments, Z. Anal. Anwendungen 15 (1996) 223–244.

[9] H.-J. Fischer, Explicit calculation of some polynomials introduced by W. Gautschi, Z. Anal. Anwendungen 17 (1998) 963–977.

[10] H.-J. Fischer, On generating orthogonal polynomials for discrete measures, Z. Anal. Anwendungen 17 (1998) 183–205.

[11] K. Gates, W.B. Gragg, Notes on TQR algorithms, J. Comput. Appl. Math. 86 (1997) 195–203.

[12] W. Gautschi, Construction of Gauss–Christoffel quadrature formulae, Math. Comput. 22 (1968) 251–270.

[13] W. Gautschi, A survey of Gauss–Christoffel quadrature formulae, in: P.L. Butzer, F. Fehér (Eds.), E.B. Christoffel: The Influence of his Work on Mathematics and the Physical Sciences, Birkhäuser, Basel, 1981, pp. 72–147.

[14] W. Gautschi, On generating orthogonal polynomials, SIAM J. Sci. Statist. Comput. 3 (1982) 289–317.

[15] W. Gautschi, How and how not to check Gaussian quadrature formulae, BIT 23 (1983) 209–216.

[16] W. Gautschi, Questions of numerical condition related to polynomials, in: G.H. Golub (Ed.), Studies in Numerical Analysis, Math. Assoc. Amer. 1984, pp. 140–177.

[17] W. Gautschi, On the sensitivity of orthogonal polynomials to perturbations in the moments, Numer. Math. 48 (1986) 369–382.

[18] W. Gautschi, Gauss–Kronrod quadrature – a survey, in: G.V. Milovanović (Ed.), Numerical Methods and Approximation Theory III, University of Niš, 1987, pp. 39–66.

[19] 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.

[20] W. Gautschi, Algorithm 726: ORTHPOL – a package of routines for generating orthogonal polynomials and Gauss-type quadrature rules, ACM Trans. Math. Software 20 (1994) 21–62.

[21] G.H. Golub, Some modified matrix eigenvalue problems, SIAM Rev. 15 (1973) 318–334.

[22] G.H. Golub, J. Kautsky, Calculation of Gauss quadratures with multiple free and fixed knots, Math. Comp. 41 (1983) 147–163.

[23] G.H. Golub, J.H. Welsch, Calculation of Gauss quadrature rules, Math. Comp. 23 (1969) 221–230.

[24] W.B. Gragg, W.J. Harrod, The numerically stable reconstruction of Jacobi matrices from spectral data, Numer. Math. 44 (1984) 317–335.

[25] M. Gu, S.C. Eisenstat, A divide-and-conquer algorithm for the bidiagonal svd, SIAM J. Matrix Anal. Appl. 16 (1995) 79–92.

[26] J. Kautsky, S. Elhay, Gauss quadratures and Jacobi matrices for weight functions not of one sign, Math. Comp. 43 (1984) 543–550.

[27] J. Kautsky, G.H. Golub, On the calculation of Jacobi matrices, Linear Algebra Appl. 52/53 (1983) 439–455.

[28] A.S. Kronrod, Nodes and Weights of Quadrature Formulas, Consultants Bureau, New York, 1965.

[29] D.P. Laurie, Stratified sequences of nested integration formulas, Quaest. Math. 15 (1992) 365–384.

[30] D.P. Laurie, Anti-Gaussian quadrature formulas, Math. Comp. 65 (1996) 739–747.

[31] D.P. Laurie, Calculation of Gauss–Kronrod quadrature formulas, Math. Comp. 66 (1997) 1133–1145.

[32] D.P. Laurie, Accurate recovery of recursion coefficients from Gaussian quadrature formulas, J. Comput. Appl. Math. 112 (1999) 165–180.

[33] D.P. Laurie, Questions related to Gaussian quadrature formulas and two-term recursions, in: W. Gautschi, G.H. Golub, G. Opfer (Eds.), Computation and Application of Orthogonal Polynomials, International Series of Numerical Mathematics, Vol. 133, Birkhäuser, Basel, 1999, pp. 133–144.

[34] S. Lewanowicz, Construction of a recurrence relation for modified moments, Technical Report, Institute of Computer Science, Wroclaw University, 1977.

[35] B.N. Parlett, The Symmetric Eigenvalue Problem, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[36] B.N. Parlett, The new qd algorithms, in: A. Iserles (Ed.), Acta Numerica 1995, Cambridge University Press, Cambridge, 1995, pp. 459–491.

[37] B.N. Parlett, I.S. Dhillon, Fernando's solution to Wilkinson's problem: an application of double factorization, Linear Algebra Appl. 267 (1997) 247–279.

[38] T.N.L. Patterson, The optimal addition of points to quadrature formulae, Math. Comp. 22 (1968) 847–856.

[39] T.N.L. Patterson, On some Gauss and Lobatto based quadrature formulae, Math. Comp. 22 (1968) 877–881.

[40] T.N.L. Patterson, Algorithm 672: generation of interpolatory quadrature rules of the highest degree of precision with pre-assigned nodes for general weight-functions, ACM Trans. Math. Software 15 (2) (1989) 137–143.

[41] L. Reichel, Fast QR decomposition of Vandermonde-like matrices and polynomial least squares approximation, SIAM J. Matrix Anal. Appl. 12 (1991) 552–564.

[42] L. Reichel, Construction of polynomials that are orthogonal with respect to a discrete bilinear form, Adv. Comput. Math. 1 (1993) 241–258.

[43] H. Rutishauser, Der Quotienten-Differenzen-Algorithmus, Birkhäuser, Basel, 1957.

[44] H. Rutishauser, On Jacobi rotation patterns, in: Experimental Arithmetic, High Speed Computation and Mathematics, American Mathematical Society, Providence, RI, 1963, pp. 219–239.

[45] R.A. Sack, A.F. Donovan, An algorithm for Gaussian quadrature given modified moments, Numer. Math. 18 (1972) 465–478.

[46] H.E. Salzer, A recurrence scheme for converting from one orthogonal expansion into another, Comm. ACM 16 (1973) 705–707.

[47] A.H. Stroud, D. Secrest, Gaussian Quadrature Formulas, Prentice-Hall, Englewood Cliffs, NJ, 1966.

Journal of Computational and Applied Mathematics 127 (2001) 219–230

JOURNAL OF
COMPUTATIONAL AND
APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Sobolev orthogonal polynomials in the complex plane

G. López Lagomasino[a, *, 1], H. Pijeira Cabrera[b], I. Pérez Izquierdo[c]

[a] *Departamento de Matematicas, Universidad Carlos III de Madrid, Av. de la Universidad, 30, 28911, Leganés, Madrid, Spain*

[b] *Universidad de Matanzas, Autopista de Varadero, Km. 3, 44740, Matanzas, Cuba*

[c] *Universidad de La Habana, San Lázaro y L, Habana 4, Cuba*

Received 15 September 1999; received in revised form 10 February 2000

Abstract

Sobolev orthogonal polynomials with respect to measures supported on compact subsets of the complex plane are considered. For a wide class of such Sobolev orthogonal polynomials, it is proved that their zeros are contained in a compact subset of the complex plane and their asymptotic-zero distribution is studied. We also find the nth-root asymptotic behavior of the corresponding sequence of Sobolev orthogonal polynomials. © 2001 Elsevier Science B.V. All rights reserved.

1. Introduction

Let $\{\mu_k\}_{k=0}^m$ be a set of $m+1$ finite positive Borel measures. For each $k=0,\ldots,m$ the support $S(\mu_k)$ of μ_k is a compact subset of the complex plane $\mathbb{C}$. We will assume that $S(\mu_0)$ contains infinitely many points. If p,q are polynomials, we define

$$\langle p,q\rangle_S=\sum_{k=0}^{m}\int p^{(k)}(x)\overline{q^{(k)}(x)}\,\mathrm{d}\mu_k(x)=\sum_{k=0}^{m}\langle p^{(k)},q^{(k)}\rangle_{L_2(\mu_k)}. \tag{1.1}$$

As usual, $f^{(k)}$ denotes the kth derivative of a function f and the bar complex conjugation. Obviously, (1.1) defines an inner product on the linear space of all polynomials. Therefore, a unique sequence of monic orthogonal polynomials is associated with it containing a representative for each degree. By Q_n we will denote the corresponding monic orthogonal polynomial of degree n. The sequence $\{Q_n\}$ is called the sequence of general monic Sobolev orthogonal polynomials relative to (1.1).

[*] Corresponding author.

E-mail addresses: lago@math.uc3m.es (G. López Lagomasino), pijeira@cdict.umtz.edu.cu (H. Pijeira Cabrera), ignacio@matcom. uh.cu (I. Pérez Izquierdo).

[1] Partially supported by Dirección General de Enseñanza Superior under Grant PB 96-0120-C03-01.

PII: S0377-0427(00)00498-2

Sobolev orthogonal polynomials have attracted considerable attention in the past decade, but only recently has there been a breakthrough in the study of their asymptotic properties for sufficiently general classes of defining measures. In this connection, we call attention to the papers [4,5,7], in which some of the first results of general character were obtained regarding nth-root, ratio, and strong asymptotics, respectively, of Sobolev orthogonal polynomials. The first two deal with measures supported on the real line and the third with measures supported on arcs and closed Jordan curves.

In this paper, we consider the nth-root asymptotic behavior; therefore, we will only comment on [4]. In [4], for measures supported on the real line and with $m = 1$, the authors assume that $\mu_0, \mu_1 \in \mathbf{Reg}$ (in the sense defined in [10]) and that their supports are regular sets with respect to the solution of the Dirichlet problem. Under these conditions, they find the asymptotic zero distribution of the zeros of the derivatives of the Sobolev orthogonal polynomials and of the sequence of Sobolev orthogonal polynomials themselves when additionally it is assumed that $S(\mu_0) \supset S(\mu_1)$. In [6], these questions were considered for arbitrary m and additional information was obtained on the location of the zeros which allowed to derive the nth-root asymptotic behavior of the Sobolev orthogonal polynomials outside a certain compact set.

The object of the present paper is to extend the results of [6] to the case when the measures involved in the inner product are supported on compact subsets of the complex plane. Under a certain domination assumption on the measures involved in the Sobolev inner product, we prove in Section 2 that the zeros of general Sobolev orthogonal polynomials are contained in a compact subset of the complex plane. For Sobolev inner products on the real line, we also study in Section 2 the case when the supports of the measures are mutually disjoint and give a sufficient condition for the boundedness of the zeros of the Sobolev orthogonal polynomials. Section 3 is dedicated to the study of the asymptotic zero distribution and nth-root asymptotic behavior of general Sobolev orthogonal polynomials. For this purpose, methods of potential theory are employed.

In order to state the corresponding results, let us fix some assumptions and additional notation. As above, (1.1) defines an inner product on the space $\mathscr{P}$ of all polynomials. The norm of $p \in \mathscr{P}$ is

$$||p||_S = \left(\sum_{k=0}^{m} \int |p^{(k)}(x)|^2 \, \mathrm{d}\mu_k(x) \right)^{1/2} = \left(\sum_{k=0}^{m} ||p^{(k)}||^2_{L_2(\mu_k)} \right)^{1/2}. \tag{1.2}$$

We say that the Sobolev inner product (1.1) is *sequentially dominated* if

$$S(\mu_k) \subset S(\mu_{k-1}), \quad k = 1, \ldots, m,$$

and

$$\mathrm{d}\mu_k = f_{k-1} \, \mathrm{d}\mu_{k-1}, \qquad f_{k-1} \in L_\infty(\mu_{k-1}), \quad k = 1, \ldots, m.$$

For example, if all the measures in the inner product are equal, then it is sequentially dominated. The concept of a sequentially dominated Sobolev inner product was introduced in [6] for the real case (when the supports of the measures are contained in the real line).

Theorem 1.1. *Assume that the Sobolev inner product* (1.1) *is sequentially dominated; then for each* $p \in \mathscr{P}$ *we have that*

$$||xp||_S \leqslant C||p||_S, \tag{1.3}$$

where

$$C=(2[C_1^2+(m+1)^2C_2])^{1/2}, \tag{1.4}$$

and

$$C_1=\max_{x\in S(\mu_0)}|x|, \qquad C_2=\max_{k=0,\ldots,m-1}||f_k||_{L_\infty(\mu_k)}.$$

As usual, two norms $||\cdot||_1$ and $||\cdot||_2$ on a given normed space E are said to be equivalent if there exist positive constants c_1,c_2 such that

$$c_1||x||_1\leqslant||x||_2\leqslant c_2||x||_1, \quad x\in E.$$

If a Sobolev inner product defines a norm on $\mathscr{P}$ which is equivalent to that defined by a sequentially dominated Sobolev inner product, we say that the Sobolev inner product is *essentially sequentially dominated.* It is immediate from the previous theorem that a Sobolev inner product which is essentially sequentially dominated also satisfies (1.3) (in general, with a constant C different from (1.4)). Whenever (1.3) holds, we say that the multiplication operator is bounded on the space of all polynomials. This property implies in turn the uniform boundedness of the zeros of Sobolev orthogonal polynomials.

Theorem 1.2. *Assume that for some positive constant C we have that*

$$||xp||_S\leqslant C||p||_S, \quad p\in\mathscr{P}.$$

Then all the zeros of the Sobolev orthogonal polynomials are contained in the disk $\{z:|z|\leqslant C\}$. In particular, this is true if the Sobolev inner product is essentially sequentially dominated.

In a recent paper, see Theorem 4.1 in [9] (for related questions see also [1]), the author proves for Sobolev inner products supported on the real line that the boundedness of the multiplication operator implies that the corresponding Sobolev inner product is essentially sequentially dominated. Therefore, in terms of the boundedness of the multiplication operator on the space of polynomials, we cannot obtain more information on the uniform boundedness of the zeros of the Sobolev orthogonal polynomials than that expressed in the theorem above. It is well known that in the case of usual orthogonality the uniform boundedness of the zeros implies that the multiplication operator is bounded. In general, this is not the case for Sobolev inner products, as the following result illustrates. In the sequel, $\mathrm{Co}(K)$ denotes the convex hull of a compact set K.

Theorem 1.3. *For $m=1$, assume that $S(\mu_0)$ and $S(\mu_1)$ are contained in the real line and*

$$\mathrm{Co}(S(\mu_0))\cap\mathrm{Co}(S(\mu_1))=\emptyset.$$

Then for all $n\geqslant 2$ the zeros of Q_n' are simple, contained in the interior of $\mathrm{Co}(S(\mu_0)\cup S(\mu_1))$, and the zeros of the Sobolev orthogonal polynomials lie in the disk centered at the extreme point of $\mathrm{Co}(S(\mu_1))$ furthest away from $S(\mu_0)$ and radius equal to twice the diameter of $\mathrm{Co}(S(\mu_0)\cup S(\mu_1))$.

The statements of this theorem will be complemented below. As Theorem 1.3 clearly indicates, Theorem 1.2 is far from giving an answer to the question of uniform boundedness of the zeros of

Sobolev orthogonal polynomials. The main question remains; that is, prove or disprove that for any compactly supported Sobolev inner product the zeros of the corresponding Sobolev orthogonal polynomials are uniformly bounded. This question is of vital importance in the study of the asymptotic behaviour of Sobolev orthogonal polynomials.

We mention some concepts needed to state the result on the nth-root asymptotic behaviour of Sobolev orthogonal polynomials. For any polynomial q of exact degree n, we denote

$$v(q):=\frac{1}{n}\sum_{j=1}^{n}\delta_{z_j},$$

where $z_1,\dots,z_n$ are the zeros of q repeated according to their multiplicity, and δ_{z_j} is the Dirac measure with mass one at the point z_j. This is the so-called normalized zero counting measure associated with q. In [10], the authors introduce a class **Reg** of regular measures. For measures supported on a compact set of the complex plane, they prove that (see Theorem 3.1.1) $\mu \in \mathbf{Reg}$ if and only if

$$\lim_{n\to\infty} \|Q_n\|_{L_2(\mu)}^{1/n} = \operatorname{cap}(S(\mu)),$$

where Q_n denotes the nth monic orthogonal polynomials (in the usual sense) with respect to μ and $\operatorname{cap}(S(\mu))$ denotes the logarithmic capacity of $S(\mu)$. In case that $S(\mu)$ is a regular compact set with respect to the solution of the Dirichlet problem on the unbounded connected component of the complement of $S(\mu)$ in the extended complex plane, the measure μ belongs to **Reg** (see Theorem 3.2.3 in [10]) if and only if

$$\lim_{n\to\infty}\left(\frac{\|p_n\|_{S(\mu)}}{\|p_n\|_{L_2(\mu)}}\right)^{1/n} = 1 \tag{1.5}$$

for every sequence of polynomials $\{p_n\}, \deg p_n \leqslant n, p_n \not\equiv 0$. Here and in the following, $\|\cdot\|_{S(\mu)}$ denotes the supremum norm on $S(\mu)$.

Set

$$\Delta = \bigcup_{k=0}^{m} S(\mu_k).$$

We call this set the support of the Sobolev inner product. Denote by $g_\Omega(z;\infty)$ the Green's function of the region Ω with singularity at infinity, where Ω is the unbounded connected component of the complement of Δ in the extended complex plane. When Δ is regular, then the Green's function is continuous up to the boundary, and we extend it continuously to all of $\mathbb{C}$ assigning to it the value zero on the complement of Ω. By ω_Δ we denote the equilibrium measure of Δ. Assume that there exists $l \in \{0,\dots,m\}$ such that $\bigcup_{k=0}^{l} S(\mu_k) = \Delta$, where $S(\mu_k)$ is regular, and $\mu_k \in \mathbf{Reg}$ for $k=0,\dots,l$. Under these assumptions, we say that the Sobolev inner product (1.1) is *l-regular.*

The next result is inspired by Theorem 1 and Corollary 13 of [4].

Theorem 1.4. *Let the Sobolev inner product* (1.1) *be l-regular. Then for each fixed* $k=0,\dots,l$ *and for all* $j \geqslant k$

$$\limsup_{n\to\infty} \|Q_n^{(j)}\|_{S(\mu_k)}^{1/n} \leqslant \operatorname{cap}(\Delta). \tag{1.6}$$

For all $j \geqslant l$

$$\lim_{n\to\infty} ||Q_n^{(j)}||_{\Delta}^{1/n} = \operatorname{cap}(\Delta). \tag{1.7}$$

Furthermore, if the interior of Δ is empty and its complement connected, then for all $j \geqslant l$

$$\lim_{n\to\infty} v(Q_n^{(j)}) = \omega_{\Delta} \tag{1.8}$$

in the weak star topology of measures.

The following example illustrates that (1.8) is not a direct consequence of (1.7). On the unit circle, take $\mu_j, j = 0, \ldots, m$, equal to the Lebesgue measure. This Sobolev inner product is 0-regular and thus (1.7) holds for all $j \geqslant 0$. Obviously, $\{z^n\}$ is the corresponding sequence of monic Sobolev orthogonal polynomials whose sequence of normalized zero counting measures converges in the weak star topology to the Dirac measure with mass one at zero. When (1.7) is true then (1.8) holds if it is known that $\lim_{n\to\infty} v(Q_n^{(j)})(A) = 0$ for every compact set A contained in the union of the bounded components of $\mathbb{C} \setminus S(\omega_{\Delta})$ (see Theorem 2.1 in [2]). But finding general conditions on the measures involved in the inner product which would guarantee this property is, in general, an open problem already in the case of usual orthogonality.

If the inner product is sequentially dominated, then $S(\mu_0) = \Delta$; therefore, if $S(\mu_0)$ and μ_0 are regular, the corresponding inner product is 0-regular. In the sequel, $\mathbb{Z}_+ = \{0, 1, \ldots\}$. An immediate consequence of Theorems 1.2 and 1.4 is the following.

Theorem 1.5. *Assume that for some positive constant C we have that*

$$||xp||_S \leqslant C||p||_S, \quad p \in \mathscr{P},$$

and that the Sobolev inner product is l-regular. Then, for all $j \geqslant l$

$$\limsup_{n\to\infty} |Q_n^{(j)}(z)|^{1/n} \leqslant \operatorname{cap}(\Delta) e^{g_{\Omega}(z;\infty)}, \quad z \in \mathbb{C}. \tag{1.9}$$

Furthermore,

$$\lim_{n\to\infty} |Q_n^{(j)}(z)|^{1/n} = \operatorname{cap}(\Delta) e^{g_{\Omega}(z;\infty)}, \tag{1.10}$$

uniformly on each compact subset of $\{z\colon |z| > C\} \cap \Omega$. Finally, if the interior of Δ is empty and its complement connected, we have equality in (1.9) for all $z \in \mathbb{C}$ except for a set of capacity zero, $S(\omega_{\Delta}) \subset \{z\colon |z| \leqslant C\}$, and

$$\lim_{n\to\infty} \frac{Q_n^{(j+1)}(z)}{nQ_n^{(j)}(z)} = \int \frac{d\omega_{\Delta}(x)}{z-x}, \tag{1.11}$$

uniformly on compact subsets of $\{z\colon |z| > C\}$.

These results will be complemented in the sections below. In the rest of the paper, we maintain the notations and definitions introduced above.

2. Zero location

The proof of Theorem 1.1 is simple and can be carried out following the same procedure as for an analogous result in [6], so we leave it to the reader. Theorem 1.2 also has an analogue in [6], but we have found a nice short proof which we include here.

Proof of Theorem 1.2. Let Q_n denote the nth Sobolev orthogonal polynomial. Since it cannot be orthogonal to itself, it is of degree n. Let x_0 denote one of its zeros. It is obvious that there exists a polynomial q of degree $n-1$ such that $xq=x_0q+Q_n$. Since Q_n is orthogonal to q, and using the boundednes of the multiplication operator, we obtain

$$|x_0|\,||q||_S=||x_0q||_S\leqslant||xq||_S\leqslant C||q||_S.$$

Simplifying $||q||_S(\neq 0)$ in the inequality above, we obtain the bound claimed on $|x_0|$ independent of n. The rest of the statements follow from Theorem 1.1. □

Now, let us consider the special case refered to in Theorem 1.3. For the proof of the corresponding result we need some auxiliary lemmas. Let I be a given interval of the real line (open or closed) and q a polynomial. By $c(q;I)$ and $\kappa(q;I)$ we denote the number of zeros and the number of changes of sign, respectively, that the polynomial q has on the interval I.

Lemma 2.1. *Let I be an interval of the real line and q a polynomial such that* $\deg q=l\geqslant 1$. *We have that*

$$c(q;I)+c(q';\mathbb{C}\setminus I)\leqslant l.$$

Proof. By Rolle's Theorem, it follows that

$$c(q;I)\leqslant c(q';I)+1.$$

Therefore,

$$c(q;I)+c(q';\mathbb{C}\setminus I)\leqslant c(q';I)+1+c(q';\mathbb{C}\setminus I)=c(q';\mathbb{C})+1=l,$$

as we wanted to prove. □

As above, let Q_n denote the nth monic Sobolev orthogonal polynomial with respect to (1.1), where all the measures are supported on the real line. In the rest of this section, we denote by $(\cdot)^{\mathrm{o}}$ the interior of the set in parentheses with the Euclidean topology on $\mathbb{R}$.

Lemma 2.2. *Assume that $n\geqslant 1$. Then*

$$\kappa(Q_n;(\mathrm{Co}(S(\mu_0)))^{\mathrm{o}})\geqslant 1.$$

Proof. If, on the contrary, Q_n does not change sign on the indicated set, we immediately obtain a contradiction from the fact that Q_n is orthogonal to 1, since then

$$0=\langle Q_n,1\rangle_S=\int Q_n(x)\,\mathrm{d}\mu_0(x)\neq 0. \qquad \square$$

Unless otherwise stated, in the rest of this section we restrict our attention to the case presented in Theorem 1.3. That is, $m = 1$, the supports of μ_0 and μ_1 are contained in the real line and their convex hulls do not intersect.

Lemma 2.3. *Under the hypothesis of Theorem* 1.3, *for* $n \geqslant 1$, *we have that*

$$\kappa(Q_n;(\mathrm{Co}(S(\mu_0)))^o) + \kappa(Q_n';(\mathrm{Co}(S(\mu_1)))^o) \geqslant n-1. \tag{2.12}$$

Proof. For $n = 1,2$ the statement follows from Lemma 2.2. Let $n \geqslant 3$ and assume that (2.12) does not hold. That is,

$$\kappa(Q_n;(\mathrm{Co}(S(\mu_0)))^o) + \kappa(Q_n';(\mathrm{Co}(S(\mu_1)))^o) = l \leqslant n-2. \tag{2.13}$$

Without loss of generality, we can assume that

$$\mathrm{Co}(S(\mu_0)) = [a,b], \qquad \mathrm{Co}(S(\mu_1)) = [c,d], \quad b < c.$$

This reduction is always possible by means of a linear change of variables.

Let x_0 be the point in (a,b) closest to $[c,d]$ where Q_n changes sign. This point exists by Lemma 2.2. There are two possibilities: either

$$Q_n'(x_0+\varepsilon)Q_n'(c+\varepsilon) > 0 \tag{2.14}$$

for all sufficiently small $\varepsilon > 0$, or

$$Q_n'(x_0+\varepsilon)Q_n'(c+\varepsilon) < 0 \tag{2.15}$$

for all sufficiently small $\varepsilon > 0$. Let us consider separately each case.

Assume that (2.14) holds. Let q be a polynomial of degree $\leqslant l$ with real coefficients, not identically equal to zero, which has a zero at each point of (a,b) where Q_n changes sign and whose derivative has a zero at each point of (c,d) where Q_n' changes sign. The existence of such a polynomial q reduces to solving a system of l equations in $l+1$ unknowns (the coefficients of q). Thus, a nontrivial solution always exists. Notice that

$$l \leqslant c(q;(a,b)) + c(q';(c,d))$$

with strict inequality if either q (resp. q') has on (a,b) (resp. (c,d)) zeros of multiplicity greater than one or distinct from those assigned by construction. On the other hand, because of Lemma 2.2, the degree of q is at least 1; therefore, using Lemma 2.1, we have that

$$c(q;(a,b)) + c(q';(c,d)) \leqslant \deg q \leqslant l.$$

The last two inequalities imply that

$$l = c(q;(a,b)) + c(q';(c,d)) = \deg q.$$

Hence, qQ_n and $q'Q_n'$ have constant sign on $[a,b]$ and $[c,d]$, respectively. We can choose q in such a way that $qQ_n \geqslant 0$ on $[a,b]$ (if this were not so replace q by $-q$). With this selection, for all sufficiently small $\varepsilon > 0$, we have that $q'(x_0+\varepsilon)Q_n'(x_0+\varepsilon) > 0$. All the zeros of q' are contained in $(a,x_0) \cup (c,d)$, so q' preserves its sign along the interval $(x_0, c+\varepsilon)$, for all sufficiently small $\varepsilon > 0$.

On the other hand, we are in case (2.14) where Q_n' has the same sign to the right of x_0 and of c. Therefore, $q'Q_n' \geqslant 0$ on $[c,d]$. Since $\deg q \leqslant n-2$, using orthogonality, we obtain the contradiction

$$0 = \int q(x)Q_n(x)\,\mathrm{d}\mu_0(x) + \int q'(x)Q_n'(x)\,\mathrm{d}\mu_1(x) > 0.$$

So, (2.14) cannot hold if (2.13) is true.

Let us assume that we are in the situation (2.15). The difference is that to the right of x_0 and c the polynomial Q_n' has different signs. Notice (see (2.13)) that we have at least one degree of freedom left to use orthogonality. Here, we construct q of degree $\leqslant l+1$ with real coefficients and not identically equal to zero with the same interpolation conditions as above plus $q'(c) = 0$. Following the same line of reasoning as above, we have that qQ_n and $q'Q_n'$ preserve their sign on $[a,b]$ and $[c,d]$, respectively. Taking q so that $qQ_n \geqslant 0$ on $[a,b]$, one can see that also $q'Q_n' \geqslant 0$ on $[c,d]$. Since $\deg q = l+1 \leqslant n-1$, using orthogonality, we obtain that (2.15) is not possible under (2.13). But either (2.14) or (2.15) must hold, thus (2.12) must be true. □

Corollary 2.4. *Set* $I = \mathrm{Co}(S(\mu_0) \cup S(\mu_1)) \setminus (\mathrm{Co}(S(\mu_0)) \cup \mathrm{Co}(S(\mu_1)))$. *Under the conditions of Theorem* 1.3, *we have that*

$$c(Q_n; I) + c(Q_n'; I) \leqslant 1.$$

Proof. This is an immediate consequence of Lemmas 2.1 and 2.3 applied to Q_n. □

Proof of Theorem 1.3. We will employ the notation introduced for the proof of Lemma 2.3. According to Lemmas 2.1 and 2.3,

$$n-1 \leqslant l = \kappa(Q_n; (a,b)) + \kappa(Q_n'; (c,d)) \leqslant n.$$

If $l = n$, then by Rolle's Theorem we have that all the zeros of Q_n' are simple and contained in $(a,b) \cup (c,d)$, which implies our first statement.

Suppose that $l = n-1$. We consider the same two cases (2.14) and (2.15) analyzed in the proof of Lemma 2.3. Following the arguments used in the proof of Lemma 2.3, we can easily see that (2.14) is not possible with $l = n-1$. If (2.15) holds, then Q_n' has an extra zero in the interval $[x_0, c]$, and again by use of Rolle's Theorem we have that all the zeros of Q_n' are simple and contained in (a,d).

In order to prove the second part of Theorem 1.3, we use the following remarkable result known as Grace's Apollarity Theorem. (We wish to thank T. Erdelyi for drawing our attention to this simple proof of the second statement.) Let q be a polynomial of degree greater than or equal to two. Take any two zeros of q in the complex plane and draw the straight line which cuts perpendicularly the segment joining the two zeros at its midpoint. Then q' has at least one zero in each of the closed half planes into which the line divides the complex plane. For the proof of this result see Theorem 1.4.7 in [8] (see also [3, pp. 23–24]).

For $n = 1$ the second statement is certainly true, because from Lemma 2.1 we know that for all $n \geqslant 1, Q_n$ has a zero on (a,b). Let $n \geqslant 2$. If Q_n had a zero outside the circle with center at d and radius equal to $|a-d|$, then by Grace's Apollarity Theorem Q_n' would have a zero outside the segment (a,d), which contradicts the first statement of the theorem. Therefore, all the zeros of Q_n lie in the indicated set. □

Remark 1. The arguments used in the proof of Lemma 2.3 allow us to deduce some other interesting properties which resemble those satisfied by usual orthogonal polynomials. For example, the interval joining any two consecutive zeros of Q_n on (a,b) intersects $S(\mu_0)$. Analogously, the interval joining any two consecutive zeros of Q_n' on (c,d) intersects $S(\mu_1)$. In order to prove this, notice that if any one of these statements were not true, then in the construction of the polynomial q in Lemma 2.3 we could disregard the corresponding zeros which gives us some extra degrees of freedom to use orthogonality, and arrive at a contradiction, as was done there. From the proof of Theorem 1.3 it is also clear that the zeros of Q_n in (a,b) are simple and interlace with the zeros of Q_n' on that set.

Remark 2. The key to the proof of Theorem 1.3 is Lemma 2.1. Its role is to guarantee that in the construction of q in Lemma 2.3 no extra zeros of q or q' fall on (a,b) or (c,d), respectively. Lemma 2.1 can be used in order to cover more general Sobolev inner products supported on the real line, as long as the supports of the measures appear in a certain order. To be more precise, following essentially the same ideas, we can prove the following result.

Consider a Sobolev inner product (1.1) supported on the real line such that for each $k=0,\dots,m-1$

$$\mathrm{Co}\left(\bigcup_{j=0}^{k} S(\mu_j)\right) \cap S(\mu_{k+1}) = \emptyset.$$

Then for all $n \geqslant m$ the zeros of $Q_n^{(m)}$ are simple and they are contained in the interior of $\mathrm{Co}(\bigcup_{j=0}^{m} S(\mu_j))$. The zeros of $Q_n^{(j)}$, $j=0,\dots,m-1$, lie in the disk centered at z_0 and radius equal to $3^{m-j}r$, where z_0 is the center of the interval $\mathrm{Co}(\bigcup_{j=0}^{m} S(\mu_j))$ and r is equal to half the length of that interval.

For $m=1$ this statement is weaker than that contained in Theorem 1.3 regarding the location of the zeros of the Q_n, because in the present conditions we allow that the support of $S(\mu_1)$ have points on both sides of $\mathrm{Co}(S(\mu_0))$.

3. Regular asymptotic zero distribution

For the proof of Theorem 1.4, we need the following lemma, which is proved in [6] and is easy to verify.

Lemma 3.1. *Let E be a compact regular subset of the complex plane and $\{P_n\}$ a sequence of polynomials such that $\deg P_n \leqslant n$ and $P_n \not\equiv 0$. Then, for all $k \in \mathbb{Z}_+$,*

$$\limsup_{n\to\infty} \left(\frac{||P_n^{(k)}||_E}{||P_n||_E}\right)^{1/n} \leqslant 1. \tag{3.16}$$

The first result of general character for the nth-root asymptotic of Sobolev orthogonal polynomials appeared in [4] (for measures supported on the real line and $m=1$). Minor details allowed two of us to extend that result to the case of arbitrary m (see [6]). Now, we present the case of Sobolev orthogonal polynomials in the complex plane. Since the proof remains essentially the same, we will only outline the main aspects.

Proof of Theorem 1.4. Let T_n denote the monic Chebyshev polynomial of degree n for the support Δ of the given Sobolev inner product. For simplicity of notation, we write $||\cdot||_{L_2(\mu_k)} = ||\cdot||_k$. By the minimizing property of the Sobolev norm of the polynomial Q_n, we have

$$||Q_n||_k \leqslant ||Q_n||_{\mathrm{S}}^2 \leqslant ||T_n||_{\mathrm{S}}^2 = \sum_{k=0}^{m} ||T_n^{(k)}||_k^2 \leqslant \sum_{k=0}^{m} \mu_k(S(\mu_k))||T_n^{(k)}||_{\Delta}^2. \tag{3.17}$$

It is well known that $\lim_{n\to\infty}||T_n||_{\Delta}^{1/n} = \operatorname{cap}(\Delta)$. Since Δ is a regular compact set, by Lemma 3.1 (applied to T_n) and (3.17) it follows that

$$\limsup_{n\to\infty} ||Q_n||_k^{1/n} \leqslant \limsup_{n\to\infty} ||Q_n||_{\mathrm{S}}^{1/n} \leqslant \operatorname{cap}(\Delta). \tag{3.18}$$

Since the measure μ_k and its support are regular, we can combine (1.5) and (3.18) to obtain that for each $k = 0,\ldots,l$,

$$\limsup_{n\to\infty} ||Q_n^{(k)}||_{S(\mu_k)}^{1/n} \leqslant \operatorname{cap}(\Delta). \tag{3.19}$$

By virtue of Lemma 3.1, relation (1.6) follows from (3.19).

If $j \geqslant l$, then (1.6) holds for each $k = 0,\ldots,l$. Since

$$||Q_n^{(j)}||_{\Delta} = \max_{k=0,\ldots,l} ||Q_n^{(j)}||_{S(\mu_k)},$$

using (1.6), we obtain

$$\limsup_{n\to\infty} ||Q_n^{(j)}||_{\Delta}^{1/n} \leqslant \operatorname{cap}(\Delta).$$

On the other hand,

$$\liminf_{n\to\infty} ||Q_n^{(j)}||_{\Delta}^{1/n} \geqslant \operatorname{cap}(\Delta)$$

is true for any sequence $\{Q_n\}$ of monic polynomials. Hence (1.7) follows.

If the compact set Δ has empty interior and connected complement, it is well known (see [2, Theorem 2.1]) that (1.7) implies (1.8). □

Remark 3. We wish to point out that in Theorem 1.4 eventually some of the measures μ_k, $k = 2,\ldots,m-1$, may be the null measure, in which case μ_k and $S(\mu_k) = \emptyset$ are considered to be regular and $||Q_n^{(j)}||_{\emptyset} = 0$. With these conventions, Theorem 1.4 remains in force.

The so-called discrete Sobolev orthogonal polynomials have attracted particular attention in the past years. They are of the form

$$\langle f, g\rangle_{\mathrm{S}} = \int f(x)\overline{g(x)}\,\mathrm{d}\mu_0(x) + \sum_{i=1}^{m}\sum_{j=0}^{N_i} A_{i,j} f^{(j)}(c_i)\overline{g^{(j)}(c_i)}, \tag{3.20}$$

where $A_{i,j} \geqslant 0, A_{i,N_i} > 0$. If any of the points c_i lie in the complement of the support $S(\mu_0)$ of μ_0, the corresponding Sobolev inner product cannot be l-regular. Nevertheless, a simple modification of the proof of Theorem 1.4 allows to consider this case. For details see [6], here we only state the corresponding result.

Theorem 3.2. *Let the discrete Sobolev inner product* (3.20) *be such that* $S(\mu_0)$ *is regular and* $\mu_0 \in \mathbf{Reg}$. *Then,* (1.7) *holds for all* $j \geqslant 0$, *with* $\Delta = S(\mu_0)$, *and so does* (1.8) *under the additional assumption that* $S(\mu_0)$ *has empty interior and connected complement.*

The proof of Theorem 1.5 contains some new elements with respect to the analogous result for Sobolev inner products supported on the real line, so we include it.

Proof of Theorem 1.5. Fix $j \in \mathbb{Z}_+$. Set

$$v_n(z) = \frac{1}{n}\log\frac{|Q_n^{(j)}(z)|}{||Q_n^{(j)}||_\Delta} - g_\Omega(z;\infty).$$

We show that

$$v_n(z) \leqslant 0, \quad z \in \mathbb{C} \cup \{\infty\}. \tag{3.21}$$

This function is subharmonic in Ω and on the boundary of Ω it is $\leqslant 0$. By the maximum principle for subharmonic functions it is $\leqslant 0$ on all Ω. On the complement of Ω we also have that $v_n(z) \leqslant 0$ because by definition (and the regularity of Δ) Green's function is identically equal to zero on this set and the other term which defines v_n is obviously at most zero by the maximum principle of analytic functions. These remarks imply (1.9) by taking the upper limit in (3.21) and using (1.7) (for this inequality no use is made of the boundedness of the multiplication operator on $\mathscr{P}$).

From Theorem 1.2, we have that for all $n \in \mathbb{Z}_+$, the zeros of the Sobolev orthogonal polynomials are contained in $\{z\colon |z| \leqslant C\}$. It is well known that the zeros of the derivative of a polynomial lie in the convex hull of the set of zeros of the polynomial itself. Therefore, the zeros of $Q_n^{(j)}$ for all $n \in \mathbb{Z}_+$ lie in $\{z\colon |z| \leqslant C\}$. Using this, we have that $\{v_n\}$ forms a family of uniformly bounded harmonic functions on each compact subset of $\{z\colon |z| > C\} \cap \Omega$ (including infinity). Take a sequence of indices Λ such that $\{v_n\}_{n\in\Lambda}$ converges uniformly on each compact subset of $\{z\colon |z| > C\} \cap \Omega$. Let v_Λ denote its limit. Obviously, v_Λ is harmonic and $\leqslant 0$ in $\{z\colon |z| > C\} \cap \Omega$, and because of (1.7), $v_\Lambda(\infty) = 0$. Therefore, $v_\Lambda \equiv 0$ in $\{z\colon |z| > C\} \cap \Omega$. Since this is true for every convergent subsequence of $\{v_n\}$, we get that the whole sequence converges to zero uniformly on each compact subset of $\{z\colon |z| > C\} \cap \Omega$. This is equivalent to (1.10).

If in addition the interior of Δ is empty and its complement connected, we can use (1.8). The measures $v_{n,j} = v(Q_n^{(j)})$, $n \in \mathbb{Z}_+$, and ω_Δ have their support contained in a compact subset of $\mathbb{C}$. Using this and (1.8), from the Lower Envelope Theorem (see [10, p. 223]), we obtain

$$\liminf_{n\to\infty} \int \log\frac{1}{|z-x|}\,\mathrm{d}v_{n,j}(x) = \int \log\frac{1}{|z-x|}\,\mathrm{d}\omega_\Delta(x),$$

for all $z \in \mathbb{C}$ except for a set of zero capacity. This is equivalent to having equality in (1.9) except for a set of capacity zero, because (see [10, p. 7])

$$g_\Omega(z;\infty) = \log\frac{1}{\operatorname{cap}(\Delta)} - \int \log\frac{1}{|z-x|}\,\mathrm{d}\omega_\Delta(x).$$

Let $x_{n,i}^j, i = 1,\ldots,n-j$, denote the $n-j$ zeros of $Q_n^{(j)}$. As mentioned above, all these zeros are contained in $\{z\colon |z| \leqslant C\}$. From (1.8), each point of $S(\omega_\Delta)$ must be a limit point of zeros of $\{Q_n^{(j)}\}$; therefore, $S(\omega_\Delta) \subset \{z\colon |z| \leqslant C\}$. Decomposing in simple fractions and using the definition of $v_{n,j}$, we obtain

$$\frac{Q_n^{(j+1)}(z)}{nQ_n^{(j)}(z)} = \frac{1}{n}\sum_{i=1}^{n-j}\frac{1}{z-x_{n,i}^{j}} = \frac{n-j}{n}\int\frac{\mathrm{d}v_{n,j}(x)}{z-x}. \tag{3.22}$$

Therefore, for each fixed $j \in \mathbb{Z}_+$, the family of functions

$$\left\{\frac{Q_n^{(j+1)}(z)}{nQ_n^{(j)}(z)}\right\}, \quad n \in \mathbb{Z}_+, \tag{3.23}$$

is uniformly bounded on each compact subset of $\{z\colon |z| > C\}$.

On the other hand, all the measures $v_{n,j}$, $n \in \mathbb{Z}_+$, are supported in $\{z\colon |z| \leqslant C\}$ and for $z, |z| > C$, fixed, the function $(z-x)^{-1}$ is continuous on $\{x\colon |x| \leqslant C\}$ with respect to x. Therefore, from (1.8) and (3.22), we find that any subsequence of (3.23) which converges uniformly on compact subsets of $\{z\colon |z| > C\}$ converges pointwise to $\int (z-x)^{-1}\,\mathrm{d}\omega_{\Delta}(x)$. Thus, the whole sequence converges uniformly to this function on compact subsets of $\{z\colon |z| > C\}$, as stated in (1.11). □

References

[1] V. Alvarez, D. Pestana, J.M. Rodríguez, E. Romera, Generalized weighted Sobolev spaces and applications to Sobolev orthogonal polynomials, preprint.

[2] H.P. Blatt, E.B. Saff, M. Simkani, Jentzsch-Szegő type theorems for the zeros of best approximants, J. London Math. Soc. 38 (1988) 192–204.

[3] P. Borwein, T. Erdelyi, in: Polynomials and Polynomial Inequalities, Graduate Texts in Mathematics, Vol. 161, Springer, New York, 1995.

[4] W. Gautschi, A.B.J. Kuijlaars, Zeros and critical points of Sobolev Orthogonal Polynomials, J. Approx. Theory 91 (1997) 117–137.

[5] G. López Lagomasino, F. Marcellán, W. Van Assche, Relative asymptotics for orthogonal polynomials with respect to a discrete Sobolev inner product, Constr. Approx. 11 (1995) 107–137.

[6] G. López Lagomasino, H. Pijeira Cabrera, Zero location and nth-root asymptotics of Sobolev orthogonal polynomials, J. Approx. Theory 99 (1999) 30–43.

[7] A. Martínez Finkelshtein, Bernstein-Szegő's theorem for Sobolev orthogonal polynomials, Constr Approx. 16 (2000) 73–84.

[8] G.V. Milovanović, D.S. Mitrinović, Th.M. Rassias, Topics in Polynomials: Extremal problems, inequalities, zeros, World Scientific, Singapore, 1994.

[9] J.M. Rodríguez, Multiplication operator in Sobolev spaces with respect to measures, preprint.

[10] H. Stahl, V. Totik, General Orthogonal Polynomials, Cambridge University Press, Cambridge, 1992.

ELSEVIER

Journal of Computational and Applied Mathematics 127 (2001) 231–254

JOURNAL OF
COMPUTATIONAL AND
APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

On the “Favard theorem” and its extensions ☆

Francisco Marcellán[a,*], Renato Álvarez-Nodarse[b,c]

[a]Departamento de Matemáticas, Universidad Carlos III de Madrid, Ave. de la Universidad 30, E-28911 Leganés, Madrid, Spain

[b,c]Departamento de Análisis Matemático, Universidad de Sevilla, Apdo. 1160, E-41080 Sevilla, Spain

[c]Instituto Carlos I de Física Teórica y Computacional, Universidad de Granada, E-18071, Granada, Spain

Received 22 July 1999; received in revised form 11 January 2000

Abstract

In this paper we present a survey on the “Favard theorem” and its extensions. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Favard theorem; Recurrence relations

1. Introduction

Given a sequence $\{P_n\}_{n=0}^{\infty}$ of monic polynomials satisfying a certain recurrence relation, we are interested in finding a general inner product, if one exists, such that the sequence $\{P_n\}_{n=0}^{\infty}$ is orthogonal with respect to it.

The original “classical” result in this direction is due to Favard [10] even though his result seems to be known to different mathematicians. The first who obtained a similar result was Stieltjes in 1894 [23]. In fact, from the point of view of J-continued fractions obtained from the contraction of an S-continued fraction with positive coefficients, Stieltjes proved the existence of a positive linear functional such that the denominators of the approximants are orthogonal with respect to it [23, Section 11]. Later on, Stone gave another approach using the spectral resolution of a self-adjoint operator associated with a Jacobi matrix [24, Theorem 10.23]. In his paper [21, p. 454] Shohat

☆ This work has been partially supported by the European project INTAS-93-219-ext and by the Spanish Dirección General de Enseñanza Superior (DGES) grant PB-96-0120-C03-01.

* Corresponding author. Tel. +34-91-624-94-42, fax: +34-91-624-91-51.

E-mail addresses: pacomarc@ing.uc3m.es (F. Marcellán), ran@cica.es (R. Álvarez-Nodarse).

PII: S0377-0427(00)00497-0

claims "We have been in possession of this proof for several years. Recently Favard published an identical proof in the *Comptes Rendus*". Also Natanson in his book [17, p. 167] said "This theorem was also discovered (independent of Favard) by the author (Natanson) in the year 1935 and was presented by him in a seminar led by S.N. Bernstein. He then did not publish the result since the work of Favard appeared in the meantime". The "same" theorem was also obtained by Perron [19], Wintner [28] and Sherman [20], among others.

Favard's result essentially means that if a sequence of monic polynomials $\{P_n\}_{n=0}^{\infty}$ satisfies a three-term recurrence relation

$$xP_n(x) = P_{n+1}(x) + a_nP_n(x) + b_nP_{n-1}(x), \tag{1.1}$$

with a_n, $b_n \in \mathbb{R}$, $b_n > 0$, then there exists a positive Borel measure μ such that $\{P_n\}_{n=0}^{\infty}$ is orthogonal with respect to the inner product

$$\langle p, q\rangle = \int_{\mathbb{R}} pq \, \mathrm{d}\mu. \tag{1.2}$$

This formulation is equivalent to the following: Given the linear operator $t : \mathbb{P} \to \mathbb{P}$, $p(t) \to tp(t)$, characterize an inner product such that the operator t is Hermitian with respect to the inner product.

A first extension of this problem is due to Chihara [5]. If $\{P_n\}_{n=0}^{\infty}$ satisfies a three-term recurrence relation like (1.1) with a_n, $b_n \in \mathbb{C}$, $b_n \neq 0$, find a linear functional $\mathscr{L}$ defined on $\mathbb{P}$, the linear space of polynomials with complex coefficients, such that $\{P_n\}_{n=0}^{\infty}$ is orthogonal with respect to the general inner product $\langle p, q\rangle = \mathscr{L}[pq]$, where $p, q \in \mathbb{P}$. Notice that in the case analyzed by Favard [10] the linear functional has an integral representation

$$\mathscr{L}[p] = \int_{\mathbb{R}} p \, \mathrm{d}\mu.$$

Favard's Theorem is an inverse problem in the sense that from information about polynomials we can deduce what kind of inner product induces orthogonality for such polynomials. The aim of this contribution is to survey some extensions of the Favard Theorem when a sequence of monic polynomials $\{P_n\}_{n=0}^{\infty}$ satisfies recurrence relations of a different form than (1.1).

In the first place, in [8] a similar problem is studied relating to polynomials orthogonal with respect to a positive Borel measure v supported on the unit circle, which satisfy a recurrence relation

$$\Phi_n(z) = z\Phi_{n-1}(z) + \Phi_n(0)\Phi_{n-1}^*(z), \quad |\Phi_n(0)| < 1, \tag{1.3}$$

where $\Phi_n^*(z) = z^n\overline{\Phi_n(1/\bar{z})}$.

Thus, a Favard Theorem means, in this case, to identify an inner product in $\mathbb{P}$ such that $\{\Phi_n\}_{n=0}^{\infty}$ satisfying (1.3) is the corresponding sequence of orthogonal polynomials.

The structure of the paper is as follows. In Section 2 we present a survey of results surrounding the Favard Theorem when a sequence of polynomials satisfies a linear relation like (1.1). In particular, we show that the interlacing property for the zeros of two consecutive polynomials gives basic information about the preceding ones in the sequence of polynomials.

In Section 3, an analogous approach is presented in the case of the unit circle in a more general situation when $|\Phi_n(0)| \neq 1$. Furthermore, an integral representation for the corresponding inner product is given. The connection with the trigonometric moment problem is stated when we assume that the nth polynomial Φ_n is coprime with Φ_n^*.

In Section 4, we present some recent results about a natural extension of the above Favard theorems taking into account their interpretation in terms of operator theory. Indeed, the multiplication by t is a Hermitian operator with respect to (1.2) and a unitary operator with respect to the inner product

$$\langle p, q\rangle = \int_{\mathbb{R}} p(e^{i\theta})\overline{q(e^{i\theta})}\,dv(\theta). \tag{1.4}$$

Thus, we are interested in characterizing inner products such that the multiplication by a fixed polynomial is a Hermitian or a unitary operator. The connection with matrix orthogonal polynomials is stated, and some examples relating to Sobolev inner products are given.

2. The Favard theorem on the real line

2.1. Preliminaries

In this subsection we summarize some definitions and preliminary results that will be useful throughout the work. Most of them can be found in [5].

Definition 2.1. Let $\{\mu_n\}_{n=0}^{\infty}$ be a sequence of complex numbers (moment sequence) and $\mathscr{L}$ a functional acting on the linear space of polynomials $\mathbb{P}$ with complex coefficients. We say that $\mathscr{L}$ is a moment functional associated with $\{\mu_n\}_{n=0}^{\infty}$ if $\mathscr{L}$ is linear, i.e., for all polynomials π_1 and π_2 and any complex numbers α_1 and α_2,

$$\mathscr{L}[\alpha_1\pi_1 + \alpha_2\pi_2] = \alpha_1\mathscr{L}[\pi_1] + \alpha_2\mathscr{L}[\pi_2] \quad \text{and} \quad \mathscr{L}[x^n] = \mu_n, \quad n = 0, 1, 2, \ldots .$$

Definition 2.2. Given a sequence of polynomials $\{P_n\}_{n=0}^{\infty}$, we say that $\{P_n\}_{n=0}^{\infty}$ is a sequence of orthogonal polynomials (SOP) with respect to a moment functional $\mathscr{L}$ if for all nonnegative integers n and m the following conditions hold:
(1) P_n is a polynomial of exact degree n,
(2) $\mathscr{L}[P_nP_m] = 0$, $m \neq n$,
(3) $\mathscr{L}[P_n^2] \neq 0$.
Usually, the last two conditions are replaced by

$$\mathscr{L}[x^mP_n(x)] = K_n\delta_{n,m}, \quad K_n \neq 0, \quad 0 \leqslant m \leqslant n,$$

where $\delta_{n,m}$ is the Kronecker symbol.

The next theorems are direct consequences of the above definition [5, Chapter I, Sections 2 and 3, pp. 8–17].

Theorem 2.3. *Let $\mathscr{L}$ be a moment functional and $\{P_n\}_{n=0}^{\infty}$ a sequence of polynomials. Then the following are equivalent*:
(1) *$\{P_n\}_{n=0}^{\infty}$ is an SOP with respect to $\mathscr{L}$.*
(2) *$\mathscr{L}[\pi P_n] = 0$ for all polynomials π of degree $m < n$, while $\mathscr{L}[\pi P_n] \neq 0$ if the degree of π is n.*
(3) *$\mathscr{L}[x^mP_n(x)] = K_n\delta_{n,m}$, where $K_n \neq 0$, for $m = 0, 1, \ldots, n$.*

Theorem 2.4. *Let $\{P_n\}_{n=0}^{\infty}$ be an SOP with respect to $\mathscr{L}$. Then, for every polynomial π of degree n*

$$\pi(x) = \sum_{k=0}^{n} d_k P_k(x) \quad \textit{where } d_k = \frac{\mathscr{L}[\pi P_k]}{\mathscr{L}[P_k^2]}, \quad k = 0, 1, \ldots, n. \tag{2.1}$$

A simple consequence of the above theorem is that an SOP is uniquely determined if we impose an additional condition that fixes the leading coefficient k_n of the polynomials ($P_n(x) = k_n x^n +$ lower-order terms). When $k_n = 1$ for all $n = 0, 1, 2, \ldots$ the corresponding SOP is called a monic SOP (MSOP). If we choose $k_n = (\mathscr{L}[P_n^2])^{-1/2}$, the SOP is called an orthonormal SOP (SONP).

The next question which obviously arises is the existence of an SOP. To answer this question, it is necessary to introduce the Hankel determinants Δ_n,

$$\Delta_n = \begin{vmatrix} \mu_0 & \mu_1 & \cdots & \mu_n \\ \mu_1 & \mu_2 & \cdots & \mu_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_n & \mu_{n+1} & \cdots & \mu_{2n} \end{vmatrix}.$$

Theorem 2.5. *Let $\mathscr{L}$ be a moment functional associated with the sequence of moments $\{\mu_n\}_{n=0}^{\infty}$. Then, the sequence of polynomials $\{P_n\}_{n=0}^{\infty}$ is an SOP with respect to $\mathscr{L}$ if and only if $\Delta_n \neq 0$ for all nonnegative n. Moreover, the leading coefficient k_n of the polynomial P_n is given by $k_n = K_n \Delta n - 1/\Delta_n$.*

Definition 2.6. A moment functional $\mathscr{L}$ is called positive definite if for every nonzero and nonnegative real polynomial π, $\mathscr{L}[\pi] > 0$.

The following theorem characterizes the positive-definite functionals in terms of the moment sequences $\{\mu_n\}_{n=0}^{\infty}$. The proof is straightforward.

Theorem 2.7. *A moment functional $\mathscr{L}$ is positive definite if and only if their moments are real and $\Delta_n > 0$ for all $n \geqslant 0$.*

Using the above theorem, we can define a positive-definite moment functional $\mathscr{L}$ entirely in terms of the determinants Δ_n. In other words, a moment functional $\mathscr{L}$ is called positive definite if all its moments are real and $\Delta_n > 0$ for all $n \geqslant 0$. Notice also that for a MSOP, it is equivalent to say that $K_n > 0$ for all $n \geqslant 0$. This, and the fact that an SOP exists if and only if $\Delta_n \neq 0$, leads us to define more general moment functionals: the so-called *quasi-definite* moment functionals.

Definition 2.8. A moment functional $\mathscr{L}$ is said to be quasi-definite if and only if $\Delta_n \neq 0$ for all $n \geqslant 0$.

We can write the explicit expression of the MOP in terms of the moments of the corresponding functional:

$$P_n(x) = \frac{1}{\Delta_{n-1}} \begin{vmatrix} \mu_0 & \mu_1 & \cdots & \mu_n \\ \mu_1 & \mu_2 & \cdots & \mu_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{n-1} & \mu_n & \cdots & \mu_{2n-1} \\ 1 & x & \cdots & x^n \end{vmatrix}, \quad \Delta_{-1} \equiv 1, \quad n = 0, 1, 2, \ldots . \tag{2.2}$$

One of the simplest characteristics of orthogonal polynomials is the so-called three-term recurrence relation (TTRR) that connects every three consecutive polynomials of the SOP.

Theorem 2.9. *If $\{P_n\}_{n=0}^{\infty}$ is an MSOP with respect to a quasi-definite moment functional, then the polynomials P_n satisfy a three-term recurrence relation*

$$P_n(x) = (x - c_n)P_{n-1}(x) - \lambda_n P_{n-2}(x), \quad n = 1, 2, 3, \ldots, \tag{2.3}$$

where $\{c_n\}_{n=0}^{\infty}$ and $\{\lambda_n\}_{n=0}^{\infty}$ are given by

$$c_n = \frac{\mathscr{L}[xP_{n-1}^2]}{\mathscr{L}[P_{n-1}^2]}, \quad n \geqslant 1, \quad \textit{and} \quad \lambda_n = \frac{\mathscr{L}[xP_{n-1}P_{n-2}]}{\mathscr{L}[P_{n-2}^2]} = \frac{\mathscr{L}[P_{n-1}^2]}{\mathscr{L}[P_{n-2}^2]}, \quad n \geqslant 2,$$

respectively, and $P_{-1}(x) \equiv 0$, $P_0(x) \equiv 1$.

The proof of the above theorem is a simple consequence of the orthogonality of the polynomials and Theorem 2.2. A straightforward calculation shows that ($\lambda_1 = \mathscr{L}[1]$)

$$\lambda_{n+1} = \frac{K_n}{K_{n-1}} = \frac{\Delta_{n-2}\Delta_n}{\Delta_{n-1}^2}, \quad n = 1, 2, 3, \ldots,$$

and $\Delta_{-1} \equiv 1$. From Theorem 2.7 and Definition 2.8 it follows that, if $\lambda_n \neq 0$, then $\mathscr{L}$ is quasi-definite whereas, if $\lambda_n > 0$, then $\mathscr{L}$ is positive definite. Notice also that from the above expression we can obtain the square norm $K_n \equiv \mathscr{L}[P_n^2]$ of the polynomial P_n as

$$K_n \equiv \mathscr{L}[P_n^2] = \lambda_1\lambda_2 \cdots \lambda_{n+1}. \tag{2.4}$$

A useful consequence of Theorem 2.5 are the Christoffel–Darboux identities.

Theorem 2.10. *Let $\{P_n\}_{n=0}^{\infty}$ be an MSOP which satisfies (2.3) with $\lambda_n \neq 0$ for all nonnegative n. Then*

$$\sum_{m=0}^{n} \frac{P_m(x)P_m(y)}{K_m} = \frac{1}{K_n} \frac{P_{n+1}(x)P_n(y) - P_{n+1}(y)P_n(x)}{x - y}, \quad n \geqslant 0, \tag{2.5}$$

and

$$\sum_{m=0}^{n} \frac{P_m^2(x)}{K_m} = \frac{1}{K_n}[P'_{n+1}(x)P_n(x) - P_{n+1}(x)P'_n(x)], \quad n \geqslant 0. \tag{2.6}$$

For an arbitrary normalization (not necessarily the monic one) of the polynomials P_n, the three-term recurrence relation becomes

$$xP_{n-1}(x) = \alpha_n P_n(x) + \beta_n P_{n-1}(x) + \gamma_n P_{n-2}(x). \tag{2.7}$$

In this case, the coefficients α_n and β_n can be obtained comparing the coefficients of x^n and x^{n-1}, respectively, in both sides of (2.7) and γ_n is given by $\mathscr{L}[xP_{n-1}P_{n-2}]/\mathscr{L}[P_{n-2}^2]$. This leads to

$$\alpha_n = \frac{k_{n-1}}{k_n}, \quad \beta_n = \frac{b_{n-1}}{k_{n-1}} - \frac{b_n}{k_n}, \quad \gamma_n = \frac{k_{n-2}}{k_{n-1}}\frac{K_{n-1}}{K_{n-2}}, \tag{2.8}$$

where k_n is the leading coefficient of P_n and b_n denotes the coefficient of x^{n-1} in P_n, i.e., $P_n(x) = k_n x^n + b_n x^{n-1} + \cdots$. Notice also that knowing two of the coefficients α_n, β_n, and γ_n, one can find the third one using (2.7) provided, for example, that $P_n(x_0) \neq 0$ for some x_0 (usually $x_0 = 0$) and for all $n = 1, 2, 3, \ldots$.

The above TTRR (2.7) can be written in matrix form,

$$x\boldsymbol{P}_{n-1} = J_n \boldsymbol{P}_{n-1} + \alpha_n P_n(x)\boldsymbol{e}_n, \tag{2.9}$$

where

$$\boldsymbol{P}_{n-1} = \begin{bmatrix} P_0(x) \\ P_1(x) \\ P_2(x) \\ \vdots \\ P_{n-2}(x) \\ P_{n-1}(x) \end{bmatrix}, \quad J_n = \begin{bmatrix} \beta_1 & \alpha_1 & 0 & \ldots & 0 & 0 \\ \gamma_2 & \beta_2 & \alpha_2 & \ldots & 0 & 0 \\ 0 & \gamma_3 & \beta_3 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \beta_{n-1} & \alpha_{n-1} \\ 0 & 0 & 0 & \ldots & \gamma_n & \beta_n \end{bmatrix}, \quad \boldsymbol{e}_n = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \tag{2.10}$$

Denoting by $\{x_{n,j}\}_{1 \leqslant j \leqslant n}$ the zeros of the polynomial P_n, we see from (2.9) that each $x_{n,j}$ is an eigenvalue of the corresponding tridiagonal matrix of order n and $[P_0(x_{n,j}), \ldots, P_{n-1}(x_{n,j})]^{\mathrm{T}}$ is the associated eigenvector. From the above representation many useful properties of zeros of orthogonal polynomials can be found.

2.2. The zeros of orthogonal polynomials

Definition 2.11. Let $\mathscr{L}$ be a moment functional. The support of the functional $\mathscr{L}$ is the largest interval $(a, b) \subset \mathbb{R}$ where $\mathscr{L}$ is positive definite.

The following theorem holds.

Theorem 2.12. *Let (a, b) be the support of the positive-definite functional $\mathscr{L}$, and let $\{P_n\}_{n=0}^{\infty}$ be the MSOP associated with $\mathscr{L}$. Then,*

(1) *All zeros of P_n are real, simple, and located inside (a, b).*

(2) *Two consecutive polynomials P_n and P_{n+1} have no common zeros.*

(3) *Let* $\{x_{n,j}\}_{j=1}^{n}$ *denote the zeros of the polynomial* P_n, *with* $x_{n,1} < x_{n,2} < \cdots < x_{n,n}$. *Then,*

$$x_{n+1,j} < x_{n,j} < x_{n+1,j+1}, \quad j = 1,2,3,\ldots,n.$$

The last property is usually called the interlacing property.

Proof. Notice that, in the case when the SOP is an SONP, i.e., $K_n = 1$ for all n, then the matrix J_n is a symmetric real matrix ($J_n = J_n^{\mathrm{T}}$, where J_n^{T} denotes the transposed matrix of J_n). So its eigenvalues, and thus, the zeros of the orthogonal polynomials are real. To prove that all zeros are simple, we can use the Christoffel–Darboux identity (2.6). Let x_k be a multiple zero of P_n, i.e., $P_n(x_k) = P_n'(x_k) = 0$. Then (2.6) gives

$$0 < \sum_{m=0}^{n} \frac{P_m^2(x_k)}{K_m} = \frac{1}{K_n}[P_{n+1}'(x_k)P_n(x_k) - P_{n+1}(x_k)P_n'(x_k)] = 0.$$

This contradiction proves the statement. Let $\{x_k\}_{k=1}^{p}$ be the zeros of P_n inside (a,b). Then, $P_n(x)\prod_{k=1}^{p}(x - x_k)$ does not change sign in (a,b) and $\mathscr{L}[P_n(x)\prod_{k=1}^{p}(x - x_k)] \neq 0$, so $p = n$, i.e., all the zeros of P_n are inside (a,b). Thus, the statement 1 is proved. To prove 2, we use the TTRR. In fact, if x_k is a zero of P_n and P_{n+1}, then it must be a zero of P_{n-1}. Continuing this process by induction, we get that x_k must be a zero of $P_0(x) \equiv 1$, which is a contradiction. Before proving the interlacing property 3 we will prove a theorem due to Cauchy [22, p. 197].

Theorem 2.13. *Let* B *be a principal* $(n-1) \times (n-1)$ *submatrix of a real symmetric* $n \times n$ *matrix* A, *with eigenvalues* $\mu_1 \geqslant \mu_2 \geqslant \cdots \geqslant \mu_{n-1}$. *Then, if* $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$ *are the eigenvalues of* A,

$$\lambda_1 \geqslant \mu_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \mu_{n-1} \geqslant \lambda_n.$$

Proof. Let A be the $n \times n$ matrix

$$A = \begin{pmatrix} B & a \\ a^{\mathrm{T}} & b \end{pmatrix},$$

and assume that the theorem is not true, i.e., $\mu_i > \lambda_i$ or $\lambda_{i+1} > \mu_i$ (since the matrix A is real symmetric, all its eigenvalues are real). Let i be the first such index. If $\mu_i > \lambda_i$ (the other case is similar), there exists a real number τ such that $\mu_i > \tau > \lambda_i$. Then, $B - \tau I_{n-1}$, where I_k denotes the identity $k \times k$ matrix, is nonsingular ($\det(B - \tau I_{n-1}) \neq 0$), and the matrix

$$H = \begin{pmatrix} B - \tau I_{n-1} & 0 \\ 0 & b - \tau - a^{\mathrm{T}}(B - \tau I_{n-1})^{-1}a \end{pmatrix}$$

$$= \begin{pmatrix} I_{n-1} & 0 \\ -a^{\mathrm{T}}(B - \tau I_{n-1})^{-1} & 1 \end{pmatrix} \begin{pmatrix} B - \tau I_{n-1} & a \\ a^{\mathrm{T}} & b - \tau \end{pmatrix} \begin{pmatrix} I & -(B - \tau I_{n-1})^{-1}a \\ 0 & 1 \end{pmatrix}$$

is congruent to $A - \tau I_n$. Then, by the inertia theorem, the matrix H has the same number of positive eigenvalues as $A - \tau I_n$, i.e., $i - 1$. But H has at least as many positive eigenvalues as $B - \tau I_{n-1}$, i.e., i. The contradiction proves the theorem. □

Obviously, the interlacing property 3 can be obtained as a simple corollary of the Cauchy Theorem, since the matrix J_n associated with the SONP is a real symmetric matrix and we can choose as A the matrix J_{n+1} whose eigenvalues are the zeros of the polynomial P_{n+1} and then, the principal submatrix B is the matrix J_n whose eigenvalues coincide with the zeros of P_n. This completes the proof of Theorem 2.12. □

2.3. The Favard Theorem and some applications

In this subsection we will prove the so-called Favard Theorem.

Theorem 2.14. *Let $\{c_n\}_{n=0}^{\infty}$ and $\{\lambda_n\}_{n=0}^{\infty}$ be two arbitrary sequences of complex numbers, and let $\{P_n\}_{n=0}^{\infty}$ be a sequence of polynomials defined by the relation*

$$P_n(x) = (x - c_n)P_{n-1}(x) - \lambda_n P_{n-2}(x), \quad n = 1,2,3,\ldots, \tag{2.11}$$

where $P_{-1}(x) = 0$ and $P_0(x) = 1$. Then, there exists a unique moment functional $\mathscr{L}$ such that

$$\mathscr{L}[1] = \lambda_1, \qquad \mathscr{L}[P_n\, P_m] = 0 \quad if\ n \neq m.$$

Moreover, $\mathscr{L}$ is quasi-definite and $\{P_n\}_{n=0}^{\infty}$ is the corresponding MSOP if and only if $\lambda_n \neq 0$, and $\mathscr{L}$ is positive definite if and only if c_n are real numbers and $\lambda_n > 0$ for all $n = 1,2,3,\ldots$.

Proof. To prove the theorem, we will define the functional $\mathscr{L}$ by induction on $\mathbb{P}_n$, the linear subspace of polynomials with degree at most n. We put

$$\mathscr{L}[1] = \mu_0 = \lambda_1, \qquad \mathscr{L}[P_n] = 0, \quad n = 1,2,3,\ldots\ . \tag{2.12}$$

So, using the three-term recurrence relation (2.11), we can find all the moments in the following way: Since $\mathscr{L}[P_n] = 0$, the TTRR gives

$$0 = \mathscr{L}[P_1] = \mathscr{L}[x - c_1] = \mu_1 - c_1\lambda_1, \quad \text{then } \mu_1 = c_1\lambda_1,$$

$$0 = \mathscr{L}[P_2] = \mathscr{L}[(x - c_2)P_1 - \lambda_2 P_0] = \mu_2 - (c_1 + c_2)\mu_1 + (c_1c_2 - \lambda_2)\lambda_1,$$

then we can find μ_2, etc. Continuing this process, we can find, recursively, μ_{n+1} by using the TTRR, and they are uniquely determined. Next, using (2.11) and (2.12), we deduce that

$$x^k P_n(x) = \sum_{i=n-k}^{n+k} d_{n,i} P_i(x).$$

Then, $\mathscr{L}[x^k\, P_n] = 0$ for all $k = 0,1,2,\ldots,n-1$. Finally,

$$\mathscr{L}[x^n P_n] = \mathscr{L}[x^{n-1}(P_{n+1} + c_{n+1}P_n + \lambda_{n+1}P_{n-1})] = \lambda_{n+1}\mathscr{L}[x^{n-1}P_{n-1}],$$

so, $\mathscr{L}[x^n P_n] = \lambda_{n+1}\lambda_n \cdots \lambda_1$.

Moreover, $\mathscr{L}$ is quasi-definite and $\{P_n\}_{n=0}^{\infty}$ is the corresponding MSOP if and only if for all $n \geqslant 1$, $\lambda_n \neq 0$, while $\mathscr{L}$ is positive definite and $\{P_n\}_{n=0}^{\infty}$ is the corresponding MSOP if and only if for all $n \geqslant 1$, $c_n \in \mathbb{R}$ and $\lambda_n > 0$. □

Next, we will discuss some results dealing with the zeros of orthogonal polynomials.

The following theorem is due to Wendroff [27] (for a different point of view using the Bézoutian matrix see [2]).

Theorem 2.15. *Let P_n and P_{n-1} be two monic polynomials of degree n and $n-1$, respectively. If $a < x_1 < x_2 < \cdots < x_n < b$ are the real zeros of P_n and $y_1 < y_2 < \cdots < y_{n-1}$ are the real zeros of P_{n-1}, and they satisfy the interlacing property, i.e.,*

$$x_i < y_i < x_{i+1}, \quad i = 1,2,3,\ldots,n-1,$$

then there exists a family of polynomials $\{P_k\}_{k=0}^{n}$ orthogonal on $[a,b]$ such that the above polynomials P_n and P_{n-1} belong to it.

Proof. Let $c_n = x_1 + x_2 + \cdots + x_n - y_1 - y_2 - \cdots - y_{n-1}$. Then, the polynomial $P_n(x) - (x - c_n)P_{n-1}(x)$ is a polynomial of degree at most $n-2$, i.e.,

$$P_n(x) - (x - c_n)P_{n-1}(x) \equiv -\lambda_n R(x),$$

where R is a monic polynomial of degree r at most $n-2$. Since

$$x_1 - c_n = (y_1 - x_2) + \cdots + (y_{n-1} - x_n) < 0,$$

and $P_{n-1}(x_1) \neq 0$ (this is a consequence of the interlacing property), then $\lambda_n \neq 0$ and $R(x_1) \neq 0$. Moreover, $P_n(y_i) = -\lambda_n R(y_i)$. Now, using the fact that $P_n(y_i)P_n(y_{i+1}) < 0$ (again this is a consequence of the interlacing property), we conclude that also $R(y_i)R(y_{i+1}) < 0$, and this immediately implies that R has exactly $n-2$ real zeros and they satisfy $y_i < z_i < y_{i+1}$ for $i = 1,2,\ldots,n-2$.

If we now define the polynomial P_{n-2} of degree exactly $n-2$, $P_{n-2} \equiv R$, whose zeros interlace with the zeros of P_{n-1}, we can construct, just repeating the above procedure, a polynomial of degree $n-3$ whose zeros interlace with the ones of P_{n-2}, etc. So we can find all polynomials P_k for $k = 1,2,\ldots,n$.

Notice also that, by construction,

$$P_n(x) = (x - c_n)P_{n-1}(x) - \lambda_n P_{n-2}(x),$$

so

$$\lambda_n = \frac{(x_1 - c_n)P_{n-1}(x_1)}{P_{n-2}(x_1)} > 0,$$

because $\operatorname{sign} P_{n-1}(x_1) = (-1)^{n-1}$ and $\operatorname{sign} P_{n-2}(x_1) = (-1)^{n-2}$, which is a consequence of the interlacing property $x_1 < y_1 < z_1$. □

We point out here that it is possible to complete the family $\{P_k\}_{k=0}^{n}$ to obtain a MSOP. To do this, we can define the polynomials P_k for $k = n+1, n+2, \ldots$ recursively by the expression

$$P_{n+j}(x) = (x - c_{n+j})P_{n+j-1}(x) - \lambda_{n+j}P_{n+j-2}(x), \quad j = 1,2,3,\ldots,$$

where c_{n+j} and λ_{n+j} are real numbers chosen such that $\lambda_{n+j} > 0$ and the zeros of P_{n+j} lie on (a,b). Notice also that, in such a way, we have defined, from two given polynomials P_{n-1} and P_n, a sequence of polynomials satisfying a three-term recurrence relation of the form (2.11). So Theorem 2.14 states that the corresponding sequence is an orthogonal polynomial sequence with respect to a quasi-definite functional. Moreover, since the coefficients in (2.11) are real and $\lambda_{n+j} > 0$, the corresponding functional is positive definite.

Theorem 2.16 (Vinuesa and Guadalupe [26]; Nevai and Totik [18]). *Let $\{x_n\}_{n=1}^{\infty}$ and $\{y_n\}_{n=1}^{\infty}$ be two sequences of real numbers such that*

$$\cdots < x_3 < x_2 < x_1 = y_1 < y_2 < y_3 < \cdots .$$

Then there exists a unique system of monic polynomials $\{P_n\}_{n=0}^{\infty}$ orthogonal with respect to a positive definite functional on the real line such that $P_n(x_n) = P_n(y_n) = 0$ and $P_n(t) \neq 0$ for $t \notin [x_n, y_n]$, $n = 1, 2, \dots$.

Proof. Set $P_0 = 1$, $\lambda_0 = 0$ and $c_0 = x_1$. Define $\{P_n\}_{n=1}^{\infty}$, $\{c_n\}_{n=1}^{\infty}$ and $\{\lambda_n\}_{n=1}^{\infty}$ by

$$P_n(x) = (x - c_n)P_{n-1}(x) - \lambda_n P_{n-2}(x), \quad n \geqslant 1,$$

$$\lambda_n = (x_n - y_n)\left[\frac{P_{n-2}(x_n)}{P_{n-1}(x_n)} - \frac{P_{n-2}(y_n)}{P_{n-1}(y_n)}\right]^{-1}, \quad c_n = x_n - \lambda_n \frac{P_{n-2}(x_n)}{P_{n-1}(x_n)}. \tag{2.13}$$

The above two formulas come from the TTRR and from the requirement $P_n(x_n) = P_n(y_n) = 0$. By induction one can show that $P_n(x) \neq 0$ if $x \notin [x_n, y_n]$, $P_n(x_n)=P_n(y_n)=0$ and $\lambda_{n+1} > 0$ for $n=0,1,2,\dots$. Then, from Theorem 2.14 $\{P_n\}_{n=0}^{\infty}$ is a MSOP with respect to a positive definite moment functional. □

Notice that, in the case $x_n = -y_n$, for $n = 1, 2, 3, \dots$, the expression (2.13) for λ_n and c_n reduces to

$$\lambda_n = x_n \frac{P_{n-2}(x_n)}{P_{n-1}(x_n)}, \quad c_n = 0.$$

3. The Favard Theorem on the unit circle

3.1. Preliminaries

In this subsection we will summarize some definitions and results relating to orthogonal polynomials on the unit circle $\mathbb{T} = \{|z| = 1, z \in \mathbb{C}\}$. See [13].

Definition 3.1. Let $\{\mu_n\}_{n\in\mathbb{Z}}$ be a bisequence of complex numbers (moment sequence) such that $\mu_{-n} = \bar{\mu}_n$ and $\mathscr{L}$ be a functional on the linear space of Laurent polynomials $\Lambda = \mathrm{Span}\{z^k\}_{k\in\mathbb{Z}}$. We say that $\mathscr{L}$ is a moment functional associated with $\{\mu_n\}$ if $\mathscr{L}$ is linear and $\mathscr{L}(x^n) = \mu_n$, $n \in \mathbb{Z}$.

Definition 3.2. Given a sequence of polynomials $\{\Phi_n\}_{n=0}^{\infty}$ we say that $\{\Phi_n\}_{n=0}^{\infty}$ is a sequence of orthogonal polynomials (SOP) with respect to a moment functional $\mathscr{L}$ if

(i) Φ_n is a polynomial of exact degree n,

(ii) $\mathscr{L}(\Phi_n(z)\cdot z^{-m}) = 0$, if $0 \leqslant m \leqslant n-1$, $\mathscr{L}(\Phi_n(z)\cdot z^{-n}) = S_n \neq 0$, for every $n = 0, 1, 2, \dots$.

For such a linear functional $\mathscr{L}$ we can define a Hermitian bilinear form in $\mathbb{P}$ (the linear space of polynomials with complex coefficients) as follows:

$$\langle p(z), q(z)\rangle = \mathscr{L}(p(z)\cdot \overline{q(1/\bar{z})}), \tag{3.1}$$

where $\overline{q(z)}$ denotes the complex conjugate of the polynomial $q(z)$.

Notice that Definition 3.2 means that $\{\Phi_n\}_{n=0}^{\infty}$ is an SOP with respect to the above bilinear form, and thus the idea of orthogonality appears, as usual, in the framework of Hermitian bilinear forms. Furthermore,

$$\langle zp(z), zq(z)\rangle = \langle p(z), q(z)\rangle, \tag{3.2}$$

i.e., the shift operator is unitary with respect to the bilinear form (3.1). In particular, the Gram matrix for the canonical basis $\{z^n\}_{n=0}^{\infty}$ is a structured matrix of Toeplitz type, i.e.,

$$\langle z^m, z^n\rangle = \langle z^{m-n}, 1\rangle = \langle 1, z^{n-m}\rangle = \mu_{m-n}, \quad m,n \in \mathbb{N}.$$

In this case the entries (m,n) of the Gram matrix depend of the difference $m-n$.

In the following we will denote $T_n = [\mu_{k-j}]_{k,j=0}^{n}$.

Now we will deduce some recurrence relations for the respective sequence of monic orthogonal polynomials.

Theorem 3.3. *Let $\mathscr{L}$ be a moment functional associated with the bisequence $\{\mu_n\}_{n\in\mathbb{Z}}$. The sequence of polynomials $\{\Phi_n\}_{n=0}^{\infty}$ is an SOP with respect to $\mathscr{L}$ if and only if* $\det T_n \neq 0$ *for every $n = 0,1,2,\ldots$. Furthermore, the leading coefficient of Φ_n is $s_n = \det T_{n-1}/\det T_n$.*

Definition 3.4. $\mathscr{L}$ is said to be a positive-definite moment functional if for every Laurent polynomial $q(z) = p(z)\overline{p(1/\bar{z})}$, $\mathscr{L}(q) > 0$.

Theorem 3.5. *$\mathscr{L}$ is a positive-definite functional if and only if* $\det T_n > 0$ *for every $n = 0,1,2,\ldots$.*

Definition 3.6. $\mathscr{L}$ is said to be a quasi-definite moment functional if $\det T_n \neq 0$ for every $n = 0,1,2,\ldots$.

Remark. Compare the above definitions with those of Section 2.1.

In the following we will assume that the SOP $\{\Phi_n\}_{n=0}^{\infty}$ is normalized using the fact that the leading coefficient is one, i.e., we have a sequence of monic orthogonal polynomials (MSOP) given by $(n = 0,1,2,\ldots)$

$$\Phi_n(x) = \frac{1}{\det T_{n-1}} \begin{vmatrix} \mu_0 & \mu_1 & \cdots & \mu_n \\ \mu_{-1} & \mu_0 & \cdots & \mu_{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{-n+1} & \mu_{-n+2} & \cdots & \mu_1 \\ 1 & z & \cdots & z^n \end{vmatrix}, \quad \det T_{-1} \equiv 1. \tag{3.3}$$

Unless stated otherwise, we will suppose the linear functional $\mathscr{L}$ is quasi-definite.

Theorem 3.7 (Geronimus [11]). *If $\{\Phi_n\}_{n=0}^{\infty}$ is an MSOP with respect to a quasi-definite moment functional, it satisfies two recurrence relations*:

(i) $\Phi_n(z) = z\Phi_{n-1}(z) + \Phi_n(0)\Phi_{n-1}^{*}(z)$, $\Phi_0(z) = 1$ (*forward recurrence relation*),

(ii) $\Phi_n(z) = (1 - |\Phi_n(0)|^2)z\Phi_{n-1}(z) + \Phi_n(0)\Phi_n^*(z)$, $\Phi_0(z) = 1$ (*backward recurrence relation*), *where* $\Phi_n^*(z) = z^n\overline{\Phi_n(1/\bar{z})}$ *is called the reciprocal polynomial of* Φ_n.

Proof. (i) Let $R_{n-1}(z) = \Phi_n(z) - z\Phi_{n-1}(z)$. Thus, from orthogonality and (3.2)

$$\langle R_{n-1}(z), z^k\rangle = \mathscr{L}(z^k \cdot \overline{R_{n-1}(1/\bar{z})}) = \mathscr{L}(z^{k-n+1} \cdot z^{n-1}\overline{R_{n-1}(1/\bar{z})}) = 0,$$

for $k = 1, 2, \dots, n-1$, and $\mathscr{L}(z^{-j} \cdot z^{n-1}\overline{R_{n-1}(1/\bar{z})}) = 0$, $j = 0, 1, \dots, n-2$.

This means that the polynomial of degree at most $n-1$, $z^{n-1}\overline{R_{n-1}(1/\bar{z})}$, with leading coefficient $\overline{\Phi_n(0)}$, is orthogonal to $\mathbb{P}_{n-2}$, i.e.,

$$z^{n-1}\overline{R_{n-1}(1/\bar{z})} = \overline{\Phi_n(0)}\Phi_{n-1}(z).$$

Thus, $R_{n-1}(z) = \Phi_n(0)\Phi_{n-1}^*(z)$.

(ii) From (i) we deduce

$$\Phi_n^*(z) = \Phi_{n-1}^*(z) + \overline{\Phi_n(0)}z\Phi_{n-1}(z).$$

Then, the substitution of $\Phi_{n-1}^*(z)$ in (i), using the above expression, leads to (ii). □

Remark. Notice that, if we multiply both sides of (ii) by $1/z^n$, use the orthogonality of Φ_n as well as the explicit expression (3.3), we get the following identity:

$$\frac{\det T_n}{\det T_{n-1}} = (1 - |\Phi_n(0)|^2)\frac{\det T_{n-1}}{\det T_{n-2}}. \tag{3.4}$$

The values $\{\Phi_n(0)\}_{n=1}^{\infty}$ are called reflection coefficients or Schur parameters for the MSOP. Notice that the main difference with the recurrence relation analyzed in Section 2 is that here only two consecutive polynomials are involved and the reciprocal polynomial is needed. On the other hand, the basic parameters which appear in these recurrence relations are the value at zero of the orthogonal polynomial.

Theorem 3.8. $\mathscr{L}$ *is a quasi-definite moment functional if and only if* $|\Phi_n(0)| \neq 1$ *for every* $n = 1, 2, 3, \dots$.

Proof. If $\mathscr{L}$ is quasi-definite the corresponding MSOP satisfies both (i) and (ii). If for some $n \in \mathbb{N}$, $|\Phi_n(0)| = 1$, then from (ii), $\Phi_n(z) = \Phi_n(0)\Phi_n^*(z)$. Thus,

$$\begin{aligned}\langle\Phi_n(z), z^n\rangle &= \Phi_n(0)\langle\Phi_n^*(z), z^n\rangle = \Phi_n(0)\langle z^n\overline{\Phi_n(1/\bar{z})}, z^n\rangle \\ &= \Phi_n(0)\langle\overline{\Phi_n(1/\bar{z})}, 1\rangle = \Phi_n(0), \Phi_n(z)\rangle = 0,\end{aligned}$$

which is a contradiction with the fact that $\{\Phi_n\}_{n=1}^{\infty}$ is an MSOP.

Assume now that a sequence of polynomials is defined by (i) with $|\Phi_n(0)| \neq 1$. We will prove by induction that there exists a moment functional $\mathscr{L}$ which is quasi-definite and such that $\{\Phi_n\}_{n=1}^{\infty}$ is the corresponding sequence of MOP.

Let $\Phi_1(z) = z + \Phi_1(0)$. We define $\mu_1 = \mathscr{L}(z) = -\Phi_1(0)\mu_0$. Thus

$$T_1 = \begin{pmatrix} \mu_0 & \mu_1 \\ \overline{\mu_1} & \mu_0 \end{pmatrix}$$

is such that $\det T_1 = \mu_0^2(1 - |\Phi_1(0)|^2) \neq 0$.

Furthermore,

$$\langle \Phi_1(z), z\rangle = \mathscr{L}(\Phi_1(z)\cdot 1/z) = \mu_0 + \Phi_1(0)\overline{\mu_1} = \mu_0(1-|\Phi_1(0)|^2) \neq 0,$$

i.e., Φ_1 is a monic polynomial of degree 1 such that $\langle \Phi_1(z), 1\rangle = \mu_1 + \Phi_1(0)\mu_0 = 0$, i.e., is orthogonal to $\mathbb{P}_0$.

Assume $\{\Phi_0, \Phi_1, \dots, \Phi_{n-1}\}$ are monic and orthogonal. Let $a_n = \Phi_n(0)$, $|a_n| \neq 1$, and construct a polynomial Φ_n of degree n such that

$$\Phi_n(z) = z\Phi_{n-1}(z) + \underbrace{\Phi_n(0)}_{a_n}\Phi^*_{n-1}(z).$$

If $\Phi_n(z) = z^n + c_{n,1}z^{n-1} + \cdots + c_{n,n-1}z + a_n$, we define $\mu_n = -c_{n,1}\mu_{n-1} - \cdots - c_{n,n-1}\mu_1 - a_n\mu_0$. Notice that this means that $\langle \Phi_n(z), 1\rangle = 0$.

On the other hand, for $1 \leqslant k \leqslant n-1$, using the recurrence relation (i)

$$\langle \Phi_n(z), z^k\rangle = \langle \Phi_{n-1}(z), z^{k-1}\rangle + a_n\langle \Phi^*_{n-1}(z), z^k\rangle = 0,$$

where the last term in the above sum vanishes since

$$\langle \Phi^*_{n-1}(z), z^k\rangle = \langle z^{n-k-1}, \Phi_{n-1}(z)\rangle.$$

Finally, using (3.4), we have

$$\langle \Phi_n(z), z^n\rangle = \frac{\det T_n}{\det T_{n-1}} = (1-|\Phi_n(0)|^2)\frac{\det T_{n-1}}{\det T_{n-2}},$$

and thus, because of the induction hypothesis, $\langle \Phi_n(z), z^n\rangle \neq 0$. □

Corollary 3.9. *The functional $\mathscr{L}$ is positive definite if and only if $|\Phi_n(0)| < 1$, for $n = 1, 2, \dots$.*

3.2. The zeros of the orthogonal polynomials

In the following we will analyze the existence of an integral representation for a moment functional.

First, we will consider the case of positive definiteness.

Proposition 3.10 (Landau [12]). *If α is a zero of $\Phi_n(z)$, then $|\alpha| < 1$.*

Proof. Let $\Phi_n(z) = (z-\alpha)q_{n-1}(z)$, where q_{n-1} is a polynomial of degree $n-1$. Then,

$$\begin{aligned}0 < \langle \Phi_n(z), \Phi_n(z)\rangle &= \langle (z-\alpha)q_{n-1}(z), \Phi_n(z)\rangle = \langle zq_{n-1}(z), \Phi_n(z)\rangle\\ &= \langle zq_{n-1}(z), zq_{n-1}(z) - \alpha q_{n-1}(z)\rangle = \langle q_{n-1}(z), q_{n-1}(z)\rangle - \bar{\alpha}\langle zq_{n-1}(z), q_{n-1}(z)\rangle\\ &= \langle q_{n-1}(z), q_{n-1}(z)\rangle - \bar{\alpha}[\langle \Phi_n(z), q_{n-1}(z)\rangle + \alpha\langle q_{n-1}(z), q_{n-1}(z)\rangle]\\ &= (1-|\alpha|^2)\langle q_{n-1}(z), q_{n-1}(z)\rangle,\end{aligned}$$

and the result follows. □

Corollary 3.11 (Montaner and Alfaro [16]). *If β is a zero of $\Phi^*_n(z)$, then $|\beta| > 1$.*

Remark. Notice that, in the quasi-definite case, we only can guarantee that $|\alpha| \neq 1$.

Next, we will define an absolutely continuous measure such that the induced inner product in $\mathbb{P}_n$ agrees with the restriction to $\mathbb{P}_n$ of our inner product associated with the positive-definite linear functional. In order to do this, we need some preliminary result.

Lemma 3.12 (Erdélyi et al. [8]). *Let ϕ_n be the nth orthonormal polynomial with respect to a positive definite linear functional. Then,*

$$\frac{1}{2\pi}\int_0^{2\pi} \phi_k(\mathrm{e}^{\mathrm{i}\theta})\overline{\phi_j(\mathrm{e}^{\mathrm{i}\theta})}\frac{\mathrm{d}\theta}{|\phi_n(\mathrm{e}^{\mathrm{i}\theta})|^2} = \delta_{j,k}, \quad 0 \leqslant j \leqslant k \leqslant n < \infty.$$

Proof. Notice that

$$\frac{1}{2\pi}\int_0^{2\pi} \phi_n(\mathrm{e}^{\mathrm{i}\theta})\overline{\phi_n(\mathrm{e}^{\mathrm{i}\theta})}\frac{\mathrm{d}\theta}{|\phi_n(\mathrm{e}^{\mathrm{i}\theta})|^2} = 1, \tag{3.5}$$

and, for $j < n$,

$$\begin{aligned}\frac{1}{2\pi}\int_0^{2\pi} \phi_n(\mathrm{e}^{\mathrm{i}\theta})\overline{\phi_j(\mathrm{e}^{\mathrm{i}\theta})}\frac{\mathrm{d}\theta}{|\phi_n(\mathrm{e}^{\mathrm{i}\theta})|^2} &= \frac{1}{2\pi}\int_0^{2\pi} \overline{\left[\frac{\phi_j(\mathrm{e}^{\mathrm{i}\theta})}{\phi_n(\mathrm{e}^{\mathrm{i}\theta})}\right]}\mathrm{d}\theta \\ &= \frac{1}{2\pi}\int_0^{2\pi} \frac{\mathrm{e}^{\mathrm{i}(n-j)\theta}\phi_j^*(\mathrm{e}^{\mathrm{i}\theta})}{\phi_n^*(\mathrm{e}^{\mathrm{i}\theta})}\,\mathrm{d}\theta \\ &= \frac{1}{2\pi i}\int_{\mathbb{T}} \frac{z^{n-j-1}\phi_j^*(z)}{\phi_n^*(z)}\,\mathrm{d}z = 0, \end{aligned} \tag{3.6}$$

because of the analyticity of the function in the last integral (see Corollary 3.11). Then, $\phi_n(z)$ is the nth orthonormal polynomial with respect to both, a positive linear functional and the absolutely continuous measure $\mathrm{d}v_n = \mathrm{d}\theta/|\phi_n(\mathrm{e}^{\mathrm{i}\theta})|^2$. By virtue of the backward recurrence relation (Theorem 3.7(ii)) for the orthonormal case, the polynomials $\{\phi_j\}_{j=0}^{n-1}$, which are uniquely defined by this recurrence relation, are orthogonal with respect to both, the linear functional and the measure $\mathrm{d}v_n$. Thus, the result follows. □

Remark. In [8], an induction argument is used in order to prove the previous result. Indeed, assuming that for a fixed $k \leqslant n$,

$$\frac{1}{2\pi}\int_0^{2\pi} \phi_k(\mathrm{e}^{\mathrm{i}\theta})\overline{\phi_j(\mathrm{e}^{\mathrm{i}\theta})}\frac{\mathrm{d}\theta}{|\phi_n(\mathrm{e}^{\mathrm{i}\theta})|^2} = \delta_{j,k}, \quad 0 \leqslant j \leqslant k,$$

they proved that

$$\frac{1}{2\pi}\int_0^{2\pi} \phi_{k-1}(\mathrm{e}^{\mathrm{i}\theta})\overline{\phi_l(\mathrm{e}^{\mathrm{i}\theta})}\frac{\mathrm{d}\theta}{|\phi_n(\mathrm{e}^{\mathrm{i}\theta})|^2} = \delta_{k-1,l}, \quad 0 \leqslant l \leqslant k-1.$$

Notice that the nth orthogonal polynomial defines in a unique way the previous ones; thus, the proof of the second statement (the induction) is not necessary. Of course, here we need not do this since we are using the backward recurrence relation for the orthogonal polynomials ϕ_n.

Notice also that the measure $\mathrm{d}\nu_n = \mathrm{d}\theta/|\phi_n(\mathrm{e}^{i\theta})|^2$ defines an MSOP $\{\Psi_n\}_{n=0}^{\infty}$ such that $\Psi_m(z) = z^{m-n}\Phi_n(z)$, for $m \geqslant n$, where Φ_n is the monic polynomial corresponding to ϕ_n. Moreover, the sequence of reflection coefficients corresponding to this MSOP $\{\Psi_n\}_{n=0}^{\infty}$ is $\{\Phi_1(0), \ldots, \Phi_n(0), 0, 0, \ldots\}$. Usually, in the literature of orthogonal polynomials, this measure $\mathrm{d}\nu_n$ is called a Bernstein–Szegö measure (see [25]).

In Section 2, Theorem 2.15, we proved that the interlacing property for the zeros of two polynomials P_{n-1} and P_n of degree $n-1$ and n, respectively, means that they are the $(n-1)$st and nth orthogonal polynomials of an MSOP. Indeed, the three-term recurrence relation for a MSOP plays a central role in the proof. In the case of the unit circle, we have an analogous result, which is known in the literature as the Schur–Cohn–Jury criterion [4].

Theorem 3.13. *A monic polynomial p of degree n has its n zeros inside the unit circle if and only if the family of parameters $\{a_k\}_{k=0}^{n}$ defined by the following backward algorithm*

$$q_n(z) = p(z), \quad q_n(0) = a_n,$$

$$q_k(z) = \frac{q_{k+1}(z) - a_{k+1}q_{k+1}^*(z)}{z(1-|a_{k+1}|^2)}, \quad a_k = q_k(0), \quad k = n-1, n-2, \ldots, 0,$$

satisfies $|a_k| < 1,\ k = 1, 2, \ldots, n$.

Proof. Notice that the polynomials $\{q_k\}_{k=1}^{n}$, $q_0 = 1$, satisfy a backward recurrence relation like the polynomials orthogonal on the unit circle with truncated Schur parameters $\{a_k\}_{k=1}^{\infty}$. Because $\{a_1, a_2, \ldots, a_n, 0, 0, \ldots\}$ is induced by the measure $\mathrm{d}\nu_n = \mathrm{d}\theta/|q_n(\mathrm{e}^{\mathrm{i}\theta})|^2 = \mathrm{d}\theta/|p(\mathrm{e}^{\mathrm{i}\theta})|^2$, up to a constant factor, then $p = q_n(z)$ is the nth monic orthogonal polynomial with respect to the measure $\mathrm{d}\nu_n$. According to Proposition 3.10 its zeros are located inside the unit disk.

Conversely, if the polynomial p has its zeros inside the unit disk, then $|a_n| = |q_n(0)| < 1$. On the other hand, since

$$q_{n-1}(z) = \frac{q_n(z) - a_n q_n^*(z)}{z(1-|a_n|^2)},$$

if α is a zero of q_{n-1} with $|\alpha| \geqslant 1$, then $q_n(\alpha) = a_n q_n^*(\alpha)$, and $0 < |q_n(\alpha)| < |q_n^*(\alpha)|$. This means that $|q_n(\alpha)/q_n^*(\alpha)| < 1$, but this is in contradiction with the fact that the zeros of $q_n(z)$ are inside the unit disk and thus, by the maximum modulus principle, $|q_n(z)/q_n^*(z)| \leqslant 1$ if $|z| < 1$, which is equivalent to $|q_n(z)/q_n^*(z)| \geqslant 1$ for $|z| \geqslant 1$. The same procedure applied to all $1 \leqslant k \leqslant n-2$ leads to the result. □

Remark. The above criterion is a very useful qualitative result in the stability theory for discrete linear systems [4]. In fact, given the characteristic polynomial of the matrix of a linear system, we do not need to calculate its zeros (the eigenvalues of the matrix) in order to prove that they are located inside the unit disk, and then to prove the stability of the system.

3.3. The trigonometric moment problem revisited

Next, we can state our main result.

Theorem 3.14 (Erdélyi et al. [8]). *Let $\{a_n\}_{n=1}^{\infty}$ be a sequence of complex numbers such that $|a_n| < 1$, $n = 1, 2, \dots$. Let*

$$\Phi_0(z) = 1, \quad \Phi_n(z) = z\,\Phi_{n-1}(z) + a_n \Phi_{n-1}^*(z), \quad n \geqslant 1.$$

Then, there exists a unique positive and finite Borel measure ν supported on $\mathbb{T}$ such that $\{\Phi_n\}_{n=0}^{\infty}$ is the corresponding MSOP. In other words, the positive-definite linear functional associated with the reflection coefficients $\{a_n\}_{n=0}^{\infty}$ can be represented as

$$\mathscr{L}[p(z)] = \int_0^{2\pi} p(\mathrm{e}^{\mathrm{i}\theta})\,\mathrm{d}\nu(\theta).$$

Proof. Let

$$\nu_n(\theta) = \int_0^{\theta} \mathrm{d}\nu_n(t) = \int_0^{\theta} \frac{\mathrm{d}t}{|\phi_n(\mathrm{e}^{\mathrm{i}t})|^2},$$

where ϕ_n denotes the nth orthonormal polynomial with respect to $\mathscr{L}$. The function ν_n is monotonic increasing in $[0, 2\pi]$ and according to Lemma 3.12,

$$|\nu_n(\theta)| \leqslant \int_0^{2\pi} \frac{\mathrm{d}\theta}{|\phi_n(\mathrm{e}^{\mathrm{i}\theta})|^2} \leqslant 2\pi d_0 < +\infty \quad \forall n \in \mathbb{N},\ \theta \in [0, 2\pi].$$

From Helly's selection principle (see, e.g., [5]) there exists a subsequence $\{\nu_{n_k}\}_{n_k=0}^{\infty}$ and a monotonic increasing function ν such that $\lim_{n_k \to \infty} \nu_{n_k}(\theta) = \nu(\theta)$. Furthermore, for every continuous function f on $\mathbb{T}$,

$$\lim_{n_k \to \infty} \frac{1}{2\pi} \int_0^{2\pi} f(\mathrm{e}^{\mathrm{i}\theta})\,\mathrm{d}\nu_{n_k}(\theta) = \frac{1}{2\pi} \int_0^{2\pi} f(\mathrm{e}^{\mathrm{i}\theta})\,\mathrm{d}\nu(\theta).$$

Finally,

$$\frac{1}{2\pi} \int_0^{2\pi} \phi_k(\mathrm{e}^{\mathrm{i}\theta})\overline{\phi_j(\mathrm{e}^{\mathrm{i}\theta})}\,\mathrm{d}\nu(\theta) = \lim_{n_l \to \infty} \frac{1}{2\pi} \int_0^{2\pi} \phi_k(\mathrm{e}^{\mathrm{i}\theta})\overline{\phi_j(\mathrm{e}^{\mathrm{i}\theta})}\,\mathrm{d}\nu_{n_l}(\theta) = \delta_{j,k},$$

taking $n_l > \max\{k, j\}$. □

To conclude the study of the positive definite case, we will show an analog of Theorem 2.16 of Section 2 in the following sense.

Theorem 3.15 (Alfaro and Vigil [1]). *Let $\{z_n\}_{n=1}^{\infty}$ be a sequence of complex numbers such that $|z_n| < 1$. Then, there exists a unique sequence of monic polynomials Φ_n orthogonal with respect to a positive-definite moment functional such that $\Phi_n(z_n) = 0$.*

Proof. Since $\Phi_1(z) = z + \Phi_1(0) = z - z_1$, then $\Phi_1(0) = -z_1$, and $|\Phi_1(0)| < 1$. Using induction, assume that z_{n-1} is a zero of Φ_{n-1} and $|\Phi_{n-1}(0)| < 1$. Let $\Phi_n(z) = z\Phi_{n-1}(z) + \Phi_n(0)\Phi_{n-1}^*(z)$, for $n > 1$, and z_n be a zero of Φ_n. Then, substituting z_n in the above expression, we deduce

$$z_n \Phi_{n-1}(z_n) = -\Phi_n(0)\Phi_{n-1}^*(z_n).$$

But $\Phi_{n-1}^*(z_n) \neq 0$ (otherwise z_n would be a zero of Φ_{n-1}, which is a contradiction). Thus,

$$\Phi_n(0) = -z_n \frac{\Phi_{n-1}(z_n)}{\Phi_{n-1}^*(z_n)}, \quad \text{but then } |\Phi_n(0)| = |z_n| \left| \frac{\Phi_{n-1}(z_n)}{\Phi_{n-1}^*(z_n)} \right| < |z_n| < 1,$$

since $|\Phi_{n-1}(z_n)/\Phi^*_{n-1}(z_n)| < 1$ by the maximum modulus principle (see the proof of Theorem 3.13). Then, the sequence $\{z_n\}_{n=1}^{\infty}$ defines uniquely a sequence of complex numbers $\{a_n\}_{n=1}^{\infty}$, with $a_n = \Phi_n(0)$, and this sequence, according to Theorem 3.14, uniquely defines a sequence of orthogonal polynomials $\{\Phi_n\}_{n=0}^{\infty}$ with reflection parameters a_n such that $\Phi_n(z_n) = 0$. □

In the quasi-definite case, as we already pointed out after Proposition 3.10, if ϕ_n is the nth orthonormal polynomial with respect to a quasi-definite moment functional $\mathscr{L}$, then the polynomials $z\phi_n(z)$ and $\phi_n^*(z)$ have no zeros in common. They are coprime, and by the Bézout identity [4], there exist polynomials $r(z)$ and $s(z)$ such that

$$z\, r(z)\phi_n(z) + s(z)\phi_n^*(z) = 1,$$

or, equivalently, if $u(z) = z\, r(z)$, i.e., $u(0) = 0$,

$$u(z)\phi_n(z) + s(z)\phi_n^*(z) = 1.$$

The next result is analogous to that stated in Lemma 3.12.

Theorem 3.16 (Atzmon [3]). *There exists a unique real trigonometric polynomial $f(\theta)$ of degree at most n, such that*

$$\frac{1}{2\pi}\int_0^{2\pi} \phi_n(\mathrm{e}^{\mathrm{i}\theta})\mathrm{e}^{-\mathrm{i}k\theta} f(\theta)\,\mathrm{d}\theta = 0, \quad 0 \leqslant k \leqslant n-1, \tag{3.7}$$

$$\frac{1}{2\pi}\int_0^{2\pi} |\phi_n(\mathrm{e}^{\mathrm{i}\theta})|^2 f(\theta)\,\mathrm{d}\theta = 1, \tag{3.8}$$

if and only if there exist u, $v \in \mathbb{P}_n$, with $u(0) = 0$, such that $u(z)\phi_n(z) + v(z)\phi_n^*(z) = 1$. Furthermore,

$$f(\theta) = |u(\mathrm{e}^{\mathrm{i}\theta})|^2 - |v(\mathrm{e}^{\mathrm{i}\theta})|^2.$$

Proof. If f satisfies (3.7) and (3.8), consider the function $g(\theta) = f(\theta)\phi_n(\mathrm{e}^{\mathrm{i}\theta})$, which is a trigonometric polynomial of degree at most $2n$. The conditions mean that the Fourier coefficients $\hat{g}(k)$ of $g(\theta)$ are $\hat{g}(j) = 0$, $j = 0, 1, \dots, n-1$, and $\hat{g}(n)\phi_n^*(0) = 1$. Then, there exist polynomials u, $v \in \mathbb{P}_n$, such that $u(0) = 0$, $v(0)\phi_n^*(0) = 1$ and $g(\theta) = \mathrm{e}^{\mathrm{i}n\theta}v(\mathrm{e}^{\mathrm{i}\theta}) - \overline{u(\mathrm{e}^{\mathrm{i}\theta})}$. In fact,

$$u(z) = -\sum_{j=1}^{n} \overline{\hat{g}(-j)} z^j \quad \text{and} \quad v(z) = \sum_{j=0}^{n} \hat{g}(j+n) z^j.$$

Now we introduce the trigonometric polynomial of degree at most $3n$, $h(\theta) = \phi_n(\mathrm{e}^{\mathrm{i}\theta}) f(\theta)\overline{\phi_n(\mathrm{e}^{\mathrm{i}\theta})}$. Notice that

$$h(\theta) = \phi_n^*(\mathrm{e}^{\mathrm{i}\theta})v(\mathrm{e}^{\mathrm{i}\theta}) - \overline{u(\mathrm{e}^{\mathrm{i}\theta})\phi_n(\mathrm{e}^{\mathrm{i}\theta})},$$

and h is a real-valued function. Then,

$$\phi_n^*(\mathrm{e}^{\mathrm{i}\theta})v(\mathrm{e}^{\mathrm{i}\theta}) - \overline{u(\mathrm{e}^{\mathrm{i}\theta})\phi_n(\mathrm{e}^{\mathrm{i}\theta})} = \overline{\phi_n^*(\mathrm{e}^{\mathrm{i}\theta})v(\mathrm{e}^{\mathrm{i}\theta})} - u(\mathrm{e}^{\mathrm{i}\theta})\phi_n(\mathrm{e}^{\mathrm{i}\theta}),$$

or, equivalently,

$$s(\theta) = u(\mathrm{e}^{\mathrm{i}\theta})\phi_n(\mathrm{e}^{\mathrm{i}\theta}) + v(\mathrm{e}^{\mathrm{i}\theta})\phi_n^*(\mathrm{e}^{\mathrm{i}\theta}) \in \mathbb{R}.$$

This means that the algebraic polynomial of degree at most $2n$,

$$q(z) = u(z)\phi_n(z) + v(z)\phi_n^*(z),$$

is real-valued on the unit circle, and thus $\hat{q}(j) = \overline{\hat{q}(-j)} = 0$, i.e.,

$$q(z) = q(0) = u(0)\phi_n(0) + v(0)\phi_n^*(0) = 1.$$

This yields our result.

Conversely, assume there exist polynomials u, $v \in \mathbb{P}_n$ with $u(0) = 0$, such that

$$u(z)\phi_n(z) + v(z)\phi_n^*(z) = 1. \tag{3.9}$$

Let $f(\theta) = v(e^{i\theta})\overline{v(e^{i\theta})} - u(e^{i\theta})\overline{u(e^{i\theta})}$, a trigonometric polynomial of degree at most n. We will prove that the orthogonality conditions (3.7) and (3.8) hold.

Indeed, let $g(\theta) = f(\theta)\phi_n(e^{i\theta})$. Taking into account (3.9), we have

$$\overline{u(e^{i\theta})}\,\overline{\phi_n(e^{i\theta})} + \overline{v(e^{i\theta})}e^{-in\theta}\phi_n(e^{i\theta}) = 1, \quad \text{i.e.,} \quad e^{in\theta} = \overline{u(e^{i\theta})}\phi_n^*(e^{i\theta}) + \overline{v(e^{i\theta})}\phi_n(e^{i\theta}).$$

Then, using (3.9) as well as the last expression, we obtain

$$g(\theta) = \phi_n(e^{i\theta})[v(e^{i\theta})\overline{v(e^{i\theta})} - u(e^{i\theta})\overline{u(e^{i\theta})}] = e^{in\theta}v(e^{i\theta}) - \overline{u(e^{i\theta})}, \tag{3.10}$$

which yields our orthogonality conditions

$$\hat{g}(j) = 0, \quad j = 0, 1, \ldots, n-1 \quad \text{and} \quad \hat{g}(n)\phi_n^*(0) = 1.$$

In order to prove uniqueness of f, notice that if u, $v \in \mathbb{P}_n$ satisfy (3.9) together with $u(0) = 0$, then $f(\theta) = u(e^{i\theta})\phi_n(e^{i\theta})f(\theta) + v(e^{i\theta})\phi_n^*(e^{i\theta})f(\theta)$. By (3.10), we get

$$f(\theta)\phi_n(e^{i\theta}) = e^{in\theta}\,v(e^{i\theta}) - \overline{u(e^{i\theta})},$$

and

$$f(\theta)\phi_n^*(e^{i\theta}) = \overline{v(e^{i\theta})} - e^{in\theta}u(e^{i\theta}).$$

Thus, $f(\theta) = |v(e^{i\theta})|^2 - |u(e^{i\theta})|^2$. The uniqueness of f follows from the uniqueness of u, v. □

To conclude this section, we will show with two simple examples how to find the function f explicitly.

Example 3.17. Let $\phi_3(z) = 2z^3 + 1$. Notice that because the zeros are inside the unit circle, we are in a positive-definite case. Moreover, $\phi_3^*(z) = z^3 + 2$. Using the Euclidean algorithm for $z\phi_3(z)$ and $\phi_3^*(z)$, we find

$$2z^4 + z = 2z(z^3 + 2) - 3z, \quad \text{and} \quad z^3 + 2 = -3z(-\tfrac{1}{3}z^2) + 2.$$

Thus,

$$\tfrac{1}{6}z^2(2z^4 + z) + (z^3 + 2)(\tfrac{1}{2} - \tfrac{1}{3}z^3) = 1, \quad \text{and} \quad u(z) = \tfrac{1}{6}z^3, \quad v(z) = \tfrac{1}{2} - \tfrac{1}{3}z^3.$$

Then

$$f(\theta) = |\tfrac{1}{2} - \tfrac{1}{3}e^{3i\theta}|^2 - \tfrac{1}{36} = \tfrac{1}{3}(1 - \cos 3\theta) = \tfrac{1}{6}|e^{3i\theta} - 1|^2 \geqslant 0.$$

Example 3.18. Let $\phi_3(z) = z(z^2+4)$. Notice that now there are two zeros outside the unit circle. In this case, $\phi_3^*(z) = 4z^2+1$. An analogous procedure leads to

$$z\phi_3(z) = z^4 + 4z^2 = \tfrac{1}{4}z^2(4z^2+1) + \tfrac{15}{4}z^2, \quad \phi_3^*(z) = \tfrac{16}{15}(\tfrac{15}{4}z^2) + 1.$$

Thus

$$-\tfrac{16}{15}z^2(z^2+4) + (\tfrac{4}{15}z^2+1)(4z^2+1) = 1, \quad u(z) = -\tfrac{16}{15}z^2, \quad v(z) = \tfrac{4}{15}z^2+1,$$

so

$$f(\theta) = |\tfrac{4}{15}\mathrm{e}^{2\mathrm{i}\theta} - 1|^2 - \tfrac{256}{225} = -\tfrac{1}{15}(1 + 8\cos 2\theta),$$

which gives rise to a nonpositive case, i.e., to a signed measure on $[-\pi, \pi]$.

4. The Favard Theorem for nonstandard inner products

To conclude this work, we will survey some very recent results concerning the Favard theorem for Sobolev-type orthogonal polynomials.

First of all, we want to point out that the Favard Theorem on the real line can be considered in a functional-analytic framework as follows.

Theorem 4.1 (Duran [6]). *Let $\mathbb{P}$ be the linear space of real polynomials and B an inner product on $\mathbb{P}$. Then, the following conditions are equivalent*:

(1) *The multiplication operator t, i.e., the operator $t : \mathbb{P} \to \mathbb{P}$, $p(t) \to t\,p(t)$, is Hermitian for B, that is, $B(tf, g) = B(f, tg)$ for every polynomial f, g.*

(2) *There exists a nondiscrete positive measure μ such that $B(f,g) = \int f(t)g(t)\,\mathrm{d}\mu(t)$.*

(3) *For any set of orthonormal polynomials (q_n) with respect to B the following three-term recurrence holds*:

$$tq_n(t) = a_{n+1}q_{n+1}(t) + b_nq_n(t) + a_nq_{n-1}(t), \quad n \geqslant 0, \tag{4.1}$$

with $q_{-1}(t) = 0$, $q_0(t) = 1$ and $\{a_n\}_{n=0}^{\infty}$, $\{b_n\}_{n=0}^{\infty}$ real sequences such that $a_n > 0$ for all n.

Notice that from the three-term recurrence relation (4.1) we get

$$\begin{aligned} t^2q_n(x) &= a_{n+2}a_{n+1}q_{n+2}(t) + (b_{n+1}a_{n+1} + b_na_{n+1})q_{n+1}(t) \\ &\quad + (a_{n+1}^2 + a_n^2 + b_n^2)q_n(t) + (a_nb_n + a_nb_{n-1})q_{n-1}(t) + a_na_{n-1}q_{n-2}(t), \end{aligned}$$

i.e., the sequence $\{q_n\}_{n=0}^{\infty}$ satisfies a five-term recurrence relation, which is a simple consequence of the symmetry of the operator $t^2 \equiv t \cdot t$.

Here we are interested in the converse problem, which is a natural extension of the Favard Theorem: To characterize the real symmetric bilinear forms such that the operator t^2 is a Hermitian operator. A nonstandard example of such an inner products is

$$B(f,g) = \int f(t)g(t)\,\mathrm{d}\mu(t) + Mf'(0)g'(0), \quad f,g \in \mathbb{P},$$

for which t^2 is Hermitian, i.e., $B(t^2f, g) = B(f, t^2g)$.

Theorem 4.2. *Let B be a real symmetric bilinear form on the linear space $\mathbb{P}$. Then the following conditions are equivalent*:

(1) *The operator t^2 is Hermitian for B, that is, $B(t^2 f, g) = B(f, t^2 g)$ for every polynomial f, g.*

(2) *There exist two functions μ and v such that*

$$B(f,g) = \int f(t)g(t)\,\mathrm{d}\mu(t) + 4\int f_0(t)g_0(t)\,\mathrm{d}v(t), \tag{4.2}$$

where f_0 and g_0 denote the odd components of f and g, respectively, i.e.,

$$f_0(t) = \frac{f(t)-f(-t)}{2}, \quad g_0(t) = \frac{g(t)-g(-t)}{2}.$$

Moreover, if we put $\alpha_n = \int t^n\,\mathrm{d}\mu(t)$ and $\beta_n = 4\int t^n\,\mathrm{d}v(t)$, then the matrix

$$a_{n,k} = \begin{cases} \alpha_{n+k} & \text{if } n \text{ or } k \text{ are even,} \\ \alpha_{n+k} + \beta_{n+k} & \text{otherwise,} \end{cases}$$

is positive definite if and only if B is an inner product. In this case the set of orthonormal polynomials with respect to an inner product of the form (4.2) *satisfies a five-term recurrence relation*

$$t^2 q_n(x) = A_{n+2}q_{n+2}(t) + B_{n+1}q_{n+1}(t) + C_n q_n(t) + B_n q_{n-1}(t) + A_n q_{n-2}(t), \quad n \geqslant 0, \tag{4.3}$$

where $\{A_n\}_{n=0}^{\infty}$, $\{B_n\}_{n=0}^{\infty}$, and $\{C_n\}_{n=0}^{\infty}$ are real sequences such that $A_n \neq 0$ for all n.

Also we get a generalization of the Favard Theorem.

Theorem 4.3. *Let $\{q_n\}_{n=0}^{\infty}$ be a set of polynomials satisfying the initial conditions $q_{-1}(t)=q_{-2}(t)=0$, $q_0(t)=1$ and the five-term recurrence relation* (4.3). *Then, there exist two functions μ and v such that the bilinear form* (4.2) *is an inner product and the polynomials $\{q_n\}_{n=0}^{\infty}$ are orthonormal with respect to B.*

Remark. The above theorem does not guarantee the positivity of the measures μ and v. In fact in [6] some examples of inner products of type (4.2) where both measures cannot be chosen to be positive, or μ is positive and v cannot be chosen to be positive, are shown.

All the previous results can be extended to real symmetric bilinear forms such that the operator "multiplication by $h(t)$", where h is a fixed polynomial, is Hermitian for B, i.e., $B(hf,g)=B(f,hg)$.

The basic idea consists in the choice of an adequate basis of $\mathbb{P}$ which is associated with the polynomial h. Assume that $\deg h = N$, and let $E_h = \mathrm{span}[1, h, h^2, \dots]$; then

$$\mathbb{P} = E_h \oplus t\,E_h \oplus \cdots \oplus t^{N-1}E_h.$$

If π_k denotes the projector operator in $t^k E_h$, then $\pi_k(p) = t^k q[h(t)]$. We introduce a new operator $\tilde{\pi}_k : \mathbb{P} \to \mathbb{P}$, $p \to q$, where q denotes a polynomial such that $\pi_k(p) = t^k q[h(t)]$. Then we obtain the following extension of Theorem 4.2:

Theorem 4.4. *Let B be a real symmetric bilinear form in $\mathbb{P}$. Then the following statements are equivalent*:

(1) *The operator "multiplication by h'' is Hermitian for B, i.e., $B(hf,g) = B(f,hg)$ for every polynomial f, g, where h is a polynomial of degree N.*

(2) *There exist functions $\mu_{m,m'}$ for $0 \leqslant m \leqslant m' \leqslant N-1$ such that B is defined as follows*:

$$B(f,g) = \int (\pi_0(f),\dots,\pi_{N-1}(f)) \begin{pmatrix} d\mu_{0,0} & \cdots & d\mu_{0,N-1} \\ \vdots & \ddots & \vdots \\ d\mu_{N-1,0} & \cdots & d\mu_{N-1,N-1} \end{pmatrix} \begin{pmatrix} \pi_0(g) \\ \vdots \\ \pi_{N-1}(g) \end{pmatrix}.$$

(3) *There exist functions μ_0* and *$\mu_{m,m'}$ for $1 \leqslant m \leqslant m' \leqslant N-1$ such that B is defined as follows*:

$$B(f,g) = \int f\,g\,d\mu_0 + \int (\pi_1(f),\dots,\pi_{N-1}(f)) \begin{pmatrix} d\mu_{1,1} & \cdots & d\mu_{1,N-1} \\ \vdots & \ddots & \vdots \\ d\mu_{N-1,1} & \cdots & d\mu_{N-1,N-1} \end{pmatrix} \begin{pmatrix} \pi_1(g) \\ \vdots \\ \pi_{N-1}(g) \end{pmatrix}.$$

(4) *There exist functions $\tilde{\mu}_{m,m'}$ for $0 \leqslant m \leqslant m' \leqslant N-1$ such that B is defined as follows*:

$$B(f,g) = \int (\tilde{\pi}_0(f),\dots,\tilde{\pi}_{N-1}(f)) \begin{pmatrix} d\tilde{\mu}_{0,0} & \cdots & d\tilde{\mu}_{0,N-1} \\ \vdots & \ddots & \vdots \\ d\tilde{\mu}_{N-1,0} & \cdots & d\tilde{\mu}_{N-1,N-1} \end{pmatrix} \begin{pmatrix} \tilde{\pi}_0(g) \\ \vdots \\ \tilde{\pi}_{N-1}(g) \end{pmatrix}.$$

(5) *There exist functions $\tilde{\mu}_0$ and $\tilde{\mu}_{m,m'}$ for $1 \leqslant m \leqslant m' \leqslant N-1$ such that B is defined as follows*:

$$B(f,g) = \int fg\,d\tilde{\mu}_0 + \int (\tilde{\pi}_1(f),\dots,\tilde{\pi}_{N-1}(f)) \begin{pmatrix} d\tilde{\mu}_{1,1} & \cdots & d\tilde{\mu}_{1,N-1} \\ \vdots & \ddots & \vdots \\ d\tilde{\mu}_{N-1,1} & \cdots & d\tilde{\mu}_{N-1,N-1} \end{pmatrix} \begin{pmatrix} \tilde{\pi}_1(g) \\ \vdots \\ \tilde{\pi}_{N-1}(g) \end{pmatrix}.$$

Proof. The equivalence $1 \Leftrightarrow 2 \Leftrightarrow 3$ was proved in [6]. 4 and 5 are a straightforward reformulation of the above statements 2 and 3, respectively. □

In a natural way, matrix measures appear in connection with this extension of the Favard Theorem. This fact was pointed out in [7, Section 2]. Even more, if B is an inner product of Sobolev type,

$$B(f,g) = \int f(t)\,g(t)\,d\mu(t) + \sum_{i=1}^{N} \int f^{(i)}(t) g^{(i)}(t)\,d\mu_i(t), \tag{4.4}$$

where $\{\mu_i\}_{i=1}^N$ are atomic measures, it is straightforward to prove that there exists a polynomial h of degree depending on N and mass points such that h induces a Hermitian operator with respect to B. As an immediate consequence we get a higher-order recurrence relation of type

$$h(t)q_n(t) = c_{n,0}q_n(t) + \sum_{k=1}^{M} [c_{n,k}q_{n-k}(t) + c_{n+k,k}q_{n+k}(t)], \tag{4.5}$$

where M is the degree of h and $\{q_n\}_{n=0}^\infty$ is the sequence of orthogonal polynomials relative to B.

Furthermore, extra information about the measures $\{\mu_i\}_{i=1}^N$ in (4.4) is obtained in [9] when the corresponding sequence of orthonormal polynomials satisfies a recurrence relation like (4.5).

Theorem 4.5. *Assume that there exists a polynomial h of* $\deg h \geqslant 1$ *such that $B(hf,g)=B(f,hg)$, where B is defined by* (4.4). *Then the measures $\{\mu_i\}_{i=1}^N$ are necessarily of the form*

$$\mu_i(t)=\sum_{k=1}^{j(i)} \alpha_{i,k}\delta(t-t_{i,k}),$$

for some positive integers $j(i)$, where
(1) $\alpha_{i,k}\geqslant 0$, $k=1,2,\ldots,j(i)$, $i=1,2,\ldots,N$.
(2) $R_i=\{t_{i,k}\}_{k=1}^{j(i)}\neq\emptyset$ *are the distinct real zeros of* $h^{(i)}$, $i=1,2,\ldots,N$.
(3) $\operatorname{supp}\mu_i\subset\bigcap_{k=1}^{i} R_k$, $k=1,2,\ldots,N$.
(4) *The degree of h is at least $N+1$ and there exists a unique polynomial H of minimal degree $m(H)$ satisfying $H(0)=0$ and $B(Hf,g)=B(f,Hg)$.*

The above situation corresponds to the so-called diagonal case for Sobolev-type orthogonal polynomials.

Finally, we state a more general result, which was obtained in [6].

Theorem 4.6. *Let $\mathbb{P}$ be the space of real polynomials and B a real symmetric bilinear form defined on $\mathbb{P}$. If $h(t)=(t-t_1)^{n_1}\cdots(t-t_k)^{n_k}$ and $N=\deg h$, then the following statements are equivalent*:
(1) *The operator "multiplication by h'' is Hermitian for B and $B(hf,tg)=B(tf,hg)$, i.e., the operators " multiplication by h'' and "multiplication by t'' commute with respect to B.*
(2) *There exist a function μ and constant real numbers $M_{i,j,l,l'}$ with $0\leqslant i\leqslant n_{l-1}$, $0\leqslant j\leqslant n_{l'}-1$, $1\leqslant l,l'\leqslant k$ and $M_{i,j,l,l'}=M_{j,i,l',l}$, such that*

$$B(f,g)=\int f(t)g(t)\,\mathrm{d}\mu(t)+\sum_{l,l'=1}^{k}\sum_{i=0}^{n_l-1}\sum_{j=0}^{n_{l'}-1} M_{i,j,l,l'}f^{(i)}(t_l)g^{(i)}(t_{l'}).$$

To conclude, in view of the fact that the operator "multiplication by h" is Hermitian with respect to the complex inner product

$$\langle f,g\rangle=\int_\Gamma f(z)\overline{g(z)}\,\mathrm{d}\mu(z), \tag{4.6}$$

where Γ is a harmonic algebraic curve defined by $\Im h(z)=0$ and h a complex polynomial (see [15]), it seems natural to ask:

Problem 1. *To characterize the sesquilinear forms $B:\mathbb{P}\times\mathbb{P}\to\mathbb{C}$ such that the operator "multiplication by h" satisfies $B(hf,g)=B(f,hg)$ for every polynomial f, $g\in\mathbb{P}$, the linear space of polynomials with complex coefficients.*

In the same way (see [14]), given an inner product like (4.6), if Γ is an equipotential curve $|h(z)|=1$, where h is a complex polynomial, then the operator "multiplication by h" is isometric with respect to (4.6). Thus, it is natural to formulate

Problem 2. *To characterize the sesquilinear forms $B:\mathbb{P}\times\mathbb{P}\to\mathbb{C}$ such that the operator "multiplication by h" satisfies $B(hf,hg)=B(f,g)$ for every polynomial f, $g\in\mathbb{P}$, the linear space of polynomials with complex coefficients.*

The connection between these problems and matrix polynomials orthogonal with respect to matrix measures supported on the real line and on the unit circle, respectively, has been shown in [15,14].

Acknowledgements

We thank Walter Gautschi for his remarks which helped us to improve the paper and also an unknown referee for the historical remarks and suggestions, especially for pointing out the work of Natanson [17].

References

[1] M.P. Alfaro, L. Vigil, Solution of a problem of P. Turán on zeros of orthogonal polynomials on the unit circle, J. Approx. Theory 53 (1988) 195–197.

[2] M. Álvarez, G. Sansigre, On polynomials with interlacing zeros, in: C. Brezinski, et al. (Eds.), Polynômes Orthogonaux et Applications. Proceedings, Bar-le-Duc 1984, Springer, Berlin, 1985, pp. 255–258.

[3] A. Atzmon, *n*-Orthonormal operator polynomials, in: I. Gohberg (Ed.), Orthogonal Matrix-valued Polynomials and Applications, Birkhäuser, Basel, 1998, pp. 47–63.

[4] S. Barnett, Polynomials and Linear Control Systems, Marcel Dekker, New York, 1983.

[5] T.S. Chihara, An Introduction to Orthogonal Polynomials, Gordon and Breach, New York, 1978.

[6] A.J. Duran, A generalization of Favard's Theorem for polynomials satisfying a recurrence relation, J. Approx. Theory 74 (1993) 83–109.

[7] A.J. Duran, W. Van Assche, Orthogonal matrix polynomials and higher-order recurrence relations, Linear Algebra Appl. 219 (1995) 261–280.

[8] T. Erdélyi, P. Nevai, J. Zhang, J.S. Geronimo, A simple proof of "Favard's theorem" on the unit circle, Atti. Sem. Mat. Fis. Univ. Modena 39 (1991) 551–556.

[9] W.D. Evans, L.L. Littlejohn, F. Marcellán, C. Markett, A. Ronveaux, On recurrence relations for Sobolev orthogonal polynomials, SIAM J. Math. Anal. 26 (1995) 446–467.

[10] J. Favard, Sur les polynômes de Tchebicheff, C.R. Acad. Sci. Paris 200 (1935) 2052–2053.

[11] Ya. L. Geronimus, Polynomials Orthogonal on a Circle and Their Applications, Amer. Math. Soc. Translations, 1954.

[12] H.J. Landau, Maximum entropy and the moment problem, Bull. Amer. Math. Soc. (N.S.) 16 (1987) 47–77.

[13] F. Marcellán, Orthogonal polynomials and Toeplitz matrices: some applications, In: M. Alfaro (Eds.), Rational Approximation and Orthogonal Polynomials, Sem. Mat. García de Galdeano, Univ. Zaragoza, 1989, pp. 31–57.

[14] F. Marcellán, I. Rodríguez, A class of matrix orthogonal polynomials on the unit circle, Linear Algebra Appl. 121 (1989) 233–241.

[15] F. Marcellán, G. Sansigre, On a class of matrix orthogonal polynomials on the real line, Linear Algebra Appl. 181 (1993) 97–109.

[16] J. Montaner, M. Alfaro, On zeros of polynomials orthogonal with respect to a quasi-definite inner product on the unit circle, Rend. Circ. Mat. Palermo (2) 44 (1995) 301–314.

[17] I.P. Natanson, Constructive Function Theory, Vol. II, Approximation in Mean, Frederick Ungar, New York, 1965.

[18] P. Nevai, V. Totik, Orthogonal polynomials and their zeros, Acta Sci. Math. (Szeged) 53 (1989) 99–104.

[19] O. Perron, Die Lehre von den Kettenbrüchen, 2nd Edition, Teubner, Leipzig, 1929.

[20] J. Sherman, On the numerators of the Stieltjes continued fractions, Trans. Amer. Math. Soc. 35 (1933) 64–87.

[21] J. Shohat, The relation of the classical orthogonal polynomials to the polynomials of Appell, Amer. J. Math. 58 (1936) 453–464.

[22] G.W. Stewart, J.G. Sun, Matrix Perturbation Theory, Academic Press, Boston, MA, 1990.

[23] T.J. Stieltjes, Recherches sur les fractions continues, Ann. Fac. Sci Toulouse 8 (1894) J1-122, 9 (1895) A1-10.

[24] M.H. Stone, Linear Transformations in Hilbert Spaces, Colloquium Publications, Vol. 15, Amer. Math. Soc., Providence, RI, 1932.

[25] G. Szegő, Orthogonal Polynomials 4th Edition, Colloquium Publications, Vol. 23, Amer. Math. Soc., Providence, RI, 1975.

[26] J. Vinuesa, R. Guadalupe, Zéros extrémaux de polynômes orthogonaux, in: C. Brezinski, et al. (Eds.), Polynômes Orthogonaux et Applications, Proceedings, Bar-le-Duc 1984, Springer, Berlin, 1985, pp. 291–295.

[27] B. Wendroff, On orthogonal polynomials, Proc. Amer. Math. Soc. 12 (1961) 554–555.

[28] A. Wintner, Spektraltheorie der unendlichen Matrizen, Einführung in den analytischen Apparat der Quantenmechanik, Hitzel, Leipzig, 1929.

Journal of Computational and Applied Mathematics 127 (2001) 255–266

JOURNAL OF
COMPUTATIONAL AND
APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Analytic aspects of Sobolev orthogonal polynomials revisited ☆

A. Martínez-Finkelshtein

*University of Almería and Instituto Carlos I de Física Teórica y Computacional, Granada University,
04120 Almeria, Spain*

Received 10 June 1999; received in revised form 14 October 1999

Abstract

This paper surveys some recent achievements in the analytic theory of polynomials orthogonal with respect to inner products involving derivatives. Asymptotic properties, zero location, approximation and moment theory are some of the topic considered. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Sobolev orthogonal polynomials; Asymptotics; Zeros; Coherent pairs; Moments; Approximation

1. Introduction

During the VIII Symposium on Orthogonal Polynomials and Applications, held in September 1997 in Sevilla (Spain), a survey on some new results and tools in the study of the analytic properties of Sobolev orthogonal polynomials was presented, which was published later in [22]. This is a modest attempt to update that survey, including some topics where progress has been made in the two intervening years.

In general terms, we refer to Sobolev orthogonal polynomials when the underlying inner product involves derivatives (in the classical or distributional sense). Here we restrict ourselves to the following setups, which are general enough to exhibit the main features of the subject (for a more general definition, see e.g. [4]):

- *Diagonal case*: let $\{\mu_k\}_{k=0}^{m}$, with $m \in \mathbb{Z}_+$, be a set of $m+1$ finite positive Borel measures such that at least one of the measures, say μ_j, has infinitely many points of increase, in which case μ_k

☆ This work was partially supported by a research grant of Dirección General de Enseñanza Superior (DGES) of Spain, project code PB95–1205 and by Junta de Andalucía, Grupo de Investigación FQM 0229.

PII: S0377-0427(00)00499-4

has at least $k+1$ points of increase, when $0 \leqslant k < j$. If $f^{(k)}$ is the kth derivative of the function f, then we denote

$$(f, g)_S = \sum_{k=0}^{m} \int f^{(k)} \overline{g^{(k)}} \, \mathrm{d}\mu_k. \tag{1}$$

- *Nondiagonal case*: for a function f put $\boldsymbol{f} = (f, f', \dots, f^{(m)})$. Given a finite positive Borel measure μ on $\mathbb{C}$ with infinitely many points of increase and an $(m+1) \times (m+1)$ Hermitian positive-definite matrix A, set

$$(f, g)_S = \int \boldsymbol{f} A \overline{(\boldsymbol{g}^{\mathrm{T}})} \mu_k. \tag{2}$$

Then either (1) or (2) defines an inner product in the linear space $\mathbb{P}$ of polynomials with complex coefficients. The Gram–Schmidt process applied to the canonical basis of $\mathbb{P}$ generates the orthonormal sequence of polynomials $\{q_n\}$, $n=0,1,\dots, \deg q_n = n$; we denote the corresponding monic polynomials by $Q_n(x) = x^n +$ lower degree terms, so that

$$q_n(x) = \frac{Q_n(x)}{||Q_n||_S}, \quad n = 0, 1, \dots, \tag{3}$$

where $||f||_S = \sqrt{(f, f)_S}$. As usual, we will call these polynomials *Sobolev orthogonal polynomials.*

In addition, we will use the following notation. If μ is a measure, then supp(μ) is its support and $P_n(\cdot; \mu)$ is the corresponding nth monic polynomial (if it exists) orthogonal with respect to the inner product

$$\langle f, g \rangle_\mu = \int_{\mathrm{supp}(\mu)} f(z) \overline{g(z)} \, \mathrm{d}\mu(z).$$

We have

$$P_n(z; \mu) = z^n + \sum_{k=0}^{n-1} c_{n,k} z^k, \quad \langle P_n(z; \mu), z^k \rangle_\mu = 0, \ \ k = 0, \dots, n-1.$$

If Γ is a compact set in the complex plane, we denote by $C(\Gamma)$ its logarithmic capacity and by Ω the unbounded component of $\bar{\mathbb{C}} \setminus \Gamma$; for simplicity we assume in what follows that Ω is regular with respect to the Dirichlet problem. Also, φ is the conformal mapping of Ω onto the exterior of a disc $|z| = r$, normalized by

$$\lim_{z \to \infty} \varphi(z)/z = 1,$$

so that the radius $r = C(\Gamma)$. Finally, ω_Γ stands for the equilibrium measure of the compact set Γ. For details, see e.g. [35] or [36].

During the 1990s very active research on Sobolev orthogonal polynomials was in progress. Nevertheless, most of the results are connected with the algebraic aspects of the theory and classical measures in the inner product. For a historical review of this period the reader is referred to [22] (see also [26]). Here we are mainly interested in the analytic theory: asymptotics, Fourier series, approximation properties, etc.

2. Strong and comparative asymptotics

To study the asymptotic properties of the sequence $\{Q_n\}$ as $n \to \infty$, a natural approach is to compare it with the corresponding behavior of the sequence $\{P_n(\cdot,\mu)\}$, where μ is one of the measures involved in the Sobolev inner product (1) or (2). If μ is "good", the asymptotic properties of the standard orthogonal polynomials are well known, so we can arrive at conclusions about Q_n.

Probably, the simplest case of a nontrivial Sobolev inner product is

$$(f,g)_S = \int f(x)g(x)\,\mathrm{d}\mu_0(x) + \lambda f'(\xi)g'(\xi).$$

It corresponds to (1) with $m=1$ and $\mu_1 = \lambda\delta_\xi$, where as usual δ_ξ is the Dirac delta (unit mass) at ξ. The case when the measures corresponding to derivatives are finite collections of Dirac deltas is known as *discrete*, although it is implicitly assumed that μ_0 has a nonzero absolutely continuous component. The asymptotic properties of $\{Q_n\}$ in such a situation have been thoroughly studied in [16]. As it was shown there, a discrete measure μ_1 cannot "outweigh" the absolutely continuous μ_0, and the asymptotic behavior of the polynomials Q_n is identical to the standard orthogonal polynomials corresponding to a mass-point modification of μ_0.

For example, assume that μ_0 is supported on an interval $[a,b] \subset \mathbb{R}$ and belongs to the Nevai class $M(a,b)$ (see [29]), and that $\lambda > 0$ and $\xi \in \mathbb{R}$. Then,

$$\lim_{n\to\infty} \frac{Q_n(z)}{P_n(z;\mu_0+\mu_1)} = 1$$

holds uniformly on compact subsets of $\bar{\mathbb{C}} \setminus ([a,b] \cup \{\xi\})$ (see [16]).

We also may try to construct an analogue of the classical theory when we have derivatives in the inner product, considering Szegő or Nevai classes of measures or weights.

Assume that all the measures μ_k in the inner product (1) are supported on the same Jordan curve or arc $\Gamma \subset \mathbb{C}$. If we recall that one of the motivations for introducing Sobolev orthogonal polynomials is a least-squares fitting of differentiable functions, this seems to be the most natural situation in practice. On the support Γ we impose a restriction: the natural (arclength) parametrization of Γ belongs to the class C^{2+}, which is the subclass of functions of C^2 whose second derivatives satisfy a Lipschitz condition.

In order to extend Szegő's theory to Sobolev orthogonality, we assume first that *all* the measures μ_k belong to the Szegő class on Γ.

The experience accumulated so far shows that the right approach to the asymptotics of $\{Q_n\}$ consists in a "decoupling" of terms in (1). Observe that after taking derivatives the polynomials involved are no longer monic. The factor $O(n^k)$ which multiples $Q_n^{(k)}$ plays a crucial role as $n \to \infty$. Decoupling here means that we can restrict our attention to the last term of (1) and show that only $\{Q_n^{(m)}\}$ "matters". This can be done by comparing the Sobolev norm $||Q_n^{(m)}||_S$ with the standard $L^2(\mu_m)$ norm of $Q_n^{(m)}$ and using the extremality of the $L^2(\mu_m)$ norm for $P_{n-m}(.;\mu_m)$. This allows us to find the asymptotics of $\{Q_n^{(m)}\}$ described by the Szegő theory. The second step is to "recover" the behavior of the sequences $\{Q_n^{(k)}\}$ for $k=0,\dots,m-1$.

By means of this scheme, Bernstein–Szegő type theorems were established in [23] (case $m=1$) and in [25] (for $m>1$). A combination of some of these results can be stated in the form of comparative asymptotics:

Theorem 1. *With the above-mentioned assumptions on the measures μ_k, $k=0,\dots,m$, we have*

$$\lim_{n\to\infty} \frac{Q_n^{(k)}(z)}{n^k P_{n-k}(z;\mu_m)} = \frac{1}{[\varphi'(z)]^{m-k}} \quad for\ k=0,1,\dots,m, \tag{4}$$

uniformly on compact subsets of Ω.

By considering a slightly different extremal problem, we can extend this result to the case when μ_m has an absolutely continuous part from the Szegő class on Γ plus a finite number of mass points in $\mathbb{C}$. The condition that the measures μ_k, $k=0,\dots,m-1$, belong to the Szegő class, is introduced because of some technicalities in the proof: at some stage we must derive strong asymptotics from the L^2 one. Indeed, this condition is clearly not necessary for (4). An easy example is produced in [23]. Later, in [3], a case of Sobolev orthogonality (1) on the unit circle with $m=1$ was studied, where, assuming more restrictive conditions on μ_1, the same asymptotics was established for a wider class of measures μ_0.

Thus, a necessary condition for (4) or, on the contrary, any nontrivial examples when this asymptotics does not hold, is still an open problem.

The case of noncoincident supports of the measures μ_k is very interesting and, for the time being, practically unexplored. An insight into the difficulties inherent in this situation is given in the paper [14] on the weak asymptotics of Q_n. Any advance in the study of the strong asymptotics of Q_n when integrals in (1) are taken in different supports is of great interest.

3. Recent extensions or the importance to be coherent

One approach to the study of asymptotics has not yet been mentioned, namely the coherence of measures μ_k. Although its scope is limited, it has played an important role during the last few years.

Historically, the coherence of measures was introduced in connection with Sobolev orthogonality and was essential in establishing first results on asymptotics in the nondiscrete case. Briefly, we say that two measures, μ and v, form a *coherent pair* if there exists a fixed constant $k\in\mathbb{N}_0$ such that for each $n\in\mathbb{N}$ the monic orthogonal polynomial $P_n(.;v)$ can be expressed as a linear combination of the set

$$P'_{n+1}(.;\mu),\dots P'_{n-k}(.;\mu).$$

Coherence is then classified in terms of k.

Coherent pairs of measures on $\mathbb{R}$ have been known for several years, but the complete classification (for $k=0$) was given only in [27]. This work was a basis for [24], where the first more or less general result on nondiscrete asymptotics was obtained by means of a very simple but successful technique: establishing an algebraic relation between the sequence $\{Q_n\}$ and the corresponding sequence $\{P_n(.;\mu)\}$, having a fixed number of terms, and studying the asymptotic behavior of the parameters involved. This leads easily to the comparative outer asymptotics.

As recent results show, this idea can be exploited in a variety of contexts. For instance, the case of a coherent pair with $m=1$ and unbounded support of the measures was studied in [28] (where a result of [20] was extended). According to [27], in this case, either one of the measures μ_0 or μ_1 in (1) is given by the classical Laguerre weight. Following the path described above one can show that for a suitable parameter α of the Laguerre polynomials $L_n^{(\alpha)}$, the ratio $Q_n(x)/L_n^{(\alpha)}(x)$ tends to a constant for $x \in \mathbb{C} \setminus [0,+\infty)$, which in turn shows that the zeros of Q_n accumulate on $[0,+\infty)$.

The nondiagonal Sobolev inner product can also be dealt with if we have coherence. In [21], the authors take (2) with a 2×2 matrix A and the measure $\mathrm{d}\mu(x) = x^\alpha \mathrm{e}^{-x}\,\mathrm{d}x$ on $[0,+\infty)$. Observe that this case can be reduced to the diagonal one (1) but with a sign-varying weight in the first integral. Thus, the results in [21] are not immediate consequences of those in [20]. Exploiting the algebraic relation between the two families (classical and Sobolev), once again one can derive the outer comparative asymptotics in $\mathbb{C} \setminus [0,+\infty)$, which gives in the limit a constant. Perhaps more informative is the scaled asymptotics $Q_n(nx)/L_n^{(\alpha-1)}(nx)$ holding uniformly in $\mathbb{C} \setminus [0,4]$, which leads directly to an analogue of the Plancherel–Rotach asymptotics for Q_n. Finally, we can use the well-known Mehler–Heine asymptotic formula for $L_n^{(\alpha-1)}(nx)$ in order to relate it (and its zeros) to the Bessel functions. Analogous research for μ given by the Hermite weight on $\mathbb{R}$ was carried out in [2].

More complicated coherent pairs have been studied in [19], which yield similar asymptotic results. At any rate, after the work [27] it became apparent that coherence cannot lead us very far from the classical weights on $\mathbb{R}$. Some attempts to extend this notion to supports in the complex plane (say, on the unit circle) have not given any important results so far.

From the discussion above, it becomes clear that a discretization of the measures μ_k in (1) for $k \geqslant 1$ changes the asymptotic behavior of the corresponding Sobolev polynomials. An alternative approach could be to get rid of derivatives in the inner product by replacing them with a suitable finite-difference scheme.

One of the first problems in this direction was considered in [9] for

$$(f,g)_S = \int_{\mathbb{R}} fg\,\mathrm{d}\mu_0 + \lambda \Delta f(c) \Delta g(c), \quad \lambda > 0,$$

where $\Delta f(x) = f(x+1) - f(x)$ and the support of μ_0 is disjoint with $(c, c+1)$. The paper is devoted mainly to algebraic properties and zero location of polynomials Q_n orthogonal with respect to this inner product.

Thus, we have the following problem: given two measures on $\mathbb{R}$, μ_0 and μ_1, describe the properties of the sequence of polynomials Q_n orthogonal with respect to the inner product

$$(f,g)_S = \langle f, g \rangle_{\mu_0} + \langle \Delta f, \Delta g \rangle_{\mu_1}. \tag{5}$$

As far as I know, only the cases of discrete measures μ_1 have so far been studied, both from algebraic and analytic points of view. For instance, in [15] the construction of the corresponding Sobolev space is discussed.

We can generalize in some sense the classical families of orthogonal polynomials of a discrete variable if in (5) we take both measures μ_k discrete. In a series of papers [5,7], the so-called Meixner–Sobolev polynomials are studied, which are orthogonal with respect to (5) and

$$\mu_0 = \sum_{i=0}^{\infty} \binom{\gamma + i - 1}{i} t^i \delta_i, \quad 0 < t < 1, \quad \gamma > 0, \quad \mu_1 = \lambda \mu_0, \quad \lambda > 0.$$

The Sobolev inner product $(\cdot,\cdot)_S$ obtained in this way fits in both schemes (1) and (2).

The asymptotic properties of the corresponding Sobolev orthogonal polynomials Q_n are studied in [7]. Once again we observe the use of coherence in the proof: the authors show that Q_n, Q_{n+1}, $P_n(\cdot;\mu_0)$

and $P_{n+1}(\cdot;\mu_0)$ are linearly dependent and find expressions for the (nonzero) coefficients of a vanishing linear combination. This yields bounds or recurrence relations for the Sobolev norms of Q_n, which in turn allows us to establish comparative asymptotics of the coefficients above. As usual, an analogue of Poincaré's theorem does the rest.

As the support of the Meixner discrete measure is unbounded, it is more interesting to study a contracted asymptotics obtained by a scaling of the variable. The authors of [7] find the behavior of the ratio $Q_n(nz)/P_n(nz;\mu_0)$ for $z \notin [0,(1+\sqrt{t})^2/(1-t)]$ and show that the zeros accumulate on the complement of this interval.

One step further in the direction of discretizing of derivatives in (1) is to consider a nonuniform mesh on $\mathbb{R}$. In particular, we could take discretization knots of the form q^k and substitute the differential operator D in (1) by the q-difference operator D_q:

$$D_q f(x) = \frac{f(qx)-f(x)}{(q-1)x}, \quad x \neq 0, \;\; q \neq 1, \;\; D_q f(0) = f'(0).$$

In [6], the little q-Laguerre measure was considered,

$$\mu_0 = \sum_{k \geqslant 0} \frac{(aq)^k (aq;q)_\infty}{(q;q)_k} \delta_{q^k}, \quad 0 < aq, \;\; q < 1,$$

where as usual,

$$(b;q)_0 = 1, \quad (b;q)_k = \prod_{j=1}^{k} (1 - bq^{j-1}), \quad 0 < k \leqslant \infty.$$

Then (1) becomes

$$(f,g)_S = \langle f,g \rangle_{\mu_0} + \langle D_q f, D_q g \rangle_{\mu_1};$$

for $\mu_1 = \lambda\mu_0$, $\lambda > 0$, the corresponding Q_n are called *little q-Laguerre–Sobolev* polynomials.

Once again, the "coherent" scheme works perfectly. In particular, the properties of Laguerre–Sobolev polynomials [20] can be recovered by taking appropriate limits in the little q-Laguerre–Sobolev family when $q \uparrow 1$.

4. Balanced Sobolev orthogonal polynomials

The role of the derivatives in the last term in (1), which introduce large factors as $n \to \infty$, was discussed above along with the idea of "decoupling" the study of each derivative $Q_n^{(k)}$. These considerations motivate the idea of balancing the terms of the Sobolev inner product by considering only monic polynomials. In other words, we can look for monic polynomials Q_n of degree n, which minimize the norm

$$||Q_n||^2 = \langle Q_n, Q_n \rangle_{\mu_0} + \langle Q_n'/n, Q_n'/n \rangle_{\mu_1}$$

in the class of all monic polynomials of degree n. In a more general setting, we can study orthogonality with respect to (1), where each term is multiplied by a parameter which depends on the degree of the polynomial.

In [1], we made use once again of coherence of measures supported on $[-1,1]$ in order to explore the following situation: let $\{\lambda_n\}$ be a decreasing sequence of real positive numbers such that

$$\lim_{n\to\infty} n^2\lambda_n = L, \quad 0 < L < +\infty, \tag{6}$$

and let Q_n now stand for the monic polynomial of degree n orthogonal to all polynomials of degree $< n$ with respect to the inner product

$$(p,q)_{S,n} = \langle p,q\rangle_{\mu_0} + \lambda_n\langle p',q'\rangle_{\mu_1}. \tag{7}$$

Observe that the inner product varies with the degree n; thus we could speak also of *varying Sobolev orthogonality*.

If we introduce, in addition, the measure μ^* on $[-1,1]$,

$$d\mu^*(x) = \{\mu_0'(x) + 4L|\varphi'(x)|^2\mu_1'(x)\}\,\mathrm{d}x, \quad x \in [-1,1], \tag{8}$$

we have the following theorem.

Theorem 2 (Alfaro et al. [1]). *Let* (μ_0,μ_1) *be a coherent pair of measures satisfying Szegő's condition on* $[-1,1]$, *and let the sequence* $\{\lambda_n\}$ *be as in* (6). *Then*,

$$\lim_{n\to\infty} \frac{Q_n(z)}{P_n(z;\mu^*)} = 1, \tag{9}$$

locally uniformly in $\mathbb{C}\setminus[-1,1]$.

In other words, the sequence $\{Q_n\}$ asymptotically behaves like the monic orthogonal polynomial sequence corresponding to the measure (8).

The study of polynomials orthogonal with respect to a varying inner product (7) under assumption (6) should be extended to a wider class of measures μ_k. Coherence still can give something new for unbounded support, but the general case of bounded $\operatorname{supp}(\mu_k)$ probably must be attacked with the help of the Szegő type theory as described above.

5. Moments and approximation properties of Sobolev polynomials

It is clear that the moment theory plays an essential role in the study of the properties of standard orthogonal polynomials. At the same time, first works in this direction for the Sobolev orthogonality are very recent. In [8] (see also [30]) the diagonal case (1) is considered, when all the measures μ_k are supported on $\mathbb{R}$. As usual, the moment problem associated with (1) looks for the inversion of the mapping

$$\boldsymbol{\mu} = (\mu_0,\ldots,\mu_m) \to \mathscr{M} = (s_{i,j})_{i,j=0}^{\infty}, \quad s_{i,j} = (x^i,x^j)_S. \tag{10}$$

To be more precise, the *Sobolev moment problem* (or the *S-moment problem*) is the following: given an infinite matrix $\mathscr{M} = (s_{i,j})_{i,j=0}^{\infty}$ and $m+1$ subsets of $\mathbb{R}$, Γ_k, $k = 0,\ldots,m$, find a set of measures $\mu_0,\ldots,\mu_m$ ($\mu_m \neq 0$) such that

$$\operatorname{supp}\mu_k \subset \Gamma_k, \quad k = 0,\ldots,m \quad \text{and} \quad s_{i,j} = (x^i,x^j)_S \quad \text{for } i,j = 0,\ldots\,.$$

As usual, the problem is considered "definite" if it has a solution, and "determinate" if this solution is unique.

In this setting we can observe once again the phenomenon of "decoupling" mentioned above. Indeed, using (1), we can see that if $\mathscr{M}$ is given by the right-hand side of (10), then

$$\mathscr{M} = \sum_{k=0}^{m} \mathscr{D}_k \mathscr{M}_k \mathscr{D}_k^{\mathrm{T}}, \quad \mathscr{M}_m \neq 0, \tag{11}$$

where

$$\mathscr{D}_k = (d_{i,j}^k)_{i,j=0}^{\infty}, \quad d_{i,j}^k = \frac{i!}{(i-k)!}\delta_{k,i-j}, \;\; k = 0,\ldots,m,$$

and $\mathscr{M}_k$ are infinite Hankel matrices. Thus, questions about S-moment problem can be tackled by means of the classical tools of moment theory. In fact, in [8] it was shown that if $\mathscr{M}$ has a decomposition (11) then definiteness (determinacy) of the S-moment problem is equivalent of that for the classical moment problem for each measure μ_k, $k = 0,\ldots,m$. A characterization of all the matrices $\mathscr{M}$ which for a fixed m admit (11) can be found in [30, Theorem 2.1], where they are called *Hankel–Sobolev matrices.*

As moments and recurrence go hand in hand in the theory of orthogonal polynomials, it is natural to explore this path here. It is immediate to see that the Sobolev inner products (1) and (2) lack of an essential feature of the standard inner product, namely

$$(xp(x), q(x))_S \neq (p(x), xq(x))_S.$$

As a consequence, we cannot expect a three-term recurrence relation for Q_n (neither any recurrence relation with a fixed number of terms, except in the case that the measures corresponding to derivatives are discrete, see [10]). Nevertheless, expanding polynomials $xq_n(x)$ in the basis $q_0,\ldots,q_n$ we obtain the Hessenberg matrix of coefficients,

$$\mathscr{R} = (r_{i,j})_{i,j=0}^{\infty}, \quad r_{i,j} = (xq_i(x), q_j(x))_S. \tag{12}$$

As in the standard case, the zeros of the Sobolev orthogonal polynomial q_n are the eigenvalues of the $n \times n$ principal minor of $\mathscr{R}$. This shows that the zeros are connected with the operator of multiplication by the variable (or shift operator).

Now we can try to use the tools of operator theory. But here special care is needed in defining the appropriate function space. For the time being we can work it out formally: since (1) or (2) define an inner product in the space $\mathbb{P}$, we can take its completion identifying all the Cauchy sequences of polynomials whose difference tends to zero in the norm $||\cdot||_S$. Let us denote the resulting Hilbert space by $\mathbb{P}_S = \mathbb{P}_S(\boldsymbol{\mu})$, $\boldsymbol{\mu} = (\mu_0,\ldots,\mu_m)$.

Considerations above show that by means of the matrix (12) we can define in $\mathbb{P}$ a linear operator R such that

$$Rp(x) = xp(x). \tag{13}$$

By continuity, it can be extended to the multiplication operator in $\mathbb{P}_S$.

Recall that the location of zeros of Sobolev orthogonal polynomials is not a trivial problem. Simple examples show that they do not necessarily remain in the convex hull of the union of the supports of the measures μ_k and can be complex even when all the μ_k are supported on $\mathbb{R}$. Some accurate numerical results in this regard can be found in [14]. In particular, the following question is open: is it true that whenever the measures $\mu_0,\ldots,\mu_m$ are compactly supported in $\mathbb{C}$, the zeros of Q_n are uniformly bounded?

The first benefit from the interpretation of the recurrence for q_n in terms of operator theory was obtained in [17; 30, Theorem 3.6] and this result was improved in [18]: if R is bounded and $||R||$ is its operator norm, then all the zeros of the Sobolev orthogonal polynomials Q_n are contained in the disc

$$\{z \in \mathbb{C}\colon |z| \leqslant ||R||\}.$$

Indeed, if x_0 is a zero of Q_n then $xp(x) = x_0 p(x) + Q_n(x)$ for a $p \in \mathbb{P}_{n-1}$. Since p and Q_n are orthogonal,

$$|x_0|^2 ||p||_S^2 = ||xp(x)||_S^2 - ||Q_n(x)||_S^2 \leqslant ||xp(x)||_S^2 = ||Rp(x)||_S^2 \leqslant ||R||^2 ||p||_S^2$$

which yields the result.

Thus, the question whether or not the multiplication operator R is bounded turns out to be a key for the location of zeros and, as it was shown in [14], to asymptotic results for the nth root of Q_n. Clearly enough, without a thorough knowledge of the space $\mathbb{P}_S$ this condition (R is bounded) lacks of any practical application.

But we can have a simple and verifiable sufficient condition for $||R|| < \infty$, introduced also in [17]: the *sequential domination* of the Sobolev inner product (1). It means that for $k = 1, \dots, m$,

$$\operatorname{supp} \mu_k \subset \operatorname{supp} \mu_{k-1}, \quad \mu_k \ll \mu_{k-1}, \quad \frac{\mathrm{d}\mu_k}{\mathrm{d}\mu_{k-1}} \in L^\infty(\mu_{k-1}), \tag{14}$$

where $\mu \ll v$ means that μ is absolutely continuous with respect to v and $\mathrm{d}\mu/\mathrm{d}v$ stands for the Radon–Nikodym derivative. A bound for $||R||$ in terms of $\max_{x \in \operatorname{supp}(\mu_0)} |x|$ and the sup-norm of the derivatives above can be obtained (see [18]).

At a second look, the assumption of sequential domination seems more natural. Indeed, we have seen above that in part owing to derivatives in the integrals defining (1), the last measure, μ_m, plays the leading role and determines the behavior of $Q_n^{(m)}$, while the other measures are bound to “control” the proper asymptotic behavior of $Q_n^{(k)}$, for $k = 0, \dots, m-1$. This can be achieved by assigning more weight to measures with smaller index, like in (14).

It turns out that the condition of sequential domination is in some sense not far from being also necessary. This comes as a result of a series of works [31–34], aiming in particular at a full understanding of the structure and properties of the space $\mathbb{P}_S$ defined above. We are talking here about a general theory of Sobolev spaces.

Weighted Sobolev spaces are studied from several points of view, motivated mainly by the analysis of differential equations. Their extension to general measures is less explored. Some examples have been considered in [11–13], but the beginnings of a systematic study can be found in the papers mentioned above. In particular, two key questions are discussed in [32,33]: what is a reasonable extension of the definition of a Sobolev space of functions with respect to a vectorial measure $\boldsymbol{\mu} = (\mu_0, \dots, \mu_m)$? For example, we could define it as the largest space where the Sobolev norm $|| \cdot ||_S$ has sense and is finite. Then, the second question arises: what is the relation of this space to $\mathbb{P}_S(\boldsymbol{\mu})$? In other words, we should study the possibility of approximation of a class of functions by polynomials in the norm $|| \cdot ||_S$.

A good description of $\mathbb{P}_S(\boldsymbol{\mu})$ led in [31,34] to a proof of both necessary and sufficient conditions for the multiplication operator R to be bounded in $\mathbb{P}_S$. Rather remarkable is the result that the sufficient condition of sequential domination is not far from being necessary. Roughly speaking, Theorem 4.1

in [31] says that if all the measures μ_k are compactly supported on $\mathbb{R}$ and R is bounded, then there exists a vectorial measure $\boldsymbol{v}$, also compactly supported on $\mathbb{R}$, whose components are sequentially dominated, and such that the Sobolev norms induced by $\boldsymbol{\mu}$ and $\boldsymbol{v}$ are equivalent.

Let us go back to asymptotics. If we have located the zeros of $\{Q_n\}$, which turn out to be uniformly bounded, we can apply the ideas of [14] to obtain the zero distribution along with the nth root asymptotics of the Sobolev polynomials. This is the second part of [17,18].

For any polynomial Q of exact degree n we denote by

$$v(Q) = \frac{1}{n} \sum_{Q(z)=0} \delta_z,$$

the normalized zero counting measure associated with Q. The (weak) zero distribution of the polynomials Q_n studies the convergence of the sequence $v(Q_n)$ in the weak-* topology. The class of regular measures $\mu \in \mathbf{Reg}$, compactly supported on $\mathbb{R}$, has been introduced in [36] and is characterized by the fact that

$$\lim_{n\to\infty} ||P_n(\cdot;\mu)||_{L^2(\mu)}^{1/n} = C(\operatorname{supp}(\mu)).$$

Consider again the Sobolev inner product (1). Assume that there exists an $l \in \{0,\ldots,m\}$ such that

$$\bigcup_{k=0}^{l} \operatorname{supp}(\mu_k) = \bigcup_{k=0}^{m} \operatorname{supp}(\mu_k)$$

and $\mu_0,\ldots,\mu_l \in \mathbf{Reg}$, with their supports regular with respect to the Dirichlet problem. Following [17], we call this inner product *l-regular*. For example, if it is sequentially dominated and the support of μ_0 is regular with respect to the Dirichlet problem, then the condition $\mu_0 \in \mathbf{Reg}$ is equivalent to 0-regularity.

If (1) is l-regular, then the derivatives $Q_n^{(k)}$ for $l \leqslant k \leqslant m$ exhibit regular behavior. Indeed, let $\Gamma = \bigcup_{k=0}^{m} \operatorname{supp}(\mu_k)$; then we have

Theorem 3 (Lopez et al. [18]). *If* (1) *is l-regular, then for* $l \leqslant k \leqslant m$,

$$\lim_{n\to\infty} \left(\max_{z\in\Gamma} |Q_n^{(k)}(z)| \right)^{1/n} = C(\Gamma)$$

and, if the interior of Γ *is empty and* $\mathbb{C}\backslash\Gamma$ *is connected,*

$$\lim_{n\to\infty} v(Q_n^{(k)}) = \omega_\Gamma,$$

the equilibrium measure of Γ.

In particular, if (1) *is sequentially dominated and* 0*-regular, then for* $k = 0,\ldots,m$,

$$\lim_{n\to\infty} |Q_n^{(k)}(z)|^{1/n} = |\varphi(z)|$$

holds locally uniformly in the intersection of $\{z \in \mathbb{C}\colon |z| > ||R||\}$ *with the unbounded connected component of* $\mathbb{C} \setminus \Gamma$. *As before, R is the multiplication operator* (13) *in* $\mathbb{P}_S(\boldsymbol{\mu})$.

This exposition shows that the analytic theory of Sobolev orthogonal polynomials, though very abundant in results and conjectures, is still in its beginning. New approaches and fresh nonstandard ideas are needed. Moreover, in spite of its "numerical" motivation, the development of the theory up to now has obeyed more its own internal logic than the needs of the practitioner. Thus, a good stimulus outside to this field would be more than welcome and could help to state the right questions leading to beautiful answers.

Acknowledgements

Criticism of several colleagues who read the first version of the manuscript was very helpful in improving its readability. I would like also to thank many of the authors of the original papers mentioned here for helpful discussions. Last but not least, I am especially grateful to Prof. F. Marcellán for getting me started in this subject and keeping me going with his advice and enthusiasm.

References

[1] M. Alfaro, A. Martínez-Finkelshtein, M.L. Rezola, Asymptotic properties of balanced extremal Sobolev polynomials: coherent case, J. Approx. Theory 100 (1999) 44–59.

[2] M. Álvarez de Morales, J.J. Moreno-Balcázar, T.E. Pérez, M.A. Piñar, Non-diagonal Hermite–Sobolev orthogonal polynomials, Acta Appl. Math., in press.

[3] A. Aptekarev, E. Berriochoa, A. Cachafeiro, Strong asymptotics for the continuous Sobolev orthogonal polynomials on the unit circle, J. Approx. Theory 100 (1999) 381–391.

[4] A.I. Aptekarev, G. López, F. Marcellán, Orthogonal polynomials with respect to a differential operator, Existence and uniqueness, 1999, unpublished.

[5] I. Area, E. Godoy, F. Marcellán, Inner product involving differences: The Meixner–Sobolev polynomials, J. Differential Equations Appl. (1999), in press.

[6] I. Area, E. Godoy, F. Marcellán, J.J. Moreno-Balcázar, Inner products involving q-differences: The little q-Laguerre–Sobolev polynomials, J. Comput. Appl. Math. 118 (2000) 1–22.

[7] I. Area, E. Godoy, F. Marcellán, J.J. Moreno-Balcázar, Ratio and Plancherel–Rotach asymptotics for Meixner–Sobolev orthogonal polynomials, J. Comput. Appl. Math. 116 (2000) 63–75.

[8] D. Barrios, G. López Lagomasino, H. Pijeira, The moment problem for a Sobolev inner product, J. Approx. Theory 100 (1999) 364–380.

[9] H. Bavinck, On polynomials orthogonal with respect to an inner product involving differences, J. Comput. Appl. Math. 57 (1995) 17–27.

[10] W.D. Evans, L.L. Littlejohn, F. Marcellán, C. Markett, A. Ronveaux, On recurrence relations for Sobolev orthogonal polynomials, SIAM J. Math. Anal. 26 (2) (1995) 446–467.

[11] W.N. Everitt, L.L. Littlejohn, The density of polynomials in a weighted Sobolev space, Rendiconti di Matematica (Roma) Serie VII 10 (1990) 835–852.

[12] W.N. Everitt, L.L. Littlejohn, S.C. Williams, Orthogonal polynomials in weighted Sobolev spaces, in: J. Vinuesa (Ed.), Orthogonal Polynomials and Their Applications, Lecture Notes in Pure and Applied Math., Vol. 117, Marcel Dekker, New York, 1989, pp. 53–72.

[13] W.N. Everitt, L.L. Littlejohn, S.C. Williams, Orthogonal polynomials and approximation in Sobolev spaces, J. Comput. Appl. Math. 48 (1–2) (1993) 69–90.

[14] W. Gautschi, A.B.J. Kuijlaars, Zeros and critical points of Sobolev orthogonal polynomials, J. Approx. Theory 91 (1) (1997) 117–137.

[15] O. Kováčik, Sobolev spaces defined by differences, Proceedings of Seminar on Orthogonal Polynomials and their Applications (SOPOA), Department of Mathematics, Faculty of Civil Engineering, University of Transport and Communications, Žilina, Slovakia, 1993, pp. 19–24.

[16] G. López, F. Marcellán, W. Van Assche, Relative asymptotics for polynomials orthogonal with respect to a discrete Sobolev inner product, Constr. Approx. 11 (1995) 107–137.

[17] G. López, H. Pijeira-Cabrera, Zero location and *n*th root asymptotics of Sobolev orthogonal polynomials, J. Approx. Theory 99 (1) (1999) 30–43.

[18] G. López Lagomasino, H. Pijeira-Cabrera, I. Pérez Izquierdo, Sobolev orthogonal polynomials in the complex plane, this issue, J. Comput. Appl. Math. 127 (2001) 219–230.

[19] F. Marcellán, A. Martínez-Finkelshtein, J.J. Moreno-Balcázar, *k*-coherence of measures with non-classical weights, manuscript 1999.

[20] F. Marcellán, H.G. Meijer, T.E. Pérez, M.A. Piñar, An asymptotic result for Laguerre–Sobolev orthogonal polynomials, J. Comput. Appl. Math. 87 (1997) 87–94.

[21] F. Marcellán, J.J. Moreno-Balcázar, Strong and Plancherel–Rotach asymptotics of non-diagonal Laguerre–Sobolev orthogonal polynomials, J. Approx. Theory, in press.

[22] A. Martínez-Finkelshtein, Asymptotic properties of Sobolev orthogonal polynomials, J. Comput. Appl. Math. 99 (1–2) (1998) 491–510.

[23] A. Martínez-Finkelshtein, Bernstein–Szegő's theorem for Sobolev orthogonal polynomials, Constr. Approx. 16 (2000) 73–84.

[24] A. Martínez-Finkelshtein, J.J. Moreno-Balcázar, T.E. Pérez, M.A. Piñar, Asymptotics of Sobolev orthogonal polynomials for coherent pairs, J. Approx. Theory 92 (1998) 280–293.

[25] A. Martínez-Finkelshtein, H. Pijeira-Cabrera, Strong asymptotics for Sobolev orthogonal polynomials, J. Anal. Math. 78 (1999) 143–156.

[26] H.G. Meijer, A short history of orthogonal polynomials in a Sobolev space I. The non-discrete case, Nieuw Arch. Wisk. 14 (1996) 93–113.

[27] H.G. Meijer, Determination of all coherent pairs, J. Approx. Theory 89 (1997) 321–343.

[28] H.G. Meijer, T.E. Pérez, M. Piñar, Asymptotics of Sobolev orthogonal polynomials for coherent pairs of Laguerre type, Report 97-45, T.U. Delft, 1997.

[29] P. Nevai, Orthogonal Polynomials, Memoirs American Mathematical Society, Vol. 213, American Mathematical Society, Providence, RI, 1979.

[30] H.E. Pijeira Cabrera, Teoría de Momentos y Propiedades Asintóticas para Polinomios Ortogonales de Sobolev, Doctoral Dissertation, Universidad Carlos III de Madrid, October 1998.

[31] J.M. Rodríguez, Multiplication operator in Sobolev spaces with respect to measures, 1998, submitted for publication.

[32] J.M. Rodríguez, Approximation by polynomials and smooth functions in Sobolev spaces with respect to measures, 1999, submitted for publication.

[33] J.M. Rodríguez, Weierstrass' theorem in weighted Sobolev spaces, J. Approx. Theory, in press.

[34] J.M. Rodríguez, V. Álvarez, E. Romera, D. Pestana, Generalized weighted Sobolev spaces and applications to Sobolev orthogonal polynomials, 1998, submitted for publication.

[35] E.B. Saff, V. Totik, Logarithmic Potentials with External Fields, Grundlehren der Mathematischen Wissenschaften, Vol. 316, Springer, Berlin, 1997.

[36] H. Stahl, V. Totik, General Orthogonal Polynomials, Encyclopedia of Mathematics and its Applications, Vol. 43, Cambridge University Press, Cambridge, 1992.

Journal of Computational and Applied Mathematics 127 (2001) 267–286

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Quadratures with multiple nodes, power orthogonality, and moment-preserving spline approximation ☆

Gradimir V. Milovanović

Faculty of Electronic Engineering, Department of Mathematics, University of Niš, P.O. Box 73, 18000 Niš, Serbia, Yugoslavia

Received 29 October 1999

Abstract

Quadrature formulas with multiple nodes, power orthogonality, and some applications of such quadratures to moment-preserving approximation by defective splines are considered. An account on power orthogonality (s- and σ-orthogonal polynomials) and generalized Gaussian quadratures with multiple nodes, including stable algorithms for numerical construction of the corresponding polynomials and Cotes numbers, are given. In particular, the important case of Chebyshev weight is analyzed. Finally, some applications in moment-preserving approximation of functions by defective splines are discussed. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Quadratures with multiple nodes; Gauss–Turán-type quadratures; Error term; Convergence; Orthogonal polynomials; s- and σ-orthogonal polynomials; Nonnegative measure; Extremal polynomial; Weights; Nodes; Degree of precision; Stieltjes procedure; Chebyshev polynomials; Spline function; Spline defect; Moments

1. Introduction

More than 100 years after Gauss published his famous method of approximate integration, which was enriched by significant contributions of Jacobi and Christoffel, there appeared the idea of numerical integration involving multiple nodes. Taking any system of n distinct points $\{\tau_1,\ldots,\tau_n\}$ and n nonnegative integers $m_1,\ldots,m_n$, and starting from the Hermite the interpolation formula, Chakalov (Tschakaloff in German transliteration) [8] in 1948 obtained the quadrature formula

$$\int_{-1}^{1} f(t)\,\mathrm{d}t = \sum_{\nu=1}^{n} [A_{0,\nu}f(\tau_\nu) + A_{1,\nu}f'(\tau_\nu) + \ldots + A_{m_\nu-1,\nu}f^{(m_\nu-1)}(\tau_\nu)], \tag{1.1}$$

☆ This work was supported in part by the Serbian Ministry of Science and Technology, grant number 04M03.

E-mail address: grade@elfak.ni.ac.yu (G.V. Milovanović).

PII: S0377-0427(00)00500-8

which is exact for all polynomials of degree at most $m_1 + \ldots + m_n - 1$. Precisely, he gave a method for computing the coefficients $A_{i,v}$ in (1.1). Such coefficients (Cotes numbers of higher order) are evidently $A_{i,v} = \int_{-1}^{1} \ell_{i,v}(t)\,\mathrm{d}t$ $(v = 1,\ldots,n;\ i = 0,1,\ldots,m_v - 1)$, where $\ell_{i,v}(t)$ are the fundamental functions of Hermite interpolation.

In 1950, specializing $m_1 = \ldots = m_n = k$ in (1.1), Turán [90] studied numerical quadratures of the form

$$\int_{-1}^{1} f(t)\,\mathrm{d}t = \sum_{i=0}^{k-1}\sum_{v=1}^{n} A_{i,v} f^{(i)}(\tau_v) + R_{n,k}(f). \tag{1.2}$$

Let $\mathscr{P}_m$ be the set of all algebraic polynomials of degree at most m. It is clear that formula (1.2) can be made exact for $f \in \mathscr{P}_{kn-1}$, for any given points $-1 \leqslant \tau_1 \leqslant \ldots \leqslant \tau_n \leqslant 1$. However, for $k = 1$ formula (1.2), i.e.,

$$\int_{-1}^{1} f(t)\,\mathrm{d}t = \sum_{v=1}^{n} A_{0,v} f(\tau_v) + R_{n,1}(f)$$

is exact for all polynomials of degree at most $2n - 1$ if the nodes τ_v are the zeros of the Legendre polynomial P_n, and it is the well-known Gauss–Legendre quadrature rule.

Because of Gauss's result it is natural to ask whether nodes τ_v can be chosen so that the quadrature formula (1.2) will be exact for algebraic polynomials of degree not exceeding $(k + 1)n - 1$. Turán [90] showed that the answer is negative for $k = 2$, and for $k = 3$ it is positive. He proved that the nodes τ_v should be chosen as the zeros of the monic polynomial $\pi_n^*(t) = t^n + \ldots$ which minimizes the integral $\int_{-1}^{1} [\pi_n(t)]^4\,\mathrm{d}t$, where $\pi_n(t) = t^n + a_{n-1}t^{n-1} + \ldots + a_1 t + a_0$.

In the general case, the answer is negative for even, and positive for odd k, and then τ_v must be the zeros of the polynomial minimizing $\int_{-1}^{1} [\pi_n(t)]^{k+1}\,\mathrm{d}t$. When $k = 1$, then π_n is the monic Legendre polynomial $\hat{P}_n$.

Because of the above, we assume that $k = 2s + 1$, $s \geqslant 0$. Instead of (1.2), it is also interesting to consider a more general *Gauss–Turán-type* quadrature formula

$$\int_{\mathbb{R}} f(t)\,\mathrm{d}\lambda(t) = \sum_{i=0}^{2s}\sum_{v=1}^{n} A_{i,v} f^{(i)}(\tau_v) + R_{n,2s}(f), \tag{1.3}$$

where $\mathrm{d}\lambda(t)$ is a given nonnegative measure on the real line $\mathbb{R}$, with compact or unbounded support, for which all moments $\mu_k = \int_{\mathbb{R}} t^k\,\mathrm{d}\lambda(t)$ $(k = 0,1,\ldots)$ exist and are finite, and $\mu_0 > 0$. It is known that formula (1.3) is exact for all polynomials of degree not exceeding $2(s + 1)n - 1$, i.e., $R_{n,2s}(f) = 0$ for $f \in \mathscr{P}_{2(s+1)n-1}$. The nodes τ_v $(v = 1,\ldots,n)$ in (1.3) are the zeros of the monic polynomial $\pi_{n,s}(t)$, which minimizes the integral

$$F(a_0, a_1, \ldots, a_{n-1}) = \int_{\mathbb{R}} [\pi_n(t)]^{2s+2}\,\mathrm{d}\lambda(t), \tag{1.4}$$

where $\pi_n(t) = t^n + a_{n-1}t^{n-1} + \ldots + a_1 t + a_0$. This minimization leads to the conditions

$$\frac{1}{2s+2}\frac{\partial F}{\partial a_k} = \int_{\mathbb{R}} [\pi_n(t)]^{2s+1} t^k\,\mathrm{d}\lambda(t) = 0 \qquad (k = 0,1,\ldots,n-1). \tag{1.5}$$

These polynomials $\pi_n = \pi_{n,s}$ are known as *s-orthogonal* (or *s-self associated*) *polynomials* on $\mathbb{R}$ with respect to the measure $\mathrm{d}\lambda(t)$ (for more details see [15,62,65,66]. For $s = 0$ they reduce to the standard orthogonal polynomials and (1.3) becomes the well-known Gauss–Christoffel formula.

Using some facts about monosplines, Micchelli [47] investigated the sign of the Cotes coefficients $A_{i,v}$ in the Turán quadrature.

A generalization of the Turán quadrature formula (1.3) (for $\mathrm{d}\lambda(t)=\mathrm{d}t$ on (a,b)) to rules having nodes with arbitrary multiplicities was derived independently by Chakalov [9,10] and Popoviciu [74]. Important theoretical progress on this subject was made by Stancu [82,84] (see also [88]).

In this case, it is important to assume that the nodes τ_v are ordered, say

$$\tau_1 < \tau_2 < \ldots < \tau_n, \tag{1.6}$$

with multiplicities m_1, m_2, …, m_n, respectively. A permutation of the multiplicities m_1, m_2, …, m_n, with the nodes held fixed, in general yields a new quadrature rule.

It can be shown that the quadrature formula (1.1) is exact for all polynomials of degree less than $2\sum_{v=1}^{n}[(m_v+1)/2]$. Thus, the multiplicities m_v that are even do not contribute toward an increase in the degree of exactness, so that it is reasonable to assume that all m_v be odd integers, $m_v=2s_v+1$ $(v=1,2,\ldots,n)$. Therefore, for a given sequence of nonnegative integers $\sigma=(s_1,s_2,\ldots,s_n)$ the corresponding quadrature formula

$$\int_{\mathbb{R}} f(t)\,\mathrm{d}\lambda(t)=\sum_{v=1}^{n}\sum_{i=0}^{2s_v} A_{i,v} f^{(i)}(\tau_v)+R(f) \tag{1.7}$$

has maximum degree of exactness

$$d_{\max}=2\sum_{v=1}^{n} s_v+2n-1 \tag{1.8}$$

if and only if

$$\int_{\mathbb{R}} \prod_{v=1}^{n}(t-\tau_v)^{2s_v+1} t^k\,\mathrm{d}\lambda(t)=0 \quad (k=0,1,\ldots,n-1). \tag{1.9}$$

The last *orthogonality conditions* correspond to (1.5) and they could be obtained by the minimization of the integral

$$\int_{\mathbb{R}} \prod_{v=1}^{n}(t-\tau_v)^{2s_v+2}\,\mathrm{d}\lambda(t).$$

The existence of such quadrature rules was proved by Chakalov [9], Popoviciu [74], Morelli and Verna [57], and existence and uniqueness (subject to (1.6)) by Ghizzetti and Ossicini [27].

Conditions (1.9) define a sequence of polynomials $\{\pi_{n,\sigma}\}_{n\in\mathbb{N}_0}$,

$$\pi_{n,\sigma}(t)=\prod_{v=1}^{n}(t-\tau_v^{(n,\sigma)}), \qquad \tau_1^{(n,\sigma)} < \tau_2^{(n,\sigma)} < \ldots < \tau_n^{(n,\sigma)},$$

such that

$$\int_{\mathbb{R}} \pi_{k,\sigma}(t)\prod_{v=1}^{n}(t-\tau_v^{(n,\sigma)})^{2s_v+1}\,\mathrm{d}\lambda(t)=0 \quad (k=0,1,\ldots,n-1). \tag{1.10}$$

Thus, we get now a general type of power orthogonality. These polynomials $\pi_{k,\sigma}$ are called σ-orthogonal polynomials, and they correspond to the sequence $\sigma=(s_1,s_2,\ldots)$. We will often write simple τ_ν or $\tau_\nu^{(n)}$ instead of $\tau_\nu^{(n,\sigma)}$. If we have $\sigma=(s,s,\dots)$, the above polynomials reduce to the s-orthogonal polynomials.

This paper is devoted to quadrature formulas with multiple nodes, power orthogonality, and some applications of such quadrature formulas to moment-preserving approximation by defective splines. In Section 2, we give an account on power orthogonality, which includes some properties of s- and σ-orthogonal polynomials and their construction. Section 3 is devoted to some methods for constructing generalized Gaussian formulas with multiple nodes. The important case of Chebyshev weight is analyzed in Section 4. Finally, some applications to moment-preserving approximation by defective splines are discussed in Section 5.

2. Power orthogonality

This section is devoted to power-orthogonal polynomials. We give an account on theoretical results on this subject, and we also consider methods for numerical construction of such polynomials.

2.1. Properties of s- and σ-orthogonal polynomials

The orthogonality conditions for s-orthogonal polynomials $\pi_{n,s}=\pi_{n,s}(\,\cdot\,;\mathrm{d}\lambda)$ are given by (1.5) i.e.,

$$\int_{\mathbb{R}} [\pi_{n,s}(t)]^{2s+1}\pi_{k,s}(t)\,\mathrm{d}\lambda(t)=0 \qquad (k=0,1,\dots,n-1). \tag{2.1}$$

These polynomials were investigated mainly by Italian mathematicians, especially the case $\mathrm{d}\lambda(t)=w(t)\,\mathrm{d}t$ on $[a,b]$ (e.g., Ossicini [62,63], Ghizzetti and Ossicini [23–27], Guerra [37,38], Ossicini and Rosati [67–69], Gori [29], Gori and Lo Cascio [30]). The basic result concerns related to zero distribution.

Theorem 2.1. *There exists a unique monic polynomial $\pi_{n,s}$ for which* (2.1) *is satisfied, and $\pi_{n,s}$ has n distinct real zeros which are all contained in the open interval* (a,b).

This result was proved by Turán [90] for $\mathrm{d}\lambda(t)=\mathrm{d}t$ on $[-1,1]$. It was also proved by Ossicini [62] (see also the book [24, pp. 74–75]) using different methods.

Usually, we assume that the zeros $\tau_\nu=\tau_\nu^{(n,s)}$ $(\nu=1,2,\dots,n)$ of $\pi_{n,s}$ are ordered as in (1.6).

In the symmetric case $w(-t)=w(t)$ on $[-b,b]$ $(b>0)$, it is easy to see that $\pi_{n,s}(-t)=(-1)^n\pi_{n,s}(t)$. In the simplest case of Legendre s-orthogonal polynomials $P_{n,s}(t)=a_n\prod_{\nu=1}^n(t-\tau_\nu)$, where the normalization factor a_n is taken to have $P_{n,s}(1)=1$, Ghizzetti and Ossicini [23] proved that $|P_{n,s}(t)|\leqslant 1$, when $-1\leqslant t\leqslant 1$. Also, they determined the minimum in (1.4) in this case,

$$F_{n,s}=\int_{-1}^{1}[P_{n,s}(t)]^{2s+2}\,\mathrm{d}t=\frac{2}{1+(2s+2)n}.$$

Indeed, integration by parts gives

$$F_{n,s}=[tP_{n,s}(t)^{2s+2}]_{-1}^{1}-(2s+2)\int_{-1}^{1}tP_{n,s}(t)^{2s+1}P'_{n,s}(t)\,\mathrm{d}t=2-(2s+2)nF_{n,s}$$

because $tP'_{n,s}(t)=nP_{n,s}(t)+Q(t)$ $(Q\in\mathscr{P}_{n-2}$ in this symmetric case). It would be interesting to determine this minimum for other classical weights.

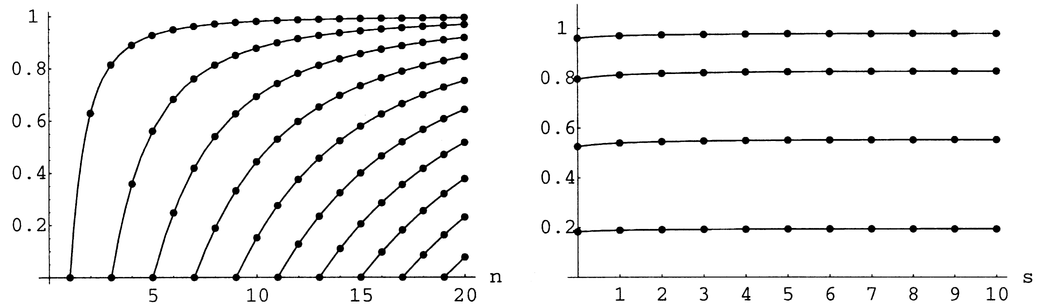


Fig. 1. Nonnegative zeros of $P_{n,s}(t)$ for $s=1$ and $n=1(1)20$ (left); Positive zeros of $P_{n,s}(t)$ for $n=8$ and $s=0(1)10$ (right).

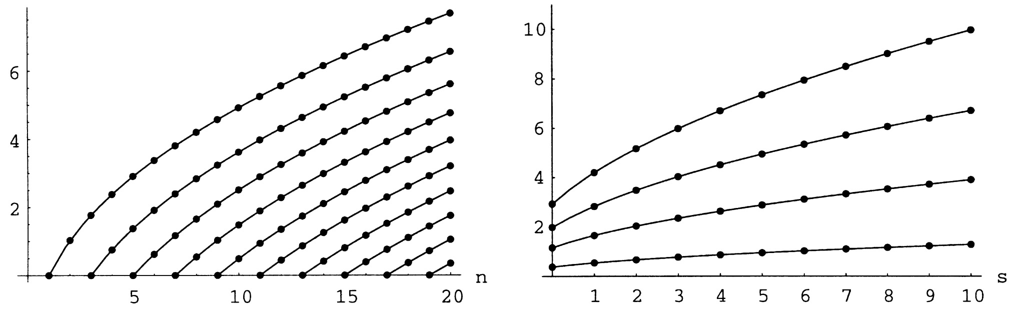


Fig. 2. Nonnegative zeros of $H_{n,s}(t)$ for $s=1$ and $n=1(1)20$ (left); Positive zeros of $H_{n,s}(t)$ for $n=8$ and $s=0(1)10$ (right).

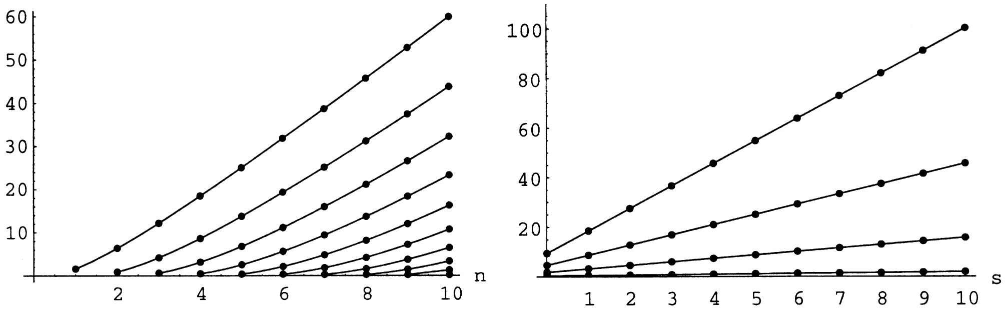


Fig. 3. Zeros of $L_{n,s}(t)$ for $s=1$ and $n=1(1)10$ (left) and for $n=4$ and $s=0(1)10$ (right).

In Fig. 1 we display the distribution of nonnegative zeros for Legendre s-orthogonal polynomials, taking $s=1$ and $n=1,2,\ldots,20$. Also, we present graphics when n is fixed $(n=8)$ and s runs up to 10. The corresponding graphics for Hermite s-orthogonal polynomials $H_{n,s}$ are given in Fig. 2.

In Fig. 3 we present all zeros of Laguerre s-orthogonal polynomials for $s=1$ and $n \leqslant 10$, and also for $n=4$ and $s \leqslant 10$. Also, we give the corresponding zero distribution of generalized Laguerre s-orthogonal polynomial $L_{n,s}^{(\alpha)}$, when $\alpha \in (-1,5]$ $(n=4,\ s=1)$ (see Fig. 4). Numerical experimentation suggests the following result.

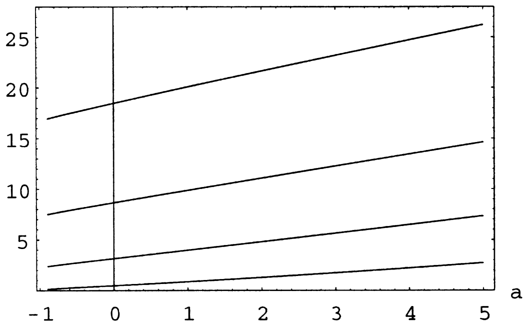


Fig. 4. Zero distribution of $L_{n,s}^{\alpha}(t)$ for $n=4$, $s=1$ and $-1<\alpha\leqslant 5$.

Theorem 2.2. *For every $s\in\mathbb{N}_0$, the zeros of $\pi_{n,s}$ and $\pi_{n+1,s}$ mutually separate each other.*

This interlacing property is well-known when $s=0$ (cf. [89, p. 46], [11, p. 28]). The proof of Theorem 2.2 can be obtained by applying a general result on interlacing properties of the zeros of the error functions in best L^p-approximations, given by Pinkus and Ziegler [73, Theorem 1.1]. Precisely, we put $u_\nu(t)=t^{\nu-1}$ $(\nu=1,\ldots,n+2)$, $p=2s+2$, and then use Corollary 1.1 from [73]. In the notation of this paper, $q_{n,p}=\pi_{n,s}$ and $q_{n+1,p}=\pi_{n+1,s}$, and their zeros strictly interlace for each $s\geqslant 0$.

A particularly interesting case is the Chebyshev measure

$$\mathrm{d}\lambda_1(t)=(1-t^2)^{-1/2}\mathrm{d}t.$$

In 1930, Bernstein [3] showed that the monic Chebyshev polynomial $\hat{T}_n(t)=T_n(t)/2^{n-1}$ minimizes all integrals of the form

$$\int_{-1}^{1}\frac{|\pi_n(t)|^{k+1}}{\sqrt{1-t^2}}\,\mathrm{d}t \qquad (k\geqslant 0).$$

Thus, the Chebyshev polynomials T_n are s-orthogonal on $[-1,1]$ for each $s\geqslant 0$. Ossicini and Rosati [65] found three other measures $\mathrm{d}\lambda_k(t)$ $(k=2,3,4)$ for which the s-orthogonal polynomials can be identified as Chebyshev polynomials of the second, third, and fourth kind: S_n, V_n, and W_n, which are defined by

$$S_n(\cos\theta)=\frac{\sin(n+1)\theta}{\sin\theta},\qquad V_n(\cos\theta)=\frac{\cos(n+\frac{1}{2})\theta}{\cos\frac{1}{2}\theta},\qquad W_n(\cos\theta)=\frac{\sin(n+\frac{1}{2})\theta}{\sin\frac{1}{2}\theta},$$

respectively (cf. [18]). However, these measures depend on s,

$$\mathrm{d}\lambda_2(t)=(1-t^2)^{1/2+s}\mathrm{d}t,\qquad \mathrm{d}\lambda_3(t)=\frac{(1+t)^{1/2+s}}{(1-t)^{1/2}}\mathrm{d}t,\qquad \mathrm{d}\lambda_4(t)=\frac{(1-t)^{1/2+s}}{(1+t)^{1/2}}\mathrm{d}t.$$

Notice that $W_n(-t)=(-1)^nV_n(t)$.

Considering the set of Jacobi polynomials $P_n^{(\alpha,\beta)}$, Ossicini and Rosati [69] showed that the only Jacobi polynomials which are s-orthogonal for a positive integer s are the Chebyshev polynomials of the first kind, which occur when $\alpha=\beta=-\frac{1}{2}$. Recently, Shi [77] (see also [78]) has proved that the Chebyshev weight $w(t)=(1-t^2)^{-1/2}$ is the only weight (up to a linear transformation) having the property: *For each fixed n, the solutions of the extremal problem*

$$\int_{-1}^{1}\left(\prod_{\nu=1}^{n}(t-\tau_\nu)\right)^{m} w(t)\,\mathrm{d}t=\min_{\pi(t)=t^n+\cdots}\int_{-1}^{1}[\pi(t)]^m w(t)\,\mathrm{d}t \tag{2.2}$$

for every even m are the same. Precisely, he proved the following result.

Theorem 2.3. *Let w be a weight supported on $[-1,1]$ such that $\int_{-1}^{1}w(t)\,\mathrm{d}t=1$. If (2.2) holds for the following pairs (m,n):*

$$m=m_1,m_2,\ldots,\ if\ n=1,2,4,\quad and\quad m=2,4,\ if\ n=3,5,6,\ldots,$$

where $\{m_k\}_{k\in\mathbb{N}}$ is a strictly increasing sequence of even natural numbers such that $m_1=2$ and $\sum_{k=1}^{+\infty}(1/m_k)=+\infty$, then there exist two numbers α and β such that $w=v_{\alpha,\beta}$, where

$$v_{\alpha,\beta}(t)=\begin{cases}\dfrac{1}{\pi\sqrt{(t-\alpha)(\beta-t)}}, & t\in(\alpha,\beta),\\ 0, & t\notin(\alpha,\beta).\end{cases}$$

Recently, Gori and Micchelli [33] have introduced for each n a class of weight functions $\mathscr{W}_n$ defined on $[-1,1]$ for which explicit n-point Gauss–Turán quadrature formulas of all orders can be found. In other words, these classes of weight functions have the peculiarity that the corresponding s-orthogonal polynomials, of the same degree, are independent of s. The class $\mathscr{W}_n$ includes certain generalized Jacobi weight functions $w_{n,\mu}(t)=|S_{n-1}(t)/n|^{2\mu+1}(1-t^2)^{\mu}$, where $S_{n-1}(\cos\theta)=\sin n\theta/\sin\theta$ (Chebyshev polynomial of the second kind) and $\mu>-1$. In this case, the Chebyshev polynomials T_n appear as s-orthogonal polynomials. For $n=2$ the previous weight function reduces to the weight $w_{2,\mu}(t)=|t|^{2\mu+1}(1-t^2)^{\mu}$, which was studied in [30,31,36].

Very little is known about σ-orthogonal polynomials. Except for Rodrigues' formula, which has an analogue for these polynomials (see [25,26]), no general theory is available. Some particular results on zeros of σ-orthogonal polynomials and their asymptotic behavior are known (cf. [59–61]). The Legendre case with $\sigma=(0,s)$ was considered by Morelli and Verna [59], and they proved that

$$\lim_{s\to+\infty}\tau_1=-1\quad\text{and}\quad\lim_{s\to+\infty}\tau_2=0.$$

2.2. Numerical construction of power-orthogonal polynomials

An iterative process for computing the coefficients of s-orthogonal polynomials in a special case, when the interval $[a,b]$ is symmetric with respect to the origin and the weight w is an even function, was proposed by Vincenti [93]. He applied his process to the Legendre case. When n and s increase, the process becomes numerically unstable.

At the Third Conference on Numerical Methods and Approximation Theory (Niš, August 18–21, 1987) (see [51]) we presented a stable method for numerically constructing s-orthogonal polynomials and their zeros. It uses an iterative method with quadratic convergence based on a discretized Stieltjes procedure and the Newton–Kantorovič method.

Table 1

n	$\mathrm{d}\mu^{(n,s)}(t)$	Orthogonal polynomials			
0	$(\pi_0^{(0,s)}(t))^{2s}\mathrm{d}\lambda(t)$	$\boxed{\pi_0^{(0,s)}}$			
1	$(\pi_1^{(1,s)}(t))^{2s}\mathrm{d}\lambda(t)$	$\pi_0^{(1,s)}$	$\boxed{\pi_1^{(1,s)}}$		
2	$(\pi_2^{(2,s)}(t))^{2s}\mathrm{d}\lambda(t)$	$\pi_0^{(2,s)}$	$\pi_1^{(2,s)}$	$\boxed{\pi_2^{(2,s)}}$	
3	$(\pi_3^{(3,s)}(t))^{2s}\mathrm{d}\lambda(t)$	$\pi_0^{(3,s)}$	$\pi_1^{(3,s)}$	$\pi_2^{(3,s)}$	$\boxed{\pi_3^{(3,s)}}$
$\vdots$					

The basic idea for our method to numerically construct s-orthogonal polynomials with respect to the measure $\mathrm{d}\lambda(t)$ on the real line $\mathbb{R}$ is a reinterpretation of the "*orthogonality conditions*" (2.1). For given n and s, we put $\mathrm{d}\mu(t)=\mathrm{d}\mu^{(n,s)}(t)=(\pi_{n,s}(t))^{2s}\mathrm{d}\lambda(t)$. The conditions can then be written as

$$\int_{\mathbb{R}} \pi_k^{(n,s)}(t)t^{\nu}\mathrm{d}\mu(t)=0 \quad (\nu=0,1,\dots,k-1),$$

where $\{\pi_k^{(n,s)}\}$ is a sequence of monic orthogonal polynomials with respect to the new measure $\mathrm{d}\mu(t)$. Of course, $\pi_{n,s}(\cdot)=\pi_n^{(n,s)}(\cdot)$. As we can see, the polynomials $\pi_k^{(n,s)}$ $(k=0,1,\dots)$ are implicitly defined, because the measure $\mathrm{d}\mu(t)$ depends of $\pi_n^{(n,s)}(t)$. A general class of such polynomials was introduced and studied by Engels (cf. [12, pp. 214–226]). We will write simply $\pi_k(\cdot)$ instead of $\pi_k^{(n,s)}(\cdot)$. These polynomials satisfy a three-term recurrence relation

$$\begin{gathered}\pi_{\nu+1}(t)=(t-\alpha_\nu)\pi_\nu(t)-\beta_\nu\pi_{\nu-1}(t), \quad \nu=0,1,\dots,\\ \pi_{-1}(t)=0, \quad \pi_0(t)=1,\end{gathered} \tag{2.3}$$

where because of orthogonality

$$\begin{aligned}\alpha_\nu&=\alpha_\nu(n,s)=\frac{(t\pi_\nu,\pi_\nu)}{(\pi_\nu,\pi_\nu)}=\frac{\int_{\mathbb{R}} t\pi_\nu^2(t)\,\mathrm{d}\mu(t)}{\int_{\mathbb{R}} \pi_\nu^2(t)\,\mathrm{d}\mu(t)},\\ \beta_\nu&=\beta_\nu(n,s)=\frac{(\pi_\nu,\pi_\nu)}{(\pi_{\nu-1},\pi_{\nu-1})}=\frac{\int_{\mathbb{R}} \pi_\nu^2(t)\,\mathrm{d}\mu(t)}{\int_{\mathbb{R}} \pi_{\nu-1}^2(t)\,\mathrm{d}\mu(t)}\end{aligned} \tag{2.4}$$

and by convention, $\beta_0=\int_{\mathbb{R}}\mathrm{d}\mu(t)$.

The coefficients α_ν and β_ν are the fundamental quantities in the constructive theory of orthogonal polynomials. They provide a compact way of representing orthogonal polynomials, requiring only a linear array of parameters. The coefficients of orthogonal polynomials, or their zeros, in contrast need two-dimensional arrays. Knowing the coefficients α_ν, β_ν $(\nu=0,1,\dots,n-1)$ gives us access to the first $n+1$ orthogonal polynomials π_0, $\pi_1,\dots,\pi_n$. Of course, for a given n, we are interested only in the last of them, i.e., $\pi_n\equiv\pi_n^{(n,s)}$. Thus, for $n=0,1,\dots,$ the diagonal (boxed) elements in Table 1 are our s-orthogonal polynomials $\pi_n^{(n,s)}$.

A stable procedure for finding the coefficients α_ν, β_ν is the discretized Stieltjes procedure, especially for infinite intervals of orthogonality (see [15,16,20]). Unfortunately, in our case this procedure cannot be applied directly, because the measure $\mathrm{d}\mu(t)$ involves an unknown polynomial

$\pi_n^{(n,s)}$. Consequently, we consider the system of nonlinear equations in the unknowns $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$, $\beta_0, \beta_1, \ldots, \beta_{n-1}$

$$\begin{aligned} f_0 &\equiv \beta_0 - \int_{\mathbb{R}} \pi_n^{2s}(t)\,\mathrm{d}\lambda(t) = 0, \\ f_{2v+1} &\equiv \int_{\mathbb{R}} (\alpha_v - t)\pi_v^2(t)\pi_n^{2s}(t)\,\mathrm{d}\lambda(t) = 0 \quad (v = 0, 1, \ldots, n-1), \\ f_{2v} &\equiv \int_{\mathbb{R}} (\beta_v \pi_{v-1}^2(t) - \pi_v^2(t))\pi_n^{2s}(t)\,\mathrm{d}\lambda(t) = 0 \quad (v = 1, \ldots, n-1), \end{aligned} \tag{2.5}$$

which follows from (2.4), and then we apply the Newton–Kantorovič method for determining the coefficients of the recurrence relation (2.3) (see [51,22]). If sufficiently good starting approximations are chosen, the convergence of this method is quadratic. The elements of the Jacobian can be easily computed using the recurrence relation (2.3), but with other (delayed) initial values (see [51,22]). All integrals in (2.5), as well as the integrals in the elements of the Jacobian, can be computed exactly, except for rounding errors, by using a Gauss–Christoffel quadrature formula with respect to the measure $\mathrm{d}\lambda(t)$:

$$\int_{\mathbb{R}} g(t)\,\mathrm{d}\lambda(t) = \sum_{v=1}^{N} A_v^{(N)} g(\tau_v^{(N)}) + R_N(g), \tag{2.6}$$

taking $N = (s+1)n$ nodes. This formula is exact for all polynomials of degree at most $2N - 1 = 2(s+1)n - 1 = 2(n-1) + 2ns + 1$.

Thus, all calculations in this method are based on using only the fundamental three-term recurrence relation (2.3) and the Gauss–Christoffel quadrature formula (2.6). The problem of finding sufficiently good starting approximations for $\alpha_v^{[0]} = \alpha_v^{[0]}(n,s)$ and $\beta_v^{[0]} = \beta_v^{[0]}(n,s)$ is the most serious one. In [51,22] we proposed to take the values obtained for $n-1$, i.e., $\alpha_v^{[0]} = \alpha_v(s, n-1)$, $\beta_v^{[0]} = \beta_v(s, n-1)$, $v \leqslant n-2$, and the corresponding extrapolated values for $\alpha_{n-1}^{[0]}$ and $\beta_{n-1}^{[0]}$. In the case $n = 1$ we solve the equation

$$\phi(\alpha_0) = \phi(\alpha_0(s,1)) = \int_{\mathbb{R}} (t - \alpha_0)^{2s+1}\,\mathrm{d}\lambda(t) = 0,$$

and then determine $\beta_0 = \beta_0(s,1) = \int_{\mathbb{R}} (t - \alpha_0)^{2s}\,\mathrm{d}\lambda(t)$.

The zeros $\tau_v = \tau_v(n,s)$ $(v = 1, \ldots, n)$ of $\pi_n^{(n,s)}$, i.e., the nodes of the Gauss–Turán-type quadrature formula (1.3), can be obtained very easily as eigenvalues of a (symmetric tridiagonal) Jacobi matrix J_n using the QR algorithm, namely

$$J_n = \begin{bmatrix} \alpha_0 & \sqrt{\beta_1} & & & \mathrm{O} \\ \sqrt{\beta_1} & \alpha_1 & \sqrt{\beta_2} & & \\ & \sqrt{\beta_2} & \alpha_2 & \ddots & \\ & & \ddots & \ddots & \sqrt{\beta_{n-1}} \\ \mathrm{O} & & & \sqrt{\beta_{n-1}} & \alpha_{n-1} \end{bmatrix},$$

where $\alpha_v = \alpha_v(n,s)$, $\beta_v = \beta_v(n,s)$ $(v = 0, 1, \ldots, n-1)$.

Table 2

σ	$(1,1,3)$	$(1,3,1)$	$(3,1,1)$
$\tau_1^{(3,\sigma)}$	−2.30298348189811	−2.26862030544612	−1.57815506119966
$\tau_2^{(3,\sigma)}$	−0.62210813435576	0	0.62210813435576
$\tau_3^{(3,\sigma)}$	1.57815506119966	2.26862030544612	2.30298348189811

An iterative method for the construction of σ-orthogonal polynomials was developed by Gori et al. [32]. In this case, the corresponding reinterpretation of the "*orthogonality conditions*" (1.10) leads to conditions

$$\int_{\mathbb{R}} \pi_k^{(n,\sigma)}(t)t^\nu \, \mathrm{d}\mu(t) = 0 \qquad (\nu = 0,1,\ldots,k-1),$$

where

$$\mathrm{d}\mu(t) = \mathrm{d}\mu^{(n,\sigma)}(t) = \prod_{\nu=1}^{n} (t - \tau_\nu^{(n,\sigma)})^{2s_\nu} \, \mathrm{d}\lambda(t). \tag{2.7}$$

Therefore, we conclude that $\{\pi_k^{(n,\sigma)}\}$ is a sequence of (standard) orthogonal polynomials with respect to the measure $\mathrm{d}\mu(t)$. Evidently, $\pi_n^{(n,\sigma)}(\,\cdot\,)$ is the desired σ-orthogonal polynomial $\pi_{n,\sigma}(\,\cdot\,)$. Since $\mathrm{d}\mu(t)$ is given by (2.7), we cannot apply here the same procedure as in the case of s-orthogonal polynomials. Namely, the determination of the Jacobian requires the partial derivatives of the zeros $\tau_\nu^{(n,\sigma)}$ with respect to α_k and β_k, which is not possible in an analytic form. Because of that, in [32] a discrete analogue of the Newton–Kantorovič method (a version of the secant method) was used. The convergence of this method is superlinear and strongly depends on the choice of the starting points. Recently, Milovanović and Spalević [56] have considered an iterative method for determining the zeros of σ-orthogonal polynomials.

As we mentioned in Section 1, σ-orthogonal polynomials are unique when (1.6) is imposed, with corresponding multiplicities $m_1, m_2, \ldots, m_n$. Otherwise, the number of distinct σ-polynomials is $n!/(k_1!k_2!\cdots k_q!)$ for some q $(1 \leqslant q \leqslant n)$, where k_i is the number of nodes of multiplicity $m_j = i$, each node counted exactly once, $\sum_{i=1}^{q} k_i = n$. For example, in the case $n = 3$, with multiplicities 3, 3, 7, we have three different Hermite σ-polynomials ($w(t) = \mathrm{e}^{-t^2}$ on $\mathbb{R}$), which correspond to $\sigma = (1,1,3)$, $(1,3,1)$, and $(3,1,1)$ (see Table 2).

3. Generalized Gaussian quadrature with multiple nodes

3.1. A theoretical approach

In order to construct a quadrature formula of form (1.7), with multiple nodes τ_ν (whose multiplicities are $m_\nu = 2s_\nu + 1$), Stroud and Stancu [88] (see also Stancu [80,84]) considered ℓ distinct real numbers $\alpha_1, \ldots, \alpha_\ell$ and assumed that none of these coincide with any of the τ_ν. The Lagrange–Hermite interpolation polynomial for the function f at simple nodes α_μ and the multiple nodes τ_ν,

$$L(t;f) \equiv L\begin{pmatrix} \tau_1 & & \tau_n & \alpha_1 & & \alpha_\ell & \\ 2s_1+1 & ,\dots, & 2s_n+1, & 1 & ,\dots, & 1 & ; \quad f|t \end{pmatrix}$$

can be expressed in the form

$$L(t;f) = \omega(t)L\begin{pmatrix} \tau_1 & & \tau_n & \\ 2s_1+1 & ,\dots, & 2s_n+1 & ; \quad f_1|t \end{pmatrix} + \Omega(t)L\begin{pmatrix} \alpha_1 & & \alpha_\ell & \\ 1 & ,\dots, & 1 & ; \quad f_2|t \end{pmatrix},$$

where $\omega(t) = (t-\alpha_1)\cdots(t-\alpha_\ell)$, $\Omega(t) = (t-\tau_1)^{2s_1+1}\cdots(t-\tau_n)^{2s_n+1}$, and $f_1(t) = f(t)/\omega(t)$, $f_2(t) = f(t)/\Omega(t)$. Since the remainder $r(t;f)$ of the interpolation formula $f(t) = L(t;f) + r(t;f)$ can be expressed as a divided difference,

$$r(t;f) = \Omega(t)\omega(t)\begin{bmatrix} \tau_1 & & \tau_n & \alpha_1 & & \alpha_\ell & t & \\ 2s_1+1 & ,\dots, & 2s_n+1, & 1 & ,\dots, & 1, & 1 & ;f \end{bmatrix}, \tag{3.1}$$

we obtain the quadrature formula

$$\int_{\mathbb{R}} f(t)\,\mathrm{d}\lambda(t) = Q(f) + \varphi(f) + \varrho(f), \tag{3.2}$$

where $Q(f)$ is the quadrature sum in (1.7), $\varrho(f) = \int_{\mathbb{R}} r(t;f)\,\mathrm{d}\lambda(t)$ and $\varphi(f)$ has the form $\varphi(f) = \sum_{\mu=1}^{\ell} B_\mu f(\alpha_\mu)$. Since the divided difference in (3.1) is of order $M + \ell = \sum_{\nu=1}^{n}(2s_\nu + 1) + \ell$, it follows that the quadrature formula (3.2) has degree of exactness $M + \ell - 1$.

For arbitrary $\alpha_1,\dots,\alpha_\ell$ it was proved [88] that it is possible to determine the nodes $\tau_1,\dots,\tau_n$ (with the m_ν given) so that $B_1 = \cdots = B_\ell = 0$. For this, the *necessary and sufficient condition* is that $\Omega(t)$ be *orthogonal* to $\mathscr{P}_{\ell-1}$ with respect to the measure $\mathrm{d}\lambda(t)$, i.e.,

$$\int_{\mathbb{R}} t^k \Omega(t)\,\mathrm{d}\lambda(t) = 0 \quad (k = 0, 1, \dots, \ell - 1). \tag{3.3}$$

If $\ell = n$, system (3.3) has at least one real solution consisting of the n distinct real nodes $\tau_1,\dots,\tau_n$. The case $\ell < n$ was considered by Stancu [85]. Stancu [81–86] also generalized the previous quadrature formulas using the quadrature sum with multiple Gaussian nodes τ_ν and multiple preassigned nodes α_μ in the form

$$Q(f) = \sum_{\nu=1}^{n}\sum_{i=0}^{m_\nu-1} A_{i,\nu} f^{(i)}(\tau_\nu) + \sum_{\mu=1}^{\ell}\sum_{j=0}^{k_\mu-1} B_{j,\mu} f^{(j)}(\alpha_\mu).$$

A particular case with simple Gaussian nodes and multiple fixed nodes was considered by Stancu and Stroud [87]. The existence and uniqueness of the previous quadratures exact for an extended complete Chebyshev (ETC) system were proved by Karlin and Pinkus [41,42] without using a variational principle. Barrow [2] gave a different proof using the topological degree of a mapping. On the other hand, Barrar et al. [1] obtained the results entirely via a variational principle. Namely, they considered the problem of finding the element of minimal L_p norm ($1 \leqslant p < +\infty$) from a family of generalized polynomials, where the multiplicities of the zeros are specified. As an application, they obtained Gaussian quadrature formulas exact for extended Chebyshev systems. The L_1 case was studied in [4,6] (see also [40]).

Using a result from [80], Stancu [84] determined the following expression for Cotes coefficients in (1.7):

$$A_{i,v} = \frac{1}{i!(2s_v - i)!}\left[\frac{1}{\Omega_v(t)}\int_{\mathbb{R}}\frac{\Omega(t)-\Omega(x)}{t-x}\,\mathrm{d}\lambda(x)\right]_{t=\tau_v}^{(2s_v-i)},$$

where $\Omega_v(t) = \Omega(t)/(t-\tau_v)^{2s_v+1}$. An alternative expression

$$A_{i,v} = \frac{1}{i!}\sum_{k=0}^{2s_v-i}\frac{1}{k!}\left[\frac{(t-\tau_v)^{2s_v+1}}{\Omega(t)}\right]_{t=\tau_v}^{(k)}\int_{\mathbb{R}}\frac{\Omega(t)}{(t-\tau_v)^{2s_v-i-k+1}}\,\mathrm{d}\lambda(t) \tag{3.4}$$

was obtained in [55].

Some properties of Cotes numbers in the Turán quadrature (1.3), as well as some inequalities related to zeros of s-orthogonal polynomials, were investigated by Ossicini and Rosati [68] (see also [46]).

The remainder term in formulas with multiple nodes was studied by Chakalov [9], Ionescu [39], Ossicini [63], Pavel [70–72]. For holomorphic functions f in the Turán quadrature (1.3) over a finite interval $[a,b]$, Ossicini and Rosati [65] found the contour integral representation

$$R_{n,2s}(f) = \frac{1}{2\pi \mathrm{i}}\oint_{\Gamma}\frac{\rho_{n,s}(z)}{[\pi_{n,s}(z)]^{2s+1}}f(z)\,\mathrm{d}z, \quad \rho_{n,s}(z) = \int_a^b\frac{[\pi_{n,s}(z)]^{2s+1}}{z-t}\,\mathrm{d}\lambda(t),$$

where $[a,b]\subset \operatorname{int}\Gamma$ and $\pi_{n,s} = \pi_{n,s}(\cdot;\mathrm{d}\lambda)$. Taking as Γ confocal ellipses (having foci at ± 1 and the sum of semiaxes equal to $\rho > 1$), Ossicini et al. [64] considered two special Chebyshev measures $\mathrm{d}\lambda_1(t)$ and $\mathrm{d}\lambda_2(t)$ (see Section 2.1) and determined estimates for the corresponding remainders $R_{n,2s}(f)$, from which they proved the convergence and rate of convergence of the quadratures, $R_{n,s}(f) = \mathrm{O}(\rho^{-n(2s+1)})$, $n\to+\infty$. Morelli and Verna [58] also investigated the convergence of quadrature formulas related to σ-orthogonal polynomials.

3.2. Numerical construction

A stable method for determining the coefficients $A_{i,v}$ in the Gauss–Turán quadrature formula (1.3) was given by Gautschi and Milovanović [22]. Some alternative methods were proposed by Stroud and Stancu [88] (see also [84]), Golub and Kautsky [28], and Milovanović and Spalević [54]. A generalization of the method from [22] to the general case when $s_v\in\mathbb{N}_0$ $(v=1,\ldots,n)$ was derived recently in [55]. Here, we briefly present the basic idea of this method.

First, we define as in the previous subsection $\Omega_v(t) = \prod_{i\neq v}(t-\tau_i)^{2s_i+1}$ and use the polynomials

$$f_{k,v}(t) = (t-\tau_v)^k\Omega_v(t) = (t-\tau_v)^k\prod_{i\neq v}(t-\tau_i)^{2s_i+1},$$

where $0\leqslant k\leqslant 2s_v$ and $1\leqslant v\leqslant n$. Notice that $\deg f_{k,v}\leqslant 2\sum_{i=1}^n s_i + n - 1$. This means that the integration (1.7) is exact for all polynomials $f_{k,v}$, i.e., $R(f_{k,v}) = 0$, when $0\leqslant k\leqslant 2s_v$ and $1\leqslant v\leqslant n$. Thus, we have

$$\sum_{j=1}^{n}\sum_{i=0}^{2s_j}A_{i,j}f_{k,v}^{(i)}(\tau_j) = \int_{\mathbb{R}}f_{k,v}(t)\,\mathrm{d}\lambda(t),$$

that is,

$$\sum_{i=0}^{2s_\nu} A_{i,\nu} f_{k,\nu}^{(i)}(\tau_\nu) = \mu_{k,\nu}, \tag{3.5}$$

because for every $j \neq \nu$ we have $f_{k,\nu}^{(i)}(\tau_j) = 0$ $(0 \leqslant i \leqslant 2s_j)$. Here, we have put

$$\mu_{k,\nu} = \int_{\mathbb{R}} f_{k,\nu}(t)\,\mathrm{d}\lambda(t) = \int_{\mathbb{R}} (t-\tau_\nu)^k \prod_{i \neq \nu} (t-\tau_i)^{2s_i+1}\,\mathrm{d}\lambda(t).$$

For each ν we have in (3.5) a system of $2s_\nu + 1$ linear equations in the same number of unknowns, $A_{i,\nu}$ $(i = 0, 1, \ldots, 2s_\nu)$. It can be shown that each system (3.5) is upper triangular. Thus, once all zeros of the σ-orthogonal polynomial $\pi_{n,\sigma}$, i.e., the nodes of the quadrature formula (1.7), are known, the determination of its weights $A_{i,\nu}$ is reduced to solving the n linear systems of $2s_\nu + 1$ equations

$$\begin{bmatrix} f_{0,\nu}(\tau_\nu) & f'_{0,\nu}(\tau_\nu) & \ldots & f_{0,\nu}^{(2s_\nu)}(\tau_\nu) \\ & f'_{1,\nu}(\tau_\nu) & \ldots & f_{1,\nu}^{(2s_\nu)}(\tau_\nu) \\ & & \ddots & \\ & & & f_{2s_\nu,\nu}^{(2s_\nu)}(\tau_\nu) \end{bmatrix} \begin{bmatrix} A_{0,\nu} \\ A_{1,\nu} \\ \vdots \\ A_{2s_\nu,\nu} \end{bmatrix} = \begin{bmatrix} \mu_{0,\nu} \\ \mu_{1,\nu} \\ \vdots \\ \mu_{2s_\nu,\nu} \end{bmatrix}.$$

Using these systems and the normalized moments

$$\hat{\mu}_{k,\nu} = \frac{\mu_{k,\nu}}{\prod_{i \neq \nu} (\tau_\nu - t_i)^{2s_i+1}} = \int_{\mathbb{R}} (t-\tau_\nu)^k \prod_{i \neq \nu} \left(\frac{t-\tau_i}{\tau_\nu - \tau_i} \right)^{2s_i+1} \mathrm{d}\lambda(t),$$

we can prove [55]

Theorem 3.1. *For fixed ν $(1 \leqslant \nu \leqslant n)$ the coefficients $A_{i,\nu}$ in the generalized Gauss–Turán quadrature formula* (1.7) *are given by*

$$b_{2s_\nu+1} = (2s_\nu)!\,A_{2s_\nu,\nu} = \hat{\mu}_{2s_\nu,\nu},$$

$$b_k = (k-1)!\,A_{k-1,\nu} = \hat{\mu}_{k-1,\nu} - \sum_{j=k+1}^{2s_\nu+1} \hat{a}_{k,j} b_j \quad (k = 2s_\nu, \ldots, 1),$$

where

$$\hat{a}_{k,k} = 1, \quad \hat{a}_{k,k+j} = -\frac{1}{j} \sum_{l=1}^{j} u_l \hat{a}_{l,j}, \quad u_l = \sum_{i \neq \nu} (2s_i + 1)(\tau_i - \tau_\nu)^{-l}.$$

The normalized moments $\hat{\mu}_{k,\nu}$ can be computed exactly, except for rounding errors, by using the same Gauss–Christoffel formula as in the construction of σ-orthogonal polynomials, i.e., (2.6) with $N = \sum_{\nu=1}^{n} s_\nu + n$ nodes. A few numerical examples can be found in [22,52,55]. Also, in [55] an alternative approach to the numerical calculation of the coefficients $A_{i,\nu}$ was given using expression (3.4).

4. Some remarks on the Chebyshev measure

From the remarks in Section 2 about s-orthogonal polynomials with Chebyshev measure, it is easy to see that the Chebyshev–Turán formula is given by

$$\int_{-1}^{1} \frac{f(t)}{\sqrt{1-t^2}}\,\mathrm{d}t = \sum_{i=0}^{2s}\sum_{v=1}^{n} A_{i,v} f^{(i)}(\tau_v) + R_n(f), \tag{4.1}$$

where $\tau_v = \cos((2v-1)\pi/2n)$ $(v=1,\ldots,n)$. It is exact for all polynomials of degree at most $2(s+1)n-1$. Turán stated the problem of explicit determination of the $A_{i,v}$ and their behavior as $n \to +\infty$ (see Problem XXVI in [91]). In this regard, Micchelli and Rivlin [49] proved the following characterization: *If* $f \in \mathscr{P}_{2(s+1)n-1}$ *then*

$$\int_{-1}^{1} \frac{f(t)}{\sqrt{1-t^2}}\,\mathrm{d}t = \frac{\pi}{n}\left\{\sum_{v=1}^{n} f(\tau_v) + \sum_{j=1}^{s} \alpha_j f'[\tau_1^{2j},\ldots,\tau_n^{2j}]\right\},$$

where

$$\alpha_j = \frac{(-1)^j}{2j4^{(n-1)j}}\begin{pmatrix}-1/2\\ j\end{pmatrix} \quad (j=1,2,\ldots)$$

and $g[y_1^r,\ldots,y_m^r]$ *denotes the divided difference of the function* g, *where each* y_j *is repeated* r *times*. In fact, they obtained a quadrature formula of highest algebraic degree of precision for the Fourier–Chebyshev coefficients of a given function f, which is based on the divided differences of f' at the zeros of the Chebyshev polynomial T_n. A Lobatto type of Turán quadrature was considered by Micchelli and Sharma [50]. Recently, Bojanov [5] has given a simple approach to questions of the previous type and applied it to the coefficients in arbitrary orthogonal expansions of f. As an auxiliary result he obtained a new interpolation formula and a new representation of the Turán quadrature formula. Some further results can be found in [79].

For $s=1$, the solution of the Turán problem XXVI is given by

$$A_{0,v} = \frac{\pi}{n}, \quad A_{1,v} = -\frac{\pi\tau_v}{4n^3}, \quad A_{2,v} = \frac{\pi}{4n^3}(1-\tau_v^2).$$

In 1975 Riess [75], and in 1984 Varma [92], using very different methods, obtained the explicit solution of the Turán problem for $s=2$. One simple answer to Turán's question was given by Kis [43]. His result can be stated in the following form: *If* g *is an even trigonometric polynomial of degree at most* $2(s+1)n-1$, *then*

$$\int_0^{\pi} g(\theta)\,\mathrm{d}\theta = \frac{\pi}{n(s!)^2}\sum_{j=0}^{s}\frac{S_j}{4^j n^{2j}}\sum_{v=1}^{n} g^{(2j)}\left(\frac{2v-1}{2n}\pi\right),$$

where the S_{s-j} $(j=0,1,\ldots,s)$ *denote the elementary symmetric polynomials with respect to the numbers* $1^2, 2^2, \ldots, s^2$, i.e., $S_s = 1$, $S_{s-1} = 1^2+2^2+\cdots+s^2, \ldots, S_0 = 1^2\cdot 2^2\cdots s^2$. Consequently,

$$\int_{-1}^{1} \frac{f(t)}{\sqrt{1-t^2}}\,\mathrm{d}t = \frac{\pi}{n(s!)^2}\sum_{j=0}^{s}\frac{S_j}{4^j n^{2j}}\sum_{v=1}^{n}[\mathrm{D}^{2j} f(\cos\theta)]_{\theta=((2v-1)/2n)\pi}.$$

An explicit expression for the coefficients $A_{i,v}$ was recently derived by Shi [76]. The remainder $R_n(f)$ in (4.1) was studied by Pavel [70].

5. Some remarks on moment-preserving spline approximation

Solving some problems in computational plasma physics, Calder and Laframboise [7] considered the problem of approximating the Maxwell velocity distribution by a step function, i.e., by a "multiple-water-bag distribution" in their terminology, in such a way that as many of the initial moments as possible of the Maxwell distribution are preserved. They used a classical method of reduction to an eigenvalue problem for Hankel matrices, requiring high-precision calculations because of numerical instability. A similar problem, involving Dirac's δ-function instead of Heaviside's step function, was treated earlier by Laframboise and Stauffer [45], using the classical Prony's method. A stable procedure for these problems was given by Gautschi [17] (see also [19]), who found the close connection of these problems with Gaussian quadratures. This work was extended to spline approximation of arbitrary degree by Gautschi and Milovanović [21]. In this case, a spline $s_{n,m}$ of degree m with n knots is sought so as to faithfully reproduce the first $2n$ moments of a given function f. Under suitable assumptions on f, it was shown that the problem has a unique solution if and only if certain Gauss–Christoffel quadratures exist that correspond to a moment functional or weight distribution depending on f. Existence, uniqueness, and pointwise convergence of such approximations were analyzed. Frontini et al. [13] and Frontini and Milovanović [14] considered analogous problems on an arbitrary finite interval. If the approximations exist, they can be represented in terms of generalized Gauss–Lobatto and Gauss–Radau quadrature formulas relative to appropriate measures depending on f.

At the Singapore Conference on Numerical Mathematics (1988) we presented a moment-preserving approximation on $[0,+\infty)$ by defective splines of degree m, with odd defect (see [53]).

A spline function of degree $m \geqslant 1$ on the interval $0 \leqslant t < +\infty$, vanishing at $t = +\infty$, with variable positive knots τ_ν $(\nu = 1,\ldots,n)$ having multiplicities m_ν $(\leqslant m)$ $(\nu = 1,\ldots,n;\ n > 1)$ can be represented in the form

$$S_{n,m}(t) = \sum_{\nu=1}^{n} \sum_{i=0}^{m_\nu - 1} \alpha_{\nu,i}(\tau_\nu - t)_+^{m-i} \quad (0 \leqslant t < +\infty), \tag{5.1}$$

where $\alpha_{\nu,i}$ are real numbers. Under the conditions

$$\int_0^{+\infty} t^{j+d-1} S_{n,m}(t)\,\mathrm{d}t = \int_0^{+\infty} t^{j+d-1} f(t)\,\mathrm{d}t \quad (j = 0,1,\ldots,2(s+1)n-1)$$

in [53] we considered the problem of approximating a function $f(t)$ of the radial distance $t = \|x\|$ $(0 \leqslant t < +\infty)$ in $\mathbb{R}^d$ $(d \geqslant 1)$ by the spline function (5.1), where $m_\nu = 2s+1$ $(\nu = 1,\ldots,n;\ s \in \mathbb{N}_0)$. Under suitable assumptions on f, we showed that the problem has a unique solution if and only if certain generalized Turán quadratures exist corresponding to a measure depending on f. A more general case with variable defects was considered by Gori and Santi [34] and Kovačević and Milovanović [44] (see also [52]). In that case, the approximation problems reduce to quadratures of form (1.7) and σ-orthogonal polynomials.

Following [44], we discuss here two problems of approximating a function $f(t)$, $0 \leqslant t < +\infty$, by the defective spline function (5.1). Let N denote the number of the variable knots τ_ν $(\nu = 1,\ldots,n)$ of the spline function $S_{n,m}(t)$, counting multiplicities, i.e., $N = m_1 + \cdots + m_n$.

Problem 5.1. *Determine* $S_{n,m}$ *in* (5.1) *such that* $S_{n,m}^{(k)}(0)=f^{(k)}(0)$ $(k=0,1,\ldots,N+n-1;\ m\geqslant N+n-1)$.

Problem 5.2. *Determine* $S_{n,m}$ in (5.1) *such that* $S_{n,m}^{(k)}(0)=f^{(k)}(0)$ $(k=0,1,\ldots,l;\ l\leqslant m)$ *and*

$$\int_0^{+\infty} t^j S_{n,m}(t)\,\mathrm{d}t=\int_0^{+\infty} t^j f(t)\,\mathrm{d}t \quad (j=0,1,\ldots,N+n-l-2).$$

The next theorem gives the solution of Problem 5.2.

Theorem 5.3. *Let* $f\in C^{m+1}[0,+\infty)$ *and* $\int_0^{+\infty} t^{N+n-l+m}|f^{(m+1)}(t)|\,\mathrm{d}t<+\infty$. *Then a spline function* $S_{n,m}$ *of form* (5.2), *with positive knots* τ_ν, *that satisfies the conditions of Problem* 5.2 *exists and is unique if and only if the measure*

$$\mathrm{d}\lambda(t)=\frac{(-1)^{m+1}}{m!}\,t^{m-l}f^{(m+1)}(t)\,\mathrm{d}t$$

admits a generalized Gauss–Turán quadrature

$$\int_0^{+\infty} g(t)\,\mathrm{d}\lambda(t)=\sum_{\nu=1}^{n}\sum_{k=0}^{m_\nu-1}A_{\nu,k}^{(n)}g^{(k)}(\tau_\nu^{(n)})+R_n(g;\mathrm{d}\lambda) \tag{5.2}$$

with n distinct positive nodes $\tau_\nu^{(n)}$, *where* $R_n(g;d\lambda)=0$ *for all* $g\in\mathscr{P}_{N+n-1}$. *The knots in* (5.1) *are given by* $\tau_\nu=\tau_\nu^{(n)}$, *and the coefficients* $\alpha_{\nu,i}$ *by the following triangular system*:

$$A_{\nu,k}^{(n)}=\sum_{i=k}^{m_\nu-i}\frac{(m-i)!}{m!}\binom{i}{k}[\mathrm{D}^{i-k}t^{m-l}]_{t=\tau_\nu}\alpha_{\nu,i} \quad (k=0,1,\ldots,m_\nu-1).$$

If we let $l=N+n-1$, this theorem gives also the solution of Problem 5.1. The case $m_1=m_2=\cdots=m_n=1$, $l=-1$, has been obtained by Gautschi and Milovanović [21]. The error of the spline approximation can be expressed as the remainder term in (5.2) for a particular function $\sigma_t(x)=x^{-(m-l)}(x-t)_+^m$ (see [44]).

Further extensions of the moment-preserving spline approximation on $[0,1]$ are given by Micchelli [48]. He relates this approximation to the theory of monosplines. A similar problem by defective spline functions on the finite interval $[0,1]$ has been studied by Gori and Santi [35] and solved by means of monosplines.

Acknowledgements

The author is grateful to Professor Walter Gautschi for helpful comments and to the referee for pointing out that Theorem 2.2 can be shown by applying the results from [73].

References

[1] R.B. Barrar, B.D. Bojanov, H.L. Loeb, Generalized polynomials of minimal norm, J. Approx. Theory 56 (1989) 91–100.

[2] D.L. Barrow, On multiple node Gaussian quadrature formulae, Math. Comp. 32 (1978) 431–439.

[3] S. Bernstein, Sur les polynomes orthogonaux relatifs à un segment fini, J. Math. Pure Appl. 9 (9) (1930) 127–177.

[4] B.D. Bojanov, Oscillating polynomials of least L_1-norm, in: G. Hammerlin (Ed.), Numerical Integration, International Series of Numerical Mathematics, Vol. 57, 1982, pp. 25–33.

[5] B. Bojanov, On a quadrature formula of Micchelli and Rivlin, J. Comput. Appl. Math. 70 (1996) 349–356.

[6] B.D. Bojanov, D. Braess, N. Dyn, Generalized Gaussian quadrature formulas, J. Approx. Theory 48 (1986) 335–353.

[7] A.C. Calder, J.G. Laframboise, Multiple-waterbag simulation of inhomogeneous plasma motion near an electrode, J. Comput. Phys. 65 (1986) 18–45.

[8] L. Chakalov, Über eine allgemeine Quadraturformel, C.R. Acad. Bulgar. Sci. 1 (1948) 9–12.

[9] L. Chakalov, General quadrature formulae of Gaussian type, Bulgar. Akad. Nauk Izv. Mat. Inst. 1 (1954) 67–84 (Bulgarian) [English transl. East J. Approx. 1 (1995) 261–276].

[10] L. Chakalov, Formules générales de quadrature mécanique du type de Gauss, Colloq. Math. 5 (1957) 69–73.

[11] T.S. Chihara, An Introduction to Orthogonal Polynomials, Gordon and Breach, New York, 1978.

[12] H. Engels, Numerical Quadrature and Cubature, Academic Press, London, 1980.

[13] M. Frontini, W. Gautschi, G.V. Milovanović, Moment-preserving spline approximation on finite intervals, Numer. Math. 50 (1987) 503–518.

[14] M. Frontini, G.V. Milovanović, Moment-preserving spline approximation on finite intervals and Turán quadratures, Facta Univ. Ser. Math. Inform. 4 (1989) 45–56.

[15] W. Gautschi, A survey of Gauss–Christoffel quadrature formulae, in: P.L. Butzer, F. Fehér (Eds.), E.B. Christoffel: The Influence of his Work in Mathematics and the Physical Sciences, Birkhäuser, Basel, 1981, pp. 72–147.

[16] W. Gautschi, On generating orthogonal polynomials, SIAM, J. Sci. Statist. Comput. 3 (1982) 289–317.

[17] W. Gautschi, Discrete approximations to spherically symmetric distributions, Numer. Math. 44 (1984) 53–60.

[18] W. Gautschi, On the remainder term for analytic functions of Gauss–Lobatto and Gauss–Radau quadratures, Rocky Mountain J. Math. 21 (1991) 209–226.

[19] W. Gautschi, Spline approximation and quadrature formulae, Atti Sem. Mat. Fis. Univ. Modena 40 (1992) 169–182.

[20] W. Gautschi, G.V. Milovanović, Gaussian quadrature involving Einstein and Fermi functions with an application to summation of series, Math. Comp. 44 (1985) 177–190.

[21] W. Gautschi, G.V. Milovanović, Spline approximations to spherically symmetric distributions, Numer. Math. 49 (1986) 111–121.

[22] W. Gautschi, G.V. Milovanović, s-orthogonality and construction of Gauss–Turán-type quadrature formulae, J. Comput. Appl. Math. 86 (1997) 205–218.

[23] A. Ghizzetti, A. Ossicini, Su un nuovo tipo di sviluppo di una funzione in serie di polinomi, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (8) 43 (1967) 21–29.

[24] A. Ghizzetti, A. Ossicini, Quadrature Formulae, Akademie Verlag, Berlin, 1970.

[25] A. Ghizzetti, A. Ossicini, Polinomi s-ortogonali e sviluppi in serie ad essi collegati, Mem. Accad. Sci. Torino Cl. Sci. Fis. Mat. Natur. (4) 18 (1974) 1–16.

[26] A. Ghizzetti, A. Ossicini, Generalizzazione dei polinomi s-ortogonali e dei corrispondenti sviluppi in serie, Atti Accad. Sci. Torino Cl. Sci. Fis. Mat. Natur. 109 (1975) 371–379.

[27] A. Ghizzetti, A. Ossicini, Sull' esistenza e unicità delle formule di quadratura gaussiane, Rend. Mat. (6) 8 (1975) 1–15.

[28] G. Golub, J. Kautsky, Calculation of Gauss quadratures with multiple free and fixed knots, Numer. Math. 41 (1983) 147–163.

[29] L. Gori Nicolo-Amati, On the behaviour of the zeros of some s-orthogonal polynomials, in: Orthogonal Polynomials and Their Applications, Second International Symposium, Segovia, 1986, Monogr. Acad. Cienc. Exactas, Fis., Quim., Nat., Zaragoza, 1988, pp. 71–85.

[30] L. Gori, M.L. Lo Cascio, On an invariance property of the zeros of some s-orthogonal polynomials, in: C. Brezinski, L. Gori, A. Ronveaux (Eds.), Orthogonal Polynomials and Their Applications, IMACS Ann. Comput. Appl. Math., Vol. 9, J.C. Baltzer AG, Scientific Publ. Co, Basel, 1991, pp. 277–280.

[31] L. Gori, M.L. Lo Cascio, A note on a class of Turán type quadrature formulas with generalized Gegenbauer weight functions, Studia Univ. Babeş-Bolyai Math. 37 (1992) 47–63.

[32] L. Gori, M.L. Lo Cascio, G.V. Milovanović, The σ-orthogonal polynomials: a method of construction, in: C. Brezinski, L. Gori, A. Ronveaux (Eds.), Orthogonal Polynomials and Their Applications, IMACS Ann. Comput. Appl. Math., Vol. 9, J.C. Baltzer AG, Scientific Publ. Co, Basel, 1991, pp. 281–285.

[33] L. Gori, C.A. Micchelli, On weight functions which admit explicit Gauss–Turán quadrature formulas, Math. Comp. 65 (1996) 1567–1581.

[34] L. Gori Nicolo-Amati, E. Santi, On a method of approximation by means of spline functions, in: A.G. Law, C.L. Wang (Eds.), Approximation, Optimization and Computing — Theory and Application, North-Holland, Amsterdam, 1990, pp. 41–46.

[35] L. Gori, E. Santi, Moment-preserving approximations: a monospline approach, Rend. Mat. (7) 12 (1992) 1031–1044.

[36] L. Gori, E. Santi, On the evaluation of Hilbert transforms by means of a particular class of Turán quadrature rules, Numer. Algorithms 10 (1995) 27–39.

[37] S. Guerra, Polinomi generati da successioni peso e teoremi di rappresentazione di una funzione in serie di tali polinomi, Rend. Ist. Mat. Univ. Trieste 8 (1976) 172–194.

[38] S. Guerra, Su un determinante collegato ad un sistema di polinomi ortogonali, Rend. Ist. Mat. Univ. Trieste 10 (1978) 66–79.

[39] D.V. Ionescu, Restes des formules de quadrature de Gauss et de Turán, Acta Math. Acad. Sci. Hungar. 18 (1967) 283–295.

[40] D.L. Johnson, Gaussian quadrature formulae with fixed nodes, J. Approx. Theory 53 (1988) 239–250.

[41] S. Karlin, A. Pinkus, Gaussian quadrature formulae with multiple nodes, in: S. Karlin, C.A. Micchelli, A. Pinkus, I.J. Schoenberg (Eds.), Studies in Spline Functions and Approximation Theory, Academic Press, New York, 1976, pp. 113–141.

[42] S. Karlin, A. Pinkus, An extremal property of multiple Gaussian nodes, in: S. Karlin, C.A. Micchelli, A. Pinkus, I.J. Schoenberg (Eds.), Studies in Spline Functions and Approximation Theory, Academic Press, New York, 1976, pp. 143–162.

[43] O. Kis, Remark on mechanical quadrature, Acta Math. Acad. Sci. Hungar. 8 (1957) 473–476 (in Russian).

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

[45] J.G. Laframboise, A.D. Stauffer, Optimum discrete approximation of the Maxwell distribution, AIAA J. 7 (1969) 520–523.

[46] M.R. Martinelli, A. Ossicini, F. Rosati, Density of the zeros of a system of s-orthogonal polynomials, Rend. Mat. Appl. (7) 12 (1992) 235–249 (in Italian).

[47] C.A. Micchelli, The fundamental theorem of algebra for monosplines with multiplicities, in: P. Butzer, J.P. Kahane, B.Sz. Nagy (Eds.), Linear Operators and Approximation, International Series of Numerical Mathematics, Vol. 20, Birkhäuser, Basel, 1972, pp. 372–379.

[48] C.A. Micchelli, Monosplines and moment preserving spline approximation, in: H. Brass, G. Hämmerlin (Eds.), Numerical Integration III, Birkhäuser, Basel, 1988, pp. 130–139.

[49] C.A. Micchelli, T.J. Rivlin, Turán formulae and highest precision quadrature rules for Chebyshev coefficients, IBM J. Res. Develop. 16 (1972) 372–379.

[50] C.A. Micchelli, A. Sharma, On a problem of Turán: multiple node Gaussian quadrature, Rend. Mat. (7) 3 (1983) 529–552.

[51] G.V. Milovanović, Construction of s-orthogonal polynomials and Turán quadrature formulae, in: G.V. Milovanović (Ed.), Numerical Methods and Approximation Theory III, Niš, 1987, Univ. Niš, Niš, 1988, pp. 311–328.

[52] G.V. Milovanović, S-orthogonality and generalized Turán quadratures: construction and applications, in: D.D. Stancu, Ch. Coman, W.W. Breckner, P. Blaga (Eds.), Approximation and Optimization, Vol. I, Cluj-Napoca, 1996, Transilvania Press, Cluj-Napoca, Romania, 1997, pp. 91–106.

[53] G.V. Milovanović, M.A. Kovačević, Moment-preserving spline approximation and Turán quadratures, in: R.P. Agarwal, Y.M. Chow, S.J. Wilson (Eds.), Numerical Mathematics Singapore, 1988, International Series of Numerical Mathematics, Vol. 86, Birkhäuser, Basel, 1988, pp. 357–365.

[54] G.V. Milovanović, M.M. Spalević, A numerical procedure for coefficients in generalized Gauss–Turán quadratures, FILOMAT (formerly Zb. Rad.) 9 (1995) 1–8.

[55] G.V. Milovanović, M.M. Spalević, Construction of Chakalov–Popoviciu's type quadrature formulae, Rend. Circ. Mat. Palermo (2) II (Suppl. 52) (1998) 625–636.

[56] G.V. Milovanović, M.M. Spalević, On the computation of the zeros of σ-orthogonal polynomials, in preparation.

[57] A. Morelli, I. Verna, 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 (1969) 91–98.

[58] A. Morelli, I. Verna, The convergence of quadrature formulas associated with σ-orthogonal polynomials, Note Mat. 6 (1986) 35–48 (Italian).

[59] A. Morelli, I. Verna, Some properties of the zeros of σ-orthogonal polynomials, Rend. Mat. Appl. (7) 7 (1987) 43–52 (in Italian).

[60] A. Morelli, I. Verna, Some observations on asymptotic behaviour of zeros of particular sequences of σ-orthogonal polynomials, Rend. Mat. Appl. (7) Ser. 11 (3) (1991) 417–424 (in Italian).

[61] A. Morelli, I. Verna, A particular asymptotic behaviour of the zeros of σ-orthogonal polynomials, Jñānābha 22 (1992) 31–40.

[62] A. Ossicini, Costruzione di formule di quadratura di tipo Gaussiano, Ann. Mat. Pura Appl. (4) 72 (1966) 213–237.

[63] A. Ossicini, Le funzioni di influenza nel problema di Gauss sulle formule di quadratura, Matematiche (Catania) 23 (1968) 7–30.

[64] A. Ossicini, M.R. Martinelli, F. Rosati, Characteristic functions and s-orthogonal polynomials, Rend. Mat. Appl. (7) 14 (1994) 355–366 (in Italian).

[65] A. Ossicini, F. Rosati, Funzioni caratteristiche nelle formule di quadratura gaussiane con nodi multipli, Boll. Un. Mat. Ital. (4) 11 (1975) 224–237.

[66] A. Ossicini, F. Rosati, Sulla convergenza dei funzionali ipergaussiani, Rend. Mat. (6) 11 (1978) 97–108.

[67] A. Ossicini, F. Rosati, Comparison theorems for the zeros of s-orthogonal polynomials, Calcolo 16 (1979) 371–381 (in Italian).

[68] A. Ossicini, F. Rosati, Numeri di Christoffel e polinomi s-ortogonali, in: P.L. Butzer, F. Fehér (Eds.), E.B. Christoffel, Birkhäuser, Basel, 1981, pp. 148–157.

[69] A. Ossicini, F. Rosati, s-orthogonal Jacobi polynomials, Rend. Mat. Appl. (7) 12 (1992) 399–403 (in Italian).

[70] P. Pavel, On the remainder of some Gaussian formulae, Studia Univ. Babeş-Bolyai Ser. Math.-Phys. 12 (1967) 65–70.

[71] P. Pavel, On some quadrature formulae of Gaussian type, Studia Univ. Babeş-Bolyai Ser. Math.-Phys. 13 (1968) 51–58 (in Romanian).

[72] P. Pavel, On the remainder of certain quadrature formulae of Gauss–Christoffel type, Studia Univ. Babeş-Bolyai Ser. Math.-Phys. 13 (1968) 67–72 (in Romanian).

[73] A. Pinkus, Z. Ziegler, Interlacing properties of the zeros of the error functions in best L^p-approximations, J. Approx. Theory 27 (1979) 1–18.

[74] T. Popoviciu, Sur une généralisation de la formule d'intégration numérique de Gauss, Acad. R.P. Romîne Fil. Iaşi Stud. Cerc. Şti. 6 (1955) 29–57 (in Romanian).

[75] R.D. Riess, Gauss–Turán quadratures of Chebyshev type and error formulae, Computing 15 (1975) 173–179.

[76] Y.G. Shi, A solution of problem 26 of P. Turán, Sci. China, Ser. A 38 (1995) 1313–1319.

[77] Y.G. Shi, On Turán quadrature formulas for the Chebyshev weight, J. Approx. Theory 96 (1999) 101–110.

[78] Y.G. Shi, On Gaussian quadrature formulas for the Chebyshev weight, J. Approx. Theory 98 (1999) 183–195.

[79] Y.G. Shi, On some problems of P. Turán concerning L_m extremal polynomials and quadrature formulas, J. Approx. Theory 100 (1999) 203–220.

[80] D.D. Stancu, On the interpolation formula of Hermite and some applications of it, Acad. R.P. Romîne Fil. Cluj Stud. Cerc. Mat. 8 (1957) 339–355 (in Romanian).

[81] D.D. Stancu, Generalization of the quadrature formula of Gauss–Christoffel, Acad. R.P. Romîne Fil. Iaşi Stud. Cerc. Şti. Mat. 8 (1957) 1–18 (in Romanian).

[82] D.D. Stancu, On a class of orthogonal polynomials and on some general quadrature formulas with minimum number of terms, Bull. Math. Soc. Sci. Math. Phys. R.P. Romîne (N.S) 1 (49) (1957) 479–498.

[83] D.D. Stancu, A method for constructing quadrature formulas of higher degree of exactness, Com. Acad. R.P. Romîne 8 (1958) 349–358 (in Romanian).

[84] D.D. Stancu, On certain general numerical integration formulas, Acad. R.P. Romîne. Stud. Cerc. Mat. 9 (1958) 209–216 (in Romanian).

[85] D.D. Stancu, Sur quelques formules générales de quadrature du type Gauss–Christoffel, Mathematica (Cluj) 1 (24) (1959) 167–182.

[86] D.D. Stancu, An extremal problem in the theory of numerical quadratures with multiple nodes, Proceedings of the Third Colloquium on Operations Research Cluj-Napoca, 1978, Univ. "Babeş-Bolyai", Cluj-Napoca, 1979, pp. 257–262.

[87] D.D. Stancu, A.H. Stroud, Quadrature formulas with simple Gaussian nodes and multiple fixed nodes, Math. Comp. 17 (1963) 384–394.

[88] A.H. Stroud, D.D. Stancu, Quadrature formulas with multiple Gaussian nodes, J. SIAM Numer. Anal. Ser. B 2 (1965) 129–143.

[89] G. Szegő, Orthogonal Polynomials, American Mathematical Society Colloquium Publications, Vol. 23, 4th Edition American Mathematical Society, Providence, RI, 1975.

[90] P. Turán, On the theory of the mechanical quadrature, Acta Sci. Math. Szeged 12 (1950) 30–37.

[91] P. Turán, On some open problems of approximation theory, J. Approx. Theory 29 (1980) 23–85.

[92] A.K. Varma, On optimal quadrature formulae, Studia Sci. Math. Hungar. 19 (1984) 437–446.

[93] G. Vincenti, On the computation of the coefficients of s-orthogonal polynomials, SIAM J. Numer. Anal. 23 (1986) 1290–1294.

ELSEVIER

Journal of Computational and Applied Mathematics 127 (2001) 287–296

JOURNAL OF
COMPUTATIONAL AND
APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

The double-exponential transformation in numerical analysis [☆]

Masatake Mori[a, *], Masaaki Sugihara[b]

[a] *Research Institute for Mathematical Sciences, Kyoto University, Kyoto 606-8502, Japan*
[b] *Graduate School of Engineering, Nagoya University, Chikusa-ku, Nagoya 464-8603, Japan*

Received 26 May 1999; received in revised form 15 October 1999

Abstract

The double-exponential transformation was first proposed by Takahasi and Mori in 1974 for the efficient evaluation of integrals of an analytic function with end-point singularity. Afterwards, this transformation was improved for the evaluation of oscillatory functions like Fourier integrals. Recently, it turned out that the double-exponential transformation is useful not only for numerical integration but also for various kinds of Sinc numerical methods. The purpose of the present paper is to review the double-exponential transformation in numerical integration and in a variety of Sinc numerical methods. © 2001 Elsevier Science B.V. All rights reserved.

MSC: 65D30; 65D32

Keywords: Numerical integration; Quadrature formula; Double-exponential transformation; Sinc method; Fourier-type integral

1. Numerical integration and the double-exponential transformation

The double-exponential transformation was first proposed by Takahasi and Mori in 1974 in order to compute the integrals with end-point singularity such as

$$I = \int_{-1}^{1} \frac{\mathrm{d}x}{(2-x)(1-x)^{1/4}(1+x)^{3/4}} \tag{1.1}$$

with high efficiency [7,10,11,30].

[☆] Partially supported by a Grant-in Aid for Scientific Research from the Ministry of Education, Sports, Culture and Science, and by the Japan Society for Promotion of Science.

[*] Correspondence address. Research Institute for Mathematical Sciences, Kyoto University, Sakyo-ku, Kyoto 606-8502, Japan.

PII: S0377-0427(00)00501-X

The double-exponential formula for numerical integration based on this transformation can be derived in the following way. Let the integral under consideration be

$$I=\int_a^b f(x)\,\mathrm{d}x. \tag{1.2}$$

The interval (a,b) of integration may be finite, half-infinite $(0,\infty)$ or infinite $(-\infty,\infty)$. The integrand $f(x)$ must be analytic on the interval (a,b) but may have a singularity at the end point $x=a$ or b or both.

Now, we apply a variable transformation

$$x=\phi(t),\quad a=\phi(-\infty),\quad b=\phi(\infty), \tag{1.3}$$

where $\phi(t)$ is analytic on $(-\infty,\infty)$, and have

$$I=\int_{-\infty}^{\infty} f(\phi(t))\phi'(t)\,\mathrm{d}t. \tag{1.4}$$

A crucial point is that we should employ a function $\phi(t)$ such that after the transformation the decay of the integrand be double exponential, i.e.,

$$|f(\phi(t))\phi'(t)|\approx\exp(-c\exp|t|),\quad |t|\to\infty. \tag{1.5}$$

On the other hand, it is known that, for an integral like (1.4) of an analytic function over $(-\infty,\infty)$, the trapezoidal formula with an equal mesh size gives an optimal formula [5,28]. Accordingly, we apply the trapezoidal formula with an equal mesh size h to (1.4), which gives

$$I_h=h\sum_{k=-\infty}^{\infty} f(\phi(kh))\phi'(kh). \tag{1.6}$$

In actual computation of (1.6) we truncate the infinite summation at $k=-N_-$ and $k=N_+$ and obtain

$$I_h^{(N)}=h\sum_{k=-N_-}^{N_+} f(\phi(kh))\phi'(kh),\quad N=N_++N_-+1, \tag{1.7}$$

where N is the number of function evaluations. Since the integrand after the transformation decays double exponentially like (1.5), we call the formula obtained in this way the double-exponential formula, abbreviated as the DE formula.

For the integral over $(-1,1)$

$$I=\int_{-1}^{1} f(x)\,\mathrm{d}x \tag{1.8}$$

the transformation

$$x=\phi(t)=\tanh\left(\frac{\pi}{2}\sinh t\right) \tag{1.9}$$

will give a double-exponential formula

$$I_h^{(N)}=h\sum_{k=-N_-}^{N_+} f\left(\tanh\left(\frac{\pi}{2}\sinh kh\right)\right)\frac{\pi/2\cosh kh}{\cosh^2(\pi/2\sinh kh)}. \tag{1.10}$$

The double-exponential formula is designed so that it gives the most accurate result by the minimum number of function evaluations. In this sense, we call it an optimal formula [30]. For example, in

Table 1
Comparison of the efficiency of DEFINT and DQAGS. The absolute error tolerance is 10^{-8}. N is the number of function evaluations and abs.error is the actual absolute error of the result

	DEFINT		DQAGS	
Integral	N	abs.error	N	abs.error
I_1	25	$2.0 \cdot 10^{-11}$	315	$6.7 \cdot 10^{-16}$
I_2	387	$4.8 \cdot 10^{-14}$	189	$3.1 \cdot 10^{-15}$
I_3	387	$1.0 \cdot 10^{-13}$	567	$1.1 \cdot 10^{-13}$
I_4	259	$4.8 \cdot 10^{-12}$	651	$1.4 \cdot 10^{-17}$

case of (1.1), it gives an approximate value which is correct up to 16 significant digits by only about $N = 50$ function evaluations.

The merits of the double-exponential formula are as follows.

First, if we write the error of (1.6) in terms of the mesh size h of the trapezoidal formula, we have [30]

$$|\Delta I_h| = |I - I_h| \approx \exp\left(-\frac{c_1}{h}\right). \tag{1.11}$$

From this we see that the error converges to 0 very quickly as the mesh size h becomes small. On the other hand, if we write the error in terms of the number N of function evaluations, we have [30]

$$|\Delta I_h^{(N)}| = |I - I_h^{(N)}| \approx \exp\left(-c_2 \frac{N}{\log N}\right). \tag{1.12}$$

A single-exponential transformation

$$x = \tanh t \tag{1.13}$$

for the integral over $(-1, 1)$ will give [30]

$$|\Delta I_h^{(N)}| \approx \exp(-c_3\sqrt{N}). \tag{1.14}$$

We can see that as N becomes large, (1.12) converges to 0 much more quickly than (1.14).

Second, if the integrand has a singularity at the end point like (1.1), it will be mapped onto infinity. On the other hand, the integrand after the transformation decays double exponentially toward infinity, and hence we can truncate the infinite summation at a moderate value of k in (1.6). In addition, we can evaluate integrals with different orders of singularity using the same formula (1.7). In that sense, we can say that the double-exponential formula is robust with regard to singularities.

Third, since the base formula is the trapezoidal formula with an equal mesh size, we can make use of the result of the previous step with the mesh size h when we improve the value by halving the mesh size to $h/2$. Therefore, the present formula is suitable for constructing an automatic integrator. In addition, the points $\phi(kh)$ and the weights $h\phi'(kh)$ can easily be computed as seen in (1.10).

In Table 1 we show numerical examples to compare the efficiency of an automatic integrator DEFINT in [8] based on the DE transformation (1.9) and DQAGS in QUADPACK [24] for the following four integrals:

$$I_1 = \int_0^1 x^{-1/4} \log(1/x)\,\mathrm{d}x, \qquad I_2 = \int_0^1 \frac{1}{16(x - \pi/4)^2 + 1/16}\,\mathrm{d}x,$$

$$I_3 = \int_0^{\pi} \cos(64 \sin x)\, \mathrm{d}x, \qquad I_4 = \int_0^1 \exp(20(x-1)) \sin(256x)\, \mathrm{d}x.$$

From Table 1 we see that DEFINT is more efficient than DQAGS for I_1, I_3 and I_4, but that it is less efficient for I_2 because its integrand has a sharp peak at $x = \pi/4$ and DEFINT regards it as almost not analytic at this point.

The double-exponential transformation can be applied not only to an integral with end-point singularity over a finite interval, but also to other kinds of integrals such as an integral over a half-infinite interval [9]. Some useful transformations for typical types of integrals are listed below:

$$I = \int_{-1}^{1} f(x)\, \mathrm{d}x \Rightarrow x = \tanh\left(\frac{\pi}{2} \sinh t\right), \tag{1.15}$$

$$I = \int_0^{\infty} f(x)\, \mathrm{d}x \Rightarrow x = \exp\left(\frac{\pi}{2} \sinh t\right), \tag{1.16}$$

$$I = \int_0^{\infty} f_1(x) \exp(-x)\, \mathrm{d}x \Rightarrow x = \exp(t - \exp(-t)), \tag{1.17}$$

$$I = \int_{-\infty}^{\infty} f(x)\, \mathrm{d}x \Rightarrow x = \sinh\left(\frac{\pi}{2} \sinh t\right). \tag{1.18}$$

Before the double-exponential formula was proposed, a formula by Iri et al. [2], abbreviated as IMT formula, based on the transformation which maps $(-1, 1)$ onto itself had been known. This transformation gave a significant hint for the discovery of the double-exponential formula [29]. However, the error of the IMT formula behaves as $\exp(-c\sqrt{N})$, which is equivalent to the behavior of a formula based on the single-exponential transformation like (1.13). Also, there have been some attempts to improve the efficiency of the IMT formula [6,12].

In a mathematically more rigorous manner, the optimality of the double-exponential formula is established by Sugihara [26]. His approach is functional analytic. The basis of this approach is due to Stenger, which is described in full detail in his book [25]. Stenger there considers the integral $\int_{-\infty}^{\infty} g(w)\, \mathrm{d}w$ as the complex integral along the real axis in the w-plane and supposes that the integrand $g(w)$ is analytic and bounded in the strip region $|\mathrm{Im}\, w| < d$ of the w-plane. Under some additional conditions, he proves that the trapezoidal rule with an equal mesh size is optimal for the integral of a function which decays single exponentially as $w \to \pm\infty$ along the real axis. He also shows that the error behaves like (1.14). Sugihara proceeds analogously. He supposes that the integrand $g(w)$ be analytic and bounded in the strip region $|\mathrm{Im}\, w| < d$ of the w-plane, and proves the optimality of the trapezoidal rule with an equal mesh size for the integral of a function enjoying the double-exponential decay, together with an error estimate like (1.12). This result shows that the double-exponential transformation provides a more efficient quadrature formula than the single exponential one. Sugihara further shows that, except the identically vanishing function, there exists no function that is analytic and bounded in the strip region $|\mathrm{Im}\, w| < d$ and that decays more rapidly than $\exp(-\exp(\pi/2d|w|))$ as $w \to \pm\infty$. Thus, he concludes that the double-exponential formula is optimal.

2. Evaluation of Fourier-type integrals

Although the double-exponential transformation is useful for various kinds of integrals, it does not work well for Fourier-type integrals of a slowly decaying oscillatory function like

$$\begin{aligned} I_s &= \int_0^\infty f_1(x)\sin\omega x\,\mathrm{d}x, \\ I_c &= \int_0^\infty f_1(x)\cos\omega x\,\mathrm{d}x. \end{aligned} \tag{2.1}$$

In 1991 Ooura and Mori proposed a variable transformation suitable for such kinds of integrals [21]. Choose a function $\phi(t)$ satisfying

$$\phi(-\infty)=0, \quad \phi(+\infty)=\infty, \tag{2.2}$$

$$\phi'(t)\to 0 \quad \text{double exponentially as } t\to-\infty, \tag{2.3}$$

$$\phi(t)\to t \quad \text{double exponentially as } t\to+\infty, \tag{2.4}$$

and transform I_s and I_c using

$$\begin{cases} I_s : x = M\phi(t)/\omega \\ I_c : x = M\phi\left(t-\dfrac{\pi}{2M}\right)\Big/\omega \end{cases} \qquad (M=\text{const.}). \tag{2.5}$$

Then we have a new kind of the double-exponential formula useful for the integrals such as (2.1). M is a constant which will be determined as shown later. This transformation is chosen in such a way that as x becomes large in the positive direction the points of the formula approach double exponentially the zeros of $\sin\omega x$ or $\cos\omega x$, so that we do not have to evaluate the integrand for large value of x.

Ooura and Mori first proposed a transformation

$$\phi(t)=\frac{t}{1-\exp(-k\sinh t)}, \tag{2.6}$$

which satisfies the condition mentioned above [21]. Afterwards, Ooura proposed

$$\phi(t)=\frac{t}{1-\exp(-2t-\alpha(1-\mathrm{e}^{-t})-\beta(\mathrm{e}^{t}-1))} \tag{2.7}$$

$$\beta=\tfrac{1}{4}, \quad \alpha=\beta/\sqrt{1+M\log(1+M)/(4\pi)} \tag{2.8}$$

as a more efficient transformation [18,20,23].

Table 2
Comparison of the efficiency of a DE integrator and DQAWF. While 10^{-8} is given as the absolute error tolerance for DQAWF, it is given as the relative error tolerance for the DE integrator. N is the number of function evaluations, and rel.error and abs.error are the actual relative and absolute errors of the result

	DE integrator			DQAWF	
Integral	N	rel.error	abs.error	N	abs.error
I_5	72	$1.0 \cdot 10^{-14}$	$1.6 \cdot 10^{-14}$	430	$1.8 \cdot 10^{-11}$
I_6	308	$3.6 \cdot 10^{-11}$	$2.0 \cdot 10^{-11}$	445	$2.7 \cdot 10^{-11}$
I_7	70	$1.2 \cdot 10^{-10}$	$6.9 \cdot 10^{-11}$	570	$6.8 \cdot 10^{-12}$
I_8	68	$1.0 \cdot 10^{-14}$	$1.3 \cdot 10^{-14}$	615	$1.4 \cdot 10^{-11}$

If we substitute $\phi(t)$ into x of I_s in (2.1) we have

$$I_s = M\int_{-\infty}^{\infty} f_1(M\phi(t)/\omega)\sin(M\phi(t))\phi'(t)/\omega\,\mathrm{d}t. \tag{2.9}$$

Then, we apply the trapezoidal formula with an equal mesh size h and have

$$I_{s,h}^{(N)} = Mh\sum_{k=-N_-}^{N_+} f_1(M\phi(kh)/\omega)\sin(M\phi(kh))\phi'(kh)/\omega. \tag{2.10}$$

The situation is similar in the case of I_c. Here we choose M and h in such a way that

$$Mh = \pi. \tag{2.11}$$

Then for I_s as well as I_c

$$\sin(M\phi(kh)) \sim \sin Mkh = \sin \pi k = 0,$$

$$\cos\left(M\phi\left(kh - \frac{\pi}{2M}\right)\right) \sim \cos\left(Mkh - \frac{\pi}{2}\right) = \cos\left(\pi k - \frac{\pi}{2}\right) = 0 \tag{2.12}$$

hold, and we see that as k becomes large the points approach the zeros of $\sin \omega x$ or $\cos \omega x$ double exponentially.

The formula gives a good result even when it is applied to an integral

$$I = \int_0^{\infty} \log x \sin x\,\mathrm{d}x = -\gamma \tag{2.13}$$

whose integrand has a divergent function $\log x$ [22]. Although this integral should be defined as

$$\lim_{\varepsilon\to 0}\int_0^{\infty} \exp(-\varepsilon x)\log x \sin x\,\mathrm{d}x = -\gamma, \tag{2.14}$$

we will get an approximate value of $-\gamma$ that is correct up to 10 significant digits with only 70 function evaluations of $f_1(x) = \log x$ in the formula (2.10).

In Table 2, we show numerical examples of Fourier-type integrals to compare the efficiency of a DE automatic integrator based on the DE transformation (2.7) and (2.8), and DQAWF in QUADPACK [24] for the following four integrals:

$$I_5 = \int_0^{\infty} \frac{\sin x}{x}\,\mathrm{d}x, \qquad I_6 = \int_0^{\infty} \frac{\cos x}{(x-2)^2+1}\,\mathrm{d}x,$$

$$I_7 = \int_0^\infty \log x \sin x \,\mathrm{d}x, \qquad I_8 = \int_0^\infty \frac{\cos x}{\sqrt{x}}\,\mathrm{d}x.$$

3. Application of the double-exponential transformation to other types of integrals

The double-exponential transformation can be used to evaluate other kinds of integrals.

Ogata et al. proposed a method to evaluate the Cauchy principal value integral

$$I = \text{p.v.} \int_{-1}^{1} \frac{f(x)}{x-\lambda}\,\mathrm{d}x \tag{3.1}$$

and the Hadamard finite-part integral

$$I = \text{f.p.} \int_{-1}^{1} \frac{f(x)}{(x-\lambda)^n}\,\mathrm{d}x \tag{3.2}$$

by means of the double-exponential transformation [16].

Ogata and Sugihara also proposed a quadrature formula for oscillatory integrals involving Bessel functions such as

$$I = \int_0^\infty \frac{x}{x^2+1} J_0(x)\,\mathrm{d}x, \tag{3.3}$$

employing the same idea as mentioned in Section 2 [13–15]. It is noted here that, while developing the quadrature formula, they achieved an extremely high-precision quadrature formula of interpolatory type for antisymmetric integrals, i.e.,

$$I = \int_{-\infty}^{\infty} (\operatorname{sign} x) f(x)\,\mathrm{d}x = \left(\int_0^\infty - \int_{-\infty}^{0}\right) f(x)\,\mathrm{d}x. \tag{3.4}$$

The abscissae of the quadrature are zeros of Bessel functions.

Ooura devised a transformation which can be regarded as a continuous version of the Euler transformation. By this transformation, together with the double exponential one, we can evaluate integrals of a slowly decaying oscillatory function like

$$I = \int_0^\infty J_0\left(\sqrt{2x+x^2}\right)\mathrm{d}x \tag{3.5}$$

whose distribution of the zeros is not equidistant [17,20].

Also Ooura combined his continuous Euler transformation with FFT to give a method for efficient evaluation of a Fourier transform [19,20] like

$$I = \frac{1}{2\pi} \int_{-\infty}^{\infty} \log(1+x^2)\,\mathrm{e}^{-\mathrm{i}\omega x}\,\mathrm{d}x. \tag{3.6}$$

4. Sinc numerical methods and the double-exponential transformation

Recently, it turned out that the double-exponential transformation is useful not only for numerical integration but also for a variety of so-called Sinc numerical methods.

The Sinc numerical methods are based on an approximation over the doubly infinite interval $(-\infty,\infty)$, which is written as

$$f(x) \approx \sum_{k=-n}^{n} f(kh)S(k,h)(x), \tag{4.1}$$

where the basis functions $S(k,h)(x)$ are the Sinc functions defined by

$$S(k,h)(x) = \frac{\sin \pi/h(x-kh)}{\pi/h(x-kh)}, \quad k = 0, \pm 1, \pm 2, \ldots, \tag{4.2}$$

with a positive constant h. The approximation (4.1) is called the Sinc approximation.

The Sinc approximation and numerical integration are closely related through an identity

$$\int_{-\infty}^{\infty} \left(\sum_{k=-n}^{n} f(kh)S(k,h)(x) - f(x) \right) \mathrm{d}x = h \sum_{k=-n}^{n} f(kh) - \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x \tag{4.3}$$

between the approximation error of the Sinc approximation and the one of integration by the trapezoidal rule. This identity implies that the class of the functions for which the Sinc approximation gives highly accurate approximations is almost identical to the class of functions for which the trapezoidal rule gives highly accurate results. This fact suggests that the applicability of the transformation technique developed in the area of numerical integration of the Sinc approximation, even further to the Sinc numerical methods. In fact, in [25], the standard treatise of the Sinc numerical methods, the single-exponential transformation is assumed to be employed. But why not the double-exponential transformation? Recently, Sugihara and his colleagues have started to examine the applicability of the double-exponential transformation to a variety of Sinc numerical methods.

In the most fundamental case, i.e., in the Sinc approximation, Sugihara makes a full study of the error, thereby proving that when the double-exponential transformation is employed, the optimal result is obtained just as in numerical integration [27].

Horiuchi and Sugihara combine the double-exponential transformation with the Sinc-Galerkin method for the second-order two-point boundary problem [1]. To be specific, consider

$$\begin{aligned} &\tilde{y}''(x) + \tilde{\mu}(x)\tilde{y}'(x) + \tilde{v}(x)\tilde{y}(x) = \tilde{\sigma}(x), \quad a < x < b, \\ &\tilde{y}(a) = \tilde{y}(b) = 0. \end{aligned} \tag{4.4}$$

Application of the variable transformation

$$x = \phi(t), \quad a = \phi(-\infty), \quad b = \phi(\infty), \tag{4.5}$$

together with the change of notation

$$y(t) = \tilde{y}(\phi(t)), \tag{4.6}$$

transforms the problem to

$$\begin{aligned} &y''(t) + \mu(t)y'(t) + v(t)y(t) = \sigma(t), \quad -\infty < t < \infty, \\ &y(-\infty) = y(\infty) = 0. \end{aligned} \tag{4.7}$$

The Sinc-Galerkin method approximates the solution of the transformed problem (4.7) by a linear combination of the Sinc functions:

$$y_N(t) = \sum_{k=-n}^{n} w_k S(k,h)(t), \quad N = 2n+1. \tag{4.8}$$

It is shown by both theoretical analysis and numerical experiments that the approximation error can be estimated by

$$|y(t) - y_N(t)| \leqslant c'N^{5/2}\exp(-c\sqrt{N}) \tag{4.9}$$

if the true solution $y(t)$ of the transformed problem decays single exponentially like

$$|y(t)| \leqslant \alpha\exp(-\beta|t|). \tag{4.10}$$

It is also shown that the approximation error can be estimated by

$$|y(t) - y_N(t)| \leqslant c'N^2\exp\left(-\frac{cN}{\log N}\right) \tag{4.11}$$

if the true solution $y(t)$ of the transformed problem decays double exponentially like

$$|y(t)| \leqslant \alpha\exp(-\beta\exp(\gamma|t|)). \tag{4.12}$$

Evidently, the error estimates (4.9) and (4.11) show the superiority of the double-exponential transformation. By the analogy with the case of the Sinc approximation we believe that the error estimate (4.11) should be best possible, i.e., the double-exponential transformation should be optimal, though it has not been proved yet.

Koshihara and Sugihara study the performance of the double-exponential transformation when used in the Sinc-Collocation method for the Sturm–Liouville eigenvalue problems. It is shown that the error behaves like (4.11) [3].

Matsuo applies the double-exponential transformation to the Sinc-pseudospectral method for the nonlinear Schrödinger equation and reports that a highly accurate numerical solution is obtained [4].

As seen above, the double-exponential transformation has proved to be a useful tool in numerical analysis in a number of areas. We believe that the double-exponential transformation should prove to be effective even in wider areas of numerical analysis.

References

[1] K. Horiuchi, M. Sugihara, Sinc-Galerkin method with the double exponential transformation for the two point boundary problems, Technical Report 99-05, Dept. of Mathematical Engineering, University of Tokyo, 1999.

[2] M. Iri, S. Moriguti, Y. Takasawa, On a certain quadrature formula, J. Comput. Appl. Math. 17 (1987) 3–20 (translation of the original paper in Japanese from Kokyuroku RIMS Kyoto Univ. No. 91, 1970, pp. 82–119.

[3] T. Koshihara, M. Sugihara, A numerical solution for the Sturm–Liouville type eigenvalue problems employing the double exponential transformation, Proceedings of the 1996 Annual Meeting of the Japan Society for Industrial and Applied Mathematics, 1996, pp. 136–137 (in Japanese).

[4] T. Matsuo, On an application of the DE transformation to a Sinc-type pseudospectral method, Proceedings of the 1997 Annual Meeting of the Japan Society for Industrial and Applied Mathematics, 1997, pp. 36–37 (in Japanese).

[5] M. Mori, On the superiority of the trapezoidal rule for the integration of periodic analytic functions, Mem. Numer. Math. (1) (1974) 11–19.

[6] M. Mori, An IMT-type double exponential formula for numerical integration, Publ. RIMS. Kyoto Univ. 14 (1978) 713–729.

[7] M. Mori, Quadrature formulas obtained by variable transformation and the DE-rule, J. Comput. Appl. Math. 12 & 13 (1985) 119–130.

[8] M. Mori, Numerical Methods and FORTRAN 77 Programming, Iwanami Shoten, Tokyo, 1986, pp. 168–186 (in Japanese).

[9] M. Mori, The double exponential formula for numerical integration over the half infinite interval, in: R.P. Agarwal et al. (Eds.), Numerical Mathematics (Singapore 1988), International Series of Numerical Mathematics, Vol. 86, Birkhäuser, Basel, 1988, pp. 367–379.

[10] M. Mori, An error analysis of quadrature formulas obtained by variable transformation, in: M. Kashiwara, T. Kawai (Eds.), Algebraic Analysis, Vol. 1, Academic Press, Boston, 1988, pp. 423–437.

[11] M. Mori, Developments in the double exponential formulas for numerical integration, Proceedings of the International Congress of Mathematicians, Kyoto, 1990, Springer, Tokyo, 1991, pp. 1585–1594.

[12] K. Murota, M. Iri, Parameter tuning and repeated application of the IMT-type transformation in numerical quadrature, Numer. Math. 38 (1982) 327–363.

[13] H. Ogata, T. Ooura, Theoretical error analysis of a DE-type quadrature formula for oscillatory integrals, Proceedings of the 1996 Annual Meeting of the Japan Society for Industrial and Applied Mathematics, 1996, pp. 18–19 (in Japanese).

[14] H. Ogata, M. Sugihara, Interpolation and quadrature formulae whose abscissae are the zeros of the Bessel functions, Trans. Japan Soc. Ind. Appl. Math. 6 (1996) 39–66 (in Japanese).

[15] H. Ogata, M. Sugihara, Quadrature formulae for oscillatory infinite integrals involving the Bessel functions, Trans. Japan Soc. Ind. Appl. Math. 8 (1998) 223–256 (in Japanese).

[16] H. Ogata, M. Sugihara, M. Mori, A DE-type quadrature rule for Cauchy principal-value integrals and Hadamard finite-part integrals, Trans. Japan Soc. Ind. Appl. Math. 3 (1993) 309–322 (in Japanese).

[17] T. Ooura, An extension of the Euler transformation, Proceedings of 1993 Annual Meeting of the Japan Society for Industrial and Applied Mathematics, 1993, pp. 111–112 (in Japanese).

[18] T. Ooura, A new variable transformation of Fourier-type integrals, Proceedings of the 1994 Annual Meeting of the Japan Society for Industrial and Appl. Mathematics, 1994, pp. 260–261 (in Japanese).

[19] T. Ooura, A method of computation of oscillatory integrals using a continuous Euler transformation and the DE transformation, Proceedings of 1996 Annual Meeting of the Japan Society for Industrial and Applied Mathematics, 1996, pp. 22–23 (in Japanese).

[20] T. Ooura, Study on numerical integration of Fourier type integrals, Doctoral Thesis, University of Tokyo, 1997 (in Japanese).

[21] T. Ooura, M. Mori, The double exponential formula for oscillatory functions over the half infinite interval, J. Comput. Appl. Math. 38 (1991) 353–360.

[22] T. Ooura, M. Mori, Double exponential formula for Fourier type integrals with a divergent integrand, in: R.P. Agarwal (Ed.), Contributions in Numerical Mathematics, World Scientific Series in Applicable Analysis, Vol. 2, World Scientific, Singapore, 1993, pp. 301–308.

[23] T. Ooura, M. Mori, A robust double exponential formula for Fourier type integrals, J. Comput. Appl. Math. 112 (1999) 229–241.

[24] R. Piessens, E. de Doncker-Kapenga, C.W. Überhuber, D.K. Kahaner, QUADPACK – A Subroutine Package for Automatic Integration, Springer, Berlin, 1983.

[25] F. Stenger, Numerical Methods Based on Sinc and Analytic Functions, Springer, New York, 1993.

[26] M. Sugihara, Optimality of the double exponential formula – functional analysis approach. Numer. Math. 75 (1997) 379–395. (The original paper in Japanese appeared in Kokyuroku RIMS Kyoto Univ. No. 585 (1986) 150–175).

[27] M. Sugihara, Near-optimality of the Sinc approximation, Technical Report 99-04, Dept. of Mathematical Engineering, University of Tokyo, 1999.

[28] H. Takahasi, M. Mori, Error estimation in the numerical integration of analytic functions, Report Comput. Centre, Univ. Tokyo, Vol. 3, 1970, pp. 41–108.

[29] H. Takahasi, M. Mori, Quadrature formulas obtained by variable transformation, Numer. Math. 21 (1973) 206–219.

[30] H. Takahasi, M. Mori, Double exponential formulas for numerical integration, Publ. RIMS Kyoto Univ. 9 (1974) 721–741.

Journal of Computational and Applied Mathematics 127 (2001) 297–315

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Orthogonal and L_q-extremal polynomials on inverse images of polynomial mappings ☆

Franz Peherstorfer *, Robert Steinbauer

Institut für Analysis und Numerik, Johannes Kepler Universität Linz, 4040 Linz-Auhof, Austria

Received 9 June 1999; received in revised form 30 November 1999

Abstract

Let $\mathscr{T}$ be a polynomial of degree N and let K be a compact set with $\mathbb{C}$. First it is shown, if zero is a best approximation to f from $\mathbb{P}_n$ on K with respect to the $L_q(\mu)$-norm, $q \in [1,\infty)$, then zero is also a best approximation to $f \circ \mathscr{T}$ on $\mathscr{T}^{-1}(K)$ with respect to the $L_q(\mu^{\mathscr{T}})$-norm, where $\mu^{\mathscr{T}}$ arises from μ by the transformation $\mathscr{T}$. In particular, $\mu^{\mathscr{T}}$ is the equilibrium measure on $\mathscr{T}^{-1}(K)$, if μ is the equilibrium measure on K. For $q = \infty$, i.e., the sup-norm, a corresponding result is presented. In this way, polynomials minimal on several intervals, on lemniscates, on equipotential lines of compact sets, etc. are obtained. Special attention is given to $L_q(\mu)$-minimal polynomials on Julia sets. Next, based on asymptotic results of Widom, we show that the minimum deviation of polynomials orthogonal with respect to a positive measure on $\mathscr{T}^{-1}(\partial K)$ behaves asymptotically periodic and that the orthogonal polynomials have an asymptotically periodic behaviour, too. Some open problems are also given. © 2001 Elsevier Science B.V. All rights reserved.

MSC: 33C45; 42C05

Keywords: Orthogonal- and extremal polynomials; Asymptotics; Composition of orthogonal polynomials; Linear (definite) functionals; Julia sets

1. Introduction and preliminaries

Let $(c_{jk})_{j,k\in\mathbb{N}_0}, \mathbb{N}_0 := \mathbb{N} \cup \{0\}$, be an infinite matrix of complex numbers and denote by $\mathscr{P}$ the space of polynomials in z and $\bar{z}$ with complex coefficients. Further, let the linear functional $\mathscr{L} : \mathscr{P} \to \mathbb{C}$

☆ This work was supported by the Austrian Fonds zur Förderung der wissenschaftlichen Forschung, project-number P12985-TEC.

* Corresponding author.

E-mail addresses: franz.peherstorfer@jk.uni-linz.ac.at (F. Peherstorfer), robert.steinbauer@jk.uni-linz.ac.at (R. Steinbauer).

PII: S0377-0427(00)00502-1

be given by, $n \in \mathbb{N}$,

$$\mathscr{L}\left(\sum_{j,k=0}^{n} d_{jk} z^j \bar{z}^k\right) = \sum_{j,k=0}^{n} d_{jk} c_{jk}. \tag{1.1}$$

In the following let us assume that there is a compact set $K \subset \mathbb{C}$ such that $\mathscr{L}$ is extendable to $C(K)$. In view of Horn [17, Theorem 4.2] the functional $\mathscr{L}$ is extendable to a bounded linear functional on $C(K)$, if $|\sum c_{jk} a_{jk}| \leqslant \text{const} \sup_{z\in K} |\sum a_{jk} z^j \bar{z}^k|$ for all 2-indexed sequences of complex numbers $\{a_{jk}\}$ with only finitely many nonzero terms. The last condition is equivalent to the fact that there exists a complex measure μ with support in K such that

$$c_{jk} = \mathscr{L}(z^j \bar{z}^k) = \int_K z^j \bar{z}^k \, \mathrm{d}\mu(z), \quad j,k \in \mathbb{N}_0. \tag{1.2}$$

We say that a function $g \in C(K)$ is orthogonal to $\mathbb{P}_{n-1}$ (as usual, $\mathbb{P}_n$ denotes the space of polynomials in z of degree less than or equal to n with complex coefficients) if

$$\mathscr{L}(z^j g(z)) = 0 \quad \text{for } j = 0,1,\ldots,n-1$$

and hermitian orthogonal to $\mathbb{P}_{n-1}$ with respect to $\mathscr{L}$ if

$$\mathscr{L}(z^j \overline{g(z)}) = 0 \quad \text{for } j = 0,1,\ldots,n-1.$$

Example 1. (a) Let K be a rectifiable curve or arc in the complex plane, w a real nonnegative integrable function on K, and set

$$c_{jk} = \int_K z^j \bar{z}^k w(z)\,|\mathrm{d}z| \quad \text{for } j,k \in \mathbb{N}_0.$$

Then the hermitian orthogonality of a polynomial p_n to $\mathbb{P}_{n-1}$ with respect to $\mathscr{L}$ becomes the usual orthogonality

$$\int_K z^j \overline{p_n(z)} w(z)\,|\mathrm{d}z| = 0 \quad \text{for } j = 0,\ldots,n-1.$$

Naturally, if K is a subset of the real line, then there is no difference in the above definitions of orthogonality.

(b) Let μ be a complex (not necessary real and/or positive) measure on the curve or arc K, and let

$$c_{jk} = c_{j+k} = \int_K z^{j+k} \, \mathrm{d}\mu(z). \tag{1.3}$$

Then the polynomial $p_n \in \mathbb{P}_n$ orthogonal with respect to $\mathscr{L}$ is the denominator of the so-called $[n/n]$ Padé approximant of the function $\int_K \mathrm{d}\mu(z)/(y-z)$, i.e.,

$$\frac{p_n^{[1]}(y)}{p_n(y)} = \int_K \frac{\mathrm{d}\mu(z)}{y-z} + \mathcal{O}\left(\frac{1}{y^{2n+1}}\right) \quad \text{as } y \to \infty,$$

where $p_n^{[1]}$ is the polynomial of the second kind given by

$$p_n^{[1]}(y) = \int_K \frac{p_n(y) - p_n(z)}{y-z} \, \mathrm{d}\mu(z).$$

Notation. For $\mathscr{T} \in \mathbb{P}_N \backslash \mathbb{P}_{N-1}$ and $S \in \mathbb{P}_{N-1}$, we put

$$\mathscr{L}^{\mathscr{T},S}(f(z)\overline{g(z)}) := \mathscr{L}\left(\sum_{j=1}^{N} \frac{S(\mathscr{T}_j^{-1}(z))}{\mathscr{T}'(\mathscr{T}_j^{-1}(z))} f(\mathscr{T}_j^{-1}(z))\overline{g(\mathscr{T}_j^{-1}(z))}\right). \tag{1.4}$$

Here, $\{\mathscr{T}_j^{-1}: j = 1,\dots,N\}$ denotes the complete assignment of branches of $\mathscr{T}^{-1}$. Definition (1.4) of the linear functional $\mathscr{L}^{\mathscr{T},S}$ is quite natural and can be understood in the following way: By partial fraction expansion we have

$$\frac{S(y)}{\mathscr{T}(y) - z} = \sum_{j=1}^{N} \frac{S(\mathscr{T}_j^{-1}(z))}{\mathscr{T}'(\mathscr{T}_j^{-1}(z))} \frac{1}{y - \mathscr{T}_j^{-1}(z)} \tag{1.5}$$

from which we get for large $y \in \mathbb{C}$ and for all $k \in \mathbb{N}_0$

$$\begin{aligned}\mathscr{L}^{\mathscr{T},S}\left(\frac{\overline{[\mathscr{T}(z)]}^k}{y - z}\right) &= \sum_{v=0}^{\infty} y^{-(v+1)} \mathscr{L}^{\mathscr{T},S}(z^v \overline{[\mathscr{T}(z)]}^k)\\ &= \sum_{v=0}^{\infty} y^{-(v+1)} \mathscr{L}\left(\sum_{j=1}^{N} \frac{S(\mathscr{T}_j^{-1}(z))}{\mathscr{T}'(\mathscr{T}_j^{-1}(z))}[\mathscr{T}_j^{-1}(z)]^v \bar{z}^k\right)\\ &= \mathscr{L}\left(\bar{z}^k \sum_{j=1}^{N} \frac{S(\mathscr{T}_j^{-1}(z))}{\mathscr{T}'(\mathscr{T}_j^{-1}(z))} \frac{1}{y - \mathscr{T}_j^{-1}(z)}\right) = S(y)\mathscr{L}\left(\frac{\bar{z}^k}{\mathscr{T}(y) - z}\right).\end{aligned} \tag{1.6}$$

In the same way we also get the relation

$$\mathscr{L}^{\mathscr{T},S}\left(\frac{[\mathscr{T}(z)]^k}{y - z}\right) = S(y)\mathscr{L}\left(\frac{z^k}{\mathscr{T}(y) - z}\right)$$

$$\text{especially } \mathscr{L}^{\mathscr{T},S}\left(\frac{1}{y - z}\right) = S(y)\mathscr{L}\left(\frac{1}{\mathscr{T}(y) - z}\right). \tag{1.7}$$

In particular, we have

$$\mathscr{L}^{\mathscr{T},1}(z^k) = 0 \quad \text{for } k = 0,\dots,N-2 \quad \text{and thus } S(y)\mathscr{L}\left(\frac{1}{\mathscr{T}(y) - z}\right) = \mathscr{L}\left(\frac{S(z)}{\mathscr{T}(y) - z}\right). \tag{1.8}$$

Now, we are ready to show how to get orthogonality properties for $\mathscr{T}$-compositions.

Theorem 2. *Let $\mathscr{T}$ and S be polynomials of degree N and $m \leqslant N - 1$, respectively.*

(a) *Suppose that $\mathscr{L}(z^j\overline{g(z)}) = 0$ for $j = 0,1,\dots,n-1$. Then*

$$\mathscr{L}^{\mathscr{T},S}(z^j\overline{(g \circ \mathscr{T})(z))} = 0 \quad for\ j = 0,1,\dots(n+1)N - m - 2.$$

(b) *Suppose that $\mathscr{L}(z^j g(z)) = 0$ for $j = 0,1,\dots,n-1$. Then*

$$\mathscr{L}^{\mathscr{T},S}(z^j(g \circ \mathscr{T})(z)) = 0 \quad for\ j = 0,1,\dots(n+1)N - m - 2.$$

Proof. (a) From (1.6) and the linearity of the functionals $\mathscr{L}^{\mathscr{T},S}$ and $\mathscr{L}$ we get

$$\mathscr{L}^{\mathscr{T},S}\left(\frac{\overline{(g\circ\mathscr{T})(z)}}{y-z}\right)=S(y)\mathscr{L}\left(\frac{\overline{g(z)}}{\mathscr{T}(y)-z}\right).$$

Let us now expand both sides in a power series:

$$\begin{aligned}\sum_{v=0}^{\infty} y^{-(v+1)}\mathscr{L}^{\mathscr{T},S}(z^v\overline{(g\circ\mathscr{T})(z)})&=S(y)\sum_{v=0}^{\infty}[\mathscr{T}(y)]^{-(v+1)}\mathscr{L}(z^v\overline{g(z)})\\&=S(y)\sum_{v=n}^{\infty}[\mathscr{T}(y)]^{-(v+1)}\mathscr{L}(z^v\overline{g(z)})=\mathscr{O}(y^{-(n+1)N+m}),\end{aligned}$$

where the second identity follows by the orthogonality property of g. Hence,

$$\mathscr{L}^{\mathscr{T},S}(z^j\overline{(g\circ\mathscr{T})(z)})=0 \quad \text{for } j=0,\ldots,(n+1)N-m-2,$$

which proves part (a) of the theorem.

Part (b) follows in the same way, just by using relation (1.7). □

Now of special interest are functionals with an integral representation (1.2). Then transformation (1.4) defines a measure $\mathrm{d}\mu^{\mathscr{T},S}$ on the inverse image $\mathscr{T}^{-1}(K)$ by

$$\begin{aligned}\mathscr{L}^{\mathscr{T},S}(f(z)\overline{g(z)})&=\sum_{j=1}^{N}\int_K f(\mathscr{T}_j^{-1}(z))\overline{g(\mathscr{T}_j^{-1}(z))}\frac{S(\mathscr{T}_j^{-1}(z))}{\mathscr{T}'(\mathscr{T}_j^{-1}(z))}\,\mathrm{d}\mu(z)\\&=:\int_{\mathscr{T}^{-1}(K)} f(z)\overline{g(z)}\,\mathrm{d}\mu^{\mathscr{T},S}(z).\end{aligned}\tag{1.9}$$

Here, we assume that $S(y)/\mathscr{T}'(y)$ does not have poles on $\mathscr{T}^{-1}(K)$. A sufficient, but not a necessary, condition is that there are no critical points of $\mathscr{T}$ on $\mathscr{T}^{-1}(K)$. But if there are critical points on $\mathscr{T}^{-1}(K)$, then they have to be canceled out by the zeros of S. Measure transformations of the kind (1.9) have been studied, e.g., in [5,6,14].

The following example will be of importance in Section 2. Let K be a complex curve, w a real integrable function on K, and let

$$\mathrm{d}\mu(z):=w(z)|\mathrm{d}z|$$

be a real measure. Then

$$\mathrm{d}\mu^{\mathscr{T},S}(z)=\frac{S(z)}{\mathscr{T}'(z)}|\mathscr{T}'(z)|w(\mathscr{T}(z))|\mathrm{d}z|=\overline{\operatorname{sgn}\mathscr{T}'(z)}S(z)w(\mathscr{T}(z))|\mathrm{d}z|,\tag{1.10}$$

in particular we obtain for $S=\mathscr{T}'$ that

$$\mathrm{d}\mu^{\mathscr{T},\mathscr{T}'}(z)=|\mathscr{T}'(z)|w(\mathscr{T}(z))|\mathrm{d}z|.$$

In this paper we will study how polynomial measure transformations of form (1.9) resp. (1.10) can be applied to L_q-approximation, $q\in[1,\infty]$, on curves, arcs, Julia sets, etc. Special attention is given to a simple representation of the necessary ingredients like Green's function, equilibrium measure, etc. and in particular to the presentation of many examples, see Sections 2.1 and 2.2, but also the Julia set example in Section 3. Furthermore, for the case $q=2$ it is demonstrated how to get from Widom's theory asymptotic statements for polynomials orthogonal on inverse images of polynomial mappings.

2. Best approximations and minimal polynomials with respect to the L_q-norm, $q \in [1,\infty]$

In this section we assume that $K \subset \mathbb{C}$ is a compact set and μ a positive Borel measure on K. Let us recall some well-known facts from approximation theory (see, e.g., [1]): A $L_q(\mu)$-integrable, $q \in [1,\infty)$, function f has 0 as a best approximation from $\mathbb{P}_{n-1}$ with respect to the $L_q(\mu)$-norm on K, i.e.,

$$\|f\|_{q,\mu,K} = \inf_{p\in\mathbb{P}_{n-1}} \|f-p\|_{q,\mu,K},$$

where $\|f\|_{q,\mu,K} = (\int_K |f(z)|^q \, d\mu(z))^{1/q}$, if and only if

$$\int_K z^j |f(z)|^{q-2}\overline{f(z)}\, d\mu(z) = 0 \quad \text{for } j = 0,\dots,n-1. \tag{2.1}$$

If f satisfies (2.1), then we say that f is $L_q(\mu)$-orthogonal on K with respect to $\mathbb{P}_{n-1}$.

Note that Theorem 2 (put $g = |f|^{q-2}\bar{f}$ there) and definition (1.9) imply immediately that $f \circ \mathscr{T}$ is $L_q(\mu^{\mathscr{T},S})$-orthogonal on $\mathscr{T}^{-1}(K)$, i.e.,

$$\int_{\mathscr{T}^{-1}(K)} z^j |f(\mathscr{T}(z))|^{q-2}\overline{f(\mathscr{T}(z))}\, d\mu^{\mathscr{T},S}(z) = 0 \tag{2.2}$$

for $j = 0,\dots,(n+1)N-m-2$, where m is the degree of the polynomial S.

As usual, a monic polynomial $p_n(z) = z^n + \cdots$ of degree n is called a $L_q(\mu)$-*minimal polynomial*, $q \in [1,\infty]$, with respect to the measure μ if p_n has 0 as a best approximation from $\mathbb{P}_{n-1}$ with respect to the $L_q(\mu)$-norm on K. Note that p_n is a L_∞-minimal polynomial if

$$\max_{z\in K} |p_n(z)| = \inf_{\substack{Q_n\in\mathbb{P}_n \\ Q_n(z)=z^n+\cdots}} \max_{z\in K} |Q_n(z)|.$$

Moreover, the monic $L_q(\mu)$-minimal polynomial, $q \in [1,\infty)$, p_n of degree n is characterized by the orthogonality condition (2.1), where f is to be replaced by p_n.

Theorem 3. *Let $K \subset \mathbb{C}$ be a compact set, μ a positive Borel measure on K, $\mathscr{T}(z) = \kappa z^N + \cdots$ a complex polynomial of degree N, and S a complex polynomial of degree $m \leqslant N-1$. Furthermore, let the measure $\mu^{\mathscr{T},S}$ be defined as in* (1.9). *Suppose that $\mu^{\mathscr{T},S}$ is a positive measure on $\mathscr{T}^{-1}(K)$ and $q \in [1,\infty)$. Then the following statements hold*:

If $f \in L_q(\mu)$ has 0 as a best approximation from $\mathbb{P}_{n-1}$ with respect to the $L_q(\mu)$-norm on K, then $(f \circ \mathscr{T})$ has 0 as a best approximation from $\mathbb{P}_{(n+1)N-m-2}$ with respect to the $L_q(\mu^{\mathscr{T},S})$-norm on $\mathscr{T}^{-1}(K)$.

If f is continuous on K and has 0 as a best approximation from $\mathbb{P}_{n-1}$ with respect to the sup-norm and the positive continuous weight function $w(z)$ on K, then $f \circ \mathscr{T}$ has 0 as a best approximation with respect to the sup-norm and weight function $w \circ \mathscr{T}$ on $\mathscr{T}^{-1}(K)$.

Moreover, if p_n is a monic $L_q(\mu)$-minimal polynomial on K, $q \in [1,\infty]$, then $(p_n \circ \mathscr{T})(z)/\kappa^n = z^{nN} + \cdots$ is a monic $L_q(\mu^{\mathscr{T},S})$-minimal polynomial on $\mathscr{T}^{-1}(K)$.

Proof. For $q \in [1,\infty)$ the assertion follows from (2.2) and Theorem 1. Concerning the statement with respect to the sup-norm, we first observe that it follows from above that for every

$q \in [1,\infty)$, $h_{n-1,q} \circ \mathscr{T}$ is a best approximation from $\mathbb{P}_{(n+1)N-m-2}$ of $f \circ \mathscr{T}$ with respect to the $L_q((w \circ \mathscr{T})^q|\mathscr{T}'||\mathrm{d}z|)$-norm if $h_{n-1,q}$ is a best approximation from $\mathbb{P}_{n-1}$ of f with respect to the $L_q(w^q|\mathrm{d}z|)$-norm. Now it is known (see [20]) that for $f, w \in C(K)$, $h_{n-1,q}(z) \to h_{n-1,\infty}(z)$ as $q \to \infty$ uniformly on K, where $h_{n-1,\infty}$ is the unique best approximation of f from $\mathbb{P}_{n-1}$ with respect to the sup-norm and weight function w. This gives the assertion. □

Thus, if we know the sequence of minimal polynomials on K, we know a subsequence of minimal polynomials on $\mathscr{T}^{-1}(K)$ at least. Let us mention that by (1.9) $\mu^{\mathscr{T},S}$ is a positive measure on $\mathscr{T}^{-1}(K)$ if $\operatorname{sgn} S = \operatorname{sgn} \mathscr{T}'$ on $\mathscr{T}^{-1}(K)$, where $\operatorname{sgn} z = z/|z|$. By different methods, various special cases of Theorem 2 have been proved in [12,19,22,23].

Before we are going to consider some examples, let us give the definition and notations for Green's function, equipotential lines, equilibrium measure, etc. of a compact set $K \subset \mathbb{C}$ with $\operatorname{cap}(K) > 0$. We denote by $g(z,K,\infty)$ Green's function for $\bar{\mathbb{C}}\backslash K$ with pole at infinity. Recall that $g(\cdot,K,\infty) : \mathbb{C}\backslash K \to \mathbb{R}^+$ is defined by being harmonic on $\mathbb{C}\backslash K$ with

$$g(z,K,\infty) = \ln|z| + \mathcal{O}(1) \quad \text{as } z \to \infty$$

and

$$g(z,K,\infty) \to 0 \quad \text{quasi-everywhere} \quad \text{as } z \to z_0 \in \partial K.$$

The equipotential lines of K for a value $\rho > 0$ are given by

$$A(K,\rho) := g^{-1}(\{\rho\},K,\infty) = \{z \in \mathbb{C}\backslash K : g(z,K,\infty) = \rho\}. \tag{2.3}$$

The equilibrium measure $\mu_{K,e}$ of K (if there is no danger of confusion we omit K and write shortly μ_e) is the measure which satisfies

$$g(z,K,\infty) + \log(\operatorname{cap}(K)) = \int_K \log|z-y|\,\mathrm{d}\mu_{K,e}(z), \tag{2.4}$$

where

$$-\log(\operatorname{cap}(K)) = \lim_{z\to\infty} (g(z,K,\infty) - \log|z|), \tag{2.5}$$

and the constant $\operatorname{cap}(K)$ is called the logarithmic capacity of K. Finally, let $\tilde{g}(z,K,\infty)$ be a harmonic conjugate of $g(z,K,\infty)$,

$$G(z,K,\infty) := g(z,K,\infty) + \mathrm{i}\tilde{g}(z,K,\infty)$$

the complex Green's function and

$$\Phi(z,K,\infty) := \exp(G(z,K,\infty)) \tag{2.6}$$

the mapping which maps the exterior of K onto the exterior of the unit circle.

Next let us show, if we know Green's function, equilibrium measure etc. for K, then we know them for $\mathscr{T}^{-1}(K)$ also.

Lemma 4. *Let* $\mathscr{T}(z) = \kappa z^N + \cdots$ *then the following relations hold*:

(a) $g(z,\mathscr{T}^{-1}(K),\infty) = g(\mathscr{T}(z),K,\infty)/N$.
(b) $\Phi(z,\mathscr{T}^{-1}(K),\infty) = \exp(G(\mathscr{T}(z),K,\infty))^{1/N}$.

(c) $\mu_{K,e}^{\mathcal{T},\mathcal{T}'}/N = \mu_{\mathcal{T}^{-1}(K),e}$; *recall definition* (1.9).
(d) $\operatorname{cap}(\mathcal{T}^{-1}(K)) = (\operatorname{cap}(K)/\kappa)^{1/N}$.
(e) $A(\mathcal{T}^{-1}(K),\rho) = \{z \in \mathbb{C}\backslash\mathcal{T}^{-1}(K){:}g(\mathcal{T}(z),K,\infty) = \rho N\}$.

Proof. Concerning relation (a) see [23], this property follows immediately from the definition of Green's function. Since $\tilde{g}(\mathcal{T}(z),K,\infty)/N$ is a harmonic conjugate of $g(\mathcal{T}(z),K,\infty)/N$, as can easily be checked by the Cauchy–Riemann differential equations, part (b) follows. Parts (c) and (d) follow from (2.4) and (2.5) with the help of (a), the relation

$$\frac{1}{N}\log|\mathcal{T}(z)-y| = \frac{1}{N}\log\prod_{\mathcal{T}(\xi)=y}|z-\xi| + \frac{1}{N}\log\kappa = \frac{1}{N}\sum_{j=1}^{N}\log|z-\mathcal{T}_j^{-1}(y)| + \frac{1}{N}\log\kappa$$

and the definition of $\mu_{K,e}^{\mathcal{T},\mathcal{T}'}$. Part (e) is again an immediate consequence of (a). □

The above lemma is more or less known, at least for special cases (see [14,23] and also [27, Chapter 6.5], as pointed out by one of the referees).

Let us now give some examples showing how useful Theorem 3 is.

2.1. Extremal polynomials on disconnected sets and equipotential lines

For $\rho \in [0,\infty)$ let $K = E_\rho := \{z \in \mathbb{C}\colon \log|z+\sqrt{z^2-1}| = \rho\}$, where the branch of $\sqrt{z}$ is chosen such that $\operatorname{sign}\sqrt{x^2-1} = \operatorname{sign}(x-1)$ for $x \in \mathbb{R}\setminus[-1,1]$. Then $E_0 = [-1,1]$ while for each $\rho > 0$ the set E_ρ is the ellipse with foci ± 1 and semi-axes equal to $(\mathrm{e}^\rho \pm \mathrm{e}^{-\rho})/2$. As it is known and can easily be seen that

$$g(z,[-1,1],\infty) = \ln|z+\sqrt{z^2-1}|. \tag{2.7}$$

Therefore, the ellipses E_ρ are the equipotential lines of E_0. Now, let as usual $T_n(x) = (1/2^{n-1})\cos n$ $(\arccos x)$, $x \in [-1,1]$, denote the classical monic Chebyshev polynomial of degree n. Then it is known that the T_n's are the monic L_∞- as well as the $L_2(\mu)$-minimal polynomials on E_ρ, where

$$\mathrm{d}\mu(z) = \mathrm{d}\mu_{E_\rho,e}(z) = \frac{|\mathrm{d}z|}{\sqrt{|1-z^2|}}. \tag{2.8}$$

Furthermore, it can be proved almost analogously as in the case $q = 2$ by using (2.1), see [10, pp. 240–241], that T_n is also a $L_q(\mu)$-, $q \in [1,\infty)$, minimal polynomial on the equipotential lines E_ρ, $\rho \in [0,\infty)$. By the way, for $\rho = 0$ this is a well-known fact. Now, let $\mathcal{T}(z) = \kappa z^N + \cdots$, $\kappa \neq 0$, be a real polynomial which has N simple zeros in $(-1,1)$ and which satisfies

$$\min\{|\mathcal{T}(x)|\colon \mathcal{T}'(x) = 0\} \geqslant 1. \tag{2.9}$$

For the simple case of $N = 2$ compare the left picture in Fig. 1. Further, let l be the number of points x such that $|\mathcal{T}(x)| = 1$ and $\mathcal{T}'(x) = 0$. Then

$$\mathcal{T}^{-1}(E_\rho) = \{z \in \mathbb{C}\colon \log|\mathcal{T}(z)+\sqrt{\mathcal{T}^2(z)-1}| = \rho\}, \quad \text{for } \rho \in [0,\infty).$$

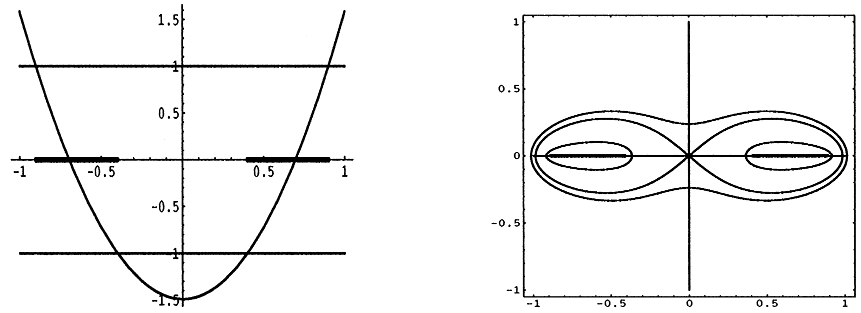


Fig. 1. Left picture: polynomial $\mathcal{T}$ of degree 2 and set $\mathcal{T}^{-1}(E_0)$. Right picture: sets $\mathcal{T}^{-1}(E_\rho)$ for $\rho = 0, 0.4, 0.9$, and 1.1.

In particular, $\mathcal{T}^{-1}(E_0)$ is the set of $N - l$ disjoint real intervals

$$\{x \in [-1, 1]: |\mathcal{T}(x)| \leqslant 1\}.$$

Recall that by Lemma 1 the sets $\mathcal{T}^{-1}(E_\rho)$, $\rho > 0$, are the equipotential lines of the $N - l$ disjoint real intervals $\mathcal{T}^{-1}(E_0)$; compare again Fig. 1 (right picture).

Hence, by Theorem 3 the polynomials $T_n(\mathcal{T}(z))/\kappa^n$ are the L_∞- resp. $L_q(\mu^{\mathcal{T},S})$-minimal polynomials, $q \in [1, \infty)$, on $\mathcal{T}^{-1}(E_\rho)$, where by (1.10)

$$\mathrm{d}\mu^{\mathcal{T},S}(z) = \frac{S(z)}{\operatorname{sgn} \mathcal{T}'(z)} \frac{|\mathrm{d}z|}{\sqrt{|1 - \mathcal{T}^2(z)|}};$$

S satisfies the assumption of Theorem 3. Recall that for $S = \mathcal{T}'$, $\mu^{\mathcal{T},\mathcal{T}'} = \mu_{\mathcal{T}^{-1}(K),e}$ is the equilibrium measure of $\mathcal{T}^{-1}(K)$. By completely different methods the case $\rho = 0$ has been treated in [22,23].

Using the fact that the Chebyshev polynomials of the second kind $U_n(z) = (1/2^n)(1 - z^2)^{-1/2} \sin[(n+1)\arccos z]$ are $L_2(1)$-minimal polynomials on the interior of the ellipses E_ρ (compare again [10, Ex. 4, p. 241]), then an analogous procedure as above gives a (sub)sequence of orthogonal polynomials on the union of complex areas with respect to a measure of form (1.9) with $\mathrm{d}\mu(z) = \mathrm{d}z$.

2.2. Extremal polynomials on “stars”

Let the sets E_ρ and the measure μ be given as at the beginning of Section 2.1 and let $\mathcal{T}(z) = z^N$, $N \geqslant 1$ be fixed. Then $\mathcal{T}^{-1}(E_0)$ is the 2N-star

$$S_{2N} := \{r\mathrm{e}^{k\pi\mathrm{i}/N}: r \in [0, 1],\ k = 0, \ldots, 2N - 1\}$$

and $\mathcal{T}^{-1}(E_\rho)$ is a smooth “curve around” this star for $\rho > 0$; compare the left picture in Fig. 2.

Again, by Theorem 3 the compositions with Chebyshev polynomials $T_n(z^N)$ give L_∞- resp. $L_q(\mu^{\mathcal{T},\mathcal{T}'})$-minimal polynomials on $\mathcal{T}^{-1}(E_\rho)$, $\rho \geqslant 0$, where by (2.8)

$$\mathrm{d}\mu^{\mathcal{T},\mathcal{T}'} = \mathrm{d}\mu_{\mathcal{T}^{-1}(E_\rho),e} = \frac{|z^{N-1}|\,|\mathrm{d}z|}{\sqrt{|1 - z^{2N}|}}$$

(here, we took $S(z) = \mathcal{T}'(z)/N = z^{N-1}$). For the L_∞-case see [23].

Naturally, we would like to know the whole sequence of $L_q(\mu^{\mathcal{T},\mathcal{T}'})$-minimal polynomials and not only a subsequence. For the 2N-star S_{2N} the whole sequence can be obtained as follows.

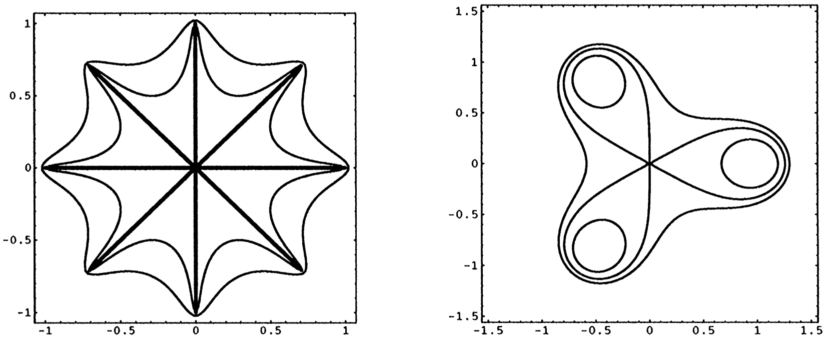


Fig. 2. Left picture: sets $\mathscr{T}^{-1}(E_\rho)$ for $\rho=0, 0.1$, and 0.4; here $N=4$. Right picture: lemniscate with $N=3$ and $r=1.2, 1, 0.7$.

Proposition 5. *Let* $n\in\mathbb{N}_0$, $1\leqslant N\in\mathbb{N}$, $m\in\{0,\dots,2N-1\}$, $q\in[1,\infty)$, $w:[0,1]\to\mathbb{R}^+$ *integrable, and let* $p_{n,m}(x)=x^n+\cdots\in\mathbb{P}_n$ *be the* $L_q(x^{-1/2+(mq/2N)}w(x)\,\mathrm{d}x)$*-minimal polynomial on* $[0,1]$*. Then*

$$\mathscr{P}_{2nN+m}(z):=z^m p_{n,m}(z^{2N})$$

is the $L_q(w(z^{2N})|z^{N-1}|\,|\mathrm{d}z|)$*-minimal polynomial on the* $2N$*-star* S_{2N}*.*

Proof. Put

$$\mathrm{d}\tilde{\mu}(z)=w(z^{2N})|z^{N-1}|\,|\mathrm{d}z|.$$

Suppose that

$$\left\|\sum_{v=0}^{2nN+m}c_v z^{2nN+m-v}\right\|_{q,\tilde{\mu},S_{2N}}^q=\inf_{b_v\in\mathbb{C},b_0=1}\left\|\sum_{v=0}^{2nN+m}b_v z^{2nN+m-v}\right\|_{q,\tilde{\mu},S_{2N}}^q.$$

Since $z\in S_{2N}$ implies $\mathrm{e}^{\mathrm{i}j\pi/N}z\in S_{2N}$ for $j=0,\dots,2N-1$, it follows that

$$\left\|\sum_{v=0}^{2nN+m}c_v z^{2nN+m-v}\right\|_{q,\tilde{\mu},S_{2N}}^q=\left\||\mathrm{e}^{\mathrm{i}j\pi m/N}|\sum_{v=0}^{2nN+m}c_v\mathrm{e}^{-\mathrm{i}vj\pi/N}z^{2nN+m-v}\right\|_{q,\tilde{\mu},S_{2N}}^q,$$

which gives by the uniqueness of the best approximation that

$$c_v=c_v\mathrm{e}^{-\mathrm{i}vj\pi/N}\quad\text{for } j=0,\dots,2N-1.$$

This last fact implies that

$$c_v=0\quad\text{for } v\neq 2Nk,\quad k\in\{0,\dots,n-1\}.$$

Hence, the minimal polynomial is of the form $z^m r_{n,m}(z^{2N})$, where $r_{n,m}(z)$ is a monic polynomial of degree n. By characterization (2.1) it follows moreover, using the transformation $x=z^N$, that

$$\int_{S_{2N}}z^{2N\kappa}|z^m|^q|r_{n,m}(z^{2N})|^{q-2}\overline{r_{n,m}(z^{2N})}\,\mathrm{d}\tilde{\mu}(z)$$
$$=2N\int_0^1 x^{2\kappa}|x|^{mq/N}|r_{n,m}(x^2)|^{q-2}\overline{r_{n,m}(x^2)}w(x^2)\,|\mathrm{d}x|=0\quad\text{for }\kappa=0,\dots,n-1.$$

From these conditions we get, by using the transformation $y=x^2$, that $r_{n,m}(x)=p_{n,m}(x)$. □

2.3. Extremal polynomials on special lemniscates

Let $K = \mathbb{U}_r := \{|z| = r\}$, $r \in [1,\infty)$ fixed, and (for simplicity) $\mathrm{d}\mu(z) = |\mathrm{d}z|$. Then, by (2.1), $\{z^n\}_{n\in\mathbb{N}_0}$ is the sequence of monic L_∞- and $L_q(\mu)$-, $q \in [1,\infty)$, minimal polynomials on $\mathbb{U}_r$. Let $\mathscr{T}(z) = z^N - r$, $N \geqslant 2$ fixed. Then the inverse image of $\mathbb{U}_r$ under $\mathscr{T}$, i.e., $\mathscr{T}^{-1}(\mathbb{U}_r)$, gives lemniscates; compare the right picture in Fig. 2. By Theorem 3 the polynomials

$$\mathscr{T}^n(z) = (z^N - r)^n$$

are L_∞- and $L_q(\mu^{\mathscr{T},\mathscr{T}'})$-minimal polynomials on the lemniscates $\mathscr{T}^{-1}(\mathbb{U}_r)$ with respect to the measure

$$\mathrm{d}\mu^{\mathscr{T},\mathscr{T}'}(z) := |z^{N-1}|\,|\mathrm{d}z|.$$

But in fact, we are able to construct the whole sequence of orthogonal polynomials again, using the structure of the inverse image, i.e., of the lemniscate.

Proposition 6. *Let* $n \in \mathbb{N}_0$, $2 \leqslant N \in \mathbb{N}$, $m \in \{0,\dots,N-1\}$, $q \in [1,\infty)$, $r \in (0,\infty)$, $w : \{|z| = r\} \to \mathbb{R}^+$ *integrable, and let* $p_{n,m}(z) = z^n + \cdots \in \mathbb{P}_n$ *be the* $L_q((r + \operatorname{Re} z)^{qm/2N} w(z)\,|\mathrm{d}z|)$*-minimal polynomial on* $\{|z| = r\}$. *Then*

$$\mathscr{P}_{nN+m}(z) := z^m p_{n,m}(z^N - r)$$

is the $L_q(w(z^N - r)|z^{N-1}|\,|\mathrm{d}z|)$*-minimal polynomial on the lemniscate*

$$\mathscr{L}_{N,r} = \{z \in \mathbb{C}\colon |z^N - r| = r\}.$$

Proof. Obviously, $z \in \mathscr{L}_{N,r}$ implies that $\mathrm{e}^{2k\mathrm{i}\pi/N} z \in \mathscr{L}_{N,r}$ for $k \in \mathbb{Z}$. Thus it follows quite similarly as in the proof of Proposition 1 that the coefficients of the L_q-minimal polynomial $\sum_{v=0}^{nN+m} c_v z^{nN+m-v}$ on $\mathscr{L}_{N,r}$ with respect to the weight $|z^{N-1}| w(z^N - r)$ vanish for $v \neq N\kappa$, $\kappa \in \{0,\dots,n-1\}$. Hence,

$$\sum_{v=0}^{nN+m} c_v z^{nN+m-v} = z^m \Psi_{n,m}(z^N - r),$$

where $\Psi_{n,m}(z) = z^n + \cdots \in \mathbb{P}_n$. Now, using the fact that for $z \in \mathscr{L}_{N,r}$

$$z^N - r = r\mathrm{e}^{\mathrm{i}\varphi} \quad \text{i.e.,}\ z = \mathrm{e}^{\mathrm{i}\varphi/2 + 2k\mathrm{i}\pi/N}(2r\cos(\varphi/2))^{1/N}$$

for a $k \in \{0,\dots,N-1\}$, and thus

$$|z^2|^{mq} = \left((2r)^2 \frac{(1+\cos\varphi)}{2}\right)^{mq/N},$$

we obtain by (2.1) that

$$\int_{\mathscr{L}_{N,r}} (z^N - r)^j z^m \overline{z^m \Psi_{n,m}(z^N - r)}\,|z^m \Psi_{n,m}(z^N - r)|^{q-2} w(z^N - r)|z^{N-1}|\,|\mathrm{d}z|$$
$$= \frac{1}{N}\int_{|v|=r} v^j \overline{\Psi_{n,m}(v)}\,|\Psi_{n,m}(v)|^{q-2}(2r(r + \operatorname{Re} v))^{mq/2N} w(v)|\mathrm{d}v| = 0 \quad \text{for } j = 0,\dots,n-1,$$

which proves, by the uniqueness of the minimal polynomials, the assertion. □

For $q=2$, $r=1$, $w(z)=1$, and $N=2$, the result has been shown by Godoy and Marcellán [16] using completely different methods. Let us note that for the case $q=2$, $r=1$ the orthogonal polynomials $\Psi_{n,m}(\mathrm{e}^{\mathrm{i}\varphi})$ are explicitly known (see [30]) if for instance $(1+\operatorname{Re} z)^{m/N}w(z)$, $z=\mathrm{e}^{\mathrm{i}\varphi}$, $\varphi\in[-\pi,\pi]$, is a Jacobi weight.

2.4. Extremal polynomials on $\mathscr{T}$-invariant sets; in particular Julia sets

Now we apply our theory on L_q-minimal polynomials from Section 2 to Julia sets, i.e., to $\mathscr{T}$-invariant complex sets. For the case $q=2$ compare also [3,6,26] and for $q=\infty$ [19].

It is well known that for polynomials the Julia set $\mathscr{J}_{\mathscr{T}}$ is the boundary of the basin of attraction of the point infinity, i.e.,

$$\mathscr{J}_{\mathscr{T}}=\partial A(\infty):=\partial\{z\in\mathbb{C}:\mathscr{T}^{(n)}(z)\to\infty \text{ as } n\to\infty\},$$

where $\mathscr{T}^{(n)}=\mathscr{T}\circ\mathscr{T}\circ\cdots\circ\mathscr{T}$ (n-times). The complement of the Julia set

$$\mathscr{F}_{\mathscr{T}}:=\bar{\mathbb{C}}\backslash\mathscr{J}_{\mathscr{T}}$$

is called the Fatou set of the polynomial $\mathscr{T}$.

The Julia set is a completely $\mathscr{T}$-invariant set, i.e.,

$$\mathscr{T}^{(n)}(\mathscr{J}_{\mathscr{T}})=\mathscr{T}^{(-n)}(\mathscr{J}_{\mathscr{T}})=\mathscr{J}_{\mathscr{T}}\quad\text{for all } n\in\mathbb{N}.$$

Here, $\mathscr{T}^{(-n)}:=[\mathscr{T}^{(n)}]^{-1}$. Moreover, Barnsley et al. showed [3] that on every Julia set $\mathscr{J}_{\mathscr{T}}$ there exists a unique invariant measure μ, namely the equilibrium measure $\mu_{\mathscr{J}_{\mathscr{T}},e}=:\mu_e$, which satisfies

$$\mu=\mu^{\mathscr{T},\mathscr{T}'}=\mu_e;$$

recall definition (1.9). Hence, by Theorem 3 the following relations hold:

$$p_{nN,q}(z)=p_{n,q}(\mathscr{T}(z))\quad\text{and}\quad p_{N^n,q}(z)=\mathscr{T}^{(n)}(z)+\text{const}(q),\tag{2.10}$$

where, $(p_{n,q})$ denotes the sequence of $L_q(\mu)$-minimal polynomials.

Here, we are not only interested in the Julia sets and their L_q-minimal polynomials. Because of the difficult structure of Julia sets one is interested in sets as simple as possible by which the Julia set can be generated resp. approximated, and in the L_q-minimal polynomials on these simpler sets. The latter minimal polynomials should be approximants of the minimal polynomials on the Julia set. The idea is now the following: Let $\mathscr{T}(z)=\kappa z^N+\cdots$, $\kappa\neq 0$, be a (complex) polynomial of degree $N\geqslant 2$ which generates the Julia set $\mathscr{J}_{\mathscr{T}}$. In order to get approximants of $\mathscr{J}_{\mathscr{T}}$, let $M^{(0)}\subseteq\mathscr{J}_{\mathscr{T}}$ be an arbitrary subset and $\mu^{(0)}$ a positive Borel measure on $M^{(0)}$. Then we define iteratively for every $n\in\mathbb{N}$,

$$M^{(n)}:=\mathscr{T}^{-1}(M^{(n-1)})=\mathscr{T}^{(-n)}(M^{(0)}).\tag{2.11}$$

By [9, Corollary 2.2] we have $\mathscr{J}_{\mathscr{T}}=\overline{\bigcup_{n=0}^{\infty}M^{(n)}}$. With the help of the measure from (1.9) we can define measures $\mu^{(n)}$ on $M^{(n)}$ recursively by

$$\int_{\mathscr{T}^{-1}(B)}f(z)\,\mathrm{d}\mu^{(n)}(z):=\frac{1}{N}\sum_{j=1}^{N}\int_B f(\mathscr{T}_j^{-1}(z))\,\mathrm{d}\mu^{(n-1)}(z),\tag{2.12}$$

$f \in L_1(\mu^{(n)})$ and B a $\mu^{(n-1)}$-measurable subset of $M^{(n-1)}$. Here, for simplicity we took $S = \mathscr{T}'/N$. For each sequence of measures $\{\mu^{(n)}\}$ so defined, one can show that there holds

$$\mu^{(n)} \overset{\text{weakly}}{\to} \mu_{\mathscr{J}_{\mathscr{T}},e}, \tag{2.13}$$

cf., e.g., [6, Section VI] but also [3, Remark, p. 381]. Especially, $\mu_{\mathscr{J}_{\mathscr{T}},e}$ is independent of the initial measure $\mu^{(0)}$. Once we know the linear L_∞- resp. the $L_q(\mu^{(0)})$-, $q \in [1,\infty)$, minimal polynomial on $M^{(0)}$, i.e., $p_1(z)$, then by (2.10) we know the L_∞- resp. $L_q(\mu^{(n)})$-minimal polynomial of degree N^n on $M^{(n)}$. Certainly, the case of orthogonal polynomials, i.e., $q = 2$, is again of special interest.

Let us demonstrate this approach in some more detail at the example of a special class of Julia sets, the so-called *dendrites*.

Definition. A Julia set $\mathscr{J}_{\mathscr{T}}$ is called a dendrite if and only if both $\mathscr{J}_{\mathscr{T}}$ and the Fatou set $\mathscr{F}_{\mathscr{T}}$ are connected.

Some examples of dendrites, generated by polynomials of the form $\mathscr{T}(z) = z^2 + c$, e.g., for $c \in \{-2, 1\}$, can be found in [11, Chapter 14].

A known sufficient condition such that the Julia set is a dendrite is that all the finite critical points of $\mathscr{T}$ are strictly preperiodic, see, e.g., [4].

In what follows let $\mathscr{T}$ be a real polynomial with a dendrite as its Julia set. Then $\mathscr{J}_{\mathscr{T}}$ is symmetric with respect to the real axis and thus it always contains real points. Hence, the values

$$\alpha := \inf\{x \in \mathscr{J}_{\mathscr{T}}\colon x \text{ real}\} \quad \text{and} \quad \beta := \max\{x \in \mathscr{J}_{\mathscr{T}}\colon x \text{ real}\} \tag{2.14}$$

exist in $\mathbb{R}$. Moreover, by the symmetry with respect to the real axis and the connectivity of the Fatou set, there holds $[\alpha,\beta] \subseteq \mathscr{J}_{\mathscr{T}}$. In the special case when $\mathscr{T}$ is an odd-degree polynomial with positive leading coefficient then α is the largest and β the smallest fixed point of $\mathscr{T}$. Anyway, we will always assume that $(\alpha,\beta) \neq \emptyset$, Following the idea from (2.11) we put

$$M^{(0)} = [\alpha,\beta] \in \mathscr{J}_{\mathscr{T}} \quad \text{and} \quad M^{(n)} := \mathscr{T}^{(-n)}([\alpha,\beta]) = \{z \in \mathbb{C}\colon \mathscr{T}^{(n)}(z) \in [\alpha,\beta]\}. \tag{2.15}$$

Then

$$M^{(n)} \subseteq M^{(n+1)} \tag{2.16}$$

and by [9, Corollary 2.2]

$$\mathscr{J}_{\mathscr{T}} = \overline{\bigcup_{n=1}^{\infty} M^{(n)}} = \lim_{n\to\infty} M^{(n)} =: M^{(\infty)}. \tag{2.17}$$

Fig. 3 shows the graph of the polynomial $\mathscr{T}(z) = 2z^4 - 1$ and its Julia set $\mathscr{J}_{\mathscr{T}}$; here $[\alpha,\beta] = [-1,1]$. The first three sets $M^{(1)}$, $M^{(2)}$, and $M^{(3)}$ are plotted in Fig. 4.

For the rate of convergence of the $M^{(n)}$'s towards the Julia set $\mathscr{J}_{\mathscr{T}}$ compare Corollary 8 below. The following theorem describes the minimal polynomials on the sets $M^{(n)} = \mathscr{T}^{(-n)}([\alpha,\beta])$ as well as on the Julia set $\mathscr{J}_{\mathscr{T}}$ (i.e., on $M^{(\infty)}$) but also on the equipotential lines $A(M^{(n)},\rho)$ of $M^{(n)}$ and of $\mathscr{J}_{\mathscr{T}}$.

Theorem 7. *Suppose that $\mathscr{T}(z) = \kappa z^N + \cdots$, $\kappa \neq 0$ and $N \geqslant 2$, is a real polynomial given as above, let $q \in [1,\infty]$, $\rho \geqslant 0$ be fixed, and let $\mu^{(0)}$ be a given real Borel measure on $A([\alpha,\beta],\rho)$, where α*

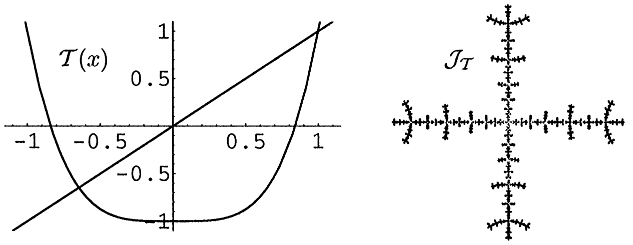


Fig. 3. Polynomial $\mathscr{T}(z)=2z^4-1$ and its Julia set $\mathscr{J}_{\mathscr{T}}$.

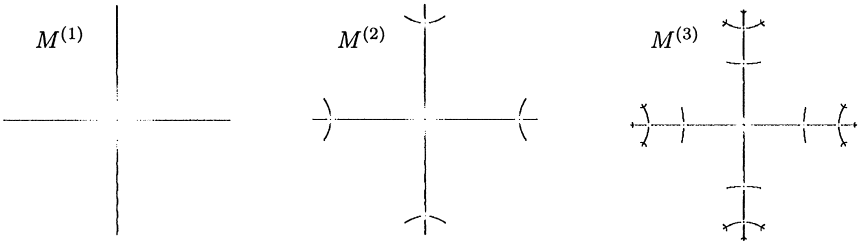


Fig. 4. Approximants $M^{(i)}$, $i=1,2$, and 3, for the Julia set $\mathscr{J}_{\mathscr{T}}$, where $\mathscr{T}(z)=2z^4-1$.

and β are defined by (2.14). *Furthermore, let $\mu^{(k)}$ be given by* (2.12) *and let $p_{1,k}(z)=z-c_k$ denote the monic linear L_∞- (if $q=\infty$) resp. $L_q(\mu^{(k)})$-minimal polynomial on $A(M^{(k)},\rho/N^k)$. Then for all $n\in\mathbb{N}$ and $k\in\mathbb{N}\cup\{\infty\}$*

$$\mathscr{P}_{n,k}(z)=\frac{\mathscr{T}^{(n)}(z)-c_k}{\kappa^{(N^n-1)/(N-1)}}=z^{N^n}+\cdots \tag{2.18}$$

is the monic L_∞- (if $q=\infty$) resp. $L_q(\mu^{(n+k)})$-minimal polynomial on the equipotential line $A(M^{(n+k)},\rho/N^{n+k})$. Moreover,

$$\lim_{k\to\infty}A(M^{(k)},\rho)=A(\mathscr{J}_{\mathscr{T}},\rho)\quad\textit{and}\quad\lim_{k\to\infty}c_k=c_\infty, \tag{2.19}$$

where c_k depends on q,ρ, and $\mu^{(0)}$, but c_∞ on q only. All the constants c_k are real.

Remark. Let us point out again that for $\rho=0$ all the equipotential lines in Theorem 7 are of the form $A(M^{(\nu)},0)$, $\nu\in\{k,k+n\}$, and can be replaced simply by $M^{(\nu)}$.

Proof of Theorem 7. Relation (2.18) follows immediately from Theorem 3 and

$$g(\mathscr{T}^{(n)}(z),M^{(k)},\infty)=N^n g(z,M^{(n+k)},\infty)$$

(compare also (2.20) below), which implies

$$\mathscr{T}^{(-n)}\left(A\left(M^{(k)},\frac{\rho}{N^k}\right)\right)=A\left(M^{(n+k)},\frac{\rho}{N^{n+k}}\right).$$

Concerning relation (2.19), let us note that (2.7) and Lemma 4(a) give

$$g(z,M^{(k)},\infty)=\frac{1}{N^k}\ln|l(\mathscr{T}^{(k)}(z))+\sqrt{[l(\mathscr{T}^{(k)}(z))]^2-1}|, \tag{2.20}$$

where l denotes the linear transformation from $[\alpha,\beta]$ to $[-1,1]$, i.e.,

$$l(x)=\frac{2x}{\beta-\alpha}-\frac{\beta+\alpha}{\beta-\alpha}.$$

Moreover, it is known that, see, e.g., [7,8],

$$g(z,\mathcal{J}_{\mathcal{T}},\infty)=\lim_{k\to\infty}\frac{1}{N^k}\ln|\mathcal{T}^{(k)}(z)|. \tag{2.21}$$

From these explicit representations one obtains the limit relation

$$\lim_{k\to\infty}[g(z,M^{(k)},\infty)-g(z,\mathcal{J}_{\mathcal{T}},\infty)]=0$$

uniformly on compact subsets of $\mathbb{C}\backslash\mathcal{J}_{\mathcal{T}}$ (note that $\mathcal{T}^{(k)}(z)\to\infty$ as $k\to\infty$ for all $z\in\mathbb{C}\backslash\mathcal{J}_{\mathcal{T}}$) and, consequently, by definition (2.3),

$$\lim_{k\to\infty}A(M^{(k)},\rho)=A(\mathcal{J}_{\mathcal{T}},\rho)\quad\text{for all }\rho>0 \tag{2.22}$$

with respect to the Hausdorff-metric. This is the first part of relation (2.19). For the remaining assertion concerning the coefficients c_k let us point out that all the sets $A(M^{(k)},\rho/N^k)$ and all the measures $\mu^{(k)}$ are symmetric with respect to the real axis. Hence, the values c_k have to be real. Furthermore, the constants c_k are convergent. Denote the limit by c_∞; then $p_{0,\infty}(z)=z-c_\infty$ is the L_∞- (if $q=\infty$) resp. the $L_q(\mu_e)$-minimal polynomial on $A(M^{(\infty)},0)=\mathcal{J}_{\mathcal{T}}$. Hence, c_∞ is independent both of ρ and $\mu^{(0)}$. □

Remark. For the case of L_q-minimal polynomials, $q\in[1,\infty)$, let us point out the nice fact that on the one hand, these minimal polynomials $\mathcal{P}_{n,k}(z)=:\mathcal{P}_{n,k}(z;\mu^{(0)})$ strongly depend on the initial measure $\mu^{(0)}$. But on the other hand, this $\mu^{(0)}$-dependence can be described completely, and in a simple and explicit way, only by the constants $c_k=c_k(q,\mu^{(0)})$.

The following corollary gives a feeling on how fast the "finite" sets $M^{(n)}$, which are approximants of the Julia set $\mathcal{J}_{\mathcal{T}}$, converge towards $\mathcal{J}_{\mathcal{T}}$, or to use different words, how good $\mathcal{J}_{\mathcal{T}}$ is describable by the $M^{(n)}$'s.

Corollary 8. *Let $\mathcal{T}(z)=\kappa z^N+\cdots$, $\kappa\neq 0$ and $N\geqslant 2$, by the polynomial from Theorem* 7. *Then*

$$\operatorname{cap}(M^{(n)})=\left(\frac{\beta-\alpha}{4}\right)^{1/N^n}|\kappa|^{-(N^n-1)/[N^n(N-1)]},\quad n=1,2,3,\ldots$$

and

$$\operatorname{cap}(\mathcal{J}_{\mathcal{T}})=|\kappa|^{-1/(N-1)}. \tag{2.23}$$

Proof. The first assertion follows from definition (2.5) and from the explicit representation of $g(z,M^{(n)},\infty)$ in (2.20). Next, from the identity

$$g(z,\mathcal{J}_{\mathcal{T}},\infty)=\ln|z|-\ln(\operatorname{cap}(\mathcal{J}_{\mathcal{T}}))+o(1)\quad\text{as } z\to\infty$$

and

$$g(\mathcal{T}^{(n)}(z),\mathcal{J}_{\mathcal{T}},\infty)=N^n g(z,\mathcal{J}_{\mathcal{T}},\infty),$$

which follows from Lemma 4 and the invariance of the Julia set, we obtain

$$\ln|\mathscr{T}^{(k)}(z)| - \ln(\operatorname{cap}(\mathscr{J}_{\mathscr{T}})) = N^k(\ln|z| - \ln(\operatorname{cap}(\mathscr{J}_{\mathscr{T}}))) + \mathrm{o}(1).$$

Now, the limiting process $z \to \infty$ gives

$$\ln|\kappa^{(N^k-1)/(N-1)}| = \ln([\operatorname{cap}(\mathscr{J}_{\mathscr{T}})]^{1-N^k}),$$

which is relation (2.23). □

Let us note that relation (2.23) is known (see, e.g., [27, Theorem 6.5.1]).

3. Asymptotics of polynomials orthogonal on inverse images of polynomials

In Propositions 5 and 6 we were able to get the whole sequence of orthogonal polynomials by symmetry reasons. In the general case this will certainly not be possible. Let us point out at this stage that polynomials orthogonal on curves or arcs in the complex plane (up to real intervals or arcs on the unit circle) do not satisfy a three-term recurrence relation in general. This is one of the reasons that they are so difficult to handle. Therefore, we are interested in asymptotics. To be able to obtain the asymptotics with the help of Widom's result, we have to derive some properties of harmonic measures. Let us recall the known fact that if K_j is a component of a compact nonpolar set $K \subset \mathbb{C}$, then the harmonic measure of K_j with respect to K at $z = \infty$, denoted by $\omega(\partial K_j, K, \infty)$, is given by

$$\omega(\partial K_j, K, \infty) = \mu_{K,e}(K_j), \tag{3.1}$$

see, e.g., [27, Theorem 4.3.14]. Again $\mu_{K,e}$ is the equilibrium measure on K.

The following lemma will play a crucial role. For the definition of a proper map and Riemann–Hurwitz formula see, e.g., [29, pp. 4–10].

Lemma 9. *Let* $\operatorname{int}(K)$ (*i.e.*, *the interior of* K) *be simply connected*, *let* $\operatorname{int}(Q_1), \ldots, \operatorname{int}(Q_l)$ *be the components of* $\mathscr{T}^{-1}(\operatorname{int}(K))$, *and assume that* $\mathscr{T}: \operatorname{int}(Q_j) \to \operatorname{int}(K)$ *is a* k_j*-fold proper map with at most* $k_j - 1$ *critical points in* Q_j. *Then the following relation holds for the harmonic measure*:

$$\omega(\partial Q_j, \mathscr{T}^{-1}(K), \infty) = \frac{k_j}{N}\mu_{K,e}(K). \tag{3.2}$$

Proof. By the Riemann–Hurwitz formula we have

$$l_{\operatorname{int}(Q_j)} - 2 = k_j(l_{\operatorname{int}(K)} - 2) + r_{\operatorname{int}(Q_j)}, \tag{3.3}$$

where $l_{\operatorname{int}(Q_j)}$ and $l_{\operatorname{int}(K)}$ denote the number of connectedness of $\operatorname{int}(Q_j)$ and $\operatorname{int}(K)$, respectively, and $r_{\operatorname{int}(Q_j)}$ the number of critical points of $\mathscr{T}$ in $\operatorname{int}(Q_j)$. Since by assumption $l_{\operatorname{int}(K)} = 1$, $r_{\operatorname{int}(Q_j)} \leqslant k_j - 1$, it follows that $l_{\operatorname{int}(Q_j)} = 1$, i.e., $\operatorname{int}(Q_j)$ is simply connected. Further, it is known that $\mathscr{T}: \partial Q_j \to \partial K$ is also a k_j-fold mapping, which implies by Lemma 4(c) and (1.7) that $\mu_{\mathscr{T}^{-1}(K),e}(Q_j) = \mu_{K,e}^{\mathscr{T},\mathscr{T}'}(Q_j) = (k_j/N)\mu_{K,e}(K)$. In view of (3.1) we have $\omega(\partial Q_j, \mathscr{T}^{-1}(K), \infty) = \mu_{K,e}^{\mathscr{T},\mathscr{T}'}(Q_j)$, which proves the statement. □

Remark. Let K be a Jordan arc and let the components $Q_1,\ldots,Q_l$ of $\mathscr{T}^{-1}(K)$ be Jordan arcs, too, and assume that $\mathscr{T}\colon Q_j \to K$, $j=1,\ldots,l$, is a k_j-fold proper map. Then relation (3.2) holds true obviously.

Naturally, Lemma 9 could also be stated for the more general case that K consists of several components, by considering the inverse image of each component of K. (But the assumptions become a little bit involved.)

Now we have the necessary ingredients to show with the help of Widom's theory on asymptotics of orthogonal polynomials [31] that polynomials orthogonal on Jordan curves or arcs, which are the inverse polynomial images of a Jordan curve or arc, have an asymptotically periodic behaviour. So far, this fact is known only for the case that the inverse polynomial image of $[-1,1]$ consists of several disjoint real intervals (see [2,22]). By the way, an asymptotically periodic behaviour of polynomials orthogonal on several arcs of the unit circle which are the inverse image of a trigonometric polynomial has been demonstrated by the authors in [24,25].

We say that the weight function ρ satisfies the generalized Szegő condition on $\mathscr{T}^{-1}(\partial K)$ if

$$\oint_{\mathscr{T}^{-1}(\partial K)} \ln \rho(\xi) \frac{\partial g(\xi, \mathscr{T}^{-1}(K), \infty)}{\partial n_\xi} |\mathrm{d}\xi| > -\infty \tag{3.4}$$

and we put

$$m_{n,\rho} := m_{n,\rho}(\mathscr{T}^{-1}(\partial K)) = \min_{\alpha_1,\ldots,\alpha_n\in\mathbb{C}} \int_{\mathscr{T}^{-1}(\partial K)} |\xi^n + \alpha_1\xi^{n-1} + \cdots + \alpha_n|^2 \rho(\xi)\,|\mathrm{d}\xi|.$$

Theorem 10. *Suppose that K and the components Q_j of $\mathscr{T}^{-1}(K)$ satisfy the assumptions of Lemma* 9 *resp. of the above remark. Furthermore, assume that the weight function ρ satisfies Szegő's condition* (3.4) *and let* (p_n) *be the monic polynomials of degree n hermitian orthogonal with respect to* $\rho(\xi)|\mathrm{d}\xi|$ *on* ∂K. *Then we have the following asymptotic behaviour with respect to n*:

$$m_{nN+j,p} \sim \mathrm{cap}(\mathscr{T}^{-1}(K))^{2(nN+j)} v_j,$$
$$p_{nN+j}(z)(\mathrm{cap}(\mathscr{T}^{-1}(K)))^{-(nN+j)}\Phi^{-(nN+j)}(z,\mathscr{T}^{-1}(K),\infty) \sim F_j(z)$$

for $j=0,\ldots,N-1$, *where the* v_j*'s and* F_j*'s are certain constants and functions, respectively* (*for details see the proof below*), Φ *is defined in* (2.6) *and* $a_n \sim b_n$ *means* $1-\varepsilon_n \leqslant a_n/b_n \leqslant 1+\varepsilon_n$, $\varepsilon_n \to 0^+$.

Proof. Let $\Upsilon = \bar{\mathbb{C}}\backslash\mathscr{T}^{-1}(K)$ and let F be a function analytic in Υ. Note that the standard analytic functions defined for the multi-connected region Υ have multi-valued argument in general. The ambiguity of the argument of a function in Υ is characterized as follows (compare [2, p. 237]): Let $\gamma=(\gamma_1,\ldots,\gamma_l)$ be a vector in $\mathbb{R}^l$. Take the coordinates of γ to be the increments in the argument of a multi-valued function $F(z)$ on marking circuits of the Q_j's, i.e.,

$$\gamma(F) = \left(\ldots, \frac{1}{2\pi} \underset{\partial Q_j}{\Delta} \arg F(z), \ldots\right). \tag{3.5}$$

We take the quotient of the functions analytic in Υ by the equivalence relation $F_1(z) \approx F_2(z) \Leftrightarrow \gamma(F_1)=\gamma(F_2)$. Note that $\gamma(F_1)=\gamma(F_2)$ is equivalent to the fact that $\arg(F_1-F_2)$ is single valued in

Υ. The classes obtained are denoted by Σ_γ, i.e.,

$$F(z) \in \Sigma_\gamma \quad \text{if } \gamma = \left(\dots, \frac{1}{2\pi} \mathop{\Delta}_{\partial Q_j} \arg F(z), \dots \right). \tag{3.6}$$

Next, let

$$\Phi(z) = \exp(g(z,\infty) + \mathrm{i}\tilde{g}(z,\infty)).$$

Here, $g(z,\infty)$ is Green's function for the set $\bar{\mathbb{C}} \backslash \Gamma$ with pole at ∞ and $\tilde{g}(z,\infty)$ is a harmonic conjugate. Further, let us set

$$\Sigma_k := -k\Sigma_{\gamma(\Phi)} \tag{3.7}$$

and recall (see (2.6)) that Φ denotes the conformal mapping of Υ onto the exterior of the unit disk. Note that by definition (3.5)

$$\gamma(\Phi) = (\omega(\partial Q_1, \mathscr{T}^{-1}(K), \infty), \dots, \omega(\partial Q_l, \mathscr{T}^{-1}(K), \infty)), \tag{3.8}$$

see, e.g., [31, p. 141], where $\omega(\partial Q_j, \mathscr{T}^{-1}(K), \infty)$ is the harmonic measure at $z=\infty$ of the jth component Q_j. Furthermore, for $\rho \in L_1(\Gamma)$ let $H_2(\Upsilon, \rho, \Sigma_\gamma)$ be the set of functions F from Σ_γ which are analytic on Υ and for which $|F(z)^2 \mathscr{R}(z)|$ has a harmonic majorant. Here, $\mathscr{R}(z)$ is the analytic function without zeros or poles in Υ whose modulus on Υ is single-valued and which takes the value $\rho(\xi)$ on Γ (see, e.g., [31, p. 155] or [2, p. 237]).

For weight functions ρ satisfying the Szegő condition (3.4), Widom has given the following asymptotic representation of the minimum deviation $m_{k,\rho}$ of monic polynomials $p_k(z)$ of degree k orthogonal with respect to $\rho(\xi)|\mathrm{d}\xi|$ [31, Theorem 12.3]:

$$m_{k,\rho} \sim (\operatorname{cap}(\mathscr{T}^{-1}(K)))^{2k} v(\rho, \Sigma_k)$$

and

$$p_k(z)(\operatorname{cap}(\mathscr{T}^{-1}(K)))^{-k} \Phi^{-k}(z, \mathscr{T}^{-1}(K), \infty) \sim F_k(z) \quad \text{for } A \subset \Upsilon,$$

A compact, where $F_k \in H_2(\Upsilon, \rho, \Sigma_k)$ is the unique solution of the following extremal problem:

$$v(\rho, \Sigma_k) = \inf_{\substack{F \in H_2(\Upsilon, \rho, \Sigma_k) \\ F(\infty)=1}} \int_{\mathscr{T}^{-1}(K)} |F(\xi)|^2 \rho(\xi)\, |\mathrm{d}\xi|, \tag{3.9}$$

hence,

$$v(\rho, \Sigma_k) = \int_{\mathscr{T}^{-1}(K)} |F_k(\xi)|^2 \rho(\xi)\, |\mathrm{d}\xi|.$$

Now, in the case under consideration we have by Lemma 9 that $\omega(\partial Q_j, \mathscr{T}^{-1}(K), \infty) = k_j/N$, $k_j \in \mathbb{N}$, for all $j = 1, \dots, l$, and thus we obtain from (3.5)–(3.8)

$$\Sigma_{j+nN} = \Sigma_j \bmod 1 \quad \text{for all } n \in \mathbb{N} \quad \text{and} \quad j = 0, 1, \dots, N-1$$

and therefore, we have by (3.9) and the uniqueness of the extremal function,

$$v(p, \Sigma_{nN+j}) = v(\rho, \Sigma_j),$$
$$F_{j+nN} \equiv F_j \quad \text{for all } n \in \mathbb{N} \quad \text{and} \quad j = 0, \dots, N-1. \qquad \square$$

For instance, the assumptions of Theorem 10 resp. Lemma 9 can be easily checked in the lemniscate case, that is, if the Q_j's are the components of $\mathscr{T}^{-1}(\{|z| = r|\})$. In particular, if r is sufficiently large then $\mathscr{T}^{-1}(\{|z| = r|\})$ contains all zeros of $\mathscr{T}'$ and consists of one component only. On the other hand, $\mathscr{T}^{-1}(\{|z| = r|\})$ consists of $\partial\mathscr{T} = N$ components containing no zero of $\mathscr{T}'$ if r, $r > 0$, is sufficiently small and $\mathscr{T}$ has simple zeros only.

Note that the expressions appearing in the above asymptotic formula can be simplified with the help of Lemma 9. Let us also point out that the only assumption on the weight function was that it satisfies the Szegő condition (3.4). If the weight function is in addition of form (1.10), then we know p_{nN} and m_{nN} and thus F_0 and v_0 explicitly. This leads us directly to our first problem.

Problem 1. It would be of great interest to know the remaining functions F_j, $j = 1, \ldots, N-1$, in Theorem 10 or, in other words, to find an explicit asymptotic expression for the remaining orthogonal polynomials. For the case that $\mathscr{T}^{-1}(K)$ is a subset of the real line or of the unit circumference, i.e., if the p_n's satisfy a recurrence relation, it is possible to derive (at least) asymptotic expressions for these functions (see [13,25]).

Problem 2. If $\mathscr{T}_N$ is a real polynomial which has N simple zeros in $(-1, 1)$ and satisfies (2.9) and $\mathscr{T}_N(\pm 1) = (\pm 1)^N$, then it can be shown easily that $\mathscr{J}_{\mathscr{T}_N}$ is a real Cantor set. Hence the orthonormal polynomials satisfy a three-term recurrence relation. Are the recurrence coefficients limit periodic, i.e., is the sequence of recurrence coefficients the limit of periodic sequences? For the special case $\mathscr{T}_N(x) = \alpha^N T_N(x/\alpha)$, $\alpha > 1$, $N \in \mathbb{N}\backslash\{1\}$, this is known [6].

Problem 3. Asymptotics of minimal polynomials with respect to the L_q-norm, $0 < q < \infty$, are so far known only in case of an interval or of a closed curve [15,21]. Is it possible to carry over Theorem 10 to the L_q-case in a suitable way (in this respect compare [18])? Recall what is needed are asymptotics of the extremal polynomials of degree $nN + j$, $j \in \{1, \ldots, N-1\}$, $n \in \mathbb{N}$.

Problem 4. Is there a bounded compact set $K \subset \mathbb{R}$ and a sequence of polynomials $p_n(x) = x^n + \cdots$, $(p_n) \neq (T_n)$, such that for each $n \in \mathbb{N}$, p_n is an L_q-extremal polynomial for every $q \in [1, \infty]$ with respect to the same weight function? What if we replace the condition "for every $q \in [1, \infty]$" by "for $q = 2$ and $q = \infty$", which certainly would be of foremost interest? We expect that there is no such sequence, that is, the classical Chebyshev polynomials T_n of the first kind with weight function $1/\sqrt{1 - x^2}$ are the only polynomials which have this property. Note that subsequences of polynomials with the above properties can easily be found, as we have shown in Section 2.1 in the statement after Fig. 1 or in [22, Theorem 2.4]. As we have learned in the meantime, the answer to the first question can be found in [28], which appeared very recently.

References

[1] N.I. Akhiezer, Theory of Approximation, Ungar, New York, 1956.

[2] A.I. Aptekarev, Asymptotic properties of polynomials orthogonal on a system of contours, and periodic motions of Toda lattices, Math. USSR Sb. 53 (1986) 233–260.

[3] M. Barnsley, J.S. Geronimo, H. Harrington, Orthogonal polynomials associated with invariant measures on Julia sets, Bull. Amer. Math. Soc. 7 (1982) 381–384.

[4] A.F. Beardon, Iteration of Rational Functions, Springer, New York, 1991.

[5] D. Bessis, Orthogonal polynomials, Padé approximations and Julia sets, in: P. Nevai (Ed.), Orthogonal Polynomials: Theory and Practice, NATO ASI Series, Series C: Mathematical and Physical Sciences, vol. 294, 1990, pp. 55–97.

[6] D. Bessis, P. Moussa, Orthogonality properties of iterated polynomial mappings, Comm. Math. Phys. 88 (1983) 503–529.

[7] B. Branner, The Mandelbrot set, in: R.L. Devaney and L. Keen (Eds.), Chaos and Fractals. The Mathematics behind the Computer Graphics, Proceedings of Symposia in Applied Mathematics, Vol. 39, AMS, Providence, RI, 1989, pp. 75–105.

[8] B. Branner, Puzzles and para-puzzles of quadratic and cubic polynomials, in: R.L. Devaney (Ed.), Complex Dynamical Systems. The Mathematics behind the Mandelbrot and Julia Sets, Proceedings of Symposia in Applied Mathematics, Vol. 49, AMS, Providence, RI, 1999, pp. 31–69.

[9] H. Brolin, Invariant sets under iteration of rational functions, Arkiv Mat. 6 (1965) 103–144.

[10] P.J. Davis, Interpolation and Approximation, Blaisdell Publishing Company, A Division of Ginn and Company, New York, 1963.

[11] K. Falconer, Fractal Geometry. Mathematical Foundations and Applications, Wiley, Chichester, 1990.

[12] B. Fischer, F. Peherstorfer, Comparing the convergence rate of Krylov subspace methods via polynomial mappings, submitted for publication.

[13] J.S. Geronimo, W. Van Assche, Orthogonal polynomials with asymptotically periodic recurrence coefficients, J. Approx. Theory 46 (1986) 251–283.

[14] J.S. Geronimo, W. Van Assche, Orthogonal polynomials on several intervals via a polynomial mapping, Trans. Amer. Math. Soc. (2) 308 (1988) 559–581.

[15] Ya.L. Geronimus, Some extremal problems in $L_p(\sigma)$ spaces, Math. Sb. 31 (1952) 3–23.

[16] E. Godoy, F. Marcellán, Orthogonal polynomials and rational modifications of measures, Canad. J. Math. 45 (1993) 930–943.

[17] R.A. Horn, On the moments of complex measures, Math. Z. 156 (1977) 1–11.

[18] V.A. Kaliaguine, On asympototics of L_p extremal polynomials on a complex curve $(0 < p < \infty)$, J. Approx. Theory 74 (1993) 226–236.

[19] S.O. Kamo, P.A. Borodin, Chebyshev polynomials for Julia sets, Mosc. Univ. Math. Bull. 49 (1994) 44–45.

[20] B.R. Kripke, Best approximation with respect to nearby norms, Numer. Math. 6 (1964) 103–105.

[21] D.S. Lubinsky, E.B. Saff, Strong Asymptotics for L_p-Extremal Polynomials $(1 < p)$ Associated with Weights on $[-1,1]$, Lecture Notes in Mathematics, Vol. 1287, Springer, Berlin, 1987, pp. 83–104.

[22] F. Peherstorfer, Orthogonal and extremal polynomials on several intervals, J. Comput. Appl. Math. 48 (1993) 187–205.

[23] F. Peherstorfer, Minimal polynomials for compact sets of the complex plane, Constr. Approx. 12 (1996) 481–488.

[24] F. Peherstorfer, R. Steinbauer, Asymptotic behaviour of orthogonal polynomials on the unit circle with asymptotically periodic reflection coefficients, J. Approx. Theory 88 (1997) 316–353.

[25] F. Peherstorfer, R. Steinbauer, Strong asymptotics of orthonormal polynomials with the aid of Green's function, SIAM J. Math. Anal., to be published.

[26] T.S. Pitcher, J.R. Kinney, Some connections between ergodic theory and the iteration of polynomials, Arkiv Mat. 8 (1968) 25–32.

[27] T. Ransford, Potential Theory in the Complex Plane, Cambridge University Press, Cambridge, 1995.

[28] Y.G. Shi, On Turán quadrature formulas for the Chebyshev weight, J. Approx. Theory 96 (1999) 101–110.

[29] N. Steinmetz, Rational Iterations. Complex Analytic Dynamical Systems, de Gruyter Studies in Mathematics, Vol. 16, de Gruyter, Berlin, 1993.

[30] G. Szegő, Orthogonal Polynomials, 4th Edition, American Mathematical Society Colloquim Publ., Vol. 23, Amer. Math. Soc., Providence, RI, 1975.

[31] H. Widom, Extremal polynomials associated with a system of curves in the complex plane, Adv. Math. 3 (1969) 127–232.

Journal of Computational and Applied Mathematics 127 (2001) 317–347

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Some classical multiple orthogonal polynomials[☆]

Walter Van Assche[*], Els Coussement

Department of Mathematics, Katholieke Universiteit Leuven, Celestijnenlaan 200B, B-3001 Leuven, Belgium

Received 14 September 1999; received in revised form 14 January 2000

1. Classical orthogonal polynomials

One aspect in the theory of orthogonal polynomials is their study as special functions. Most important orthogonal polynomials can be written as terminating hypergeometric series and during the twentieth century people have been working on a classification of all such hypergeometric orthogonal polynomial and their characterizations.

The *very classical orthogonal polynomials* are those named after Jacobi, Laguerre, and Hermite. In this paper we will always be considering monic polynomials, but in the literature one often uses a different normalization. Jacobi polynomials are (monic) polynomials of degree n which are orthogonal to all lower degree polynomials with respect to the weight function $(1-x)^\alpha(1+x)^\beta$ on $[-1,1]$, where $\alpha,\beta>-1$. The change of variables $x\mapsto 2x-1$ gives Jacobi polynomials on $[0,1]$ for the weight function $w(x)=x^\beta(1-x)^\alpha$, and we will denote these (monic) polynomials by $P_n^{(\alpha,\beta)}(x)$. They are defined by the orthogonality conditions

$$\int_0^1 P_n^{(\alpha,\beta)}(x)x^\beta(1-x)^\alpha x^k\,\mathrm{d}x=0,\quad k=0,1,\ldots,n-1. \tag{1.1}$$

The monic Laguerre polynomials $L_n^{(\alpha)}(x)$ (with $\alpha>-1$) are orthogonal on $[0,\infty)$ to all polynomials of degree less than n with respect to the weight $w(x)=x^\alpha\mathrm{e}^{-x}$ and hence satisfy the orthogonality conditions

$$\int_0^\infty L_n^{(\alpha)}(x)x^\alpha\mathrm{e}^{-x}x^k\,\mathrm{d}x=0,\quad k=0,1,\ldots,n-1. \tag{1.2}$$

Finally, the (monic) Hermite polynomials $H_n(x)$ are orthogonal to all lower degree polynomials with respect to the weight function $w(x)=\mathrm{e}^{-x^2}$ on $(-\infty,\infty)$, so that

$$\int_{-\infty}^\infty H_n(x)\mathrm{e}^{-x^2}x^k\,\mathrm{d}x=0,\quad k=0,1,\ldots,n-1. \tag{1.3}$$

[☆] The first author is a Research Director of the Belgian Fund for Scientific Research (FWO). Research supported by INTAS and FWO research project G.0278.97.

[*] Corresponding author.

PII: S0377-0427(00)00503-3

These three families of orthogonal polynomials can be characterized in a number of ways:

- Their weight functions w satisfy a first order differential equation with polynomial coefficients
$$\sigma(x)w'(x) = \rho(x)w(x), \tag{1.4}$$
with σ of degree at most two and ρ of degree one. This equation is known as *Pearson's equation* and also appears in probability theory, where the corresponding weights (densities) are known as the beta density (Jacobi), the gamma density (Laguerre), and the normal density (Hermite). Note however that for probability density functions one needs to normalize these weights appropriately. For the Jacobi weight we have $\sigma(x) = x(1-x)$, for the Laguerre weight we have $\sigma(x) = x$, and for the Hermite weight we see that $\sigma(x) = 1$, so that each family corresponds to a different degree of the polynomial σ.
- The derivatives of the very classical polynomials are again orthogonal polynomials of the same family but with different parameters (Sonin, 1887; Hahn, 1949). Indeed, integration by parts of the orthogonality relations and the use of Pearson's equation show that
$$\frac{\mathrm{d}}{\mathrm{d}x}P_n^{(\alpha,\beta)}(x) = nP_{n-1}^{(\alpha+1,\beta+1)}(x),$$
$$\frac{\mathrm{d}}{\mathrm{d}x}L_n^{(\alpha)}(x) = nL_{n-1}^{(\alpha+1)}(x),$$
$$\frac{\mathrm{d}}{\mathrm{d}x}H_n(x) = nH_{n-1}(x).$$
The differential operator $\mathrm{D} = \mathrm{d}/\mathrm{d}x$ therefore acts as a *lowering operator* that lowers the degree of the polynomial.
- Pearson's equation also gives rise to a *raising operator* that raises the degree of the polynomials. Indeed, integration by parts shows that
$$\frac{\mathrm{d}}{\mathrm{d}x}[x^\beta(1-x)^\alpha P_n^{(\alpha,\beta)}(x)] = -(\alpha+\beta+n)x^{\beta-1}(1-x)^{\alpha-1}P_{n+1}^{(\alpha-1,\beta-1)}(x), \tag{1.5}$$
$$\frac{\mathrm{d}}{\mathrm{d}x}[x^\alpha \mathrm{e}^{-x}L_n^{(\alpha)}(x)] = -x^{\alpha-1}\mathrm{e}^{-x}L_{n+1}^{(\alpha-1)}(x), \tag{1.6}$$
$$\frac{\mathrm{d}}{\mathrm{d}x}[\mathrm{e}^{-x^2}H_n(x)] = -2\mathrm{e}^{-x^2}H_{n+1}(x). \tag{1.7}$$
The raising operator is therefore of the form $\sigma(x)/w(x)\mathrm{D}w(x)$. Using this raising operation repeatedly gives the *Rodrigues formula* for these orthogonal polynomials:
$$\frac{\mathrm{d}^n}{\mathrm{d}x^n}[x^{\beta+n}(1-x)^{\alpha+n}] = (-1)^n(\alpha+\beta+n+1)_n x^\beta(1-x)^\alpha P_n^{(\alpha,\beta)}(x), \tag{1.8}$$
$$\frac{\mathrm{d}^n}{\mathrm{d}x^n}[x^{\alpha+n}\mathrm{e}^{-x}] = (-1)^n x^\alpha \mathrm{e}^{-x}L_n^{(\alpha)}(x), \tag{1.9}$$
$$\frac{\mathrm{d}^n}{\mathrm{d}x^n}\mathrm{e}^{-x^2} = (-1)^n 2^n \mathrm{e}^{-x^2}H_n(x). \tag{1.10}$$

The Rodrigues formula is therefore of the form

$$\frac{\mathrm{d}^n}{\mathrm{d}x^n}[\sigma^n(x)w(x)] = C_n w(x)P_n(x),$$

where C_n is a normalization constant (Hildebrandt, 1931).

- Combining the lowering and the raising operator gives a *linear second-order differential equation* for these orthogonal polynomials, of the form

$$\sigma(x)y''(x) + \tau(x)y'(x) = \lambda_n y(x), \tag{1.11}$$

where σ is a polynomial of degree at most 2 and τ a polynomial of degree at most 1, both independent of the degree n, and λ_n is a constant depending on n (Bochner, 1929).

The Laguerre polynomials and the Hermite polynomials are limiting cases of the Jacobi polynomials. Indeed, one has

$$\lim_{\alpha\to\infty} \alpha^n P_n^{(\alpha,\beta)}(x/\alpha) = L_n^{(\beta)}(x), \tag{1.12}$$

and

$$\lim_{\alpha\to\infty} 2^n \alpha^{n/2} P_n^{(\alpha,\alpha)}\left(\frac{x+\sqrt{\alpha}}{2\sqrt{\alpha}}\right) = H_n(x). \tag{1.13}$$

The Hermite polynomials are also a limit case of the Laguerre polynomials:

$$\lim_{\alpha\to\infty} (2\alpha)^{-n/2} L_n^{(\alpha)}(\sqrt{2\alpha}x + \alpha) = H_n(x). \tag{1.14}$$

In this respect the Jacobi, Laguerre and Hermite polynomials are in a hierarchy, with Jacobi leading to Laguerre and Laguerre leading to Hermite, and with a shortcut for Jacobi leading to Hermite. This is just a very small piece in a large table known as Askey's table which also contains classical orthogonal polynomials of a discrete variable (Hahn, Meixner, Kravchuk, and Charlier) for which the differential operator D needs to be replaced by difference operators Δ and ∇ on a linear lattice (a lattice with constant mesh, see [30]). Finally, allowing a quadratic lattice also gives Meixner–Pollaczek, dual Hahn, continuous Hahn, continuous dual Hahn, Racah, and Wilson polynomials, which are all in the Askey table. These polynomials have a number of q-extensions involving the q-difference operator and leading to the q-extension of the Askey table. In [2] Andrews and Askey suggest to define the classical orthogonal polynomials as those polynomials that are a limiting case of the ${}_4\varphi_3$-polynomials

$$R_n(\lambda(x);a,b,c,d;q) = {}_4\varphi_3\left(\begin{matrix} q^{-n}, q^{n+1}ab, q^{-x}, q^{x+1}cd \\ aq, bdq, cq \end{matrix}; q, q\right),$$

with $\lambda(x) = q^{-x} + q^{x+1}cd$ and $bdq = q^{-N}$ (these are the q-Racah polynomials) or the ${}_4\varphi_3$-polynomials

$$\frac{a^n W_n(x;a,b,c,d|q)}{(ab;q)_n(ac;q)_n(ad;q)_n} = {}_4\varphi_3\left(\begin{matrix} q^{-n}, q^{n-1}abcd, a\mathrm{e}^{\mathrm{i}\theta}, a\mathrm{e}^{-\mathrm{i}\theta} \\ ab, ac, ad \end{matrix}; q, q\right),$$

with $x = \cos\theta$ (these are the Askey–Wilson polynomials). All these classical orthogonal polynomials then have the following properties:

- they have a Rodrigues formula,

- an appropriate divided difference operator acting on them gives a set of orthogonal polynomials,
- they satisfy a second-order difference or differential equation in x which is of Sturm–Liouville type.

The classical orthogonal polynomials in this wide sense have been the subject of intensive research during the twentieth century. We recommend the report by Koekoek and Swarttouw [26], the book by Andrews et al. [3], and the books by Nikiforov and Uvarov [31], and Nikiforov, Suslov and Uvarov [30] for more material. Szegő's book [44] is still a very good source for the very classical orthogonal polynomials of Jacobi, Laguerre, and Hermite. For characterization results one should consult a survey by Al-Salam [1].

2. Multiple orthogonal polynomials

Recently, there has been a renewed interest in an extension of the notion of orthogonal polynomials known as multiple orthogonal polynomials. This notion comes from simultaneous rational approximation, in particular from Hermite–Padé approximation of a system of r functions, and hence has its roots in the nineteenth century. However, only recently examples of multiple orthogonal polynomials appeared in the (mostly Eastern European) literature. In this paper we will introduce multiple orthogonal polynomials using the orthogonality relations and we will only use weight functions. The extension to measures is straightforward.

Suppose we are given r weight functions $w_1, w_2, \ldots, w_r$ on the real line and that the support of each w_i is a subset of an interval Δ_i. We will often be using a multi-index $\boldsymbol{n} = (n_1, n_2, \ldots, n_r) \in \mathbb{N}^r$ and its length $|\boldsymbol{n}| = n_1 + n_2 + \cdots + n_r$.

- The r-vector of *type I multiple orthogonal polynomials* $(A_{\boldsymbol{n},1}, \ldots, A_{\boldsymbol{n},r})$ is such that each $A_{\boldsymbol{n},i}$ is a polynomial of degree $n_i - 1$ and the following orthogonality conditions hold:
$$\int x^k \sum_{j=1}^{r} A_{\boldsymbol{n},j}(x) w_j(x)\,\mathrm{d}x = 0, \quad k = 0, 1, 2, \ldots, |\boldsymbol{n}| - 2. \tag{2.15}$$
Each $A_{\boldsymbol{n},i}$ has n_i coefficients so that the type I vector is completely determined if we can find all the $|\boldsymbol{n}|$ unknown coefficients. The orthogonality relations (2.15) give $|\boldsymbol{n}| - 1$ linear and homogeneous relations for these $|\boldsymbol{n}|$ coefficients. If the matrix of coefficients has full rank, then we can determine the type I vector uniquely up to a multiplicative factor.
- The type II multiple orthogonal polynomial $P_{\boldsymbol{n}}$ is the polynomial of degree $|\boldsymbol{n}|$ that satisfies the following orthogonality conditions:
$$\int_{\Delta_1} P_{\boldsymbol{n}}(x) w_1(x) x^k\,\mathrm{d}x = 0, \quad k = 0, 1, \ldots, n_1 - 1, \tag{2.16}$$
$$\int_{\Delta_2} P_{\boldsymbol{n}}(x) w_2(x) x^k\,\mathrm{d}x = 0, \quad k = 0, 1, \ldots, n_2 - 1, \tag{2.17}$$
$$\vdots$$
$$\int_{\Delta_r} P_{\boldsymbol{n}}(x) w_r(x) x^k\,\mathrm{d}x = 0, \quad k = 0, 1, \ldots, n_r - 1. \tag{2.18}$$

This gives $|\boldsymbol{n}|$ linear and homogeneous equations for the $|\boldsymbol{n}|+1$ unknown coefficients of $P_{\boldsymbol{n}}(x)$. We will choose the type II multiple orthogonal polynomials to be monic so that the remaining $|\boldsymbol{n}|$ coefficients can be determined uniquely by the orthogonality relations, provided the matrix of coefficients has full rank.

In this paper the emphasis will be on type II multiple orthogonal polynomials. The unicity of multiple orthogonal polynomials can only be guaranteed under additional assumptions on the r weights. Two distinct cases for which the type II multiple orthogonal polynomials are given are as follows.

1. In an *Angelesco system* (Angelesco, 1918) the intervals Δ_i, on which the weights are supported, are disjoint, i.e., $\Delta_i \cap \Delta_j = \emptyset$ whenever $i \neq j$. Actually, it is sufficient that the open intervals $\overset{\circ}{\Delta}_i$ are disjoint, so that the closed intervals Δ_i are allowed to touch.

Theorem 1. *In an Angelesco system the Type II multiple orthogonal polynomial $P_{\boldsymbol{n}}(x)$ factors into r polynomials $\prod_{j=1}^r q_{n_j}(x)$, where each q_{n_j}* has exactly *n_j zeros on Δ_j.*

Proof. Suppose $P_{\boldsymbol{n}}(x)$ has $m_j < n_j$ sign changes on Δ_j at the points $x_1,\dots,x_{m_j}$. Let $Q_{m_j}(x) = (x-x_1)\cdots(x-x_{m_j})$, then $P_{\boldsymbol{n}}(x)Q_{m_j}(x)$ does not change sign on Δ_j, and hence

$$\int_{\Delta_j} P_{\boldsymbol{n}}(x)Q_{m_j}(x)w_j(x)\,\mathrm{d}x \neq 0.$$

But this is in contradiction with the orthogonality relation on Δ_j. Hence $P_{\boldsymbol{n}}(x)$ has at least n_j zeros on Δ_j. Now all the intervals Δ_j $(j=1,2,\dots,r)$ are disjoint, hence this gives at least $|\boldsymbol{n}|$ zeros of $P_{\boldsymbol{n}}(x)$ on the real line. The degree of this polynomial is precisely $|\boldsymbol{n}|$, so there are exactly n_j zeros on each interval Δ_j. □

2. For an *AT system* all the weights are supported on the same interval Δ, but we require that the $|\boldsymbol{n}|$ functions

$$w_1(x), xw_1(x), \dots, x^{n_1-1}w_1(x), w_2(x), xw_2(x), \dots, x^{n_2-1}w_2(x), \dots, w_r(x), xw_r(x), \dots, x^{n_r-1}w_r(x)$$

form a Chebyshev system on Δ for each multi-index $\boldsymbol{n}$. This means that every linear combination

$$\sum_{j=1}^{r} Q_{n_j-1}(x)w_j(x),$$

with Q_{n_j-1} a polynomial of degree at most n_j-1, has at most $|\boldsymbol{n}|-1$ zeros on Δ.

Theorem 2. *In an AT system the Type II multiple orthogonal polynomial $P_{\boldsymbol{n}}(x)$ has exactly $|\boldsymbol{n}|$ zeros on Δ. For the Type I vector of multiple orthogonal polynomials, the linear combination $\sum_{j=1}^r A_{\boldsymbol{n},j}(x)w_j(x)$ has exactly $|\boldsymbol{n}|-1$ zeros on Δ.*

Proof. Suppose $P_{\boldsymbol{n}}(x)$ has $m < |\boldsymbol{n}|$ sign changes on Δ at the points $x_1,\ldots,x_m$. Take a multi-index $\boldsymbol{m}=(m_1,m_2,\ldots,m_r)$ with $|\boldsymbol{m}| = m$ such that $m_i \leqslant n_i$ for every i and $m_j < n_j$ for some j and construct the function

$$Q(x) = \sum_{i=1}^{r} Q_i(x)w_i(x),$$

where each Q_i is a polynomial of degree $m_i - 1$ whenever $i \neq j$, and Q_j is a polynomial of degree m_j, satisfying the interpolation conditions

$$Q(x_k) = 0, \quad k = 1,2,\ldots,m,$$

and $Q(x_0) = 1$ for an additional point $x_0 \in \Delta$. This interpolation problem has a unique solution since we are dealing with a Chebyshev system. The function Q has already m zeros, and since we are in a Chebyshev system, it can have no additional sign changes. Furthermore, the function does not vanish identically since $Q(x_0) = 1$. Obviously $P_{\boldsymbol{n}}(x)Q(x)$ does not change sign on Δ, so that

$$\int_{\Delta} P_{\boldsymbol{n}}(x)Q(x)\,\mathrm{d}x \neq 0,$$

but this is in contrast with the orthogonality relations for the Type II multiple orthogonal polynomial. Hence $P_{\boldsymbol{n}}(x)$ has exactly $|\boldsymbol{n}|$ zeros on Δ.

The proof for the Type I multiple orthogonal polynomials is similar. First of all, since we are dealing with an AT system, the function

$$A(x) = \sum_{j=1}^{r} A_{\boldsymbol{n},j}(x)w_j(x)$$

has at most $|\boldsymbol{n}| - 1$ zeros on Δ. Suppose it has $m < |\boldsymbol{n}| - 1$ sign changes at the points $x_1,x_2,\ldots,x_m$, then we use the polynomial $Q_m(x) = (x - x_1)\cdots(x - x_m)$ so that $A(x)Q_m(x)$ does not change sign on Δ, and

$$\int_{\Delta} A(x)Q(x)\,\mathrm{d}x \neq 0,$$

which is in contradiciton with the orthogonality of the Type I multiple orthogonal polynomial. Hence $A(x)$ has exactly $|\boldsymbol{n}| - 1$ zeros on Δ. □

Orthogonal polynomials on the real line always satisfy a three-term recurrence relation. There are also finite-order recurrences for multiple orthogonal polynomials, and there are quite a few of recurrence relations possible since we are dealing with multi-indices. There is an interesting recurrence relation of order $r + 1$ for the Type II multiple orthogonal polynomials with nearly diagonal multi-indices. Let $n \in \mathbb{N}$ and write it as $n = kr + j$, with $0 \leqslant j < r$. The nearly diagonal multi-index $\boldsymbol{s}(n)$ corresponding to n is then given by

$$\boldsymbol{s}(n) = (\underbrace{k+1, k+1, \ldots, k+1}_{j \text{ times}}, \underbrace{k, k, \ldots, k}_{r-j \text{ times}}).$$

If we denote the corresponding multiple orthogonal polynomials by

$$P_n(x) = P_{s(n)}(x),$$

then the following recurrence relation holds:

$$xP_n(x) = P_{n+1}(x) + \sum_{j=0}^{r} a_{n,j} P_{n-j}(x), \tag{2.19}$$

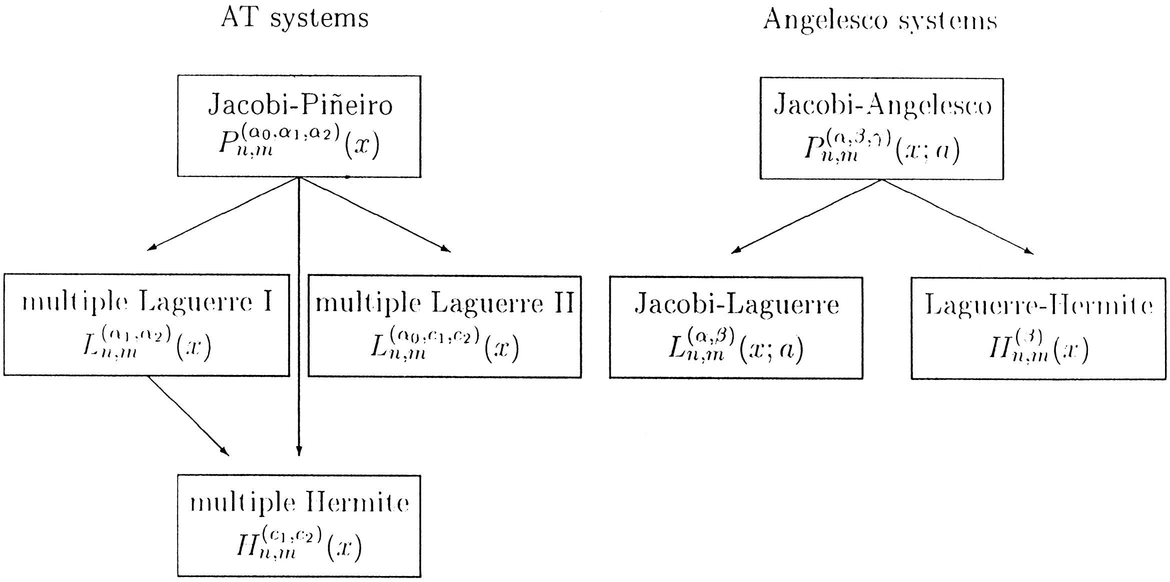
with initial conditions $P_0(x) = 1$, $P_j(x) = 0$ for $j = -1, -2, \ldots, -r$. The matrix

$$\begin{pmatrix}
a_{0,0} & 1 & & & & & & & \\
a_{1,1} & a_{1,0} & 1 & & & & & & \\
a_{2,2} & a_{2,1} & a_{2,0} & 1 & & & & & \\
\vdots & & & \ddots & \ddots & & & & \\
a_{r,r} & a_{r,r-1} & \cdots & & a_{r,0} & 1 & & & \\
 & a_{r+1,r} & \ddots & & & a_{r+1,0} & 1 & & \\
 & & \ddots & \ddots & & & \ddots & \ddots & \\
 & & & \ddots & \ddots & & & \ddots & 1 \\
 & & & & a_{n,r} & a_{n,r-1} & \cdots & a_{n,1} & a_{n,0}
\end{pmatrix}$$

has eigenvalues at the zeros of $P_{n+1}(x)$, so that in the case of Angelesco systems or AT systems we are dealing with nonsymmetric matrices with real eigenvalues. The infinite matrix will act as an operator on ℓ^2, but this operator is never self-adjoint and furthermore has not a simple spectrum, as is the case for ordinary orthogonal polynomials. Now there will be a set of r cyclic vectors and the spectral theory of this operator becomes more complicated (and more interesting). There are many open problems concerning this nonsymmetric operator.

3. Some very classical multiple orthogonal polynomials

We will now describe seven families of multiple orthogonal polynomials which have the same flavor as the very classical orthogonal polynomials of Jacobi, Laguerre, and Hermite. They certainly deserve to be called classical since they have a Rodrigues formula and there is a first-order differential operator which, when applied to these classical multiple orthogonal polynomials, gives another set of multiple orthogonal polynomials. However, these are certainly not the only families of classical multiple orthogonal polynomials (see Section 4.1). The first four families are AT systems which are connected by limit passages, the last three families are Angelesco systems which are also connected by limit passages. All these families have been introduced in the literature before. We will list some of their properties and give explicit formulas, most of which have not appeared earlier.


AT systems
Angelesco systems
Jacobi-Piñeiro $P_{n,m}^{(\alpha_0,\alpha_1,\alpha_2)}(x)$
Jacobi-Angelesco $P_{n,m}^{(\alpha,\beta,\gamma)}(x;a)$
multiple Laguerre I $L_{n,m}^{(\alpha_1,\alpha_2)}(x)$
multiple Laguerre II $L_{n,m}^{(\alpha_0;c_1,c_2)}(x)$
Jacobi-Laguerre $L_{n,m}^{(\alpha,\beta)}(x;a)$
Laguerre-Hermite $H_{n,m}^{(\beta)}(x)$
multiple Hermite $H_{n,m}^{(c_1,c_2)}(x)$

3.1. Jacobi–Piñeiro polynomials

The Jacobi–Piñeiro polynomials are multiple orthogonal polynomials associated with an AT system consisting of Jacobi weights on $[0,1]$ with different singularities at 0 and the same singularity at 1. They were first studied by Piñeiro [37] when $\alpha_0=0$. The general case appears in [34, p. 162]. Let $\alpha_0>-1$ and $\alpha_1,\ldots,\alpha_r$ be such that each $\alpha_i>-1$ and $\alpha_i-\alpha_j\notin\mathbb{Z}$ whenever $i\neq j$. The Jacobi–Piñeiro polynomial $P_{\boldsymbol{n}}^{(\alpha_0,\boldsymbol{\alpha})}$ for the multi-index $\boldsymbol{n}=(n_1,n_2,\ldots,n_r)\in\mathbb{N}^r$ and $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_r)$ is the monic polynomial of degree $|\boldsymbol{n}|=n_1+n_2+\cdots+n_r$ that satisfies the orthogonality conditions

$$\int_0^1 P_{\boldsymbol{n}}^{(\alpha_0,\boldsymbol{\alpha})}(x)x^{\alpha_1}(1-x)^{\alpha_0}x^k\,\mathrm{d}x=0,\quad k=0,1,\ldots,n_1-1, \tag{3.20}$$

$$\int_0^1 P_{\boldsymbol{n}}^{(\alpha_0,\boldsymbol{\alpha})}(x)x^{\alpha_2}(1-x)^{\alpha_0}x^k\,\mathrm{d}x=0,\quad k=0,1,\ldots,n_2-1, \tag{3.21}$$

$$\int_0^1 P_{\boldsymbol{n}}^{(\alpha_0,\boldsymbol{\alpha})}(x)x^{\alpha_r}(1-x)^{\alpha_0}x^k\,\mathrm{d}x=0,\quad k=0,1,\ldots,n_r-1. \tag{3.22}$$

Since each weight $w_i(x)=x^{\alpha_i}(1-x)^{\alpha_0}$ satisfies a Pearson equation

$$x(1-x)w_i'(x)=[\alpha_i(1-x)-\alpha_0x]w_i(x)$$

and the weights are related by

$$w_i(x)=x^{\alpha_i-\alpha_j}w_j(x),$$

one can use integration by parts on each of the r integrals (3.20)–(3.22) to find the following raising operators:

$$\frac{\mathrm{d}}{\mathrm{d}x}(x^{\alpha_j}(1-x)^{\alpha_0}P_{\boldsymbol{n}}^{(\alpha_0,\boldsymbol{\alpha})}(x)) = -(|\boldsymbol{n}|+\alpha_0+\alpha_j)x^{\alpha_j-1}(1-x)^{\alpha_0-1}P_{\boldsymbol{n}+\boldsymbol{e}_j}^{(\alpha_0-1,\boldsymbol{\alpha}-\boldsymbol{e}_j)}(x), \tag{3.23}$$

where $\boldsymbol{e}_j$ is the jth standard unit vector. Repeatedly using this raising operator gives the Rodrigues formula

$$(-1)^{|\boldsymbol{n}|}\prod_{j=1}^{r}(|\boldsymbol{n}|+\alpha_0+\alpha_j+1)_{n_j}P_{\boldsymbol{n}}^{(\alpha_0,\boldsymbol{\alpha})}(x) = (1-x)^{-\alpha_0}\prod_{j=1}^{r}\left[x^{-\alpha_j}\frac{\mathrm{d}^{n_j}}{\mathrm{d}x^{n_j}}x^{n_j+\alpha_j}\right](1-x)^{\alpha_0+|\boldsymbol{n}|}. \tag{3.24}$$

The product of the r differential operators $x^{-\alpha_j}\mathrm{D}^{n_j}x^{n_j+\alpha_j}$ on the right-hand side can be taken in any order since these operators are commuting.

The Rodrigues formula allows us to obtain an explicit expression. For the case $r=2$ we write

$$\begin{aligned}&(-1)^{n+m}(n+m+\alpha_0+\alpha_1+1)_n(n+m+\alpha_0+\alpha_2+1)_m P_{n,m}^{(\alpha_0,\alpha_1,\alpha_2)}(x)\\ &=(1-x)^{-\alpha_0}x^{-\alpha_1}\frac{\mathrm{d}^n}{\mathrm{d}x^n}x^{\alpha_1-\alpha_2+n}\frac{\mathrm{d}^m}{\mathrm{d}x^m}x^{\alpha_2+m}(1-x)^{\alpha_0+n+m}.\end{aligned} \tag{3.25}$$

The mth derivative can be worked out using the Rodrigues formula (1.8) for Jacobi polynomials and gives

$$(-1)^n(n+m+\alpha_0+\alpha_1+1)_n P_{n,m}^{(\alpha_0,\alpha_1,\alpha_2)}(x) = (1-x)^{-\alpha_0}x^{-\alpha_1}\frac{\mathrm{d}^n}{\mathrm{d}x^n}x^{\alpha_1+n}(1-x)^{\alpha_0+n}P_m^{(\alpha_0+n,\alpha_2)}(x).$$

Now use Leibniz' rule to work out the nth derivative:

$$\begin{aligned}&(-1)^n(n+m+\alpha_0+\alpha_1+1)_n P_{n,m}^{(\alpha_0,\alpha_1,\alpha_2)}(x)\\ &=(1-x)^{-\alpha_0}x^{-\alpha_1}\sum_{k=0}^{n}\binom{n}{k}\frac{\mathrm{d}^k}{\mathrm{d}x^k}x^{\alpha_1+n}\frac{\mathrm{d}^{n-k}}{\mathrm{d}x^{n-k}}(1-x)^{\alpha_0+n}P_m^{(\alpha_0+n,\alpha_2)}(x).\end{aligned}$$

In order to work out the derivative involving the Jacobi polynomial, we will use the following lemma.

Lemma 3. *Let $P_n^{(\alpha,\beta)}(x)$ be the nth degree monic Jacobi polynomial on* $[0,1]$. *Then for $\alpha>0$ and $\beta>-1$*

$$\frac{\mathrm{d}}{\mathrm{d}x}[(1-x)^{\alpha}P_n^{(\alpha,\beta)}(x)] = -(\alpha+n)(1-x)^{\alpha-1}P_n^{(\alpha-1,\beta+1)}(x), \tag{3.26}$$

and

$$\frac{\mathrm{d}^m}{\mathrm{d}x^m}[(1-x)^{\alpha}P_n^{(\alpha,\beta)}(x)] = (-1)^m(\alpha+n-m+1)_m(1-x)^{\alpha-m}P_n^{(\alpha-m,\beta+m)}(x). \tag{3.27}$$

Proof. First of all, observe that

$$\frac{\mathrm{d}}{\mathrm{d}x}[(1-x)^{\alpha}P_n^{(\alpha,\beta)}(x)] = (1-x)^{\alpha-1}(-\alpha P_n^{(\alpha,\beta)}(x)+(1-x)[P_n^{(\alpha,\beta)}(x)]'),$$

so that the right-hand side is $-(\alpha+n)(1-x)^{\alpha-1}Q_n(x)$, with Q_n a monic polynomial of degree n. Integrating by parts gives

$$\begin{aligned}&-(\alpha+n)\int_0^1 (1-x)^{\alpha-1}x^{\beta+k+1}Q_n(x)\,\mathrm{d}x\\&\quad = x^{\beta+k+1}(1-x)^{\alpha}P_n^{(\alpha,\beta)}(x)|_0^1-(\beta+k+1)\int_0^1 x^{\beta+k}(1-x)^{\alpha}P_n^{(\alpha,\beta)}(x)\,\mathrm{d}x.\end{aligned}$$

Obviously, when $\alpha>0$ and $\beta>-1$, then the integrated terms on the right-hand side vanish. The integral on the right-hand side vanishes for $k=0,1,\ldots,n-1$ because of orthogonality. Hence Q_n is a monic polynomial which is orthogonal to all polynomials of degree less than n with respect to the weight $x^{\beta+1}(1-x)^{\alpha-1}$, which proves (3.26). The more general expression (3.27) follows by applying (3.26) m times. □

By using this lemma we arrive at

$$\begin{aligned}&(n+m+\alpha_0+\alpha_1+1)_n P_{n,m}^{(\alpha_0,\alpha_1,\alpha_2)}(x)\\&\quad = n!\sum_{k=0}^{n}\binom{\alpha_1+n}{k}\binom{\alpha_0+m+n}{n-k}x^{n-k}(x-1)^k P_m^{(\alpha_0+k,\alpha_2+n-k)}(x).\end{aligned}$$

For the Jacobi polynomial we have the expansion

$$(\alpha+\beta+n+1)_n P_n^{(\alpha,\beta)}(x) = n!\sum_{j=0}^{n}\binom{\beta+n}{j}\binom{\alpha+n}{n-j}x^{n-j}(x-1)^j, \tag{3.28}$$

which can easily be obtained from the Rodrigues formula (1.8) by using Leibniz' formula, so that we finally find

$$\begin{aligned}&(n+m+\alpha_0+\alpha_1+1)_n(n+m+\alpha_0+\alpha_2+1)_m P_{n,m}^{(\alpha_0,\alpha_1,\alpha_2)}(x)\\&\quad = n!m!\sum_{k=0}^{n}\sum_{j=0}^{m}\binom{\alpha_1+n}{k}\binom{\alpha_0+m+n}{n-k}\binom{\alpha_2+n+m-k}{j}\binom{\alpha_0+k+m}{m-j}x^{n+m-k-j}(x-1)^{k+j}.\end{aligned} \tag{3.29}$$

We can explicitly find the first few coefficients of $P_{m,n}^{(\alpha_0,\alpha_1,\alpha_2)}(x)$ from this expression. We introduce the notation

$$\begin{aligned}K_{n,m} &= \frac{n!m!}{(n+m+\alpha_0+\alpha_1+1)_n(n+m+\alpha_0+\alpha_2+1)_m}\\&= \binom{\alpha_0+\alpha_1+2n+m}{n}^{-1}\binom{\alpha_0+\alpha_2+2m+n}{m}^{-1}.\end{aligned}$$

First let us check that the polynomial is indeed monic by working out the coefficient of x^{m+n}. This is given by

$$K_{n,m}\sum_{k=0}^{n}\sum_{j=0}^{m}\binom{\alpha_1+n}{k}\binom{\alpha_0+m+n}{n-k}\binom{\alpha_2+n+m-k}{j}\binom{\alpha_0+k+m}{m-j}.$$

The sum over j can be evaluated using the Chu–Vandermonde identity

$$\sum_{j=0}^{m} \binom{\alpha_2+n+m-k}{j}\binom{\alpha_0+k+m}{m-j} = \binom{\alpha_0+\alpha_2+n+2m}{m},$$

which is independent of k. The remaining sum over k can also be evaluated and gives

$$\sum_{k=0}^{n} \binom{\alpha_1+n}{k}\binom{\alpha_0+m+n}{n-k} = \binom{\alpha_0+\alpha_1+m+2n}{n},$$

and the double sum is therefore equal to $K_{n,m}^{-1}$, showing that this polynomial is indeed monic. Now let us write

$$P_{n,m}^{(\alpha_0,\alpha_1,\alpha_2)}(x) = x^{m+n} + A_{n,m}x^{n+m-1} + B_{n,m}x^{n+m-2} + C_{n,m}x^{n+m-3} + \cdots .$$

The coefficient $A_{n,m}$ of x^{m+n-1} is given by

$$-K_{n,m}\sum_{k=0}^{n}\sum_{j=0}^{m}(k+j)\binom{\alpha_1+n}{k}\binom{\alpha_0+m+n}{n-k}\binom{\alpha_2+n+m-k}{j}\binom{\alpha_0+k+m}{m-j}.$$

This double sum can again be evaluated using Chu–Vandermonde and gives

$$A_{n,m} = -\frac{n(\alpha_1+n)(\alpha_0+\alpha_2+n+m)+m(\alpha_2+n+m)(\alpha_0+\alpha_1+2n+m)}{(\alpha_0+\alpha_1+2n+m)(\alpha_0+\alpha_2+n+2m)}.$$

Similarly we can compute the coefficient $B_{n,m}$ of x^{n+m-2} and the coefficient $C_{n,m}$ of x^{m+n-3}, but the computation is rather lengthy. Once these coefficients have been determined, one can compute the coefficients in the recurrence relation

$$xP_n(x) = P_{n+1}(x) + b_nP_n(x) + c_nP_{n-1}(x) + d_nP_{n-2}(x),$$

where

$$P_{2n}(x) = P_{n,n}^{(\alpha_0,\alpha_1,\alpha_2)}(x), \quad P_{2n+1}(x) = P_{n+1,n}^{(\alpha_0,\alpha_1,\alpha_2)}(x).$$

Indeed, by comparing coefficients we have

$$b_{2n} = A_{n,n} - A_{n+1,n}, \quad b_{2n+1} = A_{n+1,n} - A_{n+1,n+1}, \tag{3.30}$$

which gives

$$\begin{aligned}
b_{2n} = {} & [36n^4 + (48\alpha_0 + 28\alpha_1 + 20\alpha_2 + 38)n^3 \\
& + (21\alpha_0^2 + 8\alpha_1^2 + 4\alpha_2^2 + 30\alpha_0\alpha_1 + 18\alpha_0\alpha_2 + 15\alpha_1\alpha_2 + 39\alpha_0 + 19\alpha_1 + 19\alpha_2 + 9)n^2 \\
& + (3\alpha_0^3 + 10\alpha_0^2\alpha_1 + 4\alpha_0^2\alpha_2 + 6\alpha_0\alpha_1^2 + 2\alpha_0\alpha_2^2 + 11\alpha_0\alpha_1\alpha_2 + 5\alpha_1^2\alpha_2 + 3\alpha_1\alpha_2^2 \\
& + 12\alpha_0^2 + 3\alpha_1^2 + 3\alpha_2^2 + 13\alpha_0\alpha_1 + 13\alpha_0\alpha_2 + 8\alpha_1\alpha_2 + 6\alpha_0 + 3\alpha_1 + 3\alpha_2)n \\
& + \alpha_0^2 + \alpha_0\alpha_1 + \alpha_2\alpha_1^2 + 2\alpha_2\alpha_1^2\alpha_0 + 2\alpha_0^2\alpha_1 + \alpha_1^2\alpha_0 + \alpha_2^2\alpha_0 + \alpha_2^2\alpha_1 + \alpha_0^3\alpha_1 \\
& + \alpha_0^2\alpha_1^2 + \alpha_2^2\alpha_0\alpha_1 + \alpha_2^2\alpha_1^2 + 2\alpha_2\alpha_0^2\alpha_1 + 3\alpha_2\alpha_1\alpha_0 + 2\alpha_2\alpha_0^2 + \alpha_1\alpha_2 + \alpha_0^3 + \alpha_0\alpha_2] \\
& \times (3n+\alpha_0+\alpha_2)^{-1}(3n+\alpha_0+\alpha_1)^{-1}(3n+\alpha_0+\alpha_2+1)^{-1}(3n+\alpha_0+\alpha_1+2)^{-1},
\end{aligned}$$

and

$$\begin{aligned}
b_{2n+1} = {} & [36n^4 + (48\alpha_0 + 20\alpha_1 + 28\alpha_2 + 106)n^3 \\
& +(21\alpha_0^2 + 4\alpha_1^2 + 8\alpha_2^2 + 18\alpha_0\alpha_1 + 30\alpha_0\alpha_2 + 15\alpha_1\alpha_2 + 105\alpha_0 + 41\alpha_1 + 65\alpha_2 + 111)n^2 \\
& +(3\alpha_0^3 + 4\alpha_0^2\alpha_1 + 10\alpha_0^2\alpha_2 + 2\alpha_0\alpha_1^2 + 6\alpha_0\alpha_2^2 + 11\alpha_0\alpha_1\alpha_2 + 3\alpha_1^2\alpha_2 + 5\alpha_1\alpha_2^2 \\
& +30\alpha_0^2 + 5\alpha_1^2 + 13\alpha_2^2 + 23\alpha_0\alpha_1 + 47\alpha_0\alpha_2 + 22\alpha_1\alpha_2 + 72\alpha_0 + 25\alpha_1 + 49\alpha_2 + 48)n \\
& +18\alpha_0\alpha_2 + 8\alpha_2\alpha_0^2 + 4\alpha_1 + 4\alpha_2^2\alpha_1 + 8\alpha_1\alpha_2 + 2\alpha_0^3 + 5\alpha_2^2\alpha_0 + 8\alpha_2\alpha_1\alpha_0 + 12\alpha_2 \\
& +7 + 15\alpha_0 + \alpha_2^2\alpha_1^2 + 10\alpha_0^2 + 6\alpha_0\alpha_1 + 2\alpha_2\alpha_1^2 + 2\alpha_0^2\alpha_1 + \alpha_1^2\alpha_0 + 5\alpha_2^2 + \alpha_2\alpha_0^3 \\
& +\alpha_2^2\alpha_0^2 + \alpha_1^2 + \alpha_2\alpha_1^2\alpha_0 + 2\alpha_2\alpha_0^2\alpha_1 + 2\alpha_2^2\alpha_0\alpha_1] \\
& \times (3n + \alpha_0 + \alpha_2 + 1)^{-1}(3n + \alpha_0 + \alpha_1 + 2)^{-1}(3n + \alpha_0 + \alpha_2 + 3)^{-1}(3n + \alpha_0 + \alpha_1 + 3)^{-1}.
\end{aligned}$$

For the recurrence coefficient c_n we have the formulas

$$c_{2n} = B_{n,n} - B_{n+1,n} - b_{2n}A_{n,n}, \quad c_{2n+1} = B_{n+1,n} - B_{n+1,n+1} - b_{2n+1}A_{n+1,n}, \tag{3.31}$$

which after some computation (and using Maple V), gives

$$\begin{aligned}
c_{2n} = {} & n(2n + \alpha_0)(2n + \alpha_0 + \alpha_1)(2n + \alpha_0 + \alpha_2) \\
& \times [54n^4 + (63\alpha_0 + 45\alpha_1 + 45\alpha_2)n^3 \\
& +(24\alpha_0^2 + 8\alpha_1^2 + 8\alpha_2^2 + 42\alpha_0\alpha_1 + 42\alpha_0\alpha_2 + 44\alpha_1\alpha_2 - 8)n^2 \\
& +(3\alpha_0^3 + \alpha_1^3 + \alpha_2^3 + 12\alpha_0^2\alpha_1 + 12\alpha_0^2\alpha_2 + 3\alpha_0\alpha_1^2 + 3\alpha_0\alpha_2^2 + 33\alpha_0\alpha_1\alpha_2 + 8\alpha_1^2\alpha_2 \\
& +8\alpha_1\alpha_2^2 - 3\alpha_0 - 4\alpha_1 - 4\alpha_2)n \\
& +\alpha_0^3\alpha_1 + \alpha_0^3\alpha_2 + 6\alpha_0^2\alpha_1\alpha_2 + \alpha_1^3\alpha_2 + \alpha_1\alpha_2^3 + 3\alpha_0\alpha_1^2\alpha_2 + 3\alpha_0\alpha_1\alpha_2^2 - \alpha_0\alpha_1 - \alpha_0\alpha_2 - 2\alpha_1\alpha_2] \\
& \times (3n + \alpha_0 + \alpha_1 + 1)^{-1}(3n + \alpha_0 + \alpha_2 + 1)^{-1}(3n + \alpha_0 + \alpha_1)^{-2}(3n + \alpha_0 + \alpha_2)^{-2} \\
& (3n + \alpha_0 + \alpha_1 - 1)^{-1}(3n + \alpha_0 + \alpha_2 - 1)^{-1}
\end{aligned}$$

and

$$\begin{aligned}
c_{2n+1} = {} & (2n + \alpha_0 + 1)(2n + \alpha_0 + \alpha_1 + 1)(2n + \alpha_0 + \alpha_2 + 1) \\
& \times [54n^5 + (63\alpha_0 + 45\alpha_1 + 45\alpha_2 + 135)n^4 \\
& + (24\alpha_0^2 + 8\alpha_1^2 + 8\alpha_2^2 + 42\alpha_0\alpha_1 + 42\alpha_0\alpha_2 + 44\alpha_1\alpha_2 + 126\alpha_0 + 76\alpha_1 + 104\alpha_2 + 120)n^3 \\
& + (3\alpha_0^3 + \alpha_1^3 + \alpha_2^3 + 12\alpha_0^2\alpha_1 + 12\alpha_0^2\alpha_2 + 3\alpha_0\alpha_1^2 + 3\alpha_0\alpha_2^2 + 33\alpha_0\alpha_1\alpha_2 + 8\alpha_1^2\alpha_2 \\
& + 8\alpha_1\alpha_2^2 + 36\alpha_0^2 + 5\alpha_1^2 + 19\alpha_2^2 + 54\alpha_0\alpha_1 + 72\alpha_0\alpha_2 + 66\alpha_1\alpha_2 + 87\alpha_0 + 39\alpha_1 \\
& + 81\alpha_2 + 45)n^2 \\
& +(\alpha_0^3\alpha_1 + \alpha_0^3\alpha_2 + 6\alpha_0^2\alpha_1\alpha_2 + \alpha_1^3\alpha_2 + \alpha_1\alpha_2^3 + 3\alpha_0\alpha_1^2\alpha_2 + 3\alpha_0\alpha_1\alpha_2^2 + 3\alpha_0^3 + 2\alpha_2^3 \\
& + 12\alpha_0^2\alpha_1 + 12\alpha_0^2\alpha_2 + 6\alpha_0\alpha_2^2 + 33\alpha_0\alpha_1\alpha_2 + 5\alpha_1^2\alpha_2 + 11\alpha_1\alpha_2^2 + 18\alpha_0^2 + 20\alpha_0\alpha_1 \\
& +38\alpha_0\alpha_2 + 14\alpha_2^2 + 26\alpha_1\alpha_2 + 24\alpha_0 + 6\alpha_1 + 24\alpha_2 + 6)n \\
& +\alpha_0^3\alpha_1 + 3\alpha_0^2\alpha_1\alpha_2 + 3\alpha_0\alpha_1\alpha_2^2 + \alpha_1\alpha_2^3 + \alpha_0^3 + \alpha_2^3 + 3\alpha_0^2\alpha_1 + 3\alpha_0^2\alpha_2 + 6\alpha_0\alpha_1\alpha_2 \\
& +3\alpha_0\alpha_2^2 + 3\alpha_1\alpha_2^2 + 3\alpha_0^2 + 3\alpha_2^2 + 2\alpha_0\alpha_1 + 6\alpha_0\alpha_2 + 2\alpha_1\alpha_2 + 2\alpha_0 + 2\alpha_2] \\
& \times (3n + \alpha_0 + \alpha_1 + 3)^{-1}(3n + \alpha_0 + \alpha_2 + 2)^{-1}(3n + \alpha_0 + \alpha_1 + 2)^{-2}(3n + \alpha_0 + \alpha_2 + 1)^{-2} \\
& (3n + \alpha_0 + \alpha_1 + 1)^{-1}(3n + \alpha_0 + \alpha_2)^{-1}.
\end{aligned}$$

Finally, for d_n we have

$$\begin{aligned} d_{2n} &= C_{n,n} - C_{n+1,n} - b_{2n}B_{n,n} - c_{2n}A_{n,n-1}, \\ d_{2n+1} &= C_{n+1,n} - C_{n+1,n+1} - b_{2n+1}B_{n+1,n} - c_{2n+1}A_{n,n}, \end{aligned} \tag{3.32}$$

giving

$$\begin{aligned} d_{2n} = {} & n(2n+\alpha_0)(2n+\alpha_0-1)(2n+\alpha_0+\alpha_1)(2n+\alpha_0+\alpha_1-1) \\ & (2n+\alpha_0+\alpha_2)(2n+\alpha_0+\alpha_2-1)(n+\alpha_1)(n+\alpha_1-\alpha_2) \\ & (3n+1+\alpha_0+\alpha_1)^{-1}(3n+\alpha_0+\alpha_1)^{-2}(3n+\alpha_0+\alpha_2)^{-1}(3n-1+\alpha_0+\alpha_1)^{-2} \\ & (3n-1+\alpha_0+\alpha_2)^{-1}(3n-2+\alpha_0+\alpha_1)^{-1}(3n-2+\alpha_0+\alpha_2)^{-1} \end{aligned}$$

and

$$\begin{aligned} d_{2n+1} = {} & n(2n+1+\alpha_0)(2n+\alpha_0)(2n+\alpha_0+\alpha_1)(2n+1+\alpha_0+\alpha_1) \\ & (2n+1+\alpha_0+\alpha_2)(2n+\alpha_0+\alpha_2)(n+\alpha_2)(n+\alpha_2-\alpha_1) \\ & (3n+2+\alpha_0+\alpha_1)^{-1}(3n+2+\alpha_0+\alpha_2)^{-1}(3n+1+\alpha_0+\alpha_1)^{-1}(3n+1+\alpha_0+\alpha_2)^{-2} \\ & (3n+\alpha_0+\alpha_1)^{-1}(3n+\alpha_0+\alpha_2)^{-2}(3n-1+\alpha_0+\alpha_2)^{-1}. \end{aligned}$$

These formulas are rather lengthy, but explicit knowledge of them will be useful in what follows. Observe that for large n we have

$$\begin{aligned} \lim_{n\to\infty} b_n &= \frac{4}{9} = 3\left(\frac{4}{27}\right), \\ \lim_{n\to\infty} c_n &= \frac{16}{243} = 3\left(\frac{4}{27}\right)^2, \\ \lim_{n\to\infty} d_n &= \frac{64}{19683} = \left(\frac{4}{27}\right)^3. \end{aligned}$$

3.2. Multiple Laguerre polynomials (first kind)

In the same spirit as for the Jacobi–Piñeiro polynomials, we can consider two different families of multiple Laguerre polynomials. The *multiple Laguerre polynomials of the first kind* $L_{\boldsymbol{n}}^{\boldsymbol{\alpha}}(x)$ are orthogonal on $[0,\infty)$ with respect to the r weights $w_j(x) = x^{\alpha_j}e^{-x}$, where $\alpha_j > -1$ for $j = 1,2,\ldots,r$. So these weights have the same exponential decrease at ∞ but have different singularities at 0. Again we assume $\alpha_i - \alpha_j \notin \mathbb{Z}$ in order to have an AT system. These polynomials were first considered by Sorokin [39,41]. The raising operators are given by

$$\frac{d}{dx}(x^{\alpha_j}e^{-x}L_{\boldsymbol{n}}^{\boldsymbol{\alpha}}(x)) = -x^{\alpha_j-1}e^{-x}L_{\boldsymbol{n}+e_j}^{\boldsymbol{\alpha}-e_j}(x), \quad j = 1,\ldots,r, \tag{3.33}$$

and a repeated application of these operators gives the Rodrigues formula

$$(-1)^{|\boldsymbol{n}|}L_{\boldsymbol{n}}^{\boldsymbol{\alpha}}(x) = e^x \prod_{j=1}^{r}\left[x^{-\alpha_j}\frac{d^{n_j}}{dx^{n_j}}x^{n_j+\alpha_j}\right]e^{-x}. \tag{3.34}$$

When $r=2$ one can use this Rodrigues formula to obtain an explicit expression for these multiple Laguerre polynomials, from which one can compute the recurrence coefficients in

$$xP_n(x)=P_{n+1}(x)+b_nP_n(x)+c_nP_{n-1}(x)+d_nP_{n-2}(x),$$

where $P_{2n}(x)=L_{n,n}^{(\alpha_1,\alpha_2)}(x)$ and $P_{2n+1}(x)=L_{n+1,n}^{(\alpha_1,\alpha_2)}(x)$. But having done all that work for Jacobi–Piñeiro polynomials, it is much easier to use the limit relation

$$L_{n,m}^{(\alpha_1,\alpha_2)}(x)=\lim_{\alpha_0\to\infty}\alpha_0^{n+m}P_{n,m}^{(\alpha_0,\alpha_1,\alpha_2)}(x/\alpha_0). \tag{3.35}$$

The recurrence coefficients can then be found in terms of the following limits of the corresponding recurrence coefficients of Jacobi–Piñeiro polynomials:

$$b_n=\lim_{\alpha_0\to\infty}b_n^{(\alpha_0,\alpha_1,\alpha_2)}\alpha_0,$$

$$c_n=\lim_{\alpha_0\to\infty}c_n^{(\alpha_0,\alpha_1,\alpha_2)}\alpha_0^2,$$

$$d_n=\lim_{\alpha_0\to\infty}d_n^{(\alpha_0,\alpha_1,\alpha_2)}\alpha_0^3,$$

giving

$$b_{2n}=3n+\alpha_1+1,$$

$$b_{2n+1}=3n+\alpha_2+2,$$

$$c_{2n}=n(3n+\alpha_1+\alpha_2),$$

$$c_{2n+1}=3n^2+(\alpha_1+\alpha_2+3)n+\alpha_1+1,$$

$$d_{2n}=n(n+\alpha_1)(n+\alpha_1-\alpha_2),$$

$$d_{2n+1}=n(n+\alpha_2)(n+\alpha_2-\alpha_1).$$

Observe that for large n we have

$$\lim_{n\to\infty}\frac{b_n}{n}=\frac{3}{2}=3\left(\frac{1}{2}\right),$$

$$\lim_{n\to\infty}\frac{c_n}{n^2}=\frac{3}{4}=3\left(\frac{1}{2}\right)^2,$$

$$\lim_{n\to\infty}\frac{d_n}{n^3}=\frac{1}{8}=\left(\frac{1}{2}\right)^3.$$

3.3. Multiple Laguerre polynomials (second kind)

Another family of multiple Laguerre polynomials is given by the weights $w_j(x)=x^{\alpha_0}e^{-c_jx}$ on $[0,\infty)$, with $c_j>0$ and $c_i\neq c_j$ for $i\neq j$. So now the weights have the same singularity at the origin but different exponential rates at infinity. These *multiple Laguerre polynomials of the second kind* $L_{\boldsymbol{n}}^{(\alpha_0,\boldsymbol{c})}(x)$ appear already in [34, p. 160]. The raising operators are

$$\frac{\mathrm{d}}{\mathrm{d}x}(x^{\alpha_0}\mathrm{e}^{-c_jx}L_{\boldsymbol{n}}^{(\alpha_0,\boldsymbol{c})}(x)) = -c_jx^{\alpha_0-1}\mathrm{e}^{-c_jx}L_{\boldsymbol{n}+\boldsymbol{e}_j}^{(\alpha_0-1,\boldsymbol{c})}(x), \quad j=1,\ldots,r, \tag{3.36}$$

and a repeated application of these operators gives the Rodrigues formula

$$(-1)^{|\boldsymbol{n}|}\prod_{j=1}^{r}c_j^{n_j}L_{\boldsymbol{n}}^{(\alpha_0,\boldsymbol{c})}(x) = x^{-\alpha_0}\prod_{j=1}^{r}\left[\mathrm{e}^{c_jx}\frac{\mathrm{d}^{n_j}}{\mathrm{d}x^{n_j}}\mathrm{e}^{-c_jx}\right]x^{|\boldsymbol{n}|+\alpha_0}. \tag{3.37}$$

These polynomials are also a limit case of the Jacobi–Piñeiro polynomials. For the case $r=2$ we have

$$L_{n,m}^{(\alpha_0,c_1,c_2)}(x) = \lim_{\alpha\to\infty}(-\alpha)^{n+m}P_{n,m}^{(\alpha_0,c_1\alpha,c_2\alpha)}(1-x/\alpha). \tag{3.38}$$

The recurrence coefficients can be obtained from the corresponding recurrence coefficients of Jacobi–Piñeiro polynomials by

$$b_n = \lim_{\alpha\to\infty}(1-b_n^{(\alpha_0,c_1\alpha,c_2\alpha)})\alpha,$$

$$c_n = \lim_{\alpha\to\infty}c_n^{(\alpha_0,c_1\alpha,c_2\alpha)}\alpha^2,$$

$$d_n = \lim_{\alpha\to\infty}-d_n^{(\alpha_0,c_1\alpha,c_2\alpha)}\alpha^3,$$

giving

$$b_{2n} = \frac{n(c_1+3c_2)+c_2+\alpha_0c_2}{c_1c_2},$$

$$b_{2n+1} = \frac{n(3c_1+c_2)+2c_1+c_2+\alpha_0c_1}{c_1c_2},$$

$$c_{2n} = \frac{n(2n+\alpha_0)(c_1^2+c_2^2)}{c_1^2c_2^2},$$

$$c_{2n+1} = \frac{2n^2(c_1^2+c_2^2)+n[c_1^2+3c_2^2+\alpha_0(c_1^2+c_2^2)]+c_2^2+\alpha_0c_2^2}{c_1^2c_2^2},$$

$$d_{2n} = \frac{n(2n+\alpha_0)(2n+\alpha_0-1)(c_2-c_1)}{c_1^3c_2},$$

$$d_{2n+1} = \frac{n(2n+\alpha_0)(2n+\alpha_0+1)(c_1-c_2)}{c_1c_2^3}.$$

Observe that for large n we have

$$\lim_{n\to\infty}\frac{b_n}{n} = \begin{cases}\dfrac{c_1+3c_2}{2c_1c_2} & \text{if } n\equiv 0\,(\mathrm{mod}\,2),\\[2ex] \dfrac{3c_1+c_2}{2c_1c_2} & \text{if } n\equiv 1\,(\mathrm{mod}\,2),\end{cases}$$

$$\lim_{n\to\infty} \frac{c_n}{n^2} = \frac{c_1^2+c_2^2}{2c_1^2c_2^2},$$

$$\lim_{n\to\infty} \frac{d_n}{n^3} = \begin{cases} \dfrac{c_2-c_1}{2c_1^3c_2} & \text{if } n \equiv 0 \,(\mathrm{mod}\, 2), \\ \dfrac{c_1-c_2}{2c_1c_2^3} & \text{if } n \equiv 1 \,(\mathrm{mod}\, 2). \end{cases}$$

3.4. Multiple Hermite polynomials

Finally we can consider the weights $w_j(x) = e^{-x^2+c_jx}$ on $(-\infty,\infty)$, for $j=1,2,\ldots,r$ and c_i different real numbers. The *multiple Hermite polynomials* $H_{\boldsymbol{n}}^{\boldsymbol{c}}(x)$ once more have raising operators and a Rodrigues formula, and they are also limiting cases of the Jacobi–Piñeiro polynomials, but also of the multiple Laguerre polynomials of the second kind. For $r=2$ this is

$$H_{n,m}^{(c_1,c_2)}(x) = \lim_{\alpha\to\infty} (2\sqrt{\alpha})^{n+m} P_{n,m}^{(\alpha,\alpha+c_1\sqrt{\alpha},\alpha+c_2\sqrt{\alpha})}\left(\frac{x+\sqrt{\alpha}}{2\sqrt{\alpha}}\right), \tag{3.39}$$

so that the recurrence coefficients can be obtained from the Jacobi–Piñeiro case by

$$b_n = \lim_{\alpha\to\infty} 2(b_n^{(\alpha,\alpha+c_1\sqrt{\alpha},\alpha+c_2\sqrt{\alpha})} - \tfrac{1}{2})\sqrt{\alpha},$$

$$c_n = \lim_{\alpha\to\infty} 4c_n^{(\alpha,\alpha+c_1\sqrt{\alpha},\alpha+c_2\sqrt{\alpha})}\alpha,$$

$$d_n = \lim_{\alpha\to\infty} 8d_n^{(\alpha,\alpha+c_1\sqrt{\alpha},\alpha+c_2\sqrt{\alpha})}(\sqrt{\alpha})^3.$$

This gives

$$b_{2n} = c_1/2,$$

$$b_{2n+1} = c_2/2,$$

$$c_n = n/2,$$

$$d_{2n} = n(c_1-c_2)/4,$$

$$d_{2n+1} = n(c_2-c_1)/4.$$

Alternatively, we can use the limit transition from the multiple Laguerre polynomials of the first kind:

$$H_{n,m}^{(c_1,c_2)}(x) = \lim_{\alpha\to\infty} \alpha_0^{n+m} L_{n,m}^{(\alpha+c_1\sqrt{\alpha/2},\alpha+c_2\sqrt{\alpha/2})}(\sqrt{2\alpha}x+\alpha). \tag{3.40}$$

The recurrence coefficients are then also given in terms of the following limits of the recurrence coefficients of the multiple Laguerre polynomials of the first kind

$$b_n = \lim_{\alpha\to\infty} \left(b_n^{(\alpha+c_1\sqrt{\alpha/2},\alpha+c_2\sqrt{\alpha/2})} - \alpha\right)/\sqrt{2\alpha},$$

$$c_n = \lim_{\alpha\to\infty} c_n^{(\alpha+c_1\sqrt{\alpha/2},\alpha+c_2\sqrt{\alpha/2})}/(2\alpha),$$

$$d_n = \lim_{\alpha\to\infty} d_n^{(\alpha+c_1\sqrt{\alpha/2},\alpha+c_2\sqrt{\alpha/2})}/(\sqrt{2\alpha})^3,$$

which leads to the same result. Observe that for large n we have

$$\lim_{n\to\infty} \frac{b_n}{\sqrt{n}} = 0,$$

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

$$\lim_{n\to\infty} \frac{d_n}{(\sqrt{n})^3} = 0.$$

3.5. Jacobi–Angelesco polynomials

The following system is probably the first that was investigated in detail [20,25]. It is an Angelesco system with weights $w_1(x) = |h(x)|$ on $[a,0]$ (with $a < 0$) and $w_2(x) = |h(x)|$ on $[0,1]$, where $h(x) = (x-a)^\alpha x^\beta (1-x)^\gamma$ and $\alpha,\beta,\gamma > -1$. Hence the same weight is used for both weights w_1 and w_2 but on two touching intervals. The Jacobi–Angelesco polynomials $P_{n,m}^{(\alpha,\beta,\gamma)}(x;a)$ therefore satisfy the orthogonality relations

$$\int_a^0 P_{n,m}^{(\alpha,\beta,\gamma)}(x;a)(x-a)^\alpha |x|^\beta (1-x)^\gamma x^k \,\mathrm{d}x = 0, \quad k = 0,1,2,\ldots,n-1, \tag{3.41}$$

$$\int_0^1 P_{n,m}^{(\alpha,\beta,\gamma)}(x;a)(x-a)^\alpha x^\beta (1-x)^\gamma x^k \,\mathrm{d}x = 0, \quad k = 0,1,2,\ldots,m-1. \tag{3.42}$$

The function $h(x)$ satisfies a Pearson equation

$$(x-a)x(1-x)h'(x) = [\alpha x(1-x) + \beta(x-a)(1-x) - \gamma(x-a)x]h(x),$$

where $\sigma(x) = (x-a)x(1-x)$ is now a polynomial of degree 3. Using this relation, we can integrate the orthogonality relations by part to see that

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}x}[(x-a)^\alpha x^\beta (1-x)^\gamma P_{n,m}^{(\alpha,\beta,\gamma)}(x;a)]\\ &\quad = -(\alpha+\beta+\gamma+n+m)(x-a)^{\alpha-1}x^{\beta-1}(1-x)^{\gamma-1}P_{n+1,m+1}^{(\alpha-1,\beta-1,\gamma-1)}(x;a),\end{aligned} \tag{3.43}$$

which raises both indices of the multi-index (n,m). Repeated use of this raising operation gives the Rodrigues formula

$$\begin{aligned}&\frac{\mathrm{d}^m}{\mathrm{d}x^m}[(x-a)^{\alpha+m}x^{\beta+m}(1-x)^{\gamma+m}P_{k,0}^{(\alpha+m,\beta+m,\gamma+m)}(x;a)]\\ &\quad = (-1)^m(\alpha+\beta+\gamma+k+2m+1)_m (x-a)^\alpha x^\beta (1-x)^\gamma P_{m+k,m}^{(\alpha,\beta,\gamma)}(x;a).\end{aligned} \tag{3.44}$$

For $k=0$ and $m=n$, this then gives

$$\frac{\mathrm{d}^n}{\mathrm{d}x^n}[(x-a)^{\alpha+n}x^{\beta+n}(1-x)^{\gamma+n}] \\ =(-1)^n(\alpha+\beta+\gamma+2n+1)_n(x-a)^\alpha x^\beta(1-x)^\gamma P_{n,n}^{(\alpha,\beta,\gamma)}(x;a). \tag{3.45}$$

Use Leibniz' formula to find

$$(-1)^n(\alpha+\beta+\gamma+2n+1)_n(x-a)^\alpha x^\beta(1-x)^\gamma P_{n,n}^{(\alpha,\beta,\gamma)}(x;a) \\ =\sum_{k=0}^{n}\binom{n}{k}\left(\frac{\mathrm{d}^k}{\mathrm{d}x^k}x^{\beta+n}(1-x)^{\gamma+n}\right)\left(\frac{\mathrm{d}^{n-k}}{\mathrm{d}x^{n-k}}(x-a)^{\alpha+n}\right).$$

Now use the Rodrigues formula for the Jacobi polynomials (1.8) to find

$$\binom{\alpha+\beta+\gamma+3n}{n}P_{n,n}^{(\alpha,\beta,\gamma)}(x;a) \\ =\sum_{k=0}^{n}(-1)^{n-k}\binom{\beta+\gamma+2n}{k}\binom{\alpha+n}{n-k}(x-a)^k x^{n-k}(1-x)^{n-k}P_k^{(\gamma+n-k,\beta+n-k)}(x).$$

Use of the expansion (3.28) for the Jacobi polynomial gives

$$\binom{\alpha+\beta+\gamma+3n}{n}P_{n,n}^{(\alpha,\beta,\gamma)}(x;a) \\ =\sum_{k=0}^{n}\sum_{j=0}^{k}\binom{\alpha+n}{n-k}\binom{\beta+n}{j}\binom{\gamma+n}{k-j}(x-a)^k x^{n-j}(x-1)^{n-k+j} \tag{3.46}$$

$$=\sum_{k=0}^{n}\sum_{j=0}^{n-k}\binom{\alpha+n}{k}\binom{\beta+n}{j}\binom{\gamma+n}{n-k-j}(x-a)^{n-k}x^{n-j}(x-1)^{k+j}, \tag{3.47}$$

where the last equation follows by the change of variable $k\mapsto n-k$. If we write this in terms of Pochhammer symbols, then

$$\binom{\alpha+\beta+\gamma+3n}{n}P_{n,n}^{(\alpha,\beta,\gamma)}(x;a) \\ =\frac{(\gamma+1)_n}{n!}\sum_{k=0}^{n}\sum_{j=0}^{n-k}\frac{(-n)_{k+j}(-\alpha-n)_k(-\beta-n)_j}{(\gamma+1)_{k+j}k!j!}(x-a)^{n-k}(x-1)^{k+j}x^{n-j} \\ =x^n(x-a)^n\binom{\gamma+n}{n}F_1\left(-n,-\alpha-n,-\beta-n,\gamma+1;\frac{x-1}{x-a},\frac{x-1}{x}\right),$$

where

$$F_1(a,b,b',c;x,y)=\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\frac{(a)_{m+n}(b)_m(b')_n}{(c)_{m+n}}\frac{x^m y^n}{m!n!}$$

is the first of Appell's hypergeometric functions of two variables.

For the polynomial $P_{n+1,n}^{(\alpha,\beta,\gamma)}(x;a)$ we have the Rodrigues formula

$$\begin{aligned}&\frac{\mathrm{d}^n}{\mathrm{d}x^n}[(x-a)^{\alpha+n}x^{\beta+n}(1-x)^{\gamma+n}P_{1,0}^{(\alpha+n,\beta+n,\gamma+n)}(x;a)]\\&\quad=(-1)^n(\alpha+\beta+\gamma+2n+2)_n(x-a)^{\alpha}x^{\beta}(1-x)^{\gamma}P_{n+1,n}^{(\alpha,\beta,\gamma)}(x;a),\end{aligned}\tag{3.48}$$

where $P_{1,0}^{(\alpha+n,\beta+n,\gamma+n)}(x;a)=x-X_n^{(\alpha,\beta,\gamma)}$ is the monic orthogonal polynomial of first degree for the weight $(x-a)^{\alpha+n}|x|^{\beta+n}(1-x)^{\gamma+n}$ on $[a,0]$. If we write down the orthogonality of this polynomial to the constant function,

$$\int_a^0 (x-X_n^{(\alpha,\beta,\gamma)})(x-a)^{\alpha+n}|x|^{\beta+n}(1-x)^{\gamma+n}\,\mathrm{d}x=0,$$

then we see that

$$X_n^{(\alpha,\beta,\gamma)}=\frac{\int_a^0 x(x-a)^{\alpha+n}|x|^{\beta+n}(1-x)^{\gamma+n}\,\mathrm{d}x}{\int_a^0 (x-a)^{\alpha+n}|x|^{\beta+n}(1-x)^{\gamma+n}\,\mathrm{d}x}.$$

A standard saddle point method gives the asymptotic behavior

$$\lim_{n\to\infty}X_n^{(\alpha,\beta,\gamma)}=x_1,\tag{3.49}$$

where x_1 is the zero of $\sigma'(x)$ in $[a,0]$, where $\sigma(x)=(x-a)x(1-x)$. Combining the Rodrigues equation in (3.48) with the Rodrigues equation (3.45) shows that

$$P_{n+1,n}^{(\alpha,\beta,\gamma)}(x;a)=xP_{n,n}^{(\alpha,\beta+1,\gamma)}(x;a)-X_n^{(\alpha,\beta,\gamma)}\frac{\alpha+\beta+\gamma+2n+1}{\alpha+\beta+\gamma+3n+1}P_{n,n}^{(\alpha,\beta,\gamma)}(x;a).\tag{3.50}$$

In order to compute the coefficients of the recurrence relation

$$xP_n(x)=P_{n+1}(x)+b_nP_n(x)+c_nP_{n-1}(x)+d_nP_{n-2}(x),$$

where

$$P_{2n}(x)=P_{n,n}^{(\alpha,\beta,\gamma)}(x;a),\qquad P_{2n+1}(x)=P_{n+1,n}^{(\alpha,\beta,\gamma)}(x;a),$$

we will compute the first few coefficients of the polynomials

$$P_{n,m}^{(\alpha,\beta,\gamma)}(x;a)=x^{m+n}+A_{n,m}x^{n+m-1}+B_{n,m}x^{m+n-2}+C_{n,m}x^{n+m-3}+\cdots.$$

First we take $n=m$. In order to check that our polynomial is monic, we see from (3.46) that the leading coefficient is given by

$$\binom{\alpha+\beta+\gamma+3n}{n}^{-1}\sum_{k=0}^{n}\sum_{j=0}^{k}\binom{\alpha+n}{n-k}\binom{\beta+n}{j}\binom{\gamma+n}{k-j}.$$

Chu–Vandermonde gives

$$\sum_{j=0}^{k} \binom{\beta+n}{j} \binom{\gamma+n}{k-j} = \binom{\beta+\gamma+2n}{k},$$

and also

$$\sum_{k=0}^{n} \binom{\alpha+n}{n-k} \binom{\beta+\gamma+2n}{k} = \binom{\alpha+\beta+\gamma+3n}{n},$$

so that the leading coefficient is indeed 1. The coefficient $A_{n,n}$ of x^{2n-1} is equal to

$$-\binom{\alpha+\beta+\gamma+3n}{n}^{-1} \sum_{k=0}^{n} \sum_{j=0}^{k} \binom{\alpha+n}{n-k} \binom{\beta+n}{j} \binom{\gamma+n}{k-j} (ak+n-k+j).$$

Working out this double sum gives

$$A_{n,n}^{(\alpha,\beta,\gamma)} = \frac{-n[\alpha+\beta+2n+a(\beta+\gamma+2n)]}{\alpha+\beta+\gamma+3n}. \tag{3.51}$$

For $P_{n+1,n}^{(\alpha,\beta,\gamma)}(x;a)$ the coefficient $A_{n+1,n}$ of x^{2n} can be obtained from (3.50)

$$A_{n+1,n}^{(\alpha,\beta,\gamma)} = A_{n,n}^{(\alpha,\beta+1,\gamma)} - X_n^{(\alpha,\beta,\gamma)} \frac{\alpha+\beta+\gamma+2n+1}{\alpha+\beta+\gamma+3n+1}. \tag{3.52}$$

The coefficient b_n in the recurrence relation can now be found from (3.30)

$$b_{2n} = \frac{n[n+\gamma+a(n+\alpha)]}{(\alpha+\beta+\gamma+3n)(\alpha+\beta+\gamma+3n+1)} + X_n^{(\alpha,\beta,\gamma)} \frac{2n+\alpha+\beta+\gamma+1}{3n+\alpha+\beta+\gamma+1},$$

$$\begin{aligned} b_{2n+1} = {} & (5n^2+(4\alpha+4\beta+3\gamma+7)n+(\alpha+\beta+\gamma+1)(\alpha+\beta+2) \\ & + a[5n^2+(3\alpha+4\beta+4\gamma+7)n+(\alpha+\beta+\gamma+1)(\beta+\gamma+2)]) \\ & \times (\alpha+\beta+\gamma+3n+1)^{-1}(\alpha+\beta+\gamma+3n+3)^{-1} \\ & - X_n^{(\alpha,\beta,\gamma)} \frac{2n+\alpha+\beta+\gamma+1}{3n+\alpha+\beta+\gamma+1}. \end{aligned}$$

The coefficient $B_{n,n}$ of x^{2n-2} in $P_{n,n}^{(\alpha,\beta,\gamma)}(x;a)$ is given by

$$\begin{aligned} B_{n,n}^{(\alpha,\beta,\gamma)} = {} & \frac{an(\alpha+\beta+\gamma+2n)(\beta+n)}{(\alpha+\beta+\gamma+3n)(\alpha+\beta+\gamma+3n-1)} \\ & + \frac{n(n-1)}{2(\alpha+\beta+\gamma+3n)(\alpha+\beta+\gamma+3n-1)} \\ & \times [(\alpha+\beta+2n)(\alpha+\beta+2n-1)+2a(\alpha+\beta+2n)(\beta+\gamma+2n) \\ & + a^2(\beta+\gamma+2n)(\beta+\gamma+2n-1)], \end{aligned}$$

and from (3.50) we also find

$$B_{n+1,n}^{(\alpha,\beta,\gamma)} = B_{n,n}^{(\alpha,\beta+1,\gamma)} - X_n^{(\alpha,\beta,\gamma)} A_{n,n}^{(\alpha,\beta,\gamma)} \frac{\alpha+\beta+\gamma+2n+1}{\alpha+\beta+\gamma+3n+1}.$$

Using (3.31) then gives

$$\begin{aligned} c_{2n} = {} & \frac{n(\alpha+\beta+\gamma+2n)}{(\alpha+\beta+\gamma+3n-1)(\alpha+\beta+\gamma+3n)^2(\alpha+\beta+\gamma+3n-1)} \\ & ((\alpha+\beta+2n)(\gamma+n) - 2a(\alpha+n)(\gamma+n) + a^2(\beta+\gamma+2n)(\alpha+n)). \end{aligned}$$

and

$$\begin{aligned} c_{2n+1} = {} & \frac{\alpha+\beta+\gamma+2n+1}{(\alpha+\beta+\gamma+3n+3)(\alpha+\beta+\gamma+3n+2)(\alpha+\beta+\gamma+3n+1)^2(\alpha+\beta+\gamma+3n)} \\ & \times (n(n+\gamma)(\alpha+\beta+2n+1)(\alpha+\beta+\gamma+3n+3) \\ & - a[24n^4 + (29\alpha+41\beta+29\gamma+48)n^3 \\ & + (10\alpha^2+39\alpha\beta+26\alpha\gamma+29\beta^2+39\beta\gamma+10\gamma^2+44\alpha+62\beta+44\gamma+30)n^2 \\ & + (\alpha^3+11\alpha^2\beta+5\alpha^2\gamma+19\alpha\beta^2+24\alpha\beta\gamma+5\alpha\gamma^2+9\beta^3+19\beta^2\gamma+11\beta\gamma^2+\gamma^3 \\ & + 11\alpha^2+39\alpha\beta+28\alpha\gamma+28\beta^2+39\beta\gamma+11\gamma^2+19\alpha+25\beta+19\gamma+6)n \\ & + (\alpha+\beta+\gamma)(\alpha+\beta+\gamma+1)(\alpha+\beta+\gamma+2)(\beta+1)] \\ & + a^2 n(n+\alpha)(\beta+\gamma+2n+1)(\alpha+\beta+\gamma+3n+3)) \\ & + \frac{\alpha+\beta+\gamma+2n+1}{(\alpha+\beta+\gamma+3n+3)(\alpha+\beta+\gamma+3n+1)^2(\alpha+\beta+\gamma+3n)} X_n^{(\alpha,\beta,\gamma)} \\ & \times (12n^3 + (16\alpha+16\beta+10\gamma+18)n^2 \\ & + [(\alpha+\beta+\gamma)(7\alpha+7\beta+2\gamma)+16\alpha+16\beta+10\gamma]n \\ & + (\alpha+\beta+\gamma)^2(\alpha+\beta) + (\alpha+\beta+\gamma)(3\alpha+3\beta+2\gamma+2) \\ & + a[12n^3 + (10\alpha+16\beta+16\gamma+18)n^2 \\ & + [(\alpha+\beta+\gamma)(2\alpha+7\beta+7\gamma)+10\alpha+16\beta+16\gamma]n \\ & + (\alpha+\beta+\gamma)^2(\beta+\gamma) + (\alpha+\beta+\gamma)(2\alpha+3\beta+3\gamma+2)]) \\ & - \frac{(\alpha+\beta+\gamma+2n+1)^2}{(\alpha+\beta+\gamma+3n+1)^2}(X_n^{(\alpha,\beta,\gamma)})^2. \end{aligned}$$

The coefficient $C_{n,n}$ of x^{2n-3} in $P_{n,n}^{(\alpha,\beta,\gamma)}(x;a)$ can be computed in a similar way, and the coefficient $C_{n+1,n}$ of x^{2n-2} in $P_{n+1,n}^{(\alpha,\beta,\gamma)}(x;a)$ is given by

$$C_{n+1,n}^{(\alpha,\beta,\gamma)} = C_{n,n}^{(\alpha,\beta+1,\gamma)} - X_n^{(\alpha,\beta,\gamma)} B_{n,n}^{(\alpha,\beta,\gamma)} \frac{\alpha+\beta+\gamma+2n+1}{\alpha+\beta+\gamma+3n+1}.$$

A lengthy but straightforward calculation, using (3.32), then gives

$$\begin{aligned}
d_{2n} = {} & \frac{-an(n+\beta)(\alpha+\beta+\gamma+2n)(\alpha+\beta+\gamma+2n-1)[n+\gamma+a(n+\alpha)]}{(\alpha+\beta+\gamma+3n-2)(\alpha+\beta+\gamma+3n-1)(\alpha+\beta+\gamma+3n)^2(\alpha+\beta+\gamma+3n+1)} \\
& + \frac{n(\alpha+\beta+\gamma+2n)(\alpha+\beta+\gamma+2n-1)X_{n-1}^{(\alpha,\beta,\gamma)}}{(\alpha+\beta+\gamma+3n-2)(\alpha+\beta+\gamma+3n-1)(\alpha+\beta+\gamma+3n)^2(\alpha+\beta+\gamma+3n+1)} \\
& \times [(n+\gamma)(\alpha+\beta+2n) - 2a(n+\gamma)(n+\alpha) + a^2(n+\alpha)(\beta+\gamma+2n)],
\end{aligned}$$

and

$$\begin{aligned}
d_{2n+1} = {} & \frac{n(\alpha+\beta+\gamma+2n+1)(\alpha+\beta+\gamma+2n)}{(\alpha+\beta+\gamma+3n+2)(\alpha+\beta+\gamma+3n+1)^2(\alpha+\beta+\gamma+3n)^2(\alpha+\beta+\gamma+3n-1)} \\
& \times ((n+\gamma)(\alpha+\beta+2n)(\alpha+\beta+2n+1) \\
& - a(n+\alpha)(n+\gamma)(2\alpha+2\beta-\gamma+3n+1) \\
& - a^2(n+\alpha)(n+\gamma)(-\alpha+2\beta+2\gamma+3n+1) \\
& + a^3(n+\alpha)(\beta+\gamma+2n)(\beta+\gamma+2n+1)) \\
& - \frac{n(\alpha+\beta+\gamma+2n+1)(\alpha+\beta+\gamma+2n)X_n^{(\alpha,\beta,\gamma)}}{(\alpha+\beta+\gamma+3n+1)^2(\alpha+\beta+\gamma+3n)^2(\alpha+\beta+\gamma+3n-1)} \\
& \times [(n+\gamma)(\alpha+\beta+2n) - 2a(n+\alpha)(n+\gamma) + a^2(n+\alpha)(\beta+\gamma+2n)].
\end{aligned}$$

The asymptotic behavior of these recurrence coefficients can easily be found using (3.49), giving

$$\lim_{n\to\infty} b_{2n} = \frac{a+1}{9} + \frac{2x_1}{3}, \qquad \lim_{n\to\infty} b_{2n+1} = \frac{5(a+1)}{9} - \frac{2x_1}{3},$$

$$\lim_{n\to\infty} c_{2n} = \frac{4}{81}(a^2-a+1), \qquad \lim_{n\to\infty} c_{2n+1} = -\frac{4}{9}x_1^2 + \frac{8}{27}x_1 + \frac{1}{81}(4a^2-a+4),$$

$$\lim_{n\to\infty} d_{2n} = \frac{4}{243}[2(a^2-a+1)x_1 - a(a+1)],$$

$$\lim_{n\to\infty} d_{2n+1} = \frac{4}{729}(4a^3-3a^2-3a+4) - \frac{8x_1}{243}(a^2-a+1),$$

where x_1 is the zero of $\sigma'(x)$ in $[a,0]$ and $\sigma(x)=(x-a)x(x-1)$. These formulas can be made more symmetric by also using the zero x_2 of $\sigma'(x)$ in $[0,1]$ and using the fact that $x_1+x_2=2(a+1)/3$:

$$\lim_{n\to\infty} b_{2n} = \frac{a+1}{9} + \frac{2x_1}{3}, \qquad \lim_{n\to\infty} b_{2n+1} = \frac{a+1}{9} - \frac{2x_2}{3},$$

$$\lim_{n\to\infty} c_n = \frac{4}{81}(a^2-a+1),$$

$$\lim_{n\to\infty} d_{2n} = -\frac{4}{27}\sigma(x_1), \qquad \lim_{n\to\infty} d_{2n} = -\frac{4}{27}\sigma(x_2).$$

3.6. Jacobi–Laguerre polynomials

When we consider the weights $w_1(x) = (x-a)^\alpha |x|^\beta \mathrm{e}^{-x}$ on $[a,0]$, with $a<0$, and $w_2(x) = (x-a)^\alpha |x|^\beta \mathrm{e}^{-x}$ on $[0,\infty)$, then we are again using one weight but on two touching intervals, one of which is the finite interval $[a,0]$ (Jacobi part), the other the infinite interval $[0,\infty)$ (Laguerre part). This system was considered by Sorokin [38]. The corresponding *Jacobi–Laguerre polynomials* $L_{n,m}^{(\alpha,\beta)}(x;a)$ satisfy the orthogonality relations

$$\int_a^0 L_{n,m}^{(\alpha,\beta)}(x;a)(x-a)^\alpha |x|^\beta \mathrm{e}^{-x} x^k \,\mathrm{d}x = 0, \quad k=0,1,\ldots,n-1,$$

$$\int_0^\infty L_{n,m}^{(\alpha,\beta)}(x;a)(x-a)^\alpha x^\beta \mathrm{e}^{-x} x^k \,\mathrm{d}x = 0, \quad k=0,1,\ldots,m-1.$$

The raising operator is

$$\frac{\mathrm{d}}{\mathrm{d}x}[(x+a)^\alpha x^\beta \mathrm{e}^{-x} L_{n,m}^{(\alpha,\beta)}(x;a)] = -(x-a)^{\alpha-1} x^{\beta-1} \mathrm{e}^{-x} L_{n+1,m+1}^{(\alpha-1,\beta-1)}(x;a), \tag{3.53}$$

from which the Rodrigues formula follows:

$$\frac{\mathrm{d}^m}{\mathrm{d}x^m}[(x-a)^{\alpha+m} x^{\beta+m} \mathrm{e}^{-x} L_{k,0}^{(\alpha+m,\beta+m)}(x;a)] = (-1)^m (x-a)^\alpha x^\beta \mathrm{e}^{-x} L_{m+k,m}^{(\alpha,\beta)}(x;a). \tag{3.54}$$

From this Rodrigues formula we can proceed as before to find an expression of the polynomials, but it is more convenient to view these Jacobi–Laguerre polynomials as a limit case of the Jacobi–Angelesco polynomials

$$L_{n,m}^{(\alpha,\beta)}(x;a) = \lim_{\gamma\to\infty} \gamma^{n+m} P_{n,m}^{(\alpha,\beta,\gamma)}(x/\gamma; a/\gamma), \tag{3.55}$$

so that (3.47) gives

$$L_{n,n}^{(\alpha,\beta)}(x;a) = \sum_{k=0}^{n} \sum_{j=0}^{n-k} \binom{\alpha+n}{k} \binom{\beta+n}{j} \frac{(-1)^{k+j}(x-a)^{n-k} x^{n-j}}{(n-k-j)!}. \tag{3.56}$$

For the recurrence coefficients in

$$xP_n(x) = P_{n+1}(x) + b_n P_n(x) + c_n P_{n-1}(x) + d_n P_{n-2}(x),$$

where $P_{2n}(x)=L_{n,n}^{(\alpha,\beta)}(x;a)$ and $P_{2n+1}(x)=L_{n+1,n}^{(\alpha,\beta)}(x;a)$ we have in terms of the corresponding recurrence coefficients of the Jacobi–Angelesco polynomials

$$b_n = \lim_{\gamma\to\infty} \gamma b_n^{(\alpha,\beta,\gamma)}(a/\gamma),$$

$$c_n = \lim_{\gamma\to\infty} \gamma^2 c_n^{(\alpha,\beta,\gamma)}(a/\gamma),$$

$$d_n = \lim_{\gamma\to\infty} \gamma^3 d_n^{(\alpha,\beta,\gamma)}(a/\gamma),$$

and

$$\lim_{\gamma\to\infty} \gamma X_n^{(\alpha,\beta,\gamma)}(a/\gamma) = \frac{\int_a^0 x(x-a)^{\alpha+n}|x|^{\beta+n}\mathrm{e}^{-x}\,\mathrm{d}x}{\int_a^0 (x-a)^{\alpha+n}|x|^{\beta+n}\mathrm{e}^{-x}\,\mathrm{d}x} := X_n^{(\alpha,\beta)}.$$

This gives

$$\begin{aligned}
&b_{2n} = n + X_n^{(\alpha,\beta)},\\
&b_{2n+1} = 3n + \alpha + \beta + 2 + a - X_n^{(\alpha,\beta)},\\
&c_{2n} = n(\alpha + \beta + 2n),\\
&c_{2n+1} = n(\alpha + \beta + 2n + 1) - a(n + \beta + 1) + (\alpha + \beta + 2n + 2 + a)X_n^{(\alpha,\beta)} - (X_n^{(\alpha,\beta)})^2,\\
&d_{2n} = -an(\beta + n) + n(\alpha + \beta + 2n)X_{n-1}^{(\alpha,\beta)},\\
&d_{2n+1} = n[(\alpha + \beta + 2n)(\alpha + \beta + 2n + 1) + a(n + \alpha)] - n(\alpha + \beta + 2n)X_n^{(\alpha,\beta)}.
\end{aligned}$$

For large n we have $X_n^{(\alpha,\beta)} = a/2 + o(1)$ so that

$$\lim_{n\to\infty} \frac{b_n}{n} = \begin{cases} 1/2 & \text{if } n \equiv 0\,(\mathrm{mod}\,2), \\ 3/2 & \text{if } n \equiv 1\,(\mathrm{mod}\,2), \end{cases}$$

$$\lim_{n\to\infty} \frac{c_n}{n^2} = 1/2,$$

$$\lim_{n\to\infty} \frac{d_n}{n^3} = \begin{cases} 0 & \text{if } n \equiv 0\,(\mathrm{mod}\,2), \\ 1/2 & \text{if } n \equiv 1\,(\mathrm{mod}\,2). \end{cases}$$

3.7. Laguerre–Hermite polynomials

Another limit case of the Jacobi–Angelesco polynomials are the multiple orthogonal polynomials $H_{n,m}^{(\beta)}(x)$ for which

$$\begin{aligned}
&\int_{-\infty}^{0} H_{n,m}^{(\beta)}(x)|x|^\beta \mathrm{e}^{-x^2} x^k \,\mathrm{d}x = 0, \quad k = 0, 1, \ldots, n-1,\\
&\int_{0}^{\infty} H_{n,m}^{(\beta)}(x)x^\beta \mathrm{e}^{-x^2} x^k \,\mathrm{d}x = 0, \quad k = 0, 1, \ldots, m-1.
\end{aligned}$$

We call these *Laguerre–Hermite polynomials* because both weights are supported on semi-infinite intervals (Laguerre) with a common weight that resembles the Hermite weight. These polynomials were already considered (for general r) by Sorokin [40]. The limit case is obtained by taking

$$H_{n,m}^{(\beta)}(x) = \lim_{\alpha\to\infty} (\sqrt{\alpha})^{n+m} P_{n,m}^{(\alpha,\beta,\alpha)}(x/\sqrt{\alpha}; -1). \tag{3.57}$$

This allows us to obtain the raising operator, the Rodrigues formula, an explicit expression, and the recurrence coefficients by taking the appropriate limit passage in the formulas for the Jacobi–Angelesco polynomials. For the recurrence coefficients this gives

$$b_n = \lim_{\alpha\to\infty} \sqrt{\alpha} b_n^{(\alpha,\beta,\alpha)}(a = -1),$$

$$c_n = \lim_{\alpha\to\infty} \alpha c_n^{(\alpha,\beta,\alpha)}(a = -1),$$

$$d_n = \lim_{\alpha\to\infty} (\sqrt{\alpha})^3 d_n^{(\alpha,\beta,\alpha)}(a = -1),$$

and

$$\lim_{\alpha\to\infty} \sqrt{\alpha} X_n^{(\alpha,\beta,\alpha)}(a=-1) = \frac{\int_{-\infty}^0 x|x|^{\beta+n} \mathrm{e}^{-x^2}\,\mathrm{d}x}{\int_{-\infty}^0 |x|^{\beta+n}\mathrm{e}^{-x^2}\,\mathrm{d}x} := X_n^{(\beta)},$$

from which we find

$$b_{2n} = X_n^{(\beta)},$$

$$b_{2n+1} = -X_n^{(\beta)},$$

$$c_{2n} = n/2,$$

$$c_{2n+1} = \frac{2n+\beta+1}{2} - (X_n^{(\beta)})^2,$$

$$d_{2n} = \frac{n}{2} X_{n-1}^{(\beta)},$$

$$d_{2n+1} = \frac{-n}{2} X_n^{(\beta)}.$$

For large n we have

$$X_n^{(\beta)} = -\sqrt{\frac{\beta+n}{2}} + \mathrm{o}(\sqrt{n}),$$

so that

$$\lim_{n\to\infty} \frac{b_n}{\sqrt{n}} = \begin{cases} -1/2 & \text{if } n \equiv 0\,(\text{mod}\,2), \\ 1/2 & \text{if } n \equiv 1\,(\text{mod}\,2), \end{cases}$$

$$\lim_{n\to\infty} \frac{c_n}{n} = 1/4,$$

$$\lim_{n\to\infty} \frac{d_n}{(\sqrt{n})^3} = \begin{cases} -1/8 & \text{if } n \equiv 0\,(\text{mod}\,2), \\ 1/8 & \text{if } n \equiv 1\,(\text{mod}\,2). \end{cases}$$

4. Open research problems

In the previous sections we gave a short description of multiple orthogonal polynomials and a few examples. For a more detailed account of multiple orthogonal polynomials we refer to Aptekarev [4] and Chapter 4 of the book of Nikishin and Sorokin [34]. Multiple orthogonal polynomials arise naturally in Hermite–Padé approximation of a system of (Markov) functions. For this kind of simultaneous rational approximation we refer to Mahler [28] and de Bruin [9,10]. Hermite–Padé approximation goes back to the nineteenth century, and many algebraic aspects have been investigated since then: existence and uniqueness, recurrences, normality of indices, etc. A study of Type II multiple orthogonal polynomials based on the recurrence relation can be found in Maroni

[29]. The more detailed analytic investigation of the zero distribution, the nth root asymptotics, and the strong asymptotics is more recent and mostly done by researchers from the schools around Nikishin [32,33] and Gonchar [18,19]. See in particular the work of Aptekarev [4], Kalyagin [20,25], Bustamante and López [11], and also the work by Driver and Stahl [15,16] and Nuttall [35]. First, one needs to understand the analysis of ordinary orthogonal polynomials, and then one has a good basis for studying this extension, for which there are quite a few possibilities for research.

4.1. Special functions

The research of orthogonal polynomials as special functions has now led to a classification and arrangement of various important (basic hypergeometric) orthogonal polynomials. In Section 3 we gave a few multiple orthogonal polynomials of the same flavor as the very classical orthogonal polynomials of Jacobi, Laguerre, and Hermite. Regarding these very classical multiple orthogonal polynomials, a few open problems arise:

(1) Are the polynomials given in Section 3 the only possible very classical multiple orthogonal polynomials? The answer very likely is no. First one needs to make clear what the notion of classical multiple orthogonal polynomial means. A possible way is to start from a Pearson type equation for the weights. If one chooses one weight but restricted to disjoint intervals, as we did for the Jacobi–Angelesco, Jacobi–Laguerre, and Laguerre–Hermite polynomials, then Aptekarev et al. [7] used the Pearson equation for this weight as the starting point of their characterization. For several weights it is more natural to study a Pearson equation for the vector of weights $(w_1, w_2, \dots, w_r)$. Douak and Maroni [13,14] have given a complete characterization of all Type II multiple orthogonal polynomials for which the derivatives are again Type II multiple orthogonal polynomials (Hahn's characterization for the Jacobi, Laguerre, and Hermite polynomials, and the Bessel polynomials if one allows moment functionals which are not positive definite). They call such polynomials classical d-orthogonal polynomials, where d corresponds to our r, i.e., the number of weights (functionals) needed for the orthogonality. Douak and Maroni show that this class of multiple orthogonal polynomials is characterized by a Pearson equation of the form

$$(\boldsymbol{\Phi w})' + \boldsymbol{\Psi w} = \mathbf{0},$$

where $\boldsymbol{w} = (w_1, \dots, w_r)^{\mathrm{t}}$ is the vector of weights, and Ψ and Φ are $r \times r$ matrix polynomials:

$$\Psi(x) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & r-1 \\ \psi(x) & c_1 & c_2 & \cdots & c_{r-1} \end{pmatrix},$$

with $\psi(x)$ a polynomial of degree one and $c_1, \dots, c_{r-1}$ constants, and

$$\Phi(x) = \begin{pmatrix} \phi_{1,1}(x) & \phi_{1,2}(x) & \cdots & \phi_{1,r}(x) \\ \phi_{2,1}(x) & \phi_{2,2}(x) & \cdots & \phi_{2,r}(x) \\ \vdots & \vdots & \cdots & \vdots \\ \phi_{r,1}(x) & \phi_{r,2}(x) & \cdots & \phi_{r,r}(x) \end{pmatrix},$$

where $\phi_{i,j}(x)$ are polynomials of degree at most two. In fact only $\phi_{r,1}$ can have degree at most two and all other polynomials are constant or of degree one, depending on their position in the matrix Φ. Douak and Maroni actually investigate the more general case where orthogonality is given by r linear functionals, rather than by r positive measures. We believe that Hahn's characterization is not the appropriate property to define classical multiple orthogonal polynomials, but gives a more restricted class. None of the seven families, given in the present paper, belong to the class studied by Douak and Maroni, but their class certainly contains several interesting families of multiple orthogonal polynomials. In fact, the matrix Pearson equation could result from a single weight (and its derivatives) satisfying a higher-order differential equation with polynomial coefficients. As an example, one can have multiple orthogonal polynomials with weights $w_1(x) = 2x^{\alpha+v/2}K_v(2\sqrt{x})$ and $w_2(x)=2x^{\alpha+(v+1)/2}K_{v+1}(2\sqrt{x})$ on $[0,\infty)$, where $K_v(x)$ is a modified Bessel function and $\alpha > -1, v \geqslant 0$ (see [47,12]).

(2) The polynomials of Jacobi, Laguerre, and Hermite all satisfy a linear second-order differential equation of Sturm–Liouville type. A possible way to extend this characterizing property is to look for multiple orthogonal polynomials satisfying a linear differential equation of order $r+1$. Do the seven families in this paper have such a differential equation? If the answer is yes, then an explicit construction would be desirable. We only worked out in detail the case where $r=2$, so the search is for a third-order differential equation for all the polynomials considered in Section 3. Such a third-order equation has been found for certain Jacobi–Angelesco systems in [25]. For the Angelesco systems in Section 3 this third-order differential equation indeed exists and it was constructed in [7]. The existence (and construction) is open for the AT systems. A deeper problem is to characterize all the multiple orthogonal polynomials satisfying a third order (order $r+1$) differential equation, extending Bochner's result for ordinary orthogonal polynomials. Observe that we already know appropriate raising operators for the seven systems described in Section 3. If one can construct lowering operators as well, then a combination of the raising and lowering operators will give the differential equation, which will immediately be in factored form. Just differentiating will usually not be sufficient (except for the class studied by Douak and Maroni): if we take $P'_{n,m}(x)$, then this is a polynomial of degree $n+m-1$, so one can write it as $P_{n-1,m}(x)+$ lower order terms, but also as $P_{n,m-1}(x)+$ lower-order terms. So it is not clear which of the multi-indices has to be lowered. Furthermore, the lower-order terms will not vanish in general since there usually are not enough orthogonality conditions to make them disappear.

(3) In the present paper we only considered the Type II multiple orthogonal polynomials. Derive explicit expressions and relevant properties of the corresponding vector $(A_{n,m}(x), B_{n,m}(x))$ of Type I multiple orthogonal polynomials. Type I and Type II multiple orthogonal polynomials are connected by

$$P_{n,m}(x) = \text{const.} \begin{vmatrix} A_{n+1,m}(x) & B_{n+1,m}(x) \\ A_{n,m+1}(x) & B_{n,m+1}(x) \end{vmatrix},$$

but from this it is not so easy to obtain the Type I polynomials.

(4) So far we limited ourselves to the very classical orthogonal polynomials of Jacobi, Laguerre, and Hermite. Discrete orthogonal polynomials, such as those of Charlier, Kravchuk, Meixner, and Hahn, can also be considered and several kinds of discrete multiple orthogonal polynomials can be worked out. It would not be a good idea to do this case by case, since these polynomials are

all connected by limit transitions, with the Hahn polynomials as the starting family. At a later stage, one could also consider multiple orthogonal polynomials on a quadratic lattice and on the general exponential lattice, leading to q-polynomials. Again, all these families are related with the Askey–Wilson polynomials as the family from which all others can be obtained by limit transitions. Do these polynomials have a representation as a (basic) hypergeometric function? Recall that we needed an Appell hypergeometric function of two variables for the Jacobi–Angelesco polynomials, so that one may need to consider (basic) hypergeometric functions of several variables.

(5) Multiple orthogonal polynomials arise naturally in the study of Hermite–Padé approximation, which is simultaneous rational approximation to a vector of r functions. In this respect it is quite natural to study multiple orthogonal polynomials as orthogonal vector polynomials. This approach is very useful in trying to extend results for the case $r = 1$ to the case $r > 1$ by looking for an appropriate formulation using vector algebra. Van Iseghem already used this approach to formulate a vector QD-algorithm for multiple orthogonal polynomials [48]. Several algebraic aspects of multiple orthogonal polynomials follow easily from the vector orthogonality [42,27]. A further generalization is to study matrix orthogonality, where the matrix need not be a square matrix [43]. Orthogonal polynomials and Padé approximants are closely related to certain continued fractions (J-fractions and S-fractions). For multiple orthogonal polynomials there is a similar relation with vector continued fractions and the Jacobi–Perron algorithm [36]. The seven families which we considered in this paper lead to seven families of vector continued fractions, which could be studied in more detail in the framework of continued fractions. Finally, one may wonder whether it is possible to use hypergeometric functions of matrix argument in the study of multiple orthogonal polynomials.

4.2. Non-symmetric banded operators

In Section 2 the connection between multiple orthogonal polynomials and banded Hessenberg operators of the form

$$\begin{pmatrix}
a_{0,0} & 1 & & & & & & & & \\
a_{1,1} & a_{1,0} & 1 & & & & & & & \\
a_{2,2} & a_{2,1} & a_{2,0} & 1 & & & & & & \\
\vdots & & & \ddots & \ddots & & & & & \\
a_{r,r} & a_{r,r-1} & \cdots & & a_{r,0} & 1 & & & & \\
& a_{r+1,r} & \ddots & & & a_{r+1,0} & 1 & & & \\
& & \ddots & \ddots & & & \ddots & \ddots & & \\
& & & \ddots & \ddots & & & \ddots & 1 & \\
& & & & a_{n,r} & a_{n,r-1} & \cdots & a_{n,1} & a_{n,0} & \ddots \\
& & & & & \ddots & \cdots & & \cdots & \ddots
\end{pmatrix}$$

was explained. For ordinary orthogonal polynomials the operator is tridiagonal and can always be made symmetric, and often it can be extended in a unique way to a self-adjoint operator (e.g., when all the coefficients are bounded). The spectrum of this tridiagonal operator corresponds to the support of the orthogonality measure, and the spectral measure is precisely the orthogonality measure.

Each tridiagonal matrix with ones on the upper diagonal and positive coefficients on the lower diagonal, corresponds to a system of orthogonal polynomials on the real line (Favard's theorem). Some preliminary work on the spectral theory of the higher-order operators ($r > 1$) was done by Kalyagin [21–23,5], but there are still quite a few open problems here.

(1) What is the proper extension of Favard's theorem for these higher-order banded Hessenberg operators? Not every banded Hessenberg operator corresponds to a system of multiple orthogonal polynomials with orthogonality relations on the real line. There needs to be additional structure, but so far this additional structure is still unknown. There is a weak version of the Favard theorem that gives multiple orthogonality with respect to linear functionals [48,24], but a stronger version that gives positive measures on the real line is needed. How do we recognize an Angelesco system, an AT system, or one of the combinations considered in [19] from the recurrence coefficients (from the operator)? The special case where all the diagonals are zero, except for the upper diagonal (which contains 1's) and the lower diagonal, has been studied in detail in [6]. They show that when the lower diagonal contains positive coefficients, the operator corresponds to multiple orthogonal polynomials on an $(r+1)$-star in the complex plane. Using a symmetry transformation, similar to the quadratic transformation that transforms Hermite polynomials to Laguerre polynomials, this also gives an AT system of multiple orthogonal polynomials on $[0,\infty)$.

(2) The asymptotic behavior of the recurrence coefficients of the seven systems described above is known. Each of the limiting operators deserves to be investigated in more detail. The limiting operator for the Jacobi–Piñeiro polynomials is a Toeplitz operator, and hence can be investigated in more detail. See, e.g., [46] for this case. Some of the other limiting operators are block Toeplitz matrices and can be investigated as well. Are there any multiple orthogonal polynomials having such recurrence coefficients? The Chebyshev polynomials of the second kind have this property when one deals with tridiagonal operators.

(3) The next step would be to work out a perturbation theory, where one allows certain perturbations of the limiting matrices. Compact perturbations would be the first step, trace class perturbations would allow us to give more detailed results.

4.3. Applications

(1) Hermite–Padé approximation was introduced by Hermite for his proof of the transcendence of e. More recently it became clear that Apéry's proof of the irrationality of $\zeta(3)$ relies on an AT system of multiple orthogonal polynomials with weights $w_1(x)=1$, $w_2(x)=-\log(x)$ and $w_3(x)=\log^2(x)$ on [0,1]. These multiple orthogonal polynomials are basically limiting cases of Jacobi–Piñeiro polynomials where $\alpha_0=0=\alpha_1=\alpha_2$. A very interesting problem is to prove irrationality of other remarkable constants, such as $\zeta(5)$, Catalan's constant, or Euler's constant. Transcendence proofs will even be better. See [4,45] for the connection between multiple orthogonal polynomials, irrationality, and transcendence.

(2) In numerical analysis one uses orthogonal polynomials when one constructs Gauss quadrature. In a similar way one can use multiple orthogonal polynomials to construct optimal quadrature formulas for jointly approximating r integrals of the same function f with respect to r weights $w_1,\ldots,w_r$. See, e.g., Borges [8], who apparently is not aware that he is using multiple orthogonal polynomials. Gautschi [17] has summarized some algorithms for computing recurrence coefficients, quadrature nodes (zeros of orthogonal polynomials) and quadrature weights (Christoffel numbers)

for ordinary Gauss quadrature. A nice problem is to modify these algorithms so that they compute recurrence coefficients, zeros of multiple orthogonal polynomials (eigenvalues of banded Hessenberg operators) and quadrature weights for simultaneous Gauss quadrature.

References

[1] W. Al-Salam, Characterization theorems for orthogonal polynomials, in: P. Nevai (Ed.), Orthogonal Polynomials: Theory and Practice, NATO ASI Series C, vol. 294, Kluwer, Dordrecht, 1990, pp. 1–24.

[2] G.E. Andrews, R. Askey, Classical orthogonal polynomials, in: C. Brezinski et al. (Eds.), Polynômes Orthogonaux et Applications, Lecture Notes in Mathematics, vol. 1171, Springer, Berlin, 1985, pp. 36–62.

[3] G.E. Andrews, R.A. Askey, R. Roy, Special Functions, Encyclopedia of Mathematics and its Applications, vol. 71, Cambridge University Press, Cambridge, 1999.

[4] A.I. Aptekarev, Multiple orthogonal polynomials, J. Comput. Appl. Math. 99 (1998) 423–447.

[5] A.I. Aptekarev, V. Kaliaguine (Kalyagin), Complex rational approximation and difference operators, Rend. Circ. Matem. Palermo, Ser. II, suppl. 52 (1998) 3–21.

[6] A.I. Aptekarev, V. Kaliaguine (Kalyagin), J. Van Iseghem, Genetic sum representation for the moments of a system of Stieltjes functions and its application, Constr. Approx., to appear.

[7] A.I. Aptekarev, F. Marcellán, I.A. Rocha, Semiclassical multiple orthogonal polynomials and the properties of Jacobi–Bessel polynomials, J. Approx. Theory 90 (1997) 117–146.

[8] C.F. Borges, On a class of Gauss-like quadrature rules, Numer. Math. 67 (1994) 271–288.

[9] M.G. de Bruin, Simultaneous Padé approximation and orthogonality, in: C. Brezinski et al. (Eds.), Polynômes Orthogonaux et Applications, Lecture Notes in Mathematics, vol. 1171, Springer, Berlin, 1985, pp. 74–83.

[10] M.G. de Bruin, Some aspects of simultaneous rational approximation, in: Numerical Analysis and Mathematical Modeling, Banach Center Publications 24, PWN-Polish Scientific Publishers, Warsaw, 1990, pp. 51–84.

[11] J. Bustamante, G. López, Hermite–Padé approximation for Nikishin systems of analytic functions, Mat. Sb. 183 (1992) 117–138; translated in Math. USSR Sb. 77 (1994) 367–384.

[12] Y. Ben Cheick, K. Douak, On two-orthogonal polynomials related to the Bateman $J_n^{u,v}$-function, manuscript.

[13] K. Douak, P. Maroni, Les Polynômes orthogonaux 'classiques' de dimension deux, Analysis 12 (1992) 71–107.

[14] K. Douak, P. Maroni, Une Caractérisation des polynômes d-orthogonaux 'classiques', J. Approx. Theory 82 (1995) 177–204.

[15] K. Driver, H. Stahl, Normality in Nikishin systems, Indag. Math., N.S. 5 (2) (1994) 161–187.

[16] K. Driver, H. Stahl, Simultaneous rational approximants to Nikishin systems, I, II, Acta Sci. Math. (Szeged) 60 (1995) 245–263; 61 (1995) 261–284.

[17] W. Gautschi, Orthogonal polynomials: applications and computation, in: A. Iserles (Ed.), Acta Numerica, Cambridge University Press, Cambridge, 1996, pp. 45–119.

[18] A.A. Gonchar, E.A. Rakhmanov, On the convergence of simultaneous Padé approximants for systems of Markov type functions, Trudy Mat. Inst. Steklov 157 (1981) 31–48: translated in Proc. Steklov Math. Inst. 3 (1983) 31–50.

[19] A.A. Gonchar, E.A. Rakhmanov, V.N. Sorokin, Hermite–Padé approximants for systems of Markov-type functions, Mat. Sb. 188 (1997) 33–58; translated in Russian Acad. Sci. Sb. Math. 188 (1997) 671–696.

[20] V.A. Kalyagin (Kaliaguine), On a class of polynomials defined by two orthogonality relations, Mat. Sb. 110 (1979) 609–627 (in Russian); translated in Math. USSR Sb. 38 (1981) 563–580.

[21] V. Kalyagin (Kaliaguine), Hermite–Padé approximants and spectral analysis of non-symmetric operators, Mat. Sb. 185 (1994) 79–100; translated in Russian Acad. Sci. Sb. Math. 82 (1995) 199–216.

[22] V.A. Kalyagin (Kaliaguine), Characteristics of the spectra of higher order difference operators and the convergence of simultaneous rational approximations, Dokl. Akad. Nauk 340 (1) (1995) 15–17; translated in Dokl. Math. 51 (1) (1995) 11–13.

[23] V.A. Kaliaguine (Kalyagin), On operators associated with Angelesco systems, East J. Approx. 1 (1995) 157–170.

[24] V. Kaliaguine (Kalyagin), The operator moment problem, vector continued fractions and an explicit form of the Favard theorem for vector orthogonal polynomials, J. Comput. Appl. Math. 65 (1995) 181–193.

[25] V.A. Kaliaguine (Kalyagin), A. Ronveaux, On a system of classical polynomials of simultaneous orthogonality, J. Comput. Appl. Math. 67 (1996) 207–217.

[26] R. Koekoek, R.F. Swarttouw, The Askey-scheme of hypergeometric orthogonal polynomials and its q-analogue, Delft University of Technology, Report 98–17, 1998. Available on-line at `http://aw.twi.tudelft.nl/˜koekoek/research.html`

[27] G. Labahn, B. Beckermann, A uniform approach for Hermite–Padé and simultaneous Padé approximants and their matrix type generalization, Numer. Algorithms 3 (1992) 45–54.

[28] K. Mahler, Perfect systems, Compositio Math. 19 (1968) 95–166.

[29] P. Maroni, L'orthogonalité et les récurrences de polynômes d'ordre supérieur à deux, Ann. Fac. Sci. Toulouse 10 (1989) 105–139.

[30] A.F. Nikiforov, S.K. Suslov, V.B. Uvarov, Classical Orthogonal Polynomials of a Discrete Variable, Springer Series in Computational Physics, Springer, Berlin, 1991.

[31] A.F. Nikiforov, V.B. Uvarov, Special Functions of Mathematical Physics, Birkhäuser, Basel, 1988.

[32] E.M. Nikishin, A system of Markov functions, Vestnik Mosk. Univ., Ser. I (4) (1979) 60–63; translated in Moscow Univ. Math. Bull. 34 (1979) 63–66.

[33] E.M. Nikishin, On simultaneous Padé approximants, Mat. Sb. 113 (115) (1980) 499–519; translated in Math. USSR Sb. 41 (1982) 409–425.

[34] E.M. Nikishin, V.N. Sorokin, Rational Approximations and Orthogonality, Translations of Mathematical Monographs, Vol. 92, Amer. Math. Soc., Providence, RI, 1991.

[35] J. Nuttall, Asymptotics of diagonal Hermite–Padé polynomials, J. Approx. Theory 42 (1984) 299–386.

[36] V.K. Parusnikov, The Jacobi–Perron algorithm and simultaneous approximation of functions. Mat. Sb. 114 (156) (1981) 322–333; translated in Math. USSR Sb. 42 (1982) 287–296.

[37] L.R. Piñeiro, On simultaneous approximations for a collection of Markov functions, Vestnik Mosk. Univ., Ser. I (2) (1987) 67–70 (in Russian); translated in Moscow Univ. Math. Bull. 42 (2) (1987) 52–55.

[38] V.N. Sorokin, Simultaneous Padé approximants for finite and infinite intervals, Izv. Vyssh. Uchebn. Zaved., Mat. (8) (1984) (267) 45–52; translated in J. Soviet Math. 28 (8) (1984) 56–64.

[39] V.N. Sorokin, A generalization of classical orthogonal polynomials and the convergence of simultaneous Padé approximants, Trudy Sem. Im. I. G. Petrovsk. 11 (1986) 125–165; translated in J. Soviet Math. 45 (1989) 1461–1499.

[40] V.N. Sorokin, A generalization of Laguerre polynomials and convergence of simultaneous Padé approximants, Uspekhi Mat. Nauk 41 (1986) 207–208; translated in Russian Math. Surveys 41 (1986) 245–246.

[41] V.N. Sorokin, Simultaneous Padé approximation for functions of Stieltjes type, Siber. Mat. Zh. 31 (5) (1990) 128–137; translated in Siber. Math. J. 31(5) (1990) 809–817.

[42] V.N. Sorokin, J. Van Iseghem, Algebraic aspects of matrix orthogonality for vector polynomials, J. Approx. Theory 90 (1997) 97–116.

[43] V.N. Sorokin, J. Van Iseghem, Matrix continued fractions, J. Approx. Theory 96 (1999) 237–257.

[44] G. Szegő, Orthogonal Polynomials, American Mathematical Society Colloquium Publ., Vol. 23, 4th Edition, AMS, Providence, RI, 1975.

[45] W. Van Assche, Multiple orthogonal polynomials, irrationality and transcendence, in: B.C. Berndt et al. (Eds.), Continued Fractions: from Analytic Number Theory to Constructive Approximation, Contemporary Mathematics, Vol. 236, Amer. Math. Soc., Providence, RI, 1999, pp. 325–342.

[46] W. Van Assche, Non-symmetric linear difference equations for multiple orthogonal polynomials, CRM Proceedings and Lecture Notes, Vol. 25, 2000, pp. 391–405.

[47] W. Van Assche, S.B. Yakubovich, Multiple orthogonal polynomials associated with Macdonald functions, Integral Transforms Special Functions 9 (2000) 229–244.

[48] J. Van Iseghem, Vector orthogonal relations, vector QD-algorithm, J. Comput. Appl. Math. 19 (1987) 141–150.

ELSEVIER Journal of Computational and Applied Mathematics 127 (2001) 349–368

JOURNAL OF
COMPUTATIONAL AND
APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Orthogonal polynomials and cubature formulae on balls, simplices, and spheres

Yuan Xu [1]

Department of Mathematics, University of Oregon, Eugene, OR 97403-1222, USA

Received 26 May 1999; received in revised form 23 September 1999

Abstract

We report on recent developments on orthogonal polynomials and cubature formulae on the unit ball B^d, the standard simplex T^d, and the unit sphere S^d. The main result shows that orthogonal structures and cubature formulae for these three regions are closely related. This provides a way to study the structure of orthogonal polynomials; for example, it allows us to use the theory of h-harmonics to study orthogonal polynomials on B^d and on T^d. It also provides a way to construct new cubature formulae on these regions. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Orthogonal polynomials in several variables; Cubature formulae; Summability; Orthogonal expansions; Symmetric group; Octahedral group

1. Introduction

The structure of orthogonal polynomials in several variables is significantly more complicated than that of orthogonal polynomials in one variable. Among the reasons for the complication, some are generic; for example, there are many distinct orders among polynomials in several variables, and consequently, the orthogonal bases are not unique. Others depend on the specific problems under consideration, for example, on weight functions and regions that define the orthogonality. The regions that have attracted most attention are regular ones, such as cubes, balls, simplices, and the surface of spheres. In the literature, orthogonal polynomials on these regions are mostly studied separately.

In the first part of the paper we report some recent results that reveal a close relation between orthogonal polynomials on the unit ball B^d, the standard simplex T^d, and the surface of the unit sphere S^d of the Euclidean space. The main results state, roughly speaking, that a basis of orthogonal polynomials on the simplex T^d is equivalent, under a simple transformation, to a basis of orthogonal

[1] Supported by the National Science Foundation under Grant DMS-9802265.

PII: S0377-0427(00)00504-5

polynomials on B^d that are invariant under sign changes; and a basis on B^d is equivalent to a basis of orthogonal polynomials on S^d that are even in one of their variables. The forerunner of the results is a relation between the spherical harmonics on S^d and a family of orthogonal polynomials on B^d that was observed and used in the work of Hermite, Didon, Appel and de Fériet, and Koschmieder (see, for example, [1,15, Chapter 12, Vol. II]). The relation between spherical harmonics and a family of orthogonal polynomials on a triangle was used in [8], and also in [6] in an application related to the method of finite elements. The general results are proved in [49,50] for large classes of weight functions subject only to mild assumptions. As an important application, this allows us to use Dunkl's theory of h-harmonics associated with the reflection groups [9–13] to derive compact formulae for the reproducing kernel and to study summability of several families of orthogonal polynomials on B^d and on T^d.

The study of the structures of polynomials on the sphere, the ball, and the simplex leads us to a close relation between cubature formulae on these regions, which is discussed in the second part of the paper. The main results state that, roughly speaking, cubature formulae on T^d are equivalent to cubature formulae that are invariant under sign changes on B^d, and formulae on B^d are equivalent to formulae that are invariant under the sign change of one fixed variable on S^d. In the literature, cubature formulae on different regions are mostly studied separately; all three regions have attracted their share of attention over decades of investigation (see, for example, [14,36,41] and the references therein). The fact that these relations are revealed only recently [49,50] seems rather remarkable. This allows us to construct formulae on one region and use the relations to obtain formulae for the other two regions. In this way, a number of new formulae on these regions have been derived [18–20], most notably a family of cubature formulae for the surface measure on S^d that exists for all d and all degrees (see Section 5). Moreover, these relations may shed light on an outstanding conjecture of Lebedev about cubature formulae on S^2, which we will discuss in some detail in the paper.

The paper is organized as follows. Section 2 is devoted to preliminaries. Section 3 deals with relations between orthogonal polynomials on the three regions. Applications to reproducing kernels and orthogonal series are discussed in Section 4. The connections between cubature formulae on the three regions are addressed in Section 5. Lebedev's conjecture is discussed in Section 6. Several open problems are discussed in Sections 4 and in 6.

2. Preliminaries

Basic notation: For $\boldsymbol{x} \in \mathbb{R}^d$ we denote by $|\boldsymbol{x}| = \sqrt{x_1^2 + \cdots + x_d^2}$ the usual Euclidean norm and by $|\boldsymbol{x}|_1 = |x_1| + \cdots + |x_d|$ the ℓ^1 norm. Let $\mathbb{N}_0$ be the set of nonnegative integers. For $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$, we write $|\alpha|_1 = \alpha_1 + \cdots + \alpha_d$, consistent with the notation $|\boldsymbol{x}|_1$. Throughout the paper we denote by B^d the unit ball of $\mathbb{R}^d$ and S^d the unit sphere on $\mathbb{R}^{d+1}$; that is

$$B^d = \{\boldsymbol{x} \in \mathbb{R}^d \colon |\boldsymbol{x}| \leqslant 1\} \quad \text{and} \quad S^d = \{\boldsymbol{y} \in \mathbb{R}^{d+1} \colon |\boldsymbol{y}| = 1\}.$$

We also denote by T^d the standard simplex in $\mathbb{R}^d$; that is

$$T^d = \{\boldsymbol{x} \in \mathbb{R}^d \colon x_1 \geqslant 0, \ldots, x_d \geqslant 0,\ 1 - |\boldsymbol{x}|_1 \geqslant 0\}.$$

For $d=2$, the ball B^2 is the unit disc and the simplex T^2 is the triangle with vertices at (0,0), (1,0) and (0,1).

Polynomial spaces: For $\alpha=(\alpha_1,\ldots,\alpha_d)\in\mathbb{N}_0^d$ and $\boldsymbol{x}=(x_1,\ldots,x_d)\in\mathbb{R}^d$ we write $\boldsymbol{x}^\alpha=x_1^{\alpha_1}\cdots x_d^{\alpha_d}$. The number $|\alpha|_1$ is called the total degree of $\boldsymbol{x}^\alpha$. We denote by Π^d the set of polynomials in d variables on $\mathbb{R}^d$ and by Π_n^d the subset of polynomials of total degree at most n. We also denote by $\mathscr{P}_n^d$ the space of homogeneous polynomials of degree n on $\mathbb{R}^d$ and we let $r_n^d=\dim\mathscr{P}_n^d$. It is well known that

$$\dim\Pi_n^d=\binom{n+d}{n}\quad\text{and}\quad r_n^d=\dim\mathscr{P}_n^d=\binom{n+d-1}{d-1}.$$

Invariance under a finite group: The region B^d and S^d are evidently invariant under the rotation group. The simplex T^d is invariant under the symmetric group of its vertices. If G is a subgroup of the rotation group of $\mathbb{R}^d$, we define the group action on a function f on $\mathbb{R}^d$ by $R(a)f(\boldsymbol{x})=f(\boldsymbol{x}a)$, $\boldsymbol{x}\in\mathbb{R}^d$, $a\in G$. If $R(a)f=f$ for all $a\in G$, we say that f is invariant under G. For example, for the simple abelian group $\mathbb{Z}_2^d$ consisting of elements $a=(\varepsilon_1,\ldots,\varepsilon_m)$, where $\varepsilon_i=\pm1$, $R(a)f(\boldsymbol{x})=f(\varepsilon_1x_1,\ldots,\varepsilon_dx_d)$. For a function f defined on T^d, we say that f is invariant on T^d if it is invariant under the symmetric group of T^d; that is, if it is invariant under permutations among the variables $\{x_1,\ldots,x_d,\ 1-|\boldsymbol{x}|_1\}$. We will deal with polynomials invariant under G. We denote by $\Pi_n^d(G)$ the space of polynomials in Π_n^d that are invariant under G, and by $\mathscr{P}_n^d(G)$ the space of homogeneous polynomials in $\mathscr{P}_n^d$ invariant under G.

Orthogonal polynomials on B^d and T^d: Let Ω denote either B^d or T^d. Let W be a weight function on Ω, which is assumed to be nonnegative and have finite moments. We often normalize W so that it has unit integral over Ω. Given an order among the monomials $\{\boldsymbol{x}^\alpha\}$, we can use the Gram–Schmidt process to generate a sequence of orthogonal polynomials with respect to the inner product of $L^2(W\,\mathrm{d}\boldsymbol{x})$. It is known that for each $n\in\mathbb{N}_0$ the set of polynomials of degree n that are orthogonal to all polynomials of lower degree forms a vector space, denoted by $\mathscr{V}_n^d(W)$, whose dimension is r_n^d. We denote by $\{P_\alpha^n\}$, $|\alpha|_1=n$ and $n\in\mathbb{N}_0$, one family of orthonormal polynomials with respect to W on Ω that forms a basis of $\mathscr{V}_n^d(W)$, where the superscript n means that $P_\alpha^n\in\Pi_n^d$. The orthonormality means that

$$\int_\Omega P_\alpha^n(\boldsymbol{x})P_\beta^m(\boldsymbol{x})W(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}=\delta_{\alpha,\beta}\delta_{m,n}.$$

We note that there are many bases of $\mathscr{V}_n^d(W)$; if Q is an invertible matrix of size r_n^d, then the components of $Q\mathbb{P}_n$ form another basis of $\mathscr{V}_n^d(W)$, where $\mathbb{P}_n$ denotes the vector $(P_\alpha^n)_{|\alpha|_1=n}$, which is orthonormal if Q is an orthogonal matrix. For results on the general structure of orthogonal polynomials in several variables, we refer to the survey [46] and the references there.

Classical orthogonal polynomials on T^d and B^d: On T^d they are orthogonal with respect to the weight functions

$$W_{\kappa,\mu}^T(\boldsymbol{x})=w_{\kappa,\mu}^Tx_1^{\kappa_1-1/2}\cdots x_d^{\kappa_d-1/2}(1-|\boldsymbol{x}|_1)^{\mu-1/2},\quad \boldsymbol{x}\in T^d,\tag{2.1}$$

where $\kappa_i>-\frac12$, $\mu>-\frac12$ and the normalized constant $w_{\kappa,\mu}^T$ is defined by

$$w_{\kappa,\mu}^T=\frac{\Gamma(|\kappa|_1+\mu+(d+1)/2)}{\Gamma(\kappa_1+\frac12)\cdots\Gamma(\kappa_d+\frac12)\Gamma(\mu+\frac12)}.$$

Related to $W^T_{\kappa,\mu}$ are weight functions $W^B_{\kappa,\mu}$ on B^d, defined by

$$W^B_{\kappa,\mu}(\boldsymbol{x}) = w^B_{\kappa,\mu}|x_1|^{2\kappa_1}\cdots|x_d|^{2\kappa_d}(1-|\boldsymbol{x}|^2)^{\mu-1/2}, \quad \boldsymbol{x}\in B^d, \tag{2.2}$$

where $\kappa_i > -\frac{1}{2}$, $\mu > -\frac{1}{2}$ and $w^B_{\kappa,\mu} = w^T_{\kappa,\mu}$. The classical orthogonal polynomials on B^d are orthogonal with respect to $W^B_\mu := W^B_{0,\mu}$; that is, the case $\kappa = 0$. For $\mu = \frac{1}{2}$, $W^B_{1/2} = 1/\mathrm{vol}\, B^d$ is the normalized Lebesgue measure.

We refer to [1,15, Chapter 12] for an account of earlier results on the classical orthogonal polynomials; they are characterized as eigenfunctions of certain second-order partial differential equation (see also [23,24]). Bases of the classical orthogonal polynomials can be constructed explicitly in terms of Jacobi polynomials, and we will see that they can be derived from orthogonal polynomials on the sphere S^d.

Ordinary spherical harmonics: The most important orthogonal polynomials on the sphere are the spherical harmonics, which are orthogonal with respect to the surface measure $\mathrm{d}\omega$ on S^d. The harmonic polynomials on $\mathbb{R}^{d+1}$ are homogeneous polynomials satisfying the Laplace equation $\Delta P = 0$, where $\Delta = \partial_1^2 + \cdots + \partial_{d+1}^2$ on $\mathbb{R}^{d+1}$ and ∂_i is the partial derivative with respect to the ith coordinate. The spherical harmonics are the restriction of harmonic polynomials on S^d. We refer to [34,40,42] for accounts of the theory of spherical harmonics.

h-harmonics associated with reflection groups: The theory of h-harmonics is established recently by Dunkl (see [9–13]). For a nonzero vector $\boldsymbol{v} \in \mathbb{R}^{d+1}$ define the reflection $\sigma_{\boldsymbol{v}}$ by $\boldsymbol{x}\sigma_{\boldsymbol{v}} := \boldsymbol{x} - 2\langle \boldsymbol{x}, \boldsymbol{v}\rangle \boldsymbol{v}/|\boldsymbol{v}|^2$, $\boldsymbol{x} \in \mathbb{R}^{d+1}$, where $\langle \boldsymbol{x}, \boldsymbol{y}\rangle$ denotes the usual Euclidean inner product. Let G be a reflection group on $\mathbb{R}^{d+1}$ with positive roots $\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_m\}$. Assume that $|\boldsymbol{v}_i| = |\boldsymbol{v}_j|$ if $\sigma_{\boldsymbol{v}_i}$ is conjugate to $\sigma_{\boldsymbol{v}_j}$. The h-harmonics are homogeneous orthogonal polynomials on S^d with respect to $h_\kappa^2\,\mathrm{d}\omega$, where the weight function h_κ is defined by

$$h_\kappa(\boldsymbol{x}) := \prod_{i=1}^m |\langle \boldsymbol{x}, \boldsymbol{v}_i\rangle|^{\kappa_i}, \quad \kappa_i \geqslant 0 \tag{2.3}$$

with $\kappa_i = \kappa_j$ whenever $\sigma_{\boldsymbol{v}_i}$ is conjugate to $\sigma_{\boldsymbol{v}_j}$ in the reflection group G generated by the reflections $\{\sigma_{\boldsymbol{v}_i}\colon 1 \leqslant i \leqslant m\}$. The function h_κ is a positively homogeneous G-invariant function. The key ingredient of the theory is a family of first-order differential–difference operators, $\mathscr{D}_i$ (Dunkl's operators), which generates a commutative algebra [11], defined by

$$\mathscr{D}_i f(\boldsymbol{x}) := \partial_i f(\boldsymbol{x}) + \sum_{j=1}^m \kappa_j \frac{f(\boldsymbol{x}) - f(\boldsymbol{x}\sigma_j)}{\langle \boldsymbol{x}, \boldsymbol{v}_j\rangle}\langle \boldsymbol{v}_j, \boldsymbol{e}_i\rangle, \quad 1 \leqslant i \leqslant d+1,$$

where $\boldsymbol{e}_1,\ldots,\boldsymbol{e}_{d+1}$ are the standard unit vectors of $\mathbb{R}^{d+1}$. The h-Laplacian is defined by $\Delta_h = \mathscr{D}_1^2 + \cdots + \mathscr{D}_{d+1}^2$, which plays the role of Laplacian in the theory of ordinary harmonics. In particular, the h-harmonics are the homogeneous polynomials satisfying the equation $\Delta_h P = 0$. The h-spherical harmonics are the restriction of h-harmonics on the sphere. The structure of the space of h-harmonics, such as dimensionality and decomposition, is parallel to that of ordinary harmonics. In particular, there is an intertwining operator V between the algebra of differential operators and the commuting algebra of Dunkl's operators, which helps us to transform certain properties of ordinary harmonics to the h-harmonics. The operator V is the unique linear operator defined by

$$V\mathscr{P}_n \subset \mathscr{P}_n, \quad V1 = 1, \quad \mathscr{D}_i V = V\partial_i, \quad 1 \leqslant i \leqslant d.$$

It is also proved in [37] that V is a positive operator. The closed form of V is known, however, only in the case of the abelian group $\mathbb{Z}_2^d$ [12,47] and the symmetric group S_3 [13]. For further properties and results of h-harmonics, we refer to [9–13] and the references therein.

Examples of reflection invariant weight functions: The group $\mathbb{Z}_2^{d+1}$ is one of the simplest reflection groups. The weight function invariant under $\mathbb{Z}_2^{d+1}$ is

$$h_{\kappa,\mu}(\boldsymbol{x}) = a_{\kappa,\mu}|x_1|^{\kappa_1}\cdots|x_d|^{\kappa_d}|x_{d+1}|^{\mu}, \quad \boldsymbol{x}\in\mathbb{R}^{d+1}, \tag{2.4}$$

where $a_{\kappa,\mu}^2 = 2w_{\kappa,\mu}^T$. We use μ for the power of the last component of $\boldsymbol{x}$ to emphasis the connection between $h_{\kappa,\mu}$ and $W_{\kappa,\mu}^B$ and $W_{\kappa,\mu}^T$ that will become clear in the next section. For $\kappa=0$ and $\mu=0$, we have that $a_0^2 = 2\Gamma((d+1)/2)/\pi^{(d+1)/2} = 1/\omega_d$, where ω_d denotes the surface area of S^d. Another interesting case is the hyper-octahedral group G generated by the reflections in $x_i=0$, $1\leqslant i\leqslant d+1$ and $x_i\pm x_j=0$, $1\leqslant i,j\leqslant d+1$; it is the Weyl group of type B_{d+1}. There are two conjugacy classes of reflections, hence two parameters for h_κ. We have

$$h_\kappa(\boldsymbol{x}) = \prod_{i=1}^{d}|x_i|^{\kappa_1}\prod_{i<j}|x_i^2-x_j^2|^{\kappa_0}. \tag{2.5}$$

The integral, hence the normalization constant, of h_κ can be computed by the use of Selberg's integral. In fact, the integral of h_κ in (2.3) for every reflection group has been computed in the work of Askey, Heckman, McDonald, Opdam and several others. See the references in [12].

3. Relation between orthogonal polynomials on the three regions

Throughout this section we fix the following notation: For $\boldsymbol{y}\in\mathbb{R}^{d+1}$, we write

$$\boldsymbol{y} = (y_1,\ldots,y_d,y_{d+1}) = (\boldsymbol{y}',y_{d+1}) = r(\boldsymbol{x},x_{d+1}), \tag{3.1}$$

where $\boldsymbol{y}'\in\mathbb{R}^d$, $r=|\boldsymbol{y}|=\sqrt{y_1^2+\cdots+y_{d+1}^2}$ and $\boldsymbol{x}=(x_1,\ldots,x_d)\in B^d$. We call a weight function H, defined on $\mathbb{R}^{d+1}$, S-symmetric if it is symmetric with respect to y_{d+1} and centrally symmetric with respect to the variables $\boldsymbol{y}'=(y_1,\ldots,y_d)$, i.e.,

$$H(\boldsymbol{y}',y_{d+1}) = H(\boldsymbol{y}',-y_{d+1}) \quad\text{and}\quad H(\boldsymbol{y}',y_{d+1}) = H(-\boldsymbol{y}',y_{d+1})$$

and we assume that H is not a zero function when restricted to S^d. For example, the weight functions of the form $H(\boldsymbol{y}) = W(y_1^2,\ldots,y_d^2)$ are S-symmetric, which include $H=h_\kappa^2$ for both h_κ in (2.4) and in (2.5).

In [49] we proved that there are orthonormal bases of homogeneous polynomials with respect to the inner product of $L^2(H\,\mathrm{d}\omega,S^d)$ for S-symmetric H. Let us denote by $\mathscr{H}_n^{d+1}(H)$ the space of homogeneous polynomials of degree n. When $H(\boldsymbol{y})=1$, we write $\mathscr{H}_n^{d+1}$, which is the space of ordinary harmonics of degree n. For $H=h_\kappa^2$ with h_κ as in (2.3), $\mathscr{H}_n^{d+1}(h_\kappa^2)$ is the space of h-harmonics. It is shown in [49] that

$$\dim\mathscr{H}_n^{d+1}(H) = \binom{n+d}{d} - \binom{n+d-2}{d} = \dim\mathscr{P}_n^{d+1} - \dim\mathscr{P}_{n-2}^{d+1}, \tag{3.2}$$

the same as for ordinary harmonics, and there is a unique decomposition of $\mathscr{P}_n^{d+1}$,

$$\mathscr{P}_n^{d+1} = \bigoplus_{k=0}^{[n/2]} |\boldsymbol{y}|^{2k} \mathscr{H}_{n-2k}^{d+1}(H). \tag{3.3}$$

These results are proved using the relation between $\mathscr{H}_n^{d+1}(H)$ and orthogonal polynomials on B^d, which we describe below.

3.1. Orthogonal polynomials on balls and on spheres

In association with a weight function H defined on $\mathbb{R}^{d+1}$, we define a weight function W_H^B on B^d by

$$W_H^B(\boldsymbol{x}) = H(\boldsymbol{x}, \sqrt{1-|\boldsymbol{x}|^2}), \quad \boldsymbol{x} \in B^d. \tag{3.4}$$

If H is S-symmetric, then the assumption that H is centrally symmetric with respect to the first d variables implies that W_H^B is centrally symmetric on B^d. Recall the notation $\mathscr{V}_n^d(W)$ for the space of orthonormal polynomials of degree n. We denote by $\{P_\alpha^n\}$ and $\{Q_\alpha^n\}$ systems of orthonormal polynomials that form a basis for $\mathscr{V}_n^d(W_1^B)$ and $\mathscr{V}_n^d(W_2^B)$ with respect to the weight functions

$$W_1^B(\boldsymbol{x}) = 2W_H^B(\boldsymbol{x})/\sqrt{1-|\boldsymbol{x}|^2} \quad \text{and} \quad W_2^B(\boldsymbol{x}) = 2W_H^B(\boldsymbol{x})\sqrt{1-|\boldsymbol{x}|^2},$$

respectively. Keeping in mind notation (3.1), we define

$$Y_{\alpha,n}^{(1)}(\boldsymbol{y}) = r^n P_\alpha^n(\boldsymbol{x}) \quad \text{and} \quad Y_{\beta,n}^{(2)}(\boldsymbol{y}) = r^n x_{d+1} Q_\beta^{n-1}(\boldsymbol{x}), \tag{3.5}$$

where $|\alpha|_1 = n$, $|\beta|_1 = n-1$ and we define $Y_{\beta,0}^{(2)}(\boldsymbol{y}) = 0$. It is proved in [49] that, as functions of $\boldsymbol{y}, Y_{\alpha,n}^{(1)}$ and $Y_{\beta,n}^{(2)}$ are, in fact, elements of $\mathscr{H}_n^{d+1}(H)$. The proof of (3.2) and (3.3) follows from this fact. On the other hand, if H is S-symmetric, then $\mathscr{H}_n^{d+1}(H)$ must have a basis that consists of homogeneous polynomials that are either even in y_{d+1} or odd in y_{d+1}. Using notation (3.1) and the fact that $x_{d+1}^2 = 1-|\boldsymbol{x}|^2$, we can write those that are even in y_{d+1} as $r^n P_\alpha^n(\boldsymbol{x})$ and those that are odd as $r^n x_{d+1} Q_\beta^{n-1}(\boldsymbol{x})$, where P_α^n and Q_β^{n-1} are polynomials in $\boldsymbol{x}$ of degree n and $n-1$, respectively. Then the polynomials P_α^n and Q_β^n are orthogonal polynomials with respect to W_1^B and W_2^B, respectively. We summarize the result as

Theorem 3.1. *Let H be an S-symmetric weight function defined on $\mathbb{R}^{d+1}$ and let W_1^B and W_2^B be defined as above. Then relation* (3.5) *defines a one-to-one correspondence between an orthonormal basis of $\mathscr{H}_n^{d+1}(H)$ and an orthonormal basis of $\mathscr{V}_n^d(W_1^B) \oplus x_{d+1}\mathscr{V}_{n-1}^d(W_2^B)$.*

In particular, if $H(\boldsymbol{y}) = 1$, then the theorem states that the ordinary spherical harmonics correspond to orthogonal polynomials with respect to the weight functions $1/\sqrt{1-|\boldsymbol{x}|^2}$ and $\sqrt{1-|\boldsymbol{x}|^2}$, respectively. In the case of $d=1$, we have $\mathscr{H}_n^2 = \text{span}\{r^n \cos n\theta, r^n \sin n\theta\}$, using the polar coordinates $y_1 = r\cos\theta$ and $y_2 = r\sin\theta$. The correspondence in the theorem is then the well-known fact that $T_n(x) = \cos n\theta$ and $U_n(x) = \sin(n+1)\theta/\sin\theta$, where $x = \cos\theta$, are orthogonal with respect to $1/\sqrt{1-x^2}$ and $\sqrt{1-x^2}$ on $[-1,1]$, respectively.

Under the correspondence, the orthogonal polynomials for the weight function $W^B_{\kappa,\mu}$ in (2.2) are related to the h-harmonics associated with $h_{\kappa,\mu}$ in (2.4); in particular, the classical orthogonal polynomials on B^d are related to h-harmonics associated with $h_\mu(\mathbf{y}) = |y_{d+1}|^\mu$. Compact formulae of an orthonormal basis of these polynomials on B^d can be obtained accordingly from the formulae in [47]. Moreover, the second-order partial differential equation satisfied by the classical orthogonal polynomials can be derived from the h-Laplacian by a simple change of variables [56].

3.2. Orthogonal polynomials on balls and on simplices

Let $W^B(\mathbf{x}) := W(x_1^2, \dots, x_d^2)$ be a weight function defined on B^d. Associated with W^B we define a weight function W^T on the simplex T^d by

$$W^T(\mathbf{u}) = W(u_1, \dots, u_d)/\sqrt{u_1 \cdots u_d}, \quad \mathbf{u} \in T^d \tag{3.6}$$

and we normalize the weight function W so that W^B has unit integral on B^d. It follows from a simple change of variables that W^T has unit integral on T^d.

A polynomial P is invariant under the group $\mathbb{Z}_2^d$, if P is even in each of its variables; such a polynomial must be of even degree. We denote by $\mathscr{V}_{2n}^d(W^B, \mathbb{Z}_2^d)$ the space of polynomials in $\mathscr{V}_{2n}^d(W^B)$ that are invariant under the group $\mathbb{Z}_2^d$. That is, $\mathscr{V}_{2n}^d(W^B, \mathbb{Z}_2^d)$ contains orthogonal polynomials of degree $2n$ that are even in their variables.

Let P_α^{2n} be a polynomial in $\mathscr{V}_{2n}^d(W^B, \mathbb{Z}_2^d)$. Since it is even in each of its variables, we can write it in the form of

$$P_\alpha^{2n}(\mathbf{x}) = R_\alpha^n(x_1^2, \dots, x_d^2), \quad |\alpha|_1 = n, \tag{3.7}$$

where R_α^n is a polynomial of degree n. It turns out that R_α^n is a polynomial in $\mathscr{V}_n^d(W^T)$. In fact, the relation defines a one-to-one correspondence.

Theorem 3.2. *Let W^B and W^T be weight functions defined as above. Then relation* (3.7) *defines a one-to-one correspondence between an orthonormal basis of $\mathscr{V}_{2n}^d(W^B, \mathbb{Z}_2^d)$ and an orthonormal basis of $\mathscr{V}_n^d(W^T)$.*

Under the correspondence, the classical orthogonal polynomials on T^d associated with the weight function $W^T_{\kappa,\mu}$ in (2.1) correspond to orthogonal polynomials with respect to $W^B_{\kappa,\mu}$ in (2.2); compact formulae of an orthonormal basis can be obtained from those in [47]. We note that the unit Lebesgue measure (unit weight function) on T^d corresponds to $|x_1 \cdots x_d|$ on B^d and the Lebesgue measure on B^d corresponds to $1/\sqrt{x_1 \cdots x_d}$ on T^d, because the Jacobian of the map $(x_1, \dots, x_d) \mapsto (x_1^2, \dots, x_d^2)$ is $|x_1 \cdots x_d|$. From the results in the previous subsection, polynomials in $\mathscr{V}_{2n}^d(W^B_{\kappa,\mu})$ satisfy a differential–difference equation that follows from the h-Laplacian and a change of variables. When we restrict to the elements of $\mathscr{V}_{2n}^d(W^B_{\kappa,\mu}, \mathbb{Z}_2^d)$, the difference part in the equation disappears owing to the invariance under $\mathbb{Z}_2^d$, and we end up with a second-order partial differential equation. Upon changing variables as in correspondence (3.7), we then recover the second-order partial differential equation satisfied by the classical orthogonal polynomials on T^d [56].

3.3. Orthogonal polynomials on spheres and on simplices

Putting the results in Sections 3.1 and 3.2 together, we also have a relation between orthogonal polynomials on S^d and those on T^d. The relation can be derived from the previous two subsections; we formulate them below for better reference.

On S^d we need to restrict to the weight function $H(\boldsymbol{y}) = W(y_1^2, \ldots, y_{d+1}^2)$, which is evidently S-symmetric. We denote by $\mathscr{H}_{2n}^{d+1}(H, \mathbb{Z}_2^{d+1})$ the space of orthogonal polynomials in $\mathscr{H}_{2n}^{d+1}(H)$ that are invariant under the group $\mathbb{Z}_2^{d+1}$. Associated with H we define a weight function on T^d by

$$W_H^T(\boldsymbol{x}) = 2W(x_1, \ldots, x_d, 1 - |\boldsymbol{x}|_1)/\sqrt{x_1 \cdots x_d(1 - |\boldsymbol{x}|_1)}, \quad \boldsymbol{x} \in T^d.$$

The constant 2 is there so that if H has unit integral on S^d then W_H^T has unit integral on T^d. Let $\{S_\alpha^{2n}\}$ denote a basis for $\mathscr{H}_{2n}^{d+1}(H, \mathbb{Z}_2^{d+1})$. Since S_α^{2n} is homogeneous and even in each of its variables, we can use notation (3.1) and the fact that $x_{d+1}^2 = 1 - |\boldsymbol{x}|^2$ to write S_α^{2n} as

$$S_\alpha^{2n}(\boldsymbol{y}) = r^{2n} R_\alpha^n(x_1^2, \ldots, x_d^2), \tag{3.8}$$

where R_α^n is a polynomial of degree n. On the other hand, given polynomials R_α^n defined on T^d, we can use (3.8) to define homogeneous polynomials on S^d. This relation connects the orthogonal polynomials on S^d and those on T^d.

Theorem 3.3. *Let H and W_H^T be weight functions defined as above. Then relation* (3.8) *defines a one-to-one correspondence between an orthonormal basis of $\mathscr{H}_{2n}^{d+1}(H, \mathbb{Z}_2^{d+1})$ and an orthonormal basis of $\mathscr{V}_n^d(W_H^T)$.*

As a consequence of this correspondence, we see that there is a unique decomposition of $\mathscr{P}_n^{d+1}(\mathbb{Z}_2^{d+1})$ in terms of $\mathscr{H}_n^d(H, \mathbb{Z}_2^{d+1})$,

$$\mathscr{P}_{2n}^{d+1}(\mathbb{Z}_2^{d+1}) = \bigoplus_{k=0}^{n} |\boldsymbol{y}|^{2k} \mathscr{H}_{2n-2k}^d(H, \mathbb{Z}_2^{d+1}).$$

Under the correspondence, the classical orthogonal polynomials on T^d associated with $W_{\kappa,\mu}^T$ in (2.1) correspond to h-spherical harmonics associated with h_κ in (2.4) with $\kappa_{d+1} = \mu$. In particular, the orthogonal polynomials with respect to the weight functions $1/\sqrt{x_1 \cdots x_d(1 - |\boldsymbol{x}|_1)}$ are related to the ordinary spherical harmonics, and those with respect to the unit weight function on T^d correspond to h-harmonics for $|x_1 \cdots x_{d+1}|$ on S^d.

4. Reproducing Kernel and Fourier orthogonal expansion

The relations stated in the previous section allow us to derive properties for orthogonal polynomials on one region from those on the other two regions. In this section, we examine the case of h-harmonics on S^d and their counterpart on B^d and on T^d. As we shall see, this approach will reveal several hidden symmetry properties of orthogonal polynomials on B^d and on T^d by relating them to the rich structure of h-harmonics. Some of the properties are new even for classical orthogonal polynomials.

We start with the definition of Fourier orthogonal expansion. Let $\{S^h_{\alpha,n}\}$ denote an orthonormal basis of h-harmonics. The reproducing kernel of $\mathscr{H}^{d+1}_n(h^2_\kappa)$ is defined by the formula

$$P_n(h^2_\kappa; \boldsymbol{x}, \boldsymbol{y}) = \sum_\alpha S^h_{\alpha,n}(\boldsymbol{x})S^h_{\alpha,n}(\boldsymbol{y}),$$

where the summation is over all h-harmonics of degree n. For $f \in L^2(h^2_\kappa, S^d)$, we consider the Fourier expansion of f in terms of the orthonormal basis $S^h_{\alpha,n}$. The partial sum of such an expansion with respect to $\mathscr{H}^{d+1}_n(h^2_\kappa)$ is given by

$$P_n(f, h^2_\kappa; \boldsymbol{x}) = \int_{S^d} f(\boldsymbol{y})P_n(h^2_\kappa; \boldsymbol{x},\boldsymbol{y})h^2_\kappa(\boldsymbol{y})\,\mathrm{d}\omega(\boldsymbol{y}) := (f * P_n(h^2_\kappa))(\boldsymbol{x}). \tag{4.1}$$

It is not hard to see that the reproducing kernel is independent of the choice of the particular bases. In fact, for h-harmonics, the kernel enjoys a compact formula in terms of the intertwining operator [48]

$$P_n(h^2_\kappa; \boldsymbol{x}, \boldsymbol{y}) = \frac{n + |\kappa|_1 + (d-1)/2}{|\kappa|_1 + (d-1)/2} V[C_n^{(|\kappa|_1+(d-1)/2)}(\langle \boldsymbol{x}, \cdot\rangle)](\boldsymbol{y}), \tag{4.2}$$

where $\boldsymbol{x}$, $\boldsymbol{y} \in S^d$ and $C_n^{(\lambda)}$ is the Gegenbauer polynomial of degree n. Here and in the following the reader may want to keep in mind that if $h_\kappa(\boldsymbol{x}) = 1$, then the h-harmonics are just the classical spherical harmonics and $V = \mathrm{id}$ is the identity operator. In that case, (4.2) is just the compact formula for the ordinary zonal polynomials (cf. [34, p. 19] or [40, p. 149]). In the case of $\mathbb{Z}_2^{d+1}$ and the weight function (2.4), the closed form of the intertwining operator V is given by [12,47]

$$Vf(\boldsymbol{x}) = \int_{[-1,1]^{d+1}} f(t_1x_1, \ldots, t_{d+1}x_{d+1}) \prod_{i=1}^{d+1} c_{\kappa_i}(1 + t_i)(1 - t_i^2)^{\kappa_i - 1}\,\mathrm{d}\boldsymbol{t}, \tag{4.3}$$

where $c_\lambda = 1/\int_{-1}^1 (1-t^2)^{\lambda-1}\,\mathrm{d}t$ for $\lambda > 0$, and we have taken $\mu = k_{d+1}$. Moreover, if some $\kappa_i = 0$, then the above formula holds under the limit relation

$$\lim_{\lambda\to 0} c_\lambda \int_{-1}^1 f(t)(1-t^2)^{\lambda-1}\,\mathrm{d}t = [f(1) + f(-1)]/2.$$

In this case, we can write down the reproducing kernel $P_n(h^2_\kappa; \boldsymbol{x},\boldsymbol{y})$ explicitly.

Although a closed form of the intertwining operator is not known in general, its average over the sphere can be computed as shown in [48].

Theorem 4.1. *Let h_k be defined as in* (2.3) *associated with a reflection group. Let V be the intertwining operator. Then*

$$\int_{S^d} Vf(\boldsymbol{x})h^2_\kappa(\boldsymbol{x})\,\mathrm{d}\omega = A_k \int_{B^{d+1}} f(\boldsymbol{x})(1 - |\boldsymbol{x}|^2)^{|\kappa|_1 - 1}\,\mathrm{d}\boldsymbol{x} \tag{4.4}$$

for $f \in \Pi^d$, where A_k is a constant that can be determined by setting $f(\boldsymbol{x}) = 1$.

Eq. (4.4) is not trivial even in the case of $\mathbb{Z}_2^{d+1}$; the reader may try to verify it with V given by (4.3). This result allows us to prove a general convergence theorem for the Fourier orthogonal expansion in h-harmonics. For $\delta > 0$, the Cesàro (C,δ) means, s_n^δ, of a sequence $\{s_n\}$ are defined by

$$s_n^\delta = \frac{1}{\binom{n+\delta}{n}} \sum_{k=0}^{n} \binom{n-k+\delta-1}{n-k} s_k = \frac{1}{\binom{n+\delta}{n}} \sum_{k=0}^{n} \binom{n-k+\delta}{n-k} c_k,$$

where the second equality holds if s_n is the nth partial sum of the series $\sum_{k=0}^{\infty} c_k$. We say that $\{s_n\}$ is Cesàro (C,δ) summable to s if s_n^δ converges to s as $n \to \infty$. Let $P_n^\delta(h_\kappa^2, f)$ denote the Cesàro (C,δ) means of the Fourier series in h-harmonics. By (4.1), we can write $P_n^\delta(f, h_\kappa^2) = f * P_n^\delta(h_\kappa^2)$, where $P_n^\delta(h_\kappa^2)$ denotes the (C,δ) means of $P_n(h_\kappa^2)$, which can be written as $V[k_n^\delta(\langle \boldsymbol{x}, \cdot \rangle)](\boldsymbol{y})$ by (4.2), where k_n^δ is the (C,δ) means of the reproducing kernel for the Gegenbauer series of order $|\kappa|_1 + (d-1)/2$. Since V is a positive operator, we have $|Vf(\boldsymbol{x})| \leqslant V(|f|)(x)$. Hence, if we apply Theorem 4.1, then we conclude that

$$\int_{S^d} |P_n^\delta(h_\kappa^2; \boldsymbol{x}, \boldsymbol{y})| h_\kappa^2(\boldsymbol{y}) \, \mathrm{d}\omega \leqslant A_\kappa \int_{B^{d+1}} |k_n^\delta(\langle \boldsymbol{x}, \boldsymbol{y} \rangle)| (1 - |\boldsymbol{y}|^2)^{|\kappa|_1 + (d-2)/2} \, \mathrm{d}\boldsymbol{y}.$$

A standard change of variables shows that the last integral can be reduced to the integral over $[-1,1]$. As a consequence, the (C,δ) summability of the h-harmonics follows from that of Gegenbauer series. We have [48].

Theorem 4.2. *Let h_κ be defined as in* (2.3). *Let $f \in L^p(h_\kappa^2, S^d)$. Then the expansion of f as the Fourier series with respect to h_κ^2 is (C,δ) summable in $L^p(h_\kappa^2, S^d)$, $1 \leqslant p \leqslant \infty$, provided $\delta > |\kappa|_1 + (d-1)/2$.*

Together with the relation between orthogonal polynomials on S^d, B^d and T^d, the above results on h-harmonics allow us to derive results for the reproducing kernel and for the summability of orthogonal expansion on B^d and on T^d. Instead of stating the results for the most general weight functions on these regions, we shall restrict ourselves to the weight functions $W_{\kappa,\mu}^T$ in (2.1) and $W_{\kappa,\mu}^B$ in (2.2); both are related to $h_{\kappa,\mu}$ in (2.3), for which all formulae can be written down in closed form. The restricted cases include those of the classical orthogonal polynomials. First, we define the reproducing kernel and the Fourier orthogonal series. Let $\{P_\alpha^n\}$ denote a sequence of orthonormal polynomials with respect to a weight function W defined on Ω, where Ω is either B^d or T^d. The reproduction kernel of $\mathscr{V}_n^d(W)$, denoted by $P_n(W; \cdot, \cdot)$, is defined by

$$P_n(W; \boldsymbol{x}, \boldsymbol{y}) = \sum_{|\alpha|_1 = n} P_\alpha^n(\boldsymbol{x}) P_\alpha^n(\boldsymbol{y}).$$

This kernel is, in fact, independent of the choice of the bases (see, for example, [46]). For f in $L^2(W, \Omega)$, we consider the Fourier orthogonal series whose nth partial sum $S_n(f, W)$ is defined by

$$S_n(f, W; \boldsymbol{x}) = \sum_{k=1}^{n} P_k(f, W; \boldsymbol{x}), \quad P_k(f, W; \boldsymbol{x}) = \int_\Omega f(\boldsymbol{y}) P_k(W; \boldsymbol{x}, \boldsymbol{y}) W(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}.$$

The relation between orthogonal polynomials on the three regions lead to relations between the reproducing kernels. In the case of $W = W_{\kappa,\mu}$ on B^d and $W = W^T_{\kappa,\mu}$ on T^d, this allows us to derive a compact formula for the reproducing kernel. First, we state the formula for the classical orthogonal polynomials on B^d, that is, for $W^B_\mu = W^B_{0,\mu}$.

Theorem 4.3. *For the classical weight function* $W^B_\mu = W_{0,\mu}$ *in* (2.2), *where* $\mu \geqslant 0$, *defined on* B^d, *we have*

$$P_n(W^B_\mu; \boldsymbol{x}, \boldsymbol{y}) = \frac{n + \mu + (d-1)/2}{\mu + (d-1)/2} \times \int_{-1}^{1} C_n^{(\mu+(d-1)/2)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle + s\sqrt{1 - |\boldsymbol{x}|^2}\sqrt{1 - |\boldsymbol{y}|^2})(1 - s^2)^{\mu-1}\,\mathrm{d}s. \tag{4.5}$$

We can also write down the compact formula for $W^B_{\kappa,\mu}$ as a multiple integral using (4.2), (4.3) and Theorem 3.1. The formula looks similar to (4.6) below. For $d = 1$, formula (4.5) reduces to the product formula of the Gegenbauer polynomials (cf. [15, Section 3.15.1, (20)]); in fact, in the first proof of (4.5) in [52], we wrote $P_n(W^B_\mu)$ as a multiple sum of an explicit orthonormal basis in terms of Gegenbauer polynomials and used the product formula repeatedly to add up the sums. Essentially, the same elementary but tedious proof is given in [47] for the compact formula of $P_n(h^2_\kappa)$, from which the formula for $W^B_{\kappa,\mu}$ can be derived. However, without using the relation between orthogonal polynomials on T^d and S^d, it is unlikely that the compact formulae for $W^T_{\kappa,\mu}$ can be discovered. The formula is given as follows.

Theorem 4.4. *For* $W^T_{\kappa,\mu}$ *in* (2.1), *where* $\kappa_i \geqslant 0$ *and* $\mu \geqslant 0$, *defined on* T^d, *we have*

$$P_n(W_{\kappa,\mu}; \boldsymbol{x}, \boldsymbol{y}) = c_\mu \frac{2n + |\kappa|_1 + \mu + (d-1)/2}{|\kappa|_1 + \mu + (d-1)/2} \times \int_{-1}^{1} \int_{[-1,1]^d} C_{2n}^{(|\kappa|_1+\mu+(d-1)/2)}(\sqrt{x_1 y_1}t_1 + \cdots + \sqrt{x_d y_d}t_d + s\sqrt{1 - |\boldsymbol{x}|_1}\sqrt{1 - |\boldsymbol{y}|_1}) \times \prod_{i=1}^{d} c_{\kappa_i}(1 - t_i^2)^{\kappa_i - 1}\,\mathrm{d}\boldsymbol{t}(1 - s^2)^{\mu-1}\,\mathrm{d}s. \tag{4.6}$$

These remarkable formulae have already been used on several occasions. In [45,53], they are used to derive asymptotics of the Christoffel functions for the ball and for the simplex. In [54], they are used to construct cubature formulae via the method of reproducing kernel. They also allow us to prove various results on summability of Fourier orthogonal expansions. For example, we have [51,52]:

Theorem 4.5. *The Cesàro* $(C, |\kappa|_1 + \mu + (d+1)/2)$ *means of the Fourier orthogonal expansion of a function with respect to* $W^\Omega_{\kappa,\mu}$, *where* $\Omega = B$ *or* T, *define a positive linear polynomial approximation identity on* $C(\Omega^d)$; *the order of summability is best possible in the sense that the* (C, δ) *means are* *not positive for* $0 < \delta < |\kappa|_1 + \mu + (d+1)/2$ *in the case of* $\Omega = T$ *and also in the case of* $\Omega = B$ *when at least one* $\kappa_i = 0$.

The sufficient part of the positivity follows from the positivity of the sums of Gegenbauer polynomials (see [16] or [2]). In fact, since V is positive, the sufficient part holds for all h-harmonics and their orthogonal polynomial counterparts on B^d and T^d. The positivity of the (C,δ) means implies that the means converge in norm. However, the positivity is not necessary for convergence. From Theorem 4.2, the compact formulae of the reproducing kernel and the relation to h-harmonics, we have the following theorem [51,52,55].

Theorem 4.6. *Let* $f \in L^p(W^{\Omega}_{\kappa,\mu}, \Omega^d)$, *where* $\Omega = B$ *or* $\Omega = T$. *Then the expansion of f as the Fourier orthogonal series with respect to* $W^{\Omega}_{\kappa,\mu}$ *is* (C,δ) *summable in* $L^p(W_{\kappa,\mu,}, \Omega^d)$, $1 \leqslant p \leqslant \infty$, *provided* $\delta > |\kappa|_1 + \mu + (d-1)/2$. *Moreover*; *if at least one* $\kappa_i = 0$, *then the expansion is not* (C,δ) *summable in* $L^p(W_{\kappa,\mu}, \Omega^d)$ *for* $p = 1$ *or* $p = \infty$ *provided* $\delta \leqslant |\kappa|_1 + \mu + (d-1)/2$.

Some remarks are in order. If at least one $\kappa_i = 0$, the index $\delta_0 := |\kappa|_1 + \mu + (d-1)/2$ is the analogy of the critical index in Fourier analysis (cf. [40]). Indeed, as the results in the previous section show, the case $\kappa = 0$ and $\mu = 0$ corresponds to expansion in the classical spherical harmonics, and the index $\delta_0 = (d-1)/2$ in this case is the critical index there. The result in the theorem states that the critical index for the classical orthogonal polynomial expansions on the ball B^d and on the simplex T^d are the same. On the other hand, the critical index for orthogonal expansions on the cube $[-1,1]^d$ is very different from these cases. Indeed, the orthogonal expansions for $1/\prod_{i=1}^d \sqrt{1-x_i^2}$ on $[-1,1]^d$ is the same as the $\ell-1$ summability of multiple Fourier series, which has no critical index; that is, $\delta_0 = 0$ ([3,4] and the reference therein). The summability of the general product Jacobi expansions is studied in [29].

There are many open questions in this direction. For example, finding a closed form of the intertwining operator V, and finding an explicit orthonormal basis for h-harmonics associated with reflection groups other than $\mathbb{Z}_2^{d+1}$. We discuss some of them below.

Question 4.1 (*Critical index*). If none of the $\kappa_i = 0$ in the setup of Theorem 4.6, then we believe that the critical index should be $\delta_0 := |\kappa|_1 - \min_i\{\kappa_i\}$, where we take $\kappa_{d+1} = \mu$. Indeed, it can be shown that (C,δ) means fail to converge in $L^1(W_{\kappa,\mu}, T^d)$ if $\delta \leqslant \delta_0$. To prove that (C,δ) means converge above this critical index, however, will require a method different from what we used to prove Theorem 4.6. For convergence below the critical index, we expect results like those for the classical spherical harmonics. We have, for example, for the Lebesgue measure on B^d the following result.

Theorem 4.7. *Let p satisfy* $|\frac{1}{2} - 1/p| \geqslant 1/(d+2)$, $d \geqslant 2$. *Then the Cesàro* (C,δ) *means of the Fourier orthogonal series with respect to the Lebesgue measure converge in* $L^p(B^d)$ *provided* $\delta > \max\{(d+1)|1/p - \frac{1}{2}| - \frac{1}{2}, 0\}$.

There is a further relation between orthogonal polynomials on spheres and on balls, which relates orthogonal polynomials on S^{d+m} to those on B^d with respect to $W^B_{(m-1)/2}$. This relation allows us to derive convergence results for $W^B_{(m-1)/2}$ from those for the spherical harmonics, leading to the above theorem. Some of the difficulties in these questions may come from the fact that we are dealing with reflection group symmetry instead of rotational symmetry, so that many techniques developed in the context of classical harmonic analysis have to be modified.

Question 4.2 (*Convergence inside the region*). The proof of Theorem 4.6 can also be modified to prove that the partial sum $S_n(f, W_{\kappa,\mu})$ converges in norm if f is in $C^{[r]}$, where $r = |\kappa|_1 + \mu + (d-1)/2$, and the modulus of continuity of $f^{[r]}$ satisfies $\omega(f^{[r]}, \varepsilon) = o(\varepsilon^{r-[r]})$, and these conditions are sharp. However, if we are interested in pointwise convergence, then we can weaken the conditions substantially. Indeed, using [44, Corollary 5.3] and an estimate of the Christoffel function, we can show that $S_n(f, W_{\kappa,\mu})$ converge in the interior of B^d or T^d when a similar condition holds with $r = d/2$. Moreover, we conjecture that the sharp condition is $r = (d-1)/2$, which is the same as the convergence of ordinary harmonics. We also conjecture that for continuous functions the (C, δ) means $S_n^\delta(f, W_{\kappa,\mu})$ will converge inside B^d and T^d if $\delta > (d-1)/2$. Naturally, we expect that the maximum of $S_n(f, W_{\kappa,\mu})$ or $S_n^\delta(f, W_{\kappa,\mu})$ is attained on the boundary of B^d or T^d, but this has yet to be proved.

Question 4.3 (*Estimate and asymptotics of the kernel*). The reproducing kernels (4.5) and (4.6) deserve a careful study. In the one-variable case, a useful form of the kernel is given in terms of the Christoffel–Darboux formula, which is no longer available in several variables. A detailed estimate of the kernel will be crucial in the study of various convergence questions. More difficult is to find the asymptotics of the kernel. For the case $\boldsymbol{x} = \boldsymbol{y}$, the asymptotics of $P_n(W; \boldsymbol{x}, \boldsymbol{x})$ have been studied for $W = W_\mu^B$ on B^d in [45], and for $W = W_{\kappa,\mu}^T$ in [53]. However, the study is incomplete in the case of $W_{\kappa,\mu}^T$.

5. Cubature formulae on the three regions

In this section we discuss the connection between cubature formulae on B^d, T^d and S^d. For a given integral $\mathscr{L}(f) := \int f W \,\mathrm{d}\boldsymbol{x}$, where W is a weight function on T^d or B^d, a cubature formula of degree M is a linear functional

$$\mathscr{I}_M(f) = \sum_{k=1}^{N} \lambda_k f(\boldsymbol{x}_k), \qquad \lambda_k \in \mathbb{R},\ \boldsymbol{x}_k \in \mathbb{R}^d,$$

defined on Π^d, such that $\mathscr{L}(f) = \mathscr{I}_M(f)$ whenever $f \in \Pi_M^d$, and $\mathscr{L}(f^*) \neq \mathscr{I}_M(f^*)$ for at least one $f^* \in \Pi_{M+1}^d$. When the weight function is supported on S^d, we need to replace Π_M^d by Π_M^{d+1} in the above formulation and require $\boldsymbol{x}_k \in S^d$. The points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ in the formula are called *nodes* and the numbers $\lambda_1, \ldots, \lambda_N$ are called *weights*. If all weights of a cubature formula are positive, we call the formula positive.

Each of the three regions B^d, T^d and S^d have drawn their share of attention over the years; a great number of cubature formulae, mostly of lower degrees, have been constructed on them; see, for example, [5,14,36,41] and reference therein. However, the fact that cubature formulae on these three regions are related, in fact are often equivalent, is revealed only recently [49,50]. We state first the correspondence between cubature formulae on balls and on spheres. For a weight function H defined on $\mathbb{R}^{d+1}$ we define W_H^B on B^d as in (3.4).

Theorem 5.1. *If there is a cubature formula of degree M on S^d given by*

$$\int_{S^d} f(\boldsymbol{y})H(\boldsymbol{y})\,\mathrm{d}\omega = \sum_{i=1}^{N} \lambda_i f(\boldsymbol{y}_i), \tag{5.1}$$

whose nodes are all located on S^d, then there is a cubature formula of degree M on B^d for W_H^B,

$$\int_{B^d} g(\boldsymbol{x})W_H^B(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \sum_{i=1}^{N} \lambda_i g(\boldsymbol{x}_i), \tag{5.2}$$

where $\boldsymbol{x}_i \in B^d$ are the first d components of $\boldsymbol{y}_i$. On the other hand, if there is a cubature formula of degree M in the form (5.2) *whose N nodes lie on B^d, then there is a cubature formula of degree M on S^d given by*

$$\int_{S^d} f(\boldsymbol{y})H(\boldsymbol{y})\,\mathrm{d}\omega = \sum_{i=1}^{N} \lambda_i \left[f\left(\boldsymbol{x}_i, \sqrt{1-|\boldsymbol{x}_i|^2}\right) + f\left(\boldsymbol{x}_i, -\sqrt{1-|\boldsymbol{x}|^2}\right)\right]/2. \tag{5.3}$$

To state the correspondence between cubature formulae on B^d and on T^d, we need the notion of invariant cubature formulae. A linear functional $\mathscr{I}(f)$ is called invariant under a group G if $\mathscr{I}(R(a)f) = \mathscr{I}(f)$ for all $a \in G$. For $\boldsymbol{u} \in \mathbb{R}^d$, we denote its *G-orbit* by $G(\boldsymbol{u})$, which is defined by $G(\boldsymbol{u}) = \{\boldsymbol{u}a | a \in G\}$; we also denote by $|G(\boldsymbol{u})|$ the number of distinct elements in $G(\boldsymbol{u})$. A cubature formula is invariant under G if the set of its nodes is a union of G-orbits and the nodes belonging to the same G-orbit have the same weight. In the case of $G = \mathbb{Z}_2^d$, the invariant cubature formula, denoted by $I_M(f, \mathbb{Z}_2^d)$, takes the form

$$\mathscr{I}_M(f; \mathbb{Z}_2^d) = \sum_{i=1}^{N} \lambda_i \sum_{\varepsilon \in \{-1,1\}^d} f(\varepsilon_1 u_{i,1}, \ldots \varepsilon_d u_{i,d})/2^{k(\boldsymbol{u}_i)},$$

where $k(\boldsymbol{u}) = |\mathbb{Z}_2^d(\boldsymbol{u})|$, which is equal to the number of nonzero elements of $\boldsymbol{u}$. We use the notation of W^B and W^T in Section 3.2.

Theorem 5.2. *If there is a cubature formula of degree M on T^d given by*

$$\int_{\Sigma^d} f(\boldsymbol{u})W^T(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} = \sum_{i=1}^{N} \lambda_i f(\boldsymbol{u}_i), \tag{5.4}$$

with all $\boldsymbol{u}_i \in \mathbb{R}_+^d$, then there is a cubature formula of degree $2M+1$ on the unit ball B^d given by

$$\int_{B^d} g(\boldsymbol{x})W^B(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \sum_{i=1}^{N} \lambda_i \sum_{\varepsilon \in \{-1,1\}^d} f(\varepsilon_1\sqrt{u_{i,1}}, \ldots, \varepsilon_d\sqrt{u_{i,d}})/2^{k(\boldsymbol{u}_i)}. \tag{5.5}$$

Moreover, a cubature formula of degree $2M+1$ in the form of (5.5) *implies a cubature formula of degree M in the form of* (5.4).

We note that the degree of formula (5.5) is $2M+1$, more than twice that of formula (5.4). From these two theorems, we can also write down a correspondence between cubature formulae on T^d and on S^d, which we shall not formulate here; see [50].

The importance of these correspondences is evident. They allow us to construct cubature formulae on one region from those on another region. Even from the existing list of cubature formulae, they lead to many new formulae. However, it should be pointed out that most of the cubature formulae in the literature are constructed for the Lebesgue measure (unit weight function). The correspondences on these regions show that the Lebesgue measure $\mathrm{d}\boldsymbol{x}$ on T^d corresponds to $|x_1\cdots x_{d+1}|\,\mathrm{d}\omega$ on S^d and $|x_1\cdots x_d|\,\mathrm{d}\boldsymbol{x}$ on B^d, and the Lebesgue measure on B^d corresponds to $|x_{d+1}|\,\mathrm{d}\omega$ on S^d and $\mathrm{d}\boldsymbol{x}/\sqrt{x_1\cdots x_d}$ on T^d. In the discussion below, we will concentrate on the cubature formulae for the surface measure $\mathrm{d}\omega$ on S^d, which corresponds to formulae for the weight function $1/\sqrt{1-|\boldsymbol{x}|^2}$ on B^d and the weight function $1/\sqrt{x_1\cdots x_d(1-|\boldsymbol{x}|_1)}$ on T^d. These two weight functions are special cases of W_μ^B and $W_{\kappa,\mu}^T$; we shall call them Chebyshev weight function on B^d and on T^d, respectively.

One way to use the correspondences is to revisit the existing methods of constructing cubature formulae on B^d or T^d for the unit weight function, use them to construct formulae for the Chebyshev weight function, and then obtain new cubature formulae for the surface measure on S^d. This approach has been used in [18–20]. We present one notable family of cubature formulae on S^d obtained in [20] below, and discuss symmetric formulae on S^2 in the following section. We need to introduce the following notation.

Let f be defined on $\mathbb{R}^{d+1}$. Since the hyper-octahedral group $\mathscr{B}_{d+1}$ is the semiproduct of the symmetric group $\mathscr{S}_{d+1}$ and $\mathbb{Z}_2^{d+1}$, we can write

$$\sum_{\sigma\in\mathscr{B}_{d+1}} f(\boldsymbol{x}\sigma)=\sum f(\pm x_{\tau_0},\pm x_{\tau_1},\ldots,\pm x_{\tau_d}),$$

where we write $\boldsymbol{x}=(x_0,\ldots,x_d)$ and the sum in the right-hand side is over all choices of signs and over all $\tau\in\mathscr{S}_{d+1}$; that is, we write $\tau=(\tau_0,\ldots,\tau_d)$ to denote a permutation of $(0,1,\ldots,d)$. Cubature formulae that consist entirely of sums as these are invariant under $\mathscr{B}_{d+1}$, they are also called fully symmetric, see [41, p. 128]. For $\alpha\in\mathbb{N}_0^{d+1}$, we will write $\alpha=(\alpha_0,\ldots,\alpha_d)$ in the rest of this section. We have the following result.

Theorem 5.3. *Let $s\in\mathbb{N}_0$ and $n=2s+1$. Then the following is a cubature formula of degree $2n+1$ on S^d:*

$$\int_{S^d} g(y)\,\mathrm{d}w=\frac{\pi^{(d+1)/2}}{2^{2s+d}}\left[\sum_{i=0}^{s}(-1)^i\frac{(n+(d-1)/2-2i)^n}{i!\Gamma(n+(d+1)/2-i)}\right.$$
$$\times\sum_{|\beta|_1=s-i,\beta_0\geqslant\cdots\geqslant\beta_d}\binom{\beta_0-1/2}{\beta_0}\cdots\binom{\beta_d-1/2}{\beta_d}$$
$$\left.\times\sum_{\sigma\in\mathscr{B}_{d+1}} g\left(\left(\frac{\sqrt{2\beta_0+1/2}}{\sqrt{n+(d-1)/2-2i}},\ldots,\frac{\sqrt{2\beta_d+1/2}}{\sqrt{n+(d-1)/2-2i}}\right)\sigma\right)\right].$$

This formula is apparently the first known family of cubature formulae that exist for higher degrees and for all d. Its number of nodes is also relatively small. The drawback of the formula, however, is that it is not positive and its condition number grows rapidly as s increases. The formula is discovered using the correspondence in Theorems 5.1 and 5.2 as follows. In [17], a family of cubature formulae was established for the unit weight function on T^d by proving a combinatorial formula. In [20], we establish the analogs of these formulae for the weight function $W^T_{\kappa,\mu}$ on T^d for general κ and μ by proving a combinatorial formula that contains a number of parameters. The combinatorial formula is as follows:

$$2^{2s}\frac{\prod_{j=0}^d \Gamma(\alpha_j+\mu_j+1)}{\prod_{j=0}^d \Gamma(\mu_j+1)} = \sum_{j=0}^{s}(-1)^j\binom{2s+|\mu|_1+d+1}{j}$$

$$\times \sum_{|\beta|_1=s-j}\binom{\beta_0+\mu_0}{\beta_0}\cdots\binom{\beta_d+\mu_d}{\beta_d}\prod_{i=0}^{d}(2\beta_i+\mu_i+1)^{\alpha_i}, \qquad (5.6)$$

which holds for all $\alpha\in\mathbb{N}_0^d$, $|\alpha|=2s+1$, and $\mu_i>-1$ for $0\leqslant i\leqslant d$. For $\mu_0=\cdots=\mu_d=\frac{1}{2}$, this is the formula proved in [17]. Let $X^\alpha=(1-|\mathbf{x}|_1)^{\alpha_0}x_1^{\alpha_1}\cdots x_d^{\alpha_d}$. Using the fact that $\{X^\alpha\}_{|\alpha|_1=n}$ forms a basis for Π_n^d and the fact that the integral of X^α with respect to $W^T_{\kappa,\mu}$ can be derived from

$$\int_{T^d} X^{\alpha+\mu}\,\mathrm{d}\mathbf{x} = \frac{\prod_{j=0}^d \Gamma(\alpha_j+\mu_j+1)}{\Gamma(|\alpha|_1+|\mu|_1+d+1)},$$

it follows that (5.6) yields a cubature formula of degree $2s+1$ for $W^T_{\kappa,\mu}$ on T^d (taking $\kappa_i=\mu_i+\frac{1}{2}$ for $0\leqslant i\leqslant d$ and $\mu=\mu_0+1$). Using the correspondence in Theorems 5.1 and 5.2, we then get cubature formulae for $W^B_{\kappa,\mu}$ on B^d and $h^2_{\kappa,\mu}$ on S^d. Theorem 5.3 is the special case of $\kappa=0$ and $\mu=0$ on S^d.

6. Cubature formulae on S^2 invariant under the octahedral group

Numerical integration on the sphere S^2 has attracted much attention; we refer to [25–28,32,36,39,41] and the references therein. Most of the cubature formulae on S^2 are constructed by solving moment equations under the assumption that the formulae are symmetric under a finite group. The symmetry helps to reduce the number of moment equations that have to be solved, owing to a fundamental result of Sobolev [38] which states that a cubature formula invariant under a finite group is exact for all polynomials in a subspace $\mathscr{P}$, if and only if, it is exact for all polynomials in $\mathscr{P}$ that are invariant under the same group. The groups that have been employed previously in this context are mainly the octahedral group and the icosahedral group. In particular, Lebedev constructed in [25–28] cubature formulae of degree up to 59, many of which have the smallest number of nodes among all formulae that are known; he also made an outstanding conjecture that we shall address in this section.

Under the correspondences in the previous section, formulae on S^2 invariant under the octahedral group (which is the symmetric group of the unit cube $\{\pm1,\pm1,\pm1\}$ in $\mathbb{R}^3$) correspond to formulae on T^2 invariant under the symmetric group of T^2. Since the octahedral group is the semi-direct product of symmetric group and $\mathbb{Z}_2^3$, and only the action of symmetric group appears on T^2 under our correspondence, there is a certain advantage in dealing with symmetric formulae on T^2 instead of octahedral symmetric formulae on S^2. We shall state Lebedev's conjecture in terms of symmetric cubature formulae on T^2. To do so, we follow the setup in [30] and use the equilateral triangle

$$\triangle=\{(x,y)\colon x\leqslant\tfrac{1}{2}, \sqrt{3}y-x\leqslant1, -\sqrt{3}y-x\leqslant1\},$$

which can be transformed to T^2 by a simple affine transformation. The symmetric group $\mathscr{S}_3(\triangle)$ of $\triangle$ is generated by a rotation through an angle $2\pi/3$ and a reflection about the x-axis. It is sometimes convenient to use the polar coordinates, $x = r\cos\theta$ and $y = r\sin\theta$, to denote points on $\triangle$. Let Λ denote the triangle

$$\Lambda = \{(x, y)\colon\ 0 \leqslant x \leqslant \tfrac{1}{2}, 0 \leqslant x \leqslant \sqrt{3}y\}.$$

Then Λ is one of the fundamental regions of $\triangle$ under $\mathscr{S}_3(\triangle)$. To describe a symmetric cubature formula on $\triangle$, it suffices to determine its nodes inside Λ. We say that a symmetric cubature formula is of type $[m_0; m_1, m_2, m_3; m_4, m_5]$, if it has m_0 nodes at the origin, m_1 nodes at the vertex $(\frac{1}{2}, \sqrt{3}/2)$, m_2 nodes at $(\frac{1}{2}, 0)$, m_3 nodes on the two sides $\theta = 0$ and $\pi/3$ (not at the vertices) of Λ, m_4 nodes on the side $x = \frac{1}{2}$ (not at the vertices) of Λ, and m_5 nodes in the interior of Λ. We also call the corresponding formula on S^2 type $[m_0; m_1, m_2, m_3; m_4, m_5]$.

Moment equations for the type $[m_0; m_1, m_2, m_3; m_4, m_5]$ formulae were set up in [30], and used to construct cubature formulae of degree up to 20 in [30,7] for the unit weight function. In [19] the moment equations are solved for the Chebyshev weight function (recall that it relates to the surface measure on S^2), which yields formulae of degree up to 41 for the surface measure on S^2, including those found by Lebedev. To set up the moment equations, one usually requires that the number of parameters matches the number of equations. For a type $[m_0; m_1, m_2, m_3; m_4, m_5]$ formula of degree M, this leads to

$$m_0 + m_1 + m_2 + 2m_3 + 2m_4 + 3m_5 = [(M^2 + 6M + 12)/12],$$

where $[x]$ denote the greatest integer less than or equal to x. For each M, there can be a number of integer solutions to the equation, leading to different types of cubature formulae. However, since the moment equations are nonlinear, many types of formulae do not exist. Based on his computation, Lebedev made the following conjecture.

Conjecture 6.1. *Cubature formulae of type* $[1; 0, 1, 3m; m, m(m-1)]$ *and* $[1; 1, 1, 3m+1; m, m^2]$ *exist.*

Formulae of these types on $\triangle$ are of degree $6m+2$ and $6m+5$, and they correspond to formulae of degree $12m+5$ and $12m+11$ on S^2, respectively. Lebedev has constructed formulae for $m = 1, 2, 3, 4$ on S^2, whose nodes turn out to be rather uniformly distributed on the sphere.

The work of [25–28], as well as that of [19], is essentially numerical computation. Being so, it gives little hint on how to prove the conjecture. The formulae invariant under the octahedral group are also called fully symmetric formulae [41]. The structure of the fully symmetric formulae or that of the associated moment equations have been studied in [21,22,30,31] and the references therein, but the study appears to be still in the initial stage. In particular, the intrinsic structure of the symmetric formulae of the types in Lebedev's conjecture has not been studied. In the following, we discuss some observations and other problems related to this conjecture.

Because of the fundamental result of Sobolev, it is essential to understand the structure of polynomials invariant under the symmetric group. Let us denote by $\Pi_n^2(\mathscr{S}_3)$ the space of polynomials of degree n invariant under the group $\mathscr{S}_3(\triangle)$. It is easy to see that

$$\Pi_n^2(\mathscr{S}_3) = \operatorname{span}\{(x^2 + y^2)^k(x^3 - 3xy^2)^j\colon\ 2k + 3j \leqslant n\}.$$

We can change variables $t = x^2 + y^2$ and $s = x^3 - 3xy^2$ so that the space $\Pi_n^2(\mathscr{S}_3)$ becomes a subspace spanned by monomials $\{t^k s^j\colon 2k + 3j \leqslant n\}$. However, the order of this polynomial subspace is messed up; for example, the multiplication by t or s is no longer a mapping from $\Pi_n^2(\mathscr{S}_3)$ to $\Pi_{n+1}^2(\mathscr{S}_3)$. On the other hand, for special values of n, we have the following observation.

Proposition 6.2. *Let $\Pi_m^* := \Pi_{3m-1}^2(\mathscr{S}_3)$ for $m = 1, 2, \ldots$. Then $s\Pi_m^* \subset \Pi_{m+1}^*$ and $t\Pi_m^* \subset \Pi_{m+1}^*$.*

Using the fact that $\Pi_0^* = \mathrm{span}\{1\}$ and $\Pi_1^* = \mathrm{span}\{1, s\}$, we can also write down the decomposition of Π_m^* in terms of the space of homogeneous polynomials. The observation suggests that the space Π_m^* has a structure similar to the space Π_m^2. What prompts us to consider this space is explained as follows. A formula in Lebedev's conjecture, if it exists, is like a Gaussian quadrature formulae in the sense that its number of parameters matches up with the number of moment equations. One possible way to establish the conjecture is to show that the nodes of the formula are common zeros of a sequence of symmetric orthogonal polynomials. We will not discuss the connection between cubature formulae and common zeros of orthogonal polynomials here, but refer to [33,35,36,43,46] and the references therein. Like the case of Gaussian quadrature, for cubature formulae of degree $2n-1$, we need to look at common zeros of orthogonal polynomials of degree n or higher. For Lebedev's formulae of degree $6m+5$, this suggests to us looking at the polynomial space of $\Pi_{3m+2}^2(\mathscr{S}_3) = \Pi_{m+1}^*$.

The mapping $t \to x^2 + y^2$ and $s \to x^3 - 3xy^2$ is nonsingular on Λ and it maps Λ to a curved triangular region which we denote by Λ^*. Symmetric cubature formulae on $\triangle$ corresponds to formulae on Λ^* for the space Π_n^*. Using a computer algebra system doing symbolic computation, we found that the degree-5 formula on Λ^*, which corresponds to the degree-11 formula ($m = 1$ of the case $6m + 5$) in Lebedev's conjecture, is indeed generated by common zeros of orthogonal polynomials.

To prove the conjecture along these lines, we need compact formulae for an orthonormal basis of Π_m^*, which can be obtained from the symmetric basis of orthogonal polynomials on the triangle, or from spherical harmonics on S^2 by the correspondence. However, a compact formula of an orthonormal basis for symmetric polynomials on T^2 is the same as a basis for the spherical harmonics invariant under the octahedral group, which is not easy to find. Only [8] seems to contain some results in this direction. Moreover, the cubature formulae in Lebedev's conjecture have nodes on the boundary of $\triangle$. Using a procedure that resembles the passage from Gauss–Radau-type quadrature to Gaussian quadrature, we can reduce the problem to finding cubature formulae with all nodes inside $\triangle$ for integrals with respect to a modified weight function. The new weight function is obtained by multiplying the Chebyshev weight with a polynomial that is quadratic on each side of the boundary of $\triangle$. The corresponding weight function on the sphere S^2 is $h(\mathbf{x}) = (x_1^2 - x_2^2)^2(x_2^2 - x_3^2)^2(x_3^2 - x_1^2)^2$, which is a special case of h_κ in (2.5). Thus, this calls for a study of the h-harmonics associated with $h^2\,\mathrm{d}\omega$ that are invariant under the octahedral group. At the time of this writing, no compact formulae of an orthonormal basis for these h-harmonics are known.

References

[1] P. Appell, J.K. de Fériet, Fonctions hypergéométriques et hypersphériques, Polynomes d'Hermite, Gauthier-Villars, Paris, 1926.

[2] R. Askey, Orthogonal polynomials and Special Functions, SIAM, Philadelphia, PA, 1975.

[3] H. Berens, Y. Xu, Fejér means for multivariate Fourier series, Math. Z. 221 (1996) 449–465.

[4] H. Berens, Y. Xu, $\ell - 1$ summability for multivariate Fourier integrals and positivity, Math. Proc. Cambridge Philos. Soc. 122 (1997) 149–172.

[5] R. Cools, P. Rabinowitz, Monomial cubature rules since "Stroud": a complication, J. Comput. Appl. Math. 48 (1992) 309–326.

[6] M. Dubiner, Spectral methods on triangles and other domains, J. Sci. Comput. 6 (4) (1991) 345–390.

[7] D.A. Dunavant, High degree efficient symmetrical Gaussian quadrature rules for the triangle, Internat. J. Numer. Methods Eng. 21 (1985) 1129–1148.

[8] C.F. Dunkl, Orthogonal polynomials with symmetry of order three, Canad. J. Math. 36 (1984) 685–717.

[9] C.F. Dunkl, Orthogonal polynomials on the sphere with octahedral symmetry, Trans. Amer. Math. Soc. 282 (1984) 555–575.

[10] C.F. Dunkl, Reflection groups and orthogonal polynomials on the sphere, Math. Z. 197 (1988) 33–60.

[11] C.F. Dunkl, Differential-difference operators associated to reflection groups, Trans. Amer. Math. Soc. 311 (1989) 167–183.

[12] C.F. Dunkl, Integral kernels with reflection group invariance, Canad. J. Math. 43 (1991) 1213–1227.

[13] C.F. Dunkl, Intertwining operators associated to the group S_3, Trans. Amer. Math. Soc. 347 (1995) 3347–3374.

[14] H. Engles, Numerical Quadrature and Cubature, Academic Press, New York, 1980.

[15] A. Erdélyi, W. Magnus, F. Oberhettinger, F.G. Tricomi, Higher Transcendental Functions, Vol. 2, McGraw-Hill, New York, 1953.

[16] G. Gasper, Positive sums of the classical orthogonal polynomials, SIAM J. Math. Anal. 8 (1977) 423–447.

[17] A. Grundmann, H.M. Möller, Invariant integration formulas for the n-simplex by combinatorial methods, SIAM J. Numer. Anal. 15 (1978) 282–290.

[18] S. Heo, Y. Xu, Constructing cubature formulae for spheres and balls, J. Comput. Appl. Math. 112 (1999) 95–119.

[19] S. Heo, Y. Xu, Constructing fully symmetric cubature formulae for the sphere, Math. Comp., accepted for publication.

[20] S. Heo, Y. Xu, Invariant cubature formulae for simplex, ball, and the surface of the sphere by combinatorial methods, SIAM J. Numer. Anal., to be published.

[21] P. Keast, Cubature formulas for the surface of the sphere, J. Comput. Appl. Math. 17 (1987) 151–172.

[22] P. Keast, J.C. Diaz, Fully symmetric integration formulas for the surface of the sphere in s dimensions, SIAM J. Numer. Anal. 20 (1983) 406–419.

[23] T. Koornwinder, Two-variable analogues of the classical orthogonal polynomials, in: R.A. Askey (Ed.), Theory and Application of Special Functions, Academic Press, New York, 1975, pp. 435–495.

[24] H.L. Krall, I.M. Sheffer, Orthogonal polynomials in two variables, Ann. Mat. Pura. Appl. 76 (4) (1967) 325–376.

[25] V.I. Lebedev, Quadrature on a sphere, USSR Comput. Math. Math. Phys. 16 (1976) 10–24.

[26] V.I. Lebedev, Spherical quadrature formulas exact to orders 25–29, Siberian Math. J. 18 (1977) 99–107.

[27] V.I. Lebedev, A quadrature formula for the sphere of 59th algebraic order of accuracy, Russian Acad. Sci. Dokl. Math. 50 (1995) 283–286.

[28] V.I. Lebedev, L. Skorokhodov, Quadrature formulas of orders 41,47 and 53 for the sphere, Russian Acad. Sci. Dokl. Math. 45 (1992) 587–592.

[29] Zh.-K. Li, Y. Xu, Summability of product Jacobi expansions, J. Approx. Theory 104 (2000) 287–301.

[30] J.N. Lyness, D. Jespersen, Moderate degree symmetric quadrature rules for the triangle, J. Inst. Math. Anal. Appl. 15 (1975) 19–32.

[31] J.I. Maeztu, E. Sainz de la Maza, Consistent structures of invariant quadrature rules for the n-simplex, Math. Comp. 64 (1995) 1171–1192.

[32] A.D. McLaren, Optimal numerical integration on a sphere, Math. Comp. 17 (1963) 361–383.

[33] H.M. Möller, Kubaturformeln mit minimaler Knotenzahl, Numer. Math. 35 (1976) 185–200.

[34] C. Müller, Analysis of Spherical Symmetries in Euclidean Spaces, Springer, New York, 1997.

[35] I.P. Mysovskikh, The approximation of multiple integrals by using interpolatory cubature formulae, in: R.A. DeVore, K. Scherer (Eds.), Quantitative Approximation, Academic Press, New York, 1980.

[36] I.P. Mysovskikh, Interpolatory Cubature Formulas. 'Nauka', Moscow, 1981 (in Russian).

[37] M. Rösler, Positivity of Dunkl's intertwining operator, Duke Math. J. 98 (1999) 445–463.

[38] S.L. Sobolev, Cubature formulas on the sphere invariant under finite groups of rotations, Soviet. Math. Dokl. 3 (1962) 1307–1310.

[39] S.L. Sobolev, V.L. Vaskevich, The Theory of Cubature Formulas, Kluwer Academic Publishers, Dordrecht, 1997.

[40] E.M. Stein, G. Weiss, Introduction to Fourier Analysis on Euclidean Spaces, Princeton University Press, Princeton, NJ, 1971.

[41] A. Stroud, Approximate Calculation of Multiple Integrals, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[42] N.J. Vilenkin, Special Functions and the Theory of Group Representations, Translation of Mathematical Monographs, Vol. 22, American Mathematical Society, Providence, RI, 1968.

[43] Y. Xu, Common Zeros of Polynomials in Several Variables and Higher Dimensional Quadrature, Pitman Research Notes in Mathematics Series, Vol. 312, Longman, Essex, 1994.

[44] Y. Xu, Christoffel functions and Fourier series for multivariate orthogonal polynomials, J. Approx. Theory 82 (1995) 205–239.

[45] Y. Xu, Asymptotics for orthogonal polynomials and Christoffel functions on a ball, Methods Appl. Anal. 3 (1996) 257–272.

[46] Y. Xu, On Orthogonal polynomials in several variables, in: Special Functions, q-Series and Related Topics, The Fields Institute for Research in Mathematical Sciences, Communications Series, Vol. 14, 1997, pp. 247–270.

[47] Y. Xu, Orthogonal polynomials for a family of product weight functions on the spheres, Canad. J. Math. 49 (1997) 175–192.

[48] Y. Xu, Integration of the intertwining operator for h-harmonic polynomials associated to reflection groups, Proc. Amer. Math. Soc. 125 (1997) 2963–2973.

[49] Y. Xu, Orthogonal polynomials on spheres and on balls, SIAM J. Math. Anal. 29 (1998) 779–793.

[50] Y. Xu, Orthogonal polynomials and cubature formulae on spheres and on simplices, Methods Appl. Anal. 5 (1998) 169–184.

[51] Y. Xu, Summability of Fourier orthogonal series for Jacobi weight functions on the simplex in $\mathbb{R}^d$, Proc. Amer. Math. Soc. 126 (1998) 3027–3036.

[52] Y. Xu, Summability of Fourier orthogonal series for Jacobi weight on a ball in $\mathbb{R}^d$, Trans. Amer. Math. Soc. 351 (1999) 2439–2458.

[53] Y. Xu, Aymptotics of the Christoffel functions on a simplex in $\mathbb{R}^d$, J. Approx. Theory 99 (1999) 122–133.

[54] Y. Xu, Constructing cubature formulae by the method of reproducing kernel, Numer. Math. 85 (2000) 155–173.

[55] Y. Xu, Orthogonal polynomials and summability on spheres and on balls, Math. Proc. Cambridge Philos. Soc., to be published.

[56] Y. Xu, Generalized classical orthogonal polynomials on the ball and on the simplex, submitted for publication.

ELSEVIER

Journal of Computational and Applied Mathematics 127 (2001) 369

JOURNAL OF
COMPUTATIONAL AND
APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Author Index Volume 127 (2001)

PII: S 0 3 7 7 - 0 4 2 7 (0 0) 0 0 6 9 3 - 2